J M SANZ-SERNA

Stability and convergence in numerical analysis – I: Linear problems – a simple, comprehensive account

In memory of Pepe

0. INTRODUCTION

The concepts of stability and convergence are crucial in the analysis of numerical methods in differential and integral equations. It is therefore with regret that we have often encountered practitioners in the field who did not possess a sufficient insight into these notions. In our opinion, two reasons account for this situation. First, a general, abstract presentation of the concepts requires, of necessity, some functional analysis, a subject which sometimes is not fully appreciated by those numerical practitioners whose basic training is not in mathematics. Secondly, and perhaps as a consequence of the previous observation, the ideas of stability and convergence tend to be treated as they apply to a given, concrete problem or set of problems (e.g. linear multistep methods in ODEs, Lax-Richtmyer theory in initial value problems in PDEs, etc.). And this is true both in the research literature and in the classroom.

The aim of this paper is to present a simple but comprehensive view of the fundamental concepts of stability and convergence as they apply to *linear* problems. Subsequent papers will treat *nonlinear* situations. Our emphasis is not in developing or creating a new abstract framework. We rather put together a number of ideas commonly used in various fields in numerical analysis and show their mutual relationships. In this sense our work is complementary, rather than alternative, to several known general theories of discretizations (e.g. Aubin [3], Stummel [43], Vainikko [46], Stetter [39], Spijker [37]).

The present paper includes examples from integral, ordinary differential and partial differential equations; initial and boundary value problems; finite-difference and Galerkin methods. We survey a number of ways of proving stability: use of maximum principles, the energy method, von Neumann analysis, regularity, collective compactness etc, thus providing an introduction to the study of the analysis of numerical methods. We have deliberately restricted

the use of functional analysis to the bare, indispensable minimum. It is hoped that in this way the paper will be beneficial to a wide audience.

The ideas have been grouped in what we call the *first* and *second paradigms*. In the first paradigm, the only elements that feature in the analysis are the discrete equations, $A_h U_h = f_h$ being studied and (a discretization of) the theoretical solution u being approximated. There is no need to introduce restriction or prolongation operators. The second paradigm investigates the *interplay* between the original problem $Au = f$ and its discretizations $A_h U_h = f_h$. The first paradigm is simpler (and therefore more versatile and less powerful) than the second. Most current research papers (particularly in nonlinear situations) are written within a first paradigm setting. On the other hand, the general theories of discretizations usually work within a second paradigm framework. An important exception is given by Spijker's thesis [35], which is limited to initial value problems.

The paper is divided into five chapters. The first describes the basic ideas of the first paradigm: consistency, stability and convergence. The chapter concludes with the (trivial) proof of the most important theorem in numerical analysis: consistency and stability imply convergence. The second chapter studies the question as to whether consistency and stability are *necessary* for convergence. Stability is not: there exist convergent discretizations that are not stable. Accordingly we introduce the stronger concept of L-convergence (i.e., convergence under perturbations) in such a way that L-convergent discretizations are necessarily stable. It is also possible to have convergence without consistency or convergence of order p with consistency of order less than p. Unknown to many, this is quite a common occurrence in practical situations. In Section 2.3 we show (following a communication of R.D. Grigorieff) that central differences on a nonuniform grid achieve second order of convergence in the sup norm, a fact some people are not aware of. The characterization of those circumstances under which the order of convergence cannot be higher than that of consistency leads to the idea of uniformly bounded discretizations.

The third chapter presents the second paradigm. The examples there contain an account of the classical ideas of the Lax-Richtmyer theory [23]. The fourth chapter examines the useful notion of regular approximation, while the last particularizes all the previous material to the highly important case of initial value problems.

The references provided are not intended as a complete survey of the existing literature, a task well beyond the author's capabilities. They rather supply illustrations to some concrete points or show the way to further material in the various fields.

It is obvious that a paper such as the present one must have been influenced by a considerable number of people. I want to express my gratitude to the Numerical Analysis Group of the University of Dundee, my former teachers. I learnt from them, among many other things, that numerical analysis is about *computing numbers*. Thanks also go to Professor R.D. Grigorieff (Berlin), who made me familiar with a number of German contributions, and to Professor Guo Ben-Yu (Shanghai), who provided much initial motivation. And, last but not least, I am indebted to my colleague Dr C. Palencia for countless valuable conversations.

1. THE FIRST PARADIGM. THE BASIC THEORY

1.1 Discrete problems

We consider a given, fixed, linear differential equation problem with solution u. In most instances u cannot be readily expressed in terms of the data of the problem and then one must obtain a 'numerical' approximation U_h to u. We have appended a subscript h in order to reflect that the numerical approximation U_h depends on a (small) parameter h such as a mesh-size, element diameter, reciprocal of number of terms retained when truncating a series etc. In what follows we always assume that h takes values in a set H of positive numbers with inf H = 0.

The numerical approximation U_h is reached by solving a *discretized problem*

$$A_h U_h = f_h, \qquad (1.1a)$$

where, for each h in H, A_h is a *fixed* linear operator mapping a vector space X_h into a vector space Y_h, and f_h is a *fixed* element in Y_h. (In this paper we assume tacitly that when several vector spaces occur simultaneously, they are either all real or all complex.)

Note that at this stage one is not concerned with endowing X_h, Y_h with norms: the discrete problem (1.1a) can be formulated and the approximation U_h obtained prior to the introduction of norms to be made later for the sake of the analysis.

A natural requirement that A_h should satisfy is that the inverse A_h^{-1} exists in order to guarantee the existence and uniqueness of U_h. However the *invertibility* of A_h is not demanded at this stage, because we shall show below that this invertibility is, under appropriate hypotheses, a consequence of the stability of (1.1a). We only assume that

$$\dim(\ker(A_h)) = \operatorname{codim}(R(A_h)) < \infty, \qquad (1.1b)$$

where ker and R denote respectively kernel and range. (Recall that $\operatorname{codim}(R(A_h))$ is, by definition, the dimension of a supplementary subspace of $R(A_h)$.) The requirement (1.1b) is very weak. It is satisfied in any of the following cases:

(i) X_h, Y_h are both finite-dimensional and $\dim(X_h) = \dim(Y_h)$.

(ii) A_h is invertible.

(iii) A_h can be written in the form $A_h = B_h + C_h$, with B_h invertible and $\dim(R(C_h)) < \infty$.

Two examples will be used throughout this paper in order to illustrate, in a simple setting, the presentation.

Example A. Two-point boundary value problem. We consider the problem

$$u''(x) = f(x), \quad 0 \leq x \leq 1, \qquad (1.2a)$$

$$u(0) = u(1) = 0, \qquad (1.2b)$$

where f is a given, fixed, real continuous function. If J is a positive integer and h = 1/J, we introduce the grid-points $x_j = jh$, $j = 0,1,\ldots,J$. Replacement of the second derivative in (1.2a) by central differences leads, on taking into account (1.2b), to the discrete problem

$$(1/h^2)\,(-2U_1 + U_2) = f(x_1),$$

$$(1/h^2)\,(U_{j-1} - 2U_j + U_{j+1}) = f(x_j),\ j = 2,3,\ldots,J-2,$$

$$(1/h^2)\,(U_{J-2} - 2U_{J-1}) = f(x_{J-1}),$$

which can be rewritten in the form (1.1) as follows:

$$h^{-2}\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ 0 & 1 & -2 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 & -2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \cdot \\ U_{J-1} \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ \cdot \\ f(x_{J-1}) \end{bmatrix} \qquad (1.3)$$

Note that, rigorously speaking, x_j, U_j depend on h. This dependence and others similar to it are not displayed in the paper to simplify the notation. Here X_h, Y_h are both identical to the space of real $(J-1)$-vectors. The requirement (1.1b) is satisfied, since $\dim(X_h) = \dim(Y_h) = J-1$.

Example B. Periodic initial-value problem for the linear convection equation. We now consider the problem

$$u_t = -u_x, \quad -\infty \leq x \leq \infty, \quad 0 \leq t \leq T < \infty, \qquad (1.4a)$$

$$u(x,0) = \eta(x), \quad -\infty < x < \infty, \qquad (1.4b)$$

$$u(x+1,t) = u(x,t), \quad -\infty < x < \infty, \quad 0 \leq t \leq T, \qquad (1.4c)$$

where the initial datum η belongs to the space L_p^2 of 1-periodic, complex valued functions which are square integrable in $0 \leq x \leq 1$. The solution of this problem is given by $u(x,t) = \eta(x-t)$ (cf. Section 3.1).

If h a positive parameter, r a fixed positive constant and square brackets denote integer part, we introduce the time levels $t_n = nk$, $n = 0,1,\ldots,N$, $N = [T/k]$, $k = rh$ and consider the discretized equations

$$U^0 = \eta(x), \quad -\infty < x < \infty,$$

$$(1/k)\,(U^{n+1}(x)-U^n(x)) = -(1/h)(U^n(x)-U^n(x-h)), \quad -\infty < x < \infty,$$

$$n = 0,1,\ldots,N-1,$$

based on replacement of the derivatives in (1.4a) by one-sided differences. These formulae enable us to compute recursively the functions $U^n \in L_p^2$, $n = 0,1,\ldots,N$. On introducing the identity operator I, the translation operator T_h given by $(T_h v)(x) = v(x-h)$, $-\infty < x < \infty$ and the operator

$$C_h = (1-r)I + rT_h,$$

the discretized equations can be written as the recursion

$$U^0 = \eta, \qquad (1.5a)$$

$$k^{-1}U^{n+1} = k^{-1}\, C_h U^n, \quad n = 0,1,\ldots,N-1. \qquad (1.5b)$$

These formulae can in turn be expressed in the form (1.1) as follows:

$$k^{-1}\begin{bmatrix} kI & 0 & 0 & \cdots & \cdot \\ -C_h & I & 0 & \cdots & \cdot \\ 0 & -C_h & I & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots -C_h & I \end{bmatrix} \begin{bmatrix} U^0 \\ U^1 \\ U^2 \\ \cdot \\ U^N \end{bmatrix} = \begin{bmatrix} \eta \\ 0 \\ 0 \\ \cdot \\ 0 \end{bmatrix} . \qquad (1.6)$$

Here X_h and Y_h are both identical to the space of $(N+1)$-vectors $[V^0, V^1, \ldots V^N]^T$, where each entry V^n is a function $V^n = V^n(x)$ which belongs to L_p^2. The operator A_h is clearly invertible, due to its bidiagonal structure. The inverse A_h^{-1} is explicitly given by

$$A_h^{-1} = k\begin{bmatrix} k^{-1}I & \cdot & \cdot & \cdots & \cdot \\ k^{-1}C_h & I & \cdot & \cdots & \cdot \\ k^{-1}C_h^2 & C_h & I & \cdots & \cdot \\ \cdot\ \cdot & \cdot & \cdot & \cdots & \cdot \\ k^{-1}C_h^N & C_h^{N-1} & C_h^{N-2} & \cdots & I \end{bmatrix} \qquad (1.7)$$

Note that, for reasons to be made clear later, we have chosen to retain the factor k^{-1} in both sides of (1.5b) (cf. also Stetter [39], para.2.2.2).

Remark The derivation of the discretizations (1.3), (1.6) from problems (1.2), (1.4) has been a motivated one, namely that of replacing derivatives by divided differences. However it is by no means necessary that the discretized problems resemble formally the original differential problem (cf. high order linear multistep or Runge-Kutta methods in ODEs). The way in which the discrete problem is derived is immaterial in the analysis and practical performance of a numerical method.

1.2 Global error, convergence

Let us suppose that we have formulated a discretized problem (1.1) in such a way that it possesses a unique solution U_h and let us also suppose that we have computed U_h. To what extent does U_h provide a good approximation to u? A first difficulty in answering this question stems from the fact that U_h can be completely dissimilar to u. Consider, for instance, Example A, where u is a function u(x), $0 \leq x \leq 1$, while U_h is a real (J-1)-vector. This difficulty is circumvented as follows. Since the numerical solution U_h yielded by (1.1) is bound to be an element in X_h, we first make up our minds as to which element u_h in X_h should be regarded as the most desirable numer- ical result. (For instance, in the context of Example A, we may decide that the (J-1)-vector $u_h = [u(x_1), u(x_2),\ldots,u(x_{J-1})]^T$ provides an 'ideal' num- erical result, so that we would be really pleased if the discretized problem gave $U_h = u_h$.) Once u_h has been chosen, we can define the vector $e_h = u_h - U_h$ as the *error* in the numerical approximation U_h. To distinguish between this concept of error and others to be introduced below, we say that e_h is the *global error* in U_h.

In order to measure the *size* of the global error we introduce, for each h in H, a norm $\|\cdot\|_{X_h}$ in X_h. (In Example A, the elements V_h in X_h are (J-1)-vectors with entries V_j and one can use the norm $\|V_h\|_{\infty,X_h} = \max\{|V_j| : 1 \leq j \leq J-1\}$.) Hereafter the subscript X_h will be omitted from the notation of the norm. Often norms in different spaces will simply be denoted by $\|\cdot\|$ without mention of the space.

We are now in a position to summarize the discussion above and to intro- duce the concept of *convergence*.

<u>Definition 1.1</u> Assume that for each h in H an element u_h in X_h and a norm in X_h have been chosen. Then if U_h is a solution of (1.1) the element $e_h = u_h - U_h \in X_h$ is called the global error in U_h. The discretization (1.1) is said to be convergent if there exists $h_o > 0$ such that, for each h in H, $h \leq h_o$, (1.1a) possesses a unique solution and, as $h \to 0$, $\lim \|e_h\| = 0$. The convergence is said to be of order p if $\|e_h\| = 0\ O(h^p)$.

Some remarks are in order:

(i) The convergence of a given approximation U_h can be investigated under different choices of norm and different choices of u_h. For instance, in the context of Example A, one can also consider the choice of the L^2 norm $\|V_h\|_2^2 = \sum_{j=1}^{J-1} h|V_j|^2$. (Note the normalizing factor h. This factor is not essential: any norm in X_h is eligible. However the introduction of the factor ensures that $\|V_h\|_2 \leq \|V_h\|_\infty$ and that, with our previous choice of u_h, $\lim_h \|u_h\|_2 = \|u\|_{L^2(0,1)}$ thus rendering the norm more meaningful.)

