stichting
mathematisch
centrum

MC

AFDELING NUMERIEKE WISKUNDE NW 161/83 SEPTEMBER
(DEPARTMENT OF NUMERICAL MATHEMATICS)

J.G. VERWER & K. DEKKER

STEP-BY-STEP STABILITY IN THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

kruislaan 413 1098 SJ amsterdam

Printed at the Mathematical Centre, Kruislaan 413, Amsterdam, The Netherlands.

The Mathematical Centre, founded 11 February 1946, is a non-profit institution for the promotion of pure and applied mathematics and computer science. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

AMS-MOS: 65XX02, 65M10, 65M20
CR-Number: 5.17

STEP-BY-STEP STABILITY IN THE NUMERICAL SOLUTION OF PARTIAL DIFFERENTIAL EQUATIONS

by

J.G. Verwer & K. Dekker

ABSTRACT

The subject of this paper is numerical stability in the time-integration of evolutionary problems in partial differential equations, primarily nonlinear problems. Following the method of lines approach, and supported by the strong developments which have taken place in the field of nonlinear stiff ordinary differential equations, the authors examine various useful numerical stability concepts for nonlinear partial differential equations, such as contractivity, monotonicity and conservation. The paper is of an expository nature. Its main objective is to illustrate the close connections between stiff problems and partial differential equations with respect to nonlinear stability. The well-known energy method plays an important role in this respect. Several examples of partial differential equations are treated so as to illustrate these connections. Now and then the authors embark upon applications which result from the nonlinear stability analysis, mainly in the two sections which deal with the well-known shallow water equations. For these equations a rigorous nonlinear stability analysis is presented.

KEY WORDS & PHRASES: *Initial Value Problems, Partial Differential Equations, Stiff Ordinary Differential Equations, Nonlinear Numerical Stability, Energy Method, Shallow Water Equations.*

1. Introduction

Many existing numerical methods for evolutionary problems in partial differential equations show a direct relation to integration methods for stiff ordinary differential equations. This relationship can be most clearly visualized via the so-called *method of lines* approach. Herewith the numerical solution process is considered as to consist of two parts, viz. *space-discretization* and *time-integration.* In the space-discretization the partial differential equation is converted into a system of ordinary differential equations by discretizing the space variables, while the time variable is left continuous. Usually, the space-discretization is performed, either by the finite difference method, or by the finite element method. Spectral methods can also be applied, however. In the time-integration the resulting system of differential equations, often called the *semi-discrete problem,* is integrated by an existing integration formula which is most appropriate for the problem at hand. Hence the relationship we are talking about lies in the time-integration of the partial differential equation.

This paper is concerned with *numerical stability* in the time-integration, primarily for *nonlinear problems.* Following the method of lines approach, and supported by the strong developments which have taken place in the field of nonlinear stiff ordinary differential equations, we examine various useful stability concepts for nonlinear partial differential equations. Herewith we should mention of course that in partial differential equations nonlinear numerical stability has been a topic much longer than in ordinary differential equations. Hampered by the large diversity the developments in partial differential equations drop behind, however. An example is furnished by the well-known energy method, the application of which still may be rather cumbersome (Richtmyer & Morton [20], Section 6.2).

Our paper is of an expository nature. Its main objective is to illustrate the close connections between stiff problems and partial differential equations with respect to *nonlinear stability.* We also take the opportunity, however, for embarking upon some new applications which result from our nonlinear stability analysis, mainly in Sections 7 and 8. We wish to mention that our paper bears a resemblance with recent work of Sanz-Serna [22], who surveys the use of stability and consistency concepts from the field of stiff equations for convergence proofs for approximations to time-dependent partial differential equations.

Section 2 is devoted to a discussion of the relevant stability concepts for semi-discrete problems, viz. *contractivity, monotonicity* and *conservation* In Section 3 we transplant these concepts to integration methods for semi-discrete problems. For many methods of practical importance, monotonicity and conservation can be established by combining the analytic tools from the field of stiff nonlinear problems with techniques from the aforementioned energy method. This is illustrated in the remaining sections which deal with several examples.

Sections 4,5 and 6 are devoted to classical examples, such as a simple nonlinear parabolic problem, a hyperbolic model problem, and a typical diffusion-convection equation. In Sections 7 and 8 we present a nonlinear stability analysis of approximations for the two-space dimensional *shallow water equations.* These equations are of direct practical importance in numerical fluid dynamics. We have chosen them as an example in our survey, since shallow water computations are often hampered by severe nonlinear instabilities. In our discussion we embark upon some possibly useful applications which result from a rigorous nonlinear stability analysis. Among others, we consider a locally one-dimensional splitting scheme, the stability of which can be *guaranteed* despite the non-linearities in the shallow water equations. We present this scheme as an alternative for two alternating direction implicit schemes proposed by Gustafsson [11] and Fairweather & Navon [7]. These alternating direction implicit schemes do suffer from nonlinear instabilities.

2. Review of some stability concepts for semi-discrete problems

Throughout this paper it is supposed that the (stiff) initial value problem

$$\dot{U} = F(t,U;\Delta), \quad \cdot = \frac{d}{dt}, \tag{2.1}$$

$$t>0, \quad U(0)=U^0,$$
$$F(t,\cdot;\Delta):\mathbb{R}^m \rightarrow \mathbb{R}^m,$$

represents a semi-discrete, time-continuous approximation to a given initial value problem or initial-boundary value problem for a partial differential equation. It will always be tacitly assumed that a unique solution exists and that F is as often differentiable as the numerical analysis requires.

The symbol Δ refers to the *grid spacing* in the space domain of the partial differential equation. We have included Δ as a parameter in the ordinary differential system (2.1) in order to emphasize that the vector function F and the vector variable U, also called the *time continuous grid function,* are always parameterized with respect to the grid spacing Δ in the space domain. The finer the grid, the larger m and the stiffer the problem. We are somewhat reluctant in calling (2.1) stiff in advance, since it is uncommon to classify semi-discrete hyperbolic equations as stiff. For our presentation, however, it is convenient to suppose that (2.1) is a stiff problem. This is justified if the stiffness of the problem is related to the grid distance in space.

It is necessary to obtain stability results which are valid uniform in the grid distance. Consequently, rather than considering one specific initial value problem (2.1), a whole family will be taken into consideration. Each particular grid distance in space defines a member of this family. In what follows, the statement *"independent of* Δ *or uniform in* Δ*"* always means that the property we are talking about applies to the whole semi-discrete family (2.1), in a uniform way. Herewith we note that in our analysis the dimension $m=m(\Delta)$ of (2.1) remains finite. We do not study convergence questions. Finally, for convenience of notation we will mostly write $\dot{U}=F(t,U)$ rather than $\dot{U}=F(t,U;\Delta)$.

Definition 2.1. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The differential equation (2.1) is called *dissipative* with respect to this norm, if every pair of solutions U and $\tilde{U}$ satisfies

$$\|\tilde{U}(t+\tau)-U(t+\tau)\| \leqslant \|\tilde{U}(t)-U(t)\|, \quad \text{all } \tau>0. \quad \square \tag{2.2}$$

Definition 2.2. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The differential equation (2.1) is called *monotone*, with respect to this norm, if every solution U satisfies

$$\|U(t+\tau)\| \leqslant \|U(t)\|, \quad \text{all } \tau>0. \quad \square \tag{2.3}$$

Definition 2.3. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The differential equation (2.1) is called *conservative*, with respect to this norm, if every solution U satisfies

$$\frac{d}{dt}\|U(t)\| = 0. \quad \square \tag{2.4}$$

Semi-discrete systems with these properties are frequently encountered in the time-integration of partial differential equations, in particular monotone and conservative systems. $\|U(t)\|$ is then related to some physical quantity, like mass, momentum or energy, which is kept monotone or is conserved by the partial differential equation. We will illustrate this in the examples of Sections 4-8. Needless to say that the above properties should hold uniform in Δ. Occasionally we will use the terminology "F is dissipative", etc.

Remark 2.4. From Definition (2.1)-(2.2) it trivially follows that

$$\text{dissipativity} \Rightarrow \text{monotonicity}$$

if $U(t)=\underline{0}$ is a solution of (2.1). The converse is not true, according to the following scalar counter example $\dot{s}=a(s)s$. This pseudo-linear scalar equation is monotone, iff $a(s)\leqslant 0$, all $s \in \mathbb{R}$. However, if $a(s)$ is such that $\partial[a(s)s]/\partial s>0$, the difference of any two solutions will increase with t, i.e. there is no contractivity. Contractivity or dissipativity deals with the growth of differences of solutions (perturbation sensitivity), whereas monotonicity deals with the growth of solutions itself. Hence, only for homogeneous linear problems

$$\dot{U} = A(t)U,$$

both concepts are identical, since here the difference of any two solutions is a solution too. The non-homogeneous linear problem

$$\dot{U} = A(t)U + G(t),$$

may very well be dissipative, but not monotone due to increasing solution components imposed by the forcing term $G(t)$. □

A fundamental role in the nonlinear stability analysis is played by the following result emanating from Dahlquist [4]:

Theorem 2.5. *Let* $\|\cdot\|$ *be a given norm on* $\mathbb{R}^m$. *Let* $\nu \in \mathbb{R}$ *be such that*

$$\mu[F'(t,\zeta)] \leqslant \nu, \quad \text{all} \quad \zeta \in \mathbb{R}^m, \quad F'(t,\zeta) = \partial F(t,\zeta) / \partial \xi,$$

where μ *denotes the logarithmic matrix norm corresponding to* $\|\cdot\|$. *Then, for any two solutions* U *and* $\tilde{U}$ *of* (2.1), *we have*

$$\|\tilde{U}(t+\tau) - U(t+\tau)\| \leqslant e^{\nu\tau} \|\tilde{U}(t) - U(t)\|, \quad \text{all} \quad \tau > 0. \quad \square \tag{2.5}$$

The proof of this important result is too long to repeat it here. It is also contained in [5], Chapter 1. That monograph surveys recent results on stability and consistency of Runge-Kutta methods for stiff nonlinear problems. Virtually all properties and results we deal with in Sections 2 and 3 of this paper are discussed in much more detail in that monograph. So, in Sections 2 and 3, we often tacitly refer to [5]. Needless to say that many results concerning nonlinear stiff problems emanate from the pioneering work of Dahlquist, Butcher, Burrage and many others.

Theorem 2.5 shows that the *logarithmic norm* of the Jacobian matrix of F can be used for investigating dissipativity, viz. if $\nu \leqslant 0$ equation (2.1) is dissipative. For a given matrix A, $\mu[A]$ is defined by

$$\mu[A] = \lim_{h \to 0+} \frac{\|I + hA\| - 1}{h}. \tag{2.6}$$

Here $\|\cdot\|$ denotes a subordinate matrix norm. For the standard norms $\mu[A]$ is known. Considering the l^1,l^2, and l^∞ norm in $\mathbb{R}^m$, we have ($A = (a_{ij})$)

$$\mu_1[A] = \max_j (a_{jj} + \sum_{i \neq j} |a_{ij}|), \tag{2.7}$$

$$\mu_2[A] = \text{maximal eigenvalue of } \frac{A + A^T}{2}, \tag{2.8}$$

$$\mu_\infty[A] = \max_i (a_{ij} + \sum_{j \neq i} |a_{ij}|). \tag{2.9}$$

For real inner product norms, $\|\zeta\| = <\zeta,\zeta>^{1/2}$, $\mu[A]$ can be written as

$$\mu[A] = \max_{\zeta \neq 0} \frac{<A\zeta,\zeta>}{\|\zeta\|^2}, \tag{2.10}$$

and is called *the one-sided Lipschitz constant* of A.

