

UNIVERSITY OF CAPE TOWN
DEPARTMENT OF MATHEMATICS

The Stability of Linear Operators

by

H.E.G. Colburn

A thesis prepared under the supervision of Dr. W. Kotzé,
in partial fulfilment of the requirements of the degree
of Master of Science in Mathematics.

Copyright by the University of Cape Town
1969

The copyright of this thesis is held by the
University of Cape Town.

Reproduction of the whole or any part
may be made for study purposes only, and
not for publication.

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

I N D E X

INTRODUCTION	(i)
NOTATION	(ii)
CHAPTER 1. Stable and strictly stable linear operators.	
Section 1. Concepts and definitions.	1.
2. Uniform boundedness and the adjoint operator.	3
3. Stable operators on a finite dimensional Banach space. The matrix theorems of Kreiss and Buchanan.	5.
4. Strictly stable linear operators.	14.
CHAPTER 2. The relation between stability and convergence.	
Section 1. The Equivalence theorem of P.D. Lax.	25.
2. Special cases of the Equivalence theorem.	28.
CHAPTER 3. The stability of finite difference equation.	
Section 1. Stability in L^2 ; the Lax-Richtmyer theory.	30.
2. Stability in L^p , $p \neq 2$.	35.
3. The variable coefficient problem.	38.
CHAPTER 4. Dahlquist's stability theory	
Section 1. Multi-step algorithms for ordinary differential equations of the first order.	44.
2. Maximum order stable difference operators.	46.
CHAPTER 5. Discretizations.	
Section 1. The concept of a discretization.	52.
2. The strong stability concept of Stetter.	54.
3. The stability of a non-linear algorithm.	56.
APPENDIX A. A well-posed Cauchy problem.	60.
APPENDIX B. Perturbations of stable operators.	62.
APPENDIX C. The uniform convergence of matrix powers.	63.
REFERENCES.	65.

I N T R O D U C T I O N

In the approximation and solution of both ordinary and partial differential equations by finite difference equations, it is well-known that for different ratios of the time interval to the spatial intervals widely differing solutions are obtained. This problem was first attacked by John von Neumann using Fourier analysis, see for example [29]. It has also been studied in the context of the theory of semi-groups of operators, see [30].

It seemed that the problem could be studied with profit if set in a more abstract structure. The concepts of the stability of a linear operator on a (complex) Banach space and the stability of a Banach sub-algebra of operators were formed in an attempt to generalize the matrix theorems of H.O. Kreiss, see [7] as applied to the L^2 stability problem.

Chapter 1 deals with the stability and strict stability of linear operators. The equivalence of stability and convergence is discussed in Chapter 2 and special cases of the Equivalence Theorem are considered in Chapters 3 and 4. In Chapter 5 a brief account of the theory of discretizations is given and used to predict instability in non-linear algorithms.

I should like to express my deep appreciation to my supervisor Dr. W. Kotzé without whose constant encouragement and many helpful suggestions, this thesis could not have been written.

NOTATION

The following symbols are used extensively in the text:

\mathbb{N} will denote the set of natural numbers,

\mathbb{I} will denote the set of integers and

\mathbb{I}^+ the set of positive integers.

If T is a mapping then $\mathcal{D}(T)$ will denote the domain of definition of T , $\mathcal{R}(T)$ the range of T .

If S is a set, then the identity map I is defined by $I(x) = x$ for all $x \in S$.

$O(h)$ means "of the order of h ".

Occasionally the abbreviation "iff" is used for "if and only if".

CHAPTER 1

Stable and strictly stable linear operators.

Section 1. Concepts and Definitions.

Let X be a Banach space of functions defined on a compact Hausdorff topological space S . Let $L(X, X)$ denote all linear operators defined on X .

A will denote a subalgebra of $L(X, X)$ with respect to the usual operations of operator addition, scalar multiplication and operator multiplication so that A is furthermore a Banach algebra with respect to the uniform (supremum) norm.

Definition 1.1 : A linear operator T on X is said to be stable if the family $\{T^n : n \in \mathbb{N}\}$ is uniformly bounded.

Definition 1.2 : A is a stable Banach subalgebra of $L(X, X)$ if there exists constant K s.t. for all $T \in A$ $\|T\| \leq K$.

Hence all members of A are bounded (in fact, stable) and we shall consider A to be a subalgebra of $B(X, X)$: all bounded linear operators on X , denoted hereafter by $B(X)$.

Definition 1.3 : Let $T \in B(X)$. The resolvent set $\rho(T)$ of T is defined by

$$\rho(T) = \{\lambda \in \mathbb{C} : (\lambda I - T)^{-1} \text{ exists as a bounded operator with dense domain}\}.$$

The spectrum $\sigma(T)$ of T is the complement of the resolvent set of T .

Note T bounded, X complete, then $B(X)$ is a Banach algebra and

$\lambda \in \rho(T)$ iff $(\lambda I - T)^{-1}$ exists in the sense that it is bounded, everywhere defined and

$$(\lambda I - T)^{-1}(\lambda I - T) = (\lambda I - T)(\lambda I - T)^{-1} = I.$$

Let us denote $(\lambda I - T)^{-1}$ by $R(\lambda, T)$, the resolvent of T ; $D(T)$ will denote the domain of the operator T .

We have the following lemma

Lemma 1.4 : (i) T a continuous linear operator on a Banach space X .

If $\lambda \in \rho(T)$, then

$$D(R(\lambda, T)) = X.$$

(ii) $\rho(T)$ is open and $R(\lambda, T)$ is holomorphic in $\rho(T)$.

(iii) If $\lambda, \mu \in \rho(T)$ then $R(\lambda, T)$ and $R(\mu, T)$ commute.

Proof. See for e.g. Dunford and Schwartz [1].

Definition 1.5 : $r(T) = \sup\{|\sigma(T)|\}$ is called the spectral radius of T .

Let us recall the well-known spectral radius theorem.

Theorem 1.6 : Let $T \in B(X)$. Then $r(T) \leq \|T^n\|^{1/n} \leq \|T\|$ and $\|T^n\|^{1/n}$ converges to $r(T)$ as $n \rightarrow \infty$.

Proof. See Taylor [2] p.262.

Corollary 1.7 : If T is a stable operator on X , then $r(T) \leq 1$.

Proof. T stable implies $\|T^n\| \leq K$, some K .

Theorem 1.8 : If T is a stable operator then there exists a constant C_R such that for all $\lambda \in \mathbb{C}$ with $|\lambda| > 1$

$(\lambda I - T)^{-1}$ exists and moreover

$$\|(\lambda I - T)^{-1}\| \leq \frac{C_R}{|\lambda| - 1}.$$

Proof. Since T is stable, $\sigma(T)$ is contained in the closed unit ball.

Hence for $|\lambda| > 1$, $\lambda \in \rho(T)$ and the resolvent exists.

$$\begin{aligned} \text{Also } \|(\lambda I - T)^{-1}\| &= \|\lambda^{-1}(I + \lambda^{-1}T + \lambda^{-2}T^2 + \dots)\| \\ &\leq \left\| \sum_{n=0}^{\infty} \lambda^{-(n+1)} T^n \right\| \\ &\leq M \frac{1}{|\lambda| - 1}. \end{aligned}$$

This theorem gives a limit on how fast the resolvent of T can grow as the unit ball is approached and is therefore called the resolvent condition.

Theorem 1.9 : T is a stable operator on X iff T^n is a Lipschitz continuous map for all n .

Proof. Necessity: $\|T^n\| \leq K$ all $n \in \mathbb{N}$

$$\text{hence } \sup_{x \in X} \frac{\|T^n x\|}{\|x\|} \leq K$$

$$\text{Let } x = y - z, \frac{\|T^n(y-z)\|}{\|y-z\|} \leq K$$

$$\text{i.e. } \|T^n(y-z)\| \leq K \|y-z\|.$$

Sufficiency : $\|T^n y - T^n z\| \leq L \|y-z\|$

$$\therefore \sup_X \frac{\|T^n y - T^n z\|}{\|y-z\|} \leq L$$

$$\text{i.e. } \|T^n\| \leq L.$$

Section 2. Uniform boundedness and the adjoint operator.

If $T \in B(X, X)$, we denote by X^* the topological dual of X and by T^* the adjoint of T .

Lemma 2.1 : (i) $\|T\| = \|T^*\|$.

(ii) T^* is cont. if X^* has the induced topology of X .

- (iii) T has a bounded inverse T^{-1} defined on all of X
 iff T^* has a bounded inverse T^{*-1} defined on all of X
 and if both exist $(T^*)^{-1} = (T^{-1})^*$
- (iv) $(T_1 + T_2)^* = T_1^* + T_2^*$
- (v) $(\alpha T)^* = \alpha T^*$
- (vi) Let A, B and C be Banach spaces, then $T \in B(A, B)$,
 $S \in B(B, C) \Rightarrow (TS)^* = S^* T^*$
- (vii) $\sigma(T) = \sigma(T^*)$
 and $R(\lambda, T) = R(\lambda, T^*)$ for all $\lambda \in \rho(T) = \rho(T^*)$.
- (viii) The map $\phi : B(A) \rightarrow B(A^*)$ defined by

$$\phi(T) = T^* \quad \text{for all } T \in B(A)$$
 is an isometric isomorphism.
- (ix) T^{**} is an extension of T and $T^{**} = T$ iff
 X is reflexive.

Proof. See for e.g. Taylor [2].

Proposition 2.2 : To each stable operator $T \in B(X)$, there corresponds a stable $T^* \in B(X^*)$.

Proof. Obviously $T^* \in B(X^*)$ and the result follows from property (vi) since T commutes with itself.

Proposition 2.3 : To each stable Banach subalgebra $A \subset B(X)$ there corresponds a stable Banach subalgebra $Y \subset B(A^*)$.

Proof. Consider the isometric isomorphism ϕ of property (viii) above.

Let $\phi(X) = Y$.

Trivially Y is a Banach subalgebra of $B(X^*)$ and for all $S \in Y$

$\|S\| \leq K$ by (i).

If X is reflexive, then the converses of the two propositions above hold true.

Proposition 2.4 : X reflexive, then the map $\phi : B(X) \rightarrow B(X^*)$ of property (viii) is 1-1 and onto.

Proof. Follows immediately from property (ix).

Section 3. Stable operators on a finite dimensional Banach space. The matrix theorems of Kreiss and Buchanan.

Suppose X is finite dimensional, of dimension p say. Let F be a family of square matrices on X .

Definition 3.1 : F is said to be a stable family of matrices if for all

$$\begin{aligned} A \in F \\ \|A^n\| \leq K \quad \text{all } n \in \mathbb{I} \end{aligned} \quad (1)$$

Three characterizations of the stability of F are given in the following theorem.

Theorem 3.2 : (Kreiss). The following statements are equivalent:

- (A) F is a stable family of matrices.
- (R) Each member A of F satisfies the resolvent condition. (2)
- (S) There exists constants C_S and C_B and to each $A \in F$ a non-singular matrix S s.t.

$$(i) \|S\|, \|S^{-1}\| \leq C_S$$

- (ii) $B = SAS^{-1}$ is upper triangular and its off-diagonal elements satisfy

$$|b_{ij}| \leq C_B (1 - |\lambda_j|) \quad (3)$$

where λ_j are the eigenvalues of A .

(H) There exists a constant C_H and to each $A \in F$ a positive definite Hermitian matrix H such that

$$C_H^{-1} I \leq H \leq C_H I$$

and $A^* H A \leq H.$ (4)

Proof. There is no loss in generality in assuming that a preliminary unitary transformation has been carried out on A to put it in upper triangular form. The diagonal elements are then the eigenvalues λ_1 , and we let $a_{1j}, j > 1$, denote the off-diagonal elements. We also set $\zeta_1 = z - \lambda_1$ and $\zeta = |z| - 1$, where z is a complex number, and use the letter C , with or without modifiers, to denote positive constants. Before proceeding with the main proofs we need two lemmas.

Lemma 1 : If a 2×2 upper triangular matrix A satisfies the resolvent condition with a constant C , then

$$|a_{12}| \leq C' \max(1 - |\lambda_2|, |\lambda_1 - \lambda_2|),$$
 (5)

where $C' \leq 16C$.