Also in Example A we could have chosen u_h to be the vector with entries

$$h^{-1} \int_{x_j-\frac{1}{2}h}^{x_j+\frac{1}{2}h} u(s)\, ds, \quad j = 1,2,\ldots,J-1;$$

i.e., we could have thought that the numerical solution obtained attempted to reproduce cell-averages rather than grid-values. This way of thinking is in fact advisable in practice when the solution u develops very steep gradients. Under those circumstances, attempts to reproduce exactly a nodal value $u(x_j)$ are doomed to fail. For further discussion of this point see Cullen & Morton [6].

(ii) There is an alternative technique for dealing with the difficulties stemming from the fact that u, U_h lie in different spaces: one can construct from U_h an element $\tilde{u}_h$ of the space X that contains u, and then compare $\tilde{u}_h$ and u_h. (In the context of Example A, one could interpolate the values U_j to obtain a function $\tilde{u}_h(x)$ defined for $0 \leq x \leq 1$.) In our opinion this technique of analysis introduces an arbitrary process of interpolation or prolongation which does not correspond to any operation actually carried out in practical implementations. Therefore the alternative technique is not to be preferred to the one previously discussed (see also Stetter [39] p. 7). For a theory based on interpolation or prolongation procedures, see Aubin [3].

In some applications (e.g. finite elements) the spaces X_h are subspaces of X and then it would be possible, in principle, to regard $u-U_h$ as error. However, note that $u - U_h = (u - u_h) + (u_h - U_h)$; the second term in the right-hand side is the global error in Definition 1.1, whereas the first term merely reflects the approximation capabilities of X_h and does not relate to the discretization (1.1). Thus, even in the case where u and U_h are directly comparable, we prefer to define the global error as $u_h - U_h$.

Before closing this section, and for further reference, we make our choices of u_h and norms in X_h in problems A and B above.

Example A. Here we set $u_h = [u(x_1), u(x_2),\dots,u(x_{J-1})]^T$ and, if $V_h = [V_1, V_2,\dots,V_{J-1}]^T$ is an element in X_h, $\|V_h\| = \max_j |V_j|$.

Example B. Now $u_h = [u(\cdot,t_0), u(\cdot,t_1),\dots,u(\cdot,t_N)]^T$. (The notation $u(\cdot,t_n)$ represents the function of x obtained when t is kept fixed $t = t_n$.) If $V_h = [V_0,V_1,\dots,V_N]^T$, with $V_n \in L_p^2$, is an element in X_h, we set $\|V_h\| = \max_n \|V_n\|_{L_p^2}$.

1.3 Local truncation error, consistency

Our aim is now to obtain *bounds* for the global error. A first step in that direction is the introduction of the *local truncation error* $l_h = A_h u_h - f_h$, an element which measures to what extent the equation (1.1a) is satisfied by u_h. The importance of l_h arises from the fact that it is often easily boundable.

Definition 1.2 Assume that for each h in H an element $u_h \in X_h$ and a norm in Y_h have been chosen. Then the element $l_h = A_h u_h - f_h$ is called the local truncation error of the discretization (1.1). The discretization is said to be consistent (resp. consistent of order $p > 0$) if, as $h \to 0$, $\lim \|l_h\| = 0$ (resp. $\|l_h\| = O(h^p)$).

Example A. With our previous choice of u_h, the j-th component of $l_h \in Y_h$ is given by

$$[l_h]_j = h^{-2}\{u(x_j-h) - 2u(x_j) + u(x_j+h)\} - f(x_j).$$

If u has four bounded derivatives in $0 \le x \le 1$, a Taylor expansion of the right-hand side, taking into account (1.2a), shows that $|[l_h]_j| \le (h^2/12)B_4$, where B_4 is a bound of $|d^4u/dx^4|$. If we choose in Y_h the maximum norm, it follows that

$$\|l_h\| = \max_j |[l_h]_j| \le (h^2/12)B_4$$

and thus the discretization is consistent of the second order.

Remark Checking consistency typically involves some sort of Taylor expansion. This demands a certain degree of smoothness in u. We shall present later (Theorem 3.4) indirect means of establishing consistency which may bypass

the need for smoothness requirements.

Example B. Here if $V_h = [V_0,V_1,\dots,V_N]^T \in Y_h$, $V_n \in L_p^2$, we employ the L^1 norm

$$\|V_h\| = \|V_0\|_{L_p^2} + \sum_{n=1}^{N} k\,\|V_n\|_{L_p^2}. \qquad (1.8)$$

Note the factor k in the right-hand side, in agreement with previous discussion. The term $\|V_0\|$ is not multiplied by k, reflecting the fact that (1.5a) does not include the factor k^{-1} as distinct from (1.5b). On the advisability of using an L^1 norm in Y_h, rather than a maximum norm, see Stetter [39] p. 75.

With our previous choice of u_h and on proceeding as in Example A, the discretization (1.6) is seen to be consistent of the first order, provided that (1.4) possesses a smooth solution. In the special case $r = 1$, the local truncation error is zero, i.e. u_h satisfies exactly the discrete equations and therefore $U_h = u_h$, since A_h is invertible.

1.4 Stability. The main theorem

Once bounds of the local truncation error l_h are available, they can be transferred to the global error by means of the idea of stability.

Definition 1.3 Assume that norms in X_h and Y_h have been chosen. The discretizations (1.1) are said to be stable if positive constants h_o, L exist such that, for each $h \le h_o$, $V_h \in X_h$,

$$\|V_h\| \le L\ \|A_h V_h\|. \qquad (1.9)$$

The constant L is the stability constant of (1.1).

It is clear that the stability of (1.1) does not depend on the right-hand sides f_h. Obviously, for stable discretizations, $\ker(A_h) = \{0\}$, which in view of (1.1b) shows that A_h^{-1} exists for $h \le h_o$, thus guaranteeing the existence of the solution U_h of (1.1a). When the existence of A_h^{-1} and the uniqueness of the solution U_h of (1.1a). When the existence of A_h^{-1} has been proved, (1.9) is equivalent to $\|A_h^{-1}\| \le L$. When (1.9) holds,

$$\|e_h\| \le L\ \|A_h e_h\| = L\ \|A_h u_h - A_h U_h\| = L\ \|A_h u_h - f_h\| = L\ \|l_h\|,$$

so that e_h can be bounded in terms of l_h through the h-independent constant L. In this simple way is proved the most important single theorem in the numerical analysis of differential equations.

Theorem 1.1 If, for given choices of u_h and norms in X_h, Y_h, the discretization (1.1) is consistent and stable (with constant L), then (1.1a) possesses, for h sufficiently small, a unique solution U_h. These solutions converge. Furthermore, for h small enough, $\|e_h\| \le L \|1_h\|$, so that if the consistency is of order p, then the convergence is also of order p.

Example A. Here the stability inequality (1.9) can be derived from a *discrete maximum principle* analogous to the maximum principle for (1.2). (The latter simply asserts that if $u''(x) \ge 0$, $0 \le x \le 1$, i.e. u is convex, then $u(x) \le 0$, $0 \le x \le 1$.) We show that if the vector A_hV_h has nonnegative entries, then $V_h = [V_1,\dots,V_{J-1}]^T$ is nonpositive. In fact, assume that the i-th entry in V_h is as large as any other entry, i.e. that i is such that $V_j - V_i \le 0$ for $1 \le j \le J-1$. If $1 < i < (J-1)$, then (a subscript denotes component) $0 \le [A_hV_h]_i = h^{-2}(V_{i-1}-V_i)+h^{-2}(V_{i+1}-V_i) \le 0$, so that $V_i = V_{i-1}$. By induction, $V_i = V_{i-1} = \dots = V_1$. Thus the first entry is always the largest. But then, $0 \le [A_hV_h]_1 = h^{-2}(V_2-V_1)-h^{-2}V_1$, showing that $V_1 \le 0$ and therefore $V_j \le 0$ for each j.

We are now in a position to prove the stability inequality (1.9). Let $V_h = [V_1,\dots,V_{J-1}]^T \in X_h$ and let M be the (maximum) norm of A_hV_h. Introduce the vector $W_h \in X_h$ with j-th entry $\frac{1}{2}h^2j^2$. One has $A_hW_h = [1,\dots,1]^T$ and $\|W_h\| \le \frac{1}{2}$. Then $Z_h = \pm V_h + MW_h$ are such that A_hZ_h is nonnegative. By the maximum principle, $\pm V_j \le \frac{1}{2}M$, so that (1.9) holds with $L = \frac{1}{2}$.

Note that the maximum principle shows that A_h^{-1} (which exists by stability) is nonpositive. Further material on matrices with nonpositive inverses and their importance in the discretization of elliptic problems can be seen, e.g., in Varga [47].

We conclude from Theorem 1.1 that (1.3) is uniquely solvable and that $\max_j |u(x_j) - U_j| = O(h^2)$, provided that f possesses two bounded derivatives.

Example B. The following lemma is needed.

Lemma 1.1 Let W, Z be normed spaces, k a positive number, N a positive integer. Let X denote the space of (N+1)-vectors $V = [V_0,\dots,V_N]^T$, $V_n \in W$, with the norm $\|V\|_X = \mathrm{Max}_n \|V_n\|_W$. Let Y denote the space of (N+1)-vectors $V = [V_0,\dots,V_N]^T$, $V_n \in Z$, with the norm

$$\|V\|_Y = \|V_0\|_Z + k \sum_{n=1}^{N} \|V_n\|_Z .$$

Let $B = (B_{mn})$, $m,n = 0,1,\dots,N$, be a matrix whose entries B_{mn} are bounded operators from Z into W. Then B defines a bounded operator from Y into X with norm

$$B = \max\{ \max_{0\le m\le N} \|B_{m0}\|,\ \max_{\substack{0\le m\le N\\ 1\le n\le N}} k^{-1}\|B_{mn}\| \}.$$

Proof When W and Z are both the real line, the proof is analogous to those in Section 1.1 of Isascson and Keller [20]. The extension to general normed spaces W, Z is trivial.

On applying the lemma with $W = Z = L^2_p$ to the inverse operator A_h^{-1} in (1.7), we see that $\|A_h^{-1}\| = \max\{\|C_h^n\|: 0 \le n \le N\}$ and therefore stability means

$$\sup_h \max_{0\le n\le N} \|C_h^n\| =: L < \infty, \tag{1.10}$$

a requirement which is often taken as the *definition* of stability in the discretization of initial value problems: Richtymyer and Morton [27], Ansorge [2]. (Note that there is no need to restrict h to be less than an appropriate h_0 since C_h^n is certainly bounded independently of h for $h > h_0$, $0 \le n \le N$.) The meaning of the powers C_h^n is obvious: they transform the starting datum U^0 corresponding to $t = 0$ into the elements U^n corresponding to $t = nk$.

Remark 1 It is useful in what follows to realize that the left-most column of A_h^{-1} contains already all the powers C_h^n, $0 \le n \le N$. Therefore, to bound $\|A_h^{-1}\|$, it is enough to bound $\|A_h^{-1}f_h\|$ for f_h of the form $f_h = [f^0,0,\dots,0]^T$, $\|f^0\| \le 1$. In other words, in the investigation of stability, the attention can be restricted to perturbations of the 0-th equation of the system (1.6), i.e. perturbations of the initial condition.

In order to see whether (1.10) holds, we resort to Fourier analysis (or von Neumann analysis, as it is often called in numerical circles). Each function $\phi \in L^2_p$ possesses a unique expansion

$$\phi(x) = \sum_{m=-\infty}^{\infty} a_m e^{2\pi imx}, \tag{1.11}$$

with

$$\Sigma|a_m|^2 = \|\phi\|^2 = \int_0^1 |\phi(x)|^2 dx < \infty.$$

Conversely, each complex sequence (a_m) with $\Sigma|a_m|^2 < \infty$ defines through (1.11) a function in L_p^2. In this way, one may think of the Fourier coefficients (a_m) as coordinates describing ϕ. On letting C_h operate on the function $\exp(2\pi imx)$, we obtain

$$C_h(e^{2\pi imx}) = c_h(m)\, e^{2\pi imx}. \qquad (1.12a)$$

with $c_h(m) = 1-r+r\exp(-2\pi imh)$, the so-called symbol or amplification factor of C_h. Therefore, if ϕ has Fourier coefficients (a_m) and $\psi = C_h\phi$ has Fourier coefficients (b_m), then the operation $\phi \to \psi$ is represented in Fourier space as the diagonal matrix transformation

$$\begin{bmatrix} \cdot \\ \cdot \\ b_{-1} \\ b_0 \\ b_1 \\ \cdot \end{bmatrix} = \begin{bmatrix} \cdot & & & & \\ & \cdot & & & \\ & & c_h(-1) & & \\ & & & c_h(0) & \\ & & & & c_h(1) \\ & & & & & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ a_{-1} \\ a_0 \\ a_1 \\ \cdot \end{bmatrix} \cdot \qquad (1.12b)$$

Thus

$$\|C_h^n\| = \sup_m |c_h(m)^n| = (\sup_m |c_h(m)|)^n. \qquad (1.13)$$

From these relations, it is easily derived that (1.6) is stable, with $L = 1$, if $0 < r \leqq 1$, and is not stable if $r > 1$. (More precisely, if $r > 1$, $\max_n \|C_h^n\|$ grows exponentially as $h \to 0$.) We conclude that, if u is smooth and $r \leqq 1$,

$$\max_{0 \leqq t_n \leqq T} \|u(\cdot,t_n) - U^n\|_{L_p^2} = O(h).$$

Note that, because of the seemingly artificial factor k^{-1} in (1.5b), the local and global errors are both O(h). When (1.5b) is written in undivided

form $U^{n+1} = C_h U^n$, the local truncation error is $O(h^2)$. This would not contradict $\|e_h\| \leq \|A_h^{-1}\|\ \|l_h\|$, because then $\|A_h^{-1}\|$ behaves like h^{-1}. In this paper difference schemes are always written in divided form, cf. Stetter [39], para 2.2.2.

