Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

P.J. van der Houwen, B.P. Sommeijer

High order difference schemes with reduced dispersion
for hyperbolic differential equations

Department of Numerical Mathematics Report NM-R8408 July

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

HIGH ORDER DIFFERENCE SCHEMES WITH REDUCED DISPERSION FOR HYPERBOLIC DIFFERENTIAL EQUATIONS

P.J. VAN DER HOUWEN, B.P. SOMMEIJER
Centre for Mathematics and Computer Science, Amsterdam

We investigate difference schemes for systems of first order hyperbolic differential equations in two space dimensions, possessing the following characteristics:

(i) The spatial discretizations are fourth order accurate.
(ii) The time discretization is of explicit Runge-Kutta type and is also fourth order accurate.
(iii) The scaled stability boundary is approximately $\sqrt{2}/2$.
(iv) The weights in the space discretizations and the Runge-Kutta parameters can be adapted in order to reduce the dispersion of the dominant Fourier components in the solution.

This method is illustrated by applying it to the shallow water equations simulating the motion of water in a shallow sea due to tidal forces. Since in such problems the dominant frequencies in the solution are known in advance, the method can take fully advantage of the possibility to tune the various parameters to these dominant frequencies.

1980 MATHEMATICS SUBJECT CLASSIFICATION: Primary: 65M20, Secondary: 76B15.
KEY WORDS & PHRASES: hyperbolic equations, difference schemes, Runge-Kutta methods, dispersion.
NOTE: This report will be submitted for publication elsewhere.

Report NM-R8408
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. INTRODUCTION

Many practical problems in fluid dynamics are modelled by the quasi-linear system

$$(1.1) \qquad \frac{\partial \vec{w}}{\partial t} = A \frac{\partial}{\partial x} \vec{w} + B \frac{\partial}{\partial y} \vec{w}, \qquad (x,y) \in \Omega$$

where $A = A(\vec{w})$ and $B = B(\vec{w})$ are symmetric matrices. Equation (1.1) represents a hyperbolic system. We will assume that the initial condition at t_0 and the boundary conditions on $\partial\Omega$ determine a unique solution.

Replacing the space derivatives in (1.1) by symmetric, finite differences, we obtain a system of ordinary differential equations (ODEs) the Jacobian of which possesses *purely imaginary* eigenvalues. This system of ODEs may count several thousands of equations.

In order to integrate such systems we need an ODE solver with *low storage requirements* and a *large imaginary stability boundary, relative to the spatial mesh used in the discretization of* (1.1).

The low storage requirement excludes fully implicit methods such as implicit Runge-Kutta methods and implicit linear multistep methods (particularly when the number of back values is large). This leaves us with explicit Runge-Kutta methods, linear multistep methods using only a few back values, and the class of splitting methods (e.g. ADI or LOD).

In this paper we concentrate on *explicit Runge-Kutta methods*, mainly because such methods turn out to be economic on vector computers which have recently become available for solving large scale problems. In particular, we will study the class of four-stage, second order Runge-Kutta methods. A drawback of such explicit methods is the stability condition, in our case of the form $\Delta t \leq c\Delta x$, c = constant, which restricts the integration step Δt more severely than necessary for accuracy. To overcome this inconvenient restriction we will use fourth order space discretizations allowing us to use large Δx-values and consequently larger Δt-values.

The main purpose of this paper is to investigate how the dispersion of difference schemes employing Runge-Kutta methods can be reduced by applying *exponential fitting techniques*. Exponential fitting goes back to GAUTSCHI [1] who used the technique for solving periodic problems with given

frequency of the solution. LINIGER and WILLOUGHBY [6], who introduced the terminology "exponential fitting", applied this technique for solving stiff problems. Both Gautschi and Liniger-Willoughby considered linear multistep methods. Here, we will consider exponentially fitted Runge-Kutta methods for the solution of hyperbolic problems.

The difference schemes derived here, are formally fourth-order accurate both in space and, by virtue of the exponential fitting, also in time. In fact, asymptotically (as $\Delta t \to 0$), the difference schemes converge to a scheme consisting of the standard fourth-order spatial discretization and the standard fourth-order Runge-Kutta time discretization. However, the dispersion of our schemes is much lower than that of this conventional scheme, particularly when we can provide estimates of the space and time frequencies of the dominant Fourier components occurring in the solution.

The stability behaviour of our difference scheme strongly resembles that of the standard fourth-order Runge-Kutta method, that is the imaginary stability boundary equals $2\sqrt{2}$. Effectively, taking into account the number of stages per step, we have $\sqrt{2}/2$. Let us compare this value with the maximal attainable effective, imaginary stability boundary within the class of all explicit Runge-Kutta and linear multistep methods: in [2,4] it was shown that for both classes the effective, imaginary stability boundary can never exceed 1. Thus, our difference scheme covers already 70% of the limit value.

In our numerical experiments we used the standard fourth-order RK method, the exponentially fitted method and the widely used Lax-Wendroff scheme. The experiments were performed on relatively long t-intervals in order to illustrate the dispersive behaviour of these methods. The exponentially fitted method turned out to be markedly less dispersive than the other ones.

