Difference Approximations for Boundary and Eigenvalue Problems for Ordinary Differential Equations

By Heinz-Otto Kreiss

Abstract. The boundary value problem for ordinary differential equations is considered and a general theory for difference approximation is developed. In particular, the influence of extra boundary conditions is investigated and the eigenvalue problem is considered in detail.

1. Introduction. Consider an nth order linear system of ordinary differential equations

$$(1.1) \qquad Ly = d^n y/dx^n + \sum_{\mu=0}^{n-1} A_\mu(x) \, d^\mu y/dx^\mu = F$$

in the interval $0 \leq x \leq 1$. Here, $y = (y^{(1)}(x), \cdots, y^{(m)}(x))'$ * and

$$F = (F^{(1)}(x), \cdots, F^{(m)}(x))' \in C$$

are vector functions and the $A_i(x) \in C^1$ ** are $m \times m$ matrices. Furthermore, mn linearly independent boundary conditions

$$(1.2) \qquad B_l y = \sum_{i=0}^{l} B_{il}(0) \, d^i y(0)/dx^i + B_{il}(1) \, d^i y(1)/dx^i = g_l,$$

$$l = 0, 1, 2, \cdots, n - 1,$$

are given. Thus $B_l y = g_l$ describes the boundary conditions which contain derivatives up to order l, the B_{il} are rectangular matrices with r_l rows and n columns. Without loss of generality, we may assume that the rows of $(B_{ll}(0), B_{ll}(1))$ are linearly independent. Thus, $\sum r_l = mn$.

We shall also consider the eigenvalue problem

$$(1.3) \qquad L\phi = \lambda\phi, \qquad B_l \phi = 0, \qquad l = 0, 1, 2, \cdots, n - 1,$$

and assume that not all complex numbers λ are eigenvalues.

The aim of this paper is to develop a general theory for difference approximation of the form (2.1), (2.14). Specifically, we shall investigate the influence of extra boundary conditions on the speed of convergence. These extra boundary conditions are

* If F is a vector, then F' denotes its transpose and $|F| = (\sum |F^{(i)}|^2)^{1/2}$ its norm.

** C^ν denotes the class of functions which are ν times continuously differentiable in the interval $0 \leq x \leq 1$. C is the class of continuous functions.

necessary if $r + s$, the "width" of the difference approximation, is larger than n. Another interesting result is that the behavior of the eigenvalue problem is completely determined by the behavior of the corresponding inhomogeneous problem. For example, if one can use Richardson extrapolation for the inhomogeneous problem, then one can also use it to determine the eigenvalues and the invariant subspaces. No assumption of selfadjointness or simplicity of the eigenvalues is needed.

Remark. The assumption that the coefficients are smooth is no real restriction. H. Keller [2] has pointed out a procedure by which one can reduce the case of piecewise smooth coefficients to the case of smooth coefficients.

2. Formulation of Difference Approximations. We want to solve the problem (1.1), (1.2) by difference approximation. Let $h = N^{-1}$, N a natural number, and define gridpoints x_ν by $x_\nu = \nu h$, $\nu = 0, 1, 2, \cdots, N$. Using the notation $v_\nu = v(x_\nu)$, we approximate (1.1) by

$$(2.1) \qquad h^n L_h v_\nu = \sum_{j=-r}^{s} C_j(x_\nu, h) E^j v_\nu = h^n \tilde{F}_\nu, \qquad \nu = r, r + 1, \cdots, N - s.$$

Here, r, s are natural numbers with $r + s \geq n$, and $C_j(x_\nu, h)$ are $m \times m$ matrices which belong to C^1 as functions of x and are polynomials in h. Furthermore, E denotes the translation operator, i.e., $E^j v_\nu = v_{\nu+j}$; and $\tilde{F}_\nu$ is an approximation of F_ν such that

$$(2.2) \qquad \limsup_{h \to 0} |F_\nu - \tilde{F}_\nu| = 0.$$

Practically all of the difference approximations used are of the form (2.1). For later reference, we discuss some of them:

(1) We approximate the system of differential equations

$$(2.3) \qquad dy/dx + A(x)y = F$$

by

$$(2.4) \qquad D_+ v_\nu + \tfrac{1}{2} A(x_\nu + \tfrac{1}{2} h)(E + I)v_\nu = \tfrac{1}{2}(E + I)F_\nu, \qquad \nu = 0, 1, 2, \cdots, N - 1.$$

Here, $I = E^0$, $D_+ = h^{-1}(E - I)$ and $r = 0$, $s = 1$, i.e., $r + s = n$. This type of approximation has been thoroughly studied by Keller [2].

(2) We approximate the differential equation

$$(2.5) \qquad d^2 y/dx^2 + a_0(x)y = F$$

by the usual difference equations

$$(2.6) \qquad D_+ D_- v_\nu + a_0(x_\nu)v_\nu = F_\nu, \qquad D_- = h^{-1}(I - E^{-1}), \qquad \nu = 1, 2, \cdots, N - 1;$$

or

$$(2.7) \qquad \left(D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2 \right) v_\nu + a_0(x_\nu)v_\nu = F_\nu, \qquad \nu = 2, \cdots, N - 2.$$

In the first case, $r = s = 1$, i.e., $r + s = 2 = n$, while in the second case, $r = s = 2$ and $r + s > n$.

We want to define consistency of the difference equations (2.1). For this reason, we rewrite the equations in the form

$$(2.8) \qquad h^n L_h v_\nu = \sum_{j=0}^{r+s} \tilde{C}_j(x_\nu, h)(h D_+)^j v_{\nu-r} = h^n \tilde{F}_\nu.$$

Here, the $\tilde{C}_j(x_\nu, h)$ are linear combinations of the $C(x_\nu, h)$. For example, (2.4), (2.6) and (2.7) can be written as

$$((I + \tfrac{1}{2} h A(x_\nu + \tfrac{1}{2} h))(h D_+) + h A(x_\nu + \tfrac{1}{2} h)) v_\nu = \tfrac{1}{2} h(E + I) F_\nu,$$

$$((h D_+)^2 + h^2 a_0(x_\nu)(h D_+) + h^2 a_0(x_\nu) I) v_{\nu-1} = h^2 F_\nu,$$

and

$$(- \tfrac{1}{12}(h D_+)^4 + (h D_+)^3 + (1 + h^2 a_0(x_\nu))(h D_+)^2$$
$$+ 2 h^2 a_0(x_\nu)(h D_+) + h^2 a_0(x_\nu) I) v_{\nu-2} = h^2 F_\nu,$$

respectively.

Definition 2.1. The difference approximation (2.1) is consistent, if there is a constant $K_1 > 0$ such that, for all $h > 0$,

$$(2.9) \qquad \sup_x \left(|\tilde{C}_n(x, h) - I| + \sum_{j=0}^{n-1} |h^{j-n} \tilde{C}_j(x, h) - A_j(x)| \right) \leqq K_1 h.$$

There is no difficulty in showing that this definition is equivalent with the usual one. We leave the proof of the following lemma to the reader:

LEMMA 2.1. *The difference approximation is consistent if and only if for every* $w(x) \in C^\infty$ *there is a constant* $K(w)$ *such that, for all* $h > 0$,

$$(2.10) \qquad \sup_{r \leqq \nu \leqq N-s} |L w(x_\nu) - L_h w_\nu| \leqq K(w) h.$$

It is obvious that the difference approximations (2.4), (2.6) and (2.7) are consistent with the corresponding differential equations.

For later purposes, we write (2.8) in the form

$$(2.11) \qquad L_h v_\nu = S_0(h) D_+^n v_{\nu-r} + \sum_{j=0}^{n-1} \tilde{A}_j(x_\nu, h) D_+^j v_{\nu-r} = \tilde{F}_\nu,$$

$$\nu = r, r+1, \cdots, N-s.$$

Here, $\tilde{A}_j(x, h) = h^{j-n} \tilde{C}_j(x, h)$ and $S_0(h)$ denotes a uniformly bounded difference operator of the form

$$(2.12) \qquad S_0(h) = \sum_{k=0}^{r+s-n} S_{0k}(x_\nu, h) E^k,$$

whose coefficients are linear combinations of $\tilde{C}_n, \cdots, \tilde{C}_{r+s}$. For example, for (2.4), (2.6) and (2.7), the operator $S_0(h)$ has the form

$$S_0(h) = I + \tfrac{1}{2} h A(x_\nu + \tfrac{1}{2} h), \qquad S_0(h) = I,$$

and

$$S_0(h) = -\tfrac{1}{12}(h D_+)^2 + (h D_+) + (1 + h^2 a_0(x_\nu)) I$$
$$= -\tfrac{1}{12} E^2 + \tfrac{7}{6} E - (\tfrac{1}{12} - h^2 a_0) I,$$

respectively.

