IMA Journal of Numerical Analysis (1988) **8,** 71–84

Stability and Convergence in Numerical Analysis III: Linear Investigation of Nonlinear Stability

J. C. López-Marcos and J. M. Sanz–Serna

Departamento de Ecuaciones Funcionales, Facultad de Ciencias, Universidad de Valladolid, Valladolid, Spain

[Received 29 April 1986 and in revised form 21 January 1987]

In a previous paper, we showed that several standard definitions of stability of nonlinear discretizations are so strong that they classify as unstable a number of useful discretizations. Then a weaker definition was introduced which, however, was powerful enough to imply, together with consistency, the existence and convergence of the discrete solutions. In this paper we prove that, for smooth discretizations, stability in the new sense is equivalent to stability of its linearization around the theoretical solution. This fact does not imply that schemes with stable linearizations are automatically useful, due to the appearance of so-called stability thresholds. The abstract ideas introduced are applied to a concrete finite-element example, with a view to assessing the advantages of the new approach.

1. Introduction

This is the third and final paper in a series devoted to discussing the basic concepts of stability and convergence in numerical analysis. The paper is, however, essentially self-contained. Part I (Sanz-Serna 1985) surveyed the linear case and Part II (López-Marcos & Sanz-Serna 1985) reviewed a number of stability definitions intended to operate in nonlinear situations. Our study of the existing definitions led to the introduction of a new stability concept. The aim of the present paper is twofold. First we show that for smooth nonlinear discretizations, stability in the new sense is equivalent to the stability of the (linear) discretization obtained by linearization around the theoretical solution. As discussed later, this neat theoretical result by no means implies that nonlinear schemes with stable linearizations are automatically useful. This is due to the presence of the so-called stability thresholds (cf. Part II). Secondly we apply our abstract results to the study of a finite-element method: the so-called product approximation technique (Christie *et al.* 1981). In this respect our motivation stems from two sources: (i) We wanted to show that nonlinear Galerkin techniques can be advantageously cast in the simple general framework considered in this series of papers. (ii) We wanted to draw some attention to the fact that, even though (in convergence proofs) the Galerkin literature almost invariably bypasses the notions of stability, it is still possible to introduce a concept of stability in such a way that convergence follows from stability and consistency, and that stability ensures insensitivity to round-off errors and other perturbations.

The organization of the article is as follows. The second section presents the basic notations and definitions. The main results are stated in the third and fourth sections and proved in the fifth. The latter section also discusses the relation between our results and other available theories. Section 6 contains the application to the product approximation technique, and Section 7 is devoted to conclusions.

2. Preliminaries

2.1 *Consistency, Stability, and Convergence*

We begin by presenting, rather tersely, the definitions of consistency, stability, and convergence required later. These concepts have been discussed at length in Parts I and II and the reader is referred to these previous papers for a more critical and detailed treatment.

We consider a fixed given problem concerning a (not necessarily linear) differential or integral equation. Let u denote a solution of this problem. (Well-behaved nonlinear problems may of course possess more than one solution.) We denote by U_h a numerical approximation to u. The subscript h reflects that U_h depends on a (small) parameter h such as a mesh-size, element diameter, etc. We always assume that h takes values in a set H of positive numbers with $\inf H = 0$. The approximation U_h is reached by solving a discretized problem

$$\Phi_h(U_h) = 0, \tag{2.1}$$

where, for each h in H, the mapping Φ_h is fixed with domain $D_h \subset X_h$ and taking values in Y_h. Here X_h and Y_h are vector spaces, either both real or both complex, with

$$\dim X_h = \dim Y_h < \infty. \tag{2.2}$$

As h ranges in H, the family of discrete problems (2.1) is referred to as a discretization.

We further assume that, for each h in H, we have chosen a norm $\|\bullet\|_{X_h}$ in X_h, a norm $\|\bullet\|_{Y_h}$ in Y_h, and an element u_h in the interior of D_h which is a suitable 'discrete' representation of u in X_h. The notions of stability, consistency, and convergence of the discretization (2.1) depend on the specific choices of norms and u_h. For simplicity, the subscripts will often be omitted in the notation of the norm, and norms in different spaces will be denoted by $\|\bullet\|$.

If U_h is a solution of (2.1), then the element $e_h = u_h - U_h \in X_h$ is, by definition, the *global error* in U_h. We say that the discretization (2.1) is *convergent* if there exists $h_0 > 0$ such that, for each h in H with $h \leqslant h_0$, (2.1) possesses a solution U_h and

$$\lim_h \|u_h - U_h\| = \lim_h \|e_h\| = 0.$$

If, furthermore, $\|e_h\| = O(h^p)$ as $h \to 0$, then the convergence is said to be *of order p*.

The *local (discretization) error* in u_h is defined to be the element $l_h = \Phi_h(u_h) \in Y_h$. The discretization (2.1) is said to be *consistent* (resp. consistent of order p) if, as $h \to 0$, we have $\|l_h\| \to 0$ (resp. $\|l_h\| = O(h^p)$).

For simplicity in the presentation, we assume in this paper that no rounding errors are present, even though they are one of the main motivations behind the notion of stability (see Part II, remark 2.8).

The following definition was motivated and introduced in Part II of the present series of articles and extends an earlier definition due to H. B. Keller (1975).