2. THE CLASS OF DIFFERENCE SCHEMES

Following the method of lines approach we first replace the spatial differential operators $\partial/\partial x$ and $\partial/\partial y$ in (1.1) by finite difference operators D_x and D_y on a uniform grid $\Omega_\Delta := \{(j\Delta x, \ell\Delta y)\}_{j,\ell}$. We will restrict our analysis to space-centered approximations of the form

$$(2.1) \qquad D_x w(x,y) := \frac{1}{2\Delta x} \sum_{\ell=0}^{k} \sum_{j=1}^{k} \zeta_j^{(\ell)} (E_x^j - E_x^{-j})(E_y^\ell + E_y^{-\ell}) w(x,y)$$

$$(2.1)\qquad D_y w(x,y) := \frac{1}{2\Delta y} \sum_{j=0}^{k} \sum_{\ell=1}^{k} \zeta_\ell^{(j)} (E_y^{\ell} - E_y^{-\ell})(E_x^{j} + E_x^{-j}) w(x,y)$$

where E_x and E_y are shift operators defined by $E_x v(x) = v(x+\Delta x), E_y v(y) = v(y+\Delta y)$ and where the $\zeta_j^{(\ell)}$ are parameters determining the accuracy of the approximation.

Let $\vec{W} = (\vec{W}_{j\ell})$ be a grid function on the grid Ω_Δ. Then (1.1) can be approximated by the system of ODEs

$$(2.2)\qquad \frac{d\vec{W}_{j\ell}}{dt} = [A(\vec{W})D_x + B(\vec{W})D_y]\vec{W}_{j\ell}, \quad (j\Delta x, \ell\Delta y) \in \Omega_\Delta.$$

We solve this system by an explicit Runge-Kutta method. Writing (2.2) in the compact form

$$(2.3)\qquad \frac{d\vec{W}(t)}{dt} = \vec{F}_\Delta(\vec{W}(t)),$$

the general explicit, m-stage Runge-Kutta method is given by

$$\vec{W}^{(i)} = \vec{W}^n + \Delta t \sum_{q=1}^{i-1} a_{iq} \vec{F}_\Delta(\vec{W}^{(q)}), \qquad i = 1,\ldots,m,$$

(2.4)

$$\vec{W}^{n+1} = \vec{W}^n + \Delta t \sum_{q=1}^{m} b_q \vec{F}_\Delta(\vec{W}^{(q)}),$$

where $\vec{W}^n$ and $\vec{W}^{n+1}$ are numerical approximations to the solution $\vec{W}(t)$ of (2.3) at $t = n\Delta t$ and $t = (n+1)\Delta t$, respectively; the accuracy of these approximations is determined by the parameters a_{iq} and b_q.

3. THE ORDER CONDITIONS

Let $\vec{w}$ be an infinitely differentiable function, then the operators D_x and D_y can be expressed as

$$(3.1)\qquad D_x = \frac{\partial}{\partial x} X\left(\Delta x \frac{\partial}{\partial x}, \Delta y \frac{\partial}{\partial y}\right)$$

$$D_y = \frac{\partial}{\partial y} X\left(\Delta y \frac{\partial}{\partial y}, \Delta x \frac{\partial}{\partial x}\right),$$

where $X(x,y)$ is given by

$$X(x,y) = 2 \sum_{\ell=0}^{k} \sum_{j=1}^{k} \zeta_j^{(\ell)} \frac{\sinh jx}{x} \cosh ly. \tag{3.2}$$

We will always assume that $X(0,0) = 1$ and $\Delta x/\Delta y = c \neq 0$ as Δx, $\Delta y \to 0$; then $D_x = \partial/\partial x + 0(\Delta x)$ and $D_y = \partial/\partial y + 0(\Delta x)$. If, in addition, $X_x(0,0) = X_y(0,0) = 0(\Delta^2 x)$ then we have second order accurate difference operators, etc. Furthermore, we assume that $\Delta x = 0(\Delta t)$ and $\Delta y = 0(\Delta t)$.

As observed in the introduction, our starting point will be the family of second order accurate Runge-Kutta methods, employing fourth-order accurate discretization for $\partial/\partial x$ and $\partial/\partial y$. The following theorem provides the precise order conditions (the proof is given in the appendix to this paper):

THEOREM 3.1. *The difference scheme* {(2.1);(2.4)} *is second order accurate in time and fourth-order accurate in space if*

$$\beta_1 := \sum_{q=1}^{m} b_q = 1, \quad \beta_2 := \sum_{q=2}^{m} b_q \sum_{i=1}^{q-1} a_{qi} = \tfrac{1}{2}$$

and if

$$\gamma_1 := \sum_{\ell=0}^{k} \sum_{j=1}^{k} j\zeta_j^{(\ell)} = \tfrac{1}{2}, \quad \gamma_{21} := \sum_{\ell=0}^{k} \sum_{j=1}^{k} j^3 \zeta_j^{(\ell)} = 0(\Delta^2 t),$$

$$\gamma_{22} := \sum_{\ell=0}^{k} \sum_{j=1}^{k} j\ell^2 \zeta_j^{(\ell)} = 0(\Delta^2 t). \quad \square$$

In this paper we will pay particular attention to four-stage Runge-Kutta methods, i.e. $m = 4$, and to the conventional 8-point space discretization of $\partial/\partial x$ and $\partial/\partial y$, i.e.

$$k = 2; \quad \zeta_j^{(1)} = \zeta_j^{(2)} = 0, \quad j = 1,2; \quad \zeta_1^{(0)} = \frac{2}{3}, \quad \zeta_2^{(0)} = -\frac{1}{12}. \tag{3.3}$$

In the near future we will investigate adapted space discretizations [3].

We remark that, usually, the order equations are solved without taking into account the $0(\Delta t^2)$ terms. For our purposes, however, the addition of these terms is essential.