From (2.9), it follows that

$$(2.13) \qquad \sum_{k=0}^{r+s-n} S_{0k}(x_\nu, h) = I + O(h), \quad \text{i.e.,} \quad \sum_{k=0}^{r+s-n} S_{0k}(x_\nu, 0) = I.$$

(2.1) represents $(N - (r + s) + 1)m$ linear equations for the $(N + 1)m$ unknowns v_ν, $\nu = 0, 1, 2, \cdots, N$. Therefore, we have to add $(r + s)m$ boundary conditions. These boundary conditions are also linear expressions between the v_i near the boundary points $x = 0$ and $x = 1$. We shall write them in the form

$$(2.14) \qquad B_{lh}v = \sum_{j=0}^{l} (\tilde{B}_{jl}(0, h) \, D_+^j v_0 + \tilde{B}_{jl}(1, h) \, D_-^j v_N) + hR_{l+1} = \tilde{g}_l(h).$$

Here, $\tilde{B}_{jl}$ are rectangular matrices whose elements are polynomials in h, and $\tilde{R}_{l+1}v$ stands for a linear combination of divided differences $D_+^{l+1}v_j$ for which an estimate

$$(2.15) \qquad |R_{l+1}v| \leq K_2 \max_{0 \leq \nu \leq N-l-1} |D_+^{l+1}v_\nu|, \qquad K_2 \text{ independent of } h,$$

holds. Thus, $R_{l+1}v$ represents the higher-order terms.

Definition 2.2. The boundary conditions (2.14) are consistent if the following conditions are fulfilled:

(1) If $r + s = n$, then $l \leq n - 1$ in (2.14) and there is a constant K_3, independent of h, such that

$$(2.16) \qquad |\tilde{g}(h) - g_l| + \sum_{j=0}^{l} |\tilde{B}_{jl}(0, h) - B_{jl}(0)| + |\tilde{B}_{jl}(1, h) - B_{jl}(1)| \leq K_3 h,$$
$$l = 0, 1, \cdots, n - 1.$$

(2) If $r + s > n$, then there are still nm boundary conditions of type (2.14) with $l \leq n - 1$ for which (2.16) holds. Furthermore, there are $((r + s) - n)m$ extra boundary conditions of the same type with $l \geq n$.

We do not know of any difference approximation which cannot be written in the above form. Let us now consider the examples.

(1) The boundary conditions for the system (2.3) are given by

$$(2.3a) \qquad\qquad B_{00}(0)y(0) + B_{00}(1)y(1) = 0$$

and we use

$$(2.4a) \qquad\qquad B_{00}(0)v_0 + B_{00}(1)v_N = 0$$

as an approximation which is obviously consistent.

(2) For the differential equation (2.5), the boundary conditions shall be given by

$$(2.5a) \qquad\qquad y(0) = y(1) = 0.$$

As an approximation, we use

$$(2.6a) \qquad\qquad v_0 = v_N = 0$$

or

$$(2.7a) \qquad\qquad v_0 = v_N = 0, \qquad D_+^\tau v_0 = D_-^\tau v_N = 0,$$

where τ is a natural number with $\tau \geq 2$. Even these approximations are consistent.

Remark. $D_+^r v_0 = D_-^r v_N = 0$ defines some kind of extrapolation which is independent of the differential equation. However, one can also use the differential equation to derive the extra boundary conditions. From (2.5) and (2.5a), it follows that

$$y''(0) = F(0), \qquad y''(1) = F(1),$$

and, therefore, we can use instead of (2.7a)

(2.7b) $$v_0 = v_N = 0, \qquad D_+^2 v_0 = F(0), \qquad D_-^2 v_N = F(1).$$

The accuracy of (2.7b) is, of course, improved if $h = (N - 2)^{-1}$ and x_ν is defined by $x_\nu = (\nu - 1)h$, i.e., $x_0 = -h$, $x_N = 1 + h$.

We shall also consider the eigenvalue problem. In that case, we approximate (1.3) by

(2.17) $$L_h \psi = \lambda_h S_h \psi_\nu, \qquad \nu = r, r + 1, \cdots, N - s,$$

with boundary conditions

(2.18) $$B_{lh} \psi = 0.$$

Here, S_h is a uniformly bounded operator of the form

(2.19) $$S_h = \sum_{j=-r}^{s} S_{1j}(s_\nu, h)E^\nu.$$

The coefficients $S_{1\nu}$ are matrices which belong to C^1 as functions of x and are polynomials in h. For consistency, we assume that

(2.20) $$\sup_\nu \left| \sum_{j=-r}^{s} S_{1j}(x_\nu, h) - I \right| \leqq K_4 h, \qquad K_4 = \text{const independent of } h.$$

For our examples, we get

(2.21) $$D_+ \psi_\nu + \tfrac{1}{2} A(x_\nu + \tfrac{1}{2}h)(E + I)\psi_\nu = \tfrac{1}{2}\lambda_h(E + I)\psi_\nu,$$

(2.22) $$D_+ D_- \psi_\nu + a(x_\nu)\psi_\nu = \lambda \psi_\nu,$$

(2.23) $$\left(D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2 \right)\psi_\nu + a(x_\nu)\psi_\nu = \lambda \psi_\nu.$$

3. General Convergence and Stability Theorems. In this section, we want to prove a number of general stability and convergence theorems. Let $\mathfrak{B}_h$ denote the Banach space of all gridfunctions $v = (v_0, v_1, \cdots, v_N)'$ with the norm defined by

$$\|v\|_h = \max_{0 \leqq \nu \leqq N} |v_\nu|.$$

If $v \in \mathfrak{B}_h$, then it is not true that $w = D_+ v$ also belongs to $\mathfrak{B}_h$, because $w_\nu = D_+ v_\nu$ is only defined for $\nu = 0, 1, 2, \cdots, N - 1$. We shall, however, use the notation

$$\|D_+ v\|_h = \max_{0 \leqq \nu \leqq N-1} |D_+ v_\nu|.$$

In general, if $T v_\nu = \sum_{j=-p}^{q} T_j(x_\nu)E^j v_\nu$ is a difference operator, then

$$\|T v\|_h = \max_{p \leqq \nu \leqq N-q} |T v_\nu|.$$

Thus,

$$||D_+^i D_-^k v||_h = ||D_+^{i+k} v||_h = \max_{0 \leq \nu \leq N-i-k} |D_+^{i+k} v_\nu|.$$

We shall denote by $\mathfrak{B}$ the Banach space of all functions $f(x) \in C$ with the norm defined by

$$||f|| = \sup_{0 \leq x \leq 1} |f(x)|.$$

With these notations, we write down the following two well-known lemmas:

LEMMA 3.1. *Let j, k with $j < k$ be natural numbers. For every $\delta > 0$, there are constants $c_{jk}(\delta)$, independent of h and v, such that*

$$||D_+^j v||_h \leq \delta ||D_+^k v||_h + c_{jk}(\delta) ||v||_h.$$

LEMMA 3.2. *Let k be a natural number. For every h and every gridfunction v_h, there exists a function $w = w(x, k, h) \in C^k$ and constants d_{jk}, independent of h and v, such that*

$$w(x_\nu, h) = v_\nu, \qquad \nu = 0, 1, 2, \cdots, N,$$

$$||D_+^j v||_h \leq ||d^i w/dx^i|| \leq d_{jk}(||v||_h + ||D_+^j v||_h), \qquad j = 0, 1, 2, \cdots, k.$$

Proof. The function w of Lemma 3.2 can, for example, be constructed in the following way: Define v_ν for $\nu > N$ by $D_+^{k+1} v_{N-k+\mu} = 0$, $\mu = 0, 1, 2, \cdots$, and construct w by Hermite interpolation with $d^i w(x_\nu)/dx^i = D_+^i v_\nu$, $j = 0, 1, 2, \cdots, k$. This process we shall denote by $w = \mathrm{Int}_k v$.