DEFINITION 2.1 Suppose that, for each h in H, the value R_h lies in the range $(0, \infty]$. The discretization (2.1) is said to be *stable*, restricted to the thresholds R_h, if there exist positive constants h_0 and S (the stability constant) such that, for h in H with $h \leq h_0$, the open ball $B(u_h, R_h)$ is contained in the domain D_h, and such that, for any V_h and W_h in that ball,

$$\|V_h - W_h\| \leq S \|\Phi_h(V_h) - \Phi_h(W_h)\| . \tag{2.3}$$

Keller's definition is recovered if the thresholds R_h are independent of h. In the sequel, 'a discretization is stable' always means that it is stable in the sense of the Definition 2.1 for a suitable choice of thresholds R_h. Examples of the use of the new stability concept in practical settings were given in Part II.

The following result, also given in Part II, is crucial and uses a deep lemma due to Stetter (1973).

THEOREM 2.1 *Assume that* (2.1) *is consistent and stable with thresholds* R_h. *If* Φ_h *is continuous in* $B(u_h, R_h)$ *and* $\|l_h\| = o(R_h)$, *as* $h \to 0$, *then:*

(i) *For* h *small enough, the discrete equations* (2.1) *possess a solution in* $B(u_h, R_h)$.

(ii) *That solution is unique in the ball.*

(iii) *As* $h \to 0$, *the solutions converge. The order of convergence is not smaller than the order of consistency.*

In practice, the thresholds are often of the form $R_h = Rh^m$, where R and m are independent of h, with $R > 0$ and $m \geq 0$. Then the condition $\|l_h\| = o(R_h)$ reduces to $m < p$, with p the order of consistency. (It is still possible to prove convergence if $m \geq p$, see Part II.)

2.2 *Semistability*

It is technically advisable to introduce the notion of semistability, even though, as we will show later, this concept possesses little or no practical significance.

DEFINITION 2.2 Suppose that, for each h in H, the value R_h lies in $(0, \infty]$. Then the discretization (2.1) is said to be *semistable*, restricted to the thresholds R_h, if there exist positive constants h_0 and S (the semistability constant) such that, for h in H with $h \leq h_0$, the ball $B(u_h, R_h)$ is contained in the domain D_h and, for any W_h in that ball,

$$\|u_h - W_h\| \leq S \|\Phi_h(u_h) - \Phi_h(W_h)\| . \tag{2.4}$$

It is evident that semistability is a weaker requirement than stability: in (2.4), W_h can only be compared with the discrete representation u_h of u, while in (2.3) it may be compared with any element V_h in $B(u_h, R_h)$. In fact, the notion of semistability (together with consistency) is too weak to imply the existence of discrete solutions U_h (cf. conclusion (i) in Theorem 2.1). Therefore Theorem 2.1 must be weakened as follows.

THEOREM 2.2 *Assume that* (2.1) *is consistent and semistable with thresholds* R_h. *If, for h small enough,* (2.1) *possesses a solution* U_h *with* $\|u_h - U_h\| < R_h$, *then these solutions converge as* $h \to 0$. *The order of convergence is not smaller than the order of consistency.*

The proof of this result is trivial (see the remarks after Definition N1 in Part II) and its application is not very appealing: the existence of U_h and the a priori bound $\|u_h - U_h\| < R_h$ must be established before the theorem is applied.

3. Main results

In practice, the mapping Φ_h in (2.1) is smooth, so that the (Fréchet) derivative (i.e.—roughly—the Jacobian matrix) $\Phi_h'(u_h)$ of Φ_h at u_h exists. Furthermore, if the discretization (2.1) is successful, a solution U_h of (2.1), close to u_h, exists. Thus

$$0 = \Phi_h(U_h) \approx \Phi_h(u_h) + \Phi_h'(u_h)(U_h - u_h)$$

and one is led to consider the linearized discretizations

$$\Xi_h(U_h) \equiv \Phi_h(u_h) - \Phi_h'(u_h)u_h + \Phi_h'(u_h)U_h = 0. \tag{3.1}$$

This new discretization is stable if and only if the following condition holds (cf. Part II):

CONDITION (L) There exist positive constants h_0 and L such that, for h in H with $h \leq h_0$, the inverse $\Phi_h'(u_h)^{-1}$ exists and $\|\Phi_h'(u_h)^{-1}\| \leq L$.

Our aim is to study the relation between the stability of (2.1) and that of its linearization (3.1). The main motivation for this sort of research is that the stability or otherwise of linear discretizations is expected to be more easily investigated than that of their nonlinear counterparts.

Our first result is the following.

THEOREM 3.1 (Equivalence between nonlinear semistability and linearized stability.) *Assume that, for h in H with h sufficiently small, the mapping* Φ_h *in* (2.1) *is (Fréchet) differentiable at* u_h. *Then, the two conditions below are equivalent*:

(i) (2.1) *is semistable.*

(ii) *Condition* (L) *holds, i.e. the linearization* (3.1) *of* (2.1) *is stable.*

Proof. See Section 5.

From this theorem we conclude that, in practice, the checking of whether condition (L) holds is insufficient to show the stability of (2.1) and thence allow

the application of the powerful convergence Theorem 2.1. The next result presents conditions which, together with (L), lead to the stability of (2.1).