Proof. Applying (2) to A and considering the sole off-diagonal element of $(A - zI)^{-1}$ gives

$$\left| \frac{a_{12}}{\zeta_1 \zeta_2} \right| \leq \frac{C}{\zeta}.$$
 (6)

Then by substituting $z = 3$ here, we obtain $|a_{12}| \leq 8C$, so that (5) is clearly satisfied if $|\lambda_2| \leq \frac{1}{2}$. On the other hand, if $|\lambda_2| > \frac{1}{2}$ we put $z = t/\bar{\lambda}_2$, where $t > 1$, so that (6) yields

$$|a_{12}| \leq C \frac{(t - |\lambda_2|^2)(t - \bar{\lambda}_2 \lambda_1)}{|\lambda_2|(t - |\lambda_2|)}$$

On letting $t \rightarrow 1$, there results $|a_{12}| \leq 3C(1 - \bar{\lambda}_2 \lambda_1)$. But

$$|1 - \bar{\lambda}_2 \lambda_1| = |1 - |\lambda_2|^2 + \bar{\lambda}_2(\lambda_2 - \lambda_1)| \\ \leq 3 \max \{1 - |\lambda_2|, |\lambda_1 - \lambda_2|\},$$

so that (5) holds for this case also. By interchanging λ_1 and λ_2 we can in fact prove the stronger result

$$|a_{12}| \leq 16C \max \{\min(1 - |\lambda_1|, 1 - |\lambda_2|), |\lambda_1 - \lambda_2|\}.$$

Lemma 2: If an $m \times m$ upper triangular matrix A satisfies the resolvent condition (2) with constant C_1 , and if all its off-diagonal elements except the upper right element a_{1m} satisfy

$$|a_{ij}| \leq C_2(1 - |\lambda_j|), \quad (7)$$

then

$$|a_{1m}| \leq C_3 \max(1 - |\lambda_m|, |\lambda_1 - \lambda_m|), \quad (8)$$

where

$$C_3 \leq 16C_1 (1 + (m-2) C_2^2)^{1/2}.$$

Proof. This is an extension of Lemma 1, to which it is reduced by the following method. We permute the 2nd and the m th rows and columns of $A - zI$, which clearly leaves the resolvent condition unchanged, and then

$$E = \begin{vmatrix} E_1 & E_2 \\ E_3 & E_4 \end{vmatrix} = \begin{vmatrix} -z_1 & a_{1m} & a_{13} & \cdots & a_{1,m-1} & a_{12} \\ 0 & -z_m & 0 & \cdots & 0 & 0 \\ \hline 0 & a_{3m} & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ \cdot & \cdot & & & & \\ 0 & a_{m-1,m} & & E_4 & & \\ 0 & a_{2m} & & & & \end{vmatrix}$$

where the detailed form of E_4 is not required. Now it is clear that $E_3 E_2 = 0$ and, indeed, $E_3 E_1^{-1} E_2 = 0$. Hence if we perform a triangular decomposition of E into $E = LU$ and denote by l_p the p th order

order unit matrix, we have

$$L = \begin{vmatrix} I_2 & 0 \\ E_3 E_1^{-1} & I_{m-2} \end{vmatrix} \quad U = \begin{vmatrix} E_1 & E_2 \\ 0 & E_4 \end{vmatrix}$$

The resolvent condition therefore gives, for any vector u ,

$$|E^{-1}u|^2 = u^*(L^{-1})^*(U^{-1})^*U^{-1}L^{-1}u \leq \left| \frac{C_1}{\xi} \right|^2 |u|^2.$$

Putting $u = Lv$, where only the first two elements of v (which sub-vector we denote by v_1) are taken to be nonzero, we have

$$|E_1^{-1}v_1|^2 \leq \left| \frac{C_1}{\xi} \right|^2 |Lv|^2.$$

But

$$\begin{aligned} |Lv|^2 &= |v_1|^2 + |E_3 E_1^{-1} v_1|^2 \\ &\leq \left| 1 + \frac{m-1}{m} \frac{a_{1m}}{m} \right|^2 |v_1|^2 \\ &\leq (1 + (m-2)C_2^2) |v_1|^2, \end{aligned}$$

so that E_1 satisfies the resolvent condition

$$|E_1^{-1}| \leq \frac{C_1 (1 + (m-2)C_2^2)^{1/2}}{\xi}$$

and Lemma 1 can be applied to give the desired result.

We can now prove the main part of Kreiss' theorem.

Theorem. (R) implies (S) implies (H).

Proof. The first step is to note that if A satisfies (R) then, because it is triangular, each corner of A (upper left or lower right principal submatrix) also satisfies it with the same constant C_R . Similarly, a corner of a corner satisfies (R). Hence we can easily obtain the desired inequalities (3) for the first upper diagonal, with $C_B = 16C_R$. For each such element lies in a 2×2 corner of a corner of A , so that

Lemma 1 can be applied to give

$$|a_{i,i+1}| \leq 16C_R \max(1 - |\lambda_{i+1}|, |\lambda_i - \lambda_{i+1}|).$$

If $1 - |\lambda_{i+1}| \geq |\lambda_i - \lambda_{i+1}|$, then (3) is already satisfied: otherwise, $a_{i,i+1}$ can be annihilated by a bounded similarity transformation. For, in general, a_{ij} is annihilated by $S_{ij}AS_{ij}^{-1}$ where $S_{ij} = I + T_{ij}$, $S_{ij}^{-1} = I - T_{ij}$, and T_{ij} is a matrix all of whose elements are zero except the (i,j) th which has the value $t_{ij} = a_{ij}/(\lambda_i - \lambda_j)$. Thus, when the transformation is needed, $|t_{i,i+1}|$ is bounded by $16C_R$, and by composing at most $n - 1$ such transformations we fulfill the requirements of (S) for the first upper diagonal with $C_S \leq 1 + 16C_R$.

To continue this process to succeeding upper diagonals we use Lemma 2. When the first $m - 2$ upper diagonals have been made to satisfy (3), each element of the $(m - 1)$ st appears as the top right element of an $m \times m$ corner of a corner of A to which this lemma can be applied. The resulting inequalities for these elements then allow another set of elementary similarity transformations of the type defined above to be applied to yield (3) for the $(m - 1)$ st upper diagonal. Each such transformation $S_{ij}AS_{ij}^{-1}$ affects only elements in column j above the element a_{ij} , and in row i to the right. Thus the upper diagonals already dealt with are unchanged, and the process can be completed in a finite number of steps. (However, at each stage the constants C_1 and C_2 of the lemma are increased by further factors. As a result of the similarity transformations occurring after one stage, the constant C_1 in the resolvent condition for the next stage is larger by a factor $(1 + C_3)^2$; and C_2 is the C_3 of the previous stage.) Thus (R) implies (S) and the constants C_S and C_B can be expressed in terms of C_R and n alone.

and C_B can be expressed in terms of C_R and n alone.

To prove (H) we introduce, with Kreiss, the diagonal matrix

$$D = \begin{vmatrix} d & & & \\ & d^2 & & \\ & & \ddots & \\ & & & d^n \\ 0 & & & & 0 \end{vmatrix} \quad \text{with } d > 1.$$

Then we can choose d sufficiently large that

$$D - B^*DB \geq 0, \quad (9)$$

i.e., $G \equiv I - (D^{-1/2}B^*D^{1/2})(D^{1/2}BD^{-1/2}) \geq 0$. For the (i,j) th element of $D^{1/2}BD^{-1/2}$ is $d^{(i-j)/2}b_{ij}$, so that the diagonal elements of G are

$$g_{ii} = 1 - \sum_{\lambda} |b_{\lambda i}|^2 d^{-1} = 1 - |\lambda_i|^2 + \epsilon_i,$$

where

$$|\epsilon_i| \leq (n-1)d^{-1}C_B^2(1 - |\lambda_i|)^2,$$

and the off-diagonal elements satisfy

$$\begin{aligned} \delta_i &= \sum_{j \neq i} |g_{ij}| = \sum_{j \neq i} \sum_{\lambda} |\bar{b}_{\lambda i} b_{\lambda j}| d^{(2-i-j)/2} \\ &\leq \frac{1}{2} n^2 d^{-1/2} C_B^2 (1 - |\lambda_i|). \end{aligned}$$

Here the fact that $2\lambda - i - j \leq -1$ follows from $\lambda \leq i$, $\lambda \leq j$, and $i \neq j$.

Thus by choosing $d > n^4 C_B^4$ we make $|\epsilon_i| + \delta_i$ less than $1 - |\lambda_i|^2$ and

Gersgorin's theorem can be applied to the hermitian matrix G to yield

the desired result. Substituting for B in (9) gives

$$S^{-1*} A^* A^* D S A S^{-1} - D \leq 0,$$

i.e.

$$A^* H A - H \leq 0,$$

with $H = S^* D S$ clearly satisfying the necessary requirements.

Since (A) implies (R) has been proved in theorem 1.8, it remains to show that (H) implies (A). Suppose (H) holds and consider the iteration

$$\begin{aligned}
 w_v &= Aw_{v-1} = A^v w_0 \\
 \text{Then } w_v^* H w_v &= w_{v-1}^* A^* H A w_{v-1} \\
 &\leq w_{v-1}^* H w_{v-1} \\
 &\dots\dots\dots \\
 &\leq w_0^* H w_0
 \end{aligned}$$

Hence by (4) : $|w_v|^2 \leq C_H^2 |w_0|^2$
 i.e. $\|A^*\| \leq C_H$.

Note. (i) (H) implies the existence of a new norm defined on X by

$$\|u\|_H^2 = u^* H u.$$

(ii) As the calculation performed in the proof is of considerable complexity, (S) merely implies the existence of a similarity matrix S . Thus the next step is to consider whether stable families can be recognized by carrying out the unitary transformations which triangularize them.

Definition 3.3 : A sequence $\{\epsilon_1, \dots, \epsilon_p\}$ of complex numbers is said to

be nested, with nesting constant K if $|\epsilon_r - \epsilon_s| \leq K |\epsilon_m - \epsilon_l|$

where $l \leq r \leq s \leq m$.

Clearly any sequence can be nested with nesting constant $\leq 2^{-p}$.

Theorem 3.4 : (Buchanan). Let F be a family of matrices A in upper triangular form with eigenvalues nested along the diagonal. Then F is a stable family iff $|\lambda_1| \leq 1$, λ_1 the eigenvalues of A and the off-diagonal elements satisfy

$$|a_{ij}| \leq \text{const.} \max\{1-|\lambda_i|, 1-|\lambda_j|, |\lambda_i-\lambda_j|\}. \quad (*)$$

Proof. All we have to show is that, under the hypothesis of nesting, the similarity transformation $S_{ij}AS_{ij}^{-1}$ and its inverse $S_{ij}^{-1}AS_{ij}$, where the S_{ij} have the form given above, leave the inequality (*) inviolate. For we proved above that the resolvent condition (R) implies the stronger condition (3) after such transformations have been made - hence we need to retrace our steps by inverting them. And to prove the sufficiency of (*), we may carry out the similarity transformations to get (S) as in the proof of Kreiss' theorem.

Thus we need to check that the elements of $S_{ij}AS_{ij}^{-1}$ satisfy (*) if those of A do. Now $S_{ij}AS_{ij}^{-1}$ replaces the elements a_{iv} on the i th row by $\hat{a}_{iv} = a_{iv} + t_{ij}a_{jv}$ and elements $a_{\mu j}$ on the j th column by $\hat{a}_{\mu j} = a_{\mu j} - t_{ij}a_{\mu i}$. Thus if (*) holds for a_{ij} , we have for some constant C ,

$$|a_{\mu i}| \leq C \max(1-|\lambda_i|, |\lambda_i-\lambda_\mu|)$$

and, since $\mu \leq i \leq j \leq v$,

$$|\lambda_i-\lambda_\mu| \leq K |\lambda_j-\lambda_\mu|,$$

$$1 - |\lambda_i| \leq 1 - |\lambda_j| + |\lambda_i-\lambda_j| \leq 1 - |\lambda_j| + K|\lambda_j-\lambda_\mu|.$$

Hence

$$|\hat{a}_{\mu j}| \leq C(1 + |t_{ij}|(1 + K))\max(1 - |\lambda_j|, |\lambda_j-\lambda_\mu|),$$

and so (*) still holds with a new constant, as $|t_{ij}|$ is itself bounded by C . The same argument clearly holds for \hat{a}_{iv} and for the inverse transformation.

That triangularization can be achieved with an arbitrary ordering of eigenvalues on the diagonal can easily be proved with the aid of Schur's theorem.

It might be conjectured that inequalities of the form (*) are perhaps sufficient for the stability of the matrix. We produce now two examples to show that this is not, in fact, the case.

Counter-examples.

(i) $A = \begin{vmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{vmatrix}$ satisfies (*) but the n -power is of the same form with the element in the top right hand corner $(n-1)$, hence is unbounded.

(ii) $\begin{vmatrix} \alpha & 1 & -(2\alpha)^{-1} \\ 0 & -\alpha & 1 \\ 0 & 0 & \alpha \end{vmatrix}$ with $\frac{1}{2} < \alpha < 1$

satisfies the resolvent condition and hence is stable, but the upper right hand element does not satisfy (*).

Section 4. Strictly stable linear operators.

We strengthen our definition of stability :

Definition 4.1: An operator $T \in B(X)$ where X is not necessarily finite dimensional, is strictly stable if the family $\{T^n, n \in \mathbb{I}\}$ is uniformly bounded.

Definition 4.2: A map ϕ on a set E is affine if

$$\phi(\lambda x + \mu y) = \lambda \phi(x) + \mu \phi(y)$$

$$\lambda + \mu = 1. \quad \lambda \geq 0, \mu \geq 0, x, y \in E.$$

The following two theorems are due to A. Markov [36].

Theorem 4.3: Let B be a compact convex subspace of a locally convex linear topological space. Γ a set of commuting continuous affine maps on B . Then there exists $x \in B$ with $\phi(x) = x$, for all $\phi \in \Gamma$.

Proof. Suppose T_ϕ is the set of fixed points of ϕ in B . Then T_ϕ is convex and closed.

A finite intersection of the T_ϕ is non-void:-
 $n = 1$ holds by Schauder's fixed point theorem.

Let $k = n$.

Let $\phi_1, \dots, \phi_n \in \Gamma$.

Then $T_{\phi_1} \dots T_{\phi_{n-1}}$ is closed, convex and compact.

By hypothesis the set is non-void.

Since Γ is commutative,

$$\phi_n (T_{\phi_1} \cap \dots \cap T_{\phi_{n-1}}) \subset T_{\phi_1} \cap \dots \cap T_{\phi_{n-1}}$$

Applying Schauder's theorem again,

there exists $x \in T_{\phi_1} \cap \dots \cap T_{\phi_{n-1}}$ left invariant by ϕ_n .

Finally, by compactness of B , we have that $\bigcap_k \phi_k \neq \emptyset$

Theorem 4.4: Γ is a commutative set of affine maps on a set E .

