SUB | NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

Werk

Titel: Numerische Mathematik
Verlag: Springer Verlag
Jahr: 1970
Kollektion: Mathematica
Digitalisiert: Niedersächsische Staats- und Universitätsbibliothek Göttingen
Werk Id: PPN362160546_0015
PURL: http://resolver.sub.uni-goettingen.de/purl?PPN362160546_0015

Übergeordnetes Werk

Werk Id: PPN362160546
PURL: http://resolver.sub.uni-goettingen.de/purl?PPN362160546

Terms and Conditions

The Goettingen State and University Library provides access to digitized documents strictly for noncommercial educational, research and private purposes and makes no warranty with regard to their use for other purposes. Some of our collections are protected by copyright. Publication and/or broadcast in any form (including electronic) requires prior written permission from the Goettingen State- and University Library.
Each copy of any part of this document must contain there Terms and Conditions. With the usage of the library's online system to access or download a digitized document you accept the Terms and Conditions.
Reproductions of material on the web site may not be made for or donated to other repositories, nor may be further reproduced without written permission from the Goettingen State- and University Library.

For reproduction requests and permissions, please contact us. If citing materials, please give proper attribution of the source.

Contact

Niedersächsische Staats- und Universitätsbibliothek Göttingen
Georg-August-Universität Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen
Germany
Email: gdz@sub.uni-goettingen.de

Numerische Mathematik

Herausgegeben von
F. L. Bauer, München
A. S. Householder, Oak Ridge
K. Samelson, München
E. Stiefel, Zürich
J. Todd, Pasadena
J. H. Wilkinson, Teddington

Unter Mitwirkung von
R. Bulirsch, Köln
L. Collatz, Hamburg
G. G. Dahlquist, Stockholm
M. Fiedler, Praha
G. E. Forsythe, Stanford
N. Gastinel, Grenoble
A. Ghizzetti, Roma
W. Givens, Argonne
G. H. Golub, Stanford
H. P. Künzi, Zürich
J. Kuntzmann, Grenoble
N. J. Lehmann, Dresden
R. D. Richtmyer, Boulder
H. Rutishauser, Zürich
R. Sauer †, München
J. Schröder, Köln
S. Sobolev, Novosibirsk
H. J. Stetter, Wien
J. Stoer, Würzburg
R. S. Varga, Cleveland
A. van Wijngaarden, Amsterdam

Band 15 · 1970

Springer-Verlag · Berlin · Heidelberg · New York

Alle Rechte, einschließlich das der Übersetzung in fremde Sprachen und das der fotomechanischen Wiedergabe oder einer sonstigen Vervielfältigung, vorbehalten. Jedoch wird gewerblichen Unternehmen für den innerbetrieblichen Gebrauch nach Maßgabe des zwischen dem Börsenverein des Deutschen Buchhandels e.V. und dem Bundesverband der Deutschen Industrie abgeschlossenen Rahmenabkommens die Anfertigung einer fotomechanischen Vervielfältigung gestattet. Wenn für diese Zeitschrift kein Pauschalabkommen mit dem Verlag vereinbart worden ist, ist eine Wertmarke im Betrage von DM 0.30 pro Seite zu verwenden. *Der Verlag läßt diese Beträge den Autorenverbänden zufließen*

Springer-Verlag / Berlin · Heidelberg · New York
Printed in Germany — Copyright © by Springer-Verlag Berlin Heidelberg 1970
Druck der Universitätsdruckerei H. Stürtz AG., Würzburg

Inhalt des 15. Bandes

BAKER, CHR. T. H.: The Error in Polynomial Interpolation 315

BARRAR, R. B., LOEB, H. L.: On the Remez Algorithm for Non-linear Families . 382

CIARLET, P. G., NATTERER, F., VARGA, R. S.: Numerical Methods of High-Order Accuracy for Singular Nonlinear Boundary Value Problems 87

DESCLOUX, J.: On the Numerical Integration of the Heat Equation 371

DÖRING, B.: Das Tschebyscheff-Verfahren in Banach-Räumen 175

DUBRULLE, A.: A Short Note on the Implicit *QL* Algorithm for Symmetric Tridiagonal Matrices 450

EHLICH, H., HAUSSMANN, W.: Konvergenz mehrdimensionaler Interpolation . . 165

FINCK VON FINCKENSTEIN, K. GRAF: Potenzbeschränktheit und Wertebereich einer Matrix 329

GRIGORIEFF, R. D.: Die Konvergenz des Rand- und Eigenwertproblems linearer gewöhnlicher Differenzengleichungen 15

— Über die Koerzitivität gewöhnlicher Differenzenoperatoren und die Konvergenz von Mehrschrittverfahren 196

HAUSSMANN, W., s. EHLICH, H. 165

HERTLING, J.: Approximation of Piecewise Continuous Functions by a Modification of Piecewise Hermite Interpolation 404

JEROME, J. W.: Linear Self-Adjoint Multipoint Boundary Value Problems and Related Approximation Schemes. 433

LAMPRECHT, G.: Über die Abhängigkeit der Fehlerfunktion bei der rationalen L_p-Approximation von den Koeffizienten des Nennerpolynoms bei minimalem Zählerpolynom 392

LAURENT, P. J., PHAM-DINH-TUAN: Global Approximation of a Compact Set by Elements of a Convex Set in a Normed Space 137

LOEB, H. L., s. BARRAR, R. B. 382

LUCAS, TH. R.: Generalization of *L*-Splines. 359

LUTHER, H. A., SIERRA, H. G.: On the Optimal Choice of Fourth-Order Runge-Kutta Formulas 354

LYNESS, J. N., MCHUGH, B. J. J.: On the Remainder Term in the *N*-Dimensional Euler Maclaurin Expansion 333

MANGASARIAN, O. L.: A Convergent Splitting of Matrices 351

MARSAL, D.: Eingabelung der Lösungen von Integralgleichungen 348

MARTENSEN, E.: Zur Restglieddarstellung der Gregoryschen Quadraturformel ungerader Ordnung 229

MCHUGH, B. J. J., s. LYNESS, J. N. 333

MIKLOŠKO, J.: Asymptotic Properties and the Convergence of Numerical Quadratures 234

NATTERER, F., s. CIARLET, P. G., *et al.* 87

NITSCHE, J.: Zur Konvergenz von Näherungsverfahren bezüglich verschiedener Normen 224

Oliveira, G. N. de: Note on an Inverse Characteristic Value Problem 345

Pham-Dinh-Tuan, s. Laurent, P. J. 137

Rakotch, E., Steinberg, J.: Error Estimate for the Numerical Solution of Fredholm's Integral Equation 320

Ruhe, A.: Properties of a Matrix with a Very Ill-conditioned Eigenproblem . . 57

Rund, H.: The Reduction of Certain Boundary Value Problems to Variational Problems by Means of Transversality Conditions 49

Schaefer, H. H.: Eine Abschätzung der nichttrivialen Eigenwerte stochastischer Matrizen 219

Schäfke, F. W.: Zum Anwendungsbereich einiger Iterationsverfahren 109

Schmidt, J. W., Vetters, K.: Ableitungsfreie Verfahren für nichtlineare Optimierungsprobleme 263

Schröder, J.: Proving Inverse-Positivity of Linear Operators by Reduction . . 100

Shah, J. M.: Two-Dimensional Polynomial Splines 1

Sierra, H. G., s. Luther, H. A. 354

Stallmann, F. W.: Numerical Solution of Integral Equations 297

Steinberg, J., s. Rakotch, E. 320

Stewart, G. W.: On Samelson's Iteration for Factoring Polynomials 306

Stoss, H.-J.: Monotonie-Eigenschaften bei Differentialungleichungen über nichtkompaktem Grundbereich 61

Urabe, M.: An Implicit One-Step Method of High-Order Accuracy for the Numerical Integration of Ordinary Differential Equations 151

Van der Sluis, A.: Condition, Equilibration and Pivoting in Linear Algebraic Systems 74

— Upperbounds for Roots of Polynomials 250

Varga, R. S., s. Ciarlet, P. G., *et al.* 87

Vetters, K., s. Schmidt, J. W. 263

Wetterling, W.: Definitheitsbedingungen für relative Extrema bei Optimierungs- und Approximationsaufgaben 122

Zenišek, A.: Interpolation Polynomials on the Triangle 283

Zettl, G.: Ein Verfahren zum Minimieren einer Funktion bei eingeschränktem Variationsbereich der Parameter 415

Indexed in Current Contents

Erratum

Numer. Math. **14**, 317—324 (1970)

CHRISTIANSEN, J., Numerical Solution of Ordinary Simultaneous Differential Equations of the 1st Order Using a Method for Automatic Step Change

Auf S. 319 muß die 26. Zeile richtig lauten:

On p. 319 the 26. line should read:

bh:= *br*:= *bx*:= **true**; *bc*:= *e*<1; *e*:= *abs* (*e*); *e5*:= 5 × *e*;

Numer. Math. 15, 1—14 (1970)

Two-Dimensional Polynomial Splines

JAYANT M. SHAH*

Received April 9, 1969

1. Introduction

This paper presents a method for constructing polynomial splines which can approximate functions over two-dimensional Euclidean domains. If a given function $u \in H_{k+1}$ then we show that one can find polynomial splines $\{\hat{u}_i\}$ such that $\{\hat{u}_i\}$ and their derivatives of order k or less converge uniformly to u and its corresponding derivatives. We show that these splines can be used to obtain a numerical method for solving elliptic boundary value problems. The method of construction can be easily extended to higher dimensional Euclidean spaces. For one dimensional splines, see, for example, references [2] and [3].

We denote by G a bounded open domain of class C^0 (see reference [1] for definition) in the Euclidean plane with boundary ∂G and closure $\overline{G}$. For a function $u \in C^j(\overline{G})$ denote by $\|\|u\|\|_{j,G}$ the maximum norm $\max_{|\alpha| \leq j} \max_{\bar{x} \in \overline{G}} |D^\alpha u(\bar{x})|$ and denote by $\|u\|_{j,G}$ the norm $\left(\int_G \sum_{|\alpha| \leq j} |D^\alpha u(\bar{x})|^2 d\bar{x} \right)^{\frac{1}{2}}$. Here $\alpha = (\alpha_1, \alpha_2)$ is the multiple index of length $|\alpha| = \alpha_1 + \alpha_2$ and D^α is the corresponding partial derivative. We denote by $|u|_{j,G}$ the semi-norm $\left(\int_G \sum_{|\alpha|=j} |D^\alpha u|^2 d\bar{x} \right)^{\frac{1}{2}}$.

2. Two-Dimensional Spline Functions

Fix an integer $m \geqq 1$ and an integer $p \geqq 0$. Subdivide the domain $\overline{G}$ by a grid $\mathfrak{h}$ into a finite number of subdomains such that each subdomain is a polygon with infinitely differentiable curvilinear sides. We call any such subdomain a cell. Let $\{S_i\}$ be the set of grid lines forming the sides of cells. We require that the interior of any grid line S_i must not contain a vertex. Let $\sup_i$(Length of S_i) $= h$. We denote by $\{G_i\}$ the disjoint collection of interiors of the cells and let $G_0 = \bigcup_i G_i$. Assume that each grid line, S_i, is a non-singular curve, defined by an equation of the form

$$\begin{aligned} x &= f_i(\xi_i) \\ y &= g_i(\xi_i) \end{aligned} \quad , \tag{1}$$

where ξ_i is the distance measured from one of the two vertices on S_i along the straight line connecting the two vertices. The unit tangent vector $\boldsymbol{t}_i = (t^1, t^2)$

* Work reported herein was supported in part by Project MAC, an M.I.T. research project sponsored by the Advanced Research Projects Agency, Department of Defense, under Office of Naval Research Contract Nonr-410 (01). Reproduction of this report, in whole or in part, is permitted for any purpose of the United States Government.

at a point on the grid line and the unit normal $\boldsymbol{n}=(-t_2, t_1)$ are defined by the equation:

$$t_1=\frac{df_i/d\xi_i}{\left[\left(\frac{df_i}{d\xi_i}\right)^2+\left(\frac{dg_i}{d\xi_i}\right)^2\right]^{\frac{1}{2}}}$$

$$t_2=\frac{dg_i/d\xi_i}{\left[\left(\frac{df_i}{d\xi_i}\right)^2+\left(\frac{dg_i}{d\xi_i}\right)^2\right]^{\frac{1}{2}}}. \tag{2}$$

If $u^{(i)}\in C^{\infty}(\overline{G}_i)$, let $D^{(j)}_{n,r}u^{(i)}$ denote the r-th derivative of $u^{(i)}$ in the direction of the normal $\boldsymbol{n}_i$, defined if S_j is one of the sides of G_i. For every cell, we define the following quantities:

$$u^{ij}_{kr}=\frac{1}{h^{k+1}}\int_{S_j}\xi_j^k D^{(j)}_{n,r}u^{(i)}\,d\xi_j \tag{3}$$

defined for all k and r whenever $0\leqq r\leqq m-1$ and $0\leqq k\leqq p+r-(m-1)$.

Also if w is a vertex of a cell G_i, define $u^{i,w}_{\alpha}=D_{\alpha}u^{(i)}(w)$ for all α such that $0\leqq|\alpha|\leqq m-2$.

We define a class $\mathfrak{F}_{m,p}(\mathfrak{h})$ of functions with domain G as follows: Let (x_i, y_i) denote the local coordinates of a cell G_i, obtained by translation of the origin to a point in G_i. If $u\in\mathfrak{F}_{m,p}(\mathfrak{h})$, then $u^{(i)}=u$ restricted to G_i is given by an equation of the form

$$u^{(i)}=a_{00}+a_{10}x_i+a_{01}y_i+\cdots+a_{0,n_i}y_i^{n_i} \tag{4}$$

such that the following compatibility conditions are satisfied:

For all k and r and whenever defined

$$u^{ij}_{kr}=u^{i'j}_{kr}\quad\text{if}\quad s_j=\partial G_i\cap\partial G_{i'}. \tag{5}$$

At a vertex w, for all $|\alpha|\leqq m-2$

$$u^{i,w}_{\alpha}=u^{i',w}_{\alpha}\quad\text{if } w \text{ belongs to both } G_i \text{ and } G_{i'}. \tag{6}$$

We also require that

$$n_i\geqq m+p. \tag{7}$$

We need to impose the following five conditions on the class of admissible grids:

(a) There exists a positive constant δ_1 such that for any admissible grid,

$$\inf_i\frac{\text{length of } S_i}{h}\geqq\delta_1.$$

(b) There exists a positive constant δ_2 such that for every admissible grid, and for every cell, there exists a point (x_0, y_0) in that cell such that the square $\{(x, y)\,|\,|x-x_0|\leqq\delta_2 h,\ |y-y_0|\leqq\delta_2 h\}$ is contained in the cell.

(c) There exists a constant C_0 such that the number of cells that any straight line through G intersects is less than C_0/h.

In order to state the last two conditions, we introduce non-dimensional coordinates $X_i=x_i/h$, $Y_i=y_i/h$, $\Xi_i=\xi_i/h$ and quantities $A_{jk}=a_{jk}h^{j+k}$, $U^{ij}_{kr}=u^{ij}_{kr}h^r$

and $U_\alpha^{i,w} = u_\alpha^{i,w} h^{|\alpha|}$ so that

$$u^{(i)} = A_{00} + A_{10} X_i + \cdots A_{0n_i} Y_i^{n_i}, \tag{8}$$

$$U_{kr}^{ij} = \int_{s_j} \Xi_j^k (h^r D_{n,r}^{(j)})\, u^{(i)}\, d\Xi_j, \qquad U_\alpha^{i,w} = (h^{|\alpha|} D_\alpha)\, u^i(w). \tag{9}$$

For a given cell G_i, let V_i be the vector space spanned by the quantities U_{kr}^{ij} and $U_\alpha^{i,w}$ defined by (9). Let A_i be the vector space spanned by the coefficients A_{jk}. Then we have a linear transformation

$$T_i \colon A_i \to V_i. \tag{10}$$

Now let S_j be a grid line which connects vertices w_1 and w_2 in the cell G_i. Assume that w_1 is the origin and that the x-axis passes through w_2. Consider a subspace $\mathfrak{a}$ of the vector space A_i spanned by $\{A_{00}, A_{10}, A_{01}, \ldots, A_{0,m-1}\}$ and a subspace $\mathfrak{b}_0$ of V_i spanned by $\{U_\alpha^{i,w_1}\}_{|\alpha|\leqq m-2}$, $\{U_\alpha^{i,w_2}\}_{|\alpha|\leqq m-2}$ and $U_{0,m-1}^{ij}$. (If $m=1$, then $\mathfrak{b}_0$ is one-dimensional, spanned by $U_{0,0}^{ij}$.) Let $j \colon \mathfrak{a} \to A_i$ be the inclusion map and $p_1 \colon V_i \to \mathfrak{b}_0$ be the projection map. Then it is easy to check that the map $\tau_i \colon \mathfrak{a} \to \mathfrak{b}_0$ given by the composition $p_1 \circ T_i \circ j$ is an isomorphism.

Now we can state the last two conditions:

(d) Introduce the absolute norm on the spaces $\mathfrak{a}$ and $\mathfrak{b}_0$. We require that the induced norm of τ_i^{-1} is bounded uniformly with respect to all $\mathfrak{h}$.

(e) We assume that n_i is chosen sufficiently large so that the map T_i is onto. Let k_i be the dimension of the kernel of T_i. We require that it is possible to choose a k_i-dimensional subspace K_i of A_i and that it is possible to extend the map T_i to an isomorphism of the normed vector spaces

$$\overline{T}_i \colon A_i \xrightarrow{\sim} V_i \oplus K_i \tag{11}$$

such that the induced norm of $\overline{T}_i^{-1}$ is bounded uniformly with respect to all $\mathfrak{h}$. We also require that n_i is bounded uniformly with respect to all grids.

Condition (e) guarantees that there exist polynomials $u^{(i)}$ such that the conditions (5), (6) and (7) are satisfied.

The vector space $\mathfrak{F}_{m,p}(\mathfrak{h})$ is constructed as follows. For each cell G_i, we have an isomorphism $A_i \xrightarrow{\sim} V_i \oplus K_i$. Then a function $u \in \mathfrak{F}_{m,p}(\mathfrak{h})$ is defined by specifying the quantities

$$u_{kr}^j = u_{kr}^{ij} = u_{kr}^{i'j} \qquad \text{for each grid line } j \text{ between cells } i \text{ and } i',$$

$$u_\alpha^w = u_\alpha^{i,w} = u_\alpha^{i',w} = \cdots \qquad \text{for each vertex } w$$

and an element of $K^{(i)}$ for each cell G_i.

We will mostly omit mentioning the suffix (m, p).

We extend the (semi) norms $|\cdot|, \|\cdot\|, |||\cdot|||$ to $\mathfrak{F}(\mathfrak{h})$ by restricting the integrals to G_0.

Let V be a vector space spanned by the quantities u_{kr}^j, u_α^w with the basis $\{e_i\}$. Let $K = \bigoplus_i K_i$ with basis $\{k_i\}$. Let $W = V \oplus K$. Then we have an isomorphism $I \colon \mathfrak{F}(\mathfrak{h}) \xrightarrow{\sim} W$. The images of e_i and k_j under I^{-1} form a basis for $\mathfrak{F}(\mathfrak{h})$. Denote this basis by $\{u_i\}$.

Let $\overset{\circ}{\mathfrak{F}}(\mathfrak{h})$ be the subspace of $\mathfrak{F}(\mathfrak{h})$ defined by the condition that $u^j_{k,r}=0$ if S_j lies in ∂G and $u^w_\alpha=0$ is w lies in ∂G.

3. Existence

In practice, conditions (d) and (e) are difficult to verify. One can usually manage to satisfy these conditions for a finite number of cells. We now show that under certain conditions, a cell satisfying conditions (d) and (e) continues to do so after certain transformations.

Theorem 1. If a cell satisfies conditions (d) and (e), then its transform under translation, an orthonormal transformation and scale change satisfies these conditions provided that condition (b) is not violated. If, in addition, we assume that the cell has straight sides, then any of its non- singular affine transforms which do not violate conditions (a) and (b) satisfies conditions (d) and (e).

Proof. We give the proof only for the special case of triangular cells with straight sides. Other cases are similar. Since the condition (d) depends only on the shape of the grid line (defined by Eq. (1)) modulo translation, rotation and scale change, we have to verify only condition (e).

Consider a cell G_i in the form of an isosceles right-angled triangle with hypothenuse h units long and with origin of the local coordinate system at one of the vertices. Since the transformation T_i for this particular cell does not depend upon h, we can find an integer n for n_i in Eq. (4) such that condition (e) is satisfied for all values of h. Now consider any other triangular cell $G_{i'}$ which satisfies conditions (a) and (b). Set $n_{i'}=n$. Without loss of generality we can assume that $G_{i'}$ has the origin of the local coordinate system at one of its vertices. There exists a non-singular linear transformation S mapping G_i onto $G_{i'}$. Now specification of the quantities $u^{i,w}_\alpha$ and $u^{ij}_{k\,r}$ (if defined) implies specification of all quantities of the form

$$\frac{1}{h^{k+1}}\int_{S_j}\xi_j^k D^{(j)}_{n,r} D^{(j)}_{i,r'} u^{(i)}\, d\xi_i$$

$$0\leqq r+r'\leqq m-1,\quad r'>0,\quad 0\leqq k\leqq p+(r+r')-(m-1)$$

where $D^{(j)}_{i,r'}u^{(i)}$ denotes the r'-th derivative of $u^{(i)}$ in the direction of the tangent, defined if S_j is one of the sides of G_i.

Therefore S induces an isomorphism

$$S_2\colon V_i \tilde{\to} V_{i'}. \tag{12}$$

Also S induces an isomorphism

$$S_1\colon A_i \tilde{\to} A_{i'}. \tag{13}$$

Let K_i and $K_{i'}$ be the kernels of the maps T_i and $T_{i'}$ respectively. Therefore we have a commutative diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K_i & \longrightarrow & A_i & \xrightarrow{T_i} & V_i & \longrightarrow & 0 \\ & & \downarrow & & \downarrow{\scriptstyle S_1} & & \downarrow{\scriptstyle S_2} & & \\ 0 & \longrightarrow & K_{i'} & \longrightarrow & A_{i'} & \xrightarrow[T_{i'}]{} & V_{i'} & \longrightarrow & 0 \end{array} \tag{14}$$

where the vertical arrows are isomorphisms because S is invertible. Thus a choice of basis in K_i induces a choice of basis in $K_{i'}$.

Therefore we have a commutative diagram

$$\begin{array}{ccc} A_i & \xrightarrow[\overline{T}_i]{} & V_i \oplus K_i \\ \downarrow{\scriptstyle S_1} & & \downarrow{\scriptstyle S_2} \\ A_{i'} & \xrightarrow[\overline{T}_{i'}]{} & V_{i'} \oplus K_{i'} \end{array} \tag{15}$$

where $\overline{T}_{i'}$ is nonsingular. Now we claim that condition (e) is satisfied. For, if not, there exists a sequence of triangular cells $\{G_k\}$ such that the corresponding sequence of the lowest absolute eigenvalues of the transformations $\overline{T}_k$ tends to zero. But any triangular cell is specified by six quantities, namely the nondimensional coordinates of its vertices. Therefore the sequence of triangles can be embedded as a closed subspace lying in the closed annulus in $\mathbb{R}^6$, lying between the unit sphere and the sphere of radius δ_1. Since the annulus is compact, the sequence G_k has an accumulation point. By continuity, the triangular cell G_{k0} corresponding to the accumulation point satisfies conditions (a) and (b) and has singular $\overline{T}_{k0}$ which contradicts (15). Q.E.D.

Let G_i and $G_{i'}$ be two cells which are superimposed such that their vertices coincide. Let S_j and $S_{j'}$ be sides of G_i and $G_{i'}$ respectively, spanning the same vertices w_1 and w_2.

Let

$$\gamma_j = \sup_P \{z \mid z = |(t_1 - t_1')_P| \text{ or } z = |(t_2 - t_2')_P|\}$$

where P denotes a point between w_1 and w_2 on the ξ-axis, $(t_1, t_2)_P$ is the unit tangent vector to S_j at the point on S_j corresponding to P, and $(t_1', t_2')_P$ is the corresponding tangent vector to $S_{j'}$. Then we say that G_i and $G_{i'}$ approximate each other within η if $\gamma_j \leqq \eta$ for all pairs of corresponding sides of G_i and $G_{i'}$.

Lemma 1. Let G_i be a cell in a grid $\mathfrak{h}$ such that $\|\overline{T}_i^{-1}\| < \lambda$ and $\|\tau_j^{-1}\| < \nu$. If ε is a positive constant less than unity, then there exists a positive constant η which is less than unity and which depends only on n_i, m, p, ε and λ such that the collection of cells which approximate G_i within η satisfies conditions (d) and (e).

Proof. Let $G_{i'}$ be a cell which approximates G_i within some δ. Let $n_{i'} = n_i$ and identify A_i with $A_{i'}$. We have maps

$$\psi = T_{i'} \circ \overline{T}_i^{-1} : V_i \oplus K_i \xrightarrow{\sim} A_i \to V_{i'}$$

and the projection

$$\pi : V_i \oplus K_i \to K_i.$$

Define

$$\theta = \psi \oplus \pi : V_i \oplus K_i \to V_{i'} \oplus K_i.$$

We will prove that θ is an isomorphism, if δ is sufficiently small.

In the following, $\{M_j(k_1, k_2, \ldots)\}$ will denote positive integers which depend only on the integers $k_1, k_2, \ldots$.

Let $b = v \oplus k$, $|b| = 1$.

Let $a = \overline{T}_i^{-1}(b)$; so $|a| < \lambda$.

Let $a = (A_{00}, A_{10}, \ldots, A_{0,n_i})$ and $u = u^{(i)} = u^{(i')} = \sum A_{jk} x^j y^k / h^{j+k}$.

Then $\Delta_r = \sum_{\alpha=r} |||D_\alpha u|||_{0, G_i \cup G_{i'}} < M_1(r, n_i) \lambda h^{-r}$.

We already have $U_\alpha^{i,w} = U_\alpha^{i'w}$ for all vertices w of G_i and all α, $|\alpha| \leqq m-2$.

Consider now a side S_j and its replacement $S_{j'}$.

Observe that

$$\beta_j = \sup_\xi \{z \mid z = |f(\xi) - f'(\xi)| \text{ or } z = |g(\xi) - g'(\xi)|, \ (f(\xi), g(\xi)) \in S_j\}$$

$$\leqq 3\delta h,$$

$$D'_{n,r} - D_{n,r} = (-t'_2 D_x + t'_1 D_y)^r - (-t_2 D_x + t_1 D_y)^r$$

$$= \sum_{k=0}^{r} \binom{r}{k} h_{k,r} D_{\alpha_k}$$

where $\alpha_k = (k, r-k)$ and $|h_{k,r}| \leqq M_2(k, r)\, \delta$.

Therefore

$$|D'_{n,r} u - D_{n,r} u)\,(f(\xi), g(\xi))| \leqq M_3(r)\, \Delta_r \delta$$

and

$$|D'_{n,r} u\,(f'(\xi), g'(\xi)) - D_{n,r} u\,(f(\xi), g(\xi))| < \Delta_{r+1} \beta_j + M_3(r)\, \Delta_r \delta$$

$$\leqq \lambda h^{-r} M_4(r, n_i)\, \delta.$$

Therefore

$$|U_{k,r}^{i',j} - U_{k,r}^{i,j}| \leqq M_5(k, r, n_i)\, \lambda \delta.$$

Let

$$M_6 = \operatorname*{Max}_{k,r} \{M_5(k, r, n_i) \mid U_{k,r}^{i,j} \in V_i\}$$

and put

$$\eta = \frac{1-\varepsilon}{\lambda M_6}.$$

If $\delta \leqq \eta$ then $|U_{k,r}^{i'j} - U_{k,r}^{i,j}| \leqq 1 - \varepsilon$ for all $U_{k,r}^{i,j} \in V_i$ which implies that the kernel of the map θ is zero. Therefore θ is an isomorphism and $\|\theta\| < 2$ and $\|\theta^{-1}\| \leqq 1/\varepsilon$

$$\|\overline{T}_{i'}^{-1}\| = \|\overline{T}_i^{-1} \theta^{-1}\| \leqq \|\overline{T}_i^{-1}\| \, \|\theta^{-1}\| < \frac{\lambda}{\varepsilon}.$$

In the same fashion, it is easy to show that

$$\|\tau_j^{-1}\| < \frac{\nu}{\varepsilon}. \quad \text{Q.E.D.}$$

As a corollary we get the following theorem.

Theorem 2. Let $\mathfrak{C}$ be the collection of cells belonging to some grid $\mathfrak{h}$ which satisfies conditions (a), (d) and (e). Let ε be a positive constant less than unity. Then corresponding to each cell G_i, there exist constants η_i such that if the collection $\mathfrak{C}$ is enlarged by adding cells each of which approximates some cell $G_i \in \mathfrak{C}$ within η_i, then the new collection satisfies conditions (a), (d), (e).

4. Some Properties

Lemma 2. Given $u \in C^{m+p+1}(G)$ there exists $\hat{u} \in \mathfrak{F}(\mathfrak{h})$ such that

$$|||u - \hat{u}|||_{k,G} < C\, h^{m+p+1-k} \|u\|_{m+p+1,G}$$

where C does not depend on $\mathfrak{h}$ and u and where $k \leqq m+p$.

Proof. In the following, $C_1, C_2, \ldots$ represent positive constants which do not depend on $\mathfrak{h}$ and u.

Using Taylor's theorem within each cell, write u as

$$u = {}'u(x_i, y_i) + R(x_i, y_i) \quad \text{where} \quad {}'u(x_i, y_i)$$

is a polynomial of degree $m+p$ and

$$|||D_\alpha R(x_i, y_i)|||_{0,G_i} < C_1 h^{m+p+1-|\alpha|} \|u\|_{m+p+1}, \qquad |\alpha| \leqq m+p.$$

Let

$${}''U_{st}^{ij} = U_{st}^{ij} - {}'U_{st}^{ij} \quad \text{and} \quad {}''U_{\alpha}^{i,w} = U_{\alpha}^{i,w} - {}'U_{\alpha}^{i,w}.$$

Then

$$|{}''U_{st}^{ij}| < C_2 h^{m+p+1} \|u\|_{m+p+1},$$

$$|{}''U_{\alpha}^{i,w}| < C_2 h^{m+p+1} \|u\|_{m+p+1}.$$

Let

$$\{{}''U_{st}^{ij}, U_{\alpha}^{i,w}\}_{\text{all } j,w,s,t,\alpha} = d \in V_i.$$

Let

$$a = \overline{T}_i^{-1}(d \oplus 0) = (A_{00}, \ldots, A_{0,n_i}).$$

Let

$${}''u(x_i, y_i) = a_{00} + a_{10} x_i + \cdots + a_{0n_i} y^{n_i}$$

where

$$a_{ij} = A_{ij}/h^{i+j}.$$

Then

$$|||D_\alpha {}''u|||_{0,G_i} < C_3 h^{m+p+1-|\alpha|} \|u\|_{m+p+1} \quad \text{for} \quad |\alpha| \leqq m+p.$$

Let

$$\hat{u} = {}'u + {}''u.$$

Then

$$\hat{u} \in \mathfrak{F}(\mathfrak{h}) \quad \text{and} \quad |||u - \hat{u}|||_{k,G_0} < C\, h^{m+p+1-k} \|u\|_{m+p+1,G}. \quad \text{Q.E.D.}$$

As a corollary, we get

Theorem 3. Given $u \in H_{m+p+1}(G)$, there exists $\hat{u} \in \mathfrak{F}(\mathfrak{h})$ such that

$$|||u - \hat{u}|||_{k,G} < C\, h^{m+p+1-k} \|u\|_{m+p+1,G}$$

where C does not depend on $\mathfrak{h}$ and u and $k \leqq m+p$.

Lemma 3. Let G_i be a cell in $\mathfrak{h}$ and let S_j be one of its sides. Let w_1, w_2 be the vertices on S_j. Let $u^{(i)}$ be a polynomial as in Eq. (4) such that

$$\begin{aligned} &u_\alpha^{i,w_1} = 0 = u_\alpha^{i,w_2} \quad \text{for} \quad |\alpha| \leqq m-2 \\ &u_{0,m-1}^{ij} = 0. \end{aligned} \tag{16}$$

Then there exists a constant $C>0$ independent of the choice of $\mathfrak{h}$ such that

$$\|\!|D_\alpha u^{(i)}|\!\|_{0,G_i} < C^{m-1-|\alpha|}\,|u^{(i)}|_{m,G_i}.$$

Proof. We can assume that x-axis coincides with ξ-axis on S_j and the origin is at w_1.

Let $\mathfrak{b}_0$ be the subspace of V_i spanned by

$$\{U_\alpha^{i,w_1}\}_{|\alpha|\leqq m-2}, \quad \{U_\alpha^{i,w_2}\}_{|\alpha|\leqq m-2} \quad \text{and} \quad U_{0,m-1}^{ij}.$$

Let

$$\mathfrak{b}_1 \text{ be the orthogonal complement of } \mathfrak{b}_0 \text{ in } V_i.$$

Let

$$\mathfrak{a}_0 = \overline{T}_i^{-1}\mathfrak{b}_0 \quad \text{and} \quad \mathfrak{a}_1 = \overline{T}_i^{-1}\mathfrak{b}_1.$$

Then

$$A_i = \mathfrak{a}_0 \oplus \mathfrak{a}_1.$$

Let

$$u^{(i)} = \sum a_{jk}\, x^j y^k.$$

Then $u^{(i)}$ satisfies (16) $\Leftrightarrow \{a_{jk}h^{j+k}\} \in \mathfrak{a}_1$.

Let $|\cdot|$ denote the absolute norm on $\mathfrak{a}_1$.

Observe that

$$h^{m-1}|u|_{m,G_i} = \left[\int \sum_{\alpha_1+\alpha_2=m} \left(\frac{\partial^m u^{(i)}}{\partial X_i^{\alpha_1}\,\partial Y_i^{\alpha_2}}\right)^2 dX_i\, dY_i\right]^{\frac{1}{2}}$$

is a positive-definite quadratic form $[,]$ on $\mathfrak{a}_1$. Let

$$0 \neq \lambda_i = \inf_{|z|=1} [z, z]^{\frac{1}{2}}.$$

λ_i depends on i and $\mathfrak{h}$. Suppose $\inf\limits_{i,\mathfrak{h}} \lambda_i = 0$.

Then there exists a sequence of cells $\{G_k\}$ such that the corresponding sequence $\lambda_k \to 0$. Now each G_k contains a point (x_k, y_k) and the square

$$\sigma_k = \{(x, y) \,|\, |x - x_k| \leqq h\,\delta_2,\ |y - y_k| \leqq h\,\delta_2\}.$$

Let $[,]'$ denote the bilinear form $[,]$ restricted to the square σ_k. Then there exists a sequence of points (X_k, Y_k), $|X_k| \leqq 1-\delta_2$, $|Y_k| \leqq 1-\delta_2$ and a corresponding sequence $\{z_k\}$ such that $\|z_k\| = 1$ and $[z_k, z_k]' \to 0$. By compactness, there exists a subsequence $(X_{k_j}, Y_{k_j}, z_{k_j})$ converging to $(X_{k_0}, Y_{k_0}, z_{k_0})$ in $\mathbb{R}^2 \oplus \mathfrak{a}_1 \oplus K_i$ such that $\|z_{k_0}\| = 1$ and $[z_{k_0}, z_{k_0}]' = 0$ which is a contradiction of the positive-definiteness of $[,]'$.

Therefore λ_i is bounded from below, and there exists a constant $C>0$ independent of i and $\mathfrak{h}$ such that if $a \in \mathfrak{a}_1$,

$$|a| < C\,h^{m-1}\,|u^{(i)}|_{m,G_i}$$

where $u^{(i)}$ is the polynomial corresponding to a.

Therefore

$$|a_{jk}| < C h^{m-1-j-k} |u^{(i)}|_{m,G_i}$$

and

$$\|\|D_\alpha u^{(i)}\|\|_{0,G_i} < C h^{m-1-|\alpha|} |u^{(i)}|_{m,G_i}. \quad \text{Q.E.D.}$$

Theorem 4. Let $u \in \mathfrak{F}(\mathfrak{h})$. Let S_j be a grid line contained in $\partial G_i \cap \partial G_{i'}$ and let $u^i = u$ restricted to G_i and $u^{i'} = u$ restricted to $G_{i'}$. Then there exists a constant $C > 0$, independent of $\mathfrak{h}$ and u such that

$$\sup_{(x,y)\in S_j} |D_\alpha u^i(x, y) - D_\alpha u^{i'}(x, y)| \leqq C h^{m-1-|\alpha|} |u|_{m,G_i \cup G_{i'}}.$$

Proof. Let x-axis coincide with the ξ-axis and let the origin be at one of the vertices on S_j.

Let $\mathfrak{b}_0, \mathfrak{b}_1 \subset V_i$, $\mathfrak{b}'_0, \mathfrak{b}'_1 \subset V_{i'}$, $\mathfrak{a}_0, \mathfrak{a}_1 \subset A_i$ and $\mathfrak{a}'_0, \mathfrak{a}'_1 \subset A_{i'}$ be defined as in Lemma 3.

Let $\mathfrak{a} \subset A_i$ be the subspace spanned by $\{A_{00}, \ldots, A_{0,m-1}\}$ and $\mathfrak{a}' \subset A_{i'}$ be the subspace spanned by $\{A'_{00}, \ldots, A'_{0,m-1}\}$.

Since $\mathfrak{b}_0$ and $\mathfrak{b}'_0$ are the same, we can identify $\mathfrak{a}$ with $\mathfrak{a}'$.

Let $a \in A_i$ defined by $u^{(i)}$ and $a' \in A_{i'}$ defined by $u^{(i')}$.

Let

$$v \oplus k = \overline{T}_i(a) \in V_i \oplus K_i$$

and

$$v' \oplus k' = \overline{T}_{i'}(a') \in V_{i'} \oplus K_{i'}.$$

Let

$$v = b_0 \oplus b_1 \quad \text{and} \quad v' = b_0 \oplus b'_1$$

and

$$a_0 = \tau_j^{-1}(b_0) \in \mathfrak{a} = \mathfrak{a}'.$$

Let

$$a_1 = a - a_0 \in \mathfrak{a}_1 \quad \text{and} \quad a'_1 = a' - a_0 \in \mathfrak{a}'_1.$$

Let z_0 be the polynomial corresponding to a_0 defined over $G_i \cup G_{i'}$.

Let z_1, z'_1 be the polynomials corresponding a_1 and a'_1 respectively. Then

$$u^{(i)} = z_0 + z_1 \quad \text{and} \quad u^{(i')} = z_0 + z'_1.$$

Also

$$|z_0|_{m,G_i \cup G_{i'}} = 0.$$

Then

$$\begin{aligned} \sup_{(x,y)\in S_j} |D_\alpha u^{(i)}(x, y) - D_\alpha u^{(i')}(x, y)| &= \sup_{(x,y)\in S_j} |D_\alpha z_1(x, y) - D_\alpha z'_1(x, y)| \\ &< \|\|D_\alpha z_1\|\|_{0,G_0} + \|\|D_\alpha z'_1\|\|_{0,G_{i'}} \\ &< C h^{m-1-|\alpha|} |u|_{m,G_i \cup G_{i'}} \quad \text{by Lemma 3.} \quad \text{Q.E.D.} \end{aligned}$$

Theorem 5. Let $u \in \mathring{\mathfrak{F}}(\mathfrak{h})$. Then there exists a constant $C > 0$ which is independent of $\mathfrak{h}$ and u such that

$$\|\|u\|\|_{m-1,G_0} < \frac{C}{h^{\frac{1}{2}}} |u|_{m,G_0}.$$

Proof. Without loss of generality we may assume that none of the grid lines is parallel to the coordinate axes. Define a function u' on $\mathbb{R}^2$ by letting $u'=u$ on G_0 and $u'=0$ elsewhere. For any line $y=y_0$ intersecting G_0, pick points (x_a, y_0) and (x_b, y_0) in ∂G such that all points having coordinates (x, y_0) and lying in G are between (x_a, y_0) and (x_b, y_0). For a fixed $\alpha=(\alpha_1, \alpha_2)$, $|\alpha|=m-1$, define a function

$$J_\alpha(t)=\begin{cases} D_\alpha u'\big(x_a+t(x_b-x_a), y_0\big) & \text{if} \quad \big(x_a+t(x_b-x_a), y_0\big)\in G_0. \\ 0 \quad \text{otherwise.} \end{cases}$$

Then $J_\alpha(t)$ is piecewise continuous and differentiable over open segments $(-\infty, t_0)$, $(t_0, t_1), \ldots, (t_k, \infty)$, where $t_0=0$, $t_k=1$.

Define

$$J_\alpha(t_i+)=\operatorname*{Lim}_{\substack{t\to t_i \\ t-t_i>0}} J_\alpha(t)$$

and

$$J_\alpha(t_i-)=\operatorname*{Lim}_{\substack{t\to t_i \\ t-t_i<0}} J_\alpha(t).$$

Let $J'_\alpha(t)=dJ_\alpha/dt$ for t in one of the open segments listed above. Then

$$J_\alpha(t_{i+1}-)=J_\alpha(t_i+)+\int_{t_i}^{t_{i+1}} J'_\alpha(\tau)\,d\tau \quad \text{for} \quad i=0, 1, \ldots, k-1.$$

$\therefore$ For any point $t\in(t_i, t_{i+1})$,

$$J_\alpha(t)=J_\alpha(t_i+)+\int_{t_i}^{t} J'_\alpha(\tau)\,d\tau,$$

$$\begin{aligned} |J_\alpha(t)|^2 &= \left[\sum_{j=0}^{i}\{J_\alpha(t_j+)-J_\alpha(t_j-)\}+\int_{t_0}^{t} J'_\alpha(\tau)\,d\tau\right]^2 \\ &\leqq 2\left[\sum_{j=0}^{i}\{J_\alpha(t_j+)-J_\alpha(t_j-)\}\right]^2+2\left[\int_{t_0}^{t} J'_\alpha(\tau)\,d\tau\right]^2 \\ &\leqq 2i\left[\sum_{j=0}^{i}\{J_\alpha(t_j+)-J_\alpha(t_j-)\}^2\right]+2t\left[\int_{t_0}^{t}|J'_\alpha(\tau)|^2 d\tau\right] \end{aligned}$$

by Cauchy-Schwarz Inequality. By Lemma 3 and Theorem 4, there exists $C_1>0$ such that

$$\sum\{J_\alpha(t_j+)-J_\alpha(t_j-)\}^2\leqq C_1\sum_k |u|^2_{m,G_k}\leqq C_1|u|^2_{m,G_0}.$$

By Lemma 3 again, there exists $C_2>0$ such that

$$|J'_\alpha(\tau)|^2\leqq\frac{C_2}{h^2}|u|^2_{m,G_k}$$

if $\big(x_a+\tau(x_b-x_a), y_0\big)\in G_k$. Therefore

$$\int_{t_0}^{t}|J'_\alpha(\tau)|^2\,d\tau\leqq C_2 h\sum_k |u|_{m,G_k}/h^2\leqq\frac{C_2}{h}|u|^2_{m,G_0}.$$

Condition (c) on the class of admissible grids implies that $i \leqq C_0/h$ and compactness of G_0 implies that $0 < C_3 < t$. Therefore there exists $C_4 > 0$ such that

$$|J_\alpha(t)|^2 \leqq \frac{C_4^2}{h} |u|^2_{m,G_0}$$

and

$$|||D_\alpha u|||_{0,G_0} = |||J_\alpha|||_{0,\mathbb{R}} \leqq \frac{C_4}{h^{\frac{1}{2}}} |u|_{m,G_0} \quad \text{for} \quad |\alpha| = m-1.$$

By successive integrations in a similar fashion,

$$|||D_\alpha u|||_{0,G_0} \leqq \frac{C_4}{h^{\frac{1}{2}}} |u|_{m,G_0} \quad \text{for} \quad |\alpha| \leqq m-1.$$

The lemma follows. Q.E.D.

5. Elliptic Boundary Value Problem

Let $L = \sum_{|\alpha| \leqq 2m} b_\alpha(\bar{x}) D^\alpha$ be a real strongly elliptic operator of order $2m$ defined in $\overline{G}$. Suppose that $b_\alpha \in H_{m+1}(\overline{G})$. Let the bilinear form associated with the operator L be given by

$$B(u, v) = \sum_{\substack{|\alpha| \leqq m \\ |\beta| \leqq m}} \int_G \{(D^\alpha u)(b_{\alpha\beta} D^\beta v)\} d\bar{x}$$

where $u, v \in H_m(\overline{G})$. We assume that $B(\cdot, \cdot)$ satisfies the Gärding's inequality in the following form:

$$|B(\varphi, \varphi)| \geqq C|\varphi|^2_{m,G} \quad \text{for all } \varphi \in C^\infty(\overline{G}') \tag{17}$$

where G' is any open subset of G.

For a given $f \in L_2(G)$, assume that there exists $\bar{u} \in H_{m+p+1}(G) \cap H_{2m}(G)$, satisfying zero Dirichlet data on the boundary and such that $\int_G v L \bar{u} = \int_G v f$ for all $v \in C^\infty(\overline{G})$.

We seek a solution $\hat{u} \in \mathring{\mathfrak{F}}(\mathfrak{h})$ such that the equation

$$B(\phi, \hat{u}) = \int_{G_0} \phi f \, d\bar{x} \tag{18}$$

is satisfied for all $\phi \in \mathring{\mathfrak{F}}(\mathfrak{h})$.

For a numerical solution of Eq. (18) we reduce it to an equivalent matrix equation as follows. We want to find $\hat{v} \in W$ such that $\hat{u} = I^{-1}(\hat{v})$ satisfies Eq. (18). Carrying out the integrations, $B(\varphi, \hat{u}) = (v, \overline{B}\,\hat{v})$ where $v = I(\varphi)$, $(,)$ is the inner product and $\overline{B}$ is a matrix. Also $\int \varphi f \, d\bar{x} = (v, F)$ where $F \in V$. Therefore Eq. (18) is equivalent to the matrix equation:

$$\overline{B}\,\hat{v} = F. \tag{19}$$

If within each cell, we approximate the coefficients $b_{\alpha\beta}$ and the function f by polynomials use of quadrature formulas for integration of polynomials over simplexes [4] simplifies the computation of $\overline{B}$. The computation is carried out by computing $B(v_i, u_i)$ cell by cell.

In the following, C, C_5, C_6, ... denote positive constants which do not depend upon a particular choice of $\mathfrak{h}$, f, $\bar{u}$.

We state without proof the following special form of the Green's formula. The proof is similar to the one given in reference [1].

Lemma 4. There exist linear differential operators $N_k(x, y, D)$ for $(x, y) \in \bigcup_i \partial G_i$ with coefficients bounded uniformly with respect to all (x, y) and all $\mathfrak{h}$, such that N_k is of order $\leqq k$ and for all $u \in C^\infty(\bar{G})$ and $\varphi \in \mathfrak{F}(\mathfrak{h})$,

$$\begin{aligned}\int_{G_0} \varphi L u = B(\varphi, u) &+ \sum_{S_k \notin \partial G} \sum_{j=0}^{m-1} \int_{S_k} (D_{n,j}\varphi^{(i_k)} - D_{n,j}\varphi^{(i')}) N_{2m-1-j}\, u\, d\xi \\ &+ \sum_{S_k \in \partial G} \sum_{j=0}^{m-1} \int_{S_k} (D_{n,j}\varphi^{(i_k)}) N_{2m-1-j}\, u\, d\xi\end{aligned} \tag{20}$$

where the grid line $S_k = \partial G_{i_k} \cap \partial G_{i'_k}$ or $\partial G \cap \partial G_{i_k}$.

Lemma 5. For a given $u \in H_{2m,\bar{G}} \cap H_{m+p+1,\bar{G}}$

$$\left| \int_{G_0} \phi L u - B(\phi, u) \right| \leqq C_{13} h^{p+1} |\phi|_{m,G_0} \|u\|_{m+p+1,G} \qquad \forall \phi \in \mathring{\mathfrak{F}}(\mathfrak{h}) \text{ and all } \mathfrak{h}.$$

Proof. If $\alpha = (\alpha_1, \alpha_2)$, let $\alpha! = \alpha_1!\, \alpha_2!$ and $(x, y)^\alpha = x^{\alpha_1} y^{\alpha_2}$. Then by Taylor's theorem

$$u = \sum_{|\alpha| \leqq p+m} \frac{1}{\alpha!} D^\alpha u(x_0, y_0)\, (x - x_0)^{\alpha_1} (y - y_0)^{\alpha_2} + R(x, y)$$

where

$$\|\!| D^\alpha R(x, y) |\!\|_G \leqq C_{10}\, d^{p+m+1-|\alpha|} \|u\|_{m+p+1,G}$$

and $d =$ distance between points (x, y) and (x_0, y_0).

Now any grid line S_k is given by the equations $x = f_k(\xi_k)$, $y = g_k(\xi_k)$. Let (x_0, y_0) be a point in S_k. Since f_k and g_k are C^∞ functions satisfying condition (d), it is possible to choose C_{10} such that

$$D_{n,j} u = \sum_{r=0}^{p+m-j} a_{n,r} \xi_k^r + R_j(\xi_k) \tag{21}$$

where

$$\|\!| R_j(\xi_k) |\!\|_{s_k} \leqq C_{10}\, \xi_k^{p+m-j+1} \|u\|_{m+p+1}.$$

Substituting (21) in Eq. (20)

$$\begin{aligned}\left| \int_{G_0} \phi L u - B(\phi, u) \right| = \sum_{s_k \notin \partial G} \Bigg[\sum_{j=0}^{m-1} \Bigg\{ \sum_{r=0}^{p+j+1-m} b_{nj}^k (\phi_{rj}^{i_k k} - \phi_{rj}^{i'_k k}) \\ + \int_{s_k} (D_{n,j}\phi^{i_k} - D_{n,j}\phi^{i'_k})\, R'_{k,j}(\xi_k) \Bigg\} \Bigg] \\ + \sum_{s_k \in \partial G} \Bigg[\sum_{j=0}^{m-1} \Bigg\{ \sum_{r=0}^{p+j+1-m} b_{nj}^k \phi_{rj}^{i_k k} + \int_{s_k} D_{n,j}\phi^{i_k} R'_{k,j}(\xi_k) \Bigg\} \Bigg]\end{aligned}$$

where

$$\|\!| R'_{k,j}(\xi_k) |\!\|_{s_k} \leqq C_{10}\, \xi_k^{p+j+2-m} \|u\|_{m+p+1}.$$

Since ϕ satisfies the compatibility conditions (5)

$$\phi_{rj}^{i_k k} = \phi_{rj}^{i'_k k} \quad \text{if} \quad s_k = \partial G_{i_k} \cap \partial G_{i'_k},$$
$$\phi_{rj}^{i_k k} = 0 \qquad \text{if} \quad s_k = \partial G_{i_k} \cap \partial G.$$

By Cauchy-Schwarz Inequality,

$$\left\{ \sum_{s_k \notin \partial G} \int_{s_k} (D_{n,j} \phi^{i_k} - D_{n,j} \phi^{i'_k}) R'_{k,j}(\xi_k) \right\}^2$$
$$\leqq \left[\sum_{s_k \notin \partial G} \int_{s_k} (D_{n,j} \phi^{i_k} - D_{n,j} \phi^{i'_k})^2 d\xi_k \right] \left[\sum_{s_k \notin \partial G} \int_{s_k} R'^2_{kj} d\xi_k \right]$$
$$\leqq C_1 C_{10} \left[h^{2(m-1-j)} \left\{ \sum_{s_k \notin \partial G} \int_{s_k} (|\phi|^2_{m,G_{i_k}} + |\phi|^2_{m,G_{i'_k}}) \right\} \right] \left[\sum_{s_k} \int_{s_k} \xi_k^{2(p+j+2-m)} d\xi_k \right]$$

(by Lemma 3 and Theorem 4)

$$\leqq 2 C_1 C_{10} \left[h^{2(m-1-j)} h \, |\phi|^2_{m,G_0} \right] \left[h^{2(p+j+2-m)} h \frac{(\text{area of } G)}{\delta^2 h^2} \right] \|u\|^2_{m+p+1}$$
$$\leqq C_{11} h^{2(p+1)} |\phi|^2_{m,G_0} \|u\|^2_{m+p+1}.$$

Similarly

$$\left\{ \sum_{s_k \in \partial G} \int_{s_k} D_{nj} \phi^{i_k} R'_{kj} d\xi_k \right\}^2 \leqq C_{11} h^{2(p+1)} |\phi|^2_{m,G_0} \|u\|^2_{m+p+1}.$$

Therefore

$$\left| \int_{G_0} \phi \, Lu - B(\phi, u) \right| \leqq C_{13} h^{p+1} |\phi|_{m,G_0} \|u\|_{m+p+1}. \qquad \text{Q.E.D.}$$

Theorem 6. Eq. (18) has a unique solution $\hat{u} \in \mathring{\mathfrak{H}}(\mathfrak{h})$ and

$$|||\hat{u} - \bar{u}|||_{m-1,G_0} \leqq C_{26} h^{p+\frac{1}{2}} \|\bar{u}\|_{m+p+1,G}$$
$$|\hat{u} - \bar{u}|_{m,G_0} \leqq C_{27} h^{p+1} \|\bar{u}\|_{m+p+1,G}.$$

Proof. Suppose for $\varphi, \hat{u} \in \mathring{\mathfrak{H}}(\mathfrak{h})$, $B(\varphi, \hat{u}) = 0$ for all φ.

$\therefore \quad 0 = |B(\hat{u}, \hat{u})| \geqq C |\hat{u}|^2_{m,G_0}$

$\therefore \quad |||\hat{u}|||_{m-1,G_0} = 0$ by Theorem 5

$\therefore \quad \hat{u} = 0$

$\therefore$ Eq. (19) has a unique $\hat{u} \in \mathring{\mathfrak{H}}(\mathfrak{h})$.

Now by Theorem 3, there exists ${}^0u \in \mathring{\mathfrak{H}}(\mathfrak{h})$

$$|||\bar{u} - {}^0u|||_{k,G} < C_{18} h^{m+p+1-k} \|\bar{u}\|_{m+p+1,G}.$$

Let $\Delta u = \hat{u} - {}^0u \in \mathring{\mathfrak{H}}(\mathfrak{h})$ and $\tilde{u} = \bar{u} - {}^0u$

$$\therefore \quad B(\Delta u, \hat{u}) = \int_{G_0} f \, \Delta u = \int (\Delta u)(L\bar{u})$$
$$= B(\Delta u, {}^0u + \tilde{u}) + \{ \int (\Delta u)(L\bar{u}) - B(\Delta u, \bar{u}) \}.$$

Therefore by Lemma 5,

$$|B(\Delta u, \Delta u)| \leqq |B(\Delta u, \tilde{u})| + C_{13} h^{p+1} |\Delta u|_{m,G_0} \|\bar{u}\|_{m+p+1,G}.$$

Now the Cauchy-Schwarz Inequality implies that

$$|B(\Delta u, \tilde{u})| \leqq C_{19}\|\Delta u\|_{m,G_0}\|\tilde{u}\|_{m,G_0} \leqq C_{20}h^{p+1}\|\Delta u\|_{m,G_0}. \tag{22}$$

Inequalities (17) and (22) imply that

$$|\Delta u|_{m,G_0} \leqq C_{24}h^{(p+1)}\|u\|_{m+p+1}.$$

By Theorem 5,

$$\begin{aligned}
|||\Delta u|||_{m-1,G_0} &\leqq C_{25}h^{p+\frac{1}{2}}\|\bar{u}\|_{m+p+1,G} \\
|||\hat{u}-\bar{u}|||_{m-1,G_0} &\leqq |||\Delta u - \tilde{u}|||_{m-1,G_0} \leqq |||\Delta u|||_{m-1,G_0} + |||\tilde{u}|||_{m-1,G_0} \\
&\leqq C_{26}h^{p+\frac{1}{2}}\|u\|_{m+p+1,G}
\end{aligned}$$

and

$$|\hat{u}-\bar{u}|_{m,G_0} \leqq C_{27}h^{(p+1)}\|\bar{u}\|_{m+p+1,G}. \quad \text{Q.E.D.}$$

References

1. Agmon, S.: Lectures on elliptic boundary value problems. New York: Van Nostrand 1965.
2. Ahlberg, J. H., Nilson, E. N., Walsh, J. L.: The theory of splines and their applications. New York: Academic Press 1967.
3. Ciarlet, P. G., Schultz, M. H., Varga, R. S.: Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem. Numer. Math. **9**, 394–430 (1967).
4. Hammer, P. C., Marlowe, O. J., Stroud, A. H.: Numerical integration over simplexes and cones. Math. Tables and other Aids to Comp. V. **10**, 130–136 (1959).

J. M. Shah
Project MAC
Mass. Institute of Technology
545 Technology Square
Cambridge, Mass. 02139, USA

Numer. Math. 15, 15—48 (1970)

Die Konvergenz des Rand- und Eigenwertproblems linearer gewöhnlicher Differenzengleichungen

ROLF DIETER GRIGORIEFF

Eingegangen am 25. Januar 1969

Summary. In this paper the convergence of finite difference approximations for the general eigenvalue and boundary value problem of ordinary differential equations is proved under the condition of consistency and stability. The eigenvalues are shown to converge preserving multiplicity. Estimates are given for the rate of convergence of difference quotients and eigenvalues.

Einleitung

In der vorliegenden Arbeit wird die Konvergenz von Differenzenapproximationen regulärer Rand- und Eigenwertprobleme bei linearen gewöhnlichen Differentialgleichungen beliebiger Ordnung bewiesen. Dabei zeigt sich ähnlich wie bei Differenzenverfahren für Anfangswertprobleme, daß konsistente und stabile Approximationen konvergent sind. Mit den hier gewonnenen Ergebnissen läßt sich z. B. die Konvergenz der in [2, 5, 6, 8] angegebenen Differenzenverfahren zeigen, womit insbesondere die bisher ungelöste Frage nach der Konvergenz von Differenzenapproximationen höherer Genauigkeit erledigt wird.

Differenzenmethoden sind in Anbetracht ihrer vielfachen Anwendung bei der numerischen Behandlung von Rand- und Eigenwertproblemen gewöhnlicher Differentialgleichungen auch im Hinblick auf ihre Konvergenz in der Literatur zahlreich untersucht worden. Dabei beschränken sich die Autoren jedoch mit Ausnahme von [6] auf die Untersuchung selbstadjungierter Gleichungen zweiter und vierter Ordnung, die bei Eigenwertaufgaben (s. [3, 4, 15, 25, 26]) halbbeschränkt und bei Randwertaufgaben (s. [1, 9, 14, 25, 26]) positiv definit vorausgesetzt werden. Wenig scheint bisher über nichtsymmetrische Probleme und solche ungerader und höherer als vierter Ordnung bekannt zu sein, wie auch über den Fall, daß in den Randbedingungen die beiden Intervallenden nicht separiert sind. Vom trivialen Fall der Multiplikation mit einer positiven Funktion abgesehen, ist anscheinend auch das allgemeine Eigenwertproblem nicht untersucht worden, in dem der Eigenwertparameter mit einem Differentialoperator gekoppelt ist, der gegebenenfalls nicht einmal positiv definit zu sein braucht. Offen ist auch die Frage nach der Konvergenz allgemeinerer Differenzenapproximationen, insbesondere von Verfahren höherer Annäherung (s. [16, S. 81 u. 104]). Ebenso scheint die Konvergenz der Eigenwerte der Vielfachheit nach nur für das Dirichletproblem des Laplaceschen Δ-Operators (s. [18]) bewiesen zu sein. Diese und andere Fragen sollen in dieser Arbeit beantwortet werden.

Unsere Untersuchungen bauen auf einer Approximationstheorie von Eigenwertproblemen und Gleichungen zweiter Art in Hilbertschen Räumen auf, die wir

in [12] entwickelt haben. Dort haben wir in einem abstrakten Rahmen die Konvergenz von Rand- und Eigenwertproblemen bewiesen, wobei wir uns einer in [24] entwickelten Konvergenztheorie bedienten, die sich insbesondere auch für Differenzenverfahren als sehr geeignet erweist. Die in [12] gewonnenen Ergebnisse beruhen auf sog. Koerzitivitätsungleichungen für die Differenzenoperatoren in Verbindung mit Kompaktheitsaussagen in einer Klasse von Hilbertschen Räumen. Analoge Methoden sind in [23] bei der Behandlung des Dirichletproblems elliptischer Differenzengleichungen verwendet worden. Auch in [1, 14, 20, 25 — 27] werden in diesem Zusammenhang a priori Abschätzungen benutzt, allerdings unter etwas anderen Gesichtspunkten. Beim Eigenwertproblem werden außerdem Ergebnisse aus der Spektraltheorie linearer Operatoren in Banachschen Räumen herangezogen, wobei sich im Hinblick auf die vorliegende Anwendung ein enger Zusammenhang der Differenzenverfahren mit Problemen der Störungstheorie zeigt.

Die aus [12] benötigten wichtigsten Begriffe und Ergebnisse stellen wir in § 1 zusammen. Hier definieren wir u. a. durch die Eigenschaften (H 1) — (H 4) die Räume $\mathbb{H}^m$ und eine Schar von einem Parameter $h \in \Lambda_0$ abhängige Räume H^m und führen die von uns verwendeten Konvergenzbegriffe ein. Im zweiten Paragraphen identifizieren wir $\mathbb{H}^m$ mit dem Sobolewschen Funktionenraum $W^{m,2}(c_0, d_0)$ und H^m mit einem Raum von Gitterfunktionen, der mit einem analogen Skalarprodukt wie $\mathbb{H}^m$ versehen ist. Die Eigenschaften (H 1) — (H 4) werden für diese Räume in § 2 und 3 mit Hilfe von Restriktions- und Einbettungsoperatoren nachgewiesen. Besonders wichtig ist hier die Bedingung (H 4), in der die schwache und starke Kompaktheit von in H^m beschränkten Folgen gezeigt wird. Mit Hilfe von Einbettungsoperatoren kann man auch konvergente Folgen von Gitterfunktionen zu in $\mathbb{H}^m$ konvergenten Funktionenfolgen fortsetzen.

Der nächste Paragraph enthält unseren Hauptkonvergenzsatz. In einem endlichen Intervall aus $\mathbb{R}^1$ betrachten wir für einen regulären gewöhnlichen Differentialoperator m-ter Ordnung das allgemeine Rand- und Eigenwertproblem, das wir mit (I) bezeichnen wollen. Der Eigenwertparameter λ tritt dabei mit einem Differentialoperator $(m-1)$-ter Ordnung auf. Zu (I) wird eine Schar von Differenzenapproximationen (II) konstruiert. Dabei unterliegen die Approximationen der Randbedingungen und Ableitungen bis zur Ordnung $m-1$ nur der Konsistenzforderung. Bei der Ableitung m-ter Ordnung kommt man mit der Konsistenz alleine nicht aus, wie Beispiele in [8, 10, 22] zeigen, vielmehr muß außerdem noch eine Stabilitätsbedingung erfüllt sein. Hierfür werden in (Z 3) algebraische Kriterien angegeben, unter denen man die für diese Fragen grundlegende Koerzitivitätsungleichung aus [13] zur Verfügung hat. Die Koeffizienten der Differenzenoperatoren und das inhomogene Glied kann man mit Hilfe der Restriktionsoperatoren aus § 2 aus den entsprechenden Größen in (I) erhalten. Unser Hauptergebnis ist der Satz (4.10), der besagt, daß die Lösungen von (II) einschließlich der Differenzenquotienten bis zur Ordnung $m-1$ gleichmäßig konvergieren, falls λ kein Eigenwert von (I) ist. Die Konvergenz der Eigenwerte von (II) erfolgt der algebraischen Vielfachheit nach. Den Beweis von (4.10) erbringen wir im nächsten Paragraphen durch Zurückführung auf die Sätze in § 1.

In § 6 zeigen wir in Satz (6.1) die Konvergenz der Eigenfunktionen bei Vorliegen eines einfachen Eigenwertes von (I). Anschließend geben wir hinreichende

Bedingungen an, daß die Konvergenz der Eigenwerte der geometrischen Vielfachheit nach erfolgt. In den Sätzen (6.5), (6.7) und (6.8) wird dann bewiesen, daß die Lösungen von (II) und bei geeigneter Normierung auch die Eigenfunktionen mindestens mit der Ordnung des Abschneidefehlers konvergieren, während die Konvergenzordnung der Eigenwerte nicht kleiner als die n-te Wurzel aus der Ordnung des Abschneidefehlers ist, wo n die Ordnung des Pols der Resolvente von (I) ist. Der letzte Paragraph enthält zwei Beispiele, ein Mehrstellenverfahren und eine Differenzenapproximation höherer Annäherung.

Mit den vorliegenden Methoden läßt sich auch die Konvergenz von Differenzenverfahren bei Rand- und Eigenwertproblemen für Systeme gewöhnlicher Differentialgleichungen beweisen. Schließlich möchte ich hiermit Herrn Prof. Dr. F. Stummel für die Anregung zu dieser Arbeit danken.

1. Zusammenstellung einiger Ergebnisse

1.1. In [12] haben wir eine Approximationstheorie entwickelt, deren für die vorliegende Arbeit wichtigsten Ergebnisse wir in diesem Paragraphen zusammenstellen.

Unseren Untersuchungen liegen die Hilbertschen Räume $\mathbb{H}^j$ und H^j zugrunde, die durch die nachfolgenden Eigenschaften (H 1)—(H 4) charakterisiert werden.

(H 1) *Es sei $\mathbb{H}^0$ ein Hilbertscher Raum mit dem Skalarprodukt $((\cdot,\cdot))_0$ und der zugehörigen Norm $\|\cdot\|_0$. Es sei m eine natürliche Zahl, und für jedes $j=1,\ldots,m$ seien $\mathbb{H}^j$ lineare Teilräume von $\mathbb{H}^0$ mit $\mathbb{H}^j \subseteq \mathbb{H}^{j-1}$, $j=1,\ldots,m$, und $Z_j^0\colon \mathbb{H}^j \to \mathbb{H}^0$, $j=0,\ldots,m$, lineare Operatoren mit $Z_0^0=I$, der identischen Abbildung von $\mathbb{H}^0$ auf sich. Die Räume $\mathbb{H}^j$ seien mit dem Skalarprodukt und der Norm*

$$((u,v))_j=\sum_{k=0}^{j}((Z_k^0u,Z_k^0v))_0, \qquad \|u\|_j=((u,u))_j^{\frac{1}{2}}, \qquad u,v\in\mathbb{H}^j.$$

Hilbertsche Räume, und $\mathbb{H}$ sei ein in jedem $\mathbb{H}^j$, $j=0,\ldots,m$, enthaltener und dicht liegender Teilraum von $\mathbb{H}^0$.

(H 2) *Für jedes $h\in\Lambda_0$, einer abzählbar unendlichen Indexmenge, sei ein Hilbertscher Raum H^0 gegeben mit dem Skalarprodukt $(\cdot,\cdot)_0$ und der zugehörigen Norm $|\cdot|_0$. Es seien $H^j\subseteq H^{j-1}$, $j=1,\ldots,m$, lineare Teilräume von H^0 und $Z_j\colon H^j\to H^0$, $j=0,\ldots,m$, lineare Operatoren mit $Z_0=I$, der identischen Abbildung von H^0 auf sich. Für jedes $j=1,\ldots,m$ sei H^j mit dem Skalarprodukt und der Norm*

$$(u,v)_j=\sum_{k=0}^{j}(Z_ku,Z_kv)_0, \qquad |u|_j=(u,u)_j^{\frac{1}{2}}, \qquad u,v\in H^j,\ h\in\Lambda_0,$$

ein Hilbertscher Raum.

Im allgemeinen wird sich aus dem jeweiligen Zusammenhang ergeben, welche Größen vom Parameter $h\in\Lambda_0$ abhängen, so daß wir ihn in der Schreibweise oft nicht besonders sichtbar machen, zumal von h unabhängige Größen im Regelfall mit einem Index „0" versehen werden. Mit Λ bezeichnen wir stets eine Teilfolge von Λ_0, nicht jedesmal notwendig dieselbe. Für die Konvergenz einer Folge $\{u\}$,

$h \in \Lambda$, gegen ein Element u_0 schreiben wir (bei Vorliegen eines Konvergenzbegriffes) $u \to u_0$ $(h \in \Lambda)$.

Die Räume $\mathbb{H}^j$ und H^j, $h \in \Lambda_0$, stehen in einer gewissen Beziehung zueinander, die durch die Eigenschaften (H 3) und (H 4) festgelegt wird.

(H 3) *Für jedes $h \in \Lambda_0$ existiere ein linearer Operator $R: \mathbb{H} \to H^m$ mit der Eigenschaft, daß für alle $\varphi, \psi \in \mathbb{H}$*

$$(Z_j R\varphi, Z_k R\psi)_0 \to ((Z_j^0 \varphi, Z_k^0 \psi))_0 \quad (h \in \Lambda_0), \quad j, k = 0, \ldots, m,$$

konvergiert.

(H 4) *Ist K eine positive Zahl und $\{u\}$, $h \in \Lambda$, eine Folge von Elementen $u \in H^m$ mit $|u|_m \leqq K$, $h \in \Lambda$, dann existiert ein Element $u_0 \in \mathbb{H}^m$ und eine Teilfolge $\Lambda' \leqq \Lambda$ mit*

$$u \rightharpoonup u_0 \quad in \quad \{H^m \to \mathbb{H}^m\}, \quad u \to u_0 \quad in \quad \{H^{m-1} \to \mathbb{H}^{m-1}\} \qquad (h \in \Lambda').$$

Die vorstehende Konvergenz der Folge $\{u\}$, $h \in \Lambda'$, ist dabei wie folgt definiert.

(1) *Sei j eine nichtnegative ganze Zahl, sei $\{u\}$, $h \in \Lambda$, eine Folge von Elementen $u \in H^j$, $h \in \Lambda$, und sei $u_0 \in \mathbb{H}^j$. Dann konvergiert*

$$u \rightharpoonup u_0 \quad in \quad \{H^j \to \mathbb{H}^j\} \qquad (h \in \Lambda),$$

genau wenn eine Zahl K existiert, so daß für jedes $k = 0, \ldots, j$ und jedes $\varphi \in \mathbb{H}$ gilt

$$|Z_k u|_0 \leqq K, \quad h \in \Lambda, \quad (R\varphi, Z_k u)_0 \to ((\varphi, Z_k^0 u_0))_0 \quad (h \in \Lambda).$$

Es konvergiert

$$u \to u_0 \quad in \quad \{H^j \to \mathbb{H}^j\} \qquad (h \in \Lambda),$$

genau wenn für jedes $k = 0, \ldots, j$ gilt

$$u \rightharpoonup u_0 \quad in \quad \{H^j \to \mathbb{H}^j\}, \quad |Z_k u|_0 \to \|Z_k^0 u_0\|_0 \qquad (h \in \Lambda).$$

Mit $(\mathbb{H}^m)^*$ bzw. $(H^m)^*$ bezeichnen wir den Banachraum der linearen beschränkten Funktionale auf $\mathbb{H}^m$ bzw. H^m und geben die folgende Definition.

(2) *Sei $l_0 \in (\mathbb{H}^m)^*$ und $\{l\}$, $h \in \Lambda$, eine Folge von Funktionalen $l \in (H^m)^*$, $h \in \Lambda$. Dann konvergiert*

$$l \to l_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \qquad (h \in \Lambda),$$

genau wenn für jede Folge $\{u\}$, $h \in \Lambda$, von Elementen $u \in H^m$, $h \in \Lambda$, mit $u \rightharpoonup u_0$ in $\{H^m \to \mathbb{H}^m\}$ folgt $l(u) \to l_0(u_0)$ $(h \in \Lambda)$.

Ist $B(\mathbb{H}^0)$ bzw. $B(H^0)$ der Banachraum der linearen beschränkten Operatoren von $\mathbb{H}^0 \to \mathbb{H}^0$ bzw. $H^0 \to H^0$, so vereinbaren wir die nachstehende Definition.

(3) *Sei $b_0 \in B(\mathbb{H}^0)$ und $\{b\}$, $h \in \Lambda$, eine Folge von Operatoren $b \in B(H^0)$, $h \in \Lambda$. Dann konvergiert*

$$b \to b_0 \qquad (h \in \Lambda),$$

genau wenn für jede Folge $\{u\}$, $h \in \Lambda$, von Elementen $u \in H^0$, $h \in \Lambda$, mit $u \to u_0$ in $\{H^0 \to \mathbb{H}^0\}$ $(h \in \Lambda)$ folgt

$$bu \to b_0 u_0 \quad in \quad \{H^0 \to \mathbb{H}^0\} \qquad (h \in \Lambda).$$

Durch eine Widerspruchsannahme beweist man leicht, daß unter der Voraussetzung von (2) (bzw. (3)) eine Konstante K existiert mit $|l|_m \leqq K$ (bzw. $|b| \leqq K$), $h \in \Lambda$.

Steht b_0^* bzw. b^* für den zu b_0 bzw. b adjungierten Operator, so kann man zeigen, daß $b \to b_0$ ($h \in \Lambda$) genau dann gilt, wenn für jede Folge $\{u\}$, $h \in \Lambda$, von Elementen $u \in H^0$, $h \in \Lambda$, mit $u \rightharpoonup u_0$ in $\{H^0 \to \mathbb{H}^0\}$ ($h \in \Lambda$) gilt

$$b^* u \rightharpoonup b_0^* u_0 \quad in \quad \{H^0 \to \mathbb{H}^0\} \qquad (h \in \Lambda).$$

(4) *Eine Folge* $\{b\}$, $h \in \Lambda$, *von Operatoren* $b \in B(H^0)$ *heißt m-regulär, genau wenn eine Konstante* $K > 0$ *existiert mit der Eigenschaft*

$$|u|_m \leqq K(|bZ_m u|_0 + |u|_{m-1}), \qquad u \in H^m, \ h \in \Lambda.$$

Die Folge heißt regulär, genau wenn gilt

$$|u|_0 \leqq K|bu|_0, \qquad u \in H^0, \ h \in \Lambda.$$

Mit diesen Begriffen sind wir in der Lage, das Problem π_λ^0 und die Folge approximierender Probleme π_λ, $h \in \Lambda_0$, zu formulieren. Gegeben seien Operatoren a_j^0, $b_k^0 \in B(\mathbb{H}^0)$, $j = 0, \ldots, m$, $k = 0, \ldots, m-1$, und Funktionale $l_0^{(j)} \in (\mathbb{H}^m)^*$, $j = 1, \ldots, m$. Wir erklären die Operatoren L_0 und M_0: $\mathbb{H}^m \to \mathbb{H}^0$ durch

$$L_0 = \sum_{k=0}^{m} a_k^0 Z_k^0, \qquad M_0 = \sum_{k=0}^{m-1} b_k^0 Z_k^0.$$

Sind dann λ und $\gamma_0^{(j)}$, $j = 1, \ldots, m$, komplexe Zahlen und ist $f_0 \in \mathbb{H}^0$, so verstehen wir unter $\pi_\lambda^0(L_0, f_0, l_0^{(j)}, \gamma_0^{(j)})$ das Problem

(5) *Gesucht ist ein Element* $u_0 \in \mathbb{H}^m$, *das die Gleichungen*

$$(\lambda M_0 - L_0) u_0 = f_0, \qquad l_0^{(j)}(u_0) = \gamma_0^{(j)}, \qquad j = 1, \ldots, m,$$

erfüllt.

Mit anderen Worten ist die Lösung einer Gleichung zweiter Art auf einer linearen Mannigfaltigkeit von $\mathbb{H}^m$ gesucht.

Die approximierenden Probleme haben ganz entsprechende Gestalt. Für jedes $h \in \Lambda_0$ seien Operatoren a_j, $b_k \in B(H^0)$, $j = 0, \ldots, m$, $k = 0, \ldots, m-1$, und Funktionale $l^{(j)} \in (H^m)^*$, $j = 1, \ldots, m$, gegeben. Dann definieren wir die Abbildungen L und M: $H^m \to H^0$ durch

$$L = \sum_{k=0}^{m} a_k Z_k, \qquad M = \sum_{k=0}^{m-1} b_k Z_k, \qquad h \in \Lambda.$$

Sind $\lambda, \gamma^{(j)}$, $j = 1, \ldots, m$, komplexe Zahlen und ist $f \in H^0$, so steht $\pi_\lambda(L, f, l^{(j)}, \gamma^{(j)})$, $h \in \Lambda_0$, für das Problem

(6) *Gesucht ist ein Element* $u \in H^m$, *das die Gleichungen*

$$(\lambda M - L) u = f, \qquad l^{(j)}(u) = \gamma^{(j)}, \qquad j = 1, \ldots, m,$$

erfüllt.

Es sei dann noch $\{P\}$, $h \in \Lambda_0$, eine Folge von Projektionen $P \in B(H^0)$, $h \in \Lambda_0$. Wir setzen

$$X_1 = \{u \in H^m \mid l^{(j)}(u) = 0, \ j = 1, \ldots, m\}, \qquad X_2 = PH^0$$

und nehmen an, daß $LH^m \subseteq X_2$, $MH^m \subseteq X_2$ erfüllt ist. Die Resolventenmenge $\varrho(L)$ besteht dann aus allen Zahlen $\lambda \in \mathbb{C}$, der komplexen Zahlenebene, so daß die Restriktion des Operators $\lambda M - L$ auf den Teilraum X_1 eine auf X_2 erklärte beschränkte Inverse besitzt. Das Komplement $\sigma(L) = \mathbb{C} - \varrho(L)$ heißt das Spektrum von L. Ein Element $u \in X_1$, $u \neq 0$, heißt Eigenvektor bzw. Hauptvektor 1. Stufe von L zum Eigenwert $\lambda \in \mathbb{C}$, genau wenn $(\lambda M - L)\, u = 0$ gilt. Allgemeiner heißt $u \in X_1$ Hauptvektor n-ter Stufe von L zum Eigenwert λ, $n \geqq 2$, genau wenn ein Hauptvektor v $(n-1)$-ter Stufe von L zu λ existiert mit $(\lambda M - L)\, u = Mv$. Die Dimension des von den Eigenvektoren (bzw. Hauptvektoren) von L zu λ aufgespannten Unterraumes heißt die geometrische (bzw. algebraische) Vielfachheit von λ. Für das Problem π_λ^0 sind $\varrho(L_0)$, $\sigma(L_0)$, X_1^0, X_2^0 usf. analog erklärt, nur ist anstelle von P die Identität I in $\mathbb{H}^0$ zu setzen.

Wir treffen dann die folgende generelle Voraussetzung.

(S) (i) *Für jedes* $j = 0, \ldots, m$, $k = 0, \ldots, m-1$, *gilt* $a_j \to a_j^0$ *und* $b_k \to b_k^0$ $(h \in \Lambda_0)$.

(ii) *Die Folge* $\{a_m\}$ *ist m-regulär, und es konvergiert* $a_m \to a_m^*$ $(h \in \Lambda_0)$.

(iii) *Für jedes* $j = 1, \ldots, m$ *gilt* $l^{(j)} \to l_0^{(j)}$ *in* $\{H^m \to \mathbb{H}^m\}$ $(h \in \Lambda_0)$.

(iv) *Für jede Folge* $\{u\}$, $h \in \Lambda$, *von Elementen* $u \in H^m$, $h \in \Lambda$, *mit* $u \to u_0$ *in* $\{H^m \to \mathbb{H}^m\}$ $(h \in \Lambda)$ *folgt* $Mu \to M_0 u_0$ *in* $\{H^0 \to \mathbb{H}^0\}$ $(h \in \Lambda)$.

Es ist leicht zu sehen, daß dann eine Konstante K existiert mit $|Mu|_0 \leqq K |u|_m$, $h \in \Lambda$. Damit gilt der folgende Satz.

(7) *Sei* $\mathbb{C}_0$ *eine kompakte Teilmenge von* $\mathbb{C}$, *die keine Eigenwerte von* L_0 *enthält. Unter den Voraussetzungen* (S) (i)–(iv) *existiert dann eine Konstante* K, *so daß für fast alle* $h \in \Lambda_0$ *die Ungleichung besteht*

$$|u|_m \leqq K \left(|(\lambda M - L)\, u|_0 + \sum_{j=1}^{m} |l^{(j)}(u)| \right), \quad u \in H^m, \quad \lambda \in \mathbb{C}_0.$$

Eine weitere letzte Voraussetzung ist

(S) (v) *Die Operatoren* M_0 *und* M, $h \in \Lambda_0$, *sind vollstetig,* $\varrho(L)$ *ist für fast alle* $h \in \Lambda_0$ *nichtleer, es konvergiert* $P \to I$ $(h \in \Lambda_0)$, *und das homogene Problem* $\pi_\lambda^0(L_0, 0, l_0^{(j)}, 0)$ *besitzt für wenigstens ein* $\lambda \in \mathbb{C}$ *nur die Nullösung.*

Dann besteht das Spektrum $\sigma(L_0)$ und $\sigma(L)$ aus höchstens abzählbar unendlich vielen Eigenwerten endlicher algebraischer Vielfachheit, die keinen im Endlichen gelegenen Häufungspunkt besitzen, und es gelten die beiden folgenden Sätze.

(8) *Ist* λ *kein Eigenwert von* L_0, *so liegt* λ *in* $\varrho(L_0)$. *Sind für fast alle* $h \in \Lambda_0$ *die Funktionale* $l^{(j)}$, $j = 1, \ldots, m$, *linear unabhängig, ist* $f_0 \in \mathbb{H}^0$ *und* $\{f\}$, $h \in \Lambda_0$, *eine Folge von Elementen* $f \in PH^0$, $h \in \Lambda_0$, *mit* $f \to f_0$ *in* $\{H^0 \to \mathbb{H}^0\}$ $(h \in \Lambda_0)$ *und sind* $\{\gamma^{(j)}\}$, $h \in \Lambda_0$, $j = 1, \ldots, m$, *Folgen komplexer Zahlen mit* $\gamma^{(j)} \to \gamma_0^{(j)}$ $(h \in \Lambda_0)$, $j = 1, \ldots, m$, *so ist das Problem* $\pi_\lambda^0(L_0, f_0, l_0^{(j)}, \gamma_0^{(j)})$ *eindeutig lösbar, es existiert für fast alle* $h \in \Lambda_0$ *eine eindeutig bestimmte Lösung* u *des Problems* $\pi_\lambda(L, f, l^{(j)}, \gamma^{(j)})$, *und es konvergiert*

$$u \to u_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \qquad (h \in \Lambda_0)$$

gegen die Lösung u_0 *von* π_λ^0.

(9) *Jede kompakte Teilmenge von $\varrho(L_0)$ liegt für fast alle $h\in\Lambda_0$ in $\varrho(L)$. Ist λ_0 ein Eigenwert von L_0 der algebraischen Vielfachheit n_0 und ist U eine Umgebung von λ_0, in deren Abschließung keine weiteren Eigenwerte von L_0 liegen, so ist für fast alle $h\in\Lambda_0$ die Summe der algebraischen Vielfachheiten der in U gelegenen Eigenwerte von L gleich n_0, d.h. die Eigenwerte von L konvergieren gegen die Eigenwerte von L_0 der algebraischen Vielfachheit nach.*

2. Die Räume $\mathbb{H}^m$ und H^m

2.1. Mit den im vorangehenden Paragraphen aufgeführten Ergebnissen wollen wir im folgenden die Differenzenverfahren bei gewöhnlichen Differentialgleichungen behandeln. Wir beginnen damit, die dazu benötigten Räume $\mathbb{H}^j$ und H^j, $j=0,\ldots,m$, zu definieren und für sie die Eigenschaften (H1)—(H3) nachzuweisen. Den Nachweis von (H4) verschieben wir auf den nächsten Paragraphen.

Es sei $\mathbb{R}^1$ der eindimensionale euklidische Vektorraum. Ist $G\subseteq\mathbb{R}^1$ eine offene Menge, so sei $L^2(G)$ der Vektorraum der auf G erklärten komplexwertigen Funktionen, die über G im Sinne von Lebesgue absolut quadratisch integrierbar sind. In $L^2(G)$ erklären wir durch

$$((u,v))=\int_G u(x)\,\bar v(x)\,dx,\qquad \|u\|=((u,u))^{\frac12},\qquad u,v\in L^2(G),$$

ein Skalarprodukt und die zugehörige Norm, wodurch $L^2(G)$ ein Hilbertscher Raum wird.

Es seien c_0, d_0 zwei Punkte aus $\mathbb{R}^1$ mit $c_0<d_0$. Dann sei $\mathbb{H}^0=L^2(c_0,d_0)$ und $\mathbb{H}^j$, $j=1,\ldots,m$, der Teilraum der Funktionen aus $\mathbb{H}^0$, die im Sinne der Distributionstheorie Ableitungen bis zur j-ten Ordnung besitzen, die in $\mathbb{H}^0$ liegen. Wir setzen noch

$$Z_j^0=\left(\frac{d}{dx}\right)^j,\qquad j=1,\ldots,m,$$

dem Differentialquotienten j-ter Ordnung, womit $\mathbb{H}^j$ gleich dem Sobolewschen Funktionenraum $W^{j,2}(c_0,d_0)$ wird und die Bedingung (H1) aus §1 erfüllt ist. Für den Teilraum $\mathbb{H}$ nehmen wir die Menge $C^m[c_0,d_0]$ der im abgeschlossenen Intervall $[c_0,d_0]$ m-mal stetig differenzierbaren Funktionen.

2.2. Wir kommen nun zur Definition der Räume H^j, $j=0,\ldots,m$. Es sei Λ_0 eine Nullfolge positiver Zahlen und $\{y\}$, $h\in\Lambda_0$, sei eine Folge von Punkten $y\in\mathbb{R}^1$, $h\in\Lambda_0$. Wir erklären dann die Menge $\mathbb{R}_h^1$ von Gitterpunkten aus $\mathbb{R}^1$ durch

$$\mathbb{R}_h^1=\{x\in\mathbb{R}^1\,|\,x=y+jh,\ j=0,\pm1,\ldots\}.$$

Mit Ω bezeichnen wir den Vektorraum der auf $\mathbb{R}_h^1$ erklärten komplexwertigen Funktionen. Ist G eine Punktmenge aus $\mathbb{R}^1$, so sei $G_h=G\cap\mathbb{R}_h^1$ und

$$l^2(G)=\{u\in\Omega\,|\,u(x)=0,\ x\in\mathbb{R}_h^1-G_h,\ \sum_{x\in\mathbb{R}_h^1}|u(x)|^2<\infty\}.$$

In $l^2(G)$ wird durch die Vorschrift

$$(u,v)=h\sum_{x\in\mathbb{R}_h^1}u(x)\,\bar v(x),\qquad |u|=(u,u)^{\frac12},\qquad u,v\in l^2(G),$$

ein Skalarprodukt und die zugehörige Norm erklärt, wodurch $l^2(G)$ ein Hilbertscher Raum wird. Es seien $\{c\}$, $\{d\}$, $c<d$, $h\in\Lambda_0$, Folgen von Gitterpunkten $c, d\in\mathbb{R}_h^1$, $h\in\Lambda_0$, mit $c\to c_0$, $d\to d_0$ $(h\in\Lambda_0)$. Dann sei $H^0=l^2[c,d]$, $h\in\Lambda_0$.

Der Verschiebungsoperator $E:\Omega\to\Omega$ wird durch die Vorschrift

$$(Eu)(x)=u(x+h),\qquad x\in\mathbb{R}_h^1,\qquad u\in\Omega,\qquad h\in\Lambda_0,$$

definiert, wodurch wir eine umkehrbar eindeutige Abbildung von Ω auf sich erhalten. Ferner seien für jedes Paar nichtnegativer ganzer Zahlen k, l die orthogonalen Projektionen P_{kl} in H^0 erklärt durch

$$(P_{kl}u)(x)=\begin{cases}u(x), & x\in[c+kh,\ d-lh]_h\\ 0, & \text{sonst in } \mathbb{R}_h^1,\ u\in H^0,\ h\in\Lambda_0.\end{cases}$$

Die vorwärts genommenen Differenzenquotienten $D^j:\Omega\to\Omega$ sind durch

$$D^j=\left(\frac{E-I}{h}\right)^j,\qquad j=0,1,\ldots,$$

gegeben, und wir setzen dann

$$Z_j=P_{0j}D^j,\qquad j=0,1,\ldots,m.$$

Damit erhalten wir Operatoren $Z_j\in B(H^0)$, und mit der Setzung $H^j=H^0$, $j=1,\ldots,m$, erfüllen wir die Bedingung (H 2) aus §1. Der Kürze halber behalten wir die Bezeichnungen $\mathbb{H}^j$ und H^j für die vorstehend speziell gewählten Räume bei und schreiben ∂^j (bzw. D_j) für Z_j^0 (bzw. Z_j).

Für die vorwärts genommenen Differenzenquotienten beweist man leicht die folgende Leibnizsche Regel.

(1) *Für jede ganze Zahl* $n=0,1,\ldots,$ *gilt*

$$D^n(u\cdot v)=\sum_{j=0}^{n}\binom{n}{j}D^ju\cdot D^{n-j}E^jv,\qquad u,v\in\Omega.$$

2.3. In diesem Abschnitt geben wir die Definition und einige Beispiele für sog. Restriktionsoperatoren, die wir dazu verwenden, die Eigenschaft (H 3) nachzuweisen und in § 5 die Koeffizienten der Differentialoperatoren und den inhomogenen Term auf das Gitter zu restringieren.

(2) *Es sei* j *eine ganze Zahl mit* $0\leqq j\leqq m$ *und sei* $\mathbb{H}_d^j$ *ein in* $\mathbb{H}^j$ *dichter Teilraum. Eine für jedes* $h\in\Lambda_0$ *erklärte Folge von Abbildungen* $R:\mathbb{H}_d^j\to H^j$ *heißt ein Restriktionsoperator in* $\mathbb{H}^j$, *genau wenn für alle* $\varphi,\psi\in\mathbb{H}_d^j$ *gilt*

(3) $$(D_kR\varphi,\ D_lR\psi)_0\to((\partial^k\varphi,\ \partial^l\psi))_0\quad(h\in\Lambda_0),\qquad k,l=0,\ldots,j.$$

Da Funktionen aus $\mathbb{H}^0$ im allgemeinen nicht auf $[c,d]$ erklärt sind, führen wir einen Fortsetzungsoperator $\tau:\mathbb{H}^0\to L^2(\mathbb{R}^1)$ ein, von dem wir verlangen, ohne auf die genauen Vorschriften im einzelnen einzugehen, daß bei dem Fortsetzungsprozeß in $[c_0,d_0]$ etwa vorhandene Differenzierbarkeit, Stetigkeit, Beschränktheit oder Definitheit auf jedem der Intervalle $[c,d]$, $h\in\Lambda_0$, erhalten bleibt.

(4) Es sei $j=m$, $\mathbb{H}_d^m=\mathbb{H}=C^m[c_0, d_0]$. Dann definieren wir den Restriktionsoperator $R: \mathbb{H}\to H^m$ durch die Vorschrift

$$(R\varphi)(x)=\begin{cases}(\tau\varphi)(x), & x\in[c, d]_h\\ 0, & \text{sonst in } \mathbb{R}_h^1,\ \varphi\in\mathbb{H},\ h\in\Lambda_0.\end{cases}$$

Man überzeugt sich anhand der Definition von D_j leicht, daß die Bedingung (3) erfüllt ist, und damit ist auch (H3) nachgewiesen.

(5) Es sei $j=0$ sowie $\mathbb{H}_d^0=\mathbb{H}'$, die Menge der auf $[c_0, d_0]$ beschränkten, fast überall stetigen Funktionen, und es sei σ eine reelle Zahl. Dann erklären wir den Operator $R_1=\mathbb{H}'\to H^0$ durch

$$(R_1\varphi)(x)=\begin{cases}(\tau\varphi)(x+\sigma h), & x\in[c, d]_h\\ 0, & \text{sonst in } \mathbb{R}_h^1,\ \varphi\in\mathbb{H}',\ h\in\Lambda_0.\end{cases} \tag{6}$$

Bekanntlich liegt $\mathbb{H}'$ dicht in $\mathbb{H}^0$, und es konvergiert für $\varphi, \psi\in\mathbb{H}'$

$$h\sum_{x\in\mathbb{R}_h^1}(R_1\varphi)(x)(R_1\overline{\psi})(x)\to\int_{c_0}^{d_0}\varphi(x)\overline{\psi}(x)\,dx \quad (h\in\Lambda_0),$$

so daß R_1 ein Restriktionsoperator in $\mathbb{H}^0$ ist.

(7) Es sei $j=0$, σ eine reelle Zahl und $\mathbb{H}_d^0=\mathbb{H}''$ die Menge der auf $[c_0, d_0]$ stückweise stetigen Funktionen. Ist dann $Q_\sigma(x)=[x+\sigma h,\ x+(\sigma+1)h]$, so setzen wir für jedes $\varphi\in\mathbb{H}''$ und $h\in\Lambda_0$

$$\operatorname{Re}(R_2\varphi)(x)=h\Big(\int_{Q_\sigma(x)}[\operatorname{Re}(\tau\varphi)(t)]^{-1}dt\Big)^{-1}$$

für $x\in[c, d]_h$, falls $\operatorname{Re}\tau\varphi\neq 0$ und stetig in $Q_\sigma(x)$ ist, $\operatorname{Re}(R_2\varphi)(x)=1$ für $x\in[c, d]_h$, falls $\operatorname{Re}\tau\varphi$ in $Q_\sigma(x)$ nicht stetig ist und $\operatorname{Re}(R_2\varphi)(x)=0$ sonst in $\mathbb{R}_h^1$. Analog wird $\operatorname{Im}R_2\varphi$ definiert.

Man sieht leicht, daß $R_2\varphi$ gleichmäßig in h beschränkt ist und in jedem kompakten Intervall $G\leq[c_0, d_0]$, das keinen Unstetigkeitspunkt von φ enthält,

$$\max_{x\in G_h}|(\varphi-R_2\varphi)(x)|\to 0 \quad (h\in\Lambda_0), \qquad \varphi\in\mathbb{H}'',$$

gilt. Hieraus leitet man sofort das Bestehen von (3) her, und damit ist R_2 ein Restriktionsoperator in $\mathbb{H}^0$.

Im Vorgriff auf den nächsten Paragraphen bemerken wir, daß der adjungierte Operator J^* eines starken Einbettungsoperators $J: H^j\to\mathbb{H}^j$ einen Restriktionsoperator in $\mathbb{H}^j$ mit $\mathbb{H}_d^j=\mathbb{H}^j$ darstellt. Auf diese Weise gebildete Restriktionsoperatoren wollen wir mit R_3 bezeichnen.

3. Einbettungsoperatoren

3.1. In diesem Paragraphen untersuchen wir eine Klasse von Einbettungsoperatoren (s. [24]), die ein wichtiges Hilfsmittel zum Nachweis der Eigenschaft (H4) darstellen. Darüberhinaus werden durch Einbettungsoperatoren in $\{H^j\to\mathbb{H}^j\}$ konvergente Folgen in solche übergeführt, die in $\mathbb{H}^j$ konvergieren. Wir geben die folgende Definition.

(1) *Ist j eine ganze Zahl mit $0 \leqq j \leqq m$, so heißt eine für alle $h \in \Lambda_0$ erklärte Folge linearer Abbildungen $J: H^j \to \mathbb{H}^j$ ein Einbettungsoperator, genau wenn es eine Zahl μ_1 gibt, so daß die Ungleichung*

$$\|Ju\|_j \leqq \mu_1 |u|_j, \quad u \in H^j, \quad h \in \Lambda_0,$$

besteht. Der Einbettungsoperator J heißt regulär, genau wenn eine positive Zahl μ_0 existiert, so daß für fast alle $h \in \Lambda_0$ gilt

$$\|Ju\|_j \geqq \mu_0 |u|_j, \quad u \in H^j.$$

Den zu J adjungierten Operator J^* erklären wir als Abbildung von $\mathbb{H}^j \to H^j$ durch die Gleichung

$$((Ju, v))_j = (u, J^*v)_j, \quad u \in H^j, \quad v \in \mathbb{H}^j.$$

(2) *J heißt ein starker Einbettungsoperator, genau wenn für jedes $\varphi \in \mathbb{H}^j$*

$$\|JJ^*\varphi - \varphi\|_j \to 0 \quad (h \in \Lambda_0)$$

konvergiert.

Wir konstruieren nun eine Klasse von starken regulären Einbettungsoperatoren. Zu diesem Zweck studieren wir zunächst die Mittelfunktionen $\omega_h^{(j)}$, $j = 0, 1, \ldots$, die erklärt sind durch

$$\omega_h^{(0)}(x) = \begin{cases} h^{-1}, & x \in (0, h) \\ 0, & \text{sonst in } \mathbb{R}^1, \end{cases} \quad h \in \Lambda_0,$$

und

$$\omega_h^{(j)}(x) = \int_{\mathbb{R}^1} \omega_h^{(0)}(x - y)\, \omega_h^{(j-1)}(y)\, dy, \quad x \in \mathbb{R}^1, \quad j = 1, 2, \ldots.$$

Man überzeugt sich leicht, daß für jedes $j = 0, 1, \ldots$, die Funktionen $\omega_h^{(j)}$ im Innern ihres Trägers $[0, (j+1)h]$ positiv sind und ihr Integral über $\mathbb{R}^1$ gleich 1 ist. Die Fouriertransformierte $\tilde{\omega}_h^{(j)}$ von $\omega_h^{(j)}$ bestimmt man zu

$$\tilde{\omega}_h^{(j)}(y) = \sqrt{2\pi}\, \tilde{\omega}_h^{(0)}(y)\, \tilde{\omega}_h^{(j-1)}(y) = \frac{1}{\sqrt{2\pi}} \left(\frac{1 - e^{-ihy}}{ihy} \right)^{j+1} \tag{3}$$

für $y \in \mathbb{R}^1$ und $j = 0, 1, \ldots$. Offensichtlich ist $y^j \tilde{\omega}_h^{(j)} \in L^2(\mathbb{R}^1)$, und folglich besitzt $\omega_h^{(j)}$ Ableitungen im Sinne der Distributionstheorie, die bis zur j-ten Ordnung in $L^2(\mathbb{R}^1)$ liegen, so daß $\omega_h^{(j)}$ aus $\mathbb{H}^j$, $h \in \Lambda_0$, ist. Durch Fouriertransformation überzeugt man sich von der Beziehung

$$\partial^k \omega_h^{(j)} = D^k \omega_h^{(j-k)}(\cdot - kh), \quad k = 0, \ldots, j, \quad j = 0, 1, \ldots. \tag{4}$$

Eine weitere Eigenschaft der Funktionen $\omega_h^{(j)}$ gibt der Hilfssatz

(5) *Sei n eine nichtnegative ganze Zahl und seien α_j, $j = 0, \ldots, n$, komplexe Zahlen. Aus*

$$\sum_{j=0}^{n} \alpha_j \omega_1^{(n)}(x + j) = 0, \quad x \in (0, 1), \quad \textit{folgt} \quad \alpha_j = 0, \quad j = 0, \ldots, n.$$

Beweis. Wir führen den Beweis durch vollständige Induktion nach n. Für $n = 0$ ist die Behauptung richtig. Nehmen wir ihre Gültigkeit für eine ganze

Zahl $n \geqq 0$ an und sei

$$\sum_{j=0}^{n+1} \alpha_j \omega_1^{(n+1)}(x+j) = 0, \qquad x \in (0,1). \tag{6}$$

Hieraus folgt durch Differentiation nach x mit Hilfe von (3), wenn wir noch $\alpha_{-1} = \alpha_{n+2} = 0$ setzen,

$$\sum_{j=0}^{n+1} \alpha_j \left(\omega_1^{(n)}(x+j) - \omega_1^{(n)}(x+j-1)\right) = \sum_{j=-1}^{n+1} (\alpha_j - \alpha_{j+1})\, \omega_1^{(n)}(x+j) = 0$$

für $x \in (0,1)$. Nun ist $\omega_1^{(n)}(x+n+1) = \omega_1^{(n)}(x-1) = 0$, $x \in (0,1)$, und daher läßt sich die Induktionsvoraussetzung anwenden, die $\alpha_j - \alpha_{j+1} = 0$, $j = 0, \ldots, n$, ergibt. Aus (6) erhalten wir wegen der Positivität jedes Summanden auch $\alpha_0 = 0$, und das war noch zu zeigen.

Wir kommen nun zur Definition einer Folge von Einbettungsoperatoren von $H^k \to \mathbb{H}^k$, $k = 0, \ldots, m$, und setzen

$$\beta_l = \frac{d - c - (l-1)h}{d_0 - c_0}, \qquad \gamma_l = \beta_l d_0 - d - h, \qquad l = 0, 1, \ldots, h \in \Lambda_0. \tag{7}$$

Aus unseren Voraussetzungen $c \to c_0$, $d \to d_0$ $(h \in \Lambda_0)$ folgt sofort

$$\beta_l \to 1, \qquad \gamma_l \to 0 \quad (h \in \Lambda_0), \qquad l = 0, 1, \ldots. \tag{8}$$

Ist $u \in H^0$ und $\omega \in L^2(\mathbb{R}^1)$, so erklären wir die Funktion $u * \omega \in H^0$ durch

$$(u * \omega)(y) = \begin{cases} h \sum\limits_{x \in \mathbb{R}_h^1} u(x)\, \omega(\beta_l y - \gamma_l - x), & \beta_l > 0, \quad y \in [c_0, d_0] \\ 0, & \beta_l \leqq 0, \quad y \in [c_0, d_0], \end{cases}$$

und definieren die Einbettungsoperatoren $J_l \colon H^k \to \mathbb{H}^k$ durch die Vorschrift

$$J_l u = u * \omega_h^{(l)}, \qquad u \in H^k, \qquad h \in \Lambda_0, \qquad l = k, k+1, \ldots, \qquad k = 0, 1, \ldots, m. \tag{9}$$

Auf den Sonderfall $\beta_l \leqq 0$ werden wir im weiteren nicht mehr besonders eingehen, da für ihn die getroffenen Aussagen trivial sind.

(10) *Für jedes $l = k, k+1, \ldots, k = 0, 1, \ldots, m$, ist J_l ein Einbettungsoperator von $H^k \to \mathbb{H}^k$ mit $\overline{\lim\limits_{h \in \Lambda_0}} |J_l|_k \leqq 1$.*

Beweis. Wir verwenden die Beziehung (4) und erhalten für jedes $s = 0, 1, \ldots, k$, $l = k, k+1, \ldots, k = 0, 1, \ldots, m$

$$\partial^s J_l u = \beta_l^s u * (\partial^s \omega_h^{(l)}) = \beta_l^s (D^s u) * \omega_h^{(l-s)}(\cdot - s h).$$

Beachten wir Träger $(\omega_h^{(l-s)}) = [0, (l+1-s)h]$, so sehen wir, daß $\partial^s J_l u$ nur von $(D^s u)(x)$ für $x \in [c, d - s h]_h$ abhängt, und daher ist

$$\partial^s J_l u = \beta_l^s D_s u * \omega_h^{(l-s)}(\cdot - s h), \qquad u \in H^k. \tag{11}$$

Hieraus ergibt sich mit der Variablensubstitution $z = \beta_l y - \gamma_l - s h$

$$\|\partial^s J_l u\|_0^2 \leqq \beta_l^{2s-1} \int_{\mathbb{R}^1} \Big| h \sum_{x \in \mathbb{R}_h^1} (D_s u)(x)\, \omega_h^{(l-s)}(z - x) \Big|^2 dz.$$

Wir werden für alle $u \in H^0$ die Beziehung

$$\int_{\mathbb{R}^1} \left| h \sum_{x \in \mathbb{R}_h^1} u(x)\, \omega_h^{(j)}(y-x) \right|^2 dy \leqq |u|_0^2, \qquad j = 0, 1, \ldots, \tag{12}$$

zeigen, so daß wir

$$\|\partial^s J_l u\|_0^2 \leqq \beta_l^{2s-1} |D_s u|_0^2, \qquad u \in H^k,$$

erhalten, woraus nach Summation über s die Beschränktheit von J_l und mit (8) auch $\overline{\lim} |J_l|_k \leqq 1$ folgt. Es bleibt (12) zu beweisen. Für $j = 0$ läßt sich das Integral sofort ausrechnen, und wir erkennen, daß in diesem Fall in (12) sogar das Gleichheitszeichen steht. Den Fall $j > 0$ führen wir hieraus zurück. Wir haben nämlich gemäß (3) für die Fouriertransformierte von $\omega_h^{(j)}$

$$\tilde{\omega}_h^{(j)}(y) = \left(\frac{1 - e^{-ihy}}{ihy} \right)^j \tilde{\omega}_h^{(0)}(y), \qquad y \in \mathbb{R}^1, \qquad j = 0, 1, \ldots .$$

Nun ist der Faktor von $\tilde{\omega}_h^{(0)}$ höchstens vom Betrage 1, und wir erhalten daher die Behauptung durch Fouriertransformation des Integranden in (12).

(13) *Für jedes $l = k, k+1, \ldots$, $k = 0, 1, \ldots, m$ ist J_l ein regulärer Einbettungsoperator von $H^k \to \mathbb{H}^k$.*

Beweis. Aus (11) ergibt sich nach einer kleinen Rechnung mit $d - c = nh$

$$\|\partial^s J_l u\|_0^2 = \beta_l^{2s-1} h \sum_{r=l}^{n} \int_0^1 \left| \sum_{k=0}^{l-s} (D_s u)(c - (k+s-r)h)\, \omega_1^{(l-s)}(z+k) \right|^2 dz. \tag{14}$$

Mit (5) folgt nun nach einem bekannten Satz die Existenz einer Konstanten $K > 0$, so daß für alle komplexen Zahlen $\alpha_0, \alpha_1, \ldots, \alpha_{l-s}$

$$\int_0^1 \left| \sum_{k=0}^{l-s} \alpha_k \omega_1^{(l-s)}(z+k) \right|^2 dz \geqq K \sum_{k=0}^{l-s} |\alpha_k|^2$$

gilt. Verwenden wir dieses Ergebnis in (14), so erhalten wir für jedes $u \in H^k$, $s = 0, \ldots, k$, $l = k, k+1, \ldots$,

$$\|\partial^s J_l u\|_0^2 \geqq \beta_l^{2s-1} K h \sum_{j=0}^{n-s} |(D_s u)(c + jh)|^2,$$

woraus nach Summation über s die Regularität von J_l folgt.

Bevor wir nun beweisen, daß J_l auch ein starker Einbettungsoperator ist, geben wir eine Darstellung für $J_l^* : \mathbb{H}^0 \to H^0$, $l = 0, 1, \ldots$. Aus der Definitionsgleichung von J_l^* erhalten wir für $u \in H^0$, $\varphi \in \mathbb{H}^0$

$$h \sum_{x \in \mathbb{R}_h^1} u(x) \int_{c_0}^{d_0} \omega_h^{(l)}(\beta_l y - \gamma_l - x)\, \overline{\varphi}(y)\, dy = h \sum_{x \in \mathbb{R}_h^1} u(x)\, (J_l^* \overline{\varphi})(x),$$

woraus durch geeignete Wahl von u

$$(J_l^* \varphi)(x) = \begin{cases} \int_{c_0}^{d_0} \omega_h^{(l)}(\beta_l y - \gamma_l - x)\, \varphi(y)\, dy, & x \in [c, d]_h, \\ 0, \quad \text{sonst in } \mathbb{R}_h^1, & \varphi \in \mathbb{H}^0, \end{cases} \qquad l = 0, 1, \ldots, \tag{15}$$

folgt. Es ist nicht schwer zu sehen, daß

$$|J_l^* \varphi - R\varphi|_0 \to 0 \quad (h \in \Lambda_0), \qquad \varphi \in C_0^\infty(c_0, d_0), \tag{16}$$

gilt. Hiermit beweist man nun leicht den Hilfssatz

(17) *Für jedes* $l = k, k+1, \ldots, k = 0, 1, \ldots, m$ *konvergiert*

$$J_l R\varphi \rightharpoonup \varphi \quad \text{in} \quad \mathbb{H}^k \ (h \in \Lambda_0), \qquad \varphi \in \mathbb{H}.$$

Einige Folgerungen aus diesem Hilfssatz sind die Aussagen

(18) *Sei* k *eine ganze Zahl mit* $0 \leqq k \leqq m$ *und sei* $l \geqq k$. *Dann gilt für jedes* $\varphi \in \mathbb{H}$

$$|J_l^* \varphi - R\varphi|_k \to 0, \qquad \|J_l R\varphi - \varphi\|_k \to 0 \quad (h \in \Lambda_0).$$

Beweis. Wegen $\overline{\lim}\, |J_l|_k = \overline{\lim}\, \|J_l^*\|_k \leqq 1$ erhalten wir mit (2.3) und (17) für $\varphi \in \mathbb{H}$ und $(h \in \Lambda_0)$

$$\begin{aligned} |J_l^* \varphi - R\varphi|_k^2 &= |J_l^* \varphi|_k^2 + |R\varphi|_k^2 - 2\,\mathrm{Re}\,(J_l^* \varphi, R\varphi)_k \\ &\leqq \|J_l^*\|_k^2 \|\varphi\|_k^2 + |R\varphi|_k^2 - 2\,\mathrm{Re}\,((\varphi, J_l R\varphi))_k \to 0. \end{aligned}$$

Entsprechend erschließt man die Konvergenz von $\|J_l R\varphi - \varphi\|_k \to 0$.

(19) *Für jedes* $l = k, k+1, \ldots, k = 0, 1, \ldots, m$ *ist* J_l *ein starker regulärer Einbettungsoperator von* $H^k \to \mathbb{H}^k$ *mit* $|J_l|_k \to 1$ $(h \in \Lambda_0)$.

Beweis. Mit (18) erhalten wir für jedes $\varphi \in \mathbb{H}$

$$\|J_l J_l^* \varphi - \varphi\|_k \leqq |J_l|_k |J_l^* \varphi - R\varphi|_k + \|J_l R\varphi - \varphi\|_k \to 0 \quad (h \in \Lambda_0).$$

Da $\mathbb{H}$ in $\mathbb{H}^k$ dicht liegt und $J_l J_l^*$ beschränkt ist, haben wir hiermit bewiesen, daß J_l ein starker Einbettungsoperator ist. Somit gilt auch

$$|J_l^* \varphi|_k^2 = ((J_l J_l^* \varphi, \varphi))_k \to \|\varphi\|_k^2 \quad (h \in \Lambda_0), \qquad \varphi \in \mathbb{H}^k,$$

woraus $\underline{\lim}\, \|J_l^*\|_k \geqq 1$ folgt. Zusammen mit (10) sind damit die Behauptungen des Satzes bewiesen.

3.2. Die vorangehenden Ergebnisse gestatten nun, das Erfülltsein der Bedingung (H4) aus §1 nachzuweisen.

(20) *Seien* k, l *zwei ganze Zahlen mit* $0 \leqq k \leqq m$ *und* $k \leqq l$, *sei* $\{u\}$, $h \in \Lambda$, *eine Folge von Elementen* $u \in H^k$, $h \in \Lambda$, *und sei* $u_0 \in \mathbb{H}^k$. *Es konvergiert*

$$u \rightharpoonup u_0 \quad (\textit{bzw.}\ u \to u_0) \quad \text{in} \quad \{H^k \to \mathbb{H}^k\} \quad (h \in \Lambda),$$

genau wenn gilt

$$J_l u \rightharpoonup u_0 \quad (\textit{bzw.}\ J_l u \to u_0) \quad \text{in} \quad \mathbb{H}^k \ (h \in \Lambda).$$

Beweis. Mit Hilfe von (11) zeigt man unschwer, daß für jedes $s = 0, 1, \ldots, k$, $l = k, k+1, \ldots$, und für jede in H^k beschränkte Folge $\{u\}$, $h \in \Lambda$,

$$\partial^s J_l u - J_{l-s} D_s u \rightharpoonup 0 \quad \text{in} \quad \mathbb{H}^0 \quad (h \in \Lambda) \tag{21}$$

konvergiert.

Ferner ergibt sich aus der für jede in H^0 beschränkte Folge $\{u\}$, $h\in\Lambda$, und jedes $\varphi\in\mathbb{H}$ gültigen Beziehung

$$|(u, R\varphi)_0 - ((J_l u, \varphi))_0| \leqq |u|_0 \, |R\varphi - J_l^* \varphi|_0 \to 0 \quad (h\in\Lambda)$$

bei Verwendung der in (19) gewonnenen Ergebnisse die Aussage

$$u \rightharpoonup u_0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \Leftrightarrow J_l u \rightharpoonup u_0 \quad \text{in} \quad \mathbb{H}^0 \quad (h\in\Lambda).$$

Hieraus und aus (21) folgt die Behauptung des Satzes über die schwache Konvergenz.

Wir wenden uns nun dem Beweis der Behauptung über die starke Konvergenz zu und setzen $J_l u \to u_0$ in $\mathbb{H}^k$ voraus. Für jedes $\varphi\in\mathbb{H}$ gilt

$$|u - R\varphi|_k \leqq K(\|J_l u - u_0\|_k + \|u_0 - \varphi\|_k + \|\varphi - J_l R\varphi\|_k),$$

so daß sich aufgrund der Dichtheit von $\mathbb{H}$ in $\mathbb{H}^k$ und (18) zu jedem $\varepsilon > 0$ ein $\varphi\in\mathbb{H}$ finden läßt, so daß für fast alle $h\in\Lambda$ die Ungleichung $|u - R\varphi|_k < \varepsilon$ erfüllt ist, woraus

$$u \to u_0 \quad \text{in} \quad \{H^k \to \mathbb{H}^k\} \quad (h\in\Lambda) \tag{22}$$

folgt, wie wir in [12, 2.1] gezeigt haben. Setzen wir umgekehrt (22) als erfüllt voraus, so entnehmen wir der Konvergenz

$$|\|u_0\|_k - \|J_l u\|_k| \leqq |\|u_0\|_k - \|J_l J_l^* u_0\|_k| + |J_l|_k \, |u - J_l^* u_0|_k \to 0$$

für $(h\in\Lambda)$ zusammen mit der bereits bewiesenen schwachen Konvergenz der Folge $\{J_l u\}$, $h\in\Lambda$, in $\mathbb{H}^k$ die Behauptung des Satzes.

Mit Hilfe von (19) und (20) überzeugt man sich leicht davon, daß auch die Bedingung (H 4) erfüllt ist, wenn wir von der schwachen Kompaktheit beschränkter Folgen in Hilbertschen Räumen und der Kompaktheit der Einbettung von $\mathbb{H}^m$ in $\mathbb{H}^{m-1}$ Gebrauch machen.

3.3. Zum Abschluß dieses Paragraphen beweisen wir noch einige wichtige Eigenschaften diskret konvergenter Folgen.

(23) *Es sei* $\{y\}$, $h\in\Lambda$, *eine Folge von Punkten* $y\in\mathbb{R}_h^1$, *und sei der Operator* $T\in B(H^0)$ *durch*

$$(Tu)(x) = \begin{cases} u(y), & x = y, \\ 0, & \text{sonst in } \mathbb{R}_h^1, \end{cases} \quad h\in\Lambda, \; u\in H^0,$$

erklärt. Dann konvergiert $T \to 0$ $(h\in\Lambda)$.

Der Beweis dieses Satzes kann ebenso wie der des nächsten leicht mit Hilfe des Einbettungsoperators J_0 und (20) geschehen.

(24) *Es seien* k *und* l *nichtnegative ganze Zahlen und sei* $\{r\}$, $h\in\Lambda$, *eine Folge von Abbildungen von* $\mathbb{R}_h^1$ *in eine endliche Menge ganzer Zahlen. Ist dann der Operator* $T\in B(H^0)$ *erklärt durch*

$$(Tu)(x) = (P_{kl} E^{r(x)} u)(x), \quad x\in\mathbb{R}_h^1, \quad u\in H^0, \quad h\in\Lambda,$$

so konvergiert $T \to I$ $(h\in\Lambda)$.

(25) *Es existiert eine Konstante K, so daß für alle $k=1, \ldots, m$ gilt*

$$|(D_{k-1}u)(x)| \leqq K|u|_k, \quad u \in H^k, \quad x \in \mathbb{R}_h^1, \quad h \in \Lambda_0.$$

Beweis. Sei $1 \leqq k \leqq m$ gewählt. Die Behauptung ist trivial für $x \notin [c, d-(k-1)h]_h$. Zum Beweis für die verbleibenden x bemerken wir, daß mit einer Konstanten K die Ungleichung

$$\max_{y \in [c_0, d_0]} |(\partial^{k-1} J_k u)(y)| \leqq K \|J_k u\|_k, \quad u \in H^k, \quad h \in \Lambda,$$

besteht. Nun gilt gemäß (11) für $y \in [c_0, d_0]$ die Darstellung

$$(\partial^{k-1} J_k u)(y) = \beta_k^{k-1} h \sum_{x \in \mathbb{R}_h^1} (D_{k-1}u)(x)\, \omega_h^{(1)}(\beta_k y - \gamma_k - x - (k-1)h).$$

Zu gegebenem $x_0 \in [c, d-(k-1)h]_h$ wählen wir y_0 so, daß $\beta_k y_0 - \gamma_k - x_0 - kh = 0$ ist. Offenbar liegt y_0 in $[c_0, d_0]$. Unter Beachtung der Definition von $\omega_h^{(1)}$ erhalten wir

$$(\partial^{k-1} J_k u)(y_0) = \beta_k^{k-1} (D_{k-1}u)(x_0).$$

Daraus ergibt sich wegen $\beta_k \to 1$ $(h \in \Lambda)$ die Behauptung.

(26) *Sei $u_0 \in \mathbb{H}^m$ und sei $\{u\}$, $h \in \Lambda$, eine Folge von Elementen $u \in H^m$, $h \in \Lambda$, mit $u \rightharpoonup u_0$ in $\{H^m \to \mathbb{H}^m\}$ $(h \in \Lambda)$. Dann gilt für $j = 0, \ldots, m-1$*

$$\max_{x \in [c, d-jh]_h} |(D^j u - \partial^j u_0)(x)| \to 0 \quad (h \in \Lambda). \tag{27}$$

Beweis. Sei $1 \leqq k \leqq m$ gewählt. Dann folgt aus (20), daß die Folge $\{\partial^{k-1} J_l u\}$, $h \in \Lambda$, in $\mathbb{H}^1$ schwach gegen $\partial^{k-1} u_0$ konvergiert. Bekanntlich liegt $\partial^{k-1} u_0$ in $C^0[c_0, d_0]$, und es gilt

$$\max_{y \in [c_0, d_0]} |(\partial^{k-1} J_k u - \partial^{k-1} u_0)(y)| \to 0 \quad (h \in \Lambda).$$

Wählen wir y_0 wie im vorangehenden Beweis, so sehen wir, daß für $j = 0, \ldots, m-1$

$$\max_{x \in [c, d-jh]_h} |(D^j u)(x) - \partial^{k-1} u_0)(\beta_j^{-1}(x + \gamma_j + jh))| \to 0 \quad (h \in \Lambda)$$

konvergiert, und wegen $\gamma_j \to 0$, $\beta_j \to 1$ $(h \in \Lambda)$ folgt hieraus die Behauptung.

Ist die Funktion u_0 in (27) nicht von vornherein auf $[c, d]$, $h \in \Lambda$, erklärt, so hat man sie sich $(m-1)$-mal stetig differenzierbar in eine Umgebung von $[c_0, d_0]$ fortgesetzt zu denken.

4. Konvergenz der Differenzenapproximation

4.1. In diesem Paragraphen geben wir unseren Hauptsatz über die Konvergenz von Differenzenapproximationen für Rand- und Eigenwertprobleme gewöhnlicher Differentialgleichungen an.

Es seien a_j^0, $j = 0, \ldots, m$, und b_j^0, $j = 0, \ldots, m-1$, Funktionen aus $L^\infty(c_0, d_0)$ mit der Eigenschaft, daß eine Zahl $\alpha > 0$ mit $|a_m^0(x)| \geqq \alpha$, $x \in [c_0, d_0]$ existiert, und seien α_{jk}, β_{jk}, $j = 1, \ldots, m$, $k = 0, \ldots, m-1$, komplexe Zahlen. Dann setzen wir

$$L_0 = \sum_{k=0}^{m} a_k^0 \partial^k, \qquad M_0 = \sum_{k=0}^{m-1} b_k^0 \partial^k$$

und

$$l_0^{(j)} = \sum_{k=0}^{m-1} (-1)^k (\alpha_{jk}\, \delta_1^{(k)} + \beta_{jk}\, \delta_2^{(k)}), \qquad j = 1, \ldots, m,$$

wobei $\delta_1^{(k)}$ (bzw. $\delta_2^{(k)}$), $k = 0, \ldots, m-1$, die Diracschen δ-Funktionale und ihre Ableitungen bis zur $(m-1)$-ten Ordnung an der Stelle c_0 (bzw. d_0) sind. Das im ersten Paragraphen definierte Problem $\pi_\lambda^0(L_0, f_0, l_0^{(j)}, \gamma_0^{(j)})$ lautet dann bei dieser Spezialisierung

(1) *Gegeben sind eine Funktion* $f_0 \in L^2(c_0, d_0)$ *und komplexe Zahlen* $\gamma_0^{(j)}$, $j = 1, \ldots, m$. *Gesucht ist eine Lösung* $u_0 \in W^{m,2}(c_0, d_0)$ *der Differentialgleichung*

$$(L_0 - \lambda M_0)\, u_0 = f_0$$

unter den Randbedingungen

$$\sum_{k=0}^{m-1} [\alpha_{jk} (\partial^k u_0)(c_0) + \beta_{jk} (\partial^k u_0)(d_0)] = \gamma_0^{(j)}, \qquad j = 1, \ldots, m.$$

Die Begriffe Eigenwert, Hauptvektor, Spektrum und Resolventenmenge von (1) haben wir in §1 definiert. Wir setzen voraus, daß nicht jede Zahl $\lambda \in \mathbb{C}$ ein Eigenwert von (1) ist.

Als nächstes geben wir eine Klasse von Differenzenapproximationen für das Problem (1) an, für die wir mit unseren Methoden die Konvergenz des Rand- und Eigenwertproblems beweisen können. Es seien a_k, b_j Operatoren aus $B(H^0)$, $\Delta_L^{(k)}$, $\Delta_M^{(j)}$ seien lineare Abbildungen von $H^m \to H^0$, $k = 0, \ldots, m$, $j = 0, \ldots, m-1$, und m_1 und m_2 seien zwei nichtnegative ganze Zahlen mit $m_1 + m_2 = m$. Wir setzen

$$L = \sum_{k=0}^{m} a_k \Delta_L^{(k)}, \qquad M = \sum_{k=0}^{m-1} b_k \Delta_M^{(k)} \tag{2}$$

und nehmen an, daß $R(L) \subseteq R(P_{m_1 m_2})$ und $R(M) \subseteq R(P_{m_1 m_2})$ gilt, wobei $P_{m_1 m_2}$ die in § 2 definierte orthogonale Projektion ist. Seien weiter $r_l^{(j)}$, $j = 1, \ldots, m$, $l = 1, 2$, für jedes $h \in \Lambda_0$ erklärte Polynome in der Unbestimmten z mit Koeffizienten aus $\mathbb{C}$. Wir setzen für $j = 1, \ldots, m$

$$l^{(j)}(u) = \left(r_1^{(j)}(E)\, u\right)(c) + \left(r_2^{(j)}(E^{-1})\, u\right)(d), \qquad u \in H^m,\ h \in \Lambda_0, \tag{3}$$

und erhalten die folgende Schar von Differenzenapproximationen des Problems (1).

(4) *Für jedes* $h \in \Lambda_0$ *sind eine Funktion* $f \in R(P_{m_1 m_2})$ *und komplexe Zahlen* λ, $\gamma^{(j)}$, $j = 1, \ldots, m$, *gegeben. Gesucht ist eine Lösung* $u \in H^m$ *der Gleichungen*

$$(L - \lambda M)\, u = f, \qquad l^{(j)}(u) = \gamma^{(j)}, \qquad j = 1, \ldots, m,\ h \in \Lambda_0.$$

Eigenvektoren und Hauptvektoren von (4) sind analog zu (1) definiert. Wie man sich leicht überzeugt, liegt $\lambda \in \mathbb{C}$ in der Resolventenmenge $\varrho(L)$, genau wenn (4) für jedes $f \in R(P_{m_1 m_2})$ und jede Wahl von $\gamma^{(j)}$, $j = 1, \ldots, m$, lösbar ist, oder gleichbedeutend, wenn λ kein Eigenwert von (4) ist.

Wir stellen nun unsere Voraussetzungen über die Differenzenapproximation (4) zusammen und beginnen dabei mit den Randbedingungen. Hier verlangen wir nur, daß $l_0^{(j)}$ durch $l^{(j)}$, $j = 1, \ldots, m$, konsistent approximiert wird, d.h. es sei die folgende Bedingung erfüllt.

(Z1) *Der Grad der Polynome* $r_l^{(j)}$, $h\in\Lambda_0$, $j=1,\ldots,m$, $l=1,2$, *sei beschränkt und für jedes* $\varphi\in C^{m-1}(\mathbb{R}^1)$ *und* $j=1,\ldots,m$, *gelte für* $(h\in\Lambda_0)$

$$(5)\qquad \left(r_1^{(j)}(E)\,\varphi\right)(c)\to\sum_{k=0}^{m-1}\alpha_{jk}(\partial^k\varphi)(c_0),\qquad \left(r_2^{(j)}(E^{-1})\,\varphi\right)(d)\to\sum_{k=0}^{m-1}\beta_{jk}(\partial^k\varphi)(d_0).$$

Auch die Approximation $\Delta^{(j)}$ des Differentialquotienten j-ter Ordnung, $j=0,\ldots,m-1$, unterliegt im wesentlichen nur der Konsistenzforderung. Es seien r_j, s_j nichtnegative ganze Zahlen und $G_h^{(j)}=[c+r_j h,\, d-s_j h]_h$, $j=0,\ldots,m$. $\mathscr{K}$ sei eine endliche Menge ganzer Zahlen, und es sei $g_{jk}\in\Omega$, $h\in\Lambda_0$, $j=0,\ldots,m$, $k\in\mathscr{K}$. Dann setzen wir

$$\Delta^{(j)}=P_{r_j s_j}\sum_{k\in\mathscr{K}}g_{jk}E^k,\qquad h\in\Lambda_0,\ j=0,\ldots,m-1,$$

und verlangen, daß die folgende Voraussetzung erfüllt ist.

(Z2) *Für jedes* $k\in\mathscr{K}$ *und* $j=0,\ldots,m-1$ *sei* $g_{jk}(x,h)=0$, $x\in G_h^{(j)}$, $x+kh\notin[c,d]_h$, $h\in\Lambda_0$, *und für alle* $\varphi\in C^{m-1}(\mathbb{R}^1)$ *konvergiere*

$$\max_{x\in G_h^{(j)}}\left|\left(\Delta^{(j)}\varphi-\partial^j\varphi\right)(x)\right|\to 0\qquad (h\in\Lambda_0).$$

Die Differentialquotienten in L_0 und M_0 können voneinander verschieden approximiert werden, was wir durch Indizes „L" und „M" gekennzeichnet haben.

Die Approximation der m-ten Ableitung muß nicht nur konsistent sondern auch stabil sein. Um den Begriff der Stabilität einführen zu können, benötigen wir die folgende Definition.

(6) *Es sei* $k\geqq 0$ *eine ganze Zahl, und* s, s_j, $j=1,\ldots,k$, *seien Polynome in der Unbestimmten* z *mit Koeffizienten aus* $\mathbb{C}$. *Die Polynome* s_j, $j=1,\ldots,k$, *spannen* s *auf, genau wenn zu jedem Polynom* t *ein Polynom* t_1 *und Zahlen* $\mu_j\in\mathbb{C}$, $j=1,\ldots,k$, *existieren, so daß gilt*

$$t=t_1 s+\sum_{j=1}^{k}\mu_j s_j.$$

Der Fall $k=0$ soll bedeuten, daß keine Polynome s_j gegeben sind. Ist $s\neq 0$, so bezeichnen wir mit s_i das Polynom, das genau die im Innern des Einheitskreises gelegenen Wurzeln von s und den Koeffizienten 1 vor der höchsten Potenz besitzt.

Es seien dann n_1, n_2 nichtnegative ganze Zahlen, $p\neq 0$ sei ein Polynom vom Grade $n\leqq n_1+n_2$ und q_{lj}, $j=1,\ldots,n_l$, $l=1,2$, seien weitere Polynome. Wir führen zur Abkürzung die Größen $c_1=c+m_1 h$, $d_1=d-m_2 h$ ein und erklären den Operator $q\in B(H^0)$ durch die Vorschrift

$$(qu)(x)=\begin{cases}\left(q_{1j}(E)\,u\right)(c), & x=c_1+(j-1)h,\ j=1,\ldots,n_1\\ \left(q_{2j}(E^{-1})\,u\right)(d-mh), & x=d_1-(j-1)h,\ j=1,\ldots,n_2\\ \left(E^{-n_1-m_1}p(E)\,u\right)(x), & x\in[c_1+n_1h,\ d_1-n_2h]_h\\ 0, & \text{sonst in } \mathbb{R}_h^1,\ u\in H^0,\ h\in\Lambda_0.\end{cases}$$

Dann setzen wir

$$\Delta^{(m)}=qD_m+P_{r_m s_m}\sum_{k\in\mathscr{K}}g_{mk}E^k,\qquad h\in\Lambda_0,$$

und machen die folgende Konsistenz- und Stabilitätsvoraussetzung.

(Z3) (i) *Es ist* $p(1)=1$ *und für alle* $\varphi\in C^{m-1}(\mathbb{R}^1)$ *konvergiert*

$$\max_{x\in G_h^{(m)}}\Big|\sum_{k\in\mathcal{K}} g_{mk}(x,h)\,(E^k\varphi)(x)\Big|\to 0 \quad (h\in\Lambda_0).$$

(ii) *Es ist* $p(e^{i\xi})\neq 0$, $\xi\in\mathbb{R}^1$, *und die Polynome* $q_{1j}(z)$, $j=1,\ldots,n_1$ *(bzw.* $q_{2j}(z)$, $j=1,\ldots,n_2$*) spannen das Polynom* $p_i(z)$ *(bzw.* $[z^{n_1+n_2}p(z^{-1})]_i$*) auf.*

Ein Beispiel für eine Approximation des Differentialquotienten zweiter Ordnung in der obigen Gestalt (s. [6, S. 163]) ist

$$(\Delta^{(2)}u)_j=\begin{cases}\frac{1}{12h^2}(-u_{j-2}+16u_{j-1}-30u_j+16u_{j+1}-u_{j+2}), & j=1,\ldots,r-1\\ \frac{1}{h^2}(u_{j-1}-2u_j+u_{j+1}), & j=1,r\\ 0, \quad \text{sonst in } \mathbb{R}_h^1, & u\in H^2,\ h\in\Lambda_0,\end{cases}$$

wobei wir $u_j=u(c_0+jh)$, $j=0,\pm 1,\ldots$, $c=c_0-h$, $d=c_0+(r+1)h$ und $m_1=m_2=n_1=n_2=1$, $q_{11}=q_{21}=1$, $p=(-z^2+14z-1)/12$ gesetzt haben. Man prüft leicht nach, daß bei dieser Wahl alle geforderten Bedingungen erfüllt sind. Wie dieses Beispiel zeigt, werden durch $\Delta^{(m)}$ auch Differenzenapproximationen höherer Genauigkeit erfaßt. Dabei wird davon Gebrauch gemacht, daß sich die üblicherweise verwendeten Differenzenquotienten höherer Annäherung (s. z.B. [6, S. 538]) für den j-ten Differentialquotienten in der Gestalt $E^r p(E)D^j$ schreiben lassen mit einem gewissen Polynom p, das $p(1)=1$ erfüllt, und einer ganzen Zahl r. Wie zu erwarten, sind bei der Ableitung höchster Ordnung keine Approximationen zugelassen, die sich analog dem zentralen Differenzenquotienten 1. Ordnung $(2h)^{-1}(E-E^{-1})=2^{-1}E^{-1}(E+I)D^1$ verhalten, da bei diesen die Elliptizitätsbedingung $p(e^{i\xi})\neq 0$, $\xi\in\mathbb{R}^1$, verletzt ist. Wir bemerken, daß sich mit unseren Methoden auch noch eine etwas allgemeinere Gestalt von $\Delta^{(m)}$ behandeln läßt, bei der die Polynome q_{lj} und p von x und h abhängige Koeffizienten besitzen, (s. [13]), jedoch verzichten wir darauf zugunsten einer größeren Übersichtlichkeit.

Die Koeffizienten der Operatoren L und M werden wie folgt gebildet. Es seien a_m, a_k' und b_k', $k=0,\ldots,m-1$, Funktionen aus Ω, $h\in\Lambda_0$. Wir behalten für den Operator der Multiplikation mit einer Funktion aus Ω deren Namen bei und machen die folgende Voraussetzung.

(Z4) *Es existieren Konstanten* K_1 *und* K_2, $K_2>0$, *so daß für jedes* $h\in\Lambda_0$ *sowie* $k=0,\ldots,m-1$

$$K_2\leqq|a_m(x)|\leqq K_1,\quad |a_k'(x)|\leqq K_1,\quad |b_k'(x)|\leqq K_1,\qquad x\in[c,d]_h,$$

gilt und

$$J_0a_m\to a_m^0,\quad J_0a_k'\to a_k^0,\quad J_0b_k'\to b_k^0\qquad (h\in\Lambda_0)$$

dem Maße nach konvergiert. Es seien dann Q_{lk} *Operatoren von* $H^0\to R(P_{m_1m_2})$, $h\in\Lambda_0$, *mit der Eigenschaft*

$$Q_{lk}\to I\quad (h\in\Lambda_0),\qquad l=1,2,\ k=0,\ldots,m-1,$$

und es sei

$$a_k=Q_{1k}a_k',\qquad b_k=Q_{2k}b_k',\qquad k=0,\ldots,m-1,\ h\in\Lambda_0.$$

Zur Konstruktion von Koeffizienten mit den gewünschten Eigenschaften, kann man die in § 3.1 eingeführten Restriktionsoperatoren $R_1: \mathbb{H}' \to H^0$, $R_2: \mathbb{H}'' \to H^0$ und $R_3: \mathbb{H}^0 \to H^0$ verwenden. Sei zunächst $0 \leqq k \leqq m-1$. Je nachdem in welchem der drei vorangehend aufgeführten Räumen a_k^0 liegt, setze man $l = 1, 2$ oder 3 und $a_k' = R_l a_k^0$, $h \in \Lambda_0$. In gleicher Weise kann b_k' erklärt werden. Beispiele für Operatoren Q_{lk} mit den gewünschten Eigenschaften erhält man in der Gestalt

$$Q_{lk} = P_{m_1 m_2} \sum_{j=-\infty}^{\infty} q_j E^j, \qquad \sum_{j=-\infty}^{\infty} q_j = 1, \tag{7}$$

mit Zahlen $q_j \in \mathbb{C}$, von denen nur endlich viele nicht verschwinden. Liegt a_m^0 in $\mathbb{H}'$ (bzw. $\mathbb{H}''$), so kann man $a_m = R_1 a_m^0$ (bzw. $a_m = R_2 a_m^0$) setzen. Im Falle $a_m^0 \in \mathbb{H}^0$ wähle man eine reelle Zahl β mit $0 < \beta < \alpha$, eine ganze Zahl $k \geqq 0$ und definiere die Funktion a_m durch

$$a_m(x) = \begin{cases} (J_k^* a_m^0)(x), & |(J_k^* a_m^0)(x)| \geqq \beta, \quad x \in [c, d]_h \\ \beta, & \text{sonst in } \mathbb{R}_h^1, \quad h \in \Lambda_0. \end{cases} \tag{8}$$

Im allgemeinen wird man es jedoch vorziehen, die Differentialgleichung so umzuformen, daß $a_m^0 = 1$ wird. Im nächsten Paragraphen wird im Anschluß an Lemma (5.5) gezeigt, daß die vorstehenden Beispiele der Bedingung (Z 4) genügen.

Von der Funktion f und den Zahlen $\gamma^{(j)}$, $j = 1, \ldots, m$, verlangen wir schließlich das Erfülltsein der folgenden Bedingung.

(Z 5) *Es konvergiert* $J_0 f \to f_0$ *in* $\mathbb{H}^0$ *und* $\gamma^{(j)} \to \gamma_0^{(j)}$ $(h \in \Lambda_0)$, $j = 1, \ldots, m$.

Ähnlich wie bei den Koeffizienten von L und M genügt je nach Regularität von f_0 die Funktion

$$f = P_{m_1 m_2}(R_l f_0 + f'), \qquad h \in \Lambda_0, \tag{9}$$

mit $l = 1, 2$ oder 3, dieser Voraussetzung, wobei $f' \in H^0$ die Beziehung $|f'|_0 \to 0$ $(h \in \Lambda_0)$ erfüllt.

Nach diesen Vorbereitungen können wir einige der wichtigsten Ergebnisse dieser Arbeit in dem folgenden **Satz** aussprechen.

(10) *Es seien die vorangehend genannten Voraussetzungen* (Z 1)—(Z 5) *erfüllt. Dann liegt jede kompakte Teilmenge von* $\mathbb{C}$, *die keine Eigenwerte von* (1) *enthält, in* $\varrho(L_0)$ *und für fast alle* $h \in \Lambda_0$ *auch in* $\varrho(L)$. *Sind* $\{u\}$, $h \in \Lambda$, *die dann existierenden Lösungen von* (4) *und ist* u_0 *die Lösung von* (1), *so konvergiert*

$$u \to u_0 \quad \text{in} \quad \{H^m \to \mathbb{H}^m\} \qquad (h \in \Lambda)$$

und für $j = 0, \ldots, m-1$

$$\max\{|(D^j u - \partial^j u_0)(x)| \mid x \in [c, d - jh]_h \cap [c_0, d_0]\} \to 0 \quad (h \in \Lambda).$$

Die Eigenwerte von (1) *sind von endlicher algebraischer Vielfachheit und besitzen keinen im Endlichen gelegenen Häufungspunkt. Ist* λ_0 *ein Eigenwert von* (1) *und* U *eine Umgebung von* λ_0, *in deren Abschließung keine weiteren Eigenwerte von* (1) *liegen, so ist für fast alle* $h \in \Lambda_0$ *die Summe der algebraischen Vielfachheiten der in* U *gelegenen Eigenwerte von* (4) *gleich der algebraischen Vielfachheit von* λ_0, *d.h. die Eigenwerte von* (4) *konvergieren der algebraischen Vielfachheit nach gegen die Eigenwerte von* (1).

Der vorstehende Satz beantwortet die Frage nach der Konvergenz von Differenzenapproximationen des allgemeinen Rand- und Eigenwertproblems gewöhnlicher Differentialgleichungen in großer Allgemeinheit. Über die Differentialgleichung wird nur vorausgesetzt, daß sie regulär ist, d.h. daß der Koeffizient vor dem Differentialquotienten m-ter Ordnung die Beziehung $|a_m^0(x)| \geqq \alpha > 0$, $x \in [a,b]$, erfüllt, und daß die charakteristische Determinante von $L_0 - \lambda M_0$ nicht für alle $\lambda \in \mathbb{C}$ identisch verschwindet, womit nicht die gesamte komplexe Zahlenebene zum Spektrum gehört. Der Operator M_0 beim Eigenwertparameter ist ein beliebiger Differentialoperator der Ordnung kleiner m. Symmetrie oder Definitheit werden nicht benötigt, wenn sich auch in diesem Fall die etwas weitergehenden Aussagen in § 7.2 machen lassen. Mit den Voraussetzungen (Z1)—(Z5) fordern wir die Konsistenz der Differenzenapproximation, wobei in (Z3) zusätzlich eine Stabilitätsbedingung für die Approximation des Differentialquotienten m-ter Ordnung verlangt wird. Die letztere ist bei den Approximationen einfacher Genauigkeit stets erfüllt und bedeutet nur für die höheren Annäherungen eine über die Konsistenz hinausgehende Einschränkung, für die sich ein Beispiel im Anschluß an (Z3) findet. Wie sich aus dem vorangehend Gesagten bereits ergibt, existieren natürlich zu einem gegebenen Rand- und Eigenwertproblem stets beliebig viele Approximationen, die den Voraussetzungen (Z1)—(Z5) genügen.

5. Beweis des Hauptsatzes

5.1. Bevor wir zum Beweis von (4.10) kommen, stellen wir den folgenden Hilfssatz bereit, der auch den Charakter der Konsistenzbedingungen in (Z1)—(Z3) erhellt.

(1) *Es sei n eine nichtnegative ganze Zahl, es sei $\alpha_k^0 \in C^0(\mathbb{R}^1)$, $k = 0, \ldots, n$, r und s seien Funktionen aus Ω, $h \in \Lambda$, mit Werten in einer endlichen Menge nichtnegativer ganzer Zahlen und $g_k \in \Omega$, $k = 0, \pm 1, \ldots$, seien Funktionen mit der Eigenschaft*

$$g_k(x,h) = 0, \quad k < -r(x,h), \quad k > s(x,h), \quad x \in \mathbb{R}_h^1, \quad h \in \Lambda.$$

Ist dann

$$Q_0 = \sum_{k=0}^{n} \alpha_k^0 \partial^k, \quad Q = \sum_{k=-\infty}^{\infty} g_k E^k, \quad h \in \Lambda,$$

so existieren Funktionen $\alpha_j, \gamma_k \in \Omega$, $j = 0, \ldots, n$, $k = 1, 2, \ldots$, mit

$$\gamma_k(x,h) = 0, \quad k > (r+s)(x,h) - n, \quad x \in \mathbb{R}_h^1, \quad h \in \Lambda,$$

so daß Q die Darstellung

$$(2) \qquad Q = \sum_{k=1}^{\infty} \gamma_k (E-I)^k E^{-r} D^n + \sum_{k=0}^{n} \alpha_k E^{-r} D^k$$

besitzt. Es seien $\{a\}$, $\{b\}$, $h \in \Lambda$, beschränkte Folgen reeller Zahlen mit $a \leqq b$, $h \in \Lambda$. Dann konvergiert für jedes $\varphi \in C^n(\mathbb{R}^1)$

$$\max_{x \in [a,b]_h} |(Q\varphi - Q_0\varphi)(x)| \to 0 \quad (h \in \Lambda),$$

genau wenn es eine Konstante K gibt, so daß

$$|\gamma_k(x,h)| \leqq K, \quad x \in [a,b]_h, \quad h \in \Lambda, \quad k = 1, 2, \ldots,$$

und für $k=0, \ldots, n$

$$\max_{x\in[a,b]_h} |\alpha_k(x,h)-\alpha_k^0(x)| \to 0 \quad (h\in\Lambda)$$

gilt.

Beweis. Durch Entwicklung von QE^r nach Potenzen von $(E-I)$ sieht man, daß Q die gewünschte Darstellung (2) besitzt. Da für jedes $\varphi\in C^n(\mathbb{R}^1)$ und gleichmäßig für jede beschränkte Menge ganzer Zahlen k

$$\max_{x\in[a,b]_h} |(E^k D^j\varphi - \partial^j\varphi)(x)| \to 0 \quad (h\in\Lambda), \qquad j=0,\ldots,n,$$

gilt, sind die angegebenen Bedingungen offensichtlich hinreichend.

Wenn wir andererseits $\varphi = x^k$, $k=0,\ldots,n$, setzen, erhalten wir die behauptete Konvergenz von $\alpha_k(x,h)$ gegen $\alpha_k^0(x)$, $k=0,\ldots,n$.

Es bleibt die Existenz der Konstanten K zu zeigen. Wir nehmen im Widerspruch zur Behauptung an, daß eine Zahl k_0, eine Teilfolge $\{h_j\}$, $j=1,2,\ldots$, von Λ und eine Folge von Punkten $\{x_j\}$, $x_j\in[a,b]_h$, $j=1,2,\ldots$, existiert mit $|\gamma_{k_0}(x_j,h_j)| \to \infty$, $(j\to\infty)$. Wir können gegebenenfalls nach Übergang zu einer Teilfolge annehmen, daß die Folge $\{x_j\}$ gegen eine Zahl x_0 konvergiert, $r(x_j,h_j)$ und $s(x_j,h_j)$ konstant und $r+s>n$ ist und für $j=1,2,\ldots$,

$$|x_{j+1}-x_0| + (r+s+n+1)h_{j+1} < \tfrac{1}{4}h_j$$

gilt. Sei $\{\varrho_j\}$, $j=1,2,\ldots$, eine Folge komplexer Zahlen mit $\varrho_j\to 0$ und $\varrho_j\gamma_{k_0}(x_j,h_j)\to\infty$ $(j\to\infty)$. Wir werden zeigen, daß sich eine Funktion $\varphi\in C^n(\mathbb{R}^1)$ angeben läßt, die für jedes $j=1,2,\ldots$,

$$(3)\qquad \big((E-I)^k E^{-r} D^n\varphi\big)(x_j) = \begin{cases} 0, & k=1,\ldots,r+s-n,\ k\neq k_0 \\ \varrho_j, & k=k_0 \end{cases}$$

erfüllt. Dies führt uns nach Einsetzen in (2) auf einen Widerspruch zu

$$(Q\varphi)(x_j)\to(Q_0\varphi)(x_0) \quad (j\to\infty).$$

Sei $x_{jk}=x_j+kh_j$, $k=0,\pm1,\ldots,j=1,2,\ldots$. Die Beziehungen (3) stellen für jedes j ein unterbestimmtes Gleichungssystem in den Unbekannten $(D^n\varphi)(x_{jk})$, $k=-r,\ldots,s-n$, dar, das sich bei beliebiger Vorgabe einer der Unbekannten eindeutig lösen läßt. Für jedes $j=1,2,\ldots$, wählen wir $n+1$ aufeinanderfolgende Punkte x_{jk}, $k=l,\ l+1,\ldots,l+n$, $l\geqq -r$, $l+n\leqq s$, unter denen sich der (oder einer) x_0 am nächsten gelegene Punkt aus $\{x_{jk}\}$, $k=-r,\ldots,s$, befinden soll. Wir setzen $\varphi(x_{jk})=0$, $k=l,l+1,\ldots,l+n$, so daß $(D^n\varphi)(x_{jl})=0$ wird. Die Lösungen $(D^n\varphi)(x_{jk})$ verhalten sich dann wie $O(\varrho_j)=o(1)$ $(j\to\infty)$, und man berechnet für die Funktionwerte

$$(4)\qquad \varphi(x_{jk}) = o(h_j^n) \quad (j\to\infty), \qquad k=-r,\ldots,s.$$

Sei $\psi\in C^n(\mathbb{R}^1)$, $\psi(0)=1$ und $\psi(x)=0$, $|x|>1/4$. Dann setzen wir

$$\varphi(x)=\sum_{j=1}^{\infty}\sum_{k=-r}^{s}\varphi(x_{jk})\,\psi\big(h_j^{-1}(x-x_{jk})\big), \qquad x\in\mathbb{R}^1,$$

womit wir eine Funktion φ erhalten, die die vorgegebenen Werte interpoliert. Außerhalb des Punktes x_0 ist φ offensichtlich m-mal stetig differenzierbar. Dasselbe ist aber wegen (4) auch für x_0 der Fall.

5.2. Der Beweis von Satz (4.10) wird mit Hilfe der anschließenden Hilfssätze erbracht.

(5) *Unter der Bedingung* (Z 4) *konvergiert für* $j=0, \ldots, m$

$$a_j \to a_j^0, \quad (a_m)^* \to (a_m^0)^* \quad (h \in \Lambda_0).$$

Beweis. Wegen $(a_m)^* = \bar{a}_m$ und $(a_m^0)^* = \overline{a_m^0}$ erhält man die zweite der Behauptungen bereits aus der ersten, so daß es genügt, diese zu beweisen. Zu diesem Zwecke zeigen wir, daß für eine Folge $\{a\}$, $h \in \Lambda_0$, von Funktionen $a \in \Omega$ aus $|a(x)| \leqq K$, $x \in [c, d]_h$ und der Maßkonvergenz

(6) $$J_0 a \to a^0 \quad (h \in \Lambda_0)$$

folgt $a \to a^0$ $(h \in \Lambda_0)$. Es sei $\{u\}$, $h \in \Lambda_0$, eine Folge von Elementen $u \in H^0$, $h \in \Lambda_0$, mit

(7) $$u \to u_0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda_0).$$

Aufgrund der Beziehung

$$(J_0 a\, u)(x) = (J_0 a)(x)\,(J_0 u)(x), \qquad x \in [c_0, d_0],$$

hat man die Abschätzung

$$\|J_0 a\, u - a^0 u_0\| \leqq K_1 \|J_0 u - u_0\| + \|(J_0 a - a^0) u_0\|, \qquad h \in \Lambda_0,$$

aus der man mit dem Satz von Lebesgue das gewünschte Ergebnis erhält.

Wir beweisen noch, daß (Z 4) für die darauffolgend genannten Beispiele erfüllt ist. Im Falle $a = J_k^* a^0$ ergibt sich aus (3.15)

$$|a(x)| \leqq \operatorname*{ess\,sup}_{x \in [c_0, d_0]} |a^0(x)| \int_{c_0}^{d_0} \omega_k^{(k)}(\beta_k y - \gamma_k - x)\, dy,$$

woraus man die gleichmäßige Beschränktheit von a, $h \in \Lambda_0$, erhält. Da $J_k J_k^* \to I$ $(h \in \Lambda_0)$ konvergiert, folgt aus (3.20)

$$J_0 a = J_0 J_k^* a^0 \to a^0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda_0)$$

und damit auch die Maßkonvergenz (6). In den beiden anderen Fällen ist die gleichmäßige Beschränktheit der Funktionen a, $h \in \Lambda_0$, leicht zu sehen. Die Konvergenz (6) erfolgt hier sogar fast überall.

Daß Operatoren Q_{1k} der Gestalt (4.7) stark gegen die Identität konvergieren, ist eine unmittelbare Folgerung aus Satz (3.24). Es bleibt der Operator aus (4.8) zu untersuchen. In den vorangehenden Überlegungen haben wir bereits gezeigt, daß die Folge $\{J_0 J_k^* a_m^0\}$, $h \in \Lambda_0$, in $\mathbb{H}^0$ gegen a_m^0 konvergiert, was

$$\mu[\{x \in [c_0, d_0] \mid |(J_0 J_k^* a_m^0 - a_m^0)(x)| \geqq \alpha - \beta\}] \to 0 \quad (h \in \Lambda_0)$$

ergibt, wobei μ das Lebesguesche Maß ist. Für $|(J_0 J_k^* a_m^0 - a_m^0)(x)| < \alpha - \beta$ gilt

$$|(J_0 J_k^* a_m^0)(x)| \geqq |a_m^0(x)| - (\alpha - \beta) \geqq \beta,$$

und wir erhalten $a_m(x) = (J_k^* a_m^0)(x)$. Hieraus ergibt sich dann zusammen mit dem vorher Bewiesenne, daß a_m der Bedingung (Z4) genügt.

(8) *Die Folge von Operatoren* $\{a_m q\}$, $h \in \Lambda_0$, *ist unter den Bedingungen* (Z3) (ii) *und* (Z4) *regulär.*

Beweis. Offensichtlich ist die Folge der Operatoren $\{a_m\}$, $h \in \Lambda_0$, regulär, so daß es zu zeigen genügt, daß $\{q\}$, $h \in \Lambda_0$, m-regulär ist. Dies ist aber gerade die fundamentale Aussage von Satz 4 aus [13], der unter der Voraussetzung (Z3) (ii) für fast alle $h \in \Lambda_0$ das Bestehen der a priori Abschätzung

$$|u|_0 \leqq K |q u|_0, \qquad u \in l^2[c, d - mh],$$

ergibt. Da $D_m u$ für alle $u \in H^0$ in $l^2[c, d - mh]$ liegt, ist damit die Behauptung bewiesen.

(9) *Unter der Bedingung* (Z1) *konvergiert für* $j = 1, \ldots, m$

$$l^{(j)} \to l_0^{(j)} \quad in \quad \{H^m \to \mathbb{H}^m\} \qquad (h \in \Lambda_0).$$

Beweis. Aus der Voraussetzung (Z1) folgt mit (1), daß sich $r_1^{(j)}(E)$ in der Gestalt

$$r_1^{(j)}(E) = \sum_{k=1}^{n} \gamma_{jk}(E - I)^k D^{m-1} + \sum_{k=0}^{m-1} \beta_{jk} D^k, \qquad j = 1, \ldots, m,$$

darstellen läßt, wobei n eine natürliche Zahl und die Folge $\{\gamma_{jk}(c, h)\}$, $h \in \Lambda_0$, beschränkt ist sowie $\beta_{jk}(c, h) \to \alpha_{jk}$ $(h \in \Lambda_0)$ konvergiert. Sei $\{u\}$, $h \in \Lambda_0$, eine Folge von Elementen $u \in H^m$, $h \in \Lambda_0$, mit

$$u \rightharpoonup u_0 \quad \text{in} \quad \{H^m \to \mathbb{H}^m\} \qquad (h \in \Lambda_0).$$

Der Beziehung (3.27) entnehmen wir, daß $(D^k u)(c) \to (\partial^k u_0)(c_0)$, $k = 0, \ldots, m-1$, und $\big((E - I)^k D^{m-1} u\big)(c) \to 0$ $(h \in \Lambda_0)$, $k = 1, \ldots, n$, gilt, woraus sich

$$\big(r_1^{(j)}(E)\, u\big)(c) \to \sum_{k=0}^{m-1} \alpha_{jk} (\partial^k u_0)(c_0) \qquad (h \in \Lambda_0), \qquad j = 1, \ldots, m,$$

ergibt. Auf dieselbe Weise zeigt man die entsprechenden Aussagen für $r_2^{(j)}$, und damit ist die behauptete Konvergenz bewiesen.

(10) *Die Schar von Differenzenapproximationen* (4.4) *genügt unter den Voraussetzungen* (Z1)—(Z4) *den Bedingungen* (S) (i)—(iv) *aus* §1.

Beweis. Als erstes bringen wir die Operatoren L und M auf die in §1 benötigte Gestalt. Dabei beschränken wir uns auf L, da die Umformungen für M analog verlaufen. Die in gleicher Weise wie die Funktionen a_k in (Z4) genannten Koeffizienten des Operators L aus §1 wollen wir zur Unterscheidung mit A_k, $k = 0, \ldots, m$, bezeichnen.

Aus den Voraussetzungen (Z2) und (Z3) (i) ergibt sich mit (1) die Darstellung

$$\Delta_L^{(j)} = P_{r_j s_j}\left(\sum_{k=1}^{\infty} \gamma_{jk}(E - I)^k E^{-n_j} D^{m-1} + \sum_{k=0}^{m-1} \alpha_{jk} E^{-n_j} D^k\right)$$

für $j=0, \ldots, m-1$, und eine entsprechende für $\Delta_L^{(m)} - q D_m$ mit gewissen Funktionen γ_{jk}, α_{jk} und $n_j \in \Omega$, die für $j=0, \ldots, m$

$$(11) \quad |\gamma_{jk}(x,h)| \leqq K, \quad k=1,2,\ldots, \quad |n_j(x,h)| \leqq K, \quad x \in G_h^{(j)},\ h \in \Lambda_0,$$

$$(12) \quad \max_{x \in G_h^{(j)}} |\alpha_{jk}(x,h) - \delta_{jk}| \to 0 \quad (h \in \Lambda_0), \quad k=0, \ldots, m-1,$$

erfüllen, wobei δ_{jk} das Kroneckersche Symbol ist. Bei geeigneter Wahl der Funktionen n_j können wir erreichen, daß sich $\Delta_L^{(j)}$, $j=0, \ldots, m-1$, auch in der Gestalt

$$\Delta_L^{(j)} = P_{r_j s_j}\left(h \sum_{k=1}^{\infty} \gamma_{jk}(E-I)^{k-1} E^{-n_j} D_m + \sum_{k=0}^{m-1} \alpha_{jk} E^{-n_j} D_k\right)$$

schreiben läßt. Zusammen mit der entsprechenden Darstellung für $\Delta^{(m)} - q D_m$ erhalten wir dann das Ergebnis

$$(13) \quad \begin{aligned} L = &\left(a_m q + h \sum_{j=0}^{m} a_j P_{r_j s_j} \sum_{k=1}^{\infty} \gamma_{jk}(E-I)^{k-1} E^{-n_j}\right) D_m \\ &+ \sum_{k=0}^{m-1} \sum_{j=0}^{m} a_j P_{r_j s_j} \alpha_{jk} E^{-n_j} D_k, \end{aligned}$$

so daß L in der Tat die in (1.6) verlangte Gestalt besitzt mit offensichtlicher Bedeutung der Koeffizienten A_j, $j=0, \ldots, m$. Für eine Folge $\{u\}$, $h \in \Lambda_0$, die (7) genügt, haben wir in (3.24) bewiesen, daß auch $\{P_{r_j s_j} E^{-n_j} u\}$, $h \in \Lambda_0$, stark gegen u_0 konvergiert, und aus (12) erhalten wir für $k=0, \ldots, m-1$, $j=0, \ldots, m$, leicht

$$\alpha_{jk} P_{r_j s_j} E^{-n_j} u \to \begin{cases} 0, & j \neq k \\ u_0, & j=k \end{cases} \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda_0).$$

Hieraus folgt dann zusammen mit (5) auch $A_j \to a_j^0$ $(h \in \Lambda_0)$, $j=0, \ldots, m-1$. Aus (3.23) und (3.24) ergibt sich bei Beachtung von $p(1)=1$ auch die starke Konvergenz von $\{qu\}$, $h \in \Lambda_0$, gegen u_0. Mit Hilfe von (11) sieht man, daß bereits für jede beschränkte Folge $\{u\}$, $h \in \Lambda_0$, und $j=0, \ldots, m$

$$(14) \quad h \sum_{k=1}^{\infty} \gamma_{jk}(E-I)^{k-1} E^{-n_j} u \to 0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda_0)$$

konvergiert, woraus man auch $A_m \to a_m^0$ $(h \in \Lambda_0)$ erhält, so daß (S)(i) gezeigt ist.

Zum Beweis von $A_m^* \to (a_m^0)^*$ $(h \in \Lambda_0)$ machen wir von der Bemerkung vor (1.4) Gebrauch, daß dies genau dann der Fall ist, wenn A_m schwach konvergente Folgen in ebensolche überführt. Bei Verwendung von (5) reduziert sich der Beweis auf den Nachweis von

$$(15) \quad q^* \to I \quad (h \in \Lambda_0).$$

Nun ist der Operator T aus (3.23) selbstadjungiert, so daß auch $T^* \to 0$ $(h \in \Lambda_0)$ konvergiert. Ferner bestätigt man für alle ganzen Zahlen k, l, r, $k \geqq 0$, $l \geqq 0$, leicht die Beziehung $(P_{kl} E^r)^* = P_{00} E^{-r} P_{kl}$. Mit (3.24) ergibt sich daher auch $(P_{kl} E^r)^* \to I$ $(h \in \Lambda_0)$, und wir erhalten (15).

In (8) haben wir die m-Regularität der Folge $\{a_m q\}$, $h \in \Lambda_0$, bewiesen, aus der auch die von $\{A_m\}$, $h \in \Lambda_0$, folgt, da man den zusätzlichen Term für genügend kleines h in der erforderlichen Weise abschätzen kann. Damit ist auch (S) (ii) gezeigt.

Die Bedingung (S) (iii) ist in (9) nachgewiesen worden. (S) (iv) erhält man aus der Eigenschaft (H 4), da die starke Konvergenz der Koeffizienten von M schon aus (S) (i) bekannt ist.

Beweis von Satz (4.10). Wir stützen uns auf Sätze (1.8) und (1.9) und prüfen zunächst die Bedingung (S) (v) nach. M_0 ist aufgrund der Kompaktheit der natürlichen Einbettung von $\mathbb{H}^m$ in $\mathbb{H}^{m-1}$ und M als Operator in einem endlichdimensionalen Raum vollstetig. Für den Operator P in (S) (v) nehmen wir $P_{m_1 m_2}$ und haben damit gemäß (3.24) auch $P \to I$ $(h \in \Lambda_0)$. Der Bedingung (Z 4) und $R(q) \subseteq P H^0$ zufolge ist auch

$$L H^m \subseteq P H^0, \qquad M H^m \subseteq P H^0, \qquad h \in \Lambda_0,$$

gesichert. Nach Voraussetzung ist die Aufgabe $\pi_\lambda^0(L_0, 0, l_0^{(j)}, 0)$ für wenigstens eine Zahl $\lambda \in \mathbb{C}$ nur trivial lösbar. Da wir in (10) schon (S) (i)—(iv) gezeigt haben, gilt für diese Zahl λ für fast alle $h \in \Lambda_0$ die Ungleichung in (1.7). Den Sätzen über Operatoren in endlichdimensionalen Räumen zufolge ist somit für fast alle $h \in \Lambda_0$ die Resolventenmenge $\varrho(L)$ nichtleer und das System der Funktionale $\{l^{(j)}\}$, $j = 1, \ldots, m$, linear unabhängig. Die Voraussetzung (Z 5) über f ist aufgrund von (3.20) mit der starken Konvergenz der Folge $\{f\}$, $h \in \Lambda_0$, gegen f_0 äquivalent, womit alle Voraussetzungen von (1.8) und (1.9) bewiesen sind und die Behauptung folgt.

6. Weitere Konvergenzaussagen. Konvergenzordnung

6.1. Unter Bezugnahme auf weitere Ergebnisse aus [12] machen wir in diesem Paragraphen Aussagen über die Konvergenz der Eigenfunktionen, der Eigenwerte der geometrischen Vielfachheit nach und über die Konvergenzordnung bei bekanntem Abschneidefehler.

Ist λ_0 ein Eigenwert von L_0 der algebraischen Vielfachheit n_0, so haben wir in [12, 3.19] das Folgende bewiesen. Sei U eine Umgebung von λ_0, in deren Abschließung keine weiteren Eigenwerte von L_0 liegen. Ist dann N_0 der algebraische Eigenraum von L_0 zu λ_0 und N die direkte Summe der algebraischen Eigenräume von L zu den in U gelegenen Eigenwerten, so existieren in N_0 bzw. N vollständige Orthonormalsysteme $\{u_1^0, \ldots, u_{n_0}^0\}$ bzw. $\{u_1, \ldots, u_{n_0}\}$ mit

$$u_j \to u_j^0 \quad \text{in} \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda_0), \qquad j = 1, \ldots, n_0.$$

Hieraus ergibt sich die folgende Aussage.

(1) *Sei λ_0 ein algebraisch einfacher Eigenwert von L_0 und u_0 eine zugehörige Eigenfunktion mit $\|u_0\|_0 = 1$. Die Funktion $\varphi \in \mathbb{H}$ erfülle die Beziehung $((u_0, \varphi))_0 > 0$. Ist dann $\{\lambda\}$, $h \in \Lambda$, eine Folge von Eigenwerten von L mit $\lambda \to \lambda_0$ $(h \in \Lambda)$, so ist λ für fast alle $h \in \Lambda$ ein algebraisch einfacher Eigenwert von L. Sind $\{u\}$, $h \in \Lambda$, zugehörige Eigenfunktionen, die durch $|u|_0 = 1$ und $(u, R\varphi)_0 \geqq 0$ normiert sind, so gelten die Konvergenzaussagen von Satz* (5.10).

Beweis. Da die Eigenwerte von L der algebraischen Vielfachheit nach konvergieren, ist λ für fast alle $h \in \Lambda$ ein einfacher Eigenwert von L. Sei $\{v\}$ eine zugehörige Folge von Eigenfunktionen mit $|v|_0 = 1$ und $v \to u_0$ in $\{H^m \to \mathbb{H}^m\}$ $(h \in \Lambda)$. Es ist offensichtlich $u = \alpha v$, $|\alpha| = 1$, für fast alle $h \in \Lambda$ und da $(v, R\varphi)_0 \to ((u_0, \varphi))_0 > 0$ $(h \in \Lambda)$ konvergiert, erhalten wir aus der Normierung von u, daß $\alpha \to 1$ $(h \in \Lambda)$ gilt. Hieraus folgt die Behauptung.

6.2. Die Konvergenz der Eigenwerte von L kann man im allgemeinen nur der algebraischen und nicht der geometrischen Vielfachheit nach behaupten, wie etwa das folgende Beispiel zeigt. Vorgelegt sei das Eigenwertproblem

$$(\partial^2 u_0)(x) = \lambda u_0(x), \qquad x \in (0, 1)$$

$$(\partial u_0)(0) = u_0(0) - u_0(1) + \tfrac{1}{2}(\partial u_0)(1) = 0.$$

Man rechnet leicht nach, daß $\lambda = 0$ ein Eigenwert der geometrischen Vielfachheit 1 und der algebraischen Vielfachheit 2 ist. Die zugehörigen Hauptvektoren 1. und 2. Stufe sind $u_0 = 1$ und $v_0 = x^2$. Als Differenzenapproximation nehmen wir

$$\frac{1}{h^2}\big(u(x+h) - 2u(x) + u(x-h)\big) = \lambda u(x), \qquad x = jh, \quad j = 1, \ldots, r-1,$$

$$\frac{1}{h}\big(u(h) - u(0)\big) = u(0) - u(1) + \frac{1-h}{2h}\big(u(1) - u(1-h)\big) = 0,$$

wobei $rh = 1$ ist. In diesem Fall ist also $c = 0$, $d = 1$, $L = E^{-1}D_2$, $M = P = P_{11}$. Man sieht sofort, daß $\lambda = 0$ auch ein Eigenwert von L der geometrischen Vielfachheit 1 ist mit der zugehörigen Eigenfunktion $u = 1$. Zur Bestimmung eventuell vorhandener Hauptvektoren 2. Stufe von L zu $\lambda = 0$ ist die Gleichung $Lv = Mu$ zu lösen. Jede Lösung dieser Gleichung, die der linken Randbedingung genügt, läßt sich als Linearkombination von u und $w = x(x-h)/2$ darstellen. Da w jedoch nicht die zweite Randbedingung erfüllt, existieren keine Hauptvektoren 2. Stufe von L zu $\lambda = 0$, und die algebraische ist gleich der geometrischen Vielfachheit. Aus der Konvergenz der Eigenwerte der algebraischen Vielfachheit nach folgt, daß L einen Eigenwert $\lambda_1 \neq 0$ der algebraischen und geometrischen Vielfachheit 1 mit $\lambda_1 \to 0$ $(h \in \Lambda_0)$ besitzen muß. Das ist aber gerade das gewünschte Gegenbeispiel.

Für die Konvergenz der geometrischen Vielfachheit nach lassen sich hinreichende Bedingungen angeben. Es seien X_1^0 und X_1 die im Anschluß an (1.6) definierten Teilräume. Es sei L_0 symmetrisch und M_0 positiv in X_1^0 (was z.B. bei volldefiniten und allgemeiner bei symmetrischen M_0-definiten Aufgaben erfüllt ist), L sei symmetrisch in X_1, und X sei ein Teilraum von H^m mit $X_1 \leq X$ und der Eigenschaft, daß M auf X positiv ist und für jedes $u \in R(P_{m_1 m_2})$ aus $(u, v)_0 = 0$, $v \in X$, folgt $u = 0$. Dann haben wir in [12, 1.2] gezeigt, daß L_0 und L nur Hauptvektoren 1. Stufe und reelle Eigenwerte besitzen und daher für jeden Eigenwert die algebraische gleich der geometrischen Vielfachheit ist.

Nehmen wir an, daß die vorstehenden Voraussetzungen nur für L_0 erfüllt sind, und ist α eine reelle Zahl aus $\varrho(L_0)$, so kann man die Eigenwerte von L_0 nach aufsteigender Größe in der Gestalt

$$\cdots \leqq \lambda_{-2}^0 \leqq \lambda_{-1}^0 \leqq \lambda_0^0 < \alpha < \lambda_1^0 \leqq \lambda_2^0 \leqq \cdots$$

anordnen. Für eine positive Zahl β können wir die Eigenwerte λ von L mit $|\operatorname{Im}\lambda| \leqq \beta$ für fast alle $h \in \Lambda_0$ durch

$$\cdots \leqq \operatorname{Re}\lambda_{-2} \leqq \operatorname{Re}\lambda_{-1} \leqq \operatorname{Re}\lambda_0 < \alpha < \operatorname{Re}\lambda_1 \leqq \operatorname{Re}\lambda_2 \leqq \cdots$$

anordnen, und es gilt dann (s. [12, 3.3])

$$\lambda_j \to \lambda_j^0 \quad (h \in \Lambda_0), \qquad j = 0, \pm 1, \ldots .$$

6.3. In diesem Abschnitt entwickeln wir einige Aussagen über die Konvergenzordnung der Lösungen und Eigenwerte von (4.4). Wir setzen voraus, daß die Funktionen f_0, a_m^0, a_j^0 und b_j^0, $j = 0, \ldots, m-1$, wenigstens aus $C^0[c_0, d_0]$ sind. Es sei τ eine positive Zahl und u_0 eine Lösung von $\pi_\lambda^0(L_0, f_0, l_0^{(j)}, \gamma_0^{(j)})$. Wir sagen, daß der Abschneidefehler für u_0 von der Ordnung $O(h^\tau)$, $h \in \Lambda_0$, ist, genau wenn

$$|(L R u_0 - R' L_0 u_0)|_0 = O(h^\tau), \qquad h \in \Lambda_0, \tag{2}$$

$$|(M R u_0 - R' M_0 u_0)|_0 = O(h^\tau), \qquad h \in \Lambda_0, \tag{3}$$

$$|l^{(j)}(R u_0) - l_0^{(j)}(u_0)| = O(h^\tau), \qquad h \in \Lambda_0, \tag{4}$$

gilt, wobei „O" das Landausche Symbol, R der Restriktionsoperator aus 2.3 und R' eine Abbildung eines Teilraumes $\mathbb{H}_2 \subseteq \mathbb{H}^0$ in H^0 ist, der f_0, $L_0 u_0$ und $M_0 u_0$ enthält, z.B. $R' = P_{m_1 m_2} R_1$.

Es sei $\alpha_j^0 \in C^0(\mathbb{R}^1)$, $j = 0, \ldots, m-1$, und Q_0 der Differentialoperator

$$Q_0 = \sum_{j=0}^{m-1} \alpha_j^0 \partial^j.$$

Es sei $\mathcal{K}$ eine endliche Menge ganzer Zahlen, g_k, $h \in \Lambda_0$, $k \in \mathcal{K}$, seien Funktionen aus Ω und r_1, r_2 seien nichtnegative ganze Zahlen mit der Eigenschaft

$$g_k(x, h) = 0, \qquad x \in [c + r_1 h, d - r_2 h]_h, \qquad x + k h \notin [c, d]_h$$

für jedes $h \in \Lambda_0$ und $k \in \mathcal{K}$. Der Operator

$$Q = \sum_{k \in \mathcal{K}} g_k E^k$$

erfülle für jede Funktion $\varphi \in C^{m-1}(\mathbb{R}^1)$ die Konsistenzbedingung

$$\max_{x \in G_h} |(Q\varphi - Q_0 \varphi)(x)| \to 0 \quad (h \in \Lambda_0),$$

wobei wir zur Abkürzung $G = [c + r_1 h, d - r_2 h] \cap [c_0, d_0]$ gesetzt haben. Wir behaupten dann den folgenden Satz.

(5) *Sei λ kein Eigenwert von L_0, sei $f_0 \in C^0[c_0, d_0]$ und sei u_0 eine Lösung von $\pi_\lambda^0(L_0, f_0, l_0^{(j)}, \gamma_0^{(j)})$, für die der Abschneidefehler von der Ordnung $O(h^\tau)$ ist. Gilt dann noch*

$$|f - R' f_0|_0 = O(h^\tau), \qquad |\gamma^{(j)} - \gamma_0^{(j)}| = O(h^\tau), \qquad j = 1, \ldots, m,$$

sowie

$$\max_{x \in G_h} |(Q R u_0 - Q_0 u_0)(x)| = O(h^\tau), \qquad h \in \Lambda_0,$$

und sind $\{u\}$, $h\in\Lambda_0$, *Lösungen der Probleme* $\pi_\lambda(L, f, l^{(j)}, \gamma^{(j)})$, *so ist*

$$\max_{x\in G_h} |(Q u - Q_0 u_0)(x)| = O(h^\tau), \quad h\in\Lambda_0.$$

Beweis. Unter den vorangehenden Voraussetzungen haben wir in [12, 4.1] gezeigt, daß $|u - R u_0|_m = O(h^\tau)$, $h\in\Lambda_0$, gilt. Die Behauptung reduziert sich daher mit Hilfe der Dreiecksungleichung auf

$$\max_{x\in G_h} |Q(u - R u_0)(x)| \leqq K |u - R u_0|_m, \quad h\in\Lambda_0. \tag{6}$$

Aus den Voraussetzungen über Q erhalten wir aus (5.1) die Darstellung

$$Q = \sum_{k=1}^{r} \gamma_k (E - I)^k E^{-n} D_{m-1} + \sum_{k=0}^{m-1} \alpha_k E^{-n} D_k$$

mit auf G_h beschränkten Funktionen γ_k, α_k, r und n, woraus sich mit (3.25) die Behauptung (6) ergibt.

Über die Konvergenzordnung der Eigenwerte haben wir in [12, 4.1] den folgenden Satz bewiesen.

(7) *Es sei* n *eine natürliche Zahl und* λ_0 *sei ein Pol* n*-ter Ordnung der Resolvente* $(\lambda M_0 - L_0)^{-1}$. *Besitzt jeder Hauptvektor von* L_0 *zu* λ_0 *einen Abschneidefehler der Ordnung* $O(h^\tau)$, *so gilt* $|\lambda - \lambda_0|^n = O(h^\tau)$, $h\in\Lambda$, *für jede Folge* $\{\lambda\}$, $h\in\Lambda$, *von Eigenwerten von* L *mit* $\lambda\to\lambda_0$ $(h\in\Lambda)$.

Im Falle eines algebraisch einfachen Eigenwertes von L_0 können wir auch Aussagen über die Konvergenzordnung der Eigenfunktionen machen. Wir nehmen an, daß die Eigenfunktion u_0 von L_0 zu λ_0 wie in (1) normiert ist. Es sei $[\cdot, \cdot]$ ein Skalarprodukt in H^0 mit der Eigenschaft, daß eine Zahl K mit $|[u]| \leqq K |u|_0$, $u\in H_0$, $h\in\Lambda^0$, existiert und daß

$$|[R u_0, R\varphi] - ((u_0, \varphi))_0| = O(h^\tau), \quad h\in\Lambda_0,$$

ist. Dann gilt der Satz

(8) *Der Abschneidefehler für* u_0 *sei von der Ordnung* $O(h^\tau)$, $\{\lambda\}$, $h\in\Lambda$, *sei eine Folge von Eigenwerten von* L *mit* $\lambda\to\lambda_0$ $(h\in\Lambda)$, *und die zugehörige Folge von Eigenfunktionen sei durch die Vorschrift* $[u] = 1$, $[u, R\varphi] \geqq 0$ *normiert. Ist dann* Q *wie in* (5) *definiert, so gilt*

$$\max_{x\in G_h} |(Q u - Q_0 u_0)(x)| = O(h^\tau), \quad h\in\Lambda.$$

Bei genügender Differenzierbarkeit der Koeffizienten a_j^0 von L_0 und b_j^0 von M_0 und der Funktion f_0 kann man mit unseren Methoden auch die Konvergenz der Differenzenquotienten höherer als m-ter Ordnung beweisen, wie es in ähnlicher Weise in [23, 28] für partielle Differentialgleichungen im Gebietsinnern geschehen ist, jedoch gehen wir hier auf diese Fragen nicht näher ein.

7. Beispiele

7.1. In den vorangehenden drei Paragraphen haben wir für eine Klasse von Differenzenoperatoren die Konvergenz des Rand- und Eigenwertproblems bewiesen. Im folgenden werden wir die dort erhaltenen Ergebnisse am Beispiel

eines Mehrstellenverfahrens und einer Approximation höherer Annäherung erläutern.

Für das Rand- und zugehörige Eigenwertproblem

$$\begin{aligned} &-(\partial^2 u_0)(x) + \left(a^0(x) - \lambda\right) u_0(x) = f_0(x), \qquad x \in (0,1) \\ &u_0(0) = \gamma_0^{(1)}, \qquad u_0(1) = \gamma_0^{(2)} \end{aligned} \tag{1}$$

mit f_0, $a^0 \in C^4[0,1]$, $a^0(x) \geqq 0$, $x \in [0,1]$, ist in [1] (s. auch [21]) das folgende Mehrstellenverfahren angegeben worden.

$$\begin{aligned} &-\frac{1}{h^2}(u_{j-1} - 2u_j + u_{j+1}) + \frac{1}{12}\left((a^0 u)_{j+1} + 10(a^0 u)_j + (a^0 u)_{j-1}\right) \\ &\quad -\lambda \frac{1}{12}(u_{j-1} + 10u_j + u_{j+1}) = f_j + \frac{h^2}{12}(\partial^2 f)_j, \qquad j = 1, \ldots, n-1, \\ &u_0 = \gamma^{(1)}, \qquad u_n = \gamma^{(2)}, \end{aligned} \tag{2}$$

wobei $nh = 1$ und $u_j = u(jh)$, $j = 0, \pm 1, \ldots$, ist. Wir bringen (2) auf die in (4.4) betrachtete Gestalt. Es sei

$$\Lambda_0 = \{h \mid h = n^{-1},\ n = 2, 3, \ldots\},$$

$c_0 = c = 0$, $d_0 = d = 1$, so daß H^0 die Menge der auf den Gitterpunkten $x_j = jh$, $j = 0, \pm 1, \ldots$, erklärten Funktionen ist, die für $j < 0$ und $j > n$ verschwinden, versehen mit dem in 2.2 definierten Skalarprodukt $(\cdot, \cdot)_0$. Der am selben Ort eingeführte Operator P_{11} ist die orthogonale Projektion von H^0 auf den Teilraum

$$X_1 = \{u \in H^0 \mid u(0) = u(1) = 0\}, \qquad h \in \Lambda_0.$$

Seien die beiden Restriktionsoperatoren R, R': $C^2[0,1] \to H^0$ definiert durch

$$(R u_0)(x) = \begin{cases} u_0(x), & x = jh, \quad j = 0, \ldots, n \\ 0, & \text{sonst} \end{cases}$$

und

$$(R' u_0)(x) = \begin{cases} u_0(x) + \dfrac{h^2}{12}(\partial^2 u_0)(x), & x = jh, \quad j = 1, \ldots, n-1 \\ 0, \quad \text{sonst}. \end{cases}$$

Ist $a_0' \in B(H^0)$ der Operator der Multiplikation mit der Funktion $R a_0$, so sind die zu (2) gehörenden Operatoren L und M und die Funktion f aus (4.4) gegeben durch

$$L = -E^{-1} D_2 + \frac{1}{12} P_{11}(E + 10I + E^{-1}) a_0',$$

$$M = \frac{1}{12} P_{11}(E + 10I + E^{-1}) \quad \text{und} \quad f = R' f_0.$$

Im vorliegenden Fall ist $m_1 = m_2 = 1$ und $\Delta_L^{(2)} = E^{-1} D_2$, dem zentralen Differenzenquotienten zweiter Ordnung, der in (Z 3) mit $g_{mk} = 0$ und $q = P_{11} E^{-1}$ enthalten ist. In der Definition von q ist dabei $p = 1$ und $n_1 = n_2 = 0$. Weiter haben wir $\Delta_L^{(0)} = I$, womit (Z 2) und (Z 3) erfüllt ist. Die Koeffizienten

$$a_0 = \frac{1}{12} P_{11}(E + 10I + E^{-1}) a_0', \qquad a_2 = R(-1)$$

genügen der Voraussetzung (Z4) mit dem Operator

$$Q_{10} = \frac{1}{12}(E + 10I + E^{-1}).$$

Auch der Operator M erfüllt die Bedingungen (Z2) und (Z4), wie man für

$$\Delta_M^{(0)} = \frac{1}{12} P_{11}(E + 10I + E^{-1}), \qquad b_0 = P_{11} R(1)$$

leicht sieht. Da schließlich

$$f = R' f_0 = P_{11} R\left(f_0 + \frac{h^2}{12} \partial^2 f_0\right)$$

die in (4.9) gegebene Gestalt hat und die Randbedingungen $u_0 = \gamma^{(1)}$, $u_n = \gamma^{(2)}$ trivialerweise die Form (4.3) besitzen, gelten bei Annahme von $\gamma^{(j)} \to \gamma_0^{(j)}$ $(h \in \Lambda_0)$, $j = 1, 2$, die Konvergenzaussagen des Hauptsatzes (4.10).

Im vorliegenden Beispiel kann man noch weitergehende Resultate erhalten. Der Operator $L_0 = -\partial^2 + a^0$ ist im Teilraum

$$X_1^0 = \{u \in W^{2,2}(0, 1) \mid u(0) = u(1) = 0\}$$

positiv definit und dasselbe gilt trivialerweise für $M_0 = I$. In [7, S. 293] ist gezeigt worden, daß die Eigenwerte von (1) einfach sind, und daher besteht das Spektrum von (1) aus einer gegen Unendlich strebenden Folge positiver algebraisch einfacher Eigenwerte $0 < \lambda_1^0 < \lambda_2^0 < \cdots$, und jeder Eigenwert ist ein Pol erster Ordnung der Resolvente $(\lambda M_0 - L_0)^{-1}$. Der Operator M ist im Teilraum X_1 symmetrisch und erfüllt die Beziehung

$$\tfrac{2}{3} |u|_0^2 \leqq (Mu, u)_0 \leqq |u|_0^2, \qquad u \in X_1, \qquad h \in \Lambda_0.$$

Dagegen ist L i.allg. nicht symmetrisch, jedoch berechnet man für jedes $u \in X_1$ mit Hilfe partieller Summation und der Friedrichschen Ungleichung für den vorwärts genommenen Differenzenquotienten D^1 (s. [24])

$$\begin{aligned} \operatorname{Re}(Lu, u)_0 &= |D^1 u|_0^2 + \frac{10}{12}(au, u)_0 + \frac{h}{12} \sum_{j=1}^{n-1} \operatorname{Re}[(au)_{j+1} + (au)_{j-1}] \bar{u}_j \\ &\geqq |u|_0^2 + \frac{5}{6} \min_{x \in [0,1]} a^0(x) |u|_0^2 - \frac{1}{6} \max_{x \in [0,1]} a^0(x) |u|_0^2 \\ &= \mu_0 |u|_0^2, \qquad \mu_0 = 1 + \frac{5}{6} \min_{x \in [0,1]} a^0(x) - \frac{1}{6} \max_{x \in [0,1]} a^0(x) \end{aligned}$$

und

$$\begin{aligned} |\operatorname{Im}(Lu, u)_0| &= \frac{h}{12} \left| \sum_{j=1}^{n-1} \operatorname{Im}[(au)_{j+1} + (au)_{j-1}] \bar{u}_j \right| \\ &\leqq \frac{1}{12} \max_{j=1,\ldots,n-1} |a_{j+1}^0 - a_j^0| \, |u|_0^2, \end{aligned}$$

so daß sich für die Eigenwerte λ von (2) für jedes $h \in \Lambda_0$ die Abschätzung

$$\operatorname{Re} \lambda \geqq \begin{cases} \mu_0, & \mu_0 \geqq 0 \\ \frac{3}{2}\mu_0, & \mu_0 < 0 \end{cases}, \qquad |\operatorname{Im} \lambda| \leqq \tfrac{1}{8} \max_{j=1,\ldots,n-1} |(a_{j+1}^0 - a_j^0)|$$

ergibt. Die Eigenwerte bleiben also nach unten beschränkt und ihre Imaginärteile konvergieren gleichmäßig gegen Null. Ordnen wir sie in der Reihenfolge

$$\operatorname{Re}\lambda_1 \leqq \operatorname{Re}\lambda_2 \leqq \cdots \leqq \operatorname{Re}\lambda_n$$

an, so erhalten wir den Überlegungen am Ende von § 6.2 zufolge, daß λ_j für fast alle $h \in \Lambda_0$ einfach ist und

$$\lambda_j \to \lambda_j^0 \quad (h \in \Lambda_0), \qquad j = 1, 2, \ldots,$$

konvergiert. Es läßt sich darüber hinaus auch die Ordnung der Konvergenz abschätzen. In [1] wird gezeigt, daß die Beziehungen

$$|L R u_0 - R' L_0 u_0|_0 = |M R u_0 - R' M_0 u_0|_0 = O(h^4), \qquad h \in \Lambda_0,$$

bestehen. Trivialerweise gilt

$$|f - R' f_0|_0 = |l^{(j)}(R u_0) - l_0^{(j)}(u_0)| = 0, \qquad j = 1, 2, \qquad h \in \Lambda_0,$$

und (6.7) ergibt die Aussage

$$|\lambda_j - \lambda_j^0| = O(h^4), \qquad h \in \Lambda_0, \qquad j = 1, 2, \ldots,$$

über die Konvergenzordnung der Eigenwerte. Setzen wir noch $|\gamma^{(j)} - \gamma_0^{(j)}| = O(h^4)$, $h \in \Lambda_0$, voraus und ist λ kein Eigenwert von (1), so erhalten wir ferner aus (6.5) für die Konvergenzordnung der Lösungen u, $h \in \Lambda_0$, von (2) gegen die Lösung u_0 von (1)

$$(3) \quad \max_{x \in [0,1]_h} |(u - u_0)(x)| = O(h^4), \qquad \max_{x \in [2h, 1-2h]_h} |(\Delta^{(1)} u - \partial u_0)(x)| = O(h^4), \quad h \in \Lambda_0,$$

wobei $\Delta^{(1)} = (12h)^{-1}(E^{-2} - 8E^{-1} + 8E - E^2)$ eine Approximation der Ordnung $O(h^4)$ des Differentialquotienten 1. Ordnung ist. Unter der Annahme, daß die Anzahl der Gitterpunkte in $[0, 1]$ ungerade ist, führen wir mit Hilfe der Simpsonformel das Skalarprodukt

$$[u, v] = \frac{h}{3} \sum_{x \in \mathbb{R}_h^1} \varrho(x)\, u(x)\, \bar{v}(x), \qquad u, v \in H^0,$$

ein, wobei $\varrho \in \Omega$ durch

$$\varrho(x) = \begin{cases} 1, & x \leqq 0, \quad x \geqq 1 \\ 2, & x = jh, \quad j = 1, 3, \ldots, n-1 \\ 4, & x = jh, \quad j = 2, 4, \ldots, n-2 \end{cases}$$

gegeben ist. Sei λ_0 ein Eigenwert von (1), und die zugehörige Eigenfunktion u_0 sei mit einer geeigneten Funktion $\varphi \in C^4[0, 1]$ durch $\|u_0\|_0 = 1$ und $((u_0, \varphi)) > 0$ normiert. Ist $\{\lambda\}$, $h \in \Lambda$, eine Folge von Eigenwerten von (2) mit $\lambda \to \lambda_0$ $(h \in \Lambda)$ und sind $\{u\}$, $h \in \Lambda$, zugehörige Eigenfunktionen, die $[u, u] = 1$, $[u, R\varphi] \geqq 0$ erfüllen, so haben wir in (6.8) bewiesen, daß auch in diesem Fall die Konvergenzaussagen (3) gelten.