Lemma 3.2 implies that Lemma 3.1 holds if we can prove it for the continuous case, i.e.,

$$||d^i u/dx^i|| \leq \delta ||d^k u/dx^k|| + C_{jk}(\delta) ||u||.$$

Let $||u||_2 = (\int_0^1 |u|^2 \, dx)^{1/2}$ denote the L_2-norm of u. Then the above inequality follows from the corresponding Sobolev inequality

$$||d^i u/dx^i||^2 \leq \delta^2 ||d^k u/dx^k||_2^2 + C_{jk}^2(\delta) ||u||_2^2.$$

With these two lemmas, we can now prove the main result of this section.

THEOREM 3.1. *Assume that $\lambda = 0$ is not an eigenvalue of (1.3) and that there is a constant K_1 such that for all h and all solutions of (2.1), (2.14) an a priori estimate*

$$(3.1) \qquad\qquad ||D_+^n v||_h \leq K_1 \left(||v||_h + ||\tilde{F}||_h + \sum_l |\tilde{g}_l| \right)$$

holds. (Here we define $\tilde{F}_\nu = 0$ for $\nu = 0, 1, \cdots, r - 1$ and $\nu = N - s + 1, \cdots, N$.) If the Eqs. (2.1), (2.14) are consistent, then these equations have, for every $\tilde{F}$, $\tilde{g}$ and all sufficiently small h, a unique solution $v \in \mathfrak{B}_h$, and there is a constant K_2 such that

$$(3.2) \qquad\qquad ||v||_h \leq K_2 \left(||\tilde{F}||_h + \sum_l |\tilde{g}| \right).$$

Furthermore, the interpolated function $\mathrm{Int}_n v$ converges to the solution u of the differential equations, i.e.,

$$(3.3) \qquad\qquad \lim_{h \to 0} ||\mathrm{Int}_n v - u|| = 0.$$

Proof. Assume that (3.2) does not hold. Then, we can select subsequences h_j, $\bar{F}^{(j)}$, $\tilde{g}_l^{(j)}$, $j = 1, 2, \cdots$, with

$$h_j \to 0, \qquad ||\bar{F}^{(j)}||_{h_j} \to 0, \qquad |\tilde{g}_1^{(j)}| \to 0,$$

such that the equations

(3.4) $$L_{h_j} v_\nu = \bar{F}^{(j)}, \qquad B_{lh_j} v = \tilde{g}_l^{(j)},$$

have, for every j, a solution $v^{(j)}$ with $||v^{(j)}||_{h_j} = 1$.

Using Lemma 3.2 with $k = n$, we can construct a sequence of functions $w^{(j)}(x) = \text{Int}_n\, v^{(j)}$, which, by (3.1), have uniformly bounded derivatives up to the order n. Therefore, we can, without restriction, assume that $w^{(j)}(x)$ and its first $n - 1$ derivatives converge uniformly to a function $u(x)$ and its derivatives. Here $||u(x)|| = 1$. Add the Eqs. (3.4) for $\nu = r, r + 1, \cdots$, and use the notations $h = h_j$, $w_\nu = w_\nu^{(j)} = v_\nu^{(j)}$. Then

(3.5) $$\sum_{\sigma=r}^{\mu} S_0(h) D_+^n w_{\sigma-r} h + \sum_{i=0}^{n-1} \sum_{\sigma=r}^{\mu} A_i(x_\sigma, h) D_+^i w_{\sigma-r} h = \sum_{\sigma=r}^{\mu} \bar{F}_\sigma^{(j)} h,$$

$$\mu = r, \cdots, N - s.$$

By (2.12),

$$\sum_{\sigma=r}^{\mu} S_0(h) D_+^n w_{\sigma-r} h = \sum_{\sigma=r}^{\mu} (S_0(h) - I) D_+^n w_{\sigma-r} h$$

$$= D_+^{n-1} w_{\mu-r} - D_+^{n-1} w_0 + R_{1\nu} + R_{2\nu},$$

where

$$R_{1\mu} = \sum_{k} \sum_{\sigma=r}^{\mu} S_{0k}(E^k - 1) D_+^n w_{\sigma-r} h,$$

$$R_{2\mu} = \sum_{\sigma=r}^{\mu} \left(\sum_{k} S_{0k} - I \right) D_+^n w_{\sigma-r} h.$$

By assumption, the $S_{0k} = S_{0k}(x_\sigma, h)$ belong to C^1 as functions of x and are polynomials in h. Therefore, partial summation gives us

$$\max_{\mu} |R_{1\mu}| \leqq \text{const } h ||D_+^n w||_h = \text{const } h ||D_+^n v||_h \leqq \text{const } h.$$

From (2.13), it follows that the same inequality also holds for $R_{2\mu}$. Thus, (3.5) implies that $u(x)$ is the solution of

$$d^{n-1} u(x)/dx^{n-1} - d^{n-1} u(0)/dx^{n-1} + \int_0^x \sum A_i(\xi)(d^i u(\xi)/d\xi^i)\, d\xi = 0,$$

i.e., $Lu = 0$.

It is obvious that u also fulfills the boundary conditions (1.2). Therefore, $\lambda = 0$ is an eigenvalue of (1.3) which is a contradiction. We have proved that an estimate of type (3.2) holds.

Writing the difference equations (2.11) in the form (3.5), consistency and the inequalities (3.1) and (3.2) then imply (3.3).

We shall now derive algebraic conditions such that the estimate (3.1) is valid. Let

$y_\nu = D_+^n v_\nu$, $\nu = 0, 1, 2, \cdots, N - n$. Then, we can write the Eq. (2.11) in the form

(3.6) $$S_0(0)y_{\nu-r} = G_\nu, \qquad \nu = r, \cdots, N - s,$$

and it follows from (2.11) and (2.13) that

$$||G||_h \leqq \text{const}\left(\sum_{\mu=0}^{n-1} ||D_+^\mu v||_h + ||\tilde{F}||_h\right).$$

(Here we again define $G_\nu = 0$ for $\nu = 0, \cdots, r - 1$ and $\nu = N - s + 1, \cdots, N$.)

If we use the relations

$$D_+^\sigma v_i = D_+^{\sigma-n} y_i, \qquad D_-^{\sigma-n} y_{i-n}, \qquad \sigma \geqq n,$$

then, we can write the $((r + s) - n)m$ extra boundary conditions (2.14) with $l \geqq n$, after multiplication with h^{l-n}, in the form

(3.7) $$\sum_{i=0}^{N-n} H_{il}y_i = \tilde{\tilde{g}}_l.$$

Here, the H_{il} are rectangular matrices independent of h and, for $\tilde{\tilde{g}}_l$, we have an estimate

(3.7a) $$|\tilde{\tilde{g}}_l| \leqq \text{const}\left(\sum_{\mu=0}^{n-1} ||D_+^\mu v||_h + h^{l-n}\sum_{l\geqq n} |\tilde{g}|\right).$$

(3.6) and (3.7) represent (at least formally) $(N - n + 1)m$ linear equations for the $(N - n + 1)m$ unknowns $y_0, \cdots, y_{N-n}$. From Lemma 3.1, we obviously get

THEOREM 3.2. *Assume that the Eqs. (3.6) and (3.7) have, for every G_ν and $\tilde{\tilde{g}}_l$, a unique solution and that there is a constant K_3 such that, for all h, G_ν and $\tilde{\tilde{g}}_l$,*

(3.8) $$\max_{0\leqq\nu\leqq N-n} |y_\nu| \leqq K_3\left(\max_{r\leqq\nu\leqq N-s} |G_\nu| + \max_l |\tilde{\tilde{g}}_l|\right),$$

then (3.1) holds.