In fact, for inner product norms the condition $\mu[F'(t,\zeta)] \leqslant \nu$, all $\zeta \in \mathbb{R}^m$, can be replaced by the so-called *one-sided Lipschitz condition* for F:

$$<F(t,\zeta_1) - F(t,\zeta_2), \zeta_1 - \zeta_2> \leqslant \nu \|\zeta_1 - \zeta_2\|^2, \quad \text{all} \quad \zeta_1,\zeta_2 \in \mathbb{R}^m, \tag{2.11}$$

where ν is now called a one-sided Lipschitz constant of F. To see this, we denote

$$\phi(t) = \|\tilde{U}(t)-U(t)\|^2 = <\tilde{U}(t)-U(t),\tilde{U}(t)-U(t)>$$

for any two solutions U and $\tilde{U}$ of (2.1). Then

$$\dot{\phi}(t) = 2<\frac{d}{dt}[\tilde{U}(t)-U(t)],\tilde{U}(t)-U(t)> \leqslant 2\nu\phi(t).$$

Multiplying both sides of the inequality $\dot{\phi}(t)\leqslant 2\nu\phi(t)$ with

$$\eta(t) = \exp(-2\int_0^t \nu d\tau),$$

yields the inequality $\frac{d}{dt}(\phi(t)\eta(t))\leqslant 0$, from which we conclude that for $t\geqslant 0$, $\phi(t)\eta(t)$ monotonically decreases. This in turn implies relation (2.5) of Theorem 2.5.

Theorem 2.5 turns out to be very useful for semi-discrete partial differential equations, since in many applications the logarithmic norm of F' or the one-sided Lipschitz constant of F can be shown to be non-positive uniform in Δ, i.e. dissipativity of the whole semi-discrete family.

As shown in Remark 2.4, monotonicity cannot be concluded from contractivity if the zero vector is not a trivial solution of the differential equation (2.1). For inner product norms it is again easy to verify that any solution U satisfies

$$\|U(t+\tau)\|\leqslant e^{\theta\tau}\|U(t)\|, \text{ all } \tau>0, \tag{2.12}$$

if F is such that, for $\theta\in\mathbb{R}$,

$$<F(t,\zeta),\zeta>\leqslant\theta\|\zeta\|^2, \text{ all } \zeta\in\mathbb{R}^m. \tag{2.13}$$

Hence, condition (2.13) with $\theta\leqslant 0$ implies *monotonicity*. Similarly, for inner product norms the problem is *conservative,* iff

$$<F(t,\zeta),\zeta> = 0, \text{ all } \zeta\in\mathbb{R}^m. \tag{2.14}$$

We conclude this section with a result on the growth of solutions of homogeneous *pseudo-linear* systems

$$\dot{U} = A(t,U)U. \tag{2.15}$$

Theorem 2.6 (Dahlquist). *Let* $\|\cdot\|$ *be a given norm on* $\mathbb{R}^m$. *Let* $\nu\in\mathbb{R}$ *be such that*

$$\mu[A(t,\zeta)] \leqslant \nu, \text{ all } \zeta\in\mathbb{R}^m. \tag{2.16}$$

Then any solution U *of the pseudo-linear system* (2.15) *satisfies*

$$\|U(t+\tau)\| \leqslant e^{\nu\tau}\|U(t)\|, \text{ all } \tau>0. \quad \square \tag{2.17}$$

This result for the pseudo-linear system (2.15) enables us to investigate *monotonicity for arbitrary norms* without computing the Jacobian matrix which is necessary for the application of Theorem 2.5. Moreover, according to Remark 2.4, for pseudo-linear systems the condition of monotonicity is less restrictive then the condition of dissipativity.

3. Review of some stability concepts for integration formulas

Along the lines of the previous section we next define the corresponding stability properties for integration formulas for the semi-discrete system (2.1). We concentrate on one-step methods

$$U^n \rightarrow U^{n+1},$$

whose stepsize $t_{n+1}-t_n$ will be denoted by τ. $U^n\simeq U(t_n)$ is called the *fully discrete grid function* at time $t=t_n$. In what follows it is tacitly assumed that the step $U^n\rightarrow U^{n+1}$ is well-defined, i.e. we assume that U^{n+1} uniquely exists for any given τ and U^n. As familiar examples we mention the *explicit Euler* formula

$$U^{n+1} = U^n + \tau F(t_n, U^n), \tag{3.1}$$

the *implicit Euler* formula

$$U^{n+1} = U^n + \tau F(t_{n+1}, U^{n+1}), \tag{3.2}$$

and the *implicit midpoint* rule

$$U^{n+1} = U^n + \tau F(t_n + \tfrac{1}{2}\tau, \frac{U^n + U^{n+1}}{2}). \tag{3.3}$$

We will be mostly concerned with these standard schemes or variants thereof.

It is emphasized that in accordance with the approach for the time-continuous problem (2.1) all properties and results on numerical step-by-step stability should be valid uniform in the grid spacing Δ.

Definition 3.1. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The integration method is called *contractive* for (2.1), with respect to this norm, if a real number $\tau_0 = \tau_0(\Delta)$ exists such that

$$\|\tilde{U}^{n+1} - U^{n+1}\| \leq \|\tilde{U}^n - U^n\| \text{ for all } \tau \in (0, \tau_0]. \quad \square \tag{3.4}$$

This definition of numerical contractivity corresponds to Definition 2.1 which deals with contractivity of exact solutions. It is implicitly assumed that $\tau_0(\Delta)$ does not depend on specific choices of U^n and $\tilde{U}^n$. In our definitions on stability, U^n and $\tilde{U}^n$ always represent arbitrary points in $\mathbb{R}^m$.

We notice that it is often possible to take $\tau_0(\Delta) = \infty$, e.g. if the problem (2.1) is dissipative for some *inner product norm* and the integration method is an *algebraically stable* Runge-Kutta method such as (3.2) or (3.3) (cf. Burrage & Butcher [2], see also [5]). We shall give their proof for implicit midpoint and implicit Euler.

Example 3.2. Consider the 1-stage Runge-Kutta formula

$$Y = U^n + \lambda \tau F(Y),$$
$$U^{n+1} = U^n + \tau F(Y),$$

where, for convenience of notation, F is supposed to be autonomous. The coefficient values $\lambda = \tfrac{1}{2}$ and $\lambda = 1$ yield implicit midpoint and implicit Euler, respectively. Let U^n, Y, U^{n+1} and $\tilde{U}^n, \tilde{Y}, \tilde{U}^{n+1}$ denote two different numerical solutions. Further, denote $V_0 = \tilde{U}^n - U^n$, $V_1 = \tilde{Y} - Y$, $V = \tilde{U}^{n+1} - U^{n+1}$ and $W = \tau F(\tilde{Y}) - \tau F(Y)$. Then

$$V_1 = V_0 + \lambda W, \quad V = V_0 + W,$$

and

$$\|V\|^2 = \|V_0\|^2 + 2 < V_0, W > + \|W\|^2,$$
$$< V_0, W > = < V_1, W > - \lambda \|W\|^2.$$

Substitution of $< V_0, W >$ into the first formula yields

$$\|V\|^2 = \|V_0\|^2 + 2 < V_1, W > + (1 - 2\lambda)\|W\|^2.$$

The condition of dissipativity for F is $< V_1, W > \leq 0$, so that $\|V\| \leq \|V_0\|$ for all $\tau > 0$, if $\lambda \geq \tfrac{1}{2}$ (the condition of algebraic stability for the present 1-stage method). $\square$

Example 3.3. For the implicit Euler method contractivity with $\tau_0(\Delta) = \infty$ can be shown for *arbitrary* vector norms. Let $\|\cdot\|$ be a given norm. Suppose that for this norm $\mu[F'(t, \zeta)] \leq \nu$, all ζ (cf. Th. 2.5). Then for any two implicit Euler solutions U^n, U^{n+1} and $\tilde{U}^n, \tilde{U}^{n+1}$ it holds that

$$\|\tilde{U}^{n+1} - U^{n+1}\| \leq \frac{1}{1 - \nu\tau} \|\tilde{U}^n - U^n\|, \text{ all } \tau\nu < 1. \tag{3.5}$$

Hence, if $\nu \leqslant 0$ we have contractivity for all $\tau > 0$. This contractivity property of implicit Euler holds for any vector norm on $\mathbb{R}^m$. In that respect it is rather exceptional. The proof is based on logarithmic norm properties and is too long to repeat here (see [5], Section 2.4). □

Definition 3.4. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The integration method is called *monotone* for (2.1), with respect to this norm, if a real number $\tau_0 = \tau_0(\Delta)$ exists such that

$$\|U^{n+1}\| \leqslant \|U^n\| \text{ for all } \tau \in (o, \tau_0]. \quad \square \tag{3.6}$$

This definition corresponds to Definition 2.2 which deals with monotonicity of time-continuous solutions $U(t)$. Notice that

$$\text{numerical contractivity} \Rightarrow \text{numerical monotonicity}$$

if $U^{n+1} = U^n = \underline{0}$ is a solution of the integration method for all $\tau > 0$. As it is for contractivity, it is often possible to take $\tau_0(\Delta) = \infty$. For example, for inner product norms any algebraically stable Runge-Kutta method is monotone for (2.1) for all $\tau > 0$, if F satisfies the monotonicity condition $<F(t,\zeta),\zeta> \leqslant 0$, all $\zeta \in \mathbb{R}^m$ (see (2.13)). The proof of this result is completely analogous to the proof of the implication: algebraic stability $\Rightarrow$ BN-stability (Burrage & Bucher [2], Th. 2.2). For implicit Euler and implicit midpoint the proof of this implication is just the proof given in Example 3.2. For these two methods an interesting situation arises if the problem is of the *pseudo-linear* form (2.15).