4. FOURIER ANALYSIS

Throughout this section it will be assumed that the matrices $A(\vec{w})$ and $B(\vec{w})$ are slowly varying with $\vec{w}$ in the neighbourhood of $t = t_n$. Furthermore, we assume that locally the exact solution $\vec{w}(t,x,y)$ of (1.1) contains dominant components of the form

$$\vec{v}(t,x,y) = \vec{a} \exp i(\alpha t+\omega_x x+\omega_y y) =: \vec{a} \exp i(\alpha t+\vec{\omega}\cdot\vec{x}), \tag{4.1}$$

where α,ω_x and ω_y are constants and $\vec{a}$ is a constant vector. If the mode (4.1) satisfies the differential equation (1.1), then

$$(A\omega_x+B\omega_y)\vec{a} = \alpha\vec{a}$$

showing that α is an eigenvalue of the matrix $\omega_x A + \omega_y B$ with eigenvector $\vec{a}$.

The Runge-Kutta solution of the semi-discrete equation (2.2) is approximately given by

$$\vec{W}^{n+1} = R_m(\Delta tAD_x+\Delta tBD_y)\vec{W}^n, \tag{4.2}$$

where $R_m(\cdot)$ is the *stability polynomial* of the method (see e.g. [2]). This polynomial is of the form

$$R_m(z) = 1 + \beta_1 z + \beta_2 z^2 +\ldots+ \beta_m z^m, \tag{4.3}$$

where the β_j are certain expressions in terms of the Runge-Kutta parameters (see Section 5). We observe that β_1 and β_2 are the same coefficients occurring in Theorem 3.1.

We want to compare $[\vec{v}(t_{n+1},x,y)]_{\Omega_\Delta}$ with $\vec{W}^{n+1}$ if we set $\vec{W}^n = [\vec{v}(t_n,x,y)]_{\Omega_\Delta}$. For that purpose it is convenient to introduce the eigenvalues $i\delta_x$ and $i\delta_y$ of the difference operators D_x and D_y corresponding to the eigenvectors $\exp(i\vec{\omega}\cdot\vec{x})$:

$$D_x e^{i\vec{\omega}\cdot\vec{x}} = i\delta_x e^{i\vec{\omega}\cdot\vec{x}}, \quad D_y e^{i\vec{\omega}\cdot\vec{x}} = i\delta_y e^{i\vec{\omega}\cdot\vec{x}}. \tag{4.4}$$

It is easily verified that

$$\delta_x = \frac{2}{\Delta x} \sum_{\ell=0}^{k} \sum_{j=1}^{k} \zeta_j^{(\ell)} \sin(j\omega_x \Delta x)\cos(\ell\omega_y \Delta y)$$

(4.5)

$$\delta_y = \frac{2}{\Delta y} \sum_{\ell=0}^{k} \sum_{j=1}^{k} \zeta_j^{(\ell)} \sin(j\omega_y \Delta y)\cos(\ell\omega_x \Delta x).$$

Using (4.4) we may write

$$R_m(\Delta t A D_x + \Delta t B D_y)\vec{a}e^{i(\alpha t_n + \vec{\omega}\cdot\vec{x})} = R_m(i\Delta t A\delta_x + i\Delta t B\delta_y)\vec{a}e^{i(\alpha t_n + \vec{\omega}\cdot\vec{x})}.$$

Suppose that $\delta_x = \delta\omega_x$ and $\delta_y = \delta\omega_y$, then, recalling that α is an eigenvalue of $\omega_x A + \omega_y B$ with eigenvector $\vec{a}$, we find

$$\vec{W}^{n+1} = R_m(\Delta t A D_x + \Delta t B D_y)[\vec{v}(t_n,x,y)]_{\Omega_\Delta} = R_m(i\alpha\Delta t\delta)[\vec{v}(t_n,x,y)]_{\Omega_\Delta}.$$

From (4.5) it follows that the condition $\delta_x = \delta\omega_x$, $\delta_y = \delta\omega_y$ is satisfied if $|\omega_x|\Delta x = |\omega_y|\Delta y$, at the same time defining δ. Thus, we have proved:

THEOREM 4.1. *Let* $\vec{v}(t,x,y)$ *be given by* (4.1) *with* $|\omega_x|\Delta x = |\omega_y|\Delta y$. *Then*

$$\vec{v}(t_{n+1},x,y) - R_m(\Delta t A D_x + \Delta t B D_y)\vec{v}(t_n,x,y)$$

$$= [e^{i\alpha\Delta t} - R_m(i\alpha\Delta t\delta)]\vec{v}(t_n,x,y),$$

where

$$\delta := 2 \sum_{\ell=0}^{k} \sum_{j=1}^{k} \zeta_j^{(\ell)} \frac{\sin j\mu \cos \ell\mu}{\mu}, \quad \mu := \omega_x \Delta x. \qquad \square \tag{4.6}$$

The function $\delta = \delta(\mu)$ will be called the (*space*) *discretization function*.

We now define dissipation and dispersion of the difference scheme {(2.1);(2.4)} in terms of the stability polynomial of the Runge-Kutta method:

DEFINITION 4.1. The *dissipation or amplitude error* of the difference scheme is defined by the quantity

$$1 - |R_m(i\alpha\Delta t\delta)|. \tag{4.7}$$

The *dispersion* or *phase error* is defined by

$$\alpha\Delta t - \arg(R_m(i\alpha\Delta t\delta)). \tag{4.8}$$ □

4.1. Exponential fitting

In the remainder of this section we will restrict our considerations to the case where $m = 4$, $\beta_1 = 1$ and $\beta_2 = \frac{1}{2}$. Then

$$R_m(z) = R_4(z) = 1 + z + \tfrac{1}{2}z^2 + \beta_3 z^3 + \beta_4 z^4. \tag{4.9}$$