THEOREM 3.2 (A sufficient condition for nonlinear stability.) *Assume that, for each h in H with h sufficiently small, the mapping Φ_h in* (2.1) *is differentiable at each point v_h in an open ball* $B(u_h, R_h)$. *Suppose that condition* (L) *holds and that*

(C) *there exists a constant Q, with $0 \leq Q < 1$, such that, for h in H with h sufficiently small, and for each v_h in* $B(u_h, R_h)$, *we have*

$$\|\Phi_h'(v_h) - \Phi_h'(u_h)\| \leq Q/L. \tag{3.2}$$

Then (2.1) *is stable with thresholds R_h and stability constant $L/(1-Q)$.*

Proof. See Section 5.

Case where Φ_h' is Lipschitz continuous. It is of importance to study in some detail the case where Φ_h is differentiable in a ball $B(u_h, R_h)$ and the differential satisfies a Lipschitz condition at u_h:

$$\|\Phi_h'(v_h) - \Phi_h'(u_h)\| \leq C_h \|v_h - u_h\| \quad \text{for } v_h \in B(u_h, R_h). \tag{3.3}$$

(See López-Marcos (1985) for an analogous treatment of the Hölder-continuous case.) Assume that (L) holds and choose S, with $S > L$. Then, if h is sufficiently small and $\|v_h - u_h\| \leq \min\{R_h, (L^{-1} - S^{-1})C_h^{-1}\}$, we can write

$$\|\Phi_h'(v_h) - \Phi_h'(u_h)\| \leq C_h(L^{-1} - S^{-1})C_h^{-1} = [(S-L)/S]/L,$$

so that (3.2) holds with $Q = (S-L)/S$. We conclude that, under these hypotheses, (2.1) is stable with stability constant S and thresholds

$$\min\{R_h, (L^{-1} - S^{-1})C_h^{-1}\}. \tag{3.4}$$

Therefore, if we can choose, independently of h, the radius R_h of the balls $B(u_h, R_h)$ where Φ_h is differentiable, and the Lipschitz constants C_h, then (2.1) is stable with h-independent thresholds, i.e. stable in the sense of Keller (1975) (see Part II).

EXAMPLES For the Euler method for the scalar ODE $u' = f(u)$, a typical component of $\Phi_h(V_h)$ takes the form $(V^{n+1} - V^n)/h - f(V^n)$, so that the nonzero elements of a typical row of $\Phi_h'(V_h)$ are h^{-1} and $-h^{-1} - f'(V^n)$. (Here V_h is the vector of components V^0, $V^1, \ldots$). Therefore, in any discrete L^p norm, $\|\Phi_h'(V_h)\|$ grows unboundedly as $h \to 0$. However, when subtracting to form $\Phi_h'(V_h) - \Phi_h'(u_h)$, the unruly terms h^{-1} cancel: the nonzero entries of $\Phi_h'(V_h) - \Phi_h'(u_h)$ are of the form $f'(V^n) - f'(u^n)$, and (3.3) holds with C_h independent of h, provided that f is smooth. Thus, in the study of Euler's rule, there is no need for h-dependent thresholds. The present argument applies to any standard ODE method: for these methods, there is no need for h-dependent thresholds. In other words, the notion of stability in the sense of Keller is sufficient to study numerical methods for ODEs.

The situation with nonlinear PDEs is quite different. Take as an example the case of a one-step explicit discretization of a scalar evolutionary problem $u_t = Lu$. We have now to deal with expressions $(\boldsymbol{V}^{n+1} - \boldsymbol{V}^n)/h - \boldsymbol{f}_{\Delta x}(\boldsymbol{V}^n)$, where $\boldsymbol{f}_{\Delta x}$,

which contains negative powers of Δx, approximates the operator L. These expressions contribute to $\Phi_h'(V_h) - \Phi_h'(u_h)$ with terms $f_{\Delta x}(V^n) - f_{\Delta x}(u^n)$, which in general grow as $\Delta x \to 0$. As a result, $C_h \uparrow \infty$ as h is decreased, and the thresholds in (3.4) will shrink correspondingly.

A further cause for the shrinking of the thresholds in PDEs is that the radius R_h of the balls $B(u_h, R_h)$ where Φ_h is defined and differentiable with a Lipschitz continuous derivative may approach 0 as $h \to 0$. For nonlinear cases like $u_t = u_{xx} + f(u)$, the radius R_h is likely to be h-independent in the L^∞ norm (i.e. the discretization is defined and possesses a Lipschitz-continuous derivative in a uniform tube around the theoretical solution). However, to study convergence, one often uses weaker norms (the L^2 norm say) and then, as measured in the weaker norm, the radius tends to 0 with h. An example of this situation has been given in Part II, in the remarks after formula (8.12).

Case of Φ_h continuously differentiable. Note that (3.2) certainly holds if $\Phi_h'(\bullet)$ is continuous at u_h. This fact, combined with the Theorem 3.1, yields trivially the following neat result.

THEOREM 3.3 (Equivalence between linearized and nonlinear stability.) *Assume that, for h in H with h sufficiently small, the mapping Φ_h is continuously differentiable at u_h. Then the following conditions are equivalent*

(i) (2.1) *is stable.*
(ii) (2.1) *is semistable.*
(iii) *Condition* (L) *holds, i.e. the linearization* (3.1) *of* (2.1) *is stable.*

In practice, the smoothness requirements of Φ_h are almost invariably satisfied. The equivalence between (i) and (ii) shows then that, for smooth discretizations, there is no need to introduce the concept of semistability. The equivalence between (i) and (iii) implies that, when investigating stability, nonlinear discretizations may be replaced by their linearizations. However it should be emphasized that *linearized stability* implies stability of the nonlinear discretization restricted to *suitable thresholds*. The size of these thresholds must be known in order to apply the main convergence Theorem 2.1. Also, in practical computations, the thresholds of a linearly stable nonlinear scheme may be so small that, for all practical purposes, the discretization behaves in an unstable manner; cf. Richtmyer & Morton (1967: pp. 124–130) and Vadillo & Sanz-Serna (1985).