Let S be all real-valued bounded functions on E .

Then there exists a real functional M on S such that

$$(i) M(1) = 1 \quad (1(x) = 1, \text{ for all } x \in E).$$

$$(ii) M(f+g) = M(f) + M(g).$$

$$(iii) M(f) \geq 0 \quad \text{if } f(x) \geq 0, \text{ for all } x \in E.$$

$$(iv) M(f\phi) = M(f) \quad \text{for all } \phi \in \Gamma.$$

Proof: Let L denote all real-valued functionals on S .

Introduce the weak topology on L

By Tychonoff's theorem, the functionals M which satisfy (i), (ii) and (iii) form a non-void, convex compact set.

To each $\phi \in \Gamma$, there corresponds a map $\phi^*: B \rightarrow B$ defined by

$$(\phi^*M)(f) = M(f\phi).$$

The ϕ^* are affine maps and form a commutative set. Applying Theorem 4.3, there exists functional $M \in B$ s.t.

$$\phi^*(M) = M \quad \text{for all } \phi \in \Gamma.$$

Clearly M satisfies conditions (i), (ii), (iii) and (iv).

Theorem 4.5: (Larionov). A strictly stable operator T on X ,

X Banach, is similar to a unitary operator on X .

Proof. Consider the group $G = \{T^n : n \in \mathbb{I}\}$.

G is commutative and bounded by K and hence, by theorem 4.4, there exists on the space of all bounded complex-valued functionals ϕ on G a functional M such that

$$(a) M(\phi(AB)) = M(\phi(A)) \quad \text{for all } A, B \in G.$$

$$(b) M(\phi(A)) \geq 0 \quad \text{for } \phi(A) \geq 0.$$

$$(c) M(1) = 1.$$

Let X^* be the adjoint space of X .

Let $K = \{\alpha x; \alpha \in \mathbb{C}, x \in X\} \subset X$.

Let a functional ϕ be defined by

$$\phi(\alpha x) = |\alpha| \cdot \|x\|$$

then $\phi(x) = \|x\|$.

By the Hahn-Banach theorem we can extend ϕ to a linear functional x^* on X

$$\text{s.t. } x^*x = \|x\|.$$

Hence the map $x \rightarrow x^*$ defines a mapping of X into X^* and we define on X

$$[x, y] = y^*(x)$$

The form $[\]$ satisfies the axioms for a semi-scalar product.

Now $\phi_{xy}(A) = [Ax, Ay]$, $A \in G$ is a functional on G .

Clearly ϕ_{xy} is bounded.

In particular if $A = 1$

$$\phi_{xy}(1) = [x, y].$$

Define a norm $\|x\|_1$ on X by

$$\|x\|_1 = M([Ax, Ax]) = M(\phi_{xx}(A))$$

For all $B \in G$, $M([ABx, ABx])$

$$= M(\phi_{xx}(AB))$$

$$= M(\phi_{xx}(A)) \quad \text{by (a)}$$

$$= M([Ax, Ax])$$

Hence for all $U \in G$, $\|U^n\|_1 = \|x\|_1$

for $n = 1$ $\|Ux\|_1 = \|x\|_1$.

The two norms are clearly topologically equivalent.

N. Dunford [4] introduced the concept of a spectral operator on a Banach space. This concept will now be used to give an example of a class of strictly stable operators.

Denote the family of Borel sets Δ in \mathbb{C} by B .

Definition 4.6: An operator function $P(\Delta)$; $\Delta \in B$, is a spectral measure if for all $\Delta_1, \Delta_2 \in B$

$$(I) P(\Delta_1 \cap \Delta_2) = P(\Delta_1) P(\Delta_2)$$

$$(II) P(\Delta_1 \cup \Delta_2) = P(\Delta_1) \cup P(\Delta_2)$$

$$\text{with } P_1 \cup P_2 = P_1 + P_2 - P_1 P_2$$

$$(III) P(\mathbb{C} - \Delta) = 1 - P(\Delta)$$

$$(IV) |P(\Delta)| \leq K \text{ some constant } K.$$

$P(\Delta)$ is then a projection operator for all $\Delta \in B$ and $P(\Delta_1), P(\Delta_2)$ commute for all $\Delta_1, \Delta_2 \in B$.

Definition 4.7: $T \in B(X)$ is a spectral operator if there exists a spectral measure $P(\Delta)$ satisfying

$$(I) T P(\Delta) = P(\Delta) T \text{ for all } \Delta \in B.$$

(II) The spectrum $\sigma(H; P(\Delta)X)$ of the operator H in the subspace $P(\Delta)X$ is contained in the closure of Δ .

(III) There exists a dense set $\Gamma \subset X'$ such that

$$f(P(\Delta_1 \cup \Delta_2 \dots)) = f(P(\Delta_1)) + f(P(\Delta_2)) \dots$$

for all $f \in \Gamma$ and denumerably many $\Delta_k \in B$ which are pairwise disjoint.

$P(\Delta)$ defines the spectral operator T uniquely. See [4].

Definition 4.8: The spectral operator T is said to be an operator of scalar type if it is representable in the form

$$T = \int_{\alpha(T)} \lambda dP(\lambda), \quad (1)$$

$P(\lambda)$ is the spectral measure of T .

Theorem 4.9: (Dunford). An operator T is a spectral operator if and only if it is representable in the form

$$T = A + N, \quad (2)$$

where A is an operator of scalar type and N is a generalized nilpotent element which commutes with A . (N is a generalized nilpotent element if

$$\lim_{n \rightarrow \infty} \sqrt[n]{|N^n|} = 0.)$$

Remark. Decomposition (2) may be considered as the continuous analogue of the Jordan form of an operator in finite-dimensional space; T may be considered as the diagonal part and N as the above-diagonal part of the Jordan form. Evidently, in the finite-dimensional case simply some power of the operator N will equal zero. Thus, every linear operator in finite-dimensional space is spectral; but in the infinite-dimensional case, there also exist non-spectral operators. In this case, representation (2) is unique, where A and T have the same spectrum and the same spectral measure.

Proof. The idea of the proof consists in the following. Suppose T is spectral and $P(\Delta)$ its spectral function. We denote by R the minimal ring, closed with respect to the operator norm, of bounded linear operators in X which contains T and all the $P(\Delta)$, and

possessing the following property:

(3) if $B \in R$ and B^{-1} exists and is bounded, then $B^{-1} \in R$.

We set $A = \int_S \lambda dP(\lambda)$, $N = T - A$. It easily follows from the definition of the spectral measure that $T(M) = A(M)$ and, consequently, $N(M) = T(M) - A(M) = 0$ on all maximal ideals M of the ring R . From this we conclude that N is a generalized nilpotent element. In this connection, it follows from the very formula $T = \int \lambda dP(\lambda)$ that A is an operator of scalar type and that $P(\Delta)$ is its spectral measure. This also proves decomposition (2); its uniqueness follows easily from the uniqueness of the spectral measure of a given spectral operator.

The converse assertion is obtained by considering the minimal complete, with respect to the operator norm, commutative ring R of bounded linear operators in X , containing A , N and all the $P(\Delta)$ (where $P(\Delta)$ is the spectral measure of the operator A) and also possessing property (3). Namely, suppose $T = A + N$, where A and N satisfy the conditions of the theorem, and suppose M_Δ is the space of maximal ideals M of the restriction of this ring to the subspace $P(\Delta)X$; then

$$\begin{aligned} \sigma(T, P(\Delta)X) &= (\lambda: \lambda = A(M) + N(M), M \in M_\Delta) \\ &= (\lambda: \lambda = A(M), M \in M_\Delta) = \sigma(A, P(\Delta)X) \subset \bar{\Delta}. \end{aligned}$$

Hence T is spectral and has the same spectrum and spectral measure as A .

Theorem 4.10: (Foguel). A spectral operator T is a scalar operator whose spectrum lies on the unit circle iff T^{-1} is a bounded, everywhere defined operator and there exists M such that

$$\|T^n\| \leq M \quad \text{for all } n \in \mathbb{I}.$$

Proof: If $T = \int_{|\lambda|=1} \lambda I(d\lambda)$

$$\text{then } \|T^n\| = \left\| \int_{|\lambda|=1} \lambda^n I(d\lambda) \right\|$$

$$\leq 4 \sup \{ |I(\alpha)| : \alpha \text{ a Borel set} \}.$$

Conversely assume $\|T^n\| \leq M$ all $n \in \mathbb{I}$, then $R(\lambda, T)$

satisfies the resolvent condition and $\sigma(T)$ lies on the unit ball.

Then if $T = S + N$ where S is scalar and N is generalized nilpotent,

$$\text{hence } N^2 = 0.$$

$$\text{so } T^n = S^n + nNS^{n-1}$$

$$\text{and } nN = (T^n - S^n)S^{-(n-1)},$$

hence nN is bounded which implies that $N = 0$.

Thus the set F_S of all operators of scalar type with spectrum lying on the unit circle is contained in the set F of all strictly stable operators in $B(X)$.

Lemma 4.11: (Sz. Nagy). Let T be a strictly stable map on a Hilbert space X . Then there exists a self-adjoint operator Q such that

$$\frac{1}{K} I \leq Q \leq KI$$

and QTQ^{-1} is unitary.

Proof. The generalized limit of Mazur and Banach is a complex-valued functional $L(\xi(s))$ defined for all complex-valued bounded functions $\xi(s)$ of the positive real variable s . L satisfies the following properties

$$(i) L(a\xi(s) + b\eta(s)) = aL(\xi(s)) + bL(\eta(s))$$

$$(ii) L(\xi(s)) \geq 0 \quad \text{if } \xi(s) \geq 0$$

(iii) $L(\xi(s+a)) = L(\xi(s))$ for all $a > 0$.

(iv) $L(1) = 1$.

Let f and g be elements of R . The sequence $\xi(n) = (T^n f, T^n g)$ ($n = 0, 1, 2, \dots$) being bounded, $|\xi(n)| \leq k^2 \|f\| \|g\|$,
 $\langle f, g \rangle = L(T^n f, T^n g)$.

By property (i) of the generalized limit, we have

$$\begin{aligned} \langle a_1 f_1 + a_2 f_2, b_1 g_1 + b_2 g_2 \rangle &= \\ &= a_1 \bar{b}_1 \langle f_1, g_1 \rangle + a_1 \bar{b}_2 \langle f_1, g_2 \rangle + a_2 \bar{b}_2 \langle f_2, g_2 \rangle, \end{aligned}$$

i.e. $\langle f, g \rangle$ is a hermitian bilinear form of the variable elements f and g . Furthermore, the inequalities

$$\frac{1}{k} \leq \frac{\|T^n f\|}{\|T^{-n} T^n f\|} = \frac{\|T^n f\|}{\|f\|} \leq k$$

imply, by the properties (i), (ii) and (iv), that

$$\frac{1}{k^2} \|f\|^2 \leq \langle f, f \rangle \leq k^2 \|f\|^2.$$

Thus there exists a selfadjoint transformation A such that

$\langle f, g \rangle = (Af, g)$. We have,

$$\frac{1}{k^2} I \leq A \leq k^2 I,$$

and by property (iii),

$$(ATf, Tg) = L(T^{n+1}f, T^{n+1}g) = L(T^n f, T^n g) = (Af, g),$$

i.e. $T^*AT = A$.

Let Q be the positive selfadjoint square-root of A ; then

$$\frac{1}{k} I \leq Q \leq kI,$$

$$\frac{1}{k} I \leq Q^{-1} \leq kI.$$

It follows that

$$Q^{-1}(T^*QQT)Q^{-1} = Q^{-1}(QQ)Q^{-1} = I,$$

$$(QTQ^{-1})^*(QTQ^{-1}) = I.$$

Thus, $U = QTQ^{-1}$ is isometric. As it admits an inverse, namely $U^{-1} = QT^{-1}Q^{-1}$, it is also unitary.

Proposition 4.12: If X is Hilbert, $F_S = F$.

Proof. Consider $T \in F$.

By Lemma 4.11, there exists a self-adjoint linear map Q on X s.t.

$$k^{-1} I \leq Q \leq kI \quad \text{and}$$

$$QTQ^{-1} \text{ is unitary.}$$

Hence T has spectrum lying on the unit circle.

Proposition 4.13: Suppose X is a finite dimensional Banach space, then if $T \in F$, T is a scalar operator with spectrum on the unit circle.

Proof. T satisfies the conditions of Foguel since on a finite dimensional space all linear operators are spectral.

Since we know, Kakutani [6], that the sum of two commuting scalar type spectral operators is not necessarily a scalar type operator, (in the case where X is Hilbert it is), we may not prove a general perturbation theorem. See Appendix B. However we may prove the following:

Theorem 4.14: Let S and T be commuting strictly stable spectral operators in $B(X)$ such that

$$(i) \quad S = S_1 + N_1$$

$$T = T_1 + N_2$$

where $S_1, T_1, S_1 + T_1$ are of scalar type.

(ii) $S + T$ is of scalar type.

(iii) $\sigma(S+T)$ is contained on the unit ball

then $S + T$ is a strictly stable operator.

Proof. S and T commute and $S_1 + T_1$ is of scalar type iff $S + T$ is spectral, Wermer [27].

$S + T$ is of scalar type with $\sigma(S+T)$ contained in the unit ball iff $S+T$ is strictly stable by Foguel.

Remark. We show by induction that the powers of a unitary matrix U are unitary.

Assume $\|U^{n-1}x\| = \|x\|$.

$$\begin{aligned} \text{Then } \|U^n x\| &= \|U(U^{n-1}x)\| \\ &= \|U^{n-1}x\| = \|x\|. \end{aligned}$$

By Buchanan's theorem, U is a stable operator.