The coefficients β_3 and β_4 will be determined by the condition that we have zero dissipation and zero dispersion when $\alpha = \alpha_0$ and $\omega_x = \omega_0$ where (α_0,ω_0) corresponds to a point in the range of dominant frequencies, i.e. $(\alpha_0,\omega_0) \in [\underline{\alpha},\bar{\alpha}]\times[\underline{\omega}_x,\bar{\omega}_x]$. It is convenient to introduce the variables $\nu = \alpha\Delta t$ and $\mu = \omega_x\Delta x$, and to write $\nu_0 = \alpha_0\Delta t$, $\mu_0 = \omega_0\Delta x$, $\delta_0 = \delta(\mu_0)$. Then β_3 and β_4 follow from the equation $R_4(i\nu_0\delta_0) = \exp(i\nu_0)$

$$\beta_3 = \frac{\nu_0\delta_0 - \sin\nu_0}{(\nu_0\delta_0)^3} \approx \left(\frac{1}{6} + \frac{\delta_0^{-1}}{\nu_0^2} - \frac{\nu_0^2}{120}\right) / \delta_0^3$$

$$\beta_4 = \frac{\cos\nu_0 - 1 + \frac{1}{2}(\nu_0\delta_0)^2}{(\nu_0\delta_0)^4} \approx \left(\frac{1}{24} + \frac{\delta_0^{2-1}}{2\nu_0^2} - \frac{\nu_0^2}{720}\right) / \delta_0^4. \tag{4.10}$$

The Runge-Kutta method is said to be *exponentially fitted* at (ν_0,δ_0) (cf.[5,p.240]). The optimal location of the fitting point will be discussed in Section 4.5.

THEOREM 4.2. *Let* $\zeta_j^{(\ell)}$ *satisfy the conditions for fourth order accuracy of Theorem* 3.1, *and let the Runge-Kutta parameters satisfy the fitting conditions* (4.10). *Then*

$$\delta_0 := \delta(\mu_0) = 1 + O(\Delta^4 t),\ \beta_3 = \frac{1}{6} + O(\Delta^2 t),\quad \beta_4 = \frac{1}{24} + O(\Delta^2 t).$$

PROOF. From (3.2) and (4.6) it follows that

$$(4.11)\qquad \delta(\mu_0) = X\left(\frac{\mu_0}{i}, \frac{\mu_0}{i}\right) = 1 + O(\mu_0^4) = 1 + O(\Delta^4 t).$$

Substitution into (4.10) leads to the expansions of β_3 and β_4 of the theorem. □

In Section 5 it will be shown that, by virtue of this theorem, we can achieve fourth order accuracy in time.

4.2. Stability

Before discussing dissipation and dispersion of the modes (4.1) by the scheme {(2.1);(2.4)}, we derive the stability condition of the scheme. Assuming that the solution $\vec{W}^n$ can be written as a Fourier series we derive from (4.2), (4.4) and (4.5) that each Fourier component is amplified by a factor $R_m(i\Delta t\lambda)$ where λ is an eigenvalue of the matrix $(\delta_x A+\delta_y B)$. Let β_{imag} be the imaginary stability boundary of the Runge-Kutta method then we have stability if $|\lambda\Delta t| < \beta_{imag}$ for all eigenvalues λ (notice that λ is real because A and B are assumed to be symmetric). Thus we arrive at the stability condition

$$(4.12)\qquad \Delta t < \frac{\beta_{imag}}{S(\delta_x A+\delta_y B)},$$

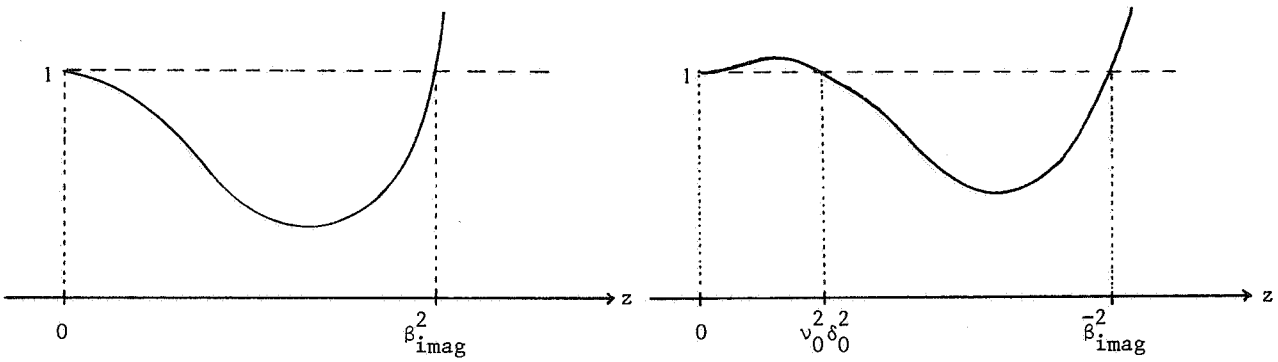
where $S(\cdot)$ denotes the spectral radius. We remark that here in δ_x, δ_y the frequencies ω_x and ω_y are independent, in contrast to the restriction made in Theorem 4.1 where we postulated $\omega_x \Delta x = \omega_y \Delta y$.

4.3. Dissipation

The concept of dissipation is strongly related to the concept of stability: if we have negative dissipation for a particular α and $\vec{\omega}$ then the mode $\vec{v}(t,x,y)$ will be amplified by the numerical scheme and may cause instabilities. The amount of dissipation is determined by the behaviour of $|R_4|$ along the imaginary axis. Let us write $z := \nu^2\delta^2$, then

$$(4.13) \qquad |R_4(i\nu\delta)|^2 = 1 + z^2(\tfrac{1}{4}+2\beta_4-2\beta_3+\beta_3^2 z-\beta_4 z+\beta_4^2 z^2).$$

In Figure 4.1 two typical situations are plotted, respectively corresponding to the cases $\beta_3 - \beta_4 \geq 1/8$ and $\beta_3 - \beta_4 < 1/8$.