A corollary is

THEOREM 3.3. *Assume that $r + s = n$, i.e., the difference scheme is as compact as possible. Then (3.8) holds.*

Proof. By (2.13), we have $S_0(0) = I$ and (3.8) follows from (3.6).

Thus, for compact difference schemes, we have only to make sure that the boundary conditions are consistent. Therefore, Theorem 3.1 implies that the solutions of our first two examples (2.4), (2.4a) and (2.6), (2.6a) converge to the solution of the differential equations if $\lambda = 0$ is not an eigenvalue of (1.3).

In most applications, the difference equation (3.6) is a scalar equation with constant coefficients, i.e., we can write it in the form

(3.9) $$\sum_{\mu=-r}^{s-n} \alpha_\mu E^\mu y_\nu = G_\nu, \qquad \nu = r, \cdots, N - s,$$

where the α_μ are constants. In this case, we can prove:

LEMMA 3.1. *Let the difference equation (3.6) be of the form (3.9) and denote by κ_i the solutions of the characteristic equation*

(3.10) $$\sum_{\mu=-r}^{s-n} \alpha_\mu \kappa^\mu = 0.$$

The estimate (3.8) *holds if and only if all* $|\kappa_i| \neq 1$ *and* (3.8) *is valid for the special case that* $G \equiv 0$.

Proof. Let $\kappa_1 = e^{i\sigma}$, σ real, be a solution of the characteristic equation, then we can write (3.9) in the form

$$(3.11) \qquad (E - e^{i\sigma})w_\nu = G_\nu, \qquad w_\nu = \alpha_{s-n} \prod_{j \neq 1} (E - \kappa_j)y_\nu.$$

If (3.8) is valid, then w_ν has to be bounded for all G_ν. This is impossible. We need only to choose $G_\nu = e^{i\nu\sigma}G_0$. Thus, the conditions of the lemma are necessary.

Assume now that all $|\kappa_i| \neq 1$. Then, there is no difficulty in showing that (3.9) has a particular solution w, i.e.,

$$\sum_{\mu=-r}^{s-n} \alpha_\mu E^\mu w_\nu = \alpha_{s-n} \prod_j (E - \kappa_j)w_\nu = G_\nu,$$

with

$$\max_{0 \leq \nu \leq N-n} |w_\nu| \leqq \text{const} \max_{r \leq \nu \leq N-s} |G_\nu|.$$

Subtracting w from y, the lemma is proved.

Let us now consider the difference equation (2.7) with the boundary conditions (2.7a). Observing that

$$\left(D_+ D_- - \frac{h^2}{12} D_+^2 D_-^2\right)v = \left(D_+^2 + h D_+^3 - \frac{h^2}{12} D_+^4\right)v_{\nu-2},$$

it follows that the homogeneous equation (3.6) has the form

$$(3.12) \qquad \left(I + h D_+ - \frac{h^2}{12} D_+^2\right)y_{\nu-2} = 0.$$

The boundary conditions (3.7) are given by

$$(3.13) \qquad h^{\tau-2} D_+^{\tau-2} y_0 = \tilde{\tilde{g}}_1, \qquad h^{\tau-2} D_-^{\tau-2} y_{N-2} = \tilde{\tilde{g}}_2.$$

The solutions of the characteristic equation $1 + (\kappa - 1) - \frac{1}{12}(\kappa - 1)^2 = 0$ are $\kappa_1 = 7 - \sqrt{48} \sim 0.07$ and $\kappa_2 = 7 + \sqrt{48} \sim 13.93$. Therefore, the general solution of (3.12) has the form

$$y_\nu = \sigma_1 \kappa_1^\nu + \sigma_2 \kappa_2^{\nu-N}.$$

The σ_i are determined by the boundary conditions (3.13) and there is no difficulty in showing that the desired estimate holds for any τ. We have thus proved that the solutions of (2.7), (2.7a) converge to the solution of the differential equation if $\lambda = 0$ is not an eigenvalue.

One can generalize the above results considerably. We state without proof:

THEOREM 3.4. *Consider the characteristic equation*

$$\text{Det} \left| \sum_{k=0}^{r+s-n} S_0(x_\nu, 0)\kappa^k \right| = 0,$$

for all fixed x_ν. If, for all its solutions κ_i, $|\kappa_i| \neq 1$, then the estimate (3.8) *is valid if it holds for the special case $G \equiv 0$.*

If the boundary conditions (3.7) can be separated into linear relations

$$\sum_{j=0}^{q} H_{il}y_i = \overset{\approx}{g}_l \quad \sum_{j=N-n-p}^{N-n} H_{il}y_i = \overset{\approx}{g}_l,$$

where q and p are some natural numbers independent of h, then one can go still further. Consider the half line problems with constant coefficients which we get from (3.6) and (3.7) by moving one of the boundaries to infinity and freezing the coefficients at the other boundary, i.e.,

$$(3.14) \qquad \sum_{k=0}^{r+s-n} S_{0k}(0, 0)E^k y_{\nu-r} = 0, \qquad \nu = r, r + 1, \cdots, N, N + 1, \cdots,$$

with boundary conditions

$$(3.15) \qquad\qquad \sum_{j=0}^{q} H_{il}y_i = 0, \qquad \sup_{r \leq \nu < \infty} |y_\nu| \leq \text{const},$$

and

$$(3.16) \qquad \sum_{k=0}^{r+s-n} S_{0k}(1, 0)E^k y_{\nu-r} = 0, \qquad \nu = N - s, N - s - 1, \cdots, 0, -1, \cdots,$$

with boundary conditions

$$(3.17) \qquad\qquad \sum_{j=N-n-p}^{N-n} H_{il}y_i = 0, \qquad \sup_{N-s \leq \nu > -\infty} |y_\nu| \leq \text{const}.$$

Then we have

THEOREM 3.5. *Assume that the conditions of Theorem 3.4 are valid and that the problems* (3.14), (3.15) *and* (3.16), (3.17) *have only the trivial solution. Then the estimate* (3.8) *holds.*

Let us consider the difference equation (2.7) with boundary conditions (2.7a) once more. The general solution of (3.12) with $\sup_\nu |y_\nu| \leq \text{const}$ is given by $y_\nu = \sigma_1 \kappa_1^\nu$ and, therefore, $h^{r-2}D_+^{r-2}y_0 = \sigma_1(\kappa_1 - 1)^{r-2} = 0$ implies that $\sigma_1 = 0$, thus, that the corresponding Eqs. (3.14), (3.15) have only the trivial solution. Obviously, the same is true for the Eqs. (3.16) and (3.17).

The stability results which we have derived here are independent of the particular differential equation, i.e., they depend only on the approximation of the nth derivative and the choice of the extra boundary conditions. If the solutions κ_i of the characteristic equations do not have the property that $|\kappa_i| \neq 1$, then this is no longer true and one cannot develop any general theory. We shall illustrate this by an example:

Consider the differential equation

$$du/dx = 0, \qquad u(0) = g_0,$$

and approximate it by

$$v_{\nu+1} - v_{\nu-1} = 0, \qquad v_0 = g_0, \qquad D_+v_0 + 2D_-v_N = g_1.$$

Then

$$v_\nu = \sigma_1 + \sigma_2(-1)^\nu,$$

where σ_1, σ_2 are determined by

$$v_0 = \sigma_1 + \sigma_2 = g_0,$$

$$h(D_+v_0 + 2D_-v_N) = (-2 + 4(-1)^N)\sigma_2 = hg_1,$$

i.e.,

(3.18) $(-2 + 4(-1)^N)v_\nu = (-2 + 4(-1)^N)g_0 + h((-1)^\nu - 1)g_1.$

Therefore, the v_ν converge to u. If we now change the differential equation to

$$du/dx - \alpha u = 0, \qquad u(0) = g_0,$$

and approximate it by

$$v_{\nu+1} - v_{\nu-1} - 2\alpha hv_\nu = 0, \qquad v_0 = g_0, \qquad D_+v_0 + 2D_-v_N = g_1,$$

then

$$v_\nu = \sigma_1 e^{\alpha x(1 + O(h^2))} + \sigma_2(-1)^\nu e^{-\alpha x(1 + O(h^2))}$$

and σ_1, σ_2 are determined by

$$\sigma_1 + \sigma_2 = 0,$$

$$h\sigma_1\alpha(e^{\alpha h/2} + 2e^{\alpha(1-h/2)} + O(h^2)) + \sigma_2(1 + 2^{-\alpha h})(-1 + 2e^{-\alpha}(-1)^N + O(h^2)) = hg_1.$$

An easy calculation shows that we do not get convergence if N is even and $\alpha = \log 2$.