Remark 2 So far, the concept of stability has been considered as a means for proving convergence. However, the idea of stability is important in its own right: in practice, because of round-off errors, inaccuracies in the data etc, the computed $\tilde{U}_h$ does not satisfy (1.1a) but rather

$$A_h \tilde{U}_h = f_h + \delta_h \qquad (1.14)$$

with δ_h a 'small' perturbation. The stability of (1.1) implies, on subtracting (1.1a), (1.14), $\|U_h - \tilde{U}_h\| \leq L\|\delta_h\|$, i.e. that $\tilde{U}_h$ is 'close' to U_h even if h is small (note that, in initial value problems, a smaller value of h means that more steps are required to integrate up to t = T and therefore there is more scope for the growth of perturbations).

Discretizations for which $\|A_h^{-1}\|$ increases exponentially as $h \to 0$ (such as that in Example B for $r > 1$) are universally considered as deprived of practical applicability. For them, $\|U_h - \tilde{U}_h\|$ may increase exponentially as $h \to 0$, even if $\|\delta_h\| = O(h^q)$, $q > 0$. On the other hand, unstable discretizations where the growth of $\|A_h^{-1}\|$ is only $O(h^{-q})$ can be of practical significance, as they can cope with perturbations $\|\delta_h\| = o(h^q)$. Such discretizations are sometimes called *weakly stable* (cf. Richtmyer and Morton [27], p. 95 and Sanz-Serna and Spijker [31]) and are often found when dealing with spectral methods [12]. Also, a number of standard finite-difference methods are weakly stable but not stable in the L^p norm, $p \neq 2$, Geveci [11]. See also Section 5.8.

Remark 3 There is another concept of stability that plays an important role in the numerical treatment of *initial-value problems* in ODEs or PDEs. This refers to the behaviour, for a given, fixed value of h, of the powers C_h^n, n increasing unboundedly. For clarity, the notion of stability in Definition 1.3 is often called *zero stability*, as it relates to an $h \to 0$ behaviour. We emphasize that the alternative, fixed-h notion of stability only applies to initial-value problems. In ODEs, Lambert [22] and Stetter [39] use respectively the terms weak and strong stability to refer to fixed-h as distinct from zero stability. In PDEs, not so much care has been exercised in distinguishing between the two notions. The terms contractivity, A-stability, B-stability etc, used in ODEs, are all fixed-h concepts (Lambert [22], Dekker and Verwer [8]). In this paper we are only concerned with zero stability. See Sanz-Serna [29] and Verwer and Sanz-Serna [48] for a study of the relationship between the two concepts of stability.

2. THE FIRST PARADIGM. REFINEMENTS

2.1 The necessity of stability. L-convergence

In Theorem 1.1, stability and consistency appear as sufficient means for proving convergence. The question arises as to whether these requirements are also necessary. From a mathematical point of view, it is clear that stability is by no means necessary for convergence, because the notion of stability depends on the norms in X_h, Y_h, whereas the convergence does not depend at all on the norm in Y_h. Thus a convergent scheme can be made unstable by a suitable change in the norm in Y_h (cf. Stetter [39] p. 14). The question remains, however, as to whether, for *numerically meaningful, reasonable* choices of norm in Y_h, convergent discretizations are stable. The answer is still negative, as we show next.

Example B. We fix $r > 1$, so that (1.6) is (exponentially) unstable and assume that the initial datum $\eta(x)$ is given by $\exp(2\pi imx)$, with m a fixed integer, leading to the solution $u(x,t) = \exp(-2\pi imt) \times \exp(2\pi imx)$. On considering (1.12), we can write

$$u(x,t_n) - U^n(x)$$
$$= \{[\exp(-2\pi imrh)]^n - [1-r+r \exp(-2\pi imh)]^n\} \exp(2\pi imx) \qquad (2.1)$$

so that, in order to bound the global error, one must bound the difference $[\exp(r\xi)]^n - [1-r+r \exp(\xi)]^n$ with $\xi = -2\pi imh$. Substitution of the exponential terms by their Taylor series and use of the binomial expansion lead, after some cancellations, to the conclusion that, for $0 \leq nk \leq T$, that difference possesses a bound $B(m)h$, with $B(m)$ independent of n and h. Therefore the discretization is convergent of the first order. (More generally, one-step consistent discretizations of periodic initial-value constant coefficient problems always converge, regardless of stability, when the initial datum

contains only a finite number of wave numbers m, cf. Thomee [45], Theorem 3.1.) The convergence of this example is, nevertheless, of no practical value, because of the exponential instability noted in the previous section, Remark 2.

In order to rule out examples like the previous one, where the convergence is merely an academic matter, it is often demanded that the convergence of U_h should persist under perturbations of the right-hand side. There are several definitions of such a stable convergence which go back, at least, to Dahlquist's thesis [7]. Among them, we only consider that of L-stability (L from Lax) stemming from the theory of numerical initial value problems in PDEs (see Ansorge [3], Palencia and Sanz-Serna [24] Sanz-Serna and Palencia [30]; the use of the term L-convergence in [3] is not always equivalent to ours).

Definition 2.1 Assume that elements $u_h \in X_h$ and norms in X_h and Y_h have been chosen. The discretization (1.1) is said to be L-convergent if, for any family $(\delta_h)_{h\in H}$, $\delta_h \in Y_h$ with $\lim \|\delta_h\| = 0$, a constant h_o exists such that the problems

$$A_h\tilde{U}_h = f_h + \delta_h, \; h \leq h_o, \qquad (2.2)$$

possess a unique solution and $\lim \|u_h-\tilde{U}_h\| = 0$.

The following characterization holds.

Theorem 2.1 The discretization (1.1) is L-convergent if and only if it is convergent and stable (for fixed choices of u_h and norms in X_h, Y_h).

Proof If (1.1) is stable, then, for h small, A_h is invertible and (2.2) uniquely solvable. Furthermore, from stability and convergence:

$$\|u_h-\tilde{U}_h\| \leq \|u_h-U_h\| + \|U_h-\tilde{U}_h\| \leq \|u_h-U_h\| + L\,\|A_hU_h-A_h\tilde{U}_h\|$$

$$\|u_h-U_h\| + L\,\|\delta_h\| \to 0.$$

Assume conversely that (1.1) is L-convergent and hence convergent. If the stability bound (1.9) does not hold, then there exist $h_j \to 0$, $\phi_{h_j} \in X_{h_j}$, $\psi_{h_j} \in Y_{h_j}$ such that (the subscript j is omitted) $\|\psi_h\| = 1$, $A_h\phi_h = \psi_h$,

$\lambda_h := \|\phi_h\| \to \infty$. If we set $\delta_h = \lambda_h^{-\frac{1}{2}}$, then $\|\delta_h\| \to 0$ and $\|U_j - \tilde{U}_j\| = \|\lambda_h^{-\frac{1}{2}}\phi_h\| \to \infty$, which contradicts the assumption of L-convergence.

On combining this result with the basic Theorem 1.1, we arrive at the following *equivalence* result.

Theorem 2.2 Assume that (1.1) is consistent. Then it is L-convergent if and only if it is stable (for given choices of u_h and norms).

The equivalence between L-convergence and stability is achievable because both concepts involve the norms in X_h *and* Y_h. The theorems above are well known in the literature (see, e.g., Stummel [43], Theorem 6, Section 1.2), but our terminology is different.

Remark For invertible A_h, it is clear that L-convergence holds if $\lim \|u_h - \tilde{U}_h\| = 0$ whenever $\|\delta_h\| \to 0$ and δ_h belongs to S_h, a subspace of Y_h with the property

$$\sup \{\|A_h^{-1} g_h\| : g_h \in S_h, \|g_h\| \le 1\} = \sup\{\|A_h^{-1} g_h\| : g_h \in Y_h, \|g_h\| \le 1\}.$$

For instance, Remark 1 in Section 1.4 shows that, in Example B, L-convergence is equivalent to convergence under null perturbations of the initial datum. We conclude that (1.6) converges (for a fixed η) whenever η in (1.5a) is replaced by approximations $\tilde{\eta}_h$, $\|\tilde{\eta}_h - \eta\| \to 0$, if and only if (1.6) is stable (i.e., r is not larger than 1).

2.2 The necessity of consistency. Uniform boundedness. L-consistency

We now consider the question as to whether consistency is necessary for convergence. Again we note that a change in the norm in Y_h alters the consistency of (1.1) but not its convergence. And again Example B provides a counterexample, as follows.

Example B. Assume now that $r < 1$ and that the initial datum is given by the step-function $\eta(x) = 0$, $0 \le x \le 1/4$ or $3/4 \le x \le 1$, $\eta(x) = 1$, $1/4 < x < 3/4$. Then $\|A_h u_h - f_h\|$ is easily computable in explicit form (recall that $u(x,t) = \eta(x-t)$) and seen to behave like $h^{-\frac{1}{2}}$ as $h \to 0$, precluding consistency. However, we shall prove in Chapter 3 that the discretization is convergent with order $p = 1/4$.

Since, as noted before, $A_h e_h = l_h$, it is clear that the local error l_h can be bounded in terms of the global error e_h uniformly in h if the operators A_h

are uniformly bounded. More precisely:

Definition 2.2 Assume that norms in X_h and Y_h have been chosen. The discertization (1.1) is said to be uniformly bounded if positive constants h_o, M exist, such that, for each $h \le h_o$, $V_h \in X_h$,

$$\|A_h V_h\| \le M \|V_h\| . \tag{2.3}$$

The constant M is called the uniform bound of (1.1).

Theorem 2.3 Assume that elements u_h in X_h and norms in X_h and Y_h have been chosen. If (1.1) is convergent (resp. convergent of order p) and uniformly bounded, then (1.1) is consistent (resp. consistent of order p). Furthermore, for h sufficiently small, $\|l_h\| \le M \|e_h\|$, where M is the uniform bound of (1.1).

We emphasize that Definition 2.2 and Theorem 2.3 are closely related to the definition of stability and the basic Theorem 1.1 respectively. In fact, if we associate to (1.1) (when A_h is invertible and norms in X_h and Y_h have been chosen) the discrete problems

$$A_h^{-1} F_h = u_h, \tag{2.4}$$

it is clear that the stability of (2.4) coincides with the uniform boundedness of (1.1), while the convergence and consistency of (1.1) with respect to the 'theoretical elements' u_h are respectively identical to the consistency and convergence of (2.4) with respect to f_h. We take further this symmetry by defining the concept of L-consistency as follows:

Definition 2.3 Assume that elements u_h in X_h and norms in X_h, Y_h have been chosen. The discretization (1.1) is said to be L-consistent if, for each family $\varepsilon_h \in X_h$ with $\lim_h \|\varepsilon_h\| = 0$, $\lim \|A_h(u_h + \varepsilon_h) - f_h\| = 0$.

The next results are 'symmetric' to Theorems 2.1, 2.2.

Theorem 2.4 The discretization (1.1) is L-consistent if and only if it is consistent and uniformly bounded (for given choices of u_h and norms).

Theorem 2.5 Assume that (1.1) is convergent; then it is L-consistent if and only if it is uniformly bounded (for given choices of u_h and norms).

We can also combine Theorems 1.1 and 2.3 to yield:

Theorem 2.6 Assume that elements u_h and norms in X_h and Y_h have been chosen. If (1.1) is stable and uniformly bounded, then it is convergent if and only if it is consistent. Furthermore, if L, M denote the stability constant and uniform bound, then, for h small enough,

$$M^{-1} \|l_h\| \leq \|e_h\| \leq L \|l_h\| \tag{2.5}$$

so that the orders of convergence and consistency coincide.

The situation in Theorem 2.6 is really convenient: $O(h^p)$ estimates of the local truncation error, which are usually easily derived by means of Taylor expansions, can be transferred to the global error and this transference is optimal, in the sense that no estimates $o(h^p)$ of the global error exist. Stummel [44] uses the term *bistable* as an abbreviation for stable and uniformly bounded. Unfortunately one often finds in practice discretizations which are *not* bistable, at least for the choices of norms that first come to mind.

Examples A and B. Neither (1.3) nor (1.6) is uniformly bounded, because of the factors h^{-2}, k^{-1} that feature in A_h. According to Theorems 2.3, 2.6, it is then possible to have convergence without consistency, and in fact we have already shown that the step-function initial datum provided an example in that direction. Also note that, after Theorem 2.4, L-consistency does not hold.

2.3 An example: central differences on a nonuniform grid

In order to summarize the main ideas presented so far, we now consider an illuminating example communicated to us by R.D. Grigorieff. Further related material can be seen in his papers [17], [18].

The problem

$$u(0) = \alpha, \quad u'(0) = \beta, \quad u''(x) = f(x), \; 0 \leq x \leq 1,$$

where f has two bounded derivatives, is discretized by central differences on a nonuniform grid: $x_0 = 0$, $x_{j+1} = x_j + \Delta_{j+1}$, $\Delta_j > 0$, $j = 0,1,\dots,J-1$, $\max_j \Delta_j = h$. More precisely, we introduce the divided difference operator D given by $DV_j = \Delta_{j+1}^{-1}(V_{j+1}-V_j)$, $j = 0,1,\dots,J-1$ and consider the system:

$$U_0 = \alpha,$$

$$DU_0 = \beta + (\Delta_1/2)\, f(0), \tag{2.6}$$

$$(2/(\Delta_j+\Delta_{j+1}))\,(DU_j - DU_{j-1}) = f(x_j), \; j = 1,\dots,J-1.$$

Here X_h, Y_h are spaces of real (J+1)-vectors and we choose u_h to be the vector of components $u(x_j)$. A naive approach to the analysis of (2.6) begins by expanding, for $j = 1,\dots,J-1$,

$$l_{j+1} = (2/(\Delta_j+\Delta_{j+1}))\,(Du_j - Du_{j-1}) - f(x_j).$$

This leads to

$$l_{j+1} = ((\Delta_{j+1}-\Delta_j)/3)\, u'''(x_j) + R_{j+1}, \tag{2.7}$$

where the remainder R_{j+1} is $O(h^2)$, uniformly in j. Therefore $l_{j+1} = O(h)$, unless the grid happens to be uniform or u is a second-degree polynomial. One is tempted to conclude that the order of convergence in, say, the maximum norm cannot be larger than one. However, the situation is quite different, as we now show.