(a) $\beta_3-\beta_4 \geq 1/8$ (b) $\beta_3-\beta_4 < 1/8$

Figure 4.1. $|R_4|^2(z)$, $z := \nu^2\delta^2$

In the case $\beta_3-\beta_4 \geq 1/8$ we have *one positive zero* of the error $1-|R_4(i\nu\delta)|^2$ at $z = \nu^2\delta^2 = \beta^2_{imag}$. This zero is just the imaginary stability boundary. We have dissipation, and therefore stability, if $\nu\delta < \beta_{imag}$, i.e.

$$(4.14a) \qquad \Delta t < \frac{\beta_{imag}}{|\alpha\delta|} .$$

This condition should be satisfied not only for those $\alpha\delta$, for which the mode (4.1) is dominant in the solution, but for all $\alpha\delta$; otherwise, instabilities will rapidly develop, because $|R_4(z)|$ is strongly increasing for $z > \beta^2_{imag}$. Since $\max|\alpha\delta|$ is usually rather large, we want a large imaginary stability boundary β_{imag}. It is easily verified that β_{imag} increases if $\beta_3-\beta_4 \downarrow 1/8$ to reach a maximal value $\beta_{imag} = 2\sqrt{2}$ for $\beta_3 = 1/6$ and $\beta_4 = 1/24$. Thus, $2\sqrt{2}$ is the maximal attainable imaginary stability boundary of all second-order, four-stage, explicit Runge-Kutta methods [2]. It should be remarked, however, that we assumed that β_3 and β_4 are defined by (4.10) from which it can be deduced that

$$\beta_3-\beta_4 \approx \frac{1}{8} - \frac{(\delta_0-1)^2}{2\nu_0^2} - \frac{5}{720}\nu_0^2$$

for small values of δ_0-1 and ν_0; hence, $\beta_3-\beta_4 < 1/8$ unless $\nu_0 = 0$.

In the case $\beta_3 - \beta_4 < 1/8$ we have *two positive zeros* of the error $1 - |R_4(i\nu\delta)|^2$ at $z = \nu_0^2\delta_0^2$ and $z = \bar{\beta}^2_{imag}$. We have dissipation, and hence stability, if

$$(4.14b) \qquad |\frac{\nu_0\delta_0}{\alpha\delta}| < \Delta t < \frac{\bar{\beta}_{imag}}{|\alpha\delta|}.$$

The right-hand inequality should be satisfied for all $\alpha\delta$ because of the same reasons as mentioned in the preceding case. The left-hand inequality is less urgent provided that $|R_4(i\nu\delta)|$ is only marginally larger than 1 for $0 < \nu^2\delta^2 < \nu_0^2\delta_0^2$. Suppose that $|\nu_0\delta_0| << 1$ then

$$|R_4|^2(z) \approx 1 + (\tfrac{1}{4}+2\beta_4-2\beta_3)z^2 + (\beta_3^2-\beta_4)z^3.$$

From this approximation we conclude that the first positive zero of $|R_4|^2 - 1$ is approximately given by

$$z = (\nu_0\delta_0)^2 \approx \frac{1-8(\beta_3-\beta_4)}{4(\beta_4-\beta_3^2)};$$

furthermore, $|R_4|^2$ assumes at $z \approx 2\nu_0^2\delta_0^2/3$ a maximum given by

$$(4.15) \qquad |R_4|^2_{max} \approx 1 + \frac{4}{27}(\beta_4-\beta_3^2)\nu_0^6\delta_0^6 \approx 1 + \frac{1}{486}\,\nu_0^6\delta_0^6,$$

where we have used (4.10). The value of the stability boundary β_{imag} is then approximately given by

$$\bar{\beta}^2_{imag} \approx \frac{\beta_4-\beta_3^2}{\beta_4^2} - \nu_0^2\delta_0^2 \approx 8 - O(\nu_0^2\delta_0^2).$$

We summarize the preceding results in the following theorem:

THEOREM 4.2. *A particular mode of the form* (4.1) *characterized by* $\alpha = \alpha_0$ *and* $\omega_x = \omega_0$ *is propagated by the numerical scheme with zero dissipation; the scheme is dissipative with respect to all modes where* $|\nu_0\delta_0| < |\nu\delta| < 2\sqrt{2} - O(\nu_0^2\delta_0^2)$; *finally, the scheme has a dissipation error of at most* $\approx \nu_0^6\delta_0^6/972 + O(\nu_0^8\delta_0^8)$ *with respect to all modes with* $|\nu\delta| < |\nu_0\delta_0|$. □

In actual computation the amplitude error in the interval $[0,\nu_0\delta_0]$ is negligible so that the method may be considered as stable if Δt satisfies the condition (4.12) with β_{imag} replaced by $\bar{\beta}_{imag}$.