4. Linearization and convergence

Theorems 3.2 (linearized stability of smooth discretizations implies stability) and 2.1 (stability and consistency imply convergence) can be combined and, together, provide a powerful means for the study of nonlinear discretizations. Essentially, if the discretization is smooth, consistent, and possesses a stable linearization, then discrete solutions exist and their global error can be bounded in terms of the local error (namely $\|e_h\| \leq S \|l_h\|$), provided that the order of consistency is high enough to satisfy the condition $\|l_h\| = o(R_h)$.

In some applications, the bound $\|e_h\| \leq S \|l_h\|$ may be too pessimistic and it is

possible for the order of convergence to be actually higher than that of consistency. This phenomenon was discussed in detail in Part I and its avoidance is a matter of choosing carefully the norm in Y_h, i.e. the norm employed to measure local errors. If this choice has been appropriate, one often can derive a bound $\|l_h\| \leq K \|e_h\|$ which shows that the order of convergence cannot exceed that of consistency.

It is possible to use information on $\Phi_h'(u_h)$ in the derivation of bounds of the form $\|l_h\| \leq K \|e_h\|$. A result in that direction is given in the next theorem.

THEOREM 4.1 *Assume that the hypotheses of the Theorem* 3.2 *hold. Suppose also that a constant* M *exists such that* $\|\Phi_h'(u_h)\| \leq M$ *for* h *small. Then, if* $v_h, w_h \in \mathrm{B}(u_h, R_h)$, *we have*

$$\|\Phi_h(v_h) - \Phi_h(w_h)\| \leq M(1+Q)\|v_h - w_h\| . \tag{4.1}$$

In particular, if U_h *is a solution of* (2.1) *with* $\|U_h - u_h\| < R_h$, *then* $\|l_h\| \leq M(1+Q)\|e_h\|$.

Proof. See Section 5.

5. Proofs and remarks

5.1 *Proof of Theorem* 3.1

Suppose first that (2.1) is semistable as in the Definition 2.2. If x_h is in $\mathrm{B}(0, R_h)$, with h sufficiently small, we can define

$$g_h(x_h) = \Phi_h(u_h + x_h) - \Phi_h(u_h) - \Phi_h'(u_h)x_h,$$

so that $\|g_h(x_h)\| = o(\|x_h\|)$ as $x_h \to 0$. Choose $\delta > 0$. If h is sufficiently small, then a positive number $R_h^{\#}(\delta) \leq R_h$ can be found so that $\|x_h\| < R_h^{\#}(\delta)$ implies $\|g_h(x_h)\| < [\delta/S(S+\delta)]\|x_h\|$. Then

$$\begin{aligned}
\|\Phi_h'(u_h)x_h\| &= \|\Phi_h(u_h + x_h) - \Phi_h(u_h) - g_h(x_h)\| \\
&\geq \|\Phi_h(u_h + x_h) - \Phi_h(u_h)\| - \|g_h(x_h)\| \\
&\geq S^{-1}\|x_h\| - \|g_h(x_h)\| \\
&\geq S^{-1}[1 - \delta/(S+\delta)]\|x_h\| = (S+\delta)^{-1}\|x_h\| ;
\end{aligned}$$

whence it is clear that $\Phi_h'(u_h)^{-1}$ exists and its norm is bounded by $S + \delta$. Since this holds for arbitrary positive δ, one even has $\|\Phi_h'(u_h)^{-1}\| \leq S$.

Assume now that condition (L) holds. As before, if h is small enough and v_h is in D_h, then

$$\begin{aligned}
\|\Phi_h(v_h) - \Phi_h(u_h)\| &= \|\Phi_h'(u_h)(v_h - u_h) + g_h(v_h - u_h)\| \\
&\geq \|\Phi_h'(u_h)(v_h - u_h)\| - \|g_h(v_h - u_h)\| \\
&\geq L^{-1}\|v_h - u_h\| - \|g_h(v_h - u_h)\| .
\end{aligned} \tag{5.1}$$

Now, given $\delta > 0$, there exists a constant $R_h'(\delta)$ such that, if $\|v_h - u_h\| < R_h'(\delta)$, then $g_h(v_h - u_h)$ is defined and has norm less than $\delta/L(L+\delta)\|v_h - u_h\|$. This easily leads to semistability with constant $L + \delta$ and thresholds $R_h'(\delta)$.