Example 3.5. Let $\|\cdot\|$ be an inner product norm. The pseudo-linear system $\dot{U} = A(t,U)U$ is then monotone, if

$$<A(t,\zeta)\zeta,\zeta> \leqslant 0, \text{ all } \zeta \in \mathbb{R}^m.$$

In that case the implicit Euler solution

$$U^{n+1} = U^n + \tau A(t_{n+1}, U^{n+1})U^{n+1} \tag{3.7}$$

and implicit midpoint solution

$$U^{n+1} = U^n + \tau A(t_n + \tfrac{1}{2}\tau, \frac{U^n + U^{n+1}}{2})\frac{U^n + U^{n+1}}{2} \tag{3.8}$$

also behave monotone for all $\tau > 0$, since both methods are algebraically stable. Furthermore, if the matrix A satisfies

$$<A(t,\tilde{\zeta})\zeta,\zeta> \leqslant 0, \text{ all } \zeta,\tilde{\zeta} \in \mathbb{R}^m, \tag{3.9}$$

their *pseudo-linear forms*

$$U^{n+1} = U^n + \tau A(t_n, U^n)U^{n+1}, \tag{3.7'}$$

$$U^{n+1} = U^n + \tau A(t_n, U^n)\frac{U^n + U^{n+1}}{2}, \tag{3.8'}$$

are monotone too, for all $\tau > 0$. This result is a trivial consequence of the monotonicity of both methods for constant coefficient monotone problems $\dot{U} = AU$. In applications the pseudo-linear forms may be more attractive because the solution U^{n+1} is no longer implicitly defined. The pseudo-linear forms are expected to be somewhat less accurate, however. The order of consistency of (3.8') is one instead of two. This order decrase can be avoided by using the slightly more complicated pseudo-linear form

$$U^{n+1} = U^n + \tau A(t_n + \tfrac{1}{2}\tau, U^n + \tfrac{1}{2}\tau A(t_n, U^n)U^n)\frac{U^n + U^{n+1}}{2}. \tag{3.8''}$$

In partial differential equations we often meet pseudo-linear semi-discrete problems $\dot{U} = A(t,U)U$ satisfying (3.9). □

Example 3.6. Like for contractivity, implicit Euler can be proved to be monotone for *arbitrary* norms when applied to monotone pseudo-linear problems. Suppose that the matrix A satisfies condition (2.16) for some given appropriate norm $||\cdot||$. From logarithmic norm properties it then follows that both the implicit Euler solution (3.7) and the pseudo-linear solution (3.7') satisfy

$$||U^{n+1}|| \leqslant \frac{1}{1-\tau\nu}||U^n||, \text{ all } \tau\nu<1.$$

The proof of this result is almost identical to the proof of inequality (3.5). □

Definition 3.7. Let $||\cdot||$ be a given norm on $\mathbb{R}^m$. The integration method is called *conservative* for (2.1), with respect to this norm, if

$$||U^{n+1}|| = ||U^n|| \text{ for all } \tau>0. \quad \square \tag{3.10}$$

This definition corresponds to Definition 2.3 which deals with conservation of time-continuous solutions $U(t)$. It is well-known that implicit midpoint preserves this property for systems (2.1) satisfying the conservation relation (2.14): consider (3.3) and form the inner product of $U^{n+1}-U^n$ with $U^{n+1}+U^n$ to obtain

$$||U^{n+1}||^2-||U^n||^2 = <U^{n+1}-U^n,U^{n+1}+U^n>=0.$$

Likewise, if the matrix A of the pseudo-linear problem (2.15) satisfies the conservation relation

$$<A(t,\tilde{\zeta})\zeta,\zeta> = 0, \text{ all } \zeta,\tilde{\zeta}\in\mathbb{R}^m, \tag{3.11}$$

the pseudo-linear midpoint rule (3.8') can be proved to be conservative too.

In the following example we will show how the classical 4-th order explicit Runge-Kutta method can be made conservative and monotone.

Example 3.8. For convenience of notation we let F be autonomous. Consider the s-stage explicit Runge-Kutta formula

$$Y_i = U^n+\tau\sum_{j=1}^{i-1}a_{ij}F(Y_j), \; i=1(1)s, \tag{3.12}$$

$$U^{n+1} = U^n+\tau\sum_{j=1}^{s}b_jF(Y_j),$$

where $Y_1=U^n$. Let $||\cdot||$ be an inner product norm. Denote $F_j=\tau F(Y_j)$. Then

$$||U^{n+1}||^2 = ||U^n||^2+2\sum_{i=1}^{s}b_i<U^n,F_i>+\sum_{i,j=1}^{s}b_ib_j<F_i,F_j>.$$

From the inner product of Y_i with F_i we have

$$<U^n,F_i> = <Y_i,F_i>-\sum_{j=1}^{i-1}a_{ij}<F_i,F_j>,$$

which on substitution into the above relation gives

$$||U^{n+1}||^2 = ||U^n||^2+2\sum_{i=1}^{s}b_i<Y_i,F_i>-Q,$$

where (compare with relation (2.6) in Burrage & Butcher [2])

$$Q = \sum_{i,j=1}^{s}m_{ij}<F_i,F_j>, \; m_{ij}=b_ia_{ij}+b_ja_{ji}-b_ib_j. \tag{3.13}$$

The idea is to make the weights b_j solution dependent, such that $Q=0$. The scheme then preserves conservation, while it preserves monotonicity if all weights are positive. Below we will show that for the 4-stage scheme with the Butcher array

$$\begin{array}{c|cccc} 0 & 0 & & & \\ ½ & ½ & 0 & & \\ ½ & 0 & ½ & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \tilde{b}_1 & \tilde{b}_2 & \tilde{b}_3 & \tilde{b}_4 \end{array} \tag{3.14}$$

the idea is feasible.

For $\tilde{b}_1=\tilde{b}_4=1/6$ and $\tilde{b}_2=\tilde{b}_3=1/3$ we obtain the classical 4-th order scheme. Let us try $b_i=\gamma\tilde{b}_i$, where γ is a solution dependent parameter used to satisfy the condition $Q=0$. Substitution of these specific weights into (3.13) reveals that $Q=0$, if

$$\gamma = 1-\frac{\delta}{\eta}, \tag{3.15}$$

where

$$\eta = \|\sum_{i=1}^{4}\tilde{b}_iF_i\|^2, \quad \delta = \eta-\tilde{b}_2<F_1,F_2>-\tilde{b}_3<F_2,F_3>-2\tilde{b}_4<F_3,F_4>.$$

For a sensible application it is required that γ is close to one. Let us examine the asymptotic behaviour of γ as $\tau\rightarrow 0$. It is trivially seen that $\eta=O(\tau^2)$, since $F_i=\tau F(Y_i)$ and $\tilde{b}_1+...+\tilde{b}_4=1$. Next, after a tedious inspection, one can see that δ can be rewritten as

$$\delta = (\|F_1-F_2-F_3+F_4\|^2+3\|F_3-F_2\|^2+6<F_3-F_2,F_1-F_4>)/36.$$

By using this expression for δ it is not difficult to prove that $\delta=O(\tau^5)$, so that $\gamma=1+O(\tau^3)$ as $\tau\rightarrow 0$.

To sum up, the Butcher array

$$\begin{array}{c|cccc} 0 & 0 & & & \\ ½ & ½ & 0 & & \\ ½ & 0 & ½ & 0 & \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{\gamma}{6} & \frac{\gamma}{3} & \frac{\gamma}{3} & \frac{\gamma}{6} \end{array} \tag{3.16}$$

where γ is defined by (3.15), defines a 4-stage explicit Runge-Kutta scheme which is conservative for conservative problems and monotone for monotone problems, provided $\gamma>0$. Because the weights b_i are third order perturbations of the constant weights $\tilde{b}_i$, the scheme is of order 3 instead of order 4, at least if U^{n+1} is meant to approximate $U(t)$ at $t=t_n+\tau$. However, if U^{n+1} is interpreted as an approximation to $U(t)$ at

$$t=t_n+\gamma\tau, \tag{3.17}$$

the order of consistency remains four. This can be understood by examining the local error expansion of U^{n+1} in powers of $\gamma\tau$ rather than τ. Consequently, in applications it is preferable to use the variable step interpretation (3.17). □

We have shown how the classical explicit 4-th order Runge-Kutta method can be modified so as to make it conservative and monotone. Naturally, scheme (3.16) is still explicit implying that in actual computation one has to be careful in selecting τ. We recommend to select τ such that γ is close to one. The performance of the new scheme is then very much alike the performance of the classical 4-th order scheme, of course. What remains is that one has an *absolute guarantee* that the computation remains stable for nonlinear conservative or monotone problems. For problems with wave like phenomena where an explicit time-integration should be considered because of accuracy requirements, this stability option may be of interest. We briefly return to this point in Section 8.

Remark 3.9. In literature, Runge-Kutta formulas with solution dependent coefficients are called rational- or nonlinear formulas (Lambert [16], Wambecq [28], Hairer [12], Calvo & Quemada [3]). A conservative and monotone second order, 2-stage scheme, similar to (3.16), has been proposed in Verwer & Dekker [26]. That scheme is closely related to the specific scheme Hairer examines for parabolic problems. Sanz-Serna [23] has investigated a nonlinear modification of the explicit midpoint rule (leap frog). His modification preserves conservation, but is not monotone. Sanz-Serna reports illustrative experiments for the nonlinear Korteweg-de Vries equation. □

We should like to conclude the present section with two more definitions on step-by-step stability. Though we will not return to these definitions very often, they cannot be missed in a survey paper like this.

Definition 3.10. Let $\|\cdot\|$ be a given norm on $\mathbb{R}^m$. The integration method is called *C-stable* for (2.1), with respect to this norm, if a real number $\tau_0=\tau_0(\Delta)$ and a real constant C_0 exist, C_0 independent of Δ, such that

$$\|\tilde{U}^{n+1}-U^{n+1}\| \leq (1+C_0\tau)\|\tilde{U}^n - U^n\| \quad \text{for all } \tau\in(0,\tau_0]. \quad \square \tag{3.18}$$

If C_0 is positive, we allow an increase in the difference $\tilde{U}^n - U^n$. C-stability is an abbreviation for *"convergence stability"* and is in fact nothing else as stability in the Lax-Richtmyer sense (see Richtmyer and Morton [20]). In [5] the authors distinguish between "convergence stability" and "computing stability". The latter property is just contractivity. In applications, especially those where long time calculations are of importance, C-stability is normally not sufficient to keep the computation stable, though it suffices for proving convergence (see also Sanz-Serna [22]). C-stability is a minimal property to be imposed on any integration method for evolutionary problems in partial differential equations. If we suppose that problem (2.1) satisfies the one-sided Lipschitz condition (2.11) for some given ν, ν independent of Δ, any algebraically stable Runge-Kutta method which is also BSI-stable, is C-stable (see [5], Section 7.4).

Our last definition deals with the classical concepts of *absolute stability, A-stability,* and the like. Suppose that our semi-discrete problem (2.1) is of the constant coefficient type

$$\dot{U} = AU. \tag{3.19}$$

For this problem any one-step integration method can be written as

$$U^{n+1} = R(\tau A)U^n, \tag{3.20}$$

where $R(\tau A)$ is a specific matrix valued polynomial or rational function. The scalar function $R(z):\mathbb{C}\to\mathbb{C}$ is the familiar *stability function.* The method is called *absolutely stable* at $z\in\mathbb{C}$, if, for this z, $|R(z)|\leq 1$. If the method is absolutely stable for all $z\in\mathbb{C}$, $\mathrm{Re}(z)\leq 0$, it is called *A-stable.* If the coefficients of the integration method are real constants, we have

$$|R(\tau\lambda[A])|<1 \iff \sigma[R(\tau A)]<1, \tag{3.21}$$

where $\lambda[A]$ stands for an arbitrary eigenvalue of A and σ denotes the spectral radius. This equivalence relates the property of absolute stability with the *spectral condition* property for (3.20):

Definition 3.11. The integration method satisfies the *spectral condition* property if a real number $\tau_0=\tau_0(\Delta)$ exists such that $\sigma[R(\tau A)]<1$ for all stepsizes $\tau\in(0,\tau_0]$. □

It is emphasized that (3.21) may not be valid for nonlinear methods such as (3.16) (cf. Hairer [12]). If the spectral condition is satisfied it holds that

$$U^{n+1}\to \underline{0} \text{ as } n\to\infty.$$

Clearly, this is a significantly weaker type of stability than strict monotonicity, which requires $\|R(\tau A)\|<1$ for some norm. However, if A is a normal matrix, we have the well-known equality

$$\|R(\tau A)\|_2 = \sigma[R(\tau A)]. \tag{3.22}$$

In each of the following sections we will examine a specific partial differential equation and a corresponding approximation. Herewith we will concentrate on the step-by-step stability of this approximation so as to illustrate the application and use of the aforementioned stability concepts. Needless to say that the accuracy of the approximations should be considered too. In our paper, however, accuracy aspects will come up for discussion in an indirect way only. Further, we will exclusively use finites differences for the space-discretization.