4.4. Dispersion

Again write $z = \nu^2\delta^2$; thus, for m = 4, the dispersion is given by

$$(4.16) \qquad \phi(z,\delta) := \nu - \arg(R_4(i\nu\delta)) = \frac{1}{\delta}\sqrt{z} - \arctan\left(\sqrt{z}\,\frac{1-\beta_3 z}{1-\frac{1}{2}z+\beta_4 z^2}\right).$$

In order to get some idea of the behaviour of this phase error we have plotted the phase of the exact solution and that of the numerical solution as a function of z. In Figure 4.2 these curves are given for

$$\delta = 1, \quad \nu_0 = .1 \quad \delta_0 = 1.0001.$$

The most interesting aspect of the phase error is its behaviour in the neighbourhood of the origin (0,0) and of the fitting point $(z_0,\delta_0) := (\nu_0^2\delta_0^2,\delta_0)$. In the region $0 \le z \le z_0$, $\delta \approx \delta_0$ we represent $\phi(z,\delta)$ by an expression of the form

$$(4.17a) \qquad \phi(z,\delta) = \sqrt{z}[\frac{1}{\delta} - 1 + d_1 z + d_2 z^2] + O(z^3\sqrt{z}).$$

By requiring that $\phi(z_0,\delta_0) = 0$ and that ϕ has a correct behaviour at z = 0 up to second derivatives, we find that

$$(4.17b) \qquad d_1 = \beta_3 - \frac{1}{6}, \quad d_2 = \frac{1-1/\delta_0-(\beta_3-1/6)z_0}{z_0^2}, \quad z_0 = \nu_0^2\delta_0^2.$$

In Figure 4.3 the function (4.17) is illustrated for a fixed $\delta > 1$ and $d_1 > 0$ (i.e. $\beta_3 > 1/6$). In addition, the phase error corresponding to the conventional time integrator (i.e. $\beta_3 = 1/6$, $\beta_4 = 1/24$, or equivalently $z_0 = 0$, $\delta_0 = 1$) is plotted. This phase error is given by

$$(4.18) \qquad \phi(z,\delta) = -(1-\frac{1}{\delta})\sqrt{z} + O(z^3\sqrt{z}).$$

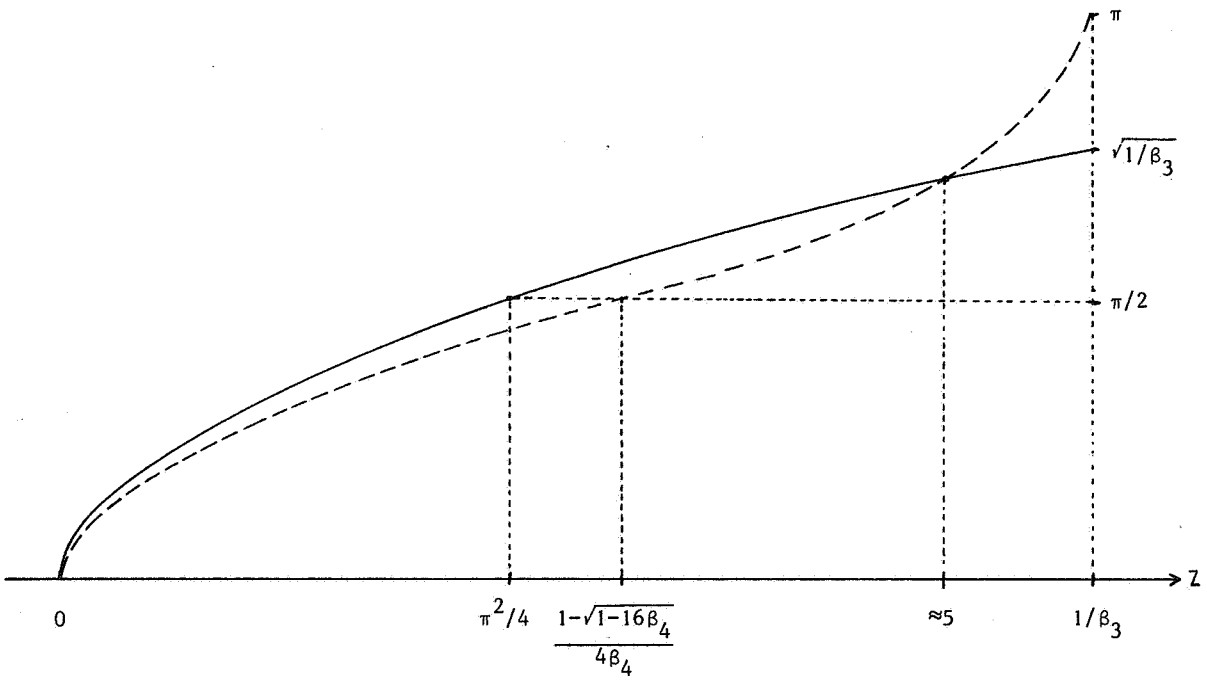


Figure 4.2. Exact (-) and numerical (--) phase for $\nu_0 = \cdot 1$, $\delta_0 = 1.0001$, $\delta = 1$