In fact, since $\|U^{-1}\| = \|U^T\| = \|U\|$, it is strictly stable.

Applying Larionov's theorem, we have a characterization of strict stability on a finite dimensional space.

This concludes the general theory for stable and strictly stable operators on a (complex) Banach space X which is not necessarily finite dimensional. Since the uniform norm dominates the L^p norms, i.e. $\|f\|_p \leq \|f\|_\infty$ for all $f \in B(X)$, (strict) stability under the uniform norm implies (strict) stability under the L^p norms. The converse of course need not hold.

If $B(X)$ is finite dimensional, which is the case when X is finite dimensional, all norms on $B(X)$ are equivalent and stability under one norm implies stability under any other. Hence any of the necessary and sufficient conditions in Section 3 or 4 may be used to prove the stability of a general linear operator on X .

In Chapter 3, Section 2, the stability of a certain kind of linear operator on the space L^p is discussed in terms of its characteristic function.

Notes and Remarks. P.D. Lax and R.D. Richtmyer, see [7], originally defined the stability of linear operators which defined finite difference equations somewhat differently and showed that if X was finite dimensional and the L^2 norm was defined on it, their concept coincided with that of definition 1.1. The proof of this assertion is given in Chapter 3, Lemma 1.2.

The proof of the Kreiss and Buchanan theorems are those given in Morton and Schechter [11].

E. Larionov [3] recently defined the concept of strict stability and used his results to consider the stability of the equation

$$\frac{dx}{dt} = A(t)x \quad -\infty < t < \infty,$$

$A(t)$ a T -periodic operator function on X , in the sense that all solutions $x(t)$ are bounded on $(-\infty, \infty)$.

Various other definitions of stability appear in the literature. An operator is weakly stable if a finite number of its powers are uniformly bounded as the time interval decreases to zero. Strong stability is defined in terms of another norm on X and the boundedness of the solution under this norm. See Stetter [26]. At a fundamental level, these concepts stem from a different concept of well-posedness of a differential equation. See Appendix A. B. Wendroff [34] studied this question and showed that no one concept of well-posedness suffices for all problems and that differential operators may be classified according to their degree of well-posedness.

C H A P T E R 2

The relation between stability and convergence.

Section 1. The equivalence theorem of P.D. Lax.

One wishes to find a one-parameter family $\{u(t)\}$ of elements of a Banach space X such that

$$\begin{aligned} \frac{d}{dt} u(t) &= Au + G(t)u \\ u(0) &= u_0 \end{aligned} \quad (1)$$

where A is a linear differential operator, $D(A) \subset X$

$G(t)$, $t \in (0, T)$ not necessarily linear

with $D(G(t)) \subset X$ s.t.

$$\|G(t)u - G(t)v\| \leq L\|u-v\| \text{ for all } u, v \in X, \text{ for all } t \in (0, 1).$$

We shall assume the problem (1) to be well-posed, in the sense that there exists a family $\{E(t)\}$ of continuous operators in X which has the following properties

(i) $u(t) = E(t)u_0$ satisfies (1) for all those initial functions for which a solution exists.

(ii) The mapping $(0, T) \xrightarrow{E(t)u_0} X$ is continuous in u_0 .

(Thompson [10] guarantees a solution to (1) if $\frac{\partial u}{\partial t} - Au = 0$)

See appendix A.

We shall approximate the solution $u(t)$ of (1) by finite difference equations.

An implicit or explicit multi-level linear finite difference equation for the solution of (1) is of form

$$\sum_{v=0}^k A_v(h) u^{n+v} = h \sum_{v=0}^k B_v G(t_{n+v}) u^{n+v} \quad (2)$$

where $vh = t_v$ and $u(t_v)$ is denoted by u_v . $A_v(h)$ and $B_v(h)$ are continuous, linear and independent of $G(t)$.

A k -component column vector, whose components are elements of X may be regarded as an element of another Banach space \tilde{X} under various obvious norms.

$$\text{Let } \tilde{\phi}_n = \begin{pmatrix} u^{n-k-1} \\ \vdots \\ u^n \end{pmatrix} \text{ and } R(t, h) = (-A_k(h) + hB_k G(t))^{-1}$$

$$\text{since } u^{n+k} = R(t_{n+k}, h) \left(\sum_0^{k-1} A_v(h) k^{n+v} - h \sum_0^{k-1} B_v G(t_{n+v}) u^{n+v} \right)$$

the finite difference equations become

$$\tilde{\phi}_{n+1} = \begin{vmatrix} R(t_{n+k}, h) & 0 & A_{k-1}(h) & \dots & A_0(h) \\ & 1 & & & \\ & & \ddots & & \\ & 0 & & 1 & 0 \\ & & & & \ddots & \\ & & & & & 1 \end{vmatrix}$$

$$-h \begin{vmatrix} B_{k-1} & \dots & B_0 \\ & & 0 \\ & & & G(t_{n+k-1}) & 0 \\ & & & & \ddots \\ & & & 0 & & G(t_n) \end{vmatrix} \tilde{\phi}_n$$

$$= \tilde{C}(t_n, h) \tilde{\phi}_n$$

$$= \prod_{v=0}^n \tilde{C}(vh, h) \phi_0$$

$$= \prod_{v=0}^n \tilde{C}(vh, h) \phi_0$$

$$\text{Let } \tilde{E}(t, h) = \begin{vmatrix} E(t+(k-1)h) & 0 \\ & \ddots \\ & & E(t) \end{vmatrix}$$

Let $X_0 = \{ \tilde{u} \mid \tilde{u} = \begin{pmatrix} u \\ \vdots \\ \dot{u} \end{pmatrix}, u \in X \} \subset \tilde{X}$. X_0 is then a Banach space.

Definition 1.1: The difference equation (2) is said to be consistent with the problem (1) if for each $G(t)$ under the above conditions, the local error satisfies

$$\begin{aligned} & \left\| R(t+kh) \left(\sum_{v=0}^{k-1} A_v(h) u(t+vh) - h \sum_{v=0}^{k-1} B_v G(t+vh) u(t+vh) \right) - u(t+kh) \right\| \\ &= \left\| (\tilde{C}(t,h) \tilde{E}(t,h) - \tilde{E}(t+h,h)) \tilde{u}_0 \right\| < \epsilon \cdot h. \end{aligned}$$

Definition 1.2: The equation (2) is convergent if for $G(t)$ as above and all $u_0 \in X$

$$\begin{aligned} & \lim_{\substack{h \rightarrow 0 \\ \tilde{u}_0 \rightarrow u_0}} \left\| \prod_{v=0}^{n_j-1} \tilde{C}(vh_j, h_j) \tilde{\phi}_0 - \tilde{E}(t, 0) u_0 \right\| \\ &= 0 \quad \text{as } n_j \rightarrow \infty, \quad n_j \in \mathbb{I} \\ & \quad (n_j+k-1)h_j \in (0, T); \quad n_j h_j \rightarrow T. \end{aligned}$$

$$\text{Put } \tilde{A}(h) = \begin{vmatrix} -A_k^{-1}(h)A_{k-1}(h) & \dots & \dots & -A_k^{-1}(h)A_0(h) \\ I & & \theta & \\ \cdot & & \cdot & \\ \dots & \dots & \dots & \dots \\ \theta & & \cdot & \cdot & I & \theta \end{vmatrix}$$

Definition 1.3: The procedure (2) is stable if the positive powers of \tilde{A}, \tilde{A}^n for $nh \in (0, T)$ are uniformly bounded, i.e.

$$\|\tilde{A}^n(h)\| \leq K \quad nh \leq T.$$

Under the preceding concepts we have the Equivalence theorem of P.D. Lax.

Theorem 1.4: (Lax). For a consistent approximation to a properly posed initial value problem, stability is necessary and sufficient for convergence.

Proof. See Ansorge [8].

For a simplified proof in the case of a single step algorithm see Richtmyer and Morton [7].

Note. We have assumed that the difference procedure is reasonable, i.e. that unique function values say $u^{n+1} = u(x, \overline{n+1}h)$ can be calculated from any previous set u^j , $j \leq n$ on which they depend continuously.

Section 2. Special cases of the Equivalence Theorem.

If $G = \theta$, null element in X then the Equivalence theorem and the various concepts reduce to those given in the Lax-Richtmyer theory.

When the problem (1) reduces to that of an ordinary first order differential equation, i.e. $A = \theta$ and X the space of reals, the Equivalence Theorem then takes the form

Theorem 2.1: Under the norm $u = u + u + \dots$ the consistent difference equation converges iff the matrices

$$\begin{vmatrix} -\frac{A_{k-1}}{A_k} & -\frac{A_0}{A_k} \\ I & 0 \\ \cdot & \cdot \\ 0 & I \end{vmatrix} \quad \text{are stable.}$$

This however is the case when the elements of \tilde{A}^n remain bounded as n increases. Thus the eigenvalues of \tilde{A} may not have moduli greater than 1 and those eigenvalues of modulus 1 are simple. Thus \tilde{A} has characteristic polynomial

$$D(\lambda) = \frac{(-1)^k}{A_k} \rho(\lambda) \quad \text{whose} \quad \rho(\lambda) = A_0 + A_1\lambda + \dots + A_k\lambda^k.$$

From the form of \tilde{A} , $\rho(\lambda) = A_k(\lambda - \lambda_1)^{e_1} \dots (\lambda - \lambda_s)^{e_s}$ (λ_1 a root of $\rho(\lambda)$ of multiplicity e_1) establishes the only invariant factor of A .

A necessary and sufficient condition for the convergence of a consistent multistep algorithm for the solution of an ordinary differential equation of first order is therefore that all roots of $\rho(\lambda)$ lie inside the unit circle and that roots on the circle are simple.

Dahlquist [22] extended this result to systems of ordinary differential equations of first order.

Notes and Remarks. Lax and Richtmyer [7] showed that the stability of multi-level difference approximations to linear Cauchy problems of the form

$$\frac{\partial u}{\partial t} = Au + g(t)$$

where $g(t)$ is piecewise uniformly continuous in t , is necessary and sufficient for convergence, provided a consistency condition is satisfied.

R.J. Thompson [10] extended this result to the quasi-linear case, i.e. when $g(t)$ is replaced by $g(t, u(t))$, defined for $0 \leq t \leq T$, continuous in t for each u and uniformly Lipschitz continuous with respect to u .

Dahlquist [22] proved that his stability concept for a system of nonlinear first order ordinary differential equations together with a consistency condition gave a necessary and sufficient condition for convergence.

Ansorge [8], whose treatment we follow, extended the Lax-Richtmyer theory to quasi-linear initial value problems including both convergence theorems.

CHAPTER 3

The Stability of finite difference equations.

Section 1. Stability in L^2 ; the Lax-Richtmyer theory.

Here we consider problems with constant coefficients and periodic boundary conditions so that the Fourier series or Fourier integral representations of the solution may be used. We could as well assume that the functions involved in the boundary conditions are quadratically integrable over all space. The representation theorems to be used are those based on the L^2 norm viz. the Riesz-Fisher theorem for Fourier series and Plancherel's theorem for integrals.

Once the function space X has been chosen, the set of Fourier coefficients or the Fourier transform determines a point in a second Banach space \tilde{X} and the Fourier transform provides a 1-1 isometric isomorphism between X and \tilde{X} . Thus all considerations proved before hold in \tilde{X} as well as X . We also note that X and \tilde{X} are Hilbert spaces.

Let us suppose that there are p dependent variables in d space variables. Then the functions corresponding to a point in X may be denoted by $u(\underline{x})$ where u is a vector of p components and \underline{x} is a vector of d components.

The general linear differential operator in X may be formally obtained by considering a $p \times p$ matrix $P(q_1, \dots, q_d)$ whose elements are polynomials in q_1, \dots, q_d and setting $\frac{\partial}{\partial x_i} = q_i$ $i = 1, \dots, d$. Let A be such an operator.

Note. We understand by the function e^M of a matrix M , the power series $1 + M + \frac{1}{2}M^2 + \dots$ which is absolutely convergent, element by element, to e^M and has the property $\frac{\partial}{\partial t} e^{tM} = M e^{tM}$. See for e.g. Dunford and Schwartz [1].

We assume that the problem is properly posed in the sense that the solution depends continuously on the initial data and that, although a solution may not exist for a particular $u_0 \in X$, u_0 may be approximated by initial functions for which a solution does exist. See Appendix A.

In an analogous way, we construct finite difference equations in \tilde{X} and obtain

$$\hat{u}^{n+1}(k) = G(\Delta t, k) \hat{u}^n(k).$$

Definition 1.1 : $G(\Delta t, k)$ is called the amplification matrix.

Lemma 1.2 : $F = \{e^{-\alpha T \log K} G(\Delta t, k)\}$ is a stable family of matrices iff there exists $T > 0$ s.t. the matrices $G(\Delta t, k)^n$ are uniformly bounded for

$$\begin{aligned} 0 < \Delta t < T \\ 0 \leq n\Delta t \leq T \\ \text{all } k \in L. \end{aligned}$$

Proof. Clearly if the powers of the amplification matrices G have a uniform bound K , then $F = \{e^{-\alpha \Delta t} G(\Delta t, k)\}$ is stable, where $\alpha = T^{-1} \log K$.