4. A pseudo-linear parabolic problem

The first example we examine is the pseudo-linear parabolic equation

$$u_t = \frac{\partial}{\partial x}(d(t,x,u)u_x), \quad 0<x<1, \quad t>0, \tag{4.1}$$

where $d(t,x,u)$ is always strictly positive. If we define $d(t,x,u)=5u^4$, we obtain the nonlinear example problem used by Richtmyer & Morton [20], Section 8.6, for illustrating the heuristic approach of the linear stability theory for nonlinear problems. We impose homogeneous Dirichlet boundary conditions

$$u(0,t) = u(1,t)=0, \quad t>1, \tag{4.2}$$

and suppose that at $t=0$ an initial function $u(x,0)$, $0\leqslant x\leqslant 1$, has been defined.

To begin with we consider the integral expression

$$E(t) = \int_0^1 u^2(x,t)dx. \tag{4.3}$$

We have, using (4.1) and (4.2),

$$\tfrac{1}{2}\dot{E}(t) = \int_0^t u_t u dx = \int_0^1 u(du_x)_x dx = -\int_0^1 du_x^2 dx < 0. \tag{4.4}$$

E is often called the *energy integral.* The reason for this name is that the physical energy of the physical system underlying to (4.1) sometimes can be expressed in this way. We see that E monotonically decreases with evolution in time. We will require a similar *monotonicity* behaviour for the semi-discrete and fully discrete approximation, respectively. Hence the approach followed is based on ideas of the so called *energy method* (see Richtmyer & Morton [20], Section 6.2).

Let us derive the semi-discrete system which we base on finite differences. The interval $0\leqslant x\leqslant 1$ is divided into $m+1$ equal subintervals of length $\Delta x=(m+1)^{-1}$. On the resulting grid $\{x_j : x_j = j\Delta x, j=1(1)m\}$ we approximate $(du_x)_x$ by means of second order central differences. If we denote $U_j(t)\simeq u(x_j,t)$ to be the resulting time-continuous approximation at $x=x_j$, we thus have

$$(du_x)_x|_{x=x_j} \simeq (\Delta x)^{-2}[D_{j-1}U_{j-1}-(D_j+D_{j-1})U_j+D_jU_{j+1}], \tag{4.5}$$

where

$$D_j = \tfrac{1}{2}[d(t,x_j,U_j)+D(t,x_{j+1},U_{j+1})], \tag{4.6}$$

and $U_0(t)=U_{m+1}(t)=0$. It follows that the time-continuous grid function $U=[U_1,\ldots,U_m]^T$ satisfies the pseudo-linear semi-discrete system

$$\dot{U} = A(U)U, \tag{4.7}$$

where $A(U)$ is the symmetric, negative definite $m\times m$ matrix

$$A(U) = \frac{1}{(\Delta x)^2} \begin{bmatrix} -(D_1+D_0) & D_1 & & \\ D_1 & -(D_2+D_1) & D_2 & \\ \ddots & \ddots & \ddots & \\ D_{m-2} & -(D_{m-1}+D_{m-2}) & D_{m-1} & \\ & D_{m-1} & -(D_m+D_{m-1}) \end{bmatrix} . \qquad (4.8)$$

The negative definiteness follows from the fact that $A(U)$ is irreducibly diagonally dominant with negative diagonal entries (Varga [25], p. 23).

From Theorem 2.6 we can directly conclude that the resulting semi-discrete problem (4.7) is monotone in the usual l^p-norms for $p=1,2$ and ∞. Alternatively, for the l^2-norm $\|\zeta\|_2^2 = \Delta x <\zeta,\zeta>_2$, strict monotonicity can be directly concluded from

$$½\frac{d}{dt}\|U\|_2^2 = \Delta x <A(U)U,U>_2 <0. \qquad (4.9)$$

We also observe that

$$\|U(t)\|_2^2 = E(t)+O((\Delta x)^2),$$

which can be seen by approximating the integral expression with the trapezoidal rule on the x-grid and by replacing $u(x_j,t)$ by $U_j(t)$. Hence, $\|U(t)\|_2^2$ represents the *semi-discrete energy*. For the present problem, $\|\cdot\|_2$ is therefore sometimes called the *energy norm*. Consequently, in view of (4.4), monotonicity in l^2 is a very natural property here.

It is emphasized that in case we space-discretize on a non-equidistant x-grid monotonicity in l^2 can be proved in the same way. The l^2-norm then is to be defined by $\|\zeta\|^2 = \Delta x <D\zeta,\zeta>_2$, where Δx stands for the maximal grid distance and D is a positive diagonal matrix whose entries are related to the various grid distances (weights in the corresponding trapezoidal rule approximation for $E(t)$).

We next consider the time-integration of system (4.7) for the explicit Euler method

$$U^{n+1} = U^n + \tau A(U^n)U^n, \qquad (4.10)$$

and for the pseudo-linear forms (3.7') and (3.8') of implicit Euler and implicit midpoint, respectively. Since, $\|U^n\|_2^2$ represents the *fully-discrete energy*, it is natural to require monotonicity for the integration method as well (see Def. 3.4).

Monoticity for (3.7') and (3.8') — for all τ and all Δx — has already been established in Example (3.5), so there remains to examine (4.10). The explicit Euler rule (4.10) is monotone in l^2, if

$$\|I+\tau A(U^n)\|_2 \leq 1. \qquad (4.11)$$

Since $A = A^T$ we have $\|I+\tau A(U^n)\|_2 \leq \|I+\tau A(U^n)\|_\infty$, so that

$$\|I+\tau A(U^n)\|_2 \leq \max_j\{r(D_{j-1}+D_j)+|1-r(D_{j-1}+D_j)|\}, \qquad (4.12)$$

where $r=\tau/(\Delta x)^2$. It follows that explicit Euler is monotone in l^2, if

$$\tau \leq \frac{(\Delta x)^2}{\max_j(D_{j-1}+D_j)}. \qquad (4.13)$$

Notice that for the linear model problem $u_t = u_{xx}$, where $D_j = 1$, the classical condition $r \leq ½$ appears (see e.g. Richtmyer & Morton [20], p. 12, for an illustration of the Fourier series method of von Neumann).

Using the terminology of Definition 3.11, we can also say that explicit Euler satisfies the spectral condition — for problem (4.7) where $D_j = 1$ — if $r \leq ½$. This is a trivial consequence of the following observation. The eigenvalues $\lambda_j[A]$, A given by (4.8) where $D_j=1$, are

$$\lambda_j[A] = (m+1)^2[-2+2\cos(\frac{j\pi}{m+1})], \qquad (4.14)$$

so that

$$\sigma[R(\tau A)] = \sigma[I+\tau A] = |1+r[-2+2\cos(\frac{m\pi}{n+1})]|. \tag{4.15}$$

We can conclude that for the linear parabolic model problem, the spectral condition, or the condition of absolute stability, is just the condition for monotonicity in l^2. This equivalence is a consequence of the normality of A (cf. (3.22)). If A is not normal, the spectral condition may be quite misleading however. The linear hyperbolic model equation of the next section has been chosen with the aim of illustrating this failure of the spectral condition.

5. The hyperbolic model problem

Following Richtmyer & Morton [20], p. 151, we consider the simple hyperbolic initial-boundary value problem

$$\begin{aligned} u_t &= -u_x, \quad 0<x\leq 1, \ t>0, \\ u(0,t) &= 0, \quad t>0, \\ u(x,0) &= u^0(x), \quad 0\leq x\leq 1, \end{aligned} \tag{5.1}$$

u^0 being an initial function. Here, the interval [0,1] is divided into m equal subintervals of length $\Delta x = m^{-1}$. On the resulting x-grid $\{x_j : x_j = j\Delta x, j = 1(1)m\}$, we approximate u_x by the first order backward difference

$$u_x|_{x=x_j} \simeq (\Delta x)^{-1}(U_j - U_{j-1}), \tag{5.2}$$

where $U_j(t) \simeq u(x_j,t)$ again denotes the time-continuous approximation at x_j. Observe that $U_0(t) = \underline{0}$. The time-continuous grid function $U = [U_1,...,U_m]^T$ satisfies the semi-discrete problem

$$\dot{U} = AU, \tag{5.3}$$

where A is the constant bidiagonal $m \times m$ matrix

$$A = \frac{1}{\Delta x}\begin{bmatrix} -1 & & & & \\ 1 & -1 & & & \\ & 1 & -1 & & \\ & & \ddots & \ddots & \\ & & & 1 & -1 \end{bmatrix}. \tag{5.4}$$

Let us integrate (5.3) with the explicit and implicit Euler method. All eigenvalues $\lambda[A]$ are equal to $(-\Delta x)^{-1}$, so that

$$\begin{aligned} \lambda[I+\tau A] &= 1-\frac{\tau}{\Delta x}, \quad \text{(explicit Euler)} \\ \lambda[(I-\tau A)^{-1}] &= (1+\frac{\tau}{\Delta x})^{-1}, \quad \text{(implicit Euler)} \end{aligned} \tag{5.5}$$

Evidently, explicit Euler satisfies the spectral condition for system (5.3), if $\tau / \Delta x < 2$, while implicit Euler admits arbitrary values for τ.

We next examine monotonicity for both methods. Define

$$E(t) = \int_0^1 |u(x,t)| dx. \tag{5.6}$$

After defining $u^0(x)=0$ for $x<0$, we can write

$$E(t) = \int_t^1 |u^0(x-t)|dx = \int_0^1 |u^0(x-t)|dx, \tag{5.7}$$

from which we conclude that $E(t+\tau)\leqslant E(t)$ for all $t,\tau>0$, i.e. problem (5.1) is monotone in l^1. Hence in $\mathbb{R}^m$ we consider the l^1-norm

$$||\zeta||_1 = \Delta x \sum_{j=1}^{m} |\zeta_j|, \tag{5.8}$$

so that $||U(t)||_1 = E(t)+0(\Delta x)$ as $\Delta x \rightarrow 0$.

The expression (2.7) for the logarithmic matrix norm μ_1 trivially yields $\mu_1[A]=0$ for matrix (5.4), so that by virtue of Theorem 2.6 monotonicity for the semi-discrete problem (5.3) has been established. Implicit Euler preserves monotonicity in l^1 for all $\tau>0$, according to Example 3.6, so there remains to examine explicit Euler. We have

$$||I+\tau A||_1 = \frac{\tau}{\Delta x}+|1-\frac{\tau}{\Delta x}|=1, \text{ iff } \frac{\tau}{\Delta x}\leqslant 1. \tag{5.9}$$

This contrasts sharply with the spectral condition $\frac{\tau}{\Delta x}\leqslant 2$. In fact, the method of lines concept of absolute stability is rather dangerous to rely on. A sample calculation illustrating this point is shown in Table 5.1. For the two ratios $\frac{\tau}{\Delta x}=\frac{12}{13},\frac{3}{2}$ the entries in the table represent the maximum absolute error

$$\max_{j,n}|u(t_n,x_j)-U_j^n|, \; j=1,...,\frac{1}{\Delta x}; \; n=1,...,\frac{1}{\tau}, \tag{5.10}$$

for a range of Δx-values, where $u(0,x)=\sin(\pi x)$. Hence for both values of Δx we integrate till $t=1$.