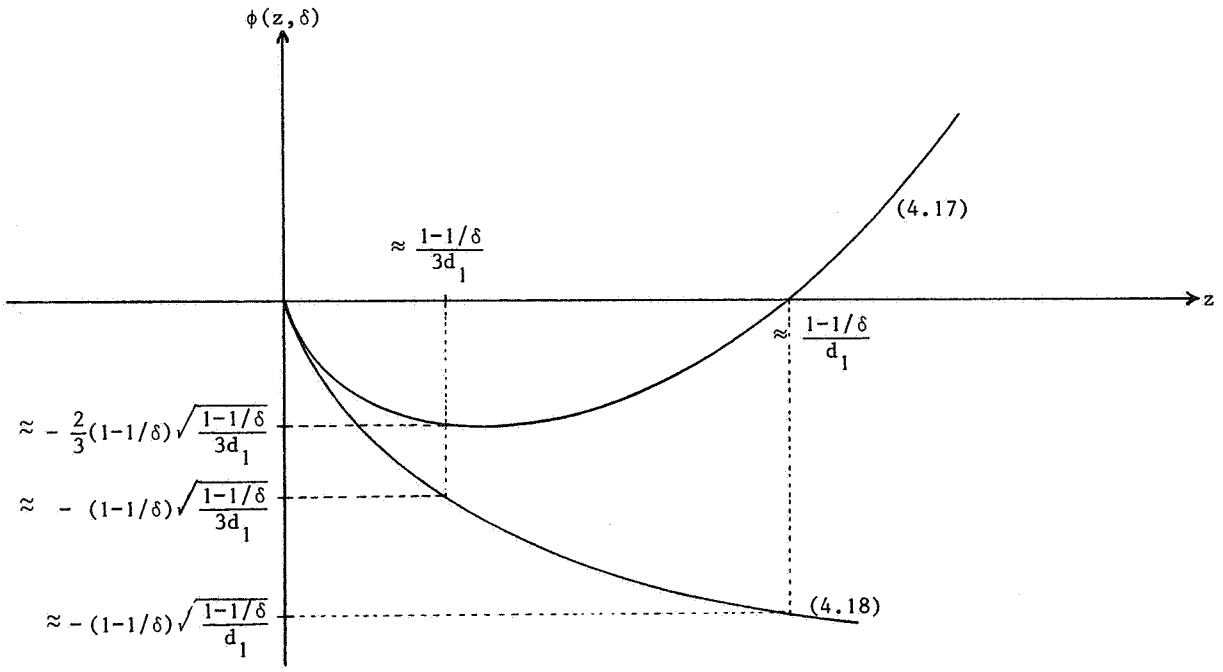


Figure 4.3. The phase error (4.17) for fixed δ and $\beta_3 = 1/6$, $\beta_3 > 1/6$

In order to compare the phase errors corresponding to the conventional and exponentially fitted methods we consider the maximum norm $\|\phi\|_\infty$ over the region

$$[\underline{z},\bar{z}] \times [\underline{\delta},\bar{\delta}] := [\underline{\nu}^2\underline{\delta}^2, \bar{\nu}^2\bar{\delta}^2] \times [\underline{\delta},\bar{\delta}].$$

Here, $[\underline{\nu},\bar{\nu}]$ denotes the interval of dominant time constants and $[\underline{\delta},\bar{\delta}]$ denotes the range of the function $\delta(\mu)$ on the interval of dominant frequencies $[\underline{\mu},\bar{\mu}]$. Since $\phi(z,\delta)$ is a monotone function of δ we have

$$(4.19) \qquad \|\phi\|_\infty = \max_{\underline{z} \le z \le \bar{z}} \{|\phi(z,\underline{\delta})|, |\phi(z,\bar{\delta})|\}.$$

Figure 4.3 indicates that for $\bar{z} \le (1-1/\delta)/d_1$, $\delta \in [\underline{\delta},\bar{\delta}]$ the phase error of

the exponentially fitted method is at least a factor $3\sqrt{3}/2$ smaller than the phase error of the conventional method, even for $\underline{z} \to 0$. However, if $\underline{z} \to \bar{z}$ an increasing gain factor is obtained. This will be illustrated in the following subsections.

4.5. The location of the optimal fitting point (ν_0, δ_0)

In order to obtain a small phase error the choice of the fitting point (ν_0, δ_0) is of crucial importance. As will be shown, the location of δ_0 - or, actually, the location of $\mu_0 \in [\underline{\mu}, \bar{\mu}]$ - is less critical than the position of ν_0 in the interval $[\underline{\nu}, \bar{\nu}]$. This is due to the fact that the range of $\delta(\mu)$ is usually very small.

To demonstrate the influence on the phase error of each parameter separately, we performed two calculations in which one of the parameters is kept fixed and the other one is chosen optimal by a straightforward numerical search. Here, optimal means that $\|\phi\|_\infty$ is minimal. In Tables 4.1 and 4.2 we have listed the gainfactors $\|\phi_c\|_\infty / \|\phi_e\|_\infty$, where the indices c and e refer to the classical method (i.e. $\beta_3 = 1/6$, $\beta_4 = 1/24$) and the exponentially fitted method (i.e. β_3 and β_4 according to (4.10)), respectively. These gainfactors are listed for several intervals $[\underline{\nu}, \bar{\nu}]$ and $[\underline{\mu}, \bar{\mu}]$ which are characterized by their respective centres $\hat{\nu}$ and $\hat{\mu}$ and the "uncertainty percentages" $\Delta\nu$ and $\Delta\mu$ defined by

$$\Delta\nu := 100\ \frac{\bar{\nu}-\hat{\nu}}{\hat{\nu}}, \qquad \Delta\mu := 100\ \frac{\bar{\mu}-\hat{\mu}}{\hat{\mu}}.$$

Moreover, we give in each case the value of the optimal fitting parameter as it was found by the numerical search. The discretization function $\delta(\mu)$ is determined by (3.3).

Table 4.1. Gainfactors $\|\phi_c\|_\infty / \|\phi_e\|_\infty$ for $\mu_0 = 0.25$ and the optimal fitting point ν_0 (in parentheses)

$\hat{\nu}$ \ $\Delta\nu$	5%	10%	20%	50%
.1	10.7(.1004)	5.8(.1015)	3.6(.1059)	3.1(.1300)
.2	13.0(.2007)	7.1(.2027)	4.6(.2101)	5.0(.2507)
.3	31.5(.2999)	18.0(.2996)	12.3(.2908)	16.2(.4500)
.4	64.6(.4068)	38.9(.4224)	22.4(.4560)	14.8(.2507)
.5	22.7(.5050)	14.4(.5190)	11.4(.5620)	8.6(.6983)

Table 4.2. Gainfactors $\|\phi_c\|_\infty / \|\phi_e\|_\infty$ for $\nu_0 = 0.25$ and the optimal fitting point μ_0 (in parentheses)

$\hat{\mu}$ \ $\Delta\mu$	5%	10%	20%	50%
.1	53.7(.1004)	27.3(.1015)	14.0(.1056)	5.8(.1263)
.2	9.0(.2007)	5.1(.2029)	3.2(.2112)	2.3(.2527)
.3	6.7(.3011)	3.9(.3044)	2.6(.3164)	2.1(.3790)
.4	6.3(.4015)	3.7(.4059)	2.5(.4219)	2.0(.5040)
.5	6.2(.5018)	3.7(.5070)	2.5(.5273)	2.0(.6300)

The larger numbers in the first Table clearly show that fitting the "time frequency parameter" ν is more advantageous than using an optimal value for the "space frequency parameter" μ.