5.2 *Proof of Theorem* 3.2

Take h sufficiently small and fix v_h and w_h in $\mathrm{B}(u_h, R_h)$. The mapping $t \mapsto A_h(t)$, defined for $0 \leq t \leq 1$ by

$$A_h(t) = (1-t)v_h + tw_h - \Phi_h'(u_h)^{-1}\big[\Phi_h\big((1-t)v_h + tw_h\big) - \Phi_h(v_h)\big],$$

is differentiable, with

$$\begin{aligned} A_h'(t) &= (w_h - v_h) - \Phi_h'(u_h)^{-1}\Phi_h'\big((1-t)v_h + tw_h\big)(w_h - v_h) \\ &= \Phi_h'(u_h)^{-1}\big[\Phi_h'(u_h) - \Phi_h'\big((1-t)v_h + tw_h\big)\big](w_h - v_h). \end{aligned}$$

Upon using (3.2) and (L), we conclude that

$$\|A_h'(t)\| \leq Q\,\|w_h - v_h\| \quad (0 \leq t \leq 1),$$

and the mean-value theorem yields then the estimate

$$\|A_h(1) - A_h(0)\| \leq Q\,\|w_h - v_h\|\,. \tag{5.2}$$

On the other hand,

$$\begin{aligned} \|A_h(1) - A_h(0)\| &= \|(w_h - v_h) - \Phi_h'(u_h)^{-1}[\Phi_h(w_h) - \Phi_h(v_h)]\| \\ &\geqslant \|w_h - v_h\| - L\,\|\Phi_h(w_h) - \Phi_h(v_h)\|\,. \end{aligned} \tag{5.3}$$

On combining (5.2) and (5.3) the proof is complete.

5.3 *Proof of Theorem* 4.1

For h small, fix v_h in $\mathrm{B}(u_h, R_h)$ and define, for x_h in $\mathrm{B}(u_h, R_h)$,

$$B_h(x_h) = x_h - \Phi_h'(u_h)^{-1}[\Phi_h(x_h) - \Phi_h(v_h)].$$

The bound $\|B_h'(x_h)\| \leq Q$ follows easily from the conditions (L) and (C), and then the mean-value theorem leads to $\|B_h(v_h) - B_h(w_h)\| \leq \|v_h - w_h\|$ for any w_h in $\mathrm{B}(u_h, R_h)$. Hence, by definition of B_h, we have

$$\|\Phi_h'(u_h)^{-1}[\Phi_h(w_h) - \Phi_h(v_h)]\| \leq (1+Q)\,\|v_h - w_h\|$$

and (4.1) follows without difficulty.

5.4 *Remarks, and Comparison with Alternative Theories*

It is useful to note in the proof of Theorem 3.1 that nonlinear semistability with constant S leads to stability of the linearization with the same constant: $\|\Phi_h'(u_h)^{-1}\| \leq S$. However, linearized stability with constant L only leads to semistability with constant $S = L + \delta$, where δ is arbitrary but positive. The choice of δ influences the size of the thresholds. The same conclusion was reached in (3.4) for the case of Lipschitz-continuous $\Phi_h'(\bullet)$.

The proof of the Theorem 3.1 makes it clear why linearized stability implies semistability rather than stability. For, if Φ_h is differentiable at u_h, then the expression

$$\Phi_h(v_h) - \Phi_h(w_h) - \Phi_h'(u_h)(v_h - w_h) \tag{5.4}$$

is generally not $o(\|v_h - w_h\|)$ as $v_h, w_h \to u_h$, and therefore the pair (u_h, v_h) in (5.1) cannot be replaced by a more general pair (v_h, w_h). In the case where (5.4) is in fact $o(\|v_h - w_h\|)$, one says that $\Phi_h'(u_h)$ is a *strong* Fréchet derivative (Ortega & Rheinbold (1970), p. 71) and linearized stability implies stability. More details can be seen in Lopez-Marcos (1985: Thm 2.1.13).

The proofs of Theorems 3.2 and 4.1 are similar to but simpler than that of Theorem (14) in §3 of Vainikko (1976). Vainikko's result shows that (nonlinear) consistency and linearized stability, with $\|\Phi_h'(u_h)\|$ bounded, imply (nonlinear) convergence with an order that equals that of consistency. His result is a particular case of our Theorems 2.1, 3.2, and 4.1 combined. However it is important to emphasize that Vainikko does not define the concept of stability in nonlinear situations. Furthermore it is not difficult to see that Vainikko's theorem is only applicable to settings that, in our terminology, originate h-independent thresholds.

Spijker (1974) has developed a theory of nonlinear discretizations which allows h-dependent thresholds. Spijker's Theorem 2 is similar to our Theorem 3.2. However, his definition of nonlinear stability demands the solvability of the discrete equations, while, in our treatment, that solvability is a *consequence* of the main Theorem 2.1 and need not be proved a priori. Furthermore, Spijker's thresholds are what we called in Part II 'right thresholds' and therefore suffer from the drawbacks outlined in Part II, §6, (F). Finally some of the technical conditions used by Spijker (notably his Condition 1) are rather strong and do not allow much freedom in the choice of norms.

In the specific field of initial-value problems, a number of nonlinear convergence theorems have been surveyed by Ansorge (1978). The most general result in Ansorge's book is due to von Dein, who shows that nonlinear convergence follows from nonlinear consistency and a suitable form of linearized stability which allows h-dependent thresholds. Von Dein's result can be recovered from our theory without much difficulty; cf. López-Marcos (1985: Ch. 3). It is perhaps useful to note that von Dein's proof is unduly complicated due to the fact that the condition $\|u_h - U_h\| < R_h$ is proved 'a priori' by an induction argument, which may be omitted by invoking our Theorem 2.1, as discussed in Part II, §6, (C).