7.2. Als zweites Beispiel wählen wir im Intervall $[0, 1]$ das Rand- und zugehörige Eigenwertproblem

$$(4) \quad \begin{aligned} &(\partial^4 + \partial^2) u_0(x) - \lambda [4x \partial^3 + (2 - 3x^2) \partial^2 - 4x \partial] u_0(x) = f_0(x), \\ &u_0(0) = \gamma_0^{(1)}, \quad u_0(1) = \gamma_0^{(2)}, \quad (\partial^2 u_0)(0) = \gamma_0^{(3)}, \quad (\partial^2 u_0)(1) = \gamma_0^{(4)}, \end{aligned}$$

das die Stabilität eines flachen Parabelbogens mit gelenkigen Enden beschreibt (s. [5, S. 44]). Es ist leicht zu sehen, daß $\lambda = 0$ kein Eigenwert von (4) ist, so daß (4) unseren Voraussetzungen aus § 5.1 genügt. Sei diesmal H^0 der Hilbertsche Raum der auf den Gitterpunkten $x_j = jh$, $j = 0, \pm 1, \ldots$, erklärten Funktionen, die für $x_j < -2h$ und $x_j > 1 + 2h$ verschwinden, versehen mit dem Skalarprodukt $(\cdot, \cdot)_0$ aus § 2.2, so daß $c = -2h$ und $d = 1 + 2h$ ist. Wir setzen wieder $u_j = u(jh)$, $j = 0, \pm 1, \ldots$, und erklären dann die Differenzenapproximationen $\Delta^{(j)}$, $j = 1, \ldots, 4$, höherer Genauigkeit für die Differentialquotienten j-ter Ordnung (s. [6, S. 538]) durch

$$(\Delta^{(1)}u)_j = \begin{cases} \frac{1}{12h}(u_{j-2} - 8u_{j-1} + 8u_{j+1} - u_{j+2}), & j = 0, \ldots, n \\ 0, \quad \text{sonst}, \end{cases}$$

$$(\Delta^{(2)}u)_j = \begin{cases} \frac{1}{12h^2}(-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}), & j = 0, \ldots, n \\ 0, \quad \text{sonst}, \end{cases}$$

$$(\Delta^{(3)}u)_j = \begin{cases} \frac{1}{8h^3}(u_{j-3} - 8u_{j-2} + 13u_{j-1} - 13u_{j+1} + 8u_{j+2} - u_{j+3}), & j = 1, \ldots, n-1 \\ \frac{1}{2h^3}(-u_{j-2} + 2u_{j-1} - 2u_{j+1} + u_{j+2}), & j = 0, n \\ 0, \quad \text{sonst}, \end{cases}$$

$$(\Delta^{(4)}u)_j = \begin{cases} \frac{1}{6h^4}(-u_{j-3} + 12u_{j-2} - 39u_{j-1} + 56u_j - 39u_{j+1} + 12u_{j+2} - u_{j+3}), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = 1, \ldots, n-1 \\ \frac{1}{h^4}(u_{j-2} - 4u_{j-1} + 6u_j - 4u_{j+1} + u_{j+2}), & j = 0, n \\ 0, \quad \text{sonst}. \end{cases}$$

Wir nehmen an, daß f_0 beschränkt und fast überall stetig ist, und setzen $f_j = f_0(x_j)$, $j = 0, \ldots, n$, und $f_j = 0$ sonst. Hiermit ergibt sich die folgende Differenzenapproximation für (4).

$$(5) \qquad \begin{aligned} &(\Delta^{(4)} + \Delta^{(2)})u - \lambda[4x\Delta^{(3)} + (2 - 3x^2)\Delta^{(2)} - 4x\Delta^{(1)}]u = f, \\ &u(0) = \gamma^{(1)}, \quad u(1) = \gamma^{(2)}, \quad (\Delta^{(2)}u)(0) = \gamma^{(3)}, \quad (\Delta^{(2)}u)(1) = \gamma^{(4)}. \end{aligned}$$

Für die vorstehende Differenzenapproximation wollen wir die Voraussetzungen von Satz (4.10) nachweisen. Es ist sofort zu sehen, daß $\Delta^{(j)}$, $j = 1, 2, 3$, die in (Z 2) betrachtete Gestalt hat, wobei $r_j = s_j = 2$ und P_{22} die orthogonale Projektion von H^0 auf den Teilraum der außerhalb von $[0, 1]_h$ verschwindenden Funktionen ist. Wir weisen noch nach, daß für alle $\varphi \in C^3(\mathbb{R}^1)$

$$(6) \qquad \max_{x \in [0,1]_h} |(\Delta^{(j)}\varphi - \partial^j\varphi)(x)| \to 0 \quad (h \in \Lambda_0), \qquad j = 1, 2, 3,$$

konvergiert. Wir beschränken uns dabei auf den Fall $j = 3$. Ist D^3 der vorwärts genommene Differenzenquotient dritter Ordnung, so besitzt $\Delta^{(3)}$ die Darstellung

$$(\Delta^{(3)}u)_j = \begin{cases} [\frac{1}{8}(-E^{-3} + 5E^{-2} + 5E^{-1} - I)D^3u]_j, & j = 1, \ldots, n-1 \\ [\frac{1}{2}(E^{-2} + E^{-1})D^3u]_j, & j = 0, n \\ 0, \quad \text{sonst}, \end{cases}$$

woraus (6) folgt. Um $\Delta^{(4)}$ in die in (Z3) definierte Klasse einzupassen, nehmen wir die Umformung

$$(\Delta^{(4)}u)_j = \begin{cases} (D^4u)(c), & j=0 \\ (D^4u)(d-4h), & j=n \\ [\frac{1}{6}(-E^2+8E-I)E^{-3}D^4u]_j, & j=1,\ldots,n-1 \\ 0, \quad \text{sonst} \end{cases}$$

vor, so daß die dort eingeführten Polynome durch

$$p=\tfrac{1}{6}(-z^2+8z-1), \qquad q_{11}=q_{21}=1$$

gegeben sind. Hieraus berechnet man

$$p_i=z-4+\sqrt{15}, \qquad \left(z^2p(z^{-1})\right)_i=1-(4+\sqrt{15})z,$$

so daß $p(e^{i\xi})\neq 0$, $\xi\in\mathbb{R}^1$, und $p(1)=1$ ist. Offenbar läßt sich zu jedem Polynom t ein Polynom t_1 und eine Zahl μ finden, so daß $t=t_1p_i+\mu\, q_{11}$ ist, und das bedeutet, daß p_i von q_{11} aufgespannt wird. Dasselbe gilt für $\left(z^2p(z^{-1})\right)_i$ und q_{21}, so daß $\Delta^{(4)}$ allen in § 4 geforderten Voraussetzungen genügt. Auch die Randbedingungen in (5) haben die geforderte Gestalt, wie man durch Umschreibung auf die Form

$$(E^2u)(c)=\gamma^{(1)}, \qquad (E^{-2}u)(d)=\gamma^{(2)},$$
$$\tfrac{1}{12}[(-E^2+14E-I)D^2u](c)=\gamma^{(3)}, \qquad \tfrac{1}{12}[(-I+14E^{-1}-E^{-2})\bar{D}^2u](d)=\gamma^{(4)}$$

ersieht, wobei $\bar{D}^2$ für den rückwärts genommenen Differenzenquotient zweiter Ordnung steht.

Die Koeffizienten der Operatoren in (5) ergeben sich durch Restriktion der Koeffizienten von (4) auf das Gitter, d.h. durch Anwendung des Operators R_1 aus § 2.3, so daß alle Voraussetzungen von Satz (4.10) erfüllt sind und wir für das vorliegende Problem die dort aufgestellten Konvergenzaussagen behaupten können.

Für $f_0\in C^4[0,1]$ ist der Abschneidefehler der Lösungen von (4) von der Ordnung $O(h^{5/2})$, $h\in\Lambda_0$, und somit gelten bei Annahme von $|\gamma^{(j)}-\gamma_0^{(j)}|=O(h^{5/2})$, $h\in\Lambda_0$, $j=1,\ldots,4$, auch die Aussagen (6.5) und (6.7) über die Konvergenzordnung der Lösungen und Eigenwerte, wenn man dort $\tau=5/2$ setzt. Ist λ kein Eigenwert von (4), so haben wir z.B.

$$\max_{x\in[0,1]_h}|(u-u_0)(x)|=O(h^{5/2}), \qquad h\in\Lambda,$$

wobei u_0 die Lösung von (4) und $\{u\}$, $h\in\Lambda$, die für fast alle $h\in\Lambda_0$ existierenden Lösungen von (5) sind.

Literatur

1. Aziz, A.K., Hubbard, B.E.: Bounds for the solution of the Sturm-Liouville problem with application to finite difference methods. J. Soc. Indust. Appl. Math. **12**, 163–178 (1964).
2. Babuška, I., Práger, M., Vitásek, E.: Numerical processes in differential equations. Prague: SNTL-Publishers of technical literature 1966.
3. Bückner, H.: Über Konvergenzsätze, die sich bei der Anwendung eines Differenzenverfahrens auf ein Sturm-Liouvillesches Eigenwertproblem ergeben. Math. Z. **51**, 423–465 (1949).

4. Collatz, L.: Konvergenzbeweis und Fehlerabschätzung für das Differenzenverfahren bei Eigenwertproblemen gewöhnlicher Differentialgleichungen zweiter und vierter Ordnung. Deutsche Math. **2**, 189—215 (1937).
5. — Eigenwertprobleme und ihre numerische Behandlung. Leipzig: Akademische Verlagsgesellschaft Becker und Erler Kom.-Ges. 1945.
6. — The numerical treatment of differential equations. Berlin-Heidelberg-New York: Springer 1966.
7. Courant, R., Hilbert, D.: Methods of mathematical physics I. New York: Interscience Publishers, Inc. 1966.
8. Fox, L.: The numerical solution of two-point boundary problems in ordinary differential equations. Oxford: Clarendon Press 1957.
9. Fujita, H.: On the error of the finite difference method in 1-dimensional boundary value problems. J. Phys. Soc. Japan **11**, 160—169 (1956).
10. Gary, J.: Computing eigenvalues of ordinary differential equations by finite differences. Math. of Comp. **19**, 365—386 (1965).
11. Grigorieff, R.D.: Über die Konvergenz von Differenzenverfahren bei linearen gewöhnlichen Differentialgleichungen. ZAMM **48**, T72—T75.
12. — Approximation von Eigenwertproblemen und Gleichungen zweiter Art in Hilbertschen Räumen. Math. Ann. **183**, 45—77 (1969).
13. — Über die Koerzitivität gewöhnlicher Differenzenoperatoren und die Kovergenz von Mehrschrittverfahren. Erscheint in Num. Math.
14. Hao Shou: Homogeneous difference schemes for a fourth order equation with discontinuous coefficients. U.S.S.R. Comp. Math. **3**, 1147—1174 (1963).
15. Hubbard, B.E.: Bouds for eigenvalues of the Sturm-Liouville problem by finite differences. Arch. Rat. Mech. Anal. **10**, 171—179 (1962).
16. Keller, H.B.: On the accuracy of finite difference approximations to the eigenvalues of differential and integral operators. Num. Math. **7**, 412—419 (1965).
17. — Numerical methods for two-point boundary-value problems. Massachusetts-Toronto-London: Blaisdell 1968.
18. Lyusternik, L.A.: On difference approximations of the Laplace operator. Am. Math. Soc. Transl. **8**, 289—351 (1958).
19. Mohr, E.: Über das Sturm-Liouvillesche Eigenwertproblem. Math. Nachr. **7**, 305—322 (1952).
20. Nitsche, J., Nitsche, J.C.C.: Error estimates for the numerical solution of elliptic equations. Arch. Rat. Mech. Anal. **5**, 293—306 (1960).
21. Schröder, J.: Über das Differenzenverfahren bei nichtlinearen Randwertaufgaben. ZAMM **36**, 319—331, 443—455 (1956).
22. Stetter, H.J.: Instability and non-monotonicity phenomena in discretizations to boundary-value problems. Num. Math. **12**, 139—145 (1968).
23. Stummel, F.: Elliptische Differenzenoperatoren unter Dirichletrandbedingungen. Math. Z. **97**, 169—211 (1967).
24. — Lineare elliptische Differenzenoperatoren. Vorlesung WS 1966/67 Universität Frankfurt am Main.
25. Tikhonov, A.N., Samarskii, A.A.: Homogeneous difference schemes. U.S.S.R. Comp. Math. **1**, 5—67 (1962).
26. — — The Sturm-Liouville difference problem. U.S.S.R. Comp. Math. **1**, 939—961 (1962).
27. Thomée, V.: Elliptic difference operators and Dirichlet's problem. Contr. Diff. Equ. **3**, 301—324 (1964).
28. — Westergren, B.: Elliptic difference equations and interior regularity. Num. Math. **11**, 190—210 (1968).

Dr. R.D. Grigorieff
Institut für Angewandte
Mathematik der Universität
6000 Frankfurt a. M.
Robert-Mayer-Straße 10

Numer. Math. 15, 49–56 (1970)

The Reduction of Certain Boundary Value Problems to Variational Problems by Means of Transversality Conditions

HANNO RUND

Received August 16, 1969

Abstract. It is shown by means of the classical theory of the transversality conditions of the calculus of variations that certain boundary value problems are equivalent to necessary conditions for the attainment of extreme values of a fundamental integral of a variational problem with variable boundaries. Systems of second order ordinary, as well as partial, differential equations are considered. The method is illustrated for both cases by means of examples, from which well-known theorems of exceptional practical importance emerge effortlessly.

1. Introduction

In the standard application of direct variational methods to certain types of boundary value problems one endeavours to construct a problem in the calculus of variations which is such that the necessary conditions for the attainment of extreme values of the fundamental integral of this problem are identical with the given boundary value problem. This procedure is motivated by the obvious consideration that, if one has found certain functions which afford extreme values to this fundamental integral (e.g., by the method of Ritz), these functions will in turn represent solutions of the given boundary value problem.

Various methods for the execution of this program are known, such as the minimum quadratic functional theorem (Mikhlin [7]) — also known as the "energy method" — which, however, is limited in its application because of the necessity of the imposition of restrictive assumptions, namely the requirement that the differential operator defining the given boundary value problem be positive definite. Another, rather different, method is described in detail in the treatise of Collatz [4], where it is shown that certain boundary value problems are equivalent to necessary conditions for extrema of integrals independently of the imposition of boundary conditions in respect of the admissible functions of comparison. *This immediately suggests a possible affinity between the method of Collatz and the theory of variational problems with variable boundaries, and thus, in particular, the application of the theory of transversality conditions.*

It is the aim of this note to reveal the relationship between these concepts. The classical theory of transversality conditions is used in the next section to display in Theorem 1 the equivalence of certain systems of second order ordinary differential equations with the necessary conditions for the attainment of extreme values of a suitably constructed integral. This theorem is applied to an example, which indicates how a basic theorem of Collatz ([4], pp. 202–207) emerges in an entirely systematic manner from our analysis. In Section 3 a similar program is carried out for boundary value problems involving systems of second order partial

differential equations (Theorem 2), which again leads to the corresponding theorem of Collatz ([4], p. 428) by specialization.

Needless to say, the method presented here is applicable solely to boundary value problems for which certain conditions are satisfied, as should be evident from the enunciation of Theorems 1 and 2; nevertheless, it is hoped that the treatment serves to shed some light on the relationship between certain boundary value problems and the calculus of variations.

2. Systems of Second Order Ordinary Differential Equations

Let us denote by $x^j(t)$ a system of n functions of a single parameter t. (In the sequel all Latin indices $j, h, \ldots$ run from 1 to n; the summation convention is operative throughout.) We shall be concerned with the boundary value problem defined by the following system of n second order ordinary differential equations:

$$\phi_k(t, x^j, \dot{x}^j, \ddot{x}^j) = 0, \tag{2.1}$$

in which $\dot{x}^j = dx^j/dt$, etc., the boundary conditions being assumed to be of the form:

$$f_k(t, x^j, \dot{x}^j) = 0 \quad \text{for} \quad t = t_1, \qquad h_k(t, x^j, \dot{x}^j) = 0 \quad \text{for} \quad t = t_2, \tag{2.2}$$

where $t_1 < t_2$. The functions ϕ_k, f_k, h_k are given, and it is supposed that they are of class C^1 in all their arguments.

We seek a Lagrange function $L(t, x^j, \dot{x}^j)$ of class C^2, which is such that the system (2.1) may be identified with the corresponding Euler-Lagrange equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}^k}\right) - \frac{\partial L}{\partial x^k} = 0. \tag{2.3}$$

Naturally, this may not always be feasible (cf. Douglas [5]); however, in many practical cases it is possible to construct such a function (as will be assumed here). It is well known that (2.3) represents a set of necessary conditions which a curve Γ: $x^j = x^j(t)$, $t_1 \leqq t \leqq t_2$, in the space R_{n+1} of the variables (t, x^j) must satisfy in order to afford an extreme value to the integral

$$I(\Gamma) = \int_{\Gamma t_1}^{t_2} L(t, x^j, \dot{x}^j)\, dt, \tag{2.4}$$

as compared with that afforded by other curves, of which, however, it is not assumed that they have common end-points $P_1(t_1, x_1^j)$, $P_2(t_2, x_2^j)$, but merely that these end-points lie respectively on a pair of given hypersurfaces of R_{n+1}:

$$s(t, x^j) = 0, \qquad S(t, x^j) = 0. \tag{2.5}$$

For these so-called "variable end-point" problems one has to augment (2.3) by a further set of necessary conditions, namely, the *transversality relations*

$$\frac{\partial L}{\partial \dot{x}^j}\,\delta x^j + \left(L - \frac{\partial L}{\partial \dot{x}^j}\,\dot{x}^j\right)\delta t = 0, \tag{2.6}$$

which must hold at both end-points of Γ for arbitrary displacements $(\delta x^j, \delta t)$ tangential to each of the hypersurfaces (2.5) (cf. Bolza [2], p. 44). On the second of these hypersurfaces, for instance, such displacements must obviously satisfy

the relations

$$\frac{\partial S}{\partial x^j}\,\delta x^j + \frac{\partial S}{\partial t}\,\delta t = 0, \tag{2.7}$$

and these must be consistent with (2.6) at the end-point P_2 of Γ. We shall now endeavour to link such consistency requirements with the second condition of (2.2) for $t = t_2$ (and similarly for $t = t_1$).

To this end we take a special case of (2.5), namely that which corresponds to two hyperplanes $t = t_1$, $t = t_2$ in R_{n+1}. Under these circumstances $\partial S/\partial x^j = 0$, $\partial S/\partial t = 1$, $\partial s/\partial x^j = 0$, $\partial s/\partial t = 1$, and for the consistency of (2.6) with (2.7) it is obviously necessary and sufficient that

$$\frac{\partial L(t, x^j, \dot{x}^j)}{\partial \dot{x}^k} = 0 \quad \text{for} \quad t = t_1, \quad \frac{\partial L(t, x^j, \dot{x}^j)}{\partial \dot{x}^k} = 0 \quad \text{for} \quad t = t_2. \tag{2.8}$$

These conditions are of the type of the given boundary conditions (2.2), but, of course, they will not in general be identical with the latter. We therefore carry out the following construction, which, under favourable circumstances, may lead to the required identification.

Let $G(t, x^j)$ denote some function of class C^2. The Lagrange function defined by

$$L^*(t, x^j, \dot{x}^j) = L(t, x^j, \dot{x}^j) + \frac{\partial G(t, x^j)}{\partial x^k}\,\dot{x}^k + \frac{\partial G(t, x^j)}{\partial t}, \tag{2.9}$$

gives rise to a problem in the calculus of variations which is said to be *equivalent* to that exemplified by (2.4). For, if we form the integral

$$I^*(\Gamma) = \int_{\Gamma t_1}^{t_2} L^*(t, x^j, \dot{x}^j)\,dt, \tag{2.10}$$

we have from (2.9) that

$$I^*(\Gamma) = \int_{\Gamma t_1}^{t_2} L(t, x^j, \dot{x}^j)\,dt + G[t_2, x^j(t_2)] - G[t_1, x^j(t_1)], \tag{2.11}$$

from which it is evident, for the special case of fixed end-points, that if Γ affords an extreme value to I^* it will do so for I and *vice-versa* (Carathéodory [3]). Clearly the Euler-Lagrange equations corresponding to L and L^* are identical. We may thus replace L throughout by L^*, while retaining (2.3); in particular, since it follows from (2.9) that

$$\frac{\partial L^*}{\partial \dot{x}^k} = \frac{\partial L}{\partial \dot{x}^k} + \frac{\partial G}{\partial x^k}, \tag{2.12}$$

we see that instead of the conditions (2.8) we now have the considerably more general relations

$$\begin{aligned} &\frac{\partial L(t, x^j, \dot{x}^j)}{\partial \dot{x}^k} + \frac{\partial G(t, x^j)}{\partial x^k} = 0 \quad \text{for} \quad t = t_1, \\ &\frac{\partial L(t, x^j, \dot{x}^j)}{\partial \dot{x}^k} + \frac{\partial G(t, x^j)}{\partial x^k} = 0 \quad \text{for} \quad t = t_2. \end{aligned} \tag{2.13}$$

Thus: in order that a curve Γ: $x^j = x^j(t)$, with end-points $P_1(t_1, x^j(t_1))$, $P_2(t_2, x^j(t_2))$, in which t_1, t_2 are fixed, but $x^j(t_1)$, $x^j(t_2)$ may be variable, afford an extreme value to the integral (2.11), it is necessary that the functions $x^j(t)$ be

solutions of the Euler-Lagrange equations (2.3) and that the boundary conditions (2.13) be satisfied.

The following is now a direct consequence:

Theorem 1. *If, for a boundary value problem* (2.1), (2.2), *there exist functions* $L(t, x^i, \dot{x}^i)$, $G(t, x^i)$, *such that* (2.1) *and* (2.2) *are respectively equivalent to* (2.3) *and* (2.13), *then the given boundary value problem represents a set of necessary conditions for the integral* (2.11) *to assume an extreme value: here* t_1, t_2 *are fixed, while* $x^i(t_1)$, $x^i(t_2)$ *may be variable.*

We remark, firstly, that our construction immediately entails that in the determination of extreme values of the integral (2.11) the admissible curves of comparison need not satisfy any boundary conditions.

Secondly, although the conditions of the theorem are no doubt fairly restrictive from the point of view of any general theory, there are many important special systems to which the theorem may well be applied. As an illustration of the method, let us briefly glance at the following example (in which $n=1$).

Example. The second order equation

$$-\frac{d}{dt}\{p(t)\dot{x}(t)\}+q(t)\,x(t)=r(t), \tag{2.14}$$

subject to the boundary conditions

$$\alpha_1 x(t_1)+\beta_1\dot{x}(t_1)=\gamma_1, \qquad \alpha_2 x(t_2)+\beta_2\dot{x}(t_2)=\gamma_2, \tag{2.15}$$

in which $p(t)$, $q(t)$, $r(t)$ are given functions, and α_1, α_2, β_1, β_2, γ_1, γ_2 are constants (with $\beta_1 \neq 0$, $\beta_2 \neq 0$).

It is easily seen that a suitable form for L, which identifies (2.14) with (2.3), is given by

$$L(t, x, \dot{x})=p(t)\dot{x}^2+q(t)\,x^2-2r(t)\,x. \tag{2.16}$$

(In fact, this form is suggested by the minimum quadratic functional theorem [7] although some of the conditions of the latter theorem need not be satisfied.) The relations (2.13) thus become identical with (2.15) (as required by the theorem), provided that the function $G(t, x)$ satisfies the requirement

$$\frac{\partial G(t_1, x(t_1))}{\partial x}=\frac{2p(t_1)}{\beta_1}[\alpha_1 x(t_1)-\gamma_1], \tag{2.17}$$

together with a similar relation for $t=t_2$. We thus construct three functions $\alpha(t)$, $\beta(t)$, $\gamma(t)$ which respectively assume the given values α_1, β_1, γ_1 for $t=t_1$, and α_2, β_2, γ_2 for $t=t_2$, while $\beta(t)\neq 0$ for $t_1 \leqq t \leqq t_2$. If we put

$$G(t, x)=p(t)\beta^{-1}(t)[\alpha(t)\,x^2-2\gamma(t)\,x], \tag{2.18}$$

the conditions (2.17) are satisfied. Theorem 1 thus yields the following result:

The given boundary value problem (2.14), (2.15) *represents a set of necessary conditions for the integral*

$$\int_{t_1}^{t_2}[p(t)\dot{x}^2+q(t)\,x^2-2r(t)\,x]\,dt+G[t_2, x(t_2)]-G[t_1, x(t_1)] \tag{2.19}$$

to assume an extreme value, the function G being given by (2.18): *where, however, the admissible curves of comparison need not satisfy any boundary conditions.*

This theorem is given by Collatz ([4], pp. 202—207), where it is derived from first principles for the special system (2.14), (2.15) without reference to transversality conditions. (If, contrary to our assumptions, $\beta_1 = 0$ in (2.15), the above theory is easily modified. For, under these conditions, it follows from (2.15) that the first end-point P_1 is fixed, i.e., $x(t_1)$ is given. Thus the expression $G[t_1, x(t_1)]$ can be omitted from (2.19) as a constant of no significance; but now all admissible curves of comparison must begin at P_1.)

3. Systems of Second Order Partial Differential Equations

Again we shall denote the n dependent functions by $x^i(t^\alpha)$, in which t^α represents the m independent variables. [In the sequel all Latin indices run from 1 to n as before, while all Greek indices assume the values $1, \ldots, m$; the summation convention applies to both sets of indices.] Let G_t denote a finite, simply-connected region in the space R_m of the variables t^α; the hypersurface in R_m which bounds G_t is denoted by ∂G_t, of which it is assumed that it is a piece-wise class C^2 manifold.

Our given boundary value problem is supposed to consist of a system of n second order partial differential equations

$$\Phi_k\left(t^\alpha, x^i, \frac{\partial x^i}{\partial t^\alpha}, \frac{\partial^2 x^i}{\partial t^\alpha \partial t^\beta}\right) = 0 \tag{3.1}$$

on G_t, subject to the boundary conditions

$$F_k\left(t^\alpha, x^i, \frac{\partial x^i}{\partial t^\alpha}\right) = 0 \quad \text{for} \quad t^\alpha \in \partial G^t, \tag{3.2}$$

the functions Φ_k, F_k being assumed to be of class C^1.

The following notation will now be adopted: For any twice differentiable subspace C_m: $x^i = x^i(t^\alpha)$ of the space R_{m+n} of the variables (t^α, x^i) we shall put $\dot{x}^i_\alpha = \partial x^i / \partial t^\alpha$, and if $\psi(t^\alpha, x^i, \dot{x}^i_\alpha)$ is a differentiable function defined on C_m, we shall write

$$\frac{d\psi}{dt^\alpha} = \frac{\partial \psi}{\partial t^\alpha} + \frac{\partial \psi}{\partial x^i}\dot{x}^i_\alpha + \frac{\partial \psi}{\partial \dot{x}^i_\beta}\frac{\partial^2 x^i}{\partial t^\beta \partial t^\alpha}.$$

In accordance with our general program we now seek a class C^2 function $L(t^\alpha, x^i, \dot{x}^i_\alpha)$ such that the Euler-Lagrange equations

$$\frac{d}{dt^\alpha}\left(\frac{\partial L}{\partial \dot{x}^i_\alpha}\right) - \frac{\partial L}{\partial x^i} = 0 \tag{3.3}$$

may be identified (if possible) with the given system (3.1). It is well known that (3.3) represents a set of necessary conditions which a subspace C_m: $x^i = x^i(t^\alpha)$ must satisfy if it is to afford an extreme value to the m-fold integral

$$I(C_m) = \int_{G_t} L(t^\alpha, x^i, \dot{x}^i_\alpha)\, d(t), \tag{3.4}$$

where $d(t)$ represents the volume element of R_m. Again we are concerned with a "variable boundary" problem in the calculus of variations; it is assumed that G_t, and hence ∂G_t, is fixed, but that the values $x^i(t^\alpha)$ may be variable for $t^\alpha \in \partial G_t$. This implies that the boundary of C_m is restricted to lie on the cylindrical sur-

face S in R_{m+n} generated by the lines $t^\alpha = \text{const.}$ through ∂G_t on R_m. If ∂G_t is represented in R_m by the equation $\sigma(t^\alpha) = 0$, the components of the unit outward normal to S in R_{m+n} are given by

$$(N_1, \ldots, N_m, 0, \ldots, 0) = (N_\alpha, 0), \tag{3.5}$$

where $N_\alpha = (\partial\sigma/\partial t^\alpha)\,|\nabla\sigma|^{-1}$. Thus any displacement $(\delta t^\alpha, \delta x^j)$ at a point P of S is tangential to S provided that

$$N_\alpha \delta t^\alpha = 0, \tag{3.6}$$

while δx^j may be quite arbitrary.

Since we are dealing with a variable boundary problem we must obviously augment the necessary conditions (3.3) for an extreme value of (3.4) by the relevant *transversality conditions*, which we shall here write as (Boerner [1], Rund [8]):

$$p_j^\alpha \delta x^j - H_\beta^\alpha \delta t^\beta = 0, \tag{3.7}$$

where, for the sake of brevity, we have put

$$p_j^\alpha = \frac{\partial L}{\partial \dot{x}_\alpha^j}, \tag{3.8}$$

and

$$H_\beta^\alpha = -L\,\delta_\beta^\alpha + p_j^\alpha \dot{x}_\beta^j. \tag{3.9}$$

Again, the systems (3.6), (3.7) must be consistent for arbitrary δx^j; this is possible if and only if the $(m+1)\times(m+n)$ matrix

$$\begin{pmatrix} -H_\beta^\alpha, & p_k^\alpha \\ N_\beta, & 0 \end{pmatrix}$$

has rank m on S. This, however, implies the existence of parameters μ, μ_α (not all of which can vanish) such that

$$-\mu_\alpha H_\beta^\alpha = \mu N_\beta, \quad \mu_\alpha p_k^\alpha = 0 \quad \text{for} \quad t^\alpha \in \partial G_t. \tag{3.10}$$

Using (3.9) and combining both entries in (3.10) we see that this implies $\mu N_\beta = L\mu_\beta$. Obviously we may assume that $L \neq 0$ for $t^\alpha \in \partial G_t$; thus $\mu \neq 0$, for otherwise $\mu_\beta = 0$ for $\beta = 1, \ldots, m$, which would render (3.10) meaningless. Hence we find, with $\mu_\beta = \mu L^{-1} N_\beta$, that the second of (3.10) yields

$$N_\alpha p_k^\alpha = 0 \quad \text{for} \quad t^\alpha \in \partial G_t. \tag{3.11}$$

Recalling the notation (3.8), we see that (3.11) is the counterpart of (2.8). Again, since this is too restrictive in general for identification with (3.2), we introduce once more a so-called equivalent problem by defining a Lagrange function of the type

$$L^*(t^\beta, x^j, \dot{x}_\beta^j) = L(t^\beta, x^j, \dot{x}_\beta^j) + \frac{\partial G^\alpha(t^\beta, x^j)}{\partial t^\alpha} + \frac{\partial G^\alpha(t^\beta, x^j)}{\partial x^h}\dot{x}_\alpha^h, \tag{3.12}$$

where $G^\alpha(t^\beta, x^j)$ denotes a set of m functions of class C^1 to be determined presently. By means of the divergence theorem applied to the region G_t in R_m we infer from (3.12) that

$$I^*(C_m) = \int_{G_t} L^*(t^\beta, x^j, \dot{x}_\beta^j)\, d(t) = \int_{G_t} L(t^\beta, x^j, \dot{x}_\beta^j)\, d(t) - \int_{\partial G_t} N_\alpha G^\alpha(t^\beta, x^j)\, d(\tau), \tag{3.13}$$

where $d(\tau)$ represents the $(m-1)$-dimensional volume element of ∂G_t, while N_α now denotes the unit *inward* normal on ∂G_t. [Clearly, for the special case of fixed boundaries for admissible subspaces of comparison, a subspace C_m which affords an extreme value to (3.4), will do so also for (3.13).] We now replace L by L^* in the above analysis, noting that the Euler-Lagrange equations (3.3) may be retained, for they are obviously identical with their counterparts in terms of L^*. Since (3.12) gives

$$\frac{\partial L^*}{\partial \dot{x}^k_\alpha} = \frac{\partial L}{\partial \dot{x}^k_\alpha} + \frac{\partial G^\alpha}{\partial x^k}, \tag{3.14}$$

it follows that we may now replace our consistency conditions (3.11) by

$$N_\alpha\left(\frac{\partial L(t^\beta, x^i, \dot{x}^i_\beta)}{\partial \dot{x}^k_\alpha} + \frac{\partial G^\alpha(t^\beta, x^i)}{\partial x^k}\right) = 0 \quad \text{for} \quad t^\alpha \in \partial G_t. \tag{3.15}$$

Thus: In order that a subspace C_m: $x^i = x^i(t^\alpha)$ afford an extreme value to the integral (3.13) over a fixed region G_t (but irrespective otherwise of any boundary conditions), it is necessary that the functions $x^i(t^\alpha)$ be solutions of the Euler-Lagrange equations (3.3) on G_t, and that the n conditions (3.15) be satisfied for $t^\alpha \in \partial G_t$. Accordingly we infer

Theorem 2. *If, for a boundary value problem* (3.1), (3.2), *there exist functions* $L(t^\beta, x^i, \dot{x}^i_\beta)$, $G^\alpha(t^\beta, x^i)$, *such that* (3.1) *and* (3.2) *are respectively equivalent to* (3.3) *and* (3.15), *then the given boundary value problem represents a set of necessary conditions for the integral* (3.13) *to assume an extreme value: for this variational problem the region* G_t *is fixed, but the admissible subspaces of comparison may be variable for* $t^\alpha \in \partial G_t$.

As a simple illustration of the method we consider the following example, in which m is arbitrary, but $n = 1$.

Example. The second order equation

$$\frac{\partial}{\partial t^\alpha}(A^{\alpha\beta}\dot{x}_\beta) - f(t^\varepsilon, x) = 0, \tag{3.16}$$

on a given region G_t, with

$$B^\beta \dot{x}_\beta + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0 = 0 \quad \text{for} \quad t^\alpha \in \partial G_t, \tag{3.17}$$

in which $\dot{x}_\beta = \partial x/\partial t^\beta$, and where $A^{\alpha\beta}$, f, B^β, $b_{m-1}, \ldots, b_0$ are given functions of t^ε, it being assumed that the $A^{\alpha\beta}$ are symmetric and of class C^1, with $B^1 \neq 0$.

Obviously a suitable Lagrangian is given by

$$L = \tfrac{1}{2}A^{\alpha\beta}\dot{x}_\alpha\dot{x}_\beta + l(t^\varepsilon, x) \tag{3.18}$$

where l is such that $\partial l/\partial x = f$. The conditions (3.15) then assume the form

$$N_\alpha(A^{\alpha\beta}\dot{x}_\beta + \partial G^\alpha/\partial x) = 0, \tag{3.19}$$

and clearly this may be identified as required with (3.17) if

$$\lambda B^\beta = N_\alpha A^{\alpha\beta}, \quad \text{with} \quad \lambda = N_\alpha A^{\alpha 1}/B^1, \tag{3.20}$$

and if the functions G^α are chosen such as to satisfy the single relation

$$N_\alpha G^\alpha = \lambda(b_0 x + \tfrac{1}{2}b_1 x^2 + \cdots + m^{-1}b_{m-1}x^m). \tag{3.21}$$

(Of course, there are many functions $G^{\alpha}(t^{\varepsilon}, x)$ which satisfy this condition.) From Theorem 2 we may now infer the following:

Suppose that the given boundary value problem (3.16), (3.17) *is such that condition* (3.20) *is satisfied. Then this boundary value problem represents a set of necessary conditions for the attainment of extreme values of the integral*

$$\int_{G_t} L\, d(t) - \int_{\partial G_t} \lambda (b_0 x + \tfrac{1}{2} b_1 x^2 + \cdots + m^{-1} b_{m-1} x^m)\, d(\tau), \tag{3.22}$$

in which L is specified by (3.18): *here the admissible functions of comparison are not subject to any boundary conditions.*

For the case that $m = 2$, the boundary conditions (3.17) being given in the form

$$a_3 \dot{x}_{(n)} + a_2 \dot{x}_{(s)} + a_1 x + a_4 = 0 \quad \text{on } \partial G_t, \tag{3.23}$$

where $\dot{x}_{(n)}, \dot{x}_{(s)}$ respectively denote the normal and tangential components of $(\dot{x}_1, \dot{x}_2)$ on ∂G_t, *this result also reduces to a fundamental theorem stated by* Collatz ([4], p. 428).

In conclusion, we remark firstly that the general method outlined in Section 2 can be generalized to the case of fourth order ordinary differential equations by invoking the theory of second order problems in the calculus of variations (Grässer [6]). Secondly, for the case of multiple integral problems, the construction of an equivalent integrand by means of a divergence as displayed in (3.12) is by no means unique (cf. [8], pp. 250—258): there are classes of functions other than divergences which can be employed for the same purpose. The use of such alternatives would necessitate an amendment of the enunciation of Theorem 2 in respect of the form of the integral (3.13); however, there are very definite indications that for certain types of boundary value problems the introduction of such alternatives may entail significant advantages.

References

1. Boerner, H.: Über die Legendresche Bedingung und die Feldtheorien in der Variationsrechnung der mehrfachen Integrale. Math. Z. **46**, 720—742 (1940).
2. Bolza, O.: Vorlesungen über Variationsrechnung. Leipzig: Koehler u. Amelang 1909; reprint (1949).
3. Carathéodory, C.: Variationsrechnung und partielle Differentialgleichungen erster Ordnung. Leipzig und Berlin: Teubner 1935.
4. Collatz, L.: The numerical treatment of differential equations, 3rd ed. Berlin-Göttingen-Heidelberg: Springer 1960.
5. Douglas, J.: Solution of the inverse problem in the calculus of variations. Trans. Amer. Math. Soc. **50**, 71—128 (1941).
6. Grässer, H. S. P.: Boundary value problems as higher order problems in the calculus of variations. Tydskr. Natuurw. **9**, No. 3 (1969).
7. Mikhlin, S. G.: Variational methods in mathematical physics. London: Pergamon 1964.
8. Rund, H.: The Hamilton-Jacobi theory in the calculus of variations. London and New York: Van Nostrand 1966.

Prof. Hanno Rund
Dept. of Applied Mathematics
Faculty of Mathematics
University of Waterloo
Waterloo, Ontario
Canada

Numer. Math. 15, 57–60 (1970)

Properties of a Matrix with a Very Ill-conditioned Eigenproblem

AXEL RUHE

Received July 16, 1969

Abstract. It is proved that a matrix that has a very ill-conditioned eigenvector matrix is close to one that has multiple eigenvalues, and an estimate of this distance is given, measured in the Euclidean matrix norm.

1. Introduction and Summary

We are going to derive a theorem that pertains to the solution of the eigenproblem for matrices for which this problem is very ill-conditioned.

It is a well known fact that a matrix which is close to one with multiple eigenvalues has an ill-conditioned eigenproblem, ([4], Ch. 2, [3]). Our first theorem will show a converse to this, that if a matrix has an ill-conditioned eigenproblem in some respects, it is also close to a matrix having multiple eigenvalues. The ill-conditioning can be indicated by nearly orthogonal left and right eigenvectors ([4], Ch. 2). A matrix having some very close eigenvalues is, of course, close to a matrix with multiple eigenvalues, but our theorem establishes that even if a matrix has well separated eigenvalues it is close to one having multiple eigenvalues, in the case that it has a very ill-conditioned eigenvector matrix ([4], p. 90). It is to be noted that we make no distinction between defective and derogatory cases, this because this distinction is not very sharp numerically ([3], [1], p. 34).

2. Notation and Preliminaries

Let $\boldsymbol{A}$ be an $n \times n$ complex matrix. The singular values $\sigma_k(\boldsymbol{A})$ are the roots of $\boldsymbol{A}^H\boldsymbol{A}$ ordered so that $\sigma_1 \geqq \sigma_2 \cdots \geqq \sigma_r \geqq 0$. We will use the Euclidean vector norm $\|\boldsymbol{x}\|_2 = (\Sigma\,|x_i|^2)^{\frac{1}{2}}$, and the matrix norms

$$\|\boldsymbol{A}\|_2 := \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \sigma_1(\boldsymbol{A}),$$

$$\|\boldsymbol{A}\|_E := \mathrm{Tr}\,(\boldsymbol{A}^H\boldsymbol{A}) = (\Sigma\,|a_{jk}|^2)^{\frac{1}{2}}.$$

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are arbitrary vectors, we denote their linear span by $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k]$. If S is a linear subspace of the n-dimensional unitary space and $\boldsymbol{x}$ is a unit vector ($\|\boldsymbol{x}\|_2 = 1$) then we define the angle between the vector and the subspace by [3]

$$\sin\theta(\boldsymbol{x}, S) := \min_{\boldsymbol{y}\in S} \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{2.1}$$

Let $\boldsymbol{\Lambda} := \mathrm{diag}\,(\lambda_i)$ be the matrix of eigenvalues of $\boldsymbol{A}$. If $\boldsymbol{A}$ is diagonalizable, then for an appropriate matrix $\boldsymbol{H}$,

$$\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{-1}. \tag{2.2}$$

The columns of $\boldsymbol{H}$ (denoted by $\boldsymbol{x}_k$) are right eigenvectors of $\boldsymbol{A}$, and the rows of $\boldsymbol{H}^{-1}$ (denoted by $\boldsymbol{y}_k^T$) are left eigenvectors of $\boldsymbol{A}$. The condition numbers $s_k(\boldsymbol{A})$ are defined by $s_k(\boldsymbol{A}) := \boldsymbol{y}_k^T \boldsymbol{x}_k$, where $\boldsymbol{y}_k$ and $\boldsymbol{x}_k$ are scaled so that $\|\boldsymbol{y}_k\|_2 = \|\boldsymbol{x}_k\|_2 = 1$. Properties of these numbers can be studied e.g. in ([4], Ch. 1 and 2).

In particular it can be demonstrated that

$$|s_k(\boldsymbol{A})| = \sin\theta(\boldsymbol{x}_k, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k+1}, \ldots, \boldsymbol{x}_n]). \tag{2.3}$$

For every matrix $\boldsymbol{A}$ there exists a unitary matrix $\boldsymbol{U}$, that transforms it into the Schur triangular form, $\boldsymbol{T}$,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^H = \boldsymbol{U}(\boldsymbol{\Lambda} + \boldsymbol{M})\boldsymbol{U}^H. \tag{2.4}$$

$\boldsymbol{M}$ is strictly upper triangular, and if $\boldsymbol{M} = 0$, $\boldsymbol{A}$ is a normal matrix. Further the eigenvalues can be made to appear in any desired order along the diagonal of $\boldsymbol{T}$.

Finally let $\mathcal{S}$ be any set of matrices. We define the distance with respect to the norm ν between a matrix $\boldsymbol{A}$ and the set $\mathcal{S}$ by

$$\operatorname{dist}_\nu(\boldsymbol{A}, \mathcal{S}) := \inf_{\boldsymbol{X} \in \mathcal{S}} \|\boldsymbol{A} - \boldsymbol{X}\|_\nu.$$

3. The Theorem

We shall now study the distance between a given matrix and the set $\mathcal{M}$ of matrices having at least two coinciding eigenvalues. We state the following theorem which points out one case when a matrix is close to $\mathcal{M}$.

Theorem 1. Let $\boldsymbol{A}$ be a $n \times n$ matrix with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and define

$$d := \max |\lambda_i - \lambda_j|. \tag{3.1}$$

Then

$$\operatorname{dist}_E(\boldsymbol{A}, \mathcal{M}) \leqq \frac{n}{4} d \tan\theta \tag{3.2}$$

where

$$\sin\theta \leqq \min_k |s_k(\boldsymbol{A})|^{\frac{1}{n-1}}.$$

Proof. Starting with $i = n$, order the normalized eigenvectors $\boldsymbol{x}_i$ so that

$$\theta(\boldsymbol{x}_i, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}]) = \min_{1 \leqq j \leqq i} \theta(\boldsymbol{x}_j, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_i]), \tag{3.3}$$

$$i = n, n-1, \ldots, 1.$$

Now we can write the eigenvector matrix $\boldsymbol{H} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ as a product of a unitary and an upper triangular matrix with positive diagonal elements ([2], p. 7)

$$\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{R} \tag{3.4}$$

and we have a Schur triangular form $\boldsymbol{T}$ of $\boldsymbol{A}$ (2.4) defined by

$$\boldsymbol{T} := \boldsymbol{Q}^H \boldsymbol{A} \boldsymbol{Q} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{-1}. \tag{3.5}$$

We note that

$$r_{ii} = \sin\theta(\boldsymbol{x}_i, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}]). \tag{3.6}$$

(2.3) and the choice (3.3) imply that

$$1 = r_{11} \geqq r_{22} \geqq \cdots \geqq r_{nn} = |s_n(\boldsymbol{A})| = \min_k |s_k(\boldsymbol{A})|. \tag{3.7}$$

We are now going to study the superdiagonal elements of $\boldsymbol{T}$, relate them to the eigenvalues and eigenvectors, and show that, after reordering some of the eigenvalues, there exists an element $t_{k\,k+1}$ that is not small. This implies that a small perturbation $\boldsymbol{E}$ is capable of producing a double eigenvalue in $\boldsymbol{T}+\boldsymbol{E}$.

Define $\theta > 0$ by

$$\sin\theta := \min_i r_{i+1\,i+1}/r_{ii}. \tag{3.8}$$

Then (3.7) implies that

$$\sin\theta \leqq \left(\frac{r_{22}}{r_{11}} \cdot \ldots \cdot \frac{r_{nn}}{r_{n-1\,n-1}}\right)^{\frac{1}{n-1}} = |s_n(\boldsymbol{A})|^{\frac{1}{n-1}}.$$

Choose k so that $r_{k+1\,k+1}/r_{kk} = \sin\theta$, and partition the matrices after the k-th row and column. The Eq. (3.5) will then read

$$\begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} \\ 0 & \boldsymbol{T}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ 0 & \boldsymbol{R}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_1 & 0 \\ 0 & \boldsymbol{\Lambda}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{R}_{11}^{-1} & -\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{R}_{22}^{-1} \\ 0 & \boldsymbol{R}_{22}^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{R}_{11}\boldsymbol{\Lambda}_1\boldsymbol{R}_{11}^{-1} & -\boldsymbol{R}_{11}(\boldsymbol{\Lambda}_1\boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12} - \boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}\boldsymbol{\Lambda}_2)\boldsymbol{R}_{22}^{-1} \\ 0 & \boldsymbol{R}_{22}\boldsymbol{\Lambda}_2\boldsymbol{R}_{22}^{-1} \end{bmatrix}.$$

Defining $\boldsymbol{Y}_{12} := \boldsymbol{R}_{11}^{-1}\boldsymbol{R}_{12}$ we see that

$$t_{k\,k+1} = -r_{kk}\, y_{k\,k+1}(\lambda_k - \lambda_{k+1})\, r_{k+1\,k+1}^{-1}. \tag{3.9}$$

We denote the first column of $\boldsymbol{Y}_{12}$ by $\boldsymbol{y}_{k+1}$, and of $\boldsymbol{R}_{12}$ by $\boldsymbol{r}_{k+1}$. If we define the angle φ by

$$\sin\varphi := r_{k+1\,k+1} \tag{3.10}$$

then

$$\|\boldsymbol{r}_{k+1}\|_2 = \cos\varphi,$$

$$\|\boldsymbol{y}_{k+1}\|_2 \geqq \|\boldsymbol{R}_{11}\|_2^{-1}\,\|\boldsymbol{r}_{k+1}\|_2 \geqq \frac{1}{\sqrt{k}}\cos\varphi$$

since the columns of $\boldsymbol{R}$ are normalized. Then there is an index j so that

$$|y_{j\,k+1}| \geqq \frac{1}{k}\cos\varphi.$$

A reordering of the eigenvalues in the Schur triangular form (2.4) can be expressed in the following way. The diagonal matrix $\boldsymbol{\Lambda}$ is replaced by $\boldsymbol{\Lambda}' = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$, where $\boldsymbol{P}$ is a permutation matrix and

$$\boldsymbol{T} = \boldsymbol{R}\boldsymbol{P}^T\boldsymbol{\Lambda}'\boldsymbol{P}\boldsymbol{R}^{-1}.$$

Now we triangularize again (3.4),

$$\boldsymbol{R}\boldsymbol{P}^T = \boldsymbol{Q}'\boldsymbol{R}' \tag{3.11}$$

and get the new Schur form

$$\boldsymbol{T}' := \boldsymbol{Q}'^H\boldsymbol{T}\boldsymbol{Q}' = \boldsymbol{R}'\boldsymbol{\Lambda}'\boldsymbol{R}'^{-1}.$$

If $\boldsymbol{P}$ involves only the first k eigenvalues, then, by partitioning (3.11), we see that

$$\begin{aligned}\boldsymbol{Y}'_{12} &= \boldsymbol{R}'^{-1}_{11}\boldsymbol{R}'_{12} = (\boldsymbol{Q}'^{H}_{11}\boldsymbol{R}_{11}\boldsymbol{P}^{T}_{11})^{-1}\boldsymbol{Q}'^{H}_{11}\boldsymbol{R}_{12}\\ &= \boldsymbol{P}_{11}\boldsymbol{R}^{-1}_{11}\boldsymbol{Q}'_{11}\boldsymbol{Q}'^{H}_{11}\boldsymbol{R}_{12} = \boldsymbol{P}_{11}\boldsymbol{Y}_{12}.\end{aligned}$$

Therefore we can choose $\boldsymbol{P}$ as $\boldsymbol{I}_{jk}$, the elementary permutation matrix ([2], p. 4) making

$$|y'_{k\,k+1}| = |y_{j\,k+1}| \geqq \frac{1}{k}\cos\varphi.$$

Now $\boldsymbol{HP}^{T} = \boldsymbol{QQ}'\boldsymbol{R}'$, and thus (3.6) applied to the new factorization gives

$$r'_{kk} = \sin\theta\,(\boldsymbol{x}_j, [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j+1}, \ldots, \boldsymbol{x}_k]).$$

The choice (3.3) will then imply $r'_{kk} \geqq r_{kk}$. The last n-k rows of $\boldsymbol{R}$ are not affected by the permutation so $r'_{k+1\,k+1} = r_{k+1\,k+1}$ and $\lambda'_{k+1} = \lambda_{k+1}$. We observe that $\lambda'_k = \lambda_j$, so the formula (3.9) applied to the new factorization gives the estimate

$$\begin{aligned}|t'_{k\,k+1}| &= |\lambda'_k - \lambda'_{k+1}|\,|y'_{k\,k+1}|\,r'_{kk}/r'_{k+1\,k+1}\\ &\geqq |\lambda_j - \lambda_{k+1}|\,\frac{\cos\varphi}{k}\,\frac{r_{kk}}{r_{k+1\,k+1}} \geqq \frac{|\lambda_j - \lambda_{k+1}|}{k}\operatorname{cotan}\theta\end{aligned} \tag{3.12}$$

since $\varphi \leqq \theta$, by (3.8) and (3.10).

By adding a matrix $\boldsymbol{E}$, being zero except for an element e in the $(k+1, k)$ position to $\boldsymbol{T}'$, the characteristic polynomial receives a quadratic factor

$$\det\begin{bmatrix}\lambda_j - \lambda & t'_{k\,k+1}\\ e & \lambda_{k+1} - \lambda\end{bmatrix}$$

which has the discriminant

$$D = (\lambda_j - \lambda_{k+1})^2 + 4e\,t'_{k\,k+1}.$$

If $\boldsymbol{T}' + \boldsymbol{E}$ has a double eigenvalue, the discriminant is zero, giving

$$e = -\frac{(\lambda_j - \lambda_{k+1})^2}{4t'_{k\,k+1}}.$$

By (3.1) and (3.12),

$$|e| \leqq \frac{n}{4}d\,\tan\theta$$

which proves (3.2).

References

1. Henrici, P.: Bounds for iterates, inverses, spectral variation and fields of values of non-normal matrices. Num. Math. **4**, 24–40 (1962).
2. Householder, A. S.: The theory of matrices in numerical analysis. New York: Blaisdell, 1st Ed., 1964.
3. Varah, J. M.: Computing invariant subspaces of a general matrix when the eigensystem is poorly conditioned, to appear.
4. Wilkinson, J. H.: The algebraic eigenvalue problem. Oxford: Clarendon Press, 1st Ed., 1965.

Dr. A. Ruhe
Lunds Universitet
Institutionen för Informationsbehandling
Sölvegatan 14 A
223 62 Lund (Sweden)

Numer. Math. 15, 61–73 (1970)

Monotonie-Eigenschaften bei Differentialungleichungen über nichtkompaktem Grundbereich

HANNS-JÖRG STOSS

Eingegangen am 14. Mai 1969

1. Einleitung

1.1. Problemstellung und Inhalt

Schröder hat in [3] Monotonie-Eigenschaften von Differentialoperatoren untersucht. Er betrachtet Differentialoperatoren F, von höchstens zweiter Ordnung, die gewissen Funktionen u aus der Menge $X(B)$ aller reellwertigen Funktionen über einem Kompaktum B des m-dimensionalen euklidischen Raumes $\mathbb{R}^m$ eine Funktion $F[u] \in X(B)$ zuordnen, und leitet hinreichende Bedingungen her, daß man für geeignete Funktionen $u, v \in X(B)$ den folgenden Schluß ziehen kann:

Aus

$$F[u](x) \leqq F[v](x) \quad (x \in B) \tag{1.1}$$

folgt

$$u(x) \leqq v(x) \qquad (x \in B). \tag{1.2}$$

In der vorliegenden Arbeit wird gezeigt, wie man durch einige Änderungen an den Überlegungen von Schröder unter recht allgemeinen Voraussetzungen Aussagen dieser Art auch dann erhalten kann, wenn man auf die Voraussetzung der Kompaktheit des Grundbereiches B verzichtet. Die hier abgeleiteten Ergebnisse gelten also für den Fall eines unbeschränkten Grundbereiches B, etwa für Außenraumaufgaben der Potentialtheorie oder für das Problem der Wärmeleitung in unendlich ausgedehnten Medien. Ferner können sie auf Differentialoperatoren angewendet werden, für die wegen ihres singulären Verhaltens am Rand von B mit den bisherigen Mitteln keine Monotonieaussagen möglich waren.

Wir erhalten Aussagen der Form:

Aus

$$F[u](x) \leqq F[v](x) \quad (x \in B)$$

und

$$\tau(u - v) \leqq 0 \tag{1.3}$$

folgt

$$u(x) \leqq v(x) \qquad (x \in B).$$

Dabei ist $\tau(u - v)$ eine Größe, die das asymptotische Verhalten der Funktion $u - v$ bei Annäherung an die nicht zu B gehörenden Randpunkte von B, bzw. für $|x| \to \infty$ beschreibt.

Schröder geht aus von einem halbgeordneten linearen Raum $(R, \leqq)$. Er beweist zunächst einen abstrakten Monotoniesatz und benutzt dabei, daß in R

ein Element z mit $0<z$ existiert, wobei $<$ eine weitere, strengere Ordnungsrelation in R bedeutet (s. Bemerkung 2.1). Wählt man bei der Anwendung auf Differentialoperatoren als $(R, \leqq)$ etwa den Raum der stetigen Funktionen über einem Bereich B mit der natürlichen Halbordnung, so ergibt sich, daß ein solches z genau dann existiert, wenn B kompakt ist. Schröder benutzt die Kompaktheit von B wesentlich nur, um die Existenz solcher im strengen Sinne positiven Funktionen zu sichern.

In Abschnitt 2 der vorliegenden Arbeit wird gezeigt, daß sich die Aussage des Monotoniesatzes von Schröder unter Benutzung einer wesentlich schwächeren Ordnungsrelation $<$ herleiten läßt, so daß die Voraussetzung der Kompaktheit von B entfallen kann.

In Abschnitt 3 wird daraus ein Monotoniesatz für Differentialoperatoren über nichtkompaktem Grundbereich abgeleitet und die Forderung (1.3) an das asymptotische Verhalten der zu vergleichenden Funktionen präzisiert.

Hinreichende Bedingungen für die Gültigkeit der Voraussetzungen dieses Satzes werden in Abschnitt 4 angegeben, wobei aus Gründen der Darstellung bewußt auf mögliche Abschwächungen der Forderungen an die Differentialoperatoren verzichtet wurde.

In Abschnitt 5 werden die Anwendungsmöglichkeiten an einfachen Beispielen gezeigt.

1.2. Literatur

Monotonie-Eigenschaften von Differentialoperatoren über nichtkompaktem Grundbereich scheinen bisher nur wenig untersucht worden zu sein. Teilergebnisse finden sich etwa bei Collatz [1], Redheffer [2] und Walter [6].

Für lineare Differentialoperatoren zweiter Ordnung hat der Verfasser in [5] Ergebnisse erhalten, die sowohl in der theoretischen Aussage, als auch in der praktischen Anwendbarkeit über das hier dargestellte hinausgehen. Bei diesen Überlegungen wird jedoch die Linearität der Operatoren so stark ausgenutzt, daß bei einer Ausdehnung auf den nichtlinearen Fall diese Vorteile wieder verloren gehen. Daher ist zur Behandlung von nichtlinearen Operatoren vor allem auf Grund der einfacheren Anwendbarkeit wohl den Ergebnissen der vorliegenden Arbeit der Vorzug zu geben.

In einer neueren Arbeit [4] hat Schröder für den Fall kompakter Bereiche die in [3] gestellten Forderungen abgeschwächt und weiterreichende Ergebnisse erhalten. Wie sich die dort verwendeten Methoden auf den Fall nichtkompakter Grundbereiche übertragen lassen, wird hier nicht untersucht.

2. Ein Ergebnis für abstrakte Operatoren

Es sei $R=\{u, v, \ldots\}$ ein reeller linearer Raum, in dem eine Ordnungsrelation $\leqq$ erklärt ist. Wir verlangen:

a) Aus $u \leqq v$, $v \leqq w$ folgt $u \leqq w$,

b) aus $u \leqq v$ folgt $u+w \leqq v+w$ für jedes $w \in R$,

c) aus $u \leqq v$ folgt $\lambda u \leqq \lambda v$ für jede Zahl $\lambda \geqq 0$.

d) Sind u, v feste Elemente, so daß $0 \leqq u+\frac{1}{n} v$ für jedes natürliche n gilt, so ist $0 \leqq u$.

Mit einem festen Element $p \geqq 0$, $p \neq 0$ definieren wir eine zweite Ordnungsrelation $u \underset{p}{<} v$ durch die folgende Festsetzung.

e) Es ist $u \underset{p}{<} v$ genau dann, wenn eine natürliche Zahl n existiert mit $p \leqq n(v-u)$.

$S = \{U, V, \ldots\}$ sei eine Menge, in der ebenfalls zwei Ordnungsrelationen $\leqq$, $<$ erklärt sind. Wir fordern nur:

$$\begin{array}{lll} \text{Aus} & U_1 \leqq U_2 < U_3 & \text{folgt} \quad U_1 < U_3, \\ \text{aus} & U_1 < U_2 \leqq U_3 & \text{folgt} \quad U_1 < U_3. \end{array} \tag{2.1}$$

M sei ein auf einer Teilmenge $D \subset R$ definierter Operator mit Werten in S.

Dann gilt der folgende

Satz 1. $v \in D$ sei fest gewählt.

I. Es gebe ein Element $z \in R$, sowie Zahlen $\alpha > 0$, $\gamma > 0$ mit folgenden Eigenschaften:

i) $z \leqq \alpha p$ (p aus Definition von $\underset{p}{<}$),

ii) $v + \lambda z \in D$ $(0 \leqq \lambda \leqq \gamma)$,

iii) $Mv < M(v + \lambda z)$ $(0 < \lambda \leqq \gamma)$.

II. Für beliebige $w, \tilde{w} \in D$ gelte:

$$\text{Aus} \quad w \leqq \tilde{w} \quad \text{und} \quad Mw < M\tilde{w} \quad \text{folgt} \quad w \underset{p}{<} \tilde{w}. \tag{2.2}$$

Ist dann $u \in D$ ein Element mit

$$u \leqq v + \gamma z \tag{2.3}$$

und

$$Mu \leqq Mv, \tag{2.4}$$

so ist

$$u \leqq v. \tag{2.5}$$

Beweis. $u \in D$ erfülle die Forderungen (2.3) und (2.4). Wir setzen $\Lambda = \{\lambda \geqq 0 : u \leqq v + \lambda z\}$. Dann ist (2.5) offenbar äquivalent mit $0 \in \Lambda$. Wegen (2.3) ist $\Lambda \neq \emptyset$ und somit existiert $\lambda_0 = \inf \Lambda \leqq \gamma$. Wir zeigen zunächst $\lambda_0 \in \Lambda$. Dazu können wir annehmen, daß eine Folge $\{\lambda_k\}$ existiert mit $\lambda_0 < \lambda_k \leqq \gamma$, $\lambda_k \to \lambda_0$, so daß für alle k gilt $v + \lambda_k z - u \geqq 0$. Dann ist nach I, i) auch

$$v + \lambda_0 z - u + (\lambda_k - \lambda_0)\alpha p = v + \lambda_k z - u + (\lambda_k - \lambda_0)(\alpha p - z) \geqq 0.$$

Da $p \geqq 0$ folgt hieraus nach Eigenschaft d) der Ordnungsrelation

$$v + \lambda_0 z - u \geqq 0, \quad \text{d.h.} \quad \lambda_0 \in \Lambda. \tag{2.6}$$

Ist $\lambda_0 = 0$, so ist $u \leqq v$.

Ist $\lambda_0 > 0$, so gilt nach (2.6) und I, iii)

$$Mu < M(v + \lambda_0 z).$$

Zusammen mit (2.6) ist also nach Voraussetzung II

$$u \underset{p}{<} v + \lambda_0 z.$$

Es existiert also eine natürliche Zahl n mit

$$\frac{1}{n} p \leqq v + \lambda_0 z - u. \tag{2.7}$$

Da $\lambda_0 > 0$ existiert eine Zahl λ_n mit $0 \leqq \lambda_n < \lambda_0$ und $(\lambda_0 - \lambda_n)\alpha \leqq 1/n$.

Dafür gilt nach (2.7)

$$\begin{aligned} v + \lambda_n z - u &= v + \lambda_0 z - u - (\lambda_0 - \lambda_n) z \\ &\geqq \frac{1}{n} p - (\lambda_0 - \lambda_n)\alpha p + (\lambda_0 - \lambda_n)(\alpha p - z) \\ &= \left(\frac{1}{n} - (\lambda_0 - \lambda_n)\alpha\right) p + (\lambda_0 - \lambda_n)(\alpha p - z). \end{aligned}$$

Hier sind beide Summanden $\geqq 0$, also $v + \lambda_n z - u \geqq 0$ oder $u \leqq v + \lambda_n z$. Da $\lambda_n < \lambda_0$ ist dies ein Widerspruch zur Minimaleigenschaft von λ_0. Also ist $\lambda_0 = 0$ und somit $0 \in \Lambda$.

Bemerkung 2.1. Der wesentliche Unterschied zu dem von Schröder aufgestellten Satz ([3], Satz 1) besteht in der hier schwächeren Forderung an die Relation $<$. Schröder verlangt statt e):

e') Es ist $u < v$ genau dann, wenn zu *jedem* $w \in R$ ein n existiert mit $w \leqq n(v - u)$. Hier wird diese Ungleichung nur für das feste Element $w = p \geqq 0$ gefordert. Ist $p > 0$ im Sinne von e'), so fallen die Relationen $u < v$ und $u \underset{p}{<} v$ zusammen.

Bemerkung 2.2. Anders als im Satz von Schröder darf Bedingung (2.3) hier nicht entfallen, selbst wenn I,ii) für alle $\gamma \geqq 0$ erfüllt werden kann.

Bemerkung 2.3. Im Beweis wird II nur für $w = u$ und $\tilde{w} = v + \lambda z$ benutzt. Bei gegebenem u und v braucht also II nur für diese w, $\tilde{w}$ nachgewiesen zu werden.

Bemerkung 2.4. Durch Umkehrung der Ungleichungen in dem eben aufgestellten Monotoniesatz lassen sich wie bei Schröder ([3], Bemerkung 1,3 ff.) auch für diesen Fall Eindeutigkeits- und Einschließungssätze gewinnen. Da hierzu keine neuen Ideen nötig sind, sei auf eine genaue Formulierung verzichtet.

3. Anwendung auf Differentialoperatoren

Es sei $G \subset \mathbb{R}^m$ ($m = 1, 2, \ldots$) ein nicht notwendig beschränktes Gebiet. Sein Rand Γ ($\Gamma = \emptyset$ zugelassen) sei disjunkt zerlegt in $\Gamma = \Gamma^1 \cup \Gamma^2$. Wir setzen $B = G \cup \Gamma^1$, wobei wir annehmen, daß B nicht kompakt ist. Es bezeichne $X(B)$ die Menge aller auf B definierten reellwertigen Funktionen und F einen auf einer Teilmenge $D_F \subset X(B)$ definierten Differentialoperator höchstens zweiter Ordnung. Wir wollen sogar $D_F \subset C_0(B)$ annehmen[1]. B werde ausgeschöpft durch kompakte Mengen B_n:

$$\emptyset \neq B_1 \subset B_2 \subset B_3 \subset \cdots \subset \bigcup_{n=1}^{\infty} B_n = B.$$

Wir setzen

$$\tilde{B}_n = B - B_n \quad (\neq \emptyset).$$

1 Genaue Regularitätsforderungen an B und an die Funktionen aus D_F sind in [3], Abschnitt 3.1 angegeben.

Mit einer Funktion $p \in C_0(B)$, $p(x) > 0$ $(x \in B)$ bilden wir für $u \in X(B)$

$$\sigma_p(u) = \lim_{n \to \infty} \left\{ \inf_{x \in \tilde{B}_n} \frac{u(x)}{p(x)} \right\}$$

$$\tau_p(u) = \lim_{n \to \infty} \left\{ \sup_{x \in \tilde{B}_n} \frac{u(x)}{p(x)} \right\}.$$

Für σ, τ sind die Werte $\pm \infty$ zugelassen.

Wir notieren einige Eigenschaften von σ und τ als

Hilfssatz 2

i) $\sigma_p(u) \leqq \tau_p(u)$,

ii) $\sigma_p(u + \lambda p) = \sigma_p(u) + \lambda, \quad \tau_p(u + \lambda p) = \tau_p(u) + \lambda$.

iii) Gilt für zwei auf B stetige Funktionen u, v

$$u(x) < v(x) \quad (x \in B) \quad \text{und} \quad \tau_p(u) < \sigma_p(v),$$

so existiert ein $\alpha > 0$, so daß

$$\alpha p(x) \leqq v(x) - u(x) \quad (x \in B)$$

gilt.

Dann gilt

Satz 3. Es sei $v \in D_F$ fest gewählt.

I. Es existiere eine Funktion $z \in C_0(B)$ und eine Zahl $\gamma > 0$ mit folgenden Eigenschaften:

i) $z(x) > 0 \qquad (x \in B)$,

ii) $v + \lambda z \in D_F \quad (0 \leqq \lambda \leqq \gamma)$,

iii) $F[v](x) < F[v + \lambda z](x) \quad (x \in B,\ 0 < \lambda \leqq \gamma)$.

II. Für beliebige w, $\tilde{w} \in D_F$ gelte:

Aus $w(x) \leqq \tilde{w}(x) \quad (x \in B)$ und $F[w](x) < F[\tilde{w}](x) \quad (x \in B)$ folgt

$$w(x) < \tilde{w}(x) \quad (x \in B).$$

Ist dann $u \in D_F$ eine Funktion mit

$$u(x) \leqq v(x) + \gamma z(x) \quad (x \in B) \tag{3.1}$$

und

$$F[u](x) \leqq F[v](x) \qquad (x \in B) \tag{3.2}$$

und

$$\tau_z(u - v) \leqq 0, \tag{3.3}$$

so gilt

$$u(x) \leqq v(x) \qquad (x \in B).$$

Zum Beweis benutzen wir Satz 1. Wir wählen $R := X(B)$ mit der natürlichen Halbordnung $\leqq$; d.h. $u \leqq v$ genau dann, wenn $u(x) \leqq v(x)$ für alle $x \in B$ gilt. Die Relation $\underset{p}{<}$ definieren wir dabei mit der Funktion $p := z$ mit dem z aus Satz 3, I.

Als Bildmenge S wählen wir $S=\{U,\ U=(u_1;\ u_2,\ u_3)\}$. Dabei sei $u_1\in X(B)$, $u_2,\ u_3\in\mathbb{R}^1$ mit $u_2\leqq u_3$. Für $u_2,\ u_3$ sind die Werte $\pm\infty$ zugelassen.

Für $U,\ V\in S$ erklären wir:

$$U\leqq V\Leftrightarrow\big(u_1(x)\leqq v_1(x)\ \ (x\in B),\ u_2\leqq v_2,\ u_3\leqq v_3\big),$$
$$U<V\Leftrightarrow\big(u_1(x)<v_1(x)\ \ (x\in B),\ u_3<v_2\big).$$

Als Operator M wählen wir mit z und v aus Satz 3:

$$M[w]=\begin{cases}F[w](x)\\ \sigma_z(w-v)\\ \tau_z(w-v)\end{cases}$$

mit dem Definitionsbereich $D:=D_F$.

Voraussetzung I von Satz 1 ist unmittelbar aus Voraussetzung I, Satz 3 erfüllt, ebenso II von Satz 1 aus II von Satz 3, wenn man noch die Aussage iii) des Hilfssatzes heranzieht.

Die Bedingungen

$$F[u](x)\leqq F[v](x)\quad(x\in B)$$

und

$$\tau_z(u-v)\leqq 0$$

liefern schließlich

$$M[u]\leqq M[v].$$

Damit ist Satz 1 anwendbar, der die behauptete Ungleichung liefert.

Bemerkung 3.1. Die Größen σ_z und τ_z müssen mit in den Operator aufgenommen werden, um aus II nicht nur $w(x)<\tilde{w}(x)$ $(x\in B)$ zu erhalten, sondern sogar $w\underset{z}{<}\tilde{w}$, wie es in Satz 1, II gefordert ist.

Bemerkung 3.2. Aus der Forderung (3.3) $\tau_z(u-v)\leqq 0$ folgt die Existenz einer Zahl $\gamma_0\geqq 0$, für die

$$u(x)\leqq v(x)+\gamma_0 z(x)\quad(x\in B)$$

gilt. Deshalb kann die Forderung (3.1) entfallen, falls man in Voraussetzung I, ii) und iii) γ beliebig groß wählen darf.

Bemerkung 3.3. Die Bemerkungen 2.3 und 2.4 gelten sinngemäß wieder.

Ohne Beweis sei noch notiert, wie Satz 3 abzuändern ist, um eine Abschätzung in der anderen Richtung zu erhalten.

Satz 4. Ersetzen wir in Satz 3 die Forderungen I, ii) durch

$$v-\lambda z\in D_F\qquad(0\leqq\lambda\leqq\gamma)$$

und I, iii) durch

$$F[v](x)>F[v-\lambda z](x)\ \ (x\in B,\ 0<\lambda\leqq\gamma),$$

so erhalten wir die Aussage:

Ist $u \in D_F$ eine Funktion mit

$$u(x) \geqq v(x) - \gamma z(x) \quad (x \in B), \tag{3.4}$$

$$F[u](x) \geqq F[v](x) \quad (x \in B) \tag{3.5}$$

und

$$\sigma_z(u - v) \geqq 0, \tag{3.6}$$

so gilt

$$u(x) \geqq v(x) \quad (x \in B).$$

Bemerkung 3.4. Die Bemerkungen 2.3 und 3.2 gelten sinngemäß wieder.

4. Hinreichende Bedingungen

Es bleibt die Aufgabe bei einem gegebenen Problem, d. h. gegebenem Bereich B, Operator F und gegebener Funktion v die beiden Voraussetzungen von Satz 3 nachzuprüfen. Wir nutzen aus, daß man für $u \in D_F$ $F[u](x)$ auffassen kann als Funktion von x, u, dem Vektor Du der ersten Ableitungen[2] von u und der Matrix D^2u der zweiten Ableitungen:

$$F[u](x) = F(x, u, Du, D^2u).$$

Fassen wir vorübergehend u, Du, D^2u als unabhängige Variable auf, so kann man durch wörtliche Wiederholung der Überlegungen von Schröder in [3], Abschnitt 4.2 und 3.2 zeigen:

Voraussetzung II von Satz 3 ist erfüllt, wenn für jedes $x \in B$ folgende Ungleichungen gelten:

$$F(x, u, Du, D^2u) \geqq F(x, u, Du + \xi, D^2u), \tag{4.1}$$

$$F(x, u, Du, D^2u) \geqq F(x, u, Du, D^2u + W). \tag{4.2}$$

Dabei ist ξ ein Vektor mit nichtnegativen Komponenten, die nur dann nicht verschwinden, wenn die entsprechende Komponente von Du Richtungsableitung ist, und W eine positiv semidefinite Matrix.

Man vergleiche dazu die Formeln (3.9) und (3.10) bei Schröder in Abschnitt 3.1.

Um Voraussetzung I nachzuweisen, ist eine Funktion z mit den geforderten Eigenschaften zu konstruieren. Dabei ist folgendes zu beachten:

Die Funktion z geht in die Größe $\tau_z(u - v)$ ein, die im allgemeinen mit nur geringer Information über $u - v$ abgeschätzt werden muß. Man wird also bemüht sein, im Einzelfall $z(x)$ möglichst „groß" zu wählen. Ein Beispiel soll dies verdeutlichen: Es sei im $\mathbb{R}^1$ $B = [0, \infty)$, $B_n = [0, n]$ und über $u - v$ bekannt

$$u(x) - v(x) = o(x) \quad (x \to \infty).$$

Können wir $z(x) \geqq x$ wählen, so gilt offenbar

$$\tau_z(u - v) = \limsup_{x \to \infty} \frac{u(x) - v(x)}{z(x)} = 0.$$

Ist dagegen etwa $z(x)$ beschränkt, so ist keine Aussage über $\tau_z(u - v)$ möglich.

2 Dabei können auch Richtungsableitungen auftreten. Siehe auch Fußnote 1.

Zur Konstruktion einer solchen Funktion z ergibt sich trivialerweise:

Ist $F(x, u, Du, D^2u)$ für jedes $x \in B$ in u streng monoton wachsend, so erfüllt $z(x) = 1$ die Voraussetzung I von Satz 3 [3].

Leider ist diese Annahme bei vielen Anwendungen nicht gegeben, so daß eine genauere Untersuchung notwendig wird. Es seien deshalb einige Hinweise zur Konstruktion eines z unter schwächeren Annahmen gegeben.

4.1. Konstruktion für parabolische Operatoren

Schröder hat in [3] Abschnitt 5.3 Möglichkeiten zur Konstruktion einer I genügenden Funktion z angegeben. Dabei wird nirgends benutzt, daß der zugrunde liegende Bereich B kompakt sein muß. Diese Ergebnisse gelten also auch im vorliegenden Fall.

4.2. Konstruktion für elliptische Operatoren

Für elliptische Operatoren benutzt Schröder die Beschränktheit von B. Wir müssen deshalb hier etwas anders vorgehen. Wir untersuchen einen einfachen Fall. Im $\mathbb{R}^m$ $(m \geqq 2)$ sei G ein unbeschränktes Gebiet mit dem Rand Γ und $B = G \cup \Gamma$. Der Nullpunkt gehöre nicht zu B. Wir betrachten den folgenden Operator:

$$F[u](x) = \begin{cases} F^1[u](x) = f\left(x, u, \dfrac{\partial u}{\partial x_i}, \dfrac{\partial^2 u}{\partial x_i \partial x_j}\right) & (x \in G) \\ F^2[u](x) = u(x) & (x \in \Gamma). \end{cases}$$

Dabei sei f für alle $x \in G$ und beliebige Werte der anderen Variablen definiert.

Es sei v eine hinreichend glatte auf B definierte Funktion. Wir setzen voraus, daß für positive λ und hinreichend glatte positive Funktionen z für alle $x \in G$ eine Ungleichung

$$\frac{1}{\lambda}\{F^1[v + \lambda z](x) - F^1[v](x)\} \geqq L[z](x) \tag{4.3}$$

gilt mit

$$L[z](x) = -\sum_{i,j}^{m} a_{ij}(x) \frac{\partial^2 z}{\partial x_i \partial x_j} - \sum_{i=1}^{m} b_i(x) \left|\frac{\partial z}{\partial x_i}\right| + c(x) z,$$

wobei die Matrix $(a_{ij}(x))$ für jedes $x \in G$ positiv semidefinit ist.

Dann ist hinreichend für I von Satz 3, daß für eine hinreichend glatte, auf B positive Funktion z

$$L[z](x) > 0 \quad (x \in G) \tag{4.4}$$

erfüllt ist. Ein solches z kann man etwa in dem hier behandelten Fall aus dem Ansatz

$$z(x) = |x|^{-\alpha} + \beta$$

mit $\alpha > 0$, $\beta \geqq 0$ gewinnen. Durch Einsetzen in (4.4) ergibt sich dann als hinreichend, daß für alle $x \in G$ die Ungleichung

$$\begin{aligned} \alpha < \frac{1}{\lambda_0(x)} \left\{ \sum_{i=1}^{m} a_{ii}(x) - 2\lambda_0(x) - |b|(x) \cdot |x| \right\} \\ + \frac{1}{\lambda_0(x)} \left\{ \frac{1}{\alpha}(1 + \beta |x|^{\alpha}) c(x) |x|^2 \right\} \end{aligned} \tag{4.5}$$

3 Es ist natürlich zu prüfen, ob bei gegebenem v für positive λ und dieses z $v + \lambda z \in D_F$. Dies sei hier angenommen.

besteht. Dabei ist $\lambda_0(x)$ der größte Eigenwert der (symmetrischen) Matrix $(a_{ij}(x))$ und $|b|(x)$ die Euklid-Norm des Vektors $b(x)$. Ist $c(x) \geqq 0$, so ist der zweite Term in (4.5) für alle in Frage kommenden α, β nichtnegativ. Wir wählen etwa $\beta = 1$ und verwenden zur Bestimmung von α nur den ersten Summanden von (4.4).

5. Beispiele

An einfachen Beispielen seien die erhaltenen Ergebnisse gezeigt.

Im $\mathbb{R}^m$ $(m \geqq 3)$ betrachten wir den über $B = \{x: |x| \geqq 1\}$ erklärten Differentialoperator

$$F[u](x) = \begin{cases} -\Delta u(x) + q(x)\, u^2(x) & (|x| > 1) \\ u(x) & (|x| = 1) \end{cases} \tag{5.1}$$

mit nichtnegativem $q(x)$.

Als Definitionsbereich D_F wählen wir alle auf $\{x: |x| \geqq 1\}$ stetigen und auf $\{x: |x| > 1\}$ zweimal stetig differenzierbaren Funktionen. F erfüllt offenbar die Monotonieforderung (4.2). (4.1) ist hier irrelevant, da keine Richtungsableitungen auftreten. II von Satz 3 gilt also für F. Für eine feste nichtnegative Funktion $v \in D_F$ ergibt sich als (4.3) entsprechende Ungleichung

$$\frac{1}{\lambda}\{F^1[v + \lambda z](x) - F^1[v](x)\} \geqq -\Delta z(x) + 2q(x)\, v(x)\, z(x). \tag{5.2}$$

Damit wird (4.5) zu

$$\alpha < \{m-2\} + \left\{\frac{1}{\alpha}(1 + \beta |x|^{\alpha})\, 2q(x)\, v(x)\, |x|^2\right\}. \tag{5.3}$$

Der zweite Term ist nach Voraussetzung nichtnegativ.

Dann ist etwa

$$z(x) = |x|^{-\frac{m-2}{2}} + 1$$

eine der Voraussetzung I von Satz 3 genügende Funktion.

Von besonderem Interesse ist der Fall $q(x) = 0$. Wir erhalten dann aus (5.1) den linearen Operator

$$F[u](x) = \begin{cases} -\Delta u(x) & (|x| > 1) \\ u(x) & (|x| = 1). \end{cases}$$

(5.2) geht dann über in

$$\alpha < m - 2,$$

so daß an $v(x)$ keine Vorzeichenbeschränkung nötig ist. Die Aussage von Satz 3 lautet dann für diesen Fall: Sind u, v zwei hinreichend glatte Funktionen mit

$$u(x) \leqq v(x) \qquad (|x| = 1), \tag{5.5}$$

$$-\Delta u(x) \leqq -\Delta v(x) \qquad (|x| > 1), \tag{5.6}$$

$$\lim_{r \to \infty} \left(\sup_{|x| > r} (u(x) - v(x))\right) \leqq 0, \tag{5.7}$$

so gilt für alle x mit $|x| \geqq 1$

$$u(x) \leqq v(x).$$

(5.7) ist dabei äquivalent zu der Bedingung $\tau_z(u-v) \leqq 0$ mit

$$z(x) = |x|^{-\frac{m-2}{2}} + 1.$$

Bei Collatz [1] wird statt (5.7)

$$u(x) - v(x) = o(|x|^{-(m-3)}) \quad (|x| \to \infty)$$

gefordert, während für (5.7)

$$u(x) - v(x) = o(1) \qquad (|x| \to \infty)$$

hinreicht, was außer im Falle $m = 3$ eine wesentlich schwächere Forderung bedeutet.

Dieses Ergebnis wäre auch mit den Überlegungen von Redheffer [2] zu erhalten gewesen. Dies gilt jedoch nicht mehr etwa für den Fall

$$2q(x)\, v(x) \geqq k^2 \quad (k > 0).$$

Wir erhalten hier statt (5.2) schärfer

$$\frac{1}{\lambda}\{F^1[v + \lambda z](x) - F^1[v](x)\} > -\Delta z(x) + k^2 z(x)$$

und können I von Satz 3 mit

$$z(x) = \frac{1}{|x|}\, e^{k|x|}$$

erfüllen. Dann ist hinreichend als Forderung an das asymptotische Verhalten der zu vergleichenden Funktionen:

$$u(x) - v(x) = o\left(\frac{1}{|x|}\, e^{k|x|}\right) \quad (|x| \to \infty).$$

An dem folgenden konkreten Beispiel sei skizziert, wie man praktisch vorgehen kann.

Im $\mathbb{R}^3$ betrachten wir den Differentialoperator

$$F[u](x) := -\Delta u(x) + \frac{12}{|x|^2}\, u(x) + \frac{6}{|x|^4}\, u^2(x) \quad (|x| > 1)$$

und fragen nach einer Abschätzung für eine „abklingende" Lösung u^* von

$$\begin{aligned} F[u^*](x) &= 0 && (|x| > 1), \\ u^*(x) &= 0{,}8\, x_1^2 + x_2^2 + 1{,}2\, x_3^2 && (|x| = 1). \end{aligned}$$

Zunächst suchen wir eine Abschätzung nach oben.

Zur Bestimmung von z erhalten wir mit einer zunächst beliebigen Funktion v die für alle $\lambda > 0$ gültige Ungleichung

$$\frac{1}{\lambda}\{F[v + \lambda z](x) - F[v](x)\} > L_v[z](x) := -\Delta z + \frac{12}{|x|^2}\left(1 + \frac{v(x)}{|x|^2}\right) z. \tag{5.8}$$

Dann ist $z(x)$ als positive Lösung von $L_v[z](x) \geqq 0$ zu bestimmen.

Um den Begriff „abklingend" zu präzisieren zeigen wir zunächst:
Gilt für $u^*(x)$ eine Abschätzung

$$u^*(x) \leqq K|x|^{\alpha-1} \quad (|x| \geqq 1) \tag{5.9}$$

mit irgendwelchen Konstanten $K>0$, $\alpha \geqq 1$, so gilt sogar

$$u^*(x) \leqq c|x|^2 \quad (|x| \geqq 1) \tag{5.10}$$

mit

$$c = \frac{\alpha(\alpha+1)}{12} + 1{,}2.$$

Zum Beweis setzen wir mit diesem c $v(x) := v_1(x) := c|x|^2$. Dann rechnet man nach:

$$F[u^*](x) = 0 \leqq F[v_1](x) \quad (|x| > 1),$$
$$u^*(x) \leqq 1{,}2 \leqq v_1(x) \quad (|x| = 1).$$

Ferner ist $z_1(x) := |x|^\alpha$ eine positive Lösung von

$$L_{v_1}[z](x) \geqq 0 \quad (|x| \geqq 1),$$

so daß $z_1(x)$ I von Satz 3 erfüllt. Mit diesem z_1 gilt nun

$$\tau_{z_1}(u-v) \leqq \lim_{|x|\to\infty} \frac{K|x|^{\alpha-1} - c|x|^2}{|x|^\alpha} \leqq 0.$$

Also gilt nach Satz 3 die Abschätzung (5.10).

Nun können wir eine bessere Schranke angeben:
Wir wählen

$$v(x) := v_2(x) := 1{,}2|x|^{-4}.$$

Dann rechnet man wieder nach:

$$F[u^*](x) = 0 \leqq F[v_2](x) \quad (|x| > 1),$$
$$u^*(x) \leqq 1{,}2 = v_2(x) \quad (|x| = 1).$$

Zu diesem $v = v_2$ ist $z_2(x) := |x|^3$ eine positive Lösung von

$$L_{v_2}[z](x) \geqq 0 \quad (|x| > 1)$$

und wir erhalten

$$\tau_{z_2}(u-v) \leqq \lim_{|x|\to\infty} \frac{c|x|^2 - 1{,}2|x|^{-4}}{|x|^3} = 0.$$

Also gilt nach Satz 3 sogar

$$u(x) \leqq 1{,}2|x|^{-4} \quad (|x| \geqq 1). \tag{5.11}$$

Beide Abschätzungen liefern zusammen:
Gilt für $u^*(x)$ eine Abschätzung[4]

$$u^*(x) \leqq K|x|^\alpha \quad (|x| \geqq 1) \quad \text{mit} \quad \alpha \geqq 0, \quad K > 0,$$

so gilt sogar

$$u^*(x) \leqq 1{,}2|x|^{-4} \quad (|x| \geqq 1). \tag{5.12}$$

4 Man kann sogar zeigen, daß das Bestehen einer Abschätzung $u^*(x) < Ke^{\alpha|x|}$ mit $K>0$, $\alpha>0$ ausreicht.

Daß diese Abschätzung nicht wesentlich verbessert werden kann, zeigt die folgende Abschätzung nach unten.

Die Funktion

$$v(x) := v_3(x) := 0{,}75\,|x|^{-4} + 0{,}05\,|x|^{-10}$$

erfüllt

$$F[u^*](x) = 0 \geqq F[v_3](x) \quad (|x| > 1),$$

$$u^*(x) \geqq 0{,}8 = v_3(x) \qquad (|x| = 1).$$

Zur Konstruktion eines z können wir hier nicht eine zu 4.3 analoge lineare Abschätzung benutzen. Die Funktion

$$z(x) := z_3(x) := |x|^2$$

erfüllt jedoch

$$\frac{1}{\lambda}\{F[v_3](x) - F[v_3 - \lambda z_3](x)\} > 0 \quad (|x| > 1), \tag{5.13}$$

aber nur für $0 < \lambda \leqq \gamma := 1$.

$$\sigma_{z_3}(u^* - v) \geqq 0$$

bedeutet hier

$$\liminf_{|x|\to\infty} \frac{u^*(x)}{|x|^2} \geqq 0.$$

Aber wir müssen hier noch wegen (5.13) die Bedingung (3.4) mit $\gamma = 1$ beachten. Wir erhalten also:

Ist

$$u^*(x) \geqq v_3(x) - |x|^2 \quad (|x| > 1) \tag{5.14}$$

und

$$\liminf_{|x|\to\infty} \frac{u^*(x)}{|x|^2} \geqq 0, \tag{5.15}$$

so ist

$$u^*(x) > v_3(x) = 0{,}75\,|x|^{-4} + 0{,}05\,|x|^{-10} \quad (|x| \geqq 1). \tag{5.16}$$

Präzisieren wir also den Begriff „abklingend" dahingehend, daß $u(x)$ den Abschätzungen (5.9), (5.14) und (5.15) genügen soll, so erhalten wir daraus die schärferen Schranken (5.12) und (5.16). In (5.14) kann man v_3 noch durch eine etwas kleinere Funktion — etwa $\varepsilon|x|^{-4}$ mit beliebigem $\varepsilon > 0$ — ersetzen. Wesentliches kann aber nicht mehr gewonnen werden. Dies sieht man am Beispiel der Funktion $w(x) := 2 - |x|^2$.

Hierfür gilt:

$$F[w](x) > 0 \geqq F[v_3](x) \quad (|x| > 1),$$

$$w(x) = 1 > v_3(x) \qquad (|x| = 1)$$

und sogar

$$w(x) \geqq v(x) - |x|^2 \qquad (|x| \geqq 1)$$

aber

$$\liminf_{|x|\to\infty} \frac{w(x)}{|x|^2} = -1 \ngeqq 0,$$

d.h. (5.15) ist verletzt. Offensichtlich ist für hinreichend große $|x|$-Werte auch

$$w(x) < v_3(x).$$

Es ist also sicher unmöglich, zur Abschätzung nach unten ein z zu finden, das stärker wächst als $|x|^2$.

Literatur

1. Collatz, L.: Fehlerabschätzungen bei Randwertaufgaben partieller Differentialgleichungen mit unendlichem Grundgebiet. ZAMP **9a**, 118—128 (1958).
2. Redheffer, R.: Eindeutigkeitssätze bei nichtlinearen Differentialgleichungen. Journ. f. reine u. angew. Mathem. **211**, 70—77 (1962).
3. Schröder, J.: Monotonie-Eigenschaften bei Differentialungleichungen. Arch. Rat. Mech. Anal. **14**, 38—60 (1963).
4. — Monotonie-Aussagen bei quasilinearen elliptischen Differentialgleichungen und anderen Problemen. ISNM Vol. 9: Numerische Mathematik, Differentialgleichungen, Approximationstheorie. Basel-Stuttgart: Birkhäuser 1968.
5. Stoß, H.-J.: Inverspositive singuläre lineare Differentialoperatoren zweiter Ordnung. Dissertation Köln (1969).
6. Walter, W.: Differential- und Integralungleichungen. Berlin-Göttingen-Heidelberg-New York: Springer 1964.

Hanns-Jörg Stoß
Universität Konstanz
Fachbereich Mathematik
775 Konstanz
Jacob-Burckhardt-Str.

Numer. Math. 15, 74—86 (1970)

Condition, Equilibration and Pivoting in Linear Algebraic Systems

A. VAN DER SLUIS

Received March 6, 1969

Introduction

This paper is concerned with some problems in relation with the numerical solution of linear algebraic systems using elimination and orthogonalization methods.

The fact that for Gaussian elimination with partial pivoting almost any pivotal strategy can be achieved by appropriate scaling of rows, raises the question how a matrix should be scaled so as to justify the use of partial pivoting. The generally accepted answer to this question is *row-equilibration*, i.e. the scaling which leads to equal norms for all rows. As far as can be gathered from literature (cf. list of references), this answer is based on mathematical intuition rather than on proofs, and little is known of its effects.

Another fact, often found puzzling, is that such stable methods as orthogonal triangularization using Givens or Householder transformations seem to require some form of scaling as well.

In Part I of this paper, on the basis of Wilkinson's error analysis a justification of equilibration for the Gaussian methods will be given; it will also be shown that on this basis equilibration must be expected to be just as useful for the orthogonalization methods.

In Part II we shall make a more detailed use of Wilkinson's error analysis to show that matrices having large condition numbers may have a host of well conditioned solutions. It will then be shown that, although equilibration may be good for the worst conditioned solutions, it may be bad for many others.

Conventions

In this paper we shall always assume that numerical computations are carried out in binary floating point with a precision of t binary digits, without accumulation of inner products in extra precision (in [8], p. 7—11, this mode is indicated by *fl.*).

Theorems 1 and 2 and their proofs are given at the end of the paper.

Part I

1. Condition Number of a Matrix

1.1. Let

$$A x = b \tag{1.1}$$

be a real or complex system of n linear algebraic equations with n unknowns with a non-singular matrix and let

$$(A + \Delta A)(x + \Delta x) = b \tag{1.2}$$

be a system which is perturbed with respect to (1.1). Then

$$\Delta x = -(I + A^{-1}\Delta A)^{-1} A^{-1}(\Delta A)\, x. \tag{1.3}$$

In stead of Δx we shall study its first order approximation

$$\delta x = -A^{-1}(\Delta A)\, x. \tag{1.4}$$

For any vector norm and its subordinate matrix norm we have

$$\frac{\|\delta x\|}{\|x\|} \leqq \|A^{-1}\|\,\|\Delta A\| \tag{1.5}$$

which is often rewritten as

$$\frac{\|\delta x\|}{\|x\|} \leqq \varkappa(A)\,\frac{\|\Delta A\|}{\|A\|} \tag{1.6}$$

where

$$\varkappa(A) = \|A^{-1}\|\,\|A\| \tag{1.7}$$

is the usual condition number of A with respect to this norm.

1.2. By *equilibration* we mean scaling of rows (and/or columns) of A such that the rows (and/or columns) obtain more or less equal norms for some vector norm. As is well known, equilibration tends to reduce the condition number of A (cf. [10], ch. 4, section 5; [5], Theorem 2.5 (a)), and, on account of (1.6) and in a qualitative sense this will make the system less perceptive to perturbations.

On the other hand, as Wilkinson has shown (see below), solving the system numerically with one of the usual methods yields a solution which is the exact solution of a system in which A is slightly perturbed.

It is thought sometimes that the two facts just mentioned would show that equilibration leads to a lower bound for the relative error $\|\Delta x\|/\|x\|$ in the calculated solution. Of course this is incorrect, since the equilibrated matrix will have different $\|A\|$ and $\|\Delta A\|$, and without knowing how these quantities behave under scaling, nothing can be said. Therefore we turn to the results of error analysis.

2. Results of Error Analysis

For the sake of convenience, constants in formulas have been rounded; reference is made to the correct formulas.

2.1. *Gaussian Elimination with Partial or Complete Pivoting* (Gpp or Gcp, for short, or Gp if either method of pivoting is allowed). These methods yield a computed solution $x + \Delta x$, satisfying (1.2) with

$$\|\Delta A\|_\infty \leqq g(A)\,(2n^2 + n^3)\,2^{-t} \quad ([8], \text{p. } 108, (25.14)) \tag{2.1}$$

where, in the notation of [8], p. 94ff.,

$$g(A) = \max_{i,j,k} |a_{ij}^{(k)}|, \tag{2.2}$$

in words, $g(A)$ is the maximum absolute value attained by any element of A at any time during the elimination process (and not, as is thought sometimes, the maximum modulus of the elements of the upper-triangular matrix which is

ultimately obtained, nor the maximum modulus of the pivots, unless complete pivoting is used).

The quantity $g(A)$ satisfies

$$g(A) \leqq 2^{n-1} \max |a_{ij}| \quad ([10], \text{p. } 212) \tag{2.3}$$

and this bound may be attained, as is shown by an example ([10], p. 212). On the basis of this bound no reasonable error analysis is possible. Fortunately, it happens only very rarely that $g(A) > 4 \max |a_{ij}|$ ([11], p. 78).

Therefore we introduce

Property W. A matrix B is said to have property W if

$$g(B) \leqq 4 \max_{i,j} |b_{ij}| \tag{2.4}$$

(where the factor 4 is rather arbitrary).

Of course one never knows beforehand whether a matrix has property W. However, a Gcp computer programme can easily test during its execution whether this is the case; in the infrequent cases of violation it might switch to a more stable (and more time consuming) procedure. For a Gpp programme to test the same it will have to do the same amount of extra work that is required for complete pivoting.

Consequently, assuming property W, for Gp we have

$$\|\Delta A\|_\infty \leqq (8n^2 + 4n^3) \max |a_{ij}| \, 2^{-t}. \tag{2.5}$$

2.2. *Orthogonal Triangularization Using Givens Transformations* (OG for short). This yields a computed solution satisfying (1.2) with

$$\|\Delta A\|_F \leqq 8n^{1.5} \|A\|_F 2^{-t} \tag{2.6}$$

($\|A\|_F$ denoting the Frobeniusnorm $\sqrt{\sum |a_{ij}|^2}$) for all n below a very liberal bound (cf. Theorem 2). In proving this result no further assumptions or restrictions or recourse to experience are made. Similar results can be derived if Householder transformations are used, though in that case a factor n^2 appears instead of $n^{1.5}$.

2.3. Relations (2.5) and (2.6) hold regardless whether or not the rows and columns of A have been subjected to some form of scaling.

These relations show that assuming property W the methods Gpp, Gcp and OG are of comparable and very good stability.

3. Row Scaling

3.1. Formulae (2.5) and (2.6) define classes of perturbations of A, and it is now a natural question how to scale the matrix such that under perturbations of these classes the system $Ax = b$ represents its solution in the best possible way.

From (1.5) and (2.5) we have for Gpp and Gcp

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leqq \varkappa'(A) \, (8n^2 + 4n^3) \, 2^{-t} \tag{3.1}$$

with the modified condition number

$$\varkappa'(A) = \|A^{-1}\|_\infty \max |a_{ij}| . \tag{3.2}$$

This has the property

$$\varkappa'(A) \geqq 1 \tag{3.3}$$

since $\varkappa'(A) = \max_j \|A^{-1}\|_\infty \|A_j\|_\infty$ and $\|A^{-1}\|_\infty \|A_j\|_\infty \geqq \|A^{-1}A_j\|_\infty = 1$, where A_j denotes the j-th column of A.

From [6], Theorem 1.9, we know that for any A and b there is a ΔA in the class of perturbations (2.5) for which the equality sign in (3.1) occurs. Therefore our question amounts to minimizing $\varkappa'(A)$ by scaling (this could also have been derived directly from [6], Theorem 2.4).

We first look at row-scaling, i.e. we consider DA, $D \in \mathfrak{D}$, $\mathfrak{D}$ the set of diagonal matrices. Then the problem is solved completely by [5], Theorem 2.5 (b), which says that $\min_{D \in \mathfrak{D}} \varkappa'(DA)$ is attained for those D for which all rows of DA have equal ∞-norm (in that case we call DA row-∞-equilibrated).

3.2. The foregoing analysis gives a surprisingly simple picture of the relation between Gp and equilibration. We might say that pivoting and row-∞-equilibration have rather independent actions. *Irrespective of equilibration, if property W holds pivoting renders the elimination process stable. Equilibration optimizes the way the system represents its solution with respect to this type of stability.*

Therefore, on the basis of Wilkinson's error analysis we should consider row-∞-equilibration justified if the equilibrated matrix possesses property W. The fact that this usually is the case is indicative for the practical success of the method. As has already been remarked (cf. § 2.1) a computer programme can easily test whether property W holds in any particular case.

3.3. For OG we have similarly

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leqq \varkappa''(A)\, 8n^{1.5} 2^{-t} \tag{3.5}$$

with

$$\varkappa''(A) = \|A^{-1}\|_2 \|A\|_F . \tag{3.6}$$

For minimizing $\varkappa''(A)$ by row-scaling we have slightly less precise results than for $\varkappa'(A)$. If A is row-2-equilibrated one can show that $\varkappa''(A)$ exceeds the best possible result by no more than a factor $\sqrt{n}$ (cf. [5], Theorem 3.5 (a)). Note, however, that here we have no additional assumptions, like property W.

This indicates that equilibration is just as useful for OG as it is for the Gaussian methods (as is already mentioned in passing in [10], p. 263, additional notes).

4. Column Scaling

Of course one might try to minimize $\varkappa'(A)$ further by column-scaling i.e. right multiplication of A by $D \in \mathfrak{D}$. If this is done by powers of 2 and one uses Gpp, this leads to scaling the calculated solution, even as regards rounding errors, by D^{-1} since the pivotal selection is not affected by column scaling. Descaling this solution by D then leads to exactly the same solution as would have been

obtained without any column-scaling. Therefore, in practical calculations, column-scaling can be skipped.

However, it is thought sometimes that a theoretical advantage can be obtained from this state of affairs in this respect, that it would imply that in (3.1) $\varkappa'(A)$ could be replaced by $\min_{D \in \mathfrak{D}} \varkappa'(AD)$, and that this quantity should be considered as the "true" or "factual" condition number, indicative for the quality of the solution.

This is fallacious, however. For, as is seen from the beginning of this section, minimizing $\varkappa'(AD)$ means minimizing $\|D^{-1}\Delta x\|_\infty / \|D^{-1}x\|_\infty$, if Δx is the error in solving the original system. This means, in fact, that the relative error is being regarded in a different norm, in which relatively large coordinates of Δx may be given smaller weights than relatively small coordinates. Therefore, one might say that minimizing $\varkappa'(A)$ by column-scaling amounts to seeking an answer to the question: "in which norm does the relative error *look* most favorable?".

The same cannot be said right away for Gcp, since then column-scaling may change the pivotal selection. But of course one has the feeling that the situation with Gcp is not different from that with Gpp. This is corroborated by the following example, for which Gpp and Gcp are identical.

As an example, we solve the system

$$\begin{aligned} x_1 - 10^{-6} x_2 &= 0.807 \\ 0.793\, x_1 + 10^{-6} x_2 &= 0.64 \end{aligned} \tag{4.1}$$

with Gpp, using 4-digit *decimal* floating arithmetic. The true solution, rounded to 4 digits, is $x_1 = 0.8070$, $x_2 = 27.33$. The calculated solution is $x_1 = 0.8070$, $x_2 = 0$, and the relative error is 100 percent.

Yet, scaling the columns of the matrix with $\tilde{D} = \begin{pmatrix} 1 & 0 \\ 0 & 10^6 \end{pmatrix}$ we obtain a matrix for which $\varkappa'$ is close to 1, and therefore $\min_D \varkappa'(AD)$ is close to 1, which shows that $\min \varkappa'(AD)$ certainly cannot serve as the factual condition number. It is also clear that $\tilde{D}^{-1}$ has exactly the effect mentioned above: it strongly reduces the ugly second component of Δx.

Part II

5. More Detailed Information about ΔA

5.1. As is shown in [5], 1.9, for any A and b there is a ΔA in the class of perturbations (2.5) for which the equality sign in (3.1) occurs. Since the error analysis in § 3 does not exclude any ΔA of this class, we cannot *count* on any solution to have a lower error than is indicated by the right hand side of (3.1).

However, more detailed information about ΔA can be given. In fact, for Gp we have

$$\|\Delta A_j\|_\infty \leqq g_j(A)\,(3n + n^2)\,2^{-t} \tag{5.1}$$

where A_j denotes the j-th column of A and

$$g_j(A) = \max_{i,k} |a_{ij}^{(k)}| \tag{5.2}$$

(cf. Theorem 1).

We now introduce

Property W'. A matrix B is said to have property W' if

$$g_j(B) \leqq 4 \max_i |b_{ij}|, \quad j = 1, \ldots, n. \tag{5.3}$$

Since with Gpp column scaling does not affect the pivotal selection nor the ratio $g_j(B)/\max_i |b_{ij}|$, a matrix has property W' if the matrix obtained from it by column-∞-equilibration has property W. Since property W is widely spread among column-∞-equilibrated matrices also, property W' must be very common for Gpp, and it is hardly to be expected that this would not be the case for Gcp. At any rate, a Gp computer programme can test whether property W' holds.

Therefore, assuming property W',

$$\|\Delta A_j\|_\infty \leqq (12n + 4n^2) \|A_j\|_\infty 2^{-t}. \tag{5.4}$$

Similarly for OG we have, independently of any property W' (cf. Theorem 2)

$$\|\Delta A_j\|_2 \leqq 8n^{1.5} \|A_j\|_2 2^{-t}. \tag{5.5}$$

Both (5.4) and (5.5) have the shape

$$\|\Delta A_j\| \leqq \chi(n) \|A_j\| 2^{-t}. \tag{5.6}$$

5.2. Formula (5.6) defines a perturbation class of type [6], (2.2), and therefore we may apply the analysis given in [6], §§ 3 and 4.

It was shown there to be relevant to write the result of substituting (5.6) in (1.5) as

$$\frac{\|\delta x\|}{\|x\|} \leqq \chi(n) \sum \varkappa_j(A) \frac{|x_j|}{\|x\|} 2^{-t} \tag{5.7}$$

where

$$\varkappa_j(A) = \|A^{-1}\| \|A_j\| \tag{5.8}$$

are the *column condition numbers*.

We recall a few properties (cf. [6], § 3)

(5.9) $\varkappa_j(A) \geqq 1$ for any Hölder norm.

(5.10) $\varkappa(A) \geqq \varkappa_j(A)$ for any Hölder norm.

(5.11) $\varkappa'(A) = \max \varkappa_j(A)$ for the ∞-norm.

(5.12) $\varkappa''(A) = \sqrt{\sum \varkappa_j(A)^2}$ for the 2-norm.

(5.13) For any A and b there is a ΔA in the perturbation class defined by (5.6) for which the equality sign in (5.7) holds.

5.3. If all $\varkappa_j(A)$ are of the same order of magnitude, i.e. all columns of A have comparable norms, the right hand side of (5.7) cannot vary more than $(\sum |x_j|)/\|x\|$ can vary; this is at most a factor n for the ∞-norm and a factor $\sqrt{n}$ for the 2-norm. On account of (5.13) we may then say that for the perturbation class (5.6) the relative perturbations of the solutions corresponding to various vectors b do not differ very much (n usually being a rather modest number).

If, however, the $\varkappa_j(A)$ have very different magnitudes, i.e. the columns of A have very different norms, we have an entirely different situation. As has been pointed out in [6], § 4 it may then happen that for a vast majority of vectors b the relative perturbations of the solutions exceed the *lowest* column condition number only by a modest factor, no matter how large the largest column condition number may be, and therefore (cf. (5.10)) how large the ordinary condition number may be. This is the case for matrices in which each column vector is well separated from the hyperplane through the other column vectors.

6. Effects of Equilibration

6.1. In the remainder we shall restrict ourselves to Gp and the ∞-norm. It will be obvious that there are similar results for OG and the 2-norm, except that for OG all restrictions regarding property W' vanish.

The quantity of interest (cf. (5.7)) is $\varkappa(A, x) = \sum \varkappa_j(A) \frac{|x_j|}{\|x\|}$, which is called the *condition number of the solution* (cf. [6]). Its maximum over all x is $\sum \varkappa_j(A)$. On behalf of (3.4) and (5.11), row-∞-equilibration minimizes $\max_j \varkappa_j$. The resulting $\sum \varkappa_j$ can, therefore, be no more than a factor n away from its minimum. Consequently, also in view of the error analysis given in § 5, row-∞-equilibration is a reasonable answer to the question how to minimize the available bound for $\max \|\Delta x\|/\|x\|$, the maximum taken over all x (i.e. the worst possible error is minimized), provided that the equilibrated matrix has property W'.

6.2. Since no longer (as in § 3) all solutions can be considered to be equally well or ill conditioned, we should now study the impact of equilibration on other solutions than the worst conditioned ones.

First, we consider an example. We define

$$A(p) = \begin{pmatrix} p & p & p \\ 1 & \varepsilon & -\varepsilon \\ 1 & -\varepsilon & \varepsilon \end{pmatrix}, \qquad |\varepsilon| \ll 1. \tag{6.1}$$

Since

$$A(p)^{-1} = \begin{pmatrix} 0 & 2p\varepsilon & 2p\varepsilon \\ 2\varepsilon & p(1-\varepsilon) & -p(1+\varepsilon) \\ 2\varepsilon & -p(1+\varepsilon) & p(1-\varepsilon) \end{pmatrix} \Big/ (4p\varepsilon) \tag{6.2}$$

we have

$$\|A(p)^{-1}\|_\infty = (1+|\varepsilon/p|)/2|\varepsilon| \quad \text{if} \quad |p| \geqq |\varepsilon| \tag{6.3}$$

and therefore, relative to the ∞-norm

$$\varkappa_1(A(\varepsilon)) = \frac{1}{|\varepsilon|}, \qquad \varkappa_2(A(\varepsilon)) = \varkappa_3(A(\varepsilon)) = 1, \tag{6.4}$$

$$\varkappa_1(A(1)) \approx \frac{1}{2|\varepsilon|}, \qquad \varkappa_2(A(1)) = \varkappa_3(A(1)) \approx \frac{1}{2|\varepsilon|}. \tag{6.5}$$

6.3. Thus, if $p = 1$ we have large bounds for the relative errors of all solutions, whereas if $p = \varepsilon$ these bounds are small for a vast majority of vectors b. However, *$A(1)$ is the matrix obtained from $A(\varepsilon)$ by row-∞-equilibration.*

This shows, that it is quite possible that equilibration transforms a system, which for the given perturbation class has a low relative error for a certain solution, into a system for which the error of this solution is *arbitrarily much greater*, the last few words meaning that there exists no bound *depending on n only* for the ratio of the errors after and before equilibration.

Moreover, we note that this is true for the solutions corresponding to a vast majority of vectors b.

Since this is a result for the errors corresponding to an entire perturbation class, of which for any given system only a single perturbation is used, one might think that the dim view given above is only due to crudeness in our analysis, and that the actual errors possibly would not suffer these ill effects (as, indeed, is often the case with analyses based on the ordinary condition numbers $\varkappa(A)$, $\varkappa'(A)$, $\varkappa''(A)$). That this is not so, is clearly seen by closer examination of matrices of type $A(p)$, where use of a_{11} as first pivot is disastrous. As is to be expected, equilibrating $A(\varepsilon)$ proves equally disastrous for OG.

6.4. The effect of equilibration on $A(\varepsilon)$ could have been predicted from § 5.3 without carrying out the calculations in § 6.2. For, if $p=\varepsilon$ none of the columns of $A(p)$ makes a small angle with the plane spanned by the other columns. However, if $p=1$, the second and third columns have nearly the same direction. Apparently, multiplying some coordinates by large factors may transform well-separated vectors into badly-separated ones.

6.5. For further clarification we consider any matrix A with $\max|a_{ij}|=1$. As follows from (3.4), row-∞-equilibration means multiplying A by a diagonal matrix $\tilde{D}$ such that $\|(\tilde{D}A)^{-1}\|_\infty$ is minimal under the additional condition $\max|(\tilde{D}A)_{ij}|=1$. However, since all diagonal elements of $\tilde{D}$ have modulus $\geqq 1$, we have $\|\tilde{D}A_j\|_\infty \geqq \|A_j\|_\infty$, and we may even have $\|\tilde{D}A_j\|_\infty \gg \|A_j\|_\infty$.

Thus, the two factors of $\varkappa_j(\tilde{D}A)$ (cf. (5.8)) have conflicting actions, and it is hard to see which one will win. In § 6.2 it is the second factor that wins: equilibration hardly decreases $\|A(p)^{-1}\|_\infty$ but greatly increases $\|A_2\|_\infty$ and $\|A_3\|_\infty$.

7. Modification of the Equilibration Process

7.1. We noted that row-∞-equilibration, though minimizing $\max_j \varkappa_j$, may cause other $\varkappa_j$ to increase by factors for which no bounds exist depending on n only.

Any search for a row scaling that minimizes all $\varkappa_j$ simultaneously, is futile unless the matrix has some very special form. For, any such scaling would certainly reduce $\max \varkappa_j$ and therefore $\varkappa'$ (cf. (5.11)) to its minimum value. We noted that row-∞-equilibration did just this, and from [5], Theorem 5.2(b), we see that any other scaling that does so can differ only by a constant factor from it unless the matrix is special.

7.2. We also have the following result: for given j, $\min_{D\in\mathfrak{D}} \|A^{-1}D^{-1}\|_\infty \cdot \|DA_j\|_\infty$ is attained for those D (and for those only unless the matrix is special) for which all coordinates of DA_j have equal modulus (cf. [5], Theorems 2.5(e) and 5.2(b)), and from this, too, we see that usually not all $\varkappa_j$ can be minimized simultaneously.

Since for any B and any $D \in \mathfrak{D}$

$$\varkappa_j(DB) \leqq \|D\|_\infty \|D^{-1}\|_\infty \varkappa_j(B) \leqq \frac{\max |d_{ii}|}{\min |d_{ii}|} \varkappa_j(B) \tag{7.1}$$

we have

$$\varkappa_j(A)/\min_{D\in\mathfrak{D}} \varkappa_j(DA) \leqq \max_i |a_{ij}| / \min_i |a_{ij}| . \tag{7.2}$$

This is a bound for the factor by which $\varkappa_j(A)$ might be reduced if A were so scaled that all coordinates of DA_j have equal modulus. This bound is very crude, of course, but, as (6.4) and (6.5) show, it may be realistic sometimes.

7.3. Our next question is whether the equilibration process can be amended in such a way that, as far as can be judged from (5.7), no solution *deteriorates* due to the equilibration process. This requires that no $\varkappa_j$ increases.

This is accomplished by taking the scaling matrix D such that

$$d_{ii} = \min_j \frac{\|A_j\|_\infty}{|a_{ij}|} . \tag{7.3}$$

Indeed for these d_{ii} we have $|d_{ii}| \geqq 1$, and therefore $\|A^{-1}D^{-1}\|_\infty \leqq \|A^{-1}\|_\infty$ (in fact this holds for any monotonic norm); on the other hand $|d_{ii}a_{ij}| \leqq \frac{\|A_j\|_\infty}{|a_{ij}|} |a_{ij}| = \|A_j\|_\infty$, so that $\|DA_j\|_\infty = \|A_j\|_\infty$.

We note that, if A is column-∞-equilibrated, these scaling factors are the same as would have been obtained by the usual row-∞-equilibration process. *Therefore, if all columns of A have the same ∞-norm, row-∞-equilibration has the desired property of not deteriorating the condition of any solution.*

In fact our amendment to the equilibration process amounts to applying column-∞-equilibration first, row-∞-equilibration thereafter and finally the inverse of the column scaling used first. Note, however, that column equilibration serves an entirely different purpose here than it did in § 5.

7.4. As has already been shown by Dekker in 1962 (cf. [3]; also cf. [2], (13)), for Gpp and direct triangular decomposition scaling need not be carried out explicitly, but the scaling factors can be taken into account when the pivot is being searched. Evidently, (7.3) allows the same feature.

For OG no such method is known. Here scaling should be carried out explicitly.

7.5. Clearly, for the matrix $A(\varepsilon)$ (cf. (6.1)) the modified equilibration process has the desired effect: it leaves the matrix unaltered.

It also leaves the following matrix unaltered:

$$A = \begin{pmatrix} 1 & \omega & \omega \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix} \quad \omega \gg 1 . \tag{7.4}$$

Now this is unfortunate, since evidently row-∞-equilibration would result in considerably better determined solutions. Since

$$A^{-1} = \begin{pmatrix} 2 & -2\omega & 0 \\ 0 & 1-\omega & -1+\omega \\ -2 & 1+\omega & 1-\omega \end{pmatrix} (2-2\omega)^{-1} \tag{7.5}$$

we have

$\|A^{-1}\| \approx 1,\ \varkappa_1 \approx 1,\ \varkappa_2 = \varkappa_3 \approx \omega$ for A as it stands

$\|A^{-1}\| \approx 2,\ \varkappa_1 \approx 2,\ \varkappa_2 = \varkappa_3 \approx 2$ for A after row-∞-equilibration such that all rows of A have ∞-norm 1.

Obviously, the modified process, in guarding off an increase of $\min_j \varkappa_j$ by a modest factor 2, prevents $\max_j \varkappa_j$ from decreasing by a factor $\omega/2$.

This experience seems to be a counterpart of our experience in § 6, where row-∞-equilibration caused a decrease of $\max_j \varkappa_j$ by a factor 2 at the cost of increasing $\min_j \varkappa_j$ by a factor $(2|\varepsilon|)^{-1}$.

In evaluating these effects, one might bear in mind that the ill effects of the usual equilibration process may spoil vast classes of solutions which are well-determined by the given system, whereas those of the modified process consist only in failure to improve solutions which are already badly determined by the given system.

Nevertheless it depends, of course, (as is so often the case in numerical analysis) on the circumstances, whether or not this modified process is to be recommended. In a computer procedure for general use, the usual equilibration process, which improves the condition of the worst conditioned solutions as much as possible, may still be the most attractive. But users should be aware of its effects.

8. Error Analysis for Gp

Theorem 1. For Gaussian elimination with partial or complete pivoting we have

$$\|\Delta A_j\|_\infty \leqq j\{2g_j + (n+1)\|U_j\|_\infty\} 2^{-t_1} \leqq j(n+3)\, g_j 2^{-t_1} \tag{8.1}$$

if $6(n^2+n+2) < 2^t$ and if g_j is defined according to (5.2) and t_1 according to [8], p. 19 (26.15). U is the upper triangular matrix to which A is ultimately reduced. The term $(n+1)\|U_j\|_\infty$ comes from the triangular systems. Note that for Gcp A_j and U_j denote the j-th pivotal column.

This result depends on the order of the calculations (cf. the proof).

Corollary. $\|\Delta A\|_\infty \leqq \frac{1}{2} n(n+1)(n+3)\, g\, 2^{-t_1}$, from which [8], p. 108 (25.14) follows.

Proof of the theorem. Gpp yields a unit lower triangular matrix L and an upper triangular matrix U such that

$$LU = A + E. \tag{8.2}$$

Similar to [8], p. 98 (16.13) we have

$$|E| < (2.01)\, 2^{-t} \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 2 & 2 & \dots & 2 & 2 \\ 1 & 2 & 3 & \dots & 3 & 3 \\ \hline 1 & 2 & 3 & \dots & (n-1) & (n-1) \end{pmatrix} \begin{pmatrix} g_1 & & & \\ & g_2 & & \\ & & \ddots & \\ & & & g_n \end{pmatrix}. \tag{8.3}$$

Subsequent numerical solution of the equations $Ly=b$, $Ux=y$, performing the calculations in the following order

$$y_r = fl(-l_{r1}y_1 - l_{r2}y_2 - \cdots - l_{rr-1}y_{r-1} + b_r), \tag{8.4}$$

$$x_r = fl\big((-u_{rn}x_n - u_{rn-1}x_{n-1} - \cdots - u_{rr+1}x_{r+1} + y_r)/u_{rr}\big) \tag{8.5}$$

yields solutions $\bar{y}$ and $\bar{x}$ satisfying

$$(L+\delta L)\bar{y} = b, \quad (U+\delta U)\bar{x} = \bar{y} \tag{8.6}$$

from which

$$\Delta A = E + L\,\delta U + \delta L\,U + \delta L\,\delta U. \tag{8.7}$$

If by $*$ we denote a matrix multiplication defined by $(a_{ij})*(b_{ij}) = (a_{ij}b_{ij})$, we have for the representative case $n=6$, similar to [8], p. 101 (19.19) (taking account of $l_{ii}=1$)

$$|\delta L| \leqq 2^{-t_1}\begin{pmatrix} 0 & & & & & \\ 1 & 1 & & & & \\ 2 & 2 & 1 & & & \\ 3 & 3 & 2 & 1 & & \\ 4 & 4 & 3 & 2 & 1 & \\ 5 & 5 & 4 & 3 & 2 & 1 \end{pmatrix} * |L|;$$

$$|\delta U| \leqq 2^{-t_1}\begin{pmatrix} 2 & 2 & 3 & 4 & 5 & 5 \\ & 2 & 2 & 3 & 4 & 4 \\ & & 2 & 2 & 3 & 3 \\ & & & 2 & 2 & 2 \\ & & & & 2 & 1 \\ & & & & & 1 \end{pmatrix} * |U|. \tag{8.8}$$

Putting $|U_j| = (u_1, \ldots, u_n)^T$, we have

$$|L\,\delta U_j| \leqq 2^{-t_1}\begin{pmatrix} j\,u_1 \\ j\,u_1 + (j-1)\,u_2 \\ \hline \end{pmatrix}; \quad |\delta L\,U_j| \leqq 2^{-t_1}\begin{pmatrix} 0 \\ u_1+u_2 \\ \hline 2u_1+2u_2+u_3 \end{pmatrix}. \tag{8.9}$$

Hence

$$\|E_j\|_\infty \leqq j\,g_j(2.01)\,2^{-t} \leqq j\,g_j(1.9)\,2^{-t_1}, \tag{8.10}$$

$$\|L\,\delta U_j\|_\infty \leqq \tfrac{1}{2}(j^2+j+2)\,\|U_j\|_\infty 2^{-t_1}, \qquad \|U_j\|_\infty \leqq g_j, \tag{8.11}$$

$$\|\delta L\,U_j\|_\infty \leqq \tfrac{1}{2}(2nj - j^2 + j - 2)\,\|U_j\|_\infty 2^{-t_1}, \quad \|U_j\|_\infty \leqq g_j, \tag{8.12}$$

$$\|\delta L\,\delta U_j\|_\infty \leqq \tfrac{1}{2}n(j^2+j+2)\,g_j 2^{-2t_1} \leqq \{\tfrac{1}{2}(n^2+n+2)\,2^{-t_1}\}\,j\,g_j 2^{-t_1} \leqq j\,g_j(0.1)\,2^{-t_1}. \tag{8.13}$$

Substitution of these inequalities in (8.7) completes the proof.

9. Error Analysis for OG

Theorem 2. For orthogonal triangularization using Givens transformations we have

$$\|\Delta A_j\|_2 \leqq 8n^{1.5}\|A_j\|_2 2^{-t} \tag{9.1}$$

if $1000n < 2^t$ and $70n^2 < 2^t$, where approximately half the error comes from the triangularization itself.

This result depends on the way the calculations are performed. In particular it has been assumed that the Givens transformations have not been multiplied explicitly.

Corollary. $\|\Delta A\|_F \leqq 8n^{1.5}\|A\|_F 2^{-t}$.

Proof of the theorem. The orthogonal triangularization method yields a sequence of matrices $R_1, \ldots, R_N$ which are very nearly orthogonal, and an upper triangular matrix U (cf. [10], p. 239). Let Q denote the exact product of the computed $R_1, \ldots, R_N$. Then

$$Q(A+E) = U. \tag{9.2}$$

If the R_i are determined as described in [10], p. 239, then from [10], p. 240 (52.2–4) with $b_0 = A_j$, $b_N = U_j$, we have

$$\|E_j\|_2 \leqq 3n^{1.5}\varrho_n\|A_j\|_2 2^{-t} \tag{9.3}$$

with

$$\varrho_n = \left[\frac{1+6\cdot 2^{-t}}{1-(4.3)\,2^{-t}}\right]^{2n-4}. \tag{9.4}$$

Subsequent numerical determination of x and y from $y = Qb$ (Q in product form) and $Ux = y$ (according to (8.5)) yields $\bar{x}$ and $\bar{y}$ satisfying

$$\bar{y} = (Q+\delta Q)\,b, \qquad (U+\delta U)\,\bar{x} = \bar{y} \tag{9.5}$$

from which

$$\Delta A = (I + Q^{-1}\delta Q)^{-1}(Q^{-1}\delta U - Q^{-1}\delta Q A + E). \tag{9.6}$$

Then

$$\|U_j\|_2 \leqq (1+6n^{1.5}2^{-t}\varrho_n)\|A_j\|_2 \quad \text{(cf. [10], p. 138 (23.20))}. \tag{9.7}$$

Since in 2-norm Q differs from an orthogonal matrix by no more than $6n^2\varrho_n 2^-$ ([10], p. 139 (25.1))

$$\|Q^{-1}\|_2 \leqq (1-6n^2\varrho_n 2^{-t})^{-1}. \tag{9.8}$$

In a similar way as [10], p. 240 (52.2) has been proved, it can be proved that $y = (Q+\delta Q)\,b$ with

$$\|\delta Q\|_2 \leqq 3n^{1.5}\varrho_n 2^{-t}. \tag{9.9}$$

From (8.8) we have

$$\|\delta U_j\|_2 \leqq n\|U_j\|_2 2^{-t_1}. \tag{9.10}$$

Substitution of these inequalities in (9.6) completes the proof.

References

1. Bauer, F. L.: Optimally scaled matrices. Num. Math. **5**, 73—87 (1963).
2. Bowdler, H. J., Martin, R. S., Peters, G., Wilkinson, J. H.: Solution of real and complex systems of linear equations. Num. Math. **8**, 217—234 (1966) (Handbook series).
3. Dekker, T. J.: Series AP 200 algorithms. Mathematisch Centrum, Amsterdam (1962).
4. Forsythe, G. E., Moler, C. B.: Computer solution of linear algebraic systems. Englewood Cliffs, N. J.: Prentice-Hall 1967.
5. Sluis, A. van der: Condition numbers and equilibration of matrices. Num. Math. **14**, 14—23 (1969).
6. — Stability of solutions of linear algebraic systems. Num. Math. **14**, 246—251 (1970).
7. Wilkinson, J. H.: Error analysis of direct methods of matrix inversion. J. Ass. Comp. Mach. **8**, 281—330 (1961).
8. — Rounding errors in algebraic processes. Notes on applied science, No. 32. London: Her Majesty's Stationery Office 1963.
9. — Rundungsfehler. Berlin-Heidelberg-New York: Springer 1969.
10. — The algebraic eigenvalue problem. London: Oxford University Press 1965.
11. — The solution of ill-conditioned linear equations. In: Mathematical methods for digital computers, ed. by A. Ralston and H. S. Wilf, vol. 2, p. 65—93. New York: John Wiley 1967.

A. van der Sluis
Rekencentrum Rijksuniversiteit
Budapestlaan 6
Utrecht, Netherlands

Numer. Math. 15, 87–99 (1970)

Numerical Methods of High-Order Accuracy for Singular Nonlinear Boundary Value Problems*

P. G. CIARLET, F. NATTERER, and R. S. VARGA

Dedicated to Professor L. Collatz on his 60th birthday

Received February 15, 1969

§ 1. Introduction

In [2] and [3], the numerical approximation of the solution of real nonlinear two-point boundary value problems of the form

$$L[u(x)]+f(x,u(x))=0, \quad 0<x<1, \tag{1.1}$$

with Dirichlet boundary conditions

$$D^k u(0)=D^k u(1)=0, \quad D\equiv\frac{d}{dx}, \quad 0\leqq k\leqq n-1, \tag{1.2}$$

was considered, where

$$L[u(x)]\equiv\sum_{j=0}^{n}(-1)^j D^j\{p_j(x)D^j u(x)\}, \quad n\geqq 1, \tag{1.3}$$

is a $2n$-th order self-adjoint linear differential operator, and it was shown that the Rayleigh-Ritz-Galerkin method is an efficient scheme, both theoretically and numerically, for solving such problems.

Our aim here is to extend the results of [2] and [3] to the case of *nonself-adjoint* linear differential operators whose coefficients have a *singularity* at one or both end points of the interval [0, 1]. For ease of exposition, we shall restrict ourselves here to second order operators, as in the particular case of

$$\mathscr{L}[u(x)]\equiv D^2u(x)+\frac{\sigma}{x}Du(x), \quad 0<x<1, \tag{1.4}$$

where σ is a constant satisfying $0\leqq\sigma<1$, and we will consider the associated real *nonlinear* Dirichlet problem

$$\mathscr{L}[u(x)]=g(x,u(x)), \quad 0<x<1, \tag{1.5}$$

$$u(0)=\alpha, \quad u(1)=\beta. \tag{1.6}$$

We remark that the one-dimensional boundary value problem of (1.4)–(1.6) can in certain cases be obtained by a separation of variables from the two-dimensional degenerate elliptic problem, treated by Gusman and Oganesyan [5].

* This research was supported in part by AEC Grant AT(11-1)-2075.

In [4], we also considered nonself-adjoint problems such as $D^2u(x) = f(x, u(x), Du(x))$, $0<x<1$, and $u(0)=\alpha$, $u(1)=\beta$. However, if we write (1.4) and (1.5) as $D^2u(x) = g(x, u(x)) - (\sigma/x)Du(x) \equiv f(x, u(x), Du(x))$, the particular dependence of $f(x, u, v)$ on its third argument is such that the analysis of [4] is not directly applicable. Thus, we feel that the problem (1.4)—(1.6) has interest in itself.

The particular singular boundary value problem (1.4)—(1.6) has been recently considered by Jamet [6, 7], and Parter [9], in the linear case only. In particular, Jamet has studied the application of a standard three-point finite difference scheme associated with a uniform mesh of mesh size h, and he has shown that the error is $O(h^{1-\sigma})$ in the max-norm, the exponent $(1-\sigma)$ of h being best possible. Jamet's method is essentially equivalent to finding the Rayleigh-Ritz-Galerkin piecewise-linear approximation of the solution Φ on a uniform mesh. By using a more suitable subspace for Rayleigh-Ritz-Galerkin's method, we shall improve Jamet's result by proving in inequality (3.16) that the error in the uniform norm for our Galerkin approximation is $O(h^{2-\sigma})$, the exponent of h again being best possible.

In order that we may apply our techniques to solving (1.5)—(1.6), we first replace the boundary conditions of (1.6) by homogeneous ones. If we were to subtract off the linear function $\alpha+(\beta-\alpha)x$ from $u(x)$ in (1.5), a singularity might be added to the function $g(x, u)$ of (1.5) for the new problem. Instead, we put $v(x) \equiv u(x) - (\alpha+(\beta-\alpha)x^2)$, so that solving (1.5)—(1.6) is equivalent to solving

$$\mathscr{L}[v(x)] = h(x, v(x)), \quad 0<x<1, \tag{1.7}$$

$$v(0) = v(1) = 0, \tag{1.8}$$

where $h(x, v) \equiv g(x, v+\alpha+(\beta-\alpha)x^2) - 2(\beta-\alpha)(1+\sigma)$.

Next, we put the boundary value problem of (1.7)—(1.8) into a self-adjoint form, using the fact that Φ is a solution of (1.7)—(1.8) if and only if it is also a solution of

$$D\{x^\sigma Du(x)\} = f(x, u(x)), \quad 0<x<1, \tag{1.9}$$

$$u(0) = u(1) = 0, \tag{1.10}$$

where

$$f(x, u) \equiv x^\sigma h(x, u). \tag{1.11}$$

At this point, we generalize the differential equation of (1.9) to

$$D\{p(x)Du(x)\} = f(x, u(x)), \quad 0<x<1, \tag{1.12}$$

where we assume that the function $p(x)$ satisfies

$$\text{(i)}\ \ p(x)>0 \ \text{ in } (0, 1),$$
$$\text{(ii)}\ \ p\in C^1(0, 1), \ \text{ and} \tag{1.13}$$
$$\text{(iii)}\ \ \frac{1}{p}\in L^1[0, 1].$$

It is easy to verify that the particular choice $p(x)=x^{\sigma}$, $0\leqq\sigma<1$, does in fact satisfy all the conditions of (1.13).

To begin our discussion, we define S to be the linear space of all real-valued functions $w\in C^0[0,1]$ satisfying the boundary conditions of (1.10), such that w is absolutely continuous on $[0,1]$, and such that

$$\sqrt{p(x)}\,Dw(x)\in L^2[0,1]. \tag{1.14}$$

Note that $H_0^1[0,1]\subset S$, where $H_0^1[0,1]$ is the Sobolev space of absolutely continuous functions defined over $[0,1]$ satisfying the boundary conditions of (1.10), and such that $Dw\in L^2[0,1]$. That we are actually obliged to extend the definition of our space of admissible functions (which, in [2], was $S=H_0^1[0,1]$ in the special case of second-order problems) to a space containing *strictly* $H_0^1[0,1]$ may be seen from the following example: The boundary value problem $D\{x^{\sigma}Du(x)\}=-3(\sigma+2)x^{1+\sigma}$, $0<x<1$, together with the homogeneous boundary conditions of (1.10), has the unique solution $\Phi(x)\equiv x^{1-\sigma}-x^3$ which belongs to S for $0\leqq\sigma<1$, but does not belong to $H_0^1[0,1]$ whenever $\frac{1}{2}\leqq\sigma\leqq 1$. Also, the space S as defined above is a special case of the so-called *weighted Sobolev spaces*, such spaces being a useful tool in studying weak solutions of singular boundary value problems (cf. Nečas [8, Chapter 6] and Gusman and Oganesyan [5]).

Next, we introduce the positive quantity (see Lemma 1)

$$\Lambda=\inf_{\substack{w\in S\\ w\not\equiv 0}}\frac{\int_0^1 p(x)\,[Dw(x)]^2\,dx}{\int_0^1 [w(x)]^2\,dx}. \tag{1.15}$$

We assume that the real function $f(x,u)$ given in (1.12) is continuous in $[0,1]\times R$, and continuously differentiable with respect to u for all $0\leqq x\leqq 1$, and all real u, and that there exists a constant γ such that

$$f_u(x,u)\equiv\frac{\partial f}{\partial u}(x,u)\geqq\gamma>-\Lambda, \quad \text{for all } 0\leqq x\leqq 1, \quad \text{and all real } u. \tag{1.16}$$

We remark that the theoretical results to follow remain valid if we replace (1.16) by a weaker assumption, based on divided differences of f (cf. [2, § 8]).

To outline the subsequent material, § 2 briefly lists the basic results concerning the application of the Rayleigh-Ritz-Galerkin method to the problem (1.12)—(1.10), and since these results are identical to those given in §§ 2—4 of [2], the corresponding proofs will be omitted. However, an essential difference arises when particular subspaces are considered in § 3. This is due to the fact that the solution of (1.12)—(1.10) has in general unbounded derivatives at the end points of $[0,1]$. Thus, the error bounds derived in [2] are no longer valid since they depended upon L^{∞}-bounds of some derivative of the solution, and this difficulty is circumvented by introducing an appropriate approximating subspace. In so doing, we employ a method introduced by Ciarlet [1] and then generalized by Perrin, Price, and Varga [10]. Finally in § 4, a numerical example is included.

We remark that by the classic change of variables,

$$z(x)=\int_0^x\frac{d\xi}{p(\xi)},$$

where $z(x)$ is a continuous strictly increasing function, the problem (1.10)—(1.12) can be reduced to the following nonsingular form:

$$\frac{d^2U(z)}{dz^2} = F\big(z, U(z)\big), \quad 0 < z < z_1 \equiv z(1), \tag{1.12'}$$

$$U(0) = U(z_1) = 0, \tag{1.10'}$$

where, if $x(z)$ denotes the inverse function of $z(x)$, then

$$U(z) \equiv u\big(x(z)\big) \quad \text{and} \quad F(z, U) \equiv p\big(x(z)\big) f\big(x(z), U\big).$$

Although this reduction is possible, we feel that it is desirable, as in [7], to directly consider the numerical approximation of the solution (1.10)—(1.12).

§ 2. Variational Formulation

We begin with the following result.

Lemma 1. The quantity

$$\|w\|_0 \equiv \left\{\int_0^1 p(x)\,[Dw(x)]^2\,dx\right\}^{\frac{1}{2}} \tag{2.1}$$

is a norm over the space S, and the following holds:

$$\|w\|_{L^\infty[0,1]} \equiv \sup\{|w(x)|:\ 0 \leqq x \leqq 1\} \leqq \sqrt{r(1)}\,\|w\|_0 \quad \text{for all} \quad w \in S, \tag{2.2}$$

where $r(x)$ is defined by

$$r(x) = \int_0^x \frac{dt}{p(t)} \quad \text{for all} \quad x \in [0, 1]. \tag{2.3}$$

Finally, the quantity Λ defined in (1.15) is positive.

Proof. Let $w \in S$. Since $\sqrt{p(x)}\,Dw(x) \in L^2[0, 1]$ and $p(x) > 0$ in $(0, 1)$ by (1.13) (i), it follows from $w(0) = w(1) = 0$ that $\|\cdot\|_0$ is a norm over the space S.

Next, since $w(x)$ is absolutely continuous and $w(0) = 0$,

$$w(x) = \int_0^x Dw(\xi)\,d\xi = \int_0^x \frac{1}{\sqrt{p(\xi)}}\big(\sqrt{p(\xi)} \cdot Dw(\xi)\big)\,d\xi \quad \text{for all} \quad x \in [0, 1],$$

so that, using the Cauchy-Schwarz inequality and the definition of $r(x)$ in (2.3), we have

$$|w(x)| \leqq \sqrt{r(1)} \cdot \|w\|_0 \quad \text{for all} \quad x \in [0, 1],$$

which proves (2.2).

Finally, it follows from the inequality of (2.2) that

$$\Lambda = \inf_{\substack{w \in S \\ w \neq 0}} \frac{\int_0^1 p(x)\,[Dw(x)]^2\,dx}{\int_0^1 [w(x)]^2\,dx} \geqq \frac{1}{r(1)} > 0,$$

which completes the proof. Q.E.D.

As in [2], we make the hypothesis that the boundary value problem (1.12)—(1—10) has a classical solution Φ (i.e., $\Phi \in C^0[0, 1] \cap C^2(0, 1)$). This implies that $\Phi \in S$, since it follows from (1.12) that

$$p(x) D\Phi(x) = \int_{x_0}^{x} f(\eta, \Phi(\eta))\, d\eta + p(x_0) D\Phi(x_0), \qquad \text{for all } x, x_0 \in (0, 1).$$

Keeping x_0 fixed, we see that $p(x) D\Phi(x)$ can be extended to a continuous function over $[0, 1]$, say $q(x)$, and thus, from (1.13) (iii), $\sqrt{p(x)}\, D\Phi(x) \in L^2[0, 1]$:

$$\int_0^1 p(x) (D\Phi(x))^2 dx = \int_0^1 \frac{(q(x))^2 dx}{p(x)} < +\infty.$$

Next, we have as in Theorem 1 of [2],

Theorem 1. With the assumptions of (1.13) and (1.16), Φ *strictly* minimizes the following functional

$$F[w] = \int_0^1 \left\{ \tfrac{1}{2} p(x) [D w(x)^2] + \int_0^{w(x)} f(x, \eta)\, d\eta \right\} dx, \tag{2.4}$$

over the space S, and thus Φ is the unique solution of (1.12)—(1.10).

Proof. It is readily verified with the above assumptions that

$$F[w] \geqq F[\Phi] + \frac{(\Lambda + \gamma)}{2} \int_0^1 [w(x) - \Phi(x)]^2 dx, \qquad \text{for all } w \in S,$$

proving that $F[w] > F[\Phi]$ unless $w \equiv \Phi$. Q.E.D.

We now briefly describe the approximation scheme. Let S_M be any finite-dimensional subspace (of dimension M) of S, and let w_i, $1 \leqq i \leqq M$, be M linearly independent functions in S_M. Then, the above inequality allows us to prove, exactly as in Theorem 2 of [2]:

Theorem 2. With the assumptions of (1.13) and (1.16), there exists a unique function $\hat{w}_M = \sum_{i=1}^{M} \hat{u}_i w_i$ in S_M which minimizes the functional $F[w]$ of (2.4) over S_M.

In what follows, we shall for brevity call $\hat{w}_M$ the *Galerkin* approximation of Φ on the subspace S_M.

As in [2, Lemma 2], we have

Lemma 2. Let g be a continuous function over $[0, 1]$ satisfying

$$-\Lambda < \gamma \leqq g(x) \leqq \Gamma, \qquad \text{for all } x \in [0, 1], \tag{2.5}$$

for some constants γ, Γ. Then

$$\|w\|_g \equiv \left\{ \int_0^1 \{p(x) [D w(x)]^2 + g(x) [w(x)]^2\} dx \right\}^{\frac{1}{2}} \tag{2.6}$$

is a norm over the space S, and moreover this norm is equivalent to the norm $\|\cdot\|_0$ of (2.1), i.e., there exist two constants $m = m(\gamma, \Gamma)$ and $M = M(\gamma, \Gamma)$ such that

$$m(\gamma, \Gamma)\,\|w\|_g \leqq \|w\|_0 \leqq M(\gamma, \Gamma)\,\|w\|_g, \quad \text{for all } w \in S. \tag{2.7}$$

As in Theorem 3 of [2], we have the following fundamental result:

Theorem 3. Let Φ be the (classical) solution of (1.12)—(1.10) subject to the assumptions of (1.13) and (1.16), let S_M be any finite-dimensional subspace of S, and let $\hat{w}_M$ be the unique Galerkin approximation which minimizes $F[w]$ over S_M. Then, there exists a constant C, which is *independent* of S_M, such that the following error bound is valid:

$$\|\hat{w}_M - \Phi\|_{L^\infty[0,1]} \leqq \sqrt{r(1)}\,\|\hat{w}_M - \Phi\|_0 \leqq C \inf\{\|w - \Phi\|_0;\ w \in S_M\}. \tag{2.8}$$

The following is then an immediate consequence of Theorem 3.

Theorem 4. Let Φ be the (classical) solution of (1.12)—(1.10) subject to the assumptions of (1.13) and (1.16), let $\{S_{M_i}\}_{i=1}^{\infty}$ be any sequence of (not necessarily nested) finite-dimensional subspaces of S, and let $\{\hat{w}_{M_i}\}_{i=1}^{\infty}$ be the corresponding sequence of Galerkin approximations obtained by minimizing $F[w]$ over the subspaces S_{M_i}. If

$$\lim_{i \to \infty} \{\inf\{\|w - \Phi\|_0;\ w \in S_{M_i}\}\} = 0, \tag{2.9}$$

then $\{\hat{w}_{M_i}\}_{i=1}^{\infty}$ converges uniformly to Φ.

§ 3. Approximating Subspaces

Let Π: $0 = x_0 < x_1 < x_2 < \cdots < x_{N+1} = 1$ denote *any* partition Π of $[0, 1]$. Then, with Π, we define the space S^Π as being the subspace of S whose functions w satisfy

$$D\{p(x)\,Dw(x)\} = 0, \quad x_j < x < x_{j+1} \quad \text{for all } 0 \leqq j \leqq N. \tag{3.1}$$

For computational purposes, a convenient basis for S^Π can be obtained in terms of the function $r(x)$ of (2.3) as follows. Let $h_i(x) \equiv r(x) - r(x_i)$ and let w_i, $1 \leqq i \leqq N$, be defined by:

$$w_i(x) \equiv \begin{cases} 0, & 0 \leqq x \leqq x_{i-1}, \\ h_{i-1}(x)/h_{i-1}(x_i), & x_{i-1} \leqq x \leqq x_i, \\ 1 - [h_i(x)/h_i(x_{i+1})], & x_i \leqq x \leqq x_{i+1}, \\ 0, & x_{i+1} \leqq x \leqq 1. \end{cases} \tag{3.2}$$

It is readily verified that each w_i, $1 \leqq i \leqq N$, belongs to the space S, and satisfies (3.1), as well as $w_i(0) = w_i(1) = 0$. Since in addition

$$w_i(x_j) = \delta_{i,j}, \quad 1 \leqq i \leqq N, \quad 0 \leqq j \leqq N+1,$$

any function $g \in S^\Pi$ can be expanded with respect to the basis $\{w_i\}_{i=1}^N$ as

$$g(x) = \sum_{i=1}^{N} g(x_i)\,w_i(x).$$

Given $\Phi \in S$, we define its S^{Π}-*interpolate* $\tilde{w}$ to be the unique element in S^{Π} which satisfies

$$\tilde{w}(x_i) = \Phi(x_i), \quad 0 \leqq i \leqq N+1. \tag{3.3}$$

We then prove

Lemma 3. Let Φ be the solution of (1.12)—(1.10), and let $\tilde{w}$ be its unique S^{Π}-interpolate. Then,

$$\|\tilde{w} - \Phi\|_{L^\infty[0,1]} \leqq M \cdot \ell(\Pi), \tag{3.4}$$

where $M \equiv \sup\{|f(x, \Phi(x))|;\ 0 \leqq x \leqq 1\}$, and

$$\ell(\Pi) \equiv \max_{0 \leqq i \leqq N} \left\{ (x_{i+1} - x_i) \int_{x_i}^{x_{i+1}} \frac{dt}{p(t)} \right\}. \tag{3.5}$$

Proof. Consider any interval $[x_i, x_{i+1}]$, $0 \leqq i \leqq N$. Since $\tilde{w} - \Phi$ vanishes at the two end points of this interval, an integration by parts gives us that

$$\int_{x_i}^{x_{i+1}} p(t) \{D(\tilde{w}(t) - \Phi(t))\}^2 dt = -\int_{x_i}^{x_{i+1}} D\{p(t) D(\tilde{w}(t) - \Phi(t))\} (\tilde{w}(t) - \Phi(t)) dt.$$

We remark that this integration by parts is valid even in the cases $i = 0, N$. Next, as $D\{p(t) D \tilde{w}(t)\} = 0$ in $[x_i, x_{i+1}]$, and $D\{p(t) D \Phi(t)\} \equiv f(t, \Phi(t))$, the integral can be expressed as

$$\int_{x_i}^{x_{i+1}} p(t) \{D(\tilde{w}(t) - \Phi(t))\}^2 dt = \int_{x_i}^{x_{i+1}} f(t, \Phi(t)) (\tilde{w}(t) - \Phi(t)) dt,$$

and thus,

$$\int_{x_i}^{x_{i+1}} p(t) \{D(\tilde{w}(t) - \Phi(t))\}^2 dt \leqq M(x_{i+1} - x_i) \|\tilde{w} - \Phi\|_{L^\infty[x_i, x_{i+1}]}. \tag{3.6}$$

Next, we can write for any $x \in [x_i, x_{i+1}]$, $0 \leqq i \leqq N$, that

$$\tilde{w}(x) - \Phi(x) = \int_{x_i}^{x} D\{\tilde{w}(t) - \Phi(t)\} dt = \int_{x_i}^{x} \frac{1}{\sqrt{p(t)}} (\sqrt{p(t)} D\{\tilde{w}(t) - \Phi(t)\}) dt,$$

and hence by applying the Cauchy-Schwarz inequality, we obtain

$$|\tilde{w}(x) - \Phi(x)| \leqq \left\{ \int_{x_i}^{x_{i+1}} \frac{dt}{p(t)} \right\}^{\frac{1}{2}} \left\{ \int_{x_i}^{x_{i+1}} p(t) \{D(\tilde{w}(t) - \Phi(t))\}^2 dt. \right\}^{\frac{1}{2}}.$$

As this holds for all $x \in [x_i, x_{i+1}]$, we have

$$\|\tilde{w} - \Phi\|_{L^\infty[x_i, x_{i+1}]} \leqq \left\{ \int_{x_i}^{x_{i+1}} \frac{dt}{p(t)} \right\}^{\frac{1}{2}} \left\{ \int_{x_i}^{x_{i+1}} p(t) \{D(\tilde{w}(t) - \Phi(t))\}^2 dt \right\}^{\frac{1}{2}}.$$

Combining this with the inequality of (3.6) then gives

$$\|\tilde{w} - \Phi\|_{L^\infty[x_i, x_{i+1}]} \leqq M(x_{i+1} - x_i) \int_{x_i}^{x_{i+1}} \frac{dt}{p(t)},$$

and thus, with the definition of $\ell(\Pi)$ of (3.5), the desired inequality of (3.4) follows. Q.E.D.

We can now prove:

Theorem 5. Let Φ be the solution of (1.12)—(1.10), subject to the assumptions of (1.13) and (1.16), and let $\hat{w}$ be the unique Galerkin approximation which minimizes $F[w]$ over the subspace S^Π. Then, there exists a constant C, independent of the partition Π such that

$$\|\hat{w} - \Phi\|_{L^\infty[0,1]} \leqq C \cdot \ell(\Pi). \tag{3.7}$$

Proof. Following Ciarlet [1] and Perrin, Price, and Varga [10], the basic idea is to compare $\hat{w}$ with the S^Π-interpolate $\tilde{w}$ of the solution Φ (cf. inequality (3.13)), using the fact that the functions of S^Π satisfy in each open interval (x_i, x_{i+1}), $0 \leqq i \leqq N$, the differential equation $D\{p(x)\, D\, w(x)\} = 0$ of (3.1).

Let

$$k_i = \int_0^1 \{p(x)\, D\tilde{w}(x)\, D w_i(x) + f(x, \tilde{w}(x))\, w_i(x)\}\, dx, \qquad 1 \leqq i \leqq N.$$

Since

$$\int_0^1 \{p(x)\, D\Phi(x)\, D w_i(x) + f(x, \Phi(x))\, w_i(x)\}\, dx = 0, \qquad 1 \leqq i \leqq N,$$

as an integration by parts shows, we may rewrite k_i as

$$k_i = \int_0^1 \{p(x)\, (D\tilde{w}(x) - D\Phi(x))\, D w_i(x) + \tilde{g}(x)\, (\tilde{w}(x) - \Phi(x))\, w_i(x)\}\, dx,$$

where $\tilde{g}(x) \equiv f_u(x, \Theta(x)\, \tilde{w}(x) + (1 - \Theta(x))\, \Phi(x))$ with $0 < \Theta(x) < 1$. By hypothesis, we know that $\tilde{g}(x)$ is a continuous function on $[0, 1]$, and moreover, from (1.16) and the fact that a priori bounds, independent of h, can be found for $\Phi(x)$ and $\tilde{w}(x)$ (cf. [2, Lemma 4]), then

$$-\Lambda < \gamma \leqq \tilde{g}(x) \leqq \tilde{\Gamma}, \qquad \text{for all} \quad 0 \leqq x \leqq 1,$$

where $\tilde{\Gamma}$ is some constant independent of h.

Next,

$$\int_0^1 p(x)\, (D\tilde{w}(x) - D\Phi(x))\, D w_i(x)\, dx = 0, \qquad 1 \leqq i \leqq N,$$

as an integration by parts shows, using (3.1) and (3.3), so that

$$k_i = \int_0^1 \tilde{g}(x)\, (\tilde{w}(x) - \Phi(x))\, w_i(x)\, dx, \qquad 1 \leqq i \leqq N. \tag{3.8}$$

As in [2], it is easy to see that the unique function $\hat{w}$ which minimizes $F[w]$ over the subspace S^{II} satisfies

$$\int_0^1 \{p(x) D\hat{w}(x) Dw_i(x) + f(x, \hat{w}(x)) w_i(x)\} dx = 0, \qquad 1 \leqq i \leqq N,$$

so that we may also express k_i as

$$k_i = \int_0^1 \{p(x)(D\tilde{w}(x) - D\hat{w}(x)) Dw_i(x) + \hat{g}(x)(\tilde{w}(x) - \hat{w}(x)) w_i(x)\} dx, \qquad 1 \leqq i \leqq N, \tag{3.9}$$

where $\hat{g}(x)$ similarly satisfies the bounds

$$-\Lambda < \gamma \leqq \hat{g}(x) \leqq \hat{\Gamma}, \qquad \text{for all} \quad 0 \leqq x \leqq 1,$$

$\hat{\Gamma}$ being some constant independent of h.

Writing $\tilde{w} = \sum_{i=1}^{N} \tilde{u}_i w_i$, and $\hat{w} = \sum_{i=1}^{N} \hat{u}_i w_i$, we obtain from (3.8)

$$\sum_{i=1}^{N} (\tilde{u}_i - \hat{u}_i) k_i = \int_0^1 \tilde{g}(x)(\tilde{w}(x) - \Phi(x))(\tilde{w}(x) - \hat{w}(x)) dx,$$

and similarly we obtain from (3.9)

$$\sum_{i=1}^{N} (\tilde{u}_i - \hat{u}_i) k_i = \int_0^1 \{p(x)[D\tilde{w}(x) - D\hat{w}(x)]^2 + \hat{g}(x)[\tilde{w}(x) - \hat{w}(x)]^2\} dx,$$

so that, using the norm of (2.6),

$$(\|\tilde{w} - \hat{w}\|_{\hat{g}})^2 = \int_0^1 \tilde{g}(x)(\tilde{w}(x) - \Phi(x))(\tilde{w}(x) - \hat{w}(x)) dx, \tag{3.10}$$

and hence from (2.7) and (3.10),

$$\begin{aligned} (\|\tilde{w} - \hat{w}\|_0)^2 &\leqq \{M(\gamma, \hat{\Gamma})\}^2 (\|\tilde{w} - \hat{w}\|_{\hat{g}})^2 \\ &\leqq C_1 \|\tilde{w} - \Phi\|_{L^2[0,1]} \|\tilde{w} - \hat{w}\|_{L^2[0,1]}, \end{aligned} \tag{3.11}$$

with $C_1 = \{M(\gamma, \hat{\Gamma})\}^2 (\max\{|\gamma|, |\tilde{\Gamma}|\})$. Since by (2.2),

$$\|\tilde{w} - \hat{w}\|_{L^2[0,1]} \leqq \|\tilde{w} - \hat{w}\|_{L^\infty[0,1]} \leqq \sqrt{r(1)}\, \|\tilde{w} - \hat{w}\|_0, \tag{3.12}$$

from (3.11), we obtain $\|\tilde{w} - \hat{w}\|_0 \leqq C_2 \|\tilde{w} - \Phi\|_{L^2[0,1]}$. Consequently, using (2.2),

$$\begin{aligned} \|\tilde{w} - \hat{w}\|_{L^\infty[0,1]} &\leqq \sqrt{r(1)}\, \|\tilde{w} - \hat{w}\|_0 \\ &\leqq C_3 \|\tilde{w} - \Phi\|_{L^2[0,1]} \leqq C_3 \|\tilde{w} - \Phi\|_{L^\infty[0,1]}. \end{aligned} \tag{3.13}$$

Thus, by combining the inequalities (3.12), (3.13), and (3.4) of Lemma 3, we finally have:

$$\begin{aligned} \|\hat{w} - \Phi\|_{L^\infty[0,1]} &\leqq \|\hat{w} - \tilde{w}\|_{L^\infty[0,1]} + \|\tilde{w} - \Phi\|_{L^\infty[0,1]} \\ &\leqq C_4 \|\tilde{w} - \Phi\|_{L^\infty[0,1]} \leqq C \cdot \ell(\Pi), \end{aligned}$$

which completes the proof. Q.E.D.

Several consequences of Theorem 5 can now be deduced. First, if the function $f(x, u)$ of (1.12) is *independent* of u, then the function $\tilde{g}(x)$ in (3.8) is necessarily zero, and it follows from (3.10) that $\tilde{w}(x) \equiv \hat{w}(x)$, and consequently, $\hat{w}(x)$ interpolates $\Phi(x)$ in the points x_i, $0 \leqq i \leqq N+1$. This gives us

Corollary 1. Let f of (1.12) be independent of u, and let $\hat{w}(x)$ be the unique Galerkin approximation in S^{Π} which minimizes the functional $F[w]$ over S^{Π}. Then, $\hat{w}(x) \equiv \tilde{w}(x)$ where $\tilde{w}(x)$ is the unique interpolate in S^{Π} of the solution $\Phi(x)$ of (1.12)—(1.10). Thus,

$$\hat{w}(ih) = \Phi(ih), \quad 0 \leqq i \leqq N+1. \tag{3.14}$$

The results of Theorem 5 and Corollary 1 make no assumptions on the partition Π of $[0, 1]$. In particular, the error bound of (3.7) of Theorem 5 does not resemble typical error bounds, in that no explicit dependence of $\ell(\Pi)$ on a mesh size h appears. For this reason, we investigate several types of partitions of $[0, 1]$. First, consider a *uniform* partition Π^h: $0 = x_0 < x_1 < \cdots < x_{N+1} = 1$ where $x_j = jh$, $0 \leqq j \leqq N+1 = h^{-1}$ of the interval $[0, 1]$. In this case,

$$\ell(\Pi^h) = h \left\{ \max_{0 \leqq i \leqq N} \int_{ih}^{(i+1)h} \frac{dt}{p(t)} \right\} \equiv h \cdot \tau(h). \tag{3.15}$$

Note that from (1.13)(iii), $\tau(h) = o(1)$ as $h \to 0$. Hence, for a uniform partition Π^h, the error bound of (3.7) of Theorem 5 is at least of order h. In the case that $p(x) \geqq \omega > 0$ for all $x \in [0, 1]$, then $\tau(h) = O(h)$ as $h \to 0$, so that $\ell(\Pi^h) = O(h^2)$, which agrees with the error bounds of [1] and [10]. On the other hand, if p vanishes at one end point of $[0, 1]$, say $x = 0$, then it is clear that for h sufficiently small,

$$\tau(h) = \int_0^h \frac{dt}{p(t)}.$$

Hence, $\ell(\Pi^h)$ is ultimately determined by the behavior of p in the neighborhood of $x = 0$. For example, if $p(x) = x^{\sigma}$, $0 \leqq \sigma < 1$, then $\tau(h) = \dfrac{h^{1-\sigma}}{1-\sigma}$ for all $h > 0$, and thus

$$\|\hat{w} - \Phi\|_{L^{\infty}[0,1]} \leqq C \cdot \ell(\Pi^h) = C_1 h^{2-\sigma}. \tag{3.16}$$

The result of (3.16) is an improvement of the recent results of Jamet [6, 7], who established by means of difference methods for *linear* problems a discrete result like that of (3.16), however, with an exponent of h reduced to $1-\sigma$. We further remark that, in the linear case, the determination of Jamet's discrete solution and the determination of the Galerkin approximation $\hat{w}(x)$ for our variational approach both depend upon the solutions of tridiagonal matrix equations, so that computationally, the methods are comparable. The numerical results of § 4 do confirm the rate of convergence of (3.16), so that the subspaces of S which correspond to the above uniform partitions $\Pi^{(h)}$ of $[0, 1]$ are attractive for such singular problems.

It is natural to ask if an error bound such as that of (3.16) is *sharp*. To show this, consider the function $\Phi(x) \equiv x^{1-\sigma} - x^{2-\sigma}$, which satisfies the boundary conditions of (1.10).

Then,

$$D\{x^{\sigma} D\Phi(x)\} = \sigma - 2, \quad 0 < x < 1,$$

and thus $\Phi(x)$ is the unique solution of (1.12)—(1.10) with $f(x, u) \equiv \sigma - 2$. Because f is independent of u in this case, then $\tilde{w}(x) = \hat{w}(x)$ from Corollary 1. Hence,

$$\|\hat{w} - \Phi\|_{L^{\infty}[0,1]} = \|\tilde{w} - \Phi\|_{L^{\infty}[0,1]} \geqq \|\tilde{w} - \Phi\|_{L^{\infty}[0,h]}.$$

By direct calculation, we find that

$$\|\tilde{w} - \Phi\|_{L^{\infty}[0,h]} = C_2 h^{2-\sigma},$$

where

$$C_2 \equiv \left\{\left(\frac{1-\sigma}{2-\sigma}\right)^{1-\sigma} - \left(\frac{1-\sigma}{2-\sigma}\right)^{2-\sigma}\right\}.$$

Hence, $\|\hat{w} - \Phi\|_{L^{\infty}[0,1]} \geqq C_2 h^{2-\sigma}$, showing that the error bound of (3.16) is sharp in this case.

Another form of partition of the interval $[0, 1]$ is suggested by the following considerations. Given a *fixed* number N of interior mesh points x_i, $1 \leqq i \leqq N$, of $(0, 1)$, then the function $\ell(\Pi)$ of (3.5), considered as a function only of the N variables $0 \leqq x_1 \leqq x_2 \leqq \cdots \leqq x_N \leqq 1$, clearly assumes its minimum for at least one *optimal* partition $\hat{\Pi}$. In addition, it is not hard to show that the associated N points $\hat{x}_i$ are distinct points in $(0, 1)$, i.e., $\hat{\Pi}$: $0 < \hat{x}_1 < \hat{x}_2 < \cdots < \hat{x}_N < 1$, and that $\ell(\hat{\Pi}) = \inf_{\Pi}\{\ell(\Pi)\}$, for a fixed N, if and only if

$$\ell(\hat{\Pi}) = (\hat{x}_{i+1} - \hat{x}_i) \int_{\hat{x}}^{\hat{x}_{i+1}} \frac{dt}{p(t)}, \quad \text{for each } i, \quad 0 \leqq i \leqq N.$$

Hence, the problem of finding an optimal partition amounts to finding N distinct points $\hat{x}_i \in (0, 1)$ with the property that all the quantities $(x_{i+1} - x_i) \int_{x_i}^{x_{i+1}} \frac{dt}{p(t)}$, $0 \leqq i \leqq N$, which occur in $\ell(\Pi)$ are *equal*. The advantage of an optimal partition is of course that for a given computational work (the solution of an $N \times N$ tridiagonal system), it is the one that minimizes the upper bound (3.7) of the error $\|\hat{w} - \Phi\|_{L^{\infty}[0,1]}$. However, searching for such an optimal partition may be time-consuming. Instead, the above characterization of optimal partitions may be used as follows: *Fix* the quantity $\tilde{\ell}(\Pi)$ in (3.7) so that the error $\|\hat{w} - \Phi\|_{L^{\infty}[0,1]}$ is less than or equal to a *given* accuracy. Then, define recursively $0 = \tilde{x}_0 < \tilde{x}_1 < \tilde{x}_2 < \cdots$, in such a way that

$$\tilde{\ell}(\Pi) = (\tilde{x}_{i+1} - \tilde{x}_i) \int_{\tilde{x}_i}^{\tilde{x}_{i+1}} \frac{dt}{p(t)}, \quad i = 0, 1, 2, \ldots.$$

If no $\tilde{x}_i$ is unity, set $\tilde{x}_{N+1} = 1$, where N is the last point of the above sequence in $(0, 1)$. Otherwise, if $\tilde{x}_j = 1$ for some j, set $\tilde{x}_{N+1} = 1$. In view of the above remarks,

such partitions $\tilde{\Pi}$: $0<\tilde{x}_1<\tilde{x}_2<\cdots<\tilde{x}_N<1$ can be called *quasi-optimal*, in that only the final quantity $(1-\tilde{x}_N)\int_{\tilde{x}_N}^{1}\frac{dt}{p(t)}$ may be smaller than $\tilde{\ell}(\Pi)$. Such quasi-optimal partitions have the advantage that for a fixed $\tilde{\ell}(\Pi)$, they lead to systems with the smallest possible number of unknowns, i.e., they minimize the computational work.

Finally, other finite-dimensional subspaces of S can also be considered. Specifically, if the subspace $H_0^1(\Pi)$ (cf. [2]) of continuous piecewise-linear functions of S is used with a uniform partition Π^h of mesh size h, then the analogue of (3.16) for the case $p(x)=x^\sigma$ is

$$\|\hat{w}-\Phi\|_{L^\infty[0,1]}\leqq C' h^{1-\sigma}, \tag{3.17}$$

which is essentially the same as Jamet's result. Moreover, the exponent of h in (3.17) can be shown to be best possible.

§ 4. Numerical Example

Consider the particular singular nonlinear boundary value problem:

$$D\{\sqrt{x}\,Du(x)\}=u^2-\left(\tfrac{3}{2}+x(1-x)^2\right),\quad 0<x<1, \tag{4.1}$$

subject to the homogeneous boundary conditions

$$u(0)=u(1)=0, \tag{4.2}$$

which corresponds to the choice $p(x)=x^\sigma$ with $\sigma=\frac{1}{2}$ in (1.12). By restricting our attention to nonnegative solutions, it is known that there exists a unique nonnegative solution $\Phi(x)$ of (4.1)—(4.2) (cf. [2, p. 419]), which is given explicitly by

$$\Phi(x)=x^{\frac{1}{2}}-x^{\frac{5}{2}},\quad 0\leqq x\leqq 1. \tag{4.3}$$

For our numerical example, we chose a *uniform* partition Π^h: $0=x_0<x_1<\cdots<x_{N+1}=1$ of $[0,1]$ with $x_j=jh$ and $h^{-1}=N+1$. Calling the resulting subspace of § 3 S^h, the nonnegative Galerkin approximation $\hat{w}_h(x)$ were determined for various values of the mesh spacing $h=1/(N+1)$. The results are given in the table below.

Table

h	dim S^h	$\|\Phi-\hat{w}_h\|_{L^\infty[0,1]}$	Exponent α
0.5	1	$1.3142\cdot 10^{-1}$	—
0.25	3	$4.6716\cdot 10^{-2}$	1.492
0.125	7	$1.6713\cdot 10^{-2}$	1.483
0.0625	15	$5.9575\cdot 10^{-3}$	1.488
0.03125	31	$2.1159\cdot 10^{-3}$	1.493
0.015625	63	$7.4962\cdot 10^{-4}$	1.497
0.0078125	127	$2.6507\cdot 10^{-4}$	1.500

The last column in the table gives the experimentally determined exponent α of h in $\|\hat{w}_h - \Phi\|_{L^\infty[0,1]} \doteqdot K h^\alpha$ for each halving of the mesh h. The theoretical value of α is, from (3.16) given by $2-\sigma = \frac{3}{2}$, which agrees well with the computed values of α.

Acknowledgment. The authors wish to express their sincere thanks to Professor Seymour V. Parter, and Dr. Christian Reinsch for many useful comments. In particular we thank Dr. Reinsch for obtaining the numerical results of the table.

References

1. Ciarlet, P. G.: An $O(h^2)$ method for a non-smooth boundary value problem. Aequat. Math. **2**, 39—49 (1968).
2. — Schultz, M. H., Varga, R. S.: Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem. Numer. Math. **9**, 394—430 (1967).
3. — — — Numerical methods of high-order accuracy for nonlinear two-point boundary value problems. Programmation en Mathématiques Numériques (Proceedings of the International Colloquium C.N.R.S., Besançon, France, Sept. 7—14, 1966), pp. 217—225. Paris: C.N.R.S. 1968.
4. — — — Numerical methods of high-order accuracy for nonlinear boundary value problems. V. Monotone operator theory. Numer. Math. **13**, 51—77 (1969).
5. Gusman, Yu. A., Oganesyan, L. A.: Inequalities for the convergence of finite difference schemes for degenerate elliptic equations. Z. Vycisl. Mat. i Mat. Fiz. **5**, 351—357 (1965).
6. Jamet, P.: Numerical methods and existence theorems for singular linear boundary-value problems. Doctoral Thesis, University of Wisconsin, 1967.
7. — On the convergence of finite-difference approximations to one-dimensional singular boundary-value problems. Numer. Math. **14**, 355—378 (1970).
8. Nečas, J.: Les Méthodes Directes en Théorie des Equations Elliptiques. Paris: Masson 1967. (351 pp.)
9. Parter, S. V.: Numerical methods for generalized axially symmetric potentials. J. Soc. Indust. Appl. Math. Ser. B Numer. Anal. **2**, 500—516 (1965).
10. Perrin, F. M., Price, H. S., Varga, R. S.: On higher-order numerical methods for nonlinear two-point boundary value problems. Numer. Math. **13**, 180—198 (1969).

Prof. R. S. Varga
Department of Mathematics
Kent State University
Kent, Ohio 44240, U.S.A.

Numer. Math. 15, 100—108 (1970)

Proving Inverse-Positivity of Linear Operators by Reduction *

JOHANN SCHRÖDER

Dedicated to Lothar Collatz on the occasion of his 60th birthday

Received December 19, 1969

1. Introduction

This paper provides tools for proving that a linear operator M is *inverse-positive*, that means: $Mu \geqq o$ *implies* $u \geqq o$. This property is frequently used in Applied Mathematics. For example, it is closely related to the boundary maximum principle for second order differential operators.

The abstract formulation of the above property, and a corresponding property for nonlinear operators, is due to Collatz (see [1] where such operators M are called "Operatoren monotoner Art", i.e. operators of monotone type). Collatz showed the importance of this property in Numerical Mathematics and, in particular, for obtaining error bounds.

An abstract theory on such operators provides sufficient and necessary conditions (see [2] and the references given there). In this paper, we develop methods for proving that these sufficient conditions are satisfied in more complicated cases where this is not immediately seen, for example, in case of higher order differential operators.

The provided methods reduce the problem of proving that M is inverse-positive to a corresponding problem for an operator $\hat{M} = LM$ (or $\hat{M} \geqq LM$). In some cases, a suitable operator L can immediately be constructed (Section 3); in others, several operators are used which theoretically can be combined in a single operator, but practically will not (Section 4). This last method is applied to operators $M = I - B$ with positive B, in Section 5. There are cases, when it is appropriate to actually carry out several reductions, one after the other (Section 6). For example, the order of differential operators can be reduced in this way, step by step.

2. Notations and Basic Assumptions

Let R and S denote real linear spaces and $M: R \to S$ a linear operator. Suppose that in both these spaces there are defined order relations, both denoted by $\leqq$.

All occurring order relations shall be reflexive, transitive and compatible with the linear structure. The space $(R, \leqq)$ will be called *partially ordered* if the relation $\leqq$ is also antisymmetric. By K_R we denote the cone of all $u \geqq o$ in R. For u in

* This work was sponsored by Boeing Scientific Research Laboratories, Document D 1-82-0938, December 1969, Seattle, Washington.

the algebraic interior (core) of K_R, we will write $u \succ o$. (In other words, $u \succ o$ if and only if, for each $v \in R$, there exists some number n with $nu + v \geqq o$.) Corresponding notations are used for S. For all occurring order relations, $> o$ shall be equivalent to $\geqq o$ and $+o$, so that, for example, $u \geqq o$ means that either $u \succ o$, or $u = o$.

We assume that $(R, \leqq)$ is *Archimedian* (that means K_R is *linearly closed*, or: for all $u, v \in R$, $nu + v \geqq o$ $(n = 1, 2, \ldots)$ implies $u \geqq o$). Suppose, moreover, that there is defined a second order relation $\trianglelefteqq$ in S which dominates the first in the following sense: For $U, V \in S$,

$$U \trianglerighteqq o \Rightarrow U \geqq o; \quad U > o, V \geqq o \Rightarrow U + V > o.$$

We are then interested in the following properties of M:

IP *$Mu \geqq o$ implies $u \geqq o$, for $u \in R$.*

IP′ *$Mu > o$ implies $u \succ o$, for $u \in R$.*

The operator M is called *inverse-positive* if **IP** holds. *If R is partially ordered and M has both properties,* **IP** *and* **IP′**, *then*

$$Mu \trianglerighteqq o \text{ implies } u \geqq o, \quad \text{for} \quad u \in R.$$

For, in this case, $Mu = o$ can only occur if $u = o$.

In case, $\succ$ is equivalent to $>$, the last property will be called *strong inverse-positivity:*

SIP *$Mu \geqq o$ implies $u \geqq o$, for $u \in R$.*

Theorem 1 in [2] contains sufficient conditions **I**, **II** for **IP** which can also be used for proving **IP′**. This paper is mainly concerned with providing methods for proving

I *$Mv \not\succ o$ for each $v \in R$ with $v \geqq o$, $v \not\succ o$.*

For describing these methods, we will need also other linear spaces with two order relations $\leqq$ and $\trianglelefteqq$ such that the second dominates the first. The letter X (also with subscript, etc.) will always denote such space. For these spaces, we will use corresponding notations as for R (K_X, $x \succ o$, etc.).

Furthermore, the letter L will always denote a linear operator mapping such space X (or S) into another such space $\hat{X}$ and satisfying the following two conditions. For each $x \in X$ $(x \in S)$,

$$x \geqq o \quad \text{implies} \quad Lx \geqq o, \tag{2.1}$$

$$x \succ o \quad \text{implies} \quad Lx \succ o. \tag{2.2}$$

(For some of the following results only one or the other of these properties is needed. Notice also that (2.1) follows from (2.2) if $(\hat{X}, \leqq)$ is Archimedian and if there exists an element $x > o$ in X.)

3. Simple Reduction

The following simple theorem is the basic tool for what is derived in this paper.

3.1. **Theorem.** *Let there exist a linear operator L mapping S into a linear space X such that* (2.1), (2.2) *hold and the following conditions are satisfied:*

I' $LMv \not> o$ *for each* $v \in R$ *with* $v \geqq o$, $v \not> o$.

II' *There exists an element* $z \in R$ *with*

$$z \geqq o, \quad LMz > o. \tag{3.2}$$

Then, M has the Properties **IP** *and* **IP'**.

Proof. Since $Mu \geqq o$ implies $LMu \geqq o$ and $Mu > o$ implies $LMu > o$, it suffices to prove the Properties **IP** and **IP'** for LM, instead of M. The above assumptions, however, are equivalent to the assumptions **I, II** in Theorem 1 [2], with M replaced by LM. According to that theorem, LM therefore has Property **IP**. This property together with **I'** yields that, $LMu > o$ implies $u \not> o$.

Remark. Instead of **II'** *one can require:*

II *There exists an element* $z \in R$ *with* $z \geqq o$, $Mz > o$.

This follows from (2.2).

The above theorem leads to the following

Method of Reduction. Find an operator $L: S \to X$ *satisfying* (2.1), (2.2) *such that* $LM = \hat{M} - N$ *where* $\hat{M}$ *has Property* **I** *and* N *is positive* $(u \geqq o \Rightarrow Nu \geqq o)$.

If then **II'** *or* **II** *holds, M has Properties* **IP** *and* **IP'**.

The last statement follows from the fact that **I'** is satisfied if $\hat{M}$ has Property **I**.

With this method, the problem involving M is reduced to a problem with the operator $\hat{M}$. This can have the following *two* advantages. First, $\hat{M}$ may be simpler, and second, Property **I** for $\hat{M}$ may be weaker than Property **I** for M. Notice, that Property **I** depends on the order relation $<$ defined in the range of the operator so that different such order relations belong to M and $\hat{M}$, respectively.

Example. Let $R = X = \mathbb{R}^n$, $S = \mathbb{R}^m$ and identify operators with matrices. Suppose that $\leqq$ denotes the natural (componentwise) order relation, in each of these spaces. For $U \in S$, define $U > o$ by, $U \not> o$ in case (i), $U > o$ in case (ii). For $x \in X$, $x > o$ and $x \not> o$ shall be equivalent.

Then, an $n \times m$-matrix L has the properties (2.1), (2.2) if and only if in case (i): each row of L contains an element > 0, in case (ii): all elements of L are > 0.

If the matrix $\hat{M}$ is diagonal, it has Property **I** (as an operator: $R \to X$). Thus, if we can find a matrix L of the prescribed type such that all off-diagonal elements of LM are $\leqq 0$, the matrix M has Properties **IP** and **IP'** (as an operator: $R \to S$), provided there exists a suitable vector z.

For example, we have

$$LM = \begin{bmatrix} 5 & 0 & 0 & 0 \\ -5 & 5 & 0 & 0 \\ 0 & -5 & 5 & 0 \\ 0 & 0 & -5 & 5 \end{bmatrix} \quad \text{for} \quad M = \begin{bmatrix} 3 & -1 & 0 & 0 \\ -3 & 3 & -1 & 0 \\ 1 & -3 & 3 & -1, \\ 0 & 1 & -3 & 2 \end{bmatrix} \quad L = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 6 & 4 & 2 \\ 2 & 4 & 6 & 3. \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

Since L has all elements >0 we consider case (ii). Then, z with $z^T=(1, 2, 3, 4)$ satisfies (3.2). Therefore, M has Property **SIP** which implies that M possesses an inverse with all elements >0.

4. Reducing with Several Operators

When a suitable operator L as needed in Theorem 3.1 cannot immediately be found the following result can be helpful.

4.1. **Theorem.** *Let the operator M have the following two properties.*

I'' *For each $v\in R$ with*

$$v\geqq o,\quad v\not> o \tag{4.2}$$

there exists a linear operator $L=L(v)$ mapping S into a linear space $X=X(v)$ such that (2.1) *and* (2.2) *hold, and*

$$LMv\not> o. \tag{4.3}$$

II *There exists an element $z\in R$ such that $z\geqq o$, $Mz> o$.*

Then, M has the Properties **IP** *and* **IP'**.

Remark. The theorem remains true if **II** *is replaced by,*

II'' *There exists an element $z\in R$ with*

$z\geqq o$, $L(v)Mz> o$ for all operators $L(v)$ occurring in **I''**.

Proof of Theorem and Remark. Let $\mathscr{V}$ be the set of all elements $v\in R$ satisfying $v\geqq o$, $v\not> o$ (i.e. $\mathscr{V}=\partial K$). We introduce the product space $\hat{X}=\prod_{v\in\mathscr{V}} X(v)$ with elements $\hat{x}=\{x(v)\}$ and define the order relations $\hat{x}\geqq o$ and $\hat{x}> o$ componentwise. Moreover, let the operator $\hat{L}$: $S\to\hat{X}$ be defined by $\hat{L}U=\{L(v)U\}$. This operator $\hat{L}$, instead of L, has the properties (2.1), (2.2). Moreover, the assumptions **I''** and **II''** imply that the conditions **I'**, **II'** of Theorem 3.1 are satisfied with L replaced by $\hat{L}$. Since **II''** is a consequence of **II**, the statements to be proven follow from Theorem 3.1.

If $X(v)$ is chosen to be $\mathbb{R}$ for all v and there exists a $U> o$ in S, **I''** obtains the following form when we write f for L.

I''' *For each $v\in R$ satisfying $v\geqq o, v\not> o$, there exists a linear functional f on S such that*

$$fU>0 \quad \textit{for all} \quad U> o, \quad \textit{and} \quad fMv\leqq 0. \tag{4.4}$$

This property has been applied earlier [4], and generalized to nonlinear problems [3]. In many cases, Property **I'''** is equivalent to **I** which can be shown by using separation theorems.

Suppose, that the set Y of all $U> o$ has the following property: Whenever a point $V\in S$ is not in Y, then there exists a linear functional f on S such that $fU>0$ for all $U\in Y$, and $fV\leqq 0$.

Under this condition, **I** *and* **I'''** *are equivalent.*

In order to show that **I** implies **I'''**, let $v\geqq o$ and $v\not> o$. Then, $Mv\not> o$ so that (4.4) follows from the above assumptions using $V=Mv$.

4.5. *Example.* Let L denote the differential operator, defined by $Lu = -u^{VI} + qu$ for $u \in C_6[0, 1]$ with given $q \in C[0, 1]$, and consider the boundary conditions,

$$u(0) = 0, \qquad u(1) = 0,$$
$$u^{IV}(0) - \alpha u'(0) = 0, \qquad u^{VI} + a u'(1) = 0,$$
$$u^{V}(0) + \beta u''(0) + \gamma u'(0), \qquad -u^{V}(1) + b u''(1) - c u'(1) = 0.$$

We assume that

$$q(s) \leqq 0 \quad (0 \leqq s \leqq 1), \quad \alpha \geqq 0, \quad a \geqq 0, \quad \beta \geqq 0, \quad b \geqq 0.$$

4.6. *Let there exist a function $z \in C_6[0, 1]$ satisfying the given boundary conditions such that $z(s) \geqq 0$, $(Lz)(s) \geqq 0$ and $(Lz)(s) \not\equiv 0$ $(0 \leqq s \leqq 1)$. Then, the following is true.*

If $u \in C_6[0, 1]$ satisfies the given boundary conditions and $L[u](s) \geqq 0$ $(0 \leqq s \leqq 1)$, then $u(s) \geqq 0$ $(0 \leqq s \leqq 1)$, and even $u \geqq o$ with the order relation $\geqq$ defined below (in (4.7)).

We will prove this statement by applying Theorem 4.1. Define $S = C[0, 1]$ and let R be the set of all $u \in C_6[0, 1]$ which satisfy the given boundary conditions. In both spaces, R and S, $\leqq$ shall denote the natural (pointwise) order relation. Then, for $u \in R$,

(4.7) $\quad u > o$ if and only if $u(s) > 0 \quad (0 < s < 1), \quad u'(0) > 0, \quad u'(1) < 0.$

For $u \in S$, let $U \gtrdot o$ be equivalent to $U \geqq o$. Moreover, define M to be the restriction of L from C_6 to R. We have then to prove Property **I''** for this operator.

Let $v \geqq o$, $v \not> o$ for some $v \in R$. Then either $v'(0) = 0$, or $v'(1) = 0$, or $v(\tau) = 0$ for some $\tau \in (0, 1)$, or several of these relations hold. For each such v, we have to find a suitable operator $L = L(v)$ mapping S into a space $X = X(v)$.

In case $v'(0) = 0$ we choose $X = \mathbb{R}$, and the linear functional $LU = \int_0^1 f(s)\, U(s)\, ds$ with $f(s) = 1 - s$. Obviously, (2.1) and (2.2) are true (for $x \in \mathbb{R}$, $x > o$ shall be equivalent to $x > 0$). By partially integrating twice, using the boundary conditions and observing $v''(0) \geqq 0$, $v'(1) \leqq 0$, we obtain $LMv \leqq 0$. Thus, (4.3) has been proved in this case.

In case $v'(1) = 0$ we proceed similarly using the functional $LU = \int_0^1 g(x)\, U(s)\, ds$ with $g(s) = s$.

Consider now the third case,

(4.8) $\qquad v(\tau) = 0 \quad$ for some $\quad \tau \in (0, 1), \quad v'(0) > 0, \quad v'(1) < 0.$

We calculate AMv with the operator A defined by

$$(AU)(t) = \int_0^1 K(s, t)\, U(s)\, ds, \qquad K(s, t) = \begin{cases} f(s)\, g(t) & \text{for} \quad 0 \leqq t \leqq s \leqq 1, \\ f(t)\, g(s) & \text{for} \quad 0 \leqq s \leqq t \leqq 1, \end{cases}$$

and again, $f(t) = 1 - t$, $g(t) = t$. Using similar means as above, we obtain $(AMv)(t) \leqq v^{IV}(t)$ $(0 \leqq t \leqq 1)$. In order to reduce the order of the occurring

derivatives further, we apply a second integral operator

$$(Bx)(r) = \int_0^1 G(t, r)\, x(t)\, dt, \qquad G(t, r) = \begin{cases} \Phi(t)\,\Psi(r) & \text{for } 0 \leqq r \leqq t \leqq 1 \\ \Phi(r)\,\Psi(t) & \text{for } 0 \leqq t \leqq r \leqq 1 \end{cases}$$

and $\Phi(t) = (1-t)^2$, $\Psi(t) = t^2$. By again partially integrating twice and observing $v'(\tau) = o$, $v''(\tau) \geqq o$, we obtain

$$\begin{aligned}(BAMv)(\tau) &\leqq (Bv^{IV})(\tau) \\ &= -2\tau(1-\tau)\, v''(\tau) + 2(1-2\tau)\, v'(\tau) - 2(1-\tau)^2 v'(0) + 2\tau^2 v'(1) < 0.\end{aligned}$$

For v satisfying (4.8), we define $X = C_0[0, 1]$ with $x \geqq o$ equivalent to the natural order relation, and $L = BA$. Then, (2.1) and (2.2) hold, and (4.3) is also true.

This proves the statement.

5. Application to Operators $M = I - B$ with Positive B

Let now the sets R and S be equal and $\leqq$ denote the same order relation, in R and S. We have, moreover, the order relations $\lesseqgtr$ in $S = R$, and $\lesssim$ in $R = S$. We write M as $M = I - B$ with the unit operator I and ask for conditions on the operator $B\colon R \to R$ such that M has Properties **IP** and **IP′**.

5.1. **Theorem.** *Suppose that the following three conditions are satisfied (for all $u \in R$):*

(i) $u \geqq o$ *implies* $Bu \geqq o$ *(B is positive).*

(ii) $\left.\begin{array}{l} u \geqq o \\ u > B^n u \quad (n = 1, 2, \ldots) \end{array}\right\}$ *implies* $u \succ o$.

(iii) *There exists an element $z \in R$ such that $z \geqq o$, $(I - B)z > o$.*

Then, the operator $M = I - B$ has the Properties **IP** *and* **IP′**.

Proof. Condition (ii) implies that for each $v \in R$ with $v \geqq o$, $v \not\succ o$, there exists a natural number n such that $(I - B^n)v \not> o$. Define then $L = L(v) = I + B + \cdots + B^{n-1}$. This operator maps $S = R$ into $X = R$ and satisfies $LMv = (I - B^n)v \not> o$ as well as (2.1) and (2.2). Thus, $M = I - B$ has Property **I″** of Theorem 4.1, and the statement to be proven here follows from that theorem.

For positive B, the assumption (ii) *is satisfied if* (a) *or* (b) *holds:*

(a) $u > o$ *is equivalent to* $u \succ o$.

(b) $u > o$ *implies* $B^n u \succ o$ *for some natural number n.*

Condition (b) in general requires that some power of B has a stronger positivity property. In view of Property **IP′** (and **SIP**) we like the second order relation $\lesseqgtr$ in S to be as weak as possible. The weaker this relation is, the stronger is condition (b).

Example. Let $R = \mathbb{R}^m$ and $\leqq$ denote the natural (componentwise) order relation. If $u > o$ is equivalent to $u \succ o$, property (b) means that some power B^n of the matrix B (the matrix associated with the operator B) has all its elements > 0. An example is

$$B = \tfrac{1}{4}\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{with} \quad B^2 = \tfrac{1}{16}\begin{bmatrix} 5 & 4 & 1 \\ 4 & 6 & 4 \\ 1 & 4 & 5 \end{bmatrix}.$$

Other definitions of $u > o$ also make sense. Let $R = \mathbb{R}^p \times \mathbb{R}^q$ with natural numbers $p, q, p+q=m$, and partition $u \in R$ correspondingly: $u^T = (u_1^T, u_2^T)$. Define then $u > o$ by $u_1 > o$, $u_2 > o$. In case $p=2$, $q=2$, the matrix

$$B = \begin{bmatrix} B_{11} & O \\ O & B_{22} \end{bmatrix} \quad \text{with} \quad B_{11} = B_{22} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

satisfies (b) with $n=1$.

Inverse-positivity of $M = I - B$ with positive B can also be proven by different means. For example, if the series $I + B + B^2 + \cdots$ converges to $(I-B)^{-1}$, in some sense to be defined, the positivity of the limit operator in general follows from the positivity of B. Under suitable assumptions (using topological terms) this convergence is guaranteed if and only if the spectral radius $\varrho(B)$ of B is <1. The above assumptions may then be considered as sufficient conditions for $\varrho(B) < 1$. Usually, $\varrho(B)$ will not be known.

6. Iterated Reductions

For certain operators, it is appropriate to perform several reductions. Let us assume that we are given linear operators L_k $(k=1, 2, \ldots, n)$ mapping a linear space X_{k-1} (with $X_0 = S$) into a linear space X_k such that all these operators have the properties (2.1), (2.2) (instead of L). Let, moreover,

$$L_k M_{k-1} = M_k - N_k \quad (k=1, 2, \ldots, n) \tag{6.1}$$

with $M_0 = M$ and linear operators M_k, N_k mapping R into X_k $(k=1, 2, \ldots, n)$. Then, we obtain the following result.

6.2. **Theorem.** *Suppose that the following three conditions are satisfied:*

(i) *All operators N_k are positive ($u \geqq o$ implies $N_k u \geqq o$).*

(ii) *$M_n v \not\geqq o$ for all $v \in R$ with $v \geqq o$, $v \not> o$.*

(iii) *There exists an element $z \in R$ such that*
$z \geqq o$, $Mz > o$, *or* $z \geqq o$, $\tilde{M} z > o$ *with* $\tilde{M} = L_n \ldots L_2 L_1 M$.

Then, M has the Properties **IP** *and* **IP'**.

Proof. We have $LM = M_n - N$ with

$$L = L_n L_{n-1} \ldots L_2 L_1, \tag{6.3}$$

$$N = N_n + L_n N_{n-1} + L_n L_{n-1} N_{n-2} + \cdots. \tag{6.4}$$

The operator $L: S \to X_n$ satisfies (2.1), (2.2), and $N: R \to X_n$ is positive. Thus, the n reductions described by (6.1) can be combined to one single reduction, and this theorem follows from Theorem 3.1.

6.5. **Corollary.** *Under the assumptions of Theorem* 6.2 *it is also true that, for all $u \in R$,*

$$\begin{aligned} &Mu \geqq o \text{ implies } u \geqq o, \quad M_k u \geqq o \quad (k=1, 2, \ldots, n), \\ &Mu > o \text{ implies } u > o, \quad M_k u > o \quad (k=1, 2, \ldots, n). \end{aligned} \tag{6.7}$$

This is an easy consequence of Theorem 6.2 and the positivity of the operators L_k and N_k.

The relation (6.7) means that M is inverse-positive with respect to the order relation $\leqq$ in S and the order relation in R, given by the set of inequalities $u \geqq o$, $M_k u \geqq o$ $(k = 1, 2, \ldots, n)$. One could very well start with this different order relation in the beginning. We will, however, not pursue this possibility here.

Although iterated reductions as in (6.1) are easily understood as one single reduction with the product operator L in (6.3), the method of iterated reductions seems to be quite fruitful, practically. We will discuss its meaning and its connection with other methods a little further.

If $(R, \leqq)$ is partially ordered and $\tilde{M} = L_n \ldots L_2 L_1 M$ is inverse-positive, we have $M^{-1} = \tilde{M}^{-1} L_n \ldots L_2 L_1$. Thus, the method of iterated reductions may be looked at as partly constructing the inverse M^{-1}. Of course, $\tilde{M}^{-1}$ is generally not known, either. What may be known is M_n^{-1} which operator is connected with M^{-1} as follows:

$$M^{-1} = M_n^{-1} L_n \ldots L_2 L_1 + \tilde{N} \quad \text{with} \quad \tilde{N} = M_n^{-1} N M^{-1}$$

and N as in (6.4). Since $\tilde{N}$ is positive (for positive N_k and inverse-positive M_n, M) we have obtained a minorant of M^{-1} in form of a product of positive operators, $M_n^{-1} L_n \ldots L_2 L_1$.

If the operators L_k have inverses P_k, the recursion formulaes (6.1) can be written as

$$M_{k-1} = P_k M_k - Q_k \quad (k = 1, 2, \ldots, n) \tag{6.8}$$

with $Q_k = P_k N_k$. Then, the whole method may be considered as one of splitting M — additively and multiplicatively — into operators P_k, M_k, Q_k as shown by the above formulaes.

To use splitting in order to prove inverse-positivity of differential operators has been suggested earlier, and this has been worked out in detail for ordinary differential operators of the fourth order (see [4] and the references given there). Some of those results have been carried over to differential operators of the third order and operators of higher order by Trottenberg [5]. The proofs in those papers are not the same as here, but related. This paper provides a more transparent approach to reduce the problem for a given operator M to a simpler problem. The following example shows how iterated reductions of type (6.1) can be used for treating a differential operator of the sixth order.

Example. Let L be as in Example 4.5 (with $q(s) \leqq 0$), but consider now the boundary conditions

$$\begin{aligned} &u(0) = 0, && u(1) = 0, \\ &u'''(0) - \alpha u''(0) + \beta u'(0) = 0, && -u'''(1) - a u''(1) - b u'(1) = 0, \\ &u^{\mathrm{IV}}(0) - \gamma u'(0), && u^{\mathrm{IV}}(1) + c u'(1) = 0 \end{aligned}$$

with nonnegative α, γ, a, c.

The statement 4.6 *holds also in this case.*

In order to prove this, define the spaces R, S, the operator M, and the order relations in R, S, as in Example 4.5. The operator $L_1 = A$ with A as given in that example maps S into

$$X_1 = \{x_1 \in C_2[0, 1]: x_1(0) = x_1(1) = 0\}.$$

For $x_1 \in X_1$, define $x_1 \geqq 0$ to hold pointwise, and $x_1 \succ o$ by $x_1 \succ o$, i.e. $x_1(s) > 0$ $(0 < s < 1)$, $x_1'(0) > 0$, $x_1'(1) < 0$. Then, L_1 has the properties (2.1), (2.2) and we calculate by partially integrating, $L_1 M = M_1 - N_1$ with

$$M_1 u = u^{\mathrm{IV}}, \qquad N_1 u = (1-s)\, u^{\mathrm{IV}}(0) + s\, u^{\mathrm{IV}}(1) - A(q\, u).$$

Because of the given boundary conditions and $q(s) \leqq 0$, the operator N_1 is positive.

If we can show that this fourth order differential operator M_1: $R \to X_1$ has Property **I**, the proof is completed. This does not depend on the way this is shown. We will apply a second reduction here.

The operator $L_2 = B$ with B defined in Example 4.5 maps X_1 into

$$X_2 = \{x_2 \in C_4[0, 1]: x_2(0) = x_2'(0) = x_2(1) = x_2'(1) = 0\}.$$

For $x_2 \in X_2$, define again $x_2 \geqq o$ to hold pointwise and $x_2 \succ o$ by $x_2 \succ o$, i.e. $x_2(s) > 0$ $(0 < t < 1)$, $x_2''(0) > 0$, $x_2''(1) > 0$. The operator L_2 satisfies (2.1), (2.2) also, and we obtain $L_2 M_1 = M_2 - N_2$ with N_2 the null operator, and

$$(M_2 u)(s) = -2s(1-s)\, u''(s) + 2(1-2s)\, u'(s) - 2(1-s)^2 u'(0) + 2s^2 u'(1).$$

It remains to be shown that M_2 has Property **I**. For that, we could try a third reduction. But, we will now use a direct proof.

If $v \in R$ with $v \geqq o$, $v \not\succ o$, then either $v(t) = 0$ for some $t \in (0, 1)$, or $v''(0) = 0$, or $v''(1) = 0$, or several of these relations hold. In the first case, we have $v'(t) = 0$, $v''(t) \geqq 0$ so that $(M_2 u)(t) \leqq 0$ and thus, $M_2 v \not\succ o$.

In the second case, $v'(0) = 0$, we calculate $(M_2 v)'(0) = 0$ and $(M_2 v)''(0) = -2v'''(0) + 4v'(1)$. This last expression is nonpositive because of the boundary conditions and $v''(0) \geqq 0$, $v'(1) \leqq 0$. Thus, $M_2 v \not\succ o$ is shown again.

The third case, $v'(1) = 1$, is handled correspondingly.

Notice, that these arguments do not carry through if a splitting $L_2 M_1 = \tilde{M}_2 - \tilde{N}_2$ with $\tilde{N}_2 v = 2(1-s)^2 v'(0) - 2s^2 v'(1)$ is used.

References

1. Collatz, L.: Aufgaben monotoner Art. Archiv Math. **3**, 365–376 (1952).
2. Schröder, J.: Differential inequalities and error bounds. Error in digital computation, Vol. 2 (Ed. L. B. Rall), pp. 141–179. John Wiley 1965.
3. — Monotonie-Aussagen bei quasilinearen elliptischen Differentialgleichungen und anderen Problemen. Numerische Mathematik, Differentialgleichungen, Approximationstheorie (Ed. L. Collatz and G. Meinardus), pp. 341–361. Birkhäuser 1968.
4. — On linear differential inequalities. J. Math. Anal. Appl. **22**, 188–216 (1968).
5. Trottenberg, U.: Über nichtnegative Greensche Funktionen bei gewöhnlichen Differentialgleichungen. Diplomarbeit, Universität Köln, 1969.

Prof. Dr. J. Schröder
Mathem. Institut der Universität Köln
D 5000 Köln 41
Weyertal 86–90

Numer. Math. 15, 109—121 (1970)

Zum Anwendungsbereich einiger Iterationsverfahren

F. W. SCHÄFKE

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 29. Dezember 1969

0. Einleitung

Schröder hat 1956 in seiner Arbeit „Das Iterationsverfahren bei allgemeinerem Abstandsbegriff" [8], in Anlehnung an Kantorovitch [4], den bekannten Banach-Weissingerschen Satz über Fixpunkte von Kontraktionen und ihre iterative Gewinnung, [11], auf „pseudometrische Räume", vgl. Collatz [2], übertragen, bei denen die Werte der Abstandsfunktion in einem reellen halbgeordneten linearen Raum liegen, in dem für gewisse Folgen ein eindeutiger limes mit bestimmten Eigenschaften gegeben ist. Schröder zeigt insbesondere, daß seine Methoden auch numerisch anpassungsfähiger sind und zu besseren Abschätzungen führen können als die gewohnte Anwendung metrischer Räume.

Die von Schröder gegebene Verallgemeinerung des Fixpunktsatzes für Kontraktionen hat gegenüber anderen Ansätzen den Vorzug, daß der geforderte Konvergenzbegriff nicht eng mit der gegebenen Halbordnung gekoppelt ist. Andererseits wird jedoch auf den Vorteil einer uniformen Topologie, vgl. Weil [10] und Köthe [5], §5, §6, die bei vielen Beispielen durchaus gegeben wäre, verzichtet. Auch erscheint es theoretisch unbefriedigend und im Hinblick auf Anwendungen einschränkend, daß der Wertebereich der „Pseudometrik" doch eine recht komplizierte Struktur darstellt, bei der u.a. auch die reellen Zahlen wieder auftreten. Eine direkte Anwendung auf Prozesse in Computern erscheint so ausgeschlossen.

Einen wesentlich einfacheren Wertebereich einer verallgemeinerten Metrik betrachten 1960 Albrecht und Karrer [1]. Sie wählen hierfür eine halbgeordnete abelsche Halbgruppe mit gewissen Eigenschaften und formulieren u.a. unter zusätzlichen Annahmen eine Verallgemeinerung des Banach-Weissingerschen Satzes. Reinermann [6], Abschnitt 3, gibt 1968 eine lokale Version des Satzes für einen solchen Raumtyp. In diesen Arbeiten wird auf Anwendungen kaum Bezug genommen. Eine genauere Untersuchung zeigt auch, daß hier die gewählte Topologie allzu eng an die Halbordnung gebunden ist und dies nicht-triviale Anwendungen weitgehend ausschließt.

Die vorliegende Note verfolgt das Ziel, wesentlich allgemeiner als in den eben genannten Arbeiten, Sätze über das Iterationsverfahren für „kontrahierende" Abbildungen und über das vereinfachte Newton-Verfahren unter möglichst minimalen Voraussetzungen zu erhalten, um damit einen wesentlich erweiterten Anwendungsbereich zu erreichen. Diese beiden Sätze sind, abgesehen von ihrer

zentralen Bedeutung, beispielhaft gewählt. Für verwandte Sätze bzw. Modifikationen oder Ergänzungen, vgl. z.B. Ehrmann [3], und Schröder [9], läßt sich ähnliches durchführen.

In Abschnitt **1.** werden die Begriffe der *quasimetrischen limes-Struktur* und des *quasimetrischen limes-Raumes* eingeführt. Dazu werden einige einfache Eigenschaften und einige abkürzende Verabredungen notiert. **2.** formuliert dann die angekündigte wesentliche Verallgemeinerung des Fixpunktsatzes für Kontraktionen. Hier wie im folgenden sind die Bezeichnungen so gewählt, daß gegenüber dem Gewohnten kaum Veränderungen auftreten. Abschnitt **3.** erklärt den Begriff der *quasinormierten limes-Gruppe* und **4.** bringt in diesem Rahmen einen Satz über das vereinfachte Newton-Verfahren. Abschnitt **5.** erklärt *quasimetrische Räume* und *quasinormierte Gruppen*. Diese Strukturen sind von Interesse, da sie einerseits noch zu uniformen Topologien führen, andererseits in kanonischer Weise quasimetrische limes-Räume bzw. quasinormierte limes-Gruppen darstellen. Die Schlußbemerkungen von **7.** geben einerseits Hinweise auf Anwendungsmöglichkeiten und andererseits auf auftretende rein mathematische Fragen.

1. Quasimetrische limes-Räume

Es werde angenommen:

(1.1) $\mathfrak{H}$ *ist eine nicht-leere Menge,*
$+$ *eine zweistellige Operation in* $\mathfrak{H}$,
$\leqq$ *eine zweistellige Relation in* $\mathfrak{H}$;

(1.2) $\mathbb{K}$ *ist eine Teilmenge von* $\mathfrak{H}^{\mathbb{N}}$, *und*
$\mathfrak{N} := \{\alpha : \alpha \in \mathfrak{H},\ \xi_n = \alpha\,(n \in \mathbb{N}) \rightsquigarrow \{\xi_n\} \in \mathbb{K}\}$ *ist nicht leer;*

(1.3) *für* $\xi_n \in \mathfrak{N}$ $(n \in \mathbb{N})$ *gilt* $\{\xi_n\} \in \mathbb{K}$;

(1.4) *aus* $x \in \mathbb{K}$ *und* $y \in \mathbb{K}$ *folgt* $x + y \in \mathbb{K}$;

(1.5) *ist* $x \in \mathbb{K}$ *und* $y \leqq x$, *so gilt* $y \in \mathbb{K}$.

Dabei sind $+$ und $\leqq$ für Folgen gliedweise definiert.

$(\mathfrak{H}, +, \leqq, \mathbb{K})$ heiße dann eine *quasimetrische limes-Struktur.*
Sei zusätzlich $\mathfrak{R}$ eine nicht-leere Menge und

$$\varrho : \mathfrak{R} \times \mathfrak{R} \to \mathfrak{H} \tag{1.6}$$

eine Abbildung mit den Eigenschaften

$$\forall a \in \mathfrak{R} \qquad \varrho(a, a) \in \mathfrak{R}, \tag{1.7}$$

$$\forall (a, b) \in \mathfrak{R}^2 \qquad \varrho(b, a) = \varrho(a, b), \tag{1.8}$$

$$\forall (a, b, c) \in \mathfrak{R}^3 \qquad \varrho(a, c) \leqq \varrho(a, b) + \varrho(b, c). \tag{1.9}$$

Dann heiße

$$(\mathfrak{R}, (\mathfrak{H}, +, \leqq, \mathbb{K}), \varrho)$$

ein *quasimetrischer limes-Raum* und ϱ eine *Quasimetrik* für $\mathfrak{R}$ bezüglich der quasimetrischen limes-Struktur $(\mathfrak{H}, +, \leqq, \mathbb{K})$.

Für einen solchen quasimetrischen limes-Raum werde definiert:

$$a \sim b :\leftrightarrows \varrho(a, b) \in \mathfrak{N}, \tag{1.10}$$

$$\{x_n\} \dot{\sim} \{y_n\} :\leftrightarrows \{\varrho(a_n, y_n)\} \in \mathbb{K}, \tag{1.11}$$

$$x_n \to x :\leftrightarrows \{\varrho(x_n, x)\} \in \mathbb{K}. \tag{1.12}$$

Es gilt dann

$$\sim \text{ ist Äquivalenzrelation in } \mathfrak{N}; \tag{1.13}$$

$$\dot{\sim} \text{ ist Äquivalenzrelation in } \mathfrak{N}^{\mathbb{N}}; \tag{1.14}$$

$$x_n \sim x \ (n \in \mathbb{N}) \rightarrow x_n \to x; \tag{1.15}$$

$$x_n \to x \rightarrow [x_n \to y \leftrightarrows x \sim y]; \tag{1.16}$$

$$x_n \to x \rightarrow [y_n \to x \leftrightarrows \{x_n\} \dot{\sim} \{y_n\}]. \tag{1.17}$$

Hier und im folgenden soll bei Folgen stets $n \in \mathbb{N}$, d.h. $n = 1, 2, 3, \ldots$, laufen. Das ist wesentlich, da keine Teilfolgeneigenschaft gefordert ist.

Beweise. Zu (1.13): Nach (1.2), (1.4), (1.5) hat $\mathfrak{N}$ die Eigenschaften $\mathfrak{N} + \mathfrak{N} \subset \mathfrak{N}$ und

$$\beta \in \mathfrak{N} \wedge \alpha \leqq \beta \rightarrow \alpha \in \mathfrak{N}.$$

(1.7), (1.8), (1.9) geben damit unmittelbar Reflexivität, Symmetrie und Transitivität.

Zu (1.14): Die Reflexivität folgt aus (1.3), die Symmetrie aus (1.8). Die Transivität folgt wegen (1.9) mit

$$\{\varrho(x_n, z_n)\} \leqq \{\varrho(x_n, y_n)\} + \{\varrho(y_n, z_n)\}$$

durch Anwendung von (1.4) und (1.5).

Zu (1.15): Unmittelbare Folge von (1.7) und (1.3).

Zu (1.16): Man hat mit (1.8) und (1.9)

$$\{\varrho(x, y)\} \leqq \{\varrho(x_n, x)\} + \{\varrho(x_n, y)\}$$

und

$$\{\varrho(x_n, y)\} \leqq \{\varrho(x_n, x)\} + \{\varrho(x, y)\}.$$

Das gibt mit (1.4), (1.5) und (1.2) die Behauptungen.

Zu (1.17): Dies folgt analog aus

$$\{\varrho(x_n, y_n)\} \leqq \{\varrho(x_n, x)\} + \{\varrho(y_n, x)\}$$

und

$$\{\varrho(y_n, x)\} \leqq \{\varrho(y_n, x_n)\} + \{\varrho(x_n, x)\}.$$

Eine Folge $\{x_n\}$ heiße *Cauchyfolge*, wenn für jede Folge $k(n)$ natürlicher Zahlen mit $k(n) > n$ $(n \in \mathbb{N})$ gilt $\{\varrho(x_n, x_{k(n)})\} \in \mathbb{K}$. Der quasimetrische limes-Raum heiße *folgenvollständig*, wenn zu jeder Cauchyfolge $\{x_n\}$ ein $x \in \mathfrak{N}$ mit $x_n \to x$ existiert.

Ist $\mathfrak{M} \subset \mathfrak{R}$, so bezeichne

$$\hat{\mathfrak{M}} := \{x: \exists\, x_n \in \mathfrak{M}\ (n \in \mathbb{N})\ x_n \to x\}. \tag{1.18}$$

Nach (1.15) gilt stets

$$\mathfrak{M} \subset \hat{\mathfrak{M}}. \tag{1.19}$$

Für eine quasimetrische limes-Struktur $(\mathfrak{H}, +, \leqq, \mathbb{K})$ verabreden wir folgende *Abkürzungen:*

$$\alpha_1 + \alpha_2 + \cdots + \alpha_n := (\alpha_1 + \cdots + \alpha_{n-1}) + \alpha_n \quad (n \geqq 3); \tag{1.20}$$

$$\sum_{\nu=1}^{\infty} \alpha_\nu \leqq \tau :\frown \forall n \in \mathbb{N} \quad \alpha_1 + \cdots + \alpha_n \leqq \tau; \tag{1.21}$$

(1.22) „$\sum_{\nu=1}^{\infty} \alpha_\nu$ konvergent" heißt: für jede Folge $k(n)$ natürlicher Zahlen mit $k(n) > n$ und $\beta_n := \alpha_n + \cdots + \alpha_{k(n)-1}$ gilt $\{\beta_n\} \in \mathbb{K}$;

(1.23) ist $\emptyset \neq \mathfrak{M} \subset \mathfrak{H}$ und

$$P: \mathfrak{M} \to \mathfrak{H},$$

so heißt P „folgenstetig", wenn für $\alpha_n \in \mathfrak{M}$

$$\{\alpha_n\} \in \mathbb{K} \frown \{P\alpha_n\} \in \mathbb{K},$$

und P „monoton", wenn für $(\alpha, \beta) \in \mathfrak{M}^2$

$$\alpha \leqq \beta \frown P\alpha \leqq P\beta.$$

2. Fixpunktsatz für kontrahierende Abbildungen

Im folgenden nehmen wir an

(2.1) $(\mathfrak{R}, (\mathfrak{H}, +, \leqq, \mathbb{K}), \varrho)$ *sei quasimetrischer limes-Raum;*

(2.2) *dieser sei folgenvollständig;*

(2.3) $\leqq$ *sei auf* $\mathfrak{H}$ *transitiv;*

(2.4) $\leqq$ *und* $+$ *erfüllen das Monotoniegesetz*

$$\forall (\alpha, \beta, \gamma) \in \mathfrak{H}^3, \quad \alpha \leqq \beta \frown \alpha + \gamma \leqq \beta + \gamma \wedge \gamma + \alpha \leqq \gamma + \beta.$$

Dann gilt

Satz 1. *Voraussetzungen.*

(2.5) *Sei entweder*

a) $\varepsilon \in \mathfrak{H}$, $x_0 \in \mathfrak{R}$, $\varrho(x_0, x_0) \leqq \varepsilon$ *und*

$$\mathfrak{K} := \{y: y \in \mathfrak{R} \wedge \varrho(y, x_0) \leqq \varepsilon\}$$

oder

b) $x_0 \in \mathfrak{R}$ *und* $\mathfrak{K} := \mathfrak{R}$;

(2.6) *sei*

$$T: \hat{\mathfrak{K}} \to \mathfrak{R}$$

und

$$\varrho(\hat{\mathfrak{R}}, \hat{\mathfrak{R}}) \subset \mathfrak{M} \subset \mathfrak{H};$$

(2.7) *mit Abbildungen*

$$P_n: \mathfrak{M} \to \mathfrak{H} \quad (n \in \mathbb{N})$$

gelte

α) $\forall (x, y) \in \mathfrak{D}_{T^n}^2 \quad \varrho(T^n x, T^n y) \leqq P_n \varrho(x, y)$,

β) $\forall \alpha \in \mathfrak{M} \quad \sum_{n=1}^{\infty} P_n \alpha$ *konvergent*,

γ) P_1 *ist folgenstetig*,

δ) *im Falle* a) *mit* $\sigma := \varrho(x_0, T x_0)$

$$\sigma + \sum_{n=1}^{\infty} P_n \sigma \leqq \varepsilon.$$

Behauptungen.

(2.8) *Man hat* $x_0 \in \mathfrak{D}_{T^n}$ *und* $T^n x_0 \in \mathfrak{R}$ *für* $n \in \mathbb{N}$.

(2.9) *Es existiert* $\hat{x} \in \hat{\mathfrak{R}}$ *mit* $T^n x_0 \to \hat{x}$.

(2.10) *Ist* $y_0 \in \mathfrak{D}_{T^n}$ $(n \in \mathbb{N})$, *so gilt* $T^n y_0 \to \hat{x}$.

(2.11) *Es gilt* $T\hat{y} \sim \hat{y} \in \hat{\mathfrak{R}}$ *genau dann, wenn* $\hat{y} \sim \hat{x}$.

(2.12) *Gilt für ein* $n \in \mathbb{N}$ *und* $\varepsilon_n \in \mathfrak{H}$

$$\sum_{m=n}^{\infty} P_m \sigma \leqq \varepsilon_n$$

und ist

$$\mathfrak{R}_n := \{y: \varrho(y, T^n x_0) \leqq \varepsilon_n\}$$

so gilt

$$\hat{x} \in \hat{\mathfrak{R}}_n.$$

Beweis. Zu (2.8): Hier ist nur für den Fall (2.5), a) etwas zu zeigen. Dazu hat man zunächst nach (2.5), (1.19) und (2.6) $x_0 \in \mathfrak{D}_T$ und nach (2.7), δ) $\varrho(x_0, T x_0) \leqq \varepsilon$, also $T x_0 \in \mathfrak{R}$ und $x_0 \in \mathfrak{D}_{T^2}$. Für $n = 2, 3, 4, \ldots$ zeigt man nun induktiv die — mit A_n bezeichnete — Aussage:

$$x_0 \in \mathfrak{D}_{T^n} \wedge \varrho(x_0, T^n x_0) \leqq \sigma + \sum_{\nu=1}^{n-1} P_\nu \sigma \leqq \varepsilon.$$

Für $n = 2$ ist der erste Teil schon bewiesen; der zweite folgt mit (1.9) und

$$\varrho(x_0, T^2 x_0) \leqq \varrho(x_0, T x_0) + \varrho(T x_0, T^2 x_0),$$

was wegen (2.3), (2.4) und (2.7), α), δ)

$$\varrho(x_0, T^2 x_0) \leqq \sigma + P_1 \sigma \leqq \varepsilon$$

ergibt. Sei nun für $2 \leqq n \in \mathbb{N}$ A_n bekannt. Dann hat man mit der Transivität (2.3) $\varrho(x_0, T^n x_0) \leqq \varepsilon$, $T^n x_0 \in \mathfrak{R}$, $x_0 \in \mathfrak{D}_{T^{n+1}}$. Weiter folgt mit (1.9), (2.3), (2.4), A_n

und (2.7), α)

$$\varrho(x_0 T^{n+1} x_0) \leqq \varrho(x_0, T^n x_0) + \varrho(T^n x_0, T^{n+1} x_0),$$

$$\varrho(x_0, T^{n+1} x_0) \leqq \left(\sigma + \sum_{\nu=1}^{n-1} P_\nu \sigma\right) + P_n \sigma$$

und damit A_{n+1}. Das ergibt sofort (2.8).

Zu (2.9): Für $k(n) > n$ hat man wegen (1.9), (2.3), (2.4) und (2.7), α)

$$(*) \qquad \varrho(T^n x_0, T^{k(n)} x_0) \leqq P_n \sigma + \cdots + P_{k(n)-1} \sigma.$$

Hier liefert nun nach (2.7), β) und (1.22) die rechte Seite für $n \in \mathbb{N}$ eine Folge aus $\mathbb{K}$, also nach (1.5) auch die linke. Das aber besagt nach **1.**, daß $\{T^n x_0\}_1^\infty$ Cauchyfolge ist. Wegen (2.2) gibt es ein $\hat{x} \in \mathfrak{R}$ mit $T^n x_0 \to \hat{x}$. Nach (1.18) gilt $\hat{x} \in \hat{\mathfrak{R}}$.

Zu (2.10): Nach (2.7), α) gilt

$$\varrho(T^n y_0, T^n x_0) \leqq P_n \varrho(y_0 x_0).$$

Mit (2.7), β) folgt, daß die rechte Seite für $n \in \mathbb{N}$ eine Folge aus $\mathbb{K}$ liefert, mit (1.5) also auch die linke. Das besagt nach (1.11) $\{T^n y_0\} \dot{\sim} \{T^n x_0\}$. Dies aber liefert mit (1.17) $T^n y_0 \to \hat{x}$.

Zu (2.11): Wegen $T x_0 \in \mathfrak{D}_{T^n}$ ($n \in \mathbb{N}$) folgt aus (2.10) $T^{1+n} x_0 \to \hat{x}$. Andererseits geben (2.7), α), γ) mit

$$\varrho(T^{1+n} x_0, T\hat{x}) \leqq P_1 \varrho(T^n x_0, \hat{x})$$

und (1.5) $T^{1+n} x_0 \to T\hat{x}$. Mit (1.16) gibt das $T\hat{x} \sim \hat{x}$. Weiter geben (2.7), α), γ) zunächst

$$\forall (a, b) \in \hat{\mathfrak{R}}^2 \qquad a \sim b \rightharpoonup Ta \sim Tb.$$

Ist nun $\hat{y} \sim \hat{x}$, so gilt wegen (1.16) $\hat{y} \in \hat{\mathfrak{R}}$ und

$$T\hat{y} \sim T\hat{x} \sim \hat{x} \sim \hat{y},$$

also $T\hat{y} \sim \hat{y}$. Ist umgekehrt

$$\hat{\mathfrak{R}} \ni \hat{y} \sim T\hat{y},$$

so hat man $\hat{y} \in \mathfrak{D}_{T^n}$ ($n \in \mathbb{N}$) und $T^n \hat{y} \sim \hat{y}$, also nach (1.15) $T^n \hat{y} \to \hat{y}$. Nach (2.10) gilt andererseits $T^n \hat{y} \to \hat{x}$, also nach (1.16) $\hat{y} \sim \hat{x}$.

Zu (2.12): Man verwendet für $m > n$

$$\varrho(T^m x_0, T^n x_0) \leqq P_n \sigma + \cdots + P_{m-1} \sigma \leqq \varepsilon_n$$

und mit $\nu = 1, 2, 3, \ldots$ als Folgenindex

$$T^{n+\nu} x_0 \to \hat{x}$$

aufgrund von $T^n x_0 \in \mathfrak{D}_{T^\nu}$ ($\nu \in \mathbb{N}$) und (2.10). Das gibt die Behauptung.

3. Quasinormierte limes-Gruppen

Sei $(\mathfrak{H}, +, \leqq, \mathbb{K})$ eine quasimetrische limes-Struktur gemäß **1.** und $(\mathfrak{G}, +)$ eine, additiv geschriebene, jedoch nicht notwendig abelsche, Gruppe.

Sei dann eine Abbildung

(3.1) $$\| \| : \mathfrak{G} \to \mathfrak{H}$$

gegeben mit den Eigenschaften

(3.2) $$\|0\| \in \mathfrak{N},$$

(3.3) $$\forall (a, b) \in \mathfrak{G}^2 \qquad \|a+b\| = \|-a-b\| \leqq \|a\| + \|b\|.$$

Wir nennen dann

$$(\mathfrak{G}, +, (\mathfrak{H}, +, \leqq, \mathbb{K}), \| \|)$$

eine *quasinormierte limes-Gruppe* und $\| \|$ eine *Quasinorm* für bezüglich der gegebenen quasimetrischen limes-Struktur.

Offenbar folgen aus $\|a+b\| = \|-a-b\|$ die beiden Eigenschaften

(3.4) $$\|-a\| = \|a\|, \qquad \|b+a\| = \|a+b\|$$

und umgekehrt.

Definiert man

(3.5) $$\varrho(a, b) := \|a-b\|,$$

so gilt:

(3.6) *ϱ ist eine Quasimetrik für $\mathfrak{G}$ bezüglich der quasimetrischen limes-Struktur $(\mathfrak{H}, +, \leqq, \mathbb{K})$. ϱ ist zweiseitig invariant: für $(a, b, c) \in \mathfrak{G}^3$ gilt*

$$\varrho(a, b) = \varrho(a+c, b+c) = \varrho(c+a, c+b).$$

Beweis. (1.7) folgt aus (3.2), (1.8) aus (3.4):

$$\varrho(b, a) = \|b-a\| = \|-(b-a)\| = \|a-b\| = \varrho(a, b).$$

(1.9) ergibt sich mit (3.3):

$$\varrho(a, c) = \|(a-b)+(b-c)\| \leqq \|a-b\| + \|b-c\| = \varrho(a, b) + \varrho(b, c).$$

Die rechtsseitige Invarianz folgt mit (3.5) sofort:

$$\varrho(a+c, b+c) = \|(a+c)-(b+c)\| = \|a-b\| = \varrho(a, b);$$

die linksseitige Invarianz benutzt noch (3.4):

$$\varrho(c+a, c+b) = \|(c+a-b)-c\| = \|a-b\| = \varrho(a, b).$$

Umgekehrt gilt

(3.7) *Ist $(\mathfrak{G}, +)$ Gruppe und ϱ eine zweiseitig invariante Quasimetrik für $\mathfrak{G}$, so liefert*

$$\|a\| := \varrho(a, 0)$$

eine Quasinorm für $\mathfrak{G}$, aus der man mit (3.5) ϱ zurückerhält.

Beweis. (3.2) ist mit (1.7) gegeben. (3.3) folgt mit der zweiseitigen Invarianz und (1.8), (1.9):

$$\|a+b\| = \varrho(a+b, 0) = \varrho(b, -a) = \varrho(0, -a-b) = \varrho(-a-b, 0) = \|-a-b\|$$

$$\|a+b\| = \varrho(a+b, 0) \leqq \varrho(a+b, b) + \varrho(b, 0) = \varrho(a, 0) + \varrho(b, 0) = \|a\| + \|b\|.$$

Schließlich wird

$$\varrho(a, b) = \varrho(a - b, 0) = \|a - b\|.$$

Wir bemerken weiter (vgl. (1.13)):

(3.8) $\sim$ *ist Kongruenzrelation in* $\mathfrak{G}$.

Beweis. Unmittelbare Folge der Definition (1.13) und der zweiseitigen Invarianz von ϱ.

Abschließend notieren wir die Folgenstetigkeit der Bildung des Inversen und der Operation $+$ in $\mathfrak{G}$:

(3.9) *Aus* $x_n \to x$ *folgt* $-x_n \to -x$. *Aus* $x_n \to x$ *und* $y_n \to y$ *folgt* $x_n + y_n \to x + y$.

Beweis. Die erste Aussage folgt wegen $\|-x_n + x\| = \|x_n - x\|$ sofort aus (1.12). Für die zweite hat man mit (1.9) und (3.6)

$$\varrho(x_n + y_n, x + y) \leqq \varrho(x_n + y_n, x + y_n) + \varrho(x + y_n, x + y) = \varrho(x_n, x) + \varrho(y_n, y).$$

Mit (1.12), (1.4), (1.5) folgt daraus in üblicher Weise die Behauptung.

4. Das vereinfachte Newton-Verfahren

Im folgenden nehmen wir an:

(4.1) $(\mathfrak{G}, +, (\mathfrak{H}, +, \leqq, \mathbb{K}), \| \|)$ *sei quasinormierte limes-Gruppe;*

(4.2) $\mathfrak{G}$ *sei als quasimetrischer limes-Raum* (vgl. (3.6)) *folgenvollständig;*

(4.3) $\leqq$ *sei auf* $\mathfrak{H}$ *transitiv;*

(4.4) $\leqq$ *und* $+$ *erfüllen das Monotoniegesetz:*

$$\forall (\alpha, \beta, \gamma) \in \mathfrak{H}^3 \qquad \alpha \leqq \beta \rightsquigarrow \alpha + \gamma \leqq \beta + \gamma \wedge \gamma + \alpha \leqq \gamma + \beta.$$

Wir benutzen die Abkürzungen (1.20), (1.21), (1.22), (1.23).

Dann gilt

Satz 2. *Voraussetzungen.*

(4.5) *Sei entweder*

a) $\varepsilon \in \mathfrak{H}$, $x_0 \in \mathfrak{G}$, $\|0\| \leqq \varepsilon$ *und*

$$\mathfrak{K} := \{y : y \in \mathfrak{G} \wedge \|y - x_0\| \leqq \varepsilon\}$$

oder

b) $x_0 \in \mathfrak{G}$ *und* $\mathfrak{K} := \mathfrak{G}$;

(4.6) $f : \hat{\mathfrak{K}} \to \mathfrak{G}$;

(4.7) α) $A : \mathfrak{G} \to \mathfrak{G}$ *sei algebraischer Isomorphismus mit der Eigenschaft*

$$\forall x \in \mathfrak{G} \qquad x \sim 0 \rightsquigarrow Ax \sim 0;$$

β) $R : \mathfrak{H} \to \mathfrak{H}$ *sei monoton;*

γ) $\forall x \in \mathfrak{G} \quad \|A^{-1}x\| \leqq R\|x\|$;

γ) *mit* $Q : \mathfrak{H} \to \mathfrak{H}$ *gelte*

$$\forall (x, y) \in \hat{\mathfrak{K}}^2 \qquad \|(f(x) - Ax) - (f(y) - Ay)\| \leqq Q\|x - y\|;$$

(4.8) α) *RQ sei folgenstetig und monoton;*

β) $\forall \alpha \in \mathfrak{H} \quad \sum_{n=1}^{\infty} (RQ)^n \alpha$ *konvergent;*

γ) *im Falle* (4.5), a) *gelte*

$$\|A^{-1} f(x_0)\| + \sum_{n=1}^{\infty} (RQ)^n \|A^{-1} f(x_0)\| \leqq \varepsilon.$$

Behauptungen.

(4.9) *Es gibt ein $\hat{x} \in \hat{\mathfrak{R}}$ mit $f(\hat{x}) \sim 0$.*

(4.10) *Es gilt $\hat{y} \in \hat{\mathfrak{R}}$ und $f(\hat{y}) \sim 0$ genau dann, wenn $\hat{y} \sim \hat{x}$.*

(4.11) *Setzt man für $x \in \hat{\mathfrak{R}}$*

$$Tx := x - A^{-1} f(x),$$

so ist für $n \in \mathbb{N}$ $T^n x_0 \in \mathfrak{R}$ definiert und man hat

$$T^n x_0 \to \hat{x}.$$

(4.12) *Gilt für ein $n \in \mathbb{N}$ und $\varepsilon_n \in \mathfrak{H}$*

$$\sum_{m=n}^{\infty} (RQ)^m \|A^{-1} f(x_0)\| \leqq \varepsilon_n$$

und ist

$$\mathfrak{R}_n := \{y : y \in \mathfrak{G} \wedge \|y - T^n x_0\| \leqq \varepsilon_n\},$$

so gilt $\hat{x} \in \hat{\mathfrak{R}}_n$.

Beweis. Wird T gemäß (4.11) definiert, so sind die Voraussetzungen für **2.**, Satz 1 erfüllt mit

$$\mathfrak{R} := \mathfrak{G}, \quad \mathfrak{M} := \mathfrak{H}, \quad P_n = (RQ)^n \quad (n \in \mathbb{N}).$$

Mit (4.1), (4.2), (4.3), (4.4) sind zunächst (2.1), (2.2), (2.3), (2.4) gegeben. (2.5) und (2.6) sind erfüllt aufgrund von (4.5), (4.6) und (4. 7), α). (2.7), α) erhält man zunächst für $n = 1$ über (4.7), α), β), γ), δ) und (4.3); man rechnet nämlich

$$\begin{aligned} \varrho(Ty, Tx) &= \|A^{-1}[(f(x) - Ax) - (f(y) - Ay)]\| \\ &\leqq R\|(f(x) - Ax) - (f(y) - Ay)\| \\ &\leqq RQ\|x - y\| = P_1 \varrho(y, x). \end{aligned}$$

Für $n = 2, 3, 4, \ldots$ erhält man dann (2.7), α) mit der Monotonie ((4.8), α)) von RQ und (4.3) durch Induktion. (2.7), β) ist gerade (4.8), β). (2.7), γ) ist mit (4.8), α) gegeben. Schließlich ist (2.7), δ) gerade die Forderung (4.8), γ) wegen

$$\sigma = \varrho(x_0, Tx_0) = \|A^{-1} f(x_0)\|.$$

Es bleibt die Umformung der Behauptungen. Dazu ist nur (3.8) zu beachten. Damit ist $T\hat{x} \sim \hat{x}$ äquivalent zu $A^{-1} f(\hat{x}) \sim 0$. Dies aber ist wegen (4.7), äquivalent zu $f(\hat{x}) \sim 0$. Damit ist Satz 2 bewiesen.

5. Quasimetrische Räume

Es werde angenommen:

(5.1) $\mathfrak{H}$ *ist eine nicht-leere Menge,*
$+$ *eine zweistellige Operation in* $\mathfrak{H}$,
$\leqq$ *eine zweistellige Relation in* $\mathfrak{H}$;

(5.2) $\mathbb{U}$ *ist ein Filter von Teilmengen von* $\mathfrak{H}$ *und* $\mathfrak{N} := \bigcap_{\mathfrak{U}\in\mathbb{U}} \mathfrak{U}$ *ist nicht leer;*

(5.3) $\forall\, \mathfrak{U}\in\mathbb{U}\ \exists\, \mathfrak{U}_1\in\mathbb{U} \qquad \mathfrak{U}_1+\mathfrak{U}_1\subset\mathfrak{U}$;

(5.4) $\forall\, \mathfrak{U}\in\mathbb{U}\ \exists\, \mathfrak{U}_0\in\mathbb{U} \qquad [\mathfrak{U}_0\subset\mathfrak{U}\wedge(\forall\beta\in\mathfrak{U}_0\,\forall\alpha\in\mathfrak{H} \quad \alpha\leqq\beta\rightsquigarrow\alpha\in\mathfrak{U}_0)]$.

Es heiße dann

$$(\mathfrak{H},\ +,\ \leqq,\ \mathbb{U})$$

eine *quasimetrische Struktur.*

Zunächst gilt

(5.5) *Eine quasimetrische Struktur wird zu einer quasimetrischen limes-Struktur durch die Definition:* $\mathbb{K}$ *sei die Menge der Folgen* $\{\xi_n\}\in\mathfrak{H}^{\mathbb{N}}$, *deren zugehöriger Frechet-Filter feiner ist als* $\mathbb{U}$. *Dabei wird das* $\mathfrak{N}$ *von* (5.2) *genau das* $\mathfrak{N}$ *von* (1.2).

Beweis. Zunächst folgt die letzte Behauptung und (1.2) sofort aus der Definition von $\mathbb{K}$. Damit ist auch (1.3) gegeben. Für den Nachweis von (1.4) sei $x=\{\xi_n\}$, $y=\{\eta_n\}$ und $\mathfrak{U}\in\mathbb{U}$. Man bestimmt hierzu $\mathfrak{U}_1$ gemäß (5.3). Hat man dann wegen $x\in\mathbb{K}$, $y\in\mathbb{K}$ für $n\geqq n_0$ $\xi_n\in\mathfrak{U}_1$, $\eta_n\in\mathfrak{U}_1$, so folgt wegen (5.3) $\xi_n+\eta_n\in\mathfrak{U}$ für $n\geqq n_0$. Da dies für jedes $\mathfrak{U}\in\mathbb{U}$ gilt, hat man $x+y\in\mathbb{K}$. Für den Nachweis von (1.5) sei $y\leqq x\in\mathbb{K}$, $x=\{\xi_n\}$, $y=\{\eta_n\}$ und $\mathfrak{U}\in\mathbb{U}$. Man bestimmt hierzu $\mathfrak{U}_0$ gemäß (5.4). Dann hat man wegen $x\in\mathbb{K}$ für $n\geqq n_0$ $\eta_n\leqq\xi_n\in\mathfrak{U}_0$, also wegen (5.4) $\eta_n\in\mathfrak{U}$ für $n\geqq n_0$. Dies gilt für jedes $\mathfrak{U}\in\mathbb{U}$; das besagt $y\in\mathbb{K}$.

Sei nun $\mathfrak{N}$ eine nicht-leere Menge, $(\mathfrak{H},\ +,\ \leqq,\ \mathbb{U})$ eine quasimetrische Struktur und ϱ eine Abbildung mit (1.6), (1.7), (1.8), (1.9). Dann heiße

$$(\mathfrak{N},\ (\mathfrak{H},\ +,\ \leqq,\ \mathbb{U}),\ \varrho)$$

quasimetrischer Raum.

Nach (5.5) ist ein solcher quasimetrischer Raum in kanonischer Weise quasimetrischer limes-Raum gemäß **1.**

Andererseits gilt

(5.6) *Ist* $(\mathfrak{N},\ (\mathfrak{H},\ +,\ \leqq,\ \mathbb{U}),\ \varrho)$ *quasimetrischer Raum, so stellt*

$$\mathbb{V}_0 := \{\mathfrak{V}\colon \mathfrak{V}\subset\mathfrak{N}\times\mathfrak{N},\ \exists\,\mathfrak{U}\in\mathbb{U}\ \ \mathfrak{V}=\varrho^{-1}(\mathfrak{U})\}$$

eine symmetrische Nachbarschaftsbasis einer uniformen Topologie (vgl. [5], §5) *für* $\mathfrak{N}$ *dar.*

Es gilt nämlich

(5.7) Jedes $\mathfrak{V}\in\mathbb{V}_0$ enthält die Diagonale von $\mathfrak{N}\times\mathfrak{N}$.

(5.8) $\mathbb{V}_0$ ist Filterbasis.

(5.9) $\forall\,\mathfrak{V}\in\mathbb{V}_0 \qquad \mathfrak{V}^{-1}=\mathfrak{V}$.

(5.10) $\forall\,\mathfrak{V}\in\mathbb{V}_0\,\exists\,\mathfrak{V}_1\in\mathbb{V}_0 \qquad \mathfrak{V}_1^2\subset\mathfrak{V}$.

Beweis. Zu (5.7): Sei $\mathfrak{V}_0 \in \mathfrak{V} = \varrho^{-1}(\mathfrak{U})$, $\mathfrak{U} \in \mathbb{U}$ und $a \in \mathfrak{R}$. Dann gilt $\varrho(a, a) \in \mathfrak{U}$, also $(a, a) \in \mathfrak{V}$.

Zu (5.8): Jedes $\mathfrak{V} \in \mathbb{V}_0$ ist nicht leer nach (5.7). Seien weiter $\mathbb{V}_0 \ni \mathfrak{V}_\nu = \varrho^{-1}(\mathfrak{U}_\nu)$, $\mathfrak{U}_\nu \in \mathbb{U}$ $(\nu = 1, 2)$. Dann gilt nach (5.2) $\mathfrak{U}_3 := \mathfrak{U}_1 \cap \mathfrak{U}_2 \in \mathbb{U}$. Damit aber hat man

$$\mathfrak{V}_3 := \mathfrak{V}_1 \cap \mathfrak{V}_2 = \varrho^{-1}(\mathfrak{U}_3) \in \mathbb{V}_0.$$

Zu (5.9): Ist $(x, y) \in \mathfrak{V} = \varrho^{-1}(\mathfrak{U})$, also $\varrho(x, y) \in \mathfrak{U}$, so hat man wegen (1.8) auch $\varrho(y, x) \in \mathfrak{U}$, also $(y, x) \in \mathfrak{V}$. Das heißt $\mathfrak{V}^{-1} = \mathfrak{V}$.

Zu (5.10). Sei $\mathbb{V}_0 \ni \mathfrak{V} = \varrho^{-1}(\mathfrak{U})$, $\mathfrak{U} \in \mathbb{U}$. Dann wähle man zunächst $\mathfrak{U}_0 \in \mathbb{U}$ gemäß (5.4). Zu $\mathfrak{U}_0$ bestimme man gemäß (5.3) ein $\mathfrak{U}_1 \in \mathbb{U}$ mit $\mathfrak{U}_1 + \mathfrak{U}_1 \subset \mathfrak{U}_0$. Dann leistet $\mathfrak{V}_1 = \varrho^{-1}(\mathfrak{U}_1)$ das Verlangte. Denn für $(a, b) \in \mathfrak{V}_1$ und $(b, c) \in \mathfrak{V}_1$ gilt $\varrho(a, b) \in \mathfrak{U}_1$ und $\varrho(b, c) \in \mathfrak{U}_1$, also gemäß (5.3) $\varrho(a, b) + \varrho(b, c) \in \mathfrak{U}_0$. Mit (1.9) und (5.4) hat man dann auch $\varrho(a, c) \in \mathfrak{U}_0 \subset \mathfrak{U}$. Das besagt $\mathfrak{V}_1^2 \subset \mathfrak{V}$.

Abschließend bemerken wir:

Faßt man einen quasimetrischen Raum einerseits mit (5.5) *als quasimetrischen limes-Raum und andererseits mit* (5.6) *als uniformen Raum auf, so gilt:*

(5.11) $x_n \to x$ *gemäß* (1.12) *bedeutet genau die topologische Konvergenz;*

(5.12) *eine Folge* $\{x_n\}$ *ist Cauchyfolge im Sinne von* **1.** *genau dann, wenn sie Cauchyfolge in der uniformen Topologie ist;*

(5.13) *für* $\mathfrak{M} \subset \mathfrak{R}$ *gilt*

$$\hat{\mathfrak{M}} \subset \overline{\mathfrak{M}}$$

wo $\overline{\mathfrak{M}}$ *die topologische Abschließung bedeutet.*

Die Beweise folgen unmittelbar aus den gegebenen Definitionen.

6. Quasinormierte Gruppen

Sei $(\mathfrak{H}, +, \leqq, \mathbb{U})$ quasimetrische Struktur im Sinne von **5.** und $(\mathfrak{G}, +)$ eine, additiv-geschriebene, nicht-notwendig abelsche Gruppe. Ist dann $\| \|$ eine Abbildung mit (3.1), (3.2), (3.3), so heißt

$$(\mathfrak{G}, +, (\mathfrak{H}, +, \leqq, \mathbb{U}), \| \|)$$

quasinormierte Gruppe.

(6.1) *Mit* (5.5) *ist eine quasinormierte Gruppe in kanonischer Weise eine quasinormierte limes-Gruppe.*

Andererseits gilt natürlich

(6.2) *Eine quasinormierte Gruppe ist mit*

$$\varrho(a, b) := \|a - b\|$$

ein quasimetrischer Raum, also mit (5.6) *ein uniformer Raum. Die Quasimetrik* ϱ *und damit die Nachbarschaftsbasis* $\mathbb{V}_0$ *sind zweiseitig invariant. Mit der erzeugten Topologie wird* $(\mathfrak{G}, +)$ *topologische Gruppe.*

Umgekehrt gilt wieder

(6.3) *Ist* $(\mathfrak{G}, +)$ *Gruppe,* $(\mathfrak{G}, (\mathfrak{H}, +, \leqq, \mathsf{U}), \varrho)$ *quasimetrischer Raum und die Quasimetrik* ϱ *zweiseitig invariant, so wird mit*

$$\|a\| := \varrho(a, 0) \quad (a \in \mathfrak{G})$$

$(\mathfrak{G}, +, (\mathfrak{H}, +, \leqq, \mathsf{U}), \| \|)$ *quasinormierte Gruppe, aus der man* ϱ *gemäß* (6.2) *zurückerhält.*

7. Schlußbemerkungen

Zu 1., 2., 3., 4.: Die Operation $+$ in $\mathfrak{H}$ ist nicht assoziativ vorausgesetzt. Das erweitert die Anwendbarkeit sowohl theoretisch, als auch praktisch. So ist eine direkte Anwendung auf Gleitkommazahlen in Computern möglich.

Die Forderungen an $\mathbb{K}$ sind sehr gering; insbesondere wird keine Teilfolgeneigenschaft verlangt. Bei Anwendungen kann etwa $\mathbb{K}$ mit Vorschriften über Konvergenzgeschwindigkeit oder mit Limitierungsverfahren oder geeignet für Prozesse in Computern gewählt werden.

Zu 5., 6.: Die Beachtung dieser Bemerkungen läßt bei vielen Anwendungen, auch bei den von Schröder [8] betrachteten, erkennen, daß uniforme Topologien vorliegen, daß also die etwas ungewohnten Definitionen von **1., 2., 3., 4.** durchaus vertraute Begriffe ergeben.

Eine genauere Untersuchung der Stellung der eingeführten quasimetrischen Räume im Zusammenhang mit zahlreichen in der Literatur diskutierten Verallgemeinerungen metrischer Räume bringt Schäfke [7]. Es zeigt sich, daß die quasimetrischen Räume hier die allgemeinsten metrikähnlichen Erzeugungsweisen uniformer Räume darstellen und sich umgekehrt jeder uniforme Raum auf einfache Weise so erhalten läßt. Ebenso wird eine einfache Charakterisierung der topologischen Gruppen gegeben, die sich als quasinormierte Gruppen darstellen lassen.

Wie mir Herr Zeidlewicz, Köln, mitteilt, ist es möglich, beim „Kontraktionssatz" (Satz 1) weitgehend auf die Symmetrie (1.8) der Quasimetrik zu verzichten. Man ersetze nämlich die „Dreiecksungleichung" (1.9) durch die (im Falle der Symmetrie äquivalente) Form:

$$\varrho(a, c) \leqq \varrho(b, c) + \varrho(b, a) \quad ((a, b, c) \in \mathfrak{R}^3). \tag{1.9'}$$

Aus (1.9') resultiert bei Beachtung von (1.3)—(1.5) und (1.7):

$$\{\varrho(x_n, y_n)\}_{n=1}^{\infty} \in \mathbb{K} \rightsquigarrow \{\varrho(y_n, x_n)\}_{n=1}^{\infty} \in \mathbb{K}$$

und hieraus durch spezielle Wahl der x_n und y_n $(n \in \mathbb{N})$

$$\varrho(a, b) \in \mathfrak{N} \rightsquigarrow \varrho(b, a) \in \mathfrak{N}.$$

Damit ist sichergestellt, daß die durch (1.10) bzw. (1.11) in $\mathfrak{R}$ bzw. $\mathfrak{R}^{\mathbb{N}}$ definierten Relationen auch im Falle der Asymmetrie der Quasimetrik Äquivalenzrelationen sind. Ebenso lassen sich die Bemerkungen (1.15), (1.16) und (1.17) in ungeänderter Form aussprechen. Bei der Formulierung des Kontraktionssatzes (Satz 1) beachte

man dann nur

$$\mathfrak{R} := \{y: y \in \mathfrak{R} \text{ und } \varrho(y, x_0) \leqq \varepsilon\} \quad \text{(s. (2.5), a))},$$
$$\sigma := \varrho(Tx_0, x_0) \quad \text{(s. (2.7), } \delta)),$$
$$\mathfrak{R}_n := \{y: y \in \mathfrak{R} \text{ und } \varrho(y, T^n x_0) \leqq \varepsilon_n\} \quad \text{(s. (2.12))}$$

und ändere die Voraussetzung (2.7), α) zu

$$\varrho(T^n x, T^n y) \leqq \begin{cases} P_n \varrho(x, y) \\ P_n \varrho(y, x) \end{cases} \quad ((x, y) \in (\mathfrak{D}_{T^n})^2,\ n \in \mathbb{N})$$

ab.

Literatur

1. Albrecht, R., Karrer, G.: Fixpunktsätze in uniformen Räumen. Math. Zeitschr. **74**, 387–391 (1960).
2. Collatz, L.: Funktionalanalysis und numerische Mathematik. Berlin-Göttingen-Heidelberg (1964).
3. Ehrmann, H.: Iterationsverfahren mit veränderlichen Operatoren. Arch. rat. Mech. Anal. **4**, 45–64 (1959).
4. Kantorovitch, L.: The method of successive approximations for functional equations. Acta math. **71**, 63–97 (1939).
5. Köthe, G.: Topologische lineare Räume. I. Berlin-Göttingen-Heidelberg (1960).
6. Reinermann, J.: Über Fixpunkte kontrahierender Abbildungen und deren Darstellung durch konvergente Iterationsverfahren. Ber. d. Ges. f. Mathem. u. Datenverarb. Bonn, Nr. 4 (1968).
7. Schäfke, F. W.: Quasimetrische Räume und quasinormierte Gruppen. Arch. d. Math. (Erscheint demnächst.)
8. Schröder, J.: Das Iterationsverfahren bei allgemeinerem Abstandsbegriff. Math. Z. **66**, 111–116 (1956).
9. — Nichtlineare Majoranten beim Verfahren der schrittweisen Näherung. Arch. d. Math. **7**, 471–484 (1956).
10. Weil, A.: Sur les espaces a structur uniforme et sur la topologie generale. Act. Sci. et Ind. Nr. 551 (1938).
11. Weissinger, J.: Zur Theorie und Anwendung des Iterationsverfahrens. Math. Nachr. **8**, 192–212 (1952).

Prof. Dr. F. W. Schäfke
Mathem. Institut der Universität Köln
D-5000 Köln 41
Weyertal 86–90

Numer. Math. 15, 122—136 (1970)

Definitheitsbedingungen für relative Extrema bei Optimierungs- und Approximationsaufgaben

W. WETTERLING

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 6. Oktober 1969

Summary. In this paper, sufficient and necessary maximum conditions are established for a class of mathematical programming problems with an infinite set of restrictions, which is described by a finite number of inequalities. The criteria may be applied to nonlinear approximation problems and to the numerical solution of boundary value problems.

§ 1. Einleitung

In diesem Beitrag werden hinreichende und notwendige Bedingungen für relative Extrema bei einem Typ zweistufiger Optimierungsaufgaben (§3) hergeleitet; das sind Aufgaben mit unendlich vielen Ungleichungen als Restriktionen, wobei die Menge der Restriktionen durch ein endliches System von Ungleichungen beschrieben wird. Die Betrachtung dieses Aufgabentyps ist durch die Anwendungsmöglichkeiten nahegelegt: Nichtlineare Probleme der Tschebyscheff-Approximation, insbesondere solche in mehreren reellen Variablen, werden hierdurch erfaßt (§8), ferner Optimierungsprobleme, die bei der numerischen Behandlung von Randwertaufgaben nach dem Prinzip der lokal optimalen Schranken [6] auftreten. In diesen Fällen entspricht die Restriktionenmenge dem Bereich, in dem die Approximations- bzw. Randwertaufgabe formuliert ist. So ist diese Arbeit als Weiterführung des von und mit Collatz in [1] und [2], §15 und weiterhin in [6—8] verfolgten Programms der Anwendung von Approximations- und Optimierungsmethoden bei Randwertaufgaben anzusehen.

Bei den hier betrachteten Aufgaben werden über die Zielfunktion und die Restriktionsfunktionen keine generellen Konvexitätsvoraussetzungen gemacht; es ergeben sich schwache Definitheitsbedingungen für die Matrix der zweiten Ableitungen einer erweiterten Lagrange-Funktion.

Hinreichend für ein Extremum ist die Definitheit, notwendig die Semidefinitheit der quadratischen Form zu dieser Matrix im Tangentialraum an der betrachteten Stelle. Während dieser im Fall von endlich vielen Restriktionen (§2) durch ein homogenes lineares Gleichungssystem beschrieben wird, in dem die ersten Ableitungen der Restriktionsfunktionen auftreten, hat man im Fall der zweistufigen Aufgaben (§3ff.) auch lineare Relationen, die die zweiten Ableitungen enthalten; diese sind als linearisierte Enveloppenbedingungen zu deuten.

Die wichtige Frage nach numerischen Methoden zur Behandlung solcher Aufgaben wird hier nicht erörtert. Offenbar erfordern die verschiedenen speziellen

Probleme, die sich der hier betrachteten allgemeinen Problemklasse unterordnen, verschiedenartige Algorithmen. Einige Hinweise findet man in [6] und [8]. Die in diesem Beitrag angegebenen Bedingungen bieten die Möglichkeit, das Ergebnis eines solchen Algorithmus, das im allgemeinen nur ein stationärer Wert ist, auf die Extremaleigenschaft zu prüfen.

§ 2. Finite Maximumaufgaben

Zunächst wird eine Aufgabe mit endlich vielen Restriktionen betrachtet; die hierüber bekannten Ergebnisse werden später benötigt:

Gesucht sind relative Maxima einer Funktion $F(x)$ auf einer Menge $M \subset R^n$, die durch

$$M = \{x;\ f_j(x) \leqq 0\ \ (j = 1, \ldots, m)\}$$

beschrieben ist. Dabei sind $F(x)$, $f_1(x), \ldots, f_m(x)$ reellwertig definiert und zweimal stetig differenzierbar für alle $x \in R^n$. Mit $F'(x)$ wird der Spaltenvektor grad $F(x)$ mit den Komponenten $\partial F/\partial x_i$ $(i = 1, \ldots, n)$ bezeichnet, mit $F''(x)$ die Matrix der zweiten partiellen Ableitungen von F. Entsprechend sind $f_j'(x)$ und $f_j''(x)$ zu verstehen.

Sei nun ein $\tilde{x} \in M$ gegeben; man möchte prüfen, ob F an der Stelle $\tilde{x}$ ein relatives Maximum bezüglich M hat. Hinreichend dafür sind folgende Bedingungen:

(a) *Lokale Kuhn-Tucker-Bedingung. Für $j \in J \subset \{1, \ldots, m\}$, aber nicht notwendig nur für diese Indizes, ist $f_j(\tilde{x}) = 0$, und es gibt reelle Zahlen $u_j > 0$ $(j \in J)$ mit*

$$F'(\tilde{x}) - \sum_{j \in J} u_j f_j'(\tilde{x}) = 0. \tag{1}$$

(b) *Sei $H \subset R^n$ der lineare Teilraum der Vektoren $\xi \in R^n$ mit $\xi^T f_j'(\tilde{x}) = 0$ $(j \in J)$* (ξ^T ist der zum Spaltenvektor ξ transponierte Zeilenvektor); *die quadratische Form*

$$Q(\xi) = \xi^T \left[F''(\tilde{x}) - \sum_{j \in J} u_j f_j''(\tilde{x})\right] \xi$$

ist negativ definit auf H.

Sind die Bedingungen (a) und (b) erfüllt, so gibt es eine Umgebung von $\tilde{x}$, in deren Durchschnitt mit M ($\tilde{x}$ selbst ausgenommen) $F(x) < F(\tilde{x})$ ist.

Die entsprechenden notwendigen Bedingungen lauten: Sei $\tilde{x} \in M$ und J nun die Menge aller Indizes j mit $f_j(\tilde{x}) = 0$. Folgende *constraint qualifications* an der Stelle $\tilde{x}$ seien erfüllt:

Ist $\xi \in R^n$ ein Vektor mit $\xi^T f_j'(\tilde{x}) \leqq 0$ $(j \in J)$, so gibt es ein $t_0 > 0$ und eine für $0 \leqq t \leqq t_0$ stetig differenzierbare vektorwertige Funktion $x(t)$ mit $x(0) = \tilde{x}$, $dx(0)/dt = \xi$ und $x(t) \in M$ $(0 \leqq t \leqq t_0)$. Ist $\xi^T f_j'(\tilde{x}) = 0$ $(j \in J)$, so gibt es ein solches $x(t)$, das überdies zweimal stetig differenzierbar ist und für das $f_j(x(t)) = 0$ $(j \in J,\ 0 \leqq t \leqq t_0)$ ist.

Ist dann $F(x) \leqq F(\tilde{x})$ im Durchschnitt einer Umgebung von $\tilde{x}$ mit M, so gilt:

(a*) *Es gibt $u_j \geqq 0$ $(j \in J)$ mit* (1).

(b*) *$Q(\xi)$ ist negativ semidefinit auf H.*

Man beachte jedoch, daß J und damit H und Q jetzt anders definiert sind als im Fall der hinreichenden Bedingungen.

Diese Aussagen sind in den Ergebnissen von McCormick [4] enthalten, der neben Ungleichungen auch Gleichungen als Restriktionen zuläßt.

Dort wird außerdem gezeigt, daß die *constraint qualifications* sicher dann erfüllt sind, wenn die Vektoren $f_j'(\tilde{x})$ $(j \in J)$ linear unabhängig sind. Für das von McCormick mit indirektem Beweis hergeleitete, hinreichende Kriterium ist in [8] unter der Voraussetzung der Stetigkeit der dritten Ableitungen der auftretenden Funktionen ein direkter Beweis angegeben, bei dem eine Kugelumgebung von $\tilde{x}$ mit der genannten Eigenschaft konstruiert wird.

§ 3. Zweistufige Maximumaufgaben, hinreichende Bedingungen

Eine ähnliche Bedingung soll nun für folgenden Aufgabentyp, dessen Struktur die Bezeichnung *zweistufig* nahelegt, formuliert werden: *Gesucht sind relative Maxima von* $F(x)$ *auf*

$$M = \{x;\ f(x, y) \leqq 0 \ \ (y \in Y)\} \subset R^n;$$

dabei ist

$$Y = \{y;\ g_\nu(y) \leqq 0 \ \ (\nu = 1, \ldots, N)\} \subset R^m.$$

Die Funktionen F, f, g_ν seien reellwertig definiert und zweimal stetig differenzierbar im R^n, $R^n \times R^m$ bzw. R^m. Wie oben werden mit $F'(x)$ und $f'(x, y)$ die Vektoren der ersten Ableitungen nach den x_i bezeichnet, ferner mit $f^{\cdot}(x, y)$ und $g_\nu^{\cdot}(y)$ die Vektoren der ersten Ableitungen nach den y_k, weiterhin mit F'', f'' $f^{\cdot\cdot}$, $f^{\cdot\cdot}$ und $g_\nu^{\cdot\cdot}$ die entsprechenden Matrizen der zweiten Ableitungen.

Wieder wird ein Punkt $\tilde{x} \in M$ betrachtet. Als hinreichend dafür, daß dort ein relatives Maximum von F bezüglich M vorliegt, erweisen sich die folgenden Bedingungen:

(A) *Lokale Kuhn-Tucker-Bedingung. Für* $y_1, y_2, \ldots, y_p \in Y$, *aber nicht notwendig nur für diese Punkte, sei* $f(\tilde{x}, y_j) = 0$, *und es gebe reelle Zahlen* $u_j > 0$ $(j = 1, \ldots, p)$ *mit*

$$F' - \sum_{j=1}^{p} u_j f_j' = 0 \tag{2}$$

(das Argument $\tilde{x}$ wird hier und im folgenden häufig weggelassen, das Argument y_j ebenfalls und durch einen Index j angedeutet).

Als Funktion von y auf der Menge Y hat $f(\tilde{x}, y)$ bei $y = y_j$ Maxima mit dem Funktionswert $f(\tilde{x}, y_j) = 0$. Diese Maxima sollen so beschaffen sein, daß ähnliche Definitheitsbedingungen wie in §2 gelten.

(B) *Für jedes* j $(1 \leqq j \leqq p)$ *gelte: Für* $\nu \in M_j \subset \{1, \ldots, N\}$ *und nur für diese Indizes ist* $g_\nu(y_j) = 0$, *die Vektoren* $g_{\nu j}^{\cdot} = g_\nu^{\cdot}(y_j)$ $(\nu \in M_j)$ *sind linear unabhängig, und es gibt reelle Zahlen* $v_{\nu j} \geqq 0$ $(\nu \in M_j)$ *mit*

$$f_j^{\cdot} - \sum_{\nu \in M_j} v_{\nu j}\, g_{\nu j}^{\cdot} = 0.$$

Auf dem linearen Teilraum

$$H_j = \{\eta;\ \eta^T g_{\nu j}^{\cdot} = 0 \ \ (\nu \in M_j)\} \subset R^m$$

ist die quadratische Form

$$Q_j(\eta) = \eta^T \left(f_j'' - \sum_{\nu \in M_j} v_{\nu j}\, g_{\nu j}''\right) \eta$$

negativ definit.

Bevor die eigentliche Definitheitsbedingung (C) formuliert werden kann, sind einige Zwischenüberlegungen nötig. Ist m_j die Anzahl der Indizes in M_j, so hat H_j die Dimension $m - m_j$.

Seien nun $\eta_j^{(1)}, \ldots, \eta_j^{(m-m_j)}$ Vektoren, die eine Basis von H_j bilden. Hiermit wird das lineare Gleichungssystem

$$\eta_j^{(k)T} \left[f_j' \xi + \left(f_j'' - \sum_{\nu \in M_j} v_{\nu j}\, g_{\nu j}''\right) \eta\right] = 0 \qquad (k = 1, \ldots, m - m_j) \tag{3}$$

aufgestellt. Es zeigt sich, daß es bei gegebenem $\xi \in R^n$ genau ein $\eta \in H_j$ als Lösung dieses Systems gibt. Der Ansatz

$$\eta = \sum_{i=1}^{m-m_j} c_i\, \eta_j^{(i)}$$

führt nämlich auf ein lineares Gleichungssystem für die Unbekannten c_i mit der Koeffizientenmatrix

$$\left(\eta_j^{(k)T} \left[f_j'' - \sum_{\nu \in M_j} v_{\nu j}\, g_{\nu j}''\right] \eta_j^{(i)}\right)_{k,\, i = 1, \ldots, m - m_j},$$

die wegen der Definitheit von $(f_j'' - \sum v_{\nu j}\, g_{\nu j}'')$ nichtsingulär ist. So sind also einem $\xi \in R^n$ als Lösungen von (3) für $j = 1, \ldots, p$ Vektoren $\eta_1 \in H_1, \ldots, \eta_p \in H_p$ eindeutig zugeordnet, und zwar linear:

$$\eta_j = P_j \xi \qquad (j = 1, \ldots, p)$$

mit von ξ unabhängigen $m \times n$-Matrizen P_j.

Die Definitheitsbedingung lautet dann

(C) *Die quadratische Form*

$$Q(\xi) = \xi^T \left[F'' - \sum_{j=1}^{p} u_j \left(f_j'' + 2 P_j^T f_j' + P_j^T \left(f_j'' - \sum_{\nu \in M_j} v_{\nu j}\, g_{\nu j}''\right) P_j\right)\right] \xi$$

sei auf

$$H = \{\xi;\, \xi^T f_j' = 0 \;\; (j = 1, \ldots, p)\}$$

negativ definit.

Diese Bedingung kann auch so formuliert werden: Es sei

$$q(\xi, \eta_j) = \xi^T (F'' - \sum u_j f_j'')\, \xi - 2 \sum u_j \eta_j^T f_j' \xi - \sum u_j \eta_j^T (f_j'' - \sum v_{\nu j}\, g_{\nu j}'')\, \eta_j < 0 \tag{4}$$

für $\xi \in H$, $\eta_j \in H_j$, sofern ξ und die η_j durch (3) verknüpft sind und solange nicht $\xi = 0$ und daher $\eta_1 = \cdots = \eta_p = 0$ ist. Wegen (3) ist hier

$$\eta_j^T f_j'' \xi + \eta_j^T (f_j'' - \sum v_{\nu j}\, g_{\nu j}'')\, \eta_j = 0$$

und daher

$$q(\xi, \eta) = \xi^T (F'' - \sum u_j f_j'')\, \xi + \sum u_j \eta_j^T (f_j'' - \sum v_{\nu j}\, g_{\nu j}'')\, \eta_j.$$

Die negative Definitheit von $\xi^T (F'' - \sum u_j f_j'')\, \xi$ (vgl. §2) ist also zusammen mit (B) hinreichend für (C), ist aber ersichtlich einschränkender als (C).

§ 4. Beweis der Maximaleigenschaft

Satz. *Sind für ein $\tilde{x}\in M$ die Bedingungen* (A), (B), (C) *erfüllt, so gibt es eine Kugel um $\tilde{x}$, in deren Durchschnitt mit M ($\tilde{x}$ selbst ausgenommen) $F(x)<F(\tilde{x})$ ist.*

Beweis (indirekt). Aus der folgenden Annahme ist ein Widerspruch herzuleiten: Es gibt eine Folge $x_1, x_2, \ldots$ von Punkten $x_k\in M$ mit $x_k \neq \tilde{x}$, $\lim\limits_{k\to\infty} x_k = \tilde{x}$ und

$$F(x_k)\geqq F(\tilde{x}). \tag{5}$$

Sei $x_k-\tilde{x}=\delta_k\xi_k$ mit $\delta_k>0$, $\|\xi_k\|=1$ mit irgendeiner Vektornorm und daher $\lim \delta_k=0$. Die Folge der ξ_k enthält eine konvergente Teilfolge. Ohne Beschränkung der Allgemeinheit sei also $\lim \xi_k=\xi$ mit $\|\xi\|=1$. Es ist

$$\frac{1}{\delta_k}\left[F(\tilde{x}+\delta_k\xi_k)-F(\tilde{x})\right]\geqq 0$$

und daher auch der Grenzwert $\xi^T F'(\tilde{x})\geqq 0$. Wegen $x_k\in M$ und $f(\tilde{x}, y_j)=0$ ist ferner

$$\frac{1}{\delta_k}\left[f(\tilde{x}+\delta_k\xi_k, y_j)-f(\tilde{x}, y_j)\right]\leqq 0$$

und daher $\xi^T f_j'\leqq 0$. Mit (2) wird

$$0\leqq \xi^T F'=\sum_{j=1}^{p} u_j \xi^T f_j'\leqq 0.$$

Wegen $u_j>0$ ist $\xi^T F'=\xi^T f_1'=\cdots=\xi^T f_p'=0$, also $\xi\in H$. Mit diesem ξ sei $\eta_j=P_j\xi$ $(j=1,\ldots,p)$.

Für jeden Index j $(1\leqq j\leqq p)$ gilt nun folgende Überlegung: Es ist $g_\nu(y_j)=0$ und $\eta_j^T \dot{g}_\nu(y_j)=0$ $(\nu\in M_j)$. In einem Intervall $0\leqq\delta\leqq\Delta$ mit hinreichend kleinem positiven Δ gibt es einen Vektor $\psi_j=\psi_j(\delta)\in R^m$ mit $\psi_j^T\eta=0$ $(\eta\in H_j)$ und

$$g_\nu\left(y_j+\delta\eta_j+\psi_j(\delta)\right)=0 \quad (\nu\in M_j), \tag{6}$$

der als Funktion von δ zweimal stetig differenzierbar und mit $\psi_j(0)=0$ eindeutig bestimmt ist. Um dies einzusehen, mache man den Ansatz

$$\psi_j(\delta)=\sum_{\nu\in M_j} b_\nu(\delta)\,\dot{g}_{\nu j}$$

und wende bekannte Sätze über implizite Funktionen an (s. etwa [3], Kap. X). Die hier auftretende Jacobi-Matrix $(\dot{g}_{\nu j}^T \dot{g}_{\mu j})_{\nu,\mu\in M_j}$ ist wegen der linearen Unabhängigkeit der Vektoren $\dot{g}_{\nu j}$ nichtsingulär. Aus (6) folgt weiterhin

$$\dot{g}_{\nu j}^T\left(\eta_j+\frac{d\psi_j(0)}{d\delta}\right)=0 \quad (\nu\in M_j);$$

wegen $\dot{g}_{\nu j}^T\eta_j=0$ und der linearen Unabhängigkeit der $\dot{g}_{\nu j}$ ist $d\psi_j(0)/d\delta=0$ und daher $\psi_j(\delta)=\delta^2\chi_j(\delta)$ mit gleichmäßig beschränktem $\chi_j(\delta)$ für $0\leqq\delta\leqq\Delta$.

Sei $\Delta>0$ weiterhin so klein gewählt, daß $g_\nu\left(y_j+\delta\eta_j+\delta^2\chi_j(\delta)\right)<0$ ist für $\nu\notin M_j$ und $0\leqq\delta\leqq\Delta$, also $y_j+\delta\eta_j+\delta^2\chi_j(\delta)\in Y$.

Die Zahlen δ_k der obigen Folge liegen für hinreichend großes k im Intervall $[0, \Delta]$. Für diese k wird mit $\chi_{jk} = \chi_j(\delta_k)$

$$
\begin{aligned}
F(\tilde{x} + \delta_k \xi_k) - \sum_{j=1}^{p} u_j \Big\{ & \underbrace{f(\tilde{x} + \delta_k \xi_k,\, y_j + \delta_k \eta_j + \delta_k^2 \chi_{jk})}_{\leqq 0} \\
& - \sum_{\nu \in M_j} v_{\nu j}\, \underbrace{g_\nu(y_j + \delta_k \eta_j + \delta_k^2 \chi_{jk})}_{=0} \Big\} \\
= F(\tilde{x}) - \sum_{j=1}^{p} u_j \Big\{ & \underbrace{f(\tilde{x},\, y_j)}_{=0} - \sum_{\nu \in M_j} \underbrace{v_{\nu j}\, g_\nu(y_j)}_{=0} \Big\} \\
& + \delta_k \Big[\xi_k^T \underbrace{(F' - \textstyle\sum u_j f_j')}_{=0} - \sum u_j (\eta_j + \delta_k \chi_{jk})^T \underbrace{(f_j' - \textstyle\sum v_{\nu j} g_{\nu j}')}_{=0} \Big] \\
& + \tfrac{1}{2} \delta_k^2 q(\xi_k, \eta_j) + o(\delta_k^2).
\end{aligned}
$$

Für $k \to \infty$ hat man hiermit einen Widerspruch zwischen (4) und (5).

Bemerkung. Aus dem Beweis ist zu erkennen, daß man auch $\eta_j = \alpha_j P_j \xi$ mit beliebigen reellen α_j hätte wählen können, und zwar bereits in der Formulierung der Bedingung (C). Man erkennt aber, daß $\alpha_j = 1$ als Maximalstelle von $\alpha_j(2 - \alpha_j)$ ausgezeichnet ist und für $\alpha_j \neq 1$ zwar auch hinreichende, aber einschränkendere Bedingungen als für $\alpha_j = 1$ herauskommen. Auch aus den folgenden notwendigen Bedingungen wird das deutlich.

§ 5. Notwendige Maximalbedingungen

Bei $\tilde{x}$ habe $F(x)$ bezüglich $M = \{x;\, f(x, y) \leqq 0 \ (y \in Y)\}$ ein schwaches relatives Maximum, d.h. im Durchschnitt einer Umgebung von $\tilde{x}$ mit M sei $F(x) \leqq F(\tilde{x})$. Sei dabei $f(\tilde{x}, y) = 0$ genau für $y = y_1, \ldots, y_p \in Y$ (es sind also nur endlich viele Nullstellen von $f(\tilde{x}, y)$ in Y zugelassen). Außer dieser Einschränkung werden noch einige weitere Bedingungen (*constraint qualifications*) gefordert, zunächst:

(I) *Ist $\xi \in R^n$ ein Vektor mit $\xi^T f'(\tilde{x}, y_j) \leqq 0$ $(j = 1, \ldots, p)$, so gibt es ein $t_0 > 0$ und eine für $0 \leqq t \leqq t_0$ stetig differenzierbare vektorwertige Funktion $x(t)$ mit $x(0) = \tilde{x}$, $dx(0)/dt = \xi$ und $x(t) \in M$ $(0 \leqq t \leqq t_0)$.*

Für ein beliebiges solches ξ ist dann

$$
\xi^T F'(\tilde{x}) = \frac{d}{dt} F\big(x(t)\big)_{(t=0)} \leqq 0,
$$

es gibt also kein $\xi \in R^n$ mit $\xi^T f'(\tilde{x}, y_j) \leqq 0$ $(j = 1, \ldots, p)$ und $\xi^T F'(\tilde{x}) > 0$.

Nach dem Alternativsatz für Systeme von linearen Ungleichungen (Lemma von Farkas, [2], § 5) gibt es $u_j \geqq 0$ $(j = 1, \ldots, p)$ mit

$$
F'(\tilde{x}) - \sum_{j=1}^{p} u_j f'(\tilde{x}, y_j) = 0. \tag{7}
$$

Weiterhin sei folgende Bedingung erfüllt:

(II) *Ist* $\xi \in R^n$ *ein Vektor mit* $\xi^T f'(\tilde{x}, y_j) = 0$ $(j = 1, \ldots, p)$, *so gibt es für* $0 \leqq t \leqq t_1$ $(t_1 > 0)$ *ein zweimal stetig differenzierbares* $x(t) \in M$ *mit* $x(0) = \tilde{x}$, $dx(0)/dt = \xi$ *und zugeordnete, ebenfalls zweimal stetig differenzierbare* $z_j(t) \in Y$ *mit* $z_j(0) = y_j$ *und*

$$f\big(x(t), z_j(t)\big) = 0 \quad (0 \leqq t \leqq t_1;\ j = 1, \ldots, p). \tag{8}$$

Seien schließlich die Maximalstellen y_j von $f(\tilde{x}, y)$ wie folgt beschaffen:

(III) *Für* $j = 1, \ldots, p$ *gilt: Es ist* $g_\nu(y_j) = 0$ *genau für* $\nu \in M_j \subset \{1, \ldots, N\}$, *die Vektoren* $g'_{\nu j}$ $(\nu \in M_j)$ *sind linear unabhängig, es ist*

$$f'_j - \sum_{\nu \in M_j} v_{\nu j}\, g'_{\nu j} = 0 \tag{9}$$

mit $v_{\nu j} > 0$ $(\nu \in M_j)$, *und die Matrix* $f''_j - \sum_{\nu \in M_j} v_{\nu j}\, g''_{\nu j}$ *ist negativ definit auf* $H_j = \{\eta;\ \eta^T g'_{\nu j} = 0\ (\nu \in M_j)\}$.

Sei nun $\xi \in H = \{\xi;\ \xi^T f'_j = 0\ (j = 1, \ldots, p)\}$ und hierzu $x(t)$ und $z_1(t), \ldots, z_p(t)$ gemäß (II) gegeben. Setzt man $dz_j(0)/dt = \eta_j$, so wird nach (8) $\xi^T f'_j + \eta_j^T f'_j = 0$, also $\eta_j^T f'_j = 0$ $(j = 1, \ldots, p)$. Wegen $z_j(t) \in Y$ ist $\eta_j^T g'_{\nu j} \leqq 0$ und wegen $v_{\nu j} > 0$ und (9) sogar

$$\eta_j^T g'_{\nu j} = 0 \quad (\nu \in M_j;\ j = 1, \ldots, p). \tag{10}$$

Für festes $t \in [0, t_1]$ sind die Punkte $y = z_j(t)$ Maximalstellen von $f\big(x(t), y\big)$ bezüglich Y. Es zeigt sich, daß wenigstens in einem Teilintervall von $[0, t_1]$ die notwendigen Bedingungen von McCormick (§2) gelten. Wegen der Stetigkeit der g_ν ist nämlich für $0 \leqq t \leqq t_2$ mit $0 < t_2 \leqq t_1$ $g_\nu\big(z_j(t)\big) < 0$ $(\nu \notin M_j)$. Sei $M_j(t) \subset M_j$ die Menge der Indizes ν, für die $g_\nu\big(z_j(t)\big) = 0$ ist $(0 \leqq t \leqq t_2)$. Auch die Vektoren $g'_\nu\big(z_j(t)\big)$ hängen stetig von t ab, und darum gibt es ein t_3 mit $0 < t_3 \leqq t_2$, so daß die $g'_\nu\big(z_j(t)\big)$ $\big(\nu \in M_j(t)\big)$ linear unabhängig sind, sofern $t \in [0, t_3]$ ist. Nach §2 gibt es Zahlen $w_{\nu j}(t) \geqq 0$ $\big(\nu \in M_j(t)\big)$ mit

$$f'\big(x(t), z_j(t)\big) - \sum_{\nu \in M_j(t)} w_{\nu j}(t)\, g'_\nu\big(z_j(t)\big) = 0. \tag{11}$$

Betrachtet man (9) und (11), so erkennt man: Für $t = 0$ liegt der Vektor $f'\big(x(t), z_j(t)\big)$ im relativen Inneren des von den Vektoren $g'_\nu\big(z_j(t)\big)$ aufgespannten Kegels. Alle diese Vektoren hängen stetig von t ab. Darum gibt es ein t_4 mit $0 < t_4 \leqq t_3$, so daß für $t \in [0, t_4]$ die Gl. (11) nur dann gelten kann, wenn $M_j(t) = M_j(0) = M_j$ und $w_{\nu j}(t) > 0$ ist. Es ist also auch $g_\nu\big(z_j(t)\big) = 0$ in $[0, t_4]$ für alle $\nu \in M_j$. Ebenso wie die Vektoren f' und g'_ν im Gleichungssystem (11) hängt auch dessen Lösung $w_{\nu j}(t)$ mit dem Anfangswert $w_{\nu j}(0) = v_{\nu j}$ im Intervall $[0, t_4]$ stetig differenzierbar von t ab. Aus (11) folgt dann

$$f'_j{}'\,\xi + \Big(f''_j - \sum_{\nu \in M_j} v_{\nu j}\, g''_{\nu j}\Big)\eta_j - \sum_{\nu \in M_j} \frac{dw_{\nu j}(0)}{dt}\, g'_{\nu j} = 0.$$

Ist $\{\eta_j^{(1)}, \ldots, \eta_j^{(m-m_j)}\}$ eine Basis von $H_j = \{\eta;\ \eta^T g'_{\nu j} = 0\ (\nu \in M_j)\}$, so wird

$$\eta_j^{(k)}\Big[f'_j{}'\,\xi + \Big(f''_j - \sum_{\nu \in M_j} v_{\nu j}\, g''_{\nu j}\Big)\eta_j\Big] = 0 \quad (k = 1, \ldots, m - m_j). \tag{12}$$

Als zugeordnete $z_j(t)$ in (II) kommen also nur solche in Frage, für die mit $\eta_j = dz_j(0)/dt$ das lineare Gleichungssystem (12) erfüllt ist. Weil $x(t) \in M$ für $0 \leqq t \leqq t_4$, F bei $\tilde{x} = x(0)$ maximal und

$$\frac{d}{dt} F\big(x(t)\big)_{(t=0)} = \xi^T F' = \sum_{j=1}^{p} u_j \xi^T f_j' = 0$$

ist, wird

$$\frac{d^2}{dt^2} F\big(x(t)\big)_{(t=0)} = \alpha^T F' + \xi^T F'' \xi \leqq 0,$$

wobei $\alpha = d^2 x(0)/dt^2$ ist. Weiterhin ist mit $\beta_j = d^2 z_j(0)/dt^2$

$$\frac{d^2}{dt^2} f\big(x(t), z_j(t)\big)_{(t=0)} = \alpha^T f_j' + \beta_j^T f_j' + \xi^T f_j'' \xi + 2\xi^T f_j'' \eta_j + \eta_j^T f_j'' \eta_j = 0 \quad (j = 1, \ldots, p)$$

und

$$\frac{d^2}{dt^2} g_\nu\big(z_j(t)\big)_{(t=0)} = \beta_j^T g_{\nu j}' + \eta_j^T g_{\nu j}'' \eta_j = 0 \quad (\nu \in M_j;\ j = 1, \ldots, p).$$

Mit (7) und (9) wird

$$\begin{aligned} \frac{d^2}{dt^2} F\big(x(t)\big)_{(t=0)} = \xi^T \Big(F'' - \sum_{j=1}^{p} u_j f_j''\Big) \xi - 2 \sum_{j=1}^{p} u_j \xi^T f_j'' \eta_j \\ - \sum_{j=1}^{p} u_j \eta_j^T \Big(f_j'' - \sum_{\nu \in M_j} v_{\nu j} g_{\nu j}''\Big) \eta_j \leqq 0 \end{aligned} \tag{13}$$

und zwar dies für alle $\xi \in H$ und die ihnen durch (12) eindeutig zugeordneten $\eta_j \in H_j$.

Satz. *Sei $F(x) \leqq F(\tilde{x})$ im Durchschnitt einer Umgebung von $\tilde{x}$ mit M, und sei $f(\tilde{x}, y) = 0$ genau für $y = y_1, \ldots, y_p \in Y$. Ist dann die Bedingung* (I) *erfüllt, so gilt die lokale Kuhn-Tucker-Bedingung* (7). *Sind außerdem die Bedingungen* (II) *und* (III) *erfüllt, so gilt die Semidefinitheitsbedingung* (13).

§ 6. Die constraint qualifications

Die Bedingungen (I) und (II) sind nicht so einschränkend, wie es zunächst scheint. Sie besagen, daß man an einer Maximalstelle nur dann die angegebenen notwendigen Bedingungen folgern kann, wenn diese Stelle nicht ein Randpunkt von M von dem Typ der aus der Theorie der linearen Optimierung bekannten entarteten Ecken ist. Hier gilt ein ähnlicher Satz wie im Fall der finiten Maximumaufgaben (§ 2).

Satz. *Sei die Menge Y beschränkt, seien ferner $f(x, y)$, $g_1(y), \ldots, g_N(y)$ dreimal stetig differenzierbar und $y_1, \ldots, y_p$ alle Punkte in Y mit $f(\tilde{x}, y) = 0$. Ist dann die Bedingung* (III) *erfüllt und sind die Vektoren $f'(\tilde{x}, y_j)$ $(j = 1, \ldots, p)$ linear unabhängig, so sind auch* (I) *und* (II) *erfüllt.*

Beweis. Sei $\xi \in R^n$ ein Vektor mit $\xi^T f'(\tilde{x}, y_j) \leqq 0$ $(j = 1, \ldots, p)$, und zwar sei

$$\xi^T f'(\tilde{x}, y_j) \begin{cases} = 0 & (j = 1, \ldots, r) \\ < 0 & (j = r+1, \ldots, p). \end{cases}$$

Zum Nachweis von (II) ist hier $r=p$ anzunehmen, während (I) für geeignetes r $(0\leqq r\leqq p)$ folgt, wenn man ohne Beschränkung der Allgemeinheit annimmt, daß die Numerierung der y_j passend gewählt ist.

Es werden vektorwertige Funktionen $x(t), z_1(t), \ldots, z_r(t)$ konstruiert, die in einem Intervall $[0, t_1]$ mit $t_1>0$ zweimal stetig differenzierbar sind und für die

$$\left.\begin{aligned} &x(0)=\tilde{x}, \qquad dx(0)/dt=\xi,\\ &z_j(0)=y_j \qquad (j=1,\ldots,r),\\ &f\big(x(t), z_j(t)\big)=0\\ &g_\nu\big(z_j(t)\big)=0 \quad (\nu\in M_j) \end{aligned}\right\} \quad (j=1,\ldots,r;\ 0\leqq t\leqq t_1) \qquad \begin{matrix}(14)\\(15)\end{matrix}$$

gilt. Bei $y=z_j(t)$ soll eine Maximalstelle von $f\big(x(t), y\big)$ bezüglich Y liegen. Dort werden daher noch die lokalen Kuhn-Tucker-Bedingungen gefordert:

$$f_y'\big(x(t), z_j(t)\big)-\sum_{\nu\in M_j} w_{\nu j}(t)\, g_\nu'\big(z_j(t)\big)=0 \quad (j=1,\ldots,r;\ 0\leqq t\leqq t_1) \tag{16}$$

mit $w_{\nu j}(0)=v_{\nu j}$. Wenn wie oben m_j die Anzahl der Indizes in M_j ist, hat man mit (14), (15) und (16)

$$r+\sum_{j=1}^{p} m_j+mr$$

Gleichungen. Für $x(t)$ wird der Ansatz

$$x(t)=\tilde{x}+t\xi+\sum_{k=1}^{r} c_k(t)\, f'(\tilde{x}, y_k) \tag{17}$$

gemacht.

Man hat dann r reellwertige Funktionen $c_k(t)$, $\sum m_j$ reellwertige Funktionen $w_{\nu j}(t)$ und mr Komponenten der vektorwertigen Funktionen $z_j(t)$ zu den Anfangswerten

$$c_k(0)=0, \qquad w_{\nu j}(0)=v_{\nu j}, \qquad z_j(0)=y_j$$

aus den Gln. (14), (15) und (16) zu bestimmen. Für die Anfangswerte sind diese Gleichungen erfüllt. Die schon in §4 benutzten Sätze über implizite Funktionen sind anwendbar, denn die Jacobi-Matrix an der Stelle $t=0$ ist nichtsingulär. Sie ist von der Form

$$\begin{pmatrix} A & 0 & B \\ 0 & 0 & C \\ \underbrace{D}_{r} & \underbrace{-C^T}_{\sum m_j} & \underbrace{G}_{mr} \end{pmatrix} \begin{matrix}\}\, r\\ \}\sum m_j\\ \}\, mr\end{matrix} \tag{18}$$

Dabei ist $A=(f_j'^T f_k')_{j,k=1,\ldots,r}$ wegen der linearen Unabhängigkeit der f_j' nichtsingulär. Weiterhin ist

$$B=\begin{pmatrix} f_1'^T & 0 & \ldots. \\ 0 & f_2'^T & \ldots. \\ . & . & . \end{pmatrix},$$

C ist entsprechend aus den Vektoren $g_{\nu j}^{\cdot T}$ aufgebaut, D aus Vektoren $f_j^{\cdot\cdot} f_k'$, und G ist eine Blockmatrix mit r $m \times m$-Matrizen $f_j^{\cdot\cdot} - \sum v_{\nu j} g_{\nu j}^{\cdot\cdot}$ längs der Hauptdiagonale, sonst mit Nullen ausgefüllt. Die Matrix (18) ist nichtsingulär, wenn es die Matrix

$$\begin{pmatrix} A & 0 & 0 \\ 0 & 0 & C \\ D & -C^T & G \end{pmatrix}$$

ist, denn diese entsteht wegen (9) aus (18), wenn man jeweils das $v_{\nu j}$-fache der Zeilen von $(0\ 0\ C)$ von den entsprechenden Zeilen $(A\ 0\ B)$ subtrahiert. Es bleibt zu zeigen, daß

$$\begin{pmatrix} 0 & C \\ -C^T & G \end{pmatrix}$$

nichtsingulär ist. Es genügt, dies für eine Matrix

$$\begin{pmatrix} 0 & \dots & 0 & & g_{1j}^{\cdot T} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \dots & 0 & & g_{mj}^{\cdot T} \\ g_{1j}^{\cdot} & \dots & g_{mj}^{\cdot} & f_j^{\cdot\cdot} - \sum v_{\nu j} g_{\nu j}^{\cdot\cdot} & \end{pmatrix}$$

unter der Annahme $M_j = \{1, \dots, m_j\}$ zu zeigen. Wäre sie singulär, so gäbe es Zahlen $a_1, \dots, a_{m_j}$ und einen Vektor $\gamma \in R^m$, nicht sämtlich $= 0$, mit

$$\begin{gathered} \sum a_\nu g_{\nu j}^{\cdot} + (f_j^{\cdot\cdot} - \sum v_{\nu j} g_{\nu j}^{\cdot\cdot})\gamma = 0, \qquad (19) \\ g_{\nu j}^{\cdot T}\gamma = 0 \quad (\nu \in M_j). \end{gathered}$$

Die $g_{\nu j}^{\cdot}$ sind in (III) als linear unabhängig vorausgesetzt, darum ist $\gamma \neq 0$. Multipliziert man (19) von links mit γ^T, so wird $\gamma^T(f_j^{\cdot\cdot} - \sum v_{\nu j} g_{\nu j}^{\cdot\cdot})\gamma = 0$ mit $\gamma \neq 0$ im Widerspruch zur Definitheitsvoraussetzung in (III). Die Sätze über implizite Funktionen besagen dann, daß es ein $t_1 > 0$ und im Intervall $[0, t_1]$ zweimal stetig differenzierbare $c_k(t), z_j(t), w_{\nu j}(t)$ gibt, die mit $c_k(0) = 0$, $z_j(0) = y_j$, $w_{\nu j}(0) = v_{\nu j}$ eindeutig bestimmt sind. Da (16) erste Ableitungen enthält, mußte im Satz die Stetigkeit der dritten Ableitungen von f und g_ν vorausgesetzt werden.

Es ist noch $dx(0)/dt = \xi$ zu zeigen. Aus (14) folgt

$$f_j^{\cdot T}\left(\xi + \sum_{k=1}^{r} \frac{dc_k(0)}{dt} f_k'\right) + f_j^{\cdot T} \frac{dz_j(0)}{dt} = 0 \quad (j = 1, \dots, r).$$

Hier ist $f_j^{\cdot T}\xi = 0$ und auch $f_j^{\cdot T} dz_j(0)/dt = 0$ wegen (9) und $g_{\nu j}^{\cdot T} dz_j(0)/dt = 0$, was aus (15) folgt. Da die Matrix $(f_j^{\cdot T} f_k')_{j,k=1,\dots,r}$ nichtsingulär ist, wird $dc_k(0)/dt = 0$ $(k = 1, \dots, r)$ und damit $dx(0)/dt = \xi$.

Für $\nu \notin M_j$ ist $g_\nu(y_j) < 0$ und darum $g_\nu(z_j(t)) < 0$ für $0 \leqq t \leqq t_2$ mit $0 < t_2 \leqq t_1$, also $z_j(t) \in Y$ für $t \in [0, t_2]$, $j = 1, \dots, r$.

Schließlich ist zu zeigen, daß $x(t) \in M$, also $f(x(t), y) \leqq 0$ $(y \in Y)$ ist in einem Intervall $[0, t^*]$ mit $t^* > 0$. Hierzu werden Kugelumgebungen von $y_1, \dots, y_r$ und von $y_{r+1}, \dots, y_p$ konstruiert, und es ist zu unterscheiden, ob y in einer solchen Umgebung liegt oder nicht.

Sei zunächst $1 \leqq j \leqq r$. Für hinreichend kleine $t > 0$ ist $y = z_j(t)$ Maximalstelle von $f(x(t), y)$ bezüglich Y zum Maximalwert $f(x(t), z_j(t)) = 0$. Nach (16) und wegen $w_{\nu j}(0) = v_{\nu j} > 0$ $(\nu \in M_j)$ ist nämlich für kleine positive t die lokale Kuhn-Tucker-Bedingung (a) von §2 erfüllt. Die für $t = 0$ durch (III) vorausgesetzte Definitheitsbedingung (b) ist ebenfalls aus Stetigkeitsgründen für kleine positive t erfüllt. Damit gibt es für $t \in [0, t_3]$ mit $0 < t_3 \leqq t_2$ eine Umgebung von $z_j(t)$, in deren Durchschnitt mit Y ($z_j(t)$ selbst ausgenommen)

$$f(x(t), y) < f(x(t), z_j(t)) = 0 \tag{20}$$

ist. Man kann hierfür die in [8] konstruierten Kugelumgebungen wählen, und zwar so, daß deren Radius stetig von t abhängt (hier wird noch einmal die Stetigkeit der dritten Ableitungen von f und g_ν benutzt). Sei also $\varrho_j(t)$ der Radius der Kugelumgebung um $z_j(t)$, wobei $\varrho_j(t) > 0$ und stetig ist in einem Intervall $[0, t_4]$ mit $t_4 > 0$, und zwar dies für alle j mit $1 \leqq j \leqq r$. Nun wird t_5 mit $0 < t_5 \leqq t_4$ so klein gewählt, daß

$$\max_{[0,t_5]} \|y_j - z_j(t)\| \leqq \tfrac{1}{3} \min_{[0,t_5]} \varrho_j(t)$$

ist für $j = 1, \ldots, r$. Dabei ist $\|\cdot\|$ die euklidische Vektornorm im R^m. Die Kugel K_j mit dem Radius $\tau_j = \frac{1}{2} \min_{[0,t_5]} \varrho_j(t)$ um y_j enthält dann y_j und $z_j(t)$ $(0 \leqq t \leqq t_5)$ in ihrem Inneren. Andererseits ist K_j enthalten in den Kugeln um $z_j(t)$ mit den Radien $\varrho_j(t)$ für alle $t \in [0, t_5]$, denn aus $\|y - y_j\| \leqq \tau_j$ folgt

$$\|y - z_j(t)\| \leqq \tau_j + \|y_j - z_j(t)\| \leqq \tfrac{5}{6} \varrho_j(t).$$

Für $y \in Y \cap (K_1 \cup \cdots \cup K_r)$ ist damit nach (20) $f(x(t), y) \leqq 0$ $(0 \leqq t \leqq t_5)$.

Nun sei $r+1 \leqq j \leqq p$. Es ist $f(\tilde{x}, y_j) = 0$ und $f(\tilde{x}, y) < 0$ für $0 < \|y - y_j\| \leqq s$, $y \in Y$ mit hinreichend kleinem $s > 0$. Wegen $\xi^T f'(\tilde{x}, y_j) < 0$ kann $s > 0$ so klein gewählt werden, daß $\xi^T f'(\tilde{x}, y) < 0$ für $\|y - y_j\| \leqq s$ wird. Im Durchschnitt von Y mit der Kugel K_j um y_j mit dem Radius s ist dann $f(x(t), y) < 0$ (abgesehen von $f(\tilde{x}, y_j) = 0$) für $t \in [0, t_6]$ mit $t_6 > 0$.

Sei schließlich $y \in L = Y - \bigcup_{j=1}^{p} K_j$; dort ist $f(\tilde{x}, y) \leqq \sigma < 0$, denn $y_1, \ldots, y_p$ sind die einzigen Nullstellen von $f(\tilde{x}, y)$ in Y und Y ist beschränkt und abgeschlossen. Darum gibt es ein $t_7 > 0$, so daß $f(x(t), y) \leqq 0$ ist für $y \in L$ und $0 \leqq t \leqq t_7$. Insgesamt ist, wenn man $t^* = \min(t_5, t_6, t_7)$ setzt,

$$f(x(t), y) \leqq 0 \quad (y \in Y,\ 0 \leqq t \leqq t^*).$$

§7. Beispiel

Im Beispiel ist $n = m = 2$ und $N = 1$ bzw. $= 2$.

$$F(x) = x_2 + c\, x_1^2,$$

$$f(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 - 4 \quad (\leqq 0),$$

$$g_1(y) = y_1^2 + y_2^2 - 1 \quad (\leqq 0).$$

Hiermit wird $M=\{x;\ x_1^2+x_2^2\leqq 1\}$. Als Maximalstelle kommt $\tilde{x}=\begin{pmatrix}0\\1\end{pmatrix}$ in Betracht. Dann wird $f(\tilde{x}, y_1)=0$ bei $y_1=\begin{pmatrix}0\\-1\end{pmatrix}$. Weiterhin ist

$$F'(\tilde{x})=\begin{pmatrix}0\\1\end{pmatrix},\quad f'(\tilde{x}, y_1)=\begin{pmatrix}0\\4\end{pmatrix},\quad u_1=\tfrac{1}{4},$$

$$f'(\tilde{x}, y_1)=\begin{pmatrix}0\\-4\end{pmatrix},\quad g_1'(y_1)=\begin{pmatrix}0\\-2\end{pmatrix},\quad v_{11}=2.$$

Es ist

$$q(\xi, \eta_1)=2c\,\xi_1^2-\tfrac{1}{2}(\xi_1^2+\xi_2^2)+(\xi_1\eta_1+\xi_2\eta_2)+\tfrac{1}{2}(\eta_1^2+\eta_2^2).$$

$\xi\in H$ besagt $\xi_2=0$, $\eta\in H_1$ besagt $\eta_2=0$. (3) besagt schließlich $\xi_1+\eta_1=0$. Damit wird $q=(2c-1)\,\xi_1^2$. Dies ist negativ definit für $c<\frac{1}{2}$, semidefinit für $c\leqq\frac{1}{2}$. Tatsächlich hat F im Falle $c\leqq\frac{1}{2}$ bei $\tilde{x}$ ein Maximum bezüglich M, für $c>\frac{1}{2}$ jedoch nicht.

Nun werde $g_2(y)=y_1\leqq 0$ als Y beschreibende Ungleichung hinzugenommen. Dann ist

$$M=\{x;\ -\sqrt{3-x_2^2-2|x_2|}\leqq x_1\leqq\sqrt{1-x_2^2}\}.$$

Es wird $g_2'=\begin{pmatrix}1\\0\end{pmatrix}$ und damit $v_{12}=0$. $\eta\in H_1$ besagt nun $\eta_1=\eta_2=0$, wegen $H_1=\{0\}$ hat man keine Bedingungen (3), und es wird $q=(2c-\frac{1}{2})\,\xi_1^2$. Negative Definitheit hat man für $c<\frac{1}{4}$, Semidefinitheit für $c\leqq\frac{1}{4}$ und ein Maximum bei $\tilde{x}$ für $c\leqq\frac{1}{4}$. Obwohl die Bedingung (III) wegen $v_{12}=0$ nicht gilt, ist die notwendige Maximalbedingung genau für die Parameterwerte c erfüllt, für die bei $\tilde{x}$ ein Maximum liegt.

§ 8. Approximationsaufgaben

Die folgende Aufgabe der Tschebyscheff-Approximation kann als Problem des in §3 behandelten Typs geschrieben werden: *Unter den Nebenbedingungen*

$$|\varphi(y)-\Phi(y, z)|\leqq\varepsilon\quad(y\in\hat{Y}\subset R^m)\tag{21}$$

soll ε möglichst klein gemacht werden durch passende Wahl von $z\in R^{n-1}$. Dabei sei $\hat{Y}$ durch endlich viele Ungleichungen $g_\nu(z)\leqq 0$ $(\nu=1, \ldots, N)$ beschrieben. Alle auftretenden Funktionen seien zweimal stetig differenzierbar im ganzen Raum.

Wenn man (21) als

$$\left.\begin{aligned}\varphi(y)-\Phi(y, z)-\varepsilon\leqq 0\\ -\varphi(y)+\Phi(y, z)-\varepsilon\leqq 0\end{aligned}\right\}\ (y\in\hat{Y})$$

schreibt, $x=(z_1, \ldots, z_{n-1}, \varepsilon)^T$ setzt und als Y zwei Exemplare von $\hat{Y}$ wählt, hat man eine Aufgabe wie in §3. Hier sollen die hinreichenden Bedingungen des §3 für ein relatives Minimum $\tilde{\varepsilon}$ von ε an einer Stelle $z=\tilde{z}$ formuliert werden. Die Bezeichnungen entsprechen den bisher verwendeten, jedoch ist $\Phi'=(\partial\Phi/\partial z_1, \ldots, \partial\Phi/\partial z_{n-1})^T$.

(A′) *Für $y_1, \ldots, y_r\in\hat{Y}$ sei $\varphi(y_j)-\Phi(y_j, \tilde{z})=\tilde{\varepsilon}$ und für $y_{r+1}, \ldots, y_p\in\hat{Y}$ sei*

$$\varphi(y_j)-\Phi(y_j, \tilde{z})=-\tilde{\varepsilon}.$$

Im übrigen sei

$$|\varphi(y) - \Phi(y, \tilde{z})| \leqq \varepsilon \qquad (y \in \hat{Y}).$$

Es gebe reelle Zahlen $u_j > 0$ $(j = 1, \ldots, r)$, $u_j < 0$ $(j = r+1, \ldots, p)$ *mit*

$$\sum_{j=1}^{p} u_j \Phi_j' = 0, \qquad \sum_{j=1}^{p} |u_j| = 1. \tag{22}$$

(B′) *Für jedes* j $(1 \leqq j \leqq p)$ *gelte: Für* $\nu \in M_j \subset \{1, \ldots, N\}$ *und nur für diese Indizes ist* $g_\nu(y_j) = 0$. *Die Vektoren* $g_{\nu j}^{\cdot}$ *sind linear unabhängig, und es gibt reelle Zahlen* $v_{\nu j}$ $(\nu \in M_j)$, *die* $\geqq 0$ *im Falle* $1 \leqq j \leqq r$ *und* $\leqq 0$ *im Falle* $r+1 \leqq j \leqq p$ *sind, mit*

$$\varphi_j^{\cdot} - \Phi_j^{\cdot} - \sum_{\nu \in M_j} v_{\nu j} g_{\nu j}^{\cdot} = 0.$$

Auf dem linearen Teilraum

$$H_j = \{\eta; \eta^T g_{\nu j}^{\cdot} = 0 \ (\nu \in M_j)\} \subset R^m$$

ist die quadratische Form

$$Q_j(\eta) = \eta^T \left[\varphi_j^{\cdot\cdot} - \Phi_j^{\cdot\cdot} - \sum_{\nu \in M_j} v_{\nu j} g_{\nu j}^{\cdot\cdot}\right] \eta$$

negativ definit im Falle $1 \leqq j \leqq r$, *positiv definit im Falle* $r+1 \leqq j \leqq p$.

(C′) *Sei* $\{\eta_j^{(k)};\ k = 1, \ldots, m - m_j\}$ *eine Basis von* H_j $(j = 1, \ldots, p)$. *Durch die linearen Gleichungssysteme*

$$\eta_j^{(k)T} \left[-\Phi_j^{\cdot\prime} \zeta + \left(\varphi_j^{\cdot\cdot} - \Phi_j^{\cdot\cdot} - \sum_{\nu \in M_j} v_{\nu j} g_{\nu j}^{\cdot\cdot}\right) \eta_j\right] = 0$$

$$(k = 1, \ldots, m - m_j;\ j = 1, \ldots, p)$$

sind jedem $\zeta = (\zeta_1, \ldots, \zeta_{n-1}) \in R^{n-1}$ *eindeutig Vektoren* $\eta_1 \in H_1, \ldots, \eta_p \in H_p$ *zugeordnet* (N.B. das ist keine Voraussetzung, sondern Folgerung aus (B′)). *Für alle* $\zeta \in R^{n-1}$ *mit*

$$\zeta^T \Phi_1' = \cdots = \zeta^T \Phi_p' = 0 \tag{23}$$

und die jeweils zugeordneten $\eta_1, \ldots, \eta_p$ *sei*

$$q(\zeta, \eta_j) = \sum_{j=1}^{p} u_j \left\{\zeta^T \Phi_j'' \zeta + \eta_j^T \left(\varphi_j^{\cdot\cdot} - \Phi_j^{\cdot\cdot} - \sum_{\nu \in M_j} v_{\nu j} g_{\nu j}^{\cdot\cdot}\right) \eta_j\right\} < 0,$$

solange nicht $\zeta = 0$ *und daher* $\eta_1 = \cdots = \eta_p = 0$ *ist.*

Sind die Bedingungen (A′), (B′), (C′) erfüllt, so liegt bei $z = \tilde{z}$ ein relatives Minimum der Maximalabweichung

$$\varrho(z) = \operatorname*{Sup}_{y \in \hat{Y}} |\varphi(y) - \Phi(y, z)|.$$

Bemerkungen. Der lineare Teilraum H von §3 wird eigentlich durch

$$\zeta^T \Phi_1' = \cdots = \zeta^T \Phi_r' = -\zeta^T \Phi_{r+1}' = \cdots = -\zeta^T \Phi_p'$$

beschrieben. Mit (22) ergibt sich jedoch (23). Auf die Formulierung der notwendigen Bedingungen von §5 für den Spezialfall der Approximationsaufgabe soll hier verzichtet werden. Jedoch sei bemerkt, daß die lokale Kuhn-Tucker-Bedingung (7) in diesem Fall mit dem von Meinardus in [5] angegebenen Satz 84

(§ 8.1) zusammenhängt. Die Aussage, aus der (7) mit dem Lemma von Farkas folgt, besagt dasselbe wie jener Satz; allerdings sind hier nur endlich viele Extremalpunkte zugelassen und es wird eine *constraint qualification* gefordert. Daß man ohne diese hier nicht auskommt, liegt daran, daß Y nicht als beschränkt vorausgesetzt ist. Im Beispiel mit

$$\varphi(y) = \frac{1}{y_1+1}, \qquad \Phi(y, z) = z_1, \qquad g_1(y) = -y_1,$$

wo also $\hat{Y} = [0, \infty)$ und $\tilde{z}_1 = \frac{1}{2}$ mit $\tilde{\varepsilon} = \frac{1}{2}$ die eindeutig bestimmte Extremalstelle ist, gilt jener Satz von Meinardus nicht. Andererseits zeigt der Beweis jenes Satzes, daß bei beschränktem Y im Falle der Approximationsaufgabe die *constraint qualification* (I) erfüllt ist.

Beispiel. $\varphi(y) = y_1$ ist für $0 \leqq y_1 \leqq 1$ durch $\Phi(y, z) = z_1 + y_1(z_2 + y_1)^2$ zu approximieren. Man findet zwei lokale Minima:

$$\tilde{z} = \begin{pmatrix} \sqrt{1/27} \\ 0 \end{pmatrix} \qquad \text{mit} \quad \tilde{\varepsilon} = \sqrt{1/27} \approx 0{,}19,$$

$$\tilde{z} = \begin{pmatrix} (10 - 7\sqrt{7})/27 \\ -2 \end{pmatrix} \quad \text{mit} \quad \tilde{\varepsilon} = \frac{7\sqrt{7}-10}{27} \approx 0{,}34.$$

In beiden Fällen sind die obigen hinreichenden Bedingungen erfüllt. Mit $\tilde{z} = (\frac{1}{2}, -1)^T$ hat man einen weiteren Punkt, an dem die lokale Kuhn-Tucker-Bedingung erfüllt ist, die notwendige Semidefinitheitsbedingung jedoch nicht.

§ 9. Verallgemeinerungen

Die in §3 angegebene Aufgabe wurde der Übersichtlichkeit wegen möglichst einfach formuliert. Einige Verallgemeinerungen liegen nahe.

a) Man kann $Y = Y_1 \cup \cdots \cup Y_s$ wählen, wobei die Y_i Teilmengen von Räumen verschiedener Dimension sind. Bei der numerischen Behandlung von Randwertaufgaben monotoner Art nach dem Prinzip der lokal optimalen Schranken [6] kann etwa Y_1 ein Bereich sein, in dem die Differentialgleichung gefordert ist, Y_2 sein Rand. Insbesondere können auch isolierte Punkte zu Y gehören oder — anders gesagt — können einzelne Ungleichungen $f_i(x) \leqq 0$ hinzutreten. Wie in diesen Fällen die Maximalbedingungen zu formulieren und zu beweisen sind, ist evident.

b) Überdies dürfen in der Beschreibung von M und Y neben Ungleichungen auch endlich viele Gleichungen vorkommen. Die zugehörigen Multiplikatoren u_i bzw. $v_{\nu,i}$ sind dann keinen Vorzeichenbeschränkungen zu unterwerfen.

c) Es ist natürlich nicht wesentlich, daß die auftretenden Funktionen im ganzen Raum definiert sind.

d) Eine naheliegende, aber wohl weder für die Anwendungen noch beweistechnisch interessante Verallgemeinerung der behandelten zweistufigen Aufgaben wären mehrstufige Aufgaben.

e) Im Hinblick auf Anwendungen u.a. in der *control theory* könnte eine Verallgemeinerung der hier gefundenen Ergebnisse auf abstrakte Räume (z.B. Banachräume statt R^n und R^m) wünschenswert sein.

Literatur

1. Collatz, L.: Approximation in partial differential equations, S. 413—422 in R. E. Langer (ed.): On numerical approximation. Madison: The University of Wisconsin Press 1959.
2. — Wetterling, W.: Optimierungsaufgaben. Berlin-Göttingen-Heidelberg-New York: Springer 1966.
3. Dieudonné, J.: Foundations of modern analysis. New York-London: Academic Press 1960.
4. McCormick, G. P.: Second order conditions for constrained minima. SIAM J. Appl. Math. **15**, 641—652 (1967).
5. Meinardus, G.: Approximation von Funktionen und ihre numerische Behandlung. Berlin-Göttingen-Heidelberg-New York: Springer 1964.
6. Wetterling, W.: Lokal optimale Schranken bei Randwertaufgaben. Computing **3**, 125—130 (1968).
7. — Lösungsschranken bei elliptischen Differentialgleichungen. International Series of Numerical Mathematics **9**, 393—401 (1968).
8. — Über Minimalbedingungen und Newton-Iteration bei nichtlinearen Optimierungsaufgaben. Erscheint in: International Series of Numerical Mathematics. Basel: Birkhäuser.

Prof. Dr. W. Wetterling
Technische Hogeschool Twente
Onderafdeling der Toegepaste Wiskunde
Postbus 217
Enschede/Niederlande

Numer. Math. 15, 137–150 (1970)

Global Approximation of a Compact Set by Elements of a Convex Set in a Normed Space

P. J. LAURENT* and PHAM-DINH-TUAN

Dedicated to Prof. Dr. Dr. h.c. L. Collatz for his 60th birthday

Received January 15, 1970

Introduction

In a paper published in 1934, Remes [12] considered the following problem: Let $I \subset R$ be a bounded set of real numbers, f a real function bounded on I, P_n the set of all polynomials of degree n. Determine $p^* \in P_n$ such that:

$$\|f - p^*\| = \min_{p \in P_n} \|f - p\| \tag{1}$$

with

$$\|f - p\| = \sup_{x \in I} |f(x) - p(x)| . \tag{2}$$

If f_1 and f_2 are the upper and lower Baire functions, i.e.:

$$f_1(x) = \limsup_{y \to x} f(y); \quad f_2(x) = \liminf_{y \to x} f(y), \tag{3}$$

then, for $x \in \bar{I}$, we have:

$$\|f - p\| = \max\left(\max_{x \in \bar{I}} (f_1(x) - p(x)); \max_{x \in \bar{I}} (p(x) - f_2(x))\right). \tag{4}$$

The functions f_1 and f_2 are respectively upper semi-continuous and lower semi-continuous on the compact set $\bar{I}$. The problem is reduced to the simultaneous approximation of two functions f_1 and f_2 by a polynomial of degree n. Remes [12] gave the following characterization of the solution which parallels the classical Chebyshev-theorem (corresponding to $f_1 = f_2$):

The polynomial $p^* \in P_n$ is a solution of (1) if and only if either: 1) there exist $n+2$ points $x_0 < x_1 < \cdots < x_{n+1}$ in $\bar{I}$ at which the two functions $f_1 - p^*$ and $p^* - f_2$ alternatively take the value $\|f - p^*\|$, or 2) $\|f - p^*\| = \delta/2$ with $\delta = \|f_1 - f_2\|$.

If $\|f - p^*\| > \delta/2$, the solution is unique. If $\|f - p^*\| = \delta/2$, it is not necessarily unique.

In 1959, Golomb [5], starting from the approximation of functions given by incomplete data, was led to consider and solve (independently from Remes) the same problem. In fact, he takes as a class of approximants either an n-dimensional linear subspace satisfying the Haar-property or a certain class of rational functions.

* Sponsored by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract No.: DA-31-124-ARO-D-462.

In 1967, Dunham [3] considered and solved again the same problem (without reference to the work of Remes and Golomb), the class of approximants being a unisolvent family of degree n.

In the present paper, the space of continuous functions with the Chebyshev-norm will be replaced by an arbitrary normed linear space E, and the set $\{f_1, f_2\}$ by a compact set F of E. We will consider the global approximation of F by an element of a closed convex set G of E. The notion of sub-differential of a convex functional will be systematically used for the characterization of the solution (cf. [10, 7, 9]). In the case when G is a finite-dimensional subspace, we obtain a generalization of Singer's theorem [13, 14]. We will consider several examples of normed spaces and of convex sets. The numerical solutions have been obtained by a modification of the Remes algorithm [6, 11].

Our practical motivation for the global approximation of a compact set is the following one. The function which is to be approximated very often depends on parameters; for example:

$$x \to f(\lambda_1, \ldots, \lambda_m; x).$$

Frequently, these parameters $\lambda_1, \ldots, \lambda_m$ are not known exactly (they can be obtained experimentally or on a computer by programmes based on interval arithmetic). We suppose that $\lambda_i \in [\alpha_i, \beta_i]$. Under some continuity-assumptions the corresponding set F of functions is compact. It seems to be better to approximate globally (i.e. simultaneously) all the functions $f \in F$ rather than one function f arbitrarily chosen in F.

I. General Characterization Theorems

Let E be a normed linear space (over the reals), the norm of $x \in E$ being denoted as $\|x\|$, and let E' be its dual space. The value of the continuous linear functional $k \in E'$ at $x \in E$ will be denoted as (k, x).

Let K be a subset of E' which is symmetrical, $\sigma(E', E)$-compact and bounded in the norm topology of E'. We define a continuous semi-norm p on E by:

$$p(x) = \max_{k \in K} (k, x). \tag{5}$$

(If $K = S'$, the unit sphere of E', then we have $p(x) = \|x\|$.) Let F be a compact set (in the norm-topology) of E. We introduce, for $g \in E$:

$$d(g) = \max_{f \in F} p(g - f). \tag{6}$$

Let G be a closed convex subset of E.

We seek an element $g^* \in G$ such that:

$$d(g^*) = \min_{g \in G} d(g). \tag{7}$$

Such an element will be called a *best global approximation to F in G*, or shorter a "solution".

The function d is in fact the supremum of a family of continuous affine functionals. If we denote:

$$d_{f,k}(g) = (k, g - f) \tag{8}$$

we have:

$$d(g) = \max_{f \in F,\, k \in K} d_{f,k}(g). \tag{9}$$

The function d therefore is convex and lower semi-continuous. We will apply several times the following theorem (Valadier[15]):

Let Ω be a compact subset and (φ_w), $w \in \Omega$, a family of convex functionals defined on a Hausdorff linear topological space. Let $\varphi = \sup_{w \in \Omega} \varphi_w$. If there exists a neighborhood $\mathcal{U}$ of x_0 such that the mapping

$$[w, x] \to \varphi_w(x)$$

is continuous on $\Omega \times \mathcal{U}$, then φ is continuous at x_0 and the sub-differential $\partial\varphi(x_0)$ of the functional φ at x_0 has the following expression:

$$\partial\varphi(x_0) = \overline{\mathrm{co}}\left(\bigcup_{w \in M(x_0)} \partial\varphi_w(x_0)\right) \tag{10}$$

with

$$M(x_0) = \{w \in \Omega\colon \varphi_w(x_0) = \varphi(x_0)\}. \tag{11}$$

($\overline{\mathrm{co}}(A)$ denotes the smallest closed convex subset containing A.)

We can now express the sub-differential of d:

Lemma 1.1. *The function d is continuous and its sub-differential $\partial d(g)$ at g is:*

$$\partial d(g) = \overline{\mathrm{co}}\left(A(g)\right) \tag{12}$$

with:

$$A(g) = \{k \in K\colon \exists f \in F,\ (k, g - f) = d(g)\}. \tag{13}$$

Proof. We have clearly:

$$\partial d_{f,k}(g) = \{k\}. \tag{14}$$

The functional

$$[f, k, g] \to (k, g - f)$$

defined on $F \times K \times E$ is continuous (for F and E we take the norm-topology and for K the $\sigma(E', E)$-topology). Therefore, applying Valadier's theorem with $\Omega = F \times K$, we obtain by (10) and (11):

$$\partial d(g) = \overline{\mathrm{co}}\left(\bigcup_{[f,k] \in M(g)} \{k\}\right)$$

with:

$$M(g) = \{[f, k] \in F \times K\colon (k, g - f) = d(g)\}.$$

Finally, we have:

$$\partial d(g) = \overline{\mathrm{co}}\{k \in K\colon \exists f \in F,\ (k, g - f) = d(g)\}$$

and the lemma is proved. Q.E.D.

We can now prove the following characterization theorem:

Theorem 1.1. *An element $g^* \in G$ is a solution if and only if there exist $l_0 \in \overline{\text{co}}\,(A(g^*))$, with $A(g^*)$ defined as in (13) and $l_1 \in E'$ satisfying $(l_1, g^*) = \max_{g \in G} (l_1, g)$ such that $l_0 + l_1 = 0$.*

Proof. Let $\varphi = d + \psi$ with

$$\psi(x) = \begin{cases} 0 & \text{if} \quad x \in G \\ \infty & \text{if} \quad x \notin G. \end{cases}$$

Then, d is minimum on G at $g^* \in G$ if and only if:

$$0 \in \partial\varphi(g^*). \tag{15}$$

Here we have (cf. [10, 9]):

$$\partial\varphi(g^*) = \partial d(g^*) + \partial\psi(g^*) \tag{16}$$

The sub-differential of ψ is given by:

$$\partial\psi(g^*) = \left\{ l \in E' : \ (l, g^*) = \max_{g \in G} (l, g) \right\}. \tag{17}$$

The Theorem 1.1 follows directly from (15), (16), (17) and Lemma 1.1. Q.E.D.

We will now consider some particular convex sets G. Let

$$G = C \cap W \tag{18}$$

where W is a translate of a closed linear subspace V and C is defined by

$$C = \{x \in E : \ c(x) \leqq 0\} \tag{19}$$

where c is a continuous convex functional on E. We assume that

$$\text{there exists } g_0 \in W \text{ such that } c(g_0) < 0. \tag{H1}$$

This assumption implies that the subset

$$I = \{x \in E : \ c(x) < 0\} \tag{20}$$

is the (non-empty) interior of C.

Therefore we have:

$$\text{Inter}\,(C) \cap W \neq \emptyset. \tag{21}$$

In the Proof of Theorem 1.2 we will give a characterization of a solution in the case where the set C would be defined by (19), but in view of the most usual examples, we particularize the function c. We define the convex subset C by

$$C = \{x \in E : \forall h \in H, \ (h, x) \leqq w(h)\} \tag{22}$$

where H is a subset of E' which is $\sigma(E', E)$-compact and bounded in the norm topology and w a weakly continuous functional defined on H. The corresponding function c is given by:

$$c(x) = \max_{h \in H} [(h, x) - w(h)]. \tag{23}$$

For $g \in G$, we define:

$$B(g) = \{h \in H : \ (h, g) - w(h) = 0\}. \tag{24}$$

The subset $B(g)$ is non-empty if and only if $c(g)=0$. We have the following characterization theorem:

Theorem 1.2. Assume (H 1). *An element $g^* \in C \cap W$ is a solution if and only if*

$$\overline{\mathrm{co}}\left(A(g^*) \cup B(g^*)\right) \cap V^{\perp} \neq \emptyset \tag{25}$$

where $A(g^*)$ and $B(g^*)$ are defined by (13) *and* (24).

Proof. Let $\varphi = d + \psi_1 + \psi_2$ with

$$\psi_1(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases} \quad \text{and} \quad \psi_2(x) = \begin{cases} 0 & \text{if } x \in W \\ \infty & \text{if } x \notin W. \end{cases}$$

An element $g^* \in C \cap W$ is a solution if and only if the condition (15) holds. The assumption (H 1) implies that ψ_1 is finite and continuous at g_0 and ψ_2 finite at this point. Therefore we have (cf. [10, 9]):

$$\partial\varphi(g^*) = \partial d(g^*) + \partial\psi_1(g^*) + \partial\psi_2(g^*). \tag{26}$$

The sub-differential $\partial d(g^*)$ is still given by Lemma 1.1. We have clearly:

$$\partial\psi_2(g^*) = V^{\perp}. \tag{27}$$

It remains to express $\partial\psi_1(g^*)$:

If $c(g^*) < 0$, i.e. $g^* \in \mathrm{Inter}(C)$, we have $\partial\psi_1(g^*) = \{0\}$. Suppose now $c(g^*) = 0$; by the assumption (H 1), the function c does not take its minimum over E at g^*, hence $0 \notin \partial c(g^*)$. Therefore, if $c(g^*) = 0$, we have:

$$\partial\psi_1(g^*) = \mathrm{cc}\left(\partial c(g^*)\right), \tag{28}$$

$\left(\mathrm{cc}(A)\right.$ denotes the smallest convex cone with vertex 0 containing $A\left.\right)$.

It follows from (15), (26), (27), (28) and Lemma 1.1 that:

An element $g^* \in C \cap W$ is a solution if and only if there exist

$$l_0 \in \overline{\mathrm{co}}\left(A(g^*)\right),$$

$$l_1 \in \begin{cases} \mathrm{cc}\left(\partial c(g^*)\right) & \text{if } c(g^*) = 0 \\ \{0\} & \text{if } c(g^*) < 0 \end{cases}$$

$$l_2 \in V^{\perp}$$

such that $l_0 + l_1 + l_2 = 0$.

We will consider two cases:

1°) $$c(g^*) = 0.$$

The subset $B(g^*)$ is non-empty. Using again Valadier's theorem one finds:

$$\partial c(g^*) = \overline{\mathrm{co}}\left(B(g^*)\right). \tag{29}$$

The assumption (H 1) implies that $\overline{\mathrm{co}}\left(B(g^*)\right) \cap V^{\perp} = \emptyset$ (otherwise, c would be minimum over V at g^*, which contradicts (H 1)). An elementary calculation

shows that the preceding necessary and sufficient condition is equivalent to:

$$\text{co}\left(\overline{\text{co}}\left(A(g^*)\right) \cup \overline{\text{co}}\left(B(g^*)\right)\right) \cap V^\perp \neq \emptyset. \tag{30}$$

$$2°) \qquad c(g^*) < 0.$$

The condition of the theorem is equivalent to

$$\overline{\text{co}}\left(A(g^*)\right) \cap V^\perp \neq \emptyset. \tag{31}$$

Since $B(g^*) = \emptyset$, this condition is equivalent to (30).

In both cases, an element $g^* \in C \cap W$ is a solution if and only if the condition (30) is satisfied.

It is not difficult to prove that:

$$\text{co}\left(\overline{\text{co}}\left(A(g^*)\right) \cup \overline{\text{co}}\left(B(g^*)\right)\right) = \overline{\text{co}}\left(A(g^*) \cup B(g^*)\right)$$

(cf. [2], p. 415), which completes the proof. Q.E.D.

II. Approximation in a Convex Subset of a Finite Dimensional Subspace

Let us suppose now that V is an n-dimensional subspace of E. We will need the following property:

Lemma 2.1. *Let Ω be a compact subset in a Hausdorff topological linear space X such that $\overline{\text{co}}(\Omega)$ is compact, and let r be a continuous linear mapping of X into a finite dimensional space. Then we have:*

$$r\left(\overline{\text{co}}(\Omega)\right) = \text{co}\left(r(\Omega)\right). \tag{32}$$

Now we can prove the following result:

Theorem 2.1. Assume (H 1). *An element $g^* \in C \cap W$ is a solution if and only if there exist: r functionals $k_1, \ldots, k_r \in K$ $(r \geqq 1)$ and r elements $f_1, \ldots, f_r \in F$ (not necessarily distinct) satisfying:*

$$(k_i, g^* - f_i) = p(g^* - f_i) = \max_{f \in F} p(g^* - f), \quad i = 1, \ldots, r \tag{33}$$

s functionals $h_1, \ldots, h_s \in H$ $(s \geqq 0)$ satisfying:

$$(h_i, g^*) = w(h_i), \quad i = 1, \ldots, s, \quad \text{with} \quad r + s \leqq n + 1 \tag{34}$$

and $r + s$ scalars $\lambda_1, \ldots, \lambda_r, \mu_1, \ldots, \mu_s > 0$ such that:

$$\sum_{i=1}^{r} \lambda_i k_i + \sum_{i=1}^{s} \mu_i h_i \in V^\perp. \tag{35}$$

Proof. Assume $g^* \in C \cap W$ to be a solution. We denote by $\tilde{l}$ the restriction to V of the functional $l \in E'$. Consider the continuous linear mapping:

$$r: l \in E' \to \tilde{l} \in V'.$$

By Theorem 1.2 we have:

$$0 \in r\left(\overline{\text{co}}\left(A(g^*) \cup B(g^*)\right)\right). \tag{36}$$

As V' is finite dimensional, Lemma 2.1 can be applied (with $\Omega = A(g^*) \cup B(g^*)$) and we obtain:

$$0 \in \operatorname{co}\left(r(A(g^*) \cup B(g^*))\right). \tag{37}$$

By Caratheodory's theorem (cf. [4]) there exist $l_1, \ldots, l_q \in A(g^*) \cup B(g^*)$ with $1 \leqq q \leqq n+1$ and $\varrho_1, \ldots, \varrho_q > 0$, $\sum_{i=1}^{q} \varrho_i = 1$ such that:

$$0 = \sum_{i=1}^{q} \varrho_i r(l_i) \tag{38}$$

i.e.:

$$\sum_{i=1}^{q} \varrho_i l_i \in V^{\perp}. \tag{39}$$

We denote by k_i the functionals l_i which belong to $A(g^*)$ and by h_i the others. The assumption (H 1) implies that $r \geqq 1$. Q.E.D.

We will now assume that:

(H 2) K and H are convex and the functional w is concave.

As in the well known Singer theorem [13, 14] we can characterize a solution by using only the functionals which are extremal points of K or H.

We recall that a non-void subset $Y \subset X$ is said to be an extremal subset of X if a proper convex combination $\lambda x_1 + (1-\lambda) x_2$, $0 < \lambda < 1$ of two points $x_1, x_2 \in X$ is in Y only if both x_1 and x_2 are in Y (an extremal subset of X consisting of just one point is called extremal point of X).

We will successively use the following lemmae:

Lemma 2.2. *Let P be a compact subset of an n-dimensional linear space. If $p \in \operatorname{co}(P)$, then there exist q extremal points $p_1, \ldots, p_q$ of P (with $q \leqq n+1$) and $\varrho_1, \ldots, \varrho_q > 0$, $\sum_{i=1}^{q} \varrho_i = 1$ such that $p = \sum_{i=1}^{q} \varrho_i p_i$.*

This lemma is a combination of Krein-Milman's theorem in R^n ([2], p. 440, [1], p. 157, Ex. 22) and of Caratheodory's theorem [4].

Lemma 2.3. *Let T be a continuous linear mapping of E_1 into E_2 (two Hausdorff local convex linear topological spaces) and X a compact subset of E_1. Then for every extremal point y of $T(X)$ there exists at least one extremal point x of X such that $T(x) = y$.*

(This property is indicated in [1] (p. 159, Ex. 32) in the case where X is convex)

Lemma 2.4. *Let φ be a convex functional defined on a subset X of a linear space. Suppose that*

$$Y = \left\{ y \in X : \varphi(y) = \sup_{x \in X} \varphi(x) \right\}$$

is non-empty. Then Y is an extremal subset of X.

Lemma 2.5 ([2], p. 440). *Let X be an extremal subset of Y and Y an extremal subset of Z (in a linear space); then X is an extremal subset of Z.*

We can now prove the following result:

Theorem 2.2. Assume (H1) and (H2). *An element $g^* \in C \cap W$ is a solution if and only if there exist: r functionals $k_1, \ldots, k_r$ extremal of K ($r \geqq 1$) and r elements $f_1, \ldots, f_r \in F$ (not necessarily distinct) satisfying*

$$(k_i, g^* - f_i) = p(g^* - f_i) = \max_{f \in F} p(g^* - f), \quad i = 1, \ldots, r \tag{40}$$

s functionals $h_1, \ldots, h_s$ extremal of H ($s \geqq 0$) satisfying:

$$(h_i, g^*) = w(h_i), \quad i = 1, \ldots, s, \quad \text{with} \quad r + s \leqq n + 1 \tag{41}$$

and $r+s$ scalars $\lambda_1, \ldots, \lambda_r, \mu_1, \ldots, \mu_s > 0$ such that:

$$\sum_{i=1}^{r} \lambda_i k_i + \sum_{i=1}^{s} \mu_i h_i \in V^\perp. \tag{42}$$

Proof. Assume that $g^* \in C \cap W$ is a solution. With the notations introduced in the proof of Theorem 2.1 we have:

$$0 \in \operatorname{co}(P) \tag{43}$$

with $P = r(\Omega)$ and $\Omega = A(g^*) \cup B(g^*)$. Hence, by Lemma 2.2:

$$0 = \sum_{i=1}^{q} \varrho_i p_i \tag{44}$$

where the p_i are extremal points of P, $\varrho_i > 0$, $\sum_{i=1}^{q} \varrho_i = 1$ and $q \leqq n+1$. By Lemma 2.3 there exist q functionals l_i which are extremal points of Ω such that:

$$p_i = r(l_i), \quad i = 1, \ldots, q. \tag{45}$$

We denote by k_i the functionals l_i which are extremal points of $A(g^*)$ and by h_i the others (which are then necessarily extremal of $B(g^*)$). Therefore we have:

$$\sum_{i=1}^{r} \lambda_i k_i + \sum_{i=1}^{s} \mu_i h_i \in V^\perp. \tag{46}$$

As $A(g^*)$ is the set of points where the convex functional:

$$k \in K \to \max_{f \in F} (k, g^* - f)$$

takes its maximum value $d(g^*)$ on K, this set $A(g^*)$ is extremal of K (Lemma 2.4).

In the same way, $B(g^*)$ is either the empty set or the set of points where the convex functional:

$$h \in H \to (h, g^*) - w(h)$$

takes its maximum value 0 on H; hence, $B(g^*)$ is an extremal subset of H. (Note that $A(g^*)$ and $B(g^*)$ are not necessarily convex.) By Lemma 2.5 the k_i are extremal points of K and the h_i extremal points of H. Q.E.D.

The following result is an immediate consequence of Theorem 2.2 in the case where G is the translate W of an n-dimensional subspace V and $p(x) = \|x\|$ (i.e. $K = S'$, the unit sphere of E').

Corollary. *An element $g^* \in W$ is a solution if and only if there exist q functionals $k_1, \ldots, k_q$ which are extremal points of S', $f_1, \ldots, f_q \in F$ (not necessarily distinct), $\lambda_1, \ldots, \lambda_q > 0$ (with $1 \leqq q \leqq n+1$) such that:*

1) $$(k_i, g^* - f_i) = \|g^* - f_i\| = \max_{f \in F} \|g^* - f\|, \quad i = 1, \ldots, q.$$

2°) $$\sum_{i=1}^{q} \lambda_i k_i \in V^{\perp}.$$

If F consists of just one element f, the corollary becomes exactly Singer's theorem.

Remark. If F is a convex subset of E, in the definition of d as well in both Theorems 2.1 and 2.2, we can replace F by $\mathscr{E}(F)$, the set of extremal points of F.

Proof. If we set

(47) $$c(k) = \min_{f \in F} (k, f)$$

we have

(48) $$d(g) = \max_{k \in K} [(k, g) - c(k)]$$

and in Theorems 2.1 and 2.2 the k_i satisfy:

(49) $$d(g^*) = (k_i, g^*) - c(k_i).$$

As we have, by Lemma 2.3:

(50) $$c(k) = \min_{f \in \mathscr{E}(F)} (k, f)$$

we can replace F by $\mathscr{E}(F)$ as indicated in the remark. Q.E.D.

III. Applications

In this section we will consider three examples of normed linear spaces. For each case we will give the characterization theorem and a numerical example.

Put:

(51) $$\delta = \sup_{f_1, f_2 \in F} p(f_1 - f_2)$$

(the diameter of F with respect to the semi-norm p) and

(52) $$\varrho = \operatorname{Inf}_{g \in G} d(g); \qquad \varrho_0 = \operatorname{Inf}_{g \in E} d(g).$$

It is easy to see that

(53) $$\delta \leqq 2\varrho_0 \leqq 2\varrho.$$

If $\delta = 2\varrho$ (which implies $\varrho_0 = \varrho$), every best global approximation to F in G is a best global approximation to F in the whole space E.

This is the second alternative in the Remes theorem (cf. introduction).

In this case the characterization theorems become very trivial (in general there are only two functionals k_1 and k_2, and they are opposite). We will choose the following examples so that the condition $\varrho_0 < \varrho$ holds.

1°) *Approximation in the Space of Continuous Functions*

Let $E = C[S]$ be the space of real continuous functions on a compact S.

We put:

$$p(f) = \max_{t \in T} |f(t)|$$

where $T \subset S$ is a compact subset.

We define the convex set C by

$$C = \{g \in C[S] : \forall u \in U,\ g(u) \leqq w(u)\}$$

when $w \in C[S]$ and $U \subset S$ is a compact subset. (One could consider many other sorts of constraints, see [11]). Let W be a translate of an n-dimensional subspace V of $C[S]$. We can apply Theorem 2.1: the subset K of $C[S]'$ is the set of all continuous linear functionals of the form:

$$f \in C[S] \to \varepsilon \cdot f(t)$$

with $\varepsilon = \pm 1$ and $t \in T$. In the same way, the subset H is the set of all continuous linear functionals of the form:

$$g \in C[S] \to g(u)$$

with $u \in U$. The subsets K and H have the required properties. Assume (H1), i.e. there exists $g_0 \in W$ such that $\forall u \in U$, $g_0(u) < w(u)$. Then the Theorem 2.1 yields the following result:

Theorem 3.1. *The element $g^* \in C \cap W$ is a solution if and only if there exist: r points $t_1, \ldots, t_r \in T$ $(r \geqq 1)$ and r functions $f_1, \ldots, f_r \in F$ (not necessarily distinct) satisfying:*

$$\varepsilon_i \cdot \big(g^*(t_i) - f_i(t_i)\big) = p(g^* - f_i) = \max_{f \in F} p(g^* - f), \qquad i = 1, \ldots, r$$

s points $u_1, \ldots, u_s \in U$ $(s \geqq 0)$ satisfying:

$$g^*(u_i) = w(u_i), \qquad i = 1, \ldots, s$$

with $r + s \leqq n + 1$ and $r + s$ scalars $\lambda_1, \ldots, \lambda_r, \mu_1, \ldots, \mu_s > 0$ such that:

$$\sum_{i=1}^{r} \lambda_i \varepsilon_i g(t_i) + \sum_{i=1}^{s} \mu_i g(u_i) = 0, \qquad \forall g \in V.$$

Example 1.

$$S = [0, 3.75]; \qquad T = [0, 2.5]; \qquad U = [2.5, 3.75].$$

As compact set F we take the set of all functions

$$f_\sigma(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-t^2/2\sigma^2)$$

with $\sigma \in [1.249, 1.251]$.

$$W = V = \text{the polynomials of degree } 4 \ (n = 5).$$

$$w(u) = f_{1.25}(u).$$

We obtain the following numerical results:

$$g^*(t) = 0.318\,38 + 0.015\,787t - 0.153\,05t^2 + 0.055\,673t^3 - 0.005\,8609t^4.$$

$$d(g^*) = 0.001\,025$$

i	λ_i	ε_i	t_i	σ_i
1	0.17451	−1	0.00000	1.249
2	0.32342	+1	0.23047	1.251
3	0.24856	−1	0.85566	1.249
4	0.17115	+1	1.7892	1.249
5	0.08236	−1	2.5000	1.251

i	μ_i	u_i
1	0.01086	3.4811

Remarks.

1) Put:

$$f_1(t) = \max_{f \in F} f(t); \qquad f_2(t) = \min_{f \in F} f(t).$$

One can prove that f_1 and f_2 are continuous. Hence we have:

$$d(g) = \max\left(p(g - f_1);\ p(g - f_2)\right)$$

and the problem is reduced to the approximation of two functions. Thus the introduction of a compact set of functions f does not really improve the generality in this particular case. (It will be different for other spaces.)

2) If S is an interval and if V has the Haar property, then the ε_i are alternatively $+1$ and -1 (for increasing abscissae) and we obtain the classical alternant theorem.

2°) *Approximation in the Space of Differentiable Functions*

Let $E = C_1[a, b]$ be the space of real differentiable functions on $[a, b]$ with the norm:

$$p(f) = \|f\|' = \max_{t \in [a,b]} |f(t)| + \nu \max_{t \in [a,b]} |f'(t)|; \qquad \nu > 0.$$

Let V be an n-dimensional subspace of $C_1[a, b]$ (we take $G = V$) and F a compact subset of $C_1[a, b]$.

We will now apply Theorem 2.2:

A continuous linear functional l on $C_1[a, b]$ is an extremal point of the unit sphere S' of $C_1[a, b]'$ if and only if it has the following form:

$$l: f \in C_1[a, b] \to (l, f) = \varepsilon \cdot f(p) + \nu \cdot \varepsilon' \cdot f'(q)$$

with $\varepsilon, \varepsilon' = \pm 1$ and $p, q \in [a, b]$; (cf. [6]).

We then obtain the following characterization theorem:

Theorem 3.2. *An element $g^* \in V$ is a solution if and only if there exist k distinct pairs of points $(p_1, q_1), \ldots, (p_k, q_k)$ of $[a, b]$ $(1 \leqq k \leqq n+1)$, k functions $f_1, \ldots, f_k \in F$*

(not necessarily distinct) and scalars $\varepsilon_i = \pm 1$, $\varepsilon_i' = \pm 1$, $\lambda_i > 0$, $i = 1, \ldots, k$ *such that*

1) $\varepsilon_i \cdot e_i(p_i) + \nu \cdot \varepsilon_i' \cdot e_i'(q_i) = \|e_i\|' = \max_{f \in F} \|g^* - f\|'$ *with* $e_i = g^* - f_i$ $i = 1, \ldots, k$.

2) $$\sum_{i=1}^{k} \lambda_i \left(\varepsilon_i h(p_i) + \nu \cdot \varepsilon_i' \cdot h'(q_i)\right) = 0, \quad \forall h \in V.$$

Remark. The condition 1) of the theorem implies that:

$$\varepsilon_i e_i(p_i) = \max_{t \in [a,b]} |e_i(t)| \quad \text{and} \quad \varepsilon_i' \cdot e_i'(q_i) = \max_{t \in [a,b]} |e_i'(t)|.$$

Example 2.

$$[a, b] = [0, 5]; \quad \nu = 0.5.$$

As compact F we take the set of all functions:

$$f_{m,\sigma,a}(t) = 0.25\, a \exp(-a t) + 0.75 \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(t-m)^2/2\sigma^2\right)$$

with $m \in [2.95, 3.05]$; $\sigma \in [0.73, 0.77]$; $a \in [0.78, 0.82]$. $V =$ the polynomials of degree 4 ($n = 5$). We obtain the following results:

$$g^*(t) = 0.155\,82 - 0.356\,14t + 0.382\,70t^2 - 0.106\,56t^3 + 0.008\,5669t^4$$

$$d(g^*) = 0.183\,11.$$

i	λ_i	ε_i	p_i	ε_i'	q_i	m_i	σ_i	a_i
1	0.40380	−1	1.8049	+1	2.4687	3.05	0.73	0.82
2	0.20205	+1	0.3478	−1	3.5234	2.95	0.73	0.82
3	0.09620	−1	4.1682	+1	5.0000	2.95	0.73	0.82
4	0.09243	+1	2.9221	−1	3.5219	2.95	0.73	0.78
5	0.19903	+1	2.9221	+1	0.0000	2.95	0.73	0.78
6	0.00648	+1	0.3478	+1	5.0000	2.95	0.73	0.82

3°) *Approximation of Integrable Functions*

Let (S, Σ, μ) be a positive σ-finite measure space [2] and let $E = L_1 = L_1(S, \Sigma, \mu)$ be the space of measurable scalar functions f on S for which the norm

$$\|f\|_1 = \int_S |f(s)|\, \mu(ds)$$

is finite. Let V be an n-dimensional subspace of L_1 (we take $G = V$) and F a compact subset of L_1.

Let $L_\infty = L_\infty(S, \Sigma, \mu)$ be the space of measurable scalar functions φ on S which are essentially bounded with the norm:

$$\|f\|_\infty = \mu \cdot \operatorname*{ess.\,sup}_{s \in S} |f(s)|.$$

Every continuous linear functional l on L_1 is represented by an element $\varphi \in L_\infty$:

$$(l, f) = \int_S \varphi(s) f(s)\, \mu(ds).$$

We have the following property:

Lemma 3.1 ([11]). *A function $\varphi \in L_\infty$ is an extremal point of the unit sphere S' of L_∞ if and only if it takes μ-almost everywhere the values ± 1.*

Theorem 3.3. *An element $g^* \in V$ is a solution if and only if there exist k functions $\varphi_1, \ldots, \varphi_k \in L_\infty$ $(1 \leqq k \leqq n+1)$ which take μ-a.e. the values ± 1, k functions $f_1, \ldots, f_k \in F$ (not necessarily distinct) and scalars $\lambda_1, \ldots, \lambda_k > 0$ such that:*

1) $\int_S \varphi_i(s) \cdot e_i(s)\, \mu(ds) = \|e_i\|_1 = \max_{f \in F} \|g^* - f\|_1$ *with* $e_i = g^* - f_i$, $i = 1, \ldots, k$.

2) $$\sum_{i=1}^{k} \lambda_i \int_S \varphi_i(s) \cdot h(s)\, \mu(ds) = 0, \qquad \forall h \in V.$$

Remark. The condition:

$$\int_S \varphi_i(s) \cdot e_i(s)\, \mu(ds) = \|e_i\|_1$$

implies that $\varphi_i(s) = \operatorname{sign}(e_i(s))$ μ-a.e. on the set $S \sim Z(e_i)$, where $Z(e_i)$ denotes the set of zeros of e_i.

Example 3.

$$S = [0, 1]$$

$$\|f\|_1 = \int_0^1 |f(s)|\, ds.$$

As compact F we take the set of all functions:

$$f_{\alpha\beta}(t) = \alpha \exp(\beta t)$$

with $\alpha, \beta \in [0.999, 1.0001]$.

$$V = \text{the polynomials of degree } 3 \quad (n = 4).$$

We obtain the following numerical results:

$$g^*(t) = 0.999\,03 + 1.0168t + 0.426\,73t^2 + 0.274\,25t^3$$

$$d(g^*) = 0.000\,3437$$

i	ϱ_i	zeros of $e_i = g^* - f_i$				sign $(e_i(0))$	α_i	β_i
1	0.54846	0.0701	0.3827	0.5472	0.9533	−1	0.9999	0.9999
2	0.45153	0.1083	0.2535	0.7544	0.8437	−1	1.0001	1.0001

By the preceding remark the zeros of $e_i = g^* - f_i$ and $\operatorname{sign}(e_i(0))$ define the function φ_i.

The authors thank Mr. F. Crary (Math. Res. Center, Madison) for his efficient help in the obtention of the numerical results.

References

1. Bourbaki, N.: Eléments de Mathématiques. Espaces vectoriels topologiques. Ch. 1 et 2, fascicule XV. Paris: Hermann 1966.
2. Dunford, N., Schwartz, J. T.: Linear operators, part I. New York: Int. Pub. 1958.
3. Dunham, C. B.: Simultaneous Chebyshev approximation of functions on an interval. Proc. Am. Math. Soc. **18**, No. 3, 472–477 (1967).

4. Eggleston, L. G.: Convexity. Camb. Univ. Press 1963.
5. Golomb, M.: On the uniformly best approximation of functions given by incomplete data. M. R. C. Technical Summary Report 121, Dec. 1959, Madison, The University of Wisconsin.
6. Laurent, P. J.: Théorèmes de caractérisation d'une meilleure approximation dans un espace normé et généralisation de l'algorithme de Rémès. Num. Math. **10**, 190—208 (1967).
7. — Cônes de déplacements admissibles, sous-gradients et approximation convexe dans un espace vectoriel normé. Nato Summer School, Venice, June 1968.
8. — Théorèmes de caractérisation en approximation convexe. Colloque «Théorie de l'approximation des fonctions» Cluj (Roumanie), 15—20 Sept. 1967. Mathematica **10** (33), 95—111 (1968).
9. — Cours de théorie de l'approximation, fascicule 5, 1968—1969, Université de Grenoble.
10. Moreau, J. J.: Fonctionnelles convexes. Séminaire du Collège de France, Paris, 1966—1967.
11. Pham-Dinh-Tuan: Approximation d'un ensemble compact de fonctions dans un espace vectoriel normé. Projet de fin d'étude, Institut Polytechnique de Grenoble, Mathématiques appliquées, 1967—1968.
12. Rémès, E.: Sur la détermination des polynomes d'approximation de degré donné. Comm. de la Soc. Math. de Kharkof, sér. 4, **10**, 41—63 (1934).
13. Singer, I.: Caractérisation des éléments de meilleure approximation dans un espace de Banach quelconque. Act. Sci. Math. Szeged **17**, 181—189 (1956).
14. — Cea mai buna aproximare in spatii vectoriale normate prin elemente din subspatii vectoriale. Ed. Acad. Rep. Soc. Romania, Bucuresti (1967).
15. Valadier, M.: Quelques propriétés des sous-gradients. Rapport de l'IRIA, Sept. 1968.

P. J. Laurent
Math. Res. Center
The University of Wisconsin
Madison, Wis., 53706 (USA)

Pham-Dinh-Tuan
Math. Appliquées, CEDEX 53
Université de Grenoble
38-Grenoble-Gare (France)

Numer. Math. 15, 151–164 (1970)

An Implicit One-Step Method of High-Order Accuracy for the Numerical Integration of Ordinary Differential Equations

MINORU URABE

Dedicated to Professor Dr. Dr. h.c. L. Collatz for his 60th birthday

Received January 9, 1970

Abstract. For a differential equation $dx/dt=f(t, x)$ with $f_t(t, x)$, $f_x(t, x)$ computable, the author presents a new one-step method of high-order accuracy. A rule of controlling the mesh size is given and the method is compared with the Runge-Kutta method in two numerical examples.

1. Introduction

Let

$$\frac{dx}{dt}=f(t, x) \tag{1.1}$$

be a differential equation such that the function

$$g(t, x)=f_t(t, x)+f_x(t, x)\, f(t, x) \tag{1.2}$$

can be evaluated on a computer. In the present paper we shall present a new one-step method of high-order accuracy for solving differential equations having the above property.

Let

$$t_i=t_0+ih \qquad (i=0, 1, 2), \tag{1.3}$$

$$f_i=f(t_i, x_i), \quad g_i=g(t_i, x_i) \qquad (i=0, 1), \tag{1.4}$$

then our integration formula reads as follows:

$$\begin{aligned} x_1&=L_1(x_0, f_0, f_1, \hat{f}_2, g_0, g_1, \hat{g}_2)\\ &\triangleq x_0+\frac{h}{240}(101f_0+128f_1+11\hat{f}_2)\\ &\quad+\frac{h^2}{240}(13g_0-40g_1-3\hat{g}_2), \end{aligned} \tag{1.5}$$

where

$$\hat{f}_2=f(t_2, \hat{x}_2), \qquad \hat{g}_2=g(t_2, \hat{x}_2) \tag{1.6}$$

and

$$\begin{aligned}\hat{x}_2 &= L_2(x_0, x_1, f_0, f_1, g_0, g_1)\\ &\triangleq -31x_0 + 32x_1 - h(14f_0 + 16f_1) + h^2(-2g_0 + 4g_1).\end{aligned} \tag{1.7}$$

Suppose that (t_0, x_0) is known. Then by means of a well-known iterative method, one can solve Eq. (1.5) numerically with respect to x_1, if necessary, using the formula

$$x_1 = x_0 + hf_0 + \frac{h^2}{2} g_0 \tag{1.8}$$

for starting the iterative process. The value of x_1 found gives a desired approximation to $x(t_1)$, where $x(t)$ is a solution of Eq. (1.1) such that $x(t_0) \approx x_0$. The value of $\hat{x}_2$ obtained in the course of solution of Eq. (1.5) gives an approximation to $x(t_2)$, but it is not adopted in our method as a final approximation to $x(t_2)$ since it is inferior in the accuracy. Once the value of x_1 has been found, one repeats the above process replacing t_1, x_1, f_1 and g_1 by t_0, x_0, f_0 and g_0 respectively, and so on. When one proceeds from any step to the next one, the value of $\hat{x}_2$ found in the first step gives an approximation to x_1 in the second step if the mesh size h is not changed. In such a case, clearly the formula (1.8) is unnecessary for starting the iterative process in the second step.

In the numerical solution of Eq. (1.5), solution x_1 should be computed as accurately as possible. As will be shown in **4**, from this requirement follows a rule of controlling the mesh size h.

In the present paper, the derivation of the integration formula (1.5) will be given in **2** and the convergence of our one-step method will be proved in **3**. A rule of controlling the mesh size h will be given in **4** and two numerical examples will be presented in **5**.

Computations in the present paper have been carried out by the use of TOSBAC 3400 at Research Institute for Mathematical Sciences, Kyoto University. The author wishes to acknowledge the assistance of Mrs. S. Asako who has written the program.

2. The Derivation of the Integration Formula

Let $x(t)$ be a solution of (1.1) and put

$$t_i = t_0 + ih \quad (i = 0, 1, 2), \tag{2.1}$$

$$x_i = x(t_i) \quad (i = 0, 1, 2). \tag{2.2}$$

If $x(t)$ is analytic in t, then we have the equality of the following form:

$$\begin{aligned}\alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 &= h(\beta_0 \dot{x}_0 + \beta_1 \dot{x}_1 + \beta_2 \dot{x}_2)\\ &\quad + h^2(\gamma_0 \ddot{x}_0 + \gamma_1 \ddot{x}_1 + \gamma_2 \ddot{x}_2) + T \quad (\cdot = d/dt),\end{aligned} \tag{2.3}$$

where T is of the form

$$T = C h^p \left(\frac{d}{dt}\right)^p x(t_0) + \cdots. \tag{2.4}$$

In order to determine the coefficients $\alpha_i, \beta_i, \gamma_i$ $(i=0, 1, 2)$, C and the positive integer p, we consider the following polynomials:

$$\begin{aligned} \varrho(\lambda) &= \alpha_0 + \alpha_1 \lambda + \alpha_2 \lambda^2, \\ \sigma(\lambda) &= \beta_0 + \beta_1 \lambda + \beta_2 \lambda^2, \\ \tau(\lambda) &= \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2. \end{aligned} \tag{2.5}$$

Let E and D be the operators such that

$$E\,x(t) = x(t+h), \qquad D\,x(t) = \dot{x}(t). \tag{2.6}$$

Then we can write equality (2.3) in the following form:

$$[\varrho(E) - h\sigma(E)\,D - h^2\tau(E)\,D^2]\,x(t_0) = C\,h^p\,D^p\,x(t_0) + \cdots.$$

Since

$$E\,x(t) = x(t+h) = e^{hD}\,x(t),$$

we then have formally

$$\varrho(e^{hD}) - \sigma(e^{hD})\,hD - \tau(e^{hD})\,h^2D^2 = C\,h^pD^p + \cdots, \tag{2.7}$$

which is equivalent formally to

$$\varrho(\zeta) - \sigma(\zeta)\log\zeta - \tau(\zeta)\log^2\zeta = C\log^p\zeta + \cdots. \tag{2.8}$$

In (2.8), clearly $\zeta \to 1$ and $\log\zeta \to 0$ as $h \to 0$. Hence in the limit we have

$$\varrho(1) = 0. \tag{2.9}$$

Put

$$\zeta = 1+\xi, \tag{2.10}$$

then clearly $\xi \to 0$ as $h \to 0$, and from (2.8) we have

$$\frac{\varrho(1+\xi)}{\log(1+\xi)} - \sigma(1+\xi) - \tau(1+\xi)\cdot\log(1+\xi) = C\log^{p-1}(1+\xi) + \cdots. \tag{2.11}$$

By (2.9), $\varrho(1+\xi)$ is of the form

$$\varrho(1+\xi) = \xi(\varrho_0 + \varrho_1\xi). \tag{2.12}$$

Clearly $\sigma(1+\xi)$ and $\tau(1+\xi)$ are of the forms

$$\begin{aligned} \sigma(1+\xi) &= \sigma_0 + \sigma_1\xi + \sigma_2\xi^2, \\ \tau(1+\xi) &= \tau_0 + \tau_1\xi + \tau_2\xi^2. \end{aligned} \tag{2.13}$$

Now it is clear that

$$\frac{\xi}{\log(1+\xi)} = \int_0^1 (1+\xi)^u\,du = \sum_{r=0}^{\infty} c_r\xi^r \tag{2.14}$$

where

$$c_r = \int_0^1 \frac{u(u-1)\cdots(u-r+1)}{r!}\,du \quad (r=0, 1, 2, \ldots). \tag{2.15}$$

By elementary calculations, it is easily seen that

$$c_0=1,\quad c_1=\frac{1}{2},\quad c_2=-\frac{1}{12},\quad c_3=\frac{1}{24},$$
$$c_4=-\frac{19}{720},\quad c_5=\frac{3}{160},\quad c_6=-\frac{863}{60,480}. \tag{2.16}$$

Now

$$\log(1+\xi)=\xi-\frac{1}{2}\xi^2+\frac{1}{3}\xi^3-\cdots,$$

hence we can write (2.11) in the following form:

$$\begin{aligned}&(\varrho_0+\varrho_1\xi)(1+c_1\xi+c_2\xi^2+\cdots)\\&\quad-(\sigma_0+\sigma_1\xi+\sigma_2\xi^2)\\&\qquad-(\tau_0+\tau_1\xi+\tau_2\xi^2)\left(\xi-\frac{1}{2}\xi^2+\frac{1}{3}\xi^3-\cdots\right)\\&\qquad=C\xi^{p-1}+\cdots.\end{aligned} \tag{2.17}$$

Let us consider the two cases.

Case I. $\varrho(\lambda)=-1+\lambda$. In this case

$$\varrho(1+\xi)=-1+(1+\xi)=\xi,$$

therefore by (2.12) we see that

$$\varrho_0=1,\quad \varrho_1=0.$$

Then from (2.17) it follows that

$$p\geqq 7 \tag{2.18}$$

if and only if the following equations are satisfied by σ_i and τ_i $(i=0, 1, 2)$:

$$\begin{aligned}&1-\sigma_0=0,\\&c_1-\sigma_1-\tau_0=0,\\&c_2-\sigma_2-\left(-\frac{1}{2}\tau_0+\tau_1\right)=0,\\&c_3-\left(\frac{1}{3}\tau_0-\frac{1}{2}\tau_1+\tau_2\right)=0,\\&c_4-\left(-\frac{1}{4}\tau_0+\frac{1}{3}\tau_1-\frac{1}{2}\tau_2\right)=0,\\&c_5-\left(\frac{1}{5}\tau_0-\frac{1}{4}\tau_1+\frac{1}{3}\tau_2\right)=0.\end{aligned} \tag{2.19}$$

Solving (2.19) with respect to σ_i and τ_i, we find that σ_i and τ_i satisfying (2.19) are

$$\sigma_0=1,\quad \sigma_1=\frac{5}{8},\quad \sigma_2=\frac{11}{240},$$
$$\tau_0=-\frac{1}{8},\quad \tau_1=-\frac{23}{120},\quad \tau_2=-\frac{1}{80}. \tag{2.20}$$

For these values of σ_i and τ_i, from (2.17) we have

$$p=7 \tag{2.21}$$

and

$$C = c_6 - \left(-\frac{1}{6}\tau_0 + \frac{1}{5}\tau_1 - \frac{1}{4}\tau_2\right) = \frac{1}{9{,}450}. \tag{2.22}$$

Since

$$\sigma(\lambda) = \sigma_0 + \sigma_1(\lambda - 1) + \sigma_2(\lambda - 1)^2,$$
$$\tau(\lambda) = \tau_0 + \tau_1(\lambda - 1) + \tau_2(\lambda - 1)^2$$

by (2.13), for the values given by (2.20) we have

$$\begin{aligned} \sigma(\lambda) &= \frac{1}{240}(101 + 128\lambda + 11\lambda^2), \\ \tau(\lambda) &= \frac{1}{240}(13 - 40\lambda - 3\lambda^2), \end{aligned} \tag{2.23}$$

which by (2.5) means

$$(\beta_0, \beta_1, \beta_2) = \frac{1}{240}(101, 128, 11),$$

$$(\gamma_0, \gamma_1, \gamma_2) = \frac{1}{240}(13, -40, -3).$$

Now for solution $x(t)$ of (1.1), it is evident that

$$\dot{x}(t) = f[t, x(t)],$$
$$\ddot{x}(t) = g[t, x(t)].$$

Hence from (2.3) we have

$$\begin{aligned} x_1 = x_0 &+ \frac{h}{240}[101 f(t_0, x_0) + 128 f(t_1, x_1) + 11 f(t_2, x_2)] \\ &+ \frac{h^2}{240}[13 g(t_0, x_0) - 40 g(t_1, x_1) - 3 g(t_2, x_2)] \\ &+ \frac{1}{9{,}450} h^7 D^7 x(t_0) + \cdots, \end{aligned} \tag{2.24}$$

from which one obtains the integration formula (1.5) by neglecting the residual terms and replacing x_2 by $\hat{x}_2$.

Case II. $\varrho(\lambda) = \alpha_0 + \alpha_1\lambda + \lambda^2$, $\beta_2 = \gamma_2 = 0$. By (2.9),

$$\alpha_1 = -(1 + \alpha_0),$$

therefore we have

$$\varrho(1 + \xi) = \xi[(1 - \alpha_0) + \xi],$$

which implies

$$\varrho_0 = 1 - \alpha_0, \quad \varrho_1 = 1 \tag{2.25}$$

in (2.12). Since $\beta_2 = \gamma_2 = 0$, we have

$$\sigma_2 = \tau_2 = 0 \tag{2.26}$$

in (2.13). Then corresponding to (2.17), we have:

$$\begin{aligned} &(\varrho_0 + \xi)(1 + c_1\xi + c_2\xi^2 + \cdots) \\ &\quad - (\sigma_0 + \sigma_1\xi) \\ &\quad - (\tau_0 + \tau_1\xi)\left(\xi - \frac{1}{2}\xi^2 + \frac{1}{3}\xi^3 - \cdots\right) \\ &\quad = C\xi^{p-1} + \cdots. \end{aligned} \tag{2.27}$$

Similarly to Case I, we then have:

$$p=6, \tag{2.28}$$

$$\begin{aligned} &\varrho_0=-30, \\ &\sigma_0=-30, \quad \sigma_1=-16, \\ &\tau_0=2, \quad \tau_1=4, \end{aligned} \tag{2.29}$$

$$C=\frac{1}{90}. \tag{2.30}$$

From (2.29) readily follows

$$\begin{aligned} &\varrho(\lambda)=31-32\lambda+\lambda^2, \\ &\sigma(\lambda)=-14-16\lambda, \\ &\tau(\lambda)=-\ 2+\ 4\lambda. \end{aligned} \tag{2.31}$$

Thus by (2.28) and (2.30), from (2.3) we have

$$\begin{aligned} x_2=&-31x_0+32x_1 \\ &-h[14f(t_0, x_0)+16f(t_1, x_1)] \\ &+h^2[-2g(t_0, x_0)+4g(t_1, x_1)] \\ &+\frac{1}{90}h^6D^6x(t_0)+\cdots, \end{aligned} \tag{2.32}$$

from which one obtains the formula (1.7) by neglecting the residual terms and replacing x_2 by $\hat{x}_2$.

3. Convergence of the Method

Let $x(t)$ be a solution of (1.1) and suppose that

$$\begin{aligned} &t_{n+1}=t_n+h_n, \\ &\tilde{x}_n=x(t_n) \qquad (n=0, 1, 2, \ldots). \end{aligned} \tag{3.1}$$

Then by (2.24) and (2.32), we have the equalities of the following forms:

$$\begin{aligned} \tilde{x}_{n+1}=&\tilde{x}_n+h_n[\beta_0 f(t_n, \tilde{x}_n)+\beta_1 f(t_{n+1}, \tilde{x}_{n+1})+\beta_2 f(t_{n+2}, \tilde{x}_{n+2})] \\ &+h_n^2[\gamma_0 g(t_n, \tilde{x}_n)+\gamma_1 g(t_{n+1}, \tilde{x}_{n+1})+\gamma_2 g(t_{n+2}, \tilde{x}_{n+2})] \\ &+T_n, \end{aligned} \tag{3.2}$$

$$\begin{aligned} \tilde{x}_{n+2}=&-31\tilde{x}_n+32\tilde{x}_{n+1} \\ &+h_n[\beta_0' f(t_n, \tilde{x}_n)+\beta_1' f(t_{n+1}, \tilde{x}_{n+1})] \\ &+h_n^2[\gamma_0' g(t_n, \tilde{x}_n)+\gamma_1' g(t_{n+1}, \tilde{x}_{n+1})] \\ &+T_n', \end{aligned} \tag{3.3}$$

where

$$T_n=\frac{1}{9{,}450}h_n^7D^7x(t_n)+\cdots, \tag{3.4}$$

$$T_n'=\frac{1}{90}h^6D^6x(t_n)+\cdots. \tag{3.5}$$

Let x_n be approximations to $\tilde{x}_n = x(t_n)$ obtained by our numerical method. Then by (1.5) and (1.7), corresponding to (3.2) and (3.3), we have

$$\begin{aligned} x_{n+1} = x_n &+ h_n[\beta_0 f(t_n, x_n) + \beta_1 f(t_{n+1}, x_{n+1}) + \beta_2 f(t_{n+2}, \hat{x}_{n+2})] \\ &+ h_n^2[\gamma_0 g(t_n, x_n) + \gamma_1 g(t_{n+1}, x_{n+1}) + \gamma_2 g(t_{n+2}, \hat{x}_{n+2})] \\ &+ R_n, \end{aligned} \tag{3.6}$$

$$\begin{aligned} \hat{x}_{n+2} = &-31 x_n + 32 x_{n+1} \\ &+ h_n[\beta_0' f(t_n, x_n) + \beta_1' f(t_{n+1}, x_{n+1})] \\ &+ h_n^2[\gamma_0' g(t_n, x_n) + \gamma_1' g(t_{n+1}, x_{n+1})], \end{aligned} \tag{3.7}$$

where R_n is a round-off error arising in computation of x_{n+1}.

We assume that functions $f(t, x)$ and $g(t, x)$ satisfy a Lipschitz condition with respect to x, that is, there are positive constants K_1 and K_2 such that

$$\begin{aligned} |f(t, x') - f(t, x'')| &\leqq K_1 |x' - x''|, \\ |g(t, x') - g(t, x'')| &\leqq K_2 |x' - x''| \end{aligned} \tag{3.8}$$

for any x' and x''.

Put

$$x_n - \tilde{x}_n = e_n \quad (n = 0, 1, 2, \ldots), \tag{3.9}$$

then subtracting (3.2) from (3.6) and making use of (3.8), we have

$$\begin{aligned} |e_{n+1}| \leqq |e_n| &+ |h_n| K_1(|\beta_0| \cdot |e_n| + |\beta_1| \cdot |e_{n+1}| \\ &+ |\beta_2| \cdot |\hat{x}_{n+2} - \tilde{x}_{n+2}|) + |h_n|^2 K_2(|\gamma_0| \cdot |e_n| \\ &+ |\gamma_1| \cdot |e_{n+1}| + |\gamma_2| \cdot |\hat{x}_{n+2} - \tilde{x}_{n+2}|) + |R_n| + |T_n|. \end{aligned} \tag{3.10}$$

However subtracting (3.3) from (3.7) and making use of (3.8), we have

$$\begin{aligned} |\hat{x}_{n+2} - \tilde{x}_{n+2}| \leqq &31 |e_n| + 32 |e_{n+1}| \\ &+ |h_n| K_1(|\beta_0'| \cdot |e_n| + |\beta_1'| \cdot |e_{n+1}|) \\ &+ |h_n|^2 K_2(|\gamma_0'| \cdot |e_n| + |\gamma_1'| \cdot |e_{n+1}|) \\ &+ |T_n'|. \end{aligned} \tag{3.11}$$

Hence substituting (3.11) into (3.10), we have

$$\begin{aligned} |e_{n+1}| \leqq |e_n| &+ |h_n| K_1(|\beta_0| \cdot |e_n| + |\beta_1| \cdot |e_{n+1}|) \\ &+ |h_n|^2 K_2(|\gamma_0| \cdot |e_n| + |\gamma_1| \cdot |e_{n+1}|) \\ &+ (|h_n| K_1 |\beta_2| + |h_n|^2 K_2 |\gamma_2|)[31 |e_n| + 32 |e_{n+1}| \\ &\quad + |h_n| K_1(|\beta_0'| \cdot |e_n| + |\beta_1'| \cdot |e_{n+1}|) \\ &\quad + |h_n|^2 K_2(|\gamma_0'| \cdot |e_n| + |\gamma_1'| \cdot |e_{n+1}|) + |T_n'|] \\ &+ |R_n| + |T_n|. \end{aligned}$$

Then we have

$$\begin{aligned}(3.12)\qquad &[1-|h_n|K_1(|\beta_1|+|\beta_2|M_n)-|h_n|^2K_2(|\gamma_1|+|\gamma_2|M_n)]|e_{n+1}|\\ &\leqq[1+|h_n|K_1(|\beta_0|+|\beta_2|N_n)+|h_n|^2K_2(|\gamma_0|+|\gamma_2|N_n)]|e_n|+\varrho_n,\end{aligned}$$

where

$$\begin{aligned}(3.13)\qquad M_n&=32+|h_n|K_1|\beta_1'|+|h_n|^2K_2|\gamma_1'|,\\ N_n&=31+|h_n|K_1|\beta_0'|+|h_n|^2K_2|\gamma_0'|,\\ \varrho_n&=|R_n|+|T_n|+|h_n|(K_1|\beta_2|+|h_n|K_2|\gamma_2|)|T_n'|.\end{aligned}$$

Suppose that

$$(3.14)\qquad |h_n|\leqq H\quad(n=0,1,2,\ldots).$$

Then for small $H>0$, by (3.4) and (3.5), we may suppose that

$$(3.15)\qquad |T_n|\leqq C_0|h_n|^7,\quad |T_n'|\leqq C_1|h_n|^6.$$

For R_n, let us suppose that

$$(3.16)\qquad |R_n|\leqq\varepsilon(|h_n|)\cdot|h_n|\quad(n=0,1,2,\ldots)$$

where $\varepsilon(h)$ is a monotonously increasing function such that $\varepsilon(h)\downarrow 0$ as $h\downarrow 0$. By (3.15) and (3.16), we then have

$$(3.17)\qquad \varrho_n\leqq\varrho(|h_n|)\cdot|h_n|,$$

where $\varrho(h)$ is a monotonously increasing function such that $\varrho(h)\downarrow 0$ as $h\downarrow 0$. Now from (3.13), we have

$$\begin{aligned}M_n&\leqq 32+H(K_1|\beta_1'|+HK_2|\gamma_1'|),\\ N_n&\leqq 31+H(K_1|\beta_0'|+HK_2|\gamma_0'|)\end{aligned}\qquad(n=0,1,2,\ldots),$$

therefore by (3.14) and (3.17), from (3.12) we have the inequality of the following form:

$$(1-|h_n|K_3)|e_{n+1}|\leqq(1+|h_n|K_4)|e_n|+\varrho(H)|h_n|,$$

that is,

$$(3.18)\qquad |e_{n+1}|\leqq\frac{1+|h_n|K_4}{1-|h_n|K_3}|e_n|+\hat{\varrho}|h_n|\quad(n=0,1,2,\ldots),$$

where K_3 and K_4 are positive constants and

$$(3.19)\qquad \hat{\varrho}=\frac{\varrho(H)}{1-HK_3}.$$

By elementary manipulations, it is easily seen that

$$\frac{1+|h_n|K_4}{1-|h_n|K_3}\leqq 1+|h_n|K_5\quad(n=0,1,2,\ldots)$$

where

$$K_5=\frac{K_3+K_4}{1-HK_3}.$$

Hence from (3.18), we have

$$|e_{n+1}| \leqq (1+|h_n| K_5)\,|e_n| + \hat{\varrho}\,|h_n| \quad (n=0, 1, 2, \ldots), \tag{3.20}$$

from which by induction readily follows

$$|e_n| \leqq \prod_{i=0}^{n-1} (1+|h_i| K_5) \cdot |e_0| + \hat{\varrho} \sum_{j=0}^{n-1} \left[|h_j| \prod_{i=j+1}^{n-1} (1+|h_i| K_5) \right] \quad (n=0, 1, 2, \ldots). \tag{3.21}$$

However

$$\begin{aligned} &\sum_{j=0}^{n-1} \left[|h_j| \prod_{i=j+1}^{n-1} (1+|h_i| K_5) \right] \\ &= \frac{1}{K_5} \sum_{j=0}^{n-1} \left[\{(1+|h_j| K_5) - 1\} \prod_{i=j+1}^{n-1} (1+|h_i| K_5) \right] \\ &= \frac{1}{K_5} \sum_{j=0}^{n-1} \left[\prod_{i=j}^{n-1} (1+|h_i| K_5) - \prod_{i=j+1}^{n-1} (1+|h_i| K_5) \right] \\ &= \frac{1}{K_5} \left[\prod_{i=0}^{n-1} (1+|h_i| K_5) - 1 \right]. \end{aligned}$$

Hence from (3.21) we have

$$\begin{aligned} |e_n| &\leqq \prod_{i=0}^{n-1} (1+|h_i| K_5) \cdot |e_0| \\ &\quad + \frac{\hat{\varrho}}{K_5} \left[\prod_{i=0}^{n-1} (1+|h_i| K_5) - 1 \right] \quad (n=0, 1, 2, \ldots). \end{aligned} \tag{3.22}$$

Let l be the length of the interval I on which the numerical integration of (1.1) is carried out. Then

$$\sum_{i=0}^{n-1} |h_i| \leqq l,$$

therefore we have

$$\begin{aligned} \prod_{i=0}^{n-1} (1+|h_i| K_5) &\leqq \prod_{i=0}^{n-1} \exp[|h_i| K_5] \\ &= \exp\left[\sum_{i=0}^{n-1} |h_i| \cdot K_5 \right] \\ &\leqq \exp(l K_5). \end{aligned}$$

Then from (3.22) follows

$$|e_n| \leqq |e_0| \cdot \exp(l K_5) + \frac{\hat{\varrho}}{K_5} [\exp(l K_5) - 1] \quad (n=0, 1, 2, \ldots), \tag{3.23}$$

which by (3.19) implies

$$|e_n| \to 0 \quad (n=0, 1, 2, \ldots) \quad \text{as} \quad |e_0|, H \to 0.$$

This proves the convergence of our integration method.

Remark 1. For the proof of the convergence of an integration method, it is easily seen that no generality is lost by the assumption (3.8).

Remark 2. By (3.4) and (3.5), from (3.13) we have

$$\begin{aligned}\varrho_n &= |R_n| + \frac{1}{9{,}450}|h_n|^7|D^7x(t_n)| + \frac{11}{240}\times\frac{1}{90}|h_n|^7\cdot K_1|D^6x(t_n)| + o(|h_n|^7)\\ &= |R_n| + \frac{1}{9{,}450}|h_n|^7\left[|D^7x(t_n)| + \frac{77}{16}K_1|D^6x(t_n)|\right] + o(|h_n|^7).\end{aligned}$$

Hence by (3.16) and (3.19), we have

$$\hat{\varrho} \leqq \varepsilon(H)[1+O(H)] + \frac{1}{9{,}450}H^6\left[\max|D^7x(t)| + \frac{77}{16}K_1\cdot\max|D^6x(t)|\right] + o(H^6), \tag{3.24}$$

by which from (3.23) we see that our integration method possesses the high accuracy.

4. A Rule of Controlling the Mesh Size

In the function in the right-hand side of (1.5), fix x_0 and substitute the right-hand side of (1.7) for $\hat{x}_2$. Let $\varphi(x_1)$ be a function obtained. Then Eq. (1.5) can be written briefly as

$$x_1 = \varphi(x_1) \tag{4.1}$$

and the iterative process for solving Eq. (1.5) can be written as

$$x_1^{(m+1)} = \varphi[x_1^{(m)}] \quad (m = 0, 1, 2, \ldots). \tag{4.2}$$

Now from (1.5) we have

$$\begin{aligned}\varphi'(x_1) = {} & \frac{8}{15}hf_x(t_1, x_1) + \frac{11}{240}hf_x(t_2, \hat{x}_2)\frac{d\hat{x}_2}{dx_1}\\ & - \frac{1}{6}h^2g_x(t_1, x_1) - \frac{1}{80}h^2g_x(t_2, \hat{x}_2)\frac{d\hat{x}_2}{dx_1},\end{aligned}$$

where $' = d/dx_1$. From (1.7) we have

$$\frac{d\hat{x}_2}{dx_1} = 32 - 16hf_x(t_1, x_1) + 4h^2g_x(t_1, x_1).$$

Hence we have

$$\begin{aligned}\varphi'(x_1) &= \frac{8}{15}hf_x(t_1, x_1) - \frac{1}{6}h^2g_x(t_1, x_1)\\ &\quad + h\left[\frac{11}{240}f_x(t_2, \hat{x}_2) - \frac{1}{80}hg_x(t_2, \hat{x}_2)\right]\\ &\quad \cdot[32 - 16hf_x(t_1, x_1) + 4h^2g_x(t_1, x_1)]\\ &= h\left(\frac{8}{15} + \frac{11}{240}\cdot 32\right)f_x(t_1, x_1) + O(h^2)\\ &= 2hf_x(t_0, x_0) + O(h^2).\end{aligned} \tag{4.3}$$

Taking a small positive number H, let us restrict the mesh size h so that

$$|h| \leqq H. \tag{4.4}$$

Then from (4.3) it follows that if

$$2|h| \cdot |f_x(t_0, x_0)| < k, \tag{4.5}$$

then

$$|\varphi(x_1') - \varphi(x_1'')| \leqq k|x_1' - x_1''| \tag{4.6}$$

for any x_1', x_1'' near x_0.

Suppose that h is chosen so that (4.5) may be valid for $k<1$. Then an approximate value to the solution x_1 of Eq. (4.1) can be indeed obtained by the iterative process (4.2) (for details, see [1]). Let ε be a bound of a round-off error in the evaluation of $\varphi(x)$ and take a positive number α such that

$$\alpha \geqq \frac{2\varepsilon}{1-k}. \tag{4.7}$$

Then as is proved in [2], we can stop the iterative process (4.2) by the criterion

$$|x_1^{(m+1)} - x_1^{(m)}| \leqq \alpha \tag{4.8}$$

and moreover, if k satisfies

$$k\alpha \leqq \varepsilon, \tag{4.9}$$

a final value $x_1^{(m+1)}$ obtained by the iterative process possesses the similar accuracy as the value in the state of oscillating numerical convergence, that is, the best approximate solution that one can get by the iterative process (4.2).

From the above results, we get a following method of controlling the mesh-size h.

Namely, taking into account inequalities (3.24), (4.7) and (4.9), we specify the numbers H, α and k before excution of computation. In the course of computation, we then determine the mesh size h by halving or doubling so that h may be a maximal mesh size satisfying both of inequalities (4.4) and (4.5).

5. Numerical Examples

Example 1.

$$\frac{dx}{dt} = \frac{1}{5}x^2, \; x(0) = 1.$$

In this example, it is evident that the exact solution is

$$x(t) = \frac{5}{5-t}.$$

Specifying H, k and α so that

$$H = 0.125, \quad k = 0.1, \quad \alpha = 10^{-8} \text{ and } 10^{-9},$$

we applied our method to the given equation. The values of the solutions obtained are shown in Table 1. The errors of the solutions multiplied by 10^9 are also shown in Table 1. In order to compare our method with the Runge-Kutta method, we computed the solution by both methods for a constant mesh size $h=0.0625$. In our method we specified α so that $\alpha=10^{-9}$. Errors multiplied by 10^9 for the solutions obtained by both methods are shown in Table 2.

Example 2.

$$\frac{dx}{dt}=5t\left(\frac{1}{2}-x\right)^{\frac{4}{5}}, \qquad x(-1)=\frac{15}{32}=0.46875.$$

In this example, it is easily seen that the exact solution is

$$x(t)=\frac{1}{2}-\left(1-\frac{1}{2}t^2\right)^5.$$

Specifying H, k and α so that

$$H=0.0625, \qquad k=0.1, \qquad \alpha=10^{-8} \quad \text{and} \quad 10^{-9},$$

we applied our method to the given equation. The values of the solutions obtained are shown in Table 3. The errors of the solutions multiplied by 10^{10} are also shown in Table 3. In order to compare our method with the Runge-Kutta method, we computed the solution by both methods for a constant mesh size $h=0.03125$. In our method we specified α so that $\alpha=10^{-9}$. Errors multiplied by 10^{10} for the solutions obtained by both methods are shown in Table 4.

Tables 1~4 show the practical usefulness of our method.

Table 1

h	t	Solutions (Errors $\times 10^9$)	
		$\alpha=10^{-8}$	$\alpha=10^{-9}$
	0	1.00000 0000	1.00000 0000
0.125	0.125	1.02564 1025 (−1)	1.02564 1026 (0)
0.0625	0.1875	1.03896 1038 (−1)	1.03896 1039 (0)
⋮	⋮	⋮	⋮
0.0625	2.5625	2.05128 2025 (−26)	2.05128 2047 (−4)
0.0312 5	2.5937 5	2.07792 2050 (−28)	2.07792 2074 (−4)
⋮	⋮	⋮	⋮
0.0312 5	3.7812 5	4.10256 3972 (−131)	4.10256 4081 (−22)
0.0156 25	3.7968 75	4.15584 4022 (−134)	4.15584 4134 (−22)
⋮	⋮	⋮	⋮
0.0156 25	4.3906 25	8.20512 7654 (−551)	8.20512 8110 (−95)
0.0078 125	4.3984 375	8.31168 7746 (−566)	8.31168 8214 (−98)
⋮	⋮	⋮	⋮
0.0078 125	4.6953 125	16.41025 417 (−2240)	16.41025 602 (−390)
0.0039 0625	4.6992 1875	16.62337 433 (−2290)	16.62337 622 (−400)
⋮	⋮	⋮	⋮
0.0039 0625	4.75	19.99999 667 (−3330)	19.99999 941 (−590)

Table 2

t	Errors $\times 10^9$	
	Our one-step method	Runge-Kutta method
0	0	0
0.5	0	0
1.0	−1	−1
1.5	−1	−1
2.0	−2	−4
2.5	−3	−10
3.0	−7	−32
3.5	−17	−138
4.0	−132	−1,051
4.5	−19,499	−33,167
4.5625	−55,860	−64,280
4.625	−194,950	−137,680
4.6875	−910,920	−337,710
4.75	−6,882,850	−1,005,330

Table 3

h	t	Solutions (Errors $\times 10^{10}$)	
		$\alpha = 10^{-8}$	$\alpha = 10^{-9}$
	−1.0	0.46875 00000	0.46875 00000
0.0039 0625	−0.99609 375	0.46751 25336 (3)	0.46751 25333 (0)
⋮	⋮	⋮	⋮
0.0039 0625	−0.91796 875	0.43511 52891 (49)	0.43511 52840 (−2)
0.0078 125	−0.91015 625	0.43101 16566 (55)	0.43101 16509 (−2)
⋮	⋮	⋮	⋮
0.0078 125	−0.63671 875	0.17782 33175 (398)	0.17782 32782 (5)
0.0156 25	−0.62109 375	0.15747 36963 (413)	0.15747 36555 (5)
⋮	⋮	⋮	⋮
0.0156 25	−0.37109 375	−0.19998 03843 (674)	−0.19998 04515 (2)
0.0312 5	−0.33984 375	−0.24274 17954 (703)	−0.24274 18654 (3)
⋮	⋮	⋮	⋮
0.0312 5	−0.18359 375	−0.41852 61361 (822)	−0.41852 62182 (1)
0.0625	−0.12109 375	−0.46387 43081 (843)	−0.46387 43934 (−10)
⋮	⋮	⋮	⋮
0.0625	0.25390 625	−0.34888 98249 (830)	−0.34888 99026 (53)
0.0312 5	0.28515 625	−0.31258 61465 (799)	−0.31258 62213 (51)
⋮	⋮	⋮	⋮
0.0312 5	0.37890 625	−0.18903 76135 (708)	−0.18903 76799 (44)
0.0156 25	0.39453 125	−0.16690 02795 (690)	−0.16690 03442 (43)
⋮	⋮	⋮	⋮
0.0156 25	0.64453 125	0.18780 92069 (373)	0.18780 91719 (23)
0.0078 125	0.65234 375	0.19766 31599 (362)	0.19766 31259 (22)
⋮	⋮	⋮	⋮
0.0078 125	0.92578 125	0.43905 38383 (99)	0.43905 38291 (7)
0.0039 0625	0.92968 75	0.44096 19448 (96)	0.44096 19359 (7)
⋮	⋮	⋮	⋮
0.0039 0625	1.0	0.46875 00053 (53)	0.46875 00004 (4)

Table 4

t	Errors $\times 10^{10}$	
	Our one-step method	Runge-Kutta method
−1	0	0
−0.75	58	39,602
−0.5	144	92,887
−0.25	216	140,186
0.0	243	159,239
0.25	214	140,193
0.5	141	92,918
0.75	58	39,237
1.0	3	− 3,849

References

1. Urabe, M.: Convergence of numerical iteration in solution of equations. J. Sci. Hiroshima Univ. Ser. A, **19**, 479−489 (1956).
2. — Error estimation in numerical solution of equations by iteration process. J. Sci. Hiroshima Univ. Ser. A−I, **26**, 77−91 (1962).

Minoru Urabe
Research Institute for
Mathematical Sciences
Kyoto University
Kitashirakawa
Kyoto, Japan

Numer. Math. 15, 165—174 (1970)

Konvergenz mehrdimensionaler Interpolation

H. Ehlich und W. Haussmann

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 15. Dezember 1969

Summary. This paper is concerned with convergence of interpolation polynomials in two variables. Some investigations have been made in the special cases of Lagrange and Fejér-Hermite interpolation. Here we give a general method to prove convergence of two-dimensional interpolation processes. The theory is based on the concept of tensor products of interpolation problems. Our main theorem shows that a two-dimensional interpolation process which is a tensor product of two one-dimensional processes converges for a function $h \in C[a, b] \otimes C[\bar{a}, \bar{b}]$ if both "factors" do for certain one-dimensional functions. In addition we prove how the order of convergence of the two-dimensional process depends on the one-dimensional processes.

The converse theorem is not true. But we can give a sufficient condition which depends on the density of the interpolation nodes to assure the converse theorem in some cases.

In an easy manner the convergence theorem may be extended to higher dimensions.

1. Einleitung

Während über Existenz, Eindeutigkeit und Darstellung mehrdimensionaler Interpolation zahlreiche Arbeiten vorliegen (Ahlin [1, 2], Haußmann [6], Milne et al. [8], Salzer [12—15], Salzer und Kimbro [16], Thacher [18, 19] sowie Thacher und Milne [20]), wurden Konvergenzuntersuchungen im Mehrdimensionalen nur in Spezialfällen vorgenommen: Sloss [17] behandelt die Konvergenz mehrdimensionaler Lagrange-Interpolation (allerdings unter starken Voraussetzungen über die zu interpolierende Funktion h), Mond und Shisha [9, 10] sowie Haußmann [7] die Fejér-Hermite- bzw. Hermite-Interpolation. Alle diese Konvergenzbetrachtungen hängen sehr stark vom jeweiligen Interpolationstyp ab.

Hier werden für eine Klasse Z von Funktionen mehrerer Veränderlicher Konvergenzaussagen über mehrdimensionale Interpolation gemacht, die unabhängig vom betreffenden Interpolationstyp gelten: Wenn nämlich gewisse (eindimensionale) Interpolationsprozesse konvergieren, so konvergiert auch der Prozeß, der durch das Tensorprodukt dieser Interpolationsprobleme gegeben ist (vgl. [6]).

Eine Umkehrung dieser Aussage stößt auf Schwierigkeiten, wie sich an einem einfachen Beispiel zeigen läßt. Wir werden jedoch im Fall der Lagrange-Interpolation hinreichende Bedingungen für eine Umkehrung in gewissen Fällen angeben. Diese Bedingungen sind nur wenig einschneidend und in vielen wichtigen Fällen erfüllt.

Bei allen Sätzen beschränken wir uns auf die zweidimensionale Interpolation. Die entsprechenden Aussagen für höhere Dimensionen lassen sich leicht gewinnen.

2. Existenz und Eindeutigkeit mehrdimensionaler Interpolation

Wir gehen aus von $C = C(I)$ mit $I = [a, b] \subset \mathbb{R}$ und betrachten einen $(m+1)$-dimensionalen Teilraum $U \subset C$. Aus U^* seien die linearen Funktionale $\varphi_0, \ldots, \varphi_m$ sowie aus C^* deren Erweiterungen $\Phi_0, \ldots, \Phi_m$ gegeben. Dann gilt nach Davis [3] folgender Hilfssatz:

Lemma 1. *Wenn* $\{\varphi_0, \ldots, \varphi_m\}$ *eine Basis von* U^* *ist, so gibt es zu jedem* $f \in C$ *genau ein* $u \in U$ *mit*

$$\varphi_\mu(u) = \Phi_\mu(f) \quad (0 \leqq \mu \leqq m). \tag{1}$$

Wir sprechen in diesem Fall vom eindeutig lösbaren Interpolationsproblem $(U, C; \Phi_0, \ldots, \Phi_m)$. Die lineare Abbildung $L: C \to U$, die jedem $f \in C$ das durch (1) eindeutig bestimmte $u \in U$ zuordnet, bezeichnen wir als Interpolationsoperator. Wir schreiben — eindeutige Lösbarkeit des Interpolationsproblems vorausgesetzt —

$$L = (U, C; \Phi_0, \ldots, \Phi_m).$$

Gegeben seien jetzt zwei (nicht notwendig eindeutig lösbare) Interpolationsprobleme $(U, C; \Phi_0, \ldots, \Phi_m)$ und $(V, \overline{C}; \Psi_0, \ldots, \Psi_n)$ wobei gilt: $V \subset \overline{C} = C(J)$ mit $J = [\bar{a}, \bar{b}]$ und $\dim V = n+1$. Wir definieren das *Tensorprodukt dieser beiden Interpolationsprobleme:*

Als *Grundraum* benutzen wir $W = U \otimes V \subset Z = C \otimes \overline{C}$. Durch Φ_μ und Ψ_ν $(0 \leqq \mu \leqq m,\ 0 \leqq \nu \leqq n)$ wird auf folgende Weise ein lineares Funktional $\Omega_{\mu\nu} \in Z^*$ festgelegt:

Die Abbildung

$$(\Phi_\mu, \Psi_\nu): C \times \overline{C} \to \mathbb{R},$$

die durch

$$(\Phi_\mu, \Psi_\nu)(f, g) = \Phi_\mu(f) \cdot \Psi_\nu(g)$$

gekennzeichnet ist, ist bilinear. Also gibt es eine lineare Abbildung

$$\Omega_{\mu\nu}: C \otimes \overline{C} \to \mathbb{R}$$

so, daß die Beziehung

$$(\Phi_\mu, \Psi_\nu)(f, g) = \Omega_{\mu\nu}(f \otimes g) \tag{2}$$

gilt (vgl. Greub [5]). In $C \otimes \overline{C}$ können wir $f \otimes g$ mit $f \cdot g$ identifizieren, wobei in dem Produkt $f g$ die Funktion f von der einen und g von der anderen Variablen abhängt. Offenbar gilt (da $\dim U$ und $\dim V$ endlich sind)

$$\Omega_{\mu\nu} | W = \varphi_\mu \otimes \psi_\nu. \tag{3}$$

Wir bezeichnen das Interpolationsproblem

$$(U \otimes V, C \otimes \overline{C};\ \Omega_{\mu\nu}:\ 0 \leqq \mu \leqq m,\ 0 \leqq \nu \leqq n)$$

als *Tensorprodukt der Interpolationsprobleme* $(U, C; \Phi_0, \ldots, \Phi_m)$ *und* $(V, \overline{C}; \Psi_0, \ldots, \Psi_n)$. Beachten wir (3), so ergibt sich aus [6, Satz 1]:

Lemma 2. $(U \otimes V, C \otimes \overline{C};\ \Omega_{\mu\nu}:\ 0 \leqq \mu \leqq m,\ 0 \leqq \nu \leqq n)$ *ist genau dann eindeutig lösbar, wenn* $(U, C; \Phi_0, \ldots, \Phi_m)$ *und* $(V, \overline{C}; \Psi_0, \ldots, \Psi_n)$ *eindeutig lösbar sind.*

3. Konvergenz von Interpolationsprozessen

Es sei m_r eine streng monotone Folge natürlicher Zahlen und W_r eine Folge von (m_r+1)-dimensionalen Teilräumen eines normierten linearen Raumes $(B, \|\cdot\|)$ mit

$$W_r \subset W_{r+1} \quad (r \in \mathbb{N}').$$

Dabei bedeutet $\mathbb{N}' = \mathbb{N} \cup \{0\}$.

Weiter sei die Folge von eindeutig lösbaren Interpolationsprozessen

$$G_r = (W_r, B; \Gamma_0^{(r)}, \ldots, \Gamma_{m_r}^{(r)})$$

gegeben. Wir interessieren uns für die Konvergenz der Operatorenfolge G_r $(r \in \mathbb{N}')$ und treffen dazu die

Definition 1. a) Die Operatorenfolge G_r heißt *konvergent für* $f \in B$, wenn

$$\|f - G_r(f)\| = o(1) \quad (r \to \infty)$$

gilt.

b) Es sei $\alpha(r)$ eine reelle Nullfolge. Gilt die Beziehung

$$\|f - G_r(f)\| = O(\alpha(r)) \quad (r \to \infty),$$

so heißt G_r *für* $f \in B$ *konvergent von der Ordnung* $\alpha(r)$.

Wir betrachten jetzt zweidimensionale Interpolationsprozesse, die sich als *Tensorprodukte von konvergenten eindimensionalen Interpolationsprozessen* ergeben, und untersuchen sie auf Konvergenz. Dabei benutzen wir an Stelle von $(B, \|\cdot\|)$ die normierten Räume $(C, \|\cdot\|_1)$, $(\overline{C}, \|\cdot\|_2)$ und $(C \otimes \overline{C}, \|\cdot\|)$ mit

$$\|f\|_1 = \sup_{x \in I} |f(x)|, \qquad \|g\|_2 = \sup_{y \in J} |g(y)|$$

und

$$\|h\| = \sup_{(x,y) \in I \times J} |h(x, y)|.$$

Als *Polynomräume* dienen

$$\Pi_k = \left\{P\colon P(x) = \sum_{i=0}^{k} c_i x^i,\ c_i \in \mathbb{R}\right\} \subset C$$

und

$$\overline{\Pi}_l = \left\{Q\colon Q(y) = \sum_{j=0}^{l} d_j y^j,\ d_j \in \mathbb{R}\right\} \subset \overline{C}$$

sowie

$$\Pi_{kl} = \left\{R\colon R(x, y) = \sum_{i=0}^{k} \sum_{j=0}^{l} a_{ij} x^i y^j,\ a_{ij} \in \mathbb{R}\right\} \subset C \otimes \overline{C}.$$

Offenbar gilt

$$\Pi_{kl} = \Pi_k \otimes \overline{\Pi}_l.$$

Für die nun folgenden Sätze 1 und 2 seien zwei monoton wachsende Folgen m_r bzw. n_r natürlicher Zahlen $(r \in \mathbb{N}')$ sowie die *Folgen von Interpolationsoperatoren*

$$L_r = (\Pi_{m_r}, C; \Phi_0^{(r)}, \ldots, \Phi_{m_r}^{(r)})$$

und

$$\overline{L}_r = (\overline{\Pi}_{n_r}, \overline{C}; \Psi_0^{(r)}, \ldots, \Psi_{n_r}^{(r)})$$

gegeben. Wird $\Omega_{\mu\nu}^{(r)}$ gemäß (2) mit Hilfe von $\Phi_\mu^{(r)}$ und $\Psi_\nu^{(r)}$ $(0 \leqq \mu \leqq m_r,\ 0 \leqq \nu \leqq n_r)$ gebildet, so ist

$$T_r = (\Pi_{m_r n_r}, C \otimes \overline{C};\ \Omega_{\mu\nu}^{(r)}\colon 0 \leqq \mu \leqq m_r,\ 0 \leqq \nu \leqq n_r)$$

der Interpolationsoperator des Tensorprodukts der beiden obigen Probleme. Jedes $h \in C \otimes \overline{C}$ hat die Darstellung

$$h(x, y) = \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} f_i(x) \cdot g_j(y).$$

Für $h \in C \otimes \overline{C}$ gilt

Satz 1. *Wenn L_r für jedes f_i $(0 \leqq i \leqq M)$ und $\overline{L}_r$ für jedes g_j $(0 \leqq j \leqq N)$ konvergiert, so konvergiert T_r für h $(r \to \infty)$.*

Beweis. Es sei

$$L_r(f_i) = P_i^{(r)} \in \Pi_{m_r} \quad (0 \leqq i \leqq M)$$

und

$$\overline{L}_r(g_j) = Q_j^{(r)} \in \overline{\Pi}_{n_r} \quad (0 \leqq j \leqq N).$$

Wir zeigen, daß dann $R^{(r)} = T_r(h)$ die Form

$$T_r(h) = \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} P_i^{(r)} \cdot Q_j^{(r)}$$

besitzt. Dazu betrachten wir für alle $(\mu, \nu) \in \{0, \ldots, m_r\} \times \{0, \ldots, n_r\}$ und beliebiges $r \in \mathbb{N}'$:

$$\begin{aligned}
\Omega_{\mu\nu}^{(r)}(h) &= \Omega_{\mu\nu}^{(r)}\left(\sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} f_i g_j\right) \\
&= \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} \Omega_{\mu\nu}^{(r)}(f_i \cdot g_j) \\
&= \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} \Phi_\mu^{(r)}(f_i) \cdot \Psi_\nu^{(r)}(g_j) \\
&= \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} \varphi_\mu^{(r)}(P_i^{(r)}) \cdot \psi_\nu^{(r)}(Q_j^{(r)}) \\
&= \varphi_\mu^{(r)} \otimes \psi_\nu^{(r)}\left(\sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} P_i^{(r)} Q_j^{(r)}\right) \\
&= \varphi_\mu^{(r)} \otimes \psi_\nu^{(r)}(R^{(r)}).
\end{aligned}$$

Mit (1) und der eindeutigen Lösbarkeit von $(\Pi_{m_r}, C;\ \Phi_0^{(r)}, \ldots, \Phi_{m_r}^{(r)})$ und $(\overline{\Pi}_{n_r}, \overline{C};\ \Psi_0^{(r)}, \ldots, \Psi_{n_r}^{(r)})$ folgt die Beziehung

$$R^{(r)} = T_r(h) \quad (r \in \mathbb{N}').$$

Also gilt

$$\|h - T_r(h)\| = \left\| \sum_{i=0}^{M} \sum_{j=0}^{N} b_{ij} (f_i g_j - P_i^{(r)} Q_j^{(r)}) \right\| \leqq \sum_{i=0}^{M} \sum_{j=0}^{N} |b_{ij}| \cdot \|f_i g_j - P_i^{(r)} Q_j^{(r)}\|. \quad (4)$$

Für $0 \leqq i \leqq M$ und $0 \leqq j \leqq N$ ergibt sich weiter

$$\|f_i g_j - P_i^{(r)} Q_j^{(r)}\| \leqq \|f_i\|_1 \cdot \|g_j - Q_j^{(r)}\|_2 + \|Q_j^{(r)}\|_2 \cdot \|f_i - P_i^{(r)}\|_1. \tag{5}$$

Wegen der Konvergenz von $\bar{L}_r$ bei g_j $(0 \leqq j \leqq N)$ folgt

$$\|Q_j^{(r)}\|_2 \leqq \|g_j\|_2 + 1 \quad \text{für} \quad r \geqq r_0. \tag{6}$$

Damit ergibt sich die Behauptung.

Satz 2. *Wenn L_r und $\bar{L}_r$ für jedes f_i $(0 \leqq i \leqq M)$ bzw. g_j $(0 \leqq j \leqq N)$ von der Ordnung $\alpha(r)$ konvergieren, so konvergiert T_r für h ebenfalls von der Ordnung $\alpha(r)$ $(r \to \infty)$.*

Beweis. Die Behauptung folgt entsprechend aus (4) mit Benutzung von (5) und (6).

Konvergieren L_r für die f_i $(0 \leqq i \leqq M)$ bzw. $\bar{L}_r$ für die g_j $(0 \leqq j \leqq N)$ *von verschiedener Ordnung* $\alpha_i(r)$ bzw. $\bar{\alpha}_j(r)$, so gilt für die Ordnung $\tilde{\alpha}(r)$ der Konvergenz von T_r für h:

$$\tilde{\alpha}(r) = \max_{i,j} \{\alpha_i(r), \bar{\alpha}_j(r)\}.$$

Die *Bedeutung der Sätze 1 und 2* liegt darin, daß sich ohne spezielle Voraussetzungen über das benutzte (zweidimensionale) Interpolationsverfahren Konvergenzeigenschaften der zugrundeliegenden Verfahren auf den zweidimensionalen Prozeß für Funktionen $h \in Z = C \otimes \overline{C}$ übertragen, sofern sich der zweidimensionale Prozeß nur als Tensorprodukt zweier eindimensionaler Interpolationsprobleme darstellen läßt.

Am Ende der Arbeit weisen wir auf einige *Anwendungsmöglichkeiten* hin.

4. Eine Umkehrung von Satz 1 für die Lagrange-Interpolation

Gegeben seien die beiden *Knotenmatrizen*

$$M_1 = \begin{pmatrix} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \quad \text{und} \quad M_2 = \begin{pmatrix} y_0^{(0)} & & & \\ y_0^{(1)} & y_1^{(1)} & & \\ y_0^{(2)} & y_1^{(2)} & y_2^{(2)} & \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

mit den Eigenschaften

1. $x_\varrho^{(r)} \in I$ bzw. $y_\varrho^{(r)} \in J$ $(0 \leqq \varrho \leqq r,\ r \in \mathbb{N}')$.
2. Für festes $r \in \mathbb{N}'$ gilt: $x_\varrho^{(r)} \neq x_\sigma^{(r)}$ $(\varrho \neq \sigma)$
   bzw. $y_\varrho^{(r)} \neq y_\sigma^{(r)}$ $(\varrho \neq \sigma)$.

Wir definieren für $r \in \mathbb{N}'$ die linearen Funktionale

$$\hat{\Phi}_\varrho^{(r)}: C \ni f \mapsto f(x_\varrho^{(r)}) \quad (0 \leqq \varrho \leqq r), \tag{7}$$

$$\hat{\Psi}_\varrho^{(r)}: \overline{C} \ni g \mapsto g(y_\varrho^{(r)}) \quad (0 \leqq \varrho \leqq r) \tag{8}$$

und damit die Interpolationsoperatoren

$$\hat{L}_r = (\Pi_r, C; \hat{\Phi}_0^{(r)}, \ldots, \hat{\Phi}_r^{(r)}) \tag{9}$$

und

$$\tilde{L}_r = (\overline{\Pi}_r, \overline{C}; \hat{\Psi}_0^{(r)}, \ldots, \hat{\Psi}_r^{(r)}). \tag{10}$$

Entsprechend wie in (2) bilden wir $\hat{\Omega}_{\varrho\sigma}^{(r)}$ mit der Eigenschaft

$$\hat{\Omega}_{\varrho\sigma}^{(r)}: C \otimes \overline{C} \ni h \mapsto h(x_\varrho^{(r)}, y_\sigma^{(r)}), \quad (0 \leqq \varrho, \sigma \leqq r).$$

Diese Beziehung ergibt sich aus (7) und (8) mit Hilfe von (2) und der Linearität von $\hat{\Omega}_{\varrho\sigma}^{(r)}$. Demnach wird der zweidimensionale Lagrange-Interpolationsoperator durch

$$L_r^{(2)} = (\Pi_{r,r}, C \otimes \overline{C}; \hat{\Omega}_{\varrho\sigma}^{(r)}: 0 \leqq \varrho, \sigma \leqq r)$$

dargestellt. Dieses Interpolationsproblem ist das Tensorprodukt der beiden obigen Lagrange-Interpolationsprobleme (9) und (10). Es wird dabei ein $r \times r$-Punktgitter des $\mathbb{R}^2$ als Menge der Interpolationsknoten verwendet. Ebenso können jedoch beliebige $r \times s$-Punktgitter des $\mathbb{R}^2$ benutzt werden.

Wir betrachten jetzt *für die Lagrange-Interpolation eine Umkehrung von Satz 1.* Zunächst zeigen wir, daß sie ohne einschränkende Bedingungen nicht möglich ist. Dazu sei

$$h_1(x, y) = f_1(x) \cdot g_1(y) \in C[-1, 1] \otimes C[-1, 1]$$

mit

$$f_1(x) = |x| \qquad (x \in [-1, 1])$$

und

$$g_1(y) = 0 \qquad (y \in [-1, 1])$$

gegeben. Die Knotenmatrizen $\overline{M}_1$ und $\overline{M}_2$ werden dadurch festgelegt, daß

$$\bar{x}_\varrho^{(r)} = \bar{y}_\varrho^{(r)} = -1 + \frac{2}{r} \cdot \varrho \quad (0 \leqq \varrho \leqq r)$$

ist. Da $h_1 = 0$ ist, folgt aus der Linearität von $L_r^{(2)}$

$$L_r^{(2)}(h_1) = 0 \quad \text{für jedes } r \in \mathbb{N}'.$$

Damit ist der zweidimensionale Interpolationsprozeß $L_r^{(2)}$ insbesondere konvergent. Nach Bernstein (vgl. Natanson [11, Band III]) konvergiert aber $\hat{L}_r$ bei äquidistanten Interpolationsknoten für f_1 nicht.

Zur Umkehrung von Satz 1 für Funktionen $h = f \cdot g$ treffen wir über die Knotenmatrix die

Definition 2. Es sei $M_1^{(i)}$ die i-te Zeile der Knotenmatrix M_1. M_1 heißt *dicht in I*, wenn zu jedem $x_0 \in I$ eine Folge

$$(x_{k_0}^{(0)}, x_{k_1}^{(1)}, x_{k_2}^{(2)}, \ldots, x_{k_r}^{(r)}, \ldots) \in \bigtimes_{i=0}^{\infty} M_1^{(i)}$$

existiert, die gegen x_0 konvergiert.

Eine entsprechende Definition gelte für M_2. Damit erhalten wir

Satz 3. *$\hat{L}_r$ und $\tilde{L}_r$ seien wie in* (9) *bzw.* (10) *definiert. Mindestens eine der Knotenmatrizen M_1 und M_2 sei dicht in I bzw. J. Dann gilt für jedes $h = f \cdot g \in C \otimes \overline{C}$ $(h \neq 0)$: Wenn $L_r^{(2)}$ für h konvergiert, so konvergieren $\hat{L}_r$ bzw. $\tilde{L}_r$ für f bzw. g.*

Beweis. Ohne Beschränkung der Allgemeinheit sei M_1 dicht in I. Wie im Beweis von Satz 1 ergibt sich, daß

$$L_r^{(2)}(h) = \hat{L}_r(f) \cdot \tilde{L}_r(g) = \hat{P}_r \cdot \hat{Q}_r$$

ist. Damit gilt

$$\|h - L_r^{(2)}(h)\| = \|f \cdot g - \hat{P}_r \cdot \hat{Q}_r\| \geqq \sup_{y \in J} |f(x_\varrho^{(r)})\, g(y) - \hat{P}_r(x_\varrho^{(r)})\, \hat{Q}_r(y)|$$

für jedes $x_\varrho^{(r)}$ aus M_1. Wegen $f(x_\varrho^{(r)}) = \hat{P}_r(x_\varrho^{(r)})$ folgt

$$\|h - L_r^{(2)}(h)\| \geqq |f(x_\varrho^{(r)})| \cdot \|g - \hat{Q}_r\|_2. \tag{11}$$

Weiter ist $f \neq 0$. Also gibt es ein $x_0 \in I$ sowie eine Umgebung $U = U(x_0)$ so, daß $f(x) \neq 0$ ist für $x \in U$. Da M_1 dicht in I ist, gibt es eine Folge

$$(x_{i_0}^{(0)}, x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots),$$

die gegen x_0 konvergiert. Aus der Stetigkeit von f folgt:

$$|f(x_{i_r}^{(r)})| \geqq \alpha > 0 \quad \text{für} \quad r \geqq r_0.$$

Damit erhalten wir aus (11):

$$\|h - L_r^{(2)}(h)\| \geqq \alpha \|g - \hat{Q}_r\|_2 \quad (r \geqq r_0).$$

Aus der Konvergenz von $L_r^{(2)}$ für h ergibt sich wegen $\alpha > 0$ die Konvergenz von $\tilde{L}_r$ für g. Wir zeigen, daß dann auch $\hat{L}_r$ für f konvergiert. Es ist nämlich

$$\|\hat{Q}_r\|_2 \cdot \|\hat{P}_r - f\|_1 \leqq \|\hat{P}_r \hat{Q}_r - f g\| + \|f\|_1 \|\hat{Q}_r - g\|_2.$$

Mit Hilfe der Konvergenz von $\tilde{L}_r$ erhalten wir

$$\|\hat{Q}_r\|_2 \geqq \frac{\|g\|_2}{2} > 0 \quad \text{für} \quad r \geqq r_1.$$

Also erhalten wir für $r \geqq r_1$:

$$\|\hat{P}_r - f\|_1 \leqq \frac{2}{\|g\|_2} (\|\hat{P}_r \hat{Q}_r - f g\| + \|f\|_1 \|\hat{Q}_r - g\|_2).$$

Dies aber besagt, daß $\hat{L}_r$ für f konvergiert.

Wir geben jetzt *zwei Klassen von Lagrange-Interpolationsprozessen* an, *deren Knotenmatrizen dicht in I sind.*

Satz 4. *Jede ϱ-normale Knotenmatrix $\hat{M}$ ($\varrho > 0$) ist dicht in I.*

(Über ϱ-normale Knotenmatrizen informiert z.B. Natanson [11].)

Beweis. Für die Lagrange-Grundpolynome $l_k^{(r)}$ (vgl. Erdös und Turan [4]) gilt bei ϱ-normaler Knotenmatrix ($\varrho > 0$):

$$\sum_{k=0}^{r} (l_k^{(r)}(x))^2 \leqq c \quad (x \in I)$$

und damit

$$|l_k^{(r)}(x)| \leqq \sqrt{c} \quad (0 \leqq k \leqq r,\ r = 0, 1, 2, \ldots).$$

Hieraus folgt nach Erdös und Turan weiter, daß die Interpolationsknoten trigonometrisch quasigleichverteilt sind. Setzen wir

$$x_k^{(r)} = \cos \vartheta_k^{(r)}$$

mit

$$0 = \vartheta_{-1}^{(r)} \leqq \vartheta_0^{(r)} < \vartheta_1^{(r)} < \cdots < \vartheta_r^{(r)} \leqq \vartheta_{r+1}^{(r)} = \pi,$$

so besagt dies, daß für passende positive Zahlen c_1 und c_2 die Abschätzungen

$$\frac{c_1}{r} \leqq \vartheta_{k+1}^{(r)} - \vartheta_k^{(r)} \leqq \frac{c_2}{r} \quad (k = 0, 1, \ldots, r-1)$$

gelten. Für $k = -1$ und $k = r$ sind die Abschätzungen nach oben ebenfalls erfüllt. Zu beliebigem $\vartheta_0 \in [0, \pi]$ läßt sich demnach eine Folge $\vartheta_{i_r}^{(r)}$ bilden, die gegen ϑ_0 konvergiert. Da die Abbildung $\omega: x \mapsto \cos x$ stetig ist, ergibt sich, daß die ϱ-normale Knotenmatrix $\hat{M}$ $(\varrho > 0)$ dicht in I ist.

Satz 5. *Jeder Lagrange-Interpolationsprozeß L_r, der für alle $f \in C$ konvergiert, die in einer Lipschitz-Klasse* Lip α $(0 < \alpha \leqq 1)$ *liegen, besitzt eine in I dichte Knotenmatrix M'.*

Beweis. Wir nehmen an, die Matrix M' sei nicht dicht in I. Dann gibt es ein $x_0 \in I$ und eine abgeschlossene Umgebung $V = V(x_0) = [A, B] \subset I$ $(A \neq B)$ so, daß gilt:

$$x_\varrho^{(r)} \notin V \quad (\text{für alle } x_\varrho^{(r)} \text{ aus } M',\ r \geqq r_2).$$

Es sei nun $f \in \mathrm{Lip}\,\alpha$. Nach Voraussetzung gilt dann

$$\|L_r(f) - f\|_1 = o(1) \quad (r \to \infty).$$

Weiter sei $F = 2 \cdot (\|f\|_1 + 1) > 0$. Wir definieren die stetige Funktion φ durch

$$\varphi(x) = \begin{cases} f(x) & \text{für} \quad x \notin [A, B] \\ f(A) + \dfrac{2(F - f(A))}{B - A}(x - A) & \text{für} \quad A \leqq x \leqq \dfrac{A+B}{2} \\ f(B) + \dfrac{2(F - f(B))}{B - A}(B - x) & \text{für} \quad \dfrac{A+B}{2} < x \leqq B \end{cases}$$

und zeigen, daß φ auch in der Klasse Lip α liegt. Für $x, y \in I - [A, B]$ gilt

$$|\varphi(x) - \varphi(y)| = |f(x) - f(y)| \leqq d_1 |x - y|^\alpha.$$

Hier wie später sind $d_1, d_2, \ldots$ positive Konstanten. Für $x, y \in [A, B]$ folgt, da φ dort stückweise geradlinig ist,

$$|\varphi(x) - \varphi(y)| \leqq d_2 |x - y| \leqq K \cdot d_2 |x - y|^\alpha$$

mit $K = \max(1, b - a)$. Wir müssen noch die Fälle

$$a \leqq x < A < y \leqq B \quad \text{und} \quad A \leqq x < B < y \leqq b$$

betrachten. Aus Symmetriegründen genügt es, den ersten Fall zu untersuchen:

$$\begin{aligned} |\varphi(x) - \varphi(y)| &\leqq |\varphi(x) - \varphi(A)| + |\varphi(A) - \varphi(y)| \\ &\leqq d_3 \cdot |x - A|^\alpha + d_4 \cdot |A - y| \leqq d_3 \cdot |x - A|^\alpha + K \cdot d_4 \cdot |A - y|^\alpha. \end{aligned}$$

Aus der Monotonie der Abbildung $\hat{\alpha}$: $z \mapsto z^{\alpha}$ $(\alpha, z > 0)$ folgt:

$$|\varphi(x) - \varphi(y)| \leqq d_3 \cdot |x - y|^{\alpha} + K \cdot d_4 \cdot |x - y|^{\alpha} = d_5 \cdot |x - y|^{\alpha},$$

d.h. $\varphi \in \mathrm{Lip}\,\alpha$.

Wir zeigen jetzt, daß — im Gegensatz zur Voraussetzung — L_r für φ nicht konvergiert. Es ist nämlich

$$L_r(f) = L_r(\varphi) = \overline{P}_r$$

und $\|\varphi\|_1 \geqq 2(\|f\|_1 + 1)$. Damit folgt

$$\begin{aligned} \|\varphi - \overline{P}_r\|_1 &\geqq \|\varphi - f\|_1 - \|f - \overline{P}_r\|_1 \\ &\geqq 2(\|f\|_1 + 1) - \|f\|_1 - \|f - \overline{P}_r\|_1 \\ &\geqq 2(\|f\|_1 + 1) - 2\|f\|_1 - 1 \\ &\geqq 1 \quad \text{für} \quad r \geqq r_3. \end{aligned}$$

Also konvergiert L_r für φ nicht; dies aber widerspricht der Voraussetzung.

Anwendungen. Die Sätze 1 und 2 lassen sich auf verschiedene Arten von mehrdimensionalen Interpolationsprozessen anwenden. Neben der *Lagrange-, Hermite- und Lücken-Interpolation in mehreren Veränderlichen* ist insbesondere die *Konvergenz mehrdimensionaler Fourier-Reihen* zu nennen.

Bei *Interpolationsprozessen mit beschränkten Normen*

$$\|G_r\| \leqq K' \quad \text{(unabhängig von } r\text{)}$$

lassen sich auch für $h \in C([a, b] \times [\bar{a}, \bar{b}])$ günstige Konvergenzaussagen gewinnen (vgl. [7, 9, 10]). Für *Interpolationsoperatoren mit unbeschränkten Normen* ergeben sich für $h \in C([a, b] \times [\bar{a}, \bar{b}])$ im allgemeinen schwächere Konvergenzaussagen als für $\bar{h} \in C[a, b] \otimes C[\bar{a}, \bar{b}]$.

Literatur

1. Ahlin, A. C.: On error bounds for Gaussian cubature. SIAM Review **4**, 25—39 (1962).
2. — A bivariative generalization of Hermite's formula. Math. Comput. **18**, 264—273 (1964).
3. Davis, P. J.: Interpolation and approximation. New York-Toronto-London: Blaisdell 1965.
4. Erdös, P., Turan, P.: On interpolation II. Annals of Math. **39**, 703—724 (1938).
5. Greub, W. H.: Linear algebra. Berlin-Göttingen-Heidelberg: Springer 1963.
6. Haußmann, W.: Tensorprodukte und mehrdimensionale Interpolation. Math. Z. **113**, 17—23 (1970).
7. — Mehrdimensionale Hermite-Interpolation. Erscheint demnächst.
8. Milne, W. E., Arntzen, W., Reynolds, N., Wheelock, J.: Mathematics for digital computers, Vol. I: Multivariative interpolation. Wright Air Development Center, Wright-Patterson Air Force Base, Ohio, 1958.
9. Mond, B., Shisha, O.: The rapidity of convergence of the Hermite-Fejér approximation to functions of one or several variables. Proc. AMS **16**, 1269—1276 (1965).
10. — — On the approximation of functions of several variables. J. Res. Nat. Bur. Standards Sect. B, **70 B**, 211—218 (1966).
11. Natanson, I. P.: Constructive function theory. Vol. I—III. New York: Ungar 1964/65.

12. Salzer, H. E.: Note on interpolation for a function of several variables. Bull. AMS **51**, 279–280 (1945).
13. – Tables of coefficients for interpolating in functions of two variables. Journ. Math. Phys. **26**, 294–305 (1948).
14. – Note on multivariative interpolation for unequally spaced arguments, with application to double summation. Journ. SIAM **5**, 254–262 (1957).
15. – Some new divided difference algorithms for two variables. Erschienen in: On numerical approximation. Proc. of a Symp., Madison, 61–98; Herausgeber: R. E. Langer, University of Wisconsin Press, 1959.
16. –, Kimbro, G. M.: Tables for bivariative osculatory interpolation over a Cartesian grid. Convair Division of General Dynamics Corporation: San Diego 1958.
17. Sloss, J. M.: Approximation by interpolation of certain real compact analytic surfaces. Math. Z. **100**, 272–281 (1967).
18. Thacher, H. C., Jr.: Generalizations of concepts related to linear dependence. Journ. SIAM **6**, 288–299 (1958).
19. – Derivation of interpolation formulas in several independent variables. Annals of the New York Acad. of Sciences **86**, 758–775 (1960).
20. –, Milne, W. E.: Interpolation in several variables. Journ. SIAM **8**, 33–42 (1960).

Prof. Dr. H. Ehlich
Dr. W. Haußmann
Institut für Mathematik der Universität
D 4630 Bochum
Buscheystr. Geb. NA

Numer. Math. 15, 175–195 (1970)

Das Tschebyscheff-Verfahren in Banach-Räumen

Boro Döring

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 15. September 1969

Summary. Two theorems on an iteration method of order three — usually referred to as Chebyshev's method — for the approximate solution of nonlinear operator equations in Banach spaces are given. Very simple and weak assumptions are obtained as well as error estimates which are much better compared with known results. Some worked examples of operator equations of different types are given.

Einleitung

Das Newton-Verfahren hat sich als eines der wirksamsten Verfahren zur Lösung nichtlinearer Operatorgleichungen, insbesondere von Systemen nichtlinearer Gleichungen, von nichtlinearen Integralgleichungen und Matrix-Eigenwertproblemen erwiesen. Trotz der quadratischen Konvergenz dieses Verfahrens erfordert jedoch die Lösung einer derartigen Operatorgleichung immer noch einen beachtlichen Rechenaufwand, den man natürlich nach Möglichkeit verringern möchte. Beispielsweise ist zur diskretisierten Lösung einer nichtlinearen Integralgleichung für jeden Newton-Schritt ein System von etwa 100 oder mehr linearen Gleichungen zu lösen. Man möchte deshalb ein Verfahren haben, daß bei möglichst wenig zusätzlichem Pro-Schritt-Aufwand mit noch weniger Iterationsschritten als das Newton-Verfahren auskommt. Das einfachste derartige Verfahren (dritter Ordnung) ist unter dem Namen verbessertes oder verschärftes Newton-Verfahren bzw. Tschebyscheff-Verfahren (TV) bekannt (s. z.B. Collatz [5], S. 252). Hierbei wird zu der Newton-Iterationsfunktion ein additives Korrekturglied hinzugefügt, das die zweite (Fréchet-)Ableitung enthält. Da wir nicht nur die Konvergenz von TV untersuchen, sondern auch Aussagen über die Existenz einer Lösung der verschiedenen Typen von Operatorgleichungen in einem Bereich machen wollen, empfiehlt es sich, die Betrachtungen in Banach-Räumen durchzuführen.

TV ist in derartigen Räumen zuerst von Nečepurenko [21] behandelt worden. Die Voraussetzungen des hierbei erhaltenen Konvergenzsatzes sind allerdings ziemlich scharf. Ferner zeigt Nečepurenko nur die quadratische, nicht aber die kubische Konvergenz des Verfahrens (s. auch Moore [20], Theorem 7). Kaazik [16] hat dieses Ergebnis verbessert, insbesondere die Schranken. Daneben beweist er einen anderen Satz, bei dem der Operator F an der Ausgangsnäherung x_0 „analytisch" vorausgesetzt wird; man braucht dann die Ableitungen nicht über einen Bereich abzuschätzen. In beiden Sätzen sind jedoch die sonstigen Voraussetzungen äußerst kompliziert. Verbesserungen dieser Sätze werden von Shafiyev [23] gegeben: Die Voraussetzungen werden etwas vereinfacht und abgeschwächt, so daß

zwar die dritte, nicht aber die zweite Fréchet-Ableitung über den ganzen Bereich abgeschätzt zu werden braucht. Černyšenko [4] arbeitet mit einer Hölderbedingung für die zweite Ableitung, so daß die dritte nicht benötigt wird. Die bewiesene Konvergenzgeschwindigkeit liegt zwischen der von Nečepurenko und Shafiyev. Varyukhin und Kas'yanyuk [27] betrachten TV für nichtlineare Gleichungssysteme. Dubovik überträgt den von Ehrmann für holomorphe Funktionen ausgesprochenen Satz ([13], Satz 5; s. auch Traub [26], Satz 2—10), wonach man zu einer bekannten Iterationsfunktion k-ter Ordnung zur Lösung der Gleichung $F(x)=0$ im Falle einer einfachen Nullstelle die allgemeinste Iterationsfunktion erhält, indem man ein gewisses Vielfaches der k-ten Potenz von $F(x)$ additiv hinzufügt, für den Fall $k=3$ auf Banach-Räume [11] und betrachtet neben TV solche Verfahren, wo die zweite Ableitung durch einen symmetrischen bilinearen Operator ersetzt ist [12] sowie Anwendungen auf Matrix-Gleichungen. Mit der Lösung einer speziellen Matrix-Gleichung zur Bestimmung der Inversen einer Matrix nach TV (= verallgemeinertes Verfahren von Schulz) befassen sich Albrecht [2] und Forster [14]. Konvergenz- und Fehleraussagen für die nach TV bestimmte Folge von Näherungslösungen ergeben sich auch aus einem allgemeinen Satz von Collatz [6] (s. auch [5], S. 253), der unter gewissen Voraussetzungen die Existenz einer Lösung sowie die Konvergenz eines Iterationsverfahrens k-ter Ordnung sichert und den Fehler abzuschätzen gestattet. Ähnliche Sätze finden sich bei Jankó [15], Kaazik [17], u.a.

Nachteilig an den zitierten bisherigen Resultaten über TV ist, daß die Voraussetzungen entweder zu kompliziert oder relativ scharf sind — meist trifft sogar beides zu —, die Fehlerschranken, besonders in Funktionenräumen, viel zu grob sind und die Eindeutigkeitsaussage entweder ganz fehlt oder nur in einem kleineren Bereich als die Existenzaussage bewiesen wird.

Diese Nachteile werden von den in der vorliegenden Note bewiesenen Sätzen vermieden. Diese wurden vor allem im Hinblick auf die Anwendungen unter folgenden Zielsetzungen aufgestellt:

1. Die Voraussetzungen sind möglichst schwach und bei der Anwendung auf konkrete Probleme der Analysis leicht nachzuprüfen.

2. Gleichzeitige (lokale) Aussagen (Existenz, Eindeutigkeit, Konvergenz, Ordnung, Fehlerschranken) unter denselben Voraussetzungen in demselben Bereich (für den der Einfachheit halber eine durch die Ausgangsnäherung x_0 allein bestimmte Kugel gewählt wird).

3. Möglichst scharfe, numerisch leicht berechenbare Fehlerschranken.

Während der erste Satz („Standardsatz", Abschn. 2) diese Bedingungen gut erfüllt, liegt der zweite Satz („Grenzsatz", Abschn. 5) bereits auf der Grenze zwischen Erfülltsein und Nichterfülltsein dieser Forderungen. In beiden Sätzen werden je eine a-priori-Schranke und zwei verschiedene a-posteriori-Schranken angegeben. Die Sätze unterscheiden sich in den Voraussetzungen, die bezüglich Existenz und Konvergenz beim Grenzsatz schwächer sind. In Abschn. 4 wird ein Satz bewiesen, der es gestattet, leicht andere Kombinationen von zulässigen Voraussetzungen zu finden. In Abschn. 6 werden die neuen Fehlerschranken an Hand von Beispielen dem wirklichen Fehler und den bisher bekannten Fehler-

schranken gegenübergestellt. Zum Schluß wird kurz über numerische Erfahrungen im Hinblick auf den Rechenaufwand und den Anwendungsbereich von TV berichtet.

1. Zur Herleitung des Tschebyscheff-Verfahrens

Im Eindimensionalen erhält man das Tschebyscheff-Verfahren (TV) zur näherungsweisen Lösung der Gleichung $F(x)=0$ zu [1]

$$x_{n+1} := x_n - \frac{F_n}{F_n'} - \frac{1}{2}\cdot\frac{F_n''}{F_n'}\cdot\left(\frac{F_n}{F_n'}\right)^2 \quad \text{(TV)} \tag{1.1}$$

als (nächst dem Newton-Verfahren) einfachsten Spezialfall des bekannten, von Schröder [22] 1869 angegebenen Iterationsverfahrens k-ter Ordnung (= nach k Gliedern abgebrochene und gleich Null gesetzte Taylorreihe der Umkehrfunktion von $y=F(x)$ für $y=0$):

$$x_{n+1} := x_n + \sum_{j=1}^{k-1} \frac{(-1)^j}{j!} F_n^j \left[\frac{1}{F_n'}\frac{d}{dx}\right]^{j-1} \frac{1}{F_n'}$$

für $k=3$. Denn

$$\left[\frac{1}{F_n'}\frac{d}{dx}\right]\frac{1}{F_n'} = \left[\frac{1}{F'(x)}\cdot\left(\frac{1}{F'(x)}\right)'\right]_{x=x_n} = -\frac{F_n''}{[F_n']^3}.$$

Eine andere, für die Beweise der folgenden Sätze wie auch im Hinblick auf eine Verallgemeinerung auf höhere Ordnung[2] nützliche Herleitung ist die folgende. Nullsetzen der beiden ersten Glieder der Taylorreihe von $F(x)$ an der Stelle x_n liefert das Newton-Verfahren

$$F_n + F_n' c_n = 0$$

und der ersten *drei* Glieder

$$F_n + F_n' \hat{d}_n + \tfrac{1}{2}\cdot F_n'' \hat{d}_n^2 = 0, \tag{1.2}$$

wobei $\hat{d}_n$ die Korrektur von x_n bedeutet. Um diese Gleichung zu linearisieren, ersetzen wir $\hat{d}_n^2$ durch c_n^2:

$$F_n + F_n'\, d_n + \tfrac{1}{2} F_n'' c_n^2 = 0. \tag{1.3}$$

(1.3) nach x_{n+1} $(= x_n + d_n)$ aufgelöst, ergibt mit der Newton-Iteration gerade (1.1) und umgekehrt. Um gleich die später in den Sätzen, wo wir aus praktischen Gründen „von $n-1$ auf n" rechnen, benötigte Form zu haben, ersetzen wir in den obigen Gleichungen n durch $n-1$. Das Verfahren lautet also (auch in Banach-Räumen)[1]:

$$F_{n-1} + F_{n-1}' c_{n-1} = \Theta, \tag{1.4a}$$

$$F_{n-1} + F_{n-1}' d_{n-1} + \tfrac{1}{2} F_{n-1}'' c_{n-1}^{(2)} = \Theta \quad \text{(TV)}, \tag{1.4b}$$

$$x_n := x_{n-1} + d_{n-1}; \quad \forall n \in \mathbb{N}; \quad x_0 \text{ zu wählen}. \tag{1.4c}$$

(1.4) und (1.1) sind also im Eindimensionalen — abgesehen von dem um 1 abweichenden Index — äquivalent. Prinzipiell gilt dies auch in Banach-Räumen,

1 Bezüglich der Bezeichnungen s. Abschn. 2.

2 Die vorliegenden Ergebnisse lassen sich auf die Ehrmann-Verfahren ([13], Satz 7) verallgemeinern (s. Döring [8]).

doch ist dann in (1.1) die Inverse von F_n' anders zu bezeichnen. Bei den folgenden Sätzen wird stets von der Form (1.4) ausgegangen. Interessanterweise ist diese Form von TV von keinem der in der Einleitung genannten Autoren, die sich speziell mit der Untersuchung von TV befaßt haben, benutzt worden. Die Aussagen des Standardsatzes wie auch die des Grenzsatzes werden der Übersichtlichkeit halber gleich in Form einer Beweisskizze gegeben.

2. Ein Standardsatz über das Tschebyscheff-Verfahren

Im folgenden bedeute F einen (nicht notwendig linearen) Operator, der eine offene Teilmenge X_F des zugrunde gelegten Banach-Raumes X in einen andern Banach-Raum abbildet. Wir führen folgende Abkürzungen ein:

$$\mathbb{N}_p := \{p, p+1, p+2, \ldots\} \Rightarrow \mathbb{N}_1 = \mathbb{N}; \quad \zeta := \|[F'(x_0)]^{-1} F(x_0)\|,$$

$$F_n := F(x_n); \quad F_n^{(j)} := F^{(j)}(x_n) \quad (j = 1, 2, 3); \quad f_n := F'(x_n) \quad \forall n \in \mathbb{N}_0,$$

wobei $F^{(j)}(x_n)$ die Fréchet-Ableitungen des Operators F an der „Stelle" x_n bedeuten. Zwei bzw. drei gleiche Argumente g des bilinearen bzw. trilinearen Operators F_n'' bzw. F_n''' werden zu $g^{(2)}$ bzw. $g^{(3)}$ zusammengefaßt. Griechische Buchstaben bedeuten stets positive reelle Zahlen mit Ausnahme von Beispiel 3 und Θ, das zur Bezeichnung des Nullelementes des jeweils in Frage kommenden Banach-Raumes verwendet wird.

Satz 1 (Standardsatz über TV).

I. *Falls* $x_0 \in X_F$ so gewählt werden kann, daß (V1)—(V3) gilt,

(V1) $\exists S := \left\{x \in X \mid \|x - x_0\| \leqq \frac{14}{9}\zeta\right\}$, wo F dreimal Fréchet-differenzierbar ist mit $\|F^{(j)}(x)\| \leqq K_j \quad \forall x \in S$ und $j = 2, 3$,

(V2) $\exists [F'(x_0)]^{-1}$ mit $\|[F'(x_0)]^{-1}\| \leqq \beta$,

(V3) $\beta K_j \zeta^{j-1} \leqq \frac{1}{3} \quad (j = 2, 3)$,

dann

(B1) existiert die durch TV bzw. (1.4) definierte Folge $\{x_n\}$, und es ist $x_n \in S$ $\forall n \in \mathbb{N}_0$. Ferner gilt:

(B2) Die Folge $\{x_n\}$ konvergiert, und zwar gilt $x^* := \lim x_n \in S$.

(B3) TV konvergiert von der Ordnung 3.

(B4) $F(x^*) = \Theta$.

(B5) In S gibt es außer x^* keine andere Lösung der Gleichung $F(x) = \Theta$.

(B6) Es gelten die Fehlerabschätzungen

$$\|x^* - x_n\| \leqq \frac{14}{9}\left(\frac{11}{18}\right)^n \zeta (\beta\gamma\zeta)^{3^n - 1}, \quad \forall n \in \mathbb{N}_0, \tag{2.1}$$

$$\|x^* - x_n\| < \frac{14}{9}\beta_n \left(M_3 + \frac{7}{3} M_2^2\right) \|x_n - x_{n-1}\|^3, \quad \forall n \in \mathbb{N}, \tag{2.2}$$

$$\|x^* - x_n\| \leqq \frac{14}{9}\beta_n [M_3 \|d_{n-1}\|^3 + M_2 \|d_{n-1} - c_{n-1}(\|d_{n-1}\| + \|c_{n-1}\|)], \quad \forall n \in \mathbb{N}, \tag{2.3}$$

wobei

$$\gamma := 2M_2 \sqrt{1+\frac{6}{5}\frac{M_3}{\beta M_2^2}}\,; \quad \eta_n := \frac{7}{3}\beta_n M_2 \zeta_n; \quad M_j := \frac{K_j}{j!} \quad (j=2,3), \tag{2.4}$$

$$\beta_n := \frac{\beta_{n-1}}{1-\eta_{n-1}} \quad \forall n \in \mathbb{N}, \quad \beta_0 := \beta; \quad \zeta_n := \|f_n^{-1} F_n\| \quad \forall n \in \mathbb{N}_0 \quad (\Rightarrow \zeta_0 = \zeta). \tag{2.5}$$

II. *Falls* außer (V1)—(V3) noch folgende Bedingung erfüllt ist

(V4) $$\frac{M_3}{M_2}\zeta \leqq \frac{1}{2}$$

dann gilt:

(B7) In den a posteriori-Schranken kann der Faktor 14/9 ersetzt werden durch

$$\mu_n := \frac{7}{6}\min\left(\lambda_n, \frac{4}{3}\right), \quad \text{wo} \quad \lambda_n := \frac{2}{1+\sqrt{1-2\eta_n}} \quad \forall n \in \mathbb{N}_0.$$

Bemerkungen. 1. Die Voraussetzungen lassen sich noch leicht dadurch abschwächen, daß man die beiden Bedingungen $\|f_0^{-1}\| \leqq \beta$ und $\|F^{(j)}(x)\| \leqq K_j$ zu *einer* zusammenfaßt:

$$\|f_0^{-1} F^{(j)}(x)\| \leqq \tilde{K}_j.$$

2. Beim Newton-Verfahren (vgl. [9], Satz 2) erhält man die zu (B7) analoge Aussage ohne eine Zusatzbedingung der Form (V4).

3. (V4) kann ersetzt werden durch die schwächere, jedoch nicht nur Ausgangsdaten enthaltende Bedingung

($\tilde{\text{V}}$4) $$\frac{M_3}{M_2}\zeta_{n-1} \leqq \frac{1}{2}.$$

4. Satz 1 ist mit Ausnahme der Schranken (2.1) und (2.3) schon in [10] mitgeteilt worden.

Beweis. Teil I. Wir beweisen (B1), indem wir durch vollständige Induktion zeigen, daß die obigen Voraussetzungen auch für jeden höheren Index n erfüllt sind, genauer, daß für alle $n \in \mathbb{N}_0$ folgendes gilt

(A1)$_n$ $\exists f_n^{-1}$ mit $\|f_n^{-1}\| \leqq \beta_n$,

(A2)$_n$ $\beta_n M_j \zeta_n^{j-1} \leqq \frac{1}{3 \cdot j!} \quad j=2,3$,

(A3)$_n$ $S_n := \left\{x \in X \,|\, \|x - x_n\| \leqq \frac{14}{9}\zeta_n\right\} \subset S_{n-1} \subset S \quad (S_0 := S_{-1} := S)$.

Aus (A3)$_n$ folgt dann unmittelbar $x_n \in S_n \subset S \quad \forall n \in \mathbb{N}_0$.

(A1)$_0$—(A3)$_0$ sind nach Voraussetzung bzw. per definitionem erfüllt.

(A1)$_{n-1}$—(A3)$_{n-1}$ $\Rightarrow$ (A1)$_n$—(A3)$_n$. (A1)$_n$: Aus (1.4b) folgt mit der Induktionsannahme und (V1)

$$\|d_{n-1}\| \leqq \zeta_{n-1}(1+\beta_{n-1} M_2 \zeta_{n-1}) \leqq \frac{7}{6}\zeta_{n-1}. \tag{2.6}$$

Statt der Existenz von f_n^{-1} zeigen wir zunächst die von $f_{n-1}^{-1} f_n$. Um hierzu ein bekanntes Lemma von Banach anwenden zu können, definieren wir den beschränkten linearen Operator

$$U_n := I - f_{n-1}^{-1} f_n = -f_{n-1}^{-1}(f_n - f_{n-1}).$$

Unter Berücksichtigung von $(A3)_{n-1}$ wenden wir hierauf den Mittelwertsatz an und erhalten unter Beachtung von $(A1)_{n-1}$, (2.6), (V2) und $(A2)_{n-1}$

$$\begin{aligned} \|U_n\| &\leqq \|f_{n-1}^{-1}\| \, \|F'(x_n) - F'(x_{n-1})\| \\ &\leqq \beta_{n-1} \|x_n - x_{n-1}\| \sup_{0 \leqq \vartheta \leqq 1} \|F''(x_{n-1} + \vartheta(x_n - x_{n-1}))\| \\ &\leqq \beta_{n-1} \cdot \frac{7}{6} \cdot \zeta_{n-1} \cdot 2M_2 = \frac{7}{3} \beta_{n-1} M_2 \zeta_{n-1} = \eta_{n-1} \leqq \frac{7}{18} < 1. \end{aligned} \tag{2.7}$$

U_n bildet den Banach-Raum X in sich ab. Wegen der Eigenschaften von U_n existiert nach dem erwähnten Banach-Lemma die beschränkte lineare Inverse $(I - U_n)^{-1} = [f_{n-1}^{-1} f_n]^{-1}$, und es besteht die Ungleichung

$$\|[f_{n-1}^{-1} f_n]^{-1}\| \leqq \frac{1}{1 - \eta_{n-1}}. \tag{2.8}$$

Daraus folgert man leicht die Existenz von f_n^{-1} und wegen $(A1)_{n-1}$ und (2.5)

$$\|f_n^{-1}\| \leqq \|[f_{n-1}^{-1} f_n]^{-1}\| \, \|f_{n-1}^{-1}\| \leqq \frac{\|f_{n-1}^{-1}\|}{1 - \eta_{n-1}} \leqq \frac{\beta_{n-1}}{1 - \eta_{n-1}} = \beta_n. \tag{2.9}$$

Aufgrund von (2.7) gilt ferner

$$\|f_n^{-1}\| \leqq \beta_n \leqq \frac{18}{11} \beta_{n-1} \quad \text{bzw.} \quad \beta_n \leqq \left(\frac{18}{11}\right)^n \beta, \quad \forall n \in \mathbb{N}. \tag{2.10}$$

Wir zeigen $(A2)_n$, indem wir die Ungleichung

$$\zeta_n < \frac{1}{4} \zeta_{n-1} \tag{2.11}$$

beweisen. Denn hieraus folgt $(A2)_n$ sofort mit (2.10) und $(A2)_{n-1}$. Wir haben also $\zeta_n = \|f_n^{-1} F_n\|$, d.h. F_n möglichst scharf durch ζ_{n-1} abzuschätzen. Dazu benutzen wir die Taylorformel (s. diesbezüglich z.B. Collatz [5] und Dieudonné [7]), die wir wegen (V1) und $x_n, x_{n-1} \in S$ anwenden können:

$$\left\| F_n - F_{n-1} - F'_{n-1} d_{n-1} - \frac{1}{2} F''_{n-1} d_{n-1}^{(2)} \right\| \leqq M_3 \|d_{n-1}\|^3.$$

Einarbeitung von (1.4b) in die linke Seite ergibt

$$\left\| F_n - \frac{1}{2} F''_{n-1} d_{n-1}^{(2)} + \frac{1}{2} F''_{n-1} c_{n-1}^{(2)} \right\| \leqq M_3 \|d_{n-1}\|^3.$$

Da F''_{n-1} ein bilinearer Operator ist, gilt

$$F''_{n-1} d_{n-1}^{(2)} - F''_{n-1} c_{n-1}^{(2)} = F''_{n-1} d_{n-1}(d_{n-1} - c_{n-1}) + F''_{n-1}(d_{n-1} - c_{n-1}) c_{n-1},$$

also

$$\|F_n\| \leqq M_3 \|d_{n-1}\|^3 + M_2 \|d_{n-1} - c_{n-1}\| (\|d_{n-1}\| + \|c_{n-1}\|). \tag{2.12}$$

Aus der Differenz von (1.4b) und (1.4a) ergibt sich mit $(A1)_{n-1}$, $(A3)_{n-1}$ und (V1)

$$\|d_{n-1} - c_{n-1}\| \leqq \beta_{n-1} M_2 \zeta_{n-1}^2. \tag{2.13}$$

Damit und mit (2.6)

$$\begin{aligned}\|F_n\| &\leqq \left(\frac{7}{6}\right)^3 M_3 \zeta_{n-1}^3 + \beta_{n-1} M_2^2 \zeta_{n-1}\left(\frac{7}{6}\zeta_{n-1} + \zeta_{n-1}\right) \\ &= \frac{1715}{1728}\frac{8}{5} M_3 \zeta_{n-1}^3 + \frac{13}{6}\beta_{n-1} M_2^2 \zeta_{n-1}^3 < \left(\frac{8}{5} M_3 + \frac{13}{6}\beta_{n-1} M_2^2\right)\zeta_{n-1}^3.\end{aligned} \tag{2.14}$$

Wegen $(A1)_n$ kann man c_n bilden. Mit (2.14), (2.10) und $(A2)_{n-1}$:

$$\begin{aligned}\zeta_n &\leqq \|f_n'^{-1}\|\,\|F_n\| < \frac{18}{11}\beta_{n-1}\left(\frac{8}{5} M_3 + \frac{13}{6}\beta_{n-1} M_2^2\right)\zeta_{n-1}^3 \\ &= \frac{18}{11}\left\{\frac{8}{5}(\beta_{n-1} M_3 \zeta_{n-1}^2) + \frac{13}{6}(\beta_{n-1} M_2 \zeta_{n-1})^2\right\}\zeta_{n-1} \\ &\leqq \frac{18}{11}\left\{\frac{8}{5}\cdot\frac{1}{3\cdot 3!} + \frac{13}{6}\left(\frac{1}{3\cdot 2!}\right)^2\right\}\zeta_{n-1} = \frac{161}{165}\cdot\frac{1}{4}\zeta_{n-1} < \frac{1}{4}\zeta_{n-1}.\end{aligned}$$

$(A3)_n$: Unter Berücksichtigung von (2.6), (2.11) und der Definition der Kugel S_n gilt

$$\begin{aligned}\|x - x_{n-1}\| &\leqq \|x - x_n\| + \|x_n - x_{n-1}\| \\ &\leqq \frac{14}{9}\zeta_n + \|d_{n-1}\| < \frac{14}{9}\cdot\frac{1}{4}\zeta_{n-1} + \frac{7}{6}\zeta_{n-1} = \frac{14}{9}\zeta_{n-1} \quad \forall\, x \in S_n,\end{aligned}$$

d.h. $S_n \subset S_{n-1}$, also wegen $(A3)_{n-1}$ auch $S_n \subset S$.

(B2): Aus (2.11) folgt

$$\zeta_{n+j} \leqq \frac{1}{4^j}\zeta_n \leqq \frac{5}{4^{n+j}}, \quad \forall\, n, j \in \mathbb{N}_0, \tag{2.15}$$

d.h. $\{\zeta_n\}$ ist eine Nullfolge. $\{x_n\}$ ist eine Cauchy-Folge, denn mit (2.6) und (2.15) wird

$$\|x_{n+p} - x_n\| \leqq \sum_{j=0}^{p-1}\|d_{n+j}\| \leqq \frac{7}{6}\sum_{j=0}^{p-1}\zeta_{n+j} \leqq \frac{7}{6}\zeta_n \sum_{j=0}^{p-1}\frac{1}{4^j} < \frac{14}{9}\zeta_n, \tag{2.16}$$

$$\forall\, n \in \mathbb{N}_0 \wedge \forall\, p \in \mathbb{N}.$$

Durch Grenzübergang folgt hieraus

$$\|x^* - x_n\| \leqq \frac{14}{9}\zeta_n, \quad \forall\, n \in \mathbb{N}_0. \tag{2.17}$$

Deswegen und wegen der Abgeschlossenheit von S_n gilt $x^* = \lim x_n \in S_n$ und damit wegen (B1) auch $x^* \in S$.

(B3): **Definition.** Ein Iterationsverfahren $\{x_n \xrightarrow{F} x^*\}$ zur näherungsweisen Lösung der Gleichung $F(x) = \Theta$ heißt (in einem Banach-Raum X) *von der Ordnung k konvergent*, falls die Ungleichung

$$\|x^* - x_n\| \leqq P_k \|x^* - x_{n-1}\|^k, \quad \forall\, n \in \mathbb{N}$$

mit einer von n unabhängigen Zahl P_k erfüllt ist. Es heißt *schwach von der Ordnung k konvergent*, falls P_k ein Polynom mit von n unabhängigen positiven Koeffizienten in $\|f_{n-1}'^{-1}\|$ ist, dessen Grad von der Ordnung k und vom Verfahren abhängt.

Falls die f_n^{-1} gleichmäßig beschränkt sind, so konvergiert das Verfahren von der Ordnung k (das ist hier mit (V3) für $k=3$ der Fall). Im anderen Falle kann es vorkommen, daß das Verfahren von geringerem Grade als k konvergiert (vgl. Satz 3). Sei z.B. $X:=\mathbb{R}$ und $F(x):=x^2$. Mit $x_0:=\frac{1}{2}$ konvergiert das Newton-Verfahren nach [9], Satz 2 ($\zeta=\frac{1}{4}$; $S\colon 0\leqq x\leqq 1$; $\beta=1$; $K_2=2$; $\beta K_2\zeta=\frac{1}{2}$) gegen $x^*=0$ und man hat

$$|x^*-x_n|\leqq\frac{1}{2}K_2|f_{n-1}^{-1}\|x^*-x_{n-1}\|^2=\frac{1}{2}\cdot 2\cdot\frac{1}{2x_{n-1}}|0-x_{n-1}|^2=\frac{1}{2}|x^*-x_{n-1}|^1.$$

Die Bezeichnung „schwach" ist also sinnvoll.

Mit (1.4b), $(A1)_{n-1}$, (V1), der Taylorformel und $d_{n-1}^*:=x^*-x_{n-1}$ erhält man

$$\begin{aligned}\|x^*-x_n\|&=\left\|f_{n-1}^{-1}\left(F_{n-1}+F'_{n-1}d_{n-1}^*+\frac{1}{2}F''_{n-1}c_{n-1}^{(2)}\right)\right\|\\&\leqq\|f_{n-1}^{-1}\|\left\{\left\|F(x^*)-F_{n-1}-F'_{n-1}d_{n-1}^*-\frac{1}{2}F''_{n-1}d_{n-1}^{*(2)}\right\|\right.\\&\quad\left.+\frac{1}{2}\|F''_{n-1}d_{n-1}^*(d_{n-1}^*-c_{n-1})+F''_{n-1}(d_{n-1}^*-c_{n-1})c_{n-1}\|\right\}\\&\leqq\|f_{n-1}^{-1}\|\{M_3\|d_{n-1}^*\|^3+M_2\|d_{n-1}^*-c_{n-1}\|(\|d_{n-1}^*\|+\|c_{n-1}\|)\}.\end{aligned}$$

Analog

$$\|d_{n-1}^*-c_{n-1}\|\leqq\|f_{n-1}^{-1}\|\,\|F(x^*)-F_{n-1}-F'_{n-1}d_{n-1}^*\|\leqq\beta_{n-1}M_2\|d_{n-1}^*\|^2.$$

Hiermit ergibt sich unter Berücksichtigung von $\|c_{n-1}\|\leqq\|d_{n-1}^*-c_{n-1}\|+\|d_{n-1}^*\|$

$$\|x^*-x_n\|\leqq M_{3,n-1}\cdot\|x^*-x_{n-1}\|^3,\tag{2.18}$$

wobei

$$M_{3,n-1}:=\|f_{n-1}^{-1}\|M_3+(\|f_{n-1}^{-1}\|M_2)^2(2+\|f_{n-1}^{-1}\|M_2\|x^*-x_{n-1}\|.$$

(B3) folgt hieraus und aus der gleichmäßigen Beschränktheit von x^*-x_{n-1} und f_{n-1}^{-1}. Letzteres sieht man analog Beweisteil $(A1)_n$ mit $U:=I-f_0^{-1}F'(x)\ \forall x\in S$.

(B4) ist klar.

(B5): Indirekter Beweis. Annahme: $\tilde{x}$ sei eine zweite, von x^* verschiedene Lösung von $F(x)=\Theta$ in S. Wir zeigen, daß $\|\tilde{x}-x_n\|\to 0$, woraus dann wegen der Eindeutigkeit des Grenzwertes $\lim x_n$ im Widerspruch zur Annahme $\tilde{x}=x^*$ folgt. Wir beweisen nun mit Hilfe von (2.18) die Ungleichung

$$(E)_n\qquad\|\tilde{x}-x_n\|<\frac{14}{9}\cdot\frac{\zeta}{4^n},\quad\forall n\in\mathbb{N}_2.$$

Man beachte, daß (2.18) für *irgendeine* Lösung x^* von $F(x)=\Theta$ in S gilt, also auch für $\tilde{x}$. Um $\|\tilde{x}-x_2\|$ nach (2.18) abzuschätzen, muß man $\|\tilde{x}-x_1\|$ selbst oder eine obere Schranke dafür kennen. Aus der Annahme $\tilde{x}\in S$ und (V3) folgt durch scharfe Abschätzung

$$M_{3,0}\zeta^2\leqq(\beta M_3\zeta^2)+(\beta M_2\zeta)^2\left(2+\beta M_2\zeta\cdot\frac{14}{9}\right)<\frac{2}{15}$$

$$\|\tilde{x}-x_1\|\leqq M_{3,0}\left(\frac{14}{9}\zeta\right)^3<\frac{7}{5}\cdot\frac{14}{9}\,\frac{\zeta}{4}.$$

Dies in $M_{3,1}$ eingesetzt, ergibt mit (V3) und (2.10)

$$M_{3,1}\zeta^2 \leqq \frac{18}{11}\,(\beta M_3 \zeta^2) + \left(\frac{18}{11}\beta M_2 \zeta\right)^2 \left(2 + \frac{18}{11}\beta M_2 \cdot \frac{7}{5} \cdot \frac{14}{9} \cdot \frac{\zeta}{4}\right) < \frac{3}{11}$$

und damit schließlich $(\mathrm{E})_2$:

$$\|\tilde{x} - x_2\| \leqq M_{3,1}\left(\frac{7}{5} \cdot \frac{14}{9}\,\frac{\zeta}{4}\right)^2 < \frac{14}{9} \cdot \frac{\zeta}{4^2}.$$

$(\mathrm{E})_{n-1} \Rightarrow (\mathrm{E})_n$: Mit (2.10), $(\mathrm{E})_{n-1}$, (2.18) und (V3) wird

$$M_{3,n-1} \leqq \left(\frac{18}{11}\right)^{n-1} \beta M_3 \zeta^2 + \left[\left(\frac{18}{11}\right)^{n-1} \beta M_2 \zeta\right]^2 \left[2 + \left(\frac{18}{11}\right)^{n-1} \beta M_2 \frac{14}{9}\,\frac{\zeta}{4^{n-1}}\right]$$

$$< \frac{2}{15}\,4^{n-1}, \quad \forall n \in \mathbb{N}_3,$$

$$\|\tilde{x} - x_n\| \leqq M_{3,n-1}\left(\frac{14}{9} \cdot \frac{\zeta}{4^{n-1}}\right)^3 < \frac{3}{32} \cdot \frac{14}{9} \cdot \frac{\zeta}{4^n} < \frac{14}{9}\,\frac{\zeta}{4^n}.$$

(B6): Die a-posteriori-Schranke (2.3) folgt unmittelbar aus $\zeta_n \leqq \beta_n \|F_n\|$, (2.12) und (2.17). Die anderen beiden Schranken erhält man hieraus durch weitere Abschätzung. Um (2.2) herzuleiten, schätzen wir die in (2.12) vorkommende Differenz $d_{n-1} - c_{n-1}$ wie folgt aus der Differenz von (1.4b) und (1.4a) ab:

$$F'_{n-1}(d_{n-1} - c_{n-1}) = -\frac{1}{2} F''_{n-1} c_{n-}^{(2)} = \frac{1}{2}\,[F''_{n-1} d_{n-1}(d_{n-1} - c_{n-1})$$

$$+ F''_{n-1}(d_{n-1} - c_{n-1})\, c_{n-1}] - \frac{1}{2} F''_{n-1} d_{n-1}^{(2)},$$

$$\|d_{n-1} - c_{n-1}\| \leqq \beta_{n-1} M_2 (\|d_{n-1}\| + \|c_{n-1}\|)\, \|d_{n-1} - c_{n-1}\| + \beta_{n-1} M_2 \|d_{n-1}\|^2.$$

Unter Beachtung von (2.7) erhält man hieraus mit (2.6)

$$\|d_{n-1} - c_{n-1}\| \leqq \frac{\beta_{n-1} M_2 \|d_{n-1}\|^2}{1 - \frac{13}{6}\beta_{n-1} M_2 \zeta_{n-1}} < \beta_n M_2 \|d_{n-1}\|^2.$$

Hiermit und mit $\|c_{n-1}\| \leqq \|d_{n-1} - c_{n-1}\| + \|d_{n-1}\|$, (2.12), (2.10), (2.7) und $(\mathrm{A}2)_{n-1}$:

$$\|F_n\| \leqq M_3 \|d_{n-1}\|^3 + \beta_n M_2^2 \|d_{n-1}\|^2 [2\|d_{n-1}\| + \beta_n M_2 \|d_{n-1}\|^2]$$

$$\leqq \left\{M_3 + \beta_n M_2^2 \left[2 + \frac{18}{11} \cdot \frac{7}{6}\,(\beta_{n-1} M_2 \zeta_{n-1})\right]\right\} \|d_{n-1}\|^3$$

$$< \left\{M_3 + \frac{7}{3}\beta_n M_2^2\right\} \|d_{n-1}\|^3.$$

Die Gültigkeit der a-priori-Schranke (2.1) ist wegen (2.17) gesichert, sobald wir die Ungleichung

$$(\mathrm{S})_n \qquad \zeta_n \leqq \left(\frac{11}{18}\right)^n \zeta (\beta\gamma\zeta)^{3^n - 1}, \quad \forall n \in \mathbb{N}_0$$

bewiesen haben, was durch vollständige Induktion geschieht. $\zeta_0 = \zeta$ per definitionem.

$(\mathrm{S})_{n-1} \Rightarrow (\mathrm{S})_n$: Mit $\beta_n \geqq \beta_{n-1} \geqq \cdots \geqq \beta$, (2.14), (2.10) und $(\mathrm{S})_{n-1}$ ergibt sich

$$\zeta_n \leqq \beta_n \beta_{n-1}\left(\frac{13}{6} M_2^2 + \frac{8}{5} \cdot \frac{M_2}{\beta}\right) \zeta_{n-1}^3 \leqq \left(\frac{18}{11}\right)^{2n-3} \beta^2 \gamma^2 \left[\frac{\zeta}{\left(\frac{18}{11}\right)^{n-1}} (\beta\zeta\gamma)^{3^{n-1}-1}\right]^3$$

$$= \left(\frac{11}{18}\right)^n \zeta (\beta\gamma\zeta)^{3^n - 1}.$$

Teil II. Zum Nachweis von (B7) zeigen wir zunächst die Gültigkeit der folgenden für alle $n \in \mathbb{N}$ geltenden Ungleichungen

$$\|F_n\| < \frac{7}{6} M_2 \zeta_{n-1}^2, \tag{2.19}$$

$$\eta_n < \frac{1}{2}\left(\frac{\eta_{n-1}}{1-\eta_{n-1}}\right)^2, \tag{2.20}$$

$$\zeta_{n-1} < \zeta_{n-1}\lambda_{n-1} - \zeta_n \lambda_n. \tag{2.21}$$

Aus (2.14) folgt mit $(A2)_{n-1}$, (2.16) und (V4) bzw. $(\tilde{V}4)$

$$\begin{aligned}\|F_n\| &< \frac{7}{6} M_2 \left\{\frac{13}{6}\cdot\frac{6}{7}(\beta_{n-1} M_2 \zeta_{n-1}) + \frac{8}{5}\cdot\frac{6}{7}\left(\frac{M_3}{M_2}\zeta_{n-1}\right)\right\}\zeta_{n-1}^2 \\ &\leqq \frac{209}{210}\cdot\frac{7}{6} M_2 \zeta_{n-1}^2 < \frac{7}{6} M_2 \zeta_{n-1}^2.\end{aligned}$$

Mit (2.19) ergibt sich (2.20) so:

$$\eta_n = \frac{7}{3}\beta_n M_2 \zeta_n < \frac{7}{3}\|F_n\|\beta_n^2 M_2 < \frac{(\frac{7}{3})^2}{2} M_2^2 \zeta_{n-1}^2 \left(\frac{\beta_{n-1}}{1-\eta_{n-1}}\right)^2 = \frac{1}{2}\left(\frac{\eta_{n-1}}{1-\eta_{n-1}}\right)^2.$$

Zum Beweis von (2.21) drücken wir $\zeta_n \lambda_n$ durch $\zeta_{n-1}\lambda_{n-1}$ aus:

$$\begin{aligned}\zeta_n \lambda_n &\leqq \beta_n \|F_n\| \lambda_n < \frac{1}{2}\frac{\frac{5}{3}\beta_{n-1} M_2 \zeta_{n-1}}{1-\eta_{n-1}} \zeta_{n-1}\lambda_n \\ &< \frac{1}{2}\frac{\eta_{n-1}\zeta_{n-1}}{1-\eta_{n-1}}\cdot\frac{2}{1+\sqrt{1-\left(\frac{\eta_{n-1}}{1-\eta_{n-1}}\right)^2}} = \zeta_{n-1}\lambda_{n-1} - \zeta_{n-1}.\end{aligned}$$

Mit Ungleichung (2.21) ergibt sich statt (2.16)

$$\|x_{n+p} - x_n\| \leqq \frac{7}{6}\sum_{j=0}^{p-1}\zeta_{n+j} < \frac{7}{6}(\zeta_n\lambda_n - \zeta_{n+p}\lambda_{n+p}), \quad \forall n \in \mathbb{N}_0 \wedge \forall p \in \mathbb{N}, \tag{2.22}$$

woraus durch Grenzübergang $p \to \infty$

$$\|x^* - x_n\| \leqq \tfrac{7}{6}\lambda_n \zeta_n, \quad \forall n \in \mathbb{N}_0 \tag{2.23}$$

folgt. Zusammen mit (2.17) wird daraus

$$\|x^* - x_n\| \leqq \mu_n \zeta_n. \tag{2.24}$$

3. Bemerkungen zu den Fehlerschranken

Die Schranke (2.1) ist als a-priori-Schranke die gröbste. Sie ist bei Operatorgleichungen zur numerischen Abschätzung des Fehlers der Näherungslösungen im allgemeinen unbrauchbar (vgl. Tabelle 2 und 3). Die Gestalt der a-posteriori-Schranke (2.2) ist analog der Ungleichung, mit deren Hilfe die Ordnung des Iterationsverfahrens definiert wird: Man erkennt an dem Exponenten der Norm auf der rechten Seite sofort die (unter gewissen Voraussetzungen gesicherte) Ordnung des Verfahrens. Diese Schranke kann man unter Berücksichtigung von (B7) als direkte Verallgemeinerung der a-posteriori-Schranke beim Newton-Verfahren (vgl. Döring [9], Satz 2) ansehen. Die a-posteriori-Schranke (2.3) ist in Verbindung

mit (B7) die schärfste der bewiesenen Schranken. Ihre Gestalt entspricht dem spezifischen Aufbau des vorliegenden Iterationsverfahrens (1.4). Durch (B7) werden die Schranken um maximal den Faktor $\frac{1}{3}$ verkleinert, denn nach einigen Iterationsschritten ist $\eta_n \ll 1$ und damit $\lambda_n \approx 1$.

Die Bedingungen (V4) bzw. ($\tilde{\text{V}}$4) sind im allgemeinen gut erfüllt; denn wenn schon (V4) nicht erfüllt sein sollte, so ist es ($\tilde{\text{V}}$4), da nach einigen Schritten $\zeta_{n-1} \ll 1$ ist. Diese Voraussetzungen bedeuten also gegenüber (V1)—(V3) keine wesentliche zusätzliche Einschränkung.

Zur Berechnung der Fehlerschranken werden im wesentlichen nur Größen verwendet, die zur iterativen Lösung der Aufgabe sowieso benötigt werden und die bis hin zur Berechnung von x_n bereits vorliegen. Es sind lediglich einige nicht ins Gewicht fallende Zusatzrechnungen wie die parallele rekursive Berechnung der β_n und η_n erforderlich.

Für die nach (B7) zu berechnenden Schranken wird η_n und damit ζ_n, d.h. F und F' an der „letzten" Näherungsstelle x_n benötigt. Dies läßt sich vermeiden, ohne die numerische Qualität dieser a-posteriori-Schranken zu beeinträchtigen. Man kann nämlich wegen $\eta_{n-1} \ll 1$ die Zahl η_n nach (2.20) günstig durch η_{n-1} abschätzen.

Die Fehlerschranken lassen sich noch ein wenig verbessern, indem man die Voraussetzungen (V1) und (V2) mit nicht getrennten Normen formuliert (vgl. Bem. 1 zu Satz 1), und indem man beachtet, daß es für die Schranken genügt, die höheren Ableitungen abzuschätzen über

$$\hat{S} := \left\{x \in X \,\middle|\, \|x - x_1\| \leqq \frac{\zeta}{2}\right\} \cup \{x := x_0 + \vartheta(x_1 - x_0) \,|\, 0 \leqq \vartheta \leqq 1\} \subset S.$$

Falls man in den Schranken mit $\|f_n^{-1}\|$ selbst und nicht mit den oberen Schranken β_n hierfür rechnen möchte — dies ist sicher im Eindimensionalen der Fall — so tritt überall $\|f_n^{-1}\|$ an die Stelle von β_n mit Ausnahme von (2.4), wo dann $\min_{m=0(1)n-1} \|f_m^{-1}\|$ statt β gesetzt wird.

Nicht nur im Hinblick auf die Voraussetzungen (V3), sondern auch bezüglich einer günstigeren Schranke erweist es sich in Funktionenräumen oft als nützlich, den Raum L_2 gegenüber dem Raum C den Vorzug zu geben (vgl. hierzu Tabelle 2 und die Beispiele zum Newton-Verfahren bei Kantorovich [18]).

4. Mögliche Abschwächungen der Voraussetzungen

Im Hinblick auf den zu Satz 1 analogen, auf Kantorowitsch zurückgehenden Satz über das Newton-Verfahren (s. z.B. Antosiewicz und Rheinboldt [3], Döring [9], Kantorowitsch und Akilov [19]) stellt sich die Frage, ob man in (V3) mit $\beta K_2 \zeta \leqq \frac{1}{2}$ auskommt. Das ist nicht der Fall. Nachdem wir dies anschließend gezeigt haben, beweisen wir im nächsten Abschnitt einen Satz, bei dem einerseits die Schranke von $\beta K_2 \zeta$ möglichst groß ist, der aber andererseits im Sinne der in der Einleitung formulierten Forderungen gerade eben noch vernünftig ist. In diesem Sinne ist der Satz 3 ein „Grenz"-Satz. Die dabei erhaltenen Schranken — der Kürze halber wird nur *eine* a-posteriori-Schranke angegeben — sind etwas gröber als in Satz 1 und die Abschätzung in (V1) ist — insbesondere in Funktionenräumen — wegen des Kugelmittelpunktes $\tilde{x}_1$ statt x_0 umständlicher zu

handhaben als in Satz 1. Deshalb wird Satz 1 als „Standard"-Satz bezeichnet. Die angekündigten Aussagen über Abschwächungen von (V3) werden zusammengefaßt in

Satz 2. Es seien die Voraussetzungen (V1) und (V2) von Satz 1 — die Kugel S aber mit noch nicht spezifiziertem Radius $\lambda\zeta$ — und

$$\beta K_j \zeta^{j-1} \leqq \sigma_j \quad (j=2,3) \tag{4.1}$$

mit gewissen positiven Zahlen λ, σ_j erfüllt. Dann ist die Ungleichung

$$\sigma_3\left(1+\frac{\sigma_2}{2}\right)^3 \leqq 6\left\{\left[1-\sigma_2\left(1+\frac{\sigma_2}{2}\right)\right]^2-\left(2+\frac{\sigma_2}{2}\right)\left(\frac{\sigma_2}{2}\right)^2\right\} \tag{4.2}$$

unabhängig von der Größe von λ eine notwendige Bedingung für die Durchführung des Beweises der Behauptungen (B1)—(B4) und Fehlerabschätzungen der Gestalt (B6) von Satz 1 in der dortigen Form. Insbesondere ist der Beweis in diesem Sinne unter den getroffenen Voraussetzungen mit $\beta K_2 \zeta \leqq \frac{1}{2}$ nicht möglich.

Beweis. Wir setzen $\varkappa_j := \sigma_j/j!$ $(j=2,3)$ und zeigen, daß wenn der Beweis von Satz 1 im dortigen Sinne durchführbar sein soll, $\varkappa_2$ und $\varkappa_3$ der zu (4.2) äquivalenten Ungleichung

$$\varkappa_3 \leqq \frac{[1-2\varkappa_2(1+\varkappa_2)]^2-(2+\varkappa_2)\,\varkappa_2^2}{(1+\varkappa_2)^3} \tag{4.3}$$

genügen muß. Dazu vollziehen wir den Beweis von (B1) nach, jetzt aber mit (4.1) statt (V3). Wir haben statt $(A2)_n$

$$(A2)'_n \qquad \beta_n M_j \zeta_n^{j-1} \leqq \varkappa_j \quad (j=2,3),$$

statt (2.6):

$$\|d_{n-1}\| \leqq (1+\varkappa_2)\,\zeta_{n-1}$$

statt (2.10):

$$\beta_n \leqq \frac{\beta_{n-1}}{1-2\varkappa_2(1+\varkappa_2)} \tag{4.4}$$

und statt (2.14)

$$\|F_n\| \leqq \{M_3(1+\varkappa_2)^3+\beta_{n-1}M_2^2(2+\varkappa_2)\}\zeta_{n-1}^3. \tag{4.5}$$

Dabei muß

$$\eta := 2\varkappa_2(1+\varkappa_2) < 1, \tag{4.6}$$

d.h. $\varkappa_2 < \frac{1}{2}(\sqrt{3}-1) \doteq 0{,}366$ [3] sein. Also

$$\zeta_n \leqq \frac{(1+\varkappa_2)^3\varkappa_3+(2+\varkappa_2)\,\varkappa_2^2}{1-2\varkappa_2(1+\varkappa_2)} \cdot \zeta_{n-1}. \tag{4.7}$$

Um die Gültigkeit von $(A2)_n$ nachzuweisen, haben wir im Beweis von Satz 1 $\beta_n M_2 \zeta_n$ durch $\beta_{n-1} M_2 \zeta_{n-1}$ abgeschätzt und $(A2)_{n-1}$ benutzt. Dabei muß der bei der Abschätzung entstehende Vorfaktor kleiner oder gleich Eins sein. Dies führt mit (4.4) statt (2.10) und (4.7) statt (2.11) nach einer leichten Umformung unmittelbar auf die Ungleichung (4.3). Für $\beta K_2 \zeta \leqq \frac{1}{2}$, d.h. $\varkappa_2 = \frac{1}{4}$ wird die rechte

3 $\doteq$ bedeutet stets *Ab*rundung, z.B. $e \doteq 2{,}71$.

Seite von (4.3), also auch σ_3 gleich Null. Dies ist aber nach Voraussetzung ausgeschlossen (weil eine obere Schranke 0 als Forderung in (V3) sinnlos wäre).

In Fig. 1 ist die Menge derjenigen Wertepaare (σ_2, σ_3) als Kurve eingetragen, für die in (4.2) das Gleichheitszeichen gilt. Für alle unterhalb dieser Grenzkurven liegenden Paare (σ_2, σ_3) ist (4.2) ebenfalls erfüllt. Während das für den folgenden Grenzsatz benutzte (σ_2, σ_3)-Paar $(\frac{2}{5}, \frac{3}{5})$ praktisch auf der Grenzkurve in Fig. 1 liegt, ist bei dem Standardsatz mit $(\frac{1}{3}, \frac{1}{3})$ für eine oder beide Schranken in (V3) noch Reserve gelassen worden (vgl. Fig. 1). Wie schon bemerkt, stellt ja (4.2) nur eine notwendige Bedingung dar, zu deren Herleitung nur *eine* Stelle im Beweis von Satz 1 benutzt wurde. Von der Wahl des Paares (σ_2, σ_3) hängt es

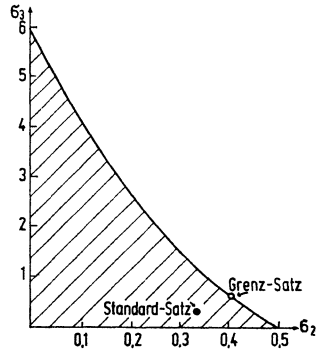

Fig. 1. Zulässige Kombinationen der Schranken in $\beta K_j \zeta^{j-1} \leqq \sigma_j$ $(j=2, 3)$ (schraffiert)

noch ab, *ob* die Behauptungen (B1)—(B5) alle nachgewiesen werden können und wie die Fehlerabschätzungen ausfallen. (4.2) bzw. Fig. 1 bieten dem Leser die Möglichkeit, in Verbindung mit dem bisher Gesagten durch Wahl von geeigneten Paaren (σ_2, σ_3) noch andere Varianten von Satz 1 über TV aufzustellen. Dazu könnte z.B. Veranlassung bestehen, wenn $K_2 \ll K_3$ ist oder umgekehrt. Es sei nur noch bemerkt, daß eine weitere Annäherung von σ_2 an $\frac{1}{2}$ bewirkt, daß die Fehlerschranken, der Radius $\lambda\zeta$ von S und damit K_2, K_3 sowie der Faktor in der Ungleichung zwischen β_n und β_{n-1} vergrößert werden. Der letztere macht dann bald einen Nachweis der Eindeutigkeit in S ohne zusätzliche Voraussetzungen unmöglich. Bei dem Grenzsatz muß für die Eindeutigkeit bereits zusätzlich $\|f_1^{-1}\| \leqq \beta$ gefordert werden, wenn der Beweis analog zu Satz 1 geführt werden soll.

5. Grenzsatz über das Tschebyscheff-Verfahren

Satz 3 (Grenzsatz über TV).

Teil I. Existenz einer Lösung. Konvergenz und Ordnung des Verfahrens. Schranken.

Falls $x_0 \in X_F$ so gewählt werden kann, daß (V1)—(V4) gilt,

(V1) $\exists \tilde{S} := \left\{x \in X \,|\, \|x - \tilde{x}_1\| \leqq \frac{3}{2}\zeta\right\}$, wo F dreimal Fréchet-differenzierbar ist mit $\tilde{x}_1 := x_0 + c_0$ und $\|F^{(j)}(x)\| \leqq K_j$, $\forall x \in \tilde{S}$ und $j = 2, 3$,

(V2) $\exists\,[F'(x_0)]^{-1}$ mit $\|[F'(x_0)]^{-1}\| \leqq \beta$,

(V3) $\beta K_2 \zeta \leqq \frac{2}{5}$,

(V4) $\beta K_3 \zeta^2 \leqq \frac{3}{5}$,

dann

(B1) existiert die durch (1.4) definierte Folge $\{x_n\}$ (TV), und es ist $x_n \in S \quad \forall n \in \mathbb{N}_0$. Ferner gilt:

(B2) Die Folge $\{x_n\}$ konvergiert, und zwar gilt $x^* := \lim x_n \in S$.

(B3) TV konvergiert schwach von der Ordnung 3.

(B4) $F(x^*) = \Theta$.

(B5) Es gelten die Fehlerabschätzungen

$$\|x^* - x_n\| \leqq \frac{5}{2}\left(\frac{13}{25}\right)^n \zeta (\beta\gamma\zeta)^{3^n-1}, \quad \forall n \in \mathbb{N}_0,$$

$$\|x^* - x_n\| \leqq \frac{5}{2}\beta_n\left(M_3 + \frac{5}{2}\beta_n M_2^2\right)\|x_n - x_{n-1}\|^3, \quad \forall n \in \mathbb{N},$$

wobei β_n und η_n wie in (2.5) definiert sind und

$$\gamma := 2M_2 \cdot \sqrt{\frac{11}{5} + \frac{7}{4}\frac{M_3}{\beta M_2^2}}; \quad \eta_n := \frac{12}{5}\beta_n M_2 \zeta_n, \quad \forall n \in \mathbb{N}_0.$$

Teil II. Eindeutigkeit der Lösung.

Falls außer (V1)—(V4) noch gilt

($\tilde{\text{V}}$2) $$\|[F'(x_1)]^{-1}\| \leqq \beta$$

dann

(B6) ist x^* in S eindeutig.

Teil III. Verschärfung der a-posteriori-Schranke.

Falls außer (V1)—(V4) folgendes gilt

(V5) $$\frac{M_3}{M_2}\zeta \leqq \frac{3}{7},$$

dann gilt mit der in Satz 1 definierten Zahl λ_n sogar die schärfere a-posteriori-Schranke

(B7) $$\|x^* - x_n\| \leqq \frac{6}{5}\lambda_n\beta_n[M_3 + 2\beta_n M_2^2(1 + 3\eta_{n-1})]\|x_n - x_{n-1}\|^3, \quad \forall n \in \mathbb{N}.$$

Die Bemerkungen zu Satz 1 gelten hier sinngemäß ebenfalls. Der *Beweis* von Satz 3 verläuft im Prinzip analog zu dem von Satz 1.

6. Beispiele

Im folgenden wenden wir den Standardsatz auf einige Beispiele an.

Beispiel 1. $F(x) := x^3 - 10 = 0$. $x_0 := 2$. $X := \mathbb{R}$. Man berechnet $\beta = \frac{1}{12}$; $\zeta = \frac{1}{6}$ $S := \{x \in \mathbb{R} \mid 1{,}74 \leqq x \leqq 2{,}26\}$; $M_2 = 6{,}78$; $M_3 = 1$. Hiermit sind (V3) und (V4) erfüllt. Also existiert nach Satz 1 im Intervall $1{,}74 \leqq x \leqq 2{,}26$ genau eine Lösung

der Gleichung $x^3-10=0$ und TV konvergiert mit $x_0=2$ von der Ordnung 3 gegen diese Lösung x^*. In Tabelle 1 sind die neuen Schranken für den Abstand der so errechneten Näherung x_2 von x^* bekannten Schranken und dem tatsächlichen Fehler gegenübergestellt.

Tabelle 1. *Vergleich verschiedener oberer Schranken für den Fehler des nach TV berechneten Näherungswertes x_2 der in $1{,}74 \leqq x \leqq 2{,}26$ liegenden eindeutigen Lösung*[a] $x^* \fallingdotseq 2{,}154434\,6900$ *von* $F(x) := x^3 - 10 = 0$ *mit* $x_0 := 2$

Autor	Lit.-Quelle	$10^9 \cdot \lvert x^* - x_2 \rvert \leqq$
Collatz (s. auch Ade)	[6] ([1])	6090
		29
Ehrmann	[13], Satz 13	Bedingung nicht erfüllt
Jankó	[15]	Bedingung nicht erfüllt
Kaazik	[16], Satz 1	71,7
Nečepurenko	[21]	113 200 000
Shafiyev	[23], Satz 2	71,2
Ulm	[24], Satz 3	122,5
Xu Li-Tse	[28], Satz 3	19000
Döring	(2.1)	1344
	(2.2)	4,43
	(2.3) + (B 7)	2,92
Wirklicher Fehler	$\lvert x^* - x_2 \rvert \fallingdotseq$	1,64

[a] $\fallingdotseq$ bedeutet die übliche Rundung; $\doteq$ stets *Auf*rundung, z.B. $\pi \doteq 3{,}15$.

Bemerkung. Im Gegensatz zu den hier angegebenen Sätzen ist der zugrunde zu legende Bereich bei Collatz durch die Ausgangsnäherung x_0 allein noch nicht festgelegt, sondern muß seinerseits passend gewählt werden. Der Wert 6090 entspricht der wie oben gewählten Kugel S. Für den Schrankenwert 29 ist der konvexe Bereich $S_c \subset S$ so gewählt worden, daß die Schranke kleinstmöglich wurde.

Beispiel 2. Lösung der einfachen nichtlinearen Integralgleichung vom Hammerstein-Typ

$$x(s) + \frac{1}{2}\int_0^1 s \cos x(t)\, dt = s. \tag{6.1}$$

Der Operator F und seine ersten beiden Fréchet-Ableitungen sind gegeben durch

$$[F\,x]\,(s) := x(s) - s + \frac{1}{2}\int_0^1 s \cos x(t)\, dt,$$

$$[F'(x)\,z]\,(s) = z(s) - \frac{1}{2}\int_0^1 s \sin x(t)\, z(t)\, dt,$$

$$[F''(x)\,z_1 z_2]\,(s) = -\frac{1}{2}\int_0^1 s \cos x(t)\, z_1(t)\, z_2(t)\, dt.$$

Das TV besteht hier darin, bei jedem Iterationsschritt das folgende Paar von linearen Fredholmschen Integralgleichungen zweiter Art einmal zu lösen ($n \in \mathbb{N}_0$):

$$c_n(s) - \frac{s}{2} \int_0^1 \sin x_n(t)\, c_n(t)\, dt = g_n(s) \Rightarrow c_n(s) \quad \big(\tilde{x}_{n+1}(s) := x_n(s) + c_n(s)\big), \tag{6.2}$$

$$d_n(s) - \frac{s}{2} \int_0^1 \sin x_n(t)\, d_n(t)\, dt = h_n(s) \Rightarrow d_n(s) \Rightarrow x_{n+1}(s) := x_n(s) + d_n(s), \tag{6.3}$$

wobei

$$g_n(s) := -x_n(s) + s - \frac{s}{2} \int_0^1 \cos x_n(t)\, dt,$$

$$h_n(s) := -x_n(s) + s - \frac{s}{2} \int_0^1 \left[1 - \frac{1}{2} c_n^2(t)\right] \cos x_n(t)\, dt.$$

Dabei ist es naheliegend, die rechte Seite als Ausgangsnäherung zu wählen: $x_0(s) := s$. Dann wird $\tilde{x}_n(s) = \tilde{\xi}_n \cdot s$ und $x_n(s) = \xi_n \cdot s$. Praktisch am einfachsten zu handhaben ist der Raum $X := C[0, 1]$ der über dem Intervall $[0, 1]$ stetigen Funktionen mit der Maximum-Norm. (V1) ist hierfür erfüllt. Man erhält

$$\zeta = \|\tilde{x}_1 - x_0\| = \max_{0 \le s \le 1} |\tilde{\xi}_1 \cdot s - \xi_0 \cdot s| = \tilde{\xi}_1 - \xi_0 \doteq 0{,}495$$

$$\Rightarrow S := \left\{x \in C[0, 1] \,\middle|\, \|x - x_0\| \le \frac{4}{5}\right\}; \quad M_2 = \frac{1}{4}; \quad M_3 = \frac{1}{12}.$$

$\exists\, f_0^{-1}$? Wir bezeichnen den Integraloperator mit dem Kern $\frac{s}{2} \sin x(t)$ mit K'_x:

$$[K'_x z](s) := \int_0^1 \frac{s}{2} \sin x(t)\, z(t)\, dt.$$

Dann gilt

$$K'_x : C[0, 1] \to C|0, 1|, \quad F'(x) = I - K'_x$$

und

$$\|K'_{x_0}\| = \max_s \frac{s}{2} \int_0^1 \sin t\, dt = \frac{1 - \cos 1}{2} \doteq 0{,}23 < 1.$$

Also folgt aus dem Banach-Lemma die Existenz der beschränkten Inversen $f_0^{-1} = (I - K'_{x_0})^{-1}$ mit

$$\|f_0^{-1}\| \le \frac{1}{1 - \|K'_{x_0}\|} \doteq \frac{1}{1 - 0{,}23} \doteq 1{,}3 =: \beta$$

(V3) und (V4) sind erfüllt. Demnach existiert in dem in Fig. 2 schraffierten Bereich genau eine in $[0, 1]$ stetige Lösung $x^*(s)$ von (6.1), TV konvergiert von

der Ordnung 3 gegen $x^*(s)$, und es gelten die in Tabelle 2 vergleichend zusammengestellten Fehlerschranken für den Abstand der Näherung $x_2(s)$ von $x^*(s)$ im Intervall $0 \leqq s \leqq 1$.

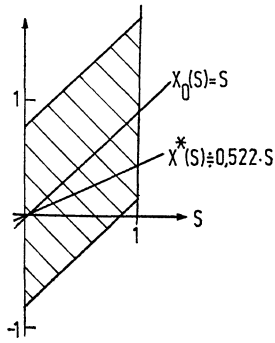

Fig. 2. Bereich, in dem die Existenz und Eindeutigkeit einer stetigen Lösung der Hammerstein-Gleichung $x(s)+\frac{1}{2}\int_0^1 s\cos x(t)\,dt=s$ nach Satz 1 gesichert ist

Tabelle 2. *Vergleich verschiedener oberer Schranken für den Fehler der nach TV berechneten Näherung $x_2(s)$ für die eindeutige Lösung $x^*(s)\doteq 0{,}52243\,66093\,99351\cdot s$ der Hammersteinschen Integralgleichung $x(s)+\frac{1}{2}\int_0^1 s\cos x(t)\,dt=s$ mit $x_0(s):=s$*

Autor	Lit.-Quelle	$10^{13}\cdot\|x^*-x_2\|\leqq$	Raum
Collatz (s. auch Ade)	[6] ([1])	1 430 000 000	$C[0,1]$
Jankó	[15]	1 440 000 000 000	
Kaazik	[16], Satz 1	19 000 000 000	
Nečepurenko	[21]	Bedingung nicht erfüllt	
Shafiyev	[23], Satz 2	82 200 000 000	
Ulm	[24], Satz 3	10 000 000 000	
Döring	(2.1)	10 000 000 000	
	(2.2)	1 490	
	(2.3) + (B 7)	992	
Wirklicher Fehler	$\|x^*-x_2\|\doteq$	5,7	
Döring	(2.3) + (B 7)	233	$L_2(0,1)$
Wirklicher Fehler	$\|x^*-x_2\|\doteq$	3,3	

Bemerkungen. 1. Für die Schranke von Collatz wurde der kleinstmögliche Bereich $S_c \subset S$ gewählt.

2. Man stellt fest, daß das Verhältnis von oberer Schranke zum wirklichen Fehler im Raum L_2 etwa um den Faktor 2,5 besser ist als im Raum C (s. auch Abschn. 3).

3. Die Gl. (6.1) kann man leicht exakt lösen: Die Lösung ist eine Gerade durch den Nullpunkt $x^*(s) = \xi^* s$, wo sich ξ^* als eindeutige Lösung einer gewissen transzendenten Gleichung bestimmen läßt.

Beispiel 3. Ein von Collatz [6] betrachtetes einfaches, leicht exakt lösbares Matrix-Eigenwert-Problem:

$$A y = \lambda B y, \quad \text{wo} \quad A := \begin{pmatrix} 2 & 3 \\ -1 & 1 \end{pmatrix} \quad B := \begin{pmatrix} -1 & 1 \\ -2 & 2 \end{pmatrix}. \tag{6.4}$$

Schon Unger [25] hat darauf hingewiesen, daß das Newton-Verfahren und damit auch dessen Verallgemeinerungen zur angenäherten Lösung von Matrix-Eigenwert-Aufgaben gut geeignet sind. Man faßt hierbei die Kombinationen aus Eigenvektor und Eigenwert

$$x := \begin{pmatrix} y \\ \lambda \end{pmatrix} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \lambda \end{pmatrix} \quad \text{mit}^4 \quad y := \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}$$

als Elemente des zugrunde zu legenden Banach-Raumes X auf mit der Norm $\|x\| := \max(|\eta_i|, |\lambda|)$. Der Operator F wird durch die Differenz der beiden Seiten der gegebenen Matrix-Gleichung plus einer Normierungsbedingung für den Eigenvektor gebildet:

$$F(x) := \begin{pmatrix} A y - \lambda B y \\ s(x) - 1 \end{pmatrix}.$$

Wir wählen mit Collatz und Ade [1] $s(x) := \eta_2$ und als Matrix-Norm das Maximum der Zeilenbetragssummen. Das Tschebyscheff-Verfahren besteht hier darin (bez. Einzelheiten s. z.B. Collatz [5, 6] und Ade [1]), für jeden Iterationsschritt ein Paar von linearen Gleichungssystemen zu lösen, die beide jeweils die gleiche Koeffizienten-Matrix haben. Dabei ist zu bemerken, daß die zweite Fréchet-Ableitung des Operators F ein konstanter bilinearer Operator ist und somit die dritte Ableitung der Null-Operator.

Im Falle eines (beliebigen endlichdimensionalen) Matrix-Eigenwert-Problems der vorliegenden Gestalt nimmt die scharfe Schranke (2.3) mit (B7) die einfache Gestalt an:

$$\|x^* - x_n\| \leqq \mu_n \beta_n M_2 \|d_{n-1} - c_{n-1}\| (\|d_{n-1}\| + \|c_{n-1}\|), \quad \forall n \in \mathbb{N}. \tag{6.5}$$

Mit

$$x_0 := \begin{pmatrix} -0{,}59 \\ 1 \\ 0{,}7 \end{pmatrix} \text{ berechnet man } \zeta \doteqdot 0{,}0351 \Rightarrow S := \{x \in \mathbb{R}^3 \mid \|x - x_0\| \leqq 0{,}055\};$$

$M_2 = 3$; $M_3 = 0$. Die Matrix f_0 ist nicht singulär, also existiert f_0^{-1}, und es ist $\beta \doteqdot 1{,}543$. Die Bedingungen (V3) und (V4) sind also ebenfalls erfüllt. Demnach gelten die Aussagen (B1)—(B6) von Satz 1 für (6.4) bezüglich S. In Tabelle 3 sind einige obere Schranken für den Fehler von x_2 zusammengestellt.

4 Eine Verwechslung mit den in Abschnitt 2 benutzten Zahlen η_n ist nicht zu befürchten.

Tabelle 3. *Vergleich verschiedener oberer Schranken für den Fehler der nach TV berechneten Näherung x_2 für die in $S=\{x\in \mathbb{R}^3 \mid \|x-x_0\|\leqq 0{,}055\}$ eindeutige Lösung*

$$x^* \doteq \begin{pmatrix} -0{,}55825\,75694\,95584 \\ 1 \\ 0{,}73623\,73841\,74027 \end{pmatrix}$$ *der Matrix-Eigenwert-Aufgabe* $\begin{pmatrix} 2 & 3 \\ -1 & 1 \end{pmatrix} y = \lambda \begin{pmatrix} -1 & 2 \\ -2 & 1 \end{pmatrix} y$ *mit*

$$x_0 := \begin{pmatrix} -0{,}59 \\ 1 \\ 0{,}7 \end{pmatrix}$$

Autor	Lit.-Quelle	$10^{12}\cdot\|x^*-x_2\|\leqq$
Collatz (bzw. Ade)	[6] ([1])	Bedingung nicht erfüllt
Jankó	[15]	Bedingung nicht erfüllt
Kaazik	[16], Satz 1	888 000 000
Nečepurenko	[21]	Bedingung nicht erfüllt
Shafiyev	[23], Satz 2	51 500
Ulm	[24], Satz 3	52 800 000
Döring	(2.1)	2 550 000
	(2.2)	55
	(2.3) + (B 7) bzw. (6.5)	30,9
Wirklicher Fehler	$\|x^*-x_2\| \doteq$	0,23

7. Rechenaufwand. Anwendungsbereich

Ehrmann ([13], Satz 11) hat gezeigt, daß unter gewissen Voraussetzungen bei der Lösung algebraischer Gleichungen n-ten Grades für $n\leqq 5$ das Newton-Verfahren, für $n>5$ das Tschebyscheff-Verfahren unter allen Verfahren höherer Ordnung vom Ehrmann-Typ (= Verallgemeinerungen von (1.4)) das günstigste ist. Einen analogen Satz für Operatorgleichungen gibt es nicht.

Die durch Anwendung des Newton- oder Tschebyscheff-Verfahrens entstehenden linearen Integralgleichungen werden im allgemeinen durch Diskretisierung gelöst. Es sind also hier wie auch beim Matrix-Eigenwert-Problem pro Iterationsschritt zwei inhomogene lineare Gleichungssysteme zu lösen, die die gleiche Koeffizienten-Matrix besitzen. Hierin liegt der rechnerische Vorteil von TV gegenüber dem Newton-Verfahren begründet: Die Inversion bzw. Dreieckszerlegung (beim Gauß-Algorithmus) der Matrix — die ja den Hauptteil des Rechenaufwandes bei der Gleichungslösung ausmacht — braucht für beide Gleichungssysteme *nur einmal* ausgeführt zu werden. Für das zweite Gleichungssystem sind nur andere „rechte Seiten" einzugeben. Die Lösung der beiden Integralgleichungen (6.2) und (6.3) nach der Quadraturformelmethode mit 70 Stützstellen — deren Einzelheiten hier weggelassen seien — auf 10 Dezimalen nach dem Komma ergab eine Rechenzeitersparnis bei TV von etwa 20% gegenüber dem Newton-Verfahren. Im vorliegenden Beispiel lassen sich (6.2) und (6.3) übrigens auch einfacher mittels Rekursionsformeln für die Koeffizienten $\tilde{\xi}_n$ (bei (6.2)) und ξ_n (bei (6.3)) lösen. Auch dabei wurden etwa 20% gespart.

Die gleiche Rechenzeitersparnis ergab sich schließlich bei der Bestimmung einer periodischen Lösung der Duffingschen Differentialgleichung, die zunächst

in eine Hammerstein-Gleichung umgewandelt und dann diskretisiert wurde. Hierbei ist der Nemycki-Operator gleich x^8. Auch in diesem Falle wollen wir darauf verzichten, auf die umfangreichen Einzelheiten der Lösung einzugehen.

Ein allgemeines Kriterium, wonach man anhand gewisser Kenndaten einer Operatorgleichung entscheiden kann, wann TV einen geringeren Aufwand zur Berechnung einer Näherungslösung auf eine vorgegebene Genauigkeit erfordert als das Newton-Verfahren, gibt es nicht. Man kann zusammenfassend nur sagen, daß TV in jenen Fällen, wo die zweite Ableitung des Operators F leicht gebildet werden kann, dem Newton-Verfahren im allgemeinen überlegen sein wird, wobei bezüglich der Rechenzeitersparnis als grobe Regel 20% angenommen werden können. Dies trifft außer auf Matrix-Eigenwert-Probleme (zweite Ableitung konstant!) z. B. auf Hammersteinsche Integralgleichungen zu, bei denen der Nemycki-Operator ein Polynom oder eine trigonometrische Funktion in x ist und auf gewisse nichtlineare Gleichungssysteme.

Literatur

1. Ade, H.: Iterationsverfahren höherer Ordnung in Banach-Räumen. Numer. Math. **13**, 39—50 (1969).
2. Albrecht, J.: Bemerkungen zum Iterationsverfahren von Schulz zur Matrix-Inversion. ZAMM **41**, 262—263 (1961).
3. Antosiewicz, H. A., Rheinboldt, W. C.: Numerical analysis and functional analysis. In: Todd, J. (Ed.): Survey of numerical analysis. New York 1962, pp. 485—517.
4. Černyšenko, D. N.: Ein Analogon der Methode von Čebyšev. [Russ.] In: Erste Republ. Math. Konf. junger Forscher. Teil I. [Russ.], pp. 691—697. Kiev: AN Ukrain. SSR Inst. Mat. 1965.
5. Collatz, L.: Funktionalanalysis und numerische Mathematik. Springer 1964.
6. — Näherungsverfahren höherer Ordnung für Gleichungen in Banach-Räumen. Arch. Rat. Mech. Anal. **2**, 66—75 (1958/59).
7. Dieudonné, J.: Foundations of modern analysis. New York: Academic Press 1960.
8. Döring, B.: A higher order iterative method for the solution of nonlinear operator equations. In: Ghizzetti, A. (Ed.): A Nato Advanced Study Institute on Theory and Applications of Monotone Operators. Gubbio: Edizioni „Oderisi" 1969, pp. 291—298.
9. — Über das Newtonsche Näherungsverfahren. Math.-Phys. Sem.-Ber. **16**, 27—40 (1969).
10. — Ein Satz über das verbesserte Newton-Verfahren. ZAMM **49** (1969), Sonderheft GAMM-Tagung.
11. Dubovik, L. I.: Allgemeine Gestalt eines Iterationsverfahrens dritter Ordnung für nichtlineare Funktionalgleichungen. [Russ.] Dieselbe Z. wie [4], pp. 219—225.
12. — Über die Konvergenz des verallgemeinerten Newton-Verfahrens und dessen Anwendung zur Lösung von Matrix-Gleichungen. [Russ.] Dieselbe Z. wie [4], pp. 225—232.
13. Ehrmann, H.: Konstruktion und Durchführung von Iterationsverfahren höherer Ordnung. Arch. Rat. Mech. Anal. **4**, 65—88 (1959/60).
14. Forster, P.: Bemerkungen zum Iterationsverfahren von Schulz zur Bestimmung der Inversen einer Matrix. Numer. Math. **12**, 211—214 (1968).
15. Jankó, B.: Sur la résolution des équations opérationnelles nonlinéaires. Mathematica (Cluj) **7** (**30**), 257—262 (1965).
16. Kaazik, Yu. Ya.: Über eine Klasse von Iterationsverfahren zur näherungsweisen Lösung von Operatorgleichungen. [Russ.] DAN SSSR **112**, 579—582 (1957).
17. — On the approximate solution of nonlinear operator equations by iterative methods. Uspehi Mat. Nauk (N.S.) **12**, no. 1 (73) 195—199 (1957) [Russ.]. Engl. Übersetzung in: AMS Transl. Ser. 2, **16**, 410—413 (1960).

18. Kantorovich, L. V.: Functional analysis and applied mathematics. [Russ.] Uspehi Mat. Nauk **3**, 6, 89—185 (1948). Transl. from Russ. by C. D. Benster, NBS Washington 1952. Editor G. E. Forsythe.
19. Kantorowitsch, L. W., Akilow, G. P.: Funktionalanalysis in normierten Räumen. Berlin 1964.
20. Moore, R. H.: Newton's method and variations. In Anselone, P. M. (Ed.): Nonlinear integral equations. Madison: The University of Wisconsin Press 1964, pp. 65—98.
21. Nečepurenko, M. I.: Über das Čebyšev-Verfahren für Funktionalgleichungen. [Russ.] Uspehi Mat. Nauk **9**, vyp. 2 (60), 163—170 (1954). Engl. Übersetzung: von L. B. Rall, MRC Tech. Summary Rept. No. 648, Mathematics Research Center, U. S. Army, University of Wisconsin, Madison 1966.
22. Schröder, E.: Ueber unendlich viele Algorithmen zur Auflösung der Gleichungen. Math. Ann. **2**, 317—365 (1870).
23. Shafiyev, R. A.: Certain iteration processes. USSR Comp. Math. Math. Phys. **4**, No. 1, 187—193 (1964).
24. Ulm, S.: Über die Konvergenz gewisser Iterationsverfahren in Banach-Räumen. [Russ.] Učen. Zap. Tartusk. GU **42**, 135—142 (1956).
25. Unger, H.: Nichtlineare Behandlung von Eigenwertaufgaben. ZAMM **30**, 281—282 (1950).
26. Traub, J. F.: Iterative methods for the solution of equations. Prentice-Hall 1964.
27. Varyukhin, V. A., Kas'yanyuk, S. A.: Method of solving non-linear systems of a special type. USSR Comp. Math. Math. Phys. **6**, No. 2, 214—221 (1966).
28. Xu Li-Zhe: Several iterative processes with a relatively high rate of convergence and a method for estimating the multiplicity of real roots. Chinese Math. **3**, 361—371 (1963).

Priv.-Doz. Dr. Boro Döring
Mathematisches Institut der
Technischen Hochschule Darmstadt
D-61 Darmstadt

Numer. Math. 15, 196–218 (1970)

Über die Koerzitivität gewöhnlicher Differenzenoperatoren und die Konvergenz von Mehrschrittverfahren

ROLF DIETER GRIGORIEFF

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 5. Dezember 1969

Summary. For linear ordinary difference operators under general boundary conditions a so called coerciveness inequality is proved. As an application the solutions and their difference quotients up to the order of the approximated differential equation of stable and consistent multistep difference methods are shown to converge.

1. Einleitung

Im ersten Teil der vorliegenden Arbeit beweisen wir eine a priori-Abschätzung für lineare gewöhnliche Differenzenoperatoren unter allgemeinen Randbedingungen. Mit ihrer Hilfe ist es z.B. möglich, die bisher noch offene Frage nach der Konvergenz von Differenzenapproximationen höherer Annäherung (s. Collatz [5]) bei Randwertaufgaben gewöhnlicher Differentialgleichungen zu beantworten, was in [10, 12, 13] geschieht. Hier verwenden wir sie im zweiten Teil der Arbeit zum Beweis der Konvergenz von Mehrschrittverfahren bei im Sinne von Carathéodory gestellten Anfangswertaufgaben gewöhnlicher Differentialgleichungen m-ter Ordnung, $m \geqq 1$. Dabei wird die Existenz von Lösungen der approximierten Probleme nicht vorausgesetzt, sondern gleichzeitig mit der Konvergenz bewiesen.

In den letzten fünf Jahren sind mit neuen Methoden wesentliche Fortschritte in der Frage nach der Konvergenz von Differenzenapproximationen bei Differentialgleichungsproblemen erzielt worden (s. die Arbeiten von Aubin [2], Céa [4], Frehse [9], Grigorieff [10, 12, 13], Stummel [26, 27], Thomée [29]). Eines der wesentlichsten Hilfsmittel sind dabei a priori-Abschätzungen für die Differenzenoperatoren mit Konstanten, die unabhängig von der Schrittweite sind, in Verbindung mit Kompaktheitsaussagen für Folgen von Gitterfunktionen bei gegen Null gehender Schrittweite. Diese a priori-Abschätzungen stehen in enger Analogie zu bekannten a priori-Abschätzungen bei elliptischen Differentialgleichungen (s. z.B. [28, Kap. 4]). Allerdings sind bei den Differenzenoperatoren bisher nur einfach zu handhabende Randbedingungen, im wesentlichen Dirichletrandbedingungen, untersucht worden. Im ersten Teil dieser Arbeit, d.h. in den Paragraphen 2 bis 4, geben wir mit Satz (4) notwendige und hinreichende Bedingungen an für das Bestehen der a priori-Abschätzung

$$(*) \qquad |u| \leqq K\left(|A_h u| + \sum_{l=1}^{2} \sum_{j=1}^{r} \langle B_{j,h}^{(l)} u \rangle_l\right), \quad u \in H_h, \quad h \in \Lambda_0.$$

Dabei ist H_h ein zu L^2 analoger Raum von Gitterfunktionen mit der Norm $|\cdot|$, und Λ_0 ist eine Nullfolge von Schrittweiten. A_h ist zwar ein gewöhnlicher Differenzenoperator, jedoch sind $B_{j,h}^{(l)}$ allgemeine Randoperatoren. Die Ungleichung (*) steht in enger Analogie zu den sog. Koerzitivitätsungleichungen von Schechter [21] für elliptische Differentialgleichungen unter allgemeinen Randbedingungen. Wie schon erwähnt, ist (*) eins der grundlegenden Hilfsmittel, um die Konvergenz von Differenzenapproximationen höherer Annäherung bei linearen (s. [10, 12]) und nichtlinearen (s. [13]) Randwertaufgaben gewöhnlicher Differentialgleichungen zu beweisen. Ein (*) verwandtes Ergebnis für Systeme von gewöhnlichen Differenzengleichungen erster Ordnung unter Verwendung einer diskreten Maximumnorm anstelle von $|\cdot|$ ist von Rjaben'kii [20] bewiesen worden.

Die Koerzitivitätsungleichung (*) wird im zweiten Teil der Arbeit herangezogen, um mit dem im vorhergehenden Absatz skizzierten Vorgehen die Konvergenz von Mehrschrittverfahren zur näherungsweisen Lösung gewöhnlicher Differentialgleichungen zu beweisen. In den Arbeiten von Dahlquist [6, 7] (s. auch die Bücher von Henrici [14, 15]) wird die Konvergenz der Mehrschrittverfahren grundsätzlich geklärt. Seitdem sind in einer Reihe weiterer Arbeiten allgemeine Konvergenztheorien aufgestellt worden, welche die Mehrschrittverfahren als Spezialfall enthalten, bzw. sind gewisse feinere Fragenstellungen behandelt worden. Von Schempp [22] wird eine distributionstheoretische Begründung der Mehrschrittverfahren gegeben. Butcher [3] und Spijker [23] lassen allgemeinere inhomogene Terme in den Differenzengleichungen zu. Spijker untersucht dabei auch Differentialgleichungen höherer als erster Ordnung. Von Ansorge [1], Stetter [25] und Thompson [30] wird der Äquivalenzsatz von Lax (s. z. B. [19]) auf inhomogene Probleme verallgemeinert, so daß sich mit der resultierenden Theorie auch die Mehrschrittverfahren bei quasilinearen Differentialgleichungen erfassen lassen. Sehr geeignet für die Behandlung von Differenzenverfahren ist ein Diskretisierungskonzept von Stetter. Ein hierauf aufbauender Satz von Pereyra [17] wird von Watt [33] bei der Untersuchung des Zusammenhanges von Konsistenz, Stabilität und Konvergenz von Mehrschrittverfahren verwendet. In den zuletzt genannten Arbeiten wie auch bei Spijker sind natürlich Differentialgleichungen in beliebigen Banachschen Räumen zugelassen.

Allen vorangehend aufgeführten Methoden sind die folgenden vier Merkmale gemeinsam. (i) Die Existenz einer Lösung des Anfangswertproblems wird vorausgesetzt. (ii) Die Inhomogenität wird als stetig vorausgesetzt. (iii) Es werden Bedingungen gestellt, welche die eindeutige Lösbarkeit sowohl der Differential- als auch der Differenzengleichungen sichern. (iv) Es wird nur die Konvergenz der Lösungen selbst aber nicht die ihrer Differenzenquotienten gezeigt. Die Forderung (i) und in einem Spezialfall auch (iii) wird von Hull und Luxemburg [16] fallengelassen, die mit der Konvergenz auch gleichzeitig die Existenz einer Lösung beweisen. Die Arbeit von Dejon [8] beschäftigt sich mit der Konvergenz der Differenzenquotienten höherer Ordnung und mit der Aufstellung notwendiger und hinreichender algebraischer Kriterien für eine Klasse von Stabilitätsbedingungen.

In der vorliegenden Arbeit lassen wir alle vier Einschränkungen (i)—(iv) fallen. Wir untersuchen in den Paragraphen 5—7 die Konvergenz von Mehrschrittverfahren bei Differentialgleichungen m-ter Ordnung, $m \geqq 1$, die wir im Sinne von Carathéodory in den Sobolewschen Funktionenräumen $W^{m,2}$ betrachten. Daher

kommt man auch mit schwacher Regularität der Inhomogenität aus, über die wir nur die übliche Voraussetzung für die Existenz einer Lösung machen (s. z.B. Walter [32]). Die Inhomogenität bei den Mehrschrittverfahren ist allgemeinerer Natur als man sie bisher zugelassen hat. Ist das Mehrschrittverfahren konsistent und stabil, so zeigen wir in Satz (55) die Existenz einer Lösung der Anfangswertaufgabe sowie der Differenzengleichungen und die Konvergenz einer Teilfolge mitsamt den Differenzenquotienten bis zur m-ten Ordnung. Bei eindeutiger Lösbarkeit konvergiert sogar die gesamte Folge. Die Existenz einer Lösung der Differenzengleichungen wird mit einem Fixpunktsatz von Petryshyn [18] erbracht. Im weiteren wird gezeigt, daß Konsistenz und Stabilität im wesentlichen auch notwendig für die Konvergenz sind. Es ergibt sich damit der interessante Zusammenhang, daß die von Stetter [24] eingeführte starke Stabilität charakteristisch ist für die Konvergenz der Differenzenquotienten bis zur Ordnung der Differentialgleichung einschließlich. Im Gegensatz dazu besagen die Ergebnisse von Dejon [8], daß in diesem Fall stets Instabilität vorliegt, was sich durch die von ihm verwendeten Normen erklärt, die von den unsrigen verschieden sind.

2. Eine Koerzitivitätsungleichung

In diesem Paragraphen führen wir die im weiteren benötigten Begriffe ein und formulieren den Hauptsatz (4) des ersten Teils dieser Arbeit. Beginnend mit der Definition der Räume H_h von Gitterfunktionen schließen wir einige Begriffsbildungen bei Polynomen an, die von Schechter [21] im Zusammenhang mit a priori-Abschätzungen bei Randwertaufgaben elliptischer Differentialgleichungen verwendet worden sind. Es werden dann der Differenzenoperator A_h und die Randoperatoren $B_{j,h}^{(l)}$, $j=1,\ldots,r$, $l=1,2$, definiert und notwendige und hinreichende Bedingungen dafür angegeben, daß die Koerzitivitätsungleichung (5) besteht.

Mit $\mathbb{R}$ bzw. $\mathbb{C}$ bezeichnen wir den Körper der reellen bzw. komplexen Zahlen und mit $\mathbb{N}$ die Menge der natürlichen Zahlen. Es sei Λ_0 eine Nullfolge positiver Zahlen und $\{y_h\}$, $h\in\Lambda_0$, sei eine Folge von Punkten $y_h\in\mathbb{R}$. Wir erklären dann die Menge $\mathbb{R}_h$ von Gitterpunkten durch

$$\mathbb{R}_h=\{x\in\mathbb{R}\mid x=y_h+jh,\ j=0,\pm1,\ldots,\},\qquad h\in\Lambda_0.$$

Es sei Ω_h der Vektorraum der auf $\mathbb{R}_h$ erklärten komplexwertigen Funktionen. In Ω_h ist für jedes $h\in\Lambda_0$ der Verschiebungsoperator E_h und seine Inverse E_h^{-1} gegeben durch

$$(E_hu)(x)=u(x+h),\qquad (E_h^{-1}u)(x)=u(x-h),\qquad x\in\mathbb{R}_h,\qquad u\in\Omega_h.$$

Für jede Teilmenge $G\subseteq\mathbb{R}$ sei $G_h=G\cap\mathbb{R}_h$ und

$$l_h^2(G)=\{u\in\Omega_h\mid u(x)=0,\ x\in\mathbb{R}_h-G_h,\ \sum_{x\in\mathbb{R}_h}|u(x)|^2<\infty\}.$$

In $l_h^2(G)$ wird durch die Vorschrift

$$(u,v)=h\sum_{x\in\mathbb{R}_h}u(x)\,\bar v(x),\qquad |u|=(u,u)^{\frac12},\qquad u,v\in l_h^2(G),$$

ein Skalarprodukt und die zugehörige Norm erklärt, wodurch $l_h^2(G)$ ein Hilbertscher Raum wird. Schließlich seien c_0, d_0 zwei Punkte aus $\mathbb{R}$ mit $c_0 < d_0$ und $\{c_h\}$, $\{d_h\}$, $h \in \Lambda_0$, seien Folgen von Gitterpunkten $c_h, d_h \in \mathbb{R}_h$ mit $c_h < d_h$, $h \in \Lambda_0$, und $c_h \to c_0$, $d_h \to d_0$ $(h \in \Lambda_0)$. Dann setzen wir $H_h = l_h^2[c_h, d_h]$, $h \in \Lambda_0$.

Es sei k eine nichtnegative ganze Zahl und s, s_j, $j = 1, \ldots, k$, seien Polynome in der Unbestimmten z mit Koeffizienten aus $\mathbb{C}$. Der Fall $k = 0$ soll bedeuten, daß keine Polynome s_j gegeben sind. Dann vereinbaren wir die folgende Definition.

(1) *Die Polynome s_j, $j = 1, \ldots, k$, spannen s auf, genau wenn zu jedem Polynom t ein Polynom t_1 und Zahlen $\mu_j \in \mathbb{C}$, $j = 1, \ldots, k$, existieren, so daß gilt*

$$t = t_1 s + \sum_{j=1}^{k} \mu_j s_j .$$

Als nächstes formulieren wir die Voraussetzung (K1)—(K4) über die Differenzenoperatoren und Randbedingungen, die wir betrachten wollen. Dabei werden die Punktmengen $M \subseteq \mathbb{R}^2$ und $N \subseteq \mathbb{R}$ verwendet, die definiert sind durch

$$M = \bigcup_{h \in \Lambda_0} [c_h, d_h]_h \times \{h\}, \qquad N = \{h \in \mathbb{R} \mid h \in \Lambda_0\}.$$

(K1) *Es seien a_k, $k = 0, \pm 1, \ldots$, Funktionen aus $C(\overline{M})$, von denen höchstens endlich viele nicht identisch verschwinden.*

(K2) *Es seien n_1, n_2 nichtnegative ganze Zahlen, es sei $a_k(c_0, 0) = 0$, $k < -n_1$, $a_k(d_0, 0) = 0$, $k > n_2$, und der Operator A_h sei definiert durch*

$$(A_h u)(x) = \begin{cases} \sum_{k=-\infty}^{\infty} a_k(x, h)\,(E_h^k u)(x), & x \in [c_h + n_1 h, d_2 - n_2 h]_h, \\ 0, \quad \text{sonst in } \mathbb{R}_h, & u \in \Omega_h, \quad h \in \Lambda_0. \end{cases}$$

(K3) *Es sei r eine natürliche Zahl und $b_{j,k}^{(l)}$, $j = 1, \ldots, r$, $k = 0, 1, \ldots$, $l = 1, 2$, seien Funktionen aus $C(\overline{N})$, von denen höchstens endlich viele nicht identisch verschwinden.*

(K4) *Die Operatoren $B_{j,h}^{(l)}$, $j = 1, \ldots, r$, $l = 1, 2$, seien definiert durch*

$$(B_{j,h}^{(1)} u)(x) = \sum_{k=0}^{\infty} b_{jk}^{(1)}(h)\,(E_h^k u)(x), \qquad x \in \mathbb{R}_h,$$

$$(B_{j,h}^{(2)} u)(x) = \sum_{k=0}^{\infty} b_{jk}^{(2)}(h)\,(E_h^{-k} u)(x), \qquad x \in \mathbb{R}_h, \qquad u \in \Omega_h, \qquad h \in \Lambda_0.$$

Im folgenden verwenden wir die Abkürzung p und $q_j^{(l)}$ für

$$p(x, h, z) = \sum_{k=-\infty}^{\infty} a_k(x, h)\, z^k, \qquad (x, h) \in \overline{M},$$

$$q_j^{(l)}(h, z) = \sum_{k=0}^{\infty} b_{jk}^{(l)}(h)\, z^k, \qquad j = 1, \ldots, r, \qquad l = 1, 2, \qquad h \in \overline{N}.$$

Für jedes Polynom $s \neq 0$ bezeichnen wir mit s_i das Polynom, das genau die im Innern des Einheitskreises gelegenen Wurzeln von s und den Koeffizienten 1 vor der höchsten Potenz besitzt. Dann geben wir die folgenden beiden Definitionen.

(2) *Der Operator A_h heißt elliptisch, genau wenn für jedes $x \in [c_0, d_0]$ und jedes $\xi \in \mathbb{R}$ gilt $p(x, 0, e^{i\xi}) \neq 0$.*

(3) *Die Operatoren $B_{j,h}^{(l)}$, $j = 1, \ldots, r$, $l = 1, 2$, überdecken den Operator A_h, genau wenn die Polynome $q_j^{(1)}(0, z)$ (bzw. $q_j^{(2)}(0, z)$), $j = 1, \ldots, r$, das Polynom $[z^{n_1} p(c_0, 0, z)]$, (bzw. $[z^{n_2} p(d_0, 0, z^{-1})]$,) aufspannen.*

Schreiben wir noch $\langle \cdot \rangle_l$, $l = 1, 2$, für die Randnormen

$$\langle u \rangle_1 = h^{\frac{1}{2}} |u(c_h)|, \quad \langle u \rangle_2 = h^{\frac{1}{2}} |u(d_h)|, \quad u \in \Omega_h, \quad h \in \Lambda_0,$$

so lautet der **Hauptsatz** über die Koerzitivität gewöhnlicher Differenzenoperatoren wie folgt.

(4) *Unter den Voraussetzungen* (K 1)–(K 4) *existiert dann und nur dann eine von u und h unabhängige Konstante K, so daß für fast alle $h \in \Lambda_0$ die Koerzitivitätsungleichung*

$$|u| \leqq K\Big(|A_h u| + \sum_{l=1}^{2} \sum_{j=1}^{r} \langle B_{j,h}^{(l)} u \rangle_l\Big), \quad u \in H_h, \tag{5}$$

besteht, wenn A_h elliptisch ist und die Operatoren $B_{j,h}^{(l)}$, $j = 1, \ldots, r$, $l = 1, 2$, den Operator A_h überdecken.

Den Beweis dieses Satzes erbringen wir in den nächsten beiden Paragraphen.

3. Der Fall konstanter Koeffizienten

Im vorliegenden Paragraphen beweisen wir die Koerzitivitätsungleichung (5) im Spezialfall, daß die Operatoren A_h und $B_{j,h}^{(l)}$ konstante Koeffizienten besitzen und ein halbunendliches Intervall zugrunde liegt, so daß nur an einem Intervallende Randbedingungen zu stellen sind. Genauer nehmen wir an, daß die Schrittweite $h = 1$ sowie $0 \in \mathbb{R}_1$ ist und a_k, $k = 0, \pm 1, \ldots, b_{jk}^{(1)}$, $j = 1, \ldots, r$, $k = 0, 1, \ldots$, komplexe Zahlen sind mit $a_k = 0$, $k < 0$, so daß $n_1 = 0$ gewählt werden kann. Der Operator A_1 ist in diesem Fall gegeben durch

$$(A_1 u)(x) = \begin{cases} \sum_{k=0}^{\infty} a_k (E_1^k u)(x), & x \in [0, \infty)_1 \\ 0 \quad , \text{ sonst in } \mathbb{R}_1, & u \in H_1, \end{cases}$$

und es wird

$$p = \sum_{k=0}^{\infty} a_k z^k, \quad q_j^{(1)} = \sum_{k=0}^{\infty} b_{jk}^{(1)} z^k, \quad j = 1, \ldots, r.$$

Wir behaupten den Satz

(6) *Es besteht dann und nur dann die Koerzitivitätsungleichung*

$$|u| \leqq K\Big(|A_1 u| + \sum_{j=1}^{r} |(B_{j,1}^{(1)} u)(0)|\Big), \quad u \in l_1^2[0, \infty), \tag{7}$$

mit einer von u unabhängigen Zahl K, wenn $p(e^{i\xi}) \neq 0$, $\xi \in \mathbb{R}$, gilt und die Polynome $q_j^{(1)}$, $j = 1, \ldots, r$, das Polynom p, aufspannen.

Bevor wir zum Beweis von (6) kommen, führen wir die diskrete Fouriertransformation F ein als Abbildung von $l_1^2(\mathbb{R}) \to L^2(-\pi, \pi)$, dem Hilbertraum

der im Sinne von Lebesgue über dem Intervall $(-\pi, \pi)$ absolut quadratisch integrablen Funktionen:

$$(F u)(\xi) = \tilde{u}(\xi) = \frac{1}{\sqrt{2\pi}} \sum_{x \in \mathbb{R}_1} u(x)\, e^{-ix\xi}, \qquad \xi \in (-\pi, \pi).$$

F bildet bekanntlich den Raum $l_1^2(\mathbb{R})$ linear und isometrisch auf $L^2(-\pi, \pi)$ ab, und es gilt

$$\int_{-\pi}^{\pi} \tilde{u}(\xi)\, e^{ij\xi}\, d\xi = \sqrt{2\pi}\, u(j), \qquad u \in l_1^2(\mathbb{R}), \qquad j = 0, \pm 1, \ldots. \tag{8}$$

Wir berechnen die diskrete Fouriertransformierte von $A_1 u$. Ist P_0 die orthogonale Projektion von $l_1(\mathbb{R})$ auf $l_1[0, \infty)$, so ergibt sich zunächst für jede natürliche Zahl j

$$\begin{aligned}(P_0 E_1^j \tilde{u})(\xi) &= \frac{1}{\sqrt{2\pi}} \left(\sum_{k=-\infty}^{\infty} u(k+j)\, e^{-ik\xi} - \sum_{k=-j}^{-1} u(k+j)\, e^{-ik\xi} \right) \\ &= e^{ij\xi} \tilde{u}(\xi) - \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{j-1} u(k)\, e^{-i(k-j)\xi}, \qquad u \in l_1^2[0, \infty).\end{aligned}$$

Damit erhalten wir für $\xi \in (-\pi, \pi)$ und $u \in l_1^2[0, \infty)$

$$(A_1 \tilde{u})(\xi) = \left(P_0\, p(E_1)\, \tilde{u}\right)(\xi) = p(e^{i\xi})\, \tilde{u}(\xi) - e^{i\xi} S(u, e^{i\xi}) \tag{9}$$

mit der Abkürzung

$$S(u, z) = \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} a_k \sum_{j=0}^{k-1} u(j)\, z^{k-j-1}, \qquad u \in l_1^2[0, \infty).$$

Beweis von (6). Um zu zeigen, daß die angegebenen Bedingungen hinreichend sind, nehmen wir im Widerspruch zur Behauptung an, daß Funktionen $u_k \in l_1^2[0, \infty)$, $k \in \mathbb{N}$, existieren mit $|u_k| = 1$ und

$$|A_1 u_k| + \sum_{j=1}^{r} |(B_{j,1}^{(1)} u_k)(0)| \to 0 \quad (k \to \infty). \tag{10}$$

Dabei können wir o.B.d.A. voraussetzen, daß für $l = 0, 1, \ldots, m-1$ die Zahlenfolgen $\{u_k(l)\}$, $k \in \mathbb{N}$, konvergent sind, wobei m so bestimmt ist, daß a_k für $k > m$ verschwindet. Damit konvergiert die Folge $\{S(u_k, e^{i\xi})\}$, $k \in \mathbb{N}$, stark in $L^2(-\pi, \pi)$. Aufgrund der Isometrie von F folgt aus (10) und (9)

$$p(e^{i\xi})\, \tilde{u}_k - e^{i\xi} S(u_k, e^{i\xi}) \to 0 \quad \text{in} \quad L^2(-\pi, \pi) \quad (k \to \infty).$$

Da $p(e^{i\xi}) \neq 0$, $\xi \in \mathbb{R}$, ist, hat dies die starke Konvergenz der Folge $\{\tilde{u}_k\}$, $k \in \mathbb{N}$, zur Folge. Damit konvergiert auch $\{u_k\}$, $k \in \mathbb{N}$, in $l_1^2(\mathbb{R})$, sagen wir gegen die Funktion $u \in l_1^2[0, \infty)$, und u genügt den Bedingungen

$$A_1 u = 0, \qquad (B_{j,1}^{(1)} u)(0) = 0, \qquad |u| = 1.$$

Wir behaupten, daß aus den beiden ersten Gleichungen $u = 0$ folgt, womit ein Widerspruch erbracht ist.

Aus $A_1 u = 0$ erhalten wir für $\tilde{u}$ die Darstellung

$$\tilde{u}(\xi) = p^{-1}(e^{i\xi})\, e^{i\xi} S(u, e^{i\xi}), \qquad \xi \in (-\pi, \pi),$$

woraus sich mit (8) nach Substitution von $e^{i\xi}=z$ auch

$$u(j)=\frac{1}{i\sqrt{2\pi}}\int_{|z|=1} z^j p^{-1}(z)\,S(u,z)\,dz,\qquad j=0,\pm 1,\ldots, \tag{11}$$

ergibt. Demnach ist die Bedingung $(B_{j,1}^{(1)}u)(0)=0$, $j=1,\ldots,r$, mit

$$\int_{|z|=1} q_j^{(1)}(z)\,p^{-1}(z)\,S(u,z)\,dz=0,\qquad j=1,\ldots,r, \tag{12}$$

äquivalent. Da die Polynome $q_j^{(1)}$, $j=1,\ldots,r$, p_i aufspannen, existieren zu jeder nichtnegativen ganzen Zahl k komplexe Zahlen μ_j, $j=1,\ldots,r$, und ein Polynom q, so daß gilt

$$z^k=q\,p_i+\sum_{j=1}^{r}\mu_j q_j^{(1)}.$$

Wegen

$$\int_{|z|=1} q(z)\,p_i(z)\,p^{-1}(z)\,S(u,z)\,dz=0$$

folgt aus (11) und (12) die Behauptung $u(k)=0$, $k=0,1,\ldots$.

Zum Beweis der Notwendigkeit bemerken wir zunächst, daß man aus der Gültigkeit von (7) leicht die Ungleichung

$$|u|\leqq K\,|p(E_1)\,u|,\qquad u\in l_1^2(\mathbb{R}),$$

herleitet. Hieraus gewinnt man durch diskrete Fouriertransformation

$$\|\tilde{u}\|\leqq K\,\|p(e^{i\xi})\,\tilde{u}\|,\qquad \tilde{u}\in L^2(-\pi,\pi),$$

wobei $\|\cdot\|$ für die Norm in $L^2(-\pi,\pi)$ steht. Dies ist aber nicht möglich, falls $p(e^{i\xi})$ in einem Punkt $\xi\in\mathbb{R}$ verschwindet.

Wir zeigen weiter, daß eine nichttriviale Lösung $u\in l_1^2[0,\infty)$ der Gleichungen

$$A_1u=0,\qquad (B_{j,1}^{(1)}u)(0)=0,\qquad j=1,\ldots,r,$$

existiert, falls die Polynome $q_j^{(1)}$, $j=1,\ldots,r$, das Polynom p_i nicht aufspannen. Wie wir vorangehend überlegt haben, können wir dabei $p(e^{i\xi})\neq 0$, $\xi\in\mathbb{R}$, voraussetzen.

Sei n_i der Grad von p_i. Wir nehmen $n_i>0$ an, da die Behauptung sonst trivial ist. Aus der Widerspruchsannahme folgt, daß es bei geeigneter Numerierung eine nichtnegative ganze Zahl $l<n_i$ gibt, so daß für jedes $j=l+1,\ldots,r$ Zahlen $\mu_k\in\mathbb{C}$, $k=1,\ldots,l$, und ein Polynom q existieren mit

$$q_j^{(1)}=q\,p_i+\sum_{k=1}^{l}\mu_k q_k^{(1)}. \tag{13}$$

Im anderen Falle wäre nämlich $r\geqq n_i$, und aus dem Bestehen der Gleichung

$$q\,p_i+\sum_{k=1}^{n_i}\mu_k q_k^{(1)}=0$$

folgte notwendig $\mu_k=0$, $k=1,\ldots,n_i$. Dann wären die Polynome $r_k^{(1)}$, $k=1,\ldots,n_i$, die sich als Reste bei der Division von $q_k^{(1)}$ durch das Polynom n_i-ten Grades p_i ergeben, linear unabhängig, so daß die Polynome $q_k^{(1)}$, $k=1,\ldots,n_i$, das Polynom p_i aufspannen müßten.

Sei p_a das durch $p = p_i\, p_a$ definierte Polynom und n_a sei der Grad von p_a. Wir wählen $n_i + n_a$ Zahlen $v(0), v(1), \ldots, v(n_i + n_a - 1)$, die das unterbestimmte lineare algebraische Gleichungssystem

$$S(v, z) \equiv 0 \mod p_a(z)$$
$$\int_{|z|=1} q_j^{(1)}(z)\, p^{-1}(z)\, S(v, z)\, dz = 0, \qquad j = 1, \ldots, l,$$

nichttrivial lösen. Für diese Lösung ist $S(v, z) \neq 0$, da $S(v, z)$ wegen $a_{n_i+n_a} \neq 0$ und $a_k = 0$, $k > n_i + n_a$, nur dann verschwindet, wenn $v(0) = \cdots = v(n_i + n_a - 1) = 0$ ist. Die Fouriertransformierte $\tilde{u}$ der gesuchten Funktion u erklären wir dann durch

$$\tilde{u}(\xi) = p^{-1}(e^{i\xi})\, e^{i\xi} S(v, e^{i\xi}), \qquad \xi \in (-\pi, \pi).$$

Aufgrund der Beziehung (8) gilt

$$u(j) = \frac{1}{i\sqrt{2\pi}} \int_{|z|=1} z^j p^{-1}(z)\, S(v, z)\, dz, \qquad j = 0, \pm 1, \ldots. \tag{14}$$

Da der Grad von S kleiner als der von p ist und $S(v, z)$ durch p_a teilbar ist, haben wir nach einem bekannten Satz aus der Funktionentheorie $u(j) = 0$, $j < 0$. Ferner erfüllt u gemäß der Wahl von v die Randbedingungen $(B_{j,1}^{(1)} u)(0)$, $j = 1, \ldots, l$. Aus (13) ersieht man die Gültigkeit dieser Beziehungen auch für $j = l + 1, \ldots, r$.

Es bleibt noch $A_1 u = 0$ zu zeigen. Berechnen wir die diskrete Fouriertransformierte dieses Ausdrucks, so sieht man, daß dies gleichbedeutend mit

$$0 = p(e^{i\xi})\, \tilde{u}(\xi) - e^{i\xi} S(u, e^{i\xi}) = e^{i\xi}\big(S(v, e^{i\xi}) - S(u, e^{i\xi})\big)$$

ist. Es genügt daher, $u(j) = v(j)$, $j = 0, \ldots, n_i + n_a - 1$, nachzuweisen. Nun ergibt sich aus (14) mit einer Zahl ϱ, für die jede Wurzel ζ von $p(z) = 0$ der Beziehung $|\zeta| < \varrho$ genügt,

$$\begin{aligned} 2\pi i\, u(j) &= \int_{|z|=\varrho} z^j p^{-1}(z) \sum_{k=1}^{n_i+n_a} a_k \sum_{l=0}^{k-1} v(k)\, z^{k-l-1}\, dz \\ &= \int_{|z|=\varrho} z^j p^{-1}(z) \sum_{k=0}^{n_i+n_a} a_k \sum_{l=0}^{n_i+n_a-1} v(k)\, z^{k-l-1}\, dz \\ &= \int_{|z|=\varrho} \sum_{l=0}^{n_i+n_a-1} v(k)\, z^{j-l-1}\, dz = 2\pi i\, v(j), \qquad j = 0, \ldots, n_i + n_a - 1, \end{aligned}$$

womit (6) vollständig bewiesen ist.

4. Der Fall variabler Koeffizienten

In diesem Paragraphen wird Satz (4) ähnlich dem bekannten Vorgehen bei partiellen Differentialgleichungen mittels einer Zerlegung der Einheit auf Satz (6) zurückgeführt.

Für jedes $y \in [c_0, d_0]$ verstehen wir unter $A_h'(y)$ den Operator

$$\big(A_h'(y)\, u\big)(x) = \begin{cases} \sum_{k=-\infty}^{\infty} a_k(y, 0)\, (E_h^k u)(x), & x \in [c_h + n_1 h, d_h - n_2 h]_h, \\ 0, \quad \text{sonst in } \mathbb{R}_h, & u \in \Omega_h, \quad h \in \Lambda_0, \end{cases}$$

und unter $B'^{(l)}_{j,h}$, $j=1,\dots,r$, die Operatoren

$$B'^{(1)}_{j,h}=\sum_{k=0}^{\infty} b^{(1)}_{jk}(0)E^k_h, \qquad B'^{(2)}_{j,h}=\sum_{k=0}^{\infty} b^{(2)}_{jk}(0)\,E^{-k}_h, \qquad h\in\Lambda_0,$$

mit konstanten Koeffizienten. Sind U, G Teilmengen von $\mathbb{R}$, so schreiben wir $U\Subset G$ falls $\bar{U}\subseteq G$ gilt. Mit K wird im folgenden stets eine von u und h unabhängige Konstante bezeichnet, nicht jedesmal notwendig dieselbe. Ein erster Hilfssatz lautet.

(15) *Zu jedem $y\in(c_0, d_0)$ existiert eine offene Umgebung $U\Subset(c_0, d_0)$ von y, so daß für fast alle $h\in\Lambda_0$ die Ungleichung*

$$|u|\leqq K|A'_h(y)\,u|, \qquad u\in l^2_h(U),$$

dann und nur dann gilt, wenn für fast alle $h\in\Lambda_0$ die Ungleichung besteht

$$|u|\leqq K|A_h u|, \qquad u\in l^2_h(U).$$

Beweis. Für jede Umgebung $U\Subset(c_0, d_0)$ von y besteht für genügend kleines $h\in\Lambda_0$ die Darstellung

$$\big(A_h-A'_h(y)\big)u=\sum_{k=-\infty}^{\infty}\big(a_k(\cdot,h)-a_k(y,0)\big)E^k_h u, \qquad u\in l^2_h(U).$$

Aufgrund der Stetigkeit von a_k kann man zu jedem $\varepsilon>0$ die Umgebung U noch so wählen, daß man für fast alle $h\in\Lambda_0$ die Abschätzung

$$\big|\big(A_h-A'_h(y)\big)\,u\big|\leqq\varepsilon|u|, \qquad u\in l^2_h(U),$$

hat. Hieraus folgt die Behauptung.

(16) *Es existiert eine offene Umgebung $U\Subset(-\infty, d_0)$ von c_0, so daß für fast alle $h\in\Lambda_0$ die Ungleichung*

$$|u|\leqq K\Big(|A'_h(c_0)\,u|+\sum_{j=1}^{r}\langle B'^{(1)}_{j,h}u\rangle_1\Big), \qquad u\in l^2_h(U)\cap H_h,$$

dann und nur dann gilt, wenn für fast alle $h\in\Lambda_0$ die Ungleichung besteht

$$|u|\leqq K\Big(|A_h u|+\sum_{j=1}^{r}\langle B^{(1)}_{j,h}u\rangle_1\Big), \qquad u\in l^2_h(U)\cap H_h.$$

Beweis. In analoger Weise wie im vorangehenden Hilfssatz kann man zu jedem $\varepsilon>0$ eine Umgebung U von c_0 finden, so daß für fast alle $h\in\Lambda_0$ die Abschätzung besteht

$$\big|\big(A_h-A'_h(c_0)\big)\,u\big|\leqq\varepsilon|u|, \qquad u\in l^2_h(U)\cap H_h.$$

Ebenso entnimmt man der Beziehung

$$(B^{(1)}_{j,h}-B'^{(1)}_{j,h})\,u=\sum_{k=0}^{\infty}\big(b^{(1)}_{jk}(h)-b^{(1)}_{jk}(0)\big)\,E^k_h u$$

aufgrund der Stetigkeit von b_{jk} für fast alle $h\in\Lambda_0$ die Ungleichung

$$\langle(B^{(1)}_{j,h}-B'^{(1)}_{j,h})\,u\rangle_1\leqq\varepsilon|u|, \qquad u\in H_h, \qquad j=1,\dots,r.$$

Bei genügend kleiner Wahl von ε erhält man somit die Behauptung des Satzes.

Bemerkung. Eine (16) entsprechende Aussage gilt auch für das rechte Intervallende.

Nach diesen Vorbereitungen können wir zum Beweis von Satz (4) kommen. Wir bemerken noch vorweg, daß zu jeder endlichen offenen Überdeckung $\{U_j\}$, $j=1,\ldots,\nu$, des Intervalls $[c_0, d_0]$ nichtnegative Funktionen $\zeta_j \in C(\mathbb{R})$ existieren mit den Eigenschaften

$$\text{(17)} \qquad \text{Träger}(\zeta_j) \subseteq U_j, \quad j=1,\ldots,\nu, \quad \sum_{j=1}^{\nu} \zeta_j(x) = 1, \quad x \in [c_h, d_h],$$

für fast alle $h \in \Lambda_0$. Die Funktionen $\{\zeta_j\}$, $j=1,\ldots,\nu$, bezeichnet man auch als eine zur Überdeckung $\{U_j\}$, $j=1,\ldots,\nu$, gehörige Zerlegung der Einheit.

Beweis von Satz (4). Wenn in der ersten der Ungleichungen in (15) bzw. (16) K die Konstante aus Satz (6) ist, so bilden die gemäß (15), (16) und der darauffolgenden Bemerkung existierenden Umgebungen eine offene Überdeckung von $[c_0, d_0]$, aus der man dem Satz von Heine-Borel zufolge eine endliche Teilüberdeckung $\{U_j\}$, $j=1,\ldots,\nu$, auswählen kann. Es seien $y_j \in [c_0, d_0]$, $j=1,\ldots,\nu$, die zu U_j gehörigen Punkte, wobei wir $y_1 = c_0$, $y_2 = d_0$ annehmen wollen, so daß $U_j \Subset (c_0, d_0)$, $j=3,\ldots,\nu$, gilt. Durch die Vorschrift

$$(T_h u)(x) = h^{\frac{1}{2}} u(c_h + h x), \quad x \in \mathbb{R}_1, \quad u \in l_h^2[c_h, \infty), \quad h \in \Lambda_0,$$

wird eine isometrische Abbildung von $l_h^2[c_h, \infty)$ auf $l_1^2[0, \infty)$ definiert. Man überzeugt sich leicht, daß $E_1 T_h = T_h E_h$, $h \in \Lambda_0$, gilt. Die Operatoren $A'(y_j): l_1^2[0, \infty) \to l_1^2[0, \infty)$, $j=1,\ldots,\nu$, seien gegeben durch

$$A'(y_j) = \sum_{k=-\infty}^{\infty} a_k(y_j, 0)\, E_1^{k+m_j} = T_h E_h^{m_j} \sum_{k=-\infty}^{\infty} a_k(y_j, 0)\, E_h^k T_h^{-1} \text{ in } [0, \infty)_1,$$

wobei $a_k(y_j, 0) = 0$, $k < -m_j$, sei. Dabei nehmen wir $m_1 = n_1$ an.

Sei zunächst j zwischen 3 und ν gewählt. Für ein geeignetes $r_j \in \mathbb{N}$ wird $A'(y_j)$ von den Operatoren E_1^k, $k=0,\ldots,r_j$, überdeckt. Da ein $h_j > 0$ existiert, so daß jedes $v \in l_h^2(U_j)$, $h \leqq h_j$, den Randbedingungen

$$(E_1^k T_h v)(0) = 0, \quad k=0,\ldots,r_j,$$

genügt, erhalten wir aus Satz (6) für fast alle $h \in \Lambda_0$ das Bestehen von

$$\text{(18)} \qquad \begin{aligned} |v| = |T_h v| &\leqq K |A'(y_j) T_h v| = K |T_h E_h^{m_j} A_h'(y_j)\, v| \\ &= K |E_h^{m_j} A_h'(y_j)\, v| = K |A_h'(y_j)\, v|, \quad v \in l_h^2(U_j). \end{aligned}$$

Verwenden wir ferner die Beziehung $|(B_{j,1}'^{(1)} T_h v)(0)| = \langle B_{j,h}'^{(1)} v \rangle_1$, so liefert Satz (6) für fast alle $h \in \Lambda_0$ die Ungleichung

$$\text{(19)} \qquad \begin{aligned} |v| &\leqq K \Big(|A'(c_0)\, T_h v| + \sum_{j=1}^{r} |(B_{j,1}'^{(1)} T_h v)(0)| \Big) \\ &= K \Big(|T_h E_h^{n_1} A_h'(c_0)\, v| + \sum_{j=1}^{r} \langle B_{j,h}'^{(1)} v \rangle_1 \Big) \\ &= K \Big(|A_h'(c_0) v| + \sum_{j=1}^{r} \langle B_{j,h}'^{(1)} v \rangle_1 \Big), \quad v \in l_h^2(U_1) \cap H_h. \end{aligned}$$

Vertauscht man E_h mit E_h^{-1}, so erhält man durch dieselbe Überlegung auch für fast alle $h \in \Lambda_0$ die Abschätzung

$$|v| \leqq K\left(|A_h'(d_0)\, v| + \sum_{j=1}^{r} \langle B_{j,h}'^{(2)} v\rangle_2\right), \qquad v \in l_h^2(U_2) \cap H_h. \tag{20}$$

Sei nun $\{\zeta_j,\ j=1, \ldots, \nu\}$ eine zur Überdeckung $\{U_j,\ j=1, \ldots, \nu\}$ gehörige Zerlegung der Einheit. Wir ersetzen zu gegebenem $u \in H_h$ in (18) v durch $\zeta_j u$, $j=3, \ldots, \nu$, und in (19) (bzw. (20)) v durch $\zeta_1 u$ (bzw. $\zeta_2 u$). Dann kann man unter Ausnutzung der Hilfssätze (15), (16) für fast alle $h \in \Lambda_0$ die folgende Rechnung durchführen

$$\begin{aligned} |u| &\leqq \sum_{j=1}^{\nu} |\zeta_j u| \leqq K\left(\sum_{j=1}^{\nu} |A_h(\zeta_j u)| + \sum_{l=1}^{2} \sum_{j=1}^{r} \langle B_{j,h}^{(l)}(\zeta_l u)\rangle_l\right) \\ &\leqq K\nu\left(|A_h u| + \sum_{l=1}^{2} \sum_{j=1}^{r} \langle B_{j,h}^{(l)} u\rangle_l + C_A + C_B\right), \end{aligned}$$

wobei wir zur Abkürzung

$$C_A = \sum_{j=1}^{\nu} |A_h(\zeta_j u) - \zeta_j A_h u|$$

$$C_B = \sum_{l=1}^{2} \sum_{j=1}^{r} \langle B_{j,h}^{(l)}(\zeta_l u) - \zeta_l B_{j,h}^{(l)}(u)\rangle_l$$

gesetzt haben. Wenn wir noch zu jedem $\varepsilon > 0$ für fast alle $h \in \Lambda_0$ die Abschätzungen

$$C_A \leqq \varepsilon |u|, \qquad C_B \leqq \varepsilon |u|, \qquad u \in H_h,$$

zeigen, so ist der hinreichende Teil der Behauptung von Satz (4) bewiesen.

Die gewünschte Abschätzung für C_A folgt wegen der Stetigkeit der ζ_j aus der für $x \in [c_h + n_1 h,\ d_h - n_2 h]_h$ gültigen Beziehung

$$[A_h(\zeta_j u) - \zeta_j A_h u](x) = \sum_{k=-\infty}^{\infty} a_k(x, h)\, [(E^k - I)\, \zeta_j](x)\, (E^k u)(x).$$

Mit demselben Argument ergibt sich auch die Abschätzung für C_B.

Die Notwendigkeit der angegebenen Bedingungen erhält man in einfacher Weise unter Verwendung der Hilfssätze (15) und (16) durch Zurückführung auf Satz (6). Wir übergehen die Einzelheiten des Beweises.

5. Definition der Mehrschrittverfahren

Im zweiten Teil der Arbeit geben wir eine Anwendung der Koerzitivitätsungleichung (4) auf die Konvergenz von Mehrschrittverfahren bei Anfangswertaufgaben gewöhnlicher Differentialgleichungen m-ter Ordnung, $m \geqq 1$. Da wir Differentialgleichungen höherer Ordnung mitbehandeln, wird der Einfachheit halber darauf verzichtet, die Ergebnisse auch für Systeme von Differentialgleichungen auszuführen, obwohl das mit den verwendeten Methoden ohne weiteres möglich wäre.

In diesem Paragraphen formulieren wir die Aufgabenstellung. Unter anderem werden die Hilbertschen Räume H^j, $j=0, \ldots, m$, von Gitterfunktionen und die Konvergenz einer Folge in $\{H^j \to \mathbb{H}^j\}$ $(h \in \Lambda)$ definiert. Wir führen die Mengen

$\mathscr{F}_0$ und $\mathscr{F}$ der zugelassenen Nichtlinearitäten ein und erklären die Randwertaufgabe $A_0(F_0, \gamma_0)$, das Mehrschrittverfahren $A(F, \eta)$ sowie die Konvergenz von $A(F, \eta)$ gegen $A_0(F_0, \gamma_0)$.

Wie in §1 seien c_0, d_0 zwei Punkte aus $\mathbb{R}$ mit $c_0 < d_0$. Dann sei $\mathbb{H}^0$ der Hilbertsche Raum der reellwertigen über (c_0, d_0) im Sinne von Lebesgue quadratisch integrablen Funktionen mit dem Skalarprodukt

$$((u, v))_0 = \int_{c_0}^{d_0} u(x)\, v(x)\, dx, \qquad u, v \in \mathbb{H}^0,$$

und der zugehörigen Norm $\|\cdot\|_0$. Für jede Zahl $j > 0$ schreiben wir ∂^j für den Differentialquotienten j-ter Ordnung und bezeichnen mit $\mathbb{H}^j$ den Sobolewschen Raum $W^{j,2}(c_0, d_0)$ der Funktionen aus $\mathbb{H}^0$, deren Ableitungen im Sinne der Distributionstheorie bis zur Ordnung j in $\mathbb{H}^0$ liegen, versehen mit dem Skalarprodukt und der zugehörigen Norm

$$((u, v))_j = \sum_{k=0}^{j} ((\partial^k u, \partial^k v))_0, \qquad \|u\|_j = ((u, u))_j^{\frac{1}{2}}, \qquad u, v \in \mathbb{H}^j.$$

Sei m eine für das Weitere fest gewählte natürliche Zahl. Die Menge $\mathscr{F}_0$ sei dann wie folgt definiert.

(21) *$\mathscr{F}_0$ ist die Menge aller Abbildungen $F_0: \mathbb{H}^m \to \mathbb{H}^0$ mit der Eigenschaft, daß eine Zahl K existiert mit $\|F_0 u_0\|_0 \leqq K$, $u_0 \in \mathbb{H}^m$.*

Zwei wichtige Teilmengen von $\mathscr{F}_0$ sind die Menge $\mathscr{K}_0$ und die wohl zuerst von Frehse [9] im Zusammenhang mit Differenzenverfahren verwendete Menge $\mathscr{K}_0'$ der im verallgemeinerten Sinne stückweise stetigen Funktionen.

(22) *Es ist $\mathscr{K}_0$ die Menge aller Abbildungen der Gestalt*

$$(F_0 u_0)(x) = f_0\big(x, u_0(x), \ldots, (\partial^{m-1} u_0)(x)\big), \qquad x \in [c_0, d_0],$$

mit einer Funktion $f_0: [c_0, d_0] \times \mathbb{R}^m \to \mathbb{R}$, die den folgenden drei Bedingungen genügt.

(i) *Für jedes $\gamma \in \mathbb{R}^m$ ist $f_0(\cdot, \gamma)$ meßbar in $[c_0, d_0]$.*

(ii) *Für fast alle $x \in [c_0, d_0]$ ist $f_0(x, \cdot) \in C(\mathbb{R}^m)$.*

(iii) *Es existiert eine Funktion $k_0 \in \mathbb{H}^0$, so daß die Abschätzung besteht*

$$|f_0(x, \gamma)| \leqq k_0(x), \qquad x \in [c_0, d_0], \qquad \gamma \in \mathbb{R}^m.$$

(23) *$\mathscr{K}_0'$ ist die Menge aller Abbildungen der Gestalt* (22), *für die eine offene Punktmenge $M \subseteq [c_0, d_0]$ vom Maß $d_0 - c_0$ und eine Zahl K existiert, so daß $f_0 \in C(M \times \mathbb{R}^m)$ ist und die Abschätzung besteht*

$$|f_0(x, \gamma)| \leqq K, \qquad x \in [c_0, d_0], \qquad \gamma \in \mathbb{R}^m. \tag{24}$$

Die Inklusionen $\mathscr{K}_0' \subseteq \mathscr{K}_0 \subseteq \mathscr{F}_0$ sind nicht schwer zu beweisen. Lipschitz-stetige Funktionen sind in $\mathscr{K}_0'$ nicht direkt enthalten, jedoch kann man sie in bekannter Weise außerhalb einer kompakten Punktmenge zu einer Funktion in $\mathscr{K}_0'$ so abändern, daß die Lösung der zugehörigen Anfangswertaufgaben dieselbe ist. $\mathscr{F}_0$ enthält außer $\mathscr{K}_0$ auch noch allgemeinere Klassen von Abbildungen, z.B. die von Spijker [23] angegebenen Integro-Differentialausdrücke.

Mit $\mathscr{F}_0$ stellen wir die folgende Klasse von Anfangswertaufgaben.

$A_0(F_0, \gamma_0)$: *Zu gegebenem* $F_0 \in \mathscr{F}_0$ *und* $\gamma_0 = (\gamma_0^{(0)}, \ldots, \gamma_0^{(m-1)}) \in \mathbb{R}^m$ *ist eine Lösung* $u_0 \in \mathbb{H}^m$ *der Anfangswertaufgabe gesucht*

$$\partial^m u_0 = F_0 u_0, \quad (\partial^j u_0)(c_0) = \gamma_0^{(j)}, \quad j = 0, \ldots, m-1.$$

Wir kommen nun zur Definition der Mehrschrittverfahren. Es seien Λ_0, Ω_h, H_h, E_h und E_h^{-1} definiert wie in § 2. Da es sich aus dem jeweiligen Zusammenhang ergibt, werden im folgenden allerdings von h abhängige Größen nicht mehr besonders gekennzeichnet, vielmehr erhalten von h unabhängige Größen i.allg. ein Index „$_0$". Wir schreiben daher nur noch Ω statt Ω_h, H statt H_h usw. Mit Λ wird stets eine Teilfolge von Λ_0 bezeichnet, nicht jedesmal notwendig dieselbe. Bei der Konvergenz einer für $h \in \Lambda$ erklärten Folge verwenden wir die Schreibweise $(h \in \Lambda)$ statt einer Durchnumerierung der Folge und des Symbols $(j \to \infty)$.

Für jede ganze Zahl $k \geqq 0$ sei P_k die orthogonale Projektion in $l^2(\mathbb{R})$ erklärt durch

$$(P_k u)(x) = \begin{cases} u(x), & x \in [c, d - k h]_h \\ 0, & \text{sonst in } \mathbb{R}_h, \end{cases} \quad u \in l^2(\mathbb{R}), \quad h \in \Lambda_0.$$

Die vorwärts genommenen Differenzenquotienten D^j: $\Omega \to \Omega$ sind durch

$$D^j = \left(\frac{E - I}{h}\right), \quad j = 0, 1, \ldots,$$

gegeben, und wir setzen

$$D_j = P_j D^j, \quad j = 0, 1, \ldots, m.$$

Damit erhalten wir Hilbertsche Räume H^j, $h \in \Lambda_0$, $j = 0, \ldots, m$, wenn wir $l^2[c, d]$, $h \in \Lambda_0$, mit dem Skalarprodukt und der zugehörigen Norm versehen

$$(u, v)_j = \sum_{k=0}^{j} (D_k u, D_k v), \quad |u|_j = (u, u)_j^{\frac{1}{2}}, \quad u, v \in l^2[c, d].$$

Um den von uns verwendeten Konvergenzbegriff definieren zu können, führen wir für jedes $h > 0$ die Funktion

$$\omega_h^{(0)}(x) = \begin{cases} h^{-1}, & x \in (0, h), \\ 0, & \text{sonst in } \mathbb{R}, \end{cases}$$

ein und setzen

$$\beta = \frac{d - c + h}{d_0 - c_0}, \quad \delta = \beta d_0 - d - h, \quad h \in \Lambda_0,$$

sowie für jedes $u \in H^0$, $h \in \Lambda_0$,

$$(J_0 u)(x) = h \sum_{y \in \mathbb{R}_h} u(y)\, \omega_h^{(0)}(\beta x - \delta - y), \quad x \in [c_0, d_0].$$

Damit ist ein Einbettungsoperator J_0: $H^0 \to \mathbb{H}^0$ $(h \in \Lambda_0)$ erklärt, mit dessen Hilfe wir die folgende Definition geben.

(25) *Sei* $0 \leqq j \leqq m$, *sei* $\{u\}$, $h \in \Lambda$, *eine Folge von Funktionen* $u \in H^j$, *und sei* $u_0 \in \mathbb{H}^j$. *Dann konvergiert*

$$u \to u_0 \quad (bzw.\ u \rightharpoonup u_0) \quad in \quad \{H^j \to \mathbb{H}^j\} \quad (h \in \Lambda)$$

genau wenn für $k=0,\ldots,j$ *gilt*

$$J_0 D_k u \to \partial^k u_0 \quad (bzw.\ J_0 D_k u \rightharpoonup \partial^k u_0) \quad in \quad \mathbb{H}^0 \quad (h\in\Lambda).$$

Zur Definition der Mehrschrittverfahren benötigen wir die Menge $\mathscr{F}$, die wie folgt gegeben ist.

(26) *Es sei* n *eine natürliche Zahl mit* $n\geqq m$ *sowie* $nh\leqq(d_0-c_0)$, $h\in\Lambda_0$, *und* $\mathscr{F}$ *sei die Menge aller Folgen* $\{F\}$, $h\in\Lambda_0$, *stetiger Abbildungen*

$$F\colon H^m\to P_n H^0\subseteq H^0, \qquad h\in\Lambda_0,$$

mit den Eigenschaften

(i) *Es existiert eine Zahl* K, *so daß gilt* $|Fu|_0\leqq K$, $u\in H^m$, $h\in\Lambda_0$.

(ii) *Zu jeder Folge* $\{F\}$, $h\in\Lambda_0$, *aus* $\mathscr{F}$ *existiert ein* $G_0\in\mathscr{F}_0$, *so daß für jede Folge* $\{u\}$, $h\in\Lambda$, *mit*

$$u\rightharpoonup u_0 \quad in \quad \{H^m\to\mathbb{H}^m\} \quad (h\in\Lambda)$$

folgt

$$Fu\to G_0 u_0 \quad in \quad \{H^0\to\mathbb{H}^0\} \quad (h\in\Lambda). \tag{27}$$

Ist $F_0\in\mathscr{K}_0'$, so kann man eine Folge $\{F\}$, $h\in\Lambda_0$, in $\mathscr{F}$ folgendermaßen konstruieren. Es sei Σ eine endliche Menge von Multiindizes der Gestalt $\mu=(\mu_0,\ldots,\mu_{m-1})$ mit ganzen Zahlen μ_j, $j=0,\ldots,m-1$, und a_μ, b_μ, $\mu\in\Sigma$, seien reelle Zahlen. Wir nehmen an, daß $x+b_\mu h\in[c_0,d_0]$ ist für $x\in[c,d-nh]$, $h\in\Lambda_0$. Dann sei $F\colon H^m\to H^0$, $h\in\Lambda_0$, erklärt durch

$$Fu=P_n\sum_{\mu\in\Sigma}a_\mu f_0\big(\cdot+b_\mu h,\, u(\cdot+\mu_0 h),\ldots,(D_{m-1}u)(\cdot+\mu_{m-1}h)\big). \tag{28}$$

Dabei setzen wir voraus, daß $f_0(x+b_\mu h,\cdot)\in C(\mathbb{R}^m)$ ist für jedes $x\in[c,d-nh]_h$, $h\in\Lambda_0$. Dann läßt sich die folgende Aussage beweisen.

(29) *Es sei* $F_0\in\mathscr{K}_0'$. *Dann liegt die Folge* $\{F\}$, $h\in\Lambda_0$, *aus* (28) *in* $\mathscr{F}$, *und es konvergiert*

$$Fu\to\sum_{\mu\in\Sigma}a_\mu F_0 u_0 \quad in \quad \{H^0\to\mathbb{H}^0\} \quad (h\in\Lambda)$$

für jede Folge $\{u\}$, $h\in\Lambda$, *mit*

$$u\rightharpoonup u_0 \quad in \quad \{H^m\to\mathbb{H}^m\} \quad (h\in\Lambda).$$

Beweis. Offensichtlich ist F für jedes feste $h\in\Lambda_0$ stetig. Sei nun J_0 der vor (25) definierte Einbettungsoperator und M die in der Definition von $\mathscr{K}_0'$ genannte offene Menge von Stetigkeitspunkten von f_0. Dann überzeugt man sich unter Verwendung von (33) leicht davon, daß für $x\in M$ und $(h\in\Lambda)$ gilt

$$(J_0 Fu)(x)\to\sum_{\mu\in\Sigma}a_\mu f_0\big(x,u_0(x),\ldots,(\partial^{m-1}u_0)(x)\big)=(F_0u_0)(x)\sum_{\mu\in\Sigma}a_\mu.$$

Aufgrund der Abschätzung (24) ist die Folge $\{J_0Fu\}$, $h\in\Lambda$, betragsmäßig beschränkt, so daß der Satz von Lebesgue die behauptete Konvergenz ergibt. Die Bedingung (26) (i) folgt unmittelbar aus (24).

Sei nun q ein Polynom vom Grade n mit reellen Koeffizienten und $\Delta^{(m)}$ der Operator

$$\Delta^{(m)} = h^{-m} P_n q(E), \quad h \in \Lambda_0. \tag{30}$$

Dann wird die Konvergenz des folgenden Mehrschrittverfahrens untersucht.

$A(F, \eta)$: Zu gegebenen Folgen $\{F\}$, $h \in \Lambda_0$, aus $\mathscr{F}$ und $\{\eta\}$, $h \in \Lambda_0$, von Funktionen $\eta: [c, c+(n-1)h]_h \to \mathbb{R}$ ist für jedes $h \in \Lambda_0$ eine Lösung $u \in H^m$ der Gleichungen gesucht

$$\Delta^{(m)} u = F u, \quad u(x) = \eta(x), \quad x \in [c, c+(n-1)h]_h. \tag{31}$$

Der Begriff der Konvergenz eines Mehrschrittverfahrens ist wie folgt definiert.

(32) *Das Mehrschrittverfahren $A(F, \eta)$ heißt konvergent gegen $A_0(F_0, \gamma_0)$, genau wenn es für jedes $h \in \Lambda_0$ eine Lösung u von (31) gibt und für jedes $\Lambda \leqq \Lambda_0$ eine Teilfolge $\Lambda' \leqq \Lambda$ und eine Lösung u_0 von $A_0(F_0, \gamma_0)$ existiert mit*

$$u \to u_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda').$$

Wenn es höchstens eine Lösung u_0 von $A_0(F_0, \gamma_0)$ gibt, so kann man mit bekannter Schlußweise zeigen, daß für ein konvergentes Mehrschrittverfahren sogar die gesamte Folge $\{u\}$, $h \in \Lambda_0$, gegen u_0 konvergiert.

Wie wir in [12] gezeigt haben, folgt aus

$$u \rightharpoonup u_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda)$$

auch

$$\max_{x \in [c, d-jh]_h} |(D^j u - \partial^j u_0)(x)| \to 0 \quad (h \in \Lambda), \quad j = 0, \ldots, m-1, \tag{33}$$

wobei u_0 gegebenenfalls in eine Umgebung von $[c_0, d_0]$ unter Erhaltung der Differenzierbarkeitseigenschaften fortgesetzt zu denken ist. Daher konvergieren die Lösungen eines konvergenten Mehrschrittverfahrens mitsamt ihren Differenzenquotienten bis zur Ordnung $m-1$ einschließlich gleichmäßig auf dem Gitter, während die Differenzenquotienten m-ter Ordnung noch im Mittel konvergent sind.

6. Stabilität und Konsistenz

Die Begriffe der Stabilität und Konsistenz eines Mehrschrittverfahrens werden in diesem Paragraphen definiert. Die Stabilität wird durch notwendige und hinreichende algebraische Bedingungen charakterisiert und einige die Konsistenzbedingung betreffende Zusammenhänge werden dargelegt.

Für die anschließende Stabilitätsbedingung benötigen wir die folgende Halbnorm

$$[u] = \left(\sum_{j=0}^{m-1} |(D^j u)(c)|^2 + h \sum_{j=1}^{n-m} |(D^m u)(c+(j-1)h)|^2 \right)^{\frac{1}{2}}, \quad u \in H^0.$$

(34) *Das Mehrschrittverfahren $A(F, \eta)$ heißt stabil, genau wenn eine Zahl K existiert, so daß für fast alle $h \in \Lambda_0$ gilt*

$$|u|_m \leqq K(|\Delta^{(m)} u|_0 + [u]), \quad u \in H^0.$$

Wenn $z=1$ eine m-fache Wurzel von q ist, so ist diese Definition mit der folgenden äquivalent.

(35) *Das Mehrschrittverfahren $A(F,\eta)$ ist unter der Voraussetzung $(\partial^j q)(1)=0$, $j=0,\ldots,m-1$, genau dann stabil, wenn die folgenden beiden Bedingungen erfüllt sind.*

(i) *Aus $q(z)=0$ und $|z|\geqq 1$ folgt $z=1$.*

(ii) *Es ist $(\partial^m q)(1)\neq 0$.*

Beweis. Da q eine m-fache Wurzel bei $z=1$ hat, läßt sich $\Delta^{(m)}$ in der Gestalt

$$\Delta^{(m)}=P_n\, p(E)\, D_m, \qquad h\in\Lambda_0, \tag{36}$$

schreiben, wobei $m!\, p(1)=(\partial^m q)(1)$ ist. Wenn wir (i), (ii) als erfüllt annehmen, so besitzt p nur Wurzeln vom Betrage kleiner Eins, so daß $p=p_i$ (s. § 2) ist. Der Grad von p_i ist $n-m$, und es ist leicht zu sehen, daß die Polynome z^j, $j=0,\ldots,n-m-1$, p_i aufspannen. Daher folgt aus Satz (4) für $u\in H^0$ und fast alle $h\in\Lambda_0$ die a priori-Abschätzung

$$|D_m u|_0^2\leqq K\left(|\Delta^{(m)}u|_0^2+h\sum_{j=1}^{n-m}|(D^m u)\,(c+(j-1)\,h)|^2\right).$$

Macht man noch von der leicht zu beweisenden diskreten Friedrichsschen Ungleichung

$$|u|_{m-1}\leqq K\left(\sum_{j=0}^{m-1}|(D^j u)\,(c)|+|D_m u|_0\right), \qquad u\in H^0, \qquad h\in\Lambda_0,$$

Gebrauch, so ist die Stabilität bewiesen.

Der Beweis der Umkehrung folgt aus dem nächsten Satz.

(37) *Ist unter der Voraussetzung $(\partial^j q)(1)=0$, $j=0,\ldots,m-1$, eine der Bedingungen (35) (i) oder (ii) verletzt, so existieren für fast alle $h\in\Lambda_0$ Lösungen $u\in H^m$ von $\Delta^{(m)}u=0$ mit der Eigenschaft*

$$|D_m u|_0=1, \qquad (D^j u)\,(c)=0, \qquad j=0,\ldots,m-1, \qquad h\in\Lambda_0,$$

und

$$h\sum_{j=1}^{n-m}|(D^m u)\,(c+(j-1)\,h)|^2\to 0 \quad (h\in\Lambda_0).$$

Beweis. Ist eine der Bedingungen (35) (i) oder (ii) verletzt, so besitzt das Polynom p aus (36) eine Wurzel ζ mit $|\zeta|\geqq 1$. Da p reelle Koeffizienten hat, ist auch die konjugiert komplexe Zahl $\bar{\zeta}$ eine Wurzel von p, und man sieht leicht, daß jede Funktion $u\in H^0$ mit

$$(D^m u)\,(c+jh)=\zeta^j+\bar{\zeta}^j, \qquad c+jh\in[c,d-mh] \tag{38}$$

eine Lösung von $\Delta^{(m)}u=0$, $h\in\Lambda_0$, ist. Wenn wir

$$(D^j u)\,(c)=0, \qquad j=0,\ldots,m-1, \tag{39}$$

vorschreiben, so ist u durch (38) und (39) eindeutig bestimmt. Eine elementare Überlegung zeigt, daß $|(D^m u)\,(x)|\geqq\frac{1}{2}$ in mindestens der Hälfte der Punkte $x\in[c,d-mh]_h$ ist, so daß man durch Normierung von u auch die restlichen Bedingungen erreichen kann.

Die Konsistenz eines Mehrschrittverfahrens wird üblicherweise unter Verwendung der Existenz einer Lösung von $A_0(F_0, \gamma_0)$ definiert. Da wir die Existenz mitbeweisen wollen, müssen wir hier eine andere Definition verwenden. Eine weitere Schwierigkeit liegt darin, daß wir nicht die eindeutige Lösbarkeit von $A_0(F_0, \gamma_0)$ voraussetzen. Zu diesem Zweck führen wir den Begriff der approximierbaren Lösung ein.

(40) *Eine Funktion $u_0 \in \mathbb{H}^m$ heißt durch $A(F, \eta)$ approximierbar, genau wenn eine Teilfolge $\{u\}$, $h \in \Lambda$, von Lösungen der Gln.* (31) *existiert mit*

$$u \rightharpoonup u_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda).$$

Dann geben wir die folgende Konsistenzdefinition (s. auch [16, 23]).

(41) *Das Mehrschrittverfahren $A(F, \eta)$ heißt konsistent mit $A_0(F_0, \gamma_0)$, genau wenn die folgenden drei Bedingungen erfüllt sind.*

(i) *Es konvergiere*

$$(D^j \eta)(c) \to \gamma_0^{(j)} \quad (h \in \Lambda_0), \qquad j = 0, \ldots, m-1, \tag{42}$$

sowie

$$h \sum_{j=1}^{n-m} |(D^m \eta)(x + (j-1)h)|^2 \to 0 \quad (h \in \Lambda_0). \tag{43}$$

(ii) *Es ist* $(\partial^j q)(1) = 0$, $j = 0, \ldots, m-1$.

(iii) *Für jede durch $A(F, \eta)$ approximierbare Funktion $u_0 \in \mathbb{H}^m$ gilt*

$$\frac{1}{m!} (\partial^m q)(1) F_0 u_0 = G_0 u_0, \tag{44}$$

wobei $G_0 \in \mathscr{F}_0$ die gemäß (27) *der Folge $\{F\}$, $h \in \Lambda_0$, zugeordnete Abbildung ist.*

Die Bedingung (41) (iii) ist natürlich unter der weiterreichenden Forderung

$$\frac{1}{m!} (\partial^m q)(1) F_0 u_0 = G_0 u_0, \qquad u_0 \in \mathbb{H}^m,$$

erfüllt. Wir verwenden hier die schwächere Variante, um später zeigen zu können, daß sie auch notwendig für die Konvergenz von $A(F, \eta)$ ist.

Statt (43) wird z.B. in [15] nur

$$(D^{m-1} \eta)(c + jh) \to \gamma_0^{(m-1)} \quad (h \in \Lambda_0), \qquad j = 1, \ldots, n-m,$$

verlangt, jedoch wird auch nur die Konvergenz der Differenzenquotienten bis zur Ordnung $m-1$ bewiesen. Wir zeigen in Satz (55) aber auch die Konvergenz der Differenzenquotienten m-ter Ordnung. Die entsprechende Eigenschaft ist dann natürlich auch von der Anlaufrechnung zu fordern mit der man die Funktionen η berechnet. In gleicher Weise benötigte man etwa für den Beweis der schwachen Konvergenz in $\{H^m \to \mathbb{H}^m\}$ der Lösungen von $A(F, \eta)$ statt (43) die Existenz einer Zahl K mit der gilt

$$h \sum_{j=1}^{n-m} |(D^m \eta)(c + (j-1)h)|^2 \leqq K, \qquad h \in \Lambda.$$

Diese Bedingung ist ihrerseits bei einer Anlaufrechnung erfüllt, die eine schwach konvergente Folge $\{D^m \eta\}$, $h \in \Lambda_0$, liefert.

Für spätere Zwecke und um die Verbindung mit der üblicherweise verwendeten Konsistenzbedingung herzustellen, wollen wir noch die Teile (ii) und (iii) der Konsistenzdefinition (41) näher analysieren. Hierzu benötigen wir die Bedingung (M).

(M) *Für jedes* $\gamma_0 \in \mathbb{R}^m$ *sei das Mehrschrittverfahren* $A(0, \eta)$ *konvergent gegen* $A_0(0, \gamma_0)$, *falls* η *die Beziehungen erfüllt*

$$(D^j \eta)(c) = \gamma_0^{(j)}, \quad j = 0, \ldots, m-1, \quad h \in \Lambda_0,$$

und

$$h \sum_{j=1}^{n-m} |(D^m \eta)(c + (j-1)h)|^2 \to 0 \quad (h \in \Lambda_0).$$

Hiermit können wir die folgende Behauptung aussprechen.

(45) *Unter der Bedingung* (M) *gilt* $(\partial^j q)(1) = 0$, $j = 0, \ldots, m-1$.

Beweis. Wir nehmen zuerst $\gamma_0 = (1, 0, \ldots, 0)$. Dann ist $u_0 = 1$ die Lösung von $A_0(\gamma_0, 0)$. Die Lösungen u von $A(0, \eta)$ genügen der Gleichung

$$(\Delta^{(m)} u)(c) = h^{-m}(q(E)\, u)(c) = 0, \quad h \in \Lambda_0. \tag{46}$$

Aufgrund der vorausgesetzten Konvergenz gilt unter Verwendung von (33)

$$(D^j u)(c) \to (\partial^j u_0)(c_0) \quad (h \in \Lambda_0), \quad j = 0, \ldots, m-1, \tag{47}$$

so daß wir nach Multiplikation mit h^m aus (46) durch Grenzübergang $q(1) = 0$ erhalten. Daher läßt sich q in die Faktoren $q(z) = (z-1)\, q^{(1)}(z)$ zerlegen. Im Falle $m > 1$ nimmt man nun $\gamma_0 = (0, 1, 0, \ldots, 0)$ und erhält aus (46) nach Multiplikation mit h^{m-1} und Verwendung von (47)

$$0 = (q^{(1)}(E)\, D^1 u)(c) \to q^{(1)}(1) \quad (h \in \Lambda_0).$$

Daher ist q sogar durch $(z-1)^2$ teilbar. Im Falle $m > 2$ verfährt man in gleicher Weise weiter und erhält so die Behauptung.

Der Zusammenhang mit der bekannten Konsistenzbedingung wird durch den folgenden Satz hergestellt.

(48) *Es sei* (41)(i) *erfüllt. Besitzt* $A_0(F_0, \gamma_0)$ *höchstens eine Lösung und ist* $A(F, \eta)$ *stabil, so gilt unter der Voraussetzung von* (41)(ii), (iii)

$$(\Delta^{(m)} - F)\, u \to 0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda) \tag{49}$$

für jede Folge $\{u\}$, $h \in \Lambda$, *für die*

$$u \to u_0 \quad \text{in} \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda) \tag{50}$$

gegen die Lösung u_0 *von* $A_0(F_0, \gamma_0)$ *strebt. Gibt es umgekehrt für jede Lösung* u_0 *von* $A_0(0, \gamma)$, $\gamma \in \mathbb{R}^m$, *eine Folge* $\{u\}$, $h \in \Lambda$, *die* (49) *und* (50) *mit* $F = 0$, $h \in \Lambda$, *erfüllt, so gilt* (41)(ii). *Besitzen darüber hinaus die Gleichungen*

$$(\partial^m q)(1)\, \partial^m u_0 = m!\, G_0 u_0, \quad (\partial^j u_0)(c_0) = \gamma_0^{(j)}, \quad j = 0, \ldots, m-1, \tag{51}$$

höchstens eine Lösung und existiert eine Lösung u_0 *von* $A_0(F_0, \gamma_0)$ *sowie eine Folge* $\{u\}$, $h \in \Lambda$, *die* (49) *und* (50) *genügen, so ist auch* (41)(iii) *erfüllt.*

Beweis. (i) Wir beginnen mit dem Nachweis des ersten Teils der Behauptung. Wie wir im nächsten Paragraphen zeigen, ist unter den gegebenen Voraussetzungen $A(F, \eta)$ gegen $A_0(F_0, \gamma_0)$ konvergent. Die eindeutig bestimmte Lösung u_0 ist daher durch $A(F, \eta)$ approximierbar. Aufgrund von (41)(ii) haben wir für $\Delta^{(m)}$ die Darstellung (36). Unter Verwendung von (54) erhält man somit

$$(\Delta^{(m)} - F)\, u \to \frac{1}{m!}\, (\partial^m q)\,(1)\, \partial^m u_0 - G_0 u_0 \quad (h \in \Lambda).$$

Setzen wir hierin (41)(iii) ein, so ergibt sich die behauptete Konvergenz (49).

(ii) Im zweiten Teil der Behauptung weist man (41)(ii) in entsprechender Weise wie in (45) nach. Ist nun u_0 die Lösung von $A_0(F_0, \gamma_0)$, so sieht man unter Ausnutzung von (49) und (50), daß u_0 auch Lösung von (51) ist. Daher gilt $(\partial^m q)\,(1)\, F_0 u_0 = m!\, G_0 u_0$. Nun ist aber auch jede durch $A(F, \eta)$ approximierbare Funktion $v_0 \in \mathbb{H}^m$ Lösung von (51), so daß $v_0 = u_0$ sein muß und (41)(iii) bewiesen ist.

Abschließend geben wir noch eine einfache Form der Konsistenzbedingung für die Folge $\{F\}$, $h \in \Lambda_0$, aus (28) an.

(52) *Es sei* $\{F\}$, $h \in \Lambda_0$, *die Folge aus* (28). *Dann ist die Bedingung*

$$(\partial^m q)\,(1) = m! \sum_{\mu \in \Sigma} a_\mu$$

hinreichend für (41)(iii), *und sie ist auch notwendig, falls eine durch* $A(F, \eta)$ *approximierbare Funktion* u_0 *mit* $F_0 u_0 \neq 0$ *existiert.*

Beweis. In (29) haben wir die Beziehung

$$G_0 u_0 = \sum_{\mu \in \Sigma} a_\mu F_0 u_0$$

gezeigt. Hieraus folgt unmittelbar die Behauptung.

Wir zitieren noch zwei Ergebnisse aus [12], die wir für den Beweis von (55) benötigen.

(53) *Ist* K *eine positive Zahl und* $\{u\}$, $h \in \Lambda$, *eine Folge von Funktionen* $u \in H^m$ *mit* $|u|_m \leqq K$, $h \in \Lambda$, *so existiert ein Element* $u_0 \in \mathbb{H}^m$ *und eine Teilfolge* $\Lambda' \subseteq \Lambda$ *mit*

$$u \rightharpoonup u_0 \quad in \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda').$$

(54) *Es seien* k, l *ganze Zahlen,* $k \geqq 0$. *Dann folgt aus*

$$u \to u_0 \quad (bzw.\ u \rightharpoonup u_0) \quad in \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda)$$

auch

$$P_k E^l u \to u_0 \quad (bzw.\ P_k E^l u \rightharpoonup u_0) \quad in \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda).$$

7. Konvergenz der Mehrschrittverfahren

Ziel dieses Paragraphen ist es, den folgenden **Hauptsatz** über die Konvergenz der Mehrschrittverfahren zu beweisen.

(55) *Ein stabiles, mit* $A_0(F_0, \gamma_0)$ *konsistentes Mehrschrittverfahren* $A(F, \eta)$ *ist konvergent gegen* $A_0(F_0, \gamma_0)$. *Unter der Bedingung* (M) *ist auch die Umkehrung richtig.*

Der Beweis dieses Satzes wird über mehrere Hilfssätze erbracht.

(56) *Die Gln.* (31) *eines Mehrschrittverfahrens sind für alle* $h \in \Lambda_0$ *lösbar.*

Beweis. Sei $h \in \Lambda_0$ fest gewählt und T die durch

$$(Tu)(x) = \begin{cases} (\Delta^{(m)} u)(x), & x = [c, d - nh]_h, \\ u(a + jh), & x = d - (n - j - 1)h, \quad j = 0, \ldots, n-1, \end{cases}$$

definierte Abbildung von $H^m \to H^0$. Die homogene Gleichung $Tu = 0$ besitzt nur die triviale Lösung, so daß gemäß den Sätzen über Operatoren in endlichdimensionalen Räumen die Inverse T^{-1}: $H^0 \to H^m$ existiert. Führen wir noch die Abbildung V: $H^m \to H^0$ ein durch

$$(Vu)(x) = \begin{cases} (Fu)(x), & x \in [c, d - nh]_h, \\ \eta(a + jh), & x = d - (n - j - 1)h, \quad j = 0, \ldots, n-1, \end{cases}$$

so sind die Gln. (31) mit

$$T^{-1} V u = u \tag{57}$$

äquivalent. Sei S_r die Sphäre vom Radius $r > 0$ in H^m. Dann existiert der Bedingung (26) (i) zufolge eine Zahl K unabhängig von r, so daß aus dem Bestehen der Gleichung $T^{-1} V u = \alpha u$ mit einer Zahl $\alpha \in \mathbb{R}$ und einer Funktion $u \in S_r$ folgt

$$|\alpha| \, r \leqq \|T^{-1}\| \, K.$$

Für genügend großes r entnimmt man hieraus $\alpha < 1$. Da $T^{-1}V$ auch stetig ist, erhält man die Existenz einer Lösung $u \in H^m$ von (57) und damit auch von (31) aus Petryshyn [18, Satz 1], womit die Behauptung bewiesen ist.

(58) *Wenn das Mehrschrittverfahren* $A(F, \eta)$ *stabil und mit* $A_0(F_0, \gamma_0)$ *konsistent ist, so ist das Anfangswertproblem* $A_0(F_0, \gamma_0)$ *lösbar und* $A(F, \eta)$ *ist konvergent gegen* $A_0(F_0, \gamma_0)$.

Beweis. Die Stabilitätsbedingung (34) ergibt, daß die Normen $|u|_m$, $h \in \Lambda_0$, der Lösungen u von $A(F, \eta)$ gleichmäßig beschränkt sind. Gemäß Satz (53) kann man daher eine Funktion $u_0 \in \mathbb{H}^m$ und eine Teilfolge $\Lambda \subseteq \Lambda_0$ finden mit

$$u \rightharpoonup u_0 \quad \text{in} \quad \{H^m \to \mathbb{H}^m\} \quad (h \in \Lambda). \tag{59}$$

Unter Verwendung der Darstellung (36) für $\Delta^{(m)}$ und von Satz (54) erhält man

$$\Delta^{(m)} u \rightharpoonup p(1) \, \partial^m u_0 \quad \text{in} \quad \{H^0 \to \mathbb{H}^0\} \quad (h \in \Lambda).$$

Der Voraussetzung (27) und der Konsistenzbedingung (41) (iii) zufolge erhält man daher durch Grenzübergang in (31), daß u_0 der Gleichung

$$p(1) \, \partial^m u_0 = G_0 u_0 = \frac{1}{m!} (\partial^m q)(1) \, F_0 u_0 = p(1) \, F_0 u_0$$

genügt. Aufgrund der Stabilität ist $p(1) \neq 0$. Da (33) und die Konsistenzbedingung (41) (i) außerdem noch $(\partial^j u_0)(c_0) = \gamma_0^{(j)}$, $j = 0, \ldots, m-1$, ergeben, ist u_0 Lösung von $A_0(F_0, \gamma_0)$.

Es bleibt die starke Konvergenz der Folge $\{u\}$, $h\in\Lambda$, zu zeigen. Dazu verwenden wir das Kriterium 2.3 aus [12] und zeigen, daß zu jedem $\varepsilon>0$ ein $\varphi\in C^m(\mathbb{R})$ existiert mit

$$|D_m(u-R\varphi)|_0\leqq\varepsilon \quad \text{für fast alle} \quad h\in\Lambda, \tag{60}$$

wobei $R\varphi$ die Restriktion von φ auf das Gitter $[c,d]_h$ bedeutet. Sei $\varphi\in C^m(\mathbb{R})$ beliebig. In [12] ist gezeigt worden, daß

$$R\varphi\to\varphi \quad \text{in } \{H^m\to\mathbb{H}^m\} \quad (h\in\Lambda)$$

konvergiert. Aus der Stabilitätsbedingung (34) erhält man die Abschätzung

$$\begin{aligned}|D_m(u-\varphi)|_0^2\leqq K|Fu-\Delta^{(m)}\varphi|_0^2+\sum_{j=0}^{m-1}|D^j(u-\varphi)(c)|^2\\ +h\sum_{j=1}^{n-m}|D^m(u-\varphi)\big(c+(j-1)h\big)|^2, \quad h\in\Lambda.\end{aligned} \tag{61}$$

Es ist leicht zu sehen, daß

$$h\sum_{j=1}^{n-m}|(D^j\varphi)\big(c+(j-1)h\big)|^2\to 0 \quad (h\in\Lambda)$$

strebt, und wie im ersten Teil des Beweises erschließt man

$$\Delta^{(m)}\varphi\to p(1)\,\partial^m\varphi \quad \text{in} \quad \{H^0\to\mathbb{H}^0\} \quad (h\in\Lambda).$$

Somit erhält man durch Grenzübergang in (61)

$$\overline{\lim_{h\in\Lambda}}\,|D_m(u-\varphi)|_0\leqq K\Big(\|\partial^m(u_0-\varphi)\|+\sum_{j=0}^{m-1}|\partial^j(u_0-\varphi)(c_0)|\Big).$$

Die rechte Seite dieser Ungleichung läßt sich bekanntlich durch $K_1\|u_0-\varphi\|_m$ abschätzen. Da $C^m[c_0,d_0]$ in $\mathbb{H}^m$ dicht liegt, kann man durch geeignete Wahl von φ erreichen, daß (60) gilt. Damit ist der noch verbliebene Teil der Behauptung bewiesen.

(62) *Unter der Bedingung* (M) *ist das Mehrschrittverfahren* $A(F,\eta)$ *stabil.*

Beweis. Wie wir in (45) gezeigt haben, ist bei Voraussetzung von (M) der Teil (ii) der Konsistenzbedingung erfüllt. Wenn $A(F,\eta)$ nicht stabil ist, so folgt aus den Sätzen (35), (37), daß Lösungen $u\in H^m$, $h\in\Lambda$, der Gleichungen $\Delta^{(m)}u=0$, $|D_mu|_0=1$, $h\in\Lambda$, sowie

$$(D^ju)(c)=0, \quad j=0,\ldots,m-1, \quad h\sum_{j=1}^{n-m}|(D^mu)\big(c+(j-1)h\big)|^2\to 0$$

für $(h\in\Lambda)$ existieren. Der Bedingung (M) zufolge müßte die Folge $\{u\}$, $h\in\Lambda$, stark in $\{H^m\to\mathbb{H}^m\}$ gegen die Funktion $u_0=0$ konvergieren, was nicht möglich ist.

(63) *Ein gegen* $A_0(F_0,\gamma_0)$ *konvergentes Mehrschrittverfahren* $A(F,\eta)$ *ist unter der Bedingung* (M) *notwendig konsistent.*

Beweis. Das Bestehen von (41) (i) ist trivial. (41) (ii) haben wir bereits in (45) bewiesen. Zum Nachweis von (41) (iii) nehmen wir im Widerspruch an, daß eine

durch $A(F, \eta)$ approximierbare Funktion $u_0 \in \mathbb{H}^m$ existiert mit

$$(\partial^m q)(1) F_0 u_0 \neq m! G_0 u_0.$$

Dann ist u_0 nicht Lösung von $A_0(F_0, \gamma_0)$, obwohl jede durch $A(F, \eta)$ approximierbare Funktion eines konvergenten Verfahrens Lösung von $A_0(F_0, \gamma_0)$ sein muß.

Abschließend bemerken wir noch, daß sich mit unseren Überlegungen auch zeigen läßt, daß die Konvergenzordnung einer konvergenten Teilfolge von Lösungen von (30) mindestens gleich der Ordnung des Abschneidefehlers ist.

Literatur

1. Ansorge, R.: Konvergenz von Mehrschrittverfahren zur Lösung halblinearer Anfangswertaufgaben. Num. Math. **10**, 209–229 (1967).
2. Aubin, J. P.: Approximation des espaces de distributions et des opérateurs différentiels. Bull. Soc. math. France **12**, 1–139 (1967).
3. Butcher, J. C.: On the convergence of numerical solutions to ordinary differential equations. Math. of Comp. **20**, 1–10 (1966).
4. Céa, J.: Approximation variationelle des problèmes aux limites. Ann. Inst. Fourier (Grenoble) **14**, 345–444 (1964).
5. Collatz, L.: The numerical treatment of differential equations. Berlin-Heidelberg-New York: Springer 1966.
6. Dahlquist, G.: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. **4**, 33–53 (1956).
7. – Stability and error bounds in the numerical integration of ordinary differential equations. Trans. Roy. Inst. Technol. Stockholm **130**, 1–85 (1959).
8. Dejon, B.: Stronger than uniform convergence of multistep difference methods. Num. Math. **8**, 29–41 (1966). Addendum: Num. Math. **9**, 268–270 (1966).
9. Frehse, J.: Existenz und Konvergenz von Lösungen nichtlinearer elliptischer Differenzengleichungen unter Dirichletrandbedingungen. Math. Z. **109**, 311–343 (1969).
10. Grigorieff, R. D.: Über die Konvergenz von Differenzenverfahren bei linearen gewöhnlichen Differentialgleichungen. ZAMM **48**, T72–T75 (1968).
11. — Approximation von Eigenwertproblemen und Gleichungen zweiter Art in Hilbertschen Räumen. Math. Ann. **183**, 45–77 (1969).
12. — Die Konvergenz des Rand- und Eigenwertproblems linearer gewöhnlicher Differenzengleichungen. Num. Math. **15**, 15–48 (1970).
13. — Verallgemeinerte approximativ kompakte Operatoren und Anwendungen. Habilitationsschrift, Frankfurt a.M. 1969.
14. Henrici, P.: Discrete variable methods in ordinary differential equations. New York-London-Sydney: J. Wiley & Sons 1962.
15. — Error propagation for difference methods. New York-London: J. Wiley & Sons 1963.
16. Hull, T. E., Luxemburg, W. A. J.: Numerical methods and existence theorems for ordinary differential equations. Num. Math. **2**, 30–41 (1960).
17. Pereyra, V.: Iterated deferred corrections for nonlinear operator equations. Num. Math. **10**, 316–323 (1967).
18. Petryshyn, W. V.: On nonlinear P-compact operators in Banach space with applications to constructive fixed-point theorems. J. Math. Anal. Appl. **15**, 228–242 (1966).
19. Richtmyer, R. D., Morton, K. W.: Difference methods for initial-value problems. New York-London-Sydney: J. Wiley & Sons 1967.
20. Ryaben'kii, V. S.: Necessary and sufficient conditions for good definition of boundary value problems for systems of ordinary difference equations. U.S.S.R. Comp. Math. Math. Phys. **4**, 2, 43–61 (1964).
21. Schechter, M.: Integral inequalities for partial differential operators and functions satisfying general boundary conditions. Comm. Pure Appl. Math. **12**, 37–66 (1959).

22. Schempp, W.: Eine funktionalanalytische Theorie der linearen Diskretisierungsalgorithmen für Anfangswertprobleme gewöhnlicher Differentialgleichungen. Dissertation, 1—139, Bochum 1968.
23. Spijker, M. N.: Convergence and stability of step-by-step methods for the numerical solution of initial-value problems. Num. Math. **8**, 161—177 (1966).
24. Stetter, H. J.: A study of strong and weak stability in discretization algorithms. J. Siam Numer. Anal. **2**, 265—280 (1965).
25. — Anwendung des Äquivalenzsatzes von P. Lax auf inhomogene Probleme. ZAMM **39**, 396—397 (1959).
26. Stummel, F.: Über die Differenzenapproximation des Dirichletproblems für eine lineare elliptische Differentialgleichung zweiter Ordnung. Math. Ann. **163**, 321—339 (1966).
27. — Elliptische Differenzenoperatoren unter Dirichletrandbedingungen. Math. Z. **97**, 169—211 (1967).
28. — Rand- und Eigenwertaufgaben in Sobolewschen Räumen. Lecture Notes in Mathematics, Bd. **102**. Berlin-Heidelberg-New York: Springer 1969.
29. Thomée, V.: Elliptic difference operators and Dirichlet's problem. Contr. Diff. Equ. **3**, 301—324 (1964).
30. Thompson, R. J.: Difference approximations for inhomogeneous and quasi-linear equations. J. Soc. Indust. Appl. Math. **12**, 189—199 (1964).
31. Urabe, M.: Theory of errors in numerical integration of ordinary differential equations. J. Sci. Hiroshima Univ. Ser. A-I, **25**, 3—62 (1961).
32. Walter, W.: Differential- und Integral-Ungleichungen. Berlin-Göttingen-Heidelberg-New York: Springer 1964.
33. Watt, J. M.: Consistency, convergence and stability of general discretizations of the initial value problem. Num. Math. **12**, 11—22 (1968).

Dr. Rolf Dieter Grigorieff
Institut für Angewandte Mathematik
der Universität
D-6000 Frankfurt a. M.
Robert-Mayer-Straße 10

Numer. Math. 15, 219–223 (1970)

Eine Abschätzung der nichttrivialen Eigenwerte stochastischer Matrizen

HELMUT H. SCHAEFER

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag in Verehrung gewidmet

Eingegangen am 28. November 1969

Es ist seit langem bekannt (Perron [6]), daß eine strikt positive $n \times n$-Matrix A genau einen Eigenwert von maximalem Absolutbetrag besitzt, nämlich den Spektralradius $r = r(A)$. Diese Eigenschaft besitzen jedoch auch weitere Klassen positiver Matrizen, z.B. alle unzerlegbaren Matrizen (im Sinne von Frobenius [3]) mit nicht verschwindender Spur. Einer Arbeit von Hopf [4] kann man für strikt positive Matrizen eine Abschätzung des Betrages der von r verschiedenen Eigenwerte λ entnehmen; für jeden solchen Eigenwert gilt

$$|\lambda| \leqq \frac{M-m}{M+m} r, \tag{1}$$

wo m das kleinste, M das größte Element von A bezeichnet. In einer ausführlichen Diskussion dieses Resultats erhält Ostrowski [5] unter anderem die etwas schwächere Abschätzung

$$|\lambda| \leqq \frac{M^2-m^2}{M^2+m^2} r. \tag{2}$$

In der vorliegenden Mitteilung geben wir eine Abschätzung (s. unten (3)) der Eigenwerte $\lambda \neq r$ für solche Matrizen $A \geqq 0$, für die A oder die Transponierte ${}^tA$ wenigstens einen strikt positiven Eigenvektor besitzt. Wie wohlbekannt und leicht einzusehen ist, ist für solche Matrizen (außer wenn $A = 0$) stets $r > 0$ und $r^{-1}A$ einer (zeilen- oder spalten-)stochastischen Matrix ähnlich. Dies schließt auch Matrizen $A \geqq 0$ mit $m = 0$ ein, für die (1) und (2) trivial werden. Die Abschätzung (1) ist insofern bestmöglich, als für gewisse strikt positive Matrizen die in (1) gegebene Schranke angenommen wird (vgl. Hopf [4]); wir zeigen indessen an einem Beispiel, daß (3) schärfer als (1) sein kann; für stochastische Matrizen ist in der Tat (1) eine Konsequenz von (3).

Für das zentrale Theorem der Hopfschen Arbeit [4], das auf dem Begriff der Oszillation beruht, wurde von Bauer [1] ein elementarer Beweis gegeben. Darüber hinaus wird der Gedanke, einen zu $r(A)$ gehörigen Eigenvektor von A zur Abschätzung der übrigen Eigenwerte heranzuziehen, in einer Dissertation von Deutsch (Bauer, Deutsch und Stoer, [2]) systematisch und in sehr allgemeinem Rahmen (nämlich für positive Endomorphismen geordneter lokalkonvexer Räume) entwickelt. Ein Vergleich der dort benutzten Methoden und erzielten Ergebnisse

mit dem in dieser Note verwendeten Ansatz (der ebenfalls nicht auf den endlichdimensionalen Fall beschränkt ist) wäre sicherlich interessant.

1. Wir legen die Bezeichnungen fest und erinnern an einige elementare Tatsachen aus der Theorie positiver Matrizen. Eine $n\times n$-Matrix $A=(a_{ik})$ heißt positiv, in Zeichen: $A\geqq 0$, wenn $a_{ik}\geqq 0$ für $i, k=1, \ldots, n$. Ist $a_{ik}>0$ für alle i, k, so heißt A strikt positiv, in Zeichen $A\gg 0$. Entsprechend nennen wir einen (komplexen) n-Vektor $x=(\xi_i)$ positiv, wenn $\xi_i\geqq 0$ für alle i, und strikt positiv, wenn $\xi_i>0$ für alle i. Wir benutzen entsprechend die Symbole $x\geqq 0$ und $x\gg 0$.

Wie üblich schreiben wir $\langle x, y\rangle=\sum_{i=1}^{n}\xi_i\eta_i$ für den Wert der kanonischen Bilinearform auf $\boldsymbol{C}^n\times\boldsymbol{C}^n$; das Tensorprodukt $x\otimes y$ zweier Vektoren bezeichnet die Matrix $(\eta_i\xi_k)$.

$\sigma(A)$ bezeichnet die Menge der komplexen Eigenwerte (Spektrum) von A, $r(A)=\sup\{|\lambda|: \lambda\in\sigma(A)\}$ den Spektralradius von A. Bekanntlich ist für $A\geqq 0$ $r=r(A)$ stets ein Eigenwert von A; in der Tat, mit $|x|=(|\xi_i|)$ und $R(\lambda)=(\lambda I-A)^{-1}$ gilt für $|\lambda|>r$ und alle $x\in\boldsymbol{C}^n$

$$|R(\lambda)x|=\left|\sum_{n=0}^{\infty}\lambda^{-(n+1)}A^n x\right|\leqq\sum_{n=0}^{\infty}|\lambda|^{-(n+1)}A^n|x|=R(|\lambda|)\,|x|.$$

Wäre also $r\notin\sigma(A)$, so wäre $R(\lambda)$ im Bereich $\{\lambda: |\lambda|>r\}$ beschränkt, während doch die meromorphe Funktion $\lambda\to R(\lambda)$ auf dem Spektralkreis $\{\lambda: |\lambda|=r\}$ wenigstens einen Pol besitzt. Es ist also $(R(\lambda)x_0)_{\lambda>r}$ für beliebiges $x_0\gg 0$ unbeschränkt; jeder Häufungspunkt der Familie $R(\lambda)x_0/\|R(\lambda)x_0\|$ für $\lambda\downarrow r$ ist ein positiver Eigenvektor zum Eigenwert r. (Hierbei kann $z\to\|z\|$ eine beliebige Norm auf $\boldsymbol{C}^n$ sein.)

Hieraus folgt insbesondere: *Besitzt $A\geqq 0$ einen strikt positiven Eigenvektor $u\gg 0$ zum Eigenwert ϱ, $\varrho u=Au$, so muß $\varrho=r(A)$ sein.* Denn es ist $rv={}^tAv$ für ein $v>0$, und aus $\varrho\neq r$ würde $\langle u, v\rangle=0$ folgen, ein Widerspruch.

2. Es sei $A\geqq 0$ und $r=r(A)>0$; ferner besitze ${}^tA$ einen strikt positiven Eigenvektor $y=(\eta_i)$. y werde durch $\langle y, e\rangle=1$ normiert, wo $e=(1, 1, \ldots, 1)$. Wir setzen

$$\mu=\inf_{x,u}\langle Ax, u\rangle,\qquad \mathsf{M}=\sup_{x,u}\langle Ax, u\rangle$$

unter den Nebenbedingungen $x\geqq 0$, $u\geqq 0$ und $\langle x, y\rangle=\langle u, e\rangle=1$. Wir behaupten:

Für jeden Eigenwert $\lambda\neq r$ von A gilt die Abschätzung

$$|\lambda|\leqq\min(r-\mu, \mathsf{M}-r). \tag{3}$$

Im Falle $\mu>0$ (d.h. $A\gg 0$) folgt hieraus

$$|\lambda|\leqq\frac{\mathsf{M}-\mu}{\mathsf{M}+\mu}r. \tag{4}$$

Beweis. Man überzeugt sich ohne Mühe, daß die Matrizen $B=A-\mu(y\otimes e)$ und $C=\mathsf{M}(y\otimes e)-A$ positiv sind; in der Tat, nach Definition von μ, M sind

die Bilinearformen $(x, u) \to \langle Bx, u\rangle$ und $(x, u) \to \langle Cx, u\rangle$ nicht negativ für alle $x \geqq 0$, $u \geqq 0$. Es gilt aber ${}^tCy = My - {}^tAy = (M-r)y$ und ${}^tBy = {}^tAy - \mu y = (r-\mu)y$. Wegen $y \gg 0$ folgt hieraus (s.o.) für die Spektralradien von B und C

$$r(B) = r - \mu, \qquad r(C) = M - r. \tag{5}$$

Ist nun $\lambda \neq r$ Eigenwert von A, $\lambda z = Az$ $(z \neq 0)$, so gilt offenbar $\langle z, y\rangle = 0$, also $Bz = Az = \lambda z$ und $Cz = -Az = -\lambda z$. Daraus folgt $|\lambda| \leqq r(B)$ und $|\lambda| \leqq r(C)$, also (3).

Zum Beweis von (4) unterscheiden wir zwei Fälle.

(a) Es ist $r(C) \geqq r(B)$, d.h. es ist $\mu + M \geqq 2r$. Es folgt

$$|\lambda| \leqq r(B) = \tfrac{1}{2}\big(r(B) - r(C)\big) + \tfrac{1}{2}\big(r(B) + r(C)\big)$$

oder

$$|\lambda|\Big(1 + \frac{r(C) - r(B)}{2|\lambda|}\Big) \leqq \tfrac{1}{2}\big(r(B) + r(C)\big) = \tfrac{1}{2}(M - \mu),$$

den trivialen Fall $\lambda = 0$ ausgeschlossen. Wegen $|\lambda| \leqq r$ (es ist immer $r = r(A)$ gesetzt) ist der Faktor auf der linken Seite nicht kleiner als

$$1 + (2r)^{-1}(M + \mu - 2r) = (2r)^{-1}(M + \mu),$$

woraus sich (4) unmittelbar ergibt.

(b) Es ist $r(C) < r(B)$, d.h. $\mu + M < 2r$. Aus $|\lambda| \leqq r(C)$ folgt

$$|\lambda| < \tfrac{1}{2}\big(r(C) + r(B)\big) = \tfrac{1}{2}(M - \mu) = \frac{M - \mu}{2r}\, r.$$

Wegen $2r > M + \mu$ ergibt sich hieraus erst recht (4).

Damit ist der Beweis von (3) und (4) erbracht.

Welches sind nun die tatsächlichen Werte von μ und M? Die durch die Nebenbedingungen für x und u festgelegten Variablenbereiche sind offenbar $(n-1)$-dimensionale Simplices

$$S_{n-1} = \{x\colon x \geqq 0,\ \langle x, y\rangle = 1\} \quad \text{und} \quad T_{n-1} = \{u\colon u \geqq 0,\ \langle u, e\rangle = 1\},$$

deren Extremalpunkte die ihnen angehörenden Vektoren x bzw. u mit jeweils genau einer von Null verschiedenen Koordinate sind. Bei festem x nimmt also die Linearform $u \to \langle Ax, u\rangle$ ihr Minimum in einem Extremalpunkt von T_{n-1} an, und Entsprechendes gilt für $x \to \langle Ax, u\rangle$ bei festem u. Daraus ergibt sich

$$\mu = \inf_{i,k} \frac{a_{ik}}{\eta_k}, \qquad M = \sup_{i,k} \frac{a_{ik}}{\eta_k}. \tag{6}$$

Ist y durch die geforderten Bedingungen noch nicht eindeutig bestimmt (in diesem Fall ist dann notwendig $\mu = 0$), so kann in der zweiten Gleichung noch das Infimum über alle zulässigen y gebildet werden.

Aus (6) läßt sich übrigens leicht die Ostrowskische Abschätzung (2) gewinnen. Wegen $ry = {}^tAy$ und $\langle y, e\rangle = 1$ gilt nämlich $r \max \eta_i \leqq M$ und $r \min \eta_i \geqq m$, wenn wie oben m bzw. M das kleinste bzw. größte Element der Matrix A bezeichnen. Hieraus folgt mit Hilfe von (6) $\mu \geqq mr/M$, $\mathsf{M} \leqq Mr/m$, also $\mu/\mathsf{M} \geqq m^2/M^2$. Aus (4) erhält man damit unmittelbar (2).

3. Die Überlegungen des vorigen Abschnitts vereinfachen sich, wenn A eine stochastische Matrix ist. Da A und ${}^tA$ die gleichen Eigenwerte haben, kann man ohne Beschränkung der Allgemeinheit A als spaltenstochastisch voraussetzen. Daher kann man in Abschnitt 2 $y = n^{-1}e$ wählen; aus (6) ergibt sich dann $\mu = nm$, $\mathsf{M} = nM$. Also:

Ist A eine stochastische $n \times n$-Matrix mit kleinstem Element m und größtem Element M, so genügt jeder nicht triviale (d.h. von 1 verschiedene) Eigenwert λ von A der Beziehung

$$|\lambda| \leqq \min(1 - nm, nM - 1). \tag{7}$$

Im Falle $m > 0$ folgt hieraus

$$|\lambda| \leqq \frac{M-m}{M+m}. \tag{8}$$

Für stochastische Matrizen ist daher (7) mindestens ebenso scharf wie die Hopfsche Abschätzung (1); wie wir noch bemerken werden, ist (7) im Falle $m + M < 2/n$ strikt schärfer als (8).

4. Wir wollen noch die Abschätzungen (3) und (4) vergleichen. Im Falle $\mu = 0$ wird (4) trivial; ist überdies $\mathsf{M} \geqq 2r$, so wird auch (3) trivial (und dies kann vorkommen, auch wenn A keinen Eigenwert $\lambda \neq r$ vom Betrag r besitzt). Nehmen wir also $\mu > 0$ an. Der Beweis von (4) zeigt, daß im Fall $\mu + \mathsf{M} = 2r$ die Aussagen (3) und (4) gleichwertig sind, während im Falle $\mu + \mathsf{M} \neq 2r$ (3) stets die schärfere Aussage liefert. Insbesondere ist für stochastische Matrizen die Aussage (7) bei $m = 0$ und $M < 2/n$ stets nichttrivial, und bei $m > 0$ und $m + M \neq 2/n$ stets schärfer als (8).

Wir betrachten ein Beispiel. Die Matrix

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

hat konstante Zeilensumme 10. Die nichttrivialen Eigenwerte sind $\lambda_{2,3} = \frac{1}{2}(1 \pm \sqrt{5})$, also gerundet $\lambda_2 = 1{,}62$ und $\lambda_3 = -0{,}62$. Man hat $r = r(A) = 10$, $\mu = 6$, $\mathsf{M} = 12$. Die Abschätzung (3) ergibt $|\lambda_{2,3}| \leqq 2$, (4) ergibt $|\lambda_{2,3}| \leqq 3{,}3$, (2) schließlich ergibt $|\lambda_{2,3}| \leqq 6$.

Literatur

1. Bauer, F. L.: An elementary proof of the Hopf inequality for positive operators. Num. Math. **7**, 331–337 (1965).
2. — Deutsch, E., Stoer, J.: Abschätzungen für die Eigenwerte positiver linearer Operatoren. Lin. Alg. Appl. **2**, 275–301 (1969).

3. Frobenius, G.: Über Matrizen aus nicht-negativen Elementen. Sitz.-Ber. Kgl. Preuß. Akad. Wiss. Berlin 1912, 456–477.
4. Hopf, E.: An inequality for positive linear integral operators. Journ. Math. Mech. **12**, 683–692 (1963).
5. Ostrowski, A. M.: Positive matrices and functional analysis. In: Recent advances in matrix theory [5], 81–101.
6. Perron, O.: Zur Theorie der Matrices. Math. Ann. **64**, 248–263 (1907).
7. Schneider, H. (ed.): Recent advances in matrix theory. Madison and Milwaukee: Univ. of Wisconsin Press 1964.

Prof. Dr. H. H. Schaefer
University of Maryland
Dept. of Mathematics, College Park
Maryland 20742
Universität Tübingen

Numer. Math. 15, 224–228 (1970)

Zur Konvergenz von Näherungsverfahren bezüglich verschiedener Normen

J. NITSCHE

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag gewidmet

Eingegangen am 19. November 1969

1. Unter einem linearen Approximationsschema in einem Banach-Raum X wollen wir eine Folge $\{\mathfrak{X}_n, P_n\}$ verstehen, wobei $\mathfrak{X}_n$ eine i.allg. aufsteigende Folge von endlichdimensionalen Teilräumen bedeute und P_n lineare Projektionsoperatoren auf $\mathfrak{X}_n$ seien. Die gleichmäßige Beschränktheit der P_n ist notwendig und hinreichend für das Bestehen einer Beziehung der Gestalt

$$\|x - P_n x\| \leqq c_1 \inf_{\xi \in \mathfrak{X}_n} \|x - \xi\|, \tag{1}$$

in welchem Fall wir das Approximationsschema fast-best-approximierend (f.b.a.) nennen (vgl. dazu Alexits [1]).

In praktischen Fällen läßt sich der vorgegebene lineare Raum, meist handelt es sich um einen Funktionenraum, in verschiedener Weise mit einer Norm ausstatten und so zu einem Banach-Raum vervollständigen. Als Beispiel seien im Raum der in einem Intervall I definierten Funktionen die Normen $(1 \leqq q \leqq \infty)$

$$\|x\|_{k,q} = \left\{ \sum_{\nu \leqq k} \int_I |x^{(\nu)}|^q \right\}^{1/q} \tag{2}$$

erwähnt, die zu den Sobolev-Räumen $W_{k,q}(I)$ führen.

In der vorliegenden Arbeit beschäftigen wir uns mit der Frage, wann aus der gleichmäßigen Beschränktheit der P_n in einer Norm dasselbe bezüglich einer anderen folgt. Diese Fragestellung ergibt sich bei Konvergenzuntersuchungen zum Ritzschen Verfahren in natürlicher Weise. Es werde etwa die Gleichung $A x = f$ in einem Hilbert-Raum H mit positivem, symmetrischen Operator A betrachtet. Der Ritz-Operator $R_n: H \to \mathfrak{X}_n$ ist die Orthogonalprojektion bezüglich der Norm $\|x\|_1 = (x, A x)^{\frac{1}{2}}$ und daher durch $\|R_n\|_1 \leqq 2$ beschränkt. Damit $A R_n x \to f$ für jedes $f \in H$ gilt, muß R_n f.b.a. im Raum mit der Norm $\|x\|_2 = \|A x\|$ sein. Die ersten hier zu nennenden Untersuchungen stammen von Polskii [20, 21], der ein zwar elegantes, in praktischen Fällen jedoch unhandliches Kriterium angab. Weitere hierher gehörende Arbeiten sind unten zusammengestellt. Zur Unterscheidung der sonst benützten Literaturstellen sind die entsprechenden Nummern durch einen Apostroph gekennzeichnet. Besonders hervorgehoben sei der Beitrag von Stephens [29]. In gewisser Weise kann unser Satz 1 als Verallgemeinerung seiner Beweismethode im Fall des Ritz-Verfahrens für Randwertprobleme 2. Ordnung bei Verwendung von Spline-Funktionen angesehen werden.

2. In diesem Abschnitt betrachten wir zwei Banach-Räume X_1 und X_2 und unterstellen, daß X_2 stetig in X_1 eingebettet ist. Gilt für das in X_1 definierte Approximationsschema $\{\mathfrak{X}_n, P_n\}$ auch $\mathfrak{X}_n \subset X_2$, so läßt es sich ebenso als solches in X_2 auffassen. Da $\mathfrak{X}_n$ endlich dimensional sein soll, existiert

$$\lambda_n = \sup_{\xi \in \mathfrak{X}_n} \frac{\|\xi\|_2}{\|\xi\|_1}. \tag{3}$$

In Anlehnung an die Bernsteinsche Ungleichung für trigonometrische Polynome wollen wir die Beziehung $\|\xi\|_2 \leqq \lambda_n \|\xi\|_1$ für $\xi \in \mathfrak{X}_n$ als Ungleichung vom Bernsteinschen Typ bezeichnen.

Weiterhin benötigen wir eine Approximationseigenschaft von $\{\mathfrak{X}_n\}$ relativ zu X_1 und X_2, charakterisiert durch die Zahlenfolge $\{\mu_n\}$, daß nämlich zu jedem $x \in X_2$ ein $\xi_x \in \mathfrak{X}_n$ gemäß

$$\begin{aligned} \|x - \xi_x\|_1 &\leqq \mu_n \|x\|_2 \\ \|x - \xi_x\|_2 &\leqq c_2 \|x\|_2 \end{aligned} \tag{4}$$

mit einer von x und n unabhängigen Konstanten c_2 existiere. Es gilt nämlich

Satz 1. Es gelte

$$\overline{\lim_{n \to \infty}} \lambda_n \mu_n < \infty. \tag{5}$$

Dann ist jedes in X_1 f.b.a. Approximationsschema auch f.b.a. in X_2.

Zum Beweis betrachten wir für ein $x \in X_2$ unter Zwischenschalten von ξ_x gemäß (4)

$$\begin{aligned} \|x - P_n x\|_2 &\leqq \|x - \xi_x\|_2 + \|P_n x - \xi_x\|_2 \\ &\leqq c_2 \|x\|_2 + \lambda_n \|P_n x - \xi_x\|_1 \\ &\leqq c_2 \|x\|_2 + \lambda_n \{\|x - P_n x\|_1 + \|x - \xi_x\|_1\}. \end{aligned}$$

Da nun nach Voraussetzung mit einer Konstanten c_3

$$\|x - P_n x\|_1 \leqq c_3 \inf_{\xi \in \mathfrak{X}_n} \|x - \xi\|_1 \leqq c_3 \|x - \xi_x\|_1$$

gilt, erhalten wir

$$\|x - P_n x\|_2 \leqq \{c_2 + (1 + c_3)\, \lambda_n \mu_n\} \|x\|_2$$

bzw. wie behauptet

$$\|P_n\|_2 \leqq 1 + c_2 + (1 + c_3) \overline{\lim}\, \lambda_n \mu_n.$$

Bemerkung. Die Größen μ_n — unter Weglassen der 2. Bedingung in (4) — bzw. λ_n werden von Aubin [3, 4] sowie Raviart [22] als Fehler- bzw. Stabilitätsfunktion bezeichnet. In Anlehnung an Brown [9] und Tihomirov [31] kann nach Aubin [4] aus (5) geschlossen werden, daß der N_n-dimensionale Durchmesser ($N_n = \dim \mathfrak{X}_n - 1$) der Einheitskugel von X_2 in X_1 durch λ_n^{-1} und μ_n nach unten und oben beschränkt ist und daß das Verfahren als Approximation der Elemente von X_2 in X_1 quasi-optimal ist.

3. Nunmehr sei H ein Hilbert-Raum mit Skalarprodukt $(\cdot, \cdot)$ und Norm $\|\cdot\|$, der stetig und dicht in einen Banach-Raum X_{-1} eingebettet ist. Der Dualraum

X_{-1}^* ist dann isometrisch dem stetig in H eingebetteten Banach-Raum

$$X_1 = \{x \mid x \in H,\ |(x, y)| \leqq \|x\|_1 \|y\|_{-1} \text{ für } y \in H\}.$$

Ist nun P ein in X_{-1} definierter linearer Operator mit Wertebereich in X_1, so läßt sich P auch als Operator in H bzw. X_1 auffassen. Bei Symmetrie von P in H gilt dann

$$\|P\|_1 = \|P\|_{-1}. \tag{6}$$

Ist umgekehrt P als beschränkter Operator in H und X_1 gegeben, so läßt sich P bei Symmetrie in H eindeutig auf ganz X_{-1} fortsetzen und es gilt wieder (6). So können Normabschätzungen in Räumen $X_{-1} \supset H$ aus denen in Räumen $X_1 \subset H$ gefolgert werden.

Der Zusammenhang wird besonders einfach, wenn auch X_{-1} ein Hilbert-Raum ist. Als Beispiel betrachten wir $H = \mathring{W}_{m.2}$. Die Räume $X_{-1} = \mathring{W}_{k.2}$ und $X_1 = W_{2m-k.2} \cap H$ sind für $0 \leqq k < m$ einander dual. Bei $\mathfrak{H}_n \subset W_{2m.2} \cap \mathring{W}_{m.2}$ sind die Orthogonalprojektionen $R_n : \mathring{W}_{m.2} \to \mathfrak{H}_n$ genau dann in $W_{k.2}$ f.b.a., wenn dasselbe in $W_{2m-k.2}$ richtig ist. Wir betrachten speziell das Dirichlet-Problem einer selbstadjungierten und koerziven elliptischen Differentialgleichung $Ax = f$ der Ordnung $2m$. Aus der obigen Bemerkung erkennen wir

Satz 2. Für die Ritz-Näherungen $x_n = R_n x$ der Gleichung $Ax = f$ gilt genau dann $Ax_n \to f$ für jedes f in $L_2(I) = W_{0.2}$, wenn R_n in $L_2(I)$ f.b.a. ist.

4. Als Beispiel wollen wir zunächst im Raum der periodischen Funktionen für $\mathfrak{X}_n$ die trigonometrischen Polynome n-ter Ordnung betrachten. Bei $X_1 = W_{r.2}^*$, $X_2 = W_{s.2}^*$ $(s > r)$ gilt

$$\lambda_n \leqq c_{r,s} n^{s-r}, \quad \mu_n \leqq c'_{r,s} n^{r-s}. \tag{7}$$

Daher gilt nach Satz 1 und 2

Satz 3. Ist $\{P_n\}$ eine Folge von linearen Projektionsoperatoren auf die Teilräume der trigonometrischen Polynome, so folgt aus der gleichmäßigen Beschränktheit der Normen in $W_{r.2}$ dasselbe in $W_{s.2}$ bei $s > r$. Ist jedes P_n darüber hinaus symmetrisch in einem $\mathring{W}_{k.2}$, so kann die Bedingung $s > r$ wegfallen.

Bemerkung. Dieser Satz läßt sich auch aus den bekannten Sätzen von Jackson und Bernstein folgern. Wesentlich scheint hier die Einfachheit der Herleitung aus (7) allein.

Als weiteres Beispiel betrachten wir polynomiale Spline-Funktionen. Zu einer Unterteilung π des Intervalles I bestehe $\mathfrak{S}_\pi$ aus den Funktionen, die in den Teilintervallen von π jeweils Polynome eines festen Grades sind und gewisse Übergangsbedingungen an den Knoten erfüllen. $\underline{\pi}$ bzw. $\bar{\pi}$ gebe die minimale bzw. maximale Gitterbreite an. Alle π mit $\bar{\pi} \leqq \gamma \underline{\pi}$ mögen γ-regulär heißen. In [17] wurde eine Ungleichung vom Bernsteinschen Typ der Form $(1 \leqq q \leqq \infty)$

$$\|\xi'\|_{L_q} \leqq c_4 \underline{\pi}^{-1} \|\xi\|_{L_q} \tag{8}$$

gezeigt. Dabei hängt c_4 nur vom Grad der benützten Funktionen, nicht hingegen von den Stetigkeitseigenschaften von $\xi \in \mathfrak{S}_\pi$ ab. Abschätzungen der Gestalt

$(1 \leqq q \leqq \infty)$

$$\|x - \xi_x\|_{L_q} \leqq c_5 \bar{\pi}^k \|x\|_{W_{k,q}}$$
$$\|x - \xi_x\|_{W_{k,q}} \leqq c_6 \|x\|_{W_{k,q}} \tag{9}$$

sind für $q=2$ und $q=\infty$ u.a. den Arbeiten von Schultz [23, 24] zu entnehmen. Der Fall $q=1$ läßt sich parallel dazu führen, so daß sich dann (9) für alle q mit Hilfe von Interpolationssätzen folgern läßt. Allerdings sind dabei gewisse Bedingungen an $\mathfrak{S}_\pi$ zu stellen, was wir als Bedingung $(A-k)$ bezeichnen wollen[1]. Es gilt daher

Satz 4. Erfüllen die Spline-Räume $\mathfrak{S}_\pi$ die Bedingung $(A-k)$, so ist jede in L_q f.b.a. Menge von Projektionsoperatoren für γ-reguläre Unterteilungen auch f.b.a. in $W_{k,q}$.

In [16] wurde für verschiedene Projektionen P_π auf den Raum der linearen Spline-Funktionen der Nachweis der Beschränktheit der Normen in $W_{0,\infty}$ und $W_{1,\infty}$ bzw. in $W_{0,2}$ und $W_{1,2}$ getrennt geführt. Satz 4 lehrt, daß jeweils der zweite Nachweis hätte wegfallen können. Eine offene Frage bleibt, ob sich bei der Kombination etwa von $W_{0,\infty}$ und $W_{1,2}$ der Satz 2 mit Vorteil heranziehen läßt.

Literatur

1. Alexits, G.: Einige Beiträge zur Approximationstheorie. Acta Scientiarum Mathematicarum **26**, 212–224 (1965).
2. Aubin, Jean-Pierre: Approximation des espaces de distributions et des opérateurs différentiels. Bull. Soc. Math. France Suppl. Mém. **12**, 139 p. (1967).
3. — Approximation of non-homogeneous Neumann problems — regularity of the convergence and estimates of errors in terms of n-width. MRC Technical Summary Report 924 August 1968 (The University of Wisconsin, Madison) Contract No.: DA-31-124-ARO-D-462.
4. — Optimal approximation and characterization of the error and stability functions in Banach spaces. MRC Technical Summary Report 935 September 1968, The University of Wisconsin, Madison, Math. Res. Center.

5'. Birkhoff, G. H., Boor, C. de: Error bounds for spline interpolation. J. Math. Mech. **13**, 827–836 (1964).

6'. — —Piecewise polynomial interpolation and approximation. Proc. of the Symp. on appr. of functions, Michigan 1964. Amsterdam: Elsevier Publishing Company 1965.

7'. Boor, C. de: On uniform approximation by splines. J. of Approximation Theory **1**, 219–235 (1968).

8'. — On the convergence of odd-degree spline interpolation. J. of Approximation Theory **1**, 452–463 (1968).

9. Brown, A. L.: Best n-dimensional approximation to sets of functions. Proc. London Math. Soc. **14**, 577–594 (1964).

10'. Cheney, E. W., Schurer, F.: A note on the operators arising in spline approximation. J. of Approximation Theory **1**, 94–102 (1968).

11'. Daugavet, I. K.: On the rate of convergence of Galerkin's method for ordinary differential equations [Russian]. Izv. Vyss. Ucebn. Zaved. Matematika 158–165 (1958).

1 Hierbei handelt es sich darum, daß einerseits der Grad der Polynome mindestens k sein muß und daß die Elemente von $\mathfrak{S}_\pi$ mindestens $(k-1)$-mal stetig differenzierbar sein müssen. Durch Modifikation der Normen ließe sich die letzte Einschränkung umgehen.

12'. Dziskariani, A. V.: On the rate of convergence of the Ritz method for eigenvalues of differential operators [Russian]. Soobsc. Akad. Nauk Gruzin. SSR **28**, 513–519 (1962).
13'. — On the methods of Ritz and Bubnov-Galerkin [Russian]. Sakharth. SSR Mecn. Akad. Moambe **45**, 11–16 (1967).
14'. Medvedev, V. A.: On the convergence of the Bubnov-Galerkin method. Prikl. Mat. Meh. **27**, 1148–1151 (1963) und J. Appl. Math. Mech. **27**, 1769–1774 (1964).
15'. Mihlin, S. G.: Über die Ritzsche Methode. Doklady Akademia Nauk SSSR **106**, 391–394 (1956).
16'. Nitsche, J.: Orthogonalreihenentwicklung nach linearen Spline-Funktionen. J. of Approximation Theory **2**, 66–78 (1969).
17. — Umkehrsätze für Spline-Approximationen. Composito Mathematica (erscheint demnächst).
18'. Nord, S.: Approximation properties of the spline fit. Nordisk Tidskr. Informations-Behandling (BIT) **7**, 132–144 (1967).
19'. Plehwe, A. v.: Eine Bemerkung zu einer Arbeit von A. V. Dzhishkariani. ZAMM **48**, 422–423 (1968).
20'. Polskii, N. I.: Über die Konvergenz gewisser Näherungsverfahren der Analysis. Ukr. mat. J. Kiew **7**, 56–70 (1955).
21'. — Projective methods in applied mathematics. Doklady Akad. Nauk SSSR **143**, 787 (1962) und Sov. Math. Doklady **3**, 488–492 (1962).
22. Raviart, P. A.: Sur l'approximation de certaines équations linéaires et non linéaires. Teil I und II. J. de Math. pures et appl. **46**, 11–107 und 109–183 (1967).
23'. Schultz, M. H.: Approximation theory of multivariate spline functions in Sobolev spaces. SIAM J. Numer. Anal. (to appear).
24'. — Multivariate L-spline interpolation. J. Approximation Theory (to appear).
25'. Schurer, F.: A note on interpolating periodic quintic splines with equally spaced nodes. J. Approximation Theory **1**, 493–500 (1968).
26'. — Cheney, E. W.: The norms of interpolating spline operators. Notices AMS **15**, 790 (1968).
27'. Sharma, A., Meir, A.: Degree of approximation of spline interpolation. J. Math. Mech. **15**, 759–767 (1966).
28'. Sobolevskii, P. E.: On equations with operators forming an acute angle. Dokl. Akad. Nauk SSSR **116**, 754–759 (1957).
29'. Stephens, A. B.: Convergence of the residual for Ritz-Galerkin approximations. Thesis, Univ. Maryland 1969.
30'. Swartz, Blair: $O(h^{2n+2-1})$ bounds on some spline interpolation errors. Bull. of the Am. Math. Soc. **74**, 1072–1078 (1968).
31. Tihomirov, V. M.: On n-dimensional diameters of certain functional classes. Dokl. Akad. Nauk SSSR **130**, 734–737 (1960) und Sov. Math. Dokl. **1**, 94–97 (1960).
32'. Vainikko, G. M.: On similar operators. Dokl. Akad. Nauk SSSR **179**, 1029–1031 (1968).

Prof. Dr. J. Nitsche
Institut für Angewandte Mathematik
der Universität Freiburg i. Br.
D-78 Freiburg, Hermann-Herder-Straße 10

Numer. Math. 15, 229–233 (1970)

Zur Restglieddarstellung der Gregoryschen Quadraturformel ungerader Ordnung

ERICH MARTENSEN

Herrn Lothar Collatz zum 60. Geburtstag am 6. 7. 1970 gewidmet

Eingegangen am 15. Dezember 1969

Die Quadraturformel von Gregory stammt aus der Frühzeit der Infinitesimalrechnung, ihr Restglied jedoch bildet bis in die jüngste Zeit einen Gegenstand mathematischer Untersuchung. Im Gegensatz zu anderen klassischen Quadraturmethoden ergab sich nämlich die Problematik, daß die zunächst bekannten Restglieder auf *mehrere* unbekannte Zwischenstellen aus dem Definitionsbereich des Integranden Bezug nehmen, die sich allenfalls auf zwei, nicht aber auf eine einzige Stelle reduzieren lassen [3]. Die damit verbundenen ungünstigen Auswirkungen auf die üblichen Fehlerabschätzungen liegen auf der Hand, und es erschien daher erstrebenswert, Restglieddarstellungen mit nur einer Zwischenstelle zu gewinnen. Ein solches Restglied ist seit einiger Zeit für die Gregorysche Quadraturformel *gerader* Ordnung bekannt [2]. Die folgenden Ausführungen beschäftigen sich mit der Quadraturformel *ungerader* Ordnung. Für diesen bislang ungelösten Fall werden zwei (nicht wesentlich voneinander verschiedene) Restglieddarstellungen angegeben, die dem genannten Kriterium genügen.

Bedeutet $f(x)$, $a \leqq x \leqq b$, eine stetige Funktion, $a = x_0 < x_1 < \cdots < x_n = b$ eine äquidistante Unterteilung des Definitionsbereiches in $n \geqq 1$ Teilintervalle der Schrittweite h und r eine ganze Zahl mit $0 \leqq r \leqq n$, so lautet die Gregorysche Quadraturformel r-ter Ordnung mit den üblichen Bezeichnungen für die Funktionswerte an den Stützstellen $x_0, x_1, \ldots, x_n$ und deren Differenzen

$$\int_a^b f(x)\,dx = h\left\{\sum_{\nu=0}^{n} f_\nu - \sum_{\varrho=0}^{r} L_{\varrho+1}\left[\nabla_n^\varrho + (-1)^\varrho \Delta_0^\varrho\right]\right\} + E_n^{(r)}; \tag{1}$$

die hier auftretenden Laplaceschen Zahlen

$$L_\mu = (-1)^{\mu-1} \int_0^1 \binom{t}{\mu} dt, \quad \mu = 1, 2, 3, \ldots, \tag{2}$$

erhält man am bequemsten aus der Rekursionsvorschrift

$$\sum_{\nu=1}^{\mu} \frac{L_\nu}{\mu - \nu + 1} = \frac{1}{\mu + 1}, \quad \mu = 1, 2, 3, \ldots. \tag{3}$$

Es ist leicht einzusehen, daß das Restglied $E_n^{(r)}$ unter den genannten Voraussetzungen bei festem r für $n \to \infty$ gegen Null konvergiert. Zur Gewinnung quantitativer Aussagen werden wir jedoch die üblichen Differenzierbarkeitsvoraussetzungen für den Integranden hinzunehmen und uns im übrigen, wie angekündigt, auf ungerade r beschränken.

Satz 1. *Es sei r eine ungerade Zahl und $f(x)$, $a \leqq x \leqq b$, eine $(r+2)$-mal stetig differenzierbare Funktion. Dann besitzt die Gregorysche Quadraturformel r-ter Ordnung die beiden Restglieddarstellungen*

$$E_n^{(r)} = -2L_{r+2}h^{r+2}\left\{f^{(r+1)}\left(\frac{a+b}{2} \mp \frac{b-a}{2}\right) \pm \frac{b-a}{2} f^{(r+2)}(\xi)\right\} \tag{4}$$

mit einer jeweils nicht näher bekannten Zwischenstelle $\xi \in [a, b]$.

Beweis durch Induktion. Da die Gregorysche Formel r-ter Ordnung alle Polynome von höchstens r-tem Grade exakt integriert [3], muß sie im Grenzfall $n = r$ aus Eindeutigkeitsgründen mit der r-ten Newton-Cotesschen Quadraturformel zusammenfallen. Daher braucht (4) für $n = r$ Teilintervalle nur mit dem Restglied jener Formel identifiziert zu werden. Hierfür kennt man im wesentlichen zwei Darstellungsformen. Immer unter der Voraussetzung ungerader r gilt nach Steffensen [5]

$$E_r^{(r)} = -2L_{r+2}h^{r+2}f^{(r+1)}(\xi) \tag{5}$$

mit einer Zwischenstelle $\xi \in [a, b]$. Zum anderen existiert nach dem Satz von Peano [4] die Darstellung

$$E_r^{(r)} = \int_a^b S(x)\, f^{(r+1)}(x)\, dx \tag{6}$$

mit einer (von r abhängenden) *Spline-Funktion* $S(x)$, $a \leqq x \leqq b$, von der in diesem Zusammenhang nur drei qualitative Eigenschaften, nämlich a) die Stetigkeit, b) der nichtpositive Wertebereich und c) die Symmetrie zur Intervallmitte, von Bedeutung sind [1, 6]. Vergleicht man (5) und (6) bei spezieller Wahl eines Polynoms $(r+1)$-ten Grades für $f(x)$ miteinander, so erhält man noch

$$\int_a^b S(x)\, dx = -2L_{r+2}h^{r+2}. \tag{7}$$

Wegen a) und der $(r+2)$-maligen stetigen Differenzierbarkeit für $f(x)$ kann (6) partiell integriert werden, und man findet, wenn

$$T(x) \equiv \int_b^x S(t)\, dt, \qquad a \leqq x \leqq b, \tag{8}$$

gesetzt wird,

$$E_r^{(r)} = [T(x)\, f^{(r+1)}(x)]_a^b - \int_a^b T(x)\, f^{(r+2)}(x)\, dx.$$

Die wegen b) mögliche Anwendung des Mittelwertsatzes liefert

$$E_r^{(r)} = -T(a)\, f^{(r+1)}(a) - f^{(r+2)}(\xi) \int_a^b T(x)\, dx \tag{9}$$

mit einer Zwischenstelle $\xi \in [a, b]$. Wiederum durch partielle Integration bekommt man

$$\int_a^b T(x)\,dx = \left[\left(x - \frac{a+b}{2}\right) T(x)\right]_a^b - \int_a^b \left(x - \frac{a+b}{2}\right) S(x)\,dx,$$

wobei rechts ein wegen c) schiefsymmetrischer Integrand erscheint:

$$\int_a^b T(x)\,dx = \frac{b-a}{2}\,T(a).$$

Durch Einsetzen in (9) entsteht

$$E_r^{(r)} = -T(a)\left\{f^{(r+1)}(a) + \frac{b-a}{2} f^{(r+2)}(\xi)\right\}$$

und damit wegen (7) und (8) die Restglieddarstellung für die r-te Newton-Cotes-Formel

$$E_r^{(r)} = -2L_{r+2}h^{r+2}\left\{f^{(r+1)}(a) + \frac{b-a}{2} f^{(r+2)}(\xi)\right\}. \tag{10}$$

Dies ist genau die Behauptung für $n = r$ Teilintervalle im Falle des oberen Vorzeichens in (4). Für das untere Vorzeichen schließt man analog, indem man die Rollen der Intervallenden miteinander vertauscht.

Wir nehmen jetzt an, die Behauptung sei für ein $n \geqq r$ Teilintervalle bewiesen, und unterteilen $[a, b]$ nunmehr in $n+1$ gleich große Intervalle der Schrittweite h mit den Teilpunkten $a = x_0 < x_1 < \cdots < x_n < x_{n+1} = b$. Dann gilt

$$\int_{x_0}^{x_n} f(x)\,dx = h\left\{\sum_{\nu=0}^{n} f_\nu - \sum_{\varrho=0}^{r} L_{\varrho+1}\left[\nabla_n^\varrho + (-1)^\varrho \Delta_0^\varrho\right]\right\}$$
$$- 2L_{r+2}h^{r+2}\left\{f^{(r+1)}(a) + \frac{x_n - x_0}{2} f^{(r+2)}(\xi_1)\right\}$$

mit einer Zwischenstelle $\xi_1 \in [x_0, x_n]$. Auf das verbleibende letzte Teilintervall wenden wir die Quadraturformel von Gregory-Newton für aufsteigende Differenzen bis zur Ordnung $r+1$ an:

$$\int_{x_n}^{x_{n+1}} f(x)\,dx = h\left\{f_{n+1} - \sum_{\varrho=0}^{r} L_{\varrho+1}\nabla_{n+1}^{\varrho+1}\right\} - L_{r+2}h^{r+3}f^{(r+2)}(\xi_2);$$

dabei gilt $\xi_2 \in [x_{n-r}, x_{n+1}]$. Addition ergibt

$$\int_{x_0}^{x_{n+1}} f(x)\,dx = h\left\{\sum_{\nu=0}^{n+1} f_\nu - \sum_{\varrho=0}^{r} L_{\varrho+1}\left[\nabla_{n+1}^\varrho + (-1)^\varrho \Delta_0^\varrho\right]\right\}$$
$$- 2L_{r+2}h^{r+2}\left\{f^{(r+1)}(a) + \frac{x_n + h - x_0}{2} f^{(r+2)}(\xi)\right\}$$

mit $\xi \in [x_0, x_{n+1}]$ und damit die Behauptung für $n+1$ Teilintervalle im Falle des oberen Vorzeichens in (4). Zum Beweis des anderen Falles wendet man die

Induktionsannahme auf das Intervall $[x_1, x_{n+1}]$ und die Gregory-Newton-Formel für absteigende Differenzen auf das Intervall $[x_0, x_1]$ an.

Durch die üblichen Abschätzungen beider Restglieddarstellungen (4) erhält man das folgende

Korollar. *Bedeute* $M_{r+2} \geqq 0$ *eine obere Schranke für* $|f^{(r+2)}(x)|$, $a \leqq x \leqq b$, *so gilt für das Restglied der Gregoryschen Quadraturformel ungerader Ordnung r die Abschätzung*

$$|E_n^{(r)}| \leqq 2 L_{r+2} h^{r+2} \left(\min\{|f^{(r+1)}(a)|, |f^{(r+1)}(b)|\} + \frac{b-a}{2} M_{r+2} \right). \tag{11}$$

Satz 2. *Besitzt die* $(r+1)$*-te Ableitung des Integranden an den Intervallenden entgegengesetztes Vorzeichen im Sinne von*

$$f^{(r+1)}(a)\, f^{(r+1)}(b) \leqq 0, \tag{12}$$

so kann das Restglied der Gregoryschen Quadraturformel ungerader Ordnung r durch

$$|E_n^{(r)}| \leqq L_{r+2}(b-a)\, h^{r+2} M_{r+2} \tag{13}$$

abgeschätzt werden. Diese Fehlerschranke ist insofern optimal, als der Faktor L_{r+2} *nicht durch eine allgemein gültige kleinere reelle Zahl ersetzt werden kann.*

Beweis. Wegen der entgegengesetzten Vorzeichen von $f^{(r+1)}(a)$ und $f^{(r+1)}(b)$ gibt es eine reelle Zahl $0 \leqq \gamma \leqq 1$ mit

$$\gamma f^{(r+1)}(a) + (1-\gamma) f^{(r+1)}(b) = 0.$$

Multipliziert man dann die erste Restglieddarstellung (4) mit γ und die zweite mit $1-\gamma$, so folgt durch Addition

$$E_n^{(r)} = -L_{r+2}(b-a)\, h^{r+2} \{\gamma f^{(r+2)}(\xi_1) - (1-\gamma) f^{(r+2)}(\xi_2)\}$$

und hieraus die behauptete Abschätzung

$$|E_n^{(r)}| \leqq L_{r+2}(b-a)\, h^{r+2} \{\gamma M_{r+2} + (1-\gamma) M_{r+2}\}.$$

Daß eine Verkleinerung des Vorfaktors allgemein nicht möglich ist, zeigt die Diskussion des Beispiels

$$f(x) = \frac{1}{(r+2)!} (x-a)^{r+2}, \qquad a \leqq x \leqq b,$$

das die gewünschte Vorzeicheneigenschaft (12) besitzt. Das Restglied lautet infolge (4)

$$E_n^{(r)} = -L_{r+2}(b-a)\, h^{r+2},$$

und (13) liefert nach Wahl der (optimalen) Schranke $M_{r+2} = 1$

$$|E_n^{(r)}| \leqq L_{r+2}(b-a)\, h^{r+2}.$$

In der Abschätzung wird also das Gleichheitszeichen angenommen, was eine Verbesserung der Fehlerschranke in dem genannten Sinne ausschließt.

Man erkennt leicht, daß die Voraussetzung (12) sicher nicht allgemein für das Zustandekommen der Restgliedabschätzung (13) entbehrt werden kann. Dies zeigt das Beispiel eines Polynoms $(r+1)$-ten Grades für $f(x)$, bei dem das Restglied (4) von Null verschieden ist, (13) aber nach Wahl von $M_{r+2}=0$ einen verschwindenden Fehler ergeben würde.

Schließlich sei noch bemerkt, daß die Fehlerschranke (13) mit der bekannten, optimalen und uneingeschränkt gültigen Fehlerabschätzung für die Gregorysche Quadraturformel *gerader* Ordnung [2] formal übereinstimmt.

Literatur

1. Kowalewski, G.: Interpolation und genäherte Quadratur. Leipzig und Berlin: Teubner 1932.
2. Martensen, E.: Optimale Fehlerschranken für die Quadraturformel von Gregory. Z. angew. Math. Mech. **44**, 159—168 (1964).
3. Milne-Thomson, L. M.: The calculus of finite differences. London: MacMillan & Co., Ltd. 1950.
4. Peano, G.: Residuo in formulas de quadratura. Mathesis **34**, 1—10 (1914).
5. Steffensen, J. F.: Interpolation. New York: Chelsea Publ. Comp. 1950.
6. Tollmien, W.: Über das Restglied der Mittelwertformeln für angenäherte Quadratur. Z. angew. Math. Mech. **29**, 193—198 (1949).

Mathematisches Institut
der Technischen Hochschule Darmstadt
D-61 Darmstadt, Hochschulstr. 1

Numer. Math. 15, 234–249 (1970)

Asymptotic Properties and the Convergence of Numerical Quadratures

JOZEF MIKLOŠKO

Received June 30, 1969

1. Introduction

The paper deals with certain asymptotic properties of interpolation quadratures and their convergence for continuous functions and for functions with singularity.

Have a Newton-Cotes type quadratures with weight function (further WF) $w(x)$

$$\int_a^b f(x)\, w(x)\, dx = \sum_{i=1}^{n} A_i^{(n)} f(x_i^{(n)}) + R_n(f) \tag{1}$$

and one of the Gauss type with positive WF $W(x)$

$$\int_a^b f(x)\, W(x)\, dx = \sum_{i=1}^{n} B_i^{(n)} f(x_i^{(n)}) + \overline{R}_n(f) \tag{2}$$

in which WF satisfy the required prerequisites, whereby in (1) to arbitrary knots $x_i^{(n)} \in [a, b]$ there will be

$$A_i^{(n)} = \int_a^b \frac{\omega_n(x)}{\omega_n'(x_i^{(n)})\,(x - x_i^{(n)})}\, w(x)\, dx \tag{3}$$

where $\omega_n(x) = \prod_{i=1}^{n} (x - x_i^{(n)})$ and in (2) again

$$B_i^{(n)} = \int_a^b \frac{Q_n(x)}{Q_n'(x_i^{(n)})\,(x - x_i^{(n)})}\, W(x)\, dx = \frac{-a_{n+1}^{(n+1)}}{a_n^{(n)} Q_n'(x_i^{(n)})\, Q_{n+1}(x_i^{(n)})} \tag{4}$$

where $\{Q_n(x)\}$ is the system of orthonormal polynomials (further OP) with WF $W(x)$ on $[a, b]$ given in (13) whereby $\{x_i^{(n)}\}$ are knots $Q_n(x)$. Of $\{Q_n(x)\}$ and $W(x)$ it will be supposed that they are uniformly bounded on $[A, B] \subset [a, b]$.

Further, the following concepts are used:

If

$$\lim_{n \to \infty} \sum_{i=1}^{n} A_i^{(n)} f(x_i^{(n)}) = \int_a^b f(x)\, w(x)\, dx \tag{5}$$

then the quadrature (1) is convergent for function $f(x)$.

The measurable function $f(x)$ is called an integrable with $w(x)$ if $\int_a^b f(x)\,w(x)\,dx<\infty$. It is obvious that if $f(x)=w(x)/W(x)$ then both $\int_a^b f^2(x)\,W(x)\,dx<\infty$.

The function $f(x)$ satisfies in point t the Dini-Lipschitz condition (further DLC) if for all sufficiently small values $|h|$ inequality

$$|f(t+h)-f(t)|\leqq M\,|\log|h||^{-\alpha-1}$$

holds (α, M are positive constants).

Let further $k(x)=w(x)/W(x)$.

2. Asymptotic Properties of Coefficients (3) and (4)

Theorem 1. Select knots $x_i^{(n)}\in[a,b]$ in (1) so as to make it for the system OP $\{Q_n(x)\}$ with WF $W(x)$ on $[a,b]$ for $k,l=0,1,\ldots,n-1$ to hold

$$\sum_{j=1}^{n} v_j^{(n)}\,Q_k(x_j^{(n)})\,Q_l(x_j^{(n)})=\begin{cases}0 & k\neq l\\ C_n & k=l\end{cases} \tag{6}$$

where $v_j^{(n)}>0$, $j=1,2,\ldots,n$ and for $n\to\infty$ $\left(\sum_{j=1}^{n} v_j^{(n)}\right)^{-1}$ is bounded. If function $k(x)$ is integrable with $w(x)$ and satisfies uniformly DLC in every point $x\in[A,B]\subset[a,b]$ then for every n and for every i such that $x_i^{(n)}$ is the inner point of $[A,B]$ it holds

$$A_i^{(n)}=\frac{v_i^{(n)}}{C_n}\,k(x_i^{(n)})+\varepsilon_i^{(n)} \tag{7}$$

where $\lim_{n\to\infty}\varepsilon_i^{(n)}=0$.

Proof. The polynomial of $(n-1)$-th degree

$$l_i(x)=\frac{\omega_n(x)}{\omega_n'(x_i^{(n)})\,(x-x_i^{(n)})}$$

from (3) can be expressed in the form

$$l_i(x)=\sum_{j=0}^{n-1} b_j^{(i)}\,Q_j(x). \tag{8}$$

If substituting $x=x_i^{(n)}$ into (8) for $i=1,2,\ldots,n$ then since $l_i(x_i^{(n)})=1$, $l_i(x_k^{(n)})=0$ a system of linear equation is obtained

$$\begin{pmatrix} Q_0(x_1^{(n)}) & Q_1(x_1^{(n)}) & \cdots & Q_{n-1}(x_1^{(n)})\\ \vdots & \vdots & & \vdots\\ Q_0(x_i^{(n)}) & Q_1(x_i^{(n)}) & \cdots & Q_{n-1}(x_i^{(n)})\\ \vdots & \vdots & & \vdots\\ Q_0(x_n^{(n)}) & Q_1(x_n^{(n)}) & \cdots & Q_{n-1}(x_n^{(n)}) \end{pmatrix}\begin{pmatrix} b_0^{(i)}\\ \vdots\\ b_i^{(i)}\\ \vdots\\ b_{n-1}^{(i)}\end{pmatrix}=\begin{pmatrix}0\\ \vdots\\ 1\\ \vdots\\ 0\end{pmatrix} \tag{9}$$

from which $b_s^{(i)}$ is to be determined, $s=0, 1, \ldots, n-1$. In calculating $b_k^{(i)}$ Eqs. (9) are successively multiplied $v_j^{(n)} Q_k(x_j^{(n)})$, $j=1, 2, \ldots, n$ and equations thus obtained, are added. We get

$$\sum_{l=0}^{n-1} b_l^{(i)} \sum_{j=1}^{n} v_j^{(n)} Q_l(x_j^{(n)}) Q_k(x_j^{(n)}) = v_i^{(n)} Q_k(x_i^{(n)})$$

which is, under the presupposition (6) simplified to

$$b_k^{(i)} = \frac{v_i^{(n)} Q_k(x_i^{(n)})}{C_n}.$$

By substituting this into (8) we obtain

$$l_i(x) = \frac{v_i^{(n)}}{C_n} \sum_{j=0}^{n-1} Q_j(x_i^{(n)}) Q_j(x)$$

thus getting from (3)

$$A_i^{(n)} = \frac{v_i^{(n)}}{C_n} S_{n-1}(x_i^{(n)}) \tag{10}$$

where

$$S_{n-1}(x_i^{(n)}) = \sum_{j=0}^{n-1} \int_a^b w(x) Q_j(x)\, dx\, Q_j(x_i^{(n)}) \tag{11}$$

is the partial sum of the development of function $k(x)$ into the Fourier series (further FS) by means of OP $\{Q_n(x)\}$. Since the presuppositions of the theorem guarantee the convergence of this series in each inner point $x_i^{(n)}$ of $[A, B]$ to the $k(x_i^{(n)})$ then

$$S_{n-1}(x_i^{(n)}) = k(x_i^{(n)}) - \sum_{j=n}^{\infty} c_j Q_j(x_i^{(n)}).$$

We show which must be the constant C_n in (6). Because of orthonormality $Q_0(x) = W^{-\frac{1}{2}}$ where $W = \int_a^b W(x)\, dx$ and thus from (6) we get for $l=k=0$ $C_n = W^{-1} \sum_{j=1}^{n} v_j^{(n)}$. If C_n^{-1} for $n \to \infty$ is bounded then from (10) we get (7).

Theorem 2. Let the knots $\{x_i^{(n)}\}$ in (1) be the roots $\{Q_n(x)\}$ i.e. OP with WF $W(x)$ on $[a, b]$. If function $k(x)$ is integrable with $w(x)$ and satisfies uniformly DLC in every point $x \in [A, B] \subset [a, b]$ then for every n and for every i such that $x_i^{(n)}$ is the inner point of $[A, B]$ it holds

$$A_i^{(n)} = B_i^{(n)} k(x_i^{(n)}) + \xi_i^{(n)} \tag{12}$$

where $\lim_{n \to \infty} \xi_i^{(n)} = 0$.

Proof. Consider OP $Q_n(x)$, $n=0, 1, \ldots$ in the form

$$Q_n(x) = \sum_{i=0}^{n} a_{n-i}^{(n)} x^{n-i} = a_n^{(n)} \omega_n(x). \tag{13}$$

Of system $\{Q_n(x)\}$ the Christoffel-Darboux identity holds, i.e.

$$(x-y)\sum_{j=0}^{n} Q_j(x)\,Q_j(y) = \frac{a_n^{(n)}}{a_{n+1}^{(n+1)}}\left(Q_{n+1}(x)\,Q_n(y) - Q_n(x)\,Q_{n+1}(y)\right).$$

If substituting in it the root $Q_n(x)$ i.e. $y = x_i^{(n)}$ then, after multiplication with $w(x)/Q_n'(x_i^{(n)})$, after integration and adjustment we obtain for (3)

$$A_i^{(n)} = \frac{-a_{n+1}^{(n+1)}}{a_n^{(n)}\,Q_n'(x_i^{(n)})\,Q_{n+1}(x_i^{(n)})}\sum_{j=0}^{n-1}\int_a^b w(x)\,Q_j(x)\,dx\,Q_j(x_i^{(n)}) \tag{14}$$

and thus considering (4) and the partial sum (11) of the uniformly convergent FS it holds (12).

Theorem 3. Let (7) or (12) be valid i.e. Theorem 1 or 2. For $n > n_1$ let $x_i^{(n)} \in I_i \subset (a, b)$ and $|w(x_i^{(n)})| \geqq |w_i|$ where sign $w(x_i^{(n)}) =$ sign w_i. If the FS of $k(x)$ by means of $\{Q_n(x)\}$ converges in I_i uniformly then from a certain n it holds

$$\text{sign } A_i^{(n)} = \text{sign } w(x_i^{(n)}). \tag{15}$$

Proof. From (7) or (12) we have

$$A_i^{(n)} = C_i^{(n)}\left(k(x_i^{(n)}) - \sum_{j=n}^{\infty} c_j Q_j(x_i^{(n)})\right) \tag{16}$$

where

$$C_i^{(n)} = \frac{v_i^{(n)}}{C_n} \quad \text{or} \quad B_i^{(n)}.$$

Let $w(x_i^{(n)}) \geqq w_i > 0$ for $n > n_1$. If on I_i is $W(x_i^{(n)}) \leqq M_i$ then there exists such $\varepsilon > 0$ that

$$w_i > M_i\,\varepsilon. \tag{17}$$

It follows from the uniform convergence of the FS of the function $k(x)$ that to a given $\varepsilon > 0$ $\left|\sum_{j=n}^{\infty} c_j Q_j(x)\right| < \varepsilon$ holds for all $x \in I_i$, $n > n_2$ and if $n > \max(n_1, n_2)$ then from (17) we get

$$w(x_i^{(n)}) > W(x_i^{(n)}) \sum_{j=n}^{\infty} c_j Q_j(x_i^{(n)}).$$

Since $C_i^{(n)}$ and $W(x_i^{(n)})$ are positive numbers then from (16) we get (15).

The proof for $w_i < 0$ is similar.

The following theorem will have an interesting consequence.

Theorem 4. Let $l_i(x)$ from (3) is given in the form (8) where $\{Q_n(x)\}$ is the system of OP with $W(x)$ on $[a, b]$. If $\{x_i^{(n)}\}$ are roots of $\{Q_n(x)\}$ then for $i = 1, 2, \ldots, n$ will be

$$B_i^{(n)} = \sum_{j=0}^{n-1} (b_j^{(i)})^2. \tag{18}$$

If for $\{x_i^{(n)}\}$ holds also (6) then for $i=1, 2, \ldots, n$

$$B_i^{(n)} = \frac{v_i^{(n)}}{C_n}. \tag{19}$$

Proof. Consider the integral

$$J_i = \int_a^b l_i^2(x)\, W(x)\, dx \tag{20}$$

for $i=1, 2, \ldots, n$. Since $l_i^2(x)$ is a polynomial of the $2n-2$ degree we obtain, by applying (2) $J_i = B_i^{(n)}$. By substituting (8) into (20) then, following rearrangement by utilizing the properties of OP $\{Q_n(x)\}$ we get (18).

If (6) holds then $b_j^{(i)} = \frac{v_i^{(n)} Q_j(x_i^{(n)})}{C_n}$ and thus from (18) we have

$$B_i^{(n)} = \left(\frac{v_i^{(n)}}{C_n}\right)^2 \sum_{j=0}^{n-1} Q_j^2(x_i^{(n)}). \tag{21}$$

Since for $B_i^{(n)}$ in (2) holds $B_i^{(n)} = \left(\sum_{j=0}^{n-1} Q_j^2(x_i^{(n)})\right)^{-1}$ then from (21) we have (19).

If in (6) for $i=1, 2, \ldots, n$ $v_i^{(n)} = c$ then also $B_i^{(n)}$ for every i are equal i.e. $B_i^{(n)} = c\, C_n^{-1} = W/n$. From this ensues

Theorem 5. If $\{x_i^{(n)}\}$ in (2) are roots of OP $\{Q_n(x)\}$ then (6) with $v_i^{(n)} = c$, $i=1, 2, \ldots, n$ holds only if $\{Q_n(x)\}$ is a normalized system of Chebyshev polynomials of 1-st kind transformed into $[a, b]$.

Proof. It is proved [6] (Posse, Sonin) that on $[-1, 1]$ there exists no WF different from $1/\sqrt{1-x^2}$ so that formula (2) would have all coefficients equal. This proof can be easily generalized for $[a, b]$ i.e. on $[a, b]$ there exists no WF different from $1/\sqrt{(x-a)(b-x)}$ $\left(\text{normalized system of } \left\{T_n\left(\frac{2x-b-a}{b-a}\right)\right\}\right)$ so that all $B_i^{(n)}$ would be equal.

Suppose the opposite of what the theorem asserts i.e. let there exists a system of OP $\{Q_n(x)\}$ on $[a, b]$ with WF $W(x) \neq 1/\sqrt{(x-a)(b-x)}$ in which, if $\{x_i^{(n)}\}$ are roots of $\{Q_n(x)\}$, holds (6) with equal $v_i^{(n)}$. On the basis of Theorem 4 (19) is valid, which contradicts the assertion already propounded.

Example 1. Let in (1) for $a=-1$, $b=1$ be given WF $w(x)$ whereby

a) $W(x) = 1/\sqrt{1-x^2}$ i.e. $\{Q_n(x)\}$ is a normalized system of Chebyshev polynomials of 1-st kind $\left(T_n(x) = \cos(n \arccos x)\right)$ with $x_i^{(n)} = \cos\varphi_i$, $\varphi_i = \frac{2i-1}{2n}\pi$, $i=1, 2, \ldots, n$,

b) $W(x) = \sqrt{1-x^2}$ i.e. $\{Q_n(x)\}$ is a normalized system of Chebyshev polynomials of 2-nd kind $\left(U_n(x) = \frac{\sin[(n+1)\arccos x]}{\sqrt{1-x^2}}\right)$ with $x_i^{(n)} = \cos\psi_i$, $\psi_i = \frac{i\pi}{n+1}$, $i=1, 2, \ldots, n$,

c) $W(x)=\sqrt{\frac{1-x}{1+x}}$ i.e. $Q_n(x)=\frac{(2n)!}{2^{2n}(n!)^2}\,\frac{T_{n+1}(x)-T_n(x)}{x-1}$ with $x_i^{(n)}=\cos 2\eta_i$, $\eta_i=\frac{i\pi}{2n+1}$, $i=1, 2, \ldots, n$.

Since in case a) is $B_i^{(n)}=\frac{\pi}{n}$ in b) $B_i^{(n)}=\frac{\pi}{n+1}\sin^2\psi_i$ and in c) $B_i^{(n)}=\frac{4\pi}{2n+1}\sin^2\eta_i$ then if the respective FS converges in $[-1, 1]$ uniformly we get for a)

$$A_i^{(n)}=\frac{\pi}{n}\,w(x_i^{(n)})\sqrt{1-(x_i^{(n)})^2}+\bar{\varepsilon}_i^{(n)}, \tag{22}$$

for b)

$$A_i^{(n)}=\frac{\pi}{n+1}\,w(x_i^{(n)})\sqrt{1-(x_i^{(n)})^2}+\bar{\bar{\varepsilon}}_i^{(n)}, \tag{23}$$

and for c)

$$A_i^{(n)}=\frac{\pi}{n+\frac{1}{2}}\,w(x_i^{(n)})\sqrt{1-(x_i^{(n)})^2}+\tilde{\varepsilon}_i^{(n)} \quad \text{since} \quad \sin^2\eta_i=\frac{1-x_i^{(n)}}{2}.$$

No attention has been devoted to the question pertaining to the existence of knots $\{x_i^{(n)}\}$ and OP $\{Q_n(x)\}$ for which $\frac{n}{2}(n+1)$ conditions (6) would be satisfied. We show two cases for which system $\{Q_n(x)\}$ is orthogonal on the set $\{x_i^{(n)}\}$ with weights $\{v_i^{(n)}\}$, i.e. for which (6) holds.

Example 2. Let $Q_0(x)=\frac{1}{\sqrt{\pi}}T_0(x)$ and $Q_n(x)=\sqrt{\frac{2}{\pi}}\,T_n(x)$, $n\geqq 1$. System $\{Q_n(x)\}$ is orthogonal:

a) on the set $\{x_i^{(n)}=\cos\varphi_i\}$ with $v_i^{(n)}=1$, $i=1, 2, \ldots, n$ and thus $C_n=n/\pi$, i.e. from (7) we get again (22).

b) on the set $\left\{x_i^{(n)}=\cos\frac{(i-1)\pi}{n-1}\right\}$ with $v_1^{(n)}=v_n^{(n)}=\frac{1}{2}$, $v_i^{(n)}=1$, $i=2, 3, \ldots, n-1$ and thus $C_n=\frac{n-1}{\pi}$.

Remarks. 1. With an concrete system of OP $\{Q_n(x)\}$ the conditions of the uniform convergence of the function's development into the FS can be formulated in a more concrete way. For the sake of generality the Theorem 1.3.6 from [1] has been adhered to.

2. From (10) and (14) it holds for every n $\operatorname{sign}A_i^{(n)}=\operatorname{sign}S_{n-1}(x_i^{(n)})$ where $S_{n-1}(x_i^{(n)})$ is in (11).

3. If in (1) coefficients $A_i^{(n)}$ are applied in their asymptotic expressions (12), then the "asymptotic formula"

$$\int_a^b f(x)\,w(x)\,dx=\sum_{i=1}^{n}B_i^{(n)}\,k(x_i^{(n)})\,f(x_i^{(n)})+\bar{R}_n\left(\frac{fw}{W}\right)$$

is obtained, where

$$\bar{R}_n\left(\frac{fw}{W}\right)=R_n(f)-\sum_{i=1}^{n}\sum_{j=n+1}^{\infty}B_i^{(n)}\,f(x_i^{(n)})\,Q_j(x_i^{(n)})\int_a^b w(x)\,Q_j(x)\,dx$$

being in fact formula (2) for the function $f(x)\,k(x)$, i.e. if $w(x)=W(x)$ then $\overline{R}_n(f)=R_n(f)$.

We see that in Example 2a) for integral $I=\int_{-1}^{1} f(x)\,w(x)\,dx=\int_0^{\pi} F(\varphi)\,d\varphi$ where $F(\varphi)=f(\cos\varphi)\,w(\cos\varphi)\sin\varphi$ we get $I\approx\frac{\pi}{n}\sum_{i=1}^{n} F(\varphi_i)$ where $\varphi_i=\frac{2i-1}{2n}\pi$ and in 2b) $I\approx\frac{\pi}{n-1}\sum_{i=1}^{n}{}'' F(\varphi_i)$ where $\varphi_i=\frac{(i-1)\pi}{n-1}$ whereby $\sum''$ means that the first and the last term of this sum is divided by 2. Thus in the first case this quadrature is on $[0,\pi]$ an asymptotically rectangular and in the second case a trapezoidal rule.

4. Relation (23) can also be obtained thus: Let (3) be

$$A_i^{(n)}=\int_0^{\pi} H(\psi)\,w(\cos\psi)\,d\psi \quad \text{where} \quad H(\psi)=l_i(\cos\psi)\sin\psi$$

is the sine polynomials of the n-th order [3], i.e. $H(\psi)=\sum_{j=1}^{n} b_j\sin j\psi$ whereby since

$$l_i(\cos\psi_j)=\begin{cases}0 & i\neq j\\ 1 & i=j\end{cases} \quad \text{will be} \quad b_j=\frac{2}{n+1}\sum_{k=1}^{n} H(\psi_k)\sin j\psi_k=\frac{2}{n+1}\sin\psi_i\sin j\psi_i.$$

$A_i^{(n)}$ will thus be

$$A_i^{(n)}=\frac{\pi}{n+1}\sin^2\psi_i\sum_{j=1}^{n}\frac{2}{\pi}\int_{-1}^{1} w(x)\,U_{j-1}(x)\,dx\,U_{j-1}(x_i^{(n)}).$$

Supposing that the FS by means of $\{U_n(x)\}$ of the function $w(x)/\sqrt{1-x^2}$ converges in point $x_i^{(n)}$ uniformly, then (23) holds.

5. In paper [2] is formulated the so called circle theorem: Let

$$W(x)=(1-x)^{\alpha}(1+x)^{\beta}, \quad \alpha,\beta>-1 \quad \text{and} \quad x_i^{(n)} \text{ and } B_i^{(n)}$$

are knots and coefficients of (2) then

$$B_i^{(n)}\sim\frac{\pi}{n}W(x_i^{(n)})\sqrt{1-(x_i^{(n)})^2} \quad \text{as} \quad n\to\infty$$

where the index $i=i(n)$ has been selected in such a way that the points $x_i^{(n)}$ remain in a fixed interval lying in the interior of $[a,b]$ $\left(u_n\sim v_n \text{ means that } \lim_{n\to\infty}\frac{u_n}{v_n}=1\right)$.

It follows from proved theorems that if FS of function $k(x)$ converges uniformly on fixed interval placed within $[a,b]$ then coefficients (3) from Example 1 a) to c) satisfy also the circle theorem.

It can be seen from (7) and (12) that under given assumptions limit curves other than circle can be obtained, either.

6. Theorem 1 does not demand $\{x_i^{(n)}\}$ to be the roots of $\{Q_n(x)\}$. Theorem 4 and 5 state that if this is presumed then only quadrature of Example 1 a) satisfies Theorem 1 and (19) whereby (2) is found to be the already mentioned Mehler's formula.

It is obvious that if integrals in (11) can be accurately calculated, precise formulae can be obtained for $A_i^{(n)}$. Here is an example of such procedure.

Example 3. Let on $[-1, 1]$ be

$$w(x) = \frac{q_k(x)}{\sqrt{1-x^2}} \quad \text{where} \quad q_k(x) = \begin{cases} \cos \pi k(x+1) \\ \sin \pi k(x+1), \end{cases} \quad k > 0,$$

integer and $W(x) = 1/\sqrt{1-x^2}$, i.e. $\{x_i^{(n)}\}$ are roots $T_n(x)$ as in Example 1 a).

Let e.g. $q_k(x) = \cos \pi k(x+1) = (-1)^k \cos \pi k x$. Coefficients (3) i.e. (10) and (14) respectively have now the form

$$A_i^{(n)} = \frac{(-1)^k \pi}{n} \left[\frac{1}{\pi} \int_{-1}^{1} \frac{\cos \pi k x}{\sqrt{1-x^2}} T_0(x)\, dx\, T_0(x_i^{(n)}) + \right. \\ \left. + \frac{2}{\pi} \sum_{j=1}^{n-1} \int_{-1}^{1} \frac{\cos \pi k x}{\sqrt{1-x^2}} T_j(x)\, dx\, T_j(x_i^{(n)}) \right]. \tag{24}$$

Since for $m > 0$ is (Granštejn, I. S., Ryžik, I. M.: Tablicy integralov, Moscow, 1963, 850)

$$\int_{-1}^{1} \frac{\cos m x}{\sqrt{1-x^2}} T_{2n}(x)\, dx = (-1)^n \pi J_{2n}(m)$$

where $J_{2n}(m)$ are Bessel functions of the first kind, then from (24) the precise formula is obtained

$$A_i^{(n)} = \frac{(-1)^k \pi}{n} \left[J_0(\pi k) + 2 \sum_{j=2(2)}^{N} (-1)^{j/2} J_j(\pi k) \cos j \varphi_i \right] \tag{25}$$

where N is the first even number $\leqq n-1$, $j = 2, 4, 6, \ldots, N$.

With $q_k(x) = \sin \pi k(x+1) = (-1)^k \sin \pi k x$ considering that for $m > 0$ there is

$$\int_{-1}^{1} \frac{\sin m x}{\sqrt{1-x^2}} T_{2n+1}(x)\, dx = (-1)^n \pi J_{2n+1}(m)$$

we obtain the relation

$$A_i^{(n)} = \frac{(-1)^k 2\pi}{n} \sum_{j=1(2)}^{N} (-1)^{\frac{j-1}{2}} J_j(\pi k) \cos j \varphi_i \tag{26}$$

where N is the first odd number $\leqq n-1$, $j = 1, 3, 5, \ldots, N$.

The sums in (25) and (26) are the known partial sums of the FS of functions $\cos \pi k x$, $\sin \pi k x$ respectively, from which it ensues that $A_i^{(n)} \approx \frac{\pi}{n} q_k(x_i^{(n)})$ which is in accord with (22).

The coefficients calculated from (25) and (26), respectively, can be applied in calculating the cosine or sine Fourier coefficients for $[0, 2\pi]$, i.e. in the quadrature

$$\int_{-1}^{1} f(x)\, q_k(x)\, dx \approx \sum_{i=1}^{n} A_i^{(n)} f(x_i^{(n)}) \sqrt{1-(x_i^{(n)})^2}$$

especially with a smaller k, when the method of the multiple use of one formula described in [7] can not be utilized. The increase of n is practically unlimited in this case.

Theorem 6. If $w(x)$ and $W(x)$ will be from Example 3 then for $A_i^{(n)}$ from (25) or (26) respectively, it holds for every i

$$|A_i^{(n)}| < \frac{2\pi}{n} \exp(p_k) \tag{27}$$

where $p_k = \pi k/2$.

Proof. Since for $[-1, 1]$ is $|T_j(x_i^{(n)})| \leqq 1$ we have in both cases

$$|A_i^{(n)}| \leqq \frac{2\pi}{n} \sum_{j=0}^{n-1} |J_j(\pi k)|.$$

Since (Bateman, H., Erdelyi, A.: Higher transcendental functions, Vol. 2, New York, 1953, 13)

$$|J_m(z)| \leqq \left|\frac{z}{2}\right|^m \frac{\exp(|v|)}{\Gamma(m+1)}$$

for $m > -\frac{1}{2}$, $z = u + iv$ (m, u, v real) and thus $|J_m(\pi k)| \leqq p_k^m/m!$ then

$$|A_i^{(n)}| < \frac{2\pi}{n} \sum_{j=0}^{\infty} \frac{p_k^j}{j!} = \frac{2\pi}{n} \exp(p_k).$$

Corollary. 1. For every n will be $\sum_{i=1}^{n} |A_i^{(n)}| < 2\pi \exp(p_k) = K$, i.e. this quadrature converges for every $f(x) \in C[-1, 1]$.

2. Since at the given i it holds for each member of the sequence $A_i^{(i)}, A_i^{(i+1)}, A_i^{(i+2)}, \ldots$. (27) i.e. $|A_i^{(n)}| < K/n$ then to each $\varepsilon > 0$ there exists such integer n_0 that for all integer $n \geqq n_0 > K/\varepsilon$ there will be $|A_i^{(n)}| < \varepsilon$ and thus $\lim_{n\to\infty} A_i^{(n)} = 0$.

3. Estimate now the inaccuracy of $A_i^{(n)}$ if instead of the exact formula its asymptotic expression is used i.e. $A_i^{(n)} \approx \frac{\pi}{n} q_k(x_i^{(n)})$.

Let $q_k(x) = \cos \pi k (x+1)$ (for $\sin \pi k (x+1)$ we can proceed analogically). From (25) we get

$$A_i^{(n)} = \frac{\pi}{n} \cos \pi k (x_i^{(n)} + 1) - \tilde{R}_n(\varphi_i)$$

where

$$\widetilde{R}_n(\varphi_i) = \frac{2\pi}{n} \sum_{j=N+2(2)}^{\infty} (-1)^{\frac{j}{2}+k} J_j(\pi k) \cos j\varphi_i.$$

We estimate $\widetilde{R}_n(\varphi_i)$ for every i. Since

$$|\widetilde{R}_n(\varphi_i)| \leqq \frac{2\pi}{n} \sum_{j=N+2(2)}^{\infty} |J_j(\pi k)| \leqq \frac{2\pi}{n} \sum_{j=N+2(2)}^{\infty} \frac{p_k^j}{j!}$$

we have

$$|\widetilde{R}_n(\varphi_i)| < \frac{2\pi p_k^{N+2}}{(N+1)(N+2)!} \sum_{j=0(2)}^{\infty} \frac{p_k^j}{j!} = Q_n$$

where

$$Q_n = \frac{p_k^{N+1} \pi^2 k \cosh p_k}{(N+1)(N+2)!}$$

and thus for $|q_i^{(n)}| \leqq 1$ holds

$$A_i^{(n)} = \frac{\pi}{n} \cos \pi k (x_i^{(n)} + 1) + q_i^{(n)} Q_n.$$

3. Convergence of (1) for Continuous Functions

We formulate now some sufficient conditions to have $\sum_{i=1}^{n} |A_i^{(n)}| \leqq K$ for $n = 1, 2, \ldots$, i.e. that quadrature (1) converge for $f(x) \in C[a, b]$.

Theorem 7. Let knots $\{x_i^{(n)}\}$ in (1) are roots $\{Q_n(x)\}$, i.e. OP with WF $W(x)$ on $[a, b]$. Let the function $k(x)$ is on (a, b) bounded. If the FS of this function by means of $\{Q_n(x)\}$ converges in (a, b) uniformly then the quadrature (1) converges for every $f(x) \in C[a, b]$.

Proof. We find to an arbitrary fixed $\varepsilon > 0$ such n_0 that for $n > n_0$

$$\left| \sum_{j=n+1}^{\infty} c_j Q_j(x_i^{(n)}) \right| < \varepsilon$$

holds. Thus from (16) for $C_i^{(n)} = B_i^{(n)}$ follows the estimate

$$|A_i^{(n)}| \leqq B_i^{(n)} (M + \varepsilon) \tag{28}$$

where $|k(x)| \leqq M$ on (a, b). By adding of inequalities (28) for $i = 1, 2, \ldots, n$ we obtain

$$\sum_{i=1}^{n} |A_i^{(n)}| \leqq W(M + \varepsilon) = K_0$$

where $W = \int_a^b W(x)\, dx$. Since for $r = 1, 2, \ldots$

$$\sum_{i=1}^{r} |A_i^{(r)}| \leqq K = \max\left(\max_{m=1(1)n_0} \sum_{i=1}^{m} |A_i^{(m)}|, K_0 \right)$$

then (1) converges for $f(x) \in C[a, b]$.

A similar theorem could be proved under similar conditions even if $\{x_i^{(n)}\}$ satisfy (6) because then $C_i^{(n)} = v_i^{(n)} W \left(\sum_{j=1}^{n} v_i^{(n)} \right)^{-1}$ and thus $\sum_{i=1}^{n} C_i^{(n)} = W$.

In proving the following theorem we use the idea inherent in the proof of theorem about the estimation of system OP [5].

Theorem 8. Let functions $\{Q_n(x)\}$ and $W(x)$ are on $[a, b]$ uniformly bounded whereby knots $\{x_i^{(n)}\}$ in (1) are roots $\{Q_n(x)\}$. If function $k(x)$ is on (a, b) bounded and $g(x) = k^{-1}(x)$ satisfies on $[a, b]$ Lipschitz condition then quadrature (1) converges for $f(x) \in C[a, b]$.

Proof. Consider $A_i^{(n)}$ in the form (14) which, since $\{x_i^{(n)}\}$ are roots $\{Q_n(x)\}$, will be

$$A_i^{(n)} = B_i^{(n)} \int_a^b \sum_{j=0}^{n} Q_j(x)\, Q_j(x_i^{(n)})\, w(x)\, dx. \tag{29}$$

If realizing that

$$w(x) = w(x) \frac{g(x_i^{(n)}) - g(x)}{g(x_i^{(n)})} + \frac{W(x)\, w(x_i^{(n)})}{W(x_i^{(n)})}$$

then from (29) we get two integrals. If substituting in the first Christoffel-Darboux identity $\left(Q_n(x_i^{(n)}) = 0\right)$ then

$$A_i^{(n)} = \frac{a_n^{(n)}}{a_{n+1}^{(n+1)}} B_i^{(n)} k(x_i^{(n)}) (I_1 + I_2) \tag{30}$$

where

$$I_1 = \int_a^b Q_n(x)\, Q_{n+1}(x_i^{(n)}) \frac{g(x) - g(x_i^{(n)})}{x - x_i^{(n)}}\, w(x)\, dx$$

and

$$I_2 = \sum_{j=0}^{n} Q_j(x_i^{(n)}) \int_a^b Q_j(x)\, W(x)\, dx.$$

Estimate $A_i^{(n)}$ from (30).

Since

$$1 = \int_a^b Q_{n+1}^2(x)\, W(x)\, dx = a_{n+1}^{(n+1)} \int_a^b x^{n+1} Q_{n+1}(x)\, W(x)\, dx$$

then from orthonormality $\{Q_n(x)\}$ follows

$$\int_a^b x Q_n(x)\, Q_{n+1}(x)\, W(x)\, dx = a_n^{(n)} \int_a^b x^{n+1} Q_{n+1}(x)\, W(x)\, dx = \frac{a_n^{(n)}}{a_{n+1}^{(n+1)}}$$

and thus

$$\begin{aligned} \frac{a_n^{(n)}}{a_{n+1}^{(n+1)}} &\leqq \max(|a|, |b|) \int_a^b |Q_n(x)|\, |Q_{n+1}(x)|\, W(x)\, dx \\ &\leqq \max(|a|, |b|) \sqrt{\int_a^b Q_n^2(x)\, W(x)\, dx} \sqrt{\int_a^b Q_{n+1}^2(x)\, W(x)\, dx} \\ &= \max(|a|, |b|). \end{aligned} \tag{31}$$

Since $g(x)$ satisfies on $[a, b]$ Lipschitz condition i.e. for $x, x_i^{(n)} \in [a, b]$ holds $|g(x) - g(x_i^{(n)})| \leqq \lambda |x - x_i^{(n)}|$ and if $\{Q_n(x)\}$ and $W(x)$ is bounded with M whereby $\int_a^b |w(x)|\, dx = w$ then $|I_1| \leqq \lambda w M^2$. Since

$$\int_a^b Q_j(x)\, W(x)\, dx = \begin{cases} 0 & j \geqq 1 \\ 1 & j = 0 \end{cases} \quad \text{then we have} \quad I_2 = \int_a^b W(x)\, dx = W.$$

Since $|k(x)| \leqq k$ from (30) we get

$$|A_i^{(n)}| \leqq B_i^{(n)} \max(|a|, |b|)\, k(\lambda w M^2 + W).$$

By adding of these inequalities for $i = 1, 2, \ldots, n$ we get for every n

$$\sum_{i=1}^{n} |A_i^{(n)}| \leqq K = \max(|a|, |b|)\, k\, W(\lambda w M^2 + W)$$

i.e. quadrature (1) converges for $f(x) \in C[a, b]$.

4. Convergence for Functions with Singularity

Now the theorem is expounded on the convergence of the general numerical quadrature with WF $p(x)$ i.e.

$$\int_{-1}^{1} f(x)\, p(x)\, dx \approx \sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)}) = G_n(f) \tag{32}$$

for functions $f(x)$ with singularity in point 1. Its presuppositions and its proof differ from theorem proved in [8] in so far that no nonnegativity will be presumed of $p(x)$ on $[-1, 1]$ and $C_i^{(n)}$.

Theorem 9. Let be given on $[-1, 1)$ the continuous functions $f(x)$ and $p(x)$ such that at a certain $\delta_1 > 0$ in $[1 - \delta_1, 1)$ the function $f(x)\, p(x)$ will be nonnegative and nondecreasing, whereby $\lim_{t \to 1^-} \int_{-1}^{t} f(x)\, p(x)\, dx = I(fp) < \infty$. Let (32) converges for every $F(x) \in C[-1, 1]$, i.e. $\lim_{n \to \infty} G_n(F) = I(Fp)$ where $-1 \leqq x_n^{(n)} < x_{n-1}^{(n)} < \cdots < x_1^{(n)} < x_0^{(n)} = 1$. If from a certain n_1 for every i such that $1 - x_i^{(n_1)} < \delta_2$ with a fixed $\delta_2 > 0$ it holds that

$$\frac{C_i^{(n_1)}}{p(x_i^{(n_1)})} \leqq K(x_{i-1}^{(n_1)} - x_i^{(n_1)}) \tag{33}$$

and if from certain n_2 there exists such $\delta_3 > 0$ that for every i such that $1 - x_i^{(n_2)} < \delta_3$ is

$$\operatorname{sign} C_i^{(n_2)} = \operatorname{sign} p(x_i^{(n_2)}) \tag{34}$$

then $\lim_{n \to \infty} G_n(f) = I(fp)$, i.e. quadrature (32) converges.

Proof. To each $\varepsilon>0$ there exists such $\delta_4>0$ that $\int_{1-\delta_4}^{1} f(x)\,p(x)\,dx \leqq \varepsilon$. Let $n=\max(n_1, n_2)$ and $\delta=\min(\delta_1, \delta_2, \delta_3, \delta_4)$. Define the function

$$g(x)=\begin{cases} f(x) & x\in[-1, 1-\delta] \\ f(1-\delta) & x\in[1-\delta, 1]. \end{cases}$$

Since for the sake of (34) there is

$$\sum_{i=1}^{n} C_i^{(n)} f(x_i^{(n)}) \geqq \sum_{i=1}^{n} C_i^{(n)} g(x_i^{(n)}),$$

i.e. $G_n(f)\geqq G_n(g)$ then, since $g(x)\in C[-1, 1]$ will be

$$\liminf_{n\to\infty} G_n(f) \geqq \liminf_{n\to\infty} G_n(g) = \lim_{n\to\infty} G_n(g) = I(gp) \geqq I(fp)-\varepsilon.$$

Consider further that

$$G_n(f)=S_1+S_2 \tag{35}$$

where

$$S_1=\sum_{x_i^{(n)}\in[-1,1-\delta]} C_i^{(n)} f(x_i^{(n)}) \quad \text{and} \quad S_2=\sum_{x_i^{(n)}\in(1-\delta,1)} C_i^{(n)} f(x_i^{(n)}).$$

Since for all i for which $1-x_i^{(n)}<\delta$ will be $C_i^{(n)}$ and $g(x_i^{(n)})$ of equal signs then

$$S_1=\sum_{x_i^{(n)}\in[-1,1-\delta]} C_i^{(n)} g(x_i^{(n)}) \leqq G_n(g).$$

For the sake of (33) we have

$$S_2=\sum_{x_i^{(n)}\in(1-\delta,1)} \frac{C_i^{(n)}}{p(x_i^{(n)})} f(x_i^{(n)})\,p(x_i^{(n)}) \leqq K \sum_{x_i^{(n)}\in(1-\delta,1)} (x_{i-1}^{(n)}-x_i^{(n)})\,f(x_i^{(n)})\,p(x_i^{(n)})$$

$$\leqq K\int_{1-\delta}^{1} f(x)\,p(x)\,dx \leqq K\varepsilon.$$

Thus we get from (35) $G_n(f)\leqq G_n(g)+K\varepsilon$ and thus

$$\limsup_{n\to\infty} G_n(f) \leqq \limsup_{n\to\infty} G_n(g)+K\varepsilon = \lim_{n\to\infty} G_n(g)+K\varepsilon$$

$$= I(gp)+K\varepsilon \leqq I(fp)+K\varepsilon.$$

From the obtained inequalities we have

$$I(fp)-\varepsilon \leqq \liminf_{n\to\infty} G_n(f) \leqq \limsup_{n\to\infty} G_n(f) \leqq I(fp)+K\varepsilon$$

i.e.

$$\lim_{n\to\infty} G_n(f) = I(fp).$$

Consider quadrature (1) for $a=-1$, $b=1$ i.e.

$$\int_{-1}^{1} f(x)\, w(x)\, dx = \sum_{i=1}^{n} A_i^{(n)} f(x_i^{(n)}) + R_n(f) \tag{36}$$

where $\{x_i^{(n)}\}$ are roots of Chebyshev polynomials $\{T_n(x)\}$. We show conditions for convergence of (36) for $f(x)$ with singularity in point 1. Now we prove theorem about validity of (33).

Theorem 10. Let in quadrature (36) WF $w(x)$ satisfy these assumption at $[-1, 1]$: $w''(x)$ exists, is integrable, $w'(x)$ has bounded variation. Let there exists such δ $(0<\delta\leqq 1)$ that on $[1-\delta, 1]$ is $w(x)\geqq s>0$. If the knots of integration $\{x_i^{(n)}\}$ are the roots of the Chebyshev polynomials $\{T_n(x)\}$, then for every n and for every i such that $x_i^{(n)}\in[1-\delta, 1)$ holds

$$\frac{A_i^{(n)}}{w(x_i^{(n)})} < K(x_{i-1}^{(n)} - x_i^{(n)}) \tag{37}$$

where $x_0^{(n)}=1$.

Proof. It ensues from (22) and (10) and (14) respectively, that for every n

$$|A_i^{(n)}| \leqq \frac{\pi}{n}\left(\sin\varphi_i |w(x_i^{(n)})| + |\tilde{R}(\varphi_i)|\right) \tag{38}$$

where

$$\tilde{R}(\varphi_i) = \frac{2}{\pi}\sum_{m=n+1}^{\infty}\int_0^{\pi} F(\varphi)\cos m\varphi\, d\varphi \cos m\varphi_i \quad \text{and} \quad F(\varphi) = w(\cos\varphi)\sin\varphi.$$

It is proved in [7] that

$$\left|\int_0^{\pi} F(\varphi)\cos m\varphi\, d\varphi\right| \leqq \frac{d}{m^2}$$

where

$$d = |w(-1)| + |w(1)| + \bigvee_0^{\pi}(F'(\varphi)).$$

Since

$$\sum_{m=n+1}^{\infty}\frac{1}{m^{k+1}} < \frac{1}{k\, n^k} \quad \text{we get the estimate} \quad |\tilde{R}(\varphi_i)| < \frac{2d}{\pi n}.$$

For the sake of $\sin\varphi_i < \varphi_i$ we get from (38)

$$|A_i^{(n)}| < \frac{\pi^2(2i-1)\,|w(x_i^{(n)})|}{2n^2} + \frac{2d}{n^2}. \tag{39}$$

Since for $i\leqq\frac{n+1}{2}$ it is $\varphi_i\in(0, \pi/2]$ and thus $\sin\varphi_i\geqq 2\varphi_i/\pi$ we get for $i>1$ that (further see [4])

$$x_{i-1}^{(n)} - x_i^{(n)} = 2\sin\frac{(i-1)\pi}{n}\sin\frac{\pi}{2n} > \frac{4(i-1)}{n^2}. \tag{40}$$

From (39) and (40) we obtain for such i that $x_i^{(n)} \in [1-\delta, 1)$

$$\frac{A_i^{(n)}}{w(x_i^{(n)})} < (x_{i-1}^{(n)} - x_i^{(n)}) \left[\frac{(2i-1)\pi^2}{8(i-1)} + \frac{d}{2(i-1)\, w(x_i^{(n)})} \right]$$

and thus

$$\frac{A_i^{(n)}}{w(x_i^{(n)})} < (x_{i-1}^{(n)} - x_i^{(n)})\, C_1$$

where

$$C_1 = \frac{3\pi^2}{8} + \frac{d}{2s}.$$

For $i=1$ we get the estimate $x_0^{(n)} - x_1^{(n)} > 1/2n^2$ that yields, having been combined with (39)

$$\frac{A_1^{(n)}}{w(x_1^{(n)})} < (x_0^{(n)} - x_1^{(n)})\, C_2$$

where

$$C_2 = \pi^2 + \frac{4d}{s}.$$

Since $C_2 > C_1$ then the assertion (37) for $K = C_2$ is proved.

Now we formulate

Theorem 11. Let Theorem 10 holds. If $f(x)$ and $w(x) = p(x)$ satisfy conditions of Theorem 9 and if from a certain n for every i such that $1 - x_i^{(n)} < \delta$ holds $\operatorname{sign} A_i^{(n)} = \operatorname{sign} w(x_i^{(n)})$ $(\delta > 0)$, then (36) converges also for $f(x)$ with singularity in point 1.

Proof. To prove the validity of Theorem 9 it suffices to prove that (36) converges for continuous functions.

It is known that in order to make the FS of function $f(x)$ by means of $\{T_n(x)\}$ converge uniformly on $[-1, 1]$ it is sufficient that $f(x)$ it be continuous and of bounded variation on this interval. Function $k(x) = w(x)\sqrt{1-x^2}$ from Theorem 10 is continuous. Since on $[-1, 1]$ is $w'(x)$ bounded and thus $w(x)$ satisfies Lipschitz's condition, thus has bounded variation. Since even function $\sqrt{1-x^2}$ is on $[0, 1]$ monotonous then has also bounded variation on $[-1, 1]$ and thus the FS of $w(x)\sqrt{1-x^2}$ converges on this interval uniformly. For the sake of validity of Theorem 7 the quadrature (36) converges for $f(x) \in C[-1, 1]$ (it is proved also in [7]).

The author make grateful acknowledgement Professor A. Talbot from University of Lancaster for helpful suggestions.

References

1. Alexits, G.: Convergence problems of orthogonal series. Akadémiai Kiadó, Budapest, 1961, 35.
2. Davis, P. J., Rabinowitz, P.: Some geometrical theorems for abscissas and weights of Gauss type. J. Math. Anal. Appl. **3**, 428–437 (1961).
3. Fejér, L.: Mechanische Quadraturen mit positiven Coteschen Zahlen. Mathem. Zeitschrift **37**, 287–309 (1933).

4. Gautschi, W.: Numerical quadrature in the presence of a singularity. SIAM J. Numer. Anal. **3**, 357–362 (1967).
5. Korous, J.: O rozvoji funkci jedné reálné proměnné v řadu jistých ortogonálních polynómů. Rozpravy České Akadémie **48**, 12 (1938).
6. Krylov, V. I.: Približennoe vyčislenie integralov. Izdatelstvo Nauka, Moskva, 1967, 195.
7. Mikloško, J.: Numerical integration with Weight Functions $\cos kx$, $\sin kx$ on $[0, 2\pi/t]$. Aplikace matematiky **3**, 179–194 (1969).
8. Rabinowitz, P.: Gaussian integration in the presence of a singularity. SIAM J. Numer. Anal. **4**, 191–201 (1967).

Dr. Jozef Mikloško
Slovak Academy of Sciences
Institut of technical Cybernetics
Bratislava, Dúbravska 1 (ČSSR)

Numer. Math. 15, 250–262 (1970)

Upperbounds for Roots of Polynomials

A. VAN DER SLUIS

Received March 6, 1969

0. Introduction

For several numerical processes it is of importance to have an upper bound for the absolute values of the roots of a polynomial. Several methods for calculating such bounds are available (cf. e.g. [2], ch. 7 and [3], ch. 5), and in particular Gershgorin's method enjoys a certain popularity and is likely to be mentioned whenever the subject of root bounds comes up.

Little seems to have been done, however, in order to get an appraisal of the practical usefulness of the methods and the quality of their results. Indeed, examples in literature may give too favourable impressions in this respect (cf. e.g. [3], p. 127, 130).

Desirable properties of such methods are, of course:

(a) *The amount of computation required should be small.* This amount should not be considerably greater than a few times n arithmetic operations, if n is the degree of the polynomial. In particular a method will have to be discarded when it requires the solution (if only roughly) of an n-th degree polynomial equation (cf. e.g. [2], theorem 27, 3).

(b) *The result should not exceed the maximum of the moduli of the roots by a large factor.*

(c) *The result should be multiplied by γ when all roots are multiplied by $\gamma > 0$.* Of course it is mathematical elegance in the first place that dictates this property, but also from a physical point of view (e.g. in the case that the roots of the polynomial denote frequencies) this scale invariance is a very natural requirement. Moreover, lack of it will usually give rise to additional computation, since one will then be interested in the value of γ for which the result will be as favourable as possible. This is particularly so since for the available methods which lack this property, property (b) is being severely violated for polynomials where the maximum absolute value of the roots is very small or very large, so that a large factor might be gained by multiplying all roots by a proper factor.

In this paper we give a few results in this area. One of the less known methods will turn out rather favourable. The results of greatest practical interest are 2.7 and 5.7.

1. Conventions and Definitions

In this paper we shall indicate definitions by **Def.**, small results by **Th.**, main results by **Theorem.**

Def. 1.1. In the sequel K_c will denote the class of monic complex polynomials of degree n, i.e. the polynomials $z^n + a_1 z^{n-1} + \cdots + a_n$, where $a_1, \ldots, a_n$ are com-

plex numbers. K will denote a subset of K_c. K_r will denote the subset of K_c consisting of real polynomials. K_ϱ will denote the subset of K_r consisting of polynomials of which all roots are real. K_0 will denote the subset of K_c consisting of the polynomials with $a_1=0$.

Def. 1.2. f will always be an element of K, and a_i will always have the meaning it has in 1.1. Also, z_i, $i=1, \ldots, n$, will always denote the roots of f.

Def. 1.3. f^γ, $\gamma>0$, will denote the polynomial defined by $f^\gamma(z)=\gamma^n f(z/\gamma)$; it is, therefore, the polynomial whose roots are γ times as large as those of f.

Def. 1.4. $\mu(f)$ will denote $\max|z_i|$.

Def. 1.5. A *root-bound-functional* (rbf) on K will be a real functional M such that $M(f)\geqq\mu(f)$ for all $f\in K$. A *root-bound* (rb) for f will be any real number m such that $m\geqq\mu(f)$.

Def. 1.6. If $f^\gamma\in K$ for all $f\in K$ and all $\gamma>0$ and M is a rbf on K then M^γ will denote the real functional on K defined by $M^\gamma(f)=M(f^\gamma)/\gamma$ for all $f\in K$ and all $\gamma>0$.

Evidently M^γ is again a rbf on K.

Def. 1.7. A rbf M on K will be called *homogeneous* or *scale-invariant* if M^γ exists for $\gamma>0$ and is independent of γ; i.e. $M(f^\gamma)=\gamma M(f)$ for all $f\in K$ and all $\gamma>0$. (Note that for M to be homogeneous it is necessary that $f\in K\Rightarrow f^\gamma\in K$ for all $\gamma>0$.)

Def. 1.8. A rbf M on K will be called *sharp* if for any $m>\inf M(K)$ there exists an $f\in K$ such that $M(f)=\mu(f)=m$.

Def. 1.9. *Norm and topology on K.* By identifying the polynomial $z^n+a_1z^{n-1}+\cdots+a_n$ with the point $(a_1, \ldots, a_n)$ of the n-dimensional complex cartesian space $\boldsymbol{C}_n$, K_c is identified with $\boldsymbol{C}_n$ and K with a subset of $\boldsymbol{C}_n$. Then any norm on $\boldsymbol{C}_n$ is also a norm on K_c. Any topological property of K and K_c will refer to any of the (equivalent) topologies on $\boldsymbol{C}_n$ induced by norms on $\boldsymbol{C}_n$ (e.g. $\sum|a_i|$).

2. Absolute Rootbounds

A number of rbf's which are well-known from literature are

Q (cf. [1], III 4):

$$Q(f)=\sum_1^n|a_i| \quad \text{if } \geqq 1, \qquad Q(f)=\sqrt[n]{\sum_1^n|a_i|} \quad \text{if } \leqq 1.$$

R (cf. [3], p. 125—126):

$$R(f)=\min\left(R_1(f), R_2(f)\right),$$

where

$$R_1(f)=\max\left(1, \sum_1^n|a_i|\right)$$

$$R_2(f)=\max\left(1+|a_1|, \ldots, 1+|a_{n-1}|, |a_n|\right)$$

(this is the rootbound derived from the Gershgorin circles).

S (cf. [2], § 30, ex. 5):

$$S(f) = 2\max\left(|a_1|, \sqrt{|a_2|}, \ldots, \sqrt[n-1]{|a_{n-1}|}, \sqrt[n]{(|a_n|/2)}\right).$$

T (cf. [2], § 30, ex. 6):

$$T(f) = \max\left(2|a_1|, 2\left|\frac{a_2}{a_1}\right|, \ldots, 2\left|\frac{a_{n-1}}{a_{n-2}}\right|, \left|\frac{a_n}{a_{n-1}}\right|\right).$$

Apparently the first two are non-homogeneous on K_c, S is homogeneous on K_c, T is homogeneous on the set of polynomials with non-vanishing $a_1, \ldots, a_{n-1}$. They all share the property that they depend on the absolute values of the coefficients only.

Def. 2.1. If for any $f = z^n + a_1 z^{n-1} + \cdots + a_n \in K$ also any $f' = z^n + a_1' z^{n-1} + \cdots + a_n'$ with $|a_i'| = |a_i|$ for all i belongs to K, and M is a rbf on K such that $M(f) = M(f')$ for any such pair f, f', then M is called an *absolute* rbf on K.

Theorem 2.2 (cf. [1], III 4 and [4], III 16—21). For any absolute rbf M on K and any $f \in K$, $f \neq z^n$, there exists a sequence of non-negative numbers $\alpha_1, \ldots, \alpha_n$ with $\sum \alpha_i \leqq 1$ such that $M(f) = \max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$, where the prime indicates that elements with $a_i = 0$ are skipped for determination of the maximum. Conversely, for each non-negative sequence $\{\alpha_i\}$ with $\sum \alpha_i \leqq 1$, $\max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$ gives a root-bound (naturally, we define $\max_i{}' = \infty$ if $a_i \neq 0$, $\alpha_i = 0$ for a certain i).

Proof of first part. There exists exactly one positive root z_0 of $f^*(z) = z^n - |a_1| z^{n-1} - \cdots - |a_n|$. Take $\alpha_i = |a_i|/z_0^i$. Then $\sum \alpha_i = 1$ and $\max_i{}' \sqrt[i]{(|a_i|/\alpha_i)} = z_0$. Since M is absolute, $M(f) = M(f^*) \geqq z_0$. If $M(f) > z_0$ we need only lower one of the non-zero α_i in order that $M(f) = \max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$.

Second part. If $|z| > \max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$ then $|a_i|/|z^i| \leqq \alpha_i$ for all i, and inequality occurs at least once, hence $\sum |a_i|/|z^i| < 1$, thus z is no root, since for any root z one has $1 = |\sum a_i/z^i| \leqq \sum |a_i|/|z^i|$. ‖

This theorem shows, that any absolute rbf may be obtained by specifying a suitable dependance of $\alpha_1, \ldots, \alpha_n$ on f. Hence, the functionals Q, R, S, T specified above can all be so obtained, which, of course, can also be verified directly (cf. 4.1, 4.2, 4.3, 4.5).

A natural question to ask is how to characterize $\alpha_1, \ldots, \alpha_n$ for a given polynomial f in such a way that $\max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$ is minimal. The answer to this question follows immediately from the proof of 2.2.

Th. 2.3. If z_0 is the positive root of $z^n - |a_1| z^{n-1} - \cdots - |a_n|$ and $a_i = |a_i|/z_0^i$ then $\max_i{}' \sqrt[i]{(|a_i|/\alpha_i)}$ is minimal and equal to z_0.

Def. 2.4. For any $f \neq z^n$ we denote the corresponding z_0 (cf. 2.3) as $P(f)$. Also we define $P(z^n) = 0$.

Th. 2.5. P is a homogeneous rbf on K_c which for any polynomial gives the best rootbound of all absolute rbf's.

Of course, there is little hope that the optimal α_i can be easily calculated. It is possible, however, to calculate *nearly* optimal α_i, as the following theorem shows.

Th. 2.6. For any f, $S(f) \leqq 2P(f)$ and this inequality is sharp.

Proof. Since S and P are homogeneous, we may assume $P(f)=1$. Then $\sum |a_i| = 1$, hence $|a_i| \leqq 1$. ‖

Therefore, for any polynomial f, $S(f)$ is never more than twice the best rootbound any absolute rbf can deliver for f. Hence, S is nearly optimal.

Now it might be thought that S lacks property (a) specified in section 0, since it seems to require the calculation of about n logarithms and n exponentials. However, in a computer where numbers have a floating point representation $a_i = m_i \cdot 2^{p_i}$, $\frac{1}{2} \leqq |m_i| < 1$, this is not so. In that case we need only calculate $p = 1 + \max [p_i/i]'$, where $[\ ]'$ indicates rounding to above; then 2^p is a very good upperbound for $S(f)$. We combine this in:

Theorem 2.7. S is not only nearly optimal but also a good approximation from above can be easily obtained. Therefore this rbf is to be recommended among all absolute rbf's.

On the other hand, this stresses the point that it is desirable to incorporate standard functions in ALGOL- and FORTRAN-like languages which for any real $a = m \cdot 2^p$, $\frac{1}{2} \leqq m < 1$, give the values of m and p and vice versa (and the application described here is not the only occasion where the need for such functions is felt).

3. Maximum Overestimation Factor

In estimating the performance of a rbf M, be it absolute or not, we shall certainly be interested in the answer to the question: when M specifies $m \neq 0$ as a rootbound for a polynomial f, by which factor at most may this rootbound be in error, i.e. exceed $\mu(f)$. Thus we are interested in the quantity $\mathrm{mof}_M(m)$:

Def. 3.1. $\mathrm{mof}_M(m) = m/\inf\{\mu(f) \mid f \in K,\ M(f) = m\}$, $m \in M(K)$, $m \neq 0$ (if the infimum happens to be 0 we define $\mathrm{mof}_M(m)$ as ∞). This quantity will be called the *maximum overestimation factor* of M at m. The subscript M will be deleted when no confusion with other rbf's is to be feared.

Th. 3.2. If M is a homogeneous rbf on K then $\mathrm{mof}(m)$ is independent of m. In that case we shall just write mof.

Def. 3.3. We shall say that for a polynomial f the mof is attained if $\mathrm{mof}(m) = m/\mu(f)$ and $M(f) = m$.

Thus, the quantity to be determined is $\mu^*_M(m)$:

Def. 3.4. $\mu^*_M(m) = \inf\{\mu(f) \mid f \in K,\ M(f) = m\}$, $m \in M(K)$.
We shall often omit the index M.

Unfortunately this quantity turns out not to be easily calculable in the cases that interest us. A close relationship between the function μ^* and the following function M^* is however to be expected:

Def. 3.5. $M^*(u) = \sup\{M(f) \mid f \in K,\ \mu(f) = u\},\ u \in \mu(K)$.

Indeed, suppose that for any $f \in K$ we plot the point $(\mu(f), M(f))$ on a rectangular coordinate system (Fig. 1); suppose that these points together constitute the shaded area in Fig. 1.

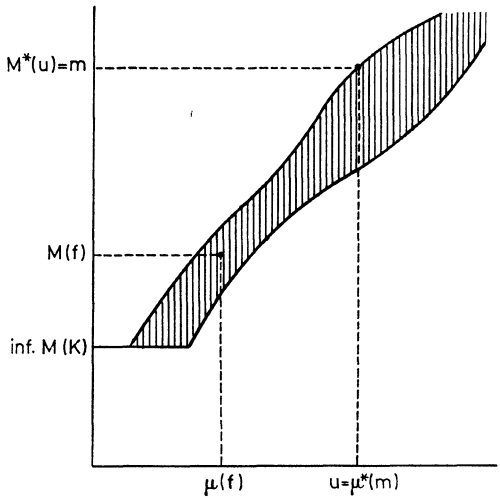

Fig. 1

Then under proper continuity and monotonicity conditions the set of all points $(u, M^*(u))$ and the set of all points $(\mu^*(m), m)$ coincide with the curve that bounds the shaded area from above. The functions M^* and μ^* should then be each other's inverses, and the curve should be the graph of M^*.

This is made explicit in the following theorem. This theorem (which expresses a kind of minimax property) allows obvious generalizations, but the present form suits our purpose.

Theorem 3.6. Let K be a closed subset of K_c (cf. 1.9) not consisting of the polynomial z^n only. Let M be continuous on K. Let $f^\gamma \in K$ for any $\gamma > 0$ and any $f \in K$. Let for any f, $M(f^\gamma)$ be an increasing function of γ for all $\gamma > 0$ for which $M(f^\gamma) > \inf M(K)$. Then

(a) $\mu^*(m) = \min\{\mu(f) \mid f \in K,\ M(f) = m\}$ for $m \in M(K)$.

(b) $M^*(u) = \max\{M(f) \mid f \in K,\ \mu(f) = u\}$ for $u \in \mu(K)$.

(c) $\min M(K)$ exists, $M(K)$ consists of all $x \geqq \min M(K)$ and $\mu(K)$ consists of all $x \geqq 0$.

(d) For all values of M^* and all arguments of μ^* exceeding $\min M(K)$ the functions M^* and μ^* are increasing continuous functions and are each other's inverses.

(e) Whenever $u = \mu^*(m)$, $m > \min M(K)$, the minimum in (a) and the maximum in (b) are attained for the same polynomials f.

(f) $\operatorname{mof}(M^*(u)) = M^*(u)/u$.

Proof. (a), (b) To any number u and any norm $\|\cdot\|$ on $\boldsymbol{C}_n$ there corresponds a number v such that $\mu(f) \leqq u \Rightarrow \|f\| \leqq v$ (for the meaning of $\|f\|$ cf. 1.9). Thus the set $\{f \mid f \in K, \mu(f) = u\}$ is bounded and closed, i.e. it is compact. Since $\mu(f) \leqq M(f)$ the set $\{f \mid f \in K, M(f) = m\}$ is bounded also, and it is closed since K is closed and M is continuous, hence this set is compact also. Existence of the max and min now follows from the continuity of M and μ respectively.

(c) Since for any $f \in K$, $\lim_{\gamma \to 0} f^\gamma = z^n$, z^n must belong to K. Since for $f = z^n$, $M(f^\gamma)$ is constant, it is impossible that $M(z^n) > \inf M(K)$. Hence $M(z^n) = \inf M(K)$. The rest is trivial.

(d) Let u be such that $M^*(u) > \min M(K)$. There exists an $f \in K$ such that $\mu(f) = u$, $M(f) = M^*(u)$. Now let $u' > u$. Then $\mu(f^\gamma) = u'$ for a certain $\gamma > 1$. Then $M^*(u') \geqq M(f^\gamma) > M(f) = M^*(u)$. Hence M^* is increasing.

We now prove that $M^*\big(\mu^*(m)\big) = m$ for all $m > \min M(K)$. We note first that $m \in M(K)$, and thus $\mu^*(m)$ is defined. Let f be such that $M(f) = m$, $\mu(f) = \mu^*(m)$. This shows that $M^*\big(\mu^*(m)\big) \geqq m$ (cf. (b)). Suppose that $M^*\big(\mu^*(m)\big) > m$. Then there would exist $g \in K$ such that $\mu(g) = \mu^*(m)$, $M(g) > m$. Then for some $\gamma < 1$ we have $M(g^\gamma) = m$, and $\mu(g^\gamma) = \gamma\, \mu(g) < \mu^*(m)$, contrary to (a). Hence $M^*\big(\mu^*(m)\big) = m$.

This shows that all $m > \min M(K)$ belong to the range of M^*, and M^* should therefore be continuous. Then μ^* is also increasing and continuous and is the inverse of M^*.

(e) The condition that f realizes the minimum in (a) reads $\mu(f) = \mu^*(m)$, $M(f) = m$. The condition that g realizes the maximum in (b) for $u = \mu^*(m)$ reads $\mu(g) = \mu^*(m)$, $M(g) = M^*\big(\mu^*(m)\big) = m$. Hence the conditions are identical. ‖

Def. 3.7. A rbf M on a set K will be called *normal* if M and K satisfy the requirements of Theorem 3.6.

We then have the following main theorem for absolute rbf's.

Theorem 3.8. Let K be K_c, K_r or K_ϱ.

(a) P is a homogeneous normal absolute rbf on K.

(b) P is sharp; for any $m > 0$, $f(z) = z^n - m z^{n-1}$ satisfies $P(f) = \mu(f) = m$.

(c) For any normal absolute rbf M on K, $\mathrm{mof}_M(m) \geqq \mathrm{mof}_P(m)$ if $m > \min M(K)$.

(d) For any normal absolute rbf M on K such that $M(f)$ depends in a non-decreasing way on each $|a_i|$ the mof is attained for any polynomial of which all roots are equal. This applies in particular to P.

(e) $\mathrm{mof}_P = 1/\big(\sqrt[n]{2} - 1\big) = n/\log(2) - \varepsilon_n$, $0 < \varepsilon_n < 0.5$ for $n = 2, 3, \ldots$ (NB. $1/\log(2) \approx 1.44$).

Proof. (c) $M^*(u) \geqq P^*(u)$ (cf. 2.5), hence $\mu_M^*(m) \leqq \mu_P^*(m)$ (cf. 3.6 (d)).

(d) In view of 3.6 (e) we need only consider the polynomials which realize the maximum in 3.6 (b). For given $\mu(f)$ all $|a_i|$ are maximal if all roots of f are equal.

(e) Because of (d) we need only consider polynomials with equal roots, and since P is homogeneous it is sufficient to consider the case that all roots are 1.

Then $|a_i| = \binom{n}{i}$ and the equation for z_0 (cf. 2.3) becomes $z^n = (1+z)^n - z^n$. Hence $z_0 = 1/(\sqrt[n]{2}-1)$. This proves the first equality sign in (e). The second equality sign follows from the observation that $x^{-1} > [\exp(x)-1]^{-1} > [x/(1-\frac{1}{2}x)]^{-1} = x^{-1} - \frac{1}{2}$. ‖

Remark 3.9. This theorem shows in particular that one cannot hope that any normal absolute rbf can have a smaller mof than approximately $1.4n$ on K_c, K_r or K_ϱ. The condition of normality being so natural, one cannot expect that dropping this condition will be of any help, and indeed it isn't. For, let f be such that mof_P is attained for f and put $m = M(f)$. Then, since $M(f) \geqq P(f)$, $\mathrm{mof}_M(m) \geqq \mathrm{mof}_P$.

Therefore, for mof's lower than $1.4n$ one should turn to non-absolute bounds.

Th. 3.10. Assertion 3.8 (c) holds, more generally, for any K on which M and P are normal, M still assumed to be absolute. With respect to normality of M a remark similar to 3.9 can be made.

4. Properties of Q, R, S, T

In this section we list some further properties of Q, R, S, T. We do this in the form of theorems.

Theorem 4.1. Properties of Q. (a) Q is an inhomogeneous normal absolute rbf on K_c, K_r and K_ϱ.

(b) Q is sharp; for $m \geqq 1$, $f(z) = z^n - mz^{n-1}$ satisfies $Q(f) = \mu(f) = m$; for $m \leqq 1$, $f(z) = z^n - m$ satisfies this relation.

(c) If for any f, γ_f is such that $Q(f^{\gamma_f}) = 1$, then $Q^{\gamma_f}(f) = P(f)$. In particular, if $Q(f) = 1$, then $M(f) \geqq 1$ for any absolute rbf M. Unfortunately, computation of γ_f requires the solution of an n-th degree equation.

(d) $\mathrm{mof}(m) = m/\{\sqrt[n]{(m+1)} - 1\}$ for $m \geqq 1$,

$\mathrm{mof}(m) = m/\{\sqrt[n]{(m^n+1)} - 1\}$ for $m \leqq 1$.

For polynomials of which all roots are equal the mof is attained.

(e) $\mathrm{mof}(m)$ has its least value for $m = 1$. This least value is $\mathrm{mof}_P \approx 1.4n$ (cf. 3.8 (e)). Also, $\mathrm{mof}(m) \sim m/\sqrt[n]{m}$ for $m \to \infty$, $\mathrm{mof}(m) \sim n/m^{n-1}$ for $m \to 0$.

Proof. (a) Apply 2.2 with $\alpha_i = |a_i| / \sum |a_j|$.

(c) Since $\gamma_f P(f) = P(f^{\gamma_f})$, the assertion is equivalent to $P(f^{\gamma_f}) = Q(f^{\gamma_f}) = 1$ which is trivial.

(d) In view of 3.8 (d) we need only consider polynomials of which all roots are equal. Therefore $Q^*(u) = \sum_1^n \binom{n}{i} u^i = (1+u)^n - 1$ if $\geqq 1$ and $Q^*(u) = \sqrt[n]{[(1+u)^n - 1]}$ if $\leqq 1$. If we take $m = Q^*(u)$, the assertion follows from 3.6 (f).

(e) If $m \geqq 1$ then substituting $m = u^n - 1$ in (d) one obtains $\mathrm{mof}(m) = (u^n - 1)/(u-1)$ which is increasing. For $m \leqq 1$, $m = \sqrt[n]{[(1+u)^n - 1]}$ yields $\mathrm{mof}(m) = \sqrt[n]{\{[(1+u)^n - 1]/u^n\}} = \sqrt[n]{\left\{\sum_1^n \binom{n}{i} \Big/ u^i\right\}}$ which is decreasing. ‖

Theorem 4.2. Properties of R. (a) R is an inhomogeneous normal absolute rbf on K_c, K_r and K_ϱ.

(b) R is sharp; for $m>1$, $f(z)=z^n-mz^{n-1}$ satisfies $R(f)=\mu(f)=m$.

(c) $\operatorname{mof}(m)=$

$$m/\max\left[\sqrt[n]{(m+1)}-1,\ \min\left\{\frac{m-1}{\binom{n}{1}},\ \sqrt{\frac{m-1}{\binom{n}{2}}},\ \ldots,\ \sqrt[n-1]{\frac{m-1}{\binom{n}{n-1}}},\ \sqrt[n]{m}\right\}\right].$$

For polynomials of which all roots are equal the mof is attained.

(d) For $m\downarrow 1$, $\lim \operatorname{mof}(m)=1/(\sqrt[n]{2}-1)\approx 1.4n$ (cf. 3.8 (e)).

For $m\lesssim 2.2$, $\operatorname{mof}(m)$ increases to $\approx 1.9n$.

For $2.2\lesssim m\lesssim 3$, $\operatorname{mof}(m)$ decreases to $\approx 1.5n$.

For $m\gtrsim 3$, $\operatorname{mof}(m)$ is increasing and is successively equal to $m\Big/\sqrt{\frac{m-1}{\binom{n}{2}}}$, $m\Big/\sqrt[3]{\frac{m-1}{\binom{n}{3}}}$, ... until for $m\gtrsim n^n$ $\operatorname{mof}(m)=m/\sqrt[n]{m}$.

(e) For any $m>1$, $\operatorname{mof}_R(m)/\operatorname{mof}_Q(m)\leqq 1$.

(f) *Conjecture.* For any $m>1$, $\operatorname{mof}_R(m)/\operatorname{mof}_Q(m)\geqq 0.6$ and for $m>10$ this ratio approaches 1 monotonically.

Proof. (a) R_1 follows from Q. For R_2 apply 2.2 with $\alpha_i=|a_i|/q^i$ with $q=R_2(f)$, for then $\alpha_i\leqq(q-1)/q^i$, $\alpha_n\leqq 1/q^{n-1}$.

(c) With a similar argument as in the proof of 4.1 (d) we have $M^*(u)=$ $\min\left[(1+u)^n-1,\ \max\left\{1+\binom{n}{1}u,\ \ldots,\ 1+\binom{n}{n-1}u^{n-1},\ u^n\right\}\right]$.

(d) For large n the numbers 2.2 and 3 are approximately the values for $M^*(u)$ where $(1+u)^n-1=1+\binom{n}{1}u$ or $1+\binom{n}{1}u=1+\binom{n}{2}u^2$ respectively.

(f) The formulae are so awkward that a rigid proof of this is difficult to obtain. The property can be proved for the values of m where the min in (c) shifts to the next term. Moreover $\operatorname{mof}_R(m)$ and $\operatorname{mof}_Q(m)$ have been calculated for many intermediary values of m, and this has been done for values of n up to 20, and for all these values of m and n the conjecture held. ‖

Theorem 4.3. Further Properties of S. (a) S is a homogeneous normal absolute rbf on K_c, K_r and K_ϱ.

(b) S is sharp; for $m=2$ (this is no restriction because of (a)) $f(z)=z^n-z^{n-1}-\cdots-z-2=(z-2)(z^n-1)/(z-1)$ satisfies $S(f)=\mu(f)=m$.

(c) $\operatorname{mof}=2n$. For polynomials of which all roots are equal the mof is attained.

Proof. (a) Apply 2.2 with $\alpha_i=2^{-i}$ for $i=1,\ldots,n-1$; $\alpha_n=2^{-n+1}$.

(c) Similarly as in the proof of Theorem 4.1 (d) we have $S^*(u)=2n\,u$. ‖

Remark 4.4. From 2.6 and 3.8 (e) one might have expected a somewhat larger value for the mof. However, the polynomials for which the mof is attained are not the same as those for which $S(f)=2P(f)$.

Theorem 4.5. Properties of T. (a) T is a homogeneous absolute rbf which is normal on any closed subset of K where $a_1, a_2, \ldots, a_{n-1}$ are non-zero.

(b) T is sharp; the same f as in 4.3 (b) satisfies $T(f) = \mu(f) = m$.

(c) mof $= \infty$ on K_c; if all roots are equal, $T(f)/\mu(f) = 2n$.

Proof. (a) Apply 2.2 with $\alpha_i = |a_i|/q^i$ with $q = T(f)$. Since

$$q^i \geqq 2|a_1| \cdot 2\left|\frac{a_2}{a_1}\right| \cdot \ldots \cdot 2\left|\frac{a_i}{a_{i-1}}\right| = 2^i|a_i|, \qquad \alpha_i \leqq 2^{-i}. \quad \|$$

Remark 4.6. Initially, one could have hoped that T, though not always defined, might do well just in those cases where S gives a severe overestimation. Because of 2.6 one can hope for no greater improvement than a factor 2. Comparison of 4.3 (c) and 4.5 (c) proves that even this hope is idle.

5. Better Bounds

In this section we consider a few rbf's with far lower mof's. First:

U (Newton): $U(f) = \sqrt{(a_1^2 - 2a_2)}$ on K_ϱ (cf. 1.1).

Theorem 5.1. Properties of U. (a) U is a homogeneous, normal, non-absolute rbf on the set K_ϱ.

(b) T is sharp; for any m, $f(z) = z^n - mz^{n-1}$ satisfies $U(f) = \mu(f) = m$.

(c) mof $= \sqrt{n}$. For polynomials of which all roots are equal the mof is attained.

Proof. (a) Since all z_i are real, $\max z_i^2 \leqq \sum z_i^2 = a_1^2 - 2a_2$.

(c) $\max z_i^2 \geqq \frac{1}{n}\sum z_i^2$. The equality sign holds if all z_i are equal. $\|$

Therefore, on the set K_ϱ as a whole, U performs essentially better than any normal absolute rbf, since even on K_ϱ any such rbf has a mof $\gtrsim 1.4n$ (cf. 3.8 (e)) (note that for an individual f, $P(f)$ may be smaller than $U(f)$, e.g. for $f = x^2 - 1$).

We now ask for a rbf which performs equally well on the entire K_c. Since P (and therefore any absolute bound) is particularly bad if all roots are equal, i.e. if a_1 is large in some sense, we expect a better behaviour if $a_1 = 0$, which can always be obtained by a shift of the variable.

Therefore, we define:

V: $V(f) = \frac{1}{2}(1+\sqrt{5}) \max \sqrt[i]{|a_i|} \approx 1.62 \max \sqrt[i]{|a_i|}$ on K_0 (cf. 1.1).

In order to give some properties of V we need two lemmas (Th. 5.2, 5.3):

Th. 5.2. Let n be any natural number and let u be any positive number.

(a) Then $\max\left\{\left|\sum_1^n u_i^2\right| \,\middle|\, |u_i| \leqq u,\ i = 1, \ldots, n;\ \sum_1^n u_i = 0\right\} = n u^2$ if n is even and $= (n-1)u^2$ if n is odd.

(b) The max is attained for the sets $\{u_i\}$, and for these sets only, where entier $(n/2)$ elements equal ωu, entier $(n/2)$ elements equal $-\omega u$ where ω is any complex number of modulus 1, and moreover, if n is odd, one element is 0.

Proof. Obviously $|\sum u_i^2| \leqq n u^2$ for any set $\{u_i\}$ with $|u_i| \leqq u$, and the $<$sign holds if not all u_i^2 are equal. This settles the problem if n is even.

Now suppose that n is odd. We first note that the max exists. Let $\{u_i\}$ be a set for which the max is attained. It is no restriction to assume that $\sum u_i^2 > 0$. Then $|\operatorname{Re}(u_i)| \leqq u$ and $\sum \operatorname{Re}(u_i) = 0$ and $\sum (\operatorname{Re}(u_i))^2 \geqq \sum \operatorname{Re}(u_i^2) = \sum u_i^2$, and the $>$ sign holds as soon as any u_i is not real. Hence all u_i must be real. We now assert that at most one of the u_i can be different from u and $-u$. For, if e.g. u_1 and u_2 satisfy $-u < u_i < u$ then replacing u_1 and u_2 by $u_1 - \delta$ and $u_2 + \delta$ respectively, where δ is a small real number, does not change $\sum u_i$, and $(u_1 - \delta)^2 + (u_2 + \delta)^2 > u_1^2 + u_2^2$ for some small values of δ. This completes the proof. ‖

Th. 5.3. If $f \in K_0$ and $\mu(f) = u$ then

(a) $\max_i \sqrt[i]{|a_i|} \leqq u \sqrt{((n-1)/2)}$ if n is odd.

(b) $\max_i \sqrt[i]{|a_i|} \leqq u \sqrt{(n/2)}$ if n is even.

(c) for any $n \geqq 2$ the equality sign in (a) or (b) holds if the roots of f form a set as in 5.2 (b).

Proof. (a) Let n be odd. We may assume that $u = \sqrt{(2/(n-1))}$. Define $S_k = \sum_1^n z_i^k$. Then $|S_2| \leqq 2$ (cf. 5.2) and $|S_k| \leqq n u^k$. We recall Newton's identity

$$a_k = -\frac{1}{k}\{S_k + a_1 S_{k-1} + \cdots + a_{k-2} S_2 + a_{k-1} S_1\}, \qquad k = 1, 2, \ldots$$

which in our case implies $a_2 = -S_2/2$, $a_3 = -S_3/3$ and

$$|a_k| \leqq \frac{n}{k}\{u^k + |a_2| u^{k-2} + \cdots + |a_{k-3}| u^3\} + \frac{2}{k}|a_{k-2}| \quad \text{for} \quad k \geqq 4.$$

Now

$$\frac{n}{k} u^k \leqq \frac{2}{k-1} u^{k-2} \leqq \frac{2}{k-1}\left(\frac{2}{k-1}\right)^{(k-2)/2} = \left(\frac{2}{k-1}\right)^{k/2} \quad \text{for} \quad k \geqq 2.$$

Hence

$$|a_2| \leqq 1; \qquad |a_3| \leqq 1;$$

$$|a_k| \leqq \left\{\left(\frac{2}{k-1}\right)^{\frac{k}{2}} + |a_2|\left(\frac{2}{k-1}\right)^{\frac{k-2}{2}} + \cdots + |a_{k-3}|\left(\frac{2}{k-1}\right)^{\frac{3}{2}}\right\} + \frac{2}{k}|a_{k-2}|, \qquad k \geqq 4.$$

This relation for $k = 4, 5, 6$ yields $|a_k| \leqq 1$ for $k = 4, 5, 6$. Now suppose that for a certain $p \geqq 7$ there holds $|a_k| \leqq 1$ for all $k < p$. Then

$$|a_p| \leqq \frac{2}{p-1} \frac{1 - \left(\frac{2}{p-1}\right)^{p-1}}{1 - \sqrt{\left(\frac{2}{p-1}\right)}} \leqq 1.$$

(b) Let n be even, and assume that $u = \sqrt{(2/n)}$. Then the same formulae as in the proof of (a) yield $|a_k| \leqq 1$ for $k \geqq 2$.

(c) If the roots are as specified, $|a_2| = \frac{1}{2}|S_2| = 1$ in parts (a) and (b) of the proof. ‖

We now have the following properties of V:

Th. 5.4. Properties of V. (a) V is a homogeneous normal absolute rbf on K_0.

(b) $\operatorname{mof} = \frac{1}{4}(1+\sqrt{5})\sqrt{(2n-2)}$ if n is odd, $\operatorname{mof} = \frac{1}{4}(1+\sqrt{5})\sqrt{(2n)}$ if n is even. The mof is attained for polynomials whose roots form sets as in 5.2 (b).

Proof. (a) Apply 2.2 with $\alpha_1 = 0$ and $\alpha_i = \alpha^i$ with $\alpha = (-1+\sqrt{5})/2$ for $i \geqq 2$. Then $\alpha_2 + \alpha_3 + \cdots + \alpha_n < \alpha^2/(1-\alpha) = 1$.

(b) Trivial consequence of 5.3. ‖

Remark 5.5. Since $\sum \alpha_i < 1$, a slightly better bound could be obtained by increasing one or more of the α_i slightly. In practice the improvement is negligeable, of course.

As a corollary of 5.4 we have the following interesting result:

Theorem 5.6. For any $f \in K_c$ all roots lie in the circle

$$|z + a_1/n| \leqq \tfrac{1}{2}(1+\sqrt{5}) \max_{i \geqq 2} \sqrt[i]{|a_i^*|}$$

where a_i^* are such that

$$f(z) = (z + a_1/n)^n + a_2^*(z + a_1/n)^{n-2} + a_3^*(z + a_1/n)^{n-3} + \cdots,$$

i.e. $a_i^* = \{f^{(n-i)}(-a_1/n)\}/(n-i)!$

If we define

W: $W(f) = |a_1/n| + \tfrac{1}{2}(1+\sqrt{5}) \max_{i \geqq 2} \sqrt[i]{|a_i^*|}$ (for a_i^* cf. 5.6)

we have

Theorem 5.7. Properties of W. (a) W is a homogeneous normal non-absolute rbf on K_c.

(b) $\mathrm{mof}_W \leqq 1 + \tfrac{1}{4}(1+\sqrt{5})\sqrt{(2n)} \approx 1 + 1.14\sqrt{n}$.

Proof. (b) We need only remember that $|a_1/n| \leqq \mu(f)$. ‖

Consequently, W performs about as well on the set of all polynomials as U does on K_ϱ. Unfortunately W does not quite satisfy requirement (a) in section 0, since calculation of the a_i^* will usually require a number of operations of order n^2. Once the a_i^* are known, $\max \sqrt[i]{|a_i^*|}$ usually can be cheaply calculated as outlined after 2.6.

6. Gershgorin Bounds for Matrices

In this section we consider $n \times n$ matrices.

For the eigenvalues of matrices we can define in an obvious way rbf's and the notions homogeneity and mof. Naturally, $\mu(A)$ will indicate the maximum modulus of the eigenvalues of A.

One rbf for matrices is then

G (cf. [3], p. 52—53):

$$G(A) = \min\left[\max_i \sum_j |a_{ij}|, \max_j \sum_i |a_{ij}|\right] = \min(\|A\|_\infty, \|A\|_1)$$

where $\|A\|_\infty$ and $\|A\|_1$ are the matrix lub-norms, subordinate to the Hölder ∞- and 1-norms, respectively. This follows immediately from the Gershgorin circles, or also from the fact that the modulus of any eigenvalue of A is not greater than any lub norm of A.

Theorem 6.1. (a) G is "homogeneous", i.e. $G(\gamma A) = |\gamma|\, G(A)$ for any complex γ and any A.

(b) G is "sharp"; the identity matrix I satisfies $G(I) = \mu(I) = 1$.

(c) $\mathrm{mof} = \sqrt{n}$ on the class of symmetric matrices;
$\mathrm{mof} = \infty$ on the class of all matrices and even on the class of all non-singular matrices.

Proof. (c) For any matrix A there holds $\|A\|_2 \geqq \|A\|_\infty/\sqrt{n}$. For a symmetric matrix $\mu(A) = \|A\|_2$ and $\|A\|_1 = \|A\|_\infty$, hence $\mu(A) \geqq G(A)/\sqrt{n}$. Therefore $\mathrm{mof} \leqq \sqrt{n}$.

To show that the equality sign occurs we note that for a real orthogonal matrix U the first column can be prescribed arbitrarily provided this column has euclidean norm 1. Choose this column as follows:

$$u_{11} = \sqrt{\{(1+\sqrt{n})(2\sqrt{n})\}}, \qquad u_{i1} = 1/\sqrt{\{2(1+\sqrt{n})\sqrt{n}\}}, \qquad i \geqq 2.$$

Now take

$$A = U \operatorname{diag}(1, -1, -1, \ldots, -1)\, U^T.$$

Then

$$a_{11} = u_{11}^2 - u_{12}^2 - \cdots - u_{1n}^2 = 2u_{11} - 1 = 1/\sqrt{n}$$

$$a_{1j} = u_{11}u_{j1} - u_{12}u_{j2} - \cdots - u_{1n}u_{jn} = 2u_{11}u_{j1} = 1/\sqrt{n}.$$

Hence $\|A\|_\infty \geqq \sum |a_{1j}| = \sqrt{n}$; obviously $\mu(A) = 1$. ‖

The rbf R follows from applying G to the "companion matrix" A_f of a polynomial f: $a_{11} = -a_1, \ldots, a_{1n} = -a_n$; $a_{21} = \cdots = a_{n,n-1} = 1$; all other $a_{ij} = 0$.

It is not surprising that G is homogeneous whereas R is not, since the relation between R and G is not "homogeneous".

Of course other bounds can be derived from G by applying G to BAB^{-1}, B any non-singular matrix. If we take $B = \operatorname{diag}(\beta_1, \ldots, \beta_n)$ and $A = A_f$ we get

Theorem 6.2. For any positive numbers $\beta_1, \ldots, \beta_n$ we define

$$G_1(f) = \max\left\{\frac{\beta_2}{\beta_1}, \frac{\beta_3}{\beta_2}, \ldots, \frac{\beta_n}{\beta_{n-1}}, \beta_1 \sum_1^n |a_i|/\beta_i\right\}$$

$$G_2(f) = \max\left\{\frac{\beta_2 + \beta_1|a_1|}{\beta_1}, \ldots, \frac{\beta_n + \beta_1|a_{n-1}|}{\beta_{n-1}}, \frac{\beta_1|a_n|}{\beta_n}\right\}.$$

Then $\tilde{G}(f) = \min(G_1(f), G_2(f))$ is a rootbound.

Obviously, since $\tilde{G}$ is absolute, no rootbounds can be derived from this theorem that could not already be derived from 2.2. It can be shown that the converse is true also. However, some rootbounds can be derived from 6.2 in a more obvious way than from 2.2. This is clearly the case for R and T; for T just take $\beta_1 = 1$, $\beta_i = |a_{i-1}|$.

For a general matrix A a natural question is to determine a diagonal matrix B such that $\|BAB^{-1}\|_\infty$ or $\|BAB^{-1}\|_1$ are minimal. For this we have the following criterion:

Th. 6.3. For a given matrix A consider $\|BAB^{-1}\|_\infty$ for all non-singular diagonal matrices B. Then $\|BAB^{-1}\|_\infty$ is minimal if all rows of BAB^{-1} have equal 1-norm

(= sum of moduli). Similarly $\|BAB^{-1}\|_1$ is minimal if all columns have equal 1-norm.

Proof of first part. It is sufficient to consider the case that all rows of A have equal 1-norm. Let $B = \mathrm{diag}(\beta_1, \ldots, \beta_n)$ and it will be no restriction to assume that $|\beta_1| = \max |\beta_i|$. Then $(BAB^{-1})_{1j} = a_{1j}\beta_1/\beta_j$. Hence $\|BAB^{-1}\|_\infty \geqq \|A\|_\infty$. The second part is similar. ‖

Of course this minimizing B cannot be found easily. In the case of A_f, $B = \mathrm{diag}(1, \varrho, \varrho^2, \ldots, \varrho^{n-1})$ with $\varrho = 1/z_0$ (cf. 2.3). For a general matrix the *vector* $b = (\beta_1^{-1}, \beta_2^{-1}, \ldots, \beta_n^{-1})^T$ should be the Perron eigenvector of $|A|: |A|\, b = \lambda\, b$, and then $\|BAB^{-1}\|_\infty = \lambda$ is the Perron eigenvalue of $|A|$.

References

1. Dieudonné, J.: La théorie analytique des polynomes d'une variable (a coefficients quelconques). Mém. des Sci. Math. fasc. XCIII. Paris 1938.
2. Marden, M.: The geometry of the zeros of a polynomial in a complex variable. Amer. Math. Soc. Math. Surv. III, New York 1949.
3. Parodi, M.: La localisation des valeurs caractéristiques des matrices et ses applications. Paris: Gauthier-Villars 1959.
4. Polya, G., Szegö, G.: Aufgaben und Lehrsätze aus der mathematischen Analysis, Teil I. Berlin: Springer 1925.

A. van der Sluis
Rekencentrum Rijksuniversiteit
Budapestlaan 6
Utrecht, Netherlands

Numer. Math. 15, 263–282 (1970)
© by Springer-Verlag 1970

Ableitungsfreie Verfahren für nichtlineare Optimierungsprobleme

JOCHEN W. SCHMIDT und KLAUS VETTERS

Herrn Prof. Dr. Dr. h.c. L. Collatz zum 60. Geburtstag am 6. 7. 1970 gewidmet

Eingegangen am 11. Juni 1969

Summary. In the paper the authors develop approximation methods for nonlinear programming problems using function-values only. The idea is to approximate the objective function by a quadratic interpolating polynomial. It is shown how to construct methods of an arbitrarily high convergence rate. By aid of the efficiency index methods are selected from the class in view which require a minimal number of function-values in order to guarantee a given exactness.

1. Einleitung

Es seien $\boldsymbol{K}$ eine konvexe, abgeschlossene Teilmenge des d-dimensionalen Vektorraumes $\boldsymbol{R}=\boldsymbol{R}_d$ und f eine reellwertige Funktion von $\boldsymbol{K}$ in $\boldsymbol{R}_1$. Betrachtet wird die Optimierungsaufgabe

$$f(\boldsymbol{x}) = \text{Extremum!} \quad \text{für} \quad \boldsymbol{x}\in\boldsymbol{K},$$

wobei auch $\boldsymbol{K}=\boldsymbol{R}$ sein kann. Es werden iterative Näherungsverfahren entwickelt und untersucht, welche ausschließlich Funktionswerte verwenden und welche im wesentlichen dadurch entstehen, daß die Funktion f mit Hilfe bereits bekannter Näherungen durch ein geeignetes quadratisches Interpolationspolynom ersetzt wird. Es wird eine einfache Konstruktionsmöglichkeit für Verfahren mit beliebig hoher Konvergenzgeschwindigkeit aufgezeigt, von denen dann diejenigen mit Hilfe des Wirkungsgrades ausgewählt werden, welche bei gleichem Genauigkeitsgewinn mit der geringsten Zahl von Funktionswertberechnungen auskommen.

Nach der Bereitstellung von geeigneten Interpolationsformeln werden zunächst Aufgaben ohne Restriktionen ($\boldsymbol{K}=\boldsymbol{R}$) betrachtet. Einmal kommt diesen Aufgaben eigenes Interesse zu. Zum anderen beruhen verschiedene Verfahren für Aufgaben mit Restriktionen ($\boldsymbol{K}\neq\boldsymbol{R}$) darauf, daß man sie in passender Weise in eine Aufgabe ohne Restriktionen überführt. Hier sei z. B. an das Verfahren von Fiacco und McCormick erinnert (s. etwa [4]). Die Ausführungen über Aufgaben ohne Nebenbedingungen knüpfen an die Arbeiten [9, 12] und Poll [7] an und übertragen die von Traub [13] sowie in [11] für die Gleichungsauflösung entwickelte Methode zur Steigerung der Konvergenzgeschwindigkeit auf Extremwertaufgaben. Außerdem wird auf eine Modifikation eingegangen, bei der die pro Schritt erforderliche Auflösung von linearen Gleichungen genähert nach einer Erweiterung des Schulz-Verfahrens erfolgt, s. [10]. Auf diese Möglichkeit wurde zuerst von Ulm [15] und Helfrich [3] beim Newton-Verfahren hingewiesen.

Bei Aufgaben mit Restriktionen, welche danach behandelt werden, ist die Approximation der Zielfunktion durch eine quadratische Funktion besonders angepaßt, falls $\boldsymbol{K}$ durch lineare Gleichungen oder Ungleichungen beschrieben wird, da für die entstehende Aufgabe der quadratischen Optimierung mehrere einfache Algorithmen vorhanden sind, (z.B. das Verfahren von Wolfe in der vereinfachten Form von Collatz und Wetterling [1]). Erste Untersuchungen in der genannten Richtung stammen von Levitin und Poljak [5] und von Ulm und Poll [16].

2. Quadratische Interpolation und Steigungen

Es wird die Bemerkung vorangestellt, daß es für die Konvergenz der zu entwickelnden Verfahren zweckmäßig ist, die Stützstellen des Interpolationspolynoms in Knoten eines (orthogonalen) Gitters zu legen (s. Fig. 1 und 2).

Der Vektor $\boldsymbol{x}\in\boldsymbol{R}$ habe die Komponenten $(\boldsymbol{x})_i = x_i$ und die Matrix $\boldsymbol{A}$ die Elemente $(\boldsymbol{A})_{ik} = a_{ik}$. Es sollen

$$\boldsymbol{x}\boldsymbol{y} = \sum_{i=1}^{d} x_i y_i, \quad \boldsymbol{A}\boldsymbol{x}\boldsymbol{y} = \sum_{i,k=1}^{d} a_{ik} x_k y_i, \quad \|\boldsymbol{x}\|^2 = \boldsymbol{x}\boldsymbol{x} \quad \text{und} \quad \|\boldsymbol{A}\|^2 = \sum_{i,k=1}^{d} |a_{ik}|^2$$

bedeuten[1]. Falls $\boldsymbol{A}^T$ die Transponierte von $\boldsymbol{A}$ bezeichnet, gilt $\boldsymbol{A}\boldsymbol{x}\boldsymbol{y} = \boldsymbol{A}^T\boldsymbol{y}\boldsymbol{x}$. Die Funktion f sei in einem achsenparallelen „Quader" $\boldsymbol{Q}$ von $\boldsymbol{R}$ erklärt und partiell differenzierbar[2]. Die partiellen Differenzenquotienten von f, auch die mit wiederholtem Argument, werden wie üblich definiert (s. etwa Willers [18]); z.B. seien

$$\delta f(\ldots, u_i v_i, \ldots) = \frac{1}{u_i - v_i}\{f(\ldots, u_i, \ldots) - f(\ldots, v_i, \ldots)\},$$

$$\delta^2 f(\ldots, u_i v_i w_i, \ldots) = \frac{1}{u_i - w_i}\{\delta f(\ldots, u_i v_i, \ldots) - \delta f(\ldots, v_i w_i, \ldots)\}.$$

Außerdem werde die Bezeichnung

$$\delta f(v_1^{i-1}, u_i v_i, u_{i+1}^d) = \delta f(v_1, \ldots, v_{i-1}, u_i v_i, u_{i+1}, \ldots, u_d)$$

und Entsprechendes für die weiteren Differenzenquotienten vereinbart. Nach [8] wird der erste Steigung genannte Vektor $\delta f(\boldsymbol{u}, \boldsymbol{v})$ durch

$$\big(\delta f(\boldsymbol{u}, \boldsymbol{v})\big)_i = \delta f(v_1^{i+1}, u_i v_i, u_{i+1}^d)$$

definiert. Man bestätigt sofort

$$(2.1) \qquad f(\boldsymbol{u}) - f(\boldsymbol{v}) = \delta f(\boldsymbol{u}, \boldsymbol{v})\,(\boldsymbol{u} - \boldsymbol{v}), \qquad \delta f(\boldsymbol{u}, \boldsymbol{u}) = f'(\boldsymbol{u}) \quad \text{für} \quad \boldsymbol{u}, \boldsymbol{v}\in\boldsymbol{Q}.$$

Die zweite Steigung $\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ ist eine Matrix, die entsprechend [14] die Elemente

$$\big(\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\big)_{ik} = \begin{cases} 0 & \text{für} \quad i < k \\ \delta^2 f(w_1^{i-1}, u_i v_i w_i, u_{i+1}^d) & \text{für} \quad i = k \\ \delta^2 f(w_1^{k-1}, v_k w_k, v_{k+1}^{i-1}, u_i v_i, u_{i+1}^d) & \text{für} \quad i > k \end{cases}$$

1 Die Matrizennorm muß, wie es hier der Fall ist, passend zur Vektornorm sein.

2 Bei den zu behandelnden Verfahren sind bestimmte Differenzierbarkeitsforderungen zu stellen, da Differenzenquotienten mit wiederholtem Argument auftreten können. Bei der praktischen Rechnung kommt dieser Fall jedoch kaum vor. Auch bleibt dann die Möglichkeit, zu einem sog. vereinfachten Verfahren überzugehen.

hat. Es gelten

$$\begin{aligned} &\delta f(\boldsymbol{u}, \boldsymbol{v}) - \delta f(\boldsymbol{u}, \boldsymbol{w}) = \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\,(\boldsymbol{v} - \boldsymbol{w}) \\ &\delta f(\boldsymbol{u}, \boldsymbol{w}) - \delta f(\boldsymbol{v}, \boldsymbol{w}) = \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T (\boldsymbol{u} - \boldsymbol{v}) \\ &\delta^2 f(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u})^T = f''(\boldsymbol{u}) \qquad \text{für} \quad \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{Q}. \end{aligned} \tag{2.2}$$

Sind außerdem die dritten partiellen Differenzenquotienten von f auf $\boldsymbol{Q}$ vorhanden und beschränkt, so existiert eine von $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z} \in \boldsymbol{Q}$ unabhängige Zahl a mit[3]

$$\begin{aligned} &\|\delta^2 f(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z}) - \delta^2 f(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z})\| \leqq a \|\boldsymbol{u} - \boldsymbol{v}\|, \\ &\|\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})\| \leqq a \|\boldsymbol{v} - \boldsymbol{w}\|, \\ &\|\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{z})\| \leqq a \|\boldsymbol{w} - \boldsymbol{z}\|. \end{aligned} \tag{2.3}$$

Der Beweis ist nicht schwierig, vgl. [7] oder [12]. Aus (2.3) lassen sich die folgenden Ungleichungen gewinnen, welche für zwei der behandelten Verfahren entscheidend sind:

$$\begin{aligned} &\|\delta f(\boldsymbol{y}, \boldsymbol{z}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\,(\boldsymbol{y} - \boldsymbol{z}) - f'(\boldsymbol{x}) \\ &\qquad - \{\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T\}(\boldsymbol{y} - \boldsymbol{x})\| \\ &\leqq p\{\|\boldsymbol{u} - \boldsymbol{x}\| + \|\boldsymbol{v} - \boldsymbol{x}\| + \|\boldsymbol{w} - \boldsymbol{x}\| + \|\boldsymbol{y} - \boldsymbol{x}\| + \|\boldsymbol{z} - \boldsymbol{x}\|\}\{\|\boldsymbol{y} - \boldsymbol{x}\| + \|\boldsymbol{z} - \boldsymbol{x}\|\}, \end{aligned} \tag{2.4}$$

$$\|f''(\boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T\| \leqq p\{\|\boldsymbol{u} - \boldsymbol{x}\| + \|\boldsymbol{v} - \boldsymbol{x}\| + \|\boldsymbol{w} - \boldsymbol{x}\|\} \quad \text{für} \quad \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{Q}. \tag{2.5}$$

Dabei ist $p = 2a$ gesetzt worden.

Beweis. Zur Ungleichung (2.5) führen die Umformung

$$\begin{aligned} \delta^2 f(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) &= \{\delta^2 f(\boldsymbol{x}, \boldsymbol{x}, \boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{x})\} \\ &\quad + \{\delta^2 f(\boldsymbol{u}, \boldsymbol{x}, \boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x})\} + \{\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\} \end{aligned}$$

und die entsprechende für die transponierte Matrix, falls man danach die Abschätzungen (2.3) anwendet. Wegen

$$\delta f(\boldsymbol{y}, \boldsymbol{z}) - f'(\boldsymbol{x}) = \delta^2 f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})^T (\boldsymbol{y} - \boldsymbol{x}) + \delta^2 f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x})\,(\boldsymbol{z} - \boldsymbol{x})$$

ist der Vektor auf der linken Seite von (2.4) gleich

$$\{\delta^2 f(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{x}) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\}(\boldsymbol{z} - \boldsymbol{x}) + \{\delta^2 f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})^T - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T\}(\boldsymbol{y} - \boldsymbol{x}),$$

woraus sich nach Einfügungen der eben vorgeführten Art die Abschätzung (2.4) ergibt.

Mit den eingeführten Steigungen wird das folgende Interpolationspolynom gebildet:

$$g(\boldsymbol{x}) = f(\boldsymbol{u}) + \delta f(\boldsymbol{u}, \boldsymbol{v})\,(\boldsymbol{x} - \boldsymbol{u}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\,(\boldsymbol{x} - \boldsymbol{v})\,(\boldsymbol{x} - \boldsymbol{u}). \tag{2.6}$$

3 Bei den angegebenen Normen folgen aus (2.3) sofort die entsprechenden Ungleichungen für die transponierten Matrizen, z.B. $\|\delta^2 f(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})^T - \delta^2 f(\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{z})^T\| \leqq a\|\boldsymbol{u} - \boldsymbol{v}\|$. Bei anderen Normen trifft dies ebenfalls zu; eventuell ist die Konstante a abzuändern.

Man bestätigt nach einfacher Rechnung, daß für die $\frac{1}{2}(d+2)(d+1)$ Vektoren

$$\boldsymbol{y}_{ik} = (w_1, \ldots, w_i, v_{i+1}, \ldots, v_k, u_{k+1}, \ldots, u_d) \quad \text{mit} \quad 0 \leqq i \leqq k \leqq d,$$

zu denen auch $\boldsymbol{u} = \boldsymbol{y}_{00}$, $\boldsymbol{v} = \boldsymbol{y}_{0d}$ und $\boldsymbol{w} = \boldsymbol{y}_{dd}$ gehören, die Interpolationsbedingung $g(\boldsymbol{y}_{ik}) = f(\boldsymbol{y}_{ik})$ erfüllt ist. Die Vektoren $\boldsymbol{y}_{ik}$ sind außerdem genau diejenigen, deren zugehörige Funktionswerte zur Angabe von g erforderlich sind. Es sind dies also genau die verwendeten Stützstellen. Für $d = 3$ sind sie in der Fig. 1 als volle Kreise eingezeichnet.

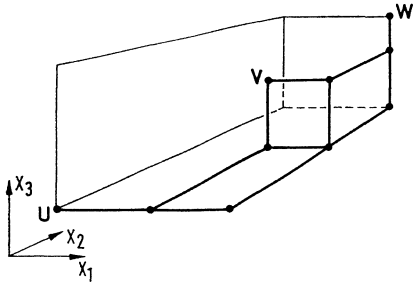

Fig. 1. Stützstellen von $g(\boldsymbol{x})$

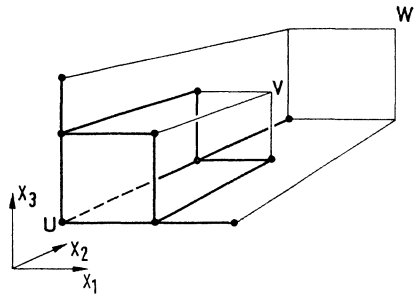

Fig. 2. Stützstellen von $\tilde{g}(\boldsymbol{x})$

Für die Ableitung von g ergibt sich

$$g'(\boldsymbol{x}) = \delta f(\boldsymbol{u}, \boldsymbol{v}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})(\boldsymbol{u} - \boldsymbol{v}) + \{\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T\}(\boldsymbol{x} - \boldsymbol{u}). \tag{2.7}$$

Es wird weiterhin mit einer Interpolationsformel gearbeitet, welche die folgenden $\frac{1}{2}(d+2)(d+1)$ Stützstellen verwendet:

$$\boldsymbol{u},\ (u_1, \ldots, u_{i-1}, v_i, u_{i+1}, \ldots, u_d),\ (u_1, \ldots, u_{i-1}, w_i, u_{i+1}, \ldots, u_d),$$

$$(u_1, \ldots, u_{i-1}, v_i, u_{i+1}, \ldots, u_{k-1}, v_k, u_{k+1}, \ldots, u_d) \quad \text{für} \quad i, k = 1, \ldots, d \text{ und } i < k.$$

Sie sind für $d = 3$ der Fig. 2 zu entnehmen. Das zugehörige Interpolationspolynom lautet

$$\begin{aligned}\tilde{g}(\boldsymbol{x}) = f(\boldsymbol{u}) &+ \sum_{i=1}^{d} \delta f(u_1^{i-1}, u_i v_i, u_{i+1}^d)(x_i - u_i) \\ &+ \sum_{i=1}^{d} \delta^2 f(u_1^{i-1}, u_i v_i w_i, u_{i+1}^d)(x_i - u_i)(x_i - v_i) \\ &+ \sum_{i=1}^{d} \sum_{k=i+1}^{d} \delta^2 f(u_1^{i-1}, u_i v_i, u_{i+1}^{k-1}, u_k v_k, u_{k+1}^d)(x_i - u_i)(x_k - u_k),\end{aligned}$$

während für seine Ableitung

$$\tilde{g}'(\boldsymbol{x}) = \Delta f(\boldsymbol{u}, \boldsymbol{v}) + \hat{\Delta}^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})(\boldsymbol{u} - \boldsymbol{v}) + \Delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})(\boldsymbol{x} - \boldsymbol{u}) \tag{2.8}$$

folgt. Dabei wurden die Abkürzungen

$$(\Delta f(\boldsymbol{u},\boldsymbol{v}))_i = \delta f(u_1^{i-1}, u_i v_i, u_{i+1}^d),$$

$$(\hat{\Delta}^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}))_{ik} = \begin{cases} \delta^2 f(u_1^{i-1}, u_i v_i w_i, u_{i+1}^d) & \text{für } i=k \\ 0 & \text{für } i\neq k, \end{cases}$$

$$(\Delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}))_{ik} = \begin{cases} 2\delta^2 f(u_1^{i-1}, u_i v_i w_i, u_{i+1}^d) & \text{für } i=k \\ \delta^2 f(u_1^{i-1}, u_i v_i, u_{i+1}^{k-1}, u_k v_k, u_{k+1}^d) & \text{für } i<k \\ \delta^2 f(u_1^{k-1}, u_k v_k, u_{k+1}^{i-1}, u_i v_i, u_{i+1}^d) & \text{für } i>k \end{cases}$$

verwendet [12]. Es gelten neben $\Delta f(\boldsymbol{u},\boldsymbol{u}) = f'(\boldsymbol{u})$ und $\Delta^2 f(\boldsymbol{u},\boldsymbol{u},\boldsymbol{u}) = f''(\boldsymbol{u})$ die später benötigten Ungleichungen

$$\begin{aligned}(2.9)\quad &\|\Delta f(\boldsymbol{y},\boldsymbol{z}) + \hat{\Delta}^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})(\boldsymbol{y}-\boldsymbol{z}) - f'(\boldsymbol{x}) - \Delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})(\boldsymbol{y}-\boldsymbol{x})\| \\ &\leqq p\{\|\boldsymbol{u}-\boldsymbol{x}\| + \|\boldsymbol{v}-\boldsymbol{x}\| + \|\boldsymbol{w}-\boldsymbol{x}\| + \|\boldsymbol{y}-\boldsymbol{x}\| + \|\boldsymbol{z}-\boldsymbol{x}\|\}\{\|\boldsymbol{y}-\boldsymbol{x}\| + \|\boldsymbol{z}-\boldsymbol{x}\|\},\end{aligned}$$

$$\begin{aligned}(2.10)\quad &\|f''(\boldsymbol{x}) - \Delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\| \leqq p\{\|\boldsymbol{u}-\boldsymbol{x}\| + \|\boldsymbol{v}-\boldsymbol{x}\| + \|\boldsymbol{w}-\boldsymbol{x}\|\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{für } \boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\boldsymbol{Q}.\end{aligned}$$

Man kann sie ebenso beweisen, wie es für die in [12] zentralen Ungleichungen dort vorgeführt wurde.

3. Verfahren für Aufgaben ohne Restriktionen

Es sei k eine nichtnegative ganze Zahl, deren zweckmäßige Bestimmung später im Abschnitt 6 beschrieben wird.

Verfahren 1. Startvektoren: $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$;
Iterationsvorschrift: $\boldsymbol{w}_{n,-1} = \boldsymbol{y}_n, \boldsymbol{w}_{n,0} = \boldsymbol{z}_n$

$$\begin{aligned}&\Delta f(\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1}) + \hat{\Delta}^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu} - \boldsymbol{w}_{n,\nu-1}) \\ &\qquad + \Delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu+1} - \boldsymbol{w}_{n,\nu}) = \boldsymbol{0} \quad (\nu = 0, \ldots, k),\end{aligned}$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{w}_{n,k-1}, \quad \boldsymbol{y}_{n+1} = \boldsymbol{w}_{n,k}, \quad \boldsymbol{z}_{n+1} = \boldsymbol{w}_{n,k+1} \quad (n = 0, 1, \ldots).$$

Verfahren 2. Startvektoren: $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$;
Iterationsvorschrift: $\boldsymbol{w}_{n,-1} = \boldsymbol{y}_n, \boldsymbol{w}_{n,0} = \boldsymbol{z}_n$

$$\begin{aligned}&\delta f(\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1}) + \delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu} - \boldsymbol{w}_{n,\nu-1}) \\ &\qquad + \{\delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n) + \delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)^T\}(\boldsymbol{w}_{n,\nu+1} - \boldsymbol{w}_{n,\nu}) = \boldsymbol{0} \quad (\nu = 0, \ldots, k),\end{aligned}$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{w}_{n,k-1}, \quad \boldsymbol{y}_{n+1} = \boldsymbol{w}_{n,k}, \quad \boldsymbol{z}_{n+1} = \boldsymbol{w}_{n,k+1} \quad (n = 0, 1, \ldots).$$

Man bestimmt also in beiden Verfahren den Vektor $\boldsymbol{w}_{n,\nu+1}$ jeweils aus einem linearen Gleichungssystem, dessen Koeffizientenmatrix unabhängig von ν ist. Für $k=0$, $d=0$ fallen beide Verfahren zusammen; für diesen Fall sind sie in [9] betrachtet worden, während für $k=0$, $d>1$ das Verfahren 1 in [12] und das Verfahren 2 in [7] behandelt worden sind. Für $k=0$ ist auch die Entstehung der Verfahren leicht zu erkennen. Der Vektor $\boldsymbol{z}_{n+1} = \boldsymbol{w}_{n,1}$ löst dann die Gleichung $g'(\boldsymbol{x}) = \boldsymbol{0}$ bzw. $\tilde{g}'(\boldsymbol{x}) = \boldsymbol{0}$, wobei das Interpolationspolynom g bzw. $\tilde{g}$ durch die

Vektoren z_n, z_{n-1} und z_{n-2} festgelegt wird. Die Gleichung $g'(x)=0$ bzw. $\tilde{g}'(x)=0$ kann als Näherung für $f'(x)=0$ angesehen werden.

Es wird in diesem und dem folgenden Abschnitt vorausgesetzt, daß ein stationärer Vektor x^* von f existiert, d.h. x^* erfülle $f'(x)=0$. Außerdem sei die zugehörige Funktionalmatrix regulär,

$$\|f''(x^*)^{-1}\| \leqq q. \tag{3.1}$$

In einer Umgebung

$$Q_\varrho(x^*)=\{x\in R: |x_i-x_i^*| \leqq \varrho \text{ für } i=1,\ldots,d\}$$

mit $\varrho>0$ mögen die dritten partiellen Differenzenquotienten von f beschränkt sein, so daß dort die Ungleichungen (2.3), (2.9) und (2.10) gelten.

Es werden jetzt für den Konvergenzsatz benötigte Zahlenfolgen eingeführt.

Zahlenfolgen 1. Startwerte: r_0, s_0, t_0.

Rekursionsformel für $k=2l$: $m_n=10pq/(1-3pqr_n)$, $q_n=m_nr_n$,

$$r_{n+1}=q_n^l s_n, \quad s_{n+1}=q_n^l t_n, \quad t_{n+1}=q_n^{l+1}s_n=q_n r_{n+1} \quad (n=0,1,\ldots).$$

Rekursionsformel für $k=2l+1$: $m_n=10pq/(1-3pqr_n)$, $q_n=m_nr_n$,

$$r_{n+1}=q_n^l t_n, \quad s_{n+1}=q_n^{l+1}s_n, \quad t_{n+1}=q_n^{l+1}t_n=q_n r_{n+1} \quad (n=0,1,\ldots).$$

Aussage 1. Es wird neben $3pqr_0<1$ die Ungleichungskette

$$q_0 s_0 \leqq t_0 \leqq s_0 \leqq r_0 \leqq \varrho \tag{3.2}$$

vorausgesetzt. Dann gelten für $n\geqq 0$

a) $r_n \leqq r_{n-1} \leqq \cdots \leqq r_0 \leqq \varrho$, b) $q_n \leqq q_{n-1} \leqq \cdots \leqq q_0 \leqq 1$,

c) $q_n s_n \leqq t_n \leqq s_n \leqq r_n$.

Außerdem ergibt sich, falls $n\geqq 2$ ist, für $k=2l$

$$r_{n+1} \leqq m_0^{2l+1} r_n^l r_{n-1}^{l+1} r_{n-2} \tag{3.3}$$

und für $k=2l+1$

$$r_{n+1} \leqq m_0^{l+1} r_n^{l+1} r_{n-1}. \tag{3.4}$$

Beweis. Für $n=0$ sind a), b) und c) nach Voraussetzung erfüllt. Der Schluß von n auf $n+1$ kann im Falle $k=2l$ wie folgt geführt werden: Wegen $q_n \leqq 1$ ist $r_{n+1} \leqq s_n \leqq r_n$ und somit auch $q_{n+1} \leqq q_n$. Weiterhin folgert man

$$q_{n+1}s_{n+1} \leqq q_n s_{n+1} = q_n^{l+1} t_n \leqq q_n^{l+1} s_n = t_{n+1} \leqq q_n^l t_n = s_{n+1} \leqq q_n^l s_n = r_{n+1}.$$

Im Falle $k=2l+1$ ergibt sich wegen $q_n \leqq 1$ sofort $r_{n+1} \leqq t_n \leqq r_n$ und $q_{n+1} \leqq q_n$, sowie

$$q_{n+1}s_{n+1} \leqq q_n s_{n+1} = q_n^{l+2} s_n \leqq q_n^{l+1} t_n = t_{n+1} \leqq q_n^{l+1} s_n = s_{n+1} \leqq q_n^l t_n = r_{n+1}.$$

Damit sind a), b) und c) bestätigt. Die Ungleichung (3.3) erhält man über

$$r_{n+1}=q_n^l s_n = q_n^l q_{n-1}^l t_{n-1} = q_n^l q_{n-1}^l t_{n-1} = q_n^l q_{n-1}^l q_{n-2} r_{n-1} \leqq m_0^{2l+1} r_n^l r_{n-1}^{l+1} r_{n-2}$$

und die Ungleichung (3.4) über

$$r_{n+1} = q_n^l t_n = q_n^l q_{n-1} r_n \leqq m_0^{l+1} r_n^{l+1} r_{n-1}.$$

Zurückgehend auf die Verfahren 1 und 2 wird verlangt, daß für ihre Startvektoren

$$\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leqq r_0, \quad \|\boldsymbol{y}_0 - \boldsymbol{x}^*\| \leqq s_0, \quad \|\boldsymbol{z}_0 - \boldsymbol{x}^*\| \leqq t_0 \tag{3.5}$$

gelten. Außerdem sei neben (3.2) die Ungleichung $13\,p\,q\,r_0 < 1$ erfüllt. Hieraus folgen $q_0 = m_0 r_0 < 1$ und über (3.3) und (3.4), daß $\lim r_n = 0$ ist.

Behauptung. Die Vektoren $\boldsymbol{w}_{n,\nu}$ seien nach dem Verfahren 1 oder nach dem Verfahren 2 bestimmt. Mit den Zahlenfolgen 1 erhält man

$$\|\boldsymbol{w}_{n,\nu} - \boldsymbol{x}^*\| \leqq \begin{cases} q_n^\lambda t_n & \text{für} \quad \nu = 2\lambda \\ q_n^{\lambda+1} s_n & \text{für} \quad \nu = 2\lambda + 1 \end{cases} \quad (\nu = -1, \ldots, k+1;\ n = 0, 1, \ldots),$$

woraus sich sowohl für $k = 2l$ als auch für $k = 2l+1$ die Abschätzungen

$$\|\boldsymbol{x}_{n+1} - \boldsymbol{x}^*\| \leqq r_{n+1}, \quad \|\boldsymbol{y}_{n+1} - \boldsymbol{x}^*\| \leqq s_{n+1}, \quad \|\boldsymbol{z}_{n+1} - \boldsymbol{x}^*\| \leqq t_{n+1} \tag{3.6}$$
$$(n = 0, 1, \ldots)$$

ergeben.

Beweis. Es genügt zu zeigen, daß aus $\|\boldsymbol{x}_n - \boldsymbol{x}^*\| \leqq r_n$, $\|\boldsymbol{y}_n - \boldsymbol{x}^*\| \leqq s_n$, $\|\boldsymbol{z}_n - \boldsymbol{x}^*\| \leqq t_n$ die genannte Abschätzung für $\nu = -1, \ldots, k+1$ folgt, wenn $n \geqq 0$ irgendeine ganze Zahl ist. Um im Beweis nicht ständig zwischen dem Verfahren 1 und dem Verfahren 2 unterscheiden zu müssen, werden die Abkürzungen

$$\begin{aligned} &\boldsymbol{h}(\boldsymbol{u}, \boldsymbol{v}) = \delta f(\boldsymbol{u}, \boldsymbol{v}) \quad \text{bzw.} \quad \boldsymbol{h}(\boldsymbol{u}, \boldsymbol{v}) = \Delta f(\boldsymbol{u}, \boldsymbol{v}) \\ &\boldsymbol{H}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \quad \text{bzw.} \quad \boldsymbol{H}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \hat{\Delta}^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \\ &\boldsymbol{G}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T \quad \text{bzw.} \quad \boldsymbol{G}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \Delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \end{aligned} \tag{3.7}$$

eingeführt.

Wenn die Matrix $\boldsymbol{F}_n$ durch

$$\boldsymbol{F}_n = f''(\boldsymbol{x}^*)^{-1}\{\boldsymbol{G}(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n) - \boldsymbol{G}(\boldsymbol{x}^*, \boldsymbol{x}^*, \boldsymbol{x}^*)\}$$

definiert wird, folgt mit (2.5) bzw. (2.10) wegen $\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{z}_n \in \boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$

$$\|\boldsymbol{F}_n\| \leqq p\,q(r_n + s_n + t_n) \leqq 3\,p\,q\,r_n < 1.$$

Es existiert also $\boldsymbol{G}(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)^{-1} = (\boldsymbol{I} + \boldsymbol{F}_n)^{-1} f''(\boldsymbol{x}^*)^{-1}$, und es ist

$$\|\boldsymbol{G}(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)^{-1}\| \leqq \frac{q}{1 - 3\,p\,q\,r_n}.$$

Insbesondere sind damit die linearen Gleichungssysteme für $\boldsymbol{w}_{n,\nu+1}$ stets eindeutig lösbar.

Die Behauptung ist für $\nu = -1$ und $\nu = 0$ wegen $\|\boldsymbol{y}_n - \boldsymbol{x}^*\| \leqq s_n$ und $\|\boldsymbol{z}_n - \boldsymbol{x}^*\| \leqq t_n$ richtig. Für den Schluß von ν auf $\nu+1$ wird neben $\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{z}_n \in \boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$ auch $\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1} \in \boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$ vorausgesetzt. Über die aus den Iterationsvorschriften zu

gewinnende Beziehung

$$\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^* = -\boldsymbol{G}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)^{-1}\{\boldsymbol{h}(\boldsymbol{w}_{n,\nu},\boldsymbol{w}_{n,\nu-1}) + \boldsymbol{H}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{w}_{n,\nu-1}) - f'(\boldsymbol{x}^*) - \boldsymbol{G}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*)\}$$

kommt man mit Hilfe von (2.4) bzw. (2.9) zu der Abschätzung

$$\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\| \leqq \frac{pq}{1-3pqr_n}\{r_n+s_n+t_n+\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|+\|\boldsymbol{w}_{n,\nu-1}-\boldsymbol{x}^*\|\} \cdot\{\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|+\|\boldsymbol{w}_{n,\nu-1}-\boldsymbol{x}^*\|\}.$$

Mit dieser Ungleichung erhält man für eine gerade Zahl $\nu=2\lambda$

$$\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\| \leqq \frac{5pqr_n}{1-3pqr_n}\{q_n^\lambda t_n+q_n^\lambda s_n\} \leqq q_n q_n^\lambda s_n = q_n^{\lambda+1}s_n.$$

Falls $\nu=2\lambda+1$ ist, folgt mit $q_n s_n \leqq t_n$

$$\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\| \leqq \frac{5pqr_n}{1-3pqr_n}\{q_n^{\lambda+1}s_n+q_n^\lambda t_n\} \leqq q_n q_n^\lambda t_n = q_n^{\lambda+1}t_n.$$

Somit ist die Behauptung bewiesen.

Da (r_n) eine Nullfolge ist, konvergieren die Folgen $(\boldsymbol{w}_{n,\nu})$ für jedes in Frage kommende ν. Insbesondere ist $\lim \boldsymbol{x}_n = \boldsymbol{x}^*$. Für die Konvergenzgeschwindigkeit der letzten Folge sind die Ungleichungen (3.3) und (3.4) maßgeblich. Die zugehörigen charakteristischen Gleichungen lauten für $k=2l$

$$\varkappa = l+\frac{l+1}{\varkappa}+\frac{1}{\varkappa^2} \tag{3.8}$$

und für $k=2l+1$

$$\varkappa = l+1+\frac{1}{\varkappa}. \tag{3.9}$$

Sie besitzen jeweils eine positive, dominante Lösung

$$\varkappa = \varkappa^*(k), \tag{3.10}$$

durch welche bekanntlich die Konvergenzgeschwindigkeit festgelegt wird[4]. Für einige k können die Zahlen $\varkappa^*(k)$ der Tabelle 1 entnommen werden. Wie bestätigt

Tabelle 1. *Konvergenzgeschwindigkeiten (3.10) und (3.12)*

k	$\varkappa^*(k)$	$\varkappa^{**}(k)$	k	$\varkappa^*(k)$	$\varkappa^{**}(k)$
0	1,32	1,41	6	4,05	4,19
1	1,62	2	7	4,24	5
2	2,15	2,30	8	5,03	5,16
3	2,41	3	9	5,19	6
4	3,08	3,24	10	6,02	6,14
5	3,30	4			

4 Wegen $\|\boldsymbol{x}_n-\boldsymbol{x}^*\| \leqq r_n$ erfolgt die Konvergenz der Folge $(\boldsymbol{x}_n)$ mindestens mit der Geschwindigkeit von (r_n).

werden kann, gilt generell für $k=2l$ und $k=2l+1$

$$l+1<\varkappa^*(k)<l+2; \tag{3.11}$$

also strebt $\varkappa^*(k)$ mit k monoton gegen Unendlich.

Beachtet man, daß bei hinreichend guten Startvektoren $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$ sich Zahlen r_0, s_0, t_0 mit (3.2), (3.5) und $13\,p\,q\,r_0<1$ finden lassen, so erhält man zusammenfassend den

Konvergenzsatz 1. *Es existiere ein stationärer Vektor $\boldsymbol{x}^*$ von f, und es sei $f''(\boldsymbol{x}^*)$ nicht singulär. In einer Umgebung von $\boldsymbol{x}^*$ seien die dritten partiellen Differenzenquotienten von f vorhanden und beschränkt. Dann konvergieren die Verfahren 1 und 2 mit der Geschwindigkeit $\varkappa^*(k)$ aus* (3.10) *gegen $\boldsymbol{x}^*$, falls nur die Startvektoren gut genug sind.*

Für die Frage, wie man die Zahl k zweckmäßig wählt, wird der Wirkungsgrad (nach Ostrowski [6]) herangezogen. Hier gehen die eben errechnete Konvergenzgeschwindigkeit und außerdem der Aufwand pro Schritt entscheidend ein. Für genauere Ausführungen sei auf den Abschnitt 6 verwiesen.

Es werde noch eine Variante der Verfahren 1 und 2 angeführt, wobei zur Vereinfachung die vereinheitlichenden Abkürzungen (3.7) benutzt werden.

Verfahren 3. Startvektoren: $\boldsymbol{x}_0, \boldsymbol{y}_0$; $\boldsymbol{u}_0=\alpha\boldsymbol{x}_0+(1-\alpha)\boldsymbol{y}_0$ mit $0<\alpha<1$.
Iterationsvorschrift: $\boldsymbol{w}_{n,-1}=\boldsymbol{u}_n$, $\boldsymbol{w}_{n,0}=\boldsymbol{y}_n$

$$\begin{aligned}&\boldsymbol{h}(\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1})+\boldsymbol{H}(\boldsymbol{y}_n, \boldsymbol{u}_n, \boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{w}_{n,\nu-1})\\ &\qquad+\boldsymbol{G}(\boldsymbol{y}_n, \boldsymbol{u}_n, \boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu+1}-\boldsymbol{w}_{n,\nu})=\boldsymbol{0} \qquad (\nu=0, \ldots, k),\\ &\boldsymbol{x}_{n+1}=\boldsymbol{w}_{n,k}, \quad \boldsymbol{y}_{n+1}=\boldsymbol{w}_{n,k+1}; \quad \boldsymbol{u}_{n+1}=\alpha\boldsymbol{x}_{n+1}+(1-\alpha)\boldsymbol{y}_{n+1} \quad (n=0, 1, \ldots).\end{aligned}$$

Für dieses Verfahren gilt, wie hier ohne Beweis mitgeteilt wird, der Konvergenzsatz 1 Wort für Wort, wenn man von dem Wert der Konvergenzgeschwindigkeit absieht. Er ändert sich zu

$$\varkappa=\varkappa^{**}(k), \tag{3.12}$$

wobei diese Zahl die positive, dominante Lösung der Gleichung $\varkappa=l+(l+2)/\varkappa$ für $k=2l$ und gleich $l+2$ für $k=2l+1$ ist.

4. Verfahren ohne Auflösung linearer Gleichungssysteme

Es werden die Verfahren 1 und 2 abgeändert, indem die dort erforderliche Auflösung linearer Gleichungssysteme in der Weise genähert vorgenommen wird, daß keine Verminderung der Konvergenzgeschwindigkeit eintritt. Die Approximation wird mit Hilfe einer Erweiterung des Schulz-Verfahrens zur Matrizeninversion durchgeführt; dabei wird die Gleichungsauflösung durch eine feste Anzahl von Matrizenmultiplikationen ersetzt. Zur Beschreibung des Verfahrens wird wieder von den Abkürzungen (3.7) Gebrauch gemacht, um beide Interpolationsansätze gleichzeitig zu erfassen.

Verfahren 4. Startvektoren: $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$; Startmatrix: $\boldsymbol{X}_0$;
Iterationsvorschrift: $\boldsymbol{w}_{n,-1}=\boldsymbol{y}_n, \quad \boldsymbol{w}_{n,0}=\boldsymbol{z}_n$

$$\boldsymbol{w}_{n,\nu+1}=\boldsymbol{w}_{n,\nu}-\boldsymbol{X}_n\{\boldsymbol{h}(\boldsymbol{w}_{n,\nu},\boldsymbol{w}_{n,\nu-1})+\boldsymbol{H}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{w}_{n,\nu-1})\}\ (\nu=0,\ldots,k),$$

$$\boldsymbol{x}_{n+1}=\boldsymbol{w}_{n,k-1}, \quad \boldsymbol{y}_{n+1}=\boldsymbol{w}_{n,k}, \quad \boldsymbol{z}_{n+1}=\boldsymbol{w}_{n,k+1},$$

$$\boldsymbol{X}_{n+1}=\boldsymbol{X}_n+\boldsymbol{X}_n\sum_{\mu=1}^{l+1}\{\boldsymbol{I}-\boldsymbol{G}(\boldsymbol{z}_{n+1},\boldsymbol{y}_{n+1},\boldsymbol{x}_{n+1})\boldsymbol{X}_n\}^{\mu} \quad (n=0,1,\ldots).$$

Die Zahlen k und l hängen über $k=2l$ oder $k=2l+1$ zusammen.

Für eine Konvergenzuntersuchung werden zuerst angepaßte Zahlenfolgen definiert.

Zahlenfolgen 2. Startwerte: r_0, s_0, t_0, ϱ_0;
Rekursionsformel für $k=2l$: $q_n=m r_n, \quad p_n=m\sqrt[l]{1+\beta}\, r_n$ mit $\beta\geqq 1, \ m>0,$

$$\varrho_{n+1}=q_n^l s_n, \quad s_{n+1}=q_n^l t_n, \quad t_{n+1}=q_n^{l+1}s_n, \quad r_{n+1}=q_n^{l+1}r_n+\beta\,\varrho_{n+1}.$$

Rekursionsformel für $k=2l+1$: q_n, p_n wie oben

$$\varrho_{n+1}=q_n^l t_n, \quad s_{n+1}=q_n^{l+1}s_n, \quad t_{n+1}=q_n^{l+1}t_n, \quad r_{n+1}=q_n^{l+1}r_n+\beta\,\varrho_{n+1}.$$

Aussage 2. Falls $p_0\leqq 1$ und

$$q_0 s_0\leqq t_0\leqq s_0\leqq\varrho_0\leqq r_0\leqq\varrho$$

erfüllt sind, folgen für $n\geqq 0$

a) $r_n\leqq r_{n-1}\leqq\cdots\leqq r_0\leqq\varrho$,

b) $q_n s_n\leqq t_n\leqq s_n\leqq\varrho_n\leqq r_n\leqq\varrho$;

und im Falle $k=2l$ gilt die Ungleichung

$$r_{n+1}\leqq m^{l+1}r_n^{l+2}+\beta\, m^{2l+1}r_n^l r_{n-1}^{l+1}r_{n-2}, \tag{4.1}$$

sowie im Falle $k=2l+1$ die Ungleichung

$$r_{n+1}\leqq m^{l+1}r_n^{l+2}+\beta\, m^{l+1}r_n^{l+1}r_{n-1}. \tag{4.2}$$

Der Beweis kann weitgehend von dem zur Aussage 1 übernommen werden. Es genügt, einige Ergänzungen für den Schluß von n auf $n+1$ zu bringen. Es wird für a) wegen $q_n\leqq p_n\leqq 1$

$$r_{n+1}\leqq q_n^{l+1}r_n+\beta\, q_n^l r_n\leqq(1+\beta)\,q_n^l r_n=p_n^l r_n\leqq r_n.$$

Für b) ist nur hinzuzufügen, daß $\varrho_{n+1}\leqq r_{n+1}$ wegen $\beta\geqq 1$ ist.

Es sei an die Größen q und ϱ aus (3.1) und die dort genannten Voraussetzungen erinnert. Zusätzlich werden

$$\tilde{q}=\|f''(\boldsymbol{x}^*)\|, \quad m=\tilde{q}+13p(q+1),$$

$$\beta=\max\left\{1,\sum_{\mu=1}^{l+1}\sum_{\sigma=1}^{\mu}\binom{\mu}{\sigma}\tilde{q}^{\mu-\sigma}(q+1)^{\sigma+1}3^{\sigma}p^{\sigma}\right\},$$

sowie $\boldsymbol{X}^*=f''(\boldsymbol{x}^*)^{-1}$ gesetzt. Es seien $\varrho\leqq 1$ und $p_0=m\sqrt[l]{1+\beta}\, r_0<1$, so daß sich (r_n) nach der Aussage 2 als eine Nullfolge erweist. Von den Startwerten zum

Verfahren 4 wird

$$(4.3)\quad \|\boldsymbol{x}_0-\boldsymbol{x}^*\|\leqq r_0,\quad \|\boldsymbol{y}_0-\boldsymbol{x}^*\|\leqq s_0,\quad \|\boldsymbol{z}_0-\boldsymbol{x}^*\|\leqq t_0,\quad \|\boldsymbol{X}_0-\boldsymbol{X}^*\|\leqq r_0$$

verlangt, d.h. sie sollen genügend gut sein. Die wesentlichen Konvergenzeigenschaften des Verfahrens ergeben sich unmittelbar aus der folgenden

Behauptung. Unter den Voraussetzungen (4.3) und $\varrho\leqq 1$, $p_0<1$ gilt für die Vektor- und Matrizenfolgen aus dem Verfahren 4 unter Verwendung der Zahlenfolgen 2

$$\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|\leqq\begin{cases} q_n^{\lambda}t_n & \text{für}\quad \nu=2\lambda\\ q_n^{\lambda+1}s_n & \text{für}\quad \nu=2\lambda+1,\end{cases}$$

$$(4.4)\qquad \|\boldsymbol{X}_{n+1}-\boldsymbol{X}^*\|\leqq r_{n+1}\quad (\nu=-1,\ldots,k+1;\ n=0,1,\ldots);$$

insbesondere hat man für gerades und ungerades k

$$(4.5)\quad \|\boldsymbol{x}_{n+1}-\boldsymbol{x}^*\|\leqq\varrho_{n+1}\leqq r_{n+1},\qquad \|\boldsymbol{y}_{n+1}-\boldsymbol{x}^*\|\leqq s_{n+1},\qquad \|\boldsymbol{z}_{n+1}-\boldsymbol{x}^*\|\leqq t_{n+1}$$
$$(n=0,1,\ldots).$$

Beweis. Es werden (4.4) und (4.5) für einen beliebigen Index $n\geqq 0$, d.h. für $\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{z}_n, \boldsymbol{X}_n$ vorausgesetzt und hieraus die genannten Abschätzungen für $\boldsymbol{w}_{n,\nu}$ $(\nu=-1,\ldots,k+1)$ und $\boldsymbol{X}_{n+1}$ hergeleitet. Für $\nu=-1$ und $\nu=0$ gilt die Behauptung. Im Induktionsschluß von ν auf $\nu+1$ geht man vom ersten Teil der Iterationsvorschrift aus; unter Verwendung von $\boldsymbol{X}^*f''(\boldsymbol{x}^*)=\boldsymbol{I}$ erhält man

$$\begin{aligned}\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^* = &-\boldsymbol{X}_n\{\boldsymbol{h}(\boldsymbol{w}_{n,\nu},\boldsymbol{w}_{n,\nu-1})+\boldsymbol{H}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{w}_{n,\nu-1})-f'(\boldsymbol{x}^*)\\ &-\boldsymbol{G}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)(\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*)\}+\boldsymbol{X}^*\{f''(\boldsymbol{x}^*)-\boldsymbol{G}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)\}(\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*)\\ &+(\boldsymbol{X}^*-\boldsymbol{X}_n)\{f''(\boldsymbol{x}^*)+[\boldsymbol{G}(\boldsymbol{z}_n,\boldsymbol{y}_n,\boldsymbol{x}_n)-f''(\boldsymbol{x}^*)]\}(\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*).\end{aligned}$$

Da neben $\boldsymbol{x}_n, \boldsymbol{y}_n, \boldsymbol{z}_n\in\boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$ auch $\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1}\in\boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$ sind, ergibt sich mit (2.4), (2.5) bzw. (2.9), (2.10) und $\|\boldsymbol{X}_n\|\leqq\|\boldsymbol{X}^*\|+\|\boldsymbol{X}_n-\boldsymbol{X}^*\|\leqq q+1$:

$$\begin{aligned}\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\|\leqq &\,p(q+1)\{r_n+s_n+t_n+\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|+\|\boldsymbol{w}_{n,\nu-1}-\boldsymbol{x}^*\|\}\\ &\cdot\{\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|+\|\boldsymbol{w}_{n,\nu-1}-\boldsymbol{x}^*\|\}+pq\{r_n+s_n+t_n\}\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|\\ &+\{\tilde{q}+p(r_n+s_n+t_n)\}r_n\|\boldsymbol{w}_{n,\nu}-\boldsymbol{x}^*\|.\end{aligned}$$

Man gewinnt hieraus bei Verwendung von $t_n\leqq s_n\leqq r_n\leqq 1$, daß für $\nu=2\lambda$

$$\begin{aligned}\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\|&\leqq 5p(q+1)r_n\{q_n^{\lambda}t_n+q_n^{\lambda}s_n\}+3pqr_nq_n^{\lambda}t_n+(\tilde{q}+3p)r_nq_n^{\lambda}t_n\\ &\leqq mr_nq_n^{\lambda}s_n=q_n^{\lambda+1}s_n\end{aligned}$$

ist, während die Behauptung für $\nu=2\lambda+1$ über

$$\begin{aligned}\|\boldsymbol{w}_{n,\nu+1}-\boldsymbol{x}^*\|&\leqq 5p(q+1)r_n\{q_n^{\lambda+1}s_n+q_n^{\lambda}t_n\}+3pqr_nq_n^{\lambda+1}s_n+(\tilde{q}+3p)r_nq_n^{\lambda+1}s_n\\ &\leqq mr_nq_n^{\lambda}t_n=q_n^{\lambda+1}t_n\end{aligned}$$

folgt. Gleichzeitig ist damit (4.5) gezeigt. Es sind daher $\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}, \boldsymbol{z}_{n+1}\in\boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$. Es ist noch (4.4) zu bestätigen. Nach einigen elementaren Umformungen, die im einzelnen nicht wiedergegeben werden sollen (vgl. [10]), bekommt man aus dem

zweiten Teil der Iterationsvorschrift

$$\boldsymbol{X}_{n+1}-\boldsymbol{X}^*=(\boldsymbol{X}_n-\boldsymbol{X}^*)\{f''(\boldsymbol{x}^*)(\boldsymbol{X}^*-\boldsymbol{X}_n)\}^{l+1}$$
$$+\sum_{\mu=1}^{l+1}\sum_{\sigma=1}^{\mu}\binom{\mu}{\sigma}\boldsymbol{X}_n\{f''(\boldsymbol{x}^*)(\boldsymbol{X}^*-\boldsymbol{X}_n)\}^{\mu-\sigma}\{(f''(\boldsymbol{x}^*)-\boldsymbol{G}(\boldsymbol{z}_{n+1},\boldsymbol{y}_{n+1},\boldsymbol{x}_{n+1}))\boldsymbol{X}_n\}^{\sigma}.$$

Somit kann unter Beachtung von (2.5) bzw. (2.10) und $t_{n+1}\leqq s_{n+1}\leqq\varrho_{n+1}\leqq r_n\leqq 1$ folgendermaßen abgeschätzt werden:

$$\|\boldsymbol{X}_{n+1}-\boldsymbol{X}^*\|\leqq\tilde{q}^{l+1}r_n^{l+2}+\sum_{\mu=1}^{l+1}\sum_{\sigma=1}^{\mu}\binom{\mu}{\sigma}\tilde{q}^{\mu-\sigma}(q+1)^{\sigma+1}r_n^{\mu-\sigma}p^{\sigma}(t_{n+1}+s_{n+1}+\varrho_{n+1})^{\sigma}$$
$$\leqq m^{l+1}r_n^{l+2}+\beta\varrho_{n+1}=r_{n+1}.$$

Damit ist der Beweis der Behauptung erbracht.

Da (r_n) bei den genannten Voraussetzungen eine Nullfolge ist, erhält man mit (4.4) und (4.5) u.a. die Aussagen $\lim \boldsymbol{x}_n=\boldsymbol{x}^*$ und $\lim \boldsymbol{X}_n=\boldsymbol{X}^*=f''(\boldsymbol{x}^*)^{-1}$. Die Konvergenzgeschwindigkeit wird wegen (4.4) und (4.5) durch die Ungleichungen (4.1) und (4.2) festgelegt. Zu dem jeweils ersten Summanden auf der rechten Seite gehört die Geschwindigkeit $\varkappa=l+2$, während die zu dem zweiten Summanden gehörige Geschwindigkeit den Wert $\varkappa=\varkappa^*(k)$ aus (3.10) hat. Da nach (3.11) stets $\varkappa^*(k)<l+2$ ist, gibt $\varkappa^*(k)$ den Ausschlag.

Der folgende Satz ist somit in allen Teilen bewiesen.

Konvergenzsatz 2. *Es sei ein stationärer Vektor $\boldsymbol{x}^*$ von f vorhanden, und es existiere $\boldsymbol{X}^*=f''(\boldsymbol{x}^*)^{-1}$. In einer Umgebung von $\boldsymbol{x}^*$ seien die dritten partiellen Differenzenquotienten von f angebbar und beschränkt. Bei hinreichend guten Startwerten konvergiert dann das Verfahren 4 mit der Geschwindigkeit $\varkappa=\varkappa^*(k)$ aus (3.10), d.h. mit der Geschwindigkeit $\varkappa=\varkappa^*(k)$ streben die Vektorfolge $(\boldsymbol{x}_n)$ gegen $\boldsymbol{x}^*$ und die Matrizenfolge $(\boldsymbol{X}_n)$ gegen $\boldsymbol{X}^*$.*

5. Verfahren für Aufgaben mit Restriktionen (Optimierungsaufgaben)

Es sei die Aufgabe

$$f(\boldsymbol{x})=\text{Minimum!}\quad\text{für}\quad\boldsymbol{x}\in\boldsymbol{K}\tag{5.1}$$

gestellt, d.h. es ist ein Vektor $\boldsymbol{x}^*$ gesucht, für den $f(\boldsymbol{x}^*)\leqq f(\boldsymbol{x})$ für alle $\boldsymbol{x}\in\boldsymbol{K}$ zutrifft. $\boldsymbol{K}$ wird als konvexe, abgeschlossene und nicht leere Teilmenge von $\boldsymbol{R}$ angenommen. Es wird vorausgesetzt, daß die im Abschnitt 2 definierte zweite Steigung $\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})$ für alle $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\boldsymbol{K}$ existiert[5] und stetig ist und daß es eine Konstante $\alpha>0$ gibt, mit der

$$\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\boldsymbol{y}\boldsymbol{y}\geqq\alpha\|\boldsymbol{y}\|^2\quad\text{für}\quad\boldsymbol{y}\in\boldsymbol{R}\quad\text{und}\quad\boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in\boldsymbol{K}\tag{5.2}$$

gilt. Für die transponierte Matrix $\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})^T$ besteht dann ebenfalls die Ungleichung, so daß mit Hilfe von (2.2) für $\boldsymbol{x}=\boldsymbol{u}=\boldsymbol{v}=\boldsymbol{w}$ die Aussage

$$f''(\boldsymbol{x})\boldsymbol{y}\boldsymbol{y}\geqq 2\alpha\|\boldsymbol{y}\|^2\quad\text{für}\quad\boldsymbol{y}\in\boldsymbol{R}\quad\text{und}\quad\boldsymbol{x}\in\boldsymbol{K}\tag{5.3}$$

5 Es werden dann u.U. auch Funktionswerte für Vektoren außerhalb von $\boldsymbol{K}$ benötigt.

folgt. Die Matrizen $f''(\boldsymbol{x})$ sind also gleichmäßig positiv definit auf $\boldsymbol{K}$. Hieraus ergibt sich die eindeutige Lösbarkeit der Aufgabe (5.1)[6]; der Minimalvektor heiße $\boldsymbol{x}^* \in \boldsymbol{K}$. Außerdem ist f streng konvex auf $\boldsymbol{K}$.

Das Näherungsverfahren basiert darauf, daß man in jedem Schritt die Zielfunktion f durch eine quadratische Funktion approximiert und dann das Optimierungsproblem mit vereinfachter Zielfunktion löst. Für $k=0$ ist das Verfahren in [16] behandelt worden.

Verfahren 5. Startvektoren: $\boldsymbol{x}_0, \boldsymbol{y}_0, \boldsymbol{z}_0$;

Iterationsvorschrift: $\boldsymbol{w}_{n,-1} = \boldsymbol{y}_n$, $\boldsymbol{w}_{n,0} = \boldsymbol{z}_n$,

$\boldsymbol{w}_{n,\nu+1} \in \boldsymbol{K}$ Minimalvektor zu $g_{n,\nu}$ auf $\boldsymbol{K}$, wobei

$$g_{n,\nu}(\boldsymbol{x}) = \delta f(\boldsymbol{w}_{n,\nu}, \boldsymbol{w}_{n,\nu-1})(\boldsymbol{x} - \boldsymbol{w}_{n,\nu}) + \delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{w}_{n,\nu-1})(\boldsymbol{x} - \boldsymbol{w}_{n,\nu})$$

$$(\nu = 0, \ldots, k)$$

$$\boldsymbol{x}_{n+1} = \boldsymbol{w}_{n,k-1}, \quad \boldsymbol{y}_{n+1} = \boldsymbol{w}_{n,k}, \quad \boldsymbol{z}_{n+1} = \boldsymbol{w}_{n,k+1} \quad (n = 0, 1, \ldots).$$

Wegen $g''_{n,\nu}(\boldsymbol{x}) = \delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n) + \delta^2 f(\boldsymbol{z}_n, \boldsymbol{y}_n, \boldsymbol{x}_n)^T$ für alle $\boldsymbol{x} \in \boldsymbol{R}$ erhält man mit (5.2)

$$g''_{n,\nu}(\boldsymbol{x})\,\boldsymbol{y}\,\boldsymbol{y} \geqq 2\alpha \|\boldsymbol{y}\|^2 \quad \text{für} \quad \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{R}.$$

Hieraus folgt die eindeutige Lösbarkeit der Näherungsaufgaben; es gibt für jedes interessierende n und ν genau einen Vektor $\boldsymbol{w}_{n,\nu+1} \in \boldsymbol{K}$ mit $g_{n,\nu}(\boldsymbol{w}_{n,\nu+1}) \leqq g_{n,\nu}(\boldsymbol{x})$ für alle $\boldsymbol{x} \in \boldsymbol{K}$.

Falls $\boldsymbol{K}$ durch lineare Gleichungen oder Ungleichungen festgelegt ist, kann man z.B. das Verfahren von Wolfe zur Lösung der Teilaufgaben heranziehen. Bekanntlich wird bei diesem Verfahren (für eine besonders kurze Form s. [1]) die Aufgabe der quadratischen Optimierung in eine Aufgabe der linearen Optimierung mit einer Zusatzbedingung überführt; diese löst man zweckmäßig nach dem dualen revidierten Simplexverfahren. Die spezielle Struktur der Zielfunktionen $g_{n,\nu}$ zieht dabei die Vereinfachung nach sich, daß man mit einer optimalen Ecke für $g_{n,\nu}$ sofort eine erste Ecke für die nächste Zielfunktion $g_{n,\nu+1}$ hat. Dagegen muß eine optimale Ecke für $g_{n,k}$ nicht notwendig eine Ecke für $g_{n+1,0}$ sein. Auf die Darstellung von Details werde hier verzichtet.

Als Hilfsmittel für einen Konvergenzsatz wird die folgende bekannte Aussage herangezogen.

---

6 Es werde ein Beweis für diese bekannte Aussage wiedergegeben. Es seien $\boldsymbol{a} \in \boldsymbol{K}$ ein fester Vektor und $\boldsymbol{L} = \{\boldsymbol{x} \in \boldsymbol{K}: f(\boldsymbol{x}) \leqq f(\boldsymbol{a})\}$. Die Menge $\boldsymbol{L}$ ist abgeschlossen. Es sei $\boldsymbol{u}$ ein beliebiger Vektor aus $\boldsymbol{L}$. Dann folgt aus der Taylorformel

$$f(\boldsymbol{u}) = f(\boldsymbol{a}) + f'(\boldsymbol{a})(\boldsymbol{u} - \boldsymbol{a}) + \tfrac{1}{2} f''(\boldsymbol{v})(\boldsymbol{u} - \boldsymbol{a})(\boldsymbol{u} - \boldsymbol{a}), \qquad \boldsymbol{v} \in \boldsymbol{K} \text{ passend},$$

mit der Schwarzschen Ungleichung und (5.3)

$$f(\boldsymbol{a}) \geqq f(\boldsymbol{u}) \geqq f(\boldsymbol{a}) - \|f'(\boldsymbol{a})\| \|\boldsymbol{u} - \boldsymbol{a}\| + \alpha \|\boldsymbol{u} - \boldsymbol{a}\|^2, \qquad \|\boldsymbol{u}\| \leqq \|\boldsymbol{a}\| + \|\boldsymbol{u} - \boldsymbol{a}\| \leqq \|\boldsymbol{a}\| + \frac{1}{\alpha} \|f'(\boldsymbol{a})\|,$$

d.h. $\boldsymbol{L}$ ist beschränkt. Somit existiert für die stetige Funktion f ein Vektor $\boldsymbol{x}^* \in \boldsymbol{L}$ mit $f(\boldsymbol{x}^*) = \min\{f(\boldsymbol{x}): \boldsymbol{x} \in \boldsymbol{L}\} = \min\{f(\boldsymbol{x}): \boldsymbol{x} \in \boldsymbol{K}\}$. Wegen der strengen Konvexität von f gibt es keinen weiteren derartigen Vektor.

Aussage. Es seien die Menge $\boldsymbol{K}$ konvex und abgeschlossen und die Funktion h auf $\boldsymbol{K}$ konvex und einmal stetig differenzierbar. Ein Vektor $\boldsymbol{y}\in\boldsymbol{K}$ ist genau dann ein Minimalvektor von h auf $\boldsymbol{K}$, wenn

$$h'(\boldsymbol{y})\,(\boldsymbol{x}-\boldsymbol{y})\geqq 0 \quad \text{für alle} \quad \boldsymbol{x}\in\boldsymbol{K}$$

gilt.

Beweis. Es sei $\boldsymbol{y}$ ein Minimalvektor und sei doch $h'(\boldsymbol{y})\,(\boldsymbol{x}-\boldsymbol{y})<0$ für ein $\boldsymbol{x}\in\boldsymbol{K}$, so wird auch $h'\big(\boldsymbol{y}+\lambda(\boldsymbol{x}-\boldsymbol{y})\big)\,(\boldsymbol{x}-\boldsymbol{y})<0$ für alle $0\leqq\lambda\leqq\lambda_0$ mit einem passenden $0<\lambda_0\leqq 1$. Folglich ist

$$h\big(\boldsymbol{y}+\lambda_0(\boldsymbol{x}-\boldsymbol{y})\big)-h(\boldsymbol{y})=\lambda_0 h'\big(\boldsymbol{y}+\lambda(\boldsymbol{x}-\boldsymbol{y})\big)\,(\boldsymbol{x}-\boldsymbol{y})<0,$$

da $0\leqq\lambda\leqq\lambda_0$ für den entsprechenden Zwischenwert λ gilt. Wegen $\boldsymbol{y}+\lambda_0(\boldsymbol{x}-\boldsymbol{y})\in\boldsymbol{K}$ ist ein Widerspruch erreicht. Gilt umgekehrt $h'(\boldsymbol{y})\,(\boldsymbol{x}-\boldsymbol{y})\geqq 0$ für alle $\boldsymbol{x}\in\boldsymbol{K}$, so ergibt sich bekanntlich mit der Konvexität von h die Ungleichung $h(\boldsymbol{x})\geqq h(\boldsymbol{y})+h'(\boldsymbol{y})\,(\boldsymbol{x}-\boldsymbol{y})$, also wie behauptet $h(\boldsymbol{x})\geqq h(\boldsymbol{y})$ für alle $\boldsymbol{x}\in\boldsymbol{K}$.

Es wird nun eine wichtige, (2.4) bzw. (2.9) entsprechende Ungleichung hergeleitet. Die quadratische Funktion

$$g(\boldsymbol{x})=\delta f(\boldsymbol{y},\boldsymbol{z})\,(\boldsymbol{x}-\boldsymbol{y})+\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\,(\boldsymbol{x}-\boldsymbol{z})\,(\boldsymbol{x}-\boldsymbol{y})$$

besitzt auf $\boldsymbol{K}$ genau einen Minimalvektor $\tilde{\boldsymbol{x}}\in\boldsymbol{K}$, denn $g''(\boldsymbol{x})$ ist gleichmäßig positiv definit auf $\boldsymbol{K}$. In einer Umgebung $\boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$ des Minimalvektors $\boldsymbol{x}^*$ von f mögen die Ungleichungen (2.3) bestehen. Dann gilt für $\boldsymbol{u},\boldsymbol{v},\boldsymbol{w},\boldsymbol{y},\boldsymbol{z}\in\boldsymbol{P}_\varrho(\boldsymbol{x}^*)=\boldsymbol{K}\cap\boldsymbol{Q}_\varrho(\boldsymbol{x}^*)$

$$\begin{aligned}\|\tilde{\boldsymbol{x}}-\boldsymbol{x}^*\|\leqq p\{&\|\boldsymbol{u}-\boldsymbol{x}^*\|+\|\boldsymbol{v}-\boldsymbol{x}^*\|+\|\boldsymbol{w}-\boldsymbol{x}^*\|+\|\boldsymbol{y}-\boldsymbol{x}^*\|+\|\boldsymbol{z}-\boldsymbol{x}^*\|\}\\&\cdot\{\|\boldsymbol{y}-\boldsymbol{x}^*\|+\|\boldsymbol{z}-\boldsymbol{x}^*\|\};\end{aligned}\tag{5.4}$$

dabei ist die Zahl $p\geqq a/2\alpha$.

Beweis von (5.4). Es ist auf Grund der Definition von g

$$\begin{aligned}g(\tilde{\boldsymbol{x}})-g(\boldsymbol{x}^*)=&\,\delta f(\boldsymbol{y},\boldsymbol{z})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)\\&-\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\{(\boldsymbol{x}^*-\boldsymbol{z})\,(\boldsymbol{x}^*-\boldsymbol{y})-(\tilde{\boldsymbol{x}}-\boldsymbol{z})\,(\tilde{\boldsymbol{x}}-\boldsymbol{y})\}.\end{aligned}\tag{5.5}$$

Beachtet man

$$g'(\boldsymbol{x})=\delta f(\boldsymbol{y},\boldsymbol{z})+\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\,(\boldsymbol{x}-\boldsymbol{z})+\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})^T\,(\boldsymbol{x}-\boldsymbol{y})$$

für $\boldsymbol{x}=\tilde{\boldsymbol{x}}$ und außerdem

$$\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})^T\,(\tilde{\boldsymbol{x}}-\boldsymbol{y})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)=\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)\,(\tilde{\boldsymbol{x}}-\boldsymbol{y}),$$

so ergibt sich

$$\begin{aligned}g(\tilde{\boldsymbol{x}})-g(\boldsymbol{x}^*)=&\,g'(\tilde{\boldsymbol{x}})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)-\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\\&\cdot\{(\boldsymbol{x}^*-\boldsymbol{z})\,(\boldsymbol{x}^*-\boldsymbol{y})-(\tilde{\boldsymbol{x}}-\boldsymbol{z})\,(\tilde{\boldsymbol{x}}-\boldsymbol{y})+(\tilde{\boldsymbol{x}}-\boldsymbol{z})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)+(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)\,(\tilde{\boldsymbol{x}}-\boldsymbol{y})\}\\=&\,g'(\tilde{\boldsymbol{x}})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)-\delta^2 f(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*).\end{aligned}$$

Da $\tilde{\boldsymbol{x}}$ Minimalvektor von g ist, gilt nach der obigen Aussage $g'(\tilde{\boldsymbol{x}})\,(\tilde{\boldsymbol{x}}-\boldsymbol{x}^*)\leqq 0$. Verwendet man weiterhin (5.2), so wird $g(\tilde{\boldsymbol{x}})-g(\boldsymbol{x}^*)\leqq-\alpha\|\tilde{\boldsymbol{x}}-\boldsymbol{x}^*\|^2$, d.h.

$$\alpha\|\tilde{\boldsymbol{x}}-\boldsymbol{x}^*\|^2\leqq g(\boldsymbol{x}^*)-g(\tilde{\boldsymbol{x}}).\tag{5.6}$$

Für eine obere Schranke zu $g(\boldsymbol{x}^*) - g(\tilde{\boldsymbol{x}})$ wird die leicht mit (2.2) zu bestätigende Beziehung

$$\delta f(\boldsymbol{y}, \boldsymbol{z}) = f'(\boldsymbol{x}^*) + \delta^2 f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*)(\boldsymbol{z} - \boldsymbol{x}^*) + \delta^2 f(\boldsymbol{y}, \boldsymbol{x}^*, \boldsymbol{x}^*)^T (\boldsymbol{y} - \boldsymbol{x}^*)$$

verwendet; es wird mit (5.5)

$$\begin{aligned} g(\boldsymbol{x}^*) - g(\tilde{\boldsymbol{x}}) = {} & f'(\boldsymbol{x}^*)(\boldsymbol{x}^* - \tilde{\boldsymbol{x}}) + \{\delta^2 f(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x}^*) - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\}(\boldsymbol{z} - \boldsymbol{x}^*)(\boldsymbol{x}^* - \tilde{\boldsymbol{x}}) \\ & + \{\delta^2 f(\boldsymbol{y}, \boldsymbol{x}^*, \boldsymbol{x}^*)^T - \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})^T\}(\boldsymbol{y} - \boldsymbol{x}^*)(\boldsymbol{x}^* - \tilde{\boldsymbol{x}}) \\ & + \delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})\{(\boldsymbol{z} - \boldsymbol{x}^*)(\boldsymbol{x}^* - \tilde{\boldsymbol{x}}) + (\boldsymbol{x}^* - \tilde{\boldsymbol{x}})(\boldsymbol{y} - \boldsymbol{x}^*) \\ & + (\boldsymbol{x}^* - \boldsymbol{z})(\boldsymbol{x}^* - \boldsymbol{y}) - (\tilde{\boldsymbol{x}} - \boldsymbol{z})(\tilde{\boldsymbol{x}} - \boldsymbol{y})\}. \end{aligned}$$

Der letzte Summand auf der rechten Seite ist gleich

$$= -\delta^2 f(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})(\tilde{\boldsymbol{x}} - \boldsymbol{x}^*)(\tilde{\boldsymbol{x}} - \boldsymbol{x}^*) \leqq -\alpha \|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2;$$

für den ersten Summanden rechts gilt nach der Aussage die Beziehung

$$f'(\boldsymbol{x}^*)(\boldsymbol{x}^* - \tilde{\boldsymbol{x}}) \leqq 0,$$

da $\boldsymbol{x}^*$ der Minimalvektor von f ist. Die geschweiften Klammern der mittleren Summanden kann man nach entsprechenden Einfügungen (vgl. Beweis von (2.5)) mit Hilfe von (2.3) jeweils durch

$$a\{\|\boldsymbol{u} - \boldsymbol{x}^*\| + \|\boldsymbol{v} - \boldsymbol{x}^*\| + \|\boldsymbol{w} - \boldsymbol{x}^*\| + \|\boldsymbol{y} - \boldsymbol{x}^*\| + \|\boldsymbol{z} - \boldsymbol{x}^*\|\}$$

abschätzen, so daß man insgesamt

$$\begin{aligned} & g(\boldsymbol{x}^*) - g(\tilde{\boldsymbol{x}}) \leqq -\alpha \|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \\ & \quad + a\{\|\boldsymbol{u} - \boldsymbol{x}^*\| + \|\boldsymbol{v} - \boldsymbol{x}^*\| + \|\boldsymbol{w} - \boldsymbol{x}^*\| + \|\boldsymbol{y} - \boldsymbol{x}^*\| + \|\boldsymbol{z} - \boldsymbol{x}^*\|\} \\ & \quad \cdot \{\|\boldsymbol{y} - \boldsymbol{x}^*\| + \|\boldsymbol{z} - \boldsymbol{x}^*\|\} \|\tilde{\boldsymbol{x}} - \boldsymbol{x}^*\| \end{aligned} \tag{5.7}$$

erhält. Aus (5.6) und (5.7) gewinnt man unmittelbar die gewünschte Formel (5.4).

Geeignete Schrankenfolgen für die Fehler der Näherungen $\boldsymbol{w}_{n,\nu}$ bekommt man durch eine leichte Modifikation der Zahlenfolgen 1.

Zahlenfolgen 3. Startwerte: r_0, s_0, t_0;
Rekursionsformel für $k = 2l$: $q_n = m r_n$ mit $m = 10p$

$$r_{n+1} = q_n^l s_n, \quad s_{n+1} = q_n^l t_n, \quad t_{n+1} = q_n r_{n+1} \quad (n = 0, 1, \ldots).$$

Rekursionsformel für $k = 2l+1$: $q_n = m r_n$ mit $m = 10p$

$$r_{n+1} = q_n^l t_n, \quad s_{n+1} = q_n^{l+1} s_n, \quad t_{n+1} = q_n r_{n+1} \quad (n = 0, 1, \ldots).$$

Völlig analog zur Aussage 1 beweist man

Aussage 3. Unter der Voraussetzung

$$q_0 s_0 \leqq t_0 \leqq s_0 \leqq r_0 \leqq \varrho$$

gelten für $n \geqq 0$

a) $$r_n \leqq r_{n-1} \leqq \cdots \leqq r_0 \leqq \varrho,$$

b) $$q_n s_n \leqq t_n \leqq s_n \leqq r_n,$$

sowie die Ungleichungen (3.3) und (3.4) mit $m_0 = m = 10p$.

Eine wichtige Vorstufe für eine Konvergenzaussage bildet die

Behauptung. Für die Startwerte zum Verfahren 5 gelte

$$\|x_0 - x^*\| \leqq r_0, \quad \|y_0 - x^*\| \leqq s_0, \quad \|z_0 - x^*\| \leqq t_0.$$

Dann bestehen die Abschätzungen

$$\|w_{n,\nu} - x^*\| \leqq \begin{cases} q_n^{\lambda} t_n & \text{für} \quad \nu = 2\lambda \\ q_n^{\lambda+1} s_n & \text{für} \quad \nu = 2\lambda + 1 \end{cases} \qquad (\nu = -1, \ldots, k+1;\ n = 0, 1, \ldots);$$

insbesondere sind

$$\|x_{n+1} - x^*\| \leqq r_{n+1}, \quad \|y_{n+1} - x^*\| \leqq s_{n+1}, \quad \|z_{n+1} - x^*\| \leqq t_{n+1} \quad (n = 0, 1, \ldots).$$

Auf Grund der Beweise zu den beiden vorangegangenen Behauptungen dieser Art genügt es hier zu erwähnen, daß über die Iterationsvorschrift mit Hilfe der Ungleichung (5.4) die Abschätzung

$$\begin{aligned} \|w_{n,\nu+1} - x^*\| \leqq p\{&\|x_n - x^*\| + \|y_n - x^*\| + \|z_n - x^*\| + \|w_{n,\nu} - x^*\| \\ &+ \|w_{n,\nu-1} - x^*\|\}\{\|w_{n,\nu} - x^*\| + \|w_{n,\nu-1} - x^*\|\} \end{aligned}$$

folgt, falls nur $w_{n,\nu}, w_{n,\nu-1}, x_n, y_n, z_n \in P_{\varrho}(x^*)$ sind.

Für $q_0 < 1$ sichert die Behauptung zusammen mit Aussage 3 die Konvergenz der Folge (x_n) gegen x^*, wobei die Geschwindigkeit wiederum $\varkappa^*(k)$ beträgt. Es besteht also der folgende

Konvergenzsatz 3. *Unter der Voraussetzung* (5.2) *hat die Aufgabe* (5.1) *genau einen Minimalvektor* $x^* \in K$. *Falls die dritten partiellen Differenzenquotienten der Zielfunktion* f *in einer Umgebung von* x^* *vorhanden und beschränkt sind, konvergiert die nach dem Verfahren* 5 *bestimmte Vektorfolge* (x_n) *bei hinreichend guten Startwerten mit der Geschwindigkeit* $\varkappa = \varkappa^*(k)$ *aus* (3.10) *gegen* x^*.

Ein Verfahren der hier behandelten Klasse, welches bei konvexen Funktionen für alle Startwerte konvergiert, ist in [17] entwickelt und untersucht worden.

6. Zweckmäßige Bestimmung von k

Nach Ostrowski [6] wird der Wirkungsgrad λ eines Verfahrens durch

$$\lambda = \varkappa^{1/\alpha} \tag{6.1}$$

definiert, wenn $\varkappa$ die Konvergenzgeschwindigkeit und α der Aufwand pro Iterationsschritt sind. Für α wird hier vereinfachend die Anzahl der pro Schritt neu zu berechnenden Funktionswerte von f genommen. Eine Möglichkeit zur Berücksichtigung weiterer Anteile des Aufwandes wie Anzahl der Matrizenbildungen, Anzahl der zu lösenden linearen Gleichungssysteme bzw. Optimierungsaufgaben, ...

ist z.B. in [10] beschrieben worden; einheitliche Aussagen für alle Funktionen f sind nach einer derartigen Verfeinerung jedoch nicht mehr zu erwarten.

Es sollen zwei Verfahren i und j miteinander verglichen werden. Die Wirkungsgrade seien λ_i und λ_j, während zur Erreichung einer bestimmten Genauigkeit insgesamt A_i und A_j Funktionswerte erforderlich seien. Dann gilt die asymptotische Beziehung

$$A_i : A_j \approx \lg \lambda_j : \lg \lambda_i. \tag{6.2}$$

Man wird daher das Verfahren mit dem größeren Wirkungsgrad auswählen. Da bei den vorliegenden Verfahren λ außer von der Dimension d noch von der Zahl k abhängt, wird man $k = k_{\max}$ bei festem d entsprechend der Forderung

$$\lambda(k_{\max}, d) = \max_{k=0,1,\ldots} \lambda(k, d) \tag{6.3}$$

bestimmen.

Für die Anzahl der Funktionswerte, die pro Schritt im laufenden Verfahren erforderlich sind, findet man beim Verfahren 1 für $d > 2$

$$\alpha_1^*(k, d) = \binom{d+2}{2} + k(d+1),$$

beim Verfahren 2

$$\alpha_2^*(k, d) = \binom{d+1}{2} + k d,$$

beim Verfahren 3 (Interpolationspolynom $\tilde{p}(x)$) für $d > 2$

$$\alpha_1^{**}(k, d) = \binom{d+2}{2} + k(d+1)$$

und beim Verfahren 3 (Interpolationspolynom $p(x)$)

$$\alpha_2^{**}(k, d) = \binom{d+2}{2} + k d - 1.$$

Beim Verfahren 4 (Interpolationspolynom $p(x)$) und beim Verfahren 5 ist diese Zahl mit $\alpha_2^*(k, d)$ gleich der kleinsten von den genannten.

Satz. *Unter den Verfahren 1, 2 und 3 hat das Verfahren 2 für jede Dimension d bei optimaler Wahl von k den größten Wirkungsgrad.*

Aus diesem Grunde wurde für Aufgaben mit Restriktionen nur das dem Verfahren 2 entsprechende Verfahren 5 untersucht.

Beweis der Aussage. Es wird behauptet, daß die Ungleichungen

$$\max_k \varkappa^*(k)^{1/\alpha_1^*(k,d)} \leqq \max_k \varkappa^*(k)^{1/\alpha_2^*(k,d)}$$

und

$$\max_k \varkappa^{**}(k)^{1/\alpha_1^{**}(k,d)} \leqq \max_k \varkappa^{**}(k)^{1/\alpha_2^{**}(k,d)} \leqq \max_k \varkappa^*(k)^{1/\alpha_1^*(k,d)}$$

gelten, falls die Maxima jeweils über alle ganzen Zahlen $k \geqq 0$ genommen werden. Die Ungleichungen führen unmittelbar zur Aussage. Wegen $\alpha_2^*(k, d) \leqq \alpha_1^*(k, d)$

und $\alpha_2^{**}(k, d) \leqq \alpha_1^{**}(k, d)$ sind die erste Ungleichung und der erste Teil der zweiten Ungleichung sofort ersichtlich. Da weiterhin $\alpha_2^{**}(k, d) = \alpha_2^{*}(k+1, d)$ und, wie leicht zu bestätigen ist, $\varkappa^{**}(k) < \varkappa^{*}(k+1)$ zutreffen und folglich

$$\varkappa^{**}(k)^{1/\alpha_1^{**}(k,d)} < \varkappa^{*}(k+1)^{1/\alpha_1^{*}(k+1,d)}$$

für jede ganze Zahl $k \geqq 0$ gilt, ist auch der zweite Teil der obigen Ungleichung und damit die Aussage bewiesen.

Die Tabelle 2 enthält zum Verfahren 2, welches auf Grund der obigen Aussage hervorzuheben ist, einige Angaben zur Illustration: In Abhängigkeit von d die optimale Stufenzahl $k_{\max}$, die Logarithmen der Wirkungsgrade $\lambda_2^{*}(k) = \varkappa^{*}(k)^{1/\alpha_1^{*}(k)}$ für $k = 0$ und $k = k_{\max}$, sowie als Maß für den verbleibenden Aufwand bei der Verwendung von $k_{\max}$ das Verhältnis von $\lg \lambda_2^{*}(0)$ zu $\lg \lambda_2^{*}(k_{\max})$ in Prozent (s. (6.2)).

Tabelle 2. *Wirkungsgrade für das Verfahren 2*

d	$k_{\max}$	$\lg \lambda_2^{*}(0)$	$\lg \lambda_2^{*}(k_{\max})$	$\lg \lambda_2^{*}(0) : \lg \lambda_2^{*}(k_{\max})$
1	1	0,12212	0,12212	100%
2	2	0,04071	0,04743	86%
3	2	0,02035	0,02767	74%
4	4	0,01221	0,01879	65%
5	4	0,00814	0,01396	58%
6	4	0,00582	0,01086	54%
7	4	0,00436	0,00872	50%
10	6	0,00222	0,00528	42%
15	8	0,00116	0,00312	37%
20	8	0,00064	0,00201	32%
30	12	0,00028	0,00106	26%

7. Numerische Beispiele

Auf einer Rechenanlage „Minsk 22" (Hochschule für Verkehrswesen „Friedrich List" zu Dresden) sind mehrere Testbeispiele nach dem Verfahren 2 durchgerechnet und der Aufwand an Funktionswerten für $k = 0$ und $k = k_{\max}$ miteinander verglichen worden. Ein Beispiel ist aus [2] übernommen worden:

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} f_i(\boldsymbol{x})^2, \qquad f_i(\boldsymbol{x}) = \sum_{k=1}^{d} (a_{ik} \sin x_k + b_{ik} \cos x_k) + e_i \quad (i = 1, \ldots, d);$$

hierbei sind die a_{ik}, b_{ik} Zufallszahlen aus $[-100, +100]$, während die e_i so gewählt werden, daß ein vorgegebener Zufallsvektor $\boldsymbol{x}^*$ mit Komponenten aus $[-\pi, +\pi]$ die Eigenschaft $f(\boldsymbol{x}^*) = 0$ hat, d.h. $\boldsymbol{x}^*$ ist Lösung der Aufgabe $f(\boldsymbol{x}) =$ Minimum! bei $\boldsymbol{x} \in \boldsymbol{R}$. Mit Zufallszahlen r_i aus $[-\pi, +\pi]$ sind die Komponenten der Startvektoren in der Form $x_i^* + \varepsilon r_i$ gewählt worden; durch die Zahl ε kann die Güte der Startvektoren gesteuert werden. Für einige Dimensionen d sind Angaben über die benötigte Anzahl $A(k)$ von Funktionswerten zur Erreichung einer bestimmten Genauigkeit in der Tabelle 3 zusammengefaßt. In den Beispielen 1 bis 7 wurde $\varepsilon = 0, 1$ gesetzt. Für $d = 6, 8, 10$ stellte sich mit diesem ε keine Konvergenz ein; die Güte der Startvektoren wurde daher verbessert, indem jetzt $\varepsilon = 0{,}01$ genommen wurde (Beispiele 8 bis 13).

Tabelle 3. *Numerische Ergebnisse*

Nr.	d	$f(x_0)$	Verfahren 2, $k=0$		Verfahren 2, $k=k_{\max}$		$A(k_{\max}):A(0)$
			$f(x_{\text{end}})$	$A(0)$	$f(x_{\text{end}})$	$A(k_{\max})$	
1	2	701	$0,^{12}63$	25	$0,^{11}11$	21	84%
2	3	890	$0,^{12}24$	47	$0,^{12}24$	38	81%
3	3	847	$0,^{13}85$	54	$0,^{11}66$	39	72%
4	4	1274	$0,^{11}30$	96	$0,^{9}14$	77	80%
5	4	943	div.		—		
6	6	2709	div.		—		
7	6	7776	div.		—		
8	4	40	$0,^{11}97$	56	$0,^{10}15$	42	75%
9	6	15	$0,^{10}15$	239	$0,^{10}65$	137	57%
10	8	34	$0,^{9}14$	334	div.		
11	10	104	$0,^{8}10$	452	$0,^{10}93$	327	72%
12	10	77	div.		$0,^{8}52$	232	
13	10	88	$0,^{9}14$	728	$0,^{9}11$	432	58%

Die genannten Fälle weisen darauf hin, daß die in den Konvergenzsätzen auftretende Voraussetzung über die Güte der Startvektoren in konkreten Beispielen durchaus eine starke Forderung bedeuten kann. Die Beschaffung von genügend guten Startvektoren kann u.U. ein zu beachtendes Zusatzproblem sein; in manchen Fällen kann man hierfür mit Erfolg langsamer konvergente, aber gegen gröbere Startvektoren weniger anfällige Verfahren heranziehen.

Es sei weiterhin das Beispiel

$$f=\frac{1}{1+(x_1-x_2)^2}+\sin\frac{\pi x_2+x_3}{2}+\exp\left(-\left(\frac{x_1+x_3}{x_2}-2\right)^2\right)=\text{Extremum!}$$

mit den stationären Vektoren $x_1=x_2=x_3=\pm\sqrt{4n+1}$ $(n=0, 1, \ldots)$ erwähnt. Um den Funktionswert f für einen der stationären Vektoren um vier Dezimalstellen zu verbessern, sind bei bestimmten Startvektoren für $k=0$ insgesamt $A(0)=89$ Funktionswerte benötigt worden, während für $k=k_{\max}=2$ nur $A(k_{\max})=38$ Funktionswerte, d.h. 43% von $A(0)$ erforderlich waren.

Es sei schließlich vermerkt, daß in den Beispielen mit singulärer Matrix $f''(x^*)$ die Konvergenz stets nur linear erfolgte.

Literatur

1. Collatz, L., Wetterling, W.: Optimierungsaufgaben. Berlin-Heidelberg-New York: Springer 1966.
2. Fletcher, D., Powell, M. J. D.: A rapidly convergent descent method for minimisation. Comput. J. **6**, 163–168 (1963).
3. Helfrich, H.-P.: Ein modifiziertes Newton-Verfahren. In „Funktionalanalytische Methoden in der numerischen Mathematik". ISNM **12**, 61–70 (1969).
4. Künzi, H. P.: Zum heutigen Stand der nichtlinearen Optimierungstheorie. Unternehmensforschung **12**, 1–22 (1968).
5. Levitin, E. S., Poljak, B. T.: Minimierungsverfahren unter Berücksichtigung von Nebenbedingungen [Russ.]. Ž. Vyčisl. Mat. i Mat. Fiz. **6**, 787–823 (1966).
6. Ostrowski, A.: Solution of equations and systems of equations. 2nd ed. New York-London: Academic Press 1966.

7. Poll, V.: Über einige Methoden zur Bestimmung von stationären Punkten bei Funktionen von mehreren Veränderlichen [Russ.] .Izv. Akad. Nauk Est. SSR Ser. Fiz. Mat. **16**, 35—44 und 157—167 (1967).
8. Schmidt, J. W.: Eine Übertragung der Regula falsi auf Gleichungen in Banachräumen. Z. Ang. Math. Mech. **41**, T 61—T 63 (1961); **43**, 1—8 und 97—110 (1963).
9. — Extremwertermittlung mit Funktionswerten. Wiss. Zeitschr. TU Dresden **12**, 1601—1605 (1963).
10. — Leder, D.: Ableitungsfreie Verfahren ohne Auflösung linearer Gleichungen. Computing **5**, 71—81 (1970).
11. — Schwetlick, H.: Ableitungsfreie Verfahren mit höherer Konvergenzgeschwindigkeit. Computing **3**, 215—226 (1968).
12. — Trinkaus, H.-F.: Extremwertermittlung mit Funktionswerten bei Funktionen von mehreren Veränderlichen. Computing **1**, 224—232 (1966).
13. Traub, J. F.: Iterative methods for the solution of equations. Englewood Cliffs, N. J.: Prentice-Hall 1964.
14. Ulm, S.: Über verallgemeinerte Steigungen [Russ.]. Izv. Akad. Nauk Est. SSR Ser. Fiz. Mat. **16**, 13—26 und 146—156 (1967).
15. — Über Iterationsverfahren mit sukzessiver Approximation des inversen Operators [Russ.]. Izv. Akad. Nauk Est. SSR Ser. Fiz. Mat. **16**, 403—411 (1967).
16. — Poll, V.: Über einige Verfahren zur Lösung von Minimierungsaufgaben [Russ.]. Izv. Akad. Nauk Est. SSR Ser. Fiz. Mat. **17**, 151—163 (1968).
17. Vetters, K.: Quadratische Approximationsmethoden zur konvexen Optimierung. Z. Ang. Math. Mech. **50**, 479—484 (1970).
18. Willers, Fr. A.: Methoden der praktischen Analysis. 3. Aufl. Berlin: Walter de Gruyter 1957.

Prof. Dr. J. W. Schmidt und Dr. K. Vetters
Bereich Numerische Mathematik
Sektion Mathematik
Technische Universität Dresden
DDR-8027 Dresden
Zellescher Weg

Numer. Math. 15, 283–296 (1970)
© by Springer-Verlag 1970

Interpolation Polynomials on the Triangle

ALEXANDER ŽENÍŠEK

Received February 28, 1969

1. Introduction

In [1] and [3] a general theory of Hermite interpolation on a rectangle is given. This theory is applicable to the solution of elliptic boundary value problems by means of the Ritz-Galerkin method when the domain is a rectangular polygon.

A similar theory of Hermite interpolation on a triangle (more applicable because any polygon can be triangulated) has not yet been systematically developed. Synge [5] proved an interpolation theorem for the polynomial of the first degree. Zlámal [6] proved interpolation theorems for polynomials of the second, third and fifth degree. These theorems can be applied in case of the second and fourth order equations.

In this paper, generalizing the methods of [6], the conditions uniquely determining a polynomial of the degree $4m+1$ are given (see Corollary of Theorem 3). These conditions generate, according to Theorem 1, functions m-times continuously differentiable on a triangulated closed domain. Further, a general interpolation theorem is proved (see Theorem 3). Making use of this theorem the rate of convergence of the Ritz procedure for an arbitrary positive definite boundary value problem of the order $2n$ can be derived.

The conditions introduced in Theorems 4, 5 and in Section 5 are a modification of the conditions of Theorem 3.

2. Prefatory Considerations

Let Ω be a simply or multiply connected bounded domain in the (x_1, x_2)-plane with the boundary Γ. Γ consists of a finite number of simple closed polygons Γ_i $(i=0, \ldots, r)$; $\Gamma_1, \ldots, \Gamma_r$ lie inside of Γ_0 and do not intersect. Let $\mathfrak{M}=\{\overline{T}_j\}$ be a triangulation of the set $\overline{\Omega}=\Omega\cup\Gamma$, i.e. a set of a finite number of triangles having the following properties: open triangles are disjoint, the union of closed triangles is $\overline{\Omega}$ and any two adjacent triangles have either a common vertex or a common side. Let d be the number of nodes (vertices) and q the number of segments (sides) in the net which is defined by the triangulation $\mathfrak{M}$. Let us denote the nodes by the symbols P_j $(j=1, \ldots, d)$ and the segments by the symbols l_s $(s=1, \ldots, q)$. Let the integers $k\geqq 0$, $m\geqq 0$ be given; inside of each segment l_s we choose $(m+1)(2k+m)/2$ points

$$Q_1^{(s,\alpha)}, \ldots, Q_{k+\alpha}^{(s,\alpha)} \quad (\alpha=0, \ldots, m).$$

If $\alpha\neq\beta$ then it may be $Q_i^{(s,\alpha)}\equiv Q_j^{(s,\beta)}$.

Let $\{\nu_s\}$ be a set of normals in the triangulation $\mathfrak{M}$. The set $\{\nu_s\}$ can be uniquely determined, for example, by the following prescription: Let P_{j_1}, P_{j_2} be the end-points of a segment l_s and let $j_1 < j_2$. We choose the direction of ν_s so that we have the point P_{j_2} on the right hand if we go through the segment l_s in this direction.

Theorem 1. Let the integers $k \geqq 0$, $m \geqq 0$, $\varkappa \geqq 0$ be given. Let the function $f(x_1, x_2)$ be $(m+\varkappa)$-times continuously differentiable on the set $\overline{\Omega}$ and let

$$n = 2(m+\varkappa) + k + 1. \tag{1}$$

Let a polynomial $p_\lambda(x_1, x_2)$ of the degree n exist on each triangle $\overline{T}_\lambda \in \mathfrak{M}$ such that

$$D^i p_\lambda(P_j) = D^i f(P_j), \quad |i| \leqq m+\varkappa,^1 \tag{2}$$

$$\frac{\partial^\alpha p_\lambda(Q_\varrho^{(s,\alpha)})}{\partial \nu_s^\alpha} = \frac{\partial^\alpha f(Q_\varrho^{(s,\alpha)})}{\partial \nu_s^\alpha} \quad (\varrho = 1, \ldots, k+\alpha;\ \alpha = 0, \ldots, m), \tag{3}$$

where P_j $(j = j_1, j_2, j_3)$ are the vertices of the triangle $\overline{T}_\lambda$ and l_s $(s = s_1, s_2, s_3)$ its sides. Then the function

$$g(x_1, x_2) = p_\lambda(x_1, x_2), \quad (x_1, x_2) \in \overline{T}_\lambda, \tag{4}$$

is m-times continuously differentiable on the set $\overline{\Omega}$.

Proof. Let us choose any two adjacent triangles $\overline{T}_{\lambda_1}, \overline{T}_{\lambda_2} \in \mathfrak{M}$ and let l_s be their common side. Let $P_{j_1} = (x_1^{(1)}, x_2^{(1)})$, $P_{j_2} = (x_1^{(2)}, x_2^{(2)})$ be the end-points of l_s. Let

$$p(x_1, x_2) = p_{\lambda_1}(x_1, x_2) - p_{\lambda_2}(x_1, x_2).$$

We have to prove

$$D^i p(x_1, x_2) = 0, \quad |i| \leqq m, \quad (x_1, x_2) \in l_s.$$

The derivative $\partial^r p(x_1, x_2)/\partial x_1^{r_1} \partial x_2^{r_2}$ $(r = r_1 + r_2)$ can be written in the form of a linear combination of $r+1$ derivatives

$$\frac{\partial^r p(x_1, x_2)}{\partial \nu^r}, \frac{\partial^r p(x_1, x_2)}{\partial \nu^{r-1} \partial \tau}, \ldots, \frac{\partial^r p(x_1, x_2)}{\partial \nu \partial \tau^{r-1}}, \frac{\partial^r p(x_1, x_2)}{\partial \tau^r},$$

$\partial/\partial\nu$ and $\partial/\partial\tau$ being the derivations in the directions ν_s and $P_{j_1} P_{j_2}$, respectively. Thus it is sufficient to prove

$$\frac{\partial^{\alpha+\beta} p(x_1, x_2)}{\partial \nu^\alpha \partial \tau^\beta} \equiv 0, \quad (x_1, x_2) \in l_s \quad (\alpha, \beta = 0, \ldots, m;\ 0 \leqq \alpha + \beta \leqq m). \tag{5}$$

Let

$$g_\alpha(\tau) = \frac{\partial^\alpha p(x_1, x_2)}{\partial \nu^\alpha}\bigg|_{l_s} = \frac{\partial^\alpha}{\partial \nu^\alpha} p\left(x_1^{(1)} + (x_1^{(2)} - x_1^{(1)})\tau,\ x_2^{(1)} + (x_2^{(2)} - x_2^{(1)})\tau\right), \tag{6}$$

where $\tau \in [0, 1]$ and $\alpha = 0, \ldots, m$. It follows from the assumptions of Theorem 1 that $g_\alpha(\tau)$ is a polynomial of the degree $n - \alpha$ and that

$$\frac{\mathrm{d}^\gamma g_\alpha(0)}{\mathrm{d}\tau^\gamma} = \frac{\mathrm{d}^\gamma g_\alpha(1)}{\mathrm{d}\tau^\gamma} = g_\alpha(\tau_1) = g_\alpha(\tau_2) = \cdots = g_\alpha(\tau_{k+\alpha}) = 0,$$

1 In (2) and in what follows we use the notation

$$i = (i_1, i_2), \quad |i| = i_1 + i_2, \quad D^i u = \partial^{|i|} u / \partial x_1^{i_1} \partial x_2^{i_2}.$$

where $\gamma = 0, \ldots, m + \varkappa - \alpha$ and $0 < \tau_1 < \tau_2 < \cdots < \tau_{k+\alpha} < 1$. Making use of the Rolle's theorem we see that

$$g_\alpha(\tau) \equiv 0 \quad (\alpha = 0, \ldots, m). \tag{7}$$

It follows from (7) that

$$\frac{\mathrm{d}^\beta g_\alpha(\tau)}{\mathrm{d}\,\tau^\beta} \equiv 0 \quad (\beta = 1,2, \ldots). \tag{8}$$

The relations (5) follow now from (6)—(8). Theorem 1 is proved.

The number of coefficients of a polynomial $p(x_1, x_2)$ of the degree n is

$$N = (n+1)(n+2)/2,$$

the number of the conditions (2) at one vertex is

$$V = (m + \varkappa + 1)(m + \varkappa + 2)/2$$

and the number of the conditions (3) on one side is

$$S = (m+1)(2k+m)/2.$$

If the conditions (2), (3) belong to the set of N conditions which determine uniquely the polynomial $p_\lambda(x_1, x_2)$ then

$$N - 3(V+S) \geqq 0. \tag{9}$$

The integers

$$\varkappa = m, \quad k = 0 \tag{10}$$

satisfy the inequality (9) and in this case

$$N - 3(V+S) = m(m-1)/2. \tag{11}$$

It follows from [5] and from [6] that in case $m = 0$ and $m = 1$ the conditions (2), (3), (10) determine uniquely the interpolation polynomial of the first and fifth degree, respectively. In case $m \geqq 2$ the integer (11) is equal to the number of the conditions

$$\frac{\partial^{\alpha+\beta} p_\lambda(P_1^{(\lambda)})}{\partial\sigma^\alpha\,\partial\tau^\beta} = \frac{\partial^{\alpha+\beta} f(P_1^{(\lambda)})}{\partial\sigma^\alpha\,\partial\tau^\beta} \quad (\alpha, \beta \geqq m+1;\ \alpha+\beta = 2m+2, \ldots, 3m) \tag{12}$$

or to the number of the conditions

$$D^i p_\lambda(P_0^{(\lambda)}) = D^i f(P_0^{(\lambda)}), \quad |i| \leqq m-2, \tag{13}$$

where $P_0^{(\lambda)}$ is the centre of gravity of the triangle $\overline{T}_\lambda$ with the vertices $P_1^{(\lambda)}$, $P_2^{(\lambda)}$, $P_3^{(\lambda)}$; $P_1^{(\lambda)}$ is the vertex with the smallest angle of the triangle $\overline{T}_\lambda$ and $\partial/\partial\sigma$ and $\partial/\partial\tau$ mean the derivations in the directions $P_1^{(\lambda)}P_2^{(\lambda)}$ and $P_1^{(\lambda)}P_3^{(\lambda)}$, respectively.

In Section 4 we shall prove that a polynomial of the degree $n = 4m+1$ is uniquely determined by the conditions (2), (3), (12) and derive an estimate for the difference $f - p_\lambda$ and its derivatives on $\overline{T}_\lambda$. In case $m = 2, 3$ the same is proved for the conditions (2), (3), (13).

3. Auxiliary Theorems

Theorem 2. Let $g(s)$ be a function of a real parameter $s \in [0, l]$, continuous on $[0, l]$ and having a bounded derivative of the order $n+1$ in $(0, l)$,

$$|g^{(n+1)}(s)| \leqq K_{n+1}, \quad s \in (0, l). \tag{14}$$

Let

$$s_0 = 0 < s_1 < s_2 < \cdots < s_r = l \tag{15}$$

and

$$|g(s_i)| \leqq \eta_i^{(0)}, \ |g'(s_i)| \leqq \eta_i^{(1)}, \ \ldots, \ |g^{(\alpha_i - 1)}(s_i)| \leqq \eta_i^{(\alpha_i - 1)} \quad (i = 0, \ldots, r) \tag{16}$$

where $\eta_i^{(k)}$ are constants and α_i given integers satisfying

$$\alpha_0 + \alpha_1 + \cdots + \alpha_r = n + 1, \quad \alpha_i \geqq 1. \tag{17}$$

Let

$$\eta = \max_{i=0,\ldots,r} \left[\max_{k=0,\ldots,\alpha_i - 1} l^k \eta_i^{(k)} \right]. \tag{18}$$

Then

$$|g^{(p)}(s)| \leqq C_{2p+1} l^{-p} \eta + C_{2p+2} K_{n+1} l^{n+1-p}, \quad s \in [0, l], \tag{19}$$

where $p = 0, 1, \ldots, n-1$. The symbols $C_1, C_2, \ldots, C_{2n}$ denote absolute constants, i.e. constants independent on the function $g(s)$ and on the interval $[0, l]$.

Proof. Let $H_n(s)$ be the Hermite interpolation polynomial of the n-th degree of the function $g(s)$ with s_i $(i = 0, \ldots, r)$ as its interpolation nodes. According to the general theorem for Hermite interpolation polynomials (see, e.g., [2]) we can write

$$\begin{aligned} H_n(s) = \sum_{i=0}^{r} \Bigg[& g(s_i) \frac{a_{i0}^{(n-\alpha_i+1)}}{l^{n-\alpha_i+1}} \Phi_{i, n-\alpha_i+1}(s) \\ & + \left(g(s_i) \frac{a_{i0}^{(n-\alpha_i+2)}}{l^{n-\alpha_i+2}} + g'(s_i) \frac{a_{i1}^{(n-\alpha_i+2)}}{l^{n-\alpha_i+1}} \right) \Phi_{i, n-\alpha_i+2}(s) \\ & + \cdots + \left(g(s_i) \frac{a_{i0}^{(n)}}{l^n} + \cdots + g^{(\alpha_i-1)}(s_i) \frac{a_{i,\alpha_i-1}^{(n)}}{l^{n-\alpha_i+1}} \right) \Phi_{i,n}(s) \Bigg] \end{aligned} \tag{20}$$

where

$$\Phi_{i,k}(s) = \frac{\omega_{n+1}(s)}{(s-s_i)^{n+1-k}} = \frac{(s-s_0)^{\alpha_0}(s-s_1)^{\alpha_1} \ldots (s-s_r)^{\alpha_r}}{(s-s_i)^{n+1-k}} \tag{21}$$

and the numbers $a_{ij}^{(k)}$ are independent on the function $g(s)$ and on the interval $[0, l]$. Let us write

$$A = \max \{a_{ij}^{(k)}\}; \tag{22}$$

it holds, according to (16), (18), (20) and (22),

$$|H_n^{(p)}(s)| \leqq A l^{-n} \eta \sum_{i=0}^{r} \sum_{k=0}^{\alpha_i - 1} (\alpha_i - k) \, |\Phi_{i, n-k}^{(p)}(s)| \, l^k.$$

If $p > n-k$ then $\Phi_{i,n-k}^{(p)} \equiv 0$. In case $p \leqq n-k$ we get from (21) by means of the Weierstrass theorem

$$\max_{[0,l]} |\Phi_{i,n-k}^{(p)}(s)| = N_{i,n-k}^{(p)} l^{n-k-p}$$

with $N_{i,n-k}^{(p)}$ as an absolute constant. Setting $N_{i,n-k}^{(p)} = 0$ if $p > n-k$ and defining

$$C_{2p+1} = A \sum_{i=0}^{r} \sum_{k=0}^{\alpha_i - 1} (\alpha_i - k) N_{i,n-k}^{(p)}$$

we obtain

$$|H_n^{(p)}(s)| \leqq C_{2p+1} l^{-p} \eta, \qquad s \in [0, l]. \tag{23}$$

As the remainder

$$R(s) = g(s) - H_n(s) \tag{24}$$

satisfies the equations

$$R(s_i) = R'(s_i) = \cdots = R^{(\alpha_i - 1)}(s_i) = 0, \qquad i = 0, \ldots, r, \tag{25}$$

we can easily prove by induction, making use of the Rolle's theorem, that there exist points $s_j^{(p)} \in [0, l]$,

$$0 \leqq s_0^{(p)} < s_1^{(p)} < s_2^{(p)} < \cdots < s_{r_p}^{(p)} \leqq l,$$

and integers $\gamma_j \geqq 1$ such that

$$\gamma_0 + \gamma_1 + \cdots + \gamma_{r_p} = n + 1 - p, \tag{26}$$

$$Q_p(s_j^{(p)}) = Q_p'(s_j^{(p)}) = \cdots = Q_p^{(\gamma_j - 1)}(s_j^{(p)}) = 0, \qquad j = 0, \ldots, r_p, \tag{27}$$

with

$$Q_p(s) = R^{(p)}(s). \tag{28}$$

Let $P_{n-p}(s)$ be the Hermite interpolation polynomial of the degree $n-p$ of the function $Q_p(s)$ having the points $s_j^{(p)}$ as its interpolation nodes. The Eqs. (27) imply $P_{n-p}(s) \equiv 0$ and from the remainder theorem for Hermite interpolation it follows:

$$R^{(p)}(s) = \frac{Q_p^{(n+1-p)}(\tau)}{(n+1-p)!} \omega_{n+1-p}(s) = \frac{R^{(n+1)}(\tau)}{(n+1-p)!} \omega_{n+1-p}(s) \tag{29}$$

where the number $\tau \in (0, l)$ depends on $s \in [0, l]$ and where

$$\omega_{n+1-p}(s) = (s - s_0^{(p)})^{\gamma_0} (s - s_1^{(p)})^{\gamma_1} \cdots (s - s_{r_p}^{(p)})^{\gamma_{r_p}}.$$

Thus, according to (26),

$$\max_{[0,l]} \left| \frac{\omega_{n+1-p}(s)}{(n+1-p)!} \right| = C_{2p+2} l^{n+1-p} \tag{30}$$

with C_{2p+2} as an absolute constant. As, according to (14) and (24),

$$R^{(n+1)}(\tau) = g^{(n+1)}(\tau) - H_n^{(n+1)}(\tau), \qquad |g^{(n+1)}(\tau)| \leqq K_{n+1}, \qquad H_n^{(n+1)}(\tau) \equiv 0$$

we get from (29) and (30)

$$|R^{(p)}(s)| \leqq C_{2p+2} K_{n+1} l^{n+1-p}, \quad s \in [0, l]. \tag{31}$$

The relations (23), (24) and (31) imply (19).

Lemma 1. Let

$$|D^i f(P)| \leqq K, \quad |i| = n,$$

P being a point in the (x_1, x_2)-plane. Then

$$\left| \frac{\partial^n f(P)}{\partial s_1 \partial s_2 \dots \partial s_n} \right| \leqq 2^{n/2} K$$

where $\partial/\partial s_j$ denotes the derivation in the direction s_j and $s_1, \dots, s_n$ are arbitrary directions.

Lemma 2. Let s_1, s_2 be two arbitrary perpendicular directions. Let

$$\left| \frac{\partial^{i_1+i_2} f(P)}{\partial s_1^{i_1} \partial s_2^{i_2}} \right| \leqq K \quad (i_1, i_2 = 0, \dots, n;\ i_1 + i_2 = n),$$

P being a point in the (x_1, x_2)-plane. Then

$$|D^i f(P)| \leqq 2^{n/2} K, \quad |i| = n.$$

Using $n(n+1)/2$-times Schwarz's inequality we obtain the assertions of both lemmas.

By $\tilde{W}_2^{(k)}(\Omega)$ we shall denote the space consisting of functions which together with all generalized derivatives of the order k belong to $L_2(\Omega)$. The norm is given by (see, e.g., [4])

$$\|u\|^2_{\tilde{W}_2^{(k)}(\Omega)} = \|u\|^2_{L_2(\Omega)} + \sum_{|i|=k} \|D^i u\|^2_{L_2(\Omega)}. \tag{32}$$

In the following we shall often need Sobolev's lemma in this special form (see [4]):

Sobolev's lemma. Let V be a domain starlike with respect to a disc. Let $0 \leqq m < k-1$ and $u(x_1, x_2) \in \tilde{W}_2^{(k)}(V)$. Then $u \in C^{(m)}(\overline{V})$ and

$$\max_{(x_1, x_2) \in \overline{V}, |i| \leqq m} |D^i u(x_1, x_2)| \leqq C \|u\|_{\tilde{W}_2^{(k)}(V)} \tag{33}$$

where the constant C does not depend on $u(x_1, x_2)$.

4. Asymptotic Estimates of the Discretization Error

Theorems 3—5 which will be now stated enable both to construct various Hermite polynomials on a triangle and to establish the rate of convergence of the finite element method. (The finite element method is a generalized Ritz method using special trial functions. In our case of triangular elements the trial functions are polynomials on each triangle of the given triangulation.)

Theorem 3. Let the function $w(x_1, x_2)$ be continuous on a closed triangle $\overline{T}$ and have bounded derivatives of the order $4m+2$ in the interior T of $\overline{T}$:

$$|D^i w(x_1, x_2)| \leqq M_{4m+2}, \quad |i| = 4m+2, \quad (x_1, x_2) \in T. \tag{34}$$

Let

$$D^i w(P_j) = 0, \quad |i| \leqq 2m, \quad j = 1, 2, 3, \tag{35}$$

$$\frac{\partial^r w(Q_k^{(r)})}{\partial \nu^r} = 0, \quad k = 1, \ldots, 3r; \; r = 1, \ldots, m, \tag{36}$$

where P_1, P_2, P_3 are the vertices of the triangle $\overline{T}$ and $Q_1^{(r)}, \ldots, Q_{3r}^{(r)}$ points dividing the sides of $\overline{T}$ into $r+1$ equal parts. Let

$$\frac{\partial^{i_1+i_2} w(P_1)}{\partial \sigma^{i_1} \partial \tau^{i_2}} = 0 \quad (i_1, i_2 \geqq m+1; \; 2m+2 \leqq i_1 + i_2 \leqq 3m) \tag{37}$$

where P_1 is the vertex with the smallest angle α of the triangle $\overline{T}$ and $\partial/\partial\sigma$ and $\partial/\partial\tau$ denote the derivations in the directions $P_1 P_2$ and $P_1 P_3$, respectively. Then it holds on $\overline{T}$

$$|D^i w(x_1, x_2)| \leqq \frac{K}{(\sin\alpha)^{|i|}} M_{4m+2} c^{4m+2-|i|}, \quad |i| \leqq 4m, \tag{38}$$

where c is the length of the largest side of the triangle $\overline{T}$ and K is a constant independent on $\overline{T}$ and on $w(x_1, x_2)$.

Proof. The method of the proof is a generalization of that one of Theorem 3 from [6].

It follows from the assumptions of Theorem 3 that the function $w(x_1, x_2)$ belongs to $\tilde{W}_2^{(4m+2)}(T)$. Thus, according to Sobolev's lemma, the function $w(x_1, x_2)$ is $4m$-times continuously differentiable on $\overline{T}$.

Let us denote the angles at the vertices P_1, P_2 and P_3 by α, β and γ, respectively. We choose the notation of the vertices in such a way that $\alpha \leqq \beta \leqq \gamma$. The lengths of the sides $P_1 P_2$, $P_1 P_3$ and $P_2 P_3$ are denoted by c, b and a, respectively. Let us construct a triangle $P_1' P_2' P_3'$ lying inside $\overline{T}$. Let the sides of $P_1' P_2' P_3'$ be parallel to the sides of $\overline{T}$ and lie in a distance δ. Choosing δ sufficiently small it holds with respect to (35), (36):

$$|D^i w(P_j')| \leqq \varepsilon, \quad |i| \leqq 2m, \quad j = 1, 2, 3, \tag{39}$$

$$\left| \frac{\partial^r w(\overline{Q}_k^{(r)})}{\partial \nu^r} \right| \leqq \varepsilon, \quad k = 1, \ldots, 3r; \; r = 1, \ldots, m, \tag{40}$$

where $\overline{Q}_1^{(r)}, \ldots, \overline{Q}_{3r}^{(r)}$ are points dividing the sides of the triangle $P_1' P_2' P_3'$ into $r+1$ equal parts. We consider the functions

$$g_k = \frac{\partial^k w}{\partial \nu^k}\bigg|_{P_1' P_2'} \quad (k = 0, 1, \ldots, m)$$

as functions of a parameter $s \in [0, c']$, c' being the length of $P_1' P_2'$. It follows from (34) and from Lemma 1 that

$$|g_k^{(4m+2-k)}(s)| \leqq 2^{2m+1} M_{4m+2}, \quad s \in [0, c'], \quad k = 0, 1, \ldots, m. \tag{41}$$

According to (39), (40) and Lemma 1

$$\begin{gathered} |g_k^{(j)}(0)| \leqq 2^{(j+k)/2} \varepsilon, \quad |g_k^{(j)}(c')| \leqq 2^{(j+k)/2} \varepsilon, \quad \left| g_r\left(\frac{q\,c'}{r+1}\right) \right| \leqq \varepsilon, \\ j = 0, \ldots, 2m-k; \quad k = 0, \ldots, m; \quad q = 1, \ldots, r; \quad r = 1, \ldots, m. \end{gathered} \tag{42}$$

Using Theorem 2 it follows from (41), (42) that we have on $P_1' P_2'$:

$$\left|\frac{\partial^{j+k} w}{\partial s^j \partial \nu^k}\right| \leqq K_{jk}\eta_k (c')^{-j} + B_{jk} M_{4m+2} c'^{4m+2-j-k}$$

where $\partial w/\partial s$ is the derivative in the direction $P_1 P_2$ and $j = 0, \ldots, m-k$, $k = 0, \ldots, m$. The symbols K_{jk}, B_{jk} denote absolute constants (i.e. constants independent on functions considered, on the interval and on the triangle $\overline{T}$) and

$$\eta_k \to 0 \quad \text{if} \quad \varepsilon \to 0+.$$

Letting $\varepsilon \to 0+$ we can see that on the side $P_1 P_2$

$$\left|\frac{\partial^{j+k} w}{\partial s^j \partial \nu^k}\right| \leqq B_{jk} M_{4m+2} c^{4m+2-j-k} \quad (j = 0, \ldots, m-k;\ k = 0, \ldots, m).$$

As $c \geqq b \geqq a$ the same estimates are true for the other sides. Let

$$A_l = 2^{l/2} \max\{B_{jk}\}, \quad j + k = l, \quad l = 0, \ldots, m.$$

Making use of Lemma 2 we get

$$|D^i w(P)| \leqq A_l M_{4m+2} c^{4m+2-l}, \quad |i| = l;\ l = 0, \ldots, m;\ P \in \overline{T} \setminus T. \tag{43}$$

Let the coordinates of the vertex P_1 be $(\bar{x}_1, \bar{x}_2)$; let (σ_1, σ_2) and (τ_1, τ_2) be the unit vectors which are parallel to the directions $P_1 P_2$ and $P_1 P_3$, respectively. Let $\overline{T}_1$ be the closed triangle in the (ξ_1, ξ_2)-plane with vertices $R_1 = (0, 0)$, $R_2 = (1, 0)$ and $R_3 = (0, 1)$. $\overline{T}_1$ is mapped one-to-one on $\overline{T}$ by the transformation

$$\begin{aligned} x_1 &= x_1(\xi_1, \xi_2) \equiv \bar{x}_1 + c\sigma_1 \xi_1 + b\tau_1 \xi_2, \\ x_2 &= x_2(\xi_1, \xi_2) \equiv \bar{x}_2 + c\sigma_2 \xi_1 + b\tau_2 \xi_2. \end{aligned} \tag{44}$$

We shall distinguish two cases: $M_{4m+2} > 0$ and $M_{4m+2} = 0$.

In the case $M_{4m+2} > 0$ let us introduce the function

$$v(\xi_1, \xi_2) = M_{4m+2}^{-1} c^{-4m-2} w\big(x_1(\xi_1, \xi_2), x_2(\xi_1, \xi_2)\big). \tag{45}$$

As

$$\frac{\partial^{i_1+i_2} v(\xi_1, \xi_2)}{\partial \xi_1^{i_1} \partial \xi_2^{i_2}} = \frac{c^{i_1} b^{i_2}}{M_{4m+2} c^{4m+2}} \frac{\partial^{i_1+i_2} w(x_1, x_2)}{\partial \sigma^{i_1} \partial \tau^{i_2}} \tag{46}$$

we can see that it holds, according to Lemma 1,

$$|D^i v(\xi_1, \xi_2)| \leqq 2^{2m+1}, \quad |i| = 4m+2, \quad (\xi_1, \xi_2) \in T_1, \tag{47}$$

$$D^i v(R_j) = 0, \qquad |i| \leqq 2m, \quad j = 1, 2, 3, \tag{48}$$

$$|D^i v(R)| \leqq 2^{l/2} A_l, \quad |i| = l;\ l = 0, \ldots, m;\ R \in \overline{T}_1 \setminus T_1, \tag{49}$$

where T_1 is the interior of $\overline{T}_1$. The function $v(\xi_1, \xi_2)$, being continuous on $\overline{T}_1$ and having bounded derivatives of the order $4m+2$ in T_1, belongs to $\tilde{W}_2^{(4m+2)}(T_1)$. Thus, according to Sobolev's lemma, $v(\xi_1, \xi_2)$ is $4m$-times continuously differentiable on $\overline{T}_1$.

Let us consider a triangle $R_1' R_2' R_3'$ lying inside $\overline{T}_1$ with sides parallel to the sides of $\overline{T}_1$ in a distance δ. Choosing δ sufficiently small it holds with respect

to (48), (49)

$$|D^i v(R_j')| \leqq \varepsilon, \quad |i| \leqq 2m, \quad j = 1, 2, 3, \tag{50}$$

$$\left|\frac{\partial^k v(\bar{S}_j^{(k)})}{\partial \xi_2^k}\right| \leqq \varepsilon + 2^{k/2} A_k \quad (j = 1, \ldots, k;\ k = 1, \ldots, m) \tag{51}$$

where the points $\bar{S}_1^{(k)}, \ldots, \bar{S}_k^{(k)}$ divide $R_1' R_2'$ into $k+1$ equal parts. The functions

$$g_r(\xi_1) = \frac{\partial^r v}{\partial \xi_2^r}\bigg|_{R_1' R_2'} \quad (r = 0, \ldots, m) \tag{52}$$

satisfy, according to (47), the inequalities

$$|g_r^{(4m+2-r)}(\xi_1)| \leqq 2^{2m+1}, \quad \xi_1 \in [0, l] \quad (r = 0, \ldots, m) \tag{53}$$

l being the length of $R_1' R_2'$. Making use of Theorem 2 and letting $\varepsilon \to 0+$ we obtain from (50)—(53)

$$\left|\frac{\partial^j v(\xi_1, 0)}{\partial \xi_1^{j-k} \partial \xi_2^k}\right| \leqq C_1, \quad \xi_1 \in [0, 1] \quad (j = 2m+1, \ldots, 3m;\ k = 0, \ldots, m). \tag{54}$$

(Here and in the following text $C_1, \ldots, C_{25}$ denote absolute constants.) Similarly we derive

$$\left|\frac{\partial^j v(0, \xi_2)}{\partial \xi_1^k \partial \xi_2^{j-k}}\right| \leqq C_2, \quad \xi_2 \in [0, 1] \quad (j = 2m+1, \ldots, 3m;\ k = 0, \ldots, m). \tag{55}$$

It follows from (37), (46), (48), (54) and (55)

$$|D^i v(R_1)| \leqq C_3 = \max(C_1, C_2), \quad |i| \leqq 3m. \tag{56}$$

Now, let R be an arbitrary point of T_1 and R_4 be the crossing point of the lines R_1, R and R_2, R_3. The length l of the segment $R_1 R_4$ is not greater than 1. By (47), (49), (56) and by Lemma 1 the function $g(s) = v|_{R_1 R_4}$ satisfies the following inequalities:

$$|g^{(j)}(0)| \leqq 2^{j/2} C_3, \quad |g^{(k)}(l)| \leqq 2^k A_k \quad (j = 0, \ldots, 3m;\ k = 0, \ldots, m)$$

$$|g^{(4m+2)}(s)| \leqq 2^{4m+2}, \quad s \in (0, l).$$

Using Theorem 2 we obtain

$$|v(R)| \leqq C_4, \quad R \in T_1. \tag{57}$$

The inequalities (49) and (57) give

$$|v(R)| \leqq \max(A_0, C_4), \quad R \in \bar{T}_1. \tag{58}$$

Thus, according to (32), (47) and (58),

$$\|v\|_{\tilde{W}_2^{(4m+2)}(T_1)} \leqq C_5. \tag{59}$$

Making use of Sobolev's lemma we get from (59)

$$|D^i v(\xi_1, \xi_2)| \leqq C_6, \quad |i| \leqq 4m, \quad (\xi_1, \xi_2) \in \bar{T}_1. \tag{60}$$

The inverse transformation to the transformation (44) reads

$$\xi_1 = \frac{\tau_2(x_1-\bar{x}_1)-\tau_1(x_2-\bar{x}_2)}{c(\sigma_1\tau_2-\sigma_2\tau_1)}, \quad \xi_2 = \frac{-\sigma_2(x_1-\bar{x}_1)+\sigma_1(x_2-\bar{x}_2)}{b(\sigma_1\tau_2-\sigma_2\tau_1)}. \tag{61}$$

As

$$1 \leqq c/b < 2, \quad |\sigma_1\tau_2-\sigma_2\tau_1| = \sin\alpha$$

the relations (45), (60) and (61) imply the inequality (38).

If $M_{4m+2}=0$ then the inequality

$$|D^i w(x_1, x_2)| \leqq \Delta, \quad |i| = 4m+2, \quad (x_1, x_2)\in T$$

holds for arbitrary $\Delta \geqq 0$. Repeating the preceding proof with

$$M_{4m+2} = \varepsilon > 0$$

where ε is arbitrarily small and letting $\varepsilon \to 0+$ Theorem 3 is proved.

Corollary. Let the function $f(x_1, x_2)$ be $3m$-times continuously differentiable on a closed triangle $\overline{T}$. Then the polynomial $p(x_1, x_2)$ of the degree $4m+1$ is uniquely determined by the conditions

$$D^i p(P_j) = D^i f(P_j), \quad |i| \leqq 2m, \quad j = 1, 2, 3, \tag{62}$$

$$\frac{\partial^r p(Q_k^{(r)})}{\partial \nu^r} = \frac{\partial^r f(Q_k^{(r)})}{\partial \nu^r}, \quad k = 1, \ldots, 3r; \ r = 1, \ldots, m, \tag{63}$$

$$\frac{\partial^{i_1+i_2} p(P_1)}{\partial \sigma^{i_1} \partial \tau^{i_2}} = \frac{\partial^{i_1+i_2} f(P_1)}{\partial \sigma^{i_1} \partial \tau^{i_2}}, \quad i_1, i_2 \geqq m+1; \ 2m+2 \leqq i_1+i_2 \leqq 3m, \tag{64}$$

where the meaning of the symbols P_j, $Q_k^{(r)}$ and $\partial/\partial\sigma$, $\partial/\partial\tau$ is the same as in Theorem 3. Further, if the function $f(x_1, x_2)$ has bounded derivatives of the order $4m+2$ in the interior T of $\overline{T}$,

$$|D^i f(x_1, x_2)| \leqq M_{4m+2}, \quad |i| = 4m+2, \quad (x_1, x_2)\in T, \tag{65}$$

then the inequality (38) holds with $w(x_1, x_2) = f(x_1, x_2) - p(x_1, x_2)$.

Proof. The number of the conditions (62)—(64) is

$$N = (2m+1)(4m+3)$$

which is the number of coefficients of a polynomial $p(x_1, x_2)$ of the degree $4m+1$:

$$p(x_1, x_2) = a_1 + a_2 x_1 + a_3 x_2 + a_4 x_1^2 + \cdots + a_N x_2^{4m+1}. \tag{66}$$

It is sufficient to prove that the N homogeneous equations

$$D^i p(P_j) = \frac{\partial^r p(Q_k^{(r)})}{\partial \nu^r} = \frac{\partial^{i_1+i_2} p(P_1)}{\partial \sigma^{i_1} \partial \tau^{i_2}} = 0 \tag{67}$$

have the trivial solution $a_1 = a_2 = \cdots = a_N = 0$ only.

As it holds, according to (66),

$$D^i p(x_1, x_2) \equiv 0, \quad |i| = 4m+2, \tag{68}$$

it follows from (67), (68) and from Theorem 3 that $p(x_1, x_2) \equiv 0$.

It follows from Theorem 1 that the function of the type (4) is m-times continuously differentiable on $\overline{\Omega}$ if $p_\lambda(x_1, x_2)$ is the polynomial of the degree $4m+1$ determined on the triangle $\overline{T}_\lambda \in \mathfrak{M}$ by the conditions (62)—(64). We can use, therefore, the functions of the type (4) as the trial functions in the finite element method for solving variational problems which arise from boundary value problems of elliptic equations of the order $2n$; $n \leqq m+1$. If the exact solution is continuous on $\overline{\Omega}$ and has bounded derivatives of the order $4m+2$ in Ω then making use of Theorem 3 the convergence of the method can be proved in the same way as in [6]. The rate of convergence is $O(h^{4m+2-n})$, h being the length of the largest segment of a triangulation.

The asymmetric condition (37) is not suitable in applications; we introduce, therefore, in cases $m=2$ and $m=3$ another condition.

Theorem 4. Let the function $w(x_1, x_2)$ be continuous on a closed triangle $\overline{T}$ and satisfy the assumptions (34)—(36) with $m=2$. Further, let

$$w(P_0) = 0, \tag{69}$$

P_0 being the centre of gravity of $\overline{T}$. Then on $\overline{T}$ the inequality (38) holds with $m=2$.

Proof. Inspecting the proof of Theorem 3 we can see that it remains to prove

$$\left| \frac{\partial^6 v(R_1)}{\partial \xi_1^3 \partial \xi_2^3} \right| \leqq C_7. \tag{70}$$

It follows from (45) and (69) that

$$v(R_0) = 0, \tag{71}$$

R_0 being the centre of gravity of $\overline{T}_1$. Let us consider the function $g = v|_{R_1 S_2}$, where S_2 is the midpoint of the side $R_2 R_3$. Using Theorem 2 we obtain from (47)—(49), (54), (55) and (71)

$$\left| \frac{\partial^6 v(R_1)}{\partial s^6} \right| \leqq C_8 \tag{72}$$

where $\partial/\partial s$ means the derivation in the direction $R_1 S_2$. We have, according to (54), (55),

$$\left| \frac{\partial^6 v(R_1)}{\partial \xi_1^{i_1} \partial \xi_2^{i_2}} \right| \leqq C_3, \quad i_1 \neq 3, \quad i_2 \neq 3, \quad i_1 + i_2 = 6. \tag{73}$$

The inequality (70) follows from (72), (73) with $C_7 = (11 C_3 + 2 C_8)/5$.

Theorem 5. Let the function $w(x_1, x_2)$ be continuous on a closed triangle $\overline{T}$ and satisfy the assumptions (34)—(36) with $m=3$. Further, let

$$D^i w(P_0) = 0, \quad |i| \leqq 1, \tag{74}$$

P_0 being the centre of gravity of $\overline{T}$. Then on $\overline{T}$ the inequality (38) holds with $m=3$.

Proof. Inspecting the proof of Theorem 3 we can see that it remains to prove

$$\left| \frac{\partial^8 v(R_1)}{\partial \xi_1^4 \partial \xi_2^4} \right| \leqq C_9, \quad \left| \frac{\partial^9 v(R_1)}{\partial \xi_1^5 \partial \xi_2^4} \right| \leqq C_{10}, \quad \left| \frac{\partial^9 v(R_1)}{\partial \xi_1^4 \partial \xi_2^5} \right| \leqq C_{11}. \tag{75}$$

Considering the functions $\partial^3 v/\partial\xi_1^3$, $\partial^3 v/\partial\xi_1^2\partial\xi_2$, $\partial^3 v/\partial\xi_1\partial\xi_2^2$ and $\partial^3 v/\partial\xi_2^3$ on $R_2'R_3'$, making use of Theorem 2 and letting $\varepsilon\to 0+$ we can prove that

$$\left|\frac{\partial^7 v(\xi_1, 1-\xi_1)}{\partial s^4\,\partial\xi_1^{i_1}\partial\xi_2^{i_2}}\right| \leqq C_{12}, \quad \xi_1\in[0,1], \quad i_1+i_2=3, \tag{76}$$

where $\partial/\partial s$ means the derivation in the direction R_2R_3. Using (54), (55) and (76) we get after elementary computation

$$\left|\frac{\partial^7 v(R_2)}{\partial\xi_1^k\,\partial\xi_2^{7-k}}\right| \leqq C_{13}, \quad \left|\frac{\partial^7 v(R_3)}{\partial\xi_1^{7-k}\,\partial\xi_2^k}\right| \leqq C_{14}, \quad k=0,\ldots,3. \tag{77}$$

Thus, according to (54), (55) and (77),

$$|D^i v(R_j)| \leqq C_{15}, \quad |i|=7, \quad j=1,2,3. \tag{78}$$

Let $S_1=(\frac{1}{2},0)$, $S_2=(\frac{1}{2},\frac{1}{2})$, $S_3=(0,\frac{1}{2})$ and let us consider the functions $g_1=v|_{R_1S_3}$, $g_2=v|_{R_2S_2}$ and $g_3=v|_{R_3S_1}$. Using Theorem 2 we obtain from (45), (47)—(49), (74) and (78)

$$\left|\frac{\partial^8 v(R_1)}{\partial t_1^8}\right| \leqq C_{16}, \quad \left|\frac{\partial^4 v(S_1)}{\partial t_3^4}\right| \leqq C_{17}, \quad \left|\frac{\partial^4 v(S_3)}{\partial t_2^4}\right| \leqq C_{18} \tag{79}$$

where $\partial/\partial t_1$, $\partial/\partial t_2$ and $\partial/\partial t_3$ denote the derivations in the directions R_1S_2, R_2S_3 and R_3S_1, respectively.

The first inequality (75) follows from the first inequality (79) and from (54), (55). Thus it holds

$$|D^i v(R_1)| \leqq C_{19}, \quad |i| \leqq 8. \tag{80}$$

In the same way as we have obtained the inequalities (54), (55) we can get

$$\left|\frac{\partial^4 v(S_1)}{\partial\xi_1^{4-k}\,\partial\xi_2^k}\right| \leqq C_{20}, \quad \left|\frac{\partial^4 v(S_3)}{\partial\xi_1^k\,\partial\xi_2^{4-k}}\right| \leqq C_{21}, \quad k=0,\ldots,3.$$

These inequalities and the second and third inequality (79) imply

$$\left|\frac{\partial^4 v(S_1)}{\partial\xi_2^4}\right| \leqq C_{22}, \quad \left|\frac{\partial^4 v(S_3)}{\partial\xi_1^4}\right| \leqq C_{23}. \tag{81}$$

Now we can prove the remaining two inequalities (75). Let us consider the functions $g_4=\partial^4 v/\partial\xi_2^4|_{R_1'R_3'}$ and $g_5=\partial^4 v/\partial\xi_1^4|_{R_1'R_3'}$. Using Theorem 2 and letting $\varepsilon\to 0+$ we obtain from (47), (48), (78), (80) and (81)

$$\left|\frac{\partial^9 v(\xi_1,0)}{\partial\xi_1^5\,\partial\xi_2^4}\right| \leqq C_{24}, \quad \xi_1\in[0,1], \quad \left|\frac{\partial^9 v(0,\xi_2)}{\partial\xi_1^4\,\partial\xi_2^5}\right| \leqq C_{25}, \quad \xi_2\in[0,1].$$

These inequalities imply the second and third inequality (75).

5. The Hierarchy of Interpolation Polynomials on the Triangle

In the preceding sections we have studied interpolation polynomials of the degree $4m+1$ and we have seen that they can be prescribed on the triangle in various ways. Let us give another example. Using the same method as in Section 4 we can prove that the following two types of conditions determine uniquely a

polynomial $p(x_1, x_2)$ of the 7-th degree (we denote $w = f - p$):

$$D^i w(P_j) = D^r w(P_0) = w(Q_k^{(2)}) = \frac{\partial w(Q_s^{(3)})}{\partial \nu} = 0 \tag{82}$$

where $|i| \leqq 2$, $|r| \leqq 1$, $j = 1, 2, 3$; $k = 1, \ldots, 6$; $s = 1, \ldots, 9$;

$$D^i w(P_j) = D^r w(P_0) = \frac{\partial w(Q_j^{(1)})}{\partial \nu} = 0 \tag{83}$$

where $|i| \leqq 3$, $|r| \leqq 1$, $j = 1, 2, 3$.

Both conditions (82) and (83) generate a function of the type (4) which is once continuously differentiable on $\overline{\Omega}$, but the conditions (83) are in applications more effective. The reason is clear: Conditions prescribed on a side are common at the most for two triangles of the given triangulation; but conditions prescribed at a vertex are common for all triangles having this vertex common. This fact implies that the total number of the conditions (83) in any triangulation is less than the total number of the conditions (82) in the same triangulation.

If we want

1. to prescribe symmetric conditions only
2. to have the maximal number of conditions at vertices

we can get the hierarchy of interpolation polynomials which are prescribed in the following way (we denote $w = f - p$, where p is a polynomial of the n-th degree and f is a function $2m$-times continuously differentiable on $\overline{\Omega}$ if $n = 4m+1$, $4m+2$ or a function $(2m+1)$-times continuously differentiable on $\overline{\Omega}$ if $n = 4m+3$, $4m+4$):

$$D^i w(P_j) = D^r w(P_0) = \frac{\partial^l w(Q_s^{(k)})}{\partial \nu^l} = 0 \tag{84}$$

where $j = 1, 2, 3$, $s = 1, \ldots, 3k$ and for $n = 4m + \varkappa$ $(1 \leqq \varkappa \leqq 4)$ the indices i, r, l, k are determined by $(85_\varkappa)$:

$$(85_1) \quad |i| \leqq 2m, \quad |r| \leqq m-2, \quad l = k = 1, \ldots, m;$$

$$(85_2) \quad |i| \leqq 2m, \quad |r| \leqq m-1, \quad l = k-1, \quad k = 1, \ldots, m+1;$$

$$(85_3) \quad |i| \leqq 2m+1, \quad |r| \leqq m, \quad l = k = 1, \ldots, m;$$

$$(85_4) \quad |i| \leqq 2m+1, \quad |r| \leqq m+1, \quad l = k-1, \quad k = 1, \ldots, m+1.$$

Some of conditions (84), $(85_\varkappa)$ lose sense in four special cases; then we do not prescribe them. Thus if $n = 1$ we do not prescribe anything at P_0 and on the sides; if $n = 2$ and $n = 5$ we do not prescribe anything at P_0; if $n = 3$ we do not prescribe anything on the sides.

It follows from Theorem 1 that the polynomial of the degree $4m + \varkappa$ $(1 \leqq \varkappa \leqq 4)$ determined by the conditions (84), $(85_\varkappa)$ generates functions of the type (4) m-times continuously differentiable on $\overline{\Omega}$.

If the function $f(x_1, x_2)$ has bounded derivatives of the order $n+1$ in the interior T of $\overline{T}$,

$$|D^i f(x_1, x_2)| \leqq M_{n+1}, \quad |i| = n+1, \quad (x_1, x_2) \in T, \tag{86}$$

we can prove that it holds on $\overline{T}$

$$|D^i w(x_1, x_2)| \leqq \frac{K}{(\sin \alpha)^{|i|}} M_{n+1} c^{n+1-|i|}, \qquad |i| \leqq n-1. \tag{87}$$

The inequality (87) is proved for $n = 2, 3, 5$ in [6] and for $n = 9, 13$ in Theorems 4, 5. For arbitrary n the inequality (87) can be proved by generalizing the method of the proof of Theorems 4, 5. This will be presented in a subsequent paper.

References

1. Ahlin, A. C.: A bivariate generalization of Hermite's interpolation formula. Math. Comp. **18**, 264–273 (1964).
2. Berezin, I. S., Židkov, N. P.: Computing methods, vol. I. English translation. Oxford: Pergamon Press 1965.
3. Birkhoff, G., Schultz, M. H., Varga, R. S.: Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. Numer. Math. **11**, 232–256 (1968).
4. Smirnov, V. I.: A course in higher mathematics, vol. V. English translation. Oxford: Pergamon Press 1964.
5. Synge, J. L.: The hypercircle in mathematical physics, pp. 209–213. London: Cambridge Univ. Press 1957.
6. Zlámal, M.: On the finite element method. Numer. Math. **12**, 394–409 (1968).

Dr. Alexander Ženíšek
Technical University
Sady Osvobození 4
Brno, Czechoslovakia

Numer. Math. 15, 297–305 (1970)
© by Springer-Verlag 1970

Numerical Solution of Integral Equations*

FRIEDMANN W. STALLMANN

Oak Ridge National Laboratory and University of Tennessee

Received April 29, 1968

Summary. A new method for the solution of integral equations is presented. The method is based on direct approximation of Dirac's delta operator by linear combination of integral operators. This avoids some pitfalls which arise in more conventional numerical procedures for integral equations.

Introduction

The work presented here originated from an attempt to solve certain types of integral equations. The characteristic feature of these equations is that the integral transformation introduces a certain amount of smoothing to the transformed function with subsequent loss of information. An attempt to restore the lost information using the conventional technique leads to oscillations and numerical instabilities in the solution.

Such integral equations have been treated frequently in the recent literature, notably by Phillips [4] and Twomey [9]. Tihonov [7, 8] gives a more theoretical discussion of similar problems. In order to obtain a stable method the problem is modified in such a way that the associated linear equations become well conditioned but the solution remains essentially unchanged as long as certain smoothness conditions are fulfilled. In the following discussion a somewhat different approach is taken: Given all available information, and that includes data as well as a priori information about the solution, what are the most likely numerical values of the solution? Based on this question a practical method is developed. Interestingly, a common feature of the methods of Phillips and of Tihonov appear also in this method: The use of a parameter which stabilizes the numerical procedure. However, the new method is not restricted to ill conditioned problems but can conveniently applied to a very broad class of operator equations.

Method

For simplicity we shall consider here only integral equations of the first kind. Extension of the idea to more general linear operator equations poses no difficulties. Let

$$F(x) = \int_{\alpha}^{\beta} (x, \tau) f(\tau) \, d\tau . \tag{1}$$

* Research sponsored by the U. S. Atomic Energy Commission under contract with the Union Carbide Corporation.

The function $F(x)$ is obtained at a discrete set of points, $x=\xi_i$, $i=1, n$, and errors ε_i are introduced in the process. That is, the values

$$\tilde{F}_i = F(\xi_i) + \varepsilon_i \tag{2}$$

are given and from this information the best approximation of $f(t)$ for given values t should be found. Since the inverse of a linear operator is also linear, it is natural to set

$$\tilde{f}(t) = f(t) + \eta(t) = \sum_{i=1}^{n} a_i \tilde{F}_i. \tag{3}$$

The error term $\eta(t)$ is, for fixed t, a linear functional in f and ε_i; using Dirac's δ-function it can be written as

$$\eta(t) = \int_{\alpha}^{\beta} \left[-\delta(t-\tau) + \sum_{i=1}^{n} a_i K(\xi_i, \tau) \right] f(\tau)\, d\tau + \sum_{i=1}^{n} a_i \varepsilon_i = E_t(f, \varepsilon). \tag{4}$$

Thus the problem of solving (1) can be considered as a problem of minimizing the linear functional E_t.

If a scalar product is defined for the linear functionals, the standard least squares method can be applied for minimization. Let f be some linear space of admissible functions $f(t)$, R^n the n-dimensional space over the reals in which the vector (ε_i) is located, and $\Omega = f \oplus R^n$ the direct sum of the two spaces. We define formally as functionals on Ω

$$\begin{aligned} K_i(f, \varepsilon) &= \int_{\alpha}^{\beta} K(\xi_i, \tau) f(\tau)\, d\tau \\ D_t(f, \varepsilon) &= f(t) \\ \Delta_i(f, \varepsilon) &= \varepsilon_i. \end{aligned} \tag{5}$$

Thus (4) can be written as

$$E_t = -D_t + \sum_{i=1}^{n} a_i (K_i + \Delta_i). \tag{6}$$

We assume that the functionals (5) belong to a scalar product space, so that the norm of E_t is given by

$$\|E_t\| = (E_t, E_t)^{\frac{1}{2}}. \tag{7}$$

By means of Schwarz inequality the error bound

$$|\eta(t)| \leqq \|E_t\| \cdot \|f, \varepsilon\|_c \tag{8}$$

is obtained, where $\|f, \varepsilon\|_c$ is the norm conjugate to $\|E_t\|$ (see Collatz [1], p. 136).

Minimizing $\|E_t\|$ leads to the set of normal equations

$$\sum_{i=1}^{n} a_i (K_i + \Delta_i, K_j + \Delta_j) = (D_t, K_j + \Delta_j), \quad j = 1, n, \tag{9}$$

for determining the coefficients a_i.

There is a wide range of choices for a scalar product of the linear functionals (5). It gives the opportunity to make use of a priori information about f and ε. This is especially simple if f and ε have a known Gaussian probability distribution in Ω. A probability distribution in Ω is Gaussian if for a finite set of linear functionals L_i, which map Ω onto the reals, the random variables $\lambda_i = L_i(f, \varepsilon)$ have a multivariate normal distribution $N(0, \sigma(\lambda_i, \lambda_j))$. (We assume, that the mean value is 0; if not, the mean value of η would not be 0 and the variance of η could be reduced by subtracting this amount from η.) The covariance matrix defines a scalar product

$$\sigma(\lambda_i, \lambda_j) = (L_i, L_j) \tag{10}$$

and the solution of the normal equation (9) leads to an error $\eta(t)$ of minimum variance.

Unfortunately a probability distribution for f is rarely available. The ε_i are usually independently distributed with known variance σ_i^2. This leads to

$$\begin{aligned} &(\Delta_i, K_j) = (\Delta_i, D_t) = 0, \\ &(\Delta_i, \Delta_j) = \sigma_i^2 \delta_{ij}, \quad \text{for all } i, j, t. \quad (\delta_{ij} \text{ is the Kronecker symbol.}) \end{aligned} \tag{11}$$

Eq. (9) becomes:

$$\sum_{i=1}^{n} a_i [(K_i, K_j) + \sigma_i^2 \delta_{ij}] = (D_t, K_j). \tag{12}$$

For lack of a probability distribution on f the selection of (K_i, K_j) and (K_i, D_t) must be guided by the error bound (8). In order to utilize this bound one has to have some a priori bound for the conjugate norm $\|f, \varepsilon\|_c$. This norm will also be a scalar product norm, this time on the space Ω. Neglecting ε, the simplest norm of this type would be

$$\|f, 0\|_c = \int_\alpha^\beta (f(t))^2 \, dt.$$

This norm does not lead to a finite norm for D_t, which means that no error bound can be given if $f(t)$ is just L_2 in (α, β). We obtain, however, finite bounds, if the derivatives are L_2, i.e.

$$\|f, 0\|_c = \int_\alpha^\beta (f^{(k)}(t))^2 \, dt$$

for some number $k > 0$. The method used is similar to the one Sard [6] has proposed for obtaining best quadrature and interpolation formulas.

We shall discuss here only the case $k = 1$. Consider a linear functional of the form

$$L(f) = \int_\alpha^\beta L(\tau) f(\tau) \, d\tau. \tag{13}$$

Assume that

$$L^{[1]}(\tau) = -\int_\alpha^\tau L(\theta) \, d\theta \tag{14}$$

is bounded and that

$$L^{[1]}(\beta) = 0. \tag{15}$$

Then

$$L(f) = \int_{\alpha}^{\beta} L^{[1]}(\tau)\, f'(\tau)\, d\tau. \tag{16}$$

If a scalar product between two such functionals L_1, L_2 is defined by

$$(L_1, L_2) = \gamma^{-2} \int_{\alpha}^{\beta} L_1^{[1]}(\tau)\, L_2^{[1]}(\tau)\, d\tau \tag{17}$$

the conjugate norm (neglecting ε) becomes

$$\|f, 0\|_c = \gamma^2 \int_{\alpha}^{\beta} (f'(\tau))^2 d\tau. \tag{18}$$

Note that $L(\tau)$ need not to be bounded in order for (17) to exist.

The functionals K_i, D_t do not in general fulfill the condition (15); but if

$$k_i = \int_{\alpha}^{\beta} K(\xi_i, \tau)\, d\tau \tag{19}$$

exists for all i, the following transformation leads to such operators. We set

$$\begin{aligned} K_i^*(\tau) &= \int_{\alpha}^{\tau} \left(\frac{K(\xi_i, \theta)}{k_i} - \frac{K(\xi_{i+1}, \theta)}{k_{i+1}} \right) d\theta, \qquad i = 1, n-1, \\ D_t^*(\tau) &= \int_{\alpha}^{\tau} \left[-\delta(t-\theta) + \frac{K(\xi_{i(t)}, \theta)}{k_{i(t)}} \right] d\theta \end{aligned} \tag{20}$$

where $i(t)$ is arbitrarily chosen among $i = 1, n$. We modify further

$$\tilde{f}(t) = f(t) + \eta(t) = \frac{\tilde{F}_{i(t)}}{k_{i(t)}} + \sum_{i=1}^{n-1} b_i \left(\frac{\tilde{F}_{i+1}}{k_{i+1}} - \frac{\tilde{F}_i}{k_i} \right); \tag{21}$$

that is the a_i in (3) are restricted in such a manner that the solution becomes exact if $f(t)$ is a constant and no errors ε_i are present. Then $\eta(t)$ can be represented as

$$\begin{aligned} \eta(t) &= \int_{\alpha}^{\beta} \left(-D_t^*(\tau) + \sum_{i=1}^{n-1} b_i K_i^*(\tau) \right) f'(\tau)\, d\tau \\ &\quad + \frac{\varepsilon_{i(t)}}{k_{i(t)}} + \sum_{i=1}^{n-1} b_i \left(\frac{\varepsilon_{i+1}}{k_{i+1}} - \frac{\varepsilon_i}{k_i} \right) = E_t(f, \varepsilon); \end{aligned} \tag{22}$$

that is

$$E_t = -\hat{D}_t + \frac{\Delta_{i(t)}}{k_{i(t)}} + \sum_{i=1}^{n-1} b_i \left(\hat{K}_i + \frac{\Delta_{i+1}}{k_{i+1}} - \frac{\Delta_i}{k_i} \right)$$

with

$$\hat{K}_i(f, \varepsilon) = \int_\alpha^\beta K_i^*(\tau) f'(\tau)\, d\tau, \quad \text{etc.}$$

The normal equations (9) become

$$\sum_{i=1}^{n-1} b_i \left[(\hat{K}_i, \hat{K}_j) + \delta_{i,j}\left(\frac{\sigma_{i+1}^2}{k_{i+1}^2} + \frac{\sigma_i^2}{k_i^2}\right) - \delta_{i+1,j}\frac{\sigma_{i+1}^2}{k_{i+1}^2} - \delta_{i,j+1}\frac{\sigma_i^2}{k_i^2} \right] \tag{23}$$
$$= (\hat{D}_t, \hat{K}_j) + (\delta_{i(t),j} - \delta_{i(t),j+1})\frac{\sigma_{i(t)}^2}{k_{i(t)}^2}, \qquad j = 1, n-1.$$

If we define the scalar product as in (17) the conjugate norm $\|f, \varepsilon\|_c$ in (8) assumes the form

$$\|f, \varepsilon\|_c = \gamma^2 \int_\alpha^\beta (f'(\tau))^2 d\tau + \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\sigma_i^2}. \tag{24}$$

The factor γ^2 regulates the relation between errors due to incorrect approximation and errors ε_i due to incorrect input values. A small value of γ^2 puts all the emphasis on good approximation, neglecting input errors whereas a large value of γ^2 smoothes the oscillation due to input errors loosing the higher frequencies of the solution in the process. Both errors should have about the same order of magnitude but the value of γ^2 which leads to this condition is difficult to determine. However, a few trials will show which values of γ^2 give a reasonable solution. The parameter γ^2 is closely related to the same parameter in the paper of Phillips [2]. Repeating the above transformation leads to norms, which contain higher derivatives of f.

Discussion

One advantage of the proposed method is that the approximation is aimed directly at the solution and not at the integral transformation, as it is done in more conventional methods. This becomes important in ill-conditioned or improperly posed problems. In such problems even a slight change, such as replacing the integral operator by a finite linear transformation, can lead to an unforseeable change in the solution. The same is true for small changes in input values, so that it becomes imperative to consider input errors, regardless how small they may be.

Of course, in improperly posed problems no reasonable solution can be obtained unless the solution is restricted in a suitable manner. Such restrictions are given in a most concise form as bounds for the norm $\|f, \varepsilon\|_c$. By proper selection of scalar products the method can be adapted to utilize different forms of restrictions. In properly posed, well-conditioned problems the inverse of a linear operator has a finite norm. If reasonable bounds of this norm are known, bounds for the norm of f can be obtained in terms of the norm of F.

The only numerical instability one may encounter in this method originates in the normal equations. Such instabilities are not primarily due to an ill-conditioned integral equation. Rather it signifies superfluous information in the input

values F_i. In such cases some of the input values could be eliminated without increasing the norm of the error functional E_t by a significant amount. In the presence of large errors ε_i the "superfluous $\tilde{F}_i$" would be needed to reduce the variance of the error. This becomes apparent in Eq. (12); large values of σ_i^2 "stabilize" the normal equation, so that more points $\tilde{F}_i$ can be included. (For Literature on ill-conditioned normal equations see [5].)

Another advantage of this method is its flexibility. There is complete freedom in the selection of the number and location of the points ξ_i and t. (Of course, sufficient information must be supplied in order to get an accurate solution.) There is also no need to treat in a special manner equations with singular or near singular kernels. The only requirement is the existence of the scalar products (17). These integrals do not contain unknown functions and can frequently be evaluated in closed form.

The price to be paid for this is an increased amount of computation. This includes the evaluation of the scalar products and a separate solution of the normal equations (23) for each different value of t. In special cases savings can be made. If the kernel is of the form $K(x, t) = K(x-t)$ and the ξ_i are equally spaced only one set of coefficients is necessary. Also, if only a few sets of $\tilde{F}_i$ need to be processed $\tilde{f}(t)$ can be computed directly by adding $\sum a_i \tilde{F}_i - y = 0$ to the normal equations and solving for y alone, (this eliminates the resubstitution with some savings in computer time).

It may be noted that the coefficients a_i are really functions of t, $a_i(t)$. A close inspection of Eq. (12) or (23) reveals that the $a_i(t)$ are linear combinations of the terms (D_t, K_j) or $(\hat{D}_t, \hat{K}_j)$ respectively, the latter set augmented by a constant. Thus the solutions obtained by this method come from a finite linear space spanned by these scalar products, considered as functions of t.

Numerical Examples

The method described in the previous section was originally developed in order to correct experimentally observed small-angle scattering of x-rays for smearing effects caused by the finite size of the slits which comprise the collimation system. It has been shown in [2] that the observed intensity $F(\theta)$, measured as a function of the scattering angle θ, can be related to the intensity $I(\theta)$ which would be observed with infinitely small slits (i.e. no instrumental smearing) by the sequence of integral equations

$$\Phi(\theta) = \int_{-\infty}^{+\infty} I\left(\sqrt{\theta^2 + u^2}\right) W(u)\, du, \tag{25}$$

$$F(x) = \int_{-\infty}^{+\infty} K(x-\theta)\, \Phi(\theta)\, d\theta. \tag{26}$$

In Eqs. (25) and (26) $W(u)$ and $K(x-\theta)$ are the slit-length and slit-width weighting functions respectively. These functions are obtained either by experimental measurement or by direct computation from known slit geometry [2, 3]. $\Phi(\theta)$ is an intermediate function, the intensity which would be observed if the slits were

infinitely thin. The two equations can be rewritten as

$$F(x) = \int_{-\infty}^{+\infty} I(\tau)\, G(x, \tau)\, d\tau \tag{27}$$

with

$$G(x, \tau) = 2 \int_{-\infty}^{+\infty} \tilde{W}(\tau, \theta)\, K(x - \theta)\, d\theta$$

and

$$\tilde{W}(\tau, \theta) = \frac{\tau W(\sqrt{\tau^2 - \theta^2})}{\sqrt{\tau^2 - \theta^2}}, \quad |\theta| < \tau$$
$$= 0, \quad |\theta| \geqq \tau.$$

Eq. (27) is the composition of two equations of different types; (25) is Abel's integral equation and (26) has a translation invariant kernel. Both can be handled simultaneously. Integration of $G(x, \theta)$ gives

$$\int_{-\infty}^{\tau} G(x, \theta)\, d\theta = 2 \int_{-\infty}^{+\infty} \tilde{W}^{[1]}(\tau, \theta)\, K(x - \theta)\, d\theta$$

$$W^{[1]}(\tau, \theta) = \int_{0}^{\sqrt{\tau^2 - \theta^2}} W(u)\, du, \quad |\theta| < \tau \tag{28}$$
$$= 0, \quad |\theta| \geqq \tau.$$

The weighting functions $W(u)$ and $K(u - \theta)$ vanish outside a finite interval. Thus the integral

$$\int_{-\infty}^{+\infty} G(x, \theta)\, d\theta = k$$

exists and is independent of x. Therefore the terms (20) can be formed and the Eqs. (23) solved.

In order to assess the accuracy of the method experimental data were simulated using various known intensity distributions $I(\tau)$ which were transformed numerically by (25), (26). The functions $W(u)$ and $K(x - \theta)$ are chosen according to the slit geometry of the instrument used in these experiments. Both are piecewise second degree polynomials with continuous first derivative. $K(x)$ approximates a Gaussian which has a σ of about 2 units; $W(u)$ is constant over the range of interest, $-60 \leqq u \leqq +60$, and goes from there rapidly to 0. The spacing of the data points ξ_i conform also to typical experiments. Random errors were added after the transformation (25), (26) in order to simulate errors which arise from the counting statistic. The method was then applied to these values and the results compared with the original values of $I(\tau)$. The parameter γ as well as the size of the random errors were varied in order to assess the influence of these parameters.

As expected the larger γ the smoother the solution, which leads to loss of detail in highly peaked curves but it also supresses oscillations due to errors in the input values. We found also, that the influence of input errors is less pronounced in highly peaked curves. This may be due to the fact that such curves give rise to smaller values $F(x)$ and errors were choosen relative to these values.

It suggests that the parameter γ should be chosen differently for different types of intensity distribution $I(t)$. It would be desirable to obtain a realistic error estimate from formula (8). Attempts to do this have not yet been successful. A typical result is shown in Fig. 1.

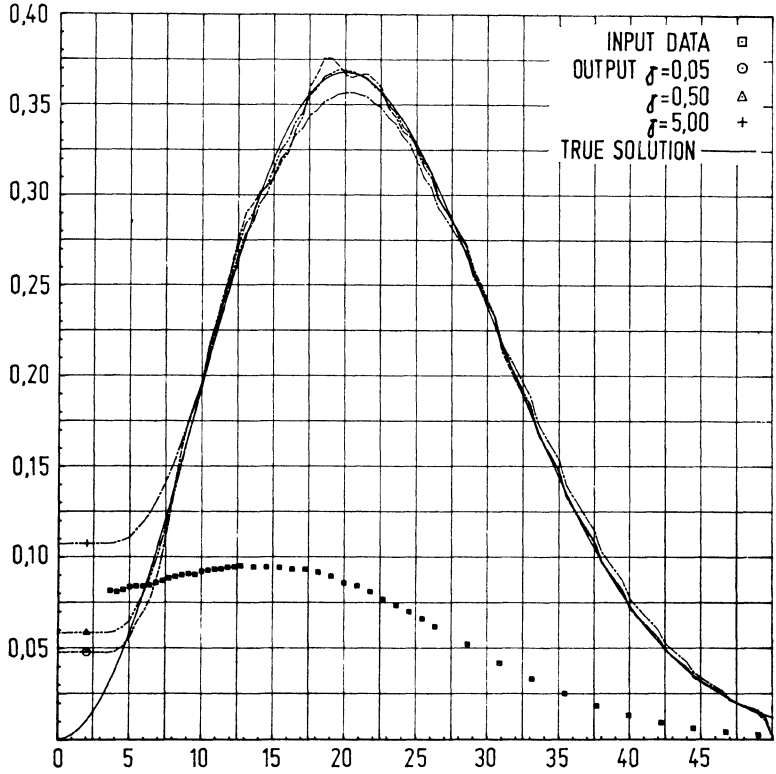

Fig. 1. This graph shows the influence of the parameter γ on the solution of the integral equation. The small squares stand for the input values $\tilde{F}_i$ on their respective locations with a random error ε_i of 2.5%. The dotted lines are the solutions computed from these $\tilde{F}_i$ with different values of $\gamma=0.05$, 0.5, 5.0. (The normal equations (23) are scaled in such a way, that the average of the diagonal terms $(\hat{K}_i, \hat{K}_i)$ is 1.) The smooth line gives the true value of $f(t)$. As shown, a value of γ which is too small introduces oscillations due to random errors; if γ is too large the height of the peak is diminished, such that the value $\gamma=0.5$ is a good compromise. Note that the input values start with $\xi_1=3.7$, so that the solutions cannot be correctly obtained for $t<3.7$

References

1. Collatz, Lothar: Functionalanalysis und Numerische Mathematik. Berlin-Göttingen-Heidelberg: Springer 1964.
2. Hendricks, R. W., Schmidt, P. W.: Calculation of weighting functions for collimation corrections in small angle scattering. Acta Physica Austriaca **26**, 97–122 (1967).
3. Buchanan, M. G., Hendricks, R. W.: A FORTRAN program for evaluation of weighting functions used in small angle scattering. ORNL-TM-1950 (1967).

4. Phillips, David L.: A technique for the numerical solution of certain integral equations of the first kind. J. ACM **9**, 84–97 (1962).
5. Rutishauser, H.: Once again: The least square problem. Linear Algebra and Applications **1**, 479–488 (1968).
6. Sard, Arthur: Linear approximation. American Mathematical Society, Providence, R. I., 1963.
7. Tihonov, A. N.: Solution of incorrectly formulated problems and the regularization method. Dokl. Akad. Nauk SSSR **151**, 501–504 (1963) — Soviet Math. Dokl. **4**, 1935–1938 (1963).
8. — Regularization of incorrectly posed problems. Dokl. Acad. Nauk SSSR 153 (1963) — Soviet Math. Dokl. **4**, 1624–1627 (1963).
9. Twomey, S.: On the numerical solution of Fredholm integral equations of the first kind by the inversion of the linear system produced by quadrature. J. ACM **10**, 97–101 (1963).

Dr. Friedmann W. Stallmann
The University of Tennessee
Mathematics Department
Ayres Hall
Knoxville 37916 USA

Numer. Math. 15, 306–314 (1970)
© by Springer-Verlag 1970

On Samelson's Iteration for Factoring Polynomials

G. W. STEWART*

Received November 7, 1969

1. Introduction

Let

$$f(z) = a_0 + a_1 z + \cdots + z^n = (z - r_1)(z - r_2) \cdots (z - r_n)$$

be a polynomial with zeros $r_1, r_2, \ldots, r_n$ over the field of complex numbers. Let

$$u(z) = (z - r_1)(z - r_2) \ldots (z - r_m)$$

and

$$v(z) = (z - r_{m+1})(z - r_{m+2}) \ldots (z - r_n)$$

be factors of f and let p and q be monic polynomials of degree m and $n - m$ approximating u and v respectively. Samelson's iteration [5] defines improved approximations p^* and q^* by the formula

$$p^* q + q^* p = f + p q. \tag{1.1}$$

If p and q are relatively prime, p^* and q^* are uniquely defined by (1.1). If u and v are relatively prime and p and q are sufficiently near u and v, the iteration defined by (1.1) converges quadratically to u and v.

Samelson's iteration was discovered independently by Stewart [7], who characterized pq^* as the linear combination of $f, q, zq, \ldots, z^{m-1}q$ that is divisible by p, a similar characterization holding for qp^*. The exact connection between these characterizations is given in [3]. Householder and Stewart [4] have derived the iteration in a way that indicates its close relation to elimination methods such as Bauer's treppen-iteration [1].

The purpose of this paper is to give another derivation of Samelson's method and to establish error bounds for the iteration. The derivation is based on a theorem of general applicability to elimination methods. The error bounds use only information that can be calculated in the course of the iteration and can in principle be used to determine when to accept the current iterate. Preliminary to the derivation of the error bounds, a generalization of the Lagrange interpolation formula will be established.

* Work on this paper was supported in part by NSF Grant GP-8442 and U.S. Army Grant DA-ARO(D)-31-124-G1050 at The University of Texas at Austin.

The notation of [7] will be used here. In particular, F_f denotes the companion matrix of the polynomial f:

$$F_f = \begin{pmatrix} 0 & 0 & \dots & 0 & -a_0 \\ 1 & 0 & \dots & 0 & -a_1 \\ 0 & 1 & \dots & 0 & -a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -a_{n-1} \end{pmatrix}.$$

The vector

$$\tilde{g} = (b_0, b_1, \dots, b_{n-1})^T$$

denotes the vector of coefficients of the polynomial

$$g(z) = b_0 + b_1 z + \cdots + b_{n-1} z^{n-1}.$$

The symbol $\|\cdot\|$ denotes the vector 1-norm or its subordinate matrix norm.

2. A New Derivation

In this section the zeros of f will be assumed distinct. The following theorem, from which Samelson's iteration will be derived, shows that under certain conditions an accurate approximation to v may be obtained by eliminating leading powers of z among a set of m suitably constructed polynomials.

Let g be a polynomial of degree k. Given the polynomial g_0 of degree not greater than $n-1$, form the polynomials $g_1, g_2, \dots, g_{m-1}$, also of degree not greater than $n-1$, by the formula

$$g_{i+1} = z g_i - \phi_i f. \tag{2.1}$$

The polynomial ϕ_i is at most of degree $k-1$ and is uniquely determined by the requirement that g_{i+1} be of degree less than n. If

$$\gamma_j = g(r_j), \qquad \delta_j = g_0(r_j) \quad (j = 1, 2, \dots, n),$$

then

$$g_i(r_j) = \gamma_j^i \delta_j \quad (i = 0, 1, \dots, m-1;\ j = 1, 2, \dots, n). \tag{2.2}$$

Theorem 2.1. *For $j = 1, 2, \dots, m$ let the γ_j be distinct and $\delta_j \neq 0$. Let*

$$\eta = \frac{\max\{|\delta_j| : j > m\}}{\min\{|\delta_j| : j \leqq m\}}.$$

If η is sufficiently small, there is a unique linear combination h_m of $g_0, g_1, \dots, g_{m-1}$ that is monic of degree $n-m$. Moreover,

$$h_m(z) = v(z) + O(\eta). \tag{2.3}$$

Proof. The proof is based on the Newton interpolation formula. Let

$$\boldsymbol{f}(z) = \begin{pmatrix} (z-r_2)(z-r_3)\dots(z-r_n) \\ (z-r_2)\dots(z-r_n) \\ \vdots \\ (z-r_n) \\ 1 \end{pmatrix}$$

and let R be the lower triangular matrix that transforms the row vector $(g(r_1), g(r_2), \ldots, g(r_n))$ into the row vector of divided differences $(g[r_1, r_2, \ldots, r_n], g[r_2, \ldots, r_n], \ldots, g[r_n])$. Then

$$g_i(z) = (g_i(r_1), g_i(r_2), \ldots, g_i(r_n))\, R\boldsymbol{f}(z). \tag{2.4}$$

Thus if

$$\boldsymbol{g}(z) = (g_0(z), g_1(z), \ldots, g_{m-1}(z))^T,$$
$$D = \operatorname{diag}(\delta_i),$$

and

$$C = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \gamma_1 & \gamma_2 & \cdots & \gamma_n \\ \vdots & \vdots & & \vdots \\ \gamma_1^{m-1} & \gamma_2^{m-1} & \cdots & \gamma_n^{m-1} \end{pmatrix},$$

then by (2.2) and (2.4)

$$\boldsymbol{g}(z) = CDR\boldsymbol{f}(z). \tag{2.5}$$

Eq. (2.5) may be partitioned in the form

$$\boldsymbol{g}(z) = (C_1D_1, C_2D_2)\begin{pmatrix} R_{11} & 0 \\ R_{21} & R_{22} \end{pmatrix}\begin{pmatrix} \boldsymbol{f}_1(z) \\ \boldsymbol{f}_2(z) \end{pmatrix}, \tag{2.6}$$

where C_1, D_1, and R_{11} are $m \times m$ matrices. The hypotheses of the theorem insure that $C_1D_1R_{11}$ is nonsingular, and, for η sufficiently small, $C_1D_1R_{11}$ dominates $C_2D_2R_{21}$, so that $C_1D_1R_{11} + C_2D_2R_{21}$ is also nonsingular.

Let

$$\begin{aligned}\boldsymbol{h}(z) &= (C_1D_1R_{11} + C_2D_2R_{21})^{-1}\boldsymbol{g}(z) \\ &= \boldsymbol{f}_1(z) + (C_1D_1R_{11} + C_2D_2R_{21})^{-1}C_2D_2R_{22}\boldsymbol{f}_2(z).\end{aligned}$$

The last component h_m of $\boldsymbol{h}$ is a linear combination of $g_0, g_1, \ldots, g_{m-1}$. It has the form

$$h_m = v + e,$$

where e is of degree less than $n-m$. Since $e(z)$ is the m-th component of $(C_1D_1R_{11} + C_2D_2R_{21})^{-1}C_2D_2R_{22}f_2(z)$, $e(z) = O(\eta)$, which establishes (2.3). The uniqueness of h_m follows from the fact that if $\alpha_0 g_0 + \alpha_1 g_1 + \cdots + \alpha_{m-1}g_{m-1}$ is of degree $n-m$, then by (2.6)

$$(\alpha_0, \alpha_1, \ldots, \alpha_{m-1})(C_1D_1R_{11} + C_2D_2R_{21}) = (0, 0, \ldots, \beta).$$

Since $C_1D_1R_{11} + C_2D_2R_{21}$ is nonsingular, the α_j are uniquely determined by the leading coefficient of h_m.

Bauer's treppen-iteration, computational points aside, may be derived from this theorem. Suppose that

$$|r_1| \geqq |r_2| \geqq \cdots \geqq |r_m| > |r_{m+1}| \geqq \cdots \geqq |r_n|.$$

Given the polynomial $g_0^{(0)}$, of degree less than n, form the sequence of polynomials $g_0^{(0)}, g_0^{(1)}, g_0^{(2)}, \ldots$ by the formula

$$g_0^{(k+1)}(z) = z\, g_0^{(k)}(z) - \alpha_k f(z) \tag{2.7}$$

where α_k is chosen so that $g_0^{(k+1)}$ is of degree less than n. Take $g(z) \equiv z$ and for each k form the polynomial $h_m^{(k)}$ of Theorem 2.1 starting with $g_0 = g_0^{(k)}$. If $g_0^{(0)}(r_i) \neq 0$ $(i = 1, 2, \ldots, m+1)$, the ratio η_k associated with $g_0^{(k)}$ is $O(|r_{m+1}/r_m|^k)$. Hence by Theorem 2.1,

$$h_m^{(k)}(z) = v(z) + O(|r_{m+1}/r_m|^k).$$

In practice, the $h_m^{(k)}$ are not formed directly by elimination among $g_0^{(k)}, g_1^{(k)}, \ldots, g_m^{(k)}$.

The treppen-iteration uses the recursion (2.7) to obtain a favorable value of η. Samelson's iteration accomplishes the same thing in a different way. Let g_0 be the polynomial of degree less than n satisfying

$$g_0(r_i) = \frac{q(r_i)}{p(r_i)}, \quad (i = 1, 2, \ldots, n). \tag{2.8}$$

Then η will be small in proportion as the zeros of p and q are near those of u and v respectively. Hence, if $g(z) \equiv z$ and h_m is formed as in Theorem 2.1, h_m should be a better approximate factor than q.

In fact, h_m is the q^* of Samelson's iteration. To see this, note that from (2.1) and the definition of h_m

$$h_m = \psi f - \theta g_0,$$

where ψ is of degree $m-2$ and θ is of degree $m-1$. Hence from (2.8)

$$p(r_i) h_m(r_i) = \theta(r_i) q(r_i), \quad (i = 1, 2, \ldots, n).$$

But θq is of degree $n-1$ and is uniquely determined by its values at r_i; hence

$$\theta q = p h_m - f.$$

Thus $p h_m = \theta q + f$ is the linear combination of $f(z), q(z), z q(z), \ldots, z^{m-1} q(z)$ that is divisible by p, which is the characterization of q^* given in §1.

Samelson notes that methods of higher order of convergence may be obtained by shifting the origin and applying a fixed number of steps of a "definite" iteration such as Bernoulli's method, and he obtains his factorization method by generalizing this observation. The above derivation also fits this pattern. The origin shift is represented by the choice of g_0; the "definite" method is the treppen-iteration.

3. An Interpolation Formula

The following interpolation theorem was established in [7].

Theorem 3.1. *Let $r_1, r_2, \ldots, r_n$ be distinct and let h be a rational function defined at the r_i. Let the coefficients of the polynomial g be defined by*

$$\tilde{g} = h(F_l)\, e_1, \tag{3.1}$$

where e_1 is the vector whose first component is unity and whose other components are zero. Then

$$g(r_i) = h(r_i), \quad (i = 1, 2, \ldots, n).$$

Even when some of the r_i coincide, the polynomial g of the above theorem is well defined. In this case, it represents the appropriate Hermite interpolant of h over the points r_i. The following theorem generalizes Theorem 3.1.

Theorem 3.2. *Let*

$$f(z) = u_1(z)\, u_2(z) \ldots u_k(z),$$

where each u_i is a monic polynomial of degree m_i and $u_1, u_2, \ldots, u_k$ are relatively prime. Let F_i be the companion matrix of u_i and let

$$\pi_i(z) = f(z)/u_i(z).$$

Then the polynomial interpolating the rational function h at the points $r_1, r_2, \ldots, r_n$ is given by

$$g(z) = \sum_{i=1}^{k} \pi_i(z)\, (1, z, \ldots, z^{m_i})\, h(F_i)\, \pi_i(F_i)^{-1} e_1. \tag{3.2}$$

Proof. It suffices to show that $g(r_i) = h(r_i)$ when the r_i are distinct, whence it follows by a continuity argument that g and the interpolant of Theorem 2.1 are identical even when some of the r_i coincide. Let r be a zero of u_i. By Theorem 2.1,

$$g_i(z) = (1, z, \ldots, z^{m_i})\, h(F_i)\, \pi_i(F_i)^{-1} e_1$$

is the interpolant for h/π_i at the zeros of u_i. Hence

$$g_i(r) = h(r)/\pi_i(r).$$

But if $j \neq i$, $\pi_j(r) = 0$. Hence

$$g(r) = \sum_{j=1}^{k} \pi_j(r)\, g_j(r) = h(r),$$

which establishes the theorem.

For $k = 1$, the formula (3.2) reduces to the formula (3.1) of Theorem 3.1. For $k = n$, formula (3.2) is the Lagrange interpolation formula, the condition that the u_i be relatively prime being in this case equivalent to the condition that the r_i be distinct.

4. Error Bounds

Let p and q be relatively prime monic polynomials of degrees m and $n-m$ and let

$$e = p\, q - f.$$

It might be expected that if e is small then there will be factors u and v of f near p and q. Moreover, the errors

$$r = p - u, \quad s = q - v$$

should tend to zero along with e. The object of this section is to confirm these expectations by exhibiting the factors u and v and deriving bounds for r and s.

If r and s are known, then from the relation

$$f = p\,q + e = (p + r)\,(q + s) \tag{4.1}$$

it follows that

$$e(F_q)\,e_1 = [p(F_q) + r(F_q)]\,s(F_q)\,e_1 \tag{4.2}$$

and

$$e(F_p)\,e_1 = [q(F_p) + s(F_p)]\,r(F_p)\,e_1. \tag{4.3}$$

Since p and q are relatively prime, Theorem 3.2 implies that e is completely determined by $e(F_q)e_1$ and $e(F_p)e_1$. Hence, if r and s satisfy (4.2) and (4.3), they also satisfy (4.1). Thus the problem of defining the factors u and v becomes one of finding polynomials r and s that satisfy

$$\tilde{r} = (Q + S)^{-1}\tilde{e}_p \tag{4.4}$$

and

$$\tilde{s} = (P + R)^{-1}\tilde{e}_q, \tag{4.5}$$

where

$$\begin{aligned} Q &= q(F_p), & P &= p(F_q), \\ S &= s(F_p), & R &= r(F_q), \\ \tilde{e}_p &= e(F_p)\,e_1, & \tilde{e}_q &= e(F_q)\,e_1. \end{aligned}$$

The polynomials r and s will be defined as the limits of two sequences of polynomials. Let

$$r_0(z) \equiv 0, \qquad s_0(z) \equiv 0.$$

Given r_i and s_i, let

$$S_i = s_i(F_p), \qquad R_i = r_i(F_q) \tag{4.6}$$

and define r_{i+1} and s_{i+1} by the formulas

$$\tilde{r}_{i+1} = (I + Q^{-1}S_i)^{-1}Q^{-1}\tilde{e}_p$$

and

$$\tilde{s}_{i+1} = (I + P^{-1}R_i)^{-1}P^{-1}\tilde{e}_q.$$

Let

$$\varrho_1 = \|Q^{-1}\tilde{e}_p\| = \|r_1\|$$

and

$$\sigma_1 = \|P^{-1}\tilde{e}_q\| = \|s_1\|.$$

It will be shown that if ϱ_1 and σ_1 are sufficiently small then the r_i and s_i are well defined and converge to limiting polynomials r and s. By continuity, these limits must satisfy (4.4) and (4.5).

From (4.6), since r_i and s_i are of degree less than m and $n-m$ respectively,

$$\|S_i\| \leqq \|F_p\|^{n-m-1}\|\tilde{s}_i\|$$

and

$$\|R_i\| \leqq \|F_q\|^{m-1} \|\tilde{r}_i\|.$$

Let

$$K = \|Q^{-1}\| \|F_p\|^{n-m-1}, \quad L = \|P^{-1}\| \|F_q\|^{m-1}.$$

By a theorem on inverses of perturbations of the identity matrix [2, p. 54], if

$$K\|\tilde{s}_i\| < \tfrac{1}{2}, \quad L\|\tilde{r}_i\| < \tfrac{1}{2}, \tag{4.7}$$

then $\tilde{r}_{i+1}$ and $\tilde{s}_{i+1}$ are defined and satisfy

$$\|\tilde{r}_{i+1}\| \leqq (1+2K\|\tilde{s}_i\|)\,\varrho_1, \tag{4.8}$$

and

$$\|\tilde{s}_{i+1}\| \leqq (1+2L\|\tilde{r}_i\|)\,\sigma_1. \tag{4.9}$$

Let

$$\alpha = \varrho_1(1+2K\sigma_1), \quad \beta = \sigma_1(1+2L\varrho_1)$$

and

$$\gamma = 4KL\varrho_1\sigma_1.$$

Then it is easily seen from (4.8) and (4.9) that if

$$\gamma < 1 \tag{4.10}$$

and $\|\tilde{r}_i\|$ and $\|\tilde{s}_i\|$ satisfy (4.7), then

$$\|\tilde{r}_i\| \leqq \varrho_1 + \frac{\alpha}{1-\gamma} = \alpha' \tag{4.11}$$

and

$$\|\tilde{s}_i\| \leqq \sigma_1 + \frac{\beta}{1-\gamma} = \beta'. \tag{4.12}$$

Hence if ϱ_1 and σ_1 are so small that

$$2K\alpha' < 1, \quad 2L\beta' < 1 \tag{4.13}$$

then the sequences of polynomials r_i and s_i are defined and satisfy (4.11) and (4.12).

To show that the r_i and s_i converge, note that

$$\tilde{r}_{i+1} - \tilde{r}_i = -(I+Q^{-1}S_i)^{-1}(I+Q^{-1}S_{i-1})^{-1}(S_i - S_{i-1})\,Q^{-2}\tilde{e}_p$$

and

$$\tilde{s}_{i+1} - \tilde{s}_i = -(I+P^{-1}R_i)^{-1}(I+P^{-1}R_{i-1})^{-1}(R_i - R_{i-1})\,P^{-2}\tilde{e}_q.$$

Since

$$\|S_i - S_{i-1}\| \leqq \|F_p\|^{n-m-1} \|\tilde{s}_i - \tilde{s}_{i-1}\|$$

and

$$\|R_i - R_{i-1}\| \leqq \|F_q\|^{m-1} \|\tilde{s}_i - \tilde{s}_{i-1}\|,$$

it follows that

$$\|\tilde{r}_{i+1} - \tilde{r}_i\| \leqq \varrho_1 K(1+2K\beta')^2 \|\tilde{s}_i - \tilde{s}_{i-1}\|$$

and

$$\|\tilde{s}_{i+1} - \tilde{s}_i\| \leqq \sigma_1 L(1+2L\alpha')^2 \|\tilde{r}_i - \tilde{r}_{i-1}\|.$$

Hence if

$$M = K(1+2K\beta')^2, \quad N = L(1+2L\alpha')^2$$

and ϱ_i and σ_i are defined by

$$\varrho_{i+1} = M\varrho_1\sigma_i, \qquad \sigma_{i+1} = N\sigma_1\varrho_i$$

then

$$\|\tilde{r}_i - \tilde{r}_{i-1}\| \leqq \varrho_i, \qquad \|\tilde{s}_i - \tilde{s}_{i-1}\| \leqq \sigma_i.$$

Now if

$$MN\sigma_1\varrho_1 < 1, \tag{4.14}$$

then

$$\sum_{i=1}^{\infty} \varrho_i = \frac{\varrho_1(1+M\sigma_1)}{1-MN\sigma_1\varrho_1} \tag{4.15}$$

and

$$\sum_{i=1}^{\infty} \sigma_i = \frac{\sigma_1(1+N\varrho_1)}{1-MN\sigma_1\varrho_1}. \tag{4.16}$$

Hence if (4.14) is satisfied, the series

$$\sum_{i=1}^{\infty} (\tilde{r}_i - \tilde{r}_{i-1}), \qquad \sum_{i=1}^{\infty} (\tilde{s}_i - \tilde{s}_{i-1})$$

are dominated by convergent series, and the r_i and s_i converge. This establishes

Theorem 4.1. *With the above definitions, if ϱ_1 and σ_1 are so small that* (4.13) *and* (4.14) *are satisfied, then there are factors u and v of f such that*

$$\|u - p\| \leqq \alpha', \qquad \|v - q\| \leqq \beta'.$$

Since $e(F_p) = f(F_p)$ and $e(F_q) = f(F_q)$, it follows from (1.1) that the polynomials p^* and q^* of Samelson's iteration are given by

$$p^* = p + r_1, \qquad q^* = q + s_1.$$

Hence,

$$\|u - p^*\| \leqq \sum_{i=2}^{\infty} \varrho_i, \qquad \|v - p^*\| \leqq \sum_{i=2}^{\infty} \sigma_i.$$

From (4.15) and (4.16) there follows

Corollary 4.2. *Under the conditions of Theorem* 4.1,

$$\|u - p^*\| \leqq M\sigma_1\varrho_1 \frac{1+N\varrho_1}{1-MN\sigma_1\varrho_1}$$

and

$$\|v - q^*\| \leqq N\sigma_1\varrho_1 \frac{1+M\sigma_1}{1-MN\sigma_1\varrho_1}.$$

The constant K, L, M, and N in the above theorems depend on the size of the matrices P^{-1} and Q^{-1}, which become unbounded as a zero of p approaches a zero of q. Thus the theorems become weaker when a divisor of p approaches one of q.

The bounds of Corollary 4.2 depend on the product of ϱ_1 and σ_1, that is, on the product of the sizes of the corrections made by Samelson's iteration. This is in accordance with the quadratic convergence of the iteration.

Finally, note that the above error bounds are of the kind found in Kantorovitch type convergence proofs, which give bounds in terms information at the current iterate and rather general information about the iteration function. However, the above bounds may be calculated knowing only f, p, and q. The reason for the simplicity of the bounds has been given by Schröder [6]. Samelson's iteration is Newton's method applied to the system of nonlinear equations

$$f = p\,q$$

in the coefficients of p and q. It is easy to verify that the second derivative of the associated operator, upon which most Kantorovitch bounds depend, is constant.

Acknowledgement. I wish to thank Professor A. S. Householder for his comments on a preliminary draft of this paper.

References

1. Bauer, F. L.: Beiträge zur Entwicklung numerischer Verfahren für programmgesteuerte Rechenanlagen. II. Direkte Faktorisierung eines Polynoms. Bayer. Akad. Wiss., Math.-Nat. Kl., Sitzber. (1965) 163–203.
2. Householder, A. S.: The theory of matrices in numerical analysis. New York: Blaisdell 1964.
3. – Stewart, G. W.: Comments on "Some iterations for factoring polynomials." Numer. Math. **13**, 458–471 (1969).
4. – – The numerical factorization of a polynomial. Submitted to SIAM Review. 1969.
5. Samelson, Klaus: Faktorisierung von Polynomen durch funktionale Iteration. Bayer. Akad. Wiss., Math.-Nat. Kl., Abhandlungen **95** (1959).
6. Schröder, J.: Factorization of polynomials by generalized Newton procedures. In: Constructive aspects of the fundamental theorem of algebra, B. Dejon and P. Henrici (eds.). New York: John Wiley 1969.
7. Stewart, G. W.: Some iterations for factoring a polynomial.

G. W. Stewart
The University of Texas
Austin, Texas 78712

Numer. Math. 15, 315–319 (1970)
© by Springer-Verlag 1970

The Error in Polynomial Interpolation

CHRISTOPHER T. H. BAKER

Received August 2, 1968

Abstract. We discuss the nature of bounds on the error in polynomial interpolation for functions of various degrees of smoothness.

1. Notation and Preliminaries

We shall examine the error $E_n(f)(x)$ associated with approximating $f(x)$ on $[-1, 1]$ by the polynomial $p_n(x)$ of degree n which interpolates $f(x)$ at the distinct sample points $x_0 > x_1 > \cdots > x_n$ in $[x_{n+1}, x_{-1}] = [-1, 1]$. The Lagrangian form of $p_n(x)$ is $\sum_{j=0}^{n} l_j(x) f(x_j)$ where $l_j(x) = \pi(x)/\pi'(x_j)(x - x_j)$ and $\pi(x) = \prod_{i=0}^{n}(x - x_i)$. We define $\lambda_n(x) = \sum_{j=0}^{n} |l_j(x)|$.

Our results depend upon the degree of smoothness of $f(x)$, and we require the following notation. If $f(x) \in C[-1, 1]$, we denote its modulus of continuity by $\omega(f; \delta) = \sup_{|x-y| \leqq \delta} |f(x) - f(y)|$. We write $f(x) \in \mathrm{Lip}_L \alpha$ if $|f(x) - f(y)| \leqq L|x - y|^{\alpha}$, $\alpha > 0$, for $x, y \in [-1, 1]$. With $\|g\| = \sup_{-1 \leqq x \leqq 1} |g(x)|$ we set $M_k[-1, 1] = \{f(x) \mid \|f^k\| \leqq M_k\}$, and if $f(x) \in C^{k-1}[-1, 1]$ and $f^k(x)$ is bounded and piecewise continuous, we write $f(x) \in PC^k[-1, 1]$.

We shall require properties of best polynomial approximation and so we define, for $f(x) \in C[-1, 1]$, $e_n(f)(x) = f(x) - p_n^*(f)(x)$ where $p_n^*(f)(x)$ is the best approximation of degree n in the norm $\|\cdot\|$. We then have the following versions of Jackson's theorem.

Theorem 1. (a) If $f(x) \in C[-1, 1]$ and $n \geqq 1$, $\|e_n(f)\| \leqq K_1 \omega(f; n^{-1})$ for some $K_1 \leqq 4$.

(b) If $f(x) \in \mathrm{Lip}_L 1$ and $n \geqq 1$, $\|e_n(f)\| \leqq KLn^{-1}$ for some $K < \pi/2$.

(c) If $n \geqq 1$ and $f(x) \in M_1[-1, 1]$, $\|e_n(f)\| \leqq \frac{1}{2}\pi M_1/(n+1)$.

(d) If $n > k$ and $f(x) \in M_{k+1}[-1, 1]$, $\|e_n(f)\| \leqq (\frac{1}{2}\pi)^{k+1} M_{k+1}(n-k)!/(n+1)!$.

Jackson's theorem is discussed in Cheney [3], Davis [4], Natanson [6], and Timan [9]. The result $K_1 \leqq 1$ for (a) (see [5]) appears to be in error.

2. The Form of Bounds on the Error

If $f(x) \in C^{n+1}[-1, 1]$, it is well-known that $|E_n(f)(x)| = |f(x) - p_n(x)| \leqq |\pi(x)| \, \|f^{n+1}\|/(n+1)!$. We generalize this result, for example by giving various functions $\gamma_{n,k}(x)$ such that $|E_n(f)(x)| \leqq \gamma_{n,k}(x) M_k$, when $x \in [-1, 1]$ and $f(x) \in M_k[-1, 1]$, $1 \leqq k \leqq n+1$. If we know $\|\gamma_{n,k}\|$ we can bound $\|E_n(f)\|$.

Definition. We say that $\gamma_{n,k}(\bar{x})$ has Property (1) (at some $\bar{x}\in[-1,1]$) if $1\leqq k\leqq n+1$ and

$$|E_n(f)(\bar{x})|\leqq\gamma_{n,k}(\bar{x})M_k \tag{1}$$

whenever $f(x)\in M_k[-1,1]$.

Theorem 2 shows how we may construct $\gamma_{n,k+1}(x)$ if such a function $\gamma_{n,k}(x)$ is known.

Theorem 2. Suppose $1\leqq k\leqq n$ and $\gamma_{n,k}(\bar{x})$ has Property (1). Then $|E_n(f)(\bar{x})|\leqq\pi\gamma_{n,k}(\bar{x})M_{k+1}/\{2(n-k+1)\}$ whenever $f(x)\in M_{k+1}[-1,1]$.

Proof. If $q_n^k(x)=p_{n-k}^*(f^k)(x)$, $E_n(f)(\bar{x})=E_n(f-q_n)(\bar{x})\leqq\gamma_{n,k}(\bar{x})\|f^k-q_n^k\|$. Since $\|f^k-q_n^k\|=\|e_{n-k}(f^k)\|$ the result follows from Theorem 1 (c). Using Theorem 1 (a) and (b) we may also show that $|E_n(f)(\bar{x})|\leqq K_1\gamma_{n,k}(\bar{x})\omega(f^k;n^{-1})$ if $f(x)\in C^k[-1,1]$ and $|E_n(f)(\bar{x})|\leqq K\gamma_{n,k}(\bar{x})L/n$ if $f^k(x)\in\mathrm{Lip}_L 1$.

It transpires that the precise conditions on $f(x)$ necessary for (1) are important. The following result is useful.

Theorem 3. Suppose that $|E_n(f)(\bar{x})|\leqq\gamma_{n,k}(\bar{x})\|f^k\|$ for some $\bar{x}\in[-1,1]$, whenever $f(x)\in C^k[-1,1]$, $1\leqq k\leqq n+1$. Then $|E_n(f)(\bar{x})|\leqq\gamma_{n,k}(\bar{x})L$ when $f^{k-1}(x)\in\mathrm{Lip}_L 1$. In particular, $\gamma_{n,k}(\bar{x})$ has Property (1).

Proof. Suppose $f(x)\in C^{k-1}[-1,1]$. Fixing $\delta>0$, define $\phi(x)$ by setting $\phi^{(r)}(0)=f^{(r)}(0)$ for $0\leqq r\leqq k-2$, and $\phi^{k-1}(x)=\int_{x-\frac{1}{2}\delta}^{x+\frac{1}{2}\delta}f^{k-1}(t)\,dt/\delta$. (For the latter, set $f^{k-1}(t)=f^{k-1}(-1)$, $t<-1$, $f^{k-1}(t)=f^{k-1}(1)$, $t>1$.) Then $\phi(x)\in C^k[-1,1]$, $\|\phi^k\|\leqq L$, and by Taylor's theorem with integral remainder, $\|f-\phi\|\leqq\frac{1}{4}L\delta/(k-1)!$. The triangle inequality gives $|E_n(f)(\bar{x})|\leqq|f(\bar{x})-\phi(\bar{x})|+|E_n(\phi)(\bar{x})|+\left|\sum_{j=0}^{n}l_j(\bar{x})\cdot\{f(x_j)-\phi(x_j)\}\right|$. By hypothesis, $|E_n(\phi)(\bar{x})|\leqq\gamma_{n,k}(\bar{x})L$. Using the bound on $\|f-\phi\|$, $|E_n(f)(\bar{x})|\leqq\frac{1}{4}(1+\|\lambda_n\|)L\delta/(k-1)!+\gamma_{n,k}(\bar{x})L$. Since this is true for any $\delta>0$ it is also true for $\delta=0$ and thus the theorem is established.

3. Some Bounds on the Error

Since $p_n(p_n^*)(x)=p_n^*(x)$ we may obtain the bound

$$|E_n(f)(\bar{x})|\leqq(1+\lambda_n(\bar{x}))\|e_n(f)\|,\quad\text{for}\quad\bar{x}\in[-1,1] \tag{2}$$

from the triangle inequality. The corresponding bound on $\|E_n(f)\|$ is in Powell [7]; a discussion appears in [6]. Values of $\|\lambda_n\|$ have been tabulated when $x_i=t_i=\cos\left(\frac{2i+1}{n+1}\frac{\pi}{2}\right)$ (see [7]) and when $x_i=h_i=1-2i/n$ [2, 8]. From Theorem 1 (c) and (2) we deduce that $\gamma_{n,1}(x)=\frac{\pi}{2(n+1)}(1+\lambda_n(x))$ has Property (1), uniformly in x, for $k=1$. Using (2) with Theorem 1 (d) we derive functions $\gamma_{n,k}(x)$ with Property 1 for $2\leqq k\leqq n+1$. This is equivalent to obtaining $\gamma_{n,k}(x)$ from Theorem 2 and $\gamma_{n,1}(x)$. The bound (2) is unsatisfactory as $\bar{x}\to x_i$ since the right-hand side does not then approach zero. We obtain better bounds by constructing the Peano kernel of the divided difference or by using Peano theory and Kowalewski's error estimate as follows.

Theorem 4. If $f(x)\in P\,C^k[-1,1]$ where $1\leqq k\leqq n+1$, then for $\bar{x}\in[-1,1]$,

$$E_n(f)(\bar{x})=\frac{1}{(k-1)!}\sum_{j=0}^{n} l_j(\bar{x})\int_{x_j}^{\bar{x}}(x_j-t)^{k-1}f^k(t)\,dt=\int_{-1}^{1}K_k(\bar{x};t)f^k(t)\,dt/(k-1)!$$

where

$$K_k(x;t)=(x-t)_+^k-\sum_{j=0}^{n} l_j(x)\,(x_j-t)_+^k.$$

Proof. The result is a modification of the case $k=n+1$ which is treated by Davis [4, pp. 70—72].

From Theorems 2, 4 and a mean value theorem, we deduce that when $d_{n,k}(x)=\sum_{j=0}^{n}|l_j(x)(x-x_j)^k|$ and $\delta_{n,k}(x)=\int_{-1}^{1}|K_k(x;t)|\,dt$, the functions $d_{n,k}(x)/k!$ and $\delta_{n,k}(x)/(k-1)!$ have Property (1) uniformly in $[-1,1]$. If $1\leqq k\leqq n+1$ the bound (1) with $\gamma_{n,k}(x)=\delta_{n,k}(x)/(k-1)!$ is attainable for some $f(x)\in PC^k[-1,1]$.

We observe that $d_{n,k}(x_i)=\delta_{n,k}(x_i)=0$, $0\leqq i\leqq n$. We find $d_{n,1}(x)=|\pi(x)|\cdot\sum_{j=0}^{n}\frac{1}{|\pi'(x_j)|}$, and can therefore relate $|E_n(f)(x)|$ to $|\pi(x)|$ and $M_1,\ldots,M_{n+1}$. We may also show that when $x_i=t_i$ then $d_{n,1}(x)=\operatorname{cosec}\left(\pi/(2n+2)\right)|T_{n+1}(x)|/(n+1)$, and this permits us to prove that uniform bounds for $\|E_n(f)\|$ in terms of $\|d_{n,k}\|$ may be worse than the corresponding bounds in terms of $\|\lambda_n\|$. The evaluation of $\delta_{n,k}(x)$ or its norm is to be preferred; we note especially that $|E_n(f)(x)|\leqq\mu_k\delta_{n,1}(x)M_{k+1}\leqq\frac{\pi}{2(n+1)}\left(1+\lambda_n(x)\right)\mu_k M_{k+1}$, where $\mu_k=(n-k)!\,(\frac{1}{2}\pi)^k/n!$ and $0\leqq k\leqq n$.

For the evaluation of $\delta_{n,k}(x)$, in general, it is necessary to integrate the modulus of a function which is piecewise polynomial. We may attempt this by quadrature or by locating the zeros of $\nu(t)=K_k(x,t)$. If $k=1$, we have the explicit expression

$$\begin{aligned}k\,\delta_{n,k}(x)=&\sum_{r=0}^{p-1}\left|\sum_{j=0}^{r}l_j(x)\left\{(x_j-x_r)^k-(x_j-x_{r+1})^k\right\}\right|\\&+\sum_{r=p+1}^{n}\left|\sum_{j=r+1}^{n}l_j(x)\left\{(x_j-x_r)^k-(x_j-x_{r+1})^k\right\}\right|\\&+\left|\sum_{j=0}^{p}l_j(x)\left\{(x_j-x)^k-(x_j-x_p)^k\right\}\right|\\&+\left|\sum_{j=p+1}^{n}l_j(x)\left\{(x_j-x)^k-(x_j-x_{p+1})^k\right\}\right|\end{aligned}$$

for $x_{p+1}<x<x_p$. If $k>1$, we replace terms in $\{\}$ by terms depending on the zeros of $\nu(t)$. For $k=2$ it is relatively easy to locate the zeros of $\nu(t)$ numerically. The accompanying table and figure display some of the results obtained in this way.

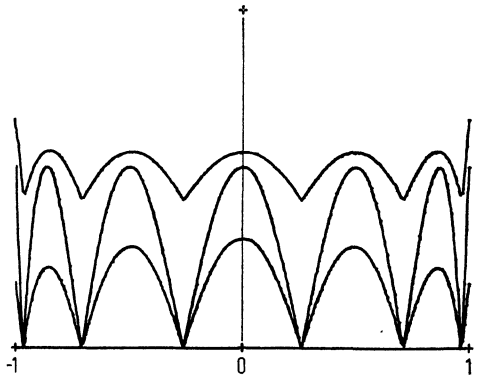

Fig. 1

The sample points associated with Fig. 1 are the zeros $\{t_i\}$ of $T_6(x)$. The bounds displayed are $\frac{1}{2}\pi(1+\lambda_n(x))/(n+1)$, $d_{n,1}(x)$ and $\delta_{n,1}(x)$, and each is better than the preceding one. Uniform bounds related to $M_1[-1, 1]$ and $M_2[-1, 1]$ are displayed in the Table for the case where the x_i are the zeros t_i of $T_{n+1}(x)$ and the case $x_i = h_i = 1-2i/n$.

Table

n	$\|\delta_{n,1}\|$	$\frac{(1+\|\lambda_n\|)\pi}{2(n+1)}$	$\|\delta_{n,2}\|$	$\frac{\|\delta_{n,1}\|\pi}{2n}$	$\frac{(1+\|\lambda_n\|)\pi^2}{4n(n+1)}$
	Sample points h_i				
2	0.632	1.179	0.151	0.496	0.926
3	0.555	1.034	0.087	0.291	0.541
4	0.565	1.008	0.068	0.222	0.396
5	0.635	1.075	0.062	0.199	0.338
6	0.771	1.246	0.065	0.202	0.326
7	1.002	1.558	0.074	0.225	0.350
8	1.381	2.085	0.091	0.271	0.409
9	1.995	2.961	0.118	0.348	0.517
10	3.001	4.413	0.160	0.471	0.693
	Sample points t_i				
2	0.547	1.397	0.113	0.430	1.097
3	0.495	1.119	0.074	0.259	0.586
4	0.425	0.940	0.050	0.167	0.369
5	0.392	0.813	0.038	0.123	0.255
6	0.351	0.719	0.029	0.092	0.188
7	0.328	0.646	0.024	0.074	0.145
8	0.301	0.587	0.020	0.059	0.115
9	0.284	0.539	0.017	0.050	0.094
10	0.265	0.499	0.015	0.042	0.078

For large values of k the method of evaluating $\delta_{n,k}(x)$ using root-finding becomes more difficult as the problem of locating zeros becomes more delicate. One may then resort to Theorem 2, or one may employ a quadrature method

to evaluate

$$\int_{-1}^{1} |K_k(x;t)|\,dt = \sum_{r=-1}^{p-1} \int_{x_{r+1}}^{x_r} \left| \sum_{j=0}^{r} L_j(x,t) \right| dt + \int_{x_{p+1}}^{x} \left| \sum_{j=p+1}^{n} L_j(x,t) \right| dt$$

$$+ \int_{x}^{x_p} \left| \sum_{j=0}^{p} L_j(x,t) \right| dt + \sum_{r=p+1}^{n} \int_{x_{r+1}}^{x_r} \left| \sum_{j=r+1}^{n} L_j(x,t) \right| dt$$

where

$$L_j(x,t) = l_j(x)\,(x_j - t)^{k-1}, \qquad x \in [x_{p+1}, x_p].$$

Numerical integration has been used to evaluate the Peano constants for quadrature rules (see references [10, 11] of [1]); this technique also presents some problems.

4. Further Developments

We have derived bounds for $\|E_n(f)\|$ from bounds for $\|e_n(f)\|$. Conversely, by adjusting the sample points to minimize $\|\delta_{n,k}\|$ we may improve on the results of Theorem 1 for particular choices of n, k. Our tabulated results permit such a limited development, and were employed in [1].

References

1. Baker, C. T. H.: On the nature of certain quadrature formulae and their errors. SIAM J. Numer. Anal. **5**, 783–804 (1968).
2. Cannon, J. R., Miller, Keith: Bounds for the ratio between the fit of interpolation and the best fit. (1964) Unpublished.
3. Cheney, E. W.: Introduction to approximation theory. New York-London: McGraw-Hill Book Co. 1967.
4. Davis, P. J.: Interpolation and approximation. (2nd Edition.) New York-London: Blaisdell Pub. Co. 1965.
5. Guerra, S.: Osservazioni su un noto teorema di Jackson. Boll. Un. Mat. Ital. (3) **18**, 57–64 (1963). (Translation by Pauline Radcliffe, University of Manchester.)
6. Natanson, I. P.: Constructive theory of functions. Vol. 3. Moscow: State Publishing House of Technical-Theoretical Literature 1949. [Translations U. S. Atomic Energy Commission (1961) AEC-tr-4503, and Frederick Ungar Pub. Co. (1965).]
7. Powell, M. J. D.: On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria. The Computer Journal **9**, no. 4, 404–407 (1967).
8. Schonhage, A.: Fehlerfortpflanzung bei Interpolation. Numerische Mathematik **3**, 62–71 (1961).
9. Timan, A. F.: Theory of approximation of functions of a real variable. Moscow: State Publishing House 1960. [Translation: Pergamon Press 1963.]

Dr. C. T. H. Baker
Department of Mathematics
The University
Manchester M 13 9 PL
England

Numer. Math. 15, 320–328 (1970)
© by Springer-Verlag 1970

Error Estimate for the Numerical Solution of Fredholm's Integral Equation

E. RAKOTCH and J. STEINBERG*

Received May 15, 1969

Abstract. A bound on the actual error incurred in the numerical solution of Fredholm's integral equations with symmetric kernel is obtained, using computed approximations of some eigenvalues.

1. Introduction

Consider the equation

$$y(x) - \lambda \int_a^b K(x, t)\, y(t)\, dt = f(x) \tag{1}$$

with

$$f(x) \in C^{(r)}(I)$$

$$K(x, t) \in C^{(r)}(I \times I)$$

where $I = [a, b]$ and λ is not an eigenvalue λ_i of $K(x, t)$, which is supposed to be a real symmetric kernel.

An approximate solution of (1) at points x_i of I is assumed to be found from the system

$$y_i^{(n)} - \lambda \sum_{j=1}^{n} w_j^{(n)} K_{ij} y_j^{(n)} = f_i, \qquad i = 1, \ldots, n \tag{2}$$

where $K_{ij} = K(x_i, x_j)$, $f_i = f(x_i)$ and $w_i^{(n)} > 0$ are coefficients of the integration method S whose error terms involve derivatives up to order r at most.

In fact, let $\tilde{y}^{(n)}$ be an approximate solution of (2) obtained by actual computation, and let

$$\varrho_i^{(n)} = \tilde{y}_i^{(n)} - \lambda \sum_{j=1}^{n} w_j^{(n)} K_{ij} \tilde{y}_j^{(n)} - f_i, \qquad i = 1, 2, \ldots, n.$$

Define the function

$$y_n(x) = \lambda \sum_{j=1}^{n} w_j^{(n)} \tilde{y}_j^{(n)} K(x, x_j) + f(x).$$

Then an estimate for the error $e_n(x) = y_n(x) - y(x)$ is to be found.

Such an estimate at points x_i was obtained in [4], but is not practicable as it involves a bound for the resolvent kernel, which in general cannot be found. Another numerical error estimate was obtained by Brown [2], but entails successful guessing of the initial bounding approximations. Kantorovich and Krylov [3] as well as Anselone and Moore [1] obtained a bound for the error in the case

* This article is a part of D. Sc. Thesis of E. Rakotch.

of a continuous kernel, provided a certain inequality involving λ is satisfied. The above references provide too general convergence theorems. In this paper a more practical estimate is derived for symmetric kernels and for $a \leqq x \leqq b$, using some results of Wielandt [6].

By (1)

$$e_n(x) = \lambda \sigma_n(x)$$

where

$$\sigma_n(x) = \sum_{i=1}^{n} w_i^{(n)} \tilde{y}_i^{(n)} K(x, x_i) - \int_a^b K(x, t)\, y(t)\, dt.$$

By (2), and the definition of $\varrho_i^{(n)}$

$$\tilde{y}_i^{(n)} = y_n(x_i) + \varrho_i^{(n)}.$$

Hence

$$\sigma_n(x) = E_n(x) + \int_a^b K(x, t)\, e_n(t)\, dt$$

where

$$E_n(x) = \sum_{i=1}^{n} w_i^{(n)} K(x, x_i)\, y_n(x_i) - \int_a^b K(x, t)\, y_n(t)\, dt + \sum_{i=1}^{n} w_i^{(n)} \varrho_i^{(n)} K(x, x_i)$$

and consequently $e_n(x)$ satisfies the integral equation

$$e_n(x) - \lambda \int_a^b K(x, t)\, e_n(t)\, dt = \lambda E_n(x). \tag{3}$$

The first sum in $E_n(x)$ is the value of the numerical quadrature for the integral coming after the sum.

Denote

$$\eta_n(x, t) = \sum_{i=1}^{n} w_i^{(n)} K(x, x_i) K(x_i, t) - \int_a^b K(x, z) K(z, t)\, dz,$$

$$\varepsilon_n(x) = \sum_{i=1}^{n} w_i^{(n)} K(x, x_i) f(x_i) - \int_a^b K(x, t) f(t)\, dt.$$

Then by the definition of $y_n(x)$ it follows that

$$E_n(x) = \lambda \sum_{i=1}^{n} w_i^{(n)} \tilde{y}_i^{(n)} \eta_n(x, x_i) + \varepsilon_n(x) + \sum_{i=1}^{n} w_i^{(n)} \varrho_i^{(n)} K(x, x_i).$$

2. Error Bound

Denote

$$a_n = \max_I |\varepsilon_n(x)|, \qquad A_n = \|\varepsilon_n\|, \qquad M = \max_{I \times I} |K(x, t)|$$

where

$$\|u\| = \left\{ \int_a^b u^2(x)\, dx \right\}^{\frac{1}{2}}.$$

Then

$$|E_n(x)| \leqq a_n + |\lambda|\, b_n + M c_n, \tag{4}$$

$$\|E_n\| \leqq A_n + |\lambda|\, B_n + \sqrt{b-a}\, M c_n \tag{5}$$

where

$$b_n = \sum_{i=1}^{n} w_i^{(n)} |\tilde{y}_i^{(n)}| \max_I |\eta_n(x, x_i)|$$

$$B_n = \sum_{i=1}^{n} w_i^{(n)} |\tilde{y}_i^{(n)}| \left\{ \int_a^b \eta_n^2(x, x_i)\, dx \right\}^{\frac{1}{2}}$$

$$c_n = \sum_{i=1}^{n} w_i^{(n)} |\varrho_i^{(n)}|$$

and the bounds for a_n, A_n, b_n and B_n are evaluated from those of the error terms for $\int_a^b K(x, t) f(t)\, dt$ and $\int_a^b K(x, z) K(z, t)\, dz$ in the method S.

Now let $\{\phi_i(x)\}$ be an orthonormal set of eigenfunctions and $\gamma_{in} = (E_n, \phi_i)$ — the corresponding expansion coefficients of E_n. The solution of (3) is then

$$e_n(x) = \lambda \left[E_n(x) + \lambda \sum_{i=1}^{\infty} \frac{\gamma_{in}}{\lambda_i - \lambda} \phi_i(x) \right].$$

By the Cauchy-Schwarz and Bessel inequalities

$$\left[\sum_{i=1}^{\infty} \frac{\gamma_{in}}{\lambda_i - \lambda} \phi_i(x) \right]^2 \leqq \sum_{i=1}^{\infty} \gamma_{in}^2 \sum_{i=1}^{\infty} \frac{\phi_i^2(x)}{(\lambda_i - \lambda)^2} \leqq \|E_n\|^2 \sum_{i=1}^{\infty} \frac{\phi_i^2(x)}{(\lambda_i - \lambda)^2}.$$

Now let $\alpha = \sup_i \left| 1 - \frac{\lambda}{\lambda_i} \right|^{-1}$; then

$$\sum_{i=1}^{\infty} \frac{\phi_i^2(x)}{(\lambda_i - \lambda)^2} \leqq \alpha^2 \sum_{i=1}^{\infty} \frac{\phi_i^2(x)}{\lambda_i^2} = \alpha^2 K^{(2)}(x, x)$$

where $K^{(2)}(x, t) = \int_a^b K(x, z) K(z, t)\, dz$ is the first iterated kernel. Hence

$$\left| \sum_{i=1}^{\infty} \frac{\gamma_{in}}{\lambda_i - \lambda} \phi_i(x) \right| \leqq \alpha \|E_n\| \sqrt{K^{(2)}(x, x)}$$

and consequently

$$|e_n(x)| \leqq |\lambda| \left(|E_n(x)| + |\lambda| \alpha \|E_n\| \sqrt{K^{(2)}(x, x)} \right) \tag{6}$$

where the bounds for $E_n(x)$ and $\|E_n\|$ are given by (4) and (5).

3. Evaluation of α

For fixed n (not necessarily equal to the order of the system (2)), let $\mu_i^{(n)}$ be the eigenvalues of the matrix $H_n = (K_{ij} w_j^{(n)})$; by similarity, they coincide with those of the symmetric matrix $D K D$, where $K = (K_{ij})$ and D is diagonal with $D_{ii} = \sqrt{w_i^{(n)}}$, and therefore obtainable by standard methods. If $\mu_i^{(n)} = 0$ is set for $i > n$ then by [6] $\mu_i^{(n)} \to \lambda_i^{-1}$ uniformly in i, i.e. there exists a sequence δ_n converging to 0 such that $|\mu_i^{(n)} - \lambda_i^{-1}| \leqq \delta_n$ for every i.

Two examples of quadrature formulae, considered in this paper, in which estimates for δ_n are known, are the following ones:

(a) Gaussian quadrature

$$\delta_n \leqq 4\left(\frac{d}{n-1}\right)^p \max_{I\times I}\left|\frac{\partial^p K}{\partial x^p}\right|, \quad p \text{ arbitrary integer}$$

where $n \geqq p$ and

$$d = \begin{cases} 1.92 & p \leqq 6 \\ 4.1 & p > 6. \end{cases}$$

(b) Trapezoidal rule with a kernel of the form

$$K(x, t) = h_1(x+t) + h_2(x-t), \quad 0 \leqq x, t \leqq 1$$

where the functions $h_r(z)$, $r = 1, 2$ are periodic having period 1.

Now let $h_r(z)$, $r = 1, 2$, be approximated by the finite Fourier series

$$h_r^{(n)}(z) = \sum_{T_n} c_{rj} e^{2\pi i j z}$$

where

$$T_n = \begin{cases} \{j \mid -m \leqq j \leqq m\}, & n = 2m+1 \\ \{j \mid -m \leqq j \leqq m+1\}, & n = 2m+2. \end{cases}$$

Then by [6], p. 276 (Theorem 11)

$$\delta_n \leqq 2(\varepsilon_1 + \varepsilon_2), \quad \text{where} \quad \varepsilon_r = \max\left|h_r^{(n)}(z) - h_r(z)\right|, \quad r = 1, 2.$$

Now take on the μ-axis, about every $\mu_i^{(n)}$ as centre, an interval I_i of length $2\delta_n$ and let U be the union of all I_i (see Fig. 1 below, where the $\mu_i^{(n)}$ are marked with crosses and the intervals with arcs). Then $\lambda_i^{-1} \in I_i$ for each i.

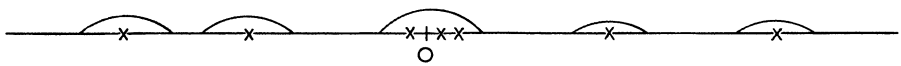

Fig. 1

Let R be the maximal neighbourhood of 0 contained in U. By the transformation $\mu \to z = \mu^{-1}$, the interval R is resolved into two infinite segments $\{z \mid z \leqq A\}$ and $\{z \mid z \geqq B\}$ with $A < 0 < B$, while the other elements of U become a finite number of closed intervals (see Fig. 2 below), and let V be their union. Every eigenvalue λ_i of $K(x, y)$ then belongs to some interval of V.

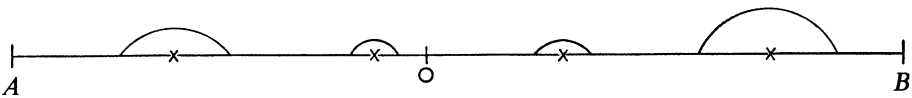

Fig. 2

Let z_i^- be the left-hand endpoint of the closed interval of V corresponding to I_i, z_i^+ — the right-hand endpoint, and let λ be any number not contained in V (for any $\lambda \neq \lambda_i$, this may be attained for n sufficiently large); then $\lambda - \lambda_i > \lambda - z_i^+ > 0$ whenever $\lambda > 1/\mu_i^{(n)}$ and $\lambda - \lambda_i < \lambda - z_i^- < 0$ wherever $\lambda < 1/\mu_i^{(n)}$.

In order to find an upper bound for $\alpha \geqq 1$, we seek a lower bound for

$$\inf_i \left|1 - \frac{\lambda}{\lambda_i}\right| = \alpha^{-1}.$$

Let first $\lambda > 0$; it is then sufficient to consider those i for which $\lambda_i > 0$, since $\alpha^{-1} \leqq 1$.

If $0 < \lambda_i < \lambda$ then

$$\left|1 - \frac{\lambda}{\lambda_i}\right| = \frac{\lambda}{\lambda_i} - 1 > \frac{\lambda}{z_i^+} - 1; \quad z_i^+ > 0$$

while if $0 < \lambda < \lambda_i$ then

$$\left|1 - \frac{\lambda}{\lambda_i}\right| = 1 - \frac{\lambda}{\lambda_i} > 1 - \frac{\lambda}{z_i^-}; \quad z_i^- > 0, \; B = z_p^- > 0.$$

Thus, letting

$$\beta^+ = \min\left(\frac{\lambda}{z_i^+} - 1 \,\middle|\, 0 < \lambda_i < \lambda\right)$$

$$\beta^- = \min\left(1 - \frac{\lambda}{z_i^-} \,\middle|\, \lambda_i > \lambda\right)$$

$$\beta = \min(\beta^+, \beta^-)$$

then $\beta \leqq \alpha^{-1}$.

Similarly, we obtain for $\lambda < 0$ the inequality $\alpha^{-1} \geqq \min(\beta^+, \beta^-)$, where

$$\beta^+ = \min\left(1 - \frac{\lambda}{z_i^+} \,\middle|\, \lambda_i < \lambda\right), \quad (z_q^+ = A < 0),$$

$$\beta^- = \min\left(\frac{\lambda}{z_i^-} - 1 \,\middle|\, \lambda < \lambda_i < 0\right).$$

This, together with an evaluation of a bound for $K^{(2)}(x, x)$, enables us to use (6) for finding a bound for $|e_n(x)|$.

4. Convergence

Consider now the function $y_n(x)$ where $\tilde{y}_i^{(n)}$, $i = 1, \ldots, n$, is the exact solution of system (2), i.e. $\varrho_i^{(n)} = 0$ for each i; then $c_n = 0$ in (4) and (5), and the uniform convergence in $[a, b]$ of $y_n(x)$ to $y(x)$ follows if the condition $\lim_{n\to\infty} b_n = 0$ is proved.

Let

$$d_n = \max\left(|\eta_n(x, t)|, \; | \; a \leqq x, t \leqq b\right).$$

Then

$$b_n \leqq d_n Y_n$$

where

$$Y_n = \sum_{i=1}^{n} w_i^{(n)} |y_i^{(n)}|$$

and it remains to prove the existence of a bound for Y_n (provided the matrix of system (2) is regular), since $\lim_{n\to\infty} d_n = 0$.

Let us assume, as it holds for every usual integration method, that $\sum_{i=1}^{n} w_i^{(n)} = b-a$; then by the Cauchy-Schwarz inequality

$$(7) \qquad Y_n^2 = \left\{\sum_{i=1}^{n} w_i^{(n)} \left|y_i^{(n)}\right|\right\}^2 \leqq \sum_{i=1}^{n} w_i^{(n)} \sum_{i=1}^{n} w_i^{(n)} (y_i^{(n)})^2 = (b-a) \sum_{i=1}^{n} w_i^{(n)} (y_i^{(n)})^2.$$

Further denote $z_i^{(n)} = \sqrt{w_i^{(n)}}\, y_i^{(n)}$; then by (2)

$$z_i^{(n)} - \lambda \sum_{j=1}^{n} \sqrt{w_i^{(n)} w_j^{(n)}}\, K_{ij} z_j^{(n)} = \sqrt{w_i^{(n)}}\, f_i = g_i^{(n)},$$

i.e.

$$(I - \lambda D K D)\, z^{(n)} = g^{(n)}$$

where $K = (K_{ij})$ and D is diagonal with $D_{ii} = \sqrt{w_i^{(n)}}$. Hence

$$(8) \qquad \left\{\sum_{i=1}^{n} w_i^{(n)} (y_i^{(n)})^2\right\}^{\frac{1}{2}} = \left|z^{(n)}\right| \leqq \left\|(I - \lambda D K D)^{-1}\right\|_2 \left|g^{(n)}\right|.$$

Now

$$\left|g^{(n)}\right| = \left\{\sum_{i=1}^{n} w_i^{(n)} f_i^2\right\}^{\frac{1}{2}}$$

$$\lim_{n\to\infty} \left|g^{(n)}\right| = \|f\|.$$

By the symmetry of K,

$$L_n = \left\|(I - \lambda D K D)^{-1}\right\|_2 = \max\left(\left|1 - \lambda \mu_i^{(n)}\right|^{-1} \middle| \, i = 1, 2, \ldots, n\right)$$

and by the uniform convergence of $\mu_i^{(n)}$ to λ_i^{-1} (excluding a finite number of values of n such that $\lambda^{-1} = \mu_i^{(n)}$ for some i)

$$\lim_{n\to\infty} L_n = \alpha.$$

Consequently, by (7) and (8)

$$\overline{\lim_{n\to\infty}}\, Y_n \leqq \lim_{n\to\infty} L_n \sqrt{b-a}\, \left|g^{(n)}\right| = \alpha \sqrt{b-a}\, \|f\|$$

which proves that Y_n is bounded.

5. Numerical Results

All the results are computed at points with constant distance h, including the endpoints. Bounds for α are those found by the procedure described in Section 3.

(a) Equation

$$y(x) - \tfrac{1}{2} \int_0^1 e^{xt} y(t)\, dt = 1 - \tfrac{1}{2} g(x)$$

where $g(x) = \dfrac{e^x - 1}{x}$; $y(x) = 1$.

$$M = e$$

$$K^{(2)}(x, x) = \frac{e^{2x} - 1}{2x} \leqq \frac{e^2 - 1}{2} = 3.1945 \ldots$$

$$\alpha < 3.1.$$

Gaussian quadrature is used. Estimates for δ_n are made for $p=6$.

By [5], p. 48

$$\int_a^b F(x)\,dx - \sum_{i=1}^{n} w_i^{(n)} F(x_i) = C_n (b-a)^{2n+1} F^{(2n)}(\xi), \qquad a<\xi<b$$

where

$$C_n = \left[\binom{2n}{n}^2 (2n+1)!\right]^{-1}.$$

Therefore

$$\varepsilon_n(x) = C_n e^{x\tau} G(x,\tau), \qquad 0<\tau<1$$

where

$$G(x,t) = \tfrac{1}{2}\sum_{k=0}^{2n}\binom{2n}{k} g^{(k)}(t)\, x^{2n-k} - x^{2n}$$

and

$$\eta_n(x,t) = -C_n (x+t)^{2n} e^{(x+t)\zeta}, \qquad 0<\zeta<1.$$

It is easy to see that if $n>1$ then $\frac{\partial G}{\partial x}(x,t) \geqq \frac{\partial G}{\partial x}(x,0) > 0$, $g(t) \leqq e-1$ and $g^{(k)}(t) < \frac{e}{k+1}$ for $k>0$ (since $g(t)$ and all its derivatives are increasing for $x \geqq 0$). Hence

$$G(x,t) < \frac{1}{2}\left[e\,\frac{(x+1)^{2n+1}-x^{2n+1}}{2n+1} - 3x^{2n}\right]$$

$$|\eta_n(x,t)| \leqq C_n (x+t)^{2n} e^{x+t}. \tag{9}$$

Therefore

$$a_n < \frac{1}{2} C_n e\left(\frac{2^{2n+1}-1}{2n+1} e - 3\right), \qquad A_n < \frac{C_n e}{4n+2}\left\{\int_0^1 e^{2x}(1+x)^{4n+2}dx\right\}^{\frac{1}{2}}.$$

For $n=2$

$$a_2 < 4.36\cdot 10^{-3}, \qquad A_2 < 2.016\cdot 10^{-3}, \qquad c_2 < 3.8\cdot 10^{-9}$$

and by (9)

$$b_2 < 7.93\cdot 10^{-3}, \qquad B_2 < 2.98\cdot 10^{-3}.$$

Hence

$$|e_2(x)| < 9.1\cdot 10^{-3}.$$

The actual maximal error at $x = j\cdot 0.01$ is about $3\cdot 10^{-4}$.

Kantorovich and Krylov [3], p. 110, obtain

$$|e_2(x)| < 7\cdot 10^{-2}$$

which is about 8 times larger than our estimate.

(b) Equation

$$y(x) - 24\int_0^{1/2} y(t)\sin x\,t\,dt = 1 - 24\,g(x)$$

where $g(x) = \frac{1}{x}\left(1-\cos\frac{x}{2}\right)$; $y(x)=1$.

$$M = \sin\frac{1}{4},$$

$$K^{(2)}(x, x) = \frac{1}{4}\left(1 - \frac{\sin x}{x}\right) \leqq \frac{1}{4}\left(1 - 2\sin\frac{1}{2}\right) < 0.0104; \quad \alpha < 268.$$

Gaussian quadrature is used. Estimates for δ_n are made for $p = 6$.

Similar considerations, using the special properties of $g(x)$ imply

$$a_n < \frac{6.4\, C_n}{2^{2n}(2n+1)}, \qquad A_n < \frac{6.4\, C_n}{2^{2n}(2n+1)\sqrt{4n+3}},$$

$$|\eta_n(x, t)| \leqq \left(\frac{1}{4}\right)^{n+1}(x+t)^{2n}.$$

For $n = 3$ the above inequalities, together with $c_3 < 1.8 \cdot 10^{-9}$, yield

$$|e_3(x)| < 6 \cdot 10^{-5}.$$

The actual maximal error at $x = j \cdot 0.01$ is $1.85 \cdot 10^{-6}$.

(c) Equation

$$y(x) - \lambda \int_{-1}^{1} \frac{1}{2}\left[\frac{1 - r\cos\pi(x-t)}{g(x-t)} - \frac{1 - r\cos\pi(x+t)}{g(x+t)}\right] y(t)\, dt$$
$$= x - \frac{2\lambda}{\pi}\operatorname{arctg}\frac{r\sin\pi x}{1 + r\cos\pi x}$$

where $0 < r < 1$ and $g(z) = 1 + r^2 - 2r\cos\pi z$; $y(x) = x$.

The trapezoidal rule is used with $h = \frac{2}{n-1}$ in each case.

The bounds for $\varepsilon_n(x)$ and $\eta_n(x, t)$ are obtained by using the inequality $1 - x\cot x \leqq \left(\frac{2x}{\pi}\right)^2$ for $|x| \leqq \frac{\pi}{2}$, the residuum theory and the following lemma.

Lemma. Let $F(t)$ and $G(t)$ be real periodic functions having period $2L$, and uniformly and absolutely convergent Fourier series

$$F(t) = \sum_{k=-\infty}^{\infty} a_k e^{\frac{ik\pi}{L}t}, \qquad G(t) = \sum_{k=-\infty}^{\infty} b_k e^{\frac{ik\pi}{L}t}.$$

Then

$$\frac{2L}{N}\sum_{j=1}^{N} F\left(2j\frac{L}{N} - L\right) G\left(2j\frac{L}{N} - L\right) - \int_{-L}^{L} F(t)\, G(t)\, dt = 4L\sum_{k=1}^{\infty} \operatorname{Re}(c_{kN})$$

where

$$c_m = \sum_{s=0}^{m} a_s b_{m-s} + \sum_{s=1}^{\infty}(\bar{a}_s b_{m+s} + \bar{b}_s a_{m+s}), \qquad m = 1, 2, \ldots.$$

Moreover, if $F(t)$ is an odd function, then

$$\frac{2L}{N}\sum_{j=1}^{N}(2j - N)\frac{L}{N} F\left(2j\frac{L}{N} - L\right) - \int_{-L}^{L} xF(x)\, dx$$
$$= 4iL^2\left[\sum_{T}\frac{(-1)^k}{k\pi} a_k\left(1 - \frac{k\pi}{N}\cot\frac{k\pi}{N}\right) + \sum_{m=1}^{\infty}\frac{(-1)^{mN}}{mN\pi} a_{mN}\right]$$

where T is the set of all natural numbers not divisible by N.

$$a_n < \alpha_n = \frac{8r + 2n^2(1-r)\,r^{n/2}}{\pi(n-1)^2(1-r)^2} + \frac{2|\lambda|\, n(r^{n-1}+r^n)}{\pi(1-r^{n-1})^2},$$

$$A_n < \alpha_n, \qquad |\eta_n(x,t)| < \frac{n(r^{n-1}+r^n)}{(1-r^{n-1})^2}$$

$$K^{(2)}(x,x) = \frac{1}{2}\left[\frac{1}{1-r^2} - \frac{1-r^2\cos 2\pi x}{1+r^4-2r^2\cos 2\pi x}\right] \leqq \frac{r^2}{1-r^4}.$$

Case 1. $r=0.5$, $\lambda=1$, $n=41$

$$\alpha < 2.1, \qquad c_n < 10^{-8}.$$

Error bound:

$$|e_n(x)| < 6.7 \cdot 10^{-3}.$$

The actual maximal error is about $2.3 \cdot 10^{-3}$.

Case 2. $r=0.5$, $\lambda=3$, $n=41$

$$\alpha < 4.1, \qquad c_n < 1.5 \cdot 10^{-8}.$$

Error bound:

$$|e_n(x)| < 7.05 \cdot 10^{-2}.$$

The actual maximal error is about $9.7 \cdot 10^{-3}$.

Case 3. $r=0.25$, $\lambda=3$, $n=21$

$$\alpha < 4.1, \qquad c_n < 4.10^{-9}.$$

Error bound:

$$|e_n(x)| < 3.5 \cdot 10^{-2}.$$

The actual maximal error is about $1.57 \cdot 10^{-2}$.

References

1. Anselone, P. M., Moore, R. H.: Approximate solutions of integral and operator equations. J. Math. Anal. Appl. **9**, 268–277 (1964).
2. Brown, R. W.: Upper and lower bounds for solutions of integral equations. Proceedings of the Symposium conducted by the Mathematical Research Center, 219–230, Madison, Wisc., 1965.
3. Kantorovich, L. V., Krylov, V. I.: Approximate methods of higher analysis. P. Noordhoff, 1958.
4. Mysovskih, I. P.: Estimate of the error involved in the numerical solution of a linear integral equation. Soviet Mathematics **2**, 1268–1270 (1961).
5. Todd, J.: Introduction to the constructive theory of functions. Basel: Birkhauser 1963.
6. Wielandt, H.: Error bounds for eigenvalues of symmetric integral equations. Proceedings of Symposia in Applied Mathematics **6**, 261–282 (1956).

E. Rakotch
Department of Computer Sciences
Technion City
Haifa, Israel

Numer. Math. 15, 329–332 (1970)
© by Springer-Verlag 1970

Potenzbeschränktheit und Wertebereich einer Matrix*

KARL Graf FINCK VON FINCKENSTEIN

Eingegangen am 20. November 1968

Bei Stabilitätsuntersuchungen von Differenzenschemata zur numerischen Behandlung von partiellen Differentialgleichungen entsteht die Frage nach Kriterien für die Potenzbeschränktheit der zu einem Differenzenschema gehörigen Amplifikationsmatrix. Dabei erweist sich die Untersuchung des sog. Toeplitzschen Wertebereiches einer Matrix als zweckmäßig.

$\mathbb{C}$ sei der Körper der komplexen Zahlen. Ist u ein Spaltenvektor mit Elementen aus $\mathbb{C}$, so sei mit u^* der zu u konjugiert komplexe Zeilenvektor bezeichnet. A sei eine quadratische $m \times m$-Matrix mit Elementen aus $\mathbb{C}$. Wir definieren:

Def. 1. Die Menge der komplexen Zahlen

$$W(A) := \{u^* A u : u^* u = 1\}$$

heiße der *Wertebereich* der Matrix A.

Folgende Tatsachen sind bekannt (vgl. [1, 2]):

1) $W(A)$ ist eine konvexe Menge, die die Eigenwerte von A umfaßt.

2) Ist A normal, so gilt für die konvexe Hülle der Eigenwerte $\lambda_1, \ldots, \lambda_m$ von A:

$$\text{konv}\{\lambda_1, \ldots, \lambda_m\} = W(A).$$

3) Ist U eine m-reihige unitäre Matrix, so gilt:

$$W(U^* A U) = W(A).$$

Mit „$\|\cdot\|$“ werde im folgenden eine beliebige, natürliche multiplikative Matrixnorm bezeichnet.

Def. 2. A heiße *potenzbeschränkt*, wenn eine Konstante K existiert, so daß gilt:

$$\|A^n\| \leqq K \quad \text{für} \quad n = 1, 2, 3, \ldots$$

Anmerkung 1. Offenbar ist Def. 2 von der speziellen Wahl der (natürlichen) Norm unabhängig.

Anmerkung 2. Man prüft sofort nach, daß die Eigenschaft einer Matrix A, potenzbeschränkt zu sein, invariant ist gegenüber Ähnlichkeitstransformationen von A.

* Die nachstehende Arbeit wurde im Rahmen des Vertrages zwischen dem Institut für Plasmaphysik GmbH und der Europäischen Atomgemeinschaft über die Zusammenarbeit auf dem Gebiete der Plasmaphysik durchgeführt.

Wir beweisen:

Satz. Folgende Aussagen sind äquivalent:

a) A ist potenzbeschränkt.

b) $W(A^n)$ bleibt beschränkt für $n=1, 2, \ldots$

c) Alle Eigenwerte $\lambda_1, \ldots, \lambda_m$ von A erfüllen $|\lambda_i| \leqq 1$; ist dabei $|\lambda_j| = 1$, so sind alle zu λ_j gehörigen Elementarteiler linear.

d) Es existiert eine reguläre Matrix T, so daß $W(T^{-1} A\, T)$ im abgeschlossenen Einheitskreis der komplexen Ebene enthalten ist.

Beweis. Wir zeigen zunächst die Äquivalenz von a) und b).

Die Richtung von a) nach b) ist trivial: Mit Hilfe der Schwarzschen Ungleichung ergibt sich nämlich

$$|(A^n u, u)| \leqq \|A^n u\|_E \leqq \|A^n\|_E \leqq K, \quad n=1, 2, \ldots,$$

woraus sofort b) folgt.

$\|\cdot\|_E$ bedeutet dabei die euklidische Vektornorm bzw. die durch diese induzierte Matrixnorm.

Umgekehrt sei A nicht potenzbeschränkt. Wegen Ergebnis 3) im Anschluß an Def. 1 können wir ohne Einschränkung annehmen, daß A eine obere Dreiecksmatrix ist. Es gibt dann eine Teilfolge $\{n_\nu\}$ der Folge der natürlichen Zahlen, so daß gilt:

$$\|A^{n_\nu}\| \to \infty \quad \text{für} \quad n \to \infty.$$

Es muß also für wenigstens ein Element $a_{jk}^{(n_\nu)}$ von A^{n_ν} gelten:

$$|a_{jk}^{(n_\nu)}| \to \infty \quad \text{für} \quad n \to \infty.$$

Wir unterscheiden jetzt zwei Fälle:

α) Es möge ein Eigenwert $\lambda_j^{n_\nu}$ von A^{n_ν} unbeschränkt wachsen. Da $\lambda_j^{n_\nu} \in W(A^{n_\nu})$, folgt, daß $W(A^{n_\nu})$ nicht beschränkt bleiben kann für $n \to \infty$.

β) Es mögen alle Eigenwerte von A^{n_ν} beschränkt bleiben. Sei $a_{jk}^{(n_\nu)}$ mit $j<k$ ein Element von A^{n_ν}, das unbeschränkt wächst. Sei u_0 der Einheitsvektor, dessen j-te und k-te Komponente $\frac{1}{2}\sqrt{2}$ ist, und dessen übrige Komponenten Null sind. Dann folgt

$$|u_0^* A^{n_\nu} u_0| = \tfrac{1}{2} \cdot |\lambda_j^{n_\nu} + \lambda_k^{n_\nu} + a_{jk}^{(n_\nu)}|.$$

Hieraus sieht man sofort, daß $W(A^{n_\nu})$ für $n \to \infty$ nicht beschränkt bleibt.

Den Rest des Beweises führen wir zyklisch:

$$\text{a)} \to \text{c)} \to \text{d)} \to \text{a)}.$$

Es gelte a) bzw. b). Dann folgt sofort, daß alle Eigenwerte $\lambda_1, \ldots, \lambda_m$ von A $|\lambda_i| \leqq 1$ erfüllen, da gilt:

$$\lambda_i^n \in W(A^n), \quad i=1, \ldots, m.$$

Weiter betrachten wir die zu A gehörige Jordansche Normalform

$$J = \begin{pmatrix} J_1 & & 0 \\ & \ddots & \\ 0 & & J_r \end{pmatrix},$$

wobei die J_k Jordanmatrizen sind. Es folgt, daß alle J_k ebenfalls potenzbeschränkt sein müssen. Man überlegt sich leicht, daß diejenigen J_k, deren Eigenwerte den Betrag 1 haben, Diagonalmatrizen sein müssen. Genau dieses aber bedeutet den zweiten Teil der Eigenschaft c).

Wir leiten jetzt aus c) die Eigenschaft d) her. Dazu betrachten wir wieder die Jordansche Normalform J von A. Für diejenigen Jordanmatrizen J_k, die Diagonalmatrizen sind, folgt:

$$W(J_k) \subset \mathfrak{E} := \{z \in \mathbb{C} : |z| \leqq 1\},$$

denn Diagonalmatrizen sind normal.

Die übrigen Jordanmatrizen haben die Gestalt

$$J_k = \begin{bmatrix} \mu_k & 1 & 0 \\ & \ddots & 1 \\ 0 & & \mu_k \end{bmatrix} \quad \text{mit} \quad |\mu_k| < 1.$$

Wir wählen jetzt ein $\varepsilon > 0$, bilden die Matrix

$$E := \begin{bmatrix} \varepsilon & & 0 \\ & \varepsilon^2 & \\ & & \ddots \\ 0 & & \varepsilon^m \end{bmatrix}$$

und bilden die Matrix:

$$J' := E^{-1} J E = \operatorname{Diag}(J'_k)_{k=1,\ldots,r}.$$

Wir behaupten: $W(J') \subset \mathfrak{E}$ für hinreichend kleines ε. Wir zeigen zunächst: $W(J'_k) \subset \mathfrak{E}$ für $k = 1, \ldots, r$.

Ist J'_k Diagonalmatrix, so ist $J_k = J'_k$ und folglich wegen obiger Betrachtungen $W(J'_k) \subset \mathfrak{E}$.

Anderenfalls hat J'_k die Gestalt:

$$J'_k = \begin{bmatrix} \mu_k & \varepsilon & 0 \\ & \ddots & \varepsilon \\ 0 & & \mu_k \end{bmatrix} \quad \text{mit} \quad |\mu_k| < 1.$$

Ist u ein beliebiger Einheitsvektor mit gleicher Reihenzahl, wie J'_k, so folgt

$$u^* J'_k u = \mu_k \cdot \sum_{\nu=1}^{l} |u_\nu|^2 + \varepsilon \cdot \sum_{\nu=2}^{l} \bar{u}_{\nu-1} \cdot u_\nu$$

also

$$|u^* J'_k u| \leqq |\mu_k| + \varepsilon \cdot K_1 \leqq 1$$

für hinreichend kleines ε, da ja $|\mu_k| < 1$ gilt. Damit ist gezeigt:

$$W(J_k') \subset \mathfrak{E} \quad \text{für} \quad k = 1, \ldots, r.$$

Es gilt jetzt folgendes:

Lemma. Sei $C := \mathrm{Diag}(C_k)_{k=1,\ldots,r}$ eine quadratische Blockdiagonalmatrix. Dann gilt:

$$W(C) = \mathrm{konv}\left\{\bigcup_{k=1}^{r} W(C_k)\right\}.$$

Aufgrund dieses Lemmas folgt also wegen der Konvexität von $\mathfrak{E}$:

$$W(J') \subset \mathfrak{E}.$$

Da aber J' aus A durch eine reguläre Transformation hervorgeht, ist somit Eigenschaft d) bewiesen. Übrig bleibt nur noch der Beweis des Lemmas:

Man zeigt sehr leicht:

$$W(C) = \left\{\sum_{k=1}^{r} \alpha_k z_k : \sum_{k=1}^{r} \alpha_k = 1,\ \alpha_k \geqq 0,\ z_k \in W(C_k)\right\}.$$

Das aber bedeutet

$$W(C) \subset \mathrm{konv}\left\{\bigcup_{k=1}^{r} W(C_k)\right\}.$$

Andererseits aber gilt:

$$W(C) \supset W(C_k) \quad \text{für} \quad k = 1, \ldots, r.$$

Wegen der Konvexität von $W(C)$ folgt damit die Behauptung. Damit ist das Lemma bewiesen.

Schließlich muß noch aus d) die Eigenschaft a) gefolgert werden. In [3] wird gezeigt: Ist $W(A) \subset \mathfrak{E}$, dann ist A potenzbeschränkt. Der Beweis in [3] läßt sich verkürzen, wir wollen dies aber aus Platzgründen hier nicht ausführen.

Wegen Anmerkung 2 bei Def. 2 folgt damit auch der letzte Teil des zyklischen Beweises, womit der Satz vollständig bewiesen ist.

Literatur

1. Hausdorff, F.: Der Wertevorrat einer Bilinearform. Math. Zeitschr. **3**, 314–316 (1919).
2. Toeplitz, O.: Das algebraische Analogon zu einem Satz von Fejer. Math. Zeitschr. **2**, 187–197 (1918).
3. Lax, P. D., Wendroff, B.: Difference schemes with high order or accuracy for solving hyperbolic equations. Comm. pure appl. math. **17**, 381–398 (1964).

Dr. Karl Graf Finck von Finckenstein
Institut für Plasmaphysik, Abt. Theorie
8046 Garching

Numer. Math. 15, 333–344 (1970)
© by Springer-Verlag 1970

On the Remainder Term in the N-Dimensional Euler Maclaurin Expansion*,**

J. N. LYNESS and B. J. J. MCHUGH

Received March 27, 1969

Abstract. The form of the remainder term in the N-dimensional Euler Maclaurin expansion is investigated. A concise formalism is developed for handling expressions which are lengthy to state using conventional notation. Conditions under which an integral representation involving only derivatives of the same total order with conventional kernel functions exists for the remainder term are derived.

1. Introduction

The Euler Maclaurin Summation formula is an expression for the difference between the $(m+1)$ point trapezoidal rule sum

$$Qf = \frac{1}{m}\sum_{j=0}^{m}{}'' f(j/m) \tag{1.1}$$

and the integral

$$If = \int_0^1 f(x)\,dx \tag{1.2}$$

to which the rule sum approximates. (The double prime next to the summation symbol in (1.1) is used to indicate that the extreme terms in the sum (in this case $j=0$ and $j=m$) are to be assigned a factor $\frac{1}{2}$.) The formula expresses this difference as a finite power series in $(1/m)$ together with a remainder term as follows:

$$\begin{aligned} Qf - If = {} & \frac{c_2}{m^2}\int_0^1 f''(x)\,dx + \frac{c_4}{m^4}\int_0^1 f^{(4)}(x)\,dx + \cdots \\ & + \frac{c_{2p-2}}{m^{2p-2}}\int_0^1 f^{(2p-2)}(x)\,dx + \frac{1}{m^{2p}}\int_0^1 g_{2p}(mx)\,f^{(2p)}(x)\,dx, \qquad p \geqq 1. \end{aligned} \tag{1.3}$$

Here c_s and $g_p(x)$ do not depend on m and are given explicitly by

$$\begin{aligned} c_{2s} &= B_{2s}/(2s)! \\ c_{2s+1} &= 0, \end{aligned} \tag{1.4}$$

$$g_p(x) = \left(B_p - \overline{B}_p(x)\right)/p! \tag{1.5}$$

* Work performed in part under the auspices of the U. S. Atomic Energy Commission.

** Part of this work was carried out by both authors at the University of N.S.W., Kensington, N.S.W., Australia.

where B_s is the Bernoulli number and $\overline{B}_p(x)$ the Bernoulli periodic function of period 1 which coincides with the Bernoulli polynomial $B_p(x)$ in the interval (0, 1) (see Abramowitz and Stegun [1]). The final term in (1.3) is known as the Poisson form of the remainder term. The kernel function $g_{2p}(mx)$ given by (1.5) is a spline polynomial of degree $2p$. We distinguish between the Euler Maclaurin summation formula, given above, and the Euler Maclaurin series, obtained from (1.3) by letting p become infinite. The Euler Maclaurin series is usually divergent.

Similar expansions exist (Lyness and Ninham [4]) in which Qf is replaced by a general quadrature rule and c_s and $g_p(x)$ are appropriately redefined.

These formulas play a fundamental role in the theory of one-dimensional Numerical Integration (Whittaker and Watson [6]). They may be used to provide the remainder term in Romberg Integration (Bulirsch [2]). They may be used to construct quadrature rules of particular polynomial degree. Generalisations of a more sophisticated nature which take into account algebraic singularities in $f(x)$ have been developed (Navot [5]). One of the important attributes of these formulas is the nature of the remainder term. This may be used in the above applications to bound the discretisation error or to show that the iterative Romberg procedure converges.

A fruitful approach to the theory of N-dimensional quadrature over a hypercube is by means of the N-dimensional analog of the Euler Maclaurin series. This corresponds to (1.3) above and may be written in the form:

$$\begin{aligned}\frac{1}{m^N}\sum_{j_1=0}^{m}{}''\sum_{j_2=0}^{m}{}''\cdots\sum_{j_N=0}^{m}{}'' f(j_1/m, j_2/m, \ldots, j_N/m)\\ -\int_0^1\int_0^1\cdots\int_0^1 f(\vec{x})\,d^N x=\sum_{s=1}^{p-1}\frac{1}{m^{2s}}\sum_{\sum s_j=s} c_{2s_1}c_{2s_2}\cdots c_{2s_N}\\ \cdot\int_0^1\int_0^1\cdots\int_0^1 f^{(2s_1,2s_2,\ldots,2s_N)}(\vec{x})\,d^N x+R_{2p}f.\end{aligned}\tag{1.6}$$

(The notation $\sum s_i=s$ indicates that the summation includes all sets of non-negative integers $s_1, s_2, \ldots, s_N$ whose sum is s. The notation $f^{(s_1,s_2,\ldots,s_N)}(\vec{x})$ stands for the mixed partial derivative of respective orders s_j in the j-th variable. $d^N x$ stands for the N-dimensional volume element.)

This equation is straightforward to derive, if one requires only the terms in the series (as specified above) and is content with the information that $R_{2p}f\sim O(m^{-2p})$. However, it is natural to enquire whether, as in the one-dimensional case, there is some simple integral representation for $R_p f$. In analogy with the terms in the series one might expect a form such as

$$R_{2p}f=\frac{1}{m^{2p}}\sum_{\sum i_j=2p}\int_0^1\int_0^1\cdots\int_0^1 g_{i_1,i_2,\ldots,i_N}(\vec{x})\,f^{(i_1,i_2,\ldots,i_N)}(\vec{x})\,d^N x\tag{1.7}$$

each kernel function $g(\vec{x})$ being defined explicitly in terms of c_s and $g_s(x)$.

This paper is concerned with the form of $R_{2p}f$. The principal difficulty with this investigation is that, at all stages, somewhat lengthy expressions occur. For example, for $N=2$, a typical, but not unique result is presented in conventional notation in Eqs. (4.5) and (4.6) below. Even this result is not complete, but refers only to p odd. This result is similar to one derived by Laurent [3] and equivalent to one derived by Bulirsch [2]. All these results, expressed in conventional notation, seem to take over half a journal page to state.

The other difficulty is, as the reader will note, that the generalisation to $N=3$ or 4 is not obvious. In fact the simplest forms are generally not symmetric under interchange of the variables. The effect of symmetrizing the results is to introduce even more terms into already lengthy expressions.

The main innovation in this paper is the introduction of a formalism, which employs a much more concise notation and which can be used to manipulate this type of expression. As an example, the content of Eq. (4.6) below is contained in Eq. (4.2) which is much shorter. In Sections 2 and 3, this formalism is developed, and in Section 4 it is applied to obtain specific results about the remainder term $R_{2p}f$.

2. One-Dimensional Operator Notation

We now introduce several linear operators and associated definitions. We work entirely with a set of orthogonal Cartesian coordinates $x^1, x^2, \ldots, x^N$; we confine our attention to the hypercube

$$H^N\colon\ 0 \leqq x^k \leqq 1, \qquad k=1, 2, \ldots, N \tag{2.1}$$

and functions $f(x^1, x^2, \ldots, x^N) \equiv f(\vec{x})$ whose partial derivatives of specified total order p are continuous in all variables within this hypercube. We denote this restriction by

$$f(\vec{x}) \in C^p[H^N]. \tag{2.2}$$

The appropriate value of p for each stated result is implied and not stated.

In this section we define a set of operators which operate in a single dimension x^k. These are indicated by a superscript $[k]$. Subsequently the superscript $[[k]]$ is used to indicate an operator which operates in the space spanned by $x^1, x^2, \ldots, x^k$.