5. CONSTRUCTION OF THE RUNGE-KUTTA SCHEME

In Theorem 3.1 the order equations for second order accuracy of the Runge-Kutta time integrator have been given. In this section it will be shown that, taking into account the exponentially fitting conditions (4.10), it is possible to construct a four-stage method which is even fourth-order accurate.

Let us denote the scheme (2.4) by the Butcher array

$$\begin{array}{c|c} \vec{c} & A \\ \hline & \vec{b}^T \end{array}, \qquad A := (a_{iq}), \quad \vec{b} = (b_q)$$

where the elements of $\vec{c}$ are the row sums of A. We will look for fourth-order schemes of the form

(5.1a)
$$\begin{array}{c|cccc} 0 & 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & b_1 & b_2 & b_3 & b_4 \end{array}$$

From the Runge-Kutta theory (e.g. [2]) it can be derived that this scheme is fourth-order accurate if the b_j satisfy the system

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1/2 & 1/2 & 1 \\ 0 & 0 & 1/4 & 1/2 \\ 0 & 1/4 & 1/4 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1/8 & 1/4 \\ 0 & 0 & 1/8 & 1/2 \\ 0 & 1/8 & 1/8 & 1 \end{pmatrix} \vec{b} = \begin{pmatrix} 1 \\ 1/2 \\ \beta_3 \\ 1/3+0(\Delta^2 t) \\ 4\beta_4 \\ 1/12+0(\Delta t) \\ 1/8 +0(\Delta t) \\ 1/4 +0(\Delta t) \end{pmatrix}$$

where β_3 and β_4 should equal $1/6 + 0(\Delta^2 t)$ and $1/24 + 0(\Delta t)$, respectively. It turns out that this system can only be solved if, and only if, β_4 equals $1/24 + 0(\Delta^2 t)$ instead of $1/24 + 0(\Delta t)$. Conveniently, by virtue of Theorem 4.2, β_4 does equal $1/24 + 0(\Delta^2 t)$. The system can now be solved to obtain

$$\vec{b}^T = (4\beta_4, 1-4\beta_3, 4\beta_3-8\beta_4, 4\beta_4). \tag{5.1b}$$

6. NUMERICAL ILLUSTRATION

6.1. A model problem

Consider the linear test equation

$$\frac{\partial w}{\partial t} = \frac{2}{10}\frac{\partial w}{\partial x} + \frac{3}{10}\frac{\partial w}{\partial y}, \qquad t \in [0,T], \quad (x,y) \in \mathbb{R}^2 \tag{6.1}$$

and the initial condition

$$w(0,x,y) = \cos(x+2y), \quad (x,y) \in \mathbb{R}^2. \tag{6.2}$$

The exact solution is given by

$$w(t,x,y) = \cos(\tfrac{4}{5}t+x+2y). \tag{6.3}$$

According to the discussion in Section 4, we have chosen a uniform grid such that $\omega_x \Delta x = \omega_y \Delta y$, i.e.

$$\Delta x = 2\Delta y. \tag{6.4}$$

Let $\partial/\partial x$ and $\partial/\partial y$ be discretized using the "line-molecules" defined by (3.3):

$$D_x = \frac{1}{\Delta x}[\frac{1}{12}, -\frac{2}{3}, 0, \frac{2}{3}, -\frac{1}{12}], \quad D_y = \frac{\Delta x}{\Delta y} D_x^T. \tag{6.5}$$

For the space meshes we have chosen $\Delta x = 2\Delta y = 2\pi/40$. By this particular choice we achieve that the solution $w(t,x,y)$ is periodic on the rectangle $[0,2\pi] \times [0,\pi]$ in the (x,y)-plane; this feature has been used in the implementation to reduce the $\mathbb{R}^2$ to this rectangle and requiring periodicity on the boundaries.

For the time integration we used the four-stage Runge-Kutta scheme (5.1) where the coefficients β_3 and β_4 are determined by the fitting point (ν_0, δ_0) (cf.(4.10)). The results of this method will be compared with the classical four-stage RK method (i.e.(5.1) with $\vec{b}^T = (1/6,1/3,1/3,1/6)$).

As the time step Δt should satisfy the stability condition (cf.(4.12))

$$\Delta t \leq \frac{2\sqrt{2}}{S(.2\delta_x + .3\delta_y)} \tag{6.6}$$

we need an estimate of the spectral radius S. Using Gerschgorin's disc theorem we find

$$S(.2\delta_x + .3\delta_y) \simeq \frac{18}{12}(\frac{.2}{\Delta x} + \frac{.3}{\Delta y}) = \frac{6}{5}\frac{1}{\Delta x} = \frac{24}{\pi} \tag{6.7}$$

resulting in $\Delta t \lesssim .37$. In our experiments we used $\Delta t = 1/3$.

Additionally, a comparison is made with a two-dimensional application of the celebrated method of Lax-Wendroff (see e.g. [7 ,p.181]). We applied this method on the same rectangle in the (x,y)-plane; however, the values of the mesh spacings are chosen somewhat smaller, viz. $\Delta x = 2\