For $v \in \mathbb{I}$, $v = m(T/\Delta t) + n$ where $0 \leq n\Delta t < T$ hence

$$\begin{aligned} \|(e^{-\alpha \Delta t} G)^v\| &\leq \|(e^{-\alpha \Delta t} G)^{T/\Delta t}\|^m \|(e^{-\alpha \Delta t} G)^n\| \\ &\leq K. \end{aligned}$$

Conversely if, for some constant α , F satisfies the stability condition, then for $0 \leq n\Delta t < T$

$$\|G^n\| \leq C e^{\alpha T} = \text{const.}$$

A necessary condition for stability of F is that

$$r(A) \leq 1 \quad \text{for all } A \in F.$$

Then it is necessary for the eigenvalues λ_i of G to satisfy

$$|\lambda_i| \leq e^{\alpha \Delta t} \quad \text{some } \alpha.$$

This condition is known as the von Neumann condition.

Note. The von Neumann condition is also a sufficient condition for the stability of $G(\Delta t, k)$ if either $G(\Delta t, k)$ is a normal matrix or $p = 1$.

The Kreiss and Buchanan theorems may of course be applied to prove the stability or otherwise of the difference scheme defined by G .

For completeness, we shall give two other sufficient conditions for stability.

Theorem 1.3 (Kato) : Suppose $G(\Delta t, k)$ is uniformly bounded and the von Neumann condition is satisfied. And, suppose that for each G a closed, rectifiable curve Γ is drawn inside the circle $|\lambda| \leq r(G)$, such that its length is uniformly bounded and its distance away from $\sigma(G)$ is uniformly bounded away from zero. Then the difference scheme is stable if there exists $\delta > 0$, independent of $\Delta t, k$ s.t. each distinct eigenvalue $\lambda_i, i = 1, \dots, q$ outside Γ has index 1 and (i) the distance of λ_i from all other eigenvalues is greater than δ or else (ii) for the set of λ_i not satisfying (i), the complete set of corresponding eigenvalues has Gram determinant Δ^2 such that

$$\Delta > \delta$$

Proof. See for e.g. [7].

Here we recall that for an eigenvalue λ of A , the set of vectors x for which $(A - \lambda I)^n x = 0, n \in \mathbb{I}$ forms the algebraic eigenspace for λ . The smallest such integer n is called the index of λ .

Definition 1.4 : If G is a square matrix then the field of values of $G, F(G) = \left\{ \sum_{i,k} g_{ik} x_i x_k \right\} = \{(Ax, x)\} = x^* Ax$ where $x = (x_1, \dots, x_n)$ and $\sum x_i \bar{x}_i = 1$.

Proposition : $F(G)$ is closed, bounded and convex.

Proof. Immediate from the definition.

Theorem 1.6 : (Lax-Wendroff). If G is such that

$$|x^* G x| \leq (1 + O(\Delta t)) \|x\|^2 \quad \text{for all } x \in X,$$

then the corresponding scheme is stable.

Proof. Choose α s.t. $A = e^{-\alpha\Delta t} G$ satisfies

$$|x^*Ax| \leq \|x\|^2.$$

Then all eigenvalues of A lie in the closed unit disc. Hence for all

$z \in \phi$, $R(z, A)$ exists.

Hence if $w \in X$ and $x = (zI - A)^{-1}w$

$$\begin{aligned} \|x\| \|w\| &\geq |x^*w| = |x^*(zI - A)x| \\ &\geq |x|^2 (|\bar{z}| - 1)^{-1} \end{aligned}$$

i.e. $\{A\}$ satisfies the resolvent condition.

Example. Consider the finite difference equations for the solution of the wave equation in one variable. The wave equation is of the second order in time so normally $p = 2$.

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad c^2 > 0 \quad (A)$$

and the conditions

$$u(x, 0) = f_1(x)$$

$$\frac{\partial u}{\partial t} u(x, 0) = f_2(x)$$

$$u(x, t) = u(x+L, t).$$

Introduce the variables

$$v = \frac{\partial u}{\partial t} \quad w = c \frac{\partial u}{\partial x}$$

then (A) becomes

$$\frac{\partial v}{\partial t} = c \frac{\partial w}{\partial x}$$

$$\frac{\partial w}{\partial t} = c \frac{\partial v}{\partial x}$$

and the norm is taken to be

$$\|u\| = \left(\int_0^L (|w|^2 + |v|^2) dx \right)^{\frac{1}{2}}$$

The problem is properly posed in this formulation. To avoid difference

quotients over the interval $2\Delta x$, we shall associate values of w with the midpoints of the interval and write $w_{j+\frac{1}{2}}^n$ etc.

The equations are

$$\frac{1}{\tau} (v_j^{n+1} - v_j^n) = c/2 \times (w_{j+\frac{1}{2}}^n - w_{j-\frac{1}{2}}^n + w_{j+\frac{1}{2}}^{n+1} - w_{j-\frac{1}{2}}^{n+1}).$$

$$\frac{1}{\tau} (w_{j-\frac{1}{2}}^{n+1} - w_{j-\frac{1}{2}}^n) = c/2 \times (v_j^{n+1} - v_{j-1}^{n+1} + v_j^n - v_{j-1}^n)$$

and have the amplification matrix

$$G = \begin{vmatrix} \frac{1-a^2/4}{1+a^2/4} & \frac{ia}{1+a^2/4} \\ \frac{ia}{1+a^2/4} & \frac{1-a^2/4}{1+a^2/4} \end{vmatrix} \quad a = (2c\Delta t / \Delta x) \sin(k\Delta x/2)$$

Both eigenvalues of G have modulus 1 and G is clearly unitary.

Hence the von Neumann condition is sufficient for stability.

Section 2. Stability on L^p , $p \neq 2$.

It should be noted that the uniform norm $\| \cdot \|_\infty$ in $B(X)$ dominates the various L^p norms, i.e.

$$\|f\|_p \leq \|f\|_\infty \quad \text{for all } f \in B(X).$$

Hence stability under L^∞ implies L^p stability should the L^p norm be defined on $B(X)$. The converse is of course not necessarily true.

When X is finite dimensional, all norms on $B(X)$ are equivalent, since $B(X)$ is then finite dimensional and stability under any norm implies stability under any other norm.

We now discuss the case when $X = L^p$ and $B(X)$ has, as before, the uniform norm.

Consider the operator A defined on real-valued functions $v(x)$ on

Lemma 2.2: A is consistent with (1) iff

$$a(\theta) = \exp(\rho\lambda(-i\theta)^{\nu} + O(\theta^{\nu})) \text{ when } \theta \rightarrow 0$$

and its order is p if there exists $\beta \neq 0$ s.t.

$$a(\theta) = \exp(\rho\lambda(-i\theta)^{\nu} - \beta\theta^{\nu+p} + O(\theta^{\nu+p})) \text{ when } \theta \rightarrow 0.$$

Theorem 2.3: A is stable in L^2 iff $|a(\theta)| \leq 1$ for all $\theta \in \mathbb{R}$.

Proof. A trivial application of Parseval's relation

$\|A\|_2 = \max|a(\theta)|$, since the characteristic function of A^n is $a(\theta)^n$.

Theorem 2.4: If A has a characteristic function analytic for all real θ , then A is stable in L^p iff it satisfies one of the following conditions:

(a) $a(\theta) = ce^{iJ\theta}$, $|c| = 1$.

(b) $|a(\theta)| < 1$ except at finitely many points, θ_q , in $|\theta| \leq \pi$ where $|a(\theta)| = 1$.

For $q = 1, \dots, N$, there exists constants α_q, β_q, μ_q where $\alpha_q \in \mathbb{R}$
 $\text{Re } \beta_q > 0$
 $\mu_q \in \mathbb{N}$ (even)

s.t. $a(\theta_q + \theta) = a(\theta_q) \exp\{i\alpha_q\theta - \beta_q\theta^{\mu_q} (1 + O(1))\}$ when $\theta \rightarrow 0$.

Proof. The sufficiency was established by Strang [17] for $p = \infty$.

For other p they follow trivially from the Minkowski inequality and

$$\|Av\|_p \leq \left(\sum_j |a_j|\right) \|v\|_p, \quad v \in L^p.$$

The necessity has been proved in Thomée [18].

Remark. Explicit operators (J finite) have characteristic functions which are trigonometric polynomials, hence are analytic.

Implicit operators have characteristic functions which are the quotient of trigonometric polynomials hence they too are analytic.

Theorem 2.5: In order that (2) admit an L^p stable consistent operator of form (1) it is necessary that

$$(i) \text{ for } \nu \text{ even ; } (-1)^{\nu/2} \operatorname{Re} \rho < 0 \quad (\text{parabolic case})$$

or

$$(ii) \nu = 1 \quad : \quad \rho \in \mathbb{R}. \quad (\text{hyperbolic case}).$$

On the other hand if A has characteristic function $a(\theta)$ s.t. $|a(\theta)| < 1$ for $0 < |\theta| < \pi$, then A is stable in case (i) if it is consistent and in case (ii) if $a(\theta)$ satisfies the order of accuracy condition in lemma 2.1 with $\nu = 1$, $\operatorname{Re} \beta > 0$ and p odd.

Proof. A simple application of Lemma 2.2 and Theorem 2.3.

Note. In case (ii) an L^p -stable operator has an odd order of accuracy, thus all the useful second order schemes for solving hyperbolic equations of form (2) are unstable in all but the L^2 norm.

The Lax-Wendroff explicit scheme, see [19] is not stable in any L^p space but L^2 .

Section 3. The Variable Coefficient Problem.

We now consider operators with variable coefficients of form

$$A_{h,t} v(x) = \sum_j a_j(x,t) v(x-jh) \quad \text{with} \quad \sum_j |a_j(x,t)| < \infty \quad (1)$$

which are used in the solution of the problem

$$\frac{\partial u}{\partial t} = \rho(x,t) \frac{\partial^\nu u}{\partial x^\nu}, \quad 0 \leq t \leq T$$

$$u(x,0) = u_0(x) \quad (2)$$

where $\frac{k}{h^\nu} = \lambda$ is kept constant.

Introduce again the periodic function

$$a(\theta, x, t) = \sum_j a_j(x, t) e^{ij\theta}.$$

Then the solution $u(x)$ at time $t = nk$ is approximated by

$$u_n(x) = A_{h,t}^n u_0(x)$$

$$\text{where } \|A_{h,t}\| = \sup_x \sum_j |a_j(x, t)|.$$

Definition 3.1: The operator (1) with coefficients "frozen" at (x_0, t_0) is

$$\tilde{A}_h(x_0, t_0) = \sum_j a_j(x_0, t_0) v(x - jh).$$

This is then an operator with constant coefficients with characteristic function

$$\tilde{a}(\theta) = a(\theta, x_0, t_0).$$

We shall refer to the stability of \tilde{A}_h at all points (x_0, t_0) as local stability. In general local stability will be neither necessary nor sufficient for stability. Kreiss and Strang have constructed examples to show that while a problem may be well-posed locally, it need not be so in a global sense. Again a problem with variable coefficients can be well-posed yet none of the corresponding constant coefficient problems need be. In this direction, Thomée [16] proves the following.

Theorem 3.2: Suppose the $a_j(x, t)$ are continuous and that $\sum_j |a_j(x, t)|$ converges uniformly in $\{0, T\}$. A necessary condition that $A_{h,t}$ is stable in $\{0, T\}$ is that $\tilde{A}_h = \tilde{A}_h(x_0, t_0)$ is uniformly stable for all $(x_0, t_0) \in \{0, T\}$.

Proof. Assume A is explicit (the other case follows). Let $n \in \mathbb{N}$ be given and let h, km vary s.t. $mk = t_0$, $k/h^{\nu-\lambda}$, $h \rightarrow 0$.

Then $A_{h,m+n,m} v(x_0) = \sum_{|j| \leq nN} a_{nj}(x_0, t_0, h) v(x_0 - jh)$.

So a_{nj} are now polynomials in $a_j(x, t)$ $|j| < N$ so only such (x_0, t) are used as to ensure

$$|x - x_0| \leq n N h \quad t \leq t \leq t_0 + n_k$$

So $\lim_{h \rightarrow 0} a_{nj}(x_0, t_0, h) = \tilde{a}_{nj}$

where \tilde{a}_{nj} are the Fourier coefficients of $a(\theta)^n$

$$\therefore \|A_{h,m+n,m} v(x_0)\| \leq C \|v\|$$

applying this to the function $v(xh^{-1} + x_0(1-h^{-1}))$

by * we get as $h \rightarrow 0$

$$\|\tilde{A}_1^n v(x_0)\| \leq C \|v\|.$$

Let $v(x) = v(x - x_0 + y)$

$$\|\tilde{A}_1^n v(y)\| \leq C \|v\|.$$

Since $\|\tilde{A}_h^n\|$ is independent of h

A is stable.

Finally, Thomée [16] has proved that under certain conditions global stability implies stability.

Note. It follows from this theorem that the local characteristic function $a(\theta)$ of $A_{h,t}$ must satisfy the conditions of Theorem 2.3. By the definition of consistency, the problem (2) must be at least locally parabolic or hyperbolic.

Theorem 3.3: A_h consistent. Then $A_{h,t}$ is stable if

(1) There exists $\delta > 0$ s.t. $a(\theta + iw, x, t)$ is analytic in $\theta + iw$ in $|w| \leq \delta$ and $\partial^l a(\theta) / \partial x^l$, $l \leq v$ are continuous and bounded for $(x, t) \in \{0, T\}$.

$$(II) \quad \sup_{\substack{\sigma < |\theta| < \pi \\ (0, T)}} |a(\theta, x, t)| < 1 \quad \text{for } 0 < \infty \leq \Pi$$

$$(III) \quad a(\theta, x, t) = \exp\{i\alpha(x, t)\theta - \beta(x, t)\theta^\mu\} (1 + o(t))$$

when θ tends uniformly to zero,

$\alpha(x, t)$ is real valued,

$$\operatorname{Re} \beta(x, t) \geq \beta > 0,$$

$\mu \in \mathbb{N}$, even.