(i) We first work with the maximum norm in both X_h and Y_h, which we denote $\|\cdot\|_0$. From (2.7), $\|l_h\|_0 = O(h)$, i.e. the order of consistency is only 1. We shall prove later that the discretization is stable. Therefore, Theorem 1.1 yields an estimate $\|u_h - U_h\|_0 = O(h)$. The norm $\|A_h\|_0$ computed according to the usual recipe (Isaacson and Keller [20], p.9) behaves like h^{-2}. Thus the discretization is not uniformly bounded, Theorem 2.6 does not apply and we are not sure as to whether the global error is of order not higher than 1.

(ii) We now analyze the *same* discrete solutions U_h, when the norms in X_h and Y_h are respectively defined as

$$\|V_h\|_1 = |V_0| + \max_{0 \leq j \leq J-1} |DV_j|, \tag{2.8}$$

$$\|F_h\|_{-1} = |F_0| + \max_{0 \leq j \leq J-1} \left|F_1 + \sum_{k=1}^{j} \tfrac{1}{2}(\Delta_k+\Delta_{k+1})F_{k+1}\right|. \tag{2.9}$$

The identity

$$V_j = V_0 + \sum_{k=0}^{j-1} \Delta_{k+1} DV_k$$

shows that, for each V_h in X_h, $\|V_h\|_0 \le \|V_h\|_1$. Therefore convergence with regard to $\|\cdot\|_1$ implies convergence with regard to $\|\cdot\|_0$ with at least the same order. Clearly $\|F_h\|_{-1} \le 3\,\|F_h\|_0$. Also note that stability in this new situation (which we prove next) certainly implies stability with respect to the old maximum norm, because $\|V_h\|_0 \le \|V_h\|_1 \le L\,\|A_hV_h\|_{-1} \le 3L\,\|V_h\|_0$.

As a result of the somewhat sophisticated choice of norms (2.8), (2.9), for them, $\|A_h\| = \|A_h^{-1}\| = 1$ and the discretization is now stable and uniformly bounded. (More precisely $\|l_h\|_{-1} = \|e_h\|_1$.) In order to see this, take V_h in X_h and set $F_h = A_hV_h$. This system of equations can be written in the form

$$V_0 = F_0,$$

$$DV_0 = F_1,$$

$$DV_j - DV_{j-1} = \tfrac{1}{2}(\Delta_{j+1} + \Delta_j)F_{j+1},\ j = 1,\dots,J-1.$$

On adding, we find, for $j = 0,1,\dots,J-1$,

$$DV_j = \sum_{k=1}^{j} \tfrac{1}{2}(\Delta_k+\Delta_{k+1})F_{k+1} + DV_0 = F_1 + \sum_{k=1}^{j} \tfrac{1}{2}(\Delta_k+\Delta_{k+1})F_{k+1},$$

so that, according to (2.8), (2.9), $\|V_h\|_1 = \|F_h\|_{-1}$. Now that we know we are dealing with a bistable discretization, we have guaranteed that a study of the local error provides full information on the global error. For the former, we easily find $l_0 = 0, |l_1| = O(h^2)$. Furthermore, from (2.7):

$$\left|\sum_{k=1}^{j} \tfrac{1}{2}(\Delta_k+\Delta_{k+1})l_{k+1}\right| = \left|\sum_{k=1}^{j} ((\Delta_{k+1}^2-\Delta_k^2)/6)u'''(x_k)\right| + O(h^2)$$

$$= (1/6)\left|\sum_{k=1}^{j-1} \Delta_{k+1}^2 (u'''(x_k) - u'''(x_{k+1})) + \Delta_{j+1}^2 u'''(x_j)\right.$$

$$\left. - \Delta_1^2 u'''(x_1)\right| + O(h^2) = O(h^2),$$

where we have summed by parts (Richtmyer and Morton [27] p. 136) and taken into account that $|u'''(x_k) - u'''(x_{k+1})| = O(h)$. Substitution of these

estimates in (2.9) leads to $\|l_h\|_{-1} = O(h^2)$, i.e. second order of consistency and hence of convergence $\|e_h\|_1 = O(h^2)$. This implies that $|u(x_j)-U_j|$ is $O(h^2)$, uniformly in j, $j = 0,\dots,J$. Note that, in view of (2.8), we have also proved that the divided differences DU_j are second-order approximations to the theoretical Du_j, uniformly in j.

The norm (2.9), which is highly useful in the derivation of sharp bounds of the local truncation error, was first introduced by Spijker in his thesis [35] (see also [36], [38]). Multivalue methods for the integration if initial value problems may also exhibit orders of convergence higher than their 'naive' order of consistency, see Skeel [33], Skeel and Jackson [34]. As done in this seciton, these authors renorm Y_h in order to avoid the discrepancy.

2.4 Modified equations

Within the framework of the first paradigm, the original problem whose solution u is being approximated plays no role in the analysis. This fact, coupled with the freedom in the choice of u_h, makes it possible to compare U_h with elements u_h that are not necessarily 'restrictions' of u. An example is given by the method of *modified equations* (Griffiths and Sanz-Serna [13]) which we now briefly illustrate in the context of Example A.

Example A. We still use the maximum norm in X_h and Y_h (for which stability was proved in Section 1.4), but now take $u_h = [v^h(x_1),\dots,v^h(x_{J-1})]^T$, where v^h is the solution of the *modified problem* (here f is supposed to have four bounded derivatives)

$$v^h(0) = v^h(1) = 0, \quad (v^h)'' = f + (h^2/12)f''. \tag{2.10}$$

A Taylor expansion shows that, with this choice of u_h, $\|l_h\| = O(h^4)$, which, according to Theorem 1.1, leads to

$$\max_j |v^h(x_j) - U_j| = O(h^4).$$

Thus the solution v^h of (2.10) is 'very close' to the numerical solution U_h and (2.10) can be used to predict the behaviour of U_h. For instance, if $f'' > 0$, then $f + (h^2/12)f'' > f$, so that $v^h < u$ and we expect that $U_j < u(x_j)$.

The paper by Griffiths and Sanz-Serna includes a long list of references to the derivation and practical use of modified equations. The idea of comparing the numerical solution U_h with a function close to, but different from, the theoretical solution u goes back to Strang [40] and is highly useful in nonlinear situations (cf. Spijker [37], Sanz-Serna [28]).

3. THE SECOND PARADIGM

3.1 Well-posed problems

In the framework of the second paradigm, the original problem whose solution u is being approximated plays a crucial role. We assume that this original problem takes the form

$$Au = f, \tag{3.1a}$$

where f represents the data, u is the sought solution and A a linear operator. More precisely, we assume that f belongs to a *normed* space Y, u is sought in a *normed* space X and A maps linearly its domain $D(A) \subset X$ onto its range $R(A) \subset Y$. The operator A may be bounded or unbounded, but we demand that

$$\ker(A) = \{0\}, \tag{3.1b}$$

so that the solution u, if it exists, is unique. Since D(A), R(A) may be smaller than X, Y respectively, there is no loss of generality in assuming that X and Y are complete (i.e. Banach) spaces. (If they were not complete, we would replace them by complete spaces $\tilde{X} \supset X$, $\tilde{Y} \supset Y$.)

We denote by A^{-1} the inverse operator mapping R(A) onto D(A). When f is in the range R(A), then $u = A^{-1}f$ is the unique solution of (3.1). We say that u is a *genuine* solution of (3.1). A very natural requirement that A^{-1} should satisfy is that of boundedness - one should be able to conclude that small changes in f lead to small changes in u. When A^{-1} is bounded, it can be uniquely extended to a bounded operator E with domain the closure $\overline{R(A)}$. Then, for $f \in \overline{R(A)} \setminus R(A)$, we say that Ef is a *generalized* solution of (3.1). This simply means that no $u \in D(A)$ exists for which Au = f, but elements $f_n \in R(A)$, $u_n \in D(A)$ exist such that $u_n \to Ef$, $f_n \to f$ and u_n is the genuine solution corresponding to the datum f_n. If R(A) is dense (i.e. $\overline{R(A)} = Y$), then such generalized solutions exist for all data in Y and we say that the problem is *well-posed*.

The previous discussion can be summarized as follows:

Definition 3.1 An original problem Au = f is given by a datum $f \in Y$ and a linear operator $A:D(A) \subset X \to R(A) \subset Y$, with $\ker(A) = \{0\}$ and X,Y Banach spaces. If $f \in R(A)$, then $A^{-1}f$ is the genuine solution of the problem. The problem is said to be well-posed if A^{-1} is bounded and R(A) dense. In this case and denoting by E the extension of A^{-1} to Y, the element $E^{-1}f$, $f \notin R(A)$ is a generalized solution cf the original problem.

Example B. Here we may take as X the Banach space of continuous mappings $u:t \to u(\cdot,t) \in L_p^2$ with the maximum norm $\|u\| = \max_{0\le t\le T} \|u(\cdot,t)\|_{L_p^2} = \max_t[\int_0^1 |u(x,t)|^2 dx]^{\frac{1}{2}}$, and Y the Banach space L_p^2. The operator A then maps each bivariate function $u(\cdot,\cdot)$ into the initial function $u(\cdot,0)$. The domain of A can be chosen to consist of the functions u of class C^1 in $-\infty < x < \infty$, $0 \le t \le T$ which satisfy (1.4a), (1.4c). (Generally speaking, in differential equation problems, D(A) always consists of functions which are smooth enough to allow the differentiations implied in the problem.) Clearly if $\eta \in L_p^2$ and is of class C^1 then $u(x,t) = \eta(x-t) \in D(A)$ and $Au = \eta$. Thus R(A) is the space of 1-periodic, C^1-functions and A^{-1} is given by $(A^{-1}\eta)(x,t) = \eta(x-t)$. It is obvious that A^{-1} is bounded and therefore possesses a bounded extension E defined everywhere in $Y = L_p^2$. This extension is still given by the formula $(E\eta)(x,t) = \eta(x-t)$. Therefore, when η is not C^1 we still regard $u(x,t) = \eta(x-t)$ as a solution to (1.4) in spite of the fact that the derivatives u_t, u_x may not exist.

3.2 Discretization of an original problem

The second paradigm relates an original problem to a family of discrete problems by means of *restriction operators*. If X is a Banach space, H is a set of positive numbers with inf H = 0 and $(X_h, \|\cdot\|_h)_{h\in H}$ is a family of normed spaces, we say that the operators $r_h:X \to X_h$ are (a family of) restriction operators if: (i) each r_h is linear, (ii) for each x in X

$$x = \lim_h \|r_h x\|_h . \tag{3.2}$$

We note that, if each r_h is a bounded operator, then the family (r_h) is in fact equicontinuous (i.e. $\sup_h \|r_h\|_h < \infty$). This follows from the generalized Banach-Steinhaus theorem, see e.g. Palencia and Sanz-Serna [25], Lemma.

Let us assume that we are given an original problem (3.1) with solution u (possibly generalized), a set of indices H, normed spaces X_h and Y_h, restriction operators $r_h:X \to X_h$, $s_h:Y \to Y_h$ and linear operators $A_h:X_h \to Y_h$ fulfilling (1.1b). On setting

$$A_h U_h = s_h f \tag{3.3}$$

we obtain a family of discrete problems like those considered in the previous chapters. If we further set $u_h = r_h u$, we possess all the necessary elements to discuss the concepts of convergence, stability, consistency, L-convergence etc. as defined before. We emphasize that those concepts were defined without reference to the original problem, i.e. within the first paradigm. However, we shall show in this chapter that the presence of the original problem and the restriction operators is very helpful in investigating stability and convergence.

Lest we miss the obvious, let us observe that the replacement of the fixed, given datum f we have been considering so far by another datum $g \in Y$ leads to a new set of discrete problems

$$A_h V_h = s_h g, \quad V_h \sim v_h = r_h v, \tag{3.4}$$

where v is the solution corresponding to g, assumed to exist. This new discretization is stable (resp. uniformly bounded) if and only if (3.3) is stable (resp. uniformly bounded). On the other hand, the concepts of consistency and convergence clearly depend on the right-hand side, i.e., (3.3) and (3.4) need not be simultaneously consistent or convergent.

Example B. The discretization (1.6) which was previously analyzed within the first paradigm may now be studied within the second. In order to do so it is enough to consider the original problem discussed in the previous section, together with the restriction operators

$$r_h v = [v(\cdot,t_0),\dots,v(\cdot,t_N)]^T, \quad v(\cdot,\cdot) \in C([0,T],L_p^2)$$

$$s_h \zeta = [\zeta,0,\dots,0]^T, \qquad \zeta \in Y = L_p^2.$$

Clearly $\|r_h\| = \|s_h\| = 1$ for each $h > 0$.

3.3 Stability implies well-posedness

The next theorem provides a first example of the potentialities of the second paradigm.

Theorem 3.1 Assume that the discretization (3.3) is stable (with stability constant L) and consistent for each f in the range of A. Then A^{-1} is bounded and $\|A^{-1}\| \le L$.

Proof From (3.2),

$$\|A^{-1}f\| = \lim_h \|r_h A^{-1} f\| = \lim_h \|u_h\|.$$

Convergence implies $\lim_h \|u_h - U_h\| = 0$ and therefore

$$\|A^{-1}f\| = \lim_h \|U_h\| = \lim_h \|A^{-1} s_h f\| \le L \lim \|s_h f\| = L\,\|f\|.$$

The 'symmetric' of this theorem is also useful:

Theorem 3.2 Assume that (3.3) is uniformly bounded (with uniform bound M) and consistent for each f in the range of A. Then A is bounded and $\|A\| \le M$.