Δx	$\frac{1}{12}$	$\frac{1}{24}$	$\frac{1}{48}$	$\frac{1}{96}$	$\frac{1}{192}$	$\frac{1}{384}$
$\frac{\tau}{\Delta x}=\frac{12}{13}$	8.9_{10}^{-2}	6.7_{10}^{-2}	4.9_{10}^{-2}	3.5_{10}^{-2}	2.5_{10}^{-2}	1.8_{10}^{-2}
$\frac{\tau}{\Delta x}=\frac{3}{2}$	6.1_{10}^{+0}	5.2_{10}^{+2}	1.2_{10}^{+7}	1.8_{10}^{+16}	1.1_{10}^{+35}	1.4_{10}^{+73}

Table 5.1

The entries of the last row — the absolutely stable calculation — clearly show that in a practical calculation with problems of the present type absolute stability is rather misleading. The explanation is simple. Strict absolute stability merely guarantees that $U^n \rightarrow \underline{0}$ as $n\rightarrow\infty$ for fixed τ and Δx. For finite n, however, error growth is not excluded by absolute stability in case of a large deviation of normality. Table 5.2, whose entries contain $\max_j |u(t_n,x_j)-U_j^n|$ for some values of n, nicely illustrate this. The observed error growth cannot occur in case of monotonicity.

n	1	10	64	192	320	448	512	640
$\Delta x=\frac{1}{96}, \frac{\tau}{\Delta x}=\frac{3}{2}$	1.6_{10}^{-2}	2.7_{10}^{+10}	1.8_{10}^{+16}	2.3_{10}^{+42}	4.1_{10}^{+30}	1.2_{10}^{+8}	9.4_{10}^{-6}	3.3_{10}^{-34}

Table 5.2

Remark 5.1. The failure of the spectral condition can also be directly explained from the well-known Courant-Friedrichs-Lewy condition for convergence, which says that the domain of dependence of the hyperbolic equation must be contained in the domain of dependence of the explicit difference scheme. If the *CFL* condition is violated, there will be no convergence as $\tau \to 0$, $\tau/\Delta x$ fixed (see e.g. Forsythe & Wasow [9], p. 25). For explicit Euler applied to (5.3) — standard forward-backward difference scheme for (5.1) — the *CFL* condition is just the l^1-monotonicity condition $\tau/\Delta x \leqslant 1$. In this connection we observe that in an analogous way monotonicity in l^∞ and l^2 can be shown under the same condition $\tau/\Delta x \leqslant 1$. □

By exploiting the general applicability of Dahlquist's Theorem 2.5 it is relatively simple to obtain results on monotonicity and contractivity for the nonlinear hyperbolic model problem

$$\begin{aligned} u_t &= -f(u)_x, \ 0<x\leqslant 1, \ t>0, \\ u(0,t) &= 0, \ t>0, \\ u(x,0) &= u^0(x), \ 0\leqslant x\leqslant 1, \end{aligned} \tag{5.11}$$

where $f'(u)>0$, all $u \in \mathbb{R}$, and $f(0)=0$. Nonlinear problems of this type (nonlinear conservation laws) have been the subject of numerous numerical studies, mainly in computational fluid dynamics. Even when $u^0(x)$ and $f(u)$ are smooth functions, typical solutions have discontinuities across curves which separate regions in which the solution is smooth. One-sided difference approximations have attractive properties to deal with such discontinuities (Engquist & Osher [6]).

Like for the linear problem (5.1), we space-discretize $f(u)_x$ with first order backward differences on a uniform grid to obtain

$$\begin{aligned} \dot{U}_1 &= -\frac{1}{\Delta x} f(U_1), \\ \dot{U}_j &= -\frac{1}{\Delta x}(f(U_j)-f(U_{j-1})), \ j=2(1)m. \end{aligned} \tag{5.12}$$

Denote $\dot{U}=F(U)$, where $U=[U_1,...,U_m]^T$. The Jacobian matrix $F'(U)$ is given by

$$F'(U) = \frac{1}{\Delta x}\begin{bmatrix} -f'(U_1) & & & \\ f'(U_1) & -f'(U_2) & & \\ & \bullet & \bullet & \\ & & \bullet \quad \bullet & \\ & & \bullet \quad \bullet & \\ & & f'(U_{m-1}) & -f'(U_m) \end{bmatrix}. \tag{5.13}$$

Because $f'(U_j)>0$, it follows that $\mu_1[F'(U)]=0$, where μ_1 corresponds to the l^1-norm (5.8). Consequently, the semi-discrete problem is dissipative in l^1 and also monotone since $F(\underline{0})=\underline{0}$. An easy calculation shows that explicit Euler is contractive and monotone in l^1, if

$$\frac{\tau}{\Delta x} \max_{U_j^n} f'(U_j^n) \leqslant 1. \tag{5.14}$$

Recall that implicit Euler preserves contractivity and monotonicity in l^1 for all $\tau>0$.

6. Diffusion-convection problems

The error growth phenomena discussed in the previous section are also observed in calculations arising from diffusion-convection problems with dominating convection terms. For such problems, however, one has to reckon with two sources of troubles, viz. the spectral condition gives misleading information, but also standard symmetrical space-differencing may yield large unwanted oscillations in the semi-discrete solution (see e.g. Siemieniuch & Gladwell [24], Morton [18], Mitchell & Griffiths [17], p. 258, and Griffiths, Christie & Mitchell [10]). It is instructive to see how one can prevent the trouble by requiring contractivity or monotonicity in combination with upwind space-differencing.

Following the aforementioned authors we consider the simple model

$$\begin{aligned} u_t &= u_{xx} - bu_x, \quad 0<x<1, \quad t>0, \\ u(0,t) &= 0, \quad u_x(1,t)=0, \quad t>0, \\ u(x,0) &= u^0(x), \end{aligned} \tag{6.1}$$

where the convection parameter b is positive and may be large. On the equidistant x-grid $\{x_j : x_j = j\Delta x, \Delta x = m^{-1}, j=1(1)m\}$ we discretize u_{xx} by means of standard second order central differences, while u_x is approximated by the generalized upwind difference quotient (cf. Griffiths et al. [10])

$$u_x|_{x=x_j} \simeq (2\Delta x)^{-1}[(1-w)U_{j+1} + 2wU_j - (1+w)U_{j-1}], \tag{6.2}$$

where $0 \leqslant w \leqslant 1$. The values $w=0$ and $w=1$ produce the central difference and backward difference (full upwinding) approximation, respectively. We thus obtain the semi-discrete system $\dot{U} = AU$, where A is the $m \times m$ matrix

$$A = \frac{1}{(\Delta x)^2} \begin{bmatrix} -2-2wq & 1-q(1-w) & & & \\ 1+q(1+w) & -2-2wq & 1-q(1-w) & & \\ & \ddots & \ddots & \ddots & \\ & & 1+q(1+w) & -2-2wq & 1-q(1-w) \\ & & & 2+2wq & -2-2wq \end{bmatrix}. \tag{6.3}$$

Here, the parameter $q = \frac{1}{2} b\Delta x$. Notice that the boundary condition $u_x(1,t)=0$ has been dealt with by substituting $U_{m+1} = U_{m-1}$ into (6.2).

As in Griffiths et al. [10], we consider the maximum norm

$$\|\zeta\|_\infty = \max_j |\zeta_j|, \zeta \in \mathbb{R}^m.$$

From the definition of μ_∞ we readily find

$$\mu_\infty[A] = \begin{matrix} 0 & , \; 0 \leqslant q(1-w) \leqslant 1, \\ 2q(1-w)-2 & , \; q(1-w)>1, \end{matrix} \tag{6.4}$$

which shows that the problem is monotone in l^∞, if and only if

$$0 \leqslant q(1-w) \leqslant 1. \tag{6.5}$$

If this condition is violated, one has to reckon with unwanted oscillations in the time-continuous solution $U(t)$ (see Siemieniuch & Gladwell [24]). Obviously, this imposes the restriction $q \leqslant 1$ for the central difference scheme for which $w=0$. By using upwinding, $0<w \leqslant 1$, it is always possible to obtain l^∞-monotone time-continuous solutions $U(t)$.

For further interesting details on the numerical solution of (6.1) the reader is referred to the aforementioned papers. Here we still observe that the l^1-norm and l^2-norm are less appropriate for the present example. For example, for $w=1$ the logarithmic norm $\mu_2[A]$ is always positive. Further,

$$\mu_1[A] = \max\{0, \pm[1-q(1-w)]\}. \tag{6.6}$$

7. The shallow water equations: conservative space differencing

A field of application where step-by-step stability is of significant practical importance is numerical fluid dynamics. Particularly in applications where a solution is sought over very *large time-intervals,* such as in long term numerical weather predictions (see e.g. Houghton, Kasahara & Washington [13]). Sections 7 and 8 are devoted to a nonlinear stability analysis of approximations for the two-space dimensional shallow water equations (the primitive equations for an incompressible, inviscid fluid with a free surface). We have chosen the shallow water equations as an example in our survey, since these equations are of direct practical importance and nonlinear instabilities frequently occur (see e.g. Arakawa [1], Houghton et al. [13], Sadourny [21], Gustafsson [11], and Fairweather & Navon [7]). Evidently, the nature of this survey set bounds to our discussion implying that accuracy will get much less attention than stability.

For the sake of comparison we will concentrate on the specific initial-boundary value problem studied by Gustafsson [11] and Fairweather & Navon [7]. As far as possible, we will also adopt their notation.

Define the vector function $w=[u,v,\phi]^T$, where each component function depends on the space coordinates x,y and on the time variable t, i.e. $w=w(x,y,t)$. Here u and v represent velocity components in the x- and y-direction, respectively, and $\phi=2\sqrt{gh}$ where h is the depth of the fluid and g is the acceleration of gravity. Our initial-boundary value problem has the form

$$w_t = A(w)w_x + B(w)w_y + C(y)w, \qquad (7.1)$$

$$0 \leqslant x \leqslant L, \; 0 \leqslant y \leqslant D, \; 0 \leqslant t$$

where

$$A = -\begin{bmatrix} u & 0 & \phi/2 \\ 0 & u & 0 \\ \phi/2 & 0 & u \end{bmatrix}, \; B = -\begin{bmatrix} v & 0 & 0 \\ 0 & v & \phi/2 \\ 0 & \phi/2 & v \end{bmatrix},$$

$$C = \begin{bmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; f=\hat{f}+\beta(y-D/2), \; \hat{f}, \beta \text{ const.} \qquad (7.2)$$

The parameter f represents the Coriolis force. We assume periodic solutions in the x-direction

$$w(x,y,t) = w(x+L,y,t). \qquad (7.3)$$

Then, with the boundary conditions

$$v(x,0,t) = v(x,D,t)=0, \qquad (7.4)$$

and initial function $w(x,y,0)$ given, the total energy

$$E = \tfrac{1}{2}\int_0^L\int_0^D (u^2+v^2+\frac{\phi^2}{4})\frac{\phi^2}{4g}dydx \qquad (7.5)$$

is independent of the time. So we are dealing with a well-posed problem. Note that no boundary conditions are necessary for u and ϕ at $y=0,D$. This conservation of energy property will be our guiding principle for the numerical stability analysis, i.e. we will analyse along the lines of the energy method.