Note. (II) is satisfied if $a(\theta, x, t) = \frac{a_1(\theta, x, t)}{a_2(\theta, x, t)}$

$$a_i = \sum_{-N_i}^{N_i} a_{i,j}(x, t) e^{ij\theta} \quad i = 1, 2$$

$a_{i,j}(x, t)$ bounded, continuous, have bounded continuous derivatives of order $\leq \nu$ and a_i is uniformly bounded for real θ .

Notes and Remarks. The question of stability in L^p of constant coefficient problems has been fairly well settled as several widely applicable conditions are known. This theory is of most use in checking the local stability of equations where the operator A depends upon the space variables, i.e. the stability of the linearised equations.

Lax and Richtmyer, see [7], considered parabolic and symmetric hyperbolic equations in the space L^2 since these are properly posed under conditions which are purely local in character. To avoid the growth of high frequency Fourier components it is required that the coefficients $a(x)$ be Lipschitz continuous and that the difference equations be dissipative, i.e. the difference scheme is designed so that the eigenvalues λ_ν of the amplification matrix satisfy $|\lambda_\nu| \leq 1 - \delta |k\Delta x|^{2r}$

when $|k\Delta x| \leq \Pi$, $\delta > 0$, $r \in \mathbb{I}^+$.

For parabolic equations this is sufficient for the stability of the difference scheme under suitable smoothness conditions on the coefficients.

In the hyperbolic case where the equation has form

$$\frac{\partial u}{\partial t} = \sum_{j=1}^d A_j(x) \frac{\partial u}{\partial x_j} \quad -\infty < x_j < \infty, \quad t \in (0,1)$$

and the $A_j(x)$ are $p \times p$ Hermitian matrices, Kreiss has proved the following theorem:

Theorem: Suppose all matrices occurring in the differential equation and the approximating difference equation are Hermitian, uniformly bounded and uniformly Lipschitz continuous in x , then if the difference equation is dissipative of order $2r$ and accurate to order $2r-1$, some positive integer r , it is stable.

Proof. See [7].

We note that this theorem is not directly applicable to the Lax-Wendroff approximating schemes which are dissipative of order 4 but accurate only to order 2. However these cases can be covered by a suitable change in variables. See for example Parlett [32].

Lax in [33] investigated one step difference schemes for time-dependent hyperbolic equations of form

$$u(t+\Delta t) = S(\Delta t) u(t)$$

$$\text{where } S(\Delta t) = \sum_j C_j T_j^{\Delta t},$$

$T_j^{\Delta t}$ is translation by an amount $j\Delta t$.

C_j are matrices dependent on x but independent of Δt .

Stability requires $\left\{ \prod_{h=1}^n S_h(\Delta t) \right\} \leq K \quad n\Delta t \leq 1$

which, if $S(\Delta t)$ is time independent, simplifies to

CHAPTER 4.

Dahlquist's stability theory

Section 1. Multi-step Algorithms for Ordinary differential equations of the 1st order.

The general linear k-step finite difference equation for the solution of the problem

$$\frac{dy}{dx} = f(x, y) \quad (1)$$

$$y(a) = \eta$$

where $f(x, y)$ is defined and continuous in the strip $a \leq x \leq b$,

$-\infty < y < \infty$, a, b finite, and Lipschitz continuous in y ,

(i.e. the existence and uniqueness of a solution of (1) is assumed)

is of the form

$$a_k y_{n+k} + \dots + a_0 y_n = h \{ b_n f_{n+k} + \dots + b_0 f_n \}$$

a_j, b_j constants independent of n (2)

($n = 0, 1, \dots$) where $f_m = f(x_m, y_m)$. This equation is known as the general linear k-step method.

With (2) we associate the two generating polynomials

$$\rho(\xi) = a_k \xi^k + \dots + a_0$$

$$\sigma(\xi) = b_k \xi^k + \dots + b_0$$

Conversely any two such polynomials define a method of the form (2).

From Chapter 2, Section 2 we see that the stability condition on (2) then takes the form that the modulus of no root of $\rho(\xi)$ exceeds one and that roots of modulus one be simple.

Definition 1.1 : The operator

$$L(y(x), h) = a_k y(x+kh) + a_{k-1} y(x+(k-1)h) + \dots + a_0 y(x) - h \{ b_k y'(x+kh) + b_{k-1} y'(x+(k-1)h) + \dots + b_0 y'(x) \} \quad (3)$$

associated with (2) is known as the difference operator.

$$\text{Since } y(x+mh) = y(x) + mhy'(x) + \frac{1}{2}m^2h^2y''(x) + \dots$$

$$hy'(x+mh) = hy'(x) + mh^2y''(x) + \dots$$

the difference operator applied to functions which have continuous derivatives of sufficiently high order becomes

$$L(y(x), h) = c_0y(x) + c_1hy'(x) + \dots$$

where

$$c_0 = \sum_{i=1}^k a_i$$

$$c_1 = \sum_{i=1}^k a_i - \sum_{i=1}^k b_i$$

.....

$$c_q = \frac{1}{q!} (a_1 + 2^q a_2 + \dots + k^q a_k) - \frac{1}{(q-1)!} (b_1 + 2^{q-1} b_2 + \dots + k^{q-1} b_k)$$

for all $q = 2, 3, \dots$

Definition 1.2 : $L(y(x), h)$ is of order p if

$$c_0 = c_1 = \dots = c_{p-1} = 0 \text{ but } c_{p+1} \neq 0.$$

It may be well to note, from a computational viewpoint, that (2) is equivalent to

$$a_n y_{n+t+k} + \dots + a_0 y_{n+t} = h \{ b_k f_{n+k+t} + \dots + b_0 f_{n+t} \}$$

and the order of $L(y(x); h)$ may be defined as the order of the first non-vanishing term in the Taylor expansion for all $t \in \mathbb{I}$

The consistency condition reduces to :

A method of form (2) is consistent if the order p of the method is at least one.

In terms of the polynomials $\rho(\xi)$ and $\sigma(\xi)$ it is expressed by the relations

$$\rho(1) = 0 \quad \rho'(1) = \sigma(1).$$

Section 2. Maximum order stable difference operators.

We now consider the problem: given a polynomial $\rho(\xi)$ of degree k such that $\rho(1) = 0$, what is the order of the associated operator L that can be achieved by choosing some suitable polynomial $\sigma(\xi)$.

With the difference operator (3) we associate the complex valued function

$$\Phi(\xi) = (\log \xi)^{-1} \rho(\xi) - \sigma(\xi) \quad \xi \in \mathbb{C}$$

Lemma 2.1 : The difference operator (3) associated with $\rho(\xi)$ and $\sigma(\xi)$ has order p iff $\Phi(\xi)$ has a zero of order p at $\xi = 1$.

Proof. With a difference operator of form (3) we associate the function of the complex variable ξ

$$\Phi(\xi) = (\log(\xi))^{-1} \rho(\xi) - \sigma(\xi).$$

The function $\log \xi$ is made singlevalued by cutting the complex plane along the negative real axis and setting $\log 1 = 0$.

Suppose the difference operator has order p . Then (3) is $O(h^{p+1})$ for any sufficiently differentiable function $y(x)$. Choose $y(x) = e^x$.

$$\begin{aligned} \text{Then } L(e^x; h) &= e^x \{ \rho(e^h) - h \sigma(e^h) \} \\ &= e^x C_{p+1} h^{p+1} + O(h^{p+2}) \end{aligned}$$

as $h \rightarrow 0$ where $C_{p+1} \neq 0$.

Hence the function $f(h) = \rho(e^h) - h \sigma(e^h)$, which is holomorphic at 0 has a zero of order $p+1$ and $h^{-1} f(h)$ a zero of order p at $h = 0$.

Since the map $\xi \rightarrow e^h$ maps a nbd. of $h = 0$ in a 1-1 manner onto a nbd. of $\xi = 1$, it follows that $\Phi(\xi) = (\log \xi)^{-1} f(\log \xi)$ has a zero of order p at $\xi = 1$.

Conversely, assume $\phi(\xi)$ has a zero of order p at $\xi = 1$.

Hence $f(h) = h \phi(e^h)$ has a zero of order $p+1$ at $h = 0$.

$$\begin{aligned} \text{Hence } L(e^x, h) &= e^x \{ \rho(e^h) - h \sigma(e^h) \} \\ &= e^x C_{p+1} h^{p+1} + O(h^{p+2}) \end{aligned}$$

for some non-zero C_{p+1} .

Hence the order of $L(y(x); h)$ is p when

$$y(x) = e^x.$$

Since the order is only dependent on the α_i, β_i , the order of L is p .

Theorem 2.2 : $\rho(\xi)$ as above, $k' \in \mathbb{N}$ with $0 \leq k' \leq k$. Then there is a unique polynomial $\sigma(\xi)$ of degree $\leq k'$ such that the order of $L(y(x); h)$ is at least $k' + 1$.

Proof. The function $(\log \xi)^{-1} \rho(\xi)$ is holomorphic at $\xi = 1$ hence

$$\frac{\rho(\xi)}{\log(\xi)} = C_0 + C_1 (\xi-1) + \dots$$

Setting $\sigma(\xi) = C_0 + C_1 (\xi-1) + \dots + C_{k'} (\xi-1)^{k'}$, $\phi(\xi)$ has a zero of multiplicity $k' + 1$ at $\xi = 1$ and by Lemma 2.1, $L(y(x); h)$ has order $\geq k' + 1$. ($> k' + 1$ if $C_{k'+1} = 0$).

Conversely if L has order $k' + 1$ then $\phi(\xi)$ has a zero of multiplicity $k' + 1$ at $\xi = 1$.

Since the Taylor expansion is unique

$$\sigma(\xi) = C_0 + C_1 (\xi-1) + \dots + C_{k'} (\xi-1)^{k'}.$$

Note. The cases of practical value are

$k' = k-1$ (best explicit operator) and $k' = k$ (best implicit operator).

Dahlquist [22] has shown that by choosing $\rho(\xi)$ suitably, the order of the operator may be as high as $2h$. However the result has little practical significance since he also proved the following results.

Theorem 2.3 : The order of a stable operator cannot exceed $k+2$. A necessary and sufficient condition for $p = k+2$ is that k be even, that all roots of $\rho(\xi)$ have modulus 1 and that $\sigma(\xi)$ be of the form

$$\sigma(\xi) = \frac{(1-z)^k}{2} \rho\left(\frac{1+\xi}{1-\xi}\right)$$

$$\text{where } \xi = \frac{1+z}{1-z}.$$

Furthermore

Theorem 2.4 : The order of a stable operator whose step number k is odd cannot exceed $k+1$.

Proof. Let $\rho(\xi)$ be a polynomial satisfying the conditions of consistency and stability. Introduce a new variable $z = x + iy$ by setting

$$z = \frac{\xi-1}{\xi+1}, \quad \xi = \frac{1+z}{1-z}$$

Instead of the polynomials $\rho(\xi)$ and $\sigma(\xi)$ we consider the functions

$$r(z) = \left(\frac{1-z}{2}\right)^k \rho\left(\frac{1+z}{1-z}\right); \quad s(z) = \left(\frac{1-z}{2}\right)^k \sigma\left(\frac{1+z}{1-z}\right)$$

which are also polynomials of degree $\leq k$. Since $\xi = 1$ is a simple root of $\rho(\xi)$, $z = 0$ is a simple root of $r(z)$.

Hence $r(z) = a_1 z + a_2 z^2 + \dots + a_k z^k$ where $a_1 \neq 0$. The a_i , $i = 1, \dots, k$ are real and without loss of generality we may assume $a_1 > 0$ (otherwise multiply $\rho(\xi)$ by a suitable constant).

Then $a_\mu \geq 0$ for $\mu = 1, \dots, k$.

Let the roots of $r(z)$ be $x_\nu + iy_\nu$ and let a_λ be the non-zero coefficient with highest index. i.e. $r(z) = a_\lambda z \prod_{\lambda} (z-x_\nu) \prod_{\mu} ((z-x_\mu)^2 + y_\mu^2)$

where λ ranges over the real roots and μ over the complex. By stability, $x_\nu \leq 0$ for all roots and so all non-zero coefficients of $r(z)$ are positive. Consider the function

$$\psi(z) = \left(\frac{1-z}{2}\right)^k \phi\left(\frac{1+z}{1-z}\right) = \frac{1}{\log\frac{1+z}{1-z}} r(z) - s(z)$$

$\psi(z)$ has a zero of order p at $z = 0$ iff $\phi(\xi)$ has a zero of order p at $\xi = 1$ and thus by Lemma 2.1 iff the operator L defined by $\rho(\xi)$ and $\sigma(z)$ has order p .

Thus if L has order p , $s(z) = b_0 + b_1 z + \dots + b_{p-1} z^{p-1}$ where

$$\frac{z}{\log\frac{1+z}{1-z}} \frac{r(z)}{z} = b_0 + b_1 z + \dots$$

Since the degree of $s(z)$ must not exceed k , for the existence of a stable operator with $p > k+1$, $b_{k+1} = \dots = b_{p-1} = 0$.

$$\text{Setting } \frac{z}{\log\frac{1+z}{1-z}} = C_0 + C_2 z^2 + C_4 z^4 + \dots$$

and defining $a_\nu = 0$ for $\nu > k$, we have

$$b_0 = C_0 a_1$$

$$b_1 = C_0 a_2$$

$$b_{2\nu} = C_0 a_{2\nu+1} + C_2 a_{2\nu-1} + \dots + C_{2\nu} a_1$$

$$b_{2\nu+1} = C_0 a_{2\nu+2} + C_2 a_{2\nu} + \dots + C_{2\nu} a_2 \quad \nu = 1, 2, \dots$$

Now $C_{2\nu}$ satisfies $C_{2\nu} < 0$ $\nu = 1, 2, \dots$

Hence if k is odd

$$b_{k+1} = C_2 a_k + C_4 a_{k-2} + \dots + C_{k+1} a_1$$

Since $a_1 > 0$ and no $a_\nu < 0$ it follows that $b_{k+1} < 0$.