There are two ways in which these results may be employed: (i) one may construct stable (resp. uniformly bounded), consistent discretizations as means for proving well-posedness (resp. boundedness) of a differential equation problem; (ii) proofs of stability (resp. uniform boundedness) must not be attempted for discrete norms which are counterparts of norms rendering the original problem not well-posed (resp. unbounded). An illustration of the last point is provided by the nonuniform-grid example in Section 2.3. There the data consist of the real numbers α,β and the function f. The maximum of $f = u''$ cannot be bounded in terms of the maximum of u (small functions can have very large derivatives). Therefore the operator A is not bounded in the maximum norm and this entails the lack of uniform boundedness of the discretization which caused trouble in Section 2.3. On the other hand, measurement of the data in the norm $|\alpha| + \max_x |\beta + \int_0^x f(s)ds|$, combined with use of the norm $|u(0)| + \max_x |u'(x)|$, ensures that the values of the norms of the datum and its corresponding solution are indentical, since $u'(x) = u'(0) + \int_0^x u''(s)ds$. Comparison of these norms with (2.8), (2.9) throws light on the bistability obtained when using the discrete norms $\|\cdot\|_1$, $\|\cdot\|_{-1}$.

3.4 Extending the convergence. Order of convergence for nonsmooth data

When the original problem is well-posed, it is possible to prove convergence for right-hand sides f for which consistency has not been checked or even does not hold (cf. Sections 2.2 - 2.3). Namely:

Theorem 3.3 Assume that the original problem (3.1) is well-posed and that the discretizations (3.3) are stable and consistent for each f in a set Y_0 dense in Y. Then (3.3) is convergent for each f in Y (and hence L-convergent for each f in Y).

Proof Given f and Y and $\varepsilon > 0$, there exist f_M in Y_0 with $\|f_M - f\| < \varepsilon$ $\|Ef_M - Ef\| < \varepsilon$. Then

$$\|r_h Ef - A_h^{-1} s_h f\| \leq \|r_h E(f-f_M)\| + \|r_h Ef_M - A_h^{-1} s_h f_M\| + \|A_h^{-1}\| \, \|s_h(f_M - f)\| . \tag{3.5}$$

When $h \to 0$, the first and third terms in the right-hand side become less than ε in view of (3.2), while the second tends to 0 according to Theorem 1.1. Therefore the discretization converges for f and also L-converges because of Theorem 2.1.

In the important case of bounded r_h, s_h one has, as noted above, $\|r_h\|$, $\|s_h\| \leq K$, with K independent of h, and (3.5) leads to

$$\|r_h Ef - A_h^{-1} s_h f\| \leq (K\|E\| + LK) \|f - f_M\| + \|r_h Ef_M - A_h^{-1} s_h f_M\| . \tag{3.6}$$

This inequality can be used to study the *order* of convergence for (i.e. the size of the left-hand side as a function of h) provided that we possess estimates of the global error for f_M (i.e., the size of $r_h Ef_M - A_h^{-1} s_h f_M$) and the degree of approximability of f by elements of Y_0 (i.e., the size of $\|f - f_M\|$). An example is now given.

Example B. Theorem 3.3 implies that, for $r \leq 1$ (stable case), the discretization (1.6) is convergent for every initial datum in L_p^2, even for those leading to generalized solutions. We investigate the order of convergence corresponding to the step initial datum of Section 2.2. (Recall that the discretization was shown there to be inconsistent.) As in Section 1.4, we work in Fourier space and write $\eta(x) = \Sigma_m b_m \exp(2\pi imx)$, $-\infty < m < \infty$. The Fourier coefficients are readily computed and seen to behave $|b_m| = O(|m|^{-1})$. Incidentally, we point out that, in order to derive estimates of the Fourier coefficients, an explicit knowledge of them is not necessary: it suffices to possess information on the differentiability of η (see, e.g., Richtmyer and Morton [27], p. 22). In the context of Theorem 3.3, we choose Y_0 equal to

the space of trigonometric polynomials of arbitrary degree M, $\sum_{m=-M}^{M} a_m \exp(2\pi imx)$. If $\eta_M = \sum_{m=-M}^{M} b_m \exp(2\pi imx)$, then $\|\eta_M - \eta\|^2 = \sum_{|m|>M} |b_m|^2 = O(1/M)$. This settles the degree of approximability of η by elements in Y_0. Turning now to the global error for the datum η_M, we denote by u_M^n, U_M^n respectively the theoretical and numerical solution at time $t = t_n$. On proceeding as in (2.1) we find

$$(\|u_M^n - U_M^n\|_{L^2})^2 = \sum_{m=-M}^{M} |b_m|^2 \, |a(mh)^n - b(mh)^n|^2 , \tag{3.7}$$

with $a(mh) = \exp(-2\pi imrh)$, $b(mh) = 1 - r + r\exp(-2\pi imh)$. The stability condition $r \leq 1$ leads to $|b| \leq 1$ (see Section 1.4) and therefore $|a^n - b^n| = |a-b| \, |a^{n-1} + ba^{n-2} + \ldots| \leq n|a-b|$. It is easy to show that $|a(mh) - b(mh)| \leq Dh^2m^2$ with D independent of m and h. On taking this estimate into (3.7), we find

$$\|u_M^n - U_M^n\|^2 \leq D \sum_{m=-M}^{M} m^{-2} n^2 h^4 m^4 \leq D\, 2M\, (M^2 h^4 n^2),$$

with D independent of h, M, n. Therefore, when $0 \leq nk \leq T$, $k = rh$,

$$\max_n \|u_M^n - U_M^n\| \leq BhM^{3/2},$$

with B independent of h and M. This shows that the order of convergence is 1 for each η_M, but that for fixed h the error is increased with M. On taking these estimates into (3.6), we obtain an $O(M^{-\frac{1}{2}} + hM^{3/2})$ bound for the global error of the step function, where M is arbitrary. Setting $M = [h^{-\frac{1}{2}}]$ minimizes the bound rendering it $O(h^{1/4})$ and we conclude that the sought order of convergence is $\frac{1}{4}$.

A more systematic approach to the technique above, together with historical references, can be seen in Ansorge [2], Section 4.5. The order of convergence for 'nonsmooth' data can also be investigated by means of interpolation theory - see Thomee [45] p. 186 and a fuller account in Brenner, Thomee and Wahlbin [4].

The 'symmetric' of Theorem 3.3 is as follows.

Theorem 3.4 Assume: (i) the operator A in the original problem (3.1) is defined everywhere (i.e. D(A) = X) and bounded, (ii) the discretizations (3.3) are uniformly bounded and consistent for each f in a set Y_0 such that the corresponding solutions u are dense in X. Then (3.3) is consistent for each f in the range of A (and hence L-consistent for each f in the range of A).

3.5 A general Lax equivalence theorem

The next result is due to Sanz-Serna and Palencia [30].

Theorem 3.5 Assume that (i) the original problem (3.1) is well-posed. (ii) the discretization (3.3) converges for each datum f in Y. (iii) the operators A_h are invertible and $A_h s_h^{-1}$ are bounded. (iv) the following condition holds:

(P) There exists a constant L such that, for each h in H and each g_h in Y_h with $\|g_h\| \leq 1$, there exists an element f in Y such that $\|f\| \leq L$ and $s_h f = g$.

Then (3.3) is stable.

Proof Let $f \in Y$. The norms $\|r_h Ef\|$ are bounded as $h \to 0$, since (3.2) applies. From the convergence assumption, $\|A_h^{-1} s_h f\|$ must also be bounded for $h < h_o$. The generalized Banach-Steinhaus Lemma (Palencia and Sanz-Serna [25]) shows that there exists a constant K such that, for $h < h_o$, $\|A_h^{-1} s_h\| \leq K$. If $g_h \in Y_h$, $\|g_h\| \leq 1$, $\|A_h^{-1} g_h\| = \|A_h^{-1} s_h f\| \leq KL$, whence $\|A_h^{-1}\| \leq KL$.

It is clear from the proof that (P) can be relaxed to read:

(P') There exist a constant L and subspaces S_h of Y_h such that $\sup\{\|A_h^{-1} g_h\| : g_h \in S_h, \|g_h\| \leq 1\} = \sup\{\|A_h^{-1} g_h\| : g_h \in Y_h, \|g_h\| \leq 1\}$ and to each g_h in S_h with $\|g_h\| \leq 1$ there corresponds an element f in Y with $\|f\| \leq L$, $s_h f = g$.

It is useful to compare the proof of the implication 'convergence ⇒ stab- ility' given here with that in Theorem 2.1. There we argued that convergence could not imply stability, since the latter could be lost by changing the norms in Y_h. Accordingly, we had to resort to the strengthened concept of L-convergence and then we were able to supply an elementary proof. Here con- vergence is assumed for the *family* of discrete problems obtained when f ranges in Y and we have employed a deep result of functional analysis. Now the possibility of altering arbitrarily the norm in Y_h is not open to us, for (3.2) and the condition (P) must hold. A further discussion of this point, together with the proof of the fact that the hypotheses of Theorem 3.5 cannot be essen- tially weakened, can be seen in [30].

Example B. (cf. Section 2.1, Remark). Here the condition (P') is veri- fied with $S_h = s_h Y = \{[\eta,0,0,\ldots,0]^T : \eta \in Y\}$ (recall Remark 1, Section 1.4). Therefore, Theorem 3.5 shows that if (1.6) converges *for each* η in L_p^2, then (1.6) shows that (1.6) is stable, i.e. $r \leq 1$. This assertion is precisely the content of the classical Lax equivalence theorem [23], as applied to this concrete situation. Note that in Section 2.1 we proved that for $r > 1$ the scheme is unstable and yet converges whenever the initial datum is a trigonometric polynomial. These polynomials are *dense* in L_p^2 (see [30] for further discussion).

3.6 Further results on restriction operators. Discrete convergence

Most of the material in this chapter would still be valid if the assumption that the restriction operators r_h are linear were relaxed and became the following asymptotic linearity requirement: for each u, v in X and scalar α, β, $\|\alpha r_h u + \beta r_h v - r_h(\alpha u + \beta v)\| \to 0$. Of course an analogous consideration applies to s_h. This asymptotic linearity was first introduced by Stummel [43] and is useful in the study of perturbations of the domain in partial differential equations and in other situations (see also Vainikko [46]).

Vainikko [46] says that two families of restrictions $r_h : X \to X_h$, $r_h' : X \to X_h$ are *equivalent* if, for each u in X, $\|r_h u - r_h' u\| \to 0$. It is clear that the convergence or otherwise of the discrete solutions U_h is not altered if r_h is replaced by an equivalent system. The corresponding order od convergence, however, does change in general. Similarly, the consistency or otherwise of a discretization is not affected by the replacement of (s_h) by an equivalent system.

Stummel has shown that if the operators $r_h : X_o \to X_h$ are linear and satisfy (3.2) for each x in a dense subspace X_o of X, then they can be extended into linear operators $r_h : X \to X_h$ which satisfy (3.2) for each x in X. The extended system is unique up to equivalences. For a proof see Vainikko [46], p. 11. The possibility of this extension is helpful in practice: consider the case $X = L^2(0,1)$. The commonly used operator $r_h u = [u(0), u(h), \ldots, u(1)]^T$ is only defined when u is continuous, since general L^2 functions are only defined almost everywhere [26].

Assume that we have introduced restriction operators $r_h : X \to X_h$, $s_h : Y \to Y_h$. Stummel [43] says that the sequence (V_h), where V_h belongs to X_h, *converges discretely* toward an element v in X if $\|r_h v - V_h\|_{X_h} \to 0$. With this terminology the convergence of (3.3) defined in Section 1.2 is nothing but the discrete convergence of the solutions U_h toward u. Furthermore, Stummel says that the *bounded* operators $B_h:X_h \to Y_h$ *converge discretely* toward the *bounded* operator $B:X \to Y$ if B_hV_h converges discretely toward Bv whenever V_h converges to v discretely: in symbols, $\|r_hv-V_h\| \to 0 \Rightarrow \|s_hBv-B_hV_h\| \to 0$. It is clear that the conclusion of Theorem 3.4 can now be expressed by saying that A_h converge discretely toward A. By analogy, the conclusion of Theorem 3.3 states the discrete convergence of A_h^{-1} toward E.

4. REGULAR AND COMPACT APPROXIMATIONS

4.1 Regular approximation

The concept of regular approximation was introduced by Grigorieff [14], [15] and provides a useful way of proving stability. We first need the notion of discrete compactness (Stummel [43]).

Definition 4.1 Let $r_h:X \to X_h$ be restriction operators as in Section 3.2. A family (indexed by h) of elements $V_h \in X_h$ is called (discretely) r_h-compact if, to each sequence h_j, $j = 0,1,\ldots$ with $h_j \to 0$, there corresponds a subsequence h_{j_r} and an element $v \in X$ with $\lim_r \|V_{h_{j_r}} - r_{h_{j_r}} v\| = 0$.

Note that in Stummel's terminology (Section 3.6) discrete compactness demands that each sequence (V_{h_j}) possesses a discretely convergent subsequence. We now place ourselves in the framework of the *second* paradigm as in Section 3.2.

Definition 4.2 Assume that A has domain D(A) = X, is bounded and satisfies (3.1b). The operators A_h satisfying (1.1b) are said to provide a regular approximation of A (with respect to the restrictions r_h, s_h) if the following conditions hold:

(R1) The discretization (3.3) is L-consistent for each f in R(A).

(R2) If (V_h) is a family such that $\|V_h\| \le$ constant and (A_hV_h) is s_h-compact, then (V_h) is r_h-compact.

In order to check (R1) see Theorem 3.4. The following result is fundamental.

Theorem 4.1 Assume that A is as in the previous definition and that A_h provide a regular approximation of A. Then A is onto (R(A) = Y) and possesses a

bounded inverse. Furthermore, for each f in Y, the discretization (3.4) is stable, uniformly bounded and L-convergent.