Remark 7.1. In many applications the matrix C has the form

$$C = -\begin{bmatrix} \lambda & -f & 0 \\ f & \lambda & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; \lambda=\lambda(x,y,w)>0, \qquad (7.6)$$

where λ is normally very small. For example, in shallow water calculations λ represents bottom friction. If $\lambda>0$, the total energy $E(t)$ will monotonically decrease with evolution in time. Hence the bottom friction has a dissipative influence. In our discussion we simply take $\lambda=0$. All stability results for the case $\lambda=0$ trivially carry over to the case $\lambda>0$. We should also mention that in many applications equation (7.1) contains an inhomogeneous term representing external forces. Step-by-step stability, however, has to do with the homogeneous part of the equation, allowing us to omit the inhomogeneous part. □

The fully continuous problem (7.1) - (7.4) is conservative with respect to the total energy E. According to the approach of our nonlinear stability analysis this property should be maintained upon space-discretization, i.e. the space-discretization should lead to a semi-discrete problem which is conservative in some suitable energy norm. If not, we must reckon with *inherent instability* of the semi-discrete system. In such a situation any step-by-step integration over large time intervals may fail due to severe instability, i.e. *exponential blow up*. For the present problem (7.1) - (7.4) this situation indeed arises after space-discretizing with standard finite differences, as in Gustafsson [11] and Fairweather & Navon [7]. The following experiment serves to illustrate this nuisance which impedes many computations in practice.

Experiment 7.2. We first describe the space-discretization employed by Gustafsson and Fairweather & Navon. The x-interval and y-interval are divided into N_x and N_y subintervals of length Δx and Δy, respectively, i.e. $N_x \Delta x = L$ and $N_y \Delta_y = D$. On the grid

$$\{(x_j,y_k) : x_j = j\Delta x,\, j=1(1)N_x \text{ and } y_k = k\Delta y,\, k=0(1)N_y\}, \tag{7.7}$$

we define $W_{jk} = [U_{jk}, V_{jk}, \Phi_{jk}]^T$ as the time-continuous approximation for $w(x_j,y_k,t)$ resulting from the application of second order symmetrical differences at all interior points and first order one-sided differences at the boundary points (x_j,y_k), $k=0,N_y$. In the x-direction symmetrical differencing is possible everywhere because of the periodicity, i.e. $W_{0k} = W_{N_xk}$ and $W_{N_x+1,k} = W_{1k}$. Note that $V_{j0} = V_{jN_y} = 0$ due to (7.4).

Upon substitution of the finite difference approximations into equation (7.1) — for the grid points (x_j,y_k), $1 \leqslant j \leqslant N_x$ and $0 \leqslant k \leqslant N_y$ — the system of ordinary differential equations in the time-continuous grid function $W(t)$ is obtained. For describing this large system, $\dot{W} = F(W)$ say, it is convenient to introduce the splitting notation

$$\dot{W} = F(W) = F^{(1)}(W) + F^{(2)}(W) + F^{(3)}(W), \tag{7.8}$$

where $F^{(1)}, F^{(2)}$ and $F^{(3)}$ correspond to $A(w)w_x$, $B(w)w_y$ and $C(y)w$, respectively. Because the expressions $A(w)w_x$ and $A(w)w_y$ are one-space dimensional, all components of $F^{(1)}$ and $F^{(2)}$ are coupled only along horizontal and vertical grid lines. This means that we can describe these splitting functions per grid line. For $k=0(1)N_y$ we have

$$\begin{aligned} F^{(1)}_{1k}(W) &= A(W_{1k})\frac{W_{2k} - W_{N_x}k}{2\Delta x}, \\ F^{(1)}_{jk}(W) &= A(W_{jk})\frac{W_{j+1,k} - W_{j-1,k}}{2\Delta x}, \quad j=2(1)N_x-1, \\ F^{(1)}_{N_xk}(W) &= A(W_{N_xk})\frac{W_{1k} - W_{N_x-1,k}}{2\Delta x}. \end{aligned} \tag{7.9}$$

and for $j=1(1)N_x$

$$\begin{aligned} F^{(2)}_{j0}(W) &= B(W_{j0})\frac{W_{j1} - W_{j0}}{\Delta y}, \\ F^{(2)}_{jk}(W) &= B(W_{jk})\frac{W_{j,k+1} - W_{j,k-1}}{2\Delta y}, \quad k=1(1)N_y-1, \\ F^{(2)}_{jN_y}(W) &= B(W_{jN_y})\frac{W_{jN_y} - W_{j,N_y-1}}{\Delta y}. \end{aligned} \tag{7.10}$$

The components of $F^{(3)}$ are not coupled, so

$$F^{(3)}_{jk}(W) = C(y_k)W_{jk}, \quad j=1(1)N_x, \quad k=0(1)N_y. \tag{7.11}$$

Assembling equations (7.9)-(7.11) yields the semi-discrete shallow water equation (7.8).

Gustafsson and Fairweather & Navon have investigated alternating direction implicit (*ADI*) methods for the time-integration of (7.8). Among others, they have performed some long runs which always ended with an "exponential blow up". They lay the blame for these "explosions" on nonlinear instabilities in the scheme, but leave undecided whether it is a failure of the *ADI* method or not. We wish to point out that these "explosions" can easily occur, for any integration method, for the simple reason that (7.8) can be shown to possess solutions, at least locally, for which the trapezoidal rule approximation E_Δ of E increases in time. More precisely, initial grid functions W^0 exist, such that

$$E_\Delta(\tau) > E_\Delta(0), \text{ all } \Delta x, \Delta y, \tag{7.12}$$

for τ sufficiently small. Hence, (7.8) is neither conservative, nor monotone with respect to the semi-discrete total energy E_Δ.

The proof of inequality (7.12) is not very illuminating, so we omit it here. A weak point in our reasoning is that this negative result has a local meaning only. It gives no information on the global behaviour of $E_\Delta(t)$ for a given W^0. Yet we think that the lack of conservation, or of monotonicity, of E_Δ causes the "exponential blow ups" reported by Gustafsson and Fairweather & Navon. To support this view we have integrated system (7.8) with the 5-th order, explicit Runge-Kutta-Fehlberg code RKF45 (Fehlberg [8]). * This code performs stepsize and local error control as it is standard nowadays in the numerical solution of ordinary differential equations. We have applied it with the tolerance parameter TOL equal to 10^{-5}, using the problem parameters (cf. Gustafsson [11])

$$\begin{aligned} L &= 6.0\ 10^6\ m,\ D = 4.4\ 10^6\ m,\ \hat{f} = 10^{-4}\ sec^{-1,}\ \beta = 1.5_{10}^{-11}\ sec^{-1}\ m^{-1}, \\ g &= 10\ m\ sec^{-2},\ H_0 = 2000\ m,\ H_1 = 220\ m,\ H_2 = 133\ m, \\ \Delta x &= \Delta y = 200.000\ m. \end{aligned} \tag{7.13}$$

The parameters H_i occur in the initial height function

$$\begin{aligned} h(x,y) = {} & H_0 + H_1 \tanh(\frac{9(D/2-y)}{2D}) + \\ & H_2 \operatorname{sech}^2(\frac{9(D/2-y)}{D}) \sin\frac{2\pi x}{L}, \end{aligned} \tag{7.14}$$

while the initial velocities are defined by $u = -gf^{-1}\partial h / \partial y$, $v = gf^{-1}\partial h / \partial x$.

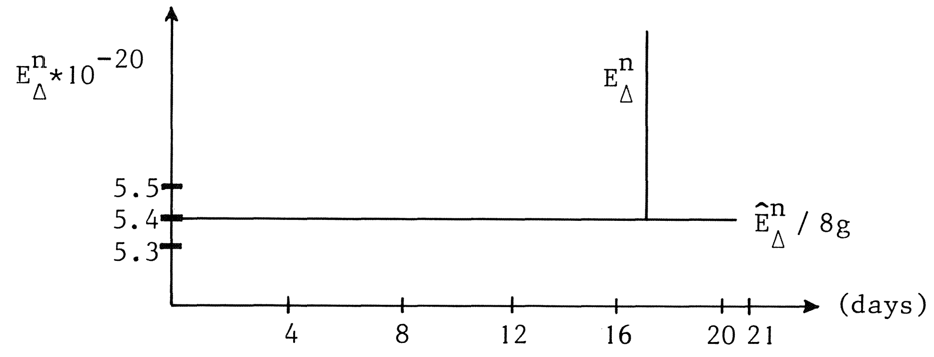

Fig. 7.1 Fully discrete total energies E_Δ^n and $\hat{E}_\Delta^n$ computed with *RKF45*.

* We have used the version implemented by Shampine and Watts in their code GERK (see TOMS 2, 172-186, 1976).

Figure 7.1 shows E_Δ^n, the fully discrete total energy. One can see that after approximately 17 days the solution suddenly blows up, despite the automatic stepsize control of *RKF*45. For stable problems the stepsize control can normally be trusted for preventing numerical instabilities. We therefore believe that inherent instability of the semi-discrete problem (7.8) underlies the "energy explosion". Up to the 17-th day the integration went smoothly with an almost constant stepsize of about 9 minutes. Some days before the "explosion" took place, the stepsize was gradually reduced by the control mechanism of *RKF*45. □

Gustafsson [11] and Fairweather & Navon [7] examine the use of artificial dissipative terms with the aim of suppressing the instabilities observed. The technique of artificial dissipation is rather ad hoc, unfortunately. A better approach, from the viewpoint of numerical stability, is to discretize in such a way that numerical stability can be *guaranteed.* In the remainder of this section we will describe how to realize this for the initial-boundary value problem (7.1)-(7.4).

In view of the constant total energy E, given by (7.5), our first task is to space-discretize problem (7.1)-(7.4) to a semi-discrete system which is conservative in an appropriate energy norm (cf. Definition 2.3). In other words, the squared norm of any semi-discrete solution should represent a constant semi-discrete total energy. Our derivation is based on a *tranformation of variables* which enables us to define a manageable norm from the integral (7.5). The new variable is $\hat{w} = [q, z, \psi]^T$, where

$$q = ½\phi u, \quad z = ½\phi v, \quad \psi = \frac{1}{4}\phi^2. \tag{7.15}$$

For these variables the integral (7.5) assumes the more attractive form

$$E = \frac{1}{8g}\int_0^L\int_0^D (q^2+z^2+\psi^2)dydx. \tag{7.16}$$

Let $\hat{W}$ denote the new time-continuous grid function which we define on the same grid as used before, viz. (7.7). Approximating E on this grid by means of the trapezoidal rule delivers (we omit the constant $(8g)^{-1}$)

$$\hat{E}_\Delta = \Delta x \Delta y \sum_{j=1}^{N_x}[\sum_{k=1}^{N_y-1} \hat{W}_{jk}^T \hat{W}_{jk} + ½(\hat{W}_{j0}^T \hat{W}_{j0} + \hat{W}_{jN_y}^T \hat{W}_{jN_y})]. \tag{7.17}$$

From this expression it directly follows that the scaled l^2-norm $\|\cdot\|$ corresponding to the inner product

$$<\zeta,\xi> = \Delta x \Delta y \sum_{j=1}^{N_x}[\sum_{k=1}^{N_y-1} \zeta_{jk}^T \xi_{jk} + ½(\zeta_{j0}^T \xi_{j0} + \zeta_{jN_y}^T \xi_{jN_y})], \tag{7.18}$$

may be a suitable energy norm, since $\|\hat{W}\|^2 = \hat{E}_\Delta$.