Theorem 2.4 is proved.

$$\text{If } k \text{ is even, } b_k = C_2 a_k + \dots + C_k a_2.$$

A necessary and sufficient condition that $b_{k+1} = 0$, is that

$$a_2 = a_4 = \dots = a_k = 0. \text{ i.e. iff } r(z) \text{ satisfies } r(-z) = -r(z).$$

Now the map $\xi \rightarrow \frac{1+z}{1-z}$ maps the unit circle onto the half-plane.

If $z \leq 0$, the stability condition on L sets the following conditions on $r(z)$:

(i) the roots of $r(z)$ have non-positive real parts

(ii) there are no multiple roots of $r(z)$ on the imaginary axis.

Hence all roots of $r(z)$ lie on the imaginary axis, which means all roots of $\rho(\xi)$ lie on $|\xi| = 1$. Since $a_k = 0$, the degree of $r(z)$ is $k-1$ and -1 is a root of $\rho(\xi)$.

Since $b_{k+2} = C_4 a_{k-1} + C_6 a_{k-3} + \dots + C_{k+2} a_1$ is negative, the order cannot exceed $k+1$.

Hence theorem 2.3 is proved.

Corollary 2.5 : If an operator of even order k is stable then the conditions

$$a_v = -a_{k-v}, \quad b_v = b_{k-v}$$

are both necessary and sufficient that it be of maximum order, i.e. $(k+2)$. All roots of $\rho(\xi)$ then lie on the unit circle.

Proof. Immediate from the definition of $r(z)$ and $s(z)$. Henrici [21] defines an operator satisfying the conditions of Theorem 2.3 to be optimal and gives two procedures for constructing optimal operators from a given $\rho(\xi)$.

Notes and Remarks. In his paper [22] Dahlquist distinguished between "strong" instability and "weak" instability. Strong instability in our terms means an unstable difference scheme, i.e. these schemes with order $p > k + 2$ which show such a growth of error and are so extremely sensitive to perturbations which must arise in the computation of the solution, either in the finite arithmetic used or the choice of values $\gamma_1, \gamma_2, \dots, \gamma_{k-1}$, that they are unsuitable for computational purposes. Weak instability occurs in formulas of order $p = k + 2$ whose effect

over some interval can be made arbitrarily small, provided h has been chosen suitably small. However it is not necessary to reject these formulas since good results may be obtained, but precautions should be taken when transients in the solution of the differential equation are damped out quickly in time, for the difference equations possess oscillating components whose amplitudes increase with the rate of damping of the transients. So when these formulas are used, the y_1, \dots, y_{k-1} should be determined to the same accuracy of the computations following and the round off errors should be kept smaller than the local truncation error. For a fuller discussion on these matters see Stetter [26].

CHAPTER 5

Discretizations

Section 1. The concept of a discretization.

Consider the problem $y = F(b)$ (1)

F a continuous map, $\mathcal{D}(F) \subset E$, $\mathcal{R}(F) \subset E^0$; E, E^0 Banach spaces. Assume the existence and uniqueness of a solution to (1).

A discretization of (1) is a family

$$\Omega = \{\phi_h, E_h, E_h^0, \Delta_E^h, \Delta_{E^0}^h\}, \quad h \in H \subset (0, h)$$

where E_h, E_h^0 are finite dimensional Banach spaces, the maps

$\phi_h : E_h \rightarrow E_h^0$ is continuous and $\Delta_E^h : E \rightarrow E_h^0$ are continuous and $\Delta_E^h : E \rightarrow E_h, \Delta_{E^0}^h : E^0 \rightarrow E_h^0$ are linear, bounded and such that

$$n^h = \phi_h(\Delta_E^h b) \text{ which is to approximate } \Delta_{E^0}^h y.$$

Definition 1.1: Ω is convergent of order p at b if the discretization error

$$\begin{aligned} n^h - \Delta_{E^0}^h y &= \phi_h(\Delta_E^h b) - \Delta_{E^0}^h G(b) \\ &= O(h^p) \quad \text{for all } h \in H. \end{aligned}$$

Definition 1.2: Ω is stable on $\{B_h : h \in H\}$, $B_h \subset E_h$ if ϕ_h satisfies a Lipschitz condition uniformly on H .

i.e. $\|\phi_h(\beta_1^h) - \phi_h(\beta_2^h)\| \leq L \|\beta_1^h - \beta_2^h\|$, for all $\beta_1^h, \beta_2^h \in B_h$, for all $h \in H$

We define with Watt [23] the inverse discretization to Ω , as the discretization Ω^* of

$$b = F^{-1}(y), \quad \Omega^* = \{\phi_h, E_h, E_h^0, \Delta_E^h, \Delta_{E^0}^h\}$$

where $\phi_h^{-1} = \phi_h^{-1}$ is assumed to exist.

Definition 1.3: Ω is consistent of order p if

Ω^* is convergent of order p .

Hence Ω is convergent if it is consistent and stable on a family of sets.

We note for future reference that Stetter 24 defines the properties of consistency and stability on Ω^* rather than on Ω . One merely applies the inverse map then to obtain the given stability condition.

Example. Consider in the unit square

$$\frac{dy}{dt} = \frac{1}{2} \frac{\partial^2}{\partial x^2} y^2$$

with the conditions

$$Fy : \frac{\partial}{\partial t} y(x,t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} (y(x,t))^2$$

$$(x,t) \in U = \{(x,t); 0 < x,t < 1\}$$

$$y(x,0) = \overline{y(x)}$$

$$y(0,t) = y_l(t)$$

$$y(1,t) = y_r(t).$$

F incorporates the conditions under suitable restrictions on its domain.

Project the usual grid onto the unit square. Then E_h and E_h^0 consist of the functions from the grid $G_h \rightarrow \mathbb{R}$

$$\|h\| = \{\max_{j,n} |n_j^n| ; n \in E_h, n_j^n = n(x_j, +n)\}$$

$$\|\delta\| = \max \{ \max_j |\delta_j^0|, \max_n |\delta_0^n|, |\delta_N^n| \} + \max_{0 < j < N} |\delta_j^n| ; \delta \in E_h^0$$

and are Banach spaces under these norms.

Δ_h, Δ_h^0 are the restriction maps,

i.e. the maps that assign functions on the unit square into \mathbb{R} to functions on the grid into \mathbb{R} .

Section 2. The strong stability concept of Stetter.

The behaviour of small perturbations of

$$Fy = 0 \quad F : E \rightarrow E^0$$

is characterized by the solution e of

$$F'(y_0)e = d \quad d \in E^0$$

F' the Fréchet derivative.

Note. The problem is said to be properly posed if e is obtained from

d by a bounded linear operator $e = G d$

$$\text{where } G = F'(y_0)^{-1}.$$

We would like the discretization to be no more sensitive to perturbations than the original problem. So we similarly define

$$\Gamma(h) = (\Phi(\Delta_h y_0))^{-1} \text{ of the variational equation.}$$

Then

Proposition 2.1: For a stable algorithm $\Gamma(h)$ exists and is uniformly bounded for $0 < h \leq h_0$, some h_0 .

Let us discretize $G : E^0 \rightarrow E$ is an asymptotic inverse of Δ_h^0 ,

i.e. $\Delta_h^0 : E_h^0 \rightarrow E^0$ s.t. $\|\Delta_h^0 \Phi_h^0 - I\| \rightarrow 0$

$$\text{and } \|\Phi_h^0\| \frac{h}{0} \rightarrow 1.$$

Then $G(h) : E_h^0 \rightarrow E$ is defined by

$$G(h) = \Delta_h^0 G \Phi_h^0 \text{ and this is unique}$$

$$\|G_1(h) - G_2(h)\| \frac{h}{0} \rightarrow 0 \text{ for } G_1 \ll \Phi_h^{*1}, G_2 \ll \Phi_h^{*2}$$

Definition 2.2: The algorithm is strongly stable for sufficiently small

h if

$$\lim_{h \rightarrow 0} \|P(h) - G(h)\| = 0.$$

Since E_h, E_h^0 are finite dimensional they have a natural basis of

functions vanishing except at one grid point. Let Γ_{mn}, G_{mn} denote the matrices corresponding to the maps $\Gamma(h), G(h)$ respectively.

Theorem 2.3: A stable algorithm $\phi_h(\eta) = 0$ is strongly stable if for sufficiently small h ,

$$\lim_{h \rightarrow 0} \|I - \phi'_h(\Delta_h y_0) G(h)\| = 0.$$

Proof. $\Gamma - G = \Gamma(I - \phi'_h G)$

and $\|\Gamma(h)\|$ is bounded as $h \rightarrow 0$ by the stability of ϕ_h .

For a system of first order differential equations it may be proved, see Stetter [26], that a multistep algorithm is characterized by the non-existence of any roots of $\rho(z)$ of modulus 1, except $z = 1$. This concept of strong stability thus precludes the possibility of errors in a convergent method from growing relative to the exact solution.

Section 3. The Stability of a non-linear algorithm.

Non-linear instability concerns the unstable behaviour of a non-linear discretization algorithm, whose linearization is stable at the true solution. We shall follow Stetter [24] in endeavouring to predict this behaviour from the relationship between a non-linear algorithm and its linearization.

The algorithm $\phi_h \eta = 0$ is consistent if

$$\lim \|(\phi_h \Delta_h - \Delta_h^0 F)Z\| = 0, \quad h \in (0, T), \quad z \in E.$$

Note. This serves to normalize $\{\phi_h\}h \rightarrow 0$.

Definition 3.1: The local discretization error

$$\lambda(h) = -\phi_h \Delta_h y_0 \in E_h^0.$$

The global error $\epsilon(h) = \eta_0(h) - \Delta_h y_0 \in E_h$

$$\text{where } \phi_h \eta_0(h) = 0.$$

The discretization is then convergent when $\lim \epsilon(h) = 0$. Obviously consistency implies convergence when small local errors imply global errors

i.e. since $\phi_h \eta(h) = \phi_h (\Delta_h y_0 + \epsilon(h)) = 0$

$$\phi_h \Delta_h y_0 = -\lambda(h)$$

that $\|\lambda(h)\| \rightarrow 0 \Rightarrow \|\epsilon(h)\| \rightarrow 0$ we must have

$$\begin{aligned} \text{If } \epsilon \in E_h \text{ satisfies } \phi_h (\Delta_h y_0 + \epsilon) - \phi_h \Delta_h y_0 & \quad (S) \\ & = \delta \in E_h^0 \end{aligned}$$

then $\|\epsilon\| \leq M \|\delta\|^\beta \quad \beta > 0, \quad M$ independent of h .

If (S) holds the algorithm is called stable since it then conforms to the basic interpretation of boundedness of the global effects of a local error.

For a linear algorithm $\Phi_h \eta = \Delta_h \eta - \gamma_h$
 where $\Delta_h : E_h \rightarrow E_h^0$ is linear, $\gamma_h \in E_h^0$
 (S) is valid if $\|\Lambda_h^{-1}\| \leq M$. See [7].

For nonlinear Φ_h we shall see under what conditions (S) is valid.

Define for $z \in E$, $\psi_h(\epsilon) = \Phi_h(\Delta_h z + \epsilon) - \Phi_h' \Delta_h z$

$$\psi_h : E_h \rightarrow E_h^0.$$

Knowing $\psi_h 0 = 0$ we must derive the existence of a family of inverse functions $\phi : E_h^0 \rightarrow E_h$ uniformly bounded for $(0, T)$.

Formulating the Inverse Function theorem suitably we obtain:

Theorem 3.2: Uniformly for $h \in H$ let

(a) Φ_h have a Fréchet-derivative $\Phi_h'(\xi)$ for

$\|\xi - \Delta_h z\| \leq S > 0$ which is Hölder-continuous at $\Delta_h z$,

i.e. $\|\Phi_h'(\xi) - \Phi_h'(\Delta_h z)\| \leq L \|\xi - \Delta_h z\|^\alpha, 0 < \alpha \leq 1$

and

(b) $\Phi_h'(\Delta_h z)^{-1}$ exist and satisfy

$$\|\Phi_h'(\Delta_h z)^{-1}\| \leq S$$

Then for $\|\delta\| \leq r$

$\Phi_h(\Delta_h z + \epsilon) - \Phi_h' \Delta_h z = \delta$ has a unique solution $\epsilon = \phi \delta$ such that
 $\|\epsilon\| \leq M \|\delta\|$ where $r = \frac{\alpha}{\alpha+1} (S(LS)^{1/\alpha})$ and $M = \frac{\alpha+1}{\alpha} S$.

Hence under the assumptions (a) and (b), (S) is valid if the perturbation is sufficiently small. It is natural to make this the definition of stability.

Definition 3.3: The discretization is stable at $z \in E$ if there exists $M, r > 0$ s.t. for all $\delta \in E_h^0$ $\|\delta\| \leq r, h \in (0, T)$

$\Phi_h(\Delta_h z + \epsilon) - \Phi_h' \Delta_h z = \delta$ has a unique solution $\epsilon = \epsilon_h(\delta)$ such that

$$\|\epsilon_h(\delta)\| \leq M \|\delta\|$$

r is called the stability threshold, M the stability bound.