Proof If (1.9) does not hold, then there exist V_{h_j}, $j = 0,1,\ldots$ with $\|V_{h_j}\| = 1$, $\lim_j \|A_{h_j}V_{h_j}\| = 0$. The condition (R2) shows then that (V_{h_j}) possesses a subsequence which converges discretely to an element $v \in X$. (This subsequence is still denoted (V_{h_j}).) The condition (R1) implies that $A_{h_j}V_{h_j}$ converge discretely to Av. But $A_{h_j}V_{h_j}$ and (3.1a) and (3.2) (with s_h replacing r_h) lead to v = 0. This is in contradiction to $\|v\| = 1$ which follows from (3.2). To see that A is onto, note first that, for each f in Y, the sequence s_hf is discretely s_h-compact. The stability and (R2) imply that $(A_h^{-1}s_hf)$ possesses a subsequence that converges discretely toward an element u in X. Then Au = f, since (R1) holds. The remaining properties follow from the general results of the previous chapters.

Example. Elliptic Galerkin methods in non-coercive situations. We consider the model boundary value problem

$$-u'' + b(x)\,u = f(x), \quad 0 \le x \le 1, \quad u(0) = u(1) = 0, \tag{4.1}$$

where b is a given real continuous function and f a datum. What follows is easily extended to more general elliptic problems with any number of independent variables. Let L^2 be the space of real, square integrable functions on $0 \le x \le 1$ with the usual inner product $(\cdot,\cdot)$. Let us further denote by H_0^1 the space of functions in L^2 whose (distributional) derivative is also in L^2 and which vanish at x = 0,1. In H_0^1 we use the norm $\|v\|_1 = \|v'\|_{L^2}$. With these definitions, the weak form of (4.1) (see, e.g., Strang and Fix [42], Ciarlet [5], Fairweather [10]) requires us to find u in H_0^1 such that, for each w in H_0^1,

$$(u',w') + (bu, w) = (f,w). \tag{4.2}$$

Here f can be an L^2 function, but (4.2) also makes sense if $(f,\cdot)$ represents a continuous, linear functional on H_0^1. We denote by H^{-1} the space of all such functionals with the usual dual norm $\|f\|_{-1} = \sup\{|(f,w)|: \|w\|_1 \le 1\}$. On introducing the operator $A:H_0^1 \to H^{-1}$ which sends each v in H_0^1 into the linear form $w \to (u',w') + (bu,w)$, equation (4.2) reads simply Au = f. We *assume* that $\ker(A) = \{0\}$. (This injectivity, which holds in particular for positive b, does not hold for arbitrary b: b could be, say, constant and equal to an eigenvalue of the operator $-u''$. On the other hand, it is clear that A is bounded.)

The weak form of (4.2) is discretized by means of Galerkin's method. If, for $0 < h < 1$, X_h is a finite-dimensional subspace of H_0^1, we seek an element U_h such that, for each W in X_h,

$$(U_h', W) + (bU_h, W) = (f, W). \tag{4.3}$$

To each v in H_0^1 there corresponds a unique Galerkin projection $r_h v$ belonging to X_h and defined by the condition that, for any W in X_h,

$$(v'-(r_h v)', W') = 0. \tag{4.4}$$

We assume that the family X_h satisfies the following condition: a constant C independent of h and an integer $m \geq 2$ exist such that, for each v in H_0^1 whose derivatives $D^j v$ are in L^2, $1 \leq j \leq m$,

$$\|v-r_h v\|_{L^2} + h\,\|(v-r_h v)'\|_{L^2} \leq C\, h^j\, \|D^j v\|_{L^2}. \tag{4.5}$$

This property holds if X_h is one of the usual spaces of polynomials of degree m-1 in a (perhaps nonuniform) grid in $0 \leq x \leq 1$ with diameter h [5], [10], [42].

Denote by Y_h the space of (continuous) linear functionals on X_h with the norm $\|F\|_{Y_h} = \sup\{|(F,W)|: \|W\|_{X_h} \leq 1\}$, by s_h the mapping which takes each $f \in H^{-1}$ into its restriction to X_h and by A_h the operator $A_h: X_h \to Y_h$ such that $A_h V = (V',\cdot) + (bV,\cdot)$. Then (4.3) takes the simple form $A_h U_h = s_h f_h$. It is easy to show that r_h, s_h satisfy (3.2) and thus we are within the framework of the second paradigm. It will be shown next that A_h provide a regular approximation to A.

We first observe that the local truncation error at a datum Av is given by the linear functional which maps $W \in X_h$ into

$$\begin{aligned}((r_h v)',W') &+ (br_h v,W) - (s_h Av,W)\\ &= (v', W') + (br_h v, W) - (Av,W) \qquad (4.6)\\ &= (b\,(v-r_h v), W).\end{aligned}$$

Therefore (4.5) implies consistency if v is smooth. Now (R1) follows from Theorem 3.4, since the uniform boundedness of A_h is easily proved.

Next assume that $V_h \in X_h$, $\|V_h\|_{H_0^1} \leq$ constant and $(A_h V_h)$ is s_h-compact. On recalling that a bounded sequence in H_0^1 possesses an L^2-convergent subsequence, we conclude that elements v in L^2, g in H^{-1} exist such that $\|V_{h_j} - v\|_{L^2} \to 0$, $\|A_{h_j} - s_h g\|_{Y_h} \to 0$, $h_j \to 0$. (The subscript j is deleted hereafter.) From these relations it is easily proved that, for any w in H_0^1 such that w'' is also in L^2,

$$-(v, w'') + (bv, w) = (g, w). \tag{4.7}$$

This shows first that $-v'' + bv = b$ (in the distributional sense) and so $v'' = bv - g \in H^{-1}$ implying $v' \in L^2$. Then integration by parts in (4.7) yields $v(0) = v(1) = 0$ and thus $v \in H_0^1$, $Av = g$. Finally, with $W_h = r_h v - V_h$, (4.2) - (4.4) imply

$$\begin{aligned}\|r_h v - V_h\|_{X_h}^2 &= ((r_h v - V_h)', W_h') = (v', W') - (V_h', W_h')\\ &= (b(v-V_h), W_h) + (g - A_h V_h, W_h).\end{aligned}$$

In the right-most term, the first inner product tends to 0 because $v - V_h = W_h$ tend to 0 in L^2, while the second inner product tends to 0 because $\|s_h g - A_h V_h\|_{Y_h} \to 0$. We conclude that V_h converges discretely toward v and so (R2) holds.

Theorem 4.1 now asserts that, under the hypotheses above, to each f in H^{-1} corresponds a unique weak solution u of the problem, that the Galerkin equations are uniquely solvable for small h and that $r_h u - U_h$ can be bounded above and below by (cf. (4.6)) $\sup\{|(b(u-r_h u),W)|: W \in X_h, \|W\| \leq 1\}$. In particular, if $D^j u$ is in L^2, $1 \leq j \leq m$, then (4.5) implies that $\|r_h u - U_h\|_{H_0^1} = O(h^m)$. On using (4.5) once more we derive $\|u - r_h u\|_{L^2} = O(h^m)$, $\|(u-r_h u)'\|_{L^2} = O(h^{m-1})$. These estimates were first obtained by Schatz [32], who employed a different technique.

Note that if $b \equiv 0$, then $s_h A = A_h r_h$ and thus the local truncation error is always 0. This in turns implies that the global error is also 0, i.e., $U_h = r_h u$. Theorem 3.5 shows that the discretization is stable, but this fact can also be easily derived from the relation $s_h A = A_h r_h$. More generally, assume that b is such that $(.',.') + (b,.)$ is a *coercive bilinear form* in H_0^1 and endow H_0^1 with the energy norm (i.e., with the norm induced by this bi- linear form). On choosing $r_h u$ to be the orthogonal projection associated with the energy norm rather than (4.4), we conclude again that $s_h A = A_h r_h$ and $U_h = r_h u$. The Galerkin solution coincides with the best approximation $r_h u$ to u in the energy norm, provided that the problem is coercive.

4.2 Compact convergence of operators

In this section we present a technique for proving (R2) in the definition of regular approximation.

Theorem 4.2 Assume that the operators $A_h : X_h \to Y_h$ are of the form $A_h = B_h + C_h$ where (i) $B_h : X_h \to Y_h$ and there exists a bounded, invertible operator B mapping X onto Y such that $B_h U_h = s_h f$ is a stable, consistent discretization of $Bu = f$ whenever $f \in Y$, (ii) $C_h : X_h \to Y_h$ and $V_h \in X_h$, $h \in H$, $\|V_h\| \le$ constant implies that $(C_h V_h)$ is discretely s_h-compact. Then (R2) in Theorem 4.1 holds.

Proof Assume that $\|V_h\| \le$ constant and $(A_h V_h)$ is discretely s_h-compact. Then $(B_h V_h) = (C_h V_h - A_h V_h)$ is also discretely compact, since (ii) holds. For an appropriate subsequence (which we still denote by (V_h)) and an appropriate f in Y, $\|s_h f - B_h V_h\| \to 0$. The identity

$$r_h B^{-1} f - V_h = -B_h^{-1}(B_h V_h - s_h f) - B_h^{-1}(s_h f - B_h r_h B^{-1} f)$$

makes it clear that V_h converges discretely toward $B^{-1}f$.

In the case where, for each h, $X_h = X$, $Y_h = Y$, $C_h = C$ and r_h, s_h are the identity mapping, the property (ii) is simply the compactness of the operator C. When $X_h = X$, $Y_h = Y$, $s_h = r_h = \mathrm{Id}$, but C_h varies with h, (ii) coincides with the notion of collective compactness considered by Anselone [1]. The generalization to spaces X_h, Y_h that vary with h is due to Stummel.

Example. Quadrature methods in integral equations (See, e.g., [9]). We consider the equation of the second kind

$$(Au)(x) = u(x) + \int_0^1 K(x,y)u(y)dy = f(x),$$

where the datum is continuous, the kernel K is twice continuously different- iable and we seek continuous solutions. We set X = Y = space of continuous

functions in $0 \le x \le 1$, with the maximum norm. It is assumed that the corresponding homogeneous equation only possesses the trivial solution.

If J is an integer, we introduce a grid $x_j = jh$, $j = 0,1,\dots,J$, $h = 1/J$ and look for approximations U_j to $u(x_j)$ by solving $(i = 0,1,\dots,J)$

$$U_i + \Sigma_j hK(x_i,x_j)U_j = f(x_j), \tag{4.8}$$

a system which originates from the replacement of the integral by the trape- zoidal rule. Throughout the example, summation is in j, $j = 0,1,\dots,J$ and the terms $j = 0,J$ must be halved. We set $X_h = Y_h$ and equal to the space of $(J+1)$-vectors with the maximum norm and $r_h = s_h$ and equal to the operator which takes each function v into the vector with entries $v(x_j)$. Note that $\|r_h\| = 1$. We have defined in this way all the elements that are needed for the second paradigm. Clearly (4.9) is uniformly bounded. The i-th component of the local truncation error is given by

$$l_i = \Sigma_j hK(x_i,x_j)u(x_j) - \int_0^1 K(x_i,y)u(y)dy. \tag{4.9}$$

Taylor expansion shows that if u is twice continuously differentiable, $|l_i|$ possesses a bound $C(u)h^2$. On applying Theorem 3.4 we conclude that the requirement (R1) in the definition of regular approximation holds. In order to prove (R2) we resort to Theorem 4.2, with $B_h = \mathrm{Id}_{X_h}$ and C_h the matrix with entries $hK(x_i,x_j)$. The hypothesis (i) is trivially satisfied with $B = \mathrm{Id}_X$ and we turn to (ii). Let $V_h = [V_0,\dots,V_J]^T \in X_h$ with $\sup_h \|V_h\| < \infty$. The family of functions $\phi_h(x) = \Sigma_j K(x,x_j)V_j$ is relatively compact in X (just apply Arzela's theorem, [26] Chapter 1). But then the property $\|r_h\| = 1$ shows that $r_h \phi_h(x) = C_h V_h$ is discretely r_h-compact.

We conclude that (4.8) is uniquely solvable for h sufficiently small and that $\max_j |u(x_j) - U_j|$ can be bounded above and below, uniformly in h, by $\max_i |l_i|$, with l_i given in (4.9). The same result is true even if K is only continuous; see [46] for this and other generalizations.

5. INITIAL VALUE PROBLEMS

5.1 One-step discretizations

This last chapter is devoted to some considerations on the definitions of stability and convergence in the important case of initial value problems. For simplicity we work within the *first* paradigm. The treatment of Example B in Chapter 3 shows the way in which extra results can be gained when employing the *second* paradigm.

Let W be a normed space. We assume that the fixed theoretical solution we try to approximate is a W-valued function u(t) of the real variable t, $0 \leq t \leq T < \infty$. In systems of s scalar ODEs, W is the space R^s or C^s. In evolutionary PDEs, W consists of scalar or vector valued functions $u(x_1,\dots,x_d)$ of d space variables (cf. Example B where $X = L^2_p$). Let k be a parameter taking values in a set K of positive numbers with inf K = 0 (in this section k, K replace h and H, so that h can be used for spatial meshsizes). We consider the discretization

$$U^0 = \eta_k \text{ (given)} \tag{5.1a}$$

$$k^{-1}U^{n+1} = k^{-1}C_k U^n, \; n = 0,1,\dots,N-1, \; N = [T/h], \tag{5.1b}$$

where C_k is a linear bounded operator mapping W into itself and whose norm depends continuously on k. The space X_k is, by definition, the space of (N+1)-vectors $V_k = [V^0,V^1,\dots,V^N]^T$, $V^n \in W$ with the maximum norm $\|V_k\| = \max_n \|V^n\|_W$. The space Y_k is also the space of (N+1)-vectors $F_k = [F^0,F^1,\dots,F^N]^T$, $F^n \in W$, but now with the L^1 norm $\|F_k\| = \|F^0\|_W + \sum_{n=1}^{N} k\,\|V^n\|_W$. On setting $u_k = [u(t_0), u(t_1),\dots,u(t_N)]^T$, $t_n = nk$, we have defined a first paradigm framework. Convergence simply means $\lim_k \max_n \|u(t_n)-U^n\|_W = 0$. Stability, just as in Section 1.4, is equivalent to the requirement

$$\sup_k \max_{0\leq n\leq N} \|C^n_k\| =: L < \infty, \tag{5.2a}$$

which can also be expressed in the form: a constant L exists such that

$$\|U^n\|_W \leq L\,\|U^0\|_W, \tag{5.2b}$$

for arbitrary $k \in K$, $0 \leq nk \leq T$, $U^0 \in W$. Turning now to consistency, the 0-th component of the local truncation error is given by

$$u(0) - \eta_k, \tag{5.3a}$$

i.e., the error in the starting value $U^0 = \eta_k$. The remaining components are $1_{n+1} = k^{-1}d_{n+1}$, n = 0,1,...,N-1, where

$$d_{n+1} = u(t_{n+1}) - C_k u(t_n). \tag{5.3b}$$

The residual d_{n+1} has a clear interpretation: it represents the difference between the exact $u(t_{n+1})$ and the element $C_k u(t_n)$ which one would have obtained from the recursion (5.1) if U^n had been correct: $U^n = u(t_n)$. This consideration explains the term 'local' truncation error. Consistency demands

$$\lim_k \|u(0) - \eta_k\|_W = 0 \tag{5.4a}$$

together with

$$\lim_k \sum_{n=1}^{N} \|d_n\| = 0 \tag{5.4b}$$

a requirement which is verified in the particular case

$$\max_n \|d_n\| = o(k). \tag{5.4c}$$

Finally, just as in the Remark in Section 2.1, L-convergence is equivalent to the demand that convergence takes place for arbitrary η_k satisfying (5.4a). Therefore Theorem 2.2 asserts that if (5.4b) holds, then the stability (5.2) is necessary and sufficient for convergence to take place for arbitrary choices of η_k satisfying (5.4a).