Now we have chosen a norm, the next step in the derivation is to define a semi-discrete approximation for the transformed problem such that it is *conservative* in the scaled l^2-norm $\|\cdot\|$. For convenience of presentation we will discuss this part of the derivation for only one space dimension.

Example 7.3. Denote $w = [v, \phi]^T$. We consider

$$w_t = B(w)w_y, \quad 0 \leqslant y \leqslant D, \quad t \geqslant 0, \tag{7.19}$$

$$B(w) = -\begin{bmatrix} v & ½\phi \\ ½\phi & v \end{bmatrix},$$

where v and ϕ do have the same meaning as in equation (7.1). For v we impose $v(0,t) = v(D,t) = 0$, $t > 0$, while no condition is imposed for ϕ. Using the new variables z and ψ (cf. (7.15)), equation (7.19) is rewritten as

$$z_t = -\frac{3}{2}z\psi^{-1/2}z_y + \frac{1}{4}\psi^{-\frac{3}{2}}z^2\psi_y - \psi^{1/2}\psi_y, \tag{7.20}$$
$$\psi_t = -\psi^{1/2}z_y - 1/2 z\psi^{-1/2}\psi_y.$$

Omitting the constant factor, E and $\hat{E}_\Delta$ are given by

$$E = \int_0^D (z^2+\psi^2)dy, \tag{7.21}$$

and

$$\hat{E}_\Delta = \Delta y[\sum_{k=1}^{N_y-1}(Z_k^2+\Psi_k^2)+1/2\Psi_0^2+1/2\Psi_{N_y}^2]. \tag{7.22}$$

As outlined above, the task we have set ourselves is to determine a semi-discrete approximation for (7.20), in the dependent variables Z_k and Ψ_k, such that $\hat{E}_\Delta$ is independent of time. For that purpose we examine

$$1/2\dot{E} = \int_0^D (zz_t+\psi\psi_t)dy = 0. \tag{7.23}$$

The clue here is to put together those terms from (7.20) for which the contribution to $\dot{E}$ is equal to zero, and to difference these collected terms in such a way that their contribution to $d\hat{E}_\Delta / dt$ is zero too. So we compute

$$1/2\dot{E} = -\int_0^D [(\frac{3}{2}z^2\psi^{-1/2}z_y - \frac{1}{4}z^3\psi^{-\frac{3}{2}}\psi_y)+(\frac{3}{2}z\psi^{1/2}\psi_y+\psi^{\frac{3}{2}}z_y)]dy \tag{7.24}$$
$$= -\int_0^D [(1/2 z^3\psi^{-1/2})_y+(z\psi^{\frac{3}{2}})_y]dy = -1/2 z^3\psi^{-1/2} - z\psi^{\frac{3}{2}}\Big|_{y=0}^{y=D}.$$

The first bracketed term corresponds to the first two terms for z_t, while the second bracketed term corresponds to the remaining terms of (7.20). Clearly, both these bracketed terms vanish in (7.24).

Let us consider the first two terms for z_t. To obtain *equal coefficients*, we rewrite the expression to

$$d = -1/2 z\psi^{-1/2}z_y - 1/2(z^2\psi^{-1/2})_y. \tag{7.25}$$

We next approximate d on the equidistant y-grid $\{y_k\}$ by second order symmetrical differences, i.e. for $k=1(1)N_y-1$ we define

$$D_k = \frac{-1}{4\Delta y}[Z_k\Psi_k^{-1/2}(Z_{k+1}-Z_{k-1})+(Z_{k+1}^2\Psi_{k+1}^{-1/2}-Z_{k-1}^2\Psi_{k-1}^{-1/2})]. \tag{7.26}$$

Recall that $Z_0=Z_{N_y}=0$. From a trivial calculation one now can see that the contribution of (7.26) to $d\hat{E}_\Delta / dt$, given by

$$2\Delta y \sum_{k=1}^{N_y-1} Z_k D_k, \tag{7.27}$$

is precisely equal to zero.

Let us consider the remaining terms of (7.20), viz. $-\psi^{1/2}\psi_y$ and the expression for ψ_t. The latter is rewritten to

$$\psi_t = -(\psi^{1/2}z)_y. \tag{7.28}$$

For $k=1(1)N_y-1$ we again approximate both expressions with second order symmetrical differences, while for $k=0$, N_y first order one-sided differences are used. Like for (7.26), an elementary calculation then reveals that the contribution to $d\hat{E}_n \,/\, dt$ of the corresponding semi-discrete expressions is precisely equal to zero.

To sum up, we first transform equation (7.19) to (7.20) and rewrite the latter to the equivalent form

$$\begin{aligned} z_t &= -\tfrac{1}{2} z \psi^{-\frac{1}{2}} z_y - \tfrac{1}{2}(z^2 \psi^{-\frac{1}{2}})_y - \psi^{\frac{1}{2}} \psi_y . \\ \psi_t &= (\psi^{\frac{1}{2}} z)_y . \end{aligned} \tag{7.29}$$

This equation then is space-discretized to

$$\begin{aligned} \dot{\Psi}_0 &= -\frac{1}{\Delta_y} \Psi_1^{\frac{1}{2}} Z_1, \\ \dot{Z}_k &= D_k - \frac{1}{2\Delta y} \Psi_k^{\frac{1}{2}} (\Psi_{k+1} - \Psi_{k-1}), \;\; k = 1(1)N_y - 1, \\ \dot{\Psi}_k &= -\frac{1}{2\Delta y} (\Psi_{k+1}^{\frac{1}{2}} Z_{k+1} - \Psi_{k-1}^{\frac{1}{2}} Z_{k-1}), \;\; k = 1(1)N_y - 1, \\ \dot{\Psi}_{N_y} &= \frac{1}{\Delta y} \Psi_{N_y - 1}^{\frac{1}{2}} Z_{N_y - 1}, \end{aligned} \tag{7.30}$$

where D_k is given by (7.26). Problem (7.30) is conservative (cf. Def. 2.3) in the energy norm corresponding to $\hat{E}_\Delta$ given by (7.22). □

We return to our discussion of the two-space dimensional flow problem. The derivation of the energy conserving semi-discrete approximation is completely analogous to the derivation for the one-space dimensional problem (7.19). First one should transform equation (7.1), according to (7.15). Then the transformed equations should be rewritten to a form similar as (7.29). After some examination, one can deduce that this specific form is given by

$$\hat{w}_t = X(\hat{w}) + Y(\hat{w}) + C(y)\hat{w}, \tag{7.31}$$

where $C(y)$ is the same matrix as in equation (7.1), and where

$$X(\hat{w}) = -\begin{bmatrix} \tfrac{1}{2}\psi^{-\frac{1}{2}} q q_x + \tfrac{1}{2}(\psi^{-\frac{1}{2}} q^2)_x + \psi^{\frac{1}{2}} \psi_x \\ \tfrac{1}{2}\psi^{-\frac{1}{2}} q z_x + \tfrac{1}{2}(\psi^{-\frac{1}{2}} q z)_x \\ (\psi^{\frac{1}{2}} q)_x \end{bmatrix}, \tag{7.32}$$

$$Y(\hat{w}) = -\begin{bmatrix} \tfrac{1}{2} z \psi^{-\frac{1}{2}} q_y + \tfrac{1}{2}(q z \psi^{-\frac{1}{2}})_y \\ \tfrac{1}{2} z \psi^{-\frac{1}{2}} z_y + \tfrac{1}{2}(z^2 \psi^{-\frac{1}{2}})_y + \psi^{\frac{1}{2}} \psi_y \\ (\psi^{\frac{1}{2}} z)_y \end{bmatrix}. \tag{7.33}$$

Standard space-differencing of this partial differential equation yields a semi-discrete system — in the dependent vector variable $\hat{W}$ — which can be proved to be *conservative* in the scaled l^2-norm corresponding to (7.18). For future reference, we denote this system by

$$\frac{d\hat{W}}{dt} = \hat{F}(\hat{W}). \tag{7.34}$$

In order to save space its actual formulation must be omitted. For the same reason we do not write down the proof of its conservation property.

Experiment 7.4. By way of comparison we have integrated system (7.34) with the Fehlberg code *RKF*45 in exactly the same way as the non-conservative system (7.8). The resulting fully discrete total energy $\hat{E}_\Delta^n$ has been computed over a period of 21 days (see Fig. 7.1). One can see that *RKF*45 keeps the energy nearly constant. More precisely, after 21 days the relative difference with the initial energy is equal to 10^{-7}, which is negligible. Notice that the explicit 5-th order Fehlberg formula in *RKF*45 is not conservative. □

Remark 7.5. A known idea in the numerical solution of the shallow water equations is the use of so-called space-staggered grids. Herewith the semi-discrete variables are defined on the grid in a staggered way, rather than on the entire grid. The aim is to reduce the computational costs by defining a minimal number of variables. Following a different approach than ours, Sadourny [21] describes an energy conserving space-disretization of the shallow water equations on such a grid. We wish to remark that the approach of this section is also applicable to space-staggered grids, of course at the cost of more complicated derivations. □

8. The shallow water equations: conservative time integration

In view of the existence of a conservative semi-discrete shallow water equation it is natural to require conservation in the time-integration as well, in order to guarantee step-by-step stability when proceeding in time. In the last section of this paper we will briefly embark upon two *conservative methods* for system (7.34).

We recall that in practice the shallow water equation normally contains a small dissipative term (see Remark 7.1). In that case our system (7.34) will become strictly monotone, though the conservative part will largely dominate. The time-integration method is therefore required to preserve not only conservation, but also monotonicity.

The first method is based on the *explicit rational Runge-Kutta formula* (3.16). This peculiar method can be made conservative and monotone by a special choice of the parameter γ. Praagman [19] has investigated the use of the classical 4-th order formula when combined with a finite element space-discretization. Obviously, our rational mode can also be combined with the finite element approach. We note that a sensible application of formula (3.16) is only possible for γ close to one, which implies a restriction on τ. Though the method is conservative for all τ, its explicitness will be charged in some way or another. In that respect the method is not different from the classical formula investigated by Praagman [19]. For other explicit methods for shallow water equations the interested reader is referred to Van der Houwen [14, 15].

The second method we have in mind is a *locally one-dimensional splitting method* (*LOD* method) based on the application of the *pseudo-linear midpoint rule* (3.8'). From the viewpoint of numerical stability we consider this method as an attractive alternative for the *ADI* methods of Gustafsson [11] and Fairweather & Navon [7]. Further, by using (3.8'), the complications of solving nonlinear systems of algebraic equations is avoided. Below we will give an outline of this second method.

Consider system (7.34). Assume, for convenience of discussion, that $C(y)$ is the zero matrix. We write

$$\frac{d\hat{W}}{dt} = \hat{F}(\hat{W}) = \hat{F}^{(1)}(\hat{W}) + \hat{F}^{(2)}(\hat{W}), \tag{8.1}$$

where $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ correspond to $X(\hat{w})$ and $Y(\hat{w})$, respectively (cf. (7.8)). Clearly, components of $\hat{F}^{(1)}$ are coupled only along horizontal grid lines. Likewise, components of $\hat{F}^{(2)}$ are coupled only along vertical grid lines. This one-space dimensional form of $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ allows the application of splitting methods such as *ADI* and *LOD*.