6. An application

6.1 *Theoretical Problem and Numerical Method*

We consider the model nonlinear problem

$$-u'' + f(x, u) = 0 \quad (0 \leq x \leq 1), \tag{6.1a}$$

$$u(0) = u(1) = 0, \tag{6.1b}$$

where $' \equiv \mathrm{d}/\mathrm{d}x$ and f is a real-valued function. Let u denote a weak solution of (6.1); i.e. assume that u belongs to $\mathrm{H}_0^1 = \mathrm{H}_0^1(0, 1)$ and that for each v in H_0^1

$$\langle u', v' \rangle + \langle f(\bullet, u), v \rangle = 0. \tag{6.2}$$

Here $\langle \bullet, \bullet \rangle$ denotes the standard inner product in $L^2 = L^2(0,1)$. We only need the following very weak hypotheses

(H1) There exists a constant δ such that f and f_u are defined and continuous in the band

$$\Omega_\delta = \{(x, v) : 0 \leqslant x \leqslant 1,\ u(x) - \delta \leqslant v \leqslant u(x) + \delta\}.$$

(H2) u is an *isolated* solution, i.e. if z is in H_0^1 and is such that, for any v in H_0^1,

$$\langle z', v' \rangle + \langle f_u(\bullet, u)z, v \rangle = 0,$$

then $z = 0$.

In order to solve (6.1) numerically, we introduce partitions $\tau : 0 = x_0 < x_1 < \cdots < x_{N(\tau)} = 1$ and set $I_j = [x_{j-1}, x_j]$ and $h_j = x_j - x_{j-1}$ for $j = 1, \ldots, N(\tau)$. We consider a family of partitions $\{\tau_h : h \in H\}$, where $h = \max_j h_j$ is the diameter of the partition τ_h and H denotes a subset of $(0, 1]$ with $\inf H = 0$. Note that this family is not assumed to be quasi-uniform. Further, if k is an integer $\geqslant 1$, we denote by $M_k(h)$ the space of real continuous functions on $[0, 1]$ which in each subinterval I_j of the partition τ_h coincide with a polynomial of degree $\leqslant k$. The subspace $M_k^0(h) \subset M_k(h)$ consists of the functions which satisfy (6.1b). Finally, we denote by Q_h the Lagrange interpolation operator which maps each function g in $C[0, 1]$ into $Q_h g$, the unique element in $M_k(h)$ which interpolates g in $k+1$ uniformly distributed points in each I_j. (Lobatto points can also be employed, see Sanz-Serna & Abia 1984.) The so-called product approximation technique approximates u by an element $U_h \in M_k^0(h)$ such that

$$\langle U_h', v \rangle + \langle Q_h f(\bullet, U_h), v \rangle = 0 \tag{6.3}$$

for each v in $M_k^0(h)$. On the advantages and disadvantages of this alternative to the standard Galerkin approximation, see Christie *et al.* (1981), Douglas & Dupont (1975), Abia & Sanz-Serna (1984), and Fletcher (1983).

Hereafter, we write $f(U_h)$ instead of $f(\bullet, U_h)$, $f_u(U_h)$ instead of $f_u(\bullet, U_h)$, etc.

6.2 *Choice of X_h, Y_h, Norms, and u_h. Consistency*

We set $X_h = M_k^0(h)$, with the norm $\|v_h\|_1 = \langle v_h', v_h' \rangle^{\frac{1}{2}}$. For the role of Y_h we take the space dual to X_h with the dual norm. Let $\Phi_h : X_h \to Y_h$ be the mapping which associates with $v_h \in X_h$ the linear form $\Phi_h(v_h)$ defined by

$$\Phi_h(v_h)\bullet = \langle v_h', \bullet' \rangle + \langle Q_h f(v_h), \bullet \rangle. \tag{6.4}$$

Here a dot represents the element in X_h on which $\Phi_h(v_h)$ acts. With these notations it is obvious that the equations (6.3), which define the product approximation U_h, take the standard form (2.1).

As representation u_h of u in X_h we take the Galerkin projection $p_h u$, i.e. the element characterized by the relation $\langle u' - (p_h u)', w' \rangle = 0$ for w in X_h.

In order to show the consistency of the discretization, the following lemma is required

LEMMA (i) There exists a constant C, independent of h, such that, for each g in $C[0, 1]$, the bound $\|Q_h v\|_\infty \leqslant C \|v\|_\infty$ holds.

(ii) If $g \in C[0,1]$, then $\|(I - Q_h)g\|_\infty \to 0$ as $h \to 0$.

(iii) If v and v_h are in $C[0,1]$ for h in H and $\|v_h - v\|_\infty \to 0$ as $h \to 0$, then $\|Q_h v_h - v\|_\infty \to 0$ as $h \to 0$.

Proof. (i) and (ii) are well known from approximation theory. (iii) follows from (i)–(ii) and the decomposition

$$Q_h v_h - v = Q_h(v_h - v) + (I - Q_h)v.$$

To study the consistency, we first note that, from (6.3)–(6.4) and the definition of Galerkin projection, we have the following expression for the local error $\Phi_h(p_h u) \in Y_h$

$$\begin{aligned}\Phi_h(p_h u)\bullet &= \langle (p_h u)', \bullet' \rangle + \langle Q_h f(p_h u), \bullet \rangle \\ &= \langle (p_h u)' - u', \bullet' \rangle + \langle Q_h f(p_h u) - f(u), \bullet \rangle \\ &= \langle Q_h f(p_h u) - f(u), \bullet \rangle. \end{aligned} \tag{6.5}$$

Therefore a straightforward application of the Cauchy–Schwartz and Poincaré inequalities yields

$$\|\Phi_h(p_h u)\|_{Y_h} \leq \pi^{-1} \|Q_h f(p_h u) - f(u)\|_{L^2}. \tag{6.6}$$

Now, as $h \to 0$, we have $p_h u \to u$ in H^1 (due to the approximation properties of X_h), so that, a fortiori, $p_h u \to u$ in L^∞. Thus $p_h u$ lies, for h small, in the set Ω_δ where the hypothesis (H1) holds. This in turn implies that $f(p_h u) \to f(u)$ in L^∞. By the lemma, $Q_h f(p_h u) \to f(u)$ in L^∞ and then (6.6) shows that $\|\Phi_h(p_h u)\| = o(1)$.