5.2 Implicit schemes

Often, the recurrence for the computation of U^n is not of the form (5.1b), but rather of the *implicit* form

$$k^{-1}C_{1k}U^{n+1} = k^{-1}C_{2k}U^n, \; n = 0,1,\dots,N-1, \tag{5.5}$$

with C_{1k}, C_{2k} bounded operators mapping W into W and whose norms depend continuously on k. We assume that for each k, C_{1k} is invertible, so that (5.5) defines U^{n+1} uniquely. There are two alternative ways of analyzing (5.1a), (5.5):

(i) On defining $C_k = C^{-1}_{1k}C_{2k}$, (5.5) takes the form (5.1b) and can be treated by the means of the previous section. Stability is identical with the requirement (5.2) and *follows* from the condition "$\max_n \|u(t_n)-U^n\| \to 0$ whenever (5.4a) holds".

(ii) The discretization (5.1a), (5.5) is written in matrix form (1.1), with

$$A_k = k^{-1}\begin{bmatrix} kI & \cdot & \cdot & \cdot & \cdot \\ -C_{2k} & C_{1k} & \cdot & \cdot & \cdot \\ \cdot & -C_{2k} & C_{1k} & \cdot & \cdot \\ \cdot & & \cdot & \cdot & \cdot \end{bmatrix}.$$

The inverse A_k^{-1} is easily found:

$$A_k^{-1} = k\begin{bmatrix} k^{-1}I & & & & \\ k^{-1}C_k & C_{1k}^{-1} & & & \\ k^{-1}C_k^2 & C_k & C_{1k}^{-1} & & \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ k^{-1}C_k^N & C_k^{N-1} & C_k^{N-2} & \cdots & C_{1k}^{-1} \end{bmatrix}$$

and computation of A_k^{-1} according to Lemma 1.1 shows that now stability is given by (5.2) *together with the extra condition*

$$\sup_k \|C_{1k}^{-1}\| < \infty. \tag{5.6}$$

In this setting, L-convergence is a strictly *stronger* requirement than the demand that convergence takes place for arbitrary η_k satisfying (5.4a). The reason for this is that now the operators C_{1k}^{-1}, which contribute to $\|A_k^{-1}\|$, do not feature in the first column of A_k^{-1}, the only column which operates on the initial datum when forming $U_k = A_k^{-1}f_k$. Consequently in this setting convergence for arbitrary consistent η_k does not imply stability (it implies (5.5) though, as we say above).

These considerations illustrate the fact that the same discretization can be written in several different ways for analytic purposes and that the stability requirements may vary with the way of writing the discretization. Here, when working within the alternative (i), stability means insensitivity with respect to small perturbations δ_k^n

$$k^{-1}\tilde{U}^{n+1} = k^{-1}C_k\tilde{U}^n + \delta_k^n,$$

i.e.

$$k^{-1}C_{1k}\tilde{U}^{n+1} = k^{-1}C_{2k}\tilde{U}^n + C_{1k}\delta_k^n, \tag{5.7}$$

whereas within alternative (ii), the perturbations are

$$k^{-1}C_{1k}\tilde{U}^{n+1} = k^{-1}C_{2k}\tilde{U}^n + \delta_k^n. \tag{5.8}$$

Clearly (5.8) accounts better than (5.7) for the sort of perturbation found in practice, where C_k is not formed.

The *uniform invertibility condition* (5.6) was first introduced by Strang [42].

5.3 Multistep schemes

For notational simplicity we restrict ourselves to the two-step case

$$U^0 = \eta_k^0,\ U^1 = \eta_k^1, \tag{5.9a}$$

$$k^{-1}U^{n+2} = k^{-1}C_{1k}U^{n+1} + k^{-1}C_{2k}U^n,\ n = 0,1,\dots,N-2. \tag{5.9b}$$

This discretization can be rewritten as a one-step recursion for the compound vectors $\underset{\sim}{U}^n = (U^{n+1}, U^n)^T \in W \times W$. Namely

$$\underset{\sim}{U}^0 = (\eta_k^1, \eta_k^0),$$

$$k^{-1}\underset{\sim}{U}^{n+1} = k^{-1}\begin{bmatrix} C_{1k} & C_{2k} \\ I & 0 \end{bmatrix}\underset{\sim}{U}^n = k^{-1}C_k\underset{\sim}{U}^n, \tag{5.10b}$$

$n = 0,1,\dots,N-2$. We endow $W \times W$ with the norm $\|(V_1,V_2)^T\|_{W\times W} = \max(\|V_1\|,\|V_2\|)$ and consider the space X_k (resp. Y_k) of N-dimensional vectors with components in $W \times W$ with the maximum (resp. L^1 norm). Finally we set $u_k = [(u(t_1),u(t_0))^T, \dots,(u(t_N), u(t_{N-1}))^T]^T$. With these definitions, convergences still means $\max_n \|u(t_n)-U^n\|_W \to 0$ and the stability condition is given by (5.2a). The formula (5.2b) can clearly be replaced by

$$\|U^n\|_W \leq L\,\|\underset{\sim}{U}^0\| \leq L\max(\|U^0\|_W, \|U^1\|_W),$$

for arbitrary k, $0 \leq nk \leq T$, U^0, U^1 in W. The local truncation error is given by

$$[(u(t_1)-\eta_k^1,\ u(t_0)-\eta_k^0)^T,\ (k^{-1}d_1,0)^T,\dots,(k^{-1}d_{N-1},0)^T]^T$$

where the elements

$$d_n = u(t_{n+1}) - C_{1k}u(t_n) - C_{2k}u(t_{n-1})$$

possess an interpretation similar to that of the one-step case. Thus, consistency is equivalent to (5.4b) together with

$$\lim_k \|u(t_0) - \eta_k^0\| = \lim_k \|u(t_1) - \eta_k^1\| = 0. \tag{5.11}$$

When (5.4b) holds, convergence for arbitrary η_k^0, η_k^1 satisfying (5.11) takes place if and only if the discretization is stable.

5.4 Time-dependent operators

Often the operator C_k in (5.1b) depends on t and the discretization takes the form

$$k^{-1}U^{n+1} = k^{-1}C_k(t_n)U^n.$$

It is assumed that C_k depends continuously on k and t, $0 \leq t \leq T$. Now the inverse A_k^{-1} is given by

$$A_k^{-1} = \begin{bmatrix} k^{-1}I & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ k^{-1}P_{1,1} & I & & & & \cdot & \cdot & \cdot & \cdot \\ k^{-1}P_{2,1} & P_{2,2} & I & & \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ k^{-1}P_{N,1} & P_{N,2} & P_{N,3} & & \cdot & \cdot & \cdot & & I \end{bmatrix}$$

where $P_{i,j}$ is the composite operator $C_k(t_{j-1})\dots C_k(t_i)C_k(t_{i-1})$. Lemma 1.1 shows that stability demands the uniform boundedness of these products. As in alternative (ii) in Section 5.2, not all the products appear in the first column. For this reason, convergence for arbitrary consistent initial data does not imply stability. A counterexample can be seen in Ansorge [2] p. 63.

5.5 A perturbation result. The Dahlquist-Henrici theory of linear multistep methods

Let us now consider discretizations of the form

$$U^0 = \eta_k,\ \text{given}, \tag{5.12a}$$

$$k^{-1}U^{n+1} = k^{-1}C_k(t_n)U^n + B_{1k}(t_n)U^n + B_{2k}(t_n)U^{n+1}, \tag{5.12b}$$

$n = 0,1,\dots,N-1$, where $C_k(t)$, $B_{1k}(t)$, $B_{2k}(t)$ are operators in W whose norms depend continuously on k and t. The following important perturbation result holds.

Theorem 5.1 Assume that W is a Banach space and that there exists a constant M such that, for each k in K and t with $0 \leq t \leq T$, $\|B_{1k}\| \leq M$, $\|B_{2k}\| \leq M$. Then (5.12) is stable if and only if the discretization given by (5.12a) and $k^{-1}U^{n+1} = k^{-1}C_k(t_n)U^n$ is stable, $n = 0,1,\dots,N-1$.

Proof See Grigorieff [16] which allows nonlinear, Lipschitz-continuous B_{1k}, B_{2k}. Earlier versions are due to Kreiss [21] and Strang [40]. A similar perturbation theorem does not hold for general discretizations (i.e., those that do not stem from initial value problems) unless the size M of the perturbation is sufficiently small. (See the discussion in Stetter [39] p. 21.)

As an application, we examine the Dahlquist-Henrici theory of linear multistep methods (see Henrici [19]). In $W = R^s$ we consider initial value problems

$$u(0)\ \text{given},\ du/dt = A(t)u(t) + f(t),\quad 0 \leq t \leq T,$$

where A(t) is a matrix depending continuously on t and f is a continuous vector-valued function. (What follows holds if the right-hand side of the equation is nonlinear, Lipschitz-continuous in u, but in this paper we deal only with linear problems.) If α_i, β_i, $i = 0,1,\dots,r$, are fixed real constants with $\alpha_r = 1$, we consider the linear r-step method

$$U^0, U^1,\dots,U^{r-1}\ \text{given},$$

$$k^{-1}\sum_{j=0}^{r}\alpha_j U^{n+j} = \sum_{j=0}^{r}\beta_j(A(t_{n+j})U^{n+j} + f(t_{n+j})).$$

For the analysis the method is rewritten as a one-step recursion (Section 5.3). On introducing the characteristic polynomials

$$\rho(z) = \sum_{j=0}^{r}\alpha_j z^j,\quad \sigma(z) = \sum_{j=0}^{r}\beta_j z^j,$$

it is easy to prove that, if $u(t)$ is smooth, then (5.4b) holds if

$$\rho(1) = 0, \quad \rho'(z) = \sigma(1). \tag{5.13}$$

(Conversely, these conditions must be fulfilled if (5.4b) is to hold for arbitrary, smooth $u(t)$.) Next recall that the stability of the discreti- zation is independent of the inhomogeneous term f. The perturbation theorem shows then that our discretization is stable if and only if the discretization

$$k^{-1} \sum_{j=0}^{r} \alpha_j U^{n+j} = 0, \quad U^0,\ldots,U^{r-1} \text{ given}, \tag{5.14}$$

is stable. The solutions of (5.14) are readily available in closed form in terms of the roots of $\rho(z) = 0$. Thus one easily concludes that stability holds if and only if ρ satisfies the *root condition*: ρ has all its roots inside the closed unit disk and roots of modulus 1 are simple. The basic theory shows that, when (5.13) is satisfied, the root condition is necessary and sufficient for convergence for arbitrary, consistent choices of $U^0,\ldots,U^{r-1}$.

5.6 Strong stability. The energy method

One of the difficulties in the investigation of the stability condition (5.2a) for a given discretization stems from the fact that (5.2a) involves the *powers* C_k^n, whereas in practice one is only given C_k in (5.1b). Unfor- tunately, in general,

$$\|C_k^n\| \neq \|C_k\|^n \tag{5.15}$$

and therefore information on $\|C_k\|$ does not necessarily yield useful inform- ation on $\|C_k^n\|$. (If W is an inner product space and C_k, $k \in K$ are self- adjoint or normal operators, then equality holds in (5.15). In the general case only $\|C_k^n\| \leq \|C_k\|^n$.)

Kreiss [21] introduced a stability definition whose checking is a given situation does not demand knowledge of the powers of C_k.

Definition 5.1 A discretization (5.1) of an initial value problem is called strongly stable if there exist positive constants k_0, L_1, L_2, L_3 such that, for each $k \leq k_0$, the space W possesses a norm $\|\cdot\|_k$ for which

(i) $L_1 \|V\|_k \leq \|V\|_W \leq L_2 \|V\|_k$, for each V in W, $k \leq k_0$.

(ii) $\|C_k\|_k \leq 1+L_3k$ for each $k \leq k_0$, i.e. $\|U^1\|_k \leq (1+L_3k)\|U^0\|_k$ for arbitrary U^0 in W, $k \leq k_0$.

Theorem 5.2 A strongly stable discretization is stable.

Proof $\|U^n\|_W \leq L_2 \|U^n\|_k \leq L_2(1+L_3k)^n \|U^0\|_k \leq L_2\exp(L_3T)\|U^0\|_k \leq L_2 \exp(L_3T)L_1^{-1}\|U^0\|_W$. Note that here C_k might depend on t as in Section 5.4.