4. Error Estimates.   We have already proved that the solutions of the difference equations converge to the solution of the differential equation. We shall now derive refined error estimates. We start with

LEMMA 4.1. *Assume that $\lambda = 0$ is not an eigenvalue of* (1.3). *Consider the equations*

(4.1) $L_h w_\nu = 0$

with boundary conditions

(4.2) $B_{lh}w = 0 \qquad \text{for } l \leqq n - 1,$

$$= h^{n-l}\tilde{g}_l \qquad \text{for } l \geqq n,$$

and assume that for the Eqs. (3.6), (3.7) *the estimate* (3.8) *holds. Then, there is a constant K_4, independent of h and $\tilde{g}_l$, such that, for the solutions of* (4.1), (4.2), *an estimate*

(4.3) $||w||_h \leqq K_4 \sum |\tilde{g}_l|$

is valid.

Proof. From (3.7a) and (3.8), it follows that an estimate of type (3.1) holds for w. Therefore, the estimate (4.3) follows by the same argument as the estimate (3.2).

In most cases, one can improve the above estimate. We can, for example, prove

LEMMA 4.2. *Assume that the conditions of Lemma* 4.1 *are fulfilled and that the Eqs.* (3.6) *have the form* (3.9) *and all solutions κ_i of the characteristic equation* (3.10) *are distinct. Then we get, instead of* (4.3), *an estimate*

(4.4) $||w||_h \leqq hK_5 \sum |\tilde{g}_l|.$

If, furthermore, $l_{\max} = \max_\nu l_\nu \leqq n - 2$ for the highest derivative appearing in the boundary conditions (1.2), *then*

(4.5) $||w||_h \leqq h^2 K_6 \cdot \sum |\tilde{g}_l|.$

Proof. Let y be the solution of

$$(4.6) \qquad S_0(0)y_{\nu-r} = 0, \qquad \sum_{j=0}^{N-n} H_{jl}y_j = \tilde{g}_l.$$

By assumption and Lemma 3.1,

$$y_\nu = \sum_{|\kappa_j|<1} \sigma_j \kappa_j^\nu + \sum_{|\kappa_j|>1} \sigma_j \kappa_j^{\nu-N},$$

with

$$|\sigma_j| \leq \text{const} \sum |\tilde{g}_l|, \qquad |\kappa_j| \neq 1.$$

Define $z_\nu^{(1)}$ by

$$(4.7) \qquad z_\nu^{(1)} = h^n \left(\sum_{|\kappa_j|<1} \sigma_j (\kappa_j - 1)^{-n} \kappa_j^\nu + \sum_{|\kappa_j|>1} \sigma_j (\kappa_j - 1)^{-n} \kappa_j^{\nu-N} \right),$$

then $D_+^n z_\nu^{(1)} = y_\nu$, and $v = w - z^{(1)}$ is the solution of

$$L_h v = hG, \qquad ||G||_h \leq \text{const} \sum |\tilde{g}_l|,$$

$$(4.8) \qquad \begin{aligned} B_{lh}v &= h^{n-l}\hat{g}_l \quad \text{for } l \leq n-1, \\ &= h^{n-l+1}\hat{g}_l \quad \text{for } l \geq n, \end{aligned} \qquad |\hat{g}_l| \leq \text{const} \sum |\tilde{g}_l|.$$

Therefore, $v = v^{\mathrm{I}} + v^{\mathrm{II}}$, where v^{I}, v^{II} are the solutions of

$$(4.9) \qquad \begin{aligned} L_h v_\nu^{\mathrm{I}} = hG_\nu, \qquad B_{lh}v^{\mathrm{I}} &= h^{n-l}\hat{g}_l \quad \text{for } l \leq n-1, \\ &= 0 \qquad \text{for } l \geq n, \end{aligned}$$

and

$$(4.10) \qquad \begin{aligned} L_h v_\nu^{\mathrm{II}} = 0, \qquad B_{lh}v_\nu^{\mathrm{II}} &= 0 \qquad \text{for } l \leq n-1, \\ &= h^{n-l+1}\hat{g}_l \quad \text{for } l \geq n, \end{aligned}$$

respectively. Then, the estimate (4.4) follows from (4.3) and (3.2).

Let $l_{\max} \leq n-2$. Then, we define $z^{(2)}$ by

$$(4.11) \qquad D_+^n z_\nu^{(2)} = \tilde{y}_\nu, \qquad z_0^{(2)} = D_+ z_0^{(2)} = \cdots = D_+^{n-1} z_0^{(2)} = 0,$$

where $\tilde{y}_\nu$ is the solution of

$$S_0(0)\tilde{y}_{\nu-r} = hG_\nu, \qquad \sum_{j=0}^{N-n} H_{jl}\tilde{y}_j = h\tilde{g}_l, \qquad \nu = r, \cdots, N-s.$$

(4.7) implies that there is a constant d with $0 \leq d < 1$ such that

$$|G_\nu| \leq \text{const}(d^\nu + d^{\nu-N}) \sum |\tilde{g}_l|$$

and therefore also

$$|\tilde{y}_\nu| \leq h\,\text{const}(d^\nu + d^{\nu-N}) \sum |\tilde{g}_l|.$$

Thus,

$$||D_+^j z^{(2)}||_h \leq h^2 \,\text{const} \sum |\tilde{g}_l| \quad \text{for } j = 0, 1, 2, \cdots, n-1.$$

$v^{(1)} = w - z^{(1)} - z^{(2)}$ is the solution of

$$L_h v^{(1)} = h^2 G_\nu^{(2)}, \qquad ||G^{(2)}||_h \leqq \text{const} \sum |\tilde{g}_l|.$$

$$B_{1h} v^{(1)} = h^2 \hat{g}_l \qquad \text{for } l \leqq n - 2,$$

$$|\hat{g}_l| \leqq \text{const} \sum |\tilde{g}_l|.$$

$$= h^{n-l+2} \hat{g}_l \quad \text{for } l \geqq n,$$

Therefore, (4.4) follows in the same way as above.

There is no doubt that Lemma 4.4 can be generalized considerably. For example, the assumption that the solutions κ_j of the characteristic equation are distinct is unnecessary. This can be seen by a simple perturbation argument.

Furthermore, it can be shown that the estimates are sharp.

It is now easy to derive error estimates: Let u be the solution of (1.1), (1.2) and substitute it into the difference equation. Then there is a natural number α, the order of the difference approximation, such that

$$L_h u_\nu = \tilde{F}_\nu + h^\alpha G_\nu,$$

(4.12)
$$B_{lh} u = \tilde{g}_l + h^\alpha f_l \quad \text{for } l \leqq n - 1,$$

$$= \tilde{g}_l + f_l \qquad \text{for } l \geqq n.$$

We get

THEOREM 4.1. *Assume that the conditions of Lemma* 4.1 *or* 4.2 *are fulfilled and let* v *denote the solution of the corresponding difference equations. Then*

$$||u - v||_h = h^\alpha K_2 \left(||G||_h + \sum_{l=0}^{n-1} |f_l| \right) + h^\sigma K_7 \sum_{l \leqq n} h^{l-n} |f_l|.$$

Here, $\sigma = 0, 1, 2$ *and* $K_7 = K_4, K_5, K_6$, *if the estimates* (4.3), (4.4), (4.5), *respectively, hold.*

Proof. The estimates follow in the usual way by writing down the difference equation for $u - v$ and then using a representation of type (4.9), (4.10).