It is obvious that stability at z ensures convergence.

If the algorithm is linear then $\|\Delta_h^{-1}\| \leq M$ and no stability threshold exists.

Assuming that a linearized discretization $\phi'_h(\xi)$ exists of ϕ_h for all ξ in a sufficiently large vicinity of $\Delta_h z$ we have the following formulation of Theorem 3.2.

Theorem 3.4: Let the linearized discretization be

- (a) Hölder-continuous at z uniformly in h
and (b) stable at z .

Then the non-linear discretization is stable at z with the above estimates for r and M .

Unfortunately this is of little practical use since negative powers of h often occur in ϕ'_h . We thus change our definition slightly:

Definition 3.5: A discretization is m -restricted stable at $z \in E$ if

there exists $M, r, m > 0$ s.t. for all $\delta \in E_h^0, \|\delta\| \leq rh^m, h \in (0, T)$

$$\phi_h(\Delta_h z + \epsilon) - \phi_h(\Delta_h z) = \delta \text{ implies } \|\epsilon\| \leq M \|\delta\|.$$

Corollary 3.6: Let the linearized discretization be such that

- (a) $\|\phi'_h(\xi) - \phi'_h(\Delta_h z)\| \leq L_h^{-h} \|\xi - \Delta_h z\|^\alpha, 0 < \alpha \leq 1$.
(b) stable at z .

Then the non-linear discretization is $\frac{n}{\alpha}$ -restricted stable at z .

Notes and Remarks. Stetter has also proved that continuity and stability of the linearization exist not only at $\Delta_h z$ but also in a norm vicinity

of $\Delta_h z (6 E_h)$ the size of which does not tend to zero with h .

We remark that this too does not give a practical result since most algorithms for the variable coefficient case require a certain smoothness of coefficients (see the discussion on this problem in Chapter 3, Section 3) and the neighbouring elements of $\Delta_h z$ need not be discretizations of given neighbours of z in the stability region as $h \rightarrow 0$.

Hence from the structure of the relationship between a nonlinear discretization and its linearization, irrespective of the problems considered, we conclude

- (a) the occurrence of stability thresholds is to be expected
- (b) stability thresholds may decrease with a power of h .
- (c) the continuity properties of the linearization may have to be considered.

Appendix A. A well-posed Cauchy problem.

Given the problem $\frac{d}{dt} u(t) = Au(t), 0 \leq t \leq T$

$$u(0) = u_0$$

In a Banach space X .

Definition: A genuine solution $u(t)$ is in the domain of A such that

$$\left\| \frac{u(t+\Delta t) - u(t)}{\Delta t} - A u(t) \right\| \xrightarrow{\Delta t} 0 \quad \text{for } t \in (0, T).$$

A problem is characterized as properly posed if the family of genuine solutions is sufficiently large and if the solutions depend uniquely and continuously on the initial data in a certain sense. See L. Payne [35] for a discussion of this point.

Let D be the set of elements $u_0 \in X$ with a genuine solution $u(t)$ exists with $u(0) = u_0$. The correspondence between u_0 and $u(t)$ determines a linear map $E_0(t)$ such that

$$E_0(t) u_0 = u(t) \quad \text{for all } u_0 \in D.$$

Definition: The initial value problem determined by A is properly posed in the sense of Hadamard if

- (I) $D(E_0(t))$ is dense in X .
- (II) $\|E_0(t)\| \leq K$ some K , for $t \in (0, T)$.

It is easy to see then that a genuine solution must be continuous and the convergence in t for $\frac{u(t+\Delta t) - u(t)}{\Delta t} \rightarrow Au$ is uniform.

In an analogous way the concept of well-posedness may be defined for the following problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= Au + g(t) \\ u(0) &= u_0 \end{aligned}$$

where $g(t)$ is piecewise uniformly continuous in t , since under some minor restrictions the solution $u(t)$ satisfies

$$u(t) = E(t) u_0 + \int_0^t E(t-s)g(s)ds.$$

$u(t)$ is then known as a generalized solution. See Thompson [10] who also discusses the question when $g(t)$ is replaced by $g(t, u(t))$ defined for $0 \leq t \leq T$, continuous in t for each u and uniformly Lipschitz continuous with respect to u .

Strang [13] has discussed the problem of finding sufficient conditions for a Cauchy problem of form

$$\frac{\partial u}{\partial t} = \sum_{|\alpha| \leq m} G_\alpha(x) D^\alpha u = Lu$$

where $D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$ to be well-posed in the sense that L which is

a closed densely defined operator generates a C_0 semi-group e^{tL} in L^2 .

Appendix B. Perturbations of stable operators.

Given $\frac{du}{dt} = Au$ where A is a linear t -independent closed, densely defined operator in X . If there exists a consistent stable difference scheme approximating the problem, then the perturbed problem

$$\frac{du}{dt} = Au + Bu \text{ has also a stable, consistent scheme if}$$

B is bounded. See Richtmyer and Morton [7].

Sunouchi [14] has extended this result to the unbounded case.

Theorem: Let B be an (unbounded) operator $D(B) \subset D(A)$ and B_n a consistent approximation to B .

i.e. $B_n u \rightarrow B u$ for all $u \in D(B)$.

$$\text{Assume } N_\lambda = \sup_{\|u\|=1} \sup_n h_n \sum_j \exp(-\lambda_j h_n) \|B_{h_n}^j u\|$$

$$< 1 \quad \text{for sufficiently large } \lambda$$

$$T_n = (T/h_n).$$

Then the scheme $u_n(\Delta t) = C_n u$

$$u_n(m\Delta t) = (C_n(1 + \Delta t B_n)) C_n^{m-1} u$$

is stable and consistent with the perturbed problem.

Theorem: Assume B consistent and $D(B) \subset D(A)$.

$$\text{Also assume } \sup \Delta t \sum_{j=1}^n \|B_n C_n^j u\| < \infty, \text{ for all } u \in X$$

$$\text{where } \Delta t = \left(\frac{1}{\Delta T}\right) + 1.$$

Then there exists $\varepsilon_0 > 0$ s.t. for all $\varepsilon : |\varepsilon| < \varepsilon_0$ the above scheme with εB_n replacing B_n is stable and consistent with

$$\frac{du}{dt} = Au + \varepsilon Bu.$$

Appendix C. The uniform convergence of matrix powers.

Oldenburger proved that for a given matrix A , $A^n \xrightarrow[n \rightarrow \infty]{} A^\infty$ iff the eigenvalues of A were less than 1 except for some eigenvalues equal to one, each of which corresponds to a linear elementary divisor

Recall Schur's theorem that any matrix A is unitarily similar to an upper triangular matrix; and an arbitrary ordering of eigenvalues can be achieved along the diagonal. Hence assume that

$$A = \begin{pmatrix} E & & C \\ \vdots & \ddots & \vdots \\ 0 & & T \end{pmatrix} \quad (1)$$

Definition: The convergence factor x_A of A is

$$x_A = \max_{\lambda_i < 1} |\lambda_i|$$

$\lambda_1, \dots, \lambda_k$ eigenvalues of $K \times K$ matrix A .

For matrices of form (1) $x_A = r(T)$.

Lemma: If A is of form (1), then A^n converges iff $E = I$.

Proof. \Rightarrow : A^n converges hence E^n converges.

By Oldenburger there exists S such that $E = SIS^{-1} = I$.

$$\begin{aligned} \Leftarrow : E = I \Rightarrow A^n &= \begin{pmatrix} I & C & \sum_{i=0}^{n-1} T^i \\ \vdots & \vdots & \vdots \\ 0 & & T^n \end{pmatrix} \\ &\xrightarrow[n \rightarrow \infty]{} \begin{pmatrix} I & C(I-T)^{-1} & \\ \vdots & \vdots & \\ 0 & & 0 \end{pmatrix} \end{aligned}$$

Theorem: (Buchanan and Parlett) [15]. F a family of $K \times K$ matrices.

Then the sequences $\{A^n\}$ converge uniformly if Oldenburger's condition holds for all $A \in F$ and there exists x, N dependent on F such that

(a) $x_A \leq x < 1$

(b) $(x_A)^N A$ are uniformly bounded.

Note. Convergence may be uniform and yet the limits may be bounded.

Theorem: A^∞ is (uniformly) bounded in F iff A is boundedly similar to the direct sum of an identity matrix and a convergent matrix.

Proof. $x_A \leq x \leq 1 \Rightarrow T$ is bounded iff $(I-T)^{-1}$ is bounded.

Now

$$A^\infty = \begin{vmatrix} E & C(I-T)^{-1} \\ 0 & S \end{vmatrix}$$

and $C(I-T)^{-1}$ is bounded if C, T (hence A) is. However when $C = 0$, then A is bounded independently of T .

Conversely if $C(I-T)^{-1} = B$ is (uniformly) bounded in F then

$$Q^*AQ = \begin{pmatrix} I & C \\ 0 & T \end{pmatrix} = \begin{pmatrix} I & -B \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ 0 & T \end{pmatrix} \begin{pmatrix} I & B \\ 0 & I \end{pmatrix}$$

some unitary matrix Q .

R E F E R E N C E S

1. Dunford N. and Schwartz J. Linear Operators Part 1. Pure and Applied Mathematics. Interscience Publishers Inc. New York 1958.
2. Taylor A.E. Introduction to Functional Analysis. John Wiley and Sons Inc. New York 1958.
3. Larionov E.A. On a stability criterion for solutions of differential equations with periodic operator coefficients in a Banach space. Soviet Math. Doklady Vol. 9 (1968) No. 6.
4. Dunford N. Spectral Operators. Pacific Journal of Mathematics 4, (1954).
5. Fougel S.R. The relations between a spectral operator and its scalar part. Pacific Journal of Mathematics, 8 (1958).
6. Kakutani S. An example concerning uniform boundedness of spectral measures. Pacific Journal of Mathematics, 4, (1954).
7. Richtmyer R and Morton K.W. Difference methods for Initial-value problems. Interscience Publishers, New York 1957.
8. Ansoerge R. Konvergenz von Mehrschrittverfahren zum Lösung halblinearer Anfangswertaufgaben. Num. Math. 10, (1967)
10. Thompson R.J. Difference approximations for inhomogeneous and quasi-linear equations. J. Soc. Indus. Appl. Math. 12, 1964
11. Morton K.W. and Schechter S. On the stability of finite difference matrices. SIAM J. Num. Anal. Series B 2, (1965).
12. Kato T. Estimation of iterated matrices with application to the von Neumann condition. Numer. Math. 22, (1960).
13. Strang G. Necessary and Insufficient conditions for well-posed Cauchy problems. J. Differential Equations, 2. (1968).
14. Sinouchi H. Perturbation theory of difference schemes. Numer. Math. 12, (1968).
15. Buchanan M.L. and Parlett B. The uniform convergence of matrix powers in Hilbert space. Numer. Math. 10, (1966)
16. Thomeé V. On maximum norm stable difference operators, Numerical Solution of Partial Differential Equations (J. Bramble ed.) Academic Press, New York 1966.

17. Strang G. Polynomial approximation of the Bernstein type.
Trans. Amer. Math. Soc. 105, (1962).
18. Thomeé V. Stability of difference schemes in L^p .
Congr. Math. Scand. 1964 (to appear).
19. Lax P.D. and Wendroff B. Systems of conservation laws.
Comm. Pure Appl. Math. 13, (1960).
20. Widlund O. Stability of Parabolic Difference Schemes in the Maximum norm. Numer. Math. 8, (1966)
21. Henrici P. Discrete Variable Methods In Ordinary Differential Equations. John Wiley and Sons, New York 1961.
22. Dahlquist G. Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. 4. (1956).
23. Watt J.M. Consistency, convergence and stability of general discretizations of the initial value problem. Numer.Math.12 (1968).
24. Stetter H.J. Stability of non-linear discretization algorithms. Numerical solution of partial differential equations (J. Bramble ed) New York. Academic Press, New York 1966.
25. Butcher J.C. On the convergence of numerical solutions to ordinary differential equations. Math. of Computation, 20, (1966).
26. Stetter H.J. A study of strong and weak stability in discretization algorithms. SIAM J. Numer. Anal. Series B, 2, (1965).
27. Wermer J. Commuting spectral measures on Hilbert space.
Pac. J. of Math. 4, (1954).
28. Naimark M.A. Normed Rings. P. Noordhoff, Groningen 1964.
29. von Neumann J. First report on the numerical calculation of flow problems. Collected Works V. Pergamon Press, New York 1963
30. Trotter H.F. Convergence of semi-groups of operators.
Pac. J. of Math. 8 (1958).
31. de Sz. Nagy B. Uniformly bounded linear transformations in Hilbert space.
Acta Sc. Math. 15 (1947).
32. Parlett B.N. Accuracy and dissipation in difference schemes.
Comm. Pure and Appl. Math. 19 (1966).
33. Lax P.D. Linear and non-linear difference schemes. Numerical solution of Partial Differential Equations. Academic Press, New York 1966.

34. Wendroff B. Well-posed problems and stable difference schemes.
SIAM J. Numer. Anal. 5 (1968).
35. Payne L. On some non well posed problems for partial differential equations. Numerical Solutions of Nonlinear Differential equations (D. Greenspan ed.) John Wiley and Sons, New York 1966.
36. Markov A. Quelques Théorèmes sur les Esembles Abéliens.
Translated from Dokl. Akad. Nauk SSSR 1(X) (1936).