Example: The energy method. The convection problem (1.4) is discretized by the *leap-frog* scheme

$$U^0,\ U^1 \text{ given in } L_p^2,$$
$$k^{-1}U^{n+2} = k^{-1}(U^n - r(T_h+T_h^{-1})U^{n+1}),\ n = 0,1,\ldots,N-2, \tag{5.16}$$

where L_p^2, r and T_h are as in Example B, Section 1.1. Recall (Section 5.3) that here C_k maps the compound vector $(U^{n+1}, U^n)^T$ into $(U^{n+2}, U^{n+1})^T$ and $\|(U^{n+1},U^n)^T\| = \max(\|U^{n+1}\|_{L_p^2}, \|U^n\|_{L_p^2})$. If $(V_1,V_2)^T \in W \times W$, we set

$$\|(V_1,V_2)^T\|_k^2 = \|V_1\|^2 + \|V_2\|^2 + r\langle(T_h + T_h^{-1})V_2,V_1\rangle,$$

where the angular brackets denote the usual L_p^2 inner product. On noting that

$$|\langle(T_h+T_h^{-1})V_2,V_1\rangle| \leq \|(T_n+T_h^{-1})V_2\|\,\|V_1\| \leq 2\,\|V_2\|\,\|V_1\|,$$

we conclude that (i) in Definition 5.1 holds provided that $r < 1$. Now take the inner product of (5.16) and $U^n + U^{n+2}$ and rearrange to get

$$\|(U^{n+2},U^{n+1})^T\|_k^2 = \|(U^{n+1},U^n)^T\|_k^2 - r\langle(T_h+T_h^{-1})U^{n+1},U^n\rangle$$
$$-r\langle(T_h+T_h^{-1})U^n,U^{n+1}\rangle.$$

Periodicity implies that the inner products cancel each other and thus $\|(U^{n+2},U^{n+1})^T\|_k = \|(U^{n+1},U^n)^T\|_k$ or $\|C_k\|_k = 1$. Therefore (5.16) is strongly stable when $r < 1$.

In more general situations the use of the energy method demands such ingenuity in the construction of an appropriate norm $\|\cdot\|_k$ (the so-called energy) and careful use of the techniques of summation and integration by parts. An excellent account can be seen in Chapter 6 of Richtmyer and Morton [27].

5.7 The von Neumann analysis

In Example B, Section 1.4, we presented a simple von Neumann analysis. In this section we comment briefly on the scope of this technique. Assume that W is the space L_p^2 of 1-periodic, C^s-valued functions of a real variable x. Note that this space may arise either when dealing with systems of PDEs having a scalar-valued dependent variables or when discretizing a single scalar PDE by means of an s-step discretization. Functions ϕ in W possess a Fourier series representation (1.11) with a_m s-dimensional complex vectors. If C_k consists of constant coefficient (i.e. x-independent) linear combinations of translations, then formulae (1.12) are still valid, but now a_m, b_m are vectors and $\hat{c}_k(m)$ suitable matrices. Clearly,

$$\|C_k^n\| = \sup_m |\hat{c}_k(m)^n|,$$

where the bars represent the matrix norm derived from the norm used in C^s. We face again the difficulty encountered in the previous section - in general, the operations of taking matrix norms and forming powers do not commute. On recalling that a matrix norm is always larger than the corresponding spectral radius, we conclude that, for each eigenvalue $\lambda_k^{(i)}(m)$, $i = 1,2,\ldots,s$, of $\hat{c}_k(m)$,

$$\|C_k^n\| \geq \sup_m |\lambda_k^{(i)}(m)^n| = (\sup_m |\lambda_k^{(i)}(m)|)^n,$$

so that the *von Neumann* condition $\sup_m |\lambda_m^{(i)}(m)| \leq 1+O(k)$, $k \to 0$ is *necessary* for stability. The condition is also *sufficient* if $s = 1$ or, more generally, if $|\hat{c}^n| = |\hat{c}|^n$. Additional hypotheses that guarantee the sufficiency of the von Neumann condition for stability in L^2 can be seen in Richtmyer and Morton [27], Chapter 4. The symbol or amplification matrix $\hat{c}$ contains full information on the discretization and can be used to derive stability conditions in other norms, see, e.g., Thomee [45] and Brenner et al [4]. Finally, the results of the von Neumann analysis can be extended to variable coefficients

situations, see Thomee [45], Richtmyer and Morton [27], Chapter 5.

5.8 Weakened stability requirements

We return again to the discretization (5.1), with W mapping W into itself. In the applications we have in mind, W consists of functions of one or more space variables. Let us assume that Z is a subspace of W such that $C_k V$ is in Z whenever V is in Z. Furthermore, we suppose that a norm $\|\cdot\|_Z$ has been defined for which a positive constant M exists such that, for all V in Z, $\|V\|_W \leq M\|V\|_Z$ (i.e., the natural injection $Z \to W$ is continuous). In the applications, Z consists of 'smooth' functions and convergence with respect to $\|\cdot\|_Z$ represents convergence of the function together with some of its derivatives.

If the starting element $U^0 = \eta_k$ lies in Z, then all the iterates U^n will also belong to Z. Therefore it is possible to consider the mapping A_k in (1.1) as an operator of the space X_k of (N+1)-vectors with components in W into the space Y_k of (N+1)-vectors with components in Z. In X_k, Y_k we consider the norms $\max_n \|U^n\|_W$, $\|F^0\|_Z + \Sigma k \|F^n\|_Z$ respectively. An application of Lemma 1.1 shows that stability is now expressed by

$$\sup_k \max_{0 \leq n \leq N} \|C_k^n\|_{L(Z,W)} = L' \tag{5.17a}$$

(where the norm is now that of bounded operators $Z \to W$) or

$$\|U^n\|_W \leq L' \|U^0\|_Z, \quad 0 \leq nk \leq T, \|U^0\| \in Z. \tag{5.17b}$$

On recalling that $\|\cdot\|_W \leq M\|\cdot\|_Z$, we conclude that if (5.2a) holds then (5.17) is satisfied with $L' \leq LM$. Thus (5.17) is a weaker requirement than (5.2). This is in agreement with the fact that (5.17) ensures insensitivity only with respect to perturbations that lie in Z (i.e., that are smooth), while (5.2) ensures insensitivity with respect to perturbations in W. We refer to [45] for an extensive collection of results on the present notion of weakened stability and to [31] for a study of the relation between (5.2) and (5.17). Earlier references are [21] and [49].

5.9 Fully discrete schemes

Throughout the present chapter, and for the sake of simplicity, it has been assumed that the theoretical $u(t_n)$ and numerical U^n elements have been members of the same space W. When practically dealing with PDEs, $u(t_n)$ is in W, but the numerical element U^n is defined only at grid points (or is sought in an appropriate finite-dimensional space) and therefore lies in a discrete space W_k that varies with k. Accordingly, C_k maps W_k into W_k and X_k, Y_k consist of (N+1)-vectors whose components belong to W_k. The theoretical element u_k is of the form $[u^0,u^1,\dots,u^N]^T$ where $u^n \in W_k$ is a suitable representation of $u(t_n) \in W$. The contents of the chapter can be easily extended to cover this new, more general situation. The reader is referred to [25] for a treatment of this case within the second paradigm.

REFERENCES

[1] Anselone, P.M. Convergence and error bounds for approximate solutions of integral and operator equations. *Proc. Symp. Wisconsin Madison* 1965, 231-252.

[2] Ansorge, R. *Differenzenapproximationen partieller Anfangswertaufgaben.* Stuttgart: Teubner 1978.

[3] Aubin, J.P. Approximations des espaces de distributions et des operateurs différentiels. *Bull. Soc. Mat. France, 12* (1967) 1-139.

[4] Brenner, P., Thomee, V. Whalbin, L.B. *Besov Spaces and Applications to Difference Methods for Initial Value Problems.* Lecture Notes in Mathematics 434. Berlin: Springer 1975.

[5] Ciarlet, P.G. *The Finite Element Method for Elliptic Problems.* Amsterdam: North Holland 1978.

[6] Cullen, M.J.P., Morton, K.W. Analysis of evolutionary error in finite element and other methods. *J. Comput. Phys. 34* (1980) 245-267.

[7] Dahlquist, G. Stability and error bounds in the numerical integration of ordinary differential equations. Uppsala: *Transactions of the Royal Institute of Technology 130*, 1959.

[8] Dekker, K., Verwer, J.G. *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations.* Amsterdam: North Holland 1984.

[9] Delves, L.M., Walsh, J. (eds.) *Numerical Solution of Integral Equations.* Oxford: Clarendon 1974.

[10] Fairweather, G. *Finite Element Galerkin Methods for Differential Equations.* New York: Marcel Dekker 1978.

[11] Geveci, T. The significance of the stability of difference schemes in different L^p spaces, *SIAM Review 24* (1982) 413-426.

[12] Gottlieb, D., Orszag, S.A. *Numerical Analysis of Spectral Methods: Theory and Applications.* CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia: SIAM 1977.

[13] Griffiths, D.F., Sanz-Serna, J.M. On the scope of the method of modified equations. University of Dundee, Department of Mathematical Sciences Report NA/83, September 1984. (Also submitted to *SIAM J. Sci. Stat. Comput.*)

[14] Grigorieff, R.D. Zur Theorie approximationsregulärer Operatoren I. *Math. Nachr. 55* (1973) 233-249.

[15] Grigorieff, R.D. Zur Theorie approximationsregulärer Operatoren II. *Math. Nachr. 55* (1973) 251-263.

[16] Grigorieff, R.D. Stability of multistep methods on variable grids. *Numer. Math. 42* (1983) 359-377.

[17] Grigorieff, R.D. Some stability inequalities for compact finite difference schemes. Preprint.

[18] Grigorieff, R.D. Einige Stabilitatsungleichungen für gewöhnliche Differenzengleichungen in nichtkompakter Form. Preprint.

[19] Henrici, P. *Discrete Variable Methods in Ordinary Differential Equations.* New York: John Wiley 1962.

[20] Isaacson, E., Keller, H.B. *Analysis of Numerical Methods.* New York: John Wiley 1966.

[21] Kreiss, H.D. Uber die Stabilitatsdefinition für Differenzengleichungen die partielle Differntialgleichungen approximieren. *BIT 2* (1962) 153-181.

[22] Lambert, J.D. *Computational Methods in Ordinary Differential Equations.* London: John Wiley 1973.

[23] Lax, P.D., Richtmyer, R.D. Survey of the stability of linear difference equations. *Comm. Pure Appl. Math. 9*, (1956) 267-293.

[24] Palencia, C., Sanz-Serna, J.M. Equivalence theorems for incomplete spaces: an appraisal. *IMA. J. Numer. Anal. 4*, (1984) 109-115.

[25] Palencia, C., Sanz-Serna, J.M. An extension of the Lax-Richtmyer theory. *Numer. Math. 44* (1984) 279-283.

[26] Reed, M., Simon, B. *Methods of Modern Mathematical Physics I: Functional Analysis.* New York: Academic Press 1980.

[27] Richtmyer, R.D., Morton, K.W. *Difference Methods for Initial Value Problems.* New York: Interscience 1967.

[28] Sanz-Serna, J.M. Convergence of the Lambert-McLeod trajectory solver and of the CELF method. *Numer. Math.* (to appear).

[29] Sanz-Serna, J.M. Convergent approximations to partial differential equations and stability concepts of methods for stiff systems of ordinary differential equations. Unpublished report, available on request.

[30] Sanz-Serna, J.M., Palencia, C. A general equivalence theorem in the theory of discretization methods. *Math. Comp.* (to appear.)

[31] Sanz-Serna, J.M., Spijker, M.N. Regions of stability, equivalence theorems and the Courant-Friedrichs-Lewy condition. Preprint.

[32] Schatz, A.H. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp. 28* (1974) 959-962.

[33] Skeel, R. Analysis of fixed-stepsize methods. *SIAM J. Numer. Anal. 13* (1976) 664-685.

[34] Skeel, R.D., Jackson, L.W. Consistency of Nordsieck methods. *SIAM J. Numer. Anal. 14* (1977) 910-924.

[35] Spijker, M.N. *Stability and Convergence of Finite-Difference Methods.* Leiden: Rijksuniversiteit 1968.

[36] Spijker, M.N. On the structure of error estimates for finite-difference methods. *Numer. Math. 18* (1971) 73-100.

[37] Spijker, M.N. Equivalence theorems for nonlinear finite-difference methods, in *Numerische Behandlung nichtlinearer Integrodifferential und Differentialgleichungen.* Lecture Notes in Mathematics 395, R. Ansorge and W. Torning eds. 109-122. Berlin: Springer 1974.

[38] Spijker, M.N. On the possibility of two-sided error bounds in the numerical solution of initial value problems. *Numer. Math. 26* (1976) 271-300.

[39] Stetter, H.J. *Analysis of Discretization Methods for Ordinary Differential Equations.* Berlin: Springer 1973.

[40] Strang, G. Accurate partial difference methods II. Nonlinear problems. *Numer. Math. 6* (1964) 37-46.

[41] Strang, G. Wiener-Hopf difference equations. *J. Math. Mech. 13* (1964) 85-96.

[42] Strang, G., Fix, G.J. *An Analysis of the Finite Element Method.* Englewood Cliffs, N.J.: Prentice Hall 1973.

[43] Stummel, F. Diskrete Konvergenz linearer Operatoren I. *Math. Ann. 190* (1970) 45-92.

[44] Stummel, F. Biconvergence, bistability and consistency of one-step methods for the numerical solution of initial value problems in ordinary differential equtions, in *Topics in Numerical Analysis* (J.J.H. Miller ed.) II, 197-211. London-New York: Academic Press 1975.

[45] Thomee, V. Stability theory for partial difference operators, *SIAM Review 11* (1969) 152-195.

[46] Vainikko, G. *Funktionalanalysis der Diskretisierungsmethoden.* Leipzig: Teubner 1976.

[47] Varga, R.S. *Matrix Iterative Analysis.* Englewood Cliffs N.J.: Prentice Hall 1962.

[48] Verwer, J.G., Sanz-Serna, J.M. Convergence of method of lines approximations to partial differential equations. *Computing* (to appear).

[49] Wendroff, B. Well-posed problems and stable difference operators. *SIAM J. Numer. Anal. 5* (1968) 71-82.

J.M. Sanz-Serna
Departamento de Ecuaciones Funcionales
Facultad de Ciencias
Universidad de Valladolid
Valladolid,
Spain