6.3 *Stability*

With a view to applying Theorem 3.2, we note again that, for h small enough, $\|u - p_h u\|_1 < \delta$ (with δ as in H1), so that, if v_h is in $B(p_h u, \delta)$, then $\|u - v_h\|_1 < 2\delta$. By Sobolev's inequality, $\|u - v_h\|_\infty < \delta$ and thus $f_u(v_h)$ is well-defined. Clearly Φ_h is then differentiable at v_h and the corresponding Fréchet derivative is the mapping which sends $w_h \in X_h$ into the linear form

$$\Phi_h'(v_h)w_h\bullet = \langle w_h', \bullet' \rangle + \langle Q_h[f_u(v_h)w_h], \bullet \rangle.$$

An argument very similar to that employed in the consistency proof shows that, to each $\mu > 0$, there corresponds $\delta_\mu < \delta$ such that, for each h sufficiently small and each v_h in $B(p_h u, \delta_\mu)$,

$$\|\Phi_h'(v_h) - \Phi_h'(u_h)\| \leq \mu.$$

Thus the condition (C) in Theorem 3.2 is certainly satisfied. It is then enough to prove linearized stability to conclude nonlinear stability with h-independent thresholds. In other words, we need only prove that there exist $h_0 > 0$ and $S > 0$ such that, for h in H with $h \leq h_0$, and for v_h in X_h, we have

$$\|v_h\|_1 \leq S \|\Phi_h'(p_h u)v_h\|. \tag{6.7}$$

If this is not true, then we can choose sequences (h_n) in H and (v_{h_n}) with $\lim_n h_n = 0$, $v_{h_n} \in X_{h_n}$, $\|v_h\|_1 = 1$, and

$$\lim_n \|\Phi'_{h_n}(p_{h_n}u)v_{h_n}\| = 0. \tag{6.8}$$

For simplicity, we write v_n for v_{h_n}, p_n for p_{h_n}, etc.

Now H_0^1 can be compactly injected in $C[0,1]$, so that we can assume that (v_n) converges in $C[0,1]$ to a function z with $z(0) = z(1) = 0$. We claim that

$$-z'' + f_u(u)z = 0, \tag{6.9}$$

where the derivative z'' must be understood in the distributional sense; i.e. we claim that, if w is C^∞ with (compact) support in $(0,1)$, then

$$-\langle z, w''\rangle + \langle f_u(u)z, w\rangle = 0. \tag{6.10}$$

This follows from the observation that the right hand-side of (6.10) can be rewritten as

$$\langle v_n - z, w''\rangle + \langle f_u(u)z - Q_n[f_u(p_nu)v_n], w\rangle + \langle Q_n[f_u(p_nu)v_n], w - p_nw\rangle + \langle \Phi'_n(p_nu)v_n, p_nw\rangle,$$

a sum where each term can be easily shown to vanish in the limit $n \to \infty$. (The arguments are similar to those used to prove consistency.)

Equation (6.9) implies that $z \in C^2$ and a fortiori $z \in H_0^1$. On invoking the hypothesis (H2), we conclude that $z = 0$ or equivalently $\|v_n\|_\infty \to 0$. This fact, together with (6.8) and with the identity

$$1 = \|v_n\|_1^2 = \langle \Phi'_n(p_nv)v_n, v_n\rangle - \langle Q_n[f_u(p_nu)v_n], v_n\rangle$$

lead to the contradiction $1 = 0$. Thus (6.7) must hold, and the discretization is stable with h-independent thresholds.

It should be pointed out that it is not difficult to show that $\|\Phi'_h(p_hu)\|$ can be bounded independently of h (see Lopez-Marcos 1985) so that the Theorem 4.1 is also applicable.

The technique used here to prove linearized stability follows to some extent ideas of Grigorieff (1973a,b). (Compare also with Part I, Section 4.1.)

6.4 *Existence and Convergence of the Product Approximation*

A straightforward application of Theorems 2.1, 3.2, and 4.1 shows that, under the very mild hypotheses (H1) and (H2), there exist positive constants h_0, R, α, and β such that, for h in H with $h \leq h_0$, there exists a product approximation U_h unique in $B(p_hu, R)$ and satisfying

$$\alpha \|\Phi_h(p_hu)\| \leq \|p_hu - U_h\|_1 \leq \beta \|\Phi_h(p_hu)\|, \tag{6.11}$$

$$\lim_h \|p_hu - U_h\|_1 = 0. \tag{6.12}$$

Note that (6.12) implies, via the triangle inequality, that $U_h \to u$ in H_0^1 and, a fortiori in L^2. After (6.11), a finer study of the rate of convergence can be carried out by analysing $\|\Phi_h(p_h u)\|$. In this connection, the bound (6.6) is useful. In fact, it is not difficult to show (Lopez-Marcos 1985) that, if $f \in C^{k+1}(\Omega_\delta)$ and the partitions are quasiuniform, then $\|Q_h f(p_h u) - f(u)\|_{L^2} = O(h^{k+1})$ and therefore

$$\|p_h u_h - U_h\|_1 = O(h^{k+1}), \qquad \|p_h u_h - U_h\|_{L^2} = O(h^{k+1}),$$
$$\|u - U_h\|_1 = O(h^k), \qquad \|u - U_h\|_{L^2} = O(h^{k+1}).$$

7. Conclusions

In Section 6 we have applied the abstract ideas developed in Sections 2–5 to the study of the product approximation technique. This technique had been analyzed previously (Sanz-Serna & Abia, 1984). Actually, we choose the product approximation method as test application in order to be able to compare the new approach suggested here with more standard Galerkin analyses like that of Sanz-Serna & Abia (1984).