Suppose, for the time being, that $\hat{F}^{(i)}$ can be written in the pseudo-linear form (cf. (2.15))

$$\hat{F}^{(i)}(\hat{W}) = P^{(i)}(\hat{W})\hat{W}, \quad i=1,2, \tag{8.2}$$

such that for the inner product (7.18) it holds that

$$<P^{(i)}(\tilde{\zeta})\zeta, \zeta> = 0, \quad \text{all } \zeta, \tilde{\zeta}. \tag{8.3}$$

Next consider the following *LOD* integration method based on the pseudo-linear midpoint rule (3.8'):

$$\hat{W}^{*} = \hat{W}^{n} + \tau P^{(1)}(\hat{W}^{n})\frac{\hat{W}^{(n)} + \hat{W}^{*}}{2}, \tag{8.4}$$
$$\hat{W}^{n+1} = \hat{W}^{*} + \tau P^{(2)}(\hat{W}^{n})\frac{\hat{W}^{*} + \hat{W}^{n+1}}{2}.$$

One complete step $\hat{W}^{n} \rightarrow \hat{W}^{n+1}$ consists of two consecutive one-dimensional steps. By virtue of (8.3), and because of the conservation property of (3.8'), it follows that the *LOD* method is conservative, viz.

$$\|\hat{W}^{n+1}\| = \|\hat{W}^{*}\| = \|\hat{W}^{n}\|. \tag{8.5}$$

We also note that (8.4) preserves monotonicity. Further, the term $C(y)\hat{w}$ can be included without any essential modification. For computational purposes it is convenient to write (8.4) into the equivalent form

$$\hat{W}^{**} = \hat{W}^{n} + ½\tau P^{(1)}(\hat{W}^{n})\hat{W}^{**}, \quad \hat{W}^{*} = 2\hat{W}^{**} - \hat{W}^{n}, \tag{8.6}$$
$$\hat{W}^{n+1,*} = \hat{W}^{*} + ½\tau P^{(2)}(\hat{W}^{n})\hat{W}^{n+1,*}, \quad \hat{W}^{n+1} = 2\hat{W}^{n+1,*} - \hat{W}^{*}.$$

This notation avoids the matrix vector operation in (8.4).

The practical value of this conservative *LOD* method must still be established of course. However, the method is based on a sound basis. Step-by-step stability is guaranteed for all $\tau > 0$, despite the non-linearities in the shallow water equation. Per integration step the computational costs are not very large. For each grid line one has to make up a block tridiagonal matrix and, subsequently, solve the corresponding system of linear algebraic equations. The storage requirements are also very modest. The order of accuracy in time is equal to one, which may be insufficient. If so, one possibly can improve the accuracy by global extrapolation. This classical technique has no interference with the stability (Verwer & de Vries [27]).

There remains to prove our assumption concerning the existence of the two matrices $P^{(1)}$ and $P^{(2)}$ satisfying (8.3). For this proof one needs of course the actual formulation of $\hat{F}$ which we have omitted in order to save space. For that reason we will again resort to the one-dimensional flow problem (7.19) and present the derivation of the conservative pseudo-linear form for system (7.30). The derivation for the functions $\hat{F}^{(1)}$ and $\hat{F}^{(2)}$ goes along the same lines, though it is somewhat more cumbersome.

Example 8.1. Consider system (7.30). Let us denote its dependent vector variable by

$$W = [Z_0, \Psi_0 \ \cdots \ Z_k, \Psi_k \ \cdots \ Z_{N_y}, \Psi_{N_y}]^{T},$$

where $Z_0 = Z_{N_y} = 0$. The inner product we are dealing with (see (7.22)) is given by

$$<\zeta,\xi> = \Delta y <G\zeta,\xi>_2, \quad G = \mathrm{diag}(½,½,1,1, \ \cdots \ 1,1,½,½). \tag{8.7}$$

Our task is to write system (7.30) in the pseudo-linear notation, $\dot{W} = P(W)W$ say, where $P(W)$ is of the 2-block tridiagonal form

$$P = \frac{-1}{2\Delta y}\begin{bmatrix} 0 & p_{01} & & & \\ p_{10} & 0 & p_{12} & & \\ & \ddots & \ddots & \ddots & \\ & & p_{N_y-1,N_y-2} & 0 & p_{N_y-1,N_y} \\ & & & p_{N_y,N_y-1} & 0 \end{bmatrix}, \tag{8.8}$$

and where $<P(\tilde{\zeta})\zeta,\zeta> = 0$ for all $\zeta,\tilde{\zeta}$. A judicious inspection of (7.30) leads us to

$$p_{k,k-1} = \begin{bmatrix} -½(\Psi_{k-1}^{-½}Z_{k-1} + \Psi_{k}^{-½}Z_{k}) & -\Psi_{k}^{½} \\ -\Psi_{k-1}^{½} & 0 \end{bmatrix}, \tag{8.9}$$

$$p_{k,k+1} = \begin{bmatrix} \frac{1}{2}(\Psi_{k+1}^{-\frac{1}{2}} Z_{k+1} + \Psi_k^{-\frac{1}{2}} Z_k) & \Psi_k^{\frac{1}{2}} \\ \Psi_{k+1}^{\frac{1}{2}} & 0 \end{bmatrix}, \tag{8.10}$$

for $k=1(1)N_y-1$, while

$$p_{01} = 2\begin{bmatrix} 0 & 0 \\ \Psi_1^{\frac{1}{2}} & 0 \end{bmatrix}, \quad p_{N_y,N_y-1} = 2\begin{bmatrix} 0 & 0 \\ -\Psi_{N_y-1}^{\frac{1}{2}} & 0 \end{bmatrix}. \tag{8.11}$$

We see that $p_{k,k-1} = -p_{k-1,k}^T$ for $k=2(1)N_y-1$. Further, since $Z_0=0$, the first column of p_{10} may be chosen zero, so that $p_{01} = -2p_{10}^T$. Likewise, we get $p_{N_y,N_y-1} = -2p_{N_y-1,N_y}$. It thus follows that

$$<P(\tilde{\zeta})\zeta,\zeta> \; = \Delta y <[GP(\tilde{\zeta}) + P^T(\tilde{\zeta})G]\zeta,\zeta>_2 = 0 \tag{8.12}$$

for arbitrary grid functions ζ and $\tilde{\zeta}$. □

Acknowledgement. We gratefully acknowledge the assistance of Joke Blom in programming the Experiments 7.2 and 7.4.

References

[1] Arakawa, A., *Computational Design for Long-Term Numerical Integration of the Equations of Fluid Motion: Two Dimensional Incompressible Flow. Part I,* J. of Comp. Physics 1, 119-143, 1966.

[2] Burrage, K & J.C. Butcher, *Stability Criteria for Implicit Runge-Kutta Methods,* SIAM J. Numer. Anal. 16, 46-57, 1979.

[3] Calvo, M. & M. Mar Quemada, *On the stability of rational Runge-Kutta methods,* J. Comp. Appl. Math. 8, 289-293, 1982.

[4] Dahlquist, G., *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations (Thesis),* Transactions of the Royal Institute of Technology, No. 130, Stockholm, 1959.

[5] Dekker, K. & J.G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations,* A forthcoming monograph.

[6] Engquist, B. & S. Osher, *One-sided Difference Approximations for Nonlinear Conservation laws,* Math. Comp. 36, 321-351, 1981.

[7] Fairweather, G. & I.M. Navon, *A Linear ADI Method for the Shallow Water Equations,* J. of Comp. Physics 37, 1-18, 1980.

[8] Fehlberg, E., *Low order Classical Runge-Kutta Formulas with Stepsize Control and their Application to some Heat Transfer Problems,* NASA Tech. Rep. TR R-315, George C. Marshall Space Flight Center, Marshall, Ala, 1969.

[9] Forsythe, G.E. & W.R. Wasow, *Finite Difference Methods for Partial Differential Equations,* John Wiley & Sons, New York, 1960.

[10] Griffiths, D.F., I. Christie & A.R. Mitchell, *Analysis of Error Growth for Explicit Difference Schemes in Conduction-Convection Problems,* Report NA/29, University of Dundee, 1978.

[11] Gustafsson, B., *An Alternating Direction Implicit Method for Solving the Shallow Water Equations,* J. of Comp. Physics 7, 239-254, 1971.

[12] Hairer, E., *Unconditionally Stable Explicit Methods for Parabolic Equations,* Numer. Math. 35, 57-68, 1980.

[13] Houghton D., A. Kasahara & W. Washington, *Long-term Integration of the Barotropic Equations*

by the Lax-Wendroff Method, Mon. Weather Rev. 94, 141-150, 1966.

[14] Houwen, P.J. van der, *Finite Difference Methods for Solving Partial Differential Equations (Thesis),* MC Tract 20, Mathematical Centre, Amsterdam, 1968.

[15] Houwen, P.J. van der, *Berekening van Waterstanden in Zeeën en Rivieren* (Dutch), MC Syllabus 33, Mathematical Centre, Amsterdam, 1977.

[16] Lambert, J.D. *Two unconventional classes of methods for stiff systems,* **in** Stiff Differential Systems, R.A. Willoughby (ed.), Plenum Press, New York, 171-186, 1974.

[17] Mitchell, A.R. & D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations,* John Wiley & Sons, Chichester, 1980.

[18] Morton, K.W., *Stability of Finite Difference Approximations to a Diffusion-Convection Equation,* Int. J. Num. Meth. Engng 15, 677-683, 1980.

[19] Praagman, N., *Numerical Solution of the Shallow Water Equations by a Finite Element Method (Thesis),* Technical University of Delft, The Netherlands, 1979.

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

[21] Sadourny, R., *The Dynamics of Finite Difference Models of The Shallow Water Equations,* J. of the Atmos. Sci. 32, 680-689, 1975.

[22] Sanz-Serna, J.M., *Convergent Approximations to Partial Differential Equations and Stability Concepts of Methods for Stiff Systems of Ordinary Differential Equations,* submitted.

[23] Sanz-Serna, J.M., *An Explicit Finite Difference Scheme with Exact Conservation Properties,* J. of Comp. Physics 47, 199-210, 1982.

[24] Siemieniuch, J.L. & I. Gladwell, *Analysis of Explicit Difference Methods for a Diffusion-Convection Equation,* Int. J. Num. Meth. Engng 12, 899-916, 1978.

[25] Varga, R.S., *Matrix Iterative Analysis,* Prentice-Hall, Englewood Cliffs, New Yersey, 1962.

[26] Verwer, J.G. & K. Dekker, *Stability in the Method of Lines,* **in** Proceedings of the seminar "Numerische Behandlung von Differentialgleichungen", ed. K. Strehmel, Martin Luther Universität Halle, to appear.

[27] Verwer, J.G. & H.B. de Vries, *Global Extrapolation of a First Order Splitting Method,* Report NW 150/83, Mathematical Centre, Amsterdam, 1983.

[28] Wambecq, A., *Rational Runge-Kutta methods for solving systems of ordinary differential equations,* Computing 20, 333-342, 1978.