It is easy to see that $\alpha = 2$ for the approximations (2.4), (2.4a) and (2.6), (2.6a). In these cases, the second sum in the error estimate does not appear. For (2.7), (2.7a), we have $\alpha = 4$ and the second sum is of order $O(h^r)$. For (2.7), (2.7b), we again have $\alpha = 4$ and the second term is of order $O(h^4)$ if we use a grid defined by $x_\nu = (\nu - 1)h$, $h = (N - 2)^{-1}$. Otherwise, it is of order $O(h^3)$.

One can always construct compact difference schemes, i.e., $r + s = n$, such that the truncation error can be expanded into power series $\sum h^2 \nu \phi_\nu(x)$ in h^2. Therefore, it is doubtful that one should use difference schemes with $r + s > n$ to increase the accuracy. Instead, one can use Fox's difference correction method [1] or Richardson extrapolation. Justifications of these methods are given in [3], [4] and depend on the estimate (3.2).

5. The Algebraic Eigenvalue Problem.

The eigenvalue problem for the differential equations can be considered as the limit of finite-dimensional problems. Therefore, we shall discuss the latter briefly.

Let X_p denote the p-dimensional vector space $x = (x_1, \cdots, x_p)'$. Denote by A a $p \times p$ matrix, and let λ_1 be an eigenvalue of A, i.e., a solution of

(5.1)
$$\text{Det } |A - \lambda I| = 0.$$

It is well known that the associated invariant subspace J is given by the projections

$$(5.2) \qquad J = P X_p, \qquad P = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta_1} (A - I)^{-1} \, dz.$$

Here, $\delta_1 > 0$ is a constant which is so small that there is no other eigenvalue λ with $|\lambda - \lambda_1| \leq \delta_1$. A basis of J can (for theoretical reasons) be constructed by solving the set of equations

$$(5.3) \qquad (A - \lambda_1)y_0 = 0, \qquad (A - \lambda_1 I)y_1 = y_0, \cdots, (A - \lambda_1 I)y_{i+1} = y_i, \cdots.$$

Now let B be another $p \times p$ matrix and let us consider the generalized eigenvalue problem

$$(5.4) \qquad\qquad\qquad Ax = \lambda Bx.$$

Without restriction, we may assume that

$$Ax = 0, \qquad Bx = 0 \quad \text{imply } x = 0.$$

Then, the eigenvalues are given by

$$(5.5) \qquad\qquad\qquad \text{Det } |(A - \lambda B)| = 0.$$

The invariant subspace J associated with an eigenvalue λ_1 is given by

$$(5.6) \qquad J = P X_p, \qquad P = \frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta_1} (A - zB)^{-1} B \, dz,$$

and a basis of J can be constructed by solving the equations

$$(5.7) \qquad (A - \lambda_1 B)y_0 = 0, \qquad (A - \lambda B)y_1 = By_0, \cdots, (A - \lambda B)y_{i+1} = By_i, \cdots.$$

If B is nonsingular, then we can write (5.4) in the form $B^{-1}Ax = \lambda x$, and (5.6), (5.7) follow directly from (5.2), (5.3). If B is singular, then we perturb B and consider

$$P(\epsilon) = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} (A - z\tilde{B})^{-1} \tilde{B} \, dz, \qquad \tilde{B} = B + \epsilon B_1.$$

Here, B_1 is chosen in such a way that $\tilde{B}$ is not singular in an interval $0 < \epsilon < \epsilon_0$. Observing that $P(\epsilon)$ is, for sufficiently small ϵ, an analytic function of ϵ, the relations (5.6) and (5.7) follow by a perturbation argument.

In applications, the generalized eigenvalue problem (5.4) is often written in the form

$$(5.8) \qquad\qquad Ax = \lambda Bx, \quad \text{with side conditions } Cx = 0.$$

Here A, B are $l \times p$ matrices, $l \leq p$ and C is a $p - l \times p$ matrix. (5.8) can, of course, be written in the form (5.4),

$$(5.9) \qquad\qquad\qquad \binom{A}{C}x = \lambda\binom{B}{0}x.$$

Let $\mathfrak{D}_1$ denote the subspace of vectors $x \in X_p$ with $Cx = 0$ and denote by $\mathfrak{X}(z)$ the operator

$$\mathfrak{X}(z)x = (A - zB)x, \qquad x \in \mathfrak{D}_1,$$

which maps $\mathfrak{D}_1$ into X_l. If z is not an eigenvalue of (5.9) then $\mathfrak{X}(z)^{-1}$ exists and we can write the invariant subspace connected with an eigenvalue λ_1 of (5.9) in the form

$$(5.10) \qquad J = PX_p, \qquad P = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta_1} \mathfrak{X}(z)^{-1} B \, dz.$$

This follows directly from (5.6) and (5.9) because

$$\left(\begin{pmatrix} A \\ C \end{pmatrix} - z \begin{pmatrix} B \\ 0 \end{pmatrix} \right)^{-1} \begin{pmatrix} B \\ 0 \end{pmatrix} x = \mathfrak{X}(z)^{-1} Bx.$$

Remark. If there is more than one eigenvalue inside the circle $|z - \lambda_1| \leqq \delta_1$, then J denotes the combined invariant subspace.

6. The Eigenvalue Problem for the Differential Equations. Let $\mathfrak{D} \in \mathfrak{B}$ denote the subspace of all $\phi \in C^n$ which fulfill the homogeneous boundary conditions (1.2), i.e., $B_l \phi = 0$. Then we can write the eigenvalue problem (1.3) in operator form

$$(6.1) \qquad \mathfrak{L}\phi = \lambda\phi, \qquad \phi \in \mathfrak{D}.$$

We assume that not all complex numbers λ are eigenvalues. Then the equation

$$(6.2) \qquad (\mathfrak{L} - zI)u = F, \qquad y \in \mathfrak{D},$$

has a unique solution for every z which is not an eigenvalue and every $F \in \mathfrak{B}$ and $\|(\mathfrak{L} - zI)^{-1}\| < \infty$ if z is not an eigenvalue. In fact, $(\mathfrak{L} - zI)^{-1}$ is a meromorphic function of z with poles at the eigenvalues. Let λ_1 be an eigenvalue and assume that $\delta > 0$ is a constant such that no other eigenvalue belongs to $\|z - \lambda_1\| \leqq \delta$. Then the invariant subspace $J(\lambda_1)$ which is associated with λ_1 is given by the projection

$$(6.3) \qquad J(\lambda_1) = P\mathfrak{B}, \qquad P = \frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} (\mathfrak{L} - zI)^{-1} \, dz.$$

It is well known that $J(\lambda_1)$ is finite-dimensional.

The operators L_h, S_h, defined by (2.1) and (2.19), respectively, map $\mathfrak{B}_h$ into the $N - (r + s) + 1$-dimensional vector space V_h consisting of the gridfunctions $(w_r, \cdots , w_{N-s})$. Let $\mathfrak{D}_h \in \mathfrak{B}_h$ denote the subspace of gridfunctions $v = (v_0, \cdots , v_N)'$ which fulfill the boundary conditions (2.14) and denote by $\mathfrak{L}_h$, $\mathfrak{S}_h$ the restriction of L_h, S_h to $\mathfrak{D}_h$. Now, we can write the discrete eigenvalue problem in the form

$$(6.4) \qquad \mathfrak{L}_h\psi = \lambda_h\mathfrak{S}_h\psi, \qquad \psi \in \mathfrak{D}_h,$$

which is a generalized eigenvalue problem of the form (5.8). For the computation of the invariant subspaces, we consider the equation

$$(6.5) \qquad (\mathfrak{L}_h - z\mathfrak{S}_h)v = S_hF, \qquad v \in \mathfrak{D}_h, \qquad S_hF = (S_hF_r, \cdots , S_hF_{N-s})' \in V_h.$$

We now make a number of assumptions:

Assumption 6.1. Let z be a complex number which is not an eigenvalue of $\mathfrak{L}$. We assume that there is constant $h_0 > 0$ such that

$$\sup_{0 < h \leqq h_0} \|(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1}\|_h = K < \infty.$$

The following lemma is valid.