It would be unfair to say that the present treatment is much easier than that given by Sanz-Serna & Abia. However:

(i) The present treatment is systematic: we now possess a list of hypotheses to work through and know distinctly the role that each hypothesis plays in the overall picture. In contrast, the analysis in Sanz-Serna & Abia 1984 is highly ad hoc and definitely demanded a nontrivial amount of ingenuity.

(ii) The present treatment operates under much weaker hypotheses than that of Sanz-Serna and Abia, which among other superfluous requirements necessitated that f and f_u were defined and continuous in $-\infty < u < \infty$ and also needed the monotonicity assumption $f_u \geq m > -\pi^2$. The last assumption forces the global uniqueness of weak solutions and rules out a number of important nonlinearities $f(x, u)$ which effectively give rise to several isolated solutions. (Our treatment copes without difficulty with the case of nonunique isolated solutions, due to the presence of thresholds, see Part II, §6, (G).)

(iii) The present technique yields not only a convergence proof but also a stability result. We now know that small round-off errors or small perturbations in f will not change substantially the numerical solution. The study of stability properties has been somewhat neglected in the Galerkin-method literature, where convergence has traditionally been proved directly, rather than through the 'stability plus consistency' approach, standard in the finite-difference literature. In this connection, we feel that the stability notion considered here is appropriate for the investigation of stability and convergence of most nonlinear discretizations, both in the finite-element and finite-difference fields. As shown in Part II, a stronger definition would classify as unstable a number of useful discretizations, while a weaker concept would probably not be powerful enough to yield meaningful existence and convergence results. The crucial idea here is that of stability threshold: for smooth discretizations, nonlinear stability with suitable thresholds is equivalent to stability of the linearization.

Finally we would like to point out that the stability of a nonlinear discretization (2.1) can be investigated not only through the stability of its linearization (3.1) but also through the stability of other discretizations $\Lambda_h(U_h) = 0$ such that Λ_h approximates Φ_h near u_h. Results in that direction can be seen in López-Marcos (1985).

References

Abia, L., & Sanz-Serna, J. M. 1984 On the use of the product approximation technique in nonlinear Galerkin methods. *Int. J. numer. Methods Eng.* **20,** 778–779.

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

Christie, I., Griffiths, D. F., Mitchell, A. R., & Sanz-Serna, J. M. 1981 Product approximation for nonlinear problems in the finite element method. *IMA. J. numer. Anal.* **1,** 252–266.

Douglas, J. & Dupont, T. 1975 The effect of interpolating the coefficients in nonlinear parabolic Galerkin procedures. *Math. Comput.* **29,** 360–389.

Fletcher, C. A. J. 1983 The group finite element formulation. *Comp. Methods Appl. Mech. Eng.* **37,** 225–243.

Grigorieff, R. D. 1973a Zur Theorie approximationsregulaerer Operatoren I. *Mat. Nach.* **55,** 233–249.

Grigorieff, R. D. 1973b Zur Theorie approximationsregulaerer Operatoren II. *Mat. Nach.* **55,** 251–263.

Keller, H. B. 1975 Approximation methods for nonlinear problems with application to two-point boundary value problems. *Math. Comput.* **130,** 464–474.

Lopez-Marcos, J. C. 1985 Estabilidad de discretizaciones no lineales. Ph.D. Thesis, Universidad de Valladolid, Valladolid.

Lopez-Marcos, J. C., & Sanz-Serna, J. M. 1985 Stability and convergence in numerical analysis II: The definition of stability in nonlinear problems, preprint.

Ortega, J. M. & Rheinboldt, W. C. 1970 *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic Press.

Richtmyer, R. D. & Morton, K. W. 1967 Difference Methods for Initial Value Problems. New York: Interscience.

Sanz-Serna, J. M. 1985 Stability and convergence in numerical analysis I: Linear problems, a simple, comprehensive account. In: *Nonlinear Differential Equations and Applications* (J. K. Hale & P. Martínez-Amores, Eds), pp. 64–113. London: Pitman.

Sanz-Serna, J. M., & Abia, L. 1984 Interpolation of the coefficients in nonlinear Galerkin procedures. *SIAM J. numer. Anal.* **21,** 77–83.

Spijker, M. N. 1974 Equivalence theorems for nonlinear finite difference methods. In: *Numerische Behandlung Nichtlinearer Integrodifferential und Differential Gleichungen* (R. Ansorge & W. Törnig, Eds). Berlin: Springer.

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

Vadillo, F. & Sanz-Serna, J. M. 1986 Studies in numerical nonlinear instability II: a new look at $u_t + uu_x = 0$. *J. Comput Phys.* **66,** 225–238.

Vainikko, G. 1976 *Funktional Analysis der Diskretisierungmethoden.* Leipzig: Teubner.