Definitions.

$$Q^{[k]}f(x^1, x^2, \ldots, x^N) = \frac{1}{m}\sum_{j=0}^{m}{}'' f\left(x^1, x^2, \ldots, x^{k-1}, \frac{j}{m}, x^{k+1}, \ldots, x^N\right), \tag{2.3}$$

$$I^{[k]}f(x^1, x^2, \ldots, x^N) = \int_0^1 f(x^1, x^2, \ldots, x^N)\, dx^k. \tag{2.4}$$

It should be noted that $Q^{[k]}f(\vec{x})$ and $I^{[k]}f(\vec{x})$ are functions of $x^1, x^2, \ldots, x^{k-1}, x^{k+1}, \ldots, x^N$. We also require differential operators

Definitions.

$$D_w^{[k]} = \frac{g_w(m\,x^k)}{m^w}\left(\frac{\partial}{\partial x^k}\right)^w, \qquad w = 1, 2, \ldots, \tag{2.5}$$

$$\delta_w^{[k]} = \frac{c_w}{m^w}\left(\frac{\partial}{\partial x^k}\right)^w, \qquad w = 1, 2, \ldots \tag{2.6}$$

$$\delta_0^{[k]} = 1.$$

In terms of these quantities, the one-dimensional Euler Maclaurin summation formula (1.3) above has the form

$$Q^{[k]} f(x^k) - I^{[k]}\,\delta_0^{[k]} f(x^k) = I^{[k]}(\delta_1^{[k]} + \delta_2^{[k]} + \cdots + \delta_{p-1}^{[k]} + D_p^{[k]})\, f(x^k), \qquad p = 1, 2, \ldots \tag{2.7}$$

and the process of integration by parts, on which this formula is based, may be written

$$I^{[k]} D_w^{[k]} f(x^k) = I^{[k]}(\delta_w^{[k]} + D_{w+1}^{[k]})\, f(x^k), \qquad w = 1, 2, \ldots. \tag{2.8}$$

Definition.

$$D_0^{[k]} f(x^k) = \frac{1}{m}\sum_{j=0}^{m}{}'' \,\delta\left(x^k - \frac{j}{m}\right) f(x^k) = g_0(m\,x^k)\, f(x^k) \tag{2.9}$$

where $\delta(x)$ is the Dirac delta function. This is applied only in conjunction with the integration operator, giving

$$\begin{aligned} I^{[k]} D_0^{[k]} f(x^k) &= \frac{1}{m}\sum_{j=0}^{m}{}'' \int_0^1 \delta\left(x^k - \frac{j}{m}\right) f(x^k)\, dx^k \\ &= \frac{1}{m}\sum_{j=0}^{m}{}'' f\left(\frac{j}{m}\right) = Q^{[k]} f(x^k). \end{aligned} \tag{2.10}$$

Eq. (2.7) with $p = 1$ may be written in the form

$$I^{[k]} D_0^{[k]} f(x^k) = I^{[k]}(\delta_0^{[k]} + D_1^{[k]})\, f(x^k). \tag{2.11}$$

This somewhat artificial definition extends Eq. (2.8) to the case $w = 0$.

In the subsequent theory we used mixed operators of the type $D_{w_1}^{[1]}\,\delta_{w_2}^{[2]}\,\delta_{w_3}^{[3]} \ldots$. We define

Definition. An operator $X_{w_1}^{[1]} X_{w_2}^{[2]} \ldots X_{w_N}^{[N]}$ where the symbol X stands for either D or δ is of *order* w where

$$w = w_1 + w_2 + \cdots + w_N. \tag{2.12}$$

An operator which is the sum of several operators each of the same order w is also of order w.

The order of an operator corresponds to the order of the partial derivatives which it contains.

Finally we define an equivalence between different operators (denoted by $\equiv$).

Definition. Two operators A and B are *equivalent*, written

$$A \equiv B \tag{2.13}$$

if and only if

$$I^{[1]}I^{[2]} \dots I^{[N]} A f(\vec{x}) = I^{[1]}I^{[2]} \dots I^{[N]} B f(\vec{x}) \tag{2.14}$$

for all functions $f(\vec{x}) \in c^{\infty}[H^N]$.

This definition satisfies all the properties normally associated with equivalence relations.

In terms of these definitions, the one-dimensional Euler Maclaurin summation formula (1.3) or (2.7) may be expressed as a set of equivalences connecting operators of different orders as follows:

$$D_0^{[k]} \equiv \delta_0^{[k]} + \delta_1^{[k]} + \dots + \delta_{p-1}^{[k]} + D_p^{[k]}, \qquad p = 1, 2, 3, \dots . \tag{2.15}$$

Alternatively this may be written recursively in the form

$$D_w^{[k]} \equiv \delta_w^{[k]} + D_{w+1}^{[k]}, \qquad w = 0, 1, 2, \dots . \tag{2.16}$$

3. N-Dimensional Operators

We now apply the one-dimensional theory of Section 2 to the N-dimensional situation. In one-dimension the Euler Maclaurin summation formula may be expressed in the form

$$D_0^{[k]} \equiv \sum_{r=0}^{p-1} \delta_r^{[k]} + D_p^{[k]}.$$

In this section, the generalisation is effected in the form

$$\prod_{k=1}^{N} D_0^{[k]} \equiv \sum_{r=0}^{p-1} S_r^{[[N]]} + R_p^{[[N]]}. \tag{3.1}$$

This section is concerned with defining these operators, establishing this equivalence and investigating the structure of the remainder operator.

We now define an operator which corresponds to the term having coefficient $(1/m)^r$ in (1.6). This is an N-dimensional symmetric sum of order r of δ operators:

Definition.

$$S_r^{[[N]]} = \sum_{\Sigma r_i = r} \delta_{r_1}^{[1]} \delta_{r_2}^{[2]} \dots \delta_{r_N}^{[N]}. \tag{3.2}$$

As indicated the sum is over all sets of positive or zero integers r_i whose sum is r.

An alternative definition of the same sum is recursive in character. This is

$$S_i^{[[N+1]]} = \sum_{j=0}^{i} S_j^{[[N]]}\, \delta_{i-j}^{[N+1]} \tag{3.3}$$

where

$$S_i^{[[1]]} = \delta_i^{[1]}. \tag{3.3a}$$

It follows from (3.3) that

$$\sum_{i=0}^{p} S_i^{[[N+1]]} = \sum_{r=0}^{p} \sum_{j=0}^{p-r} S_r^{[[N]]}\, \delta_j^{[N+1]}. \tag{3.4}$$

This identity is required in a subsequent proof.

A second operator is required to describe the remainder term. This is defined recursively.

Definition.

$$R_p^{[[1]]} = D_p^{[1]}, \quad p = 0, 1, 2, \ldots, \tag{3.5a}$$

$$R_0^{[[N]]} = D_0^{[1]} D_0^{[2]} \ldots D_0^{[N]}, \quad N = 1, 2, 3, \ldots, \tag{3.5b}$$

$$R_{p+1}^{[[i+1]]} = R_{p+1}^{[[i]]} + R_p^{[[i]]} D_1^{[i+1]} + \sum_{r=0}^{p-1} S_r^{[[i]]} D_{p-r+1}^{[i+1]}, \quad p = 1, 2, 3, \ldots, \quad i = 2, 3, 4, \ldots. \tag{3.5c}$$

We note that in the above two definitions the order of each operator is correctly given by the subscript.

Theorem 1.

$$\prod_{k=1}^{N} D_0^{[k]} \equiv \sum_{r=0}^{p-1} S_r^{[[N]]} + R_p^{[[N]]}, \quad p = 0, 1, \ldots, \quad N = 1, 2, \ldots. \tag{3.6}$$

Proof. In the case $p = 0$, all N, relation (3.6) is true by definition (3.5b).

In the case $N = 1$, all p, relation (3.6) is true since, using (3.5a) and (3.3a) it is identical with

$$D_0^{[1]} \equiv \delta_0^{[1]} + \delta_1^{[1]} + \cdots + \delta_{p-1}^{[1]} + D_p^{[1]}.$$

This is the one-dimensional Euler Maclaurin summation formula (2.15).

We prove the general result using induction on N. We assume

$$\prod_{k=1}^{i} D_0^{[k]} \equiv \sum_{r=0}^{p-1} S_r^{[[i]]} + R_p^{[[i]]}, \quad p = 0, 1, 2, \ldots \tag{3.7}$$

and establish the same equivalence with $i+1$ replacing i. First, we note a trivial consequence of (3.7), namely

$$R_p^{[[i]]} \equiv S_p^{[[i]]} + R_{p+1}^{[[i]]}. \tag{3.8}$$

In the following manipulation, we use (2.16) and (3.8). It follows from (3.7) that

$$\begin{aligned}\prod_{k=1}^{i+1} D_0^{[k]} &\equiv \sum_{r=0}^{p-1} S_r^{[[i]]} D_0^{[i+1]} + R_p^{[[i]]} D_0^{[i+1]} \\ &\equiv \sum_{r=0}^{p-1} S_r^{[[i]]} \left(\sum_{j=0}^{p-r} \delta_j^{[i+1]} + D_{p-r+1}^{[i+1]} \right) + R_p^{[[i]]} (1 + D_1^{[i+1]}) \\ &\equiv \left[\sum_{r=0}^{p-1} S_r^{[[i]]} \sum_{j=0}^{p-r} \delta_j^{[i+1]} + S_p^{[[i]]} \right] \\ &\quad + \left[\sum_{r=0}^{p-1} S_r^{[[i]]} D_{p-r+1}^{[i+1]} + R_{p+1}^{[[i]]} + R_p^{[[i]]} D_1^{[i+1]} \right].\end{aligned} \tag{3.9}$$

Using (3.4) and (3.5c) the terms in square brackets here may be identified with the required operators, giving

$$\prod_{k=1}^{i+1} D_0^{[k]} \equiv \sum_{r=0}^{p} S_r^{[[i+1]]} + R_{p+1}^{[[i+1]]}, \qquad p = 0, 1, 2, \ldots . \tag{3.10}$$

This establishes (3.6) for $N = i+1$ and $p = 1, 2, \ldots$. The inductive proof of Theorem 1 follows immediately.

Theorem 1 indicates that the operator $R_p^{[N]}$ defined in (3.5) is indeed a remainder operator of order p. We now investigate further the structure of this operator. The result of this investigation is the following Theorem.

Theorem 2.

$$R_p^{[[N]]} \equiv \sum_{\Sigma i_j = p}^{*} X_{i_1}^{[1]} X_{i_2}^{[2]} \ldots X_{i_N}^{[N]} \tag{3.11}$$

where

(a) Each $X_{i_s}^{[s]}$ stands for either $\delta_{i_s}^{[s]}$ or for $D_{i_s}^{[s]}$.

(b) In each product at least one $X_{i_s}^{[s]}$ stands for $D_{i_s}^{[s]}$.

(c) When $p \geqq N$ every zero order factor $X_0^{[s]}$ stands for $\delta_0^{[s]}$.

When $p < N$, in any product at most $N-p$ zero order factors $X_0^{[s]}$ stand for $D_0^{[s]}$; all remaining zero order factors stand for $\delta_0^{[s]}$. (3.12)

We attach an asterisk to the summation symbol to indicate these restrictions in the elements in the sum. The notation $\sum i_j = p$ attached to the summation symbol indicates, as before, that the summation is over all sets of positive integer or zero i_j which sum to p.

The proof is by induction and is based principally on (3.5c); since

$$R_p^{[[1]]} \equiv D_p^{[1]}, \qquad p = 0, 1, 2, \ldots . \tag{3.13}$$

Eq. (3.11) is clearly correct in the case $N = 1$, and each of conditions (a), (b), and (c) holds.

We assume that (3.11) is valid for $N=1, 2, \ldots, j$ and $p=0, 1, 2, \ldots$ and we use (3.5c) to establish its validity for $N=j+1$, $p=1, 2, \ldots$. Its validity for $N=j+1$, $p=0$ follows from (3.5b). Substituting into (3.5c), both (3.2) and expression (3.11) with $N=i$, and introducing $\delta_0^{[j+1]}=1$, we find

$$\begin{aligned} R_{p+1}^{[[j+1]]} \equiv & \sum_{\Sigma ij=p+1}^{*} X_{i_1}^{[1]} X_{i_2}^{[2]} \ldots X_{ij}^{[j]} \delta_0^{[j+1]} \\ & + \sum_{\Sigma ij=p}^{*} X_{i_1}^{[1]} X_{i_2}^{[2]} \ldots X_{ij}^{[j]} D_1^{[j+1]} \\ & + \sum_{r=0}^{p-1} \sum_{\Sigma ij=r} \delta_{i_1}^{[1]} \delta_{i_2}^{[2]} \ldots \delta_{ij}^{[j]} D_{p-r+1}^{[j+1]}. \end{aligned} \tag{3.14}$$

We note that this is of precisely the required form and may be written

$$R_{p+1}^{[[j+1]]} \equiv \sum_{\Sigma ij=p+1} X_{i_1}^{[1]} X_{i_2}^{[2]} \ldots X_{i_{j+1}}^{[j+1]}, \tag{3.15}$$

where conditions (a) and (b) are clearly satisfied. Remembering that the asterisk indicates that at most $j-p$ factors $X_0^{[k]}$ in any term stand for $D_0^{[k]}$, the maximum number of factors $D_0^{[k]}$ in any term in (3.14) above is also $j-p$, which is $(j+1)-(p+1)$. Thus the sum in (3.15) satisfies all conditions (a), (b), and (c) and an asterisk may be added to the summation symbol in (3.15). The theorem is now established using the principle of mathematical induction.

In the theory of this section and the previous section we have made no explicit use of the circumstance that $c_{2s+1}=0$. This is in order to be able to generalise the theory subsequently to rules which are not symmetric. For all symmetric one-dimensional rules certain simplifications occur. In particular

$$\delta_{2s+1}^{[k]}=0. \tag{3.16}$$

Theorem 3. For a symmetric rule

$$R_{2p}^{[[N]]} \equiv \sum_{\Sigma ij=p}^{*} X_{2i_1}^{[1]} X_{2i_2}^{[2]} \ldots X_{2i_N}^{[N]} \tag{3.17}$$

where the $X_{i_s}^{[s]}$ and the summation are subject to the interpretation of the previous theorem.

Theorem 2 is a direct algebraic consequence of expressing $\Pi D_0^{[k]}$ in terms of δ and D operators using

$$D_m^{[k]} \equiv \delta_m^{[k]} + D_{m+1}^{[k]}. \tag{3.18}$$

If $\delta_{2s+1}^{[k]} \equiv 0$, it follows that

$$D_{2m}^{[k]} \equiv \delta_{2m}^{[k]} + D_{2m+2}^{[k]}. \tag{3.19}$$

Thus the proof of Theorem 2 (which includes the proof of Theorem 1 and the whole of this section) can be re-written, simply replacing each subscript m by subscript $2m$. The result of doing this is Theorem 3.

A direct proof is as follows: We write Eq. (3.5c) with $2p+1$ replacing p and substitute for $R_{2p+1}^{[[i]]}$ and for $D_1^{[i+1]}$ using (3.8) and (2.16) respectively. This gives

$$R_{2p+2}^{[[i+1]]} \equiv R_{2p+2}^{[[i]]} + (R_{2p}^{[[i]]} - S_{2p}^{[[i]]})\,(\delta_1^{[i+1]} + D_2^{[i+1]}) + \sum_{r=0}^{2p} S_r^{[[i]]}\, D_{2p-r+2}^{[i+1]}. \tag{3.20}$$

For symmetric rule operators $\delta_{2s+1}^{[k]} \equiv 0$ and so

$$S_{2s+1}^{[[i]]} \equiv 0 \tag{3.21}$$

and $S_{2r}^{[[i]]}$ contains only δ operators of even order. The sum over r may consequently be restricted to include only even values of r and the term with $\delta_1^{[i+1]}$ may be disregarded. It is clear that recursion relation (3.20) retains the even character of the remainder term, and that if $R_{2p}^{[[i]]}$ contains only even terms the same is true of $R_{2p+2}^{[[i+1]]}$. Since the result is true for $i=1$, the theorem follows by induction.

4. The Remainder Term in the N-Dimensional Euler Maclaurin Expansion

The theorems of Section 3 give correctly the form of $R_p^{[[N]]}$ for any value of N and of p. However, there are generally many different expressions all of which satisfy these conditions. In specific instances in which N is known, specific use of recurrence relation (2.16) has to be made to obtain a remainder term of the precise nature required.

When $N=2$, one convenient form of $R_{2p}^{[[2]]}$ is the following

$$R_{2p}^{[[2]]} \equiv \sum_{j=0}^{\frac{p-2}{2}} (\delta_{2j}^{[1]} D_{2(p-j)}^{[2]} + D_{2(p-j)}^{[1]} \delta_{2j}^{[2]}) + D_p^{[1]} D_p^{[2]}, \qquad p \text{ even}, \tag{4.1}$$

$$R_{2p}^{[[2]]} \equiv \sum_{j=0}^{\frac{p-3}{2}} (\delta_{2j}^{[1]} D_{2(p-j)}^{[2]} + D_{2(p-j)}^{[1]} \delta_{2j}^{[2]}) + \delta_{p-1}^{[1]} D_{p+1}^{[2]} + D_{p+1}^{[1]} D_{p-1}^{[2]}, \qquad p \text{ odd}. \tag{4.2}$$

This can be verified in a straightforward manner by establishing that

$$R_{2p+2}^{[[2]]} \equiv -S_{2p}^{[[2]]} + R_{2p}^{[[2]]} \tag{4.3}$$

using the relations listed in the previous section, together with the condition (for symmetric rules) that

$$\delta_{2s+1} \equiv 0. \tag{4.4}$$

The formula which results, after operating on $f(\vec{x})$ and integrating over the hypercube has been stated by Laurent [3] for p even. For p odd it is

$$\begin{aligned} &\frac{1}{m^2} \sum_{j_1=0}^{m}{}'' \sum_{j_2=0}^{m}{}'' f(j_1/m, j_2/m) - \int_0^1\int_0^1 f(x,y)\,dx\,dy \\ &\quad = \sum_{s=1}^{p-1} \frac{1}{m^{2s}} \sum_{s_1+s_2=s} c_{2s_1} c_{2s_2} \int_0^1\int_0^1 f^{(2s_1,2s_2)}(x,y)\,dx\,dy + R_{2p} f \end{aligned} \tag{4.5}$$

where

$$
\begin{aligned}
R_{2p}f = \frac{1}{m^{2p}} \sum_{\substack{j_1=0 \\ j_2=p-j_1}}^{\frac{p-3}{2}} \Bigg\{ & c_{2j_1} \int_0^1 \int_0^1 g_{2j_2}(m\,y)\, f^{(2j_1,2j_2)}(x,\,y)\,dx\,dy \\
& + c_{2j_1} \int_0^1 \int_0^1 g_{2j_2}(m\,x)\, f^{(2j_2,2j_1)}(x,\,y)\,dx\,dy \Bigg\} \\
& + \frac{1}{m^{2p}}\, c_{p-1} \int_0^1 \int_0^1 g_{p+1}(m\,y)\, f^{(p-1,p+1)}(x,\,y)\,dx\,dy \\
& + \frac{1}{m^{2p}} \int_0^1 \int_0^1 g_{p+1}(m\,x)\, g_{p-1}(m\,y)\, f^{(p+1,p-1)}(x,\,y)\,dx\,dy .
\end{aligned}
\tag{4.6}
$$

(The comparative brevity of (4.2), when compared with (4.6) illustrates an advantage of the compact notation for practical manipulation.)

This remainder term may be expressed in a simpler, but possibly useful form. Because $g_{2s}(m\,x)$ is of definite sign, and

$$
\int_0^1 g_s(m\,x)\,dx = c_s(x) \tag{4.7}
$$

we may apply the intermediate value theorem to obtain

$$
R_{2p}f = \frac{1}{m^{2p}} \sum_{\substack{j_1=0 \\ j_2=p-j_1}}^{p} c_{2j_1} c_{2j_2} f^{(2j_1,2j_2)}(\xi_{j_1},\,\xi_{j_2}) \tag{4.8}
$$

where $(\xi_{j_1},\,\xi_{j_2}) \in H^2$.

This result is also valid for even p. Essentially, in applying the intermediate value theorem, the same form of expression is obtained whether the constant c_{2s} or the function $g_{2s}(m\,x)$ occurs under the integral sign. The difference is of course that $\vec{\xi}$ is different.

In more than two dimensions, the number of individual terms in expressions such as (4.2) becomes large. Merely to write out an expression corresponding to (4.6) would be a major undertaking.

However, information of a general nature about the form of the remainder term can be gleaned from Theorems 2 and 3 without having to write down any long expression. The remainder term $R_{2p}f$ is $I^{[[N]]} R_{2p}^{[[N]]} f(\vec{x})$ where $R_{2p}^{[[N]]}$ is a sum of products specified by (3.11) or by (3.17). Moreover, since $X_i^{[k]}$ stands for either $\delta_i^{[k]}$ or for $D_i^{[k]}$, it follows that

$$
R_{2p}f = \frac{1}{m^{2p}} \sum_{\Sigma\, \mu = 2p} \int_0^1 \cdots \int_0^1 \phi_{i_1,i_2,\ldots,i_N}(\vec{x})\, f^{(i_1,i_2,\ldots,i_N)}(\vec{x})\, d^N x \tag{4.9}
$$

where the kernel functions have the form

$$
\phi_{i_1,i_2,\ldots,i_N}(\vec{x}) = h_{i_1}(x^1)\, h_{i_2}(x^2) \ldots h_{i_N}(x^N) \tag{4.10}
$$

these functions being given by

$$h_i(x) = c_i \quad \text{or} \quad h_i(x) = g_i(m\,x). \tag{4.11}$$

However, these kernel functions need not necessarily be conventional functions. Should one of the subscripts $i_j = 0$ and $h_0(x^j)$ be $g_0(m\,x^j)$, according to definition (2.9) $g_0(m\,x)$ is a sum of delta functions. For such a term, the integration over the x^j variable is replaced by a trapezoidal rule sum. This occurs if one of the terms in (3.11) or (3.17) includes $X_0^{[j]} = D_0^{[j]}$. The conditions under which this happens are described in Section (c) of Theorem 2. Applying this theorem and Theorem 3 we find the following information about possible analytic forms for $R_{2p}f$.

Theorem 4. (a) When $p \geqq N$, a form of (4.9) is possible with all kernel functions conventional and with the sum limited to only even values of i_j.

(b) When $2p \geqq N$, a form of (4.9) is possible with all kernel functions conventional.

(c) When $2p \geqq N$, a form of (4.9) is possible with the sum limited to only even values of i_j.

The form described in (b) is of course different from, but equal in value to, the form described in (c). In all other cases known to the authors, when the remainder term is expressed in the form (4.9), some delta function terms occur and also some non-zero terms having odd subscripts.

We conclude this section by applying the intermediate value theorem $2p$ times to each of the terms in (4.9). Whichever choice for $h_{2i}(x)$ is made in (4.11), it follows that $h_{2i}(x)$ is of definite sign and that

$$\int_0^1 h_{2i}(x)\,dx = c_{2i}, \qquad i = 0, 1, 2, \ldots .$$

Consequently the formal result of applying the intermediate value theorem is independent of the functional form of $h_i(x)$. (Different forms would give different values of $\vec{\xi}$.) The result is

Theorem 5. When $2p \geqq N$

$$R_{2p}f = \frac{1}{m^{2p}} \sum_{\sum i_j = p} c_{2i_1} c_{2i_2} \cdots c_{2i_N} f^{(2i_1, 2i_2, \ldots, 2i_N)}(\vec{\xi}_{2i_1, 2i_2, \ldots, 2i_N}) \tag{4.12}$$

where

$$\vec{\xi}_{2i_1, 2i_2, \ldots, 2i_N} \in H^N.$$

5. Concluding Remarks

The Euler Maclaurin expansion (in one dimension) is perhaps the most well known of many similar expansions for the error functional corresponding to a quadrature rule Qf. In general any quadrature rule which is exact for the constant function has an error expansion of the form

$$Qf - If = c_1 If' + c_2 If'' + \cdots + c_{2p-1} If^{(2p-1)} + I(g_{2p} f^{(2p)}).$$

The properties of the coefficients are discussed in Lyness and Ninham [4]. In this context, the results of this paper, which refer only to the product trapezoidal rule may be extended with appropriate modification to other product quadrature rules, and to N-dimensional rules in general since these are linear sums of one point product rules. In such an extension of the theory, the same formalism may be used to advantage, the values of individual coefficients being modified in accordance with the rule under consideration.

An investigation along these lines is in progress and the results will be reported subsequently.

References

1. Abramowitz, M., Stegun, I. A.: Handbook of mathematical functions. National Bureau of Standards, AMS-55, Washington, 1954.
2. Bulirsch, R.: Bemerkungen zur Romberg-Integration. Numer. Math. **6**, 6—16 (1966).
3. Laurent, P. J.: Formules de quadrature approchée sur domaines rectangulaires convergentes pour toute fonction intégrable Riemann. C.R.Acad. Sci. Paris **258**, 798—801 (1964).
4. Lyness, J. N., Ninham, B. W.: Asymptotic expansions and numerical quadrature. Math. Comp. **21**, 162—178 (1967).
5. Navot, I.: An extension of the Euler-Maclaurin summation formula to function with a branch singularity. J. Math. and Phys. **41**, 271—276 (1962).
6. Whittaker, E. T., Watson, G. N.: A course of modern analysis. New York: Cambridge University Press 1958.

J. N. Lyness
Applied Mathematics Division
Argonne National Laboratory
Argonne, Illinois 60439, USA

J. B. B. McHugh
Division of Animal Genetics
C.S.I.R.O.
Epping, N.S.W., Australia

Numer. Math. 15, 345–347 (1970)
© by Springer-Verlag 1970

Note on an Inverse Characteristic Value Problem

G. N. DE OLIVEIRA

Received December 2, 1968

Summary. Given an arbitrary $n \times n$ real matrix A and n real numbers $s_1, \ldots, s_n$, we study the problem of the existence of a diagonal matrix M such that the characteristic values of $A+M$ be $s_1, \ldots, s_n$.

1. Let $A=[a_{ij}]$ be an arbitrary real $n \times n$ matrix. Under what conditions does there exist a diagonal $n \times n$ matrix M such that $A+M$ has prescribed real characteristic values $s_1, \ldots, s_n$? To study this problem we can assume without loss of generality that $a_{jj}=0$ $(j=1, \ldots, n)$.

In [1] Hadeler has proved that if A is symmetric and

$$\min_{k \neq h} |s_k - s_h| > 0, \qquad \frac{\max_i \left\{ \sum_{j=1}^{n} a_{ij}^2 \right\}^{\frac{1}{2}}}{\min_{k \neq h} |s_k - s_h|} \leqq \frac{\sqrt{3}}{6},$$

then M exists. Here, by arguments similar to those of Hadeler we prove some new results.

Let

$$P_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|.$$

Theorem 1. *Suppose there exists a reordering $s_{m_1}, \ldots, s_{m_n}$ of the real numbers $s_1, \ldots, s_n$, such that*

$$|s_{m_i} - s_{m_j}| > 2(P_i + P_j) \quad (i, j=1, \ldots, n;\ i \neq j).$$

Then a diagonal matrix M exists such that $A+M$ has characteristic values $s_1, \ldots, s_n$. Moreover the i-th element of the diagonal of M can be chosen within the closed interval $[s_{m_i} - P_i, s_{m_i} + P_i]$.

Proof. Let $v_1, \ldots, v_n$ be real numbers lying in the closed intervals

$$[s_{m_1} - P_1, s_{m_1} + P_1], \ldots, [s_{m_n} - P_n, s_{m_n} + P_n]$$

respectively. Let v denote the vector $(v_1, \ldots, v_n)$ and V the diagonal matrix whose diagonal elements are $v_1, \ldots, v_n$. We denote the characteristic values of the matrix $A+V$ by $\lambda_1(v), \ldots, \lambda_n(v)$ and consider now the Geršgorin circles corresponding to this matrix. As is well known, if a Geršgorin circle corresponding

to a real matrix does not intersect the union of the remaining circles, then it contains one and only one characteristic value of the matrix and this characteristic value is real. It follows that for any v such that $s_{m_i} - P_i \leqq v_i \leqq s_{m_i} + P_i$ $(i=1, \ldots, n)$, $\lambda_k(v)$ $(k=1, \ldots, n)$ is real and each closed interval $[v_j - P_j, v_j + P_j]$ $(j=1, \ldots, n)$ contains one and only one characteristic value of $A+V$. We may assume that the numbers $\lambda_1(v), \ldots, \lambda_n(v)$ have been numbered in such a way that $\lambda_j(v)$ is the one lying in $[v_j - P_j, v_j + P_j]$. Let $\lambda(v) = (\lambda_1(v), \ldots, \lambda_n(v))$ and let T be the operator defined by

$$T(v) = v + s - \lambda(v)$$

where $s = (s_{m_1}, \ldots, s_{m_n})$.

The i-th coordinate of $T(v)$ varies over the closed interval $[s_{m_i} - P_i, s_{m_i} + P_i]$. Therefore $T(v)$ maps the cube $s_{m_i} - P_i \leqq v_i \leqq s_{m_i} + P_i$ $(i=1, \ldots, n)$ into a subset of itself. Using the Brouwer fixed point theorem we conclude that there exists a vector $v_0 = (v_1^0, \ldots, v_n^0)$ such that $s_{m_i} - P_i \leqq v_i^0 \leqq s_{m_i} + P_i$ $(i=1, \ldots, n)$ and $T(v_0) = v_0$. Therefore $\lambda(v_0) = s$. We can now take $M = \operatorname{diag}(v_1^0, \ldots, v_n^0)$.

Theorem 1 when used for symmetric matrices is not contained in Theorem 4 of [1] and conversely. For instance the matrix

$$A = \begin{bmatrix} 0 & 4 & 4 \\ 4 & 0 & 0 \\ 4 & 0 & 0 \end{bmatrix}$$

and the numbers $s_1 = -25$, $s_2 = 0$, $s_3 = 16.1$ satisfy the hypothesis of Theorem 1 above but do not satisfy the hypothesis of Theorem 4 in [1]. We can see that the converse is also true with the same matrix and $s_1 = -22.7$, $s_2 = 0$, $s_3 = 21.4$.

2. Let T be a real non-singular diagonal matrix. If D is a diagonal matrix of the same order as T we have $TDT^{-1} = D$. We see immediately that it is sufficient to show that the hypothesis of Theorem 1 is satisfied by the matrix TAT^{-1} for some non-singular diagonal matrix T. We have seen that the matrix A above and $s_1 = -22.7$, $s_2 = 0$, $s_3 = 21.4$ do not satisfy the hypothesis of Theorem 1. However if we take $T = \operatorname{diag}(\frac{3}{4}, 1, 1)$ we conclude, without using Hadeler's theorem, that a diagonal matrix M exists such that the characteristic values of $A+M$ are -22.7, 0, 21.4.

In particular, if A is a non-negative irreducible matrix, there exists a diagonal non-negative matrix T such that TAT^{-1} has all row sums equal to the dominant characteristic value of A, as is well known. Hence we have the following result.

Theorem 2. *Let A be a non-negative irreducible matrix with dominant characteristic value r. If the real numbers $s_1, \ldots, s_n$ are such that $|s_i - s_j| > 4r$ $(i, j = 1, \ldots, n;\ i \neq j)$, there exists a diagonal matrix M such that $A+M$ has the numbers $s_1, \ldots, s_n$ as characteristic values.*

Remark. In Theorem 1 and its consequences P_i can be replaced by

$$Q_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ji}| \quad (i=1, \ldots, n).$$

Theorem 3. *If M is a diagonal matrix such that $A+M$ has as characteristic values the real numbers $s_1, \ldots, s_n$, we have*

$$\sum_{i=1}^{n} s_i^2 - \operatorname{tr}(A^2) \geqq \frac{1}{n}\left(\sum_{i=1}^{n} s_i\right)^2$$

where tr *means trace.*

Proof. If $A+M$ has real characteristic values s_i, there exists an orthogonal matrix Q such that

$$Q(A+M)Q^T = R$$

where R is upper triangular and has diagonal elements s_i.

We have

$$R^2 = Q(A^2 + AM + MA + M^2)Q^T$$

giving

$$\operatorname{tr}(R^2) = \operatorname{tr}(A^2) + \operatorname{tr}(M^2)$$

since AM and MA have zero diagonals. Let $M = \operatorname{diag}(m_1, \ldots, m_n)$.

This equality implies

$$\begin{aligned}\sum_{i=1}^{n} s_i^2 &= \operatorname{tr}(A^2) + \sum_{i=1}^{n} m_i^2 \\ &\geqq \operatorname{tr}(A^2) + \frac{1}{n}\left(\sum_{i=1}^{n} m_i\right)^2 \\ &= \operatorname{tr}(A^2) + \frac{1}{n}\left(\sum_{i=1}^{n} s_i\right)^2.\end{aligned}$$

The author wishes to thank the referee for his suggestions, especially in the proof of Theorem 3.

Reference

1. Hadeler, K. P.: Ein Inverses Eigenwertproblem. Linear algebra and its applications, vol. 1, p. 83–101 (1968).

G. N. de Oliveira
Universidade de Coimbra
Instituto de Matemática
Coimbra, Portugal

Numer. Math. 15, 348–350 (1970)
© by Springer-Verlag 1970

Eingabelung der Lösungen von Integralgleichungen

D. MARSAL

Eingegangen am 2. Januar 1969

Das hier dargestellte Verfahren verallgemeinert die Methode der klassischen Iteration in konvergenzbeschleunigender Weise, wobei an den Kern leicht übersehbare, häufig erfüllte Ansprüche gestellt werden. Man kann mit Hilfe des Verfahrens entweder unter Verzicht auf Iteration schnell einen Überblick über den Verlauf der Lösung einer Integralgleichung erhalten oder mit einigen Iterationsschritten den Hauptanteil der Lösung abspalten und die Restfunktion eingabeln.

Die Untersuchung ist ganz auf die unmittelbare numerische Anwendung ausgerichtet, so daß z.B. auf die Einordnung des hier Dargestellten in die umfassenderen Betrachtungen von Collatz, Albrecht, Schröder, Walter und anderen Autoren (s. z.B. [1] und [2]) über Operatoren monotoner Art verzichtet wird.

Linearsysteme mit nichtnegativem Kern

Gegeben sei für irgendeinen Integralbegriff eine in allen Funktionen reellwertige Integralgleichung $y(X)=f(X)+\int K(X,T)\,y(T)\,dT$, wobei sich das Integral über ein (nicht notwendig beschränktes oder zusammenhängendes) Gebiet G des R_n erstrecke, X bzw. T für $(x_1,\ldots,x_n)$ bzw. $(t_1,\ldots,t_n)$ stehe und der Kern K für $X\in G$, $T\in G$ nichtnegativ sei. Die Integralgleichung besitze ferner mindestens eine nichtnegative Lösung (Fall I) bzw. nichtpositive Lösung (Fall II)*.

Weiter seien $\phi(X)$, $\Phi(X)$, $\psi(T)$ und $\Psi(T)$ vier nichtnegative Funktionen mit

$$(\phi,\psi)<1,\quad \phi(X)\,\psi(T)\leqq K(X,T)\leqq\Phi(X)\,\Psi(T),\quad (\Phi,\Psi)<1,^{**}$$

die zwei Funktionen ζ und ξ vermittels

$$\zeta(X)=f(X)+\frac{(\psi,f)}{1-(\psi,\phi)}\,\phi(X),\qquad \xi(X)=f(X)+\frac{(\Psi,f)}{1-(\Psi,\Phi)}\,\Phi(X)$$

definieren. Dann gilt für alle Lösungen $y(X)$ der Integralgleichung und alle $X\in G$ mit

Fall I

$$\mu(X)=\begin{cases}\zeta(X) & \text{für alle } x\in G \text{ mit } \zeta\geqq 0\\ 0 & \text{sonst,}\end{cases}$$

$$\zeta(X)\leqq\mu(X)\leqq y(X)\leqq\xi(X),$$

Fall II

$$\mu(X)=\begin{cases}\zeta(X) & \text{für alle } X\in G \text{ mit } \zeta\leqq 0\\ 0 & \text{sonst,}\end{cases}$$

$$\xi(X)\leqq y(X)\leqq\mu(X)\leqq\zeta(X).$$

* Wechselt die Lösung v der Integralgleichung $v=g+\int Kv$ das Vorzeichen und ist eine Unterfunktion $w\leqq v$ bzw. eine Oberfunktion $z\geqq v$ bekannt, so gilt $v=w+y$ mit $y\geqq 0$ bzw. $v=z+y$ mit $y\leqq 0$, und Einsetzen in die ursprüngliche Integralgleichung für v liefert eine neue für y mit demselben Kern.

** (α,β) ist das Skalarprodukt $\int_G \alpha(T)\,\beta(T)\,dT$. Wir setzen stets die Existenz aller Skalarprodukte voraus.

(Ist $\zeta<0$ im Fall I bzw. $\zeta>0$ im Fall II, so ist die Trivialschranke Null besser.) Setzt man $y(X)=\zeta(X)+y_1(X)$ bzw. $y(X)=\mu(X)+y_1(X)$ mit unbekannter, einzugabelnder Funktion y_1 in die ursprüngliche Integralgleichung ein, so erhält man eine Beziehung $y_1=f_1+\int K y_1$ mit $y_1\geqq 0$ (I) bzw. $\leqq 0$ (II), auf die man die obigen Relationen wieder anwenden kann usw.:

$$\begin{array}{lll} y=f+\int K y, & y\in[\mu,\xi], & y=\mu+y_1 \\ y_1=f_1+\int K y_1, & y_1\in[\mu_1,\xi_1], & y_1=\mu_1+y_2 \\ y_2=f_2+\int K y_2, & y_2\in[\mu_2,\xi_2], & y_2=\mu_2+y_3 \end{array} \tag{i}$$

usw. und damit

$$\begin{array}{ll} \text{(I):} & \mu\leqq\mu+\mu_1\leqq\mu+\mu_1+\mu_2\leqq\cdots\leqq y\leqq\min\{\xi,\mu+\xi_1,\mu+\mu_1+\xi_2,\ldots\} \\ \text{(II):} & \max\{\xi,\mu+\xi_1,\mu+\mu_1+\xi_2,\ldots\}\leqq y\leqq\cdots\leqq\mu_2+\mu_1+\mu\leqq\mu_1+\mu\leqq\mu. \end{array} \tag{ii}$$

Ersetzt man in (i) und (ii) μ und μ_i durch ζ bzw. ζ_i, so gilt (ii) im allgemeinen nicht mehr.

Gibt es kein Funktionspaar Φ, Ψ mit $K\leqq\Phi\Psi$ und $(\Phi,\Psi)<1$, so muß man sich mit einseitiger Annäherung begnügen, die wegen $0\leqq K$ stets möglich ist. Dabei entspricht Wahl von $\phi\equiv\psi\equiv 0$ und Benutzen der ζ-Relation der klassischen Iteration. Bei Volterraschen Integralgleichungen muß man $\phi\equiv\psi\equiv 0$ wählen, da ihr Kern für $x_i>t_i$ $(i=1,2,\ldots,n)$ verschwindet.

Man beachte, daß das Verfahren bei Benutzung der μ-Relation (auf Grund elementarer Eigenschaften monotoner Folgen) stets konvergiert; und zwar ist $\mu+\mu_1+\mu_2+\cdots\leqq y$ für alle $X\in G$ im Fall I und $y\leqq\mu+\mu_1+\mu_2+\cdots$ im Falle II.

Beweis. Wir führen ihn für die Oberfunktion ξ des Falles I. Wegen $K\leqq\Phi\Psi$ und $y\geqq 0$ ist $yK\leqq\Phi\Psi y$ und damit $y=f+\int K y\leqq f+\Phi(\Psi,y)$. Multiplikation mit Ψ ergibt $\Psi y\leqq\Psi f+\Psi\Phi(\Psi,y)$ und damit $(\Psi,y)\leqq(\Psi,f)+(\Psi,\Phi)(\Psi,y)$, und folglich $(\Psi,y)\leqq(\Psi,f)/\{1-(\Psi,\Phi)\}$; und Einsetzen dieser Ungleichung in $y\leqq f+\Phi(\Psi,y)$ liefert ξ.

Beispiel 1.

$$y(x)=x+\int_0^1 K(x,t)\,y(t)\,dt, \qquad K=(1+te^x)^{-1},$$

$$\phi\equiv 1, \qquad \psi=(1+et)^{-1}, \qquad (\phi,\psi)=(1/e)\ln(1+e)<1$$

$$\Phi\equiv 1, \qquad \Psi=(1+t)^{-1}, \qquad (\Phi,\Psi)=\ln 2<1,$$

$$f=x, \qquad x+e^{-1}\leqq y\leqq x+1, \qquad y=x+e^{-1}+y_1, \qquad 0\leqq y_1\leqq 0{,}632,$$

$$f_1\simeq 0{,}1945\,(1-x), \quad f+0{,}110\leqq y_1\leqq f_1+0{,}245, \quad y_1=f_1+0{,}110+y_2, \quad 0\leqq y_2\leqq 0{,}135,$$

$$f_2\simeq 0{,}0400\,(1-x), \quad f+0{,}023\leqq y_2\leqq f_2+0{,}051, \quad y_2=f_2+0{,}023+y_3, \quad 0\leqq y_3\leqq 0{,}028,$$

$$y=0{,}7354+0{,}7655\,x+y_3.$$

(f_1 und f_2 sind transzendente Ausdrücke, die sich gut linear approximieren lassen.)

Beispiel 2.

$$y(x) = -\ln x + \int_0^1 K(x,t)\, y(t)\, dt, \qquad K = (1+x e^t)^{-1},$$

$$\phi = (1+e x)^{-1}, \quad \psi \equiv 1, \quad \Phi = (1+x)^{-1}, \quad \Psi \equiv 1, \quad (\phi,\psi), (\Phi,\Psi) \text{ wie oben.}$$

$$-\ln x + \frac{e}{e-\ln(1+e)} \cdot \frac{1}{1+e x} \leqq y \leqq -\ln x + \frac{1}{1-\ln 2} \cdot \frac{1}{1+x}.$$

Eine Klasse Hammersteinscher Integralgleichungen

Gegeben sei die Integralgleichung $y(X) = f(X) + \int K(X,T)\, F\big(y(T)\big)\, dt$ mit nichtnegativem Kern K und der Eingabelung $\psi\phi \leqq K \leqq \Psi\Phi$, nichtnegativer, monoton mit y wachsender Funktion F und nichtnegativer Lösung y.

Besitzen die Ungleichungen

$$c \geqq \int \psi(T)\, F\big(f(T) + c\phi(T)\big)\, dt, \qquad C \leqq \int \Psi(T)\, F\big(f(T) + C\Phi(T)\big)\, dt$$

je eine Lösung $c \geqq c_0$, $C \leqq C_0$, so ist

$$f(X) + c_0 \phi(X) \leqq y(X) \leqq f(X) + C_0 \Phi(X).^*$$

Existiert mehr als eine Lösung c_0 bzw. C_0, so verwende man nach Möglichkeit $\max c_0$ sowie $\min C_0$.

Beweis. Wir führen ihn für die untere Schranke. Aus $\phi\psi \leqq K$ folgt $f + \phi \int \psi F \leqq f + \int KF = y$, also mit $c = \int \psi(T)\, F\big(y(T)\big)\, dT$ die Ungleichung $f + \phi c \leqq y$ und damit wegen der Monotonie von F auch $F(f + \phi c) \leqq F(y)$. Also ist

$$\psi(T)\, F\big(f(T) + c\phi(T)\big) < \psi(T)\, F\big(y(T)\big),$$

und beiderseitige Integration ergibt die Ungleichung für c, deren Lösung $c \geqq c_0$ sei. Dann ist $y \geqq f + \phi c \geqq f + \phi c_0$.

Beispiel. Es sei $y(x) = x + \int_0^1 y^2(t)\, dt/(1+t e^x)$. Wahl von Φ und Ψ wie beim ersten Beispiel führt auf $C \leqq \int_0^1 (t+C)^2 dt/(1+t)$ und damit auf $(C-\alpha)^2 \geqq \alpha^2 - \alpha$ mit $\alpha = 1 - (2 \cdot \ln 2)^{-1} \simeq 0{,}28$. Wegen $\alpha^2 - \alpha < 0$ liefert die Ungleichung $(C-\alpha)^2 \geqq 0$ den kleinstmöglichsten C_0-Wert, und zwar $\min C_0 = \alpha$, so daß wir, wenn wir uns mit $c_0 = 0$ begnügen, die Relation $x \leqq y \leqq x + 0{,}28$ erhalten.

Literatur

1. Collatz, L.: Funktionalanalysis und numerische Mathematik. Berlin-Heidelberg-New York: Springer 1968.
2. Walter, W.: Differential- und Integralungleichungen. Berlin-Göttingen-Heidelberg-New York: Springer 1964.

* An Stelle der Voraussetzung „$(\phi,\psi) < 1$, $(\Phi,\Psi) < 1$" ist also die Forderung der Lösbarkeit gewisser Ungleichungen getreten.

Dr. D. Marsal
3 Hannover
Hammersteinstraße 2

Numer. Math. 15, 351–353 (1970)
© by Springer-Verlag 1970

A Convergent Splitting of Matrices*

O. L. MANGASARIAN

Received March 27, 1969

Abstract. Let A, M, N be $n \times n$ real matrices, let $A = M - N$, let A and M be nonsingular, let $M'y \geqq 0$ imply $N'y \geqq 0$, and let $A'y \geqq 0$ imply $N'y \geqq 0$ (where the prime denotes the transpose). Then the spectral radius $\varrho(M^{-1}N)$ of $M^{-1}N$ is less than one, and the iterative process $x^{i+1} = M^{-1}Nx^i + M^{-1}b$ converges to the solution of $Ax = b$ starting from any x^0.

Consider the problem of finding an n-vector x which satisfies

$$Ax = b$$

where A is an $n \times n$ real matrix and b is an n-vector. Many iterative techniques for solving this problem such as the point Jacobi, Gauss-Seidel and successive overrelaxation methods [6, Chap. 3] can be obtained by splitting the matrix A into the difference of two $n \times n$ real matrices M and N and then using the iteration

$$x^{i+1} = M^{-1}Nx^i + M^{-1}b.$$

The convergence of these methods hinges upon the condition that the spectral radius $\varrho(M^{-1}N)$ of $M^{-1}N$ is less than one. The following well known theorem gives sufficient conditions for $\varrho(M^{-1}N) < 1$ and hence the convergence of the above iterative method.

Theorem 1 [6, p. 89]. Let A, M, N be $n \times n$ real matrices, let $A = M - N$, let A and M be nonsingular, and let $A^{-1} \geqq 0$, $M^{-1} \geqq 0$, $N \geqq 0$. Then $\varrho(M^{-1}N) < 1$.

The purpose of this work is to generalize the above theorem as follows:

Theorem 2. Let A, M, N be $n \times n$ real matrices, let $A = M - N$, let A and M be nonsingular, and let

$$M'y \geqq 0 \Rightarrow N'y \geqq 0$$

$$A'y \geqq 0 \Rightarrow N'y \geqq 0,$$

where the prime denotes the transpose. Then $\varrho(M^{-1}N) < 1$.

We remark immediately that Theorem 1 is included in Theorem 2 but not conversely. For suppose that the assumptions of Theorem 1 are satisfied, then $(M')^{-1} \geqq 0$, $(A')^{-1} \geqq 0$, $N \geqq 0$ and

$$M'y \geqq 0 \Rightarrow (M')^{-1}M'y \geqq 0 \Rightarrow y \geqq 0 \Rightarrow Ny \geqq 0,$$

$$A'y \geqq 0 \Rightarrow (A')^{-1}A'y \geqq 0 \Rightarrow y \geqq 0 \Rightarrow Ny \geqq 0.$$

* Sponsored by the Mathematics Research Center, United States Army, Madison, Wisconsin, under Contract No. DA-31-124-ARO-D-462, and in part by the National Science Foundation under Grant NSF GP-6070.

Hence the assumptions of Theorem 2 are satisfied. To see that Theorem 2 is not included in Theorem 1 consider the matrices

$$A=\begin{pmatrix}0 & 1\\ \frac{1}{2} & -1\end{pmatrix},\quad M=\begin{pmatrix}1 & 1\\ 0 & -1\end{pmatrix},\quad N=\begin{pmatrix}1 & 0\\ -\frac{1}{2} & 0\end{pmatrix}.$$

It is easy to verify that these matrices satisfy the assumptions of Theorem 2 but not the assumptions of Theorem 1 (because M^{-1} and N are not nonnegative). However

$$\varrho(M^{-1}N)=\varrho\begin{pmatrix}\frac{1}{2} & 0\\ \frac{1}{2} & 0\end{pmatrix}=\tfrac{1}{2}<1.$$

Proof of Theorem 2. The key of the proof is the following dual characterization of the matrix pairs (M, N) and (A, N) which was established in [2, 3] for the more general case of rectangular matrices with the aid of Motzkin's transposition theorem [4, 5, 1]

$$\langle M'y\geqq 0\Rightarrow N'y\geqq 0\rangle\Leftrightarrow\langle N=MX \text{ for some } n\times n \text{ real matrix } X\geqq 0\rangle,$$

$$\langle A'y\geqq 0\Rightarrow N'y\geqq 0\rangle\Leftrightarrow\langle N=AY \text{ for some } n\times n \text{ real matrix } Y\geqq 0\rangle.$$

For the special case here of square matrices this dual characterization can be proven in a very elementary manner[1]. The backward implication is evident. The forward implication is established as follows

$$M'(M')^{-1}=I\geqq 0\Rightarrow N'(M')^{-1}\geqq 0$$

$$\text{(since } M'y\geqq 0\Rightarrow N'y\geqq 0)$$

$$\Leftrightarrow N=MX \text{ for some } n\times n \text{ real matrix } X\geqq 0.$$

Similarly we establish that $N=AY$ for some $n\times n$ real matrix $Y\geqq 0$. It follows from this dual characterization that

$$X=M^{-1}N\geqq 0,$$

$$Y=A^{-1}N\geqq 0.$$

The remainder of the proof is essentially similar to that of Varga [6, pp. 88—89].

Let τ be an eigenvalue of $A^{-1}N$ with a corresponding eigenvector x. Then $Nx=\tau Ax$ and so $\tau Ax=\frac{\tau}{1+\tau}(A+N)x$ (since $\tau\neq -1$, otherwise $0=(N+A)x=Mx$, $x\neq 0$, which contradicts the nonsingularity assumption on M). Hence $Nx=\frac{\tau}{1+\tau}(A+N)x$ or $Nx=\frac{\tau}{1+\tau}Mx$, and $\frac{\tau}{1+\tau}$ is an eigenvalue of $M^{-1}N$ with a corresponding eigenvector x.

Conversely, let μ be an eigenvalue of $M^{-1}N$ with a corresponding eigenvector z. Then $Nz=\mu Mz$ and so $Nz=\frac{\mu}{1-\mu}(M-N)z$ (since $\mu\neq 1$, otherwise $0=(M-N)z=Az$, $z\neq 0$, which contradicts the nonsingularity assumption on A). Hence $Nz=\frac{\mu}{1-\mu}Az$, and $\frac{\mu}{1-\mu}$ is an eigenvalue of $A^{-1}N$ with a corresponding eigenvector z.

1 I am indebted to a referee for pointing out the fact that this part of the proof can be established without Motzkin's theorem and who supplied a proof different from the one given above.

Hence there is a one-to-one correspondence between the eigenvalues and eigenvectors of $A^{-1}N$ and $M^{-1}N$. Since both of these matrices are nonnegative it follows by the Frobenius theorem on nonnegative matrices [6, p. 46] that the spectral radius $\varrho(A^{-1}N)$ of $A^{-1}N$ is an eigenvalue of $A^{-1}N$ with a corresponding nonnegative nonvanishing eigenvector, and $\varrho(M^{-1}N)$ is an eigenvalue of $M^{-1}N$ with a corresponding nonnegative nonvanishing eigenvector.

We will show now that $\varrho(M^{-1}N)$ corresponds to a nonnegative eigenvalue of $A^{-1}N$. Since $\varrho(M^{-1}N)$ is an eigenvalue of $M^{-1}N$ it corresponds to a real eigenvalue τ of $A^{-1}N$. We will show that if $\tau<0$ a contradiction ensues. We have $\varrho(M^{-1}N)=\frac{\tau}{1+\tau}$ and $A^{-1}Nx=\tau x$. Since x is also an eigenvector of the nonnegative matrix $M^{-1}N$, $x\neq 0$ and $x\geqq 0$, by the Frobenius theorem. We have then $\tau x=A^{-1}Nx=Yx$, which is the desired contradiction since $\tau<0$, $x\geqq 0$, $x\neq 0$ and $Y\geqq 0$. Hence $\tau\geqq 0$ and $\varrho(M^{-1}N)=\frac{\tau}{1+\tau}$. Again, since $A^{-1}N=Y\geqq 0$, it follows that the largest nonnegative eigenvalue of $A^{-1}N$ is $\varrho(A^{-1}N)$, and since $\frac{\tau}{1+\tau}$ is increasing for $\tau\geqq 0$, we have that

$$\varrho(M^{-1}N)=\frac{\varrho(A^{-1}N)}{1+\varrho(A^{-1}N)}<1. \qquad \text{Q.E.D.}$$

Because of the equivalent dual characterization of the matrix pairs (M, N) and (A, N) we can state the following equivalent theorem to Theorem 2.

Theorem 2'. Let A, M, N be $n\times n$ real matrices, let $A=M-N$, let A and M be nonsingular and let

$$N=MX \text{ for some } n\times n \text{ real matrix } X\geqq 0,$$
$$N=AY \text{ for some } n\times n \text{ real matrix } Y\geqq 0.$$

Then $\varrho(M^{-1}N)<1$.

The conditions of Theorem 2' are equivalent to $A=M-N$, $M^{-1}N\geqq 0$ and $A^{-1}N\geqq 0$. Theorem 2' can also be established directly by the method of Varga [6, pp. 88—89]. However to establish Theorem 2, the dual representation of the matrix pairs (M, N) and (A, N) must be established first.

Bibliography

1. Mangasarian, O. L.: Nonlinear programming. New York: McGraw-Hill 1969.
2. — Characterizations of real matrices of monotone kind. SIAM Review **10**, 439—441 (1968).
3. — Perron-Frobenius properties of $Ax=\lambda Bx$. Mathematics Research Center, University of Wisconsin, Madison, Report No. 952, November 1968, forthcoming in J. Math. Anal. Appl.
4. Motzkin, T. S.: Beiträge zur Theorie der linearen Ungleichungen. Inaugural-Dissertation, University of Basel, Jerusalem, 1936.
5. — Two consequences of the transposition theorem of linear inequalities. Econometrica **19**, 184—185 (1951). (The first inequality in this paper must be reversed.)
6. Varga, R. S.: Matrix iterative analysis. Englewood Cliffs, New Jersey: Prentice-Hall 1962.

O. L. Mangasarian
Computer Science Department
University of Wisconsin
1210 West Dayton St. Madison
Wisconsin 53706, U.S.A.

Numer. Math. 15, 354–358 (1970)
© by Springer-Verlag 1970

On the Optimal Choice of Fourth-Order Runge-Kutta Formulas

H. A. LUTHER and H. G. SIERRA

Received July 4, 1969

Abstract. A criterion for optimum stability of Runge-Kutta formulas is suggested. It is shown, for the fourth-order case, that Kutta's formula is optimum.

1. Introduction

The fourth-order Runge-Kutta formulas considered are four stage, and of explicit type. The parameters can be chosen for optimum effects. Because the chief value of the method is its use in starting the linear multistep methods, these parameters are normally chosen so that truncation error is minimized. This aspect has been closely studied in [3, 7, 9], and is best exemplified by the Kuntzmann optimal formula [5].

Some studies exist involving stability [1, 2]. These are based on Kutta's formula [6], but seem to give no reason for the choice. It is shown below that Kutta's formula is optimum, according to a reasonable criterion.

The criterion used is capable of extension to Runge-Kutta methods of other orders.

2. Fundamental Relations

The differential system treated is written in the form

$$dy/dx = f(x, y), \quad y(x_0) = y_0, \quad x_0 \leqq x \leqq b, \tag{2.1}$$

where y and f are column matrices of length p.

To formulate the Runge-Kutta approach to numerical solution, for $i = 1, 2, 3, 4$, let

$$k_i(h, x, w) = f\left(x + a_i h, w + h \textstyle\sum_{j=1}^{i-1} b_{ij} k_j(h, x, w)\right). \tag{2.2.1}$$

Here $a_1 = 0$ and $\sum_{j=1}^{0} b_{1j} k_j(h, x, w) = 0$. Also let

$$F\{h, x, w; f\} = \textstyle\sum_{i=1}^{4} R_i k_i(h, x, w). \tag{2.2.2}$$

Using $x_n = x_0 + n h$, define $\{z_n\}$ by

$$z_{n+1} = z_n + h F\{h, x_n, z_n; f\}. \tag{2.3}$$

The as yet undetermined R_i, a_i, b_{ij} are scalar parameters related to each other according to equations to be given below. For convenience in presentation, round-off error has been omitted in the formulation of (2.3).

We suppose h so restricted that, for some positive integer N, $hN = b - x_0$. We suppose further that the function $f(x, w)$ occurring in (2.1) belongs to a class of sufficient differentiability, and that (2.3) defines $\{z_n\}$ in terms of z_0 for $1 \leqq n \leqq N$. In the numerical solution of (2.1), $z_0 = x_0$.

The relations that the parameters of (2.2) must satisfy are well known. For the stability discussion we need to know that $a_4 = 1$ and

$$\sum_{i=1}^{4} R_i = 1; \tag{2.4.1}$$

$$\sum_{i=2}^{4} a_i^j R_i = 1/(j+1), \quad j = 1, 2, 3; \tag{2.4.2}$$

$$a_2^j b_{32} R_3 + (a_2^j b_{42} + a_3^j b_{43}) R_4 = 1/[(j+1)(j+2)], \quad j = 1, 2, 3; \tag{2.4.3}$$

$$a_i = \sum_{j=1}^{i-1} b_{ij}, \quad i = 2, 3, 4; \tag{2.4.4}$$

$$a_2 b_{32} b_{43} R_4 = 1/24. \tag{2.4.5}$$

A solution of (2.4) is called singular if the matrix of coefficients of (2.4.2), considered as equations for R_2, R_3, R_4, is singular. The only singular solutions may be written

$$\begin{aligned} &a_2 = 1/2, \quad a_3 = 0, \quad R_2 = 2/3, \quad R_1 + R_3 = 1/6, \quad R_4 = 1/6, \quad b_{21} = 1/2, \\ &b_{31} = -1/[12R_3], \quad b_{32} = 1/[12R_3], \quad b_{41} = -1/2 - 6R_3, \\ &b_{42} = 3/2, \quad b_{43} = 6R_3; \end{aligned} \tag{2.5.1}$$

$$\begin{aligned} &a_2 = 1/2, \quad a_3 = 1/2, \quad R_1 = 1/6, \quad R_2 + R_3 = 2/3, \quad R_4 = 1/6, \quad b_{21} = 1/2, \\ &b_{31} = (3R_3 - 1)/[6R_3], \quad b_{32} = 1/[6R_3], \quad b_{41} = 0, \\ &b_{42} = 1 - 3R_3, \quad b_{43} = 3R_3; \end{aligned} \tag{2.5.2}$$

$$\begin{aligned} &a_2 = 1, \quad a_3 = 1/2, \quad R_1 = 1/6, \quad R_2 + R_4 = 1/6, \quad R_3 = 2/3, \quad b_{21} = 1, \\ &b_{31} = 3/8, \quad b_{32} = 1/8, \quad b_{41} = 1, \quad b_{42} = -1/[12R_4], \quad b_{43} = 1/[12R_4]. \end{aligned} \tag{2.5.3}$$

In (2.5.1) and (2.5.2), R_3 is a nonzero parameter; in (2.5.3) R_4 is a nonzero parameter.

For the nonsingular case,

$$6a_3a_2 - 4(a_2 + a_3) + 3 \neq 0; \quad a_2 \neq a_3; \quad a_2 \neq 0, 1/2, 1; \quad a_3 \neq 0, 1; \tag{2.6.1}$$

$$\begin{aligned} R_1 &= [6a_2a_3 - 2(a_2 + a_3) + 1]/[12a_2a_3]; \\ R_2 &= (2a_3 - 1)/[12a_2(a_3 - a_2)(1 - a_2)]; \\ R_3 &= (1 - 2a_2)/[12a_3(a_3 - a_2)(1 - a_3)]; \\ R_4 &= [6a_2a_3 - 4(a_2 + a_3) + 3]/[12(1 - a_2)(1 - a_3)]; \\ b_{32} &= a_3(a_3 - a_2)/[2a_2(1 - 2a_2)]; \\ b_{42} &= (1 - a_2)[-2 + a_2 + 5a_3 - 4a_3^2] \div \\ &\quad [2a_2(a_3 - a_2)\{6a_2a_3 - 4(a_2 + a_3) + 3\}]; \\ b_{43} &= (1 - 2a_2)(1 - a_2)(1 - a_3)/[a_3(a_3 - a_2)\{6a_2a_3 - 4(a_2 + a_3) + 3\}]. \end{aligned} \tag{2.6.2}$$

3. Stability Criteria

If u is a column matrix of length p, we write $u=({}^1u, {}^2u, \ldots, {}^pu)$ and use the norm $\|u\|=\max_i|{}^iu|$. Then, if for $1\leqq i, j\leqq p$,

$$|\partial^i f(x, w)/\partial^j w|\leqq M,$$

we readily find

$$\|F\{h, x, w; f\}-F\{h, x, v; f\}\|\leqq C\|w-v\| \tag{3.1.1}$$

where

$$C=pM\{C_1+hpMC_2+(hpM)^2C_3+(hpM)^3C_4\} \tag{3.1.2}$$

and

$$\begin{aligned} C_1&=|R_1|+|R_2|+|R_3|+|R_4|,\\ C_2&=|R_2|\cdot|b_{21}|+|R_3|(|b_{31}|+|b_{32}|)+|R_4|(|b_{41}|+|b_{42}|+|b_{43}|),\\ C_3&=|R_3|\,|b_{32}b_{21}|+|R_4|(|b_{42}b_{21}|+|b_{43}b_{31}|+|b_{43}b_{32}|),\\ C_4&=|R_4|\cdot|b_{43}b_{32}b_{21}|. \end{aligned} \tag{3.1.3}$$

This means that for suitable functions $f(x, w)$ we have stability, since for two sequences $\{z_n\}$ and $\{z'_n\}$ satisfying (2.3),

$$\|z_n-z'_n\|\leqq(1+hC)^n\|z_0-z'_0\|<\exp([b-a]C)\cdot\|z_0-z'_0\|,$$

and C is bounded above for h approaching zero.

Consider now the involvement in C of C_1, C_2, C_3 and C_4. It is seen that minimizing C_1 is most important, then if that be done, C_2 should be minimized if possible, and so on. It will be shown that

$$C_i\geqq 1/(i!), \quad i=1, 2, 3, 4,$$

and that these minima are all attained for Kutta's formula and no other formula of the type considered.

It may be that fourth-order Runge-Kutta is the highest order for which this type of statement is true. When a similar optimization is attempted for fifth-order explicit Runge-Kutta of usual type [4, 8], it can be shown that $\sum_{i=1}^{6}|R_i|$ can be 1, but that the expression corresponding to C_2 cannot be 1/2.

In what follows, it is assumed first of all that R_i, a_i, b_{ij} satisfy (2.4), hence that (2.5) and (2.6) apply. Then because of (2.4.1),

$$C_1=1 \quad \text{iff} \quad R_i\geqq 0, \quad i=1, 2, 3, 4. \tag{3.2}$$

Because of (2.4.4) and the first of (2.4.2), when $R_i>0$ for $i=2, 3, 4$

$$C_2=1/2 \quad \text{iff} \quad b_{ij}\geqq 0, \quad 2\leqq j<i\leqq 4. \tag{3.3.1}$$

The establishment of (3.3.1) also proves

$$C_1=1, \quad C_2=1/2 \quad \text{imply} \quad a_2>0 \quad \text{and} \quad a_3>0. \tag{3.3.2}$$

Using (2.5) and (2.6), we find

$$C_1=1, \quad C_2=1/2 \quad \text{imply} \quad R_2>0 \quad \text{and} \quad R_3>0. \tag{3.3.3}$$

For $R_2=0$ implies $a_3=1/2$ whence $b_{42}=-1/(2a_2)$. $R_3=0$ implies $a_2=1/2$.

It is now shown that

$$C_1 = 1, \quad C_2 = 1/2 \quad \text{imply} \quad a_3 < 1, \quad a_2 < 1. \tag{3.3.4}$$

From $R_3 b_{32} > 0$ we find $a_3 < 1$. Now suppose $a_2 > 1$. Then $a_3 - a_2 < 0$, and $R_4 b_{42} > 0$ implies $a_2 - 2 + 5a_3 - 4a_3^2 \leqq 0$. This means $0 < a_2 - 1 \leqq (4a_3 - 1)(a_3 - 1)$. Hence $4a_3 - 1 < 0$, and $2a_3 - 1 < 0$. But then $R_2 < 0$.

It can now be seen from (2.4.3) and (2.4.5) that

$$C_1 = 1, \quad C_2 = 1/2 \quad \text{imply} \quad C_3 = 1/6, \quad C_4 = 1/24. \tag{3.4}$$

Thus optimizing C_1 and C_2 also optimizes C_3 and C_4.

4. Uniqueness of the Kutta Formula

Note first that both (2.5.1) and (2.5.3) necessarily have a negative b_{ij}. Moreover, so does (2.5.2) unless $R_3 = 1/3$, leading to the Kutta formula. Turning to the nonsingular case, and using $R_3 > 0$ from (3.3.3), we have

$$0 < a_2 < 1/2 \quad \text{or} \quad 1/2 < a_2 < 1; \quad 0 < a_3 < 1/2 \quad \text{or} \quad 1/2 < a_3 < 1. \tag{4.1}$$

Because $R_4 > 0$ and $R_1 \geqq 0$,

$$6a_2a_3 - 4(a_2 + a_3) + 3 > 0, \quad 6a_2a_3 - 2(a_2 + a_3) + 1 \geqq 0. \tag{4.2}$$

In addition to (2.6) we have also (see (2.4.4))

$$\begin{aligned} b_{31} &= a_3(3a_2 - 4a_2^2 - a_3)/[2a_2(1 - 2a_2)] \\ b_{41} &= \phi/[2a_2a_3\{6a_2a_3 - 4(a_2 + a_3) + 3\}] \end{aligned} \tag{4.3}$$

where

$$\phi = 12a_2^2a_3^2 - 12a_2^2a_3 - 12a_2a_3^2 + 15a_2a_3 + 4a_2^2 + 4a_3^2 - 6a_2 - 5a_3 + 2.$$

Using $b_{31} \geqq 0$ and $b_{32} > 0$, we have $a_2 < 1/2$ and $3a_2 - 4a_2^2 \geqq a_3 > a_2$ or $a_2 > 1/2$ and $3a_2 - 4a_2^2 \leqq a_3 < a_2$. From (3.2) and $b_{42} \geqq 0$, we have $a_3 \geqq a_2$ and $a_2 \geqq 2 - 5a_3 + 4a_3^2$ or $a_3 < a_2$ and $a_2 \leqq 2 - 5a_3 + 4a_3^2$. From $b_{43} > 0$, we have $a_2 < 1/2$ and $a_3 < a_2$ or $a_2 > 1/2$ and $a_2 > a_3$. Note that $a_3 = 3a_2 - 4a_2^2$ and $a_2 = 2 - 5a_3 + 4a_3^2$ are tangent at $(1/2, 1/2)$. The conclusion thus far is that a solution (for the nonsingular case), if any, is for (a_2, a_3) in the region bounded by the parabolas and using only $1/2 < a_2 < 1$, $0 < a_3 < 1/2$.

Now include the regions described by (4.2). The result is a region contained in the square $0 < a_3 < 1/2$, $1/2 < a_2 < 1$, and bounded more particularly by the two parabolas and the two hyperbolas.

Finally consider $b_{41} \geqq 0$. This means $\phi \geqq 0$ (see (4.3)). Let $a_2 = 1/2 + u$, $a_3 = 1/2 - v$. Then $0 < u < 1/2$ and $0 < v < 1/2$, while

$$\phi = 12u^2v^2 + u^2 + v^2 - 3uv - u/2 - v/2. \tag{4.4}$$

But ϕ is a continuous function of u and v, and $\phi < 0$ for $u = v = 1/4$. Also, $\phi = 0$ implies

$$u = \{3v + 1/2 \pm [(3v + 1/2)^2 + 4v(1/2 - v)(12v^2 + 1)]^{1/2}\}/(24v^2 + 2).$$

But since $0<v<1/2$, the use of the minus sign requires $u<0$, while the use of the plus sign makes $u>(6v+1)/(24v^2+2)$. For $0\leqq v\leqq 1/2$, this last is a minimum of $1/2$, hence ϕ does not vanish in the region of interest. Thus also, in this region $b_{41}<0$.

References

1. Abdel Karim, A. I.: The stability of the fourth order Runge-Kutta method for the solution of systems of differential equations. Comm. ACM **9**, 113—116 (1966).
2. Carr, J. W.: Error bounds for the Runge-Kutta single-step integration process. J. ACM **5**, 39—44 (1958).
3. Hull, T. E., Johnston, R. L.: Optimum Runge-Kutta methods. Math. Comp. **18**, 306—310 (1964).
4. Konen, H. P., Luther, H. A.: Some singular explicit fifth order Runge-Kutta solutions. SIAM J. Numer. Anal. **4**, 607—619 (1967).
5. Kuntzmann, J.: Deux formules optimales du type de Runge-Kutta. Chiffres **2**, 21—26 (1959).
6. Kutta, W.: Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. Zeit. Math. Phys. **46**, 435—453 (1901).
7. Lotkin, M.: On the accuracy of Runge-Kutta's method. MTAC **5**, 128—132 (1951).
8. Luther, H. A.: Further explicit fifth-order Runge-Kutta formulas. SIAM Review **8**, 374—379 (1966).
9. Ralston, A.: Runge-Kutta methods with minimum error bounds. Math. Comp. **16**, 431—437 (1962); Corrigendum **17**, 488 (1963).

H. A. Luther and H. G. Sierra
Mathematics Department
Texas A & M University
College Station, Texas 77843

Numer. Math. 15, 359–370 (1970)
© by Springer-Verlag 1970

A Generalization of L-Splines*

TH. R. LUCAS

Received April 14, 1969

Abstract. A theory of generalized splines is developed for all regular formally self adjoint differential operators L with real coefficients. A special case of such operators are those which may be factored in the form $L=L_1^* L_1$, such as those related to the generalized splines of Ahlberg, Nilson, and Walsh [1, 2], and the L-splines of Schultz and Varga [6]. Theorems giving unique interpolation, integral relations, and convergence rates are established. If L has a certain positivity property, a useful extremal result is proven.

1. Introduction

For each positive integer i, let $K_2^i[a, b]$ denote the collection of all real-valued functions u defined on $[a, b]$ such that $u \in C^{i-1}[a, b]$, and $D^{i-1}u(x) \equiv u^{(i-1)}(x)$ is absolutely continuous, with $D^i u \in L^2[a, b]$. Let L be any $2n$-th order formally self adjoint differential operator with real coefficients. Then [4, p. 204], L can be written as

$$L[u(x)] = \sum_{i=0}^{n} (-1)^i D^i [a_i(x)\, D^i u(x)], \tag{1.1}$$

for any $u \in K_2^{2n}[a, b]$, with $a_i \in K_2^i[a, b]$, $0 \leqq i \leqq n$, $n \geqq 1$. We shall also assume there is an ω such that

$$a_n(x) \geqq \omega > 0 \quad \text{for all} \quad x \in [a, b]. \tag{1.2}$$

Let M be the corresponding bilinear functional defined over all $u, v \in K_2^n[a, b]$ by

$$M[u(x), v(x)] = \int_a^b \sum_{i=0}^{n} a_i(x) \left(D^i u(x)\right) \left(D^i v(x)\right) dx. \tag{1.3}$$

Now let π: $a = x_0 < x_1 < \cdots < x_N < x_{N+1} = b$ denote a partition of $[a, b]$, and $\boldsymbol{z} = (z_0, z_1, \ldots, z_{N+1})$ be a corresponding incidence vector; that is, a $N+2$-vector with integer components. It is required that $1 \leqq z_i \leqq n$ for $1 \leqq i \leqq N$, and that $0 \leqq z_i \leqq n$ for $i = 0$ or $N+1$. Let $\bar{\pi} = \max_{0 \leqq i \leqq N} (x_{i+1} - x_i)$ be the partition norm of π and define $\underline{\pi} = \min_{0 \leqq i \leqq N} (x_{i+1} - x_i)$. The domains of L and M may now be generalized to include functions in $\bigcap_{i=0}^{N} K_2^{2n}[x_i, x_{i+1}]$ and $\bigcap_{i=0}^{N} K_2^n[x_i, x_{i+1}]$, respectively.

* This research was supported in part by a NASA Traineeship, at the Georgia Institute of Technology.

Let $(u, v) = \int_a^b u(x)\, v(x)\, dx$ for all $u, v \in L^2[a, b]$. Then by repeated integration by parts, the following Green's type formulas may be derived:

$$(L[u], v) - M(u, v) = \sum_{i=0}^{N} \sum_{j=0}^{n-1} [D^j v(x)]\, H_{n,j}(u; x)\Big|_{x_i}^{x_{i+1}}, \tag{1.4}$$

for all $u \in \bigcap_{i=0}^{N} K_2^{2n}[x_i, x_{i+1}]$, $v \in \bigcap_{i=0}^{N} K_2^{n}[x_i, x_{i+1}]$ where

$$H_{n,j}(u; x) = \sum_{k=0}^{n-j-1} (-1)^{k+1} D^k [a_{k+j+1}(x)\, D^{k+j+1} u(x)]. \tag{1.5}$$

This yields immediately the result:

$$(L[u], v) - (u, L[v]) = \sum_{i=0}^{N} \sum_{j=0}^{n-1} [D^j v(x)\, H_{n,j}(u; x) - D^j u(x)\, H_{n,j}(v; x)]_{x_i}^{x_{i+1}}, \tag{1.6}$$

for all $u, v \in \bigcap_{i=0}^{N} K_2^{2n}[x_i, x_{i+1}]$.

Definition 1. A real-valued function $u(x)$ defined on $[a, b]$ is said to be $(\pi, \boldsymbol{z})$-*smooth* if and only if $u \in \bigcap_{i=0}^{N} K_2^{2n}[x_i, x_{i+1}]$ and

$$D^k u(x_i^-) = D^k u(x_i^+) \quad \text{for} \quad 0 \leqq k \leqq 2n - z_i - 1, \quad 1 \leqq i \leqq N. \tag{1.7}$$

A real-valued function $s(x)$ defined on $[a, b]$ is said to be a *L-spline* for π and $\boldsymbol{z}$ if and only if s is $(\pi, \boldsymbol{z})$-smooth and

$$L[s(x)] = 0 \quad \text{for almost all} \quad x \in [a, b]. \tag{1.8}$$

The class of all L-splines for fixed π and $\boldsymbol{z}$ is denoted by $Sp(L, \pi, \boldsymbol{z})$.

The components z_0 and z_{N+1} of $\boldsymbol{z}$ will play no role until later.

Note that if $L = D^{2n}$ and $\boldsymbol{z} = (n, 1, 1, \ldots, 1, n)$ then $Sp(L, \pi, \boldsymbol{z})$ is the space of *ordinary spline functions*. If $\boldsymbol{z} = (n, n, \ldots, n)$ and $L = D^{2n}$, then $Sp(L, \pi, \boldsymbol{z})$ is the Hermite space $H^{(n)}(\pi)$ of *piecewise-polynomial functions* described in [3]. More generally, suppose

$$L_1[u(x)] = \sum_{j=0}^{n} b_j(x)\, D^j u(x) \tag{1.9}$$

with $b_j \in K_2^n[a, b]$, $0 \leqq j \leqq n$, and $b_n(x) \geqq \omega_1 > 0$ for all $x \in [a, b]$. The formal adjoint of L_1 is defined by

$$L_1^*[u(x)] = \sum_{j=0}^{n} (-1)^j D^j [b_j(x)\, u(x)]. \tag{1.10}$$

Suppose there are $a_j \in K_2^j[a, b]$ such that

$$L_1^* L_1[u(x)] = \sum_{j=0}^{n} (-1)^j D^j [a_j(x)\, D^j [u(x)]]. \tag{1.11}$$

Then defining the operator on the right to be $L[u(x)]$ as in (1.1), the space $Sp(L, \pi, \boldsymbol{z})$ becomes the space of *generalized spline functions* $\hat{S}p(L_1, \pi, \boldsymbol{z})$ defined by Schultz and Varga [6].

In fact such $a_j(x)$ can be constructively determined from the $b_j(x)$, $0 \leqq j \leqq n$. Let $a_n(x) = b_n^2(x)$. Then $a_n(x) \geqq \omega_1^2 > 0$ and $L_1^* L_1[u(x)] - (-1)^n D^n[a_n(x) D^n u(x)]$ is a self adjoint operator of order no greater than $2n-1$, since the coefficient of $D^{2n}u$ is zero. But any formally self adjoint differential operator with real coefficients must be of even order. Define $(-1)^{n-1}a_{n-1}(x)$ to be the coefficient of $D^{2n-2}u$ in the above expression. Continuing in this fashion, define $a_j(x)$ to be the coefficient of $D^{2j}u$ in the expression

$$\overline{G}_j[u(x)] = (-1)^j \left[L_1^* L_1[u(x)] - \sum_{k=j+1}^{n} (-1)^k D^k [a_k(x) D^k u(x)] \right], \tag{1.12}$$

where $\overline{G}_j$ is a differential operator of order no greater than $2j$. Then $a_j(x) \in K_2^j[a, b]$, $0 \leqq j \leqq n$ and (1.11) holds. Moreover, if the b_j are periodic of order $n-1-z_0$ for $0 \leqq j \leqq n$, then a_j is periodic of order $j-1-z_0$ for $i+z_0 \leqq j \leqq n$.

Definition 2. Let $f \in K_2^n[a, b]$, L, and (π, z) be given. A real-valued function $u(x)$ is said to be a (π, z)-*interpolate of* f if and only if $u \in K_2^n[a, b]$ and

$$D^j u(x_i) = D^j f(x_i), \quad 0 \leqq j \leqq z_i - 1, \quad 1 \leqq i \leqq N, \tag{1.13}$$

and

$$\text{if } z_i \geqq 1, \text{ then } D^j u(x_i) = D^j f(x_i), \quad 0 \leqq j \leqq z_i - 1, \quad i = 0 \text{ and } N+1. \tag{1.14}$$

A real-valued function u is said to *satisfy end conditions with respect to* f

of Type I if $z_0 = z_{N+1} = n$;

of Type II if $H_{n,j}(u; x_i) = 0$, $z_i \leqq j \leqq n-1$, $i = 0$ and $N+1$;

of Type III if $H_{n,j}(u; x_i) = H_{n,j}(f; x_i)$, $z_i \leqq j \leqq n-1$, $i = 0$ and $N+1$;

of Type IV if
(i) $D^j u(a) = D^j u(b)$, $z_0 \leqq j < 2n - z_0 - 1$,
(ii) $D^j f(a) = D^j f(b)$, $0 \leqq j \leqq n-1$,
(iii) $D^j a_i(a) = D^j a_i(b)$, $0 \leqq j \leqq i - z_0 - 1$, $1 + z_0 \leqq i \leqq n$,
(iv) $z_0 = z_{N+1}$;

and, for all types, (1.14) is valid. A real-valued function $s(x)$ is said to be a $Sp(L, \pi, z)$-*interpolate of* f *of Type* U, $U \in \{\text{I, II, III, IV}\}$, if and only if $s \in Sp(L, \pi, z)$, s is a (π, z)-interpolate of f, and s satisfies end conditions of Type U with respect to f.

2. Existence and Uniqueness

The following theorem will aid significantly in establishing that *all* formally self adjoint operators L defined by (1.1) have a unique $Sp(L, \pi, z)$-interpolate for functions $f \in K_2^n[a, b]$ when the partition norm $\bar{\pi}$ is sufficiently small. It will also lead simply and directly to the establishment of some convergence theorems for $Sp(L, \pi, z)$-interpolates.

Theorem 1. Let L be defined by (1.1) where $a_i \in K_2^i[a, b]$, $0 \leqq i \leqq n$, and $a_n(x) \geqq \omega > 0$ for $x \in [a, b]$. Then there exist positive numbers M_j, $\overline{M}_j$, $0 \leqq j \leqq n-1$, so that if u is any (π, z)-interpolate of the zero function (with $N \geqq n$), then

i) $\|D^j u\|_{L^2[a,b]} \leqq (\bar{\pi})^{n-j} M_j \|D^n u\|_{L^2[a,b]}$, $\quad 0 \leqq j \leqq n-1$,

ii) $\|D^j u\|_{L^\infty[a,b]} \leqq (\bar{\pi})^{n-j-(\frac{1}{2})} \overline{M}_j \|D^n u\|_{L^2[a,b]}$, $\quad 0 \leqq j \leqq n-1$.

Moreover, there is an $\varepsilon > 0$ such that if $\bar{\pi} < \varepsilon$, then

iii) There is an $H > 0$ such that $\|D^n u\|^2_{L^2[a,b]} \leqq H^2 M(u,u)$, and

iv) If u is a $Sp(L, \pi, \mathfrak{z})$-interpolate of the zero function of Type I, II, III, or IV, then $u \equiv 0$.

Proof. Choose any partition π and corresponding incidence vector $\mathfrak{z}$. Assume N, the number of interior partition points, is at least n. Since $u(x_i) = 0$, $1 \leqq i \leqq N$, by repeated applications of Rolle's theorem, there exist points $\{\xi_l^{(k)}\}_{l=0}^{N-k-1}$, $0 \leqq k \leqq n-1$, such that

A) $D^k u(\xi_l^{(k)}) = 0, \quad 0 \leqq l \leqq N-k-1, \quad 0 \leqq k \leqq n-1,$

B) $a \leqq \xi_0^{(k)} < \xi_1^{(k)} < \cdots < \xi_{N-k-1}^{(k)} \leqq b, \quad 0 \leqq k \leqq n-1,$

C) $\xi_l^{(k)} < \xi_l^{(k+1)} < \xi_{l+1}^{(k)}, \quad 0 \leqq l \leqq N-k-2, \quad 0 \leqq k \leqq n-2.$

Define also $\xi_{-1}^{(k)} = a$ and $\xi_{N-k}^{(k)} = b$, $0 \leqq k \leqq n-1$.

By a Wirtinger-type inequality [5, p. 184],

$$\int_{\xi_l^{(k)}}^{\xi_{l+1}^{(k)}} (D^k u(x))^2 dx \leqq \left(\frac{2(k+1)\bar{\pi}}{\pi}\right)^2 \int_{\xi_l^{(k)}}^{\xi_{l+1}^{(k)}} (D^{k+1} u(x))^2 dx \tag{2.1}$$

since $\xi_{l+1}^{(k)} - \xi_l^{(k)} \leqq (k+1)\bar{\pi}$ for $-1 \leqq l \leqq N-k-1$, $0 \leqq k \leqq n-1$. Summing both sides of (2.1) from $l = -1$ to $N-k-1$, gives

$$\int_a^b (D^k u(x))^2 dx \leqq \left(\frac{2(k+1)\bar{\pi}}{\pi}\right)^2 \int_a^b (D^{k+1} u(x))^2 dx. \tag{2.2}$$

Letting k range from j to $n-1$, and using (2.2) repeatedly, yields

$$\int_a^b (D^j u(x))^2 dx \leqq \left(\frac{2^{n-j} n!}{j!\, \pi^{n-j}}\right)^2 (\bar{\pi})^{2(n-j)} \int_a^b (D^n u(x))^2 dx. \tag{2.3}$$

Now take the square root of both sides of (2.3) and let $M_j = \dfrac{2^{n-j} n!}{j!\, \pi^{n-j}}$. This establishes (i).

Now let y_j be such that $\|D^j u\|_{L^\infty} = |D^j u(y_j)|$, $0 \leqq j \leqq n-1$. Then

$$\|D^j u\|_{L^\infty} = \left| \int_{\xi_l^{(j)}}^{y_j} D^{j+1} u(x)\, dx \right| \leqq (j+1)^{\frac{1}{2}} (\bar{\pi})^{\frac{1}{2}} \|D^{j+1} u\|_{L^2},$$

where $\xi_l^{(j)}$ is the closest of $\{\xi_l^{(j)}: 0 \leqq l \leqq N-j-1\}$ to y_j, $0 \leqq j \leqq n-1$. But by (i), $\|D^{j+1} u\|_{L^2} \leqq \bar{\pi}^{n-j-1} M_{j+1} \|D^n u\|_{L^2}$. Then letting $\bar{M}_j = (j+1)^{\frac{1}{2}} M_{j+1}$, (set $M_n = 1$), (ii) is established.

Consider the identity

$$\int_a^b a_n(x) (D^n u(x))^2 dx = M(u,u) - \sum_{i=0}^{n-1} \int_a^b a_i(x) (D^i u(x))^2 dx.$$

Then

$$\omega \|D^n u\|_{L^2}^2 \leqq M(u,u) + \sum_{i=0}^{n-1} \|a_i(x)\|_{L^2} \|(D^i u(x))^2\|_{L^2}$$
$$\leqq M(u,u) + \sqrt{b-a} \sum_{i=0}^{n-1} \|a_i(x)\|_{L^2} \|D^i u(x)\|_{L^\infty}^2$$
$$\leqq M(u,u) + \sqrt{b-a} \sum_{i=0}^{n-1} \|a_i(x)\|_{L^2} \bar{\pi}^{2n-2i-1} \bar{M}_i^2 \|D^n u\|_{L^2}^2.$$

The last inequality follows from (ii), and the earlier ones are variations of Hölder's inequality. Consider the expression

$$G(\bar{\pi}) = \omega - \sqrt{b-a} \sum_{i=0}^{n-1} \|a_i(x)\|_{L^2} \bar{\pi}^{2n-2i-1} \bar{M}_i^2. \quad \text{If } \bar{\pi} \searrow 0, \quad G(\bar{\pi}) \nearrow \omega.$$

Therefore there is an $\varepsilon > 0$, and a $H_1 > 0$ so that if $\bar{\pi} \leqq \varepsilon$, then $G(\bar{\pi}) \geqq H_1$. Then $G(\bar{\pi}) \|D^n u\|_{L^2}^2 \leqq M(u,u)$, and so $\|D^n u\|_{L^2} \leqq M^{\frac{1}{2}}(u,u) H_1^{-\frac{1}{2}}$. Letting $H = H_1^{-\frac{1}{2}}$, (iii) is established for $(\bar{\pi}) \leqq \varepsilon$.

Now suppose u is a $Sp(L, \pi, z)$-interpolate of the zero function of Type I, II, III, or IV, with $\bar{\pi} < \varepsilon$. By (1.4), after rearranging terms,

$$(L[u], u) = M(u,u) - \sum_{i=0}^{N} \sum_{j=0}^{n-1} [D^j u(x)] H_{n,j}(u; x) \Big|_{x_i^-}^{x_i^+}$$
$$+ \sum_{j=0}^{n-1} [D^j u(x)] H_{n,j}(u; x) \Big|_a^b. \tag{2.4}$$

For end conditions of Type I, II, or III (Definition 2) the last sum is clearly zero. For end conditions of Type IV, $D^j u(a) = D^j u(b) = 0$ for $0 \leqq j \leqq z_0 - 1$. When $z_0 \leqq j \leqq n-1$, $H_{n,j}$ is a differential operator of order no greater than $2n - z_0 - 1$. Also $D^k u(a) = D^k u(b)$ for $0 \leqq k \leqq 2n - 1 - z_0$ and $D^k a_{k+j+1}(a) = D^k a_{k+j+1}(b)$ for $0 \leqq k \leqq n-j-1$, $z_0 \leqq j \leqq n-1$ (from Definition 2). So recalling (1.5), $H_{n,j}(u; x)\big|_a^b = 0$, $z_0 \leqq j \leqq n-1$. Therefore the second sum in (2.4) is zero. Using the facts that u has z_i zeros at x_i, that $D^k u(x_i^+) = D^k u(x_i^-)$ for $0 \leqq k \leqq 2n - z_i - 1$, and that $H_{n,j}$ is an operator of order no more than $2n - z_i - 1$ for $z_i \leqq j \leqq n-1$, and $1 \leqq i \leqq N$, it may similarly be shown that the first summation in (2.4) is zero. Therefore $(L[u], u) = M(u,u)$. Since $u \in Sp(L, \pi, z)$, $L[u] = 0$ almost everywhere, so $(L[u], u) = 0 = M(u,u)$. Suppose $\bar{\pi} < \varepsilon$. Then, from (iii), $\|D^n u\|_{L^2} \leqq H M^{\frac{1}{2}}(u,u) = 0$. Therefore, by (ii), $u \equiv 0$. Q.E.D.

Corollary. If u is a $Sp(L, \pi, z)$-interpolate of the zero function satisfying end conditions of Type I, II, III, or IV, then $(L[u], u) = M(u,u) = 0$.

Theorem 2. Let L, π, z, and $U \in \{\text{I, II, III, IV}\}$ be given. If u is a $Sp(L, \pi, z)$-interpolate of the zero function of Type U implies that $u \equiv 0$, then for all $f \in K_2^n[a,b]$ there exists a unique $Sp(L, \pi, z)$-interpolate of f of Type U.

Proof. Recalling (1.1) and the requirement that $a_i \in K_2^i[a,b]$, $0 \leqq i \leqq n$, it follows that $L[u(x)] = \sum_{j=0}^{2n} c_j(x) D^j u(x)$, where $c_j \in L^2[a,b]$, and $c_{2n} = (-1)^n a_n \in C[a,b]$. Then [4, p. 67], there exist $2n$ linearly independent functions $v_j \in K_2^{2n}[a,b]$, $1 \leqq j \leqq 2n$ such that $L[v_j(x)] = 0$ almost everywhere in $[a,b]$. Now if $s(x) \in$

$Sp(L, \pi, z)$, then s must be of the form $s(x) = \sum_{j=1}^{2n} \alpha_{ij} v_j(x)$ for $x \in [x_i, x_{i+1}]$, $0 \leqq i \leqq N$. Thus s is determined in $[a, b]$ by $2n(N+1)$ coefficients. That s be (π, z)-smooth places $\sum_{i=1}^{N} (2n - z_i)$ linear restraints on these coefficients. That s interpolates to f at the interior points through (1.13) adds $\sum_{i=1}^{N} z_i$ restraints. The end conditions of type U add a final $2n$ restraints. Thus in all there are $2n(N+1)$ linear restraints. Suppose the corresponding homogeneous system has a non-trivial solution. Then there would be a non-trivial $Sp(L, \pi, z)$-interpolate of the zero function, contrary to the hypothesis of the theorem. Thus the nonhomogeneous system always has a unique solution, which gives a unique $Sp(L, \pi, z)$-interpolate of f. Q.E.D.

Corollary. Let L and $U \in \{\text{I, II, III, IV}\}$ be given. Then there exists an $\varepsilon > 0$ such that if $\bar{\pi} < \varepsilon$ and z is an incidence vector corresponding to π, and $f \in K_2^n[a, b]$, then there is a unique $Sp(L, \pi, z)$-interpolate of f of Type U.

This follows immediately from Theorems 1 and 2, and for interpolation of Type I or Type IV, generalizes Theorem 4 of [1].

3. Integral Relations and Some Extremal Results

Theorem 3. Let $f \in K_2^n[a, b]$, π, and z be given. If $s(x)$ is a $Sp(L, \pi, z)$-interpolate of f of Type I, II, or IV, then the following *first integral relation* is valid:

$$M(f, f) = M(f - s, f - s) + M(s, s). \tag{3.1}$$

Proof. $M(f, f) = M(f - s + s, f - s + s)$

$$= M(f - s, f - s) + 2M(s, f - s) + M(s, s).$$

Thus it will suffice to show that $M(s, f - s) = 0$. From (1.4), letting $u = s$ and $v = f - s$,

$$M(s, f - s) = (L[s], f - s) - \sum_{i=0}^{N} \sum_{j=0}^{n-1} D^j [f(x) - s(x)] H_{n,j}(s; x)\Big|_{x_i}^{x_{i+1}}.$$

But $(L[s], f - s) = 0$ since $L[s]$ is zero almost everywhere, and the double sum on the right can be shown to be zero by the argument that follows Eq. (2.4), since s is a $Sp(L, \pi, z)$-interpolate of f of Type I, II, or IV. Q.E.D.

Theorem 4. Let $f \in K_2^{2n}[a, b]$, π, and z be given. If $s(x)$ is a $Sp(L, \pi, z)$-interpolate of f of Type I, III, or IV (and in the last case if $D^k f(a) = D^k f(b)$, $0 \leqq k \leqq 2n - z_0 - 1$), then the following *second integral relation* is valid:

$$M(f - s, f - s) = (f - s, L[f]). \tag{3.2}$$

Proof. With $u = f - s$ and $v = f - s$ in (1.4), there results

$$(L[f - s], f - s) = M(f - s, f - s) + \sum_{i=0}^{N} \sum_{j=0}^{n-1} D^j [f(x) - s(x)] H_{n,j}(f - s; x)\Big|_{x_i}^{x_{i+1}}.$$

But $(L[f-s], f-s) = (L[f], f-s)$ since $L[s]$ is zero almost everywhere, and the right-hand double sum is zero as in the proofs of Theorems 1 and 3 since s is a $Sp(L, \pi, \mathbf{z})$-interpolate of f of Type I, III, or IV. Q.E.D.

For a given L, π, and $\mathbf{z}$, it need not be the case that $M(u, u) \geqq 0$ for all $(\pi, \mathbf{z})$-smooth functions. Part (iii) of Theorem 1 demonstrates at least the existence of an $\varepsilon > 0$ such that if $\bar{\pi} < \varepsilon$ then $M(u, u) \geqq 0$ for all $(\pi, \mathbf{z})$-interpolates of the zero function. The following positivity condition will also be of use:

Definition 3. Let $(\pi, \mathbf{z})$ be given. A formally self adjoint differential operator L is said to have *property* (P, I) or *property* (P, IV) if $M(u, u) \geqq 0$ for all $u(x)$ which are $(\pi, \mathbf{z})$-smooth and satisfy end conditions of either Type I or Type IV, respectively, w.r.t. the zero function.

That there is a large class of operators L with property (P, I) or property (P, IV) is demonstrated by the next theorem which includes all of the operators of [1] and [6] in this class.

Theorem 5. Let $(\pi, \mathbf{z})$ be given. If $L_1[u(x)] = \sum_{i=0}^{n} b_i(x)\, D^i u(x)$ with $b_i \in K_2^n[a, b]$ and $b_n(x) \geqq \omega_1 > 0$, then $L \equiv L_1^* L_1$ has property (P, I). If in addition the b_i's are periodic of order $n - z_0 - 1$, $0 \leqq i \leqq n$, then L has property (P, IV).

Proof. It was shown in Section 1, that there exist $a_i \in K_2^i[a, b]$ with $a_n(x) \geqq \omega_1^2 = \omega > 0$ for $x \in [a, b]$ such that $L[u] = \sum_{i=1}^{n} (-1)^i D^i [a_i D^i u]$, and that moreover the a_i can be constructively determined from the b_i by the formula

$$a_j = \text{the coefficient of } D^{2j}u \text{ in the } 2j\text{-th order (at most) expression}$$

$$(-1)^j \left\{ L_1^* L_1[u(x)] - \sum_{k=j+1}^{n} (-1)^k D^k [a_k(x)\, D^k u(x)] \right\}, \tag{3.3}$$

$$j = n-1, n-2, \ldots, 0,$$

and $a_n(x) = b_n^2(x)$.

Suppose u is $(\pi, \mathbf{z})$-smooth and satisfies end conditions of Type I or IV w.r.t. $f \equiv 0$. Then from (1.4), after rearranging terms as in (2.4), and making use of the end conditions of Type I or IV, there results

$$(L[u], u) = M(u, u) - \sum_{i=1}^{N} \sum_{k=0}^{n-1} [D^k u(x)]\, H_{n,k}(u; x)\Big|_{x_i^-}^{x_i^+}. \tag{3.4}$$

Also, for any $v_1, v_2 \in K_2^n[\alpha, \beta]$, [4, p. 86],

$$\int_{\alpha}^{\beta} v_2 L_1^*[v_1] - v_1 L_1[v_2]\, dx = \sum_{j=0}^{n-1} [D^j v_2(x)]\, \beta_{n,j}(v_1; x)\Big|_{\alpha}^{\beta}, \tag{3.5}$$

where

$$\beta_{n,j}(v_1; x) = \sum_{k=0}^{n-j-1} (-1)^{k+1} D^k [b_{k+j+1}(x)\, v_1(x)]. \tag{3.6}$$

Thus for $v_1 = L_1[u]$ and $v_2 = u$, (3.5) gives

$$(L_1^* L_1[u], u) = (L_1[u], L_1[u]) + \sum_{i=0}^{N} \sum_{j=0}^{n-1} [D^j u(x)]\, \beta_{n,j}(L_1[u]; x)\Big|_{x_i}^{x_{i+1}}. \tag{3.7}$$

After rearranging terms in (3.7), and using the condition that u satisfies end conditions of Type I or IV w.r.t. zero, there results

$$(3.8)\qquad (L_1^* L_1[u], u) = (L_1[u], L_1[u]) - \sum_{i=1}^{N} \sum_{j=0}^{n-1} [D^j u(x)]\, \beta_{n,j}(L_1[u]; x)\Big|_{x_i^-}^{x_i^+}.$$

But $L_1^* L_1 \equiv L$, so if the terms on the right in (3.4) and (3.8) can be shown to be equal it will follow that $M(u, u) = (L_1[u], L_1[u]) \geqq 0$, and the theorem will be proven.

Since u is (π, z)-smooth, $u \in \left(\bigcap_{i=0}^{N} K_2^{2n}[x_i, x_{i+1}]\right) \cap K_2^n[a, b]$, regardless of the incidence vector z. Define

$$(3.9)\qquad G_j(u; x) = \beta_{n,j}(L_1[u]; x) - H_{n,j}(u; x), \quad \text{for } 0 \leqq j \leqq n-1.$$

It is evident that to show $\sum_{i=1}^{n} \sum_{j=0}^{n-1} [D^j u(x)]\, G_j(u; x)\Big|_{x_i^-}^{x_i^+} = 0$, it suffices to show that $G_j(u; x)$ is continuous at all interior node points, for $0 \leqq j \leqq n-1$.

From (3.9), (3.6), and (1.5),

$$(3.10)\qquad G_j(u; x) = -\sum_{k=0}^{n-j-1} (-1)^k D^k \{b_{k+j+1}(x)\, L_1[u(x)] - a_{k+j+1}(x)\, D^{k+j+1} u(x)\}.$$

Or, letting $l = k + j + 1$, and operating on (3.10) by D^{j+1}

$$(3.11)\qquad D^{j+1} G_j(u; x) = (-1)^j \sum_{l=j+1}^{n} (-1)^l D^l \{b_l(x)\, L_1[u(x)] - a_l(x)\, D^l u(x)\},$$

and from (1.10), (3.11) becomes

$$(3.12)\qquad \begin{aligned} D^{j+1} G_j(u; x) = (-1)^j &\left\{ L_1^* L_1[u(x)] - \sum_{l=j+1}^{n} (-1)^l D^l [a_l(x)\, D^l u(x)] \right\} \\ &- (-1)^j \sum_{l=0}^{j} (-1)^l D^l \big(b_l(x)\, L_1[u(x)]\big). \end{aligned}$$

The last term in (3.12) contains derivatives of order up to $n + j$, while the first term on the right, by (3.3), contains after expanding, derivatives no higher than $2j$. Therefore G_j is a differential operator of order $n-1$, and hence when applied to functions $u \in K_2^n[a, b]$, gives a continuous function. Q.E.D.

For operators which have property (P, I) or property (P, IV), the following extremal theorem holds:

Theorem 6. Let $f \in K_2^n[a, b]$, π, and z be given. If L has property $(\mathrm{P, I(IV)})$, and if s_f is a $Sp(L, \pi, z)$-interpolate of f of Type I (IV), then

$$(3.13)\qquad \begin{aligned} M(f - s_f, f - s_f) = \min \{ &M(f - s, f - s): s \in Sp(L, \pi, z) \\ &\text{and } s \text{ satisfies end conditions of} \\ &\text{Type I (IV) with respect to } f \}. \end{aligned}$$

Proof. In Theorem 3 replace f by $f - s$ and s by $s_f - s$. Then $M(f-s, f-s) = M(f-s_f, f-s_f) + M(s_f - s, s_f - s)$. But $M(s_f - s, s_f - s) \geqq 0$ since $s_f - s$ satisfies

end conditions of Type I (IV) w.r.t. zero and L has property $(P, \text{I (IV)})$. Therefore $M(f-s, f-s) \geqq M(f-s_f, f-s_f)$ and the minimum is clearly achieved for $s = s_f$. Q.E.D.

The following theorem is a generalization of a result from [1], as applied to the functional $M(u, u)$.

Theorem 7. Let L and ε be given as in Theorem 1, $U \in \{\text{I, II, IV}\}$, $f \in K_2^n[a, b]$ with f periodic of order $n-1$ if $U = \text{IV}$, and let $(\pi, \boldsymbol{z})$ be such that $\bar{\pi} < \varepsilon$ (and $N \geqq n$). Then there is a unique $S\mathfrak{p}(L, \pi, \boldsymbol{z})$-interpolate of Type U to f, f_U, and

$$M(f_U, f_U) = \min\{M(g, g)\colon\ g \text{ is a } (\pi, \boldsymbol{z})\text{-interpolate of } f \text{ and if } U = \text{IV},\ g \text{ is periodic of order } n-1\}.$$

Moreover $M(g, g) > M(f_U, f_U)$ if $g \neq f_U$, and g is a $(\pi, \boldsymbol{z})$-interpolate of f.

Proof. The existence and uniqueness of f_U follows directly from Theorems 1 and 2. By Definition 2, if g is any $(\pi, \boldsymbol{z})$-interpolate of f then f_U will also be the $S\mathfrak{p}(L, \pi, \boldsymbol{z})$-interpolate of g of Type U. Therefore by the first integral relation,

$$M(g, g) = M(g-f_U, g-f_U) + M(f_U, f_U). \tag{3.14}$$

But $g-f_U$ is a $(\pi, \boldsymbol{z})$-interpolate of the zero function. Therefore by part (iii) of Theorem 1, $M(g-f_U, g-f_U) \geqq H^{-2}\|D^n(g-f_U)\|_{L^2} \geqq 0$ so $M(g, g) \geqq M(f_U, f_U)$. If $M(g, g) = M(f_U, f_U)$, then by (3.14), $M(g-f_U, g-f_U) = 0$, so $D^n(g-f_U) = 0$ almost everywhere. Since $D^{n-1}(g-f_U)$ is absolutely continuous, $D^{n-1}(g-f_U)$ must be a constant and hence $g-f_U$ is a polynomial of degree no greater than $n-1$. But since $g-f_U$ has at least n zeros in $[a, b]$, $g-f_U = 0$ in $[a, b]$. So if $g \neq f_U$ then $M(g, g) > M(f_U, f_U)$. Q.E.D.

4. Convergence

Let $\{\pi_i\}_{i=1}^{\infty}$ be any sequence of partitions of $[a, b]$, with $\lim_{i\to\infty} \bar{\pi}_i = 0$. It is then possible to prove a variety of convergence results which, in the case of Type I or Type IV interpolation, generalize Theorems 1, 2, and 3 of [2], and in the case of Type I interpolation, generalize Theorem 9 of [3], and Theorems 6, 7, 8, 9, and 10 of [6].

Theorem 8. Let $f \in K_2^n[a, b]$, let $\{\pi_i\}_{i=1}^{\infty}$ be a sequence of partitions of $[a, b]$ with $\lim_{i\to\infty} \bar{\pi}_i = 0$, and let $\{\boldsymbol{z}^{(i)}\}_{i=1}^{\infty}$ be any sequence of incidence vectors associated with $\{\pi_i\}_{i=1}^{\infty}$. Then there exists a positive integer i_0 such that the $S\mathfrak{p}(L, \pi_i, \boldsymbol{z}^{(i)})$-interpolate $s_i(x)$ of f of Type I, II, III, or IV exists and is unique for any $i \geqq i_0$, and moreover there exist constants $T_{1,j}$, $0 \leqq j \leqq n$, and $T_{2,j}$, $0 \leqq j \leqq n-1$, independent of f and i, such that

$$\text{(4.1)}\quad \text{i)}\ \|D^j(f-s_i)\|_{L^2[a,b]} \leqq T_{1,j}(\bar{\pi}_i)^{n-j}\{M(f-s_i, f-s_i)\}^{\frac{1}{2}}, \qquad 0 \leqq j \leqq n,$$

and

$$\text{(4.2)}\quad \text{ii)}\ \|D^j(f-s_i)\|_{L^\infty[a,b]} \leqq T_{2,j}(\bar{\pi}_i)^{n-j-(\frac{1}{2})}\{M(f-s_i, f-s_i)\}^{\frac{1}{2}}, \qquad 0 \leqq j \leqq n-1,$$

and for any $i \geqq i_0$.

Proof. Let ε be determined by Theorem 1. Since $\bar{\pi}_i \to 0$ there is an i_0 such that $i \geqq i_0$ implies $\bar{\pi}_i < \varepsilon$ and $N_i \geqq n$. Then by part (iv) of Theorem 1 and Theorem 2 there is a unique $Sp(L, \pi_i, z^{(i)})$-interpolate s_i of f for $i \geqq i_0$. Letting $T_{1,j} = HM_j$ and $T_{2,j} = H\bar{M}_j$, where M_j, $\bar{M}_j$ and H are given by parts (i), (ii), and (iii) of Theorem 1, $0 \leqq j \leqq n-1$, and $T_{1,n} = H$, the remainder of Theorem 8 follows immediately from Theorem 1, since $f - s_i$ is a $(\pi_i, z^{(i)})$-interpolate of zero for $i \geqq i_0$. Q.E.D.

The next theorem gives a considerable improvement in the rate of convergence, for the case where $f \in K_2^{2n}[a, b]$.

Theorem 9. Let $f \in K_2^{2n}[a, b]$, let $\{\pi_i\}_{i=1}^{\infty}$ be a sequence of partitions of $[a, b]$ with $\lim_{i \to \infty} \bar{\pi}_i = 0$, and let $\{z^{(i)}\}_{i=1}^{\infty}$ be any sequence of incidence vectors associated with $\{\pi_i\}_{i=1}^{\infty}$. Then there exists a positive integer i_0 such that the $Sp(L, \pi_i, z^{(i)})$-interpolate s_i of f of Type I, III, or IV exists and is unique for any $i \geqq i_0$, and moreover there exist constants $T_{3,j}$, $0 \leqq j \leqq n$, and $T_{4,j}$, $0 \leqq j \leqq n-1$, independent of f and i, such that

$$\text{(4.3)} \qquad \text{i)} \quad \|D^j(f - s_i)\|_{L^2[a,b]} \leqq T_{3,j}(\bar{\pi}_i)^{2n-j} \|L[f]\|_{L^2[a,b]}, \qquad 0 \leqq j \leqq n,$$

and

$$\text{(4.4)} \qquad \text{ii)} \quad \|D^j(f - s_i)\|_{L^\infty[a,b]} \leqq T_{4,j}(\bar{\pi}_i)^{2n-j-(\frac{1}{2})} \|L[f]\|_{L^2[a,b]}, \qquad 0 \leqq j \leqq n-1,$$

and for any $i \geqq i_0$.

Proof. Choose i_0 as in Theorem 8, establishing the first part of the Theorem. In the case of Type IV interpolation it will also be assumed that $D^k f(a) = D^k f(b)$, $0 \leqq k \leqq 2n - z_0^{(i)} - 1$, $i \geqq i_0$. Then by the second integral relation (Theorem 4), $M(f - s_i, f - s_i) = (f - s_i, L[f])$, for $i \geqq i_0$. Applying the Schwarz inequality to this gives

$$M(f - s_i, f - s_i) \leqq \|f - s_i\|_{L^2[a,b]} \|L[f]\|_{L^2[a,b]}.$$

Then (4.1), with $j = 0$, gives $\|f - s_i\|_{L^2[a,b]} \leqq T_{1,0}(\bar{\pi}_i)^n \{M(f - s_i, f - s_i)\}^{\frac{1}{2}}$. Thus $\{M(f - s_i, f - s_i)\}^{\frac{1}{2}} \leqq T_{1,0}(\bar{\pi}_i)^n \|L[f]\|_{L^2[a,b]}$. This last result, combined with (4.1) and (4.2), gives (4.3) and (4.4) with $T_{3,j} = T_{1,0} T_{1,j}$, $0 \leqq j \leqq n$, and $T_{4,j} = T_{1,0} T_{2,j}$, $0 \leqq j \leqq n-1$. Q.E.D.

It is possible under some circumstances to improve on the exponent of (4.4). The following, which generalizes Theorem 9 of [3], is one such improvement.

Theorem 10. Suppose $a_k \in C^k[a, b]$, $0 \leqq k \leqq n$. Let $f \in C^{2n}[a, b]$, let $\{\pi_i\}_{i=1}^{\infty}$ be a sequence of partitions of $[a, b]$ with $\lim_{i \to \infty} \bar{\pi}_i = 0$, and let $\{z^{(i)}\}_{i=1}^{\infty}$ be the sequence of incidence vectors associated with $\{\pi_i\}_{i=1}^{\infty}$ such that $z_l^{(i)} = n$, $0 \leqq l \leqq N_i + 1$. Then there exists a positive integer i_0 such that the $Sp(L, \pi_i, z^{(i)})$-interpolate $s_i(x)$ of f of Type I exists and is unique for any $i \geqq i_0$, and moreover there exists a constant T_5, independent of i, j or f such that

$$\text{(4.5)} \qquad \|D^j(f - s_i)\|_{L^\infty[a,b]} \leqq T_5(\bar{\pi}_i)^{2n-j} \|L[f]\|_{L^\infty[a,b]}, \qquad 0 \leqq j \leqq 2n,$$

and all $i \geqq i_0$, where

$$\|D^j(f - s_i)\|_{L^\infty[a,b]} \equiv \max_{0 \leqq k \leqq N_i} \{\|D^j(f - s_i)\|_{L^\infty[x_k^{(i)}, x_{k+1}^{(i)}]}\}.$$

Proof. Suppose that for some $i \geqq 1$, $u(x)$ is a $(\pi_i, z^{(i)})$-interpolate of the zero function and $u \in C^{2n}[x_l, x_{l+1}]$, $0 \leqq l \leqq N_i$. Then $D^j u(x_l^{(i)}) = 0$ for $0 \leqq j \leqq n-1$, $0 \leqq l \leqq N_i + 1$. By repeated applications of Rolle's theorem, there exist points $\{y_{j,l}\}$ in $[x_l^{(i)}, x_{l+1}^{(i)}]$, such that $D^j u(y_{j,l}) = 0$, $0 \leqq l \leqq N_i$, $0 \leqq j \leqq 2n-1$. Let w_j be such that $\|D^j u\|_{L^\infty[a,b]} = |D^j u(w_j)|$, $0 \leqq j \leqq 2n-1$. Then

$$\|D^j u\|_{L^\infty[a,b]} = \left| \int_{y_{j,1}}^{w_j} D^{j+1} u(x)\, dx \right| < \bar{\pi}_i \|D^{j+1} u\|_{L^\infty[a,b]},$$

where $w_j \in [x_l^{(i)}, x_{l+1}^{(i)}]$, $0 \leqq j \leqq 2n-1$. Therefore

$$\|D^j u\|_{L^\infty[a,b]} \leqq (\bar{\pi}_i)^{2n-j} \|D^{2n} u\|_{L^\infty[a,b]}, \qquad 0 \leqq j \leqq 2n-1. \tag{4.6}$$

Expanding $L[u]$, $L[u] = \sum_{k=0}^{2n} c_k D^k u$, where $c_k \in C[a, b]$, $0 \leqq k \leqq 2n$, and $c_{2n}(x) = (-1)^n a_n(x) \geqq \omega > 0$. Thus

$$(-1)^n a_n(x) D^{2n} u(x) = L[u(x)] - \sum_{k=0}^{2n-1} c_k(x) D^k u(x),$$

and

$$\omega |D^{2n} u(x)| \leqq \|L[u]\|_{L^\infty[a,b]} + \sum_{k=0}^{2n-1} \|c_k\|_{L^\infty[a,b]} \|D^k u(x)\|_{L^\infty[a,b]}. \tag{4.7}$$

Applying (4.6) to (4.7),

$$\omega \|D^{2n} u\|_{L^\infty} \leqq \|L[u]\|_{L^\infty} + \sum_{k=0}^{2n-1} \|c_k\|_{L^\infty} (\bar{\pi}_i)^{2n-k} \|D^{2n} u\|_{L^\infty}.$$

Define $F(\bar{\pi}_i) \equiv \omega - \sum_{k=0}^{2n-1} \|c_k\|_{L^\infty} (\bar{\pi}_i)^{2n-k}$. Then there is a $\varepsilon > 0$, and a $H_1 > 0$ such that if $\bar{\pi}_i < \varepsilon$, then $F(\bar{\pi}_i) > H_1$. Thus for $\bar{\pi}_i < \varepsilon$,

$$\|D^{2n} u\|_{L^\infty} \leqq \frac{1}{H_1} \|L[u]\|_{L^\infty}. \tag{4.9}$$

Choose i), as in Theorem 8, so that there will be a unique $Sp(L, \pi_i, z^{(i)})$-interpolate, s_i, of f of Type I for $i \geqq i_1$, and let i_2 be such that for $i \geqq i_2$, $\bar{\pi}_i < \varepsilon$. Let $i_0 = \max(i_1, i_2)$, and let $T_5 = H_1^{-1}$. Then for $i \geqq i_0$, $f - s_i$ is a $(\pi_i, z^{(i)})$-interpolate of zero and $f - s_i \in C^{2n}[x_l, x_{l+1}]$, $0 \leqq l \leqq N_i$, so (4.6) and (4.9) give

$$\|D^j (f - s_i)\|_{L^\infty} \leqq T_5 (\bar{\pi}_i)^{2n-j} \|L[f - s_i]\|_{L^\infty}, \qquad 0 \leqq j \leqq 2n.$$

But $\|L[f - s_i]\|_{L^\infty} = \|L[f]\|_{L^\infty}$ since $L[s] \equiv 0$ on $[x_l, x_{l+1}]$, $0 \leqq l \leqq N_i$. Q.E.D.

In [6], Schultz and Varga also obtain the convergence of derivatives of higher order of $(f - s_i)$, using a technique of Ahlberg, Nilson, and Walsh [2]. This was achieved in Theorem 10, by placing restrictions on the incidence vector $z^{(i)}$. For a general incidence vector $z^{(i)}$, using techniques similar to those of the two papers referred to above, the following theorem may be proven:

Theorem 11. If the hypotheses of Theorem 9 hold, and if there is a positive constant σ such that $\sigma \pi_i \geqq \bar{\pi}_i$ for all $i \geqq 1$, then there exists a positive integer i_0 and constants $T_{6,j}$, $0 \leqq j \leqq 2n-1$, dependent on f but independent of i, such that

$$\|D^j (f - s_i)\|_{L^\infty[a,b]} \leqq T_{6,j} (\bar{\pi}_i)^{n-j-(\frac{1}{2})}$$

for $0 \leqq j \leqq 2n-1$ and any $i \geqq i_0$.

References

1. Ahlberg, J. H., Nilson, E. N., Walsh, J. L.: Fundamental properties of generalized splines. Proc. Nat. Acad. Sci. U.S.A. **52**, 1412–1419 (1964).
2. – – – Convergence properties of generalized splines. Proc. Nat. Acad. Sci. U.S.A. **54**, 344–350 (1965).
3. Ciarlet, P. G., Schultz, M. H., Varga, R. S.: Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem. Numer. Math. **9**, 394–430 (1967).
4. Coddington, E. A., Levinson, N.: Theory of ordinary differential equations. New York: McGraw-Hill Book Co. 1955.
5. Hardy, G. H., Littlewood, J. E., Pólya, G.: Inequalities. Cambridge: Cambridge University Press 1952.
6. Schultz, M. H., Varga, R. S.: L-splines. Numer. Math. **10**, 345–369 (1967).

Dr. Thomas R. Lucas
Department of Mathematics
The University of North Carolina at Charlotte
Charlotte, North Carolina 28213
P. O. Box 12665

Numer. Math. 15, 371–381 (1970)
© by Springer-Verlag 1970

On the Numerical Integration of the Heat Equation

J. Descloux*

Received May 20, 1969

1. Introduction

We consider the following problem of heat conduction:

$$u_t(x,t) - u_{xx}(x,t) = f(x,t), \qquad 0 \leqq x \leqq 1, \qquad t \geqq 0, \tag{1}$$

$$u(0,t) = u(1,t) = 0, \qquad t \geqq 0, \tag{2}$$

$$u(x,0) = u_0(x), \qquad 0 \leqq x \leqq 1; \tag{3}$$

the indices t and x denote the partial derivatives with respect to t and x; f and u_0 are given.

The method of Galerkin consists in finding an approximate solution of the form

$$v(x,t) = \sum_{k=1}^{n} \alpha_k(t)\, s_k(x); \tag{4}$$

the given independent functions s_k are of class C^1 with a second piecewise continuous derivative and satisfy the boundary conditions $s_k(0) = s_k(1) = 0$. The unknown functions α_k are determined by the relations:

$$\int_0^1 s_i(x)\,\big(v_t(x,t) - v_{xx}(x,t) - f(x,t)\big)\,dx = 0, \qquad i = 1, 2, \ldots, n, \tag{6}$$

$$v(x,0) = v_0(x), \tag{7}$$

where $v_0 = \sum \alpha_{0k} s_k$ is some approximation of u_0. One replaces in (6) and (7) v by its expression (4) and gets, after an integration by parts a system of n ordinary differential equations

$$\sum_{k=1}^{n} \left(\int_0^1 s_i(x)\, s_k(x)\, dx \right) \alpha_k'(t) + \sum_{k=1}^{n} \left(\int_0^1 s_i'(x)\, s_k'(x)\, dx \right) \alpha_k(t) = \int_0^1 s_i(x)\, f(x,t)\, dx, \tag{8}$$

$$\alpha_i(0) = \alpha_{0i}, \qquad i = 1, 2, \ldots, n; \tag{9}$$

one remarks that (8) is still meaningfull if s_k is only continuous with a first piecewise continuous derivative (actually (see Section 5) piecewise linear functions give excellent results for practical computations).

We now give an abstract formulation of Eqs. (1), (2), (3) and their approximations (6), (7). Let $\mathscr{H}$ be a real Hilbert space with scalar product $(,)$ and norm $\|\ \|$,

* Département de mathématiques. Ecole polytechnique fédérale, Lausanne. Avenue de Cour 33.

L be a self-adjoint positive definite operator with domain of definition Δ; $u_0 \in \mathscr{H}$ and f, a function of $\mathbb{R}^+$ with values in $\mathscr{H}$ are given; unknown is the function u of $\mathbb{R}^+$ with values in $\mathscr{H}$ satisfying the relations

$$u'(t) + Lu(t) = f(t), \quad t > 0, \tag{10}$$

$$\lim_{t \to 0} u(t) = u_0; \tag{11}$$

u' denotes the derivative of u with respect to the norm $|| \; ||$; "lim" is also the limit with respect to the norm $|| \; ||$.

Let $[x, y] = (x, Ly)$, $[\![x]\!] = [x, x]^{\frac{1}{2}}$, $x, y \in \Delta$; $[,]$ is a positive definite bilinear form; H will denote the completion of Δ with respect to $[\![\;]\!]$; H is a Hilbert space with scalar product $[,]$ and, as a set, $H \subset \mathscr{H}$ (see [1], p 325).

Let S be a finite dimensional subspace of Δ; the approximation of Galerkin is defined by the equations

$$\begin{gathered} \big(s, v'(t) + Lv(t) - f(t)\big) = 0, \quad \forall s \in S, \\ v(0) = v_0 \in S, \end{gathered} \tag{12}$$

or equivalently

$$\begin{gathered} \big(s, v'(t)\big) + [s, v(t)] = \big(s, f(t)\big), \quad \forall s \in S \\ v(0) = v_0; \end{gathered} \tag{13}$$

(13) suggests that S can be chosen as a subspace of H.

For the particular problem (1), (2), (3) $\mathscr{H}$ is the set of square integrable functions on $0 \leqq x \leqq 1$, Δ is the set of functions satisfying the relations $g(0) = g(1) = 0$, having an absolutely continuous derivative and a second square integrable derivative, H is the set of absolutely continuous functions g with square integrable derivative and satisfying the conditions $g(0) = g(1) = 0$; writting $v(t) = \sum \alpha_k(t)\, s_k$ where $s_1, s_2, \ldots, s_n$ is a basis of S, (13) is transformed in (8), where s has to be replaced successively by $s_1, s_2, \ldots, s_n$.

We introduce the operator Z mapping S on S defined by the relation

$$(s, Zx) = [s, x], \quad \forall s \in S;$$

Z is self-adjoint positive definite; (13) is equivalent to $v'(t) + Zv(t) = Pf(t)$, where P is the orthogonal projector of $\mathscr{H}$ on S with respect to $(,)$. In what follows we shall consider a sequence $S^1, S^2, \ldots$ of subspaces of H of dimension $1, 2, \ldots$ and the corresponding problems:

$$v^{n\prime}(t) + Z^n v^n(t) = P^n f(t), \quad t \geqq 0, \tag{14}$$

$$v^n(0) = v_0^n. \tag{15}$$

Problems of type (10) have been studied extensively (see [5]); in particular Theorem 1 is proved in [2].

Theorem 1. If a) L is a self-adjoint positive definite operator with compact inverse, b) f is continuous with a piecewise continuous derivative for $t \geqq 0$, then problem (10), (11) possesses one and only one solution. If furthermore c) for

any $x \in H$ there exists $s^n \in S^n$ with $\lim_{n\to\infty} \|x - s^n\| = 0$, d) $\lim_{n\to\infty} v_0^n = u_0$ then for any $0 < \eta < T$ one has

$$\lim_{n\to\infty} \|u(t) - v^n(t)\| = 0 \quad \text{uniformly for} \quad \eta \leqq t \leqq T \tag{15a}$$

where u and v^n are defined by (10), (11), (14), (15).

For the particular problem (1), (2), (3), (15a) implies

$$\lim_{n\to\infty} |u(x,t) - v^n(x,t)| = 0 \quad \text{uniformly for} \quad 0 \leqq x \leqq 1, \quad \eta \leqq t \leqq T$$

and

$$\lim_{n\to\infty} \int_0^1 \left(u'(x,t) - v^{n\prime}(x,t)\right)^2 dx = 0 \quad \text{uniformly for} \quad \eta \leqq t \leqq T.$$

2. Numerical Integration of the Differential System

From now on, we assume that the operator L of (10) has a compact inverse. Let $0 < \lambda_1 \leqq \lambda_2 \leqq \lambda_3 \ldots$ be the eigenvalues of L, $\mu_1^n \leqq \mu_2^n \ldots \leqq \mu_n^n$ be the eigenvalues of Z^n. The Courant principle states that $\mu_j^n \geqq \lambda_j$, $j = 1, 2, \ldots, n$. It follows that $\lim_{n\to\infty} \mu_n^n = \infty$; for large n (14) is a "stiff" differential system.

Appropriate methods for integrating stiff differential equations have been proposed recently. This paper is concerned with a class of multistep methods suggested in [6].

For integrating the differential equation $y' = f(t, y)$ we consider the p-step method defined by the relation:

$$y_{j+p} + a_{p-1} y_{j+p-1} + \cdots + a_0 y_j = h\left(b_p f(t_{j+p}, y_{j+p}) + \cdots + b_0 f(t_j, y_j)\right) \tag{16}$$

where $t_j = t_0 + jh$ and y_j is an approximation of $y(t_j)$. Let

$$\varrho(\xi) = \xi^p + a_{p-1}\xi^{p-1} + \cdots + a_0, \qquad \sigma(\xi) = b_p \xi^p + \cdots + b_0.$$

Definition. Method (16) is said "positive ultrastable" (PUS) if a) $\varrho(1) = 0$, $\varrho'(1) = \sigma(1)$; b) the roots of ϱ of modulus 1 are simple; c) $m(\varkappa) < 1$ for $\varkappa > 0$ where $m(\varkappa)$ is the maximum of the modulus of the roots of the polynomial $\varrho(\xi) + \varkappa\sigma(\xi)$; d) $\lim_{\varkappa\to\infty} m(\varkappa) = 0$.

Conditions a, b, c imply the convergence of the method in the sense of Dahlquist (see [3]). From elementary properties of polynomials it follows that $b_0 = b_1 = \cdots = b_{p-1} = 0$ and that $m(\varkappa) = O(\varkappa^{-1/p})$ as $\varkappa \to \infty$. Four positive ultrastable methods called in the following PUS 1, PUS 2, PUS 3, PUS 4 have been experimented; their associated polynomials are given in the table. For $p = 1, 2, 3, 4$,

Table

Method	$\varrho(\xi)$	$\sigma(\xi)$
PUS 1	$\xi - 1$	ξ
PUS 2	$\xi^2 - \frac{4}{3}\xi + \frac{1}{3}$	$\frac{2}{3}\xi^2$
PUS 3	$\xi^3 - \frac{18}{11}\xi^2 + \frac{9}{11}\xi - \frac{2}{11}$	$\frac{6}{11}\xi^3$
PUS 4	$\xi^4 - \frac{48}{25}\xi^3 + \frac{36}{25}\xi^2 - \frac{16}{25}\xi + \frac{3}{25}$	$\frac{12}{25}\xi^4$

PUS p is a p-step method of order p; it is however not possible to construct for any p a p-step PUS method of order p (see [6]).

With the help of a PUS method, the solution v^n of (14), (15) is approximated at $t_j = jh$ by $w^n_{h,j}$ defined by the relations

$$w^n_{h,j+p} + a_{p-1} w^n_{h,j+p-1} + \cdots + a_0 w^n_{h,j} = h\, b_p \left(P^n f(t_{j+p}) - Z^n w^n_{n,j+p}\right), \tag{17}$$

$$w^n_{h,0} = v^n_0; \tag{18}$$

in order to complete this definition, we shall assume in the following, without much restriction of generality, that $w^n_{h,1}, \ldots, w^n_{h,p-1}$ are obtained from $w^n_{h,0}$ by the method PUS 1. For practical computation (17) takes the form:

$$w^n_{h,j} = \sum_{k=1}^{n} \beta_{j,k}\, s^n_k,$$

$$\sum_{k=1}^{n} (s^n_i, s^n_k)\,(\beta_{j+p,k} + a_{p-1}\beta_{j+p-1,k} + \cdots + a_0 \beta_{j,k})$$

$$= h\, b_p \left((s^n_i, f(t_{j+p})) - \sum_{k=1}^{n} [s^n_i, s^n_k]\, \beta_{j+p,k} \right), \qquad i = 1, 2, \ldots, n,$$

where $s^n_1, \ldots, s^n_n$ is a basis of S^n.

Theorem 2. If a) L is a self-adjoint positive definite operator with compact inverse, b) for any $x \in H$ there exists $s^n \in S^n$ with $\lim_{n\to\infty} \| x - s^n \| = 0$, c) $\lim_{n\to\infty} v^n_0 = u_0$, d) (17) is a PUS method, e) f is continuous for $t \geqq 0$, then for any $\varepsilon > 0$, $0 < \eta < T$ there exists $h_0 > 0$ such that for all n

$$\| w^n_{h,j} - v^n(t_j) \| < \varepsilon \quad \text{for} \quad h < h_0, \quad \eta \leqq t_j \leqq T,$$

where $w^n_{h,j}$ and v^n are defined by (17), (18), (14), (15).

Remarks 1. Theorems 1 and 2 prove the uniform $\| \ \|$ convergence on $\eta \leqq t \leqq T$ of $w^n_{h,j}$ towards u when $h \to 0$ and $n \to \infty$ independently. This kind of "stability" is seldom encountered when dealing with pure difference schemes.

2. Theorem 2 is not true for "less stable" methods; for example, for the trapezoïdal rule, Theorem 2 is only valid when one replaces $\| \ \|$ by $| \ |$; it is entirely false for Runge-Kutta methods.

3. Results similar to Theorems 1, 2 can be obtained for problems more general than (10), (11) (concerning the method of Galerkin, see [2]). Examples will be given in Section 5.

3. Intuitive Analysis

In this section we try to give for a very simple case an intuitive idea of the mechanism of a PUS method. Let A be a symmetric matrix of order n, x a function with values in $\mathbb{R}^n$ satisfying the differential system $x'(t) + A\,x(t) = 0$, $x(0) = x_0$. Let $0 \leqq \theta_1 \leqq \theta_2 \ldots \theta_n$ be the eigenvalues of A, $\varphi_1, \varphi_2, \ldots, \varphi_n$ be a corr$^{\prime}$sponding set of orthonormal eigenvectors. Writing

$$x(t) = \sum c_i(t)\varphi_i, \qquad x_0 = \sum c_{0i}\varphi_i,$$

the system is reduced to the n scalar differential equations

$$c_i'(t) + \theta_i c_i(t) = 0, \quad c_i(0) = c_{0i}, \quad i = 1, 2, \ldots, n. \tag{19}$$

For convenient starting values, integration of the differential system with a multistep method is equivalent to the integration of the n scalar equations (19) with the same method. Let h be chosen such that $h\theta_1$ is "small" and $h\theta_n$ is "large"; c_{ij} denotes the approximative value of $c_i(jh)$ obtained by a PUS p-multistep method of order p. For $i=1$, one can expect that a large value of p will give a very accurate solution of (19). However for $i=n$, one has (see Section 2):

$$c_{nj} \doteq r(h\theta_n)^{-j/p}$$

where r is some constant; although r depends on p, one can expect more realistic results with a small p.

The optimal value of p will essentially depend on the importance of the "high harmonics", i.e. of the value of c_{i0} for large values of i.

4. Elements of Proof of Theorem 2

In this section we prove Theorem 2 only for the method PUS 1 and under the hypothesis that all eigenvalues of L are simple. The general proof follows the same pattern but is much longer.

For the PUS 1 method (17) becomes simply

$$w^n_{h,j+1} - w^n_{h,j} = h\left(P^n f(t_{j+1}) - Z^n w^n_{h,j+1}\right);$$

with the initial value $w^n_{h,0} = v^n_0$, this difference equation admits the solution

$$w^n_{h,j} = (I + hZ^n)^{-j} v^n_0 + h \sum_{k=1}^{j} (I + hZ^n)^{k-j-1} P^n f(t_k). \tag{20}$$

We suppose that all the hypotheses of Theorem 2 are satisfied. Let $0<\lambda_1<\lambda_2, \ldots$ be the eigenvalues of L, $\psi_1, \psi_2, \ldots$ be a system of corresponding orthonormal eigenelements, $\mu^n_1 \leqq \mu^n_2, \ldots \leqq \mu^n_n$ be the eigenvalues of Z^n. Lemma 1 is proved in [7] or in [1], p. 369.

Lemma 1. a) $\lim_{n\to\infty} \mu^n_i = \lambda_i$, $i = 1, 2, \ldots$; b) for each value of n, there exists a system $g^n_1, \ldots, g^n_n$ of orthonormal eigenelements of Z^n corresponding to $\mu^n_1, \ldots, \mu^n_n$ such that $\lim_{n\to\infty} \|g^n_i - \psi_i\| = 0$, $i = 1, 2, \ldots$.

In the following g^n_i is supposed to have the property stated in Lemma 1. Let Ω^n_m be the orthogonal projector of $\mathscr{H}$ on the subspace spanned by $g^n_1, g^n_2, \ldots, g^n_m$; for $m \geqq n$, $\Omega^n_m = P^n$, the orthogonal projector on S^n. Clearly the restriction of Ω^n_m to S^n commutes with Z^n. Furthermore for $x, y \in S^n$ one has

$$[x, y] = (x, Z^n y) = \sum_{i=1}^{n} \mu^n_i (x, g^n_i)(y, g^n_i). \tag{21}$$

Lemma 2. Let f be a continuous function of $\mathbb{R}^+$ with values in $\mathscr{H}$. Then for given $T>0$, $\varepsilon>0$, there exists M such that for $m>M$ and all n

$$\|\Omega^n_m P^n f(\tau) - P^n f(\tau)\| < \varepsilon, \quad 0 \leqq \tau \leqq T.$$

Proof. Because of the uniform continuity of f on the interval $0 \leqq \tau \leqq T$ we can find $0 = \tau_1 < \tau_2 \ldots < \tau_q = T$ such that $\|f(\tau) - f(\tau_i)\| < \varepsilon/6$ for $\tau_i \leqq \tau \leqq \tau_{i+1}$. Hypothesis b of Theorem 2 implies the existence of N such that $\|P^n f(\tau_i) - f(\tau_i)\| < \varepsilon/6$ for $n > N$, $i = 1, 2, \ldots, q$. Since the set of ψ_i is total in $\mathscr{H}$, there exists M_1 such that $\|f(\tau_i) - \sum_{i=1}^{M_1} (f(\tau_i), \psi_i)\, \psi_i\| < \varepsilon/6$, $i = 1, 2, \ldots, q$. By Lemma 1 we can find $M \geqq \max(N, M_1)$ such that for $n > M$ and $i = 1, 2, \ldots, q$

$$\|f(\tau_i) - \sum_{i=1}^{M_1} f(\tau_i, g_i^n)\, g_i^n\| = \|f(\tau_i) - \Omega_{M_1}^n f(\tau_i)\| < \frac{\varepsilon}{3}.$$

Thus we have for $\tau_i \leqq \tau \leqq \tau_{i+1}$, $n > M$:

$$\begin{aligned}\|\Omega_{M_1}^n P^n f(\tau) - P^n f(\tau)\| \leqq{}& \|\Omega_{M_1}^n P^n (f(\tau) - f(\tau_i))\| + \|\Omega_{M_1}^n (P^n f(\tau_i) \\ &- f(\tau_i))\| + \|\Omega_{M_1}^n f(\tau_i) - f(\tau_i)\| + \|f(\tau_i) - P^n f(\tau_i)\| \\ &+ \|P^n (f(\tau_i) - f(\tau))\| < \varepsilon.\end{aligned}$$

Since $M \geqq M_1$, we have for $n > M$, $0 \leqq \tau \leqq T$: $\|\Omega_M^n P^n f(\tau) - P^n f(\tau)\| < \varepsilon$; for $n \leqq M$, $\Omega_M^n = P^n$ and the last inequality is trivial. q.e.d.

The result we must prove in this section follows from the three statements: for given $\varepsilon > 0$, $0 < \eta < T$, there exists M and $h_0 > 0$ such that for all n

$$[\!]\Omega_M^n v^n(t) - v^n(t)[\!] < \frac{\varepsilon}{3}, \quad \eta \leqq t \leqq T, \tag{22}$$

$$[\!]\Omega_M^n w_{h,j}^n - w_{h,j}^n[\!] < \frac{\varepsilon}{3}, \quad \eta \leqq t_j \leqq T, \tag{23}$$

$$[\!]\Omega_M^n v^n(t_j) - \Omega_M^n w_{h,j}^n[\!] < \frac{\varepsilon}{3}, \quad \eta \leqq t_j \leqq T, \quad h < h_0. \tag{24}$$

One first determines M such that (22), (23) are valid and then h_0. Integrating (14) one gets

$$v^n(t) = e^{-Z^n t} v_0^n + \int_0^t e^{-Z^n (t-\tau)} P^n f(\tau)\, d\tau;$$

since Ω_m^n commutes with Z^n, we can write

$$[\!]\Omega_m^n v^n(t) - v^n(t)[\!] \leqq [\!]e^{-Z^n t} a[\!] + [\!] \int_0^t e^{-Z^n (t-\tau)} b(\tau)\, d\tau [\!]$$

where $a = \Omega_m^n v_0^n - v_0^n$, $b(\tau) = \Omega_m^n P^n f(\tau) - P^n f(\tau)$; using (21), the hypotheses of Theorem 2 and Lemma 2, (22) is a consequence of the inequalities

$$[\!]e^{-Z^n t} a[\!]^2 = \sum_{i=1}^n \mu_i^n e^{-2\mu_i^n t} (a, g_i^n)^2 \leqq \max_{\mu \geqq 0} \mu e^{-2\mu\eta} \|a\|^2, \quad t \geqq \eta;$$

$$\|a\| \leqq \|\Omega_m^n (v_0^n - P^n u_0)\| + \|\Omega_m^n P^n u_0 - P^n u_0\| + \|P^n u_0 - v_0^n\|;$$

$$[\!] \int_0^t e^{-Z^n (t-\tau)} b(\tau)\, d\tau [\!]^2 = \sum_{i=1}^n \mu_i^n \left(\int_0^t e^{-\mu_i^n (t-\tau)} (b(\tau), g_i^n)\, d\tau \right)^2$$

$$\leqq \sum_{i=1}^n \mu_i^n \int_0^t e^{-2\mu_i^n (t-\tau)} d\tau \cdot \int_0^t (b(\tau), g_i^n)^2\, d\tau \leqq \frac{T}{2} \max_{0 \leqq \tau \leqq T} \|b(\tau)\|^2,$$

$0 \leqq t \leqq T$; here we used Schwarz' inequality and the relation

$$\mu \int_0^t e^{-2\mu(t-\tau)}\, d\tau < \tfrac{1}{2}.$$

We proceed similarly for (23), using the commutativity of Ω_M^n and Z^n, we get from (20):

$$\| \Omega_M^n w_{h,j}^n - w_{h,j}^n \| \leqq \| (I + hZ^n)^{-j} a \| + \| h \sum_{k=1}^{j} (I + hZ^n)^{k-j-1} b(t_k) \|,$$

where a and b are defined above. We note the inequality

$$\max_x \frac{x}{(1+x)_p} = \frac{1}{p}\left(\frac{p-1}{p}\right)^{p-1} < \frac{1}{p}, \qquad p > 0. \tag{25}$$

Using (21) and (25), (23) follows from the relations:

$$\| (I + hZ^n)^{-j} a \|^2 = \sum_{i=1}^{n} (1 + h\mu_i^n)^{-2j} \mu_i^n (a, g_i^n)^2 \leqq \frac{1}{2\eta} \|a\|^2, \qquad t_j \geqq \eta;$$

$$\begin{aligned}
\| h \sum_{k=1}^{j} (I + hZ^n)^{k-j-1} b(t_k) \|^2 &= h^2 \sum_{i=1}^{j} \mu_i^n \left(\sum_{k-1}^{j} (1 + h\mu_i^n)^{k-j-1} (b(t_k), g_i^n) \right)^2 \\
&\leqq h^2 \sum_{i=1}^{n} \mu_i^n \left(\sum_{k=1}^{j} (1 + h\mu_i^n)^{2(k-j-1)} \right) \left(\sum_{k=1}^{j} (b(t_k), g_i^n)^2 \right) \\
&\leqq \sup_{\mu > 0} \left(h\mu \sum_{k=1}^{j} (1 + h\mu)^{2(k-j-1)} \right) \left(\sum_{k=1}^{j} h \sum_{i=1}^{n} (b(t_k), g)^2 \right) \\
&\leqq 0.5\, T \max_{0 \leqq \tau \leqq T} \|b(\tau)\|^2 \quad \text{for} \quad t_j \leqq T;
\end{aligned}$$

indeed

$$\sup_{\mu > 0} \left(h\mu \sum_{k=1}^{j} (1 + h\mu)^{2(k-j-1)} \right) \leqq \sup_{x > 0} \sum_{i=1}^{\infty} \frac{x}{(1+x)^{2i}} = \sup_{x > 0} \frac{1}{2+x} = \frac{1}{2}.$$

For proving (24) we remark that, M being now fixed, there exists a constant c such that $\mu_i^n \leqq c$, $i = 1, 2, \ldots, M$ for all n; by (21), it suffices to show that there exists h_0 with

$$|(v^n(t_j), g_i^n) - (w_{h,j}^n, g_i^n)| \leqq \frac{\varepsilon}{3\sqrt{Mc}}, \qquad t_j \leqq T, \quad h < h_0, \quad i = 1, 2, \ldots, M;$$

but $(w_{h,j}^n g_i^n)$ is the PUS 1 approximation of the scalar differential equation $q'(t) + \mu_i^n q(t) = (P^n f(t), g_i^n)$ which is satisfied by $(v^n(t), g_i^n)$; at that point we use wellknown inequalities for multistep methods (see for example [3] p. 246).

5. Numerical Examples

In this section $u(x, t)$ denotes the "exact" solution and $v(x, t)$ its approximation by the method of Galerkin.

We consider the following problem:

$$u_t(x,t) - \big(p(x)\,u_x(x,t)\big)_x = f(x,t), \qquad 0 \leqq x \leqq 1, \qquad t > 0, \tag{26}$$

$$u(x,0) = u_0(x), \qquad 0 \leqq x \leqq 1, \tag{27}$$

$$u_x(0,t) = 0, \qquad A\,u_x(1,t) + B\,u(1,t) = r(t), \qquad t > 0 \tag{28}$$

p, f, u_0, A, B, r are given. We first consider the case when $p(1) = 1$, $A = 1$; for t fixed u satisfies the equation of variation:

$$\delta\left\{\tfrac{1}{2}\int_0^1 p(x)\,g^2(x)\,dx + \tfrac{1}{2}B\,g^2(1) - \int_0^1 \big(f(x,t) - u_t(x,t)\big)\,g(x)\,dx - r(t)\,g(1)\right\} = 0 \tag{29}$$

where the variation is taken on the set G of continuous functions g having a piecewise continuous derivative (no boundary condition); restricting G to a subspace S with basis $s_1, s_2, \ldots, s_n$ (and replacing in (29) u_t by v_t) we obtain the relations:

$$v(x,t) = \sum_{i=1}^{n} \alpha_i(t)\,s_i(x),$$

$$\sum_{k=1}^{n}\left(\int_0^1 s_i(x)\,s_k(x)\,dx\right)\alpha_k'(t) + \sum_{k=1}^{n}\left(\int_0^1 p(x)\,s_i'(x)\,s_k'(x)\,dx + B s_i(1)\,s_k(1)\right)\alpha_k(t)$$

$$= \int_0^1 s_i(x)\,f(x,t)\,dx + s_i(1)\,r(t), \qquad i = 1, 2, \ldots, n.$$

If $A = 0$, $B = 1$, for t fixed, u is solution of the equation of variation

$$\delta\left\{\tfrac{1}{2}\int_0^1 p(x)\,g^2(x)\,dx - \int_0^1 \big(f(x,t) - u_t(x,t)\big)\,g(x)\,dx\right\} = 0 \tag{30}$$

where the variation is taken on the set G_t of continuous functions g having a piecewise continuous derivative and satisfying the boundary condition $g(1) = r(t)$. Let $q_0, q_1, \ldots, q_n \in G$ with boundary conditions $q_0(1) = 1$, $q_k(1) = 0$, $k = 1, 2, \ldots, n$, S_t be the set of functions of the form $r(t)\,q_0 + \sum_{i=1}^{n} \beta_i\,q_i$. Restricting G_t to S_t (and replacing u_t by v_t), (30) becomes

$$v(x,t) = r(t)\,q_0(x) + \sum_{k=1}^{n} \alpha_k(t)\,q_k(x),$$

$$\sum_{k=1}^{n}\left(\int_0^1 q_i(x)\,q_k(x)\right)\alpha_k'(t) + \sum_{k=1}^{n}\left(\int_0^1 p(x)\,q_i'(x)\,q_k'(x)\right)\alpha_k(t)$$

$$= \int_0^1 q_i(x)\,f(x,t)\,dx - r(t)\int_0^1 q_i(x)\,q_0(x) - r'(t)\int_0^1 p(x)\,q_i'(x)\,q_0'(x)\,dx, \qquad i = 1, 2, \ldots, n.$$

None of the three computed examples is exactly of type (10), (11). However for them results analogous to Theorems 1, 2 are valid (see [2] and [10] for generalisations of Theorem 1). For Examples 1 and 3 the boundary conditions at $x = 1$ are not homogeneous; in Example 2, the operator L is positive, but not positive definite.

In order to specify the "basic" functions effectively used, we introduce the functions U, V, W:

$$U(x) = V(x) = W(x) = 0, \quad |x| \geqq 1,$$

$$U(x) = 1 - |x|, \quad V(x) = 2|x^3| - 3x^2 + 1, \quad W(x) = x^3 - 2x|x| + x, \quad |x| \leqq 1;$$

U is continuous piecewise linear with $U(0) = 1$; V and W are piecewise cubic, of class C^1 with $V(0) = 1$, $V'(0) = 0$, $W(0) = 0$, $W'(1) = 1$.

Let $h = 1/p$. On the interval $(0, 1)$ the function $g(x) = \sum_{i=0}^{p} c_i U\big(p(x - ih)\big)$ is characterized by the properties:

a) g is continuous;

b) g is linear on each interval $\big((i-1)h, ih\big)$, $i = 1, \ldots, p$;

c) $g(ih) = c_i$, $\quad i = 0, 1, \ldots, p$.

On the interval $(0,1)$ the function $g(x) = \sum_{i=0}^{p} \{c_i V\big(p(x - ih)\big) + d_i h W\big(p(x - ih)\big)\}$ is characterized by the properties:

a) g is of class C^1;

b) g is a cubic polynomial on each interval $\big((i-1)h, ih\big)$, $i = 1, \ldots, p$;

c) $g(ih) = c_i$, $\quad g'(ih) = d_i$, $\quad i = 0, 1, \ldots, p$.

Relations between such functions g and the condition of completeness c of Theorem 1 are studied in [4].

All computations have been realized on a IBM 7040 in double precision (16 decimals) in order to garantee that the numerical results are free of round-off errors.

Example 1. $u_t(x, t) - u_{xx}(x, t) = 4x^4 t^3 - 12x^2 t^4$; $\quad u_0(x) = 0$; $\quad u_x(0) = 0$;

$$u_x(1, t) + 2u(1, t) = 6t^4; \quad 0 \leqq x \leqq 1, \quad t \geqq 0.$$

The exact solution is $u(x, t) = x^4 t^4$. One chooses: $n = 42$,

$$s_{2i+1}(x) = V\big((x - 0.05i)/0.05\big), \quad s_{2i+2}(x) = W\big((x - 0.05i)/0.05\big),$$

$i = 0, 1, 2, \ldots, 20$. Acurate computations show:

$$\max_{0 \leqq x \leqq 1} |u(x, 1) - v(x, 1)| = 2 \cdot 10^{-7}.$$

For $i = 1, 2, 3, 4$ $w_i(x)$ denotes the discrete approximation of $v(x, 1)$ obtained in the following conditions: for $0 \leqq t \leqq 0.08$, PUS 1, $h = 0.01$; for $0.08 \leqq t \leqq 0.16$, PUS i, $h = 0.02$; for $0.16 \leqq t \leqq 1$, PUS i, $h = 0.04$; let $a_i = \max_{0 \leqq x \leqq 1} |v(x, 1) - w_i(x)|$.

The following results have been computed:

$$a_1 = 12 \cdot 10^{-3}, \quad a_2 = 7 \cdot 10^{-4}, \quad a_3 = 3 \cdot 10^{-5}, \quad a_4 = 7 \cdot 10^{-7}.$$

Comment. In this example, coefficients of the equation and of the solution are very "regular". a) Confirming the analysis of Section 3, the best results are obtained with PUS 4. b) The use of piecewise linear functions (type U) gives for v an approximation which is not so good.

Example 2. $u_t(x,t) - u_{xx}(x,t) = 0$; $u_0(x) = x$ for $0 \leqq x \leqq 0.25$, $u_0(x) = x - 1$ for $0.25 < x \leqq 1$; $u_x(0,t) = u_x(1,t) = 0$; $0 \leqq t \leqq 1$, $t > 0$.

One chooses $n = 81$; $s_{i+1}(x) = U\big((x - 0.0125\,i)\big)$, $i = 0, 1, \ldots, 80$. Accurate computations show:

$$\max_{0 \leqq x \leqq 1} u(x, 0.02) - \min_{0 \leqq x \leqq 1} u(x, 0.02) = 0.34;$$

$$\max_{0 \leqq x \leqq 1} |u(x, 0.02) - v(x, 0.02)| = 3 \cdot 10^{-4}.$$

For $i = 1, 2, 3, 4$ $w_i(x)$ denotes the discrete approximation of $v(x, 0.02)$ obtained in the following conditions: for $0 \leqq t \leqq 0.0016$, PUS 1, $h = 0.0002$; for $0.0016 \leqq t \leqq 0.0032$, PUS i, $h = 0.0004$; for $0.0032 \leqq t \leqq 0.02$, PUS i, $h = 0.0008$; let $a_i = \max_{0 \leqq x \leqq 1} |v(x, 0.02) - w_i(x)|$. The following results have been computed:

$$a_1 = 25 \cdot 10^{-4}, \quad a_2 = 0.6 \cdot 10^{-4}, \quad a_3 = 0.6 \cdot 10^{-4}, \quad a_4 = 0.5 \cdot 10^{-4}.$$

Comment. The initial function $u_0(x)$ is discontinuous.

a) In the sense of Section 3 $v(x,t)$ will contained many "high harmonics"; PUS 2, PUS 3, PUS 4 produce the same amount of error which is, however, much smaller than the error of PUS 1.

b) Because of the "irregularity" of u_0, piecewise linear functions gave, as far $v(x,t)$ is concerned, better results than piecewise cubic functions.

Example 3. $u_t(x,t) - \big(p(x)\,u_x(x,t)\big)_x = f(x,t)$; $p(x) = 1 - 3x$ for $0 \leqq x \leqq 0.25$, $p(x) = x$ for $0.25 \leqq x \leqq 1$; $u_0(x) = 1$; $u_x(0,t) = 0$, $u(1,t) = \cos 5t$; $f(x,t)$ is chosen such that the exact solution is $u(x,t) = \cos 5xt$.

One chooses $n = 200$; $q_0(x) = U\big((x-1)/0.005\big)$, $q_{i+1}(x) = U\big((x - 0.005\,i)/0.005\big)$, $i = 0, 1, 2, \ldots, 199$. Accurate computations show: $\max_{0 \leqq x \leqq 1} |u(x, 0.5) - v(x, 0.5)| = 2 \cdot 10^{-4}$. For $i = 1, 2, 3, 4$ $w_i(x)$ denotes the discrete approximation of $v(x, 0.5)$ obtained in the following conditions: for $0 \leqq t \leqq 0.02$, PUS 1, $h = 0.0025$; for $0.02 \leqq t \leqq 0.04$, PUS i, $h = 0.005$; for $0.04 \leqq t \leqq 0.5$, PUS i, $h = 0.01$; let $a_i = \max_{0 \leqq x \leqq 1} |v(x, 1) - w_i(x)|$. The following results have been computed:

$$a_1 = 4.5 \cdot 10^{-4}, \quad a_2 = 1.1 \cdot 10^{-4}, \quad a_3 = 0.5 \cdot 10^{-4}, \quad a_4 = 0.5 \cdot 10^{-4}.$$

Comment. a) Because of the lack of regularity of p, solutions of the corresponding homogeneous problem will not be very smooth. Even for the very regular solution of the present problem one can expect a great importance of the "high harmonics" (Section 3).

b) Use of piecewise cubic functions gives for this example poor results.

6. Final Remarks

1. Except for very particuliar cases, effective bounds or asymptotic evaluation of the error are difficult to obtain (see [2]). Results of Weinberger [8] can be of some help.

2. The methods and results of this pages can be applied to heat problems with several spatial dimensions (see [9]). For the choice of explicit subspaces S^n see [4].

References

1. Riesz, Nagy: Leçons d'analyse fonctionnelle. Académie des sciences de Hongrie 1952.
2. Descloux: Solution of the heat problem by the method of moments. Report. Department of Mathematics. EPF-Lausanne, 1969.
3. Henrici: Discrete variable methods in ordinary differential equation. John Wiley 1962.
4. Goël: Construction of basic functions for numerical utilisation of Ritz method. Numerische Mathematik **12**, 435–447 (1968).
5. Lions: Equations différentielles opérationnelles. Berlin-Göttingen-Heidelberg: Springer 1961.
6. Gear: The automatic integration of stiff ordinary differential equations. IFIP congress 68, Edinburgh.
7. Descloux: The solution of the heat problem by the method of moments, Part I. Technical report 12. Computer Center, University of California 12, Berkeley, 1967.
8. Weinberger, A. F.: Variational methods for eigenvalue problems. University of Minnesota, Department of Mathematics.
9. Wilson, Nickel: Application of the finite element method to heat conduction analysis. Nuclear Engineering and Design **4**, 276–286 (1966).
10. Descloux: On the heat equation. Math. Z. **113**, 376–382 (1970).

J. Descloux
Ecole Polytechnique Fédéral
de Lausanne
33 Avenue de Cour
CH 1007 Lausanne, Suisse

Numer. Math. 15, 382–391 (1970)
© by Springer-Verlag 1970

On the Remez Algorithm for Non-linear Families

RICHARD B. BARRAR* ** and HENRY L. LOEB* ***

Received April 5, 1969

I. Introduction

In this paper we give a convergence proof for the classic Remez Algorithm when it is applied to certain non-linear approximating families. We shall assume that the best approximation satisfies certain regularity conditions and that a sufficiently good estimate of the best approximation is available to begin the algorithm.

The setting for the problem will be $C[0, 1]$, where for $f \in C[0, 1]$,

$$\|f\| = \max_{x \in [0,1]} |f(x)| .$$

Several well known non-linear families can be treated under the theory. For example, ordinary rational functions form such a family; that is, if for any polynomial p we designate the degree of p by "∂p", then our approximating family is

$$R_m^n = \{r = p/q\colon\ \partial p \leqq n,\ \partial q \leqq m,\ q > 0 \text{ over } [0, 1],\ p \text{ and } q \text{ are relatively prime}\}.$$

We call $r^* = p^*/q^* \in R_m^n$ a best approximation to $f \in C[0, 1]$ iff

$$\|f - r^*\| = \operatorname{dist}(f, R_m^n) .$$

If r^* satisfies the additional condition,

$$\min\{n - \partial p, m - \partial q\} = 0 ,$$

and a $r^{(0)} \in R_m^n$ sufficiently close to r^* is available to begin the algorithm, then the sequence of iterates will converge to r^*. Another example is the so-called exponential family, where for a fixed positive integer n,

$$G = \left\{ \sum_{i=1}^{n} a_i e^{a_{n+i} x} : a_{n+i} \neq a_{n+j},\ i \neq j,\ i, j = 1, \ldots, n \right\} .$$

For $f \in C[0, 1]$ we seek the $g^* \in G$ such that

$$\|f - g^*\| = \operatorname{dist}(f, G) .$$

If g^* has the additional property that $a_i^* \neq 0$ $(i = 1, \ldots, n)$, then again if a sufficiently close estimate of g^* is available the algorithm converges. A lucid con-

* Mathematics Department, University of Oregon, Eugene, Oregon.
** Supported in part by N.S.F. Grant G.P.–9384.
*** Supported in part by N.S.F. Grant G.P.–8686.

vergence proof for the case when the Remez Algorithm is applied to a linear family can be found in [1]. For ordinary rational approximation a local convergence proof can be found in [2].

II. The Non-linear Approximating Family

We consider an approximating family of the form,

$$F = \{F(A, x): A \in P\},$$

where P is an open subset of real m space. We assume that at each $A = (a_1, \ldots, a_m) \in P$;

1. Each component of the gradient,

$$\left(\frac{\partial F(A, x)}{\partial a_1}, \ldots, \frac{\partial F(A, x)}{\partial a_m}\right),$$

is jointly continuous in A and x, where x ranges over $[0, 1]$.

2. The m components of the gradient generate a Haar Subspace[1] of $C[0, 1]$ of dimension $d(A)$ where $d(A) \geqq 1$.

3. For each $A' \in P$, $F(A) \not\equiv F(A') \Rightarrow F(A, x) - F(A', x)$ has at most $d(A) - 1$ distinct zeros in $[0, 1]$.

Then, $F(A^*, x)$ is called a best approximation to $f \in C[0, 1]$ from F iff

$$\|f - F(A^*)\| = \operatorname{dist}(f, F).$$

In order to simplify notation we assume without loss of generality that $f \equiv 0$. Hence we seek an $F(A^*)$ with the property

$$\|F(A^*)\| = \min_{A \in P} \|F(A)\|.$$

In [3] it was shown that:

Characterization Theorem. $F(A^*)$ is the best approximation iff there exists $s = d(A^*) + 1$ points, $0 \leqq x_1 < x_2 < \cdots < x_s \leqq 1$, such that

(a) $F(A^*, x_i) = -F(A^*, x_{i+1}), \quad i = 1, \ldots, s - 1;$

(b) $\|F(A^*)\| = |F(A^*, x_1)|.$

Further, a proof can be found in [5] of

De La Vallée Poussin Lemma. If for $A \in P$ there are $s = d(A) + 1$ points $0 \leqq x_1 < x_2 < \cdots < x_s \leqq 1$, such that $\operatorname{sgn} F(A, x_i) = -\operatorname{sgn} F(A, x_{i+1})$ $(i = 1, \ldots, s - 1)$ then

$$\inf_{\bar{A} \in P} \|F(\bar{A})\| \geqq \min_{1 \leqq i \leqq s} |F(A, x_i)|.$$

Throughout the remainder of the discussion we will assume that the best approximation $F(A^*, x)$ has the property

$$d(A^*) = \max_{A \in P} d(A).$$

1 I.e. if $h(x)$ is any element in the span of the components, then $h(x) \not\equiv 0 \Rightarrow h(x)$ has at most $d(A) - 1$ distinct zeros.

In the nomenclature of [2], [6], and [7], such an $F(A^*)$ is called a "normal point". Set $n = \max_{A \in P} d(A)$. In [4], the following result was proved.

Strong Uniqueness Theorem. If $F(A^*, x)$ is the best approximation and $d(A^*) = n$, there is a $\gamma > 0 \ni$ for all $A \in P$,

$$\|F(A)\| \geqq \|F(A^*)\| + \gamma \|F(A) - F(A^*)\|.$$

As indicated in the introduction we will need a close estimate of $F(A^*, x)$ to begin the algorithm. Let us try to see what this involves. We assume without loss of generality that the first n components of the gradient vector,

$$\left(\frac{\partial F(A^*, x)}{\partial a_1}, \ldots, \frac{\partial F(A^*, x)}{\partial a_m}\right),$$

are independent; that is, they are a basis for the linear span of the components. Now pick n distinct points $(x_1, \ldots, x_n)$ in $[0, 1]$ and consider the system,

$$F(A, x_i) = F(A^*, x_i) \quad (i = 1, \ldots, n).$$

The Haar hypotheses and the fact that P is open allows us to invoke the Implicit Function Theorem to show an $\varepsilon > 0$ can be found such that for any set $(\bar{a}_{n+1}, \bar{a}_{n+2}, \ldots, \bar{a}_m)$ with $|\bar{a}_{n+i} - a^*_{n+i}| < \varepsilon$ $(i = 1, \ldots, m-n)$, there is a unique solution $(a_1, \ldots, a_n)$ to the system,

$$F(a_1, \ldots, a_n, \bar{a}_{n+1}, \ldots, \bar{a}_m, x_i) = F(A^*, x_i) \quad (i = 1, \ldots, n).$$

Note then by the properties of the family F,

$$F(a_1, \ldots, a_n, \bar{a}_{n+1}, \ldots, \bar{a}_m, x) \equiv F(A^*, x).$$

We assume therefore that our initial estimate, $F(A^{(0)}, x)$ is such that the last $m-n$ components of $A^{(0)}$ lie within the "ε band", and we label the last $m-n$ components of $A^{(0)}$, $(a^*_{n+1}, \ldots, a^*_m)$. During the discussion, the last $m-n$ components of each A will be fixed at these values. Even though $F(A)$ is now a function of only n variables we do not rename $F(A)$. In the known cases there is no problem in finding such an $A^{(0)}$. For example for R^n_m the constant term in the denominator is set to one. For the exponentials there are not any "free variables".

III. Preliminary Results

The convergence proof will be based on a sequence of lemmas which will be developed in this section. To avoid having to talk about special cases in the discussion we rule out the trivial case; that is, we assume $F(A^*, x) \not\equiv 0$.

Now let

$$D = \{u = (u_0, u_1, \ldots, u_n) \in E_{n+1}\colon\ 0 \leqq u_0 < u_1 < \cdots < u_n \leqq 1\},$$

$$C = \{u \in D\colon\ F(A^*, u_i) = -F(A^*, u_{n+1}),\ i = 0, 1, \ldots, n-1;\ \|F(A^*)\| = |F(A^*, u_0)|\}.$$

For each $u \in E_{n+1}$ set $\|u\| = \max_{0 \leq i \leq n} |u_i|$. Further for each $A = (a_1, \ldots, a_m)$, let

$$\|A\| = \max_{1 \leq i \leq m} |a_i|.$$

Note that for each A and A' under discussion

$$\|A - A'\| = \max_{1 \leqq i \leqq n} |a_i - a_i'|$$

since these vectors agree in their last $m-n$ components.

Remark 1. For each $x \equiv (x_0, \ldots, x_n) \in D$, the $(n+1) \times (n+1)$ matrix,

$$\begin{pmatrix} \frac{\partial F(A^*, x_0)}{\partial a_1} \cdots \frac{\partial F(A^*, x_0)}{\partial a_n} & (-1) \\ \vdots & \vdots \\ \frac{\partial F(A^*, x_n)}{\partial a_1} \cdots \frac{\partial F(A^*, x_n)}{\partial a_n} & (-1)^{n+1} \end{pmatrix}$$

is non-singular.

Lemma 1. There is a $\delta_1 > 0$ such that for $\hat{\delta}_1 \leqq \delta_1$,

(a) $D(\hat{\delta}_1) = \{v \in D\colon \operatorname{dist}(v, C) \leqq \hat{\delta}_1\}$ is closed.

(b) $v \in D(\hat{\delta}_1) \Rightarrow$ the system of equations

$$F(A, v_i) = (-1)^i \lambda \quad (i = 0, 1, \ldots, n) \tag{1}$$

for λ and A has a unique continuous solution $(\lambda(v), A(v))$ where $A(v) \in P$ and the first n components of the gradient vector $\left(\frac{\partial F(A(v), x)}{\partial a_1}, \ldots, \frac{\partial F(A(v), x)}{\partial a_m}\right)$ are independent.

Proof. It is easy to see, since $F(A^*) \not\equiv 0$ and for each $u \in C$, $F(A^*, u_i)$ $(i = 0, 1, \ldots, n)$ alternates in sign, that a $\varrho > 0$ exists so that for each $u \in C$, $u_{i+1} - u_i \geqq \varrho$ $(i = 0, 1, \ldots, n-1)$. Conclusion (a) follows directly from this remark. Now assume conclusion (b) is false. Then there exists sequences $\{u^{(j)}\} \subset C$, $\{v^{(j)}\} \subset D$ such that $\|u^{(j)} - v^{(j)}\| \to 0$ and for each j, conclusion (b) is false. We can further assume $u^{(j)}, v^{(j)} \to u \in C$. By the Implicit Function Theorem, Remark 1, and the fact that P is open, there is a $\delta_1 > 0$ such that for $\|u - v\| \leqq \delta_1$, (1) has a unique continuous solution $(A(v), \lambda(v))$ where $A(v) \in P$. By making δ_1 smaller we can also secure the condition that the first n components of the gradient vector at $A(v)$ are independent. Now for large j, $\|u - v^{(j)}\| \leqq \delta_1$. This is a contradiction. The proof is thus complete.

Remark 2. For each $v \in D(\delta_1)$, the first n components of the gradient vector, $\left(\frac{\partial F(A(v), x)}{\partial a_1}, \ldots, \frac{\partial F(A(v), x)}{\partial a_m}\right)$, are a basis for a Haar Subspace of $C[0, 1]$. Hence for each $x \equiv (x_0, \ldots, x_n) \in D$, the $(n+1) \times (n+1)$ matrix,

$$\begin{pmatrix} \frac{\partial F(A(v), x_0)}{\partial a_1} \cdots \frac{\partial F(A(v), x_0)}{\partial a_n} & -1 \\ \vdots \quad \cdots \quad \vdots & \vdots \\ \frac{\partial F(A(v), x_n)}{\partial a_1} \cdots \frac{\partial F(A(v), x_n)}{\partial a_n} & (-1)^{n+1} \end{pmatrix}$$

is non-singular.

Remark 3. For arbitrary $\varepsilon_3 > 0$ $\exists$ $\delta_2 > 0$ so that $v \in D(\delta_2)$ implies the solution to (1), $(A(v), \lambda(v))$ satisfies

$$\|A^* - A(v)\| < \varepsilon_3.$$

For each k $(k=0, 1, \ldots, n)$ let $(k_1, k_2, \ldots, k_n)$ be the sequence defined as follows:

$$k_i = \begin{cases} i-1 & \text{for } 1 \leqq i \leqq k \\ i & \text{for } k+1 \leqq i \leqq n. \end{cases}$$

We need the following result in order to prove Lemma 3.

Lemma 2. There exist δ_3, $\varepsilon_3 > 0$ such that $\|A^{(i)} - A^*\| \leqq \varepsilon_3$ and $v \in D(\delta_3)$ $(i=0, 1, \ldots, n)$ imply

(a) $A^{(i)} \in P$ $(i=0, 1, \ldots, n)$.

(b) The $n \times n$ matrix $\left\{\frac{\partial F(A^{(k_i)}, v_{k_i})}{\partial a_j}\right.$ $\left. i, j = 1, \ldots, n\right\}$ has a non-zero determinant (always of the same sign) which is bounded uniformally away from zero $(k = 0, 1, \ldots, n)$. (This result is a generalization of the sign properties of Chebyshev systems, [1].)

Proof. The proof follows almost immediately from the compactness of C and the hypotheses dealing with the uniform continuity of $F(A, x)$ and the Haar conditions.

Let $\delta = \min(\delta_1, \delta_2, \delta_3)$.

Lemma 3. For any $u \in D(\delta)$ and $\bar{A} \in P$ with the properties,

(a) $\|\bar{A} - A^*\| < \varepsilon_3$,

(b) $\operatorname{sgn} F(\bar{A}, u_i) = -\operatorname{sgn} F(\bar{A}, u_{i+1})$ $(i=0, 1, \ldots, n-1)$

the solutions $(A(u), \lambda(u))$ to (1) satisfy

$$|\lambda(u)| = \sum_{i=0}^{n} \theta_i |F(\bar{A}, u_i)|$$

where

$$\sum_{i=0}^{n} \theta_i = 1, \quad \theta_i \geqq 1 - \theta \quad (i=0, 1, \ldots, n), \quad \theta \in (0, 1).$$

Proof. By the mean value theorem and Lemma 1,

$$\begin{aligned} F(\bar{A}, u_i) - (-1)^i \lambda(u) &= F(\bar{A}, u_i) - F(A(u), u_i) \\ &= \sum_{j=1}^{n} \frac{\partial F(A^{(i)}, u_i)}{\partial a_j} (\bar{a}_j - a_j(u)) \end{aligned} \tag{2}$$

where $\|A^{(i)} - A^*\| < \varepsilon_3$ $(i=0, 1, \ldots, n)$. Solving for $\lambda(u)$ in (2) via Cramer's rule,

$$\lambda(u) = \frac{\begin{vmatrix} F(\bar{A}, u_0) & \frac{\partial F(A^{(0)}, u_0)}{\partial a_1} & \cdots & \frac{\partial F(A^{(0)}, u_0)}{\partial a_n} \\ \vdots & & & \vdots \\ F(\bar{A}, u_n) & \frac{\partial F(A^{(n)}, u_n)}{\partial a_1} & \cdots & \frac{\partial F(A^{(n)}, u_n)}{\partial a_n} \end{vmatrix}}{\begin{vmatrix} +1 & \frac{\partial F(A^{(0)}, u_0)}{\partial a_1} & \cdots & \frac{\partial F(A^{(0)}, u_0)}{\partial a_n} \\ -1 & & \cdots & \\ \vdots & & & \vdots \\ (-1)^n & \frac{\partial F(A^{(n)}, u_n)}{\partial a_1} & \cdots & \frac{\partial F(A^{(n)}, u_n)}{\partial a_n} \end{vmatrix}}.$$

Now by Lemma 2 and the sign properties of $F(\bar{A}, u_i)$,

$$|\lambda(u)| = \sum_{i=0}^{n} \theta_i |F(\bar{A}, u_i)|$$

where $\theta_i \geqq 1-\theta$, $\sum_{i=0}^{n} \theta_i = 1$, $\theta \in (0, 1)$. Here θ depends only on δ and ε_3.

Lemma 4. Let $\{u^{(j)}\}$ be any sequence in $D(\delta)$ with the property that the corresponding solutions $\{A(u^{(j)}), \lambda(u^{(j)})\}$ to (1) satisfy

$$|\lambda(u^{(j)})| \to \|F(A^*)\|.$$

Then,

$$\|A(u^{(j)}) - A^*\| \to 0.$$

Proof. Assume the result is false. Then there is an $\varepsilon > 0$ and a sequence $\{u^{(j)}\} \subset D(\delta)$ satisfying the condition:

If $\left(A(u^{(j)}), \lambda(u^{(j)})\right)$ is the corresponding solution to (1) then,

$$|\lambda(u^{(j)})| \to \|F(A^*)\|,$$
$$\|A(u^{(j)}) - A^*\| \geqq \varepsilon \quad (j = 1, 2, \ldots).$$

Further one can assume $u^{(j)} \to u \in D(\delta)$. Now by Lemma 1, Remark 2, and the Implicit Function Theorem there is an $\varepsilon_1 > 0$ such that the system (1) has a unique and continuous solution $A(v)$ and $\lambda(v)$ for $\|u - v\| \leqq \varepsilon_1$. Hence $\|F(A^*)\| = |\lambda(u)|$. Note

$$|F(A^*, u_j)| \leqq \|F(A^*)\| = |F\left(A(u), u_j\right)| \quad (j = 0, 1, \ldots, n).$$

But since $F\left(A(u), x\right)$ alternates in sign over the set $(u_0, \ldots, u_n)$ we conclude $F(A^*, x)$ and $F\left(A(u), x\right)$ agree at n points. Hence $F(A^*, x) \equiv F\left(A(u), x\right)$. Thus $A^* = A(u)$, and $A(u^{(j)}) \to A^*$, which is a contradiction.

Lemma 5. For prescribed $\bar{\delta} > 0$ there is an $\varepsilon_2 > 0$ such that for any $A \in P$ and $v \in D$ with the properties;

(a) $\operatorname{sgn} F(A, v_i) = -\operatorname{sgn} F(A, v_{i+1}) \quad (i = 0, 1, \ldots, n-1)$,

(b) $\gamma = \min_{0 \leqq i \leqq n} |F(A, v_i)| \Rightarrow \|F(A)\| - \gamma < \varepsilon_2$, then the $\operatorname{dist}(v, C) \leqq \bar{\delta}$.

Proof. Assume the result is false. Then there exist sequences $\{A^{(j)}\} \subset P$, $\{\gamma^{(j)}\}$, $\{u^{(j)}\} \subset D$ such that

$$\gamma^{(j)} = \min_{0 \leqq i \leqq n} |F(A^{(j)}, u_i^{(j)})|,$$
$$\|F(A^{(j)})\| - \gamma^{(j)} \downarrow 0, \tag{3}$$
$$\operatorname{dist}(u^{(j)}, C) > \bar{\delta} \quad (j = 0, 1, 2, \ldots).$$

Now by the De La Vallée Poussin Lemma,

$$\|F(A^{(j)})\| - \gamma^{(j)} \geqq \|F(A^{(j)})\| - \|F(A^*)\| \geqq 0. \tag{4}$$

Hence by (3) and (4), $\|F(A^{(j)})\| \to \|F(A^*)\|$. Thus by the Strong Uniqueness Theorem,

$$\|F(A^{(j)}) - F(A^*)\| \to 0. \tag{5}$$

We can further assume $u^{(j)} \to (u_0, \ldots, u_n)$ where $u_i \leqq u_{i+1}$ $(i = 0, 1, \ldots, n-1)$. Now

$$|F(A^*, u_i) - F(A^{(j)}, u_i^{(j)})| \leqq |F(A^*, u_i) - F(A^*, u_i^{(j)})| + |F(A^*, u_i^{(j)}) - F(A^{(j)}, u_i^{(j)})|. \tag{6}$$

By (5) the right hand side of (6) goes to zero. Hence,

$$F(A^{(j)}, u_i^{(j)}) \to F(A^*, u_i) \quad (i = 0, 1, \ldots, n). \tag{7}$$

Now from (3) and (4)

$$|F(A^{(j)}, u_i^{(j)})| \to \|F(A^*)\| \quad (i = 0, 1, \ldots, n). \tag{8}$$

Finally using (7) and (8),

$$|F(A^*, u_i)| = \|F(A^*)\| \quad (i = 0, 1, \ldots, n). \tag{9}$$

It is easy to see then that,

$$\operatorname{sgn} F(A^*, u_i) = -\operatorname{sgn} F(A^*, u_{i+1}) \neq 0 \quad (i = 0, 1, \ldots, n-1). \tag{10}$$

But (9) and (10) imply $u \in C$. This is clearly a contradiction.

Lemma 6. For prescribed $\varepsilon_2 > 0$ there is an $\varepsilon_1 > 0$ such that for any $v \in D(\delta)$ with the property that the corresponding solutions $(A(v), \lambda(v))$ to (1) satisfy,

$$\|F(A^*)\| - |\lambda(v)| < \varepsilon_1,$$

then

$$\|F(A(v))\| - |\lambda(v)| < \varepsilon_2.$$

Proof. Assume the conclusion is false. Then there is an $\varepsilon_2 > 0$ and a sequence $\{u^{(j)}\} \subset D(\delta)$ where the corresponding solutions to (1) $A^{(j)}$, $\lambda^{(j)}$ satisfy

$$\|F(A^*)\| - |\lambda^{(j)}| \to 0,$$

$$\|F(A^{(j)})\| - |\lambda^{(j)}| \geqq \varepsilon_2 \quad (j = 0, 1, 2, \ldots).$$

Now by Lemma 4 $A^{(j)} \to A^*$. Hence $\|F(A^{(j)})\| \to \|F(A^*)\|$. Thus,

$$\|F(A^{(j)})\| - |\lambda^{(j)}| \to 0,$$

which is a contradiction.

Remark 4. It follows directly from the De La Vallée Poussin Lemma that

$$\max_{u \in D(\delta)} |\lambda(u)| = \|F(A^*)\|.$$

IV. Convergence of the Algorithm

The Remez Algorithm proceeds as follows. Given $A^{(i)} \in P$ and $u^{(i)} \in D$ such that

$$\operatorname{sgn} F(A^{(i)}, u_j^{(i)}) = -\operatorname{sgn} F(A^{(i)}, u_{j+1}^{(i)}) \quad (j = 0, 1, \ldots, n-1)$$

one solves the system

$$F(A, u_j^{(i)}) = (-1)^j \lambda \quad (j = 0, 1, \ldots, n)$$

to obtain $A^{(i+1)}$ and $\lambda^{(i+1)}$. Set $\alpha^{(i+1)} = |\lambda^{(i+1)}|$. Now $u^{(i+1)} \in D$ is chosen so that

(a) $\operatorname{sgn} F(A^{(i+1)}, u_j^{(i+1)}) = -\operatorname{sgn} F(A^{(i+1)}, u_{j+1}^{(i+1)}) \quad (j = 0, 1, \ldots, n-1)$.

(b) $\alpha^{(i+1)} \leqq \gamma^{(i+1)} \equiv \min_{0 \leqq j \leqq n} |F(A^{(i+1)}, u_j^{(i+1)})|$.

(c) $\|F(A^{(i+1)})\| = \max_{0 \leqq j \leqq n} |F(A^{(i+1)}, u_j^{(i+1)})|$.

We assume that the initial estimates $A^{(0)} \in P$ and $u^{(0)} \in D(\delta)$ satisfy,

1. $F(A^{(0)}, x)$ alternates in sign over $u^{(0)}$.
2. If $\gamma^{(0)} = \min_{0 \leqq i \leqq n} |F(A^{(0)}, u_i^{(0)})|$ then $\|F(A^*)\| - \gamma^{(0)} < \varepsilon_1$.
3. $\|A^{(0)} - A^*\| < \varepsilon_3$.

The key to the convergence proof for the non-linear Remez Algorithm is the following result. In the application of the preceeding lemmas we set $\bar{\delta} = \delta$ in Lemma 5.

Theorem 1. For each i $(i = 1, 2, 3, \ldots)$

(a) $\|F(A^*)\| - \gamma^{(i-1)} < \varepsilon_1$; $\quad \|F(A^*)\| - \alpha^i < \varepsilon_1$.

(b) $\|A^{(i)} - A^*\| < \varepsilon_3$.

(c) $\alpha^{(i)} = \sum_{j=0}^{n} \theta_j |F(A^{(i-1)}, u_j^{(i-1)})|$ where $\theta_j \geqq 1 - \theta$, $\theta \in (0, 1)$, $\sum_{j=0}^{n} \theta_j = 1$. Here θ is independent of i.

(d) $u^{(i)} \in D(\delta)$.

Proof. We proceed by induction. For $i = 1$, $\|F(A^*)\| - \gamma^{(0)} < \varepsilon_1$ by construction. Noting $u^{(0)} \in D(\delta)$, we surmise by Lemma 1 that the system,

$$F(A, u_j^{(0)}) = (-1)^j \lambda \quad (j = 0, 1, \ldots, n)$$

has a solution $(A^{(1)}, \lambda^{(1)})$ where $A^{(1)} \in P$. By Remark 3, $\|A^{(1)} - A^*\| < \varepsilon_3$. It follows from Lemma 3 that

$$\alpha^{(1)} \equiv |\lambda^{(1)}| = \sum_{j=0}^{n} \theta_j |F(A^{(0)}, u_j^{(0)})|$$

where $\theta_j \geqq 1 - \theta$, $\theta \in (0, 1)$ and $\sum_{j=0}^{n} \theta_j = 1$. Hence $\alpha^{(1)} \geqq \gamma^{(0)}$ which implies

$$\|F(A^*)\| - \alpha^{(1)} < \varepsilon_1.$$

Now $u^{(1)} \in D$ is chosen so that

(a) $\operatorname{sgn} F(A^{(1)}, u_j^{(1)}) = -\operatorname{sgn} F(A^{(1)}, u_{j+1}^{(1)}) \quad (j = 0, 1, \ldots, n-1)$.

(b) $\alpha^{(1)} \leqq \gamma^{(1)} \equiv \min_{0 \leqq j \leqq n} |F(A^{(1)}, u_j^{(1)})| \leqq \max_{0 \leqq j \leqq n} |F(A^{(1)}, u_j^{(1)})| = \|F(A^{(1)})\|$.

Since $\|F(A^*)\| - \alpha^{(1)} < \varepsilon_1$ and $u^{(0)} \in D(\delta)$ we infer by Lemma 6 that

$$\|F(A^{(1)})\| - \alpha^{(1)} < \varepsilon_2.$$

Hence

$$\|F(A^{(1)})\| - \gamma^{(1)} < \varepsilon_2.$$

Therefore by Lemma 5, $u^{(1)} \in D(\delta)$. This completes the first step of the induction. Now assume the results are true for $1 \leqq i \leqq k$. By induction $u^{(k)} \in D(\delta)$ and $\|F(A^*)\| - \alpha^{(k)} < \varepsilon_1$. Now by construction, $\gamma^{(k)} \geqq \alpha^{(k)}$ which implies

$$\|F(A^*)\| - \gamma^{(k)} < \varepsilon_1.$$

It follows from Lemma 1 and Remark 3 that the system

$$F(A, u_j^{(k)}) = (-1)^j \lambda \quad (j = 0, 1, \ldots, n)$$

has a solution $(A^{(k+1)}, \lambda^{(k+1)})$ such that $\|A^{(k+1)} - A^*\| < \varepsilon_3$ and $A^{(k+1)} \in P$. Again by Lemma 3, $\alpha^{k+1} \equiv |\lambda^{(k+1)}|$ has the property,

$$\alpha^{(k+1)} = \sum_{j=0}^{n} \theta_j |F(A^{(k)}, u_j^{(k)})|$$

where $\theta_j \geqq 1 - \theta$, $\theta \in (0, 1)$, $\sum_{j=0}^{n} \theta_j = 1$. Hence $\alpha^{(k+1)} \geqq \gamma^{(k)}$ which implies

$$\|F(A^*)\| - \alpha^{(k+1)} < \varepsilon_1.$$

Thus since $u^{(k)} \in D(\delta)$, Lemma 6 yields the result,

$$\|F(A^{(k+1)})\| - \alpha^{(k+1)} < \varepsilon_2.$$

Recall $u^{(k+1)} \in D$ is picked so that,

(a) $\operatorname{sgn} F(A^{(k+1)}, u_j^{(k+1)}) = -\operatorname{sgn} F(A^{(k+1)}, u_{j+1}^{(k+1)}) \quad j = 0, 1, \ldots, n-1.$

(b) $$\alpha^{(k+1)} \leqq \gamma^{(k+1)} = \min_{0 \leqq j \leqq n} |F(A^{(k+1)}, u_j^{(k+1)})| \leqq \max_{0 \leqq j \leqq n} |F(A^{(k+1)}, u_j^{(k+1)})| = \|F(A^{(k+1)})\|.$$

Hence $\|F(A^{(k+1)})\| - \gamma^{(k+1)} < \varepsilon_2$. By Lemma 5, $u^{(k+1)} \in D(\delta)$. This completes the induction.

Now using Theorem 1 together with the standard techniques employed in proving the convergence of the linear Remez Algorithm, [1], we can demonstrate the existence of constants θ and C where $\theta \in (0, 1)$ so that for all i

$$\|F(A^{(i)})\| - \|F(A^*)\| \leqq C\theta^i.$$

Hence by the Strong Uniqueness Theorem there is an additional constant A such that

$$\|F(A^{(i)}) - F(A^*)\| \leqq A\theta^i, \quad (i = 1, 2, \ldots).$$

This completes the proof of the convergence of non-linear Remez algorithm.

Bibliography

1. Cheney, E. W.: Introduction to approximation theory. McGraw-Hill Publishing Co., 1966, 97—100.
2. Werner, H.: Die Bedeutung der Normalität bei rationaler Tschebyscheff-Approximation. Computing **2**, 34—52 (1967).
3. Meinardus, G., Schwedt, D.: Nicht-lineare Approximationen. Arch. Rational Mech. Analysis **17**, 297—326 (1964).
4. Barrar, R. B., Loeb, H. L.: On the continuity of the non-linear Tschebyscheff operator. To appear in Pacific Journal of Mathematics.
5. Rice, J. R.: Tchebycheff approximations of functions unisolvent of variable degree. Trans. of Amer. Math. Soc. **99**, 298—302 (1961).
6. Cheney, E. W., Loeb, H. L.: Generalized rational approximation. SIAM Journal Series B, Numerical Analysis **1**, 11—25 (1964).
7. Loeb, H. L.: Approximation by generalized rationals. SIAM Numerical Analysis Journal **3**, 34—55 (1966).

Richard B. Barrar
Henry L. Loeb
Mathematics Department
University of Oregon
Eugene, Oregon

Numer. Math. 15, 392–403 (1970)
© by Springer-Verlag 1970

Über die Abhängigkeit der Fehlerfunktion bei der rationalen L_p-Approximation von den Koeffizienten des Nennerpolynoms bei minimalem Zählerpolynom

G. LAMPRECHT

Eingegangen am 23. Mai 1969

Einleitung

Bei der Approximation von Funktionen durch rationale Funktionen in der Tschebyscheff-Norm ist die beste Approximierende eindeutig bestimmt. Bei Zugrundelegen der L_p-Norm mit $1 \leqq p < \infty$ braucht dies nicht zu gelten[1]. Damit versagen die Verfahren zur Berechnung der besten rationalen Approximierenden, die an die Eindeutigkeit gebunden sind. In der vorliegenden Arbeit wird ein anderer Weg zur Bestimmung der besten L_p-Approximierenden vorgeschlagen.

Nach Vorgabe des Nennerpolynoms ist das Zählerpolynom eindeutig bestimmt, für das die L_p-Norm der Fehlerfunktion minimal ist. Dieses „zählerminimale Polynom" und die L_p-Norm hängen für alle Koeffizientenvektoren des Nennerpolynoms stetig von diesen ab. Für solche Vektoren, für die das Nennerpolynom keine Nullstellen im Approximationsintervall hat, ist die Abhängigkeit sogar stetig differenzierbar bis zur Ordnung $[p-1]$. Um die besten Approximierenden zu erhalten, hat man den abgeschlossenen Bereich der Nennerkoeffizientenvektoren zu untersuchen, in dem diese betragsmäßig kleiner oder gleich eins sind.

Allerdings muß man die Menge der rationalen Funktionen etwas allgemeiner als sonst üblich einführen.

Es seinen

$$\mathfrak{a} := (a_0, a_1, \ldots, a_l),$$

$$\mathfrak{b} := (b_0, b_1, \ldots, b_r)$$

und

$$\mathfrak{R}_{l,r} := \{(\mathfrak{a}, \mathfrak{b}) \,|\, |\mathfrak{b}| = 1\} \quad \text{mit} \quad |\mathfrak{b}| := \sqrt{\sum_{j=0}^{r} b_j^2}.$$

Jedem Vektor $(\mathfrak{a}, \mathfrak{b}) \in \mathfrak{R}_{l,r}$ wird eine rationale Funktion

$$R(x) = \frac{P_l(x)}{Q_r(x)}$$

zugeordnet durch

$$P_l(x) := \sum_{i=0}^{l} a_i x^i \quad \text{und} \quad Q_r(x) := \sum_{j=0}^{r} b_j x^j.$$

[1] Hierauf soll an anderer Stelle ausführlicher eingegangen werden (Computing, 1970).

Die Menge aller so gebildeten rationalen Funktionen und aller Funktionen, die sich von einer derartigen rationalen Funktion nur in einer Punktmenge vom Maße Null unterscheiden, wird mit $\mathfrak{R}_{l,r}$ bezeichnet.

Die Abbildung von $\mathfrak{R}_{l,r}$ in $\mathfrak{R}_{l,r}$ ist eindeutig. Um jeder Funktion aus $\mathfrak{R}_{l,r}$ in eindeutiger Weise einen Vektor aus $\mathfrak{R}_{l,r}$ zuordnen zu können, werden Funktionen als gleich angesehen, die sich nur in einer Punktmenge vom Maße Null unterscheiden. Die jeweilige rationale Funktion sei ferner in eine solche Form gebracht, daß Zähler- und Nennerpolynom keinen nichttrivialen Teiler besitzen und die Koeffizienten b_j des Nennerpolynoms normiert sind durch

$$\sum_{j=0}^{r} b_j^2 = 1;$$

dabei sei der höchste von Null verschiedene Koeffizient positiv.

Definition 1. Es sei f eine Funktion aus L_p mit $1 \leqq p < \infty$. Zu vorgegebenem Koeffizientenvektor $\mathfrak{b}^0 = (b_0^0, \ldots, b_r^0)$ des Nennerpolynoms Q_r^0 heißen der Koeffizientenvektor $\mathfrak{a}^0 = (a_0^0, \ldots, a_l^0)$ und das zugehörige Polynom P_l^0 zählerminimal (bezüglich f und der L_p-Norm), wenn für alle Polynome $t(x)$, deren Grad höchstens l ist, die Beziehung

$$\left\| f - \frac{P_l^0}{Q_r^0} \right\|_p \leqq \left\| f - \frac{t}{Q_r^0} \right\|_p$$

besteht; dabei wird hier und im folgenden

$$\|f\|_p := \left[\int_{-1}^{1} |f|^p \, dx \right]^{1/p}$$

gesetzt.

Satz 1. Es sei f eine Funktion aus dem Raume L_p mit $1 < p < \infty$, die nicht zur Menge $\mathfrak{R}_{l,r}$ gehört.

1. Dann ist nach Vorgabe des Koeffizientenvektors $\mathfrak{b}^0$ des Nennerpolynoms Q_r^0 das zählerminimale Polynom P_l^0 eindeutig bestimmt.

2. Das Nennerpolynom Q_r^0 sei in der Form zerlegt

$$Q_r^0(x) = T_\mu(x) \cdot Q_{r-\mu}^0(x)$$

mit

$$Q_{r-\mu}^0(x) \neq 0 \quad \text{im Intervall } [-1, 1]$$

und

$$T_\mu(x) := \prod_i (x - x_i)^{\sigma_i},$$

wobei gilt

$$\mu := \sum_i \sigma_i, \qquad \mu \in \{0, 1, \ldots, r\}$$

und

$$x_{i_1} \neq x_{i_2} \quad \text{für} \quad i_1 \neq i_2, \qquad x_i \in [-1, 1].$$

Falls $\mu \leqq l$ ist, gilt dann für das zählerminimale Polynom $P_l^0(x)$ die Beziehung

$$P_l^0(x) = T_\mu(x) \cdot P_{l-\mu}^0(x).$$

Das Polynom $P^0_{l-\mu}(x)$ ist zählerminimal zu $Q^0_{r-\mu}$ und läßt sich eindeutig aus den folgenden $(l-\mu+1)$ Gleichungen

$$(1) \qquad \int_{-1}^{1} \left| f - \frac{P^0_{l-\mu}}{Q^0_{r-\mu}} \right|^{p-1} \operatorname{sign}\left(f - \frac{P^0_{l-\mu}}{Q^0_{r-\mu}} \right) \cdot \frac{x^i}{Q^0_{r-\mu}}\, dx = 0 \quad \text{für} \quad i = 0, 1, \ldots, l-\mu$$

bestimmen.

Falls $\mu > l$ ist, erhält man als zählerminimales Polynom

$$P^0_l \equiv 0.$$

Beweis. I. Zunächst besitze das Nennerpolynom Q^0_r keine Nullstelle im Approximationsintervall $[-1, 1]$. Setzt man

$$f_1(x) := f(x) \cdot Q^0_r(x)$$

und

$$w(x) := |Q^0_r(x)|^{-p},$$

so sieht man, daß die Bestimmung des zählerminimalen Polynoms P^0_l ein lineares Approximationsproblem in der strikt konvexen Norm

$$\left[\int_{-1}^{1} |f_1 - P^0_l|\, w\, dx \right]^{1/p}$$

ist. Dieses Problem ist bekanntlich eindeutig lösbar. Setzt man

$$\Phi(a_0, a_1, \ldots, a_l) := \int_{-1}^{1} |f_1 - P_l|^p\, w\, dx,$$

so ist diese Funktion wegen $p > 1$ stetig partiell differenzierbar nach ihren Argumenten $a_0, a_1, \ldots, a_l$ (man vgl. hierzu etwa [1], § 575). Aus den für das Vorliegen eines Minimums notwendigen und hinreichenden Bedingungen (vgl. [3], S. 18)

$$\frac{\partial \Phi}{\partial a_i} = 0, \qquad i = 0, 1, \ldots, l$$

folgen die behaupteten Bestimmungsgleichungen (1).

II. Nun besitze das Nennerpolynom $Q^0_r(x)$ Nullstellen im Approximationsintervall.

Da bei der Suche nach dem zählerminimalen Polynom P^0_l das Nullpolynom eine zugelassene Konkurrenzfunktion ist, gilt

$$\left\| f - \frac{P^0_l}{Q^0_r} \right\|_p \leqq \|f\|_p,$$

woraus

$$\left\| \frac{P^0_l}{Q^0_r} \right\|_p \leqq 2\|f\|_p$$

folgt. Damit muß das zählerminimale Polynom P^0_l im Approximationsintervall an denselben Stellen Nullstellen von mindestens derselben Ordnung besitzen wie das Nennerpolynom Q^0_r. Für $\mu > l$ impliziert dieses das identische Verschwinden des zählerminimalen Polynoms P^0_l. Im anderen Fall kann man von P^0_l den

Faktor $T_\mu(x)$ abspalten, d.h. es gilt

$$P_l^0 = T_\mu \cdot P_{l-\mu}^0,$$

und das Polynom $P_{l-\mu}^0$ muß bezüglich $Q_{r-\mu}^0$ zählerminimal sein. Auf Grund des obigen Beweisteiles I ist $P_{l-\mu}^0$ eindeutig bestimmt und durch die $(l-\mu+1)$ Bestimmungsgleichungen (1) berechenbar. Damit ist die Behauptung des Satzes 1 bewiesen.

Für das Folgende sei

$$\mathfrak{D} := \left\{ \mathfrak{b} \,\middle|\, |\mathfrak{b}| = 1,\ \sum_{j=0}^{r} b_j x^j \neq 0 \text{ im Intervall } [-1, 1] \right\}.$$

Satz 2. Es sei $f \in L_p$ mit $1 < p < \infty$.

Für $\mathfrak{b} \in \mathfrak{D}$ hängen der zählerminimale Vektor $\mathfrak{a}^0$ und die Norm

$$\left\| f - \frac{P_l^0}{Q_r} \right\|_p$$

stetig von $\mathfrak{b}$ ab.

Dieser Satz 2 stellt eine Folgerung aus dem folgenden Lemma 1 dar, das man — unter Ausnutzung der Konvexität der Norm — elementar beweisen kann. Auf den Beweis sei an dieser Stelle verzichtet.

Lemma 1. Es sei f eine Funktion aus L_p mit $1 < p < \infty$ und $w(x)$ sei eine stetige Funktion, die im Intervall $[-1, 1]$ positiv ist.

Es sei

$$\psi(f, \mathfrak{a}, w) := \left[\int_{-1}^{1} |f - P_l|^p w \, dx \right]^{1/p},$$

wobei $\mathfrak{a}$ der zum Polynom P_l gehörende Koeffizientenvektor ist.

Ferner sei

$$\eta_w[f] := \min_{\mathfrak{a}} \psi(f, \mathfrak{a}, w).$$

Dann gilt

1. Die Funktion ψ hängt stetig von f, $\mathfrak{a}$ und w ab.

2. Der Koeffizientenvektor $\mathfrak{a}_{f,w}$, für den $\psi(f, \mathfrak{a}, w)$ bei vorgegebener Funktion f und w minimal wird, ändert sich stetig mit f und w.

3. Die Funktion $\eta_w[f]$ hängt stetig von f und w ab.

Satz 3. Es sei f eine Funktion aus dem Raum L_p mit $2 \leqq p < \infty$, die nicht zur Menge $\mathfrak{R}_{l,r}$ gehört.

Dann besitzen die zählerminimalen Koeffizienten $a_0^0, a_1^0, \ldots, a_l^0$ und die Norm der Fehlerfunktion $f - \frac{P_l^0}{Q_r}$ als Funktionen von $b_1, b_2, \ldots, b_r$ für alle Vektoren $\mathfrak{b} = (b_0, b_1, \ldots, b_r) \in \mathfrak{D}$ stetige partielle Ableitungen bis zur Ordnung $[p-1]$.

Ist p eine gerade Zahl, so besitzen die angegebenen Funktionen stetige partielle Ableitungen beliebig hoher Ordnung.

Beweis. Für $i=0, 1, \ldots, l$ sei

$$F_i(a_0, \ldots, a_l, b_1, \ldots, b_r) := \int_{-1}^{1} \left| f - \frac{P_l}{Q_r} \right|^{p-1} \operatorname{sign}\left(f - \frac{P_l}{Q_r}\right) \cdot \frac{x^i}{Q_r}\, dx.$$

Dann sind die zählerminimalen Koeffizienten $a_0^0, a_1^0, \ldots, a_l^0$ nach Satz 1 eindeutig bestimmt durch das Gleichungssystem

$$F_i(a_0^0, \ldots, a_l^0, b_1, \ldots, b_r) = 0 \quad \text{für} \quad i=0, 1, \ldots, l.$$

Wegen $b \in \mathfrak{D}$ sind die Funktionen F_i in allen Argumenten definiert und stetig. Wegen $[p-1] \geqq 1$ sind außerdem die partiellen Ableitungen der Funktionen F_i mit $i=0, 1, \ldots, l$ nach allen Argumenten bis zur Ordnung $[p-1]$ vorhanden und stetig. Falls p gerade ist, gilt dieses für die partiellen Ableitungen beliebig hoher Ordnung.

Die zugehörige Funktionaldeterminante

$$\det\left(\frac{\partial F_i}{\partial a_k}\right)_{i,k=0,1,\ldots,l}$$

ist gerade die Gramsche Determinante der Potenzen x^k mit $k=0, 1, \ldots, l$ zur Gewichtsfunktion

$$w(x) := -(p-1) \cdot \left| f - \frac{P_l^0}{Q_r} \right|^{p-2} \cdot \frac{1}{Q_r^2}.$$

Da die Funktion f nach Voraussetzung zu keiner Funktion aus der Menge $\mathfrak{R}_{l,r}$ äquivalent ist, existiert eine Punktmenge mit positivem Maß, in der $w(x)$ von Null verschieden ist. Die Potenzen x^k, $k=0, 1, \ldots, l$ sind linear unabhängig in dieser Punktmenge, also ist die Gramsche Determinante von Null verschieden. Nach dem sog. Satz über implizite Funktionen gilt damit die Behauptung von Satz 3.

Beschränkt man sich für die Koeffizientenvektoren $\mathfrak{b}$ nicht auf die Menge $\mathfrak{D}$, sondern läßt beliebige, zu 1 normierte Vektoren zu, so kann das zugehörige Nennerpolynom Nullstellen im Approximationsintervall $[-1, 1]$ besitzen. Man kann nicht erwarten, daß die zählerminimalen Koeffizienten $a_0^0, \ldots, a_l^0$ und die Norm der Fehlerfunktion noch differenzierbar von $b_1, \ldots, b_r$ abhängen. Immerhin läßt sich eine stetige Abhängigkeit nachweisen. Dieses soll in den folgenden beiden Sätzen geschehen.

Satz 4. Es sei f eine Funktion aus dem Raume L_p mit $1<p<\infty$, die nicht zur Menge $\mathfrak{R}_{l,r}$ gehört. Der Koeffizientenvektor $\mathfrak{b}^0$ gehöre dem Rand von $\mathfrak{D}$ an. Ferner sei eine Folge $\{\mathfrak{b}^n\}_{n=1,2,\ldots}$ von Vektoren aus $\mathfrak{D}$ gegeben, die gegen $\mathfrak{b}^0$ konvergiert. Die zugehörigen Nennerpolynome seien mit Q_r^0 und Q_r^n bezeichnet, die zugehörigen zählerminimalen Koeffizientenvektoren und Polynome mit $\mathfrak{a}^0$, P_l^0 und $\mathfrak{a}^n$, P_l^n.

Dann gilt:

Die Folge $\{\mathfrak{a}^n\}_{n=1,2,\ldots}$ konvergiert gegen $\mathfrak{a}^0$ und die Folge

$$\left\{ \left\| f - \frac{P_l^n}{Q_r^n} \right\|_p \right\}_{n=1,2,\ldots} \quad \text{gegen} \quad \left\| f - \frac{P_l^0}{Q_r^0} \right\|_p.$$

Beweis. Weil nach Voraussetzung $\mathfrak{b}^0$ zum Rand von $\mathfrak{D}$ gehört, besitzt das Polynom Q_r^0 im Approximationsintervall $[-1, 1]$ Nullstellen. Man kann Q_r^0 zerlegen in

$$Q_r^0(x) = T_m^0(x) \cdot Q_{r-m}^0(x),$$

wobei Q_{r-m}^0 im Intervall $[-1, 1]$ von Null verschieden ist und für T_m^0 gilt

$$T_m^0(x) := \prod_i (x - x_i)^{\sigma_i}$$

mit $m := \sum_i \sigma_i$ und $x_i \in [-1, 1]$, $x_i \neq x_k$ für $i \neq k$.

Die Menge der Nullstellen von Q_r^0 im Intervall $[-1, 1]$ sei mit Z bezeichnet.

Für das zählerminimale Polynom P_l^0 folgt nach Satz 1

$$P_l^0 = T_m^0 \cdot P_{l-m}^0.$$

Jedes der zu $\mathfrak{b}^n$ mit $n = 1, 2, \ldots$ gehörende Polynom Q_r^n kann man in seine Linearfaktoren zerlegen und diese so zusammenfassen, daß für

$$Q_r^n(x) = T_m^n(x) \cdot Q_{r-m}^n(x)$$

beide Faktoren reell sind, der höchste Koeffizient von $T_m^n(x)$ gleich 1 ist und daß ferner

$$T_m^n(x) \quad \text{gegen} \quad T_m^0(x)$$

und

$$Q_{r-m}^n(x) \quad \text{gegen} \quad Q_{r-m}^0(x)$$

konvergiert, wenn n über alle Grenzen wächst.

Nach Voraussetzung sind die Faktoren $T_m^n(x)$ und $Q_{r-m}^n(x)$ für alle $n = 1, 2, \ldots$ im Approximationsintervall von Null verschieden.

Die zugehörigen zählerminimalen Polynome P_l^n werden zerlegt in der Form

$$P_l^n(x) = T_m^n(x) \cdot P_{l-m}^n(x) + H_{m-1}^n(x), \tag{2}$$

wobei der Grad von $H_{m-1}^n(x)$ kleiner oder gleich $(m-1)$ ist. Gilt $m > l$, so ist $P_{l-m}^n(x) \equiv 0$ zu setzen.

Da die Polynome $P_l^n(x)$ und $T_m^n(x)$ eindeutig bestimmt sind, folgt dieses auch für P_{l-m}^n und H_{m-1}^n. Die L_p-Norm der Polynome P_l^n ist durch eine von n unabhängige Konstante beschränkt; damit gilt das gleiche auch für die Koeffizienten von P_l^n. Der höchste Koeffizient von T_m^n ist nach Konstruktion gleich 1. Also ist auch der Koeffizientenvektor von P_{l-m}^n beschränkt durch eine von n unabhängige Konstante. Die gleiche Aussage folgt anschließend für den Koeffizientenvektor von H_{m-1}^n, d. h. er liegt in einer beschränkten, abgeschlossenen Kugel des m-dimensionalen euklidischen Raumes.

Es soll nun gezeigt werden, daß die Folge der Polynome $\{H_{m-1}^n\}_{n=1,2,\ldots}$ gegen eine zum Nullpolynom äquivalente Funktion strebt, und daß die Folge der Polynome $\{T_m^n \cdot P_{l-m}^n\}_{n=1,2,\ldots}$ gegen eine Funktion strebt, die zu $T_m^0 \cdot P_{l-m}^0$ äquivalent ist.

Da $T_m^n \cdot P_{l-m}^0$ ein bei der Suche nach dem zählerminimalen Polynom P_l^n zugelassenes Konkurrenzpolynom ist, gilt für alle $n = 1, 2, \ldots$

$$\left\| f - \frac{P_l^n}{Q_r^n} \right\|_p \leqq \left\| f - \frac{P_{l-m}^0}{Q_{r-m}^n} \right\|_p . \tag{3}$$

Wegen der Gl. (2) folgt hieraus durch Anwenden der Dreiecksungleichung

$$\left\| \frac{H_{m-1}^n}{T_m^n \cdot Q_{r-m}^n} \right\|_p \leqq \left\| f - \frac{P_{l-m}^n}{Q_{r-m}^n} \right\|_p + \left\| f - \frac{P_{l-m}^0}{Q_{r-m}^n} \right\|_p .$$

Da die Folge der Polynome Q_{r-m}^n gegen das im Intervall $[-1, 1]$ von Null verschiedene Polynom Q_{r-m}^0 konvergiert, und da nach den obigen Überlegungen die Polynome P_{l-m}^n durch eine gemeinsame Konstante beschränkt sind, gilt

$$\left\| \frac{H_{m-1}^n}{T_m^n \cdot Q_{r-m}^n} \right\|_p \leqq k \tag{4}$$

mit einer von n unabhängigen Konstanten k.

Da die zur Folge $\{H_{m-1}^n\}_{n=1,2,\ldots}$ gehörenden Koeffizientenvektoren in einer beschränkten, abgeschlossenen Menge liegen, kann man eine konvergente Teilfolge der Koeffizientenvektoren angeben. Diese sei durch die Indexmenge J_1 festgelegt. Dann konvergiert die Folge $\{H_{m-1}^n\}_{n \in J_1}$ gegen ein Polynom H_{m-1}^*, dessen Grad höchstens $m-1$ ist. Wegen der Beziehung (4) muß gelten

$$\left\| \frac{H_{m-1}^*}{T_m^0 \cdot Q_{r-m}^0} \right\|_p \leqq k .$$

Das Polynom H_{m-1}^* muß also in denselben Punkten x_i aus Z Nullstellen von mindestens derselben Ordnung σ_i besitzen wie das Polynom T_m^0; denn sonst könnte die Norm nicht beschränkt sein. Da der Grad von H_{m-1}^* höchstens $(m-1)$ ist, während T_m^0 genau m Nullstellen (mit Vielfachheit) besitzt, muß H_{m-1}^* identisch verschwinden. Also ist für alle Punkte $x \in [-1, 1] - Z$

$$\lim_{\substack{n \to \infty \\ n \in J_1}} \frac{H_{m-1}^n(x)}{T_m^n(x) \cdot Q_{r-m}^n(x)} = 0, \tag{5}$$

und da Z eine Menge vom Maße Null ist, gilt

$$\left\| \lim_{\substack{n \to \infty \\ n \in J_1}} \frac{H_{m-1}^n}{T_m^n \cdot Q_{r-m}^n} \right\|_p = 0 .$$

Diese Beziehung muß für jede konvergente Teilfolge von $\{H_{m-1}^n\}_{n=1,2,\ldots}$ gelten, und damit für die gesamte Folge.

Da die Koeffizientenvektoren der Polynome P_{l-m}^n mit $n = 1, 2, \ldots$ in einer beschränkten und abgeschlossenen Menge liegen, existiert eine konvergente Teilfolge der Koeffizientenvektoren. Diese sei durch die Indexmenge J_2 festgelegt. Es sei

$$P_{l-m}^* := \lim_{\substack{n \to \infty \\ n \in J_2}} P_{l-m}^n .$$

Da das Grenzpolynom Q^0_{r-m} der Folge $\{Q^0_{r-m}\}_{n=1,2,\ldots}$ nach Konstruktion im Approximationsintervall von Null verschieden ist, gilt

$$\lim_{\substack{n\to\infty \\ n\in J_2}} \frac{P^n_{l-m}}{Q^n_{r-m}} = \frac{P^*_{l-m}}{Q^0_{r-m}}, \tag{6}$$

und diese Grenzfunktion ist offensichtlich stetig im Intervall $[-1, 1]$. Also gehört die Grenzfunktion

$$\lim_{\substack{n\to\infty \\ n\in J_2}} \frac{P^n_l}{Q^n_r} = \lim_{\substack{n\to\infty \\ n\in J_2}} \frac{P^n_{l-m}}{Q^n_{r-m}} + \lim_{\substack{n\to\infty \\ n\in J_2}} \frac{H^n_{m-1}}{T^n_m \cdot Q^n_{r-m}} \tag{7}$$

zum Raume L_p und ist wegen (5) und (6) der rationalen Funktion

$$\frac{P^*_{l-m}}{Q^0_{r-m}}$$

äquivalent, gehört also zur Menge $\mathfrak{R}_{l-m,r-m}$.

Man kann die Abschätzung (3)

$$\left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \leqq \left\| f - \frac{P^0_{l-m}}{Q^n_{r-m}} \right\|_p$$

weiter umformen

$$\left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \leqq \left\| f - \frac{P^0_{l-m}}{Q^0_{r-m}} \right\|_p + \left\| \frac{P^0_{l-m}}{Q^0_{r-m}} - \frac{P^0_{l-m}}{Q^n_{r-m}} \right\|_p. \tag{8}$$

Man kann nun eine von n unabhängige Konstante angeben, die eine obere Schranke für die rechte Seite von (8) darstellt. Damit besitzt die Folge

$$\left\{ \left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \right\}_{n=1,2,\ldots}$$

eine konvergente Teilfolge. Diese sei durch die Indexmenge J_3 festgelegt, die ohne Einschränkung in der Menge J_2 enthalten sei.

Da nach Konstruktion die Folge $\{Q^n_{r-m}\}_{n=1,2,\ldots}$ gegen das Polynom Q^0_{r-m} konvergiert, muß wegen (8) gelten

$$\lim_{\substack{n\to\infty \\ n\in J_3}} \left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \leqq \left\| f - \frac{P^0_{l-m}}{Q^0_{r-m}} \right\|_p.$$

Nach einem Satz von Fatou (vgl. [2], S. 155/156) kann man damit die Einschließung

$$\left\| f - \lim_{\substack{n\to\infty \\ n\in J_3}} \frac{P^n_l}{Q^n_r} \right\|_p \leqq \lim_{\substack{n\to\infty \\ n\in J_3}} \left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \leqq \left\| f - \frac{P^0_{l-m}}{Q^0_{r-m}} \right\|_p$$

angeben. Nach den Beziehungen (6) und (7) gilt also

$$\left\| f - \frac{P^*_{l-m}}{Q^0_{r-m}} \right\|_p \leqq \lim_{\substack{n\to\infty \\ n\in J_3}} \left\| f - \frac{P^n_l}{Q^n_r} \right\|_p \leqq \left\| f - \frac{P^0_{l-m}}{Q^0_{r-m}} \right\|_p. \tag{9}$$

Nun ist das Polynom P^0_{l-m} zählerminimal zu Q^0_{r-m} und da dieses Polynom eindeutig bestimmt ist, muß

$$P^*_{l-m} = P^0_{l-m} \tag{10}$$

sein, d.h. es gilt in der Formel (9) das Gleichheitszeichen. Also ist

$$\lim_{\substack{n\to\infty \\ n\in J_3}} \left\| f - \frac{P_l^n}{Q_r^n} \right\|_p = \left\| f - \frac{P_{l-m}^0}{Q_{r-m}^0} \right\|_p,$$

und dieses muß für jede konvergente Teilfolge und damit für die gesamte Folge gelten.

Wegen (7) und (10) muß jede konvergente Teilfolge der Polynome

$$P_l^n = T_m^n \cdot P_{l-m}^n + H_{m-1}^n$$

gegen $T_m^0 \cdot P_{l-m}^0$ konvergieren. Wegen der Eindeutigkeit der Grenzfunktion muß es für die gesamte Folge gelten. Damit ist auch die Behauptung gezeigt, daß die Koeffizientenvektoren $\mathfrak{a}^n$ von P_l^n gegen den Koeffizientenvektor $\mathfrak{a}^0$ von $T_m^0 \cdot P_{l-m}^0$ konvergiert.

Satz 5. Es sei f eine Funktion aus L_p mit $1 < p < \infty$, die nicht zur Menge $\mathfrak{R}_{l,r}$ gehört. Für alle Koeffizientenvektoren $\mathfrak{b}$ des Nennerpolynoms Q_r gelte die Normierung $|\mathfrak{b}| = 1$. Dann hängen der Koeffizientenvektor $\mathfrak{a}^0$ des zählerminimalen Polynoms P_l^0 und die Norm der Fehlerfunktion $f - \frac{P_l^0}{Q_r}$ stetig von dem Koeffizientenvektor $\mathfrak{b}$ ab.

Beweis. Für diesen Beweis werden Mengen $\mathfrak{D}_\varrho(M_1, M_2)$ für $\varrho = 0, 1, \ldots, r$ und $0 < M_1 \leqq M_2 < \infty$ in der folgenden Weise definiert

$$\mathfrak{D}_\varrho(M_1, M_2) := \Big\{ \mathfrak{b}_\varrho \,|\, \mathfrak{b}_\varrho = (b_0, \ldots, b_\varrho);\ M_1 \leqq |\mathfrak{b}_\varrho| \leqq M_2 \text{ und } \sum_{j=0}^{\varrho} b_j x^j \neq 0 \text{ im Intervall } [-1, 1] \Big\}.$$

Offensichtlich stimmt $\mathfrak{D}_r(1, 1)$ mit der Menge $\mathfrak{D}$ überein.

Es sei $\mathfrak{b}_r^0$ ein beliebiger, aber fester Vektor mit $|\mathfrak{b}_r^0| = 1$. Ferner sei $\{\mathfrak{b}_r^n\}_{n=1,2,\ldots}$ eine beliebige Folge normierter Vektoren, die gegen $\mathfrak{b}_r^0$ konvergiert. Man kann annehmen, daß $\mathfrak{b}_r^0$ nicht zu $\mathfrak{D}$ gehört, denn sonst folgt die Behauptung aus Satz 2. Das Polynom Q_r^0 besitzt damit Nullstellen im Intervall $[-1, 1]$. Man kann Q_r^0 also folgendermaßen zerlegen

$$Q_r^0 = T_m^0 \cdot Q_{r-m}^0, \tag{11}$$

wobei für T_m^0 und Q_{r-m}^0 die gleichen Vereinbarungen wie im Beweis von Satz 4 gelten.

Da der höchste Koeffizient von T_m^0 nach Konstruktion den Wert 1 hat, folgt aus der Beschränktheit des Koeffizientenvektors $\mathfrak{b}_r^0$ die Beschränktheit des Koeffizientenvektors $\mathfrak{b}_{r-m}^0$ des Polynoms Q_{r-m}^0, d.h. es existiert eine Konstante M_2 mit

$$|\mathfrak{b}_{r-m}^0| \leqq M_2.$$

Andererseits existiert eine positive untere Schranke M_1 für $|\mathfrak{b}_{r-m}^0|$, denn sonst folgte aus $|\mathfrak{b}_{r-m}^0| = 0$ über (11) das identische Verschwinden von Q_r^0, was wegen der Normierung $|\mathfrak{b}_r^0| = 1$ nicht sein kann.

Analog zu (11) kann man alle Polynome Q_r^n in der Form

$$Q_r^n = T_j^n \cdot Q_{r-j}^n \tag{12}$$

darstellen. Abgesehen von endlich vielen Ausnahmen gilt für den Grad j der Polynome T_j^n die Einschließung

$$0 \leqq j \leqq m.$$

Mit Hilfe von (12) kann man in eindeutiger Weise der Folge $\{\mathfrak{b}_r^n\}_{n=1,2,\ldots}$ die Folgen $\{\mathfrak{b}_{r-j}^k\}_{k\in J_j}$ zuordnen, wobei gewisse endliche oder unendliche Indexmengen J_j mit $j=0, 1, \ldots, m$ definiert werden. Für das Folgende sind nur die unendlichen Indexmengen zu berücksichtigen.

Da die Folge $\{\mathfrak{b}_r^n\}_{n=1,2,\ldots}$ nach Voraussetzung gegen $\mathfrak{b}_r^0$ konvergiert, existieren positive und endliche Schranken M_{1j}, M_{2j}, so daß für alle Koeffizientenvektoren $\mathfrak{b}_{r-j}^k$ die Beziehung gilt

$$M_{1j} \leqq |\mathfrak{b}_{r-j}^k| \leqq M_{2j}.$$

Damit liegen die Folgen $\{\mathfrak{b}_{r-j}^k\}_{k\in J_j}$ in den Mengen $\mathfrak{D}_{r-j}(M_{1j}, M_{2j})$. Für $j=0, 1, \ldots, m-1$ müssen alle Häufungspunkte der Folgen $\{\mathfrak{b}_{r-j}^k\}_{k\in J_j}$ auf dem Rand von $\mathfrak{D}_{r-j}(M_{1j}, M_{2j})$ liegen, und da alle Elemente der Folgen der Norm nach zwischen positiven, endlichen Schranken liegen, kann man für jeden Häufungspunkt einer jeden Folge den Satz 4 anwenden. Man erhält für jede Folge $\{\mathfrak{b}_{r-j}^k\}_{k\in J_j}$ die Beziehungen

$$\lim_{\substack{k\to\infty \\ k\in J_j}} \left\| f - \frac{T_j^k \cdot P_{l-j}^k}{T_j^k \cdot Q_{r-j}^k} \right\|_p = \left\| f - \frac{T_m^0 \cdot P_{l-m}^0}{T_m^0 \cdot Q_{r-m}^0} \right\|_p$$

und

$$\lim_{\substack{k\to\infty \\ k\in J_j}} \mathfrak{a}^k = \mathfrak{a}^0,$$

wobei für $k \in J_j$ der Koeffizientenvektor $\mathfrak{a}^k$ zu dem Polynom

$$P_l^k = T_j^k \cdot P_{l-j}^k$$

gehört. Für $j=0, 1, \ldots, m-1$ ist damit die Behauptung gezeigt.

Für $j=m$ konvergiert die Folge $\{T_m^k\}_{k\in J_m}$ gegen T_m^0. Es existiert für die Folge $\{\mathfrak{b}_{r-m}^k\}_{k\in J_m}$ nur ein Häufungspunkt, der im Innern von $\mathfrak{D}_{r-m}(M_{1m}, M_{2m})$ liegt. Nach Satz 1 gilt für die zählerminimalen Polynome

$$P_l^k = T_m^k \cdot P_{l-m}^k,$$

wobei P_{l-m}^k bezüglich Q_{r-m}^k zählerminimal ist. Man kann dann nach Satz 2 folgern

$$\lim_{\substack{k\to\infty \\ k\in J_m}} \left\| f - \frac{P_l^k}{Q_r^k} \right\|_p = \left\| f - \frac{T_m^0 \cdot P_{l-m}^0}{T_m^0 \cdot Q_{r-m}^0} \right\|_p$$

und

$$\lim_{\substack{k\to\infty \\ k\in J_m}} \mathfrak{a}^k = \mathfrak{a}^0.$$

Damit ist der Satz 5 bewiesen.

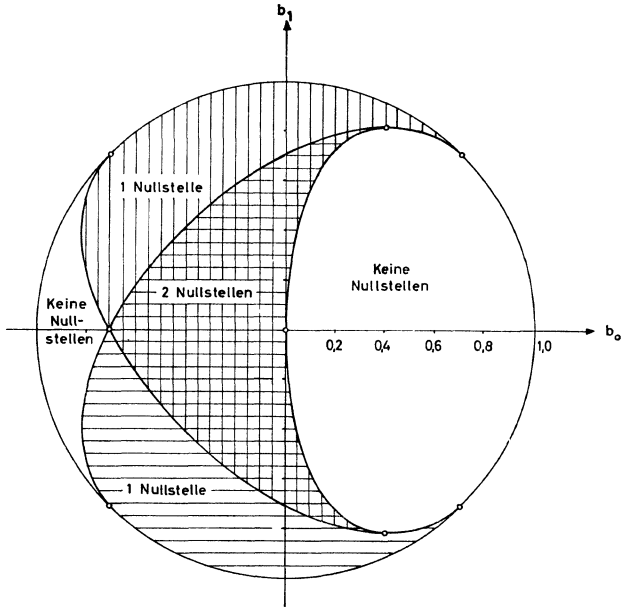

Fig. 1

Beispiel. Als Beispiel wurde auf der Rechenanlage der Universität Münster die Funktion

$$f(x) := \operatorname{sign}(x)$$

im Intervall $[-1, 1]$ durch Funktionen aus der Menge $\mathfrak{R}_{l,r}$ mit $l=r=2$ in der L_2-Norm approximiert.

Wegen der Normierung des Nennerpolynoms ergibt sich

$$b_2 = \sqrt{1 - b_1^2 - b_0^2},$$

so daß zur Bestimmung der besten Approximierenden nur der Bereich

$$\{(b_0, b_1) \mid b_0^2 + b_1^2 \leqq 1\}$$

untersucht werden muß.

In Fig. 1 wurde dieser Bereich weiter unterteilt, je nachdem wie viele Nullstellen das zugehörige Polynom

$$Q_2(x) = b_2 x^2 + b_1 x + b_0$$

im Approximationsintervall besitzt. Nur in den Teilbereichen, in denen $Q_2(x)$ keine Nullstellen hat, können die besten Approximierenden von $f(x)$ liegen.

In Fig. 2 wurden die punktweise berechneten Höhenlinien der Funktion

$$G(b_0, b_1) := \left\| f - \frac{P_2^0}{Q_2} \right\|_2$$

mit jeweils zählerminimalem Polynom P_2^0 dargestellt. An der Figur erkennt man, daß es in der Menge $\mathfrak{R}_{2,2}$ zwei beste Approximierende für die Funktion $f(x)$ gibt.

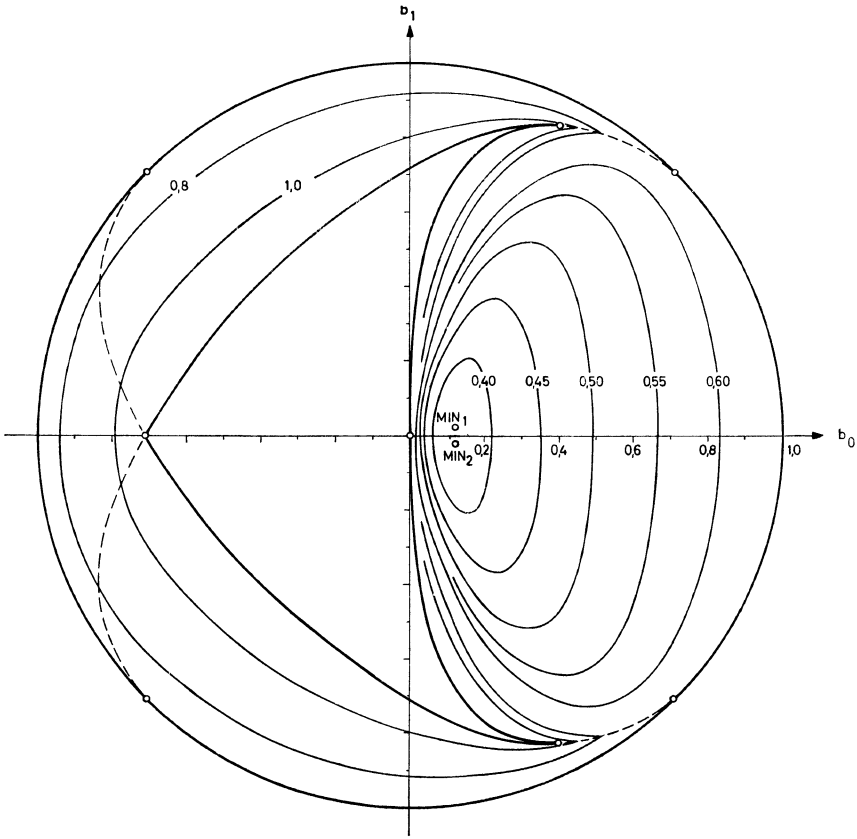

Fig. 2

Literatur

1. Caratheodory, C.: Vorlesungen über reelle Funktionen. New York: Chelsea Publ. Co. 1948.
2. Natanson, I. P.: Theorie der Funktionen einer reellen Veränderlichen. Berlin: Akademie-Verlag 1961.
3. Werner, H.: Vorlesungen über Approximationstheorie. Berlin-Heidelberg-New York: Springer 1966.

Dr. Günther Lamprecht
D-44 Münster
Rechenzentrum der Westf. Wilhelms-Universität
Roxeler Str. 64

Numer. Math. 15, 404–414 (1970)
© by Springer-Verlag 1970

Approximation of Piecewise Continuous Functions by a Modification of Piecewise Hermite Interpolation

JÖRG HERTLING

Received May 20, 1969

1. Introduction

The basic concept of the approximation of a smooth function $r(x)$ by a spline function $\omega(x)$ is the simulation of certain continuity properties in $\omega(x)$. In this paper we want to study the approximation of a function of the simple form

$$r(x) = \frac{\bar{r}(x)}{(x-s_1)^{k_1}(x-s_2)^{k_2}\dots(x-s_P)^{k_P}}$$

where $\bar{r}(x)$ is sufficiently smooth by a modification of piecewise Hermite interpolation. However, it will become clear how the same idea can be applied to spline interpolation of various types and to more complicated singularities.

2. One Dimensional Problem

Let us define two partitions of the unit interval $I_u = [0, 1]$.

$$\bar{\pi}\colon\ 0 < s_1 < s_2 < \dots < s_P < 1,$$

$$\pi\colon\ 0 = x_0 < x_1 < \dots < x_{M+1} = 1$$

and let us assume

$$s_p \neq x_0, x_1, \dots, x_{M+1}, \qquad p = 1, 2, \dots, P$$

and furthermore we choose the partition π such that every interval $[x_i, x_{i+1}]$, $i = 0, 1, \dots, M$ contains at most one point s_p. That means $M \geqq P-1$. Let us denote $\bar{h}(\pi) = \max\limits_{0 \leqq i \leqq M} h_i$ where $h_i = x_{i+1} - x_i$. We also choose the notation

$$\bar{r}_p(x) = \frac{\bar{r}(x)}{(x-s_1)^{k_1}(x-s_2)^{k_2}\dots(x-s_{p-1})^{k_{p-1}}(x-s_{p+1})^{k_{p+1}}\dots(x-s_P)^{k_P}}.$$

Theorem 1. Let $\bar{r}(x) \in \mathbb{C}^t[0, 1]$ with $t \geqq \bar{m} = m_1 + m_2 + m_3$. Let us consider an interval $[x_i, x_{i+1}]$ of π and a point $s_p \in [\underline{x}_p, \bar{x}_p]$, $\underline{x}_p \leqq x_i$, $x_{i+1} \leqq \bar{x}_p$, $s_p \neq x_i$, $s_p \neq x_{i+1}$, and let us define a polynomial $\bar{v}_i(x)$ of degree $\bar{m} - 1$ by

$$D^{l_1}\bar{v}_i(x_i) = D^{l_1}\bar{r}_p(x_i), \qquad 0 \leqq l_1 \leqq m_1 - 1,$$

$$D^{l_2}\bar{v}_i(s_p) = D^{l_2}\bar{r}_p(s_p), \qquad 0 \leqq l_2 \leqq m_2 - 1,$$

$$D^{l_3}\bar{v}_i(x_{i+1}) = D^{l_3}\bar{r}_p(x_{i+1}), \qquad 0 \leqq l_3 \leqq m_3 - 1,$$

with $0 \leqq i \leqq M$. Then

$$|D^k(\bar{v}_i(x) - \bar{r}_p(x))| \leqq \frac{\|D^{\bar{m}}\bar{r}_p(x)\|_{L^\infty[\bar{x}_p, \bar{x}_p]} h_i^{2k}}{(2k)!\,(\bar{m}-3k)!}(x-x_i)^{m_1-k}|x-s_p|^{m_2-k}(x_{i+1}-x)^{m_3-k}$$

for all k with $0 \leqq k \leqq \min(m_1, m_2, m_3)$. Hence

$$\left\| \frac{D^k(\bar{v}_i(x) - \bar{r}_p(x))}{(x-s_p)^{m_2-k}} \right\|_{L^\infty[x_i, x_{i+1}]} \leqq \frac{\|D^{\bar{m}}\bar{r}_p(x)\|_{L^\infty[\bar{x}_p, \bar{x}_p]} (\bar{h}(\pi))^{m_1+m_3}}{2^{2\min(m_1,m_3)-2k}(2k)!\,(\bar{m}-3k)!},$$
$$0 \leqq k \leqq \min(m_1, m_2, m_3)$$

for this interval $[x_i, x_{i+1}]$.

Proof. Let σ be any point $\neq s_p$ of (x_i, x_{i+1}) and consider

$$\psi(x) \equiv D^k(\bar{v}_i(x) - \bar{r}_p(x)) - \{D^k(\bar{v}_i(\sigma) - \bar{r}_p(\sigma))\} \cdot \left(\frac{x-x_i}{\sigma-x_i}\right)^{m_1-k}\left(\frac{x-s_p}{\sigma-s_p}\right)^{m_2-k}\left(\frac{x-x_{i+1}}{\sigma-x_{i+1}}\right)^{m_3-k}$$

for all $x \in [x_i, x_{i+1}]$. Then

$$D^{l_1}\psi(x_i) = 0, \quad 0 \leqq l_1 \leqq m_1 - k - 1,$$
$$D^{l_2}\psi(s_p) = 0, \quad 0 \leqq l_2 \leqq m_2 - k - 1,$$
$$D^{l_3}\psi(x_{i+1}) = 0, \quad 0 \leqq l_3 \leqq m_3 - k - 1,$$

and in addition $\psi(\sigma) = 0$. Hence, repeated applications of Rolle's Theorem shows that a point ξ in (x_i, x_{i+1}) exists such that $D^{\bar{m}-3k}\psi(\xi) = 0$. Thus

$$0 = D^{\bar{m}-3k}\psi(\xi) = D^{\bar{m}-2k}(\bar{v}_i(\xi) - \bar{r}_p(\xi)) - \frac{(\bar{m}-3k)!\,(D^k(\bar{v}_i(\sigma) - \bar{r}_p(\sigma)))}{(\sigma-x_i)^{m_1-k}(\sigma-s_p)^{m_2-k}(\sigma-x_{i+1})^{m_3-k}},$$

and substituting the resulting expression for $D^k(\bar{v}_i(\sigma) - \bar{r}_p(\sigma))$ gives

$$\psi(x) = D^k(\bar{v}_i(x) - \bar{r}_p(x)) - \frac{[D^{\bar{m}-2k}(\bar{v}_i(\xi) - \bar{r}_p(\xi))]}{(\bar{m}-3k)!}(x-x_i)^{m_1-k}(x-s_p)^{m_2-k}(x-x_{i+1})^{m_3-k}.$$

Since $\psi(\sigma) = 0$ and σ is an arbitrary point $\neq s_p$ in the open interval (x_i, x_{i+1}), then using continuity gives us that

$$D^k(\bar{v}_i(x) - \bar{r}_p(x)) = \frac{[D^{\bar{m}-2k}(\bar{v}_i(\xi) - \bar{r}_p(\xi))]}{(\bar{m}-3k)!}(x-x_i)^{m_1-k}(x-s_p)^{m_2-k}(x-x_{i+1})^{m_3-k}$$

for all $x \in [x_i, x_{i+1}]$. It then remains to get an upper bound for $|D^{\bar{m}-2k}(\bar{v}_i(\xi) - \bar{r}_p(\xi))|$ in terms of $\|D^{\bar{m}}\bar{r}_p(x)\|_{L^\infty[\bar{x}_p, \bar{x}_p]}$. If $\Omega(x) \equiv \bar{v}_i(x) - \bar{r}_p(x)$, then

$$D^{l_1}\Omega(x_i) = 0, \quad 0 \leqq l_1 \leqq m_1 - 1,$$
$$D^{l_2}\Omega(s_p) = 0, \quad 0 \leqq l_2 \leqq m_2 - 1,$$
$$D^{l_3}\Omega(x_{i+1}) = 0, \quad 0 \leqq l_3 \leqq m_3 - 1,$$

and Rolle's Theorem again gives us that there exist points η_r in $[x_i, x_{i+1}]$ such that $D^{\bar{m}-1-r}\Omega(\eta_r)=0$, $0 \leqq r \leqq \bar{m}-1$. But as $\bar{v}_i(x)$ is a polynomial of degree $\bar{m}-1$, then

$$D^{\bar{m}-1}\Omega(x) = \int_{\eta_0}^{x} D^{\bar{m}}\Omega(\varrho_1)\, d\varrho_1 = \int_{\eta_0}^{x} D^{\bar{m}}\bar{r}_p(\varrho_1)\, d\varrho_1,$$

and in general

$$D^{\bar{m}-2k}\Omega(\xi) = \int_{\eta_{2k-1}}^{\xi} d\varrho_{2k} \int_{\eta_{2k-2}}^{\varrho_{2k}} d\varrho_{2k-1} \cdots \int_{\eta_0}^{\varrho_2} D^{\bar{m}}\bar{r}_p(\varrho_1)\, d\varrho_1.$$

Thus

$$|D^{\bar{m}-2k}\Omega(\xi)| = |D^{\bar{m}-2k}(\bar{v}_i(\xi)-\bar{r}_p(\xi))| \leqq \|D^{\bar{m}}\bar{r}_p(\xi)\|_{L^\infty[x_p,\bar{x}_p]} \frac{h_i^{2k}}{(2k)!}.$$

Q.E.D.

Let us now consider a simple example. We want to approximate the function

$$r(x) = \frac{\bar{r}(x)}{(x-s_1)^{k_1}(x-s_2)^{k_2}\cdots(x-s_P)^{k_P}}, \qquad \bar{r}(x) \in \mathbb{C}^t[0,1],$$
$$t \geqq \max(k_1, k_2, \ldots, k_P) + 2m$$

by piecewise rational interpolation.

Then in those intervals $[x_i, x_{i+1}]$ which contain no point s_p we can approximate $r(x)$ by a unique interpolating $v_i(x)$ of degree $2m-1$ such that

$$\begin{aligned} D^l v_i(x_i) &= D^l r(x_i), \\ D^l v_i(x_{i+1}) &= D^l r(x_{i+1}), \quad 0 \leqq l \leqq m-1. \end{aligned}$$

According to [2] the error is given by the formula

$$\|v_i(x)-r(x)\|_{L^\infty[x_i,x_{i+1}]} \leqq \frac{\|D^{2m}r(x)\|_{L^\infty[x_i,x_{i+1}]}(\bar{h}(\pi))^{2m}}{2^{2m}(2m)!}.$$

In those intervals $[x_i, x_{i+1}]$ however which contain a point s_p the term $\|v_i(x)-r(x)\|_{L^\infty}$ usually is not bounded. In order to bound this term we choose the unique polynomial $\bar{v}_i(x)$ of degree $2m+k_p-1$ such that

$$\begin{aligned} D^l \bar{v}_i(x_i) &= D^l \bar{r}_p(x_i), & 0 \leqq l \leqq m-1, \\ D^{l_1} \bar{v}_i(s_p) &= D^{l_1} \bar{r}_p(s_p), & 0 \leqq l_1 \leqq k_p-1, \\ D^l \bar{v}_i(x_{i+1}) &= D^l \bar{r}_p(x_{i+1}), & 0 \leqq l \leqq m-1. \end{aligned}$$

Then, according to Theorem 1 with $v_i(x) = \frac{\bar{v}_i(x)}{(x-s_p)^{k_p}}$ we obtain the error estimate

$$\|v_i(x)-r(x)\|_{L^\infty[x_i,x_{i+1}]} \leqq \frac{\|D^{2m+k_p}\bar{r}_p(x)\|_{L^\infty[x_i,x_{i+1}]}(\bar{h}(\pi))^{2m}}{2^{2m}(2m+k_p)!}$$

for the interval which contains s_p. It is obvious that the approximation of $r(x)$ in those intervals which contain such a point s_p is not a polynomial but a rational function.

By fitting together the polynomials $v_i(x)$ of degree $2m-1$ and the rational functions $v_i(x)=\frac{\bar{v}_i(x)}{(x-s_p)^{k_p}}$ we obtain the approximating function $\omega(x)$. Obviously $\omega(x)\in\mathbb{C}^{m-1}$ over the whole interval $[0,1]$ except at the points s_p. In these points s_p, $\omega(x)$ has the same behaviour as $r(x)$ in the sense that the error term is bounded.

Let us now consider the same example, but we wish to obtain error estimates for $\|D^k(\omega(x)-r(x))\|_{L^\infty}$, $0\leqq k\leqq\min(m,m')$, $m'\geqq 1$. Then in those intervals $[x_i, x_{i+1}]$ which contain no point s_p we get the well known result

$$\|D^k(v_i(x)-r(x))\|_{L^\infty[x_i,x_{i+1}]}\leqq\frac{\|D^{2m}r(x)\|_{L^\infty[x_i,x_{i+1}]}(\bar{h}(\pi))^{2m-k}}{2^{2m-2k}k!\,(2m-2k)!}$$

of [2].

In those intervals $[x_i, x_{i+1}]$ which contain a point s_p we choose the unique polynomial $\bar{v}_i(x)$ of degree $2m+m_p-1$, $m_p=m'+k_p$ where k_p is again the order of the pole in s_p. We impose the conditions

$$\begin{aligned}
D^l\bar{v}_i(x_i)&=D^l\bar{r}_p(x_i), && 0\leqq l\leqq m-1,\\
D^{l_1}\bar{v}_i(s_p)&=D^{l_1}\bar{r}_p(s_p), && 0\leqq l_1\leqq m_p-1,\\
D^l\bar{v}_i(x_{i+1})&=D^l\bar{r}_p(x_{i+1}), && 0\leqq l\leqq m-1.
\end{aligned}$$

Due to repeated differentiation of $\frac{\bar{v}_i(x)-\bar{r}_p(x)}{(x-s)^{k_p}}$ the error term is more complicated in this case but can again be bounded by application of Theorem 1.

Of course, it is not necessary that the singularities $s_1, s_2, \ldots, s_P$ always lie within the open intervals (x_i, x_{i+1}). This especially cannot be done if the singularity occurs at 0 or 1. Let us consider for example the two intervals $[x_{i-1}, x_i]$ and $[x_i, x_{i+1}]$ with one singularity at x_i. Then in order to get the same error bound we have to construct polynomials $\bar{v}_{i-1}$ and $\bar{v}_i$ of degree $2m+m_j-1$ for both intervals and therefore two rational functions which might not be so efficient from the point of view of computation. This leads to the following

Corollary 1.1. Let $\bar{r}(x)\in\mathbb{C}^t[0,1]$ with $t\geqq\bar{m}=m_1+m_2+m_3$. Let us consider an interval $[x_i, x_{i+1}]$ of π and let us define a polynomial $\bar{v}_i(x)$ of degree $\bar{m}-1$ by

$$\begin{aligned}
D^{l_1}\bar{v}_i(x_i)&=D^{l_1}\bar{r}_p(x_i), && 0\leqq l_1\leqq m_1-1,\\
D^{l_3}\bar{v}_i(x_{i+1})&=D^{l_3}\bar{r}_p(x_{i+1}), && 0\leqq l_3\leqq m_2+m_3-1,
\end{aligned}$$

with $0\leqq i\leqq N$. Then

$$|D^k(\bar{v}_i(x)-\bar{r}_p(x))|\leqq\frac{\|D^{\bar{m}}\bar{r}_p(x)\|_{L^\infty[\underline{x}_p,\bar{x}_p]}h_i^{2k}}{(2k)!\,(\bar{m}-3k)!}(x-x_i)^{m_1-k}(x_{i+1}-x)^{m_2+m_3-2k}$$

for all k with $0\leqq k\leqq\min(m_1, m_2+m_3)$, $\underline{x}_p\leqq x_i$, $x_{i+1}\leqq\bar{x}_p$. Hence

$$\left\|\frac{D^k(\bar{v}_i(x)-\bar{r}_p(x))}{(x-s_p)^{m_2-k}}\right\|_{L^\infty[x_i,x_{i+1}]}\leqq\frac{\|D^{\bar{m}}\bar{r}_p(x)\|_{L^\infty[\underline{x}_p,\bar{x}_p]}(\bar{h}(\pi))^{m_1+m_2}}{2^{2\min(m_1,m_2)-2k}(2k)!\,(\bar{m}-3k)!},$$

$$0\leqq k\leqq\min(m_1, m_2, m_3)$$

for this interval $[x_i, x_{i+1}]$.

With the preceding considerations a generalization to intervals which contain more than one singularity would be easy. This however seems not reasonable from the computational point of view.

Clearly Theorem 1 also yields the following

Corollary 1.2. $\int_a^b |r(x)|\,dx < \infty$ iff $\int_a^b |\omega(x)|\,dx < \infty$.

If one wants to use this kind of interpolation in order to calculate $\int_a^b r(x)\,dx$ then in every interval $[x_i, x_{i+1}]$ one has to carry out $\int_{x_i}^{x_{i+1}} \omega(x)\,dx$ analytically. This yields no theoretical difficulty on the computer because in those intervals, which contain no singularity, the integrand is a simple polynomial. In those intervals, which contain a singularity the integrand has the $\frac{a_0 x^n + a_1 x^{n-1} + \cdots + a_n}{(x - s_p)^{k_p}}$ with $n \geqq k_p + 1$. The corresponding integrals are of the form:

$$\int \frac{x^\mu}{(c\,x+d)^{\lambda+1}}\,dx = -\sum_{\nu=0}^{\mu} \frac{\mu!\,(\lambda-\nu-1)!}{(\mu-\nu)!\,\lambda!\,c^{\nu+1}} \frac{x^{\mu-\nu}}{(c\,x+d)^{\lambda-\nu}} + C_1$$

$$= \frac{-1}{c^{\mu+1}} \sum_{\nu=0}^{\mu} \binom{\mu}{\nu} \frac{(-d)^{\mu-\nu}}{(\lambda-\nu)\,(c\,x+d)^{\lambda-\nu}} + C_2,$$

for $\lambda = 0, 1, \ldots, \mu$ in the sum the term with $\nu = \lambda$ is to be replaced by the following: $-\binom{\mu}{\nu}(-d)^{\mu-\lambda}\log(c\,x+d)$.

If some k_p' in $r(x)$ is not an integer, then in the interval $[x_i, x_{i+1}]$ which contains this singularity in s_p we have to choose a polynomial $\bar{v}_i(x)$ of degree $2m + m_p - 1$, $m_p = m' + [k_p'] + 1$, such that

$$D^l \bar{v}_i(x_i) = D^l \bar{r}_p(x_i), \qquad 0 \leqq l \leqq m-1,$$

$$D^{l_1} \bar{v}_i(s_p) = D^{l_1} \bar{r}_p(s_p), \qquad 0 \leqq l_1 \leqq m_p - 1,$$

$$D^l \bar{v}_i(x_{i+1}) = D^l \bar{r}_p(x_{i+1}), \qquad 0 \leqq l \leqq m-1.$$

Of course, we can always choose a higher order contact in s_p in some interval $[x_i, x_{i+1}]$ which contains this point s_p. That means that we choose a polynomial $\bar{v}_i(x)$ of degree $2m + m_p' - 1$, $m_p' > m_p = m' + k_p$ such that

$$D^l \bar{v}_i(x_i) = D^l \bar{r}_p(x_i), \qquad 0 \leqq l \leqq m-1,$$

$$D^{l_1} \bar{v}_i(s_p) = D^{l_1} \bar{r}_p(s_p), \qquad 0 \leqq l_1 \leqq m_p' - 1,$$

$$D^l \bar{v}_i(x_{i+1}) = D^l \bar{r}_p(x_{i+1}), \qquad 0 \leqq l \leqq m-1.$$

This increases the order of accuracy of the approximating rational function in this special interval and furthermore ensures that $\max_{x \in [x_i, x_{i+1}]} |v_i(x) - r(x)|$ occurs

only in the open intervals (x_i, s_p) respectively (s_p, x_{i+1}). But by choosing the correct signs in $v_i(x) = \pm \frac{\bar{v}_i(x)}{|x - s_p|^{k_p}}$ we always have to ensure that

$$\lim_{x \to s_p-} v_i(x) = \lim_{x \to s_p-} r(x),$$

$$\lim_{x \to s_p+} v_i(x) = \lim_{x \to s_p+} r(x).$$

We have obtained error estimates in the $\mathbb{L}^\infty$ norm and we are looking now for error estimates in the $\mathbb{L}^g$ norm.

Definition. Let $\mathbb{W}^{e,g}(I)$, where e is a positive integer and g is an extended real number $1 \leqq g \leqq +\infty$ denote the set of all real-valued functions $\bar{r}(x)$ on I such that $D^{e-1}\bar{r}(x)$ is absolutely continuous on I and $D^e\bar{r}(x) \in \mathbb{L}^g(I)$.

For simplicity let us consider at first the unit interval $I_u = [0, 1]$. We look at the approximation of some function $\bar{r}_p(x) \in \mathbb{W}^{e,g}(I_u)$ by a unique polynomial $\bar{v}_{\bar{m}}(x)$ of degree $\bar{m} - 1 = m_1 + m_2 + m_3 - 1$ defined by the conditions

$$D^{l_1}\bar{v}_{\bar{m}}(0) = D^{l_1}\bar{r}_p(0), \qquad 0 \leqq l_1 \leqq m_1 - 1,$$

$$D^{l_2}\bar{v}_{\bar{m}}(s_p) = D^{l_2}\bar{r}_p(s_p), \qquad 0 \leqq l_2 \leqq m_2 - 1,$$

$$D^{l_3}\bar{v}_{\bar{m}}(1) = D^{l_3}\bar{r}_p(1), \qquad 0 \leqq l_3 \leqq m_3 - 1,$$

with $0 < s_p < 1$. For fixed x_0, $0 \leqq x_0 \leqq 1$ the error functional

$$F(f) \equiv D^k\left(\bar{r}_p(x_0) - \bar{v}_{\bar{m}}(x_0)\right), \qquad 0 \leqq k \leqq e - 1$$

depends on k, e, g, x_0, m_1, m_2, m_3. We want to represent this linear functional by a sum of Riemann-Stieltjes integrals of the form

$$F(f) = \sum_{l=0}^{e-1} \int_0^1 D^l\bar{r}_p(x)\, d\mu_l(x; x_0; s_p), \qquad \bar{r}_p \in \mathbb{W}^{e,g}(I_u).$$

We define polynomials $S_{0,l}(x, \bar{m})$, $S_{s_p,l}(x, \bar{m})$, and $S_{1,l}(x, \bar{m})$ of degree $\bar{m} - 1$ by

$$D^{l_1}S_{0,l_1}(0, \bar{m}) = \delta_{l_1,l_1}, \quad D^{l_2}S_{0,l_1}(s_p, \bar{m}) = 0, \quad D^{l_3}S_{0,l_1}(1, \bar{m}) = 0, \quad 0 \leqq l_1, \bar{l}_1 \leqq m_1 - 1,$$

$$D^{l_1}S_{s_p,l_2}(0, \bar{m}) = 0, \quad D^{l_2}S_{s_p,l_2}(s_p, \bar{m}) = \delta_{l_2,l_2}, \quad D^{l_3}S_{s_p,l_2}(1, \bar{m}) = 0, \quad 0 \leqq l_2, \bar{l}_2 \leqq m_2 - 1,$$

$$D^{l_1}S_{1,l_3}(0, \bar{m}) = 0, \quad D^{l_2}S_{1,l_3}(s_p, \bar{m}) = 0, \quad D^{l_3}S_{1,l_3}(1, \bar{m}) = \delta_{l_3,l_3}, \quad 0 \leqq l_3, \bar{l}_3 \leqq m_3 - 1,$$

where $\delta_{l,l}$ is the Kronecker delta function. Then

$$\bar{v}_{\bar{m}}(x) = \sum_{l=0}^{m_1-1}\left(D^l\bar{r}_p(0)\right)S_{0,l}(x, \bar{m}) + \sum_{l=0}^{m_2-1}\left(D^l r_p(s_p)\right)S_{s_p,l}(x, \bar{m}) + \sum_{l=0}^{m_3-1}\left(D^l r_p(1)\right)S_{1,l}(x, \bar{m}).$$

Now, corresponding to the magnitude of m_1, m_2, and m_3 we have to distinguish several cases. Let us take for example $m_1 < m_2 < m_3$. Then, for $0 < x_0 < s_p$ we can define

$$\mu_l(x; x_0; s_p) = \begin{cases} 0 & x = 0, \\ -D^l S_{0,l}(x_0, \overline{m}), & 0 < x \leqq x_0, \\ \delta_{l,l} - D^l S_{0,l}(x_0, \overline{m}), & x_0 < x \leqq s_p, \\ \delta_{l,l} - D^l S_{0,l}(x_0, \overline{m}) - D^l S_{s_p,l}(x_0, \overline{m}), & s_p < x < 1, \\ \delta_{l,l} - D^l S_{0,l}(x_0, \overline{m}) - D^l S_{s_p,l}(x_0, \overline{m}) & \\ \quad - D^l S_{1,l}(x_0, \overline{m}), & x = 1, \\ & 0 \leqq \bar{l} \leqq m_1 - 1, \end{cases}$$

$$\mu_l(x; x_0; s_p) = \begin{cases} 0 & 0 \leqq x \leqq x_0, \\ \delta_{l\bar{l}}, & x_0 < x \leqq s_p, \\ \delta_{l\bar{l}} - D^l S_{s_p,l}(x_0, \overline{m}), & s_p < x < 1, \\ \delta_{l\bar{l}} - D^l S_{s_p,l}(x_0, \overline{m}) - D^l S_{1,l}(x_0, \overline{m}), & x = 1, \\ & m_1 \leqq \bar{l} \leqq m_2 - 1, \end{cases}$$

$$\mu_l(x; x_0; s_p) = \begin{cases} 0 & 0 \leqq x \leqq x_0, \\ \delta_{l\bar{l}}, & x_0 < x < 1, \\ \delta_{l\bar{l}} - D^l S_{1,l}(x_0, \overline{m}), & x = 1, \\ & m_2 \leqq \bar{l} \leqq m_3 - 1, \end{cases}$$

$$\mu_l(x; x_0; s_p) = \begin{cases} 0 & 0 \leqq x \leqq x_0, \\ \delta_{l\bar{l}}, & x_0 < x \leqq 1, \\ & m_3 \leqq \bar{l} \leqq e - 1. \end{cases}$$

Let us still look at the case $m_1 = m_2 = m_3$, $s_p < x_0 < 1$:

$$\mu(x; x_0; s_p) = \begin{cases} 0 & x = 0, \\ -D^l S_{0,l}(x_0, \overline{m}), & 0 < x \leqq s_p, \\ -D^l S_{0,l}(x_0, \overline{m}) - D^l S_{s_p,l}(x_0, \overline{m}), & s_p < x \leqq x_0, \\ \delta_{l\bar{l}} - D^l S_{0,l}(x_0, \overline{m}) - D^l S_{s_p,l}(x_0, \overline{m}), & x_0 < x < 1, \\ \delta_{l\bar{l}} - D^l S_{0,l}(x_0, \overline{m}) - D^l S_{s_p,l}(x_0, \overline{m}) & \\ \quad - D^l S_{1,l}(x_0, \overline{m}), & x = 1, \\ & 0 \leqq \bar{l} \leqq m_1 - 1, \end{cases}$$

$$\mu_l(x; x_0; s_p) = \begin{cases} 0 & 0 \leqq x \leqq x_0, \\ \delta_{l\bar{l}}, & x_0 < x \leqq 1, \\ & m_1 \leqq \bar{l} \leqq e - 1. \end{cases}$$

Again, we can formulate an expression for the error functional by the application of the Peano Kernel theorem:

Theorem 2. For any $\bar{r}_p(x) \in W^{e,g}(I)$, $e \geqq \max(m_1, m_2, m_3)$, and any x_0 with $0 \leqq x_0 \leqq 1$, $I_u \subset I$, $s_p \in I$, $s_p \neq 0, 1$,

$$D^k\left(\bar{r}_p(x_0) - \bar{v}_{\bar{m}}(x_0)\right) = \int_0^1 D^n f(z)\, K_{k,n,m_1,m_2,m_3}(z;\, x_0;\, s_p)\, dz,$$

$$n = \min(e, \bar{m}), \qquad 0 \leqq k \leqq n-1,$$

$$K_{k,n,m_1,m_2,m_3}(z;\, x_0;\, s_p) = F_x\left\{\frac{(x-z)_+^{n-1}}{(n-1)!}\right\} = \sum_{l=0}^{n-1} \int_z^1 d\mu_l(x;\, x_0;\, s_p)\, \frac{(x-z)^{n-l-1}}{(n-l-1)!}$$

where

$$(x-z)_+^{n-1} = \begin{cases} (x-z)^{n-1} & \text{for } z \leqq x, \\ 0 & \text{for } x < z. \end{cases}$$

For example, the kernel $K_{0,3,1,1,1}(z;\, x_0;\, s_p)$, $\frac{1}{2} = s_p \leqq x_0 \leqq 1$ is given by Green's function for the boundary value problem $D^3 u(x) = g(x)$, $0 < x < 1$, $u(0) = u(s_p) = u(1) = 0$:

$$2K_{0,3,1,1,1}(z;\, x_0;\, s_p) = \begin{cases} (2x_0^2 - 3x_0 + 1)\, z^2, & 0 \leqq z \leqq s_p = \frac{1}{2}, \\ (-2x_0^2 + x_0 + 1)\, z^2 + (4x_0^2 - 4x_0)\, z - x_0^2 + x_0, & \frac{1}{2} = s_p \leqq z \leqq x_0, \\ (x_0 - 2x_0^2)\,(z^2 - 2z + 1), & x_0 \leqq z \leqq 1. \end{cases}$$

It turns out that the kernel $K_{k,n,m_1,m_2,m_3}(z;\, x_0;\, s_p)$ is uniformly bounded for $(z;\, x_0)$ in $I_u \times I_u$. Let us denote

$$V_{k,k_p,n,m_1,m_2,m_3}(z;\, x_0;\, s_p) = \frac{K_{k,n,m_1,m_2,m_3}(z;\, x_0;\, s_p)}{(x_0 - s_p)^{k_p}}.$$

Then by means of the construction of the zero of $K_{k,n,m_1,m_2,m_3}(z;\, x_0;\, s_p)$ at $x_0 = s_p$, $V_{k,k_p,n,m_1,m_2,m_3}(z;\, x_0;\, s_p)$ is also uniformly bounded for $(z;\, x_0)$ in $I_u \times I_u$. With $1/g + 1/g' = 1$ the function

$$G_{k,k_p,n,g,m_1,m_2,m_3}(x_0;\, s_p) \equiv \left\{\int_0^1 |V_{k,k_p,n,g,m_1,m_2,m_3}(z;\, x_0;\, s_p)|^{g'}\, dz\right\}^{1/g'}$$

is an element of $\mathbb{L}^h[I_u]$ for any h with $1 \leqq h \leqq +\infty$, and we define the constants

$$C_{k,k_p,n,g,h,m_1,m_2,m_3} \equiv \left(\int_0^1 |G_{k,k_p,n,g,m_1,m_2,m_3}(x_0;\, s_p)|^h\, dx_0\right)^{1/h}.$$

By application of Hölder's inequality we obtain

$$\left|\frac{D^k(\bar{r}_p(x_0) - \bar{v}_{\bar{m}}(x_0))}{(x_0 - s_p)^{k_p}}\right| \leqq \|D^n \bar{r}_p(x_0)\|_{\mathbb{L}^g[I_u]} G_{k,k_p,n,g,m_1,m_2,m_3}(x_0;\, s_p),$$

$$0 \leqq k \leqq n-1, \qquad k_p \leqq m_2 - k.$$

This yields the following

Corollary 2.1. For any $\bar{r}_p \in W^{e,g}(I)$, $e \geqq \max(m_1, m_2, m_3)$, $n = \min(e, m)$, $I_u \subset I$, $s_p \in I$, $s_p \neq 0, 1$, we can find a finite constant $C_{k,k_p,n,g,h,m_1,m_2,m_3}$ such that

$$\left\| \frac{D^k(\bar{r}_p - \bar{v}_{\bar{m}})}{(x-s_p)^{k_p}} \right\|_{L^h[I_u]} \leqq C_{k,k_p,n,g,h,m_1,m_2,m_3} \|D^n \bar{r}_p\|_{L^g[I_u]},$$

$$0 \leqq k \leqq n-1, \quad 1 \leqq h \leqq +\infty, \quad k_p \leqq m_2 - k.$$

For an arbitrary interval this corollary would be

Corollary 2.2. For any $\bar{r}_p \in W^{e,g}[\underline{x}_p, \bar{x}_p]$, $e \geqq \max(m_1, m_2, m_3)$, $n = \min(e, \bar{m})$ we can find a finite constant $C_{k,k_p,n,g,h,m_1,m_2,m_3}$ such that

$$\left\| \frac{D^k(\bar{r}_p - \bar{v}_i)}{(x-s_p)^{k_p}} \right\|_{L^h[x_i, \bar{x}_{i+1}]} \leqq h_i^{n-k-(1/g)+(1/h)} C_{k,k_p,n,g,h,m_1,m_2,m_3} \|D^n \bar{r}_p\|_{L^g[\underline{x}_p, \bar{x}_p]},$$

$$s_p \in [\underline{x}_p, \bar{x}_p], \quad \underline{x}_p \leqq x_i, \quad x_{i+1} \leqq \bar{x}_p, \quad s_p \neq x_i, \quad s_p \neq x_{i+1}, \quad 0 \leqq k \leqq n-1,$$

$$1 \leqq h \leqq +\infty, \quad k_p \leqq m_2 - k.$$

With

$$r(x) = \frac{\bar{r}_p(x)}{(x-s_p)^{k_p}}, \qquad v_i(x) = \frac{\bar{v}_i(x)}{(x-s_p)^{k_p}}$$

we are now able to obtain an error estimate for $\|D^k(r(x) - v_i(x))\|_{L^h[x_i, x_{i+1}]}$.

With these error estimates most of the rest of the corresponding analysis in the papers [1] and [2] can be carried out without difficulty.

If the singularity is not a pole but, for example, a singularity of type $e^{-\frac{1}{\bar{r}(x)}}$ it is obvious that we cannot choose a simple polynomial for $\bar{v}_i(x)$. In this case we can construct an L-Hermite function which has the desired properties. So L-Hermites which have been introduced as a generalization of polynomials but without any particular need now become important in order to handle various kinds of singularities.

3. Uniform Approximation

Although, with the preceding considerations one can construct an interpolating function which is arbitrarily close to $r(x)$, the constants of the previous error estimates are not uniformly bounded as $\bar{h}(\pi)$ tends to zero. Namely, in this case, the term $\|D^{2m} r(x)\|_{L^\infty}$ is not bounded for those intervals which do not contain a singularity.

We will now describe a procedure which avoids this difficulty. For every singularity in s_p, $p = 1, 2, \ldots, P$, we choose an interval $[\underline{x}_p, \bar{x}_p]$, such that $s_p \in (\underline{x}_p, \bar{x}_p)$ and such that all the intervals $[\underline{x}_p, \bar{x}_p]$ are disjoint and during the rest of the procedure we keep these intervals fixed. Then we consider a sequence of partitions π such that $\bar{h}(\pi)$ tends to zero and such that every $\underline{x}_p$ resp. $\bar{x}_p$ coincides with one x_i in every partition π of this sequence of partitions.

Then in those intervals $[x_i, x_{i+1}]$ which are not contained in any of the intervals $[\underline{x}_p, \bar{x}_p]$ we choose the usual interpolating polynomial of degree $2m-1$ for those intervals which contain no singularity, and we use the error estimates in the L^g resp. L^∞ norm of the papers [1] resp. [2].

In all those intervals $[x_i, x_{i+1}]$ which are contained in $[\underline{x}_p, \bar{x}_p]$ choose rational functions with polynomials of degree $2m + k_p - 1$ in the denominator, such that

these polynomials have a higher order contact with $\bar{r}_p(x)$ at $x = s_p$. Notice that the interval $[\underline{x}_p, \bar{x}_p]$ may contain many intervals $[x_i, x_{i+1}]$, but only one of these intervals $[x_i, x_{i+1}]$ contains the singularity. Nevertheless we can construct in all these intervals rational functions and apply Theorems 1 and 2 with their Corollaries. This process now yields uniform boundedness of all constants as $\bar{h}(\pi)$ tends to zero.

4. Piecewise Bivariate Interpolation

Let us consider a rectangular domain $\bar{R} = [a, b] \times [c, d]$ and let us assume at first that there is only one pole at $s_{pq} = (s_p, s_q)$ in this domain. If s_{pq} does not lie on the boundary we know that we can choose some mesh

$$\pi_x\colon\ a = x_0 < x_1 < \cdots < x_{M+1} = b,$$
$$\pi_y\colon\ c = y_0 < y_1 < \cdots < y_{N+1} = d$$

such that

$$s_p \neq x_0, x_1, \ldots, x_{M+1},$$
$$s_q \neq y_0, y_1, \ldots, y_{N+1}.$$

Let us suppose for example that we have a pole of order k_{pq} in s_{pq}, $s_p \in [x_i, x_{i+1}]$, $s_q \in [y_j, y_{j+1}]$ and let us look for a piecewise bivariate rational interpolation in $\bar{R}$ with error estimates for $\|D^{(k,l)}(w - r)\|_{L_\infty}$, $0 \leqq k, l \leqq \min(m, m')$, $m' \geqq 0$. Again we assume

$$r(x, y) = \frac{\bar{r}(x, y)}{(\sqrt{(x - s_p)^2 + (y - s_q)^2})^{k_{pq}}}$$

with $\bar{r}(x, y)$ sufficiently smooth. We want to construct a piecewise continuous surface $w(x, y)$ for all $(x, y) \in \bar{R}$, $(x, y) \neq s_{pq}$ which satisfies in all mesh points the condition

$$\frac{\partial^{l_1+l_2} w(x, y)}{\partial x^{l_1} \partial y^{l_2}} = \frac{\partial^{l_1+l_2} r(x, y)}{\partial x^{l_1} \partial y^{l_2}}, \qquad 0 \leqq l_1, l_2 \leqq m - 1.$$

In all intervals $[x_{i'}, x_{i'+1}] \times [y_{j'}, y_{j'+1}]$ with $i' \neq i$ and $j' \neq j$ we choose the usual piecewise bivariate Hermite polynomial which is uniquely defined by the above conditions.

For the interval $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ we choose a piecewise bivariate rational interpolation which satisfies at least the following conditions:

$$\frac{\partial^{l_1+l_2} \bar{w}(x, y)}{\partial x^{l_1} \partial y^{l_2}} = \frac{\partial^{l_1+l_2} \bar{r}(x, y)}{\partial x^{l_1} \partial y^{l_2}}, \qquad 0 \leqq l_1, l_2 \leqq m - 1$$

at the mesh points and

$$\frac{\partial^{l_1+l_2} \bar{w}(s_p, s_q)}{\partial x^{l_1} \partial x^{l_2}} = \frac{\partial^{l_1+l_2} \bar{r}(s_p, s_q)}{\partial x^{l_1} \partial y^{l_2}}, \qquad 0 \leqq l_1, l_2 \leqq m_{pq} - 1$$

at the singularity. Additional conditions may be defined in order to get uniqueness of $\bar{w}(x, y)$. Analogously

$$w(x, y) = \frac{\bar{w}(x, y)}{(\sqrt{(x - s_p)^2 + (y - s_q)^2})^{k_{pq}}}.$$

The problem of uniform approximation and the case of more than one singularity is treated in a similar way as in the one-dimensional case. At first we define a set of rectangular intervals $\bar{I}$ such that each of these intervals contains only one singularity and such that these intervals are disjoint from each other. Then we consider a sequence of partitions such that the diameter of each rectangular interval of this partition tends to zero. If an interval of this partition is contained in some element of $\bar{I}$ we construct a higher order contact for $\overline{w}(x, y)$ at the corresponding singularity, whether or not this element of the partition contains the singularity. Error estimates and the problem of uniqueness will be discussed in another paper.

I would like to thank Professor Richard S. Varga for valuable discussions.

References

1. Birkhoff, G., Schultz, M. H., Varga, R. S.: Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. Numer. Math. **11**, 232–256 (1968).
2. Ciarlet, P. G., Schultz, M. H., Varga, R. S.: Numerical methods of high-order accuracy for nonlinear boundary value problems. I. One dimensional problem. Numer. Math. **9**, 394–430 (1967).
3. Gröbner, W., Hofreiter, N.: Integraltafel. Wien-New York: Springer 1965.
4. Sard, A.: Linear approximation. Mathematical Surveys 9. American Mathematical Society, Providence, Rhode Island 1963.

Prof. Dr. Jörg Hertling
Department of Mathematics
Oregon State University
Corvallis, Oregon

Numer. Math. 15, 415–432 (1970)
© by Springer-Verlag 1970

Ein Verfahren zum Minimieren einer Funktion bei eingeschränktem Variationsbereich der Parameter*

G. ZETTL

Eingegangen am 12. Dezember 1968

Summary. An algorithm is proposed for minimization of a function Φ of many variables, which is capable of handling constraints in the form of linear inequalities among the variables and which has second order convergence properties. It is assumed that Φ and its partial derivatives can be computed. The method is related to that suggested by Fletcher and Powell. However, a different algorithm is used to approximate the inverse of the Hessian matrix which enables the insertion of constraints. Examples show that convergence is as good as in the method mentioned above.

Zusammenfassung. Es wird eine neue, quadratisch konvergierende Methode vorgeschlagen zur Minimierung einer Funktion Φ von mehreren Veränderlichen, bei der den Parametern Nebenbedingungen in Form linearer Ungleichungen auferlegt werden können. Neben dem Funktionswert Φ werden die partiellen Ableitungen von Φ nach den Parametern benötigt. Die Ableitungen zweiter Ordnung werden indirekt vom Verfahren berücksichtigt. Die Methode ist mit der von Fletcher und Powell angegebenen verwandt. Zum Aufbau der Inversen der Matrix der zweiten Ableitungen wird jedoch ein anderer, neuer Algorithmus benutzt, der das Einführen von Nebenbedingungen ermöglicht. Beispiele zeigen, daß die neue Methode in ihrem Konvergenzverhalten der erwähnten ebenbürtig ist.

1. Einführung

Optimierungsprobleme der Technik lassen sich oft als eine Minimierungsaufgabe formulieren. Die Optimallösung deckt sich dabei mit dem Minimum einer Fehlerquadratsumme. Frei laufende Parameter können zu unerwünschten, weil nicht realisierbaren, Lösungen führen. Ziel dieser Arbeit ist es, eine Methode vorzustellen, die folgende Aufgabe löst:

Man bestimme das Minimum der Funktion

$$\Phi = \Phi(x_1, x_2, \ldots, x_n), \tag{1}$$

wobei die Parameter $x_1, x_2, \ldots, x_n$ den linearen Ungleichungen

$$b_{k0} + \sum_{i=1}^{n} b_{ki} x_i \geqq 0, \qquad k = 1, 2, \ldots, l; \qquad l \gtreqless n \tag{2}$$

genügen sollen.

* Mitteilung aus dem Zentrallaboratorium für Nachrichtentechnik der Siemens AG, München.

Ein Lösungsweg, mit dem sich diese Aufgabe bewältigen läßt, wurde von Fiacco und McCormick in [3] und [4] vorgeschlagen und von Lasdon und Waren in [6] angewandt. Dieser Lösungsweg gestattet die Einführung allgemeiner Nebenbedingungen, beschränkt sich also nicht auf Nebenbedingungen bezüglich der Parameter in der Form (2), erfordert dafür aber einen vergleichsweise großen Rechenaufwand.

Das hier vorzustellende Verfahren ist mit dem von Fletcher und Powell [5] angegebenen verwandt. Es arbeitet auch mit einer Näherung zweiter Ordnung für die Funktion und benötigt dazu Funktions- und Gradientenwerte.

Die Glieder zweiter Ordnung werden durch einen anderen Algorithmus als in [5] berücksichtigt. Dieser neue Algorithmus stellt keine Orthogonalitätsforderungen an den Iterationsablauf und gestattet deshalb das Einführen von Nebenbedingungen.

Konvergenzvergleiche mit [5] lassen sich nur bei frei laufenden Parametern durchführen. In den gerechneten Testbeispielen konvergierte die hier zu beschreibende Methode etwa ebenso gut wie die von Fletcher und Powell.

2. Verwendete Symbole

In dieser Arbeit wird für Vektoren und Matrizen die Diracsche Schreibweise [2] verwendet. Elemente dieser Schreibweise sind der senkrechte Strich und die nach links bzw. rechts geöffnete Ecke. Dabei bedeuten:

$\langle x \rangle$ Skalar

$|x\rangle$ Spaltenvektor

$\langle x|$ Zeilenvektor

$|x|$ Matrix.

Die Art einer zusammengesetzten Größe läßt sich einfach aus den links und rechts von ihr stehenden Strichen oder Ecken entnehmen, z.B.

$\langle x|y\rangle$ Skalarprodukt

$|x\rangle\langle y|$ dyadisches Produkt (Matrix).

Die im folgenden verwendeten Größen seien hier vorweg zusammengestellt:

α Parameter der eindimensionalen Minimumsuche

b_{ij} Koeffizient der Nebenbedingungen

$|B|$ Rechteckige Matrix mit Elementen b_{ij}

Φ Zu minimierende Funktion

F Nach den Gliedern zweiter Ordnung abgebrochene Taylorentwicklung von Φ

$|g(x)\rangle$ Gradientenvektor an der Stelle $|x\rangle$

$|\Delta g\rangle$ Differenz zweier Gradientenvektoren

$|H_i|$ Iterationsmatrix, Näherung für $|K^{-1}|$

$\|K\|$	Matrix der zweiten Ableitungen
$\|\Lambda\rangle$	Vektor der Lagrange-Parameter
m	Anzahl der erfaßten Randebenen
n	Anzahl der Parameter
Q	Unter Nebenbedingungen zu minimierende Funktion
$\|r\rangle$	Suchstrahl
$\|s\rangle$	Schrittvorschlag ohne Nebenbedingungen
$\|u\rangle$	Schrittvorschlag unter Nebenbedingungen
$\|x\rangle$	Ortsvektor
$\|\Delta x\rangle$	Differenz zweier Ortsvektoren.

3. Behandlung der Nebenbedingungen

Das Verfahren beginnt mit einer im n-dimensionalen Variationsbereich liegenden Näherung. Variationsraum ist ein von ebenen Flächen berandeter Körper, in dem sich der Parametervektor frei bewegen darf. Zunächst zielen die Parameteränderungen ohne einschränkende Nebenbedingungen auf eine möglichst starke Reduktion der Funktion Φ. Dabei kann es geschehen, daß Berandungsebenen durchstoßen werden. Das Minimum von Φ liegt dann offenbar außerhalb des Körpers. Im Körper nimmt Φ seinen kleinsten Wert auf dem Rand an, und es gilt, dieses Minimum auf dem Rand zu bestimmen. Die Reduktion erfolgt jetzt unter den durch die Berandung gegebenen Nebenbedingungen. Dabei ist normalerweise nur ein Teil der Begrenzungsebenen im Spiel. Die restlichen Randebenen sind in den vorausgehenden Iterationsschritten nicht angesprochen worden und bleiben deshalb unbeachtet. Im weiteren Verlauf der Iteration kann sich herausstellen, daß das wirkliche Minimum der Funktion doch im Innern des Körpers liegt, nur der Weg zu diesem Minimum führte über den Rand. Dann sind die betreffenden Randnebenbedingungen wieder freizugeben.

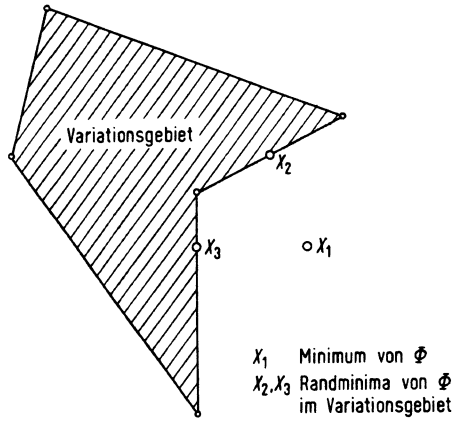

Fig. 1. Nichtkonvexes Variationsgebiet mit zwei Randminima

Daß die Nebenbedingungen (2) keinen Widerspruch enthalten dürfen, versteht sich von selbst. Wenn sie einen nicht konvexen Körper beschreiben (konvex: je zwei Punkte im Innern des Körpers müssen sich durch eine ganz im Innern des Körpers liegende Strecke verbinden lassen), dann ist zu beachten, daß auch bei einer im Gesamtraum konvexen Funktion Φ das Minimum nicht mehr eindeutig zu sein braucht. In Fig. 1 habe Φ ein Minimum außerhalb des Variationsgebietes. Im Variationsbereich existieren zwei verschiedene Minima.

4. Die Methode

Kern des Verfahrens ist eine nach den Gliedern zweiter Ordnung abgebrochene Taylorreihe, die die zu minimierende Funktion annähert:

$$F = f_0 + \langle g(x_0) | x - x_0 \rangle + \tfrac{1}{2} \langle x - x_0 | K | x - x_0 \rangle. \tag{3}$$

Diese Näherungsfunktion wird iterativ aufgebaut. Die Taylorentwicklung einer Funktion hängt ab von dem Punkt, um den entwickelt wird. Bei der Iteration bewegt sich der Parametervektor, ausgehend von der Näherung, durch den Raum zum Minimum. Die Taylorentwicklung der Funktion wandelt sich dabei von der Entwicklung um die Näherung zur Entwicklung um die Lösung. Bestimmendes Element der Näherungsfunktion ist die Matrix $|K|$ oder deren Inverse $|K^{-1}|$. Für diese Matrix wird eine Näherung ermittelt aus einander zugeordneten Änderungen des Parameter- und Gradientenvektors. Jeder Iterationsschritt liefert solch ein Paar von Vektoren. Für eine quadratische Funktion sind zur Bestimmung der Matrix n linear unabhängige Vektorpaare erforderlich. Im allgemeinen bestimmen die letzten n Vektorpaare die quadratische Näherung. In Gebieten zäher Konvergenz und in der Umgebung des Minimums wird die Matrix automatisch genauer aufgebaut als in anderen Gebieten. Dies geschieht dadurch, daß der Parametervektor mehrere Schritte lang in diesem Gebiet verweilt und mehrere Vektorpaare aus diesem Gebiet zum Aufbau der Matrix benutzt werden.

Nebenbedingungen beeinträchtigen die Konvergenz des Verfahrens nicht. Der Verbesserungsvektor unter Nebenbedingungen ergibt sich aus dem bei frei laufenden Parametern durch ein Korrekturglied.

Der Übergang von frei laufenden Parametern zur Iteration unter Nebenbedingungen geschieht ohne Informationsverlust, denn die erreichte Näherungsfunktion wird beibehalten und weiterbenutzt.

Die Näherungsmatrix strebt gegen die Inverse der Matrix der zweiten Ableitungen und braucht nicht unbedingt positiv definit zu sein wie in der Methode von Fletcher und Powell [5]. Es ergibt sich eine echte Näherung für die Funktion in der Umgebung des momentanen Parametervektors, auch wenn ein Sattelpunkt vorhanden ist. Mit Hilfe dieser Matrix wird näherungsweise der Punkt bestimmt, in dem der Gradient verschwindet. Dies ist nicht nur im Minimum, sondern auch im Sattelpunkt der Fall. In der Nähe eines Minimums zielen die Schrittvorschläge zu diesem hin. Bei einem Sattelpunkt könnte ein derartiger Schrittvorschlag in Richtung wachsender Funktionswerte zeigen. Er würde zum Sattelpunkt statt zum Minimum führen. In diesem Fall wird, damit man vom Sattel wegkommt, ein Schritt vorgeschlagen in Richtung des Eigenvektors, der dem kleinsten Eigenwert der Näherungsmatrix für die zweiten Ableitungen zugeordnet ist.

Zum Aufbau der Iterationsmatrix wird ein anderer Algorithmus verwendet als in [5]. Dieser Algorithmus verlangt nicht die Beachtung von Orthogonalitätsforderungen. Die Schritte können beliebig sein; für den Aufbau der Matrix ist lediglich erforderlich, daß die Vektorpaare den ganzen Parameterraum erfassen, also linear unabhängig sind.

Als Startmatrix kann die Einheitsmatrix verwendet werden. Bei guten Näherungen empfiehlt es sich, aus n Probebewegungen, z.B. längs den Koordinatenachsen, die Matrix vorher aufzubauen.

Damit das Verfahren stabil ist, wird bei jedem Schritt eine eindimensionale Minimumsuche vorgenommen. Das Minimum braucht hierbei nicht genau bestimmt zu werden.

5. Der Algorithmus

Zur Funktion

$$F(x) = f_0 + \langle g(x_0) \,|\, x - x_0 \rangle + \tfrac{1}{2} \langle x - x_0 \,|\, K \,|\, x - x_0 \rangle \tag{3}$$

gehört der Gradient

$$|\, g(x) \rangle = |\, g(x_0) \rangle + |\, K \,|\, x - x_0 \rangle . \tag{4}$$

Im Minimum der Funktion verschwindet der Gradient. Für den Zuschlagvektor $|\, s \rangle$, der zum Minimum führt, folgt:

$$|\, s \rangle = |\, x - x_0 \rangle = - |\, K^{-1} \,|\, g(x_0) \rangle . \tag{5}$$

Haben einer oder mehrere der Parameter den Rand errreicht, so ist das Minimum von F unter der Nebenbedingung

$$|\, b_0 \rangle + |\, B \,|\, x \rangle = 0 \tag{6}$$

zu bestimmen. Gl. (6) geht aus dem Ungleichungssystem (2) hervor. Sind m Randbedingungen zu beachten, dann ist $|\, b_0 \rangle$ ein m-dimensionaler Vektor und $|\, B \,|$ eine rechteckige Matrix aus n Spalten und m Zeilen.

Für die Minimumsuche unter Nebenbedingungen bietet sich die Lagrange-Methode an. Hierbei ist das Minimum der Funktion

$$\begin{aligned} Q(x, \Lambda) &= F(x) + \langle \Lambda \,[\, |\, b_0 \rangle + |\, B \,|\, x \rangle] \\ &= f_0 + \langle g(x_0) \,|\, x - x_0 \rangle + \tfrac{1}{2} \langle x - x_0 \,|\, K \,|\, x - x_0 \rangle + \langle \Lambda \,[\, |\, b_0 \rangle + |\, B \,|\, x \rangle] \end{aligned} \tag{7}$$

zu finden. $|\Lambda\rangle$ ist dabei der m-dimensionale Vektor der Lagrange-Parameter.

Für $|\Lambda\rangle$ und $|\, x \rangle$ gelten neben der Gl. (6) die Beziehungen

$$|\, b_0 \rangle + |\, B \,|\, x_0 \rangle = 0 , \tag{8}$$

$$|\, g(x_0) \rangle + |\, K \,|\, x - x_0 \rangle + |\, B^t \,|\Lambda\rangle = 0 . \tag{9}$$

Aus diesen Gleichungen läßt sich der zum Minimum führende Zuschlag $|\, u \rangle$ entnehmen:

$$\text{Aus (6) und (8) folgt } |\, B \,|\, x - x_0 \rangle = 0 . \tag{10}$$

$$\text{Aus (9) } |\, x - x_0 \rangle = - |\, K^{-1} \,[\, |\, g(x_0) \rangle + |\, B^t \,|\Lambda\rangle] . \tag{11}$$

(11) in (10): $$|B|K^{-1}|\,g(x_0)\rangle + |B|K^{-1}|\,B^t|\Lambda\rangle = 0.$$

Daraus $$|\Lambda\rangle = -[|B|K^{-1}|\,B^t|]^{-1}\,B|K^{-1}|\,g(x_0)\rangle. \tag{12}$$

(12) in (11):

$$|x - x_0\rangle = |u\rangle = -|K^{-1}|\,g(x_0)\rangle + |K^{-1}|\,B^t[|B|K^{-1}|\,B^t|]^{-1}\,B|K^{-1}|\,g(x_0)\rangle$$

oder mit (5)

$$|u\rangle = |s\rangle + |K^{-1}|\,B^t[|B|K^{-1}|\,B^t|]^{-1}\,B|K^{-1}|\,g(x_0)\rangle. \tag{13}$$

Wie man sieht, besteht der Zuschlag $|u\rangle$ unter Nebenbedingungen aus dem Zuschlag $|s\rangle$ bei freier Iteration und einem Korrekturglied. Gl. (13) verlangt keine vollständige Matrixinversion, wie es zunächst scheint. Es genügt, wenn das Gleichungssystem

$$|B|K^{-1}|\,B^t|z\rangle = |B|K^{-1}|\,g(x_0)\rangle \tag{14}$$

nach $|z\rangle$ aufgelöst wird.

Die Arbeitsweise des Verfahrens sei an Hand des Flußdiagramms Fig. 2 erläutert.

Bei guten Ausgangsnäherungen ist es vorteilhaft, wenn vor Iterationsbeginn die Matrix $|K^{-1}|$, die Inverse der Matrix der zweiten Ableitungen von F und damit auch von Φ, bestimmt wird. Auf diese Weise lassen sich Schritte und Funktionswertberechnungen einsparen. Zum Aufbau der Matrix wird der Gradient der Funktion Φ in $n+1$ Punkten der engeren Umgebung der Ausgangsnäherung benötigt. Die Matrix gewinnt man wie folgt:

Man bildet

$$|\Delta x_i\rangle = |x_i\rangle - |x_0\rangle, \tag{15}$$

$$|\Delta g_i\rangle = |g_i\rangle - |g_0\rangle, \tag{16}$$

$$|y_i\rangle = |\Delta x_i\rangle - |H_{i-1}|\Delta g_i\rangle, \tag{17}$$

$$|H_i| = |H_{i-1}| + \frac{|y_i\rangle\langle y_i|}{\langle y_i|\Delta g_i\rangle}. \tag{18}$$

Der Index i läuft dabei von 1 bis n. Als $|\Delta x_i\rangle$ können z.B. Änderungen längs der Koordinatenachsen gewählt werden. $|H_0|$ kann eine beliebige symmetrische Matrix sein, etwa die Einheitsmatrix.

$|H_n|$ stimmt mit $|K^{-1}|$ der Taylorentwicklung überein. Näheres über diese Vorschrift für den Matrixaufbau findet sich in Abschnitt 6.

Ist die Ausgangsnäherung schlecht, dann bringt ein Matrixaufbau am Anfang keine Vorteile. Er kostet nur unnötig Zeit. In diesem Fall startet man am besten mit der Einheitsmatrix.

Die vorliegende Matrix $|H|$ braucht nicht unbedingt positiv definit zu sein, und es kann daher geschehen, daß der Schrittvorschlag $|s\rangle$ (oder $|u\rangle$) in Richtung wachsender Funktionswerte zeigt. Die Matrix enthält dann negative Eigenwerte, ein Zeichen dafür, daß ein Sattelpunkt in der Nähe liegt und die Näherung nicht für das gesuchte Minimum gilt. In diesem Fall ist es am günstigsten, wenn der Schrittvorschlag in das vom Sattelpunkt am steilsten abwärts führende Tal weist. Dieses Tal läuft vom Sattel aus in Richtung des zum kleinsten Eigenwert ge-

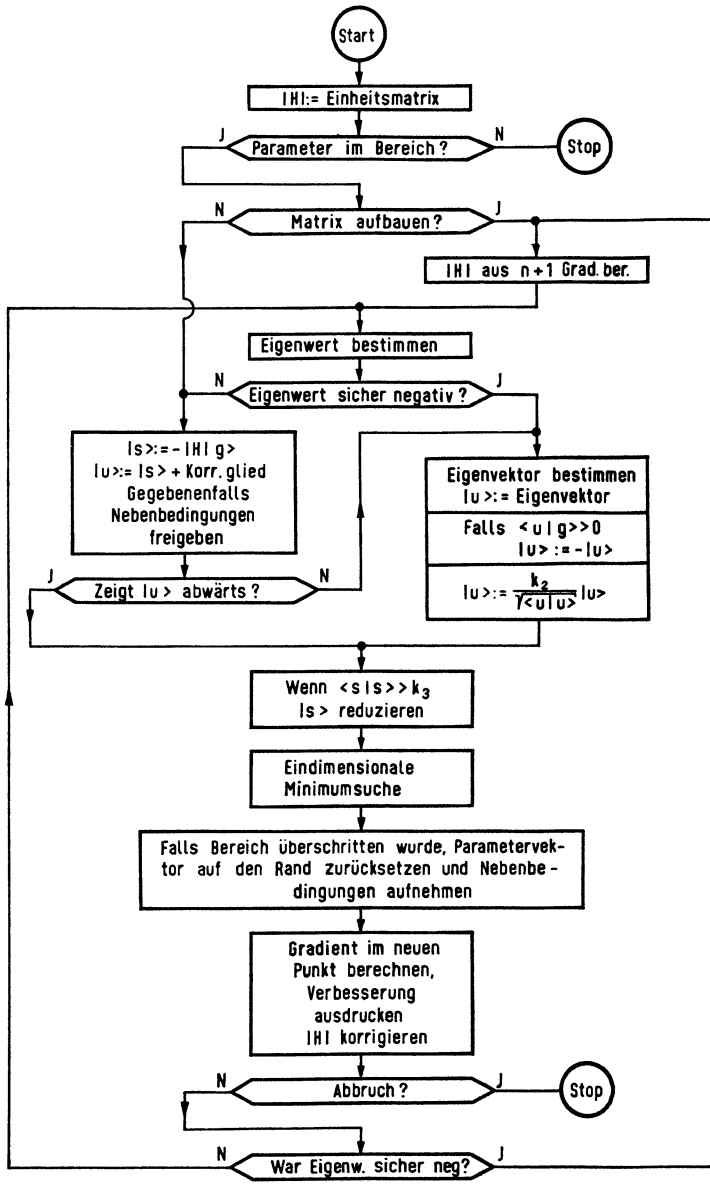

Fig. 2. Flußdiagramm

hörenden Eigenvektors der Matrix $|K|$, sofern $|K|$ die Umgebung dieses Sattels richtig beschreibt. Für $|K^{-1}|$ ist dies der größte negative Eigenwert.

Unter Nebenbedingungen tritt an Stelle von $|K^{-1}|$ die Matrix

$$|C| = |K^{-1}| - |K^{-1}|\, B^t [|B|\, K^{-1}|\, B^t|]^{-1} B |K^{-1}|, \tag{19}$$

wie man aus (13) leicht sieht.

Diese Matrix ist so aufgebaut, daß sie Änderungen des Gradienten in solche Änderungen des Parametervektors transformiert, die den Rand des Bereichs, also die m erfaßten Begrenzungsebenen, nicht verlassen. Alle Vektoren, die zu diesen Randebenen orthogonal sind, werden von $|C|$ in Null transformiert. $|C|$ ist daher singulär mit dem Rangabfall m und hat den m-fachen Eigenwert 0.

Sind die übrigen $n-m$ Eigenwerte sämtlich positiv, dann wird nach (5) $|s\rangle$ errechnet. An die Stelle von $|K^{-1}|$ tritt dabei in dieser Gleichung natürlich $|H|$. $|H|$ stimmt nur bei vorausgegangenem Matrixaufbau oder gegen Ende der Iteration mit $|K^{-1}|$ überein. Unter Nebenbedingungen ist an $|s\rangle$ noch eine Korrektur vorzunehmen. Bevor diese angebracht wird, ist zu untersuchen, ob $|s\rangle$ von einer oder mehreren der Berandungsebenen zurück in den Körper zeigt. Ist dies der Fall, dann werden die betreffenden Nebenbedingungen probeweise freigegeben. Weist auch der korrigierte Vektor $|u\rangle$ nicht aus dem Körper heraus, dann war die Freigabe berechtigt, ansonsten war sie verfrüht, und der Vektor muß wieder rückgestellt werden.

Die Eigenwerte von $|H|$ und $|C|$ können nur mit beschränkter Genauigkeit ermittelt werden. Es ist daher sinnvoll, auch bei negativen Eigenwerten, die an der Grenze des positiven Bereiches liegen, die Matrix $|H|$ oder $|C|$ zunächst einmal als positiv anzunehmen. Die Matrix wird als nicht positiv definit angesehen, wenn sie Eigenwerte in dem Intervall $k_0 < \lambda < k_1$ besitzt. Die Größen k_0 und k_1 sind dabei entsprechend der zu erwartenden Ungenauigkeit zu wählen (z.B. $k_0 = -10^{+2}$, $k_1 = 1/k_0$). Zeigt der Schrittvorschlag $|u\rangle$ (bzw. $|s\rangle$, wenn keine Nebenbedingungen vorliegen) abwärts, ist also $\langle u|-g\rangle > 0$, dann ist er brauchbar, ansonsten wird wie bei negativem Eigenwert vorgegangen.

Bei negativen Eigenwerten, genauer, wenn ein Eigenwert in das Intervall $k_0 < \lambda < k_1$ fällt, wird die Richtung des zum kleinsten dieser Eigenwerte zugehörigen Eigenvektors bestimmt. Damit dieser Vektor in Richtung fallender Funktionswerte weist, muß gegebenenfalls noch sein Vorzeichen umgekehrt werden. Der Eigenvektor gibt nur eine Richtung, sein Betrag liegt nicht fest. In dieser Richtung wird ein kleiner Schritt vom Betrag k_2 vorgeschlagen (z.B. $k_2 = 10^{-2}$).

Komplizierte Funktionen besitzen mehrere Minima. Übersteigt der Betrag von $|s\rangle$ (bzw. $|u\rangle$) eine Schranke k_3, dann wird er auf diese Schranke reduziert. Auf diese Weise soll erreicht werden, daß die Iteration in jenes Minimum mündet, das der Rohnäherung am nächsten liegt. Man verhindert so, daß die Iteration in ein „Nachbartal" springt, das zu einem anderen Minimum führt.

Der Schrittvorschlag liegt damit fest und zeigt abwärts. Um Stabilität und Konvergenz des Verfahrens zu sichern, wird längs des durch den Schrittvorschlag gegebenen Strahls das Minimum bestimmt. Dieses Minimum braucht nicht genau ermittelt zu werden, denn es bestehen keine Orthogonalitätsforderungen wie in [5]. In Abschnitt 7 wird eine Methode zur Bestimmung dieses Minimums näher beschrieben.

Im Anschluß an diese eindimensionale Minimierung wird untersucht, ob der Parametervektor noch im vorgeschriebenen Körper liegt. Wenn nicht, wird er auf die Berandungsfläche zurückgesetzt und die durchstoßene Randebene als Nebenbedingung aufgenommen.

Der Schritt ist jetzt ausgeführt. Im neuen Punkt wird der Gradient berechnet und aus der Differenz des neuen und des alten Gradienten und der Parameter-

änderung die Matrix $|H|$ korrigiert:

$$|\Delta x\rangle = |x_{j+1}\rangle - |x_j\rangle; \tag{15a}$$

$$|\Delta g\rangle = |g_{j+1}\rangle - |g_j\rangle; \tag{16a}$$

$$|y\rangle = |\Delta x\rangle - |H_j|\Delta g\rangle; \tag{17a}$$

$$|H_{j+1}| = |H_j| + \frac{|y\rangle\langle y|}{\langle y|\Delta g\rangle}. \tag{18a}$$

Für die Korrektur findet der bereits zum Matrixaufbau benutzte und in Abschnitt 6 näher beschriebene Algorithmus Verwendung.

Im nächsten Programmabschnitt wird geprüft, ob die Kriterien für den Abbruch der Iteration erfüllt sind. Als Abbruchkriterien dienen dabei die größte Parameteränderung und der Betrag der Gradientenkomponente senkrecht zu den erfaßten Randebenen.

Ist das Minimum noch nicht erreicht, dann wird der Zyklus erneut durchlaufen. War die Näherungsfunktion eindeutig eine Sattelfläche (λ im Intervall $k_0 < \lambda < k_1$) und wurde der Schritt parallel zum Tal dieser Fläche ausgeführt, dann beginnt der neue Zyklus mit einem Matrixaufbau. Diesem Vorgehen liegen folgende Überlegungen zugrunde: Bei einer quadratischen Funktion würde die Funktion längs dieser Richtung beständig abnehmen. Der Schritt in Eigenvektorrichtung führt aber bei der eindimensionalen Minimumsuche bis zu einem Minimum und damit aus der Umgebung heraus, in der die Näherungsfunktion F die zu minimierende Funktion Φ beschrieb. In der neuen Umgebung gilt die Näherung nicht mehr. Um in dieser Umgebung die Iteration mit einer brauchbaren Näherungsfunktion fortführen zu können, wird die Matrix für diese Umgebung neu bestimmt. Ein Matrixaufbau nach den Gln. (15)—(18) kostet weniger Zeit als n Schritte, in denen neben den Gradientenberechnungen noch zusätzlich Funktionswerte für die eindimensionalen Minimierungen berechnet werden müssen.

6. Beweis der Vorschrift für den Matrixaufbau

Es seien n Vektorpaare gegeben,

$$|\Delta g_i\rangle, \quad |\Delta x_i\rangle; \quad i = 1, n,$$

und es existiere eine symmetrische Matrix $|K^{-1}|$ mit der Eigenschaft

$$|K^{-1}|\Delta g_i\rangle = |\Delta x_i\rangle; \quad i = 1, n. \tag{20}$$

Diese Matrix ist aus einer beliebigen symmetrischen Rohmatrix $|H_0|$ schrittweise aufzubauen:

$$|H_j| = |H_{j-1}| + |\Delta H_j|. \tag{21}$$

Die Matrix $|H_j|$ soll die ersten j Vektoren $|\Delta g_i\rangle$ $(i = 1, j)$ in die entsprechenden $|\Delta x_i\rangle$ transformieren. Sofern die $|\Delta g_i\rangle$ linear unabhängig sind, ist dann nach n Schritten

$$|H_n| \equiv |K^{-1}|. \tag{22}$$

Der sukzessive Aufbau der Matrix $|K^{-1}|$ ist im Iterationsverfahren deshalb vorteilhaft, weil sich die Matrix der Ableitungen zweiter Ordnung, deren Inverse ja

$|K^{-1}|$ ist, bei beliebigen Funktionen während der Iteration ändert. Für eine Form zweiter Ordnung ist diese Matrix vom Ort unabhängig und konstant.

Gl. (21) ist bereits der Lösungsansatz für die gestellte Aufgabe. Ehe auf die von $|H_j|$ zu fordernden Eigenschaften eingegangen wird, zunächst noch einige allgemeine Betrachtungen.

Durch Linearkombinationen von $j-1$ linear unabhängigen Vektoren $|\Delta g_i\rangle$ wird ein $(j-1)$-dimensionaler Unterraum $R_{g,j-1}$ des n-dimensionalen Vektorraumes R aufgespannt. $|K^{-1}|$ bildet diesen Raum auf einen Raum $R_{x,j-1}$ ab. In $R_{x,j-1}$ liegen die Vektoren $|\Delta x_i\rangle$, $i=1, j-1$.

Nimmt man an, $|H_{j-1}|$ transformiere die Vektoren $|\Delta g_i\rangle$, $i=1, \ldots, j-1$ in der vorgeschriebenen Weise, dann transformiert $|H_{j-1}|$ den Raum $R_{g,j-1}$ in den Raum $R_{x,j-1}$. $|\Delta H_j|$ muß $R_{g,j-1}$ in Null transformieren, denn er wird ja von $|H_{j-1}|$ bereits richtig transformiert. Für die Verbesserungsmatrix setzt man am einfachsten eine Matrix vom Range Eins an. Der Eigenvektor dieser Matrix mit nicht verschwindendem Eigenwert sei $|y_i\rangle$. $|y_i\rangle$ muß orthogonal sein zu $R_{g,j-1}$. $R_{g,j-1}$ schneidet aus dem Gesamtraum R einen Teilraum heraus. Der verbleibende Restraum wird von $|H_{j-1}|$ in beliebiger Weise transformiert. $|\Delta g_j\rangle$ hat eine Komponente in diesem Restraum, wenn es von den $|\Delta g_i\rangle$, $i=1, \ldots, j-1$ nicht linear abhängt. Diese Komponente transformiert $|H_{j-1}|$ normalerweise falsch. $|\Delta H_j|$ soll diesen Fehler ausgleichen.

Die Differenzmatrix

$$|D_j| = |K^{-1}| - |H_{j-1}| \tag{23}$$

bildet $R_{g,j-1}$ in Null ab.

$|D_j|$ ist symmetrisch und hat deshalb orthogonale Eigenvektoren. $|D_j|$ ist auch singulär mit dem Rangabfall $j-1$. Die Eigenvektoren von $|D_j|$ mit verschwindendem Eigenwert spannen gerade den Raum $R_{g,j-1}$ auf. Denkt man sich $|D_j|$ nach den Eigenvektoren $|e_k\rangle$ zerlegt,

$$|D_j| = \sum_{k=1}^{n} \frac{|e_k\rangle\langle e_k|}{\langle e_k|e_k\rangle}\lambda_k, \tag{24}$$

und multipliziert man diese Zerlegung mit einem beliebigen Vektor, dann ist unmittelbar einzusehen, daß der Produktvektor keine Komponente in $R_{g,j-1}$ haben kann, weil die Eigenvektoren orthogonal sind und jene Eigenvektoren, die in $R_{g,j-1}$ transformieren, den Eigenwert Null haben. Das Produkt

$$|D_j|\Delta g_j\rangle = |\Delta x_j\rangle - |H_{j-1}|\Delta g_j\rangle, \tag{25}$$

das gerade den Fehler liefert, um den $|H_{j-1}|$ den Vektor $|\Delta g_j\rangle$ falsch transformiert, ist folglich orthogonal zu $R_{g,j-1}$. Es hat damit alle Eigenschaften, die vom Eigenvektor $|y_j\rangle$ von $|\Delta H_j|$ zu fordern sind:

$$|y_j\rangle = |\Delta x_j\rangle - |H_{j-1}|\Delta g_j\rangle. \tag{26}$$

Die Korrekturmatrix $|\Delta H_j|$ ist also bis auf einen Faktor γ bestimmt:

$$|\Delta H_j| = \gamma|y_j\rangle\langle y_j|. \tag{27}$$

Aus der Forderung

$$|H_j|\Delta g_j\rangle = |\Delta x_j\rangle$$

ergibt sich mit (21), (26) und (27)

$$\gamma = \frac{1}{\langle y_j | \Delta g_j\rangle} \tag{28}$$

und somit

$$|\Delta H_j| = \frac{|y_j\rangle \langle y_j|}{\langle y_j | \Delta g_j\rangle}. \tag{29}$$

Die Rohmatrix $|H_0|$ wird normalerweise keinen der Vektoren $|\Delta g_i\rangle$ richtig transformieren. $|H_1|$ transformiert aber dann $|\Delta g_1\rangle$ wie vorgeschrieben in $|\Delta x_1\rangle$. Durch Schluß von n auf $n+1$ ergibt sich: $|H_i|$ transformiert die ersten i Vektoren richtig, und $|H_n|$ stimmt schließlich mit $|K^{-1}|$ überein, denn es transformiert den ganzen Raum R.

$|H_j|$ braucht nicht positiv definit zu sein. $|H_j|$ strebt gegen $|K^{-1}|$. Besitzt $|K^{-1}|$ negative Eigenwerte, dann hat auch $|H_n|$ welche.

Ein kleines Beispiel soll die Arbeitsweise des Algorithmus zeigen. Gegeben sei die Matrix

$$|K| = \begin{vmatrix} 1 & 2 & 3 \\ 2 & 5 & 8 \\ 3 & 8 & 14 \end{vmatrix};$$

zwischen $|\Delta g\rangle$ und $|\Delta x\rangle$ gelte die Beziehung

$$|\Delta g\rangle = |K|\Delta x\rangle.$$

Anfangsmatrix sei die Einheitsmatrix. Die Tabelle zeigt den Rechengang bei willkürlich gewählten $|\Delta x_i\rangle$. Nach drei Schritten erhält man die Inverse von $|K|$.

Tabelle. *Zahlenbeispiel für Matrixaufbau*

i	H_{i-1}	$\lvert\Delta x_i\rangle$	$\lvert\Delta g_i\rangle$	$\lvert y_i\rangle$
1	$\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}$	$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 0 \\ -2 \\ -3 \end{pmatrix}$
2	$\frac{1}{13}\begin{vmatrix} 13 & 0 & 0 \\ 0 & 9 & -6 \\ 0 & -6 & 4 \end{vmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}$	$\frac{1}{13}\begin{pmatrix} -26 \\ 16 \\ -2 \end{pmatrix}$
3	$\frac{1}{3}\begin{vmatrix} 16 & -8 & 1 \\ -8 & 7 & -2 \\ 1 & -2 & 1 \end{vmatrix}$	$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 3 \\ 8 \\ 14 \end{pmatrix}$	$\frac{2}{3}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$
4	$\begin{vmatrix} 6 & -4 & 1 \\ -4 & 5 & -2 \\ 1 & -2 & 1 \end{vmatrix}$		—	

7. Minimumsuche längs des Strahls

Das Verfahren, das zur Bestimmung des Minimums längs des Strahls

$$|r\rangle = \alpha|s\rangle \tag{30}$$

verwendet wurde, setzt sich aus drei Teilen zusammen, nämlich:

a) einem Intervallaufweitungsverfahren, bei dem durch Intervallvergrößerung längs des betrachteten Strahls ein Intervall bestimmt wird, in dem das Minimum liegt;

b) einem Intervallschachtelungsverfahren, das dieses Intervall auf einen vorgegebenen Wert einengt;

c) einem Interpolationsverfahren, das in dem durch Schachtelung gewonnenen Intervall das Minimum durch Interpolation bestimmt.

Die einzelnen Teile der Minimumsuche sollen zusammenpassen. Der komplizierteste Teil ist das Schachtelungsverfahren. Es legt die übrigen Teile weitgehend fest und soll deshalb zuerst besprochen werden.

Ist in einem Intervall die Ordinate eines inneren Punktes kleiner als die der Randpunkte, dann existiert ein Minimum in dem Intervall. (Die Funktion ist dabei als konvex vorauszusetzen.) Kennt man in dem Intervall noch einen weiteren Punkt, dann läßt sich ein Teilintervall abspalten, in dem das gesuchte Minimum enthalten ist.

In Fig. 3 liegt das Minimum im Intervall (a_1, a_4). Die beiden Intervalle (a_1, a_3), (a_2, a_4) überdecken das Intervall (a_1, a_4).

Das Minimum der Funktion muß also in mindestens einem der beiden Teilintervalle liegen. In welchem es liegt, hängt davon ab, welcher der beiden mittleren Punkte die kleinere Ordinate hat.

Aus dem Intervall (a_1, a_4) entsteht so das Intervall (a_1, a_3) oder (a_2, a_4). Berechnet man in dem Teilintervall einen weiteren Punkt, dann kann aus diesem ein noch kleineres Intervall herausgespalten werden.

Die Punkte a_2, a_3 sollten das Intervall (a_1, a_4) gleichmäßig aufteilen, damit Ungenauigkeiten in der Funktionsberechnung möglichst wenig stören und die Intervallängen schnell abnehmen. Fortgesetztes Schachteln sollte die Teilverhält-

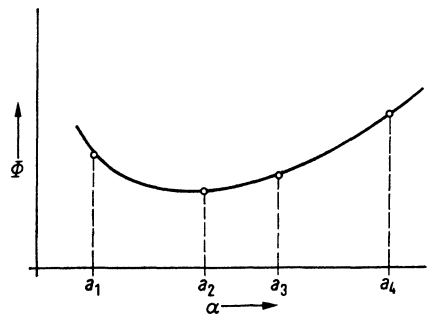

Fig. 3. Intervallschachtelung bei der eindimensionalen Minimumsuche

nisse erhalten. Aus diesen Überlegungen folgt der Ansatz:

$$\frac{a_3 - a_1}{a_4 - a_1} = \frac{a_2 - a_1}{a_3 - a_1}, \tag{31}$$

$$(a_2 - a_1) + (a_4 - a_2) = (a_4 - a_1), \tag{32}$$

$$(a_4 - a_2) = (a_3 - a_1). \tag{33}$$

Die Gln. (31), (32) und (33) führen zu der Lösung

$$(a_3 - a_1) = [-\tfrac{1}{2} \pm \tfrac{1}{2}\sqrt{5}]\,(a_4 - a_1). \tag{34}$$

$(a_3 - a_1)$ soll positiv sein. Somit bleibt nur die Lösung:

$$(a_3 - a_1) = \frac{\sqrt{5} - 1}{2}\,(a_4 - a_1). \tag{35}$$

Für die andere Teilstrecke folgt damit aus (31)

$$(a_2 - a_1) = \frac{3 - \sqrt{5}}{2}\,(a_4 - a_1). \tag{36}$$

Die Gln. (35) und (36) besagen, daß die Strecke $\overline{a_1\, a_3}$ durch die Punkte a_2 und a_4 innen und außen harmonisch geteilt wird. Im angegebenen Beispiel läßt sich aus dem Intervall (a_1, a_4) das Teilintervall (a_1, a_3) herausspalten. Wurden a_2 und a_3 in obiger Weise bestimmt, dann läßt sich (a_1, a_3) durch einen weiteren Punkt zu einem Vier-Punkte-Intervall ergänzen, das wieder harmonisch aufgeteilt ist. Nach vier Schritten, also nach Berechnung von vier weiteren Funktionswerten, ist das ursprüngliche Intervall auf etwa ein Zehntel seiner Breite zusammengeschrumpft. Die Schachtelung wird abgebrochen, wenn die Intervallbreite einen vorgegebenen Wert unterschreitet. Dieses Schachtelverfahren ist eine einfache und gute Näherung für die Fibonacci-Schachtelung [1].

Dem Schachtelverfahren ist das Aufweitungsverfahren vorgeschaltet. Bei der Aufweitung soll sich ein harmonisch geteiltes Intervall ergeben, an das die Schachtelung unmittelbar angeschlossen werden kann. Für die Aufweitung wurde deshalb folgende Methode gewählt:

Man beginnt mit zwei Punkten a_0 und a_1 und ergänzt diese durch einen dritten Punkt a_2, so daß das Intervall (a_0, a_2) von a_1 harmonisch geteilt wird. Im erstellten Programm haben die Punkte a_0 den Wert $\alpha = 0$ und a_2 den Wert $\alpha = 1$.

Aus drei Punkten läßt sich entscheiden, ob im vorliegenden Intervall ein Minimum enthalten ist. Ist $f(a_2) < f(a_1)$, dann wird ausgehend vom Intervall (a_1, a_2), ein äußerer Punkt a_3 bestimmt, so daß a_2 die Strecke $\overline{a_1\, a_3}$ harmonisch teilt. Danach ist zu untersuchen, ob im neuen Intervall ein Minimum enthalten ist $(f(a_3) < f(a_2)\,?)$ und so fort.

Die Aufweitung ist abzubrechen, wenn

$$f(a_{n+1}) > f(a_n)$$

ist.

Die Intervallbreite bei der Aufweitung wächst in geometrischer Progression. Zwei aufeinanderfolgende Intervalle verhalten sich wie $(1 + \sqrt{5})/2$. Der n-te Punkt

hat die Abszisse

$$a_n = \frac{\sqrt{5}-1}{2}\left[\left(\frac{\sqrt{5}+1}{2}\right)^n - 1\right]. \tag{37}$$

Für den dritten Teil der eindimensionalen Minimumsuche, die Interpolation, bietet sich ein Polynom dritten Grades an. Es läßt sich aus den vier Punkten des letzten Schachtelintervalls bestimmen und besitzt ein eindeutiges Minimum. Bei der Bestimmung dieses Minimums ist zu beachten, daß das Polynom dritten Grades zu einer Parabel entarten kann.

Sollte der Funktionswert an der interpolierten Stelle größer sein als der Funktionswert zu Beginn der eindimensionalen Minimumsuche (dies kann bei extrem nichtlinearen Funktionen geschehen), dann ist die Schachtelung fortzusetzen. Ergibt sich auch dabei kein kleinerer Funktionswert, dann ist die Iteration in ein höher gelegenes Nachbartal gerutscht.

In Abschnitt 5 wurde gesagt, in scheinbarem Widerspruch zu diesen Ausführungen, für die Iteration sei eine genaue Minimumbestimmung nicht erforderlich. Das hier beschriebene aufwendige Verfahren soll lediglich gewährleisten, daß die Funktion bei jedem Iterationsschritt abnimmt. Einfachere Methoden führen bei komplizierten Funktionen zur Divergenz.

Im verwendeten Programm wird nach der Intervallaufweitung gleich interpoliert. Die Schachtelung setzt nur ein, wenn der Funktionswert an der interpolierten Stelle nicht kleiner geworden ist.

8. Beispiele

Das beschriebene Verfahren wurde programmiert und an Funktionen mit maximal 50 Variablen erprobt. In einem Programm zur Optimierung linearer elektrischer Netzwerke bewährt es sich seit über einem Jahr. Für die Neben-

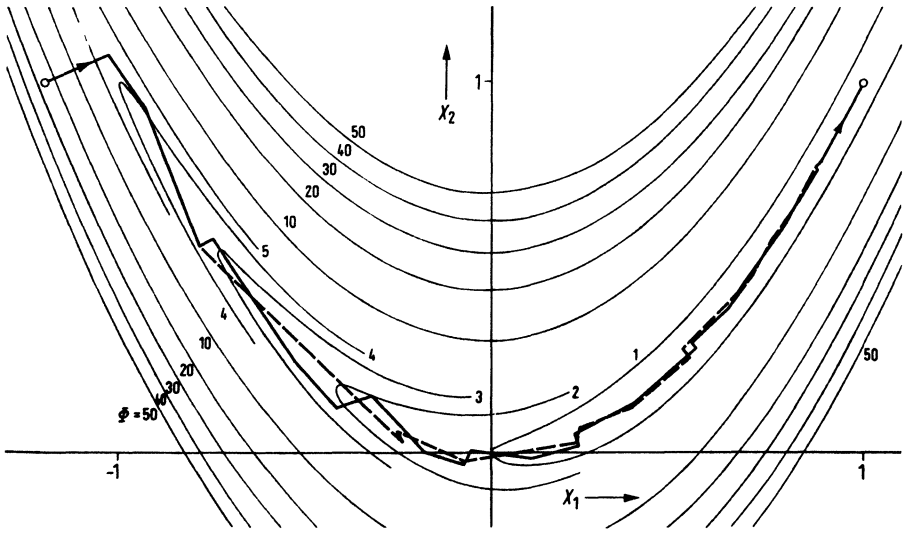

Fig. 4. Iteration an der Funktion $\Phi=(1-x_1)^2+100(x_2-x_1^2)^2$ (Bananafunction).
—— Neue Methode. --- Fletcher-Powell

bedingungen wurde dabei nicht die allgemeine Form (2) vorgesehen sondern die einfachere

$$x_{i\min} \leqq x_i \leqq x_{i\max}.$$

Weil es in diesem speziellen Fall bequemer ist, wurden die Eigenvektoren nicht aus $|H|$, sondern aus der Matrix $|H^{-1}|$ gewonnen. $|H^{-1}|$ wurde in gleicher Weise wie $|H|$ aufgebaut; im Algorithmus sind dabei nur $|\Delta g\rangle$ und $|\Delta x\rangle$ zu vertauschen. Zur Berechnung der benötigten Eigenwerte und Eigenvektoren diente ein abgewandeltes Hessenberg-Verfahren [5].

Die folgenden Beispiele sollen einen Eindruck von der Arbeitsweise des Verfahrens vermitteln. In der zum Vergleich angegebenen Fletcher-Powell-Methode wurde dasselbe oben beschriebene Unterprogramm für die eindimensionale Minimumsuche benutzt. Vergleichen lassen sich die beiden Verfahren natürlich nur für Iterationen ohne Nebenbedingungen.

Fig. 4 zeigt die Minimierung der Funktion $\Phi = (1-x_1)^2 + 100\,(x_2 - x_1^2)^2$. Diese Funktion (Bananafunction) wurde von Rosenbrock in die Minimierungsliteratur eingeführt und zeichnet sich durch eine parabelförmige, extrem schmale Rinne aus, durch die die Iteration sich durchschlängeln muß. Die neue Methode benötigt vom Startpunkt $|-1.2, 1\rangle$ zur Lösung $|1, 1\rangle$ 25 Schritte, die nach Fletcher und Powell 18.

Bei dem in Fig. 5 gezeigten Beispiel wird eine Funktion mit Sattelpunkt minimiert: $\Phi = \frac{1}{4}x_1 + (x_1^2 - 2x_1 + x_2^2)^2$. Die Funktionsfläche hat die Gestalt eines schräg gestellten Hutes mit ausladender Krempe (Sombrero). Die vom Startpunkt zur Lösung führende Rinne ist nicht so extrem schmal wie im obigen Beispiel. Hier wurde nicht mit einem Gradientenschritt, sondern mit einem Matrixaufbau begonnen. Die Matrix der zweiten Ableitungen besitzt in der Nähe des

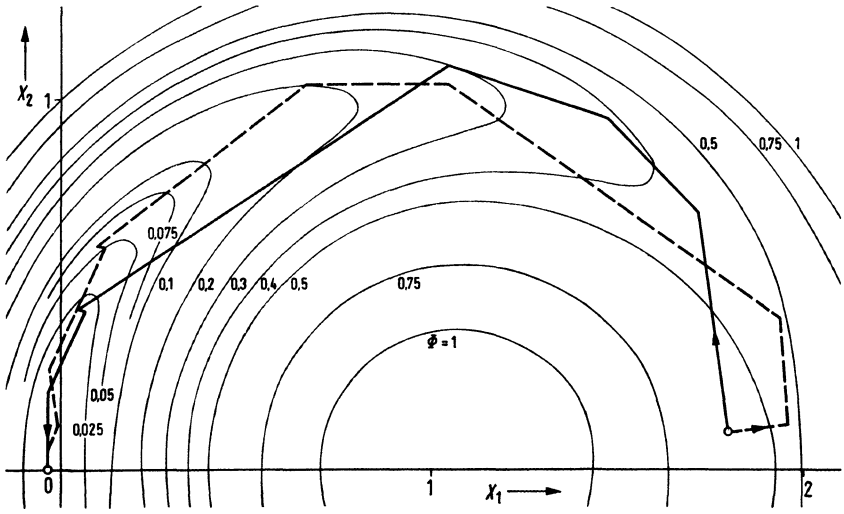

Fig. 5. Iteration an der Funktion $\Phi = \frac{1}{4}x_1 + (x_1^2 - 2x_1 + x_2^2)^2$. —— Neue Methode. --- Fletcher-Powell

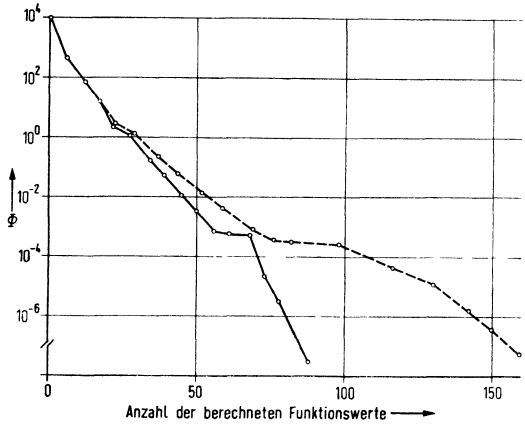

Fig. 6

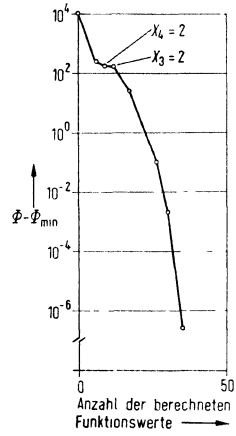

Fig. 7

Fig. 6. Iteration ohne Nebenbedingungen an der Funktion
$\Phi=(x_1+10x_2)^2+5(x_3-x_4)^2+(x_2-2x_3)^4+10(x_1-x_4)^4$.
—— Neue Methode. --- Fletcher-Powell

Fig. 7. Iteration unter den Nebenbedingungen $x_3 \geqq 2$, $x_4 \geqq 2$ an der Funktion
$\Phi=(x_1+10x_2)^2+5(x_3-x_4)^2+(x_2-2x_3)^4+10(x_1-x_4)^4$

Sattels einen negativen Eigenwert, und der erste Schritt führt parallel zur Talsohle vom Sattel weg. Die neue Methode benötigt hier 7, die von Fletcher und Powell 11 Schritte.

In Fig. 6 ist der Funktionswert über der Anzahl der berechneten Funktionswerte beim Minimieren der Funktion

$$\Phi=(x_1+10x_2)^2+5(x_3-x_4)^2+(x_2-2x_3)^4+10(x_1-x_4)^4$$

aufgetragen. Diese Funktion wurde von Fletcher und Powell angegeben. Im Minimum ($|x\rangle = |0, 0, 0, 0\rangle$) ist die Matrix der zweiten Ableitungen dieser Funktion doppelt singulär. Gestartet wurde in dem Punkt $|x\rangle = |1, 2, 3, 4\rangle$. An dieser Funktion zeigt das neue Verfahren gegenüber dem Fletcher-Powellschen deutlich Vorteile. Es konvergierte von allen angesetzten Startpunkten aus wesentlich schneller.

Eine Iteration unter Nebenbedingungen zeigt Fig. 7. Funktion und Startvektor sind die gleichen wie in Fig. 6. Nebenbedingungen sind die beiden Ungleichungen $x_3 \geqq 2$ und $x_4 \geqq 2$. Die Iteration ist hier schneller am Ziel ($|x\rangle = |1{,}27\ldots, 0{,}63\ldots, 2, 2\rangle$), weil nach Erreichen der Begrenzungsebenen effektiv nur mehr ein zweidimensionaler Raum der Variation zur Verfügung steht. Ein zweidimensionales Problem bereitet weniger Mühe als ein vierdimensionales.

Für die folgenden Testbeispiele mußte die Strategie des Verfahrens abgeändert werden. Die Funktionen erwiesen sich als so ausgeprägt nichtlinear, daß sich mit fest vorgegebenen Probebewegungen beim Aufbau der Iterationsmatrix nach einem Schritt längs des Eigenvektors keine brauchbaren Näherungsmatrizen er-

gaben. Dieser spezielle Matrixaufbau wurde deshalb fallen gelassen. Die Funktionen waren von der Form

$$\Phi = \sum_{k=1}^{1} \left[c \frac{\prod_{i=1}^{m} (1 + z_{1i} x_k + z_{2i} x_k^2)}{\prod_{j=1}^{m} (1 + n_{1j} x_k + n_{2j} x_k^2)} - f_k \right]^2 .$$

Die Sollwerte f_k wurden aus der Beziehung

$$f(x) = c \frac{\prod_{i=1}^{m} [(x+1-3i)^2 + \frac{1}{2}]}{\prod_{i=1}^{m} [(x+\frac{1}{2}-3i)^2 + \frac{1}{2}]}$$

gewonnen, wobei c gerade so gewählt war, daß f für $x=0$ den Wert 1 hatte. Variiert wurden die Parameter $z_{1i}, z_{2i}, n_{1j}, n_{2j}$ und c. Als Ausgangsnäherung wurden einfachheitshalber für die Parameter z und n die exakten Lösungswerte eingegeben, nur der Parameter c war von der Lösung verschieden. Bei der Iteration bewegten sich die Parameter z und n zunächst von den Lösungen fort und kehrten dann wieder zu ihnen zurück, ähnlich dem Parameter x_2 bei der „Bananafunction".

In den Fig. 8 und 9 sind die Iterationsabläufe für zwei Beispiele aufgezeichnet. Dabei ist der Funktionswert in logarithmischem Maßstab über der Anzahl der berechneten Funktionswerte aufgetragen. Zum Vergleich ist die Iteration nach der Methode von Fletcher und Powell gestrichelt eingezeichnet. Für $m=3$ erhält man eine 13-Parameter-Funktion. Fig. 8 zeigt die Iteration an dieser Funktion, wobei $l=24$ Stützpunkte im Bereich $0{,}5 \leqq x \leqq 12{,}0$ vorgegeben waren. Die neue Methode rechnete 61 Schritte und benötigte dazu 753 Funktionswerte. Bei

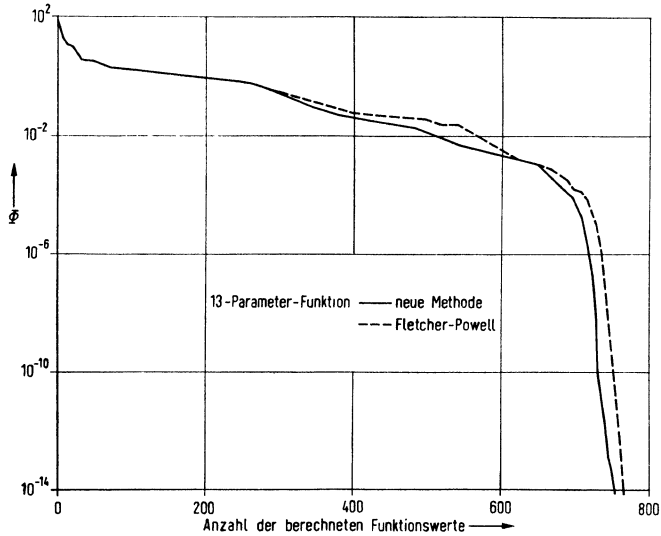

Fig. 8. Iteration an einer 13-Parameter-Funktion

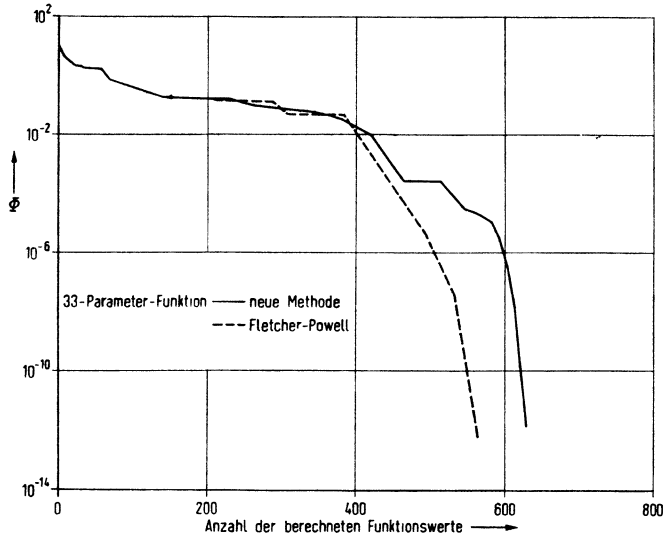

Fig. 9. Iteration an einer 33-Parameter-Funktion

Fletcher und Powell waren es 69 Schritte und 766 Funktionswerte. Setzt man $m=8$, dann besitzt die Funktion 33 Parameter. In Fig. 9 ist die Minimierung dieser Funktion für $l=52$ Stützwerte im Bereich $0{,}5 \leqq x \leqq 26{,}0$ dargestellt. Beide Verfahren rechnen 57 Schritte. Die neue Methode berechnet dazu die Funktion 629mal, Fletcher und Powells Verfahren 584mal.

Literatur

1. Bellmann, R.: Dynamic programming. Princeton: Princeton University Press 1962.
2. Dirac, P. A. M.: The principles of quantum mechanics. Oxford: Clarendon Press 1958.
3. Fiacco, A. V., McCormick, G. P.: The sequential unconstraint minimization technique for nonlinear programming, a primal-dual method. Management Science **10**, 360—366 (1964).
4. — — Computational algorithm for the sequential unconstrained minimization technique for nonlinear programming. Management Science **10**, 601—617 (1964).
5. Fletcher, R., Powell, M. J.: A rapidly convergent descent method for minimization. The Computer Journal **6**, 163 (1963).
6. Lasdon, L. S., Waren, A. D.: Optimal design of filters with bounded, lossy elements. IEEE Transactions CT-13 (June 1966), S. 175—187.
7. Zurmühl, R.: Matrizen, S. 321 ff. Berlin-Göttingen-Heidelberg: Springer 1964.

G. Zettl
Siemens AG
D-8 München 25
Postfach

Numer. Math. 15, 433–449 (1970)
© by Springer-Verlag 1970

Linear Self-Adjoint Multipoint Boundary Value Problems and Related Approximation Schemes

JOSEPH W. JEROME*

Received August 14, 1969

Summary. Linear self-adjoint multipoint boundary value problems are investigated. The case of the homogeneous equation is shown to lead to spline solutions, which are then utilized to construct a Green's function for the case of homogeneous boundary conditions. An approximation scheme is described in terms of the eigenfunctions of the inverse of the Green's operator and is shown to be optimal in the sense of the n-widths of Kolmogorov. Convergence rates are given and generalizations to more general boundary value problems are discussed.

§ 1. Introduction

Let L be a linear differential operator of the form

$$L = \sum_{j=0}^{m} a_j D^j$$

where the $a_j \in C^j[a, b]$, $a_m(x) \geqq \alpha > 0$ and $D = d/dx$ and let L^*,

$$L^* f = \sum_{j=0}^{m} (-1)^j D^j (a_j f),$$

denote the formal adjoint of L. We will consider a certain class of multipoint boundary value problems involving L^*L on $[a, b]$. Indeed, consider the partition

$$a = x_0 < x_1 < \cdots < x_k = b$$

where $k \geqq 1$. Let a $(k+1)$-row by m-column incidence matrix

$$M = (e(i, j)) \begin{cases} 0 \leqq i \leqq k \\ 0 \leqq j \leqq m-1 \end{cases}$$

be specified where $e(i, j) = 1$ or 0 depending upon whether the j^{th} derivative at x_i is specified or not. It will be assumed that

$$\sum_{j=0}^{m-1} e(i, j) \geqq 1 \quad \text{for each } i, \qquad 0 < i < k,$$

i.e., at least one derivative of order between 0 and $m-1$ is specified at each $x_i \in (a, b)$.

* Present address, Department of Mathematics, Northwestern University, Evanston, Illinois 60201, U.S.A.

We will be concerned with determining solutions of the class of multipoint boundary value problems of the form

$$
\begin{aligned}
&\text{(a)} \quad L^* L f(x) = g(x), \qquad x_i < x < x_{i+1}, \qquad 0 \leqq i \leqq k-1 \\
&\text{(b)} \quad D^j f(x_i) = r_{ij} \quad \text{if} \quad e(i,j) = 1 \\
&\text{(c)} \quad f \in AC^{m-1}[a,b] \cap C^{2m}\big(\cup (x_i, x_{i+1})\big) \\
&\text{(d)} \quad O_j L f(x_i+) = O_j L f(x_i-) \quad \text{if} \quad e(i,j) = 0, \qquad 0 \leqq i \leqq k
\end{aligned} \tag{1.1}
$$

where $O_j L f(x_0-) = O_j L f(x_k+) = 0$ by definition, where g is any integrable function on (a, b) continuous on each subinterval (x_i, x_{i+1}) and O_j is the differential operator

$$O_j v = \sum_{\mu=0}^{m-j-1} (-1)^\mu D^\mu \{a_{j+\mu+1} v\}$$

which arises in the well-known integration by parts formula

$$\int_\alpha^\beta (v L u - u L^* v) = \left\{ \sum_{j=0}^{m-1} u^{(j)}(\beta) O_j v(\beta) - \sum_{j=0}^{m-1} u^{(j)}(\alpha) O_j v(\alpha) \right\}.$$

The terminology $AC^{m-1}[a, b]$ means the class of functions f on $[a, b]$ with $f^{(m-1)}$ absolutely continous on $[a, b]$. Throughout § 1—5 we will deal exclusively with real-valued functions.

Perhaps the simplest nontrivial physical example of such a multi-point boundary value problem is the determination of the configuration of a thin elastic beam under the influence of a load $g = g(x)$ which may have discontinuities at points x_i where the beam or the angular attitude of the beam are constrained to have a certain position or orientation. In this case the operator L is of the form

$$L = \sqrt{p(x)}\, \frac{d^2}{dx^2}$$

and L^* is given by

$$L^* = \frac{d^2 (\sqrt{p(x)})}{dx^2}$$

for some twice continuously differentiable, positive function $p(x)$ which describes the cross-section of the beam. The problem considered here is that of determining a continuously differentiable function $u(x)$ satisfying conditions on $u(x_i)$ or $u'(x_i)$ for certain points $0 = x_0 < x_1 < \cdots < x_k = l$, such that $L^* L u(x) = g(x)$ for $x_i < x < x_{i+1}$ and such that u'' is continuous at x_i if $u'(x_i)$ is not specified and u''' is continuous at x_i if $u(x_i)$ is not specified.

In § 2 the case of the homogeneous equation is considered, i.e., the case $g \equiv 0$ in (1.1 a). In § 3 the case of homogeneous boundary conditions is considered, i.e., $r_{ij} = 0$ in (1.1 b). In this case a Green's function is obtained for the problem under an assumption on the incidence matrix M which guarantees uniqueness of solutions of (1.1). Thus, a solution of (1.1) is obtained by adding, respectively, a solution of the problem of § 2 and a solution of the problem of § 3. In § 4, an approximation scheme is exhibited for determining the solution f of the problem of § 3. The approximation scheme exhibited is optimal in a sense of asymptotic rate of convergence to be defined in § 5 where it is also established that the mean

square rate of convergence of the approximation scheme is of the order $O(n^{-m})$. Finally, in §6, various generalizations are indicated which widen the scope of the theory treated in §1 through §5.

§ 2. The Homogeneous Equation

Solutions of (1.1) are known to exist under any choice of the incidence matrix M when $g \equiv 0$ in (1.1 a). (See [8].) Such solutions are known as Lg-splines interpolating Hermite-Birkhoff data. They may be alternately characterized as solutions of the minimization problem

$$\min \int_a^b (Lf)^2, \quad D^j f(x_i) = r_{ij} \quad \text{if} \quad e(i,j) = 1 \tag{2.1}$$

over the class $\mathscr{H}_2^m(a, b)$ of f such that $f^{(m-1)}$ is absolutely continuous on (a, b) and $f^{(m)}$ is square integrable on (a, b). The solution of (1.1) with the homogeneous equation or, equivalently, of (2.1) is unique if and only if the matrix M and the differential operator L satisfy the condition

$$\{Lu \equiv 0,\ D^j u(x_i) = 0 \text{ if } e(i,j) = 1\} \Rightarrow u \equiv 0. \tag{2.2}$$

In this case there exist algorithms (see [8]) for the computation of the unique solution of the homogeneous equation.

Let $\xi \in (a, b) - \{x_i\}_1^{k-1}$. We denote by (1.1′) the problem which includes the specification of a function value r_ξ at ξ.

Theorem 2.1. Let (2.2) be satisfied and let s be the unique Lg-spline solution of (1.1′) with $g \equiv 0$. Then, for $0 \leqq j < m$, $D^j s$ is a uniformly continuous function of the argument $(x; \xi, r_\xi)$.

Proof. Designate by $\{\lambda_\nu\}_1^n$ the linear functionals $D^j(\cdot)(x_i)$ and $(\cdot)(\xi)$ and by $\{r_\nu\}_1^n$ the corresponding numbers $\{r_{ij}\}$ and r_ξ, i.e., $\sum_{i,j} e(i,j) = n - 1$. By Theorem 4.6 of [8] the functions

$$f_j(x) = \sum_{\nu=1}^n \beta_{j\nu} \lambda_\nu \hat{\theta}(\cdot, x), \qquad 1 \leqq j \leqq n - m \tag{2.3}$$

form a basis for the $n - m$ dimensional linear space $\{Lf : f \text{ solves (1.1′) with } g \equiv 0$ for some $\{r_\nu\}_1^n\}$, where $\beta_j = \{\beta_{j\nu}\}_{\nu=1}^n$, $1 \leqq j \leqq n - m$, are appropriate vectors and $\hat{\theta}(y, x) = \theta(y, x)$ if $y \geqq x$ and is 0 otherwise. Here $\theta(\cdot, x)$ is that element of the null space N_L of L, defined, for each fixed x, by the formula

$$D^j \theta(\cdot, x)]_x = \delta_{j, m-1}, \qquad 0 \leqq j \leqq m - 1.$$

The dot notation here represents the variable with respect to which the operations λ_ν and D^j are taken. Also, if $n \leqq m$, then $Ls \equiv 0$. Now it follows from the preceding remarks that for a fixed vector $r = \{r_\nu\}_1^n$, the Lg-spline s has the representation, for $0 \leqq j < m$,

$$s^j(x) = \sum_{\nu=1}^{n-m} C_\nu D_x^j \big(\hat{\theta}(x, y), f_\nu(y)\big)_{\mathscr{L}_2} + \sum_{\mu=1}^m q_\mu D^j u_\mu(x), \tag{2.4}$$

where $\{u_\mu\}_1^m$ span N_L, where the C_ν are uniquely determined by the system of equations

$$\sum_{\nu=1}^{n-m} C_\nu (f_\nu, f_j)_{\mathscr{L}_2} = (\beta_j, r)_{l_2}, \qquad 1 \leqq j \leqq n-m \tag{2.5}$$

and where the $\{q_\mu\}_1^m$ are uniquely determined by the system

$$\sum_{\mu=1}^{m} q_\mu \tilde{\lambda}_j u_\mu = \tilde{r}_j - \sum_{\nu=1}^{n-m} C_\nu \tilde{\lambda}_j \big(\hat{\theta}(x, y), f_\nu(y)\big)_{\mathscr{L}_2}, \qquad 1 \leqq j \leqq m. \tag{2.6}$$

The $\{\tilde{\lambda}_j\}_1^m$ are any subset of $\{\lambda_j\}_1^{n-1}$ which are linearly independent over N_L. The existence of such a subset is implied by (2.2).

Now the key to the proof is the observation that the vectors $\{\beta_j\}_1^{n-m}$ may be chosen for each set $\{\lambda_j\}_1^n$ in such a way that the components vary uniformly with (ξ, r_ξ). Indeed β_j may be taken to be fixed for $1 \leqq j \leqq n-m-1$ and β_{n-m} may be chosen as the unique solution, with $\beta_{n-m,n} = 1$, of

$$\sum_{\nu=1}^{n} \beta_{n-m,\nu} \lambda_\nu u_\mu = 0, \qquad 1 \leqq \mu \leqq m$$

with $\beta_{n-m,\nu} = 0$ if $\lambda_\nu \notin \{\tilde{\lambda}_j\}_1^m$. We refer the reader to §4 of [8] for the explicit construction of $\{\beta_j\}_1^n$. With this fact, (2.5) then implies that the coefficients $\{C_\nu\}_1^{n-m}$ vary uniformly with (ξ, r_ξ) since the functions f_j vary uniformly with $(x; \xi, r_\xi)$. Since the functions $\lambda_j\big(\hat{\theta}(y, x) f_\nu(x)\big)_{\mathscr{L}_2}$ also vary uniformly with (ξ, r_ξ), the coefficients q_μ vary uniformly with (ξ, r_ξ), by (2.6). These remarks, and the representation (2.4), imply the assertion of the theorem.

§ 3. Homogeneous Boundary Conditions

A solution to problem (1.1) with $r_{ij} = 0$ in (1.1 b) is obtained by integrating g against an appropriate Green's function:

$$f(x) = \int_a^b G(x, \xi)\, g(\xi)\, d\xi.$$

We will assume throughout this section that (2.2) is satisfied. This will enable us to construct a uniquely defined Green's function for the problem with homogeneous boundary conditions. In this case G is a real-valued function of two variables (x, ξ) where $x \in [a, b]$ and $\xi \in [a, b] - \{x_i\}_{i=0}^k$ satisfying, for each fixed ξ,

(3.1)

(a) $(L^* L)_x G(x, \xi) = 0, \qquad x \in [a, b] - \{x_i\}_0^k - \{\xi\}$

(b) $\begin{cases} D_x^j G(x_i, \xi) = 0 & \text{if} \quad e(i, j) = 1 \\ G(\xi, \xi) = c \neq 0 & \text{where} \quad c = c(\xi) \text{ is uniformly continuous} \end{cases}$

(c) $G(\cdot, \xi) \in A\, C^{m-1}[a, b] \cap C^{2m}([a, b] - \{x_i\}_0^k - \{\xi\})$

(d) $D_x^j G(x, \xi)$ is uniformly continuous on $[a, b] \times ([a, b] - \{x_i\}_0^k)$, $0 \leqq j \leqq m-1$, and bounded for $0 \leqq j \leqq 2m-1$,

(e) $\begin{cases} O_j L_x G(x_i+, \xi) = O_j L_x G(x_i-, \xi) & \text{if} \quad e(i, j) = 0 \\ G(\cdot, \xi) \quad \text{is} \quad C^{2m-2} \quad \text{in a nbd. of } \xi. \end{cases}$

Theorem 3.1. There exists a unique real-valued function $G=G(x,\xi)$ defined on $[a,b]\times([a,b]-\{x_i\}_0^k)$ satisfying (3.1 a, b, c, d, e). The nonzero function $c=c(\xi)$ may be chosen so that the identity

$$(-1)^m a_m^2(\xi)\,[G^{(2m-1)}(\xi+,\xi)-G^{(2m-1)}(\xi-,\xi)]=1 \tag{3.2}$$

holds for each $\xi\in[a,b]-\{x_i\}_0^k$. Furthermore, $G(x,\xi)=G(\xi,x)$ if $(x,\xi)\in([a,b]-\{x_i\}_0^k)\times([a,b]-\{x_i\}_0^k)$ and G may be extended to a function defined and symmetric on $[a,b]\times[a,b]$, such that $G\in C^{m-1}([a,b]\times[a,b])$.

Proof. For any prescribed ξ and $c=c(\xi)$ there exists a unique Lg-spline $G_\xi(x)=G(x,\xi)$ with knots at $\{x_i\}_0^k\cup\{\xi\}$ satisfying (3.1 a, b, c, d, e). Property (3.1 d) follows from Theorem 2.1. Now let $\Gamma=\Gamma(x,\xi)$ satisfy (3.1 a, b, c, d, e) with $c(\xi)\equiv 1$. For each $\xi\in[a,b]-\{x_i\}_0^k$ we have

$$\int_a^b\{L_x\Gamma(x,\xi)\}^2\,dx>0.$$

Indeed, if $\int_a^b\{L_x\Gamma(x,\xi)\}^2dx=0$ for some $\xi\in[a,b]-\{x_i\}_0^k$ then $L_x\Gamma(x,\xi)=0$ a.e. and since $L_x\Gamma(x,\xi)$ is continuous on $[a,b]-\{x_i\}_0^k$ as a function of x it follows that $L_x\Gamma(x,\xi)=0$ if $x\neq x_i$, $i=0,1,\ldots,k$. Since $\Gamma(\cdot,\xi)\in C^{m-1}[a,b]$ it follows that the m^{th} derivative of $\Gamma(\cdot,\xi)$ exists and is equal to zero at each x_i, $i=0,1,\ldots,k$ and hence $L_x\Gamma(x,\xi)=0$ for all $x\in[a,b]$. By (2.2) it follows that $\Gamma(x,\xi)=0$ for all $x\in[a,b]$, contradicting $\Gamma(\xi,\xi)=1$.

In order to obtain a solution of (3.1) which satisfies (3.2) we define

$$G(x,\xi)=\frac{\Gamma(x,\xi)}{\int_a^b\{L_y\Gamma(y,\xi)\}^2dy}. \tag{3.3}$$

Notice that G satisfies (3.1) with

$$G(\xi,\xi)=1/\int_a^b\{L_y\Gamma(y,\xi)\}^2>0.$$

Integrating the expression

$$\int_a^b\{L_yG(y,\xi)\}^2dy$$

by parts we conclude that

$$(-1)^m a_m^2(\xi)\,[G^{(2m-1)}(\xi+,\xi)-G^{(2m-1)}(\xi-,\xi)]=\frac{\int_a^b\{L_yG(y,\xi)\}^2dy}{G(\xi,\xi)}=1$$

by (3.3).

We now prove that

$$G(x,\xi)=G(\xi,x)\quad\text{if}\quad(x,\xi)\in([a,b]-\{x_i\}_0^k)\times([a,b]-\{x_i\}_0^k). \tag{3.4}$$

Indeed, it suffices to show that, for $\xi < x$, (3.4) holds. Now if $a < \xi < x < b$,

$$\begin{aligned}
0 &= \int_a^b \{G(t, x)\, L^* L_t G(t, \xi) - G(t, \xi)\, L^* L_t G(t, x)\}\, dt \\
&= \int_a^{\xi} \{\ \}\, dt + \int_{\xi}^{x} \{\ \}\, dt + \int_x^b \{\ \}\, dt \\
&= (-1)^m a_m^2(\xi)\, [G^{(2m-1)}(\xi+, \xi) - G^{(2m-1)}(\xi-, \xi)]\, G(\xi, x) \\
&\quad - (-1)^m a_m^2(x)\, [G^{(2m-1)}(x+, x) - G^{(2m-1)}(x-, x)]\, G(x, \xi) \\
&\quad + \int_a^b L_t G(t, x)\, L_t G(t, \xi)\, dt - \int_a^b L_t G(t, \xi)\, L_t G(t, x)\, dt \\
&= G(\xi, x) - G(x, \xi).
\end{aligned}$$

We now define $G(x, \xi) = G(\xi, x)$ if ξ is a knot and x is not. Thus far, then, we have defined G on the set

$\mathcal{U} = \{(x, y) \in [a, b] \times [a, b] : (x, y) \neq (x_i, x_j)$ for some $0 \leqq i \leqq k,\ 0 \leqq j \leqq k\}$ and established that G is symmetric on $\mathcal{U}$. Now by construction, $G(\cdot, \xi) \in C^{m-1}[a, b]$ for each fixed $\xi \notin \{x_i\}_0^k$ and by (3.1 d) $D_x^j G(x, \xi)$ is uniformly continuous in $(x, \xi) \in [a, b] \times ([a, b] - \{x_i\}_0^k)$ for each $j = 0, 1, \ldots, m-1$. Hence, by the symmetry of G on $\mathcal{U}$, it follows $G \in C^{m-1}(\mathcal{U})$ and that each partial derivative of order $\leqq m-1$ is uniformly continuous on $\mathcal{U}$. It follows that G and each partial derivative of G of order $\leqq m-1$ can be extended continuously to the closure of $\mathcal{U}$, i.e., to $[a, b] \times [a, b]$. Thus, $G \in C^{m-1}([a, b] \times [a, b])$ and is symmetric on $[a, b] \times [a, b]$ since G is symmetric on $\mathcal{U}$ and each value $G(x_i, x_j)$ is the uniform limit of $G(x, y)$ as $(x, y) \to (x_i, x_j)$ through values of $\mathcal{U}$. This concludes the proof of the theorem.

We now formally define the Green's function for the multipoint boundary value problem with homogeneous boundary conditions as the unique solution of (3.1) satisfying (3.2). We now prove that G actually defines an integral operator which will yield a solution of the boundary value problem with homogeneous boundary conditions.

Theorem 3.2. For each piecewise continuous, integrable function g on $[a, b]$, with discontinuities restricted to the points $\{x_i\}_1^{k-1}$, the function

$$f(x) = \int_a^b G(x, \xi)\, g(\xi)\, d\xi \tag{3.5}$$

satisfies the multi-point boundary value problem (1.1) with $r_{ij} = 0$ if $e(i, j) = 1$. Moreover, if f satisfies (1.1 b, c, d) with $r_{ij} = 0$ if $e(i, j) = 1$ then

$$f(x) = \int_a^b G(x, \xi)\, L^* L f(\xi)\, d\xi. \tag{3.5'}$$

Proof. If $x_i < x < x_{i+1}$ then by direct computation it follows that $f^{(2m)}(x)$ exists and

$$L^* L f(x) = g(x)\, \{(-1)^m a_m^2(x)\, [G^{(2m-1)}(x+, x) - G^{(2m-1)}(x-, x)]\}$$

where f is defined by (3.5). Since G satisfies (3.2) it follows that f satisfies (1.1 a). By (3.1 b) it follows that f satisfies (1.1 b) with $r_{ij} = 0$ if $e(i, j) = 1$. The absolute

continuity of $f^{(m-1)}$ on $[a, b]$ follows from the relations

$$\begin{aligned}\int_a^x f^{(m)}(t)\,dt &= \int_a^x \left\{\int_a^b D_t^m G(t,\xi)\,g(\xi)\,d\xi\right\} dt \\ &= \int_a^b \left\{\int_a^x D^m G(t,\xi)\,dt\right\} g(\xi)\,d\xi \\ &= \int_a^b D^{m-1}G(x,\xi)\,g(\xi)\,d\xi - \int_a^b D^{m-1}G(a,\xi)\,g(\xi)\,d\xi \\ &= f^{(m-1)}(x) - f^{(m-1)}(a).\end{aligned}$$

To see that $f \in C^{2m}\left(\bigcup (x_i, x_{i+1})\right)$ we observe that $L^*Lf = g \in C\left(\bigcup (x_i, x_{i+1})\right)$ and

$$f^{(2m)}(x) = L^*Lf(x)/(-1)^m a_m^2(x) + \text{lower order terms continuous in } x.$$

Thus f satisfies (1.1 c). Finally, to see that f satisfies (1.1 d) we compute

$$\begin{aligned}O_j L f(x_i+) - O_j L f(x_i-) &= \lim_{x\to x_i+} \int_a^b O_j L_x G(x,\xi)\,g(\xi)\,d\xi - \lim_{x\to x_i-} \int_a^b O_j L_x G(x,\xi)\,g(\xi)\,d\xi \\ &= \int_a^b \{O_j L_x G(x_i+,\xi) - O_j L_x G(x_i-,\xi)\}\,g(\xi)\,d\xi\end{aligned}$$

by (3.1 d) and Lebesgue's dominated convergence theorem and the bracketed expression is zero by (3.1 e). The proof of the theorem will be completed if we can verify (3.5′). Let

$$h(x) = \int_a^b G(x,\xi)\,L^*Lf(\xi)\,d\xi$$

and set $d(x) = f(x) - h(x)$. Then $d(x)$ satisfies

$$\begin{aligned}&\text{(a)} \quad L^*Ld(x) = 0 \quad \text{if} \quad x_i < x < x_{i+1} \\ &\text{(b)} \quad D^j d(x_i) = 0 \quad \text{if} \quad e(i,j) = 1 \\ &\text{(c)} \quad d \in A\,C^{m-1}[a,b] \cap C^{2m}\left(\cup (x_i, x_{i+1})\right) \\ &\text{(d)} \quad O_j L d(x_i+) = O_j L d(x_i-), \quad 0 \leqq i \leqq k\end{aligned} \tag{3.6}$$

and upon integrating the expression L^*Ld by parts we obtain, using (3.6),

$$0 = \int_a^b L^*Ld(x)\,dx = \int_a^b \left(Ld(x)\right)^2 dx$$

and it follows that $Ld(x) = 0$ a.e. Since $d \in C^m\left(\bigcup_i (x_i, x_{i+1})\right)$ we conclude that $Ld(x) = 0$ for $x_i < x < x_{i+1}$, i.e., $d \in N_L$ on each (x_i, x_{i+1}), $i = 0, 1, \ldots, k-1$. We assert that $d \in N_L$ on $[a, b]$. Indeed, fix $0 < i < k$ and suppose $d \equiv u$ on (x_{i-1}, x_i) and $d \equiv v$ on (x_i, x_{i+1}). Then, since $d \in A\,C^{m-1}[a, b]$ and, in particular, $\in C^{m-1}[a, b]$, we have

$$u^{(j)}(x_i) = v^{(j)}(x_i) \quad j = 0, 1, \ldots, m-1. \tag{3.7}$$

Since N_L is of dimension m, the m conditions (3.7) imply by the non-vanishing, of the Wronskian determinant, that $u \equiv v$. It follows that $d \in N_L$ on $[a, b]$ and,

since d satisfies (3.6b), condition (2.2) implies that $d \equiv 0$. This completes the proof of the theorem.

The final property of $G(x, \xi)$ needed is that it determine a positive operator on $\mathscr{L}_2(a, b)$. Indeed, we have,

Theorem 3.3. For every $g \in \mathscr{L}_2(a, b)$ we have

$$\int_a^b \left\{ \int_a^b G(x, t)\, g(t)\, dt \right\} g(x)\, dx \geqq 0. \tag{3.8}$$

Proof. Slight modifications of the arguments of Theorem 3.2 show that the operator defined by

$$u(x) = \int_a^b G(x, \xi)\, g(\xi)\, d\xi \tag{3.9}$$

maps $\mathscr{L}_2(a, b)$ one to one onto the class $\mathscr{S}$ of real-valued functions u satisfying

$$\begin{aligned} &\text{(a)} \quad D^j u(x_i) = 0 \quad \text{if} \quad e(i, j) = 1 \\ &\text{(b)} \quad u \in AC^{m-1}[a, b] \cap AC^{2m-1}\left(\bigcup (x_i, x_{i+1})\right) \\ &\text{(c)} \quad O_j L u(x_i+) = O_j L u(x_i-) \quad \text{if} \quad e(i, j) = 0 \\ &\text{(d)} \quad L^* L u \in \mathscr{L}_2(a, b) \end{aligned} \tag{3.10}$$

where $AC^{2m-1}\left(\bigcup (x_i, x_{i+1})\right)$ designates the class of functions f with $f^{(2m-1)}$ absolutely continuous on each (x_i, x_{i+1}). It is easily seen that L^*L is the inverse of the operator (3.9). Now, if $g \in \mathscr{L}_2(a, b)$ we have,

$$\int_a^b \left\{ \int_a^b G(x, t)\, g(t)\, dt \right\} g(x)\, dx = \int_a^b u(x)\, L^* L u(x)\, dx = \int_a^b \left(L u(x)\right)^2 dx > 0$$

if $g \neq 0$. This completes the proof of the theorem.

§ 4. Approximation Schemes

In this section we will assume that g is square integrable on (a, b), a stronger assumption than integrability, and that (2.2) is satisfied. We will develop approximation schemes for the solution of the problem with homogeneous boundary conditions. The approximations will be expressed in terms of the eigenfunctions of the following eigenvalue problem:

$$\begin{aligned} &\text{(a)} \quad L^* L u(x) = \lambda u(x), \qquad x_i < x < x_{i+1} \\ &\text{(b)} \quad D^j u(x_i) = 0 \qquad \text{if} \quad e(i, j) = 1 \\ &\text{(c)} \quad u \in AC^{m-1}[a, b] \cap AC^{2m-1}\left(\bigcup (x_i, x_{i+1})\right) \\ &\text{(d)} \quad O_j L u(x_i+) = O_j L u(x_i-) \quad \text{if} \quad e(i, j) = 0. \end{aligned} \tag{4.1}$$

The first result is

Theorem 4.1. The eigenvalue problem (4.1) has positive eigenvalues $0 < \lambda_1 \leqq \lambda_2 \leqq \cdots$ of finite multiplicity tending to infinity. The associated eigenfunctions form a complete orthonormal system in $\mathscr{L}_2(a, b)$.

Proof. Define $\mathscr{S}$ as the class of real-valued functions u satisfying (3.10). By the proof of Theorem 3.3 $G(x, \xi)$ determines an operator G defined by (3.9) and maps $\mathscr{L}_2(a, b)$ one to one onto $\mathscr{S}$ with inverse L^*L. Now G is compact since $G(x, \xi)$ is square integrable on $[a, b]\times[a, b]$ and, by Theorem 3.3, G is positive. Hence the eigenvalues of G consist of a sequence of positive eigenvalues $\mu_1 \geqq \mu_2 \geqq \cdots > 0$ of finite multiplicity tending to zero and G has a complete orthonormal system of eigenfunctions $\varphi_1, \varphi_2, \ldots \subset \mathscr{S}$. The reciprocals $\lambda_j = \mu_j^{-1}$ are the eigenvalues of L^*L and the φ_j are the associated eigenfunctions. This completes the proof.

The sequence of approximants Ψ_n in $\mathscr{S}$ to the solution f of the problem with homogeneous boundary conditions is defined by

$$\Psi_n = \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu} \left\{ \int_a^b g(\xi)\, \varphi_\nu(\xi)\, d\xi \right\} \varphi_\nu. \tag{4.2}$$

In the next section we will investigate the convergence of Ψ_n to f in a variety of norms and will give the asymptotic rate of convergence in some cases. In particular we will show that Ψ_n converges to f in the well known Sobolev space $\mathscr{H}_2^m(a, b)$, with norm

$$\|u\|^2_{\mathscr{H}_2^m} = \sum_{j=0}^{m} \int_a^b (u^{(j)})^2,$$

which implies uniform convergence, of the derivatives of Ψ_n through order $m-1$, to the corresponding derivatives of f. Moreover, we will show that

$$\|f - \Psi_n\|_{\mathscr{H}_2^j} = O(n^{-(m-j)}), \qquad 0 \leqq j < m$$

which implies for $j = 0$ that Ψ_n converges to f in the mean square sense with order $O(n^{-m})$.

The important property of the approximation scheme (4.2) is that it is optimal in the following sense. If it is known that the solution f satisfies

$$\int_a^b (Lf)^2 \leqq \varrho^2 \tag{4.3}$$

then the collection $\mathscr{R}$ of all $u \in \mathscr{S}$ satisfying (4.3) contains f and the scheme (4.2) is the optimal scheme for the mean square approximation of the elements $u \in \mathscr{R}$, where g in (4.2) is given by L^*Lu. The precise meaning of optimality will become clearer in the following section.

Finally, we close this section by showing that an estimate for $\int_a^b g^2(\xi)\, d\xi$ leads to an estimate for $\int_a^b (Lf(\xi))^2 d\xi$ and hence an upper bound ϱ as discussed above. Indeed, by (3.5) and the Schwarz inequality,

$$\int_a^b (Lf(\xi))^2 d\xi \leqq \left\{ \sup_{x\in[a,b]} \int_a^b (L_x G(x, \xi))^2 d\xi \right\} \left(\int_a^b g^2(\xi)\, d\xi \right) (b-a) \tag{4.4}$$

where the quantity $\int_a^b (L_x G(x, \xi))^2 d\xi = G(x, x)$ is bounded in x since $G(x, x)$ is continuous on $[a, b]\times[a, b]$.

§ 5. Convergence

In order to investigate the convergence of the approximation scheme defined by the sequence $\{\Psi_n\}$ in §4 we find it convenient to introduce the notion of n-width first defined by Kolmogorov as a means of characterizing the approximability of sets $\mathscr{A}$ in normed linear spaces $\mathscr{X}$ by finite dimensional subspaces of dimension n. Indeed, to obtain the n-width, $d_n(\mathscr{A})$ of $\mathscr{A}$ in $\mathscr{X}$, we first calculate the dispersion

$$E(\mathscr{A},\mathscr{M}) = \sup_{f\in\mathscr{A}} \inf_{g\in\mathscr{M}} \|f-g\|_{\mathscr{X}}$$

of $\mathscr{A}$ from a fixed subspace $\mathscr{M}$ of dimension n and then take the infimum

$$d_n(\mathscr{A}) = \inf\{E(\mathscr{A},\mathscr{M}):\ \mathscr{M}\subset\mathscr{X},\ \dim\mathscr{M}=n\}$$

over all such subspaces in $\mathscr{X}$. $\mathscr{M}$ is called an optimal approximating subspace if $\dim\mathscr{M}\leqq n$ and $d_n(\mathscr{A}) = E(\mathscr{A},\mathscr{M})$. Specializing $\mathscr{A}$ to the class

$$\mathscr{R}_m = \{f\in\mathscr{H}_2^m:\ \|Lf\|_{\mathscr{L}_2}\leqq 1\}$$

and $\mathscr{X}$ to $\mathscr{L}_2(a,b)$ we consider the eigenvalue problem

$$\text{(5.1)}\qquad \begin{aligned} &\text{(a)}\quad L^*Lu(x) = \lambda u(x), && a<x<b\\ &\text{(b)}\quad (Lu)^{(j)}(a) = (Lu)^{(j)}(b) = 0, && 0\leqq j\leqq m-1. \end{aligned}$$

It is known that the eigenvalues of (5.1) are nonnegative, and converge to infinity. We have the following

Theorem 5.1 (Lorentz) [9]. If $\{\lambda_j\}_{j=1}^{\infty}$ are the eigenvalues of (5.1) counted according to multiplicity and arranged in increasing order then

$$d_n(\mathscr{R}_m) = 1/\sqrt{\lambda_{n+1}},\qquad n=0,1,\ldots.$$

In particular, $d_n(\mathscr{R}_k)=\infty$ for $n=0,1,\ldots,m-1$ and

$$d_n(\mathscr{R}_m) = \left(\frac{b-a}{\pi}\right)n^{-m}[1+O(n^{-1})],\qquad n\geqq m.$$

Finally, optimal approximating subspaces of dimension n are spanned by the first n eigenfunctions of (5.1).

Lorentz's result can be generalized to the setting of interest to us in this paper. Specifically, we define

$$\mathscr{S}_m = \{f\in\mathscr{H}_2^m(a,b):\ \|Lf\|_{\mathscr{L}_2}\leqq 1,\ D^if(x_i)=0 \text{ if } e(i,j)=1\}$$

and we notice that $\mathscr{S}_m$ reduces to $\mathscr{R}_m$ if $k=1$, $a=x_0$, $b=x_1$, and $e(i,j)=0$ for all i and j. Consider again the eigenvalue problem (4.1), which reduces to (5.1) if $k=1$, $a=x_0$, $b=x_1$, and $e(i,j)=0$ for all i and j. With the assumption that $\lambda=0$ is not an eigenvalue, i.e., (2.2), the eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ are of finite multiplicity, positive and converge to infinity, which follows from Theorem 4.1. We state now

Theorem 5.2. The $\mathscr{L}_2$ n-widths of $\mathscr{S}_m$ satisfy

$$d_n(\mathscr{S}_m) = 1/\sqrt{\lambda_{n+1}}, \quad n = 0, 1, \ldots, \tag{5.2}$$

$$d_n(\mathscr{S}_m) = (C/n)^{-m}\big(1 + o(1)\big) \quad \text{as} \quad n \to \infty. \tag{5.3}$$

where the constant C in (5.3) is computable. Moreover, optimal approximating subspaces of dimension n are spanned by the first n eigenfunctions of (4.2).

Proof. Let $\mathscr{H}_0$ denote the functions f in $\mathscr{H}_2^m(a, b)$ satisfying $D^j f(x_i) = 0$ if $e(i, j) = 1$. Throughout the proof we will use, as an inner product on $\mathscr{H}_0$,

$$(f, g)_{\mathscr{H}_0} = \int_a^b f g + \int_a^b L f \, L g. \tag{5.4}$$

By Theorem 4.1, $\mathscr{H}_0$ is dense in $\mathscr{L}_2(a, b)$, and, moreover, $\mathscr{H}_0$ is a Hilbert space in its own right with inner product (5.4). By Theorem 3.2 of [6] there exists a positive self adjoint operator R on $\mathscr{H}_0$ satisfying

$$\int_a^b L f \, L g = \int_a^b R f \cdot g, \quad f \in \mathscr{D}_R, \quad g \in \mathscr{H}_0 \tag{5.5}$$

and by Theorem 3.1 of [6] the set $\mathscr{R}_m = \{f \in \mathscr{D}_R : \int_a^b R f \cdot f \leqq 1\}$ has the same closure and hence the same n-widths as $\mathscr{S}_m$. But the n-widths of $\mathscr{R}_m$ satisfy (5.2). Indeed, the operator R has the same eigenvalues and eigenfunctions as L^*L as is seen from (5.5). Hence,

$$R = \sum_{\nu=1}^{\infty} \lambda_\nu P_\nu,$$

where P_ν represents the projector onto the one-dimensional space spanned by φ_ν. Theorem 2.1 of [6] then implies that $d_n(\mathscr{R}_m)$ satisfy (5.2). To verify (5.3) we first observe that the eigenvalues λ_ν of L^*L have the asymptotic distribution

$$N(\lambda) = C\,\lambda^{+2m}\big(1 + o(1)\big) \quad \text{as} \quad \lambda \to \infty \tag{5.6}$$

for the number of eigenvalues $\leqq \lambda$ with $C = (2\pi)^{-1} \int_a^b \text{measure}\,\{\xi : 0 < a_m^2(x)\,\xi < 1\} dx$. This follows directly from Theorem 14.6 of [1] with $\Omega = \bigcup_i (x_i, x_{i+1})$, $m' = 2m$, $A = L^*L$. The self-adjoint extension of L^*L, as required in Theorem 14.6, can be established easily from Friedrich's extension theorem [10, Sec. 124].

We will show now that (5.6) immediately implies (5.3). Indeed, let $\{\mu_\nu\}_{\nu \geqq 1}$ denote the distinct eigenvalues of L^*L and let N_ν denote the corresponding multiplicities. Then,

$$\overline{\lim}\, d_n(\mathscr{S}_m)\, n^m \leqq \overline{\lim}\, \mu_{\nu+1}^{-\frac{1}{2}} N_{\nu+1}^m = C,$$

i.e., $d_n(\mathscr{S}_m) \leqq (C/n)^{-m}\big(1 + o(1)\big)$ as $n \to \infty$. To see that C is sharp, we remark first that Lemma 3.2 of [7] implies that $N_{\nu+1} = N_\nu\big(1 + o(1)\big)$. With this remark,

we have,

$$\begin{aligned}\underline{\lim}\, d_n(\mathscr{S}_m)\, n^m &\geqq \underline{\lim}\, \mu_{\nu+1}^{-\frac{1}{2}} N_\nu^m \\ &= \underline{\lim}\, (N_\nu/N_{\nu+1})^m N_{\nu+1}^m \mu_{\nu+1}^{-\frac{1}{2}} \\ &= \underline{\lim}\, N_{\nu+1}^m \mu_{\nu+1}^{-\frac{1}{2}} \\ &= C,\end{aligned}$$

i.e., $d_n(\mathscr{S}_m) \geqq (C/n)^{-m}\big(1+o(1)\big)$ as $n\to\infty$. This concludes the proof of Theorem 5.2.

The next theorem provides convergence results in norms other than the $\mathscr{L}_2$-norm.

Theorem 5.3. The approximations Ψ_n defined by (4.2) converge in $\mathscr{H}_2^m(a,b)$ to the solution f of the boundary value problem (1.1) with $r_{ij}=0$. More generally,

$$\|f-\Psi_n\|_{\mathscr{H}_2^j} = o(n^{-(m-j)}) \quad \text{as} \quad n\to\infty, \quad 0<j\leqq m, \quad \|f-\Psi_n\|_{\mathscr{L}_2} O(n^{-m}). \tag{5.7}$$

Proof. To show that $\|f-\Psi_n\|_{\mathscr{H}_2^m}\to 0$ as $n\to\infty$ we consider the equivalent norm

$$|u|^2_{\mathscr{H}_2^j} = \sum_{e(i,j)=1} \big(D^i u(x_i)\big)^2 + \int_a^b (Lu)^2. \tag{5.8}$$

That (5.8) defines a true norm follows from (2.2). The inequality $|u|^2_{\mathscr{H}_2^m} \leqq C_1 \|u\|^2_{\mathscr{H}_2^m}$ for some positive constant C_1 independent of u in $\mathscr{H}_2^m(a,b)$ follows from Sobolev's inequality, the inequality

$$\Big(\sum_{j=0}^m \eta_j\Big)^2 \leqq 2^m \sum_{j=0}^m \eta_j^2$$

and the boundedness of the coefficients of L.

The inequality $\|u\|^2_{\mathscr{H}_2^m} \leqq C_2 |u|^2_{\mathscr{H}_2^m}$ follows from the representation

$$u(x) = \varphi(x) + \int_a^x \theta(x,\xi)\, Lu(\xi)\, d\xi$$

where $\varphi \in N_L$ and $\theta(\cdot,\xi)\in N_L$ satisfies $D^j\theta(\cdot,\xi)|_{x=\xi} = \delta_{j,m-1}$, $j=0,1,\ldots,m-1$, and from the equivalence of any two norms on the finite dimensional space N_L.

We will show now that $|f-\Psi_n|_{\mathscr{H}_2^m}\to 0$ as $n\to\infty$. Since the subspace $\mathscr{H}_0$ of $\mathscr{H}_2^m(a,b)$, consisting of all $h\in\mathscr{H}_2^m(a,b)$ such that $D^i h(x_i)=0$ if $e(i,j)=1$, is a closed subsapce of $\mathscr{H}_2^m(a,b)$, and since f and $\{\varphi_\nu\}_{\nu\geqq 1}$ are contained in $\mathscr{H}_0$, it suffices to show that $\{\varphi_\nu\}_{\nu\geqq 1}$ is a complete system in $\mathscr{H}_0$, orthogonal with respect to the inner product determined by (5.8), and that the Ψ_n are the Fourier expansions of f along the $\{\varphi_\nu\}_{\nu\geqq 1}$. But, if $\nu\neq\mu$,

$$\sum_{e(i,j)=1} D^i\varphi_\nu(x_i)\, D^i\varphi_\mu(x_i) + \int_a^b L\varphi_\nu\, L\varphi_\mu = \int_a^b L^*L\varphi_\nu\,\varphi_\mu = \lambda_\nu \int_a^b \varphi_\nu\varphi_\mu = 0$$

and hence $\{\varphi_\nu\}_{\nu\geqq 1}$ are orthogonal in $\mathscr{H}_0$. If $h\in\mathscr{H}_0$, and if, for each ν,

$$\sum_{e(i,j)=1} D^i h(x_i)\, D^i\varphi_\nu(x_i) + \int_a^b L h\, L\varphi_\nu = 0$$

then it follows that

$$0 = \int_a^b L h \, L\varphi_\nu = \int_a^b h \, L^* L \varphi_\nu = \lambda_\nu \int_a^b h \, \varphi_\nu .$$

Since the φ_ν are complete in $\mathscr{L}_2(a, b)$ and since $\lambda_\nu > 0$ for all ν it follows that $h(x) = 0$ a.e. and, in fact, $h(x) \equiv 0$ since h is continuous. Thus $\{\varphi_\nu\}_{\nu \geqq 1}$ is a complete orthogonal system in $\mathscr{H}_0$. Now, for any ν,

$$|\varphi_\nu|^2_{\mathscr{H}_2^m} = \int_a^b (L\varphi_\nu)^2 = \lambda_\nu .$$

Hence, $\{\varphi_\nu / \sqrt{\lambda_\nu}\}$ is a complete orthonormal system in $\mathscr{H}_0$ and the n^{th} partial sum of f is given by

$$\begin{aligned} S_n f &= \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu} \left\{ \int_a^b L f \, L\varphi_\nu \right\} \varphi_\nu \\ &= \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu} \left\{ \int_a^b L^* L f \, \varphi_\nu \right\} \varphi_\nu \\ &= \sum_{\nu=1}^{n} \frac{1}{\lambda_\nu} \left\{ \int_a^b g \, \varphi_\nu \right\} \varphi_\nu \\ &= \Psi_n . \end{aligned}$$

To complete the proof we will verify (5.7). We have already treated the case $j = m$. Now the case $j = 0$ follows from Theorem 5.2. Indeed, by (4.4), $f \in \varrho \|g\|_{\mathscr{L}_2} \mathscr{S}_m$ where

$$\varrho = \sqrt{b-a} \left\{ \sup_{x \in [a,b]} \int_a^b [L_x G(x, \xi)]^2 \, d\xi \right\}^{\frac{1}{2}}$$

and it follows that the n-widths of $\varrho \|g\|_{\mathscr{L}_2} \mathscr{S}_m$ are simply those of $\mathscr{S}_m$ multiplied by $\varrho \|g\|_{\mathscr{L}_2}$. Thus (5.7), with $j = 0$, is implied by (5.3). If now $0 < j < m$, it follows from the theory of interpolation spaces (see [4]) that $\mathscr{H}_2^j(a, b)$ interpolates $\mathscr{H}_2^m(a, b)$ and $\mathscr{L}_2(a, b)$ and we have, for each $u \in \mathscr{H}_2^m(a, b)$,

$$\|u\|_{\mathscr{H}_2^j} \leqq (\|u\|_{\mathscr{H}_2^m})^{j/m} (\|u\|_{\mathscr{L}_2})^{1-j/m},$$

and hence (5.7) for $0 < j < m$ follows from the result for $j = 0$ and $j = m$, upon setting $u = f - \Psi_n$. This concludes the proof of Theorem 5.3.

§ 6. Generalizations

The preceding theory is capable of further generalization. In particular, let $L_1, L_2, \ldots, L_q$ be linear differential operators of order m_p, $1 \leqq p \leqq q$, of the form

$$L_p = \sum_{j=0}^{m_p} a_{pj} D^j$$

with $a_{pj} \in C^j[a, b]$, $0 \leqq j \leqq m_p$, $1 \leqq p \leqq q$ where the a_{pj} may be complex-valued. We will assume that one of L_p, say L_q, is of order $m_q = m$ with $m_p < m$ for $1 \leqq p < q$. Also, we assume that $a_{qm}(x) \neq 0$ on $[a, b]$, i.e., the coefficient of the

highest order derivative of L_q does not vanish. Consider now the operator

$$\Lambda = \sum_{p=1}^{q} L_p^* L_p, \tag{6.1}$$

where $L_p^* f = \sum_{j=0}^{m_p} (-1)^j D^j(\bar{a}_{jp} f)$, $p = 1, \ldots, q$. It is well-known that any linear, formally self-adjoint, differential operator with continuous coefficients on $[a, b]$ of one sign can be written in the form (6.1) where the L_p, in general, will have complex-valued coefficients. We consider the following generalization of (1.1):

$$\begin{aligned}
&\text{(a)} \quad \Lambda f(x) = g(x), \qquad x_i < x < x_{i+1} \\
&\text{(b)} \quad D^j f(x_i) = r_{ij} \qquad \text{if} \quad e(i,j) = 1 \\
&\text{(c)} \quad f \in A\,C^{m-1}[a,b] \cap C^{2m}\big(\textstyle\bigcup (x_i, x_{i+1})\big) \\
&\text{(d)} \quad \sum_{p:\, j < m_p} \overline{O_{jp} L_p f(x_i+)} = \sum_{p:\, j < m_p} \overline{O_{jp} L_p f(x_i-)} \quad \text{if} \quad e(i,j) = 0, \\
&\qquad 0 \leqq i \leqq k \quad \text{and} \quad 0 \leqq j \leqq m-1,
\end{aligned} \tag{6.2}$$

where O_{jp} is associated with L_p, $1 \leqq p \leqq q$, as in §1.

We make the following assumption:

$$L_p f \equiv 0, \quad p = 1, 2, \ldots, q \quad \text{and} \quad D^j f(x_i) = 0 \quad \text{if} \quad e(i,j) = 1 \Rightarrow f \equiv 0. \tag{6.3}$$

Our first result provides a generalization of the spline functions of §2.

Theorem 6.1. Let $g \equiv 0$ in (6.1 a). Then there exists, for any prescribed complex numbers r_{ij}, a unique complex-valued function f satisfying (6.1).

Proof. For convenience, we merely sketch the proof. We define $\mathscr{H}_\Lambda(a, b)$ to be the collection of functions $h \in A\,C^{m-1}[a, b]$ such that $\sum_{p=1}^{q} \int_a^b |L_p h|^2 < \infty$. It is easily seen that $\mathscr{H}_\Lambda(a, b)$ is a Hilbert space under the inner product

$$(u, v)_{\mathscr{H}_\Lambda} = \sum_{e(i,j)=1} D^j u(x_i)\, \overline{D^j v(x_i)} + \int_a^b \sum_{p=1}^{q} L_p u\, \overline{L_p v}.$$

Now the variational problem

$$\inf_{u \in \mathscr{U}} \|u\|^2_{\mathscr{H}_\Lambda} = \|f\|^2_{\mathscr{H}_\Lambda} \tag{6.4}$$

where $\mathscr{U} = \{u \in \mathscr{H}_\Lambda(a, b)\colon D^j u(x_i) = r_{ij} \text{ if } e(i,j) = 1\}$ is equivalent to the problem

$$\inf_{u \in \mathscr{U}} \sum_{p=1}^{q} \int_a^b |L_p u|^2 = \sum_{p=1}^{q} \int_a^b |L_p f|^2 \tag{6.5}$$

and since (6.4) has a unique solution f, it follows that (6.5) has (the same) unique solution f. We claim that f satisfies the homogeneous problem (6.2), i.e., with $g \equiv 0$. Indeed, f satisfies the orthogonality relation

$$\sum_{p=1}^{q} \int_a^b L_p f\, \overline{L_p h} = 0 \tag{6.6}$$

for all $h \in \mathscr{H}_\Lambda(a, b)$ such that $D^j h(x_i) = 0$ if $e(i,j) = 1$.

Now $f \in A\,C^{m-1}[a, b]$ since $f \in \mathscr{H}_\Lambda(a, b)$ and f satisfies (6.2b) since $f \in \mathscr{U}$. The fact that f satisfies the remaining properties of (6.2) follows from (6.6). Indeed, suppose i is fixed, $0 \leqq i \leqq k$, and let $\varphi \in C_c^\infty(x_i, x_{i+1})$, i.e., φ is an infinitely differentiable function with compact support on (x_i, x_{i+1}). Then by (6.6) and integration by parts,

$$0 = \sum_{p=1}^{q} \int_a^b L_p f \overline{L_p \varphi} = \int_a^b f \Big(\sum_{p=1}^{q} L_p^* L_p \varphi \Big),$$

i.e., f is a distribution solution of $\sum_{p=1}^{q} L_p^* L_p f = 0$ on (a, b). By [5, §8] it follows that $f \in C^{2m}\big(\bigcup_i (x_i, x_{i+1})\big)$ and $L^* L f(x) = 0$ for $x_i < x < x_{i+1}$.

It remains only to show that f satisfies (6.2d). And indeed this follows from the integration by parts formula, valid if φ is an infinitely differentiable function with compact support in an interval containing only $x_* = x_i$ and satisfying $D^j \varphi(x_*) = 0$ if $e(i, j) = 1$,

$$\sum_{p=1}^{q} \int_a^b L_p f \overline{L_p \varphi} = \sum_{j=0}^{m-1} \big(\overline{D^j \varphi(x_*)}\big) \Big\{ \sum_{p:\, j < m_p} [\overline{O_{jp} L_p f}]_{x_*} \Big\} \tag{6.7}$$

where the notation $[\Psi]_{x_*}$ means $\Psi(x_* +) - \Psi(x_* -)$. That f satisfies (6.2d) follows from (6.7). This completes the proof.

The preceding theorem may be easily generalized to the situation where the boundary functionals are generalized to linear combinations of derivatives through order $m-1$ evaluated at some point x_i. We need only assume that the linear combinations associated with a given x_i are linearly independent and that the generalized version of (6.3) holds. Of course the higher order continuity conditions of (6.2d) are replaced by more complicated conditions (see [8] where this case is treated for $\Lambda = L^* L$). Moreover, a Green's function can be constructed, as before, for the problem (6.2) with homogeneous boundary conditions defining a positive, symmetric compact operator on $\mathscr{L}_2(a, b)$. Results similar to the previous ones can be obtained for the eigenvalue problem and the optimal convergence scheme. In the eigenvalue problem, $L^* L$ is replaced by Λ and O_j by $\sum_{p:\, j < m_p} O_{jp}$. In the optimal convergence schemes, the convergence takes place in $\mathscr{H}_\Lambda(a, b)$, the same set as $\mathscr{H}_2^m(a, b)$ and equivalent to it as a Hilbert space, and in $\mathscr{L}_2(a, b)$ with the order $O(n^{-m})$. The other convergence results are thus also of the same order. It should be pointed out, however, that the class for which the scheme is optimal is determined no longer by $\|Lf\|_{\mathscr{L}_2}^2 \leqq \varrho^2$ but by

$$\sum_{p=1}^{q} \|L_p f\|_{\mathscr{L}_2}^2 \leqq \varrho^2.$$

Other generalizations can be made also. For example, the boundary set $\{x_i\}_0^k$ can be enlarged to an infinite set (see [3]). Finally, non-linear equations of the form

$$\Lambda u(x) = f\big(x, u(x)\big)$$

can be treated with suitable assumptions on f. The details of this will be deferred to another paper.

§ 7. Remarks

The optimal approximation scheme discussed in §4 and §5 is quite closely related to the schemes appearing in [2, Sec. 4.6], though in this latter case only two point boundary value problems are considered and order of convergence results are not presented.

We also want to mention the work of Zettl and others (see [11] and references therein) in connection with self-adjoint multi-point boundary value problems. These authors, however, do not appear to have investigated the convergence properties of approximations to the solutions of the multipoint problems.

Finally, we include an example to illustrate the results developed in the previous sections. We will take L to be the differential operator D^2 and the interval $[a, b]$ to be the interval $[-\tau, \tau]$ for some $\tau > 0$. Thus, $m = 2$ and $L^*L = D^4$. We choose the partition $-\tau < 0 < \tau$ of $[-\tau, \tau]$ and specify the three by two incidence matrix

$$M = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

Given a function g continuous on $(-\tau, 0) \cup (0, \tau)$, the homogeneous boundary value problem consists, in this case, of determining a function $f \in C^2[-\tau, \tau] \cup C^4[(-\tau, 0) \cup (0, \tau)]$ such that $D^4 f$ coincides with g on $(-\tau, 0) \cup (0, \tau)$ and such that $f(-\tau) = Df(-\tau) = f(0) = f(\tau) = Df(\tau) = 0$. Physically, this corresponds to determining the configuration of a beam of length 2τ, clamped at both ends, pinned at the midpoint and supporting a load g, continuously distributed in each of the two halves.

The Green's function is readily determined and is given by:

$$G(x, \xi) = A(\xi) + B(\xi)\, x + C(\xi)\, x^2 + D(\xi)\, x^3 + E(\xi)\, (x)_+^3 + \tfrac{1}{6}(x - \xi)_+^3,$$

where, in general,

$$t_+^3 = \begin{cases} t^3 & \text{if} \quad t \geqq 0 \\ 0 & \text{if} \quad t < 0, \end{cases}$$

and where

$$A(\xi) = -\tfrac{1}{6}(-\xi)_+^3,$$

$$B(\xi) = \frac{\xi}{8\tau}(\tau - \xi)^2,$$

$$C(\xi) = \frac{1}{4\tau^2}[2(-\xi)_+^3 + (\tau - \xi)^2(2\tau - \xi)],$$

$$D(\xi) = \frac{1}{8\tau^3}[\tfrac{8}{3}(-\xi)_+^3 + \xi(\tau - \xi)^2],$$

$$E(\xi) = \frac{-1}{6\tau^3}[4(-\xi)_+^3 + (\tau - \xi)^2(\tau + 2\xi)].$$

G is easily seen to satisfy (3.1) and (3.2). In particular, $G(\cdot, \xi)$ satisfies the homogeneous boundary conditions and is in $C^2[-\tau, \tau]$ as a function of the first argument and $G^{(3)}(\xi+, \xi) - G^{(3)}(\xi-, \xi) = 1$. Also, $G \in C^1([-\tau, \tau] \times [-\tau, \tau])$ and is symmetric by Theorem 3.1.

The eigenvalue problem (4.1) in this case reduces to determining the positive real numbers λ such that the corresponding eigenfunctions u satisfy the equation $D^4u = \lambda u$ on each of the subintervals $(-\tau, 0)$ and $(0, \tau)$, u satisfies the boundary conditions homogeneously and is in $C^2[-\tau, \tau]$, with absolutely continuous third derivatives on each of the intervals $(-\tau, 0)$ and $(0, \tau)$. A lengthy computation yields the result that the eigenvalues are the positive roots λ of the equation

$$\cosh \lambda\tau \cos \lambda\tau = 1$$

and the corresponding eigenfunctions are given by:

$$u(x) = \sin \lambda x + \sigma \cos \lambda x - \sinh \lambda x - \sigma \cosh \lambda x + 2\left(\sinh \lambda(x)_+ - \sin \lambda(x)_+\right)$$

where

$$\sigma = \frac{\cosh \lambda\tau - \cos \lambda\tau}{\sinh \lambda\tau + \sin \lambda\tau}.$$

References

1. Agmon, S.: Lectures on elliptic boundary value problems. Princeton: Van Nostrand 1965.
2. Babuska, I., Prager, M., Vitasek, E.: Numerical processes in differential processes in differential equations. London: Interscience, Wiley 1966.
3. Golomb, M., Jerome, J.: Ordinary differential equations with boundary conditions on arbitrary point sets (to appear), TAMS.
4. Grisvard, P.: Characterisation de quelques espaces d'interpolation. Arch. Rat. Mech. Anal. **25**, 40–63 (1967).
5. Halperin, I.: Introduction to the theory of distributions. Toronto: University of Toronto Press 1952.
6. Jerome, J.: On the $\mathscr{L}_2$ n-width of certain classes of functions of several variables. J. Math. Anal. Appl. **20**, 110–123 (1967).
7. — On n-widths in Sobolev spaces and applications to elliptic boundary value problems. J. Math. Anal. Appl. **29**, 201–215 (1970).
8. — Schumaker, L.: On Lg splines. J. Approx. Th. **2**, 29–49 (1969).
9. Lorentz, G.: Approximation of functions. New York: Holt 1966.
10. Riesz, F., Sz.-Nagy, B.: Functional analysis. New York: Ungar 1955.
11. Zettl, A.: Adjoint and self-adjoint boundary value problems with interface conditions. M.R.C. Technical Summary Report 827, Madison, Wisconsin, 1967.

Joseph W. Jerome
Department of Mathematics
Northwestern University
Evanston, Illinois 60201

Numer. Math. 15, 450 (1970)
© by Springer-Verlag 1970

A Short Note on the Implicit QL Algorithm for Symmetric Tridiagonal Matrices

A. DUBRULLE

Received October 9, 1969

The following is a formulation of the QL algorithm with implicit shift which requires fewer operations than the explicit and implicit algorithms described in [1] and [2].

Let $d_i^{(s)}$ $(i=1, \ldots, n)$ and $e_i^{(s)}$ $(i=1, \ldots, n-1)$ be the diagonal and codiagonal elements of the matrix at the s-th iteration, and k_s the shift. Then the $(s+1)$-st iteration can be expressed as follows.

$$
\begin{aligned}
& c_n = s_n = 1, \quad q_n = d_n^{(s)} - k_s, \quad u_n = 0 \\
& \left.\begin{aligned}
h_i &= c_{i+1} e_i^{(s)} \\
p_i &= s_{i+1} e_i^{(s)} \\
e_{i+1}^{(s+1)} &= (p_i^2 + q_{i+1}^2)^{\frac{1}{2}} \\
s_i &= p_i / e_{i+1}^{(s+1)} \\
c_i &= q_{i+1} / e_{i+1}^{(s+1)} \\
g_i &= d_{i+1}^{(s)} - u_{i+1} \\
t_i &= (d_i^{(s)} - g_i)\, s_i + 2 c_i h_i \\
u_i &= s_i t_i \\
d_{i+1}^{(s+1)} &= g_i + u_i \\
q_i &= c_i t_i - h_i
\end{aligned}\right\} i = n-1, \ldots, 1 \\
& d_1^{(s+1)} = d_1^{(s)} - u_1, \quad e_1^{(s+1)} = q_1.
\end{aligned}
$$

Acknowledgements. The author wishes to thank Dr. J. H. Wilkinson for his helpful comments.

References

1. Bowdler, H., Martin, R. S., Reinsch, C., Wilkinson, J. H.: The QR and QL algorithms for symmetric matrices. Numerische Mathematik **11**, 293–306 (1968).
2. Martin, R. S., Wilkinson, J. H.: The implicit QL algorithm. Numerische Mathematik **12**, 377–383 (1968).

A. Dubrulle
IBM Corporation
Industry Development-Scientific Applications
6900 Fannin
Houston, Texas 77025, U.S.A.