LEMMA 6.1. *Let Ω denote a compact set in the complex z-plane which does not contain any eigenvalue of $\mathfrak{L}$. If Assumption 6.1 holds, then there is a constant $h_0 > 0$ such that*

$$\sup_{z \in \Omega, 0 < h \leq h_0} ||(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1}||_h = K_\Omega < \infty .$$

Proof. Let $z_0 \in \Omega$. By Assumption 6.1,

$$(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} = (\mathfrak{L}_h - z_0\mathfrak{S}_h)^{-1}(I + (z - z_0)\mathfrak{S}_h(\mathfrak{L}_h - z_0\mathfrak{S}_h))^{-1}$$

is uniformly bounded in a whole neighborhood of z_0. The lemma follows from the theorem of Heine-Borel.

Let $z \in \Omega$, $F \in \mathfrak{B}$ be a fixed function and define $\tilde{F} = S_h F \in V_h$ by $\tilde{F}_\nu = S_h F_\nu$, $\nu = r, r + 1, \cdots , N - s$. By Assumption 6.1, the function $v = (\mathfrak{L}_h - z\mathfrak{S}_h)^{-1}S_h F$ exists for all $h \leq h_0$ and, by Lemma 3.2, we can construct the interpolated function $w = \text{Int}_n v$. We now make

Assumption 6.2. For every $z \in \Omega$ and every $F \in \mathfrak{B}$,

$$\lim_{h \to 0} ||w - u|| = \lim_{h \to 0} || \text{Int}_n(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} S_h F - (\mathfrak{L} - zI)^{-1}F|| = 0.$$

LEMMA 6.2.

$$\lim_{h \to 0} \sup_{z \in \Omega} ||w - u|| = 0.$$

Proof. From Assumption 6.1, it follows that the w's are analytic functions of z and that the first derivatives $\partial w / \partial z$ are uniformly bounded for all $z \in \Omega$ and all $h < h_0$. Therefore, the lemma follows from Assumption 6.2.

LEMMA 6.3. *Let $h_\mu \to 0$, $F^{(\mu)} \in \mathfrak{B}$ be sequences with $\lim_{\mu \to \infty} ||F^{(\mu)} - F|| = 0$. Denote by u the solution of $(\mathfrak{L} - zI)u = F$ and by $w^{(\mu)}$ the solution*

$$w^{(\mu)} = \text{Int}_n(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} S_h F^{(\mu)}, \qquad h = h_\mu.$$

Then

$$\lim_{h \to 0} \sup_{z \in \Omega} ||u - w^{(\mu)}|| = 0.$$

Proof. The lemma obviously follows from Assumption 6.1 and Lemma 6.1.

Assumption 6.3. Let $h_\mu \to 0$, $F^{(\mu)} \in \mathfrak{B}$ with $\sup_\mu ||F^{(\mu)}|| < \infty$ be sequences. Then, we assume that the solutions of the discrete problem

$$w^{(\mu)} = \text{Int}_n(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} S_h F^{(\mu)}, \qquad z \in \Omega, h = h_\mu,$$

form a compact sequence.

Assumptions 6.1, 6.2 are the usual stability and consistency requirements. They guarantee that the eigenvalues of the discrete problem converge to the eigenvalues of the differential equations. However, the invariant subspaces need not converge. For that, the last assumption is essential. Consider, for example, the eigenvalue problem

$$dy/dx - \lambda y = 0, \qquad y(0) = y(1),$$

and approximate it by the leap-frog scheme

$$v_{\nu+1} - v_{\nu-1} - 2h\lambda v_\nu = 0, \qquad v_0 = v_N, \qquad v_2 - v_0 = 0.$$

It is easy to see that Assumptions 6.1 and 6.2 are fulfilled and $\lambda = 0$ is an eigenvalue for both problems. However, the corresponding invariant subspaces are $y = \text{const}$ and $v_\nu = \text{const} + \sigma(-1)^\nu$, respectively.

The above assumptions are natural in our framework because we get from Theorem 3.1 without difficulty:

THEOREM 6.1. *Assume that for every z there is an a priori estimate of type* (3.1) *for the solutions of* (6.5). *Then, the Assumptions* 6.1–6.3 *are fulfilled.*

Let λ_1 be an eigenvalue of the differential equation (6.1) and $J(\lambda_1)$ the corresponding invariant subspace. (5.10) shows that

$$J_h(\lambda_{\nu h}) = P_h \mathfrak{L}_h, \qquad P_h = \int_{|z-\lambda_1|=\delta} (\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} S_h \, dz$$

denotes the combined invariant subspace of the eigenvalues λ_{jh} of the discrete eigenvalue problem which lie inside the circle $|z - \lambda_1| \leq \delta$. We shall now imbed $J_h(\lambda_{\nu h})$ in $\mathfrak{B}$ by interpolation. We let

$$J(\lambda_{\nu h}) = \text{Int}_n \, J_h(\lambda_{\nu h})$$

and prove

THEOREM 6.2. $\lim_{h \to 0} J(\lambda_{\nu h}) = J(\lambda_1)$, *i.e., for sufficiently small h the dimension of $J(\lambda_{\nu h})$ is the same as that of $J(\lambda_1)$, and there is a basis of $J(\lambda_{\nu h})$ which converges to a basis of $J(\lambda_1)$.*

Proof. Let $\phi_1, \cdots, \phi_p$ be a basis of $J(\lambda_1)$. Then, $\psi_i(x, h) = \text{Int}_n P_h \phi_i$ belongs to $J(\lambda_{\nu h})$. By Assumption 6.2 and Lemma 6.2,

$$\lim_{h \to 0} \psi_i(x, h) = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} (\mathfrak{L} - zI)^{-1} \, dz\phi_i = \phi_i.$$

Therefore, we need only show that the $\psi_i(x, h)$ define, for sufficiently small h, a basis of $J(\lambda_{\nu h})$. Assume that there is a sequence $h_\mu \to 0$ such that, for every $h = h_\mu$, there is a function $\psi(x, h) \in J(\lambda_{\nu h})$ with

$$\|\psi(x, h)\| = 1 \quad \text{and} \quad \int_0^1 \psi_j^*(x, h)\psi(x, h) \, dx = 0, \qquad j = 1, 2, \cdots, p.$$

Then

$$\psi(x, h) = -\text{Int}_n \frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} (\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} \, dz S_h \psi(x, h).$$

By Assumption 6.1, the sequence $(\mathfrak{L}_h - z\mathfrak{S}_h)^{-1} S_h \psi(x, h)$, $h = h_\mu$, is equicontinuous with respect to z. Therefore, Assumption 6.3 implies that the sequence $\psi(x, h)$, $h = h_\mu$, is compact and we can, without loss of generality, assume that $\lim \psi(x, h) = \phi(x)$. Then, it follows from Assumption 6.2 and Lemma 6.2, that

$$\phi(x) = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} (\mathfrak{L} - zI)^{-1} \, dz\phi(x) \in J(\lambda_1).$$

Furthermore, $\|\phi\| = 1$ and ϕ is orthogonal to all ϕ_i. This is impossible, and the theorem is proved.

We have thus proved that the eigenvalues and corresponding invariant subspaces of the discrete problem converge to the eigenvalues and invariant subspaces of the

continuous problem. In most cases, one can derive sharper error estimates and can show that Richardson extrapolation is possible. We want to show

THEOREM 6.3. *Consider any compact set Ω which does not contain any eigenvalue and assume that there are natural numbers α and q such that (6.2) has, for every $F \in C^{\alpha p}$, $p \leq q$ and any $z \in \Omega$, a solution $u(x, z) \in C^{\alpha p}$. Assume, furthermore, that for the corresponding solution of (6.5), there is an expansion*

$$(6.6) \qquad \begin{aligned} v &= v(x, h, z) \\ &= u(x, z) + h^{\alpha} u_1(x, z) + \cdots + h^{\alpha(p-1)} u_{p-1}(x, z) + O(h^{\alpha p}), \qquad x = x_{\nu}. \end{aligned}$$

Here, the $u_i(x, z) \in C^{\alpha(p-i)}$ as functions of x and are continuous functions of z. Then, there exists a basis $\{\psi_i(x_{\nu}, h)\}$ of $J_h(\lambda_{\nu h})$ such that

$$(6.7) \quad \psi_i(x, h) = \phi_i(x) + h^{\alpha}\phi_{i1}(x) + \cdots + h^{\alpha(q-1)}\phi_{iq-1}(x) + O(h^{\alpha q}), \qquad x = x_{\nu},$$

where $\phi_i \in C^{\alpha q}$ denotes a basis of $J(\lambda_1)$ and $\phi_{i\nu} \in C^{\alpha(q-\nu)}$.

Furthermore, if λ_1 has multiplicity τ, then there are precisely τ eigenvalues $\lambda_{\nu h}$, counted according to their multiplicity, with $|\lambda_{\nu h} - \lambda_1| < \infty$ and

$$(6.8) \qquad \frac{1}{\tau} \sum_{\nu=1}^{\tau} \lambda_{\nu h} = \lambda_1 + h^{\alpha}\lambda_{11} + \cdots + h^{\alpha(q-1)}\lambda_{1q-1} + O(h^{\alpha q}).$$

Proof. Let $F = \phi_i \in J(\lambda_1)$. Then $\phi_i \in C^{\alpha q}$ and, by (6.6),

$$\psi_i(x, h) = P_h \phi_i = -\frac{1}{2\pi i} \int_{|z-\lambda_1|=\delta} v(x, h, z) \, dz$$

$$= \phi_i(x) - \frac{h^{\alpha}}{2\pi i} \int_{|z-\lambda_1|=\delta} u_1(x, z) \, dz + \cdots$$

and (6.7) follows directly.

Let $z_0 \in \Omega$ and $\phi_i \in J(\lambda_1)$. Then also $(\mathfrak{L} - z_0 I)^{-1}\phi_i \in J(\lambda_1)$ and, therefore,

$$(6.9) \qquad (\mathfrak{L} - z_0 I)^{-1}\phi_i = \sum_{k=1}^{\tau} a_{ik}\phi_k.$$

The $\tau \times \tau$ matrix $A = (a_{ik})$ has the sole eigenvalue $\lambda = (\lambda_1 - z_0)^{-1}$. In the same way, if $\psi_i \in J_h(\lambda_{\nu h})$, then

$$(6.10) \qquad (\mathfrak{L}_h - z_0\mathfrak{S}_h)^{-1}S_h\psi_i = \sum_{k=1}^{\tau} b_{ik}\psi_k$$

and the $\tau \times \tau$ matrix $B = (b_{ik})$ has the eigenvalues $(\lambda_{\nu h} - z_0)^{-1}$, $\nu = 1, 2, \cdots, \tau$. We want to derive a relation between A and B. By (6.7),

$$\sum_{k=1}^{\tau} b_{ik}\psi_k = (\mathfrak{L}_h - z_0\mathfrak{S}_h)^{-1}S_h\phi_i + h^{\alpha}(\mathfrak{L}_h - z_0\mathfrak{S}_h)^{-1}S_h\phi_{i1} + \cdots.$$

Now, apply (6.6) to all of the terms on the right-hand side. Then, (6.9) gives us

$$\sum_{k=1}^{\tau} b_{ik}\psi_k = (\mathfrak{L} - z_0 I)^{-1}\phi_i + h^{\alpha}w_{1i}(x) + h^{2\alpha}w_{2i}(x) + \cdots$$

$$= \sum_{k=1}^{\tau} a_{ik}\phi_k + h^{\alpha}w_{1i}(x) + h^{2\alpha}w_{2i}(x) + \cdots.$$

We can now use (6.7) to express ϕ_k in terms of the ψ_k and get

$$\sum_{\kappa=1}^{\tau} b_{ik}\psi_k = \sum_{k=1}^{\tau} a_{ik}\psi_k + R_i.$$

$R_i = h^\alpha \bar{w}_{1i}(x) + h^{2\alpha} \bar{w}_{2i}(x) + \cdots \in J_h(\lambda_{\nu h})$ and therefore (6.7) implies that

$$R_i = \sum_{\sigma=1}^{q-1} h^{\alpha\sigma} c_{ik}^{(\sigma)} \psi_k + O(h^{\alpha q}).$$

Therefore,

$$B = A + hC_1 + h^{2\alpha}C_2 + \cdots + h^{\alpha(q-1)}C_{q-1} + O(h^{\alpha q}), \qquad C_\sigma = (c_{ik}^{(\sigma)}),$$

and (6.8) follows without difficulty.

We have thus shown that the error behavior of the eigenvalue problem is the same as that of the inhomogeneous problem, the estimates for our examples are obvious. If the dimensions of the invariant subspaces are always one, then there are no practical problems either. Otherwise, we have to cope with two difficulties:

(1) If there is an eigenvalue of multiplicity $\tau > 1$, then it might be difficult to decide which are the τ eigenvalues $\lambda_{\nu h}$ which converge to λ. In general, this difficulty can only be overcome by some a priori information of the differential equation.

(2) Assume that (6.7) holds and that we have constructed a basis $\{\psi_1(x, h), \cdots, \psi_\tau(x, h)\}$ for the invariant subspace $J_h(\lambda_{\nu h})$ for a number of values $h = h_0, h_1, \cdots$ with $h_i/h_0 = \rho_i =$ natural number. In general, Richardson extrapolation will not work directly. Though there is a basis in $J_h(\lambda_{\nu h})$ as described by (6.7), the particular bases we have constructed need not have that property. We can, however, proceed in the following way.

(1) Compute bases for $h = h_0, h_1, \cdots$ and consider these bases for $x = \bar{x}_\nu = \nu h_0$.

(2) Now change the bases for $h = h_1, h_2, \cdots$ to bases $\{\tilde{\psi}_i(x, h_\sigma)\}$, $x = \bar{x}_\nu$, by demanding

$$(6.11) \qquad ||\tilde{\psi}_i(x, h_\sigma) - \psi_i(x, h_0)||_2 = \sum_{\nu=0}^{N_0} |\tilde{\psi}_i(\bar{x}_\nu, h_\sigma) - \psi_i(\bar{x}_\nu, h_0)|^2 h_0 = \min.$$

We have

THEOREM 6.4. *Richardson extrapolation is possible for the bases*

$$\{\psi_i(x, h_0)\}, \{\tilde{\psi}_i(x, h_1)\}, \cdots, x = \tilde{x}_\nu.$$

Proof. By (6.7), there are bases $\{\hat{\psi}_k(x, h_\sigma)\}$ for which

$$\hat{\psi}_k(x, h_\sigma) = \phi_k(x) + h^\alpha \hat{\phi}_{k1}(x) + \cdots + h^{\alpha(q-1)} \hat{\phi}_{kq-1}(x) + O(h^{\alpha q}), \qquad x = \tilde{x}_\nu$$

and

$$(\hat{\psi}_k(x, h_\sigma), \hat{\psi}_l(x, h_\sigma))_2 = \sum_{\nu=0}^{N_0} \hat{\psi}_k(x_\nu, h_\sigma)\hat{\psi}_l(\tilde{x}_\nu, h_\sigma)h_0 = 0 \quad \text{for } k \neq l,$$

$$= 1 \quad \text{for } k = l.$$

Therefore,

$$\tilde{\psi}_i(x, h_\sigma) = \sum a_{ik}\hat{\psi}_k(x, h_\sigma)$$

and (6.11) implies

$$a_{ik} = (\hat{\psi}_k(x, h), \ \psi_i(x, h_0))_2.$$

This proves the theorem.

Computer Sciences Department
University of Uppsala
Uppsala, Sweden

1. L. Fox, *The Numerical Solution of Two-Point Boundary Problems in Ordinary Differential Equations*, Oxford Univ. Press, Oxford, 1957. MR **21** #972.

2. H. B. Keller, "Accurate difference methods for linear ordinary differential systems subject to linear constraints," *SIAM J. Numer. Anal.*, v. 6, 1969, pp. 8–30. MR **40** #6776.

3. V. Pereyra, "Iterated deferred corrections for nonlinear operator equations," *Numer. Math.*, v. 10, 1967, pp. 316–323. MR **36** #4812.

4. H. J. Stetter, "Asymptotic expansions for the error of discretization algorithms for non-linear functional equations," *Numer. Math.*, v. 7, 1965, pp. 18–31. MR **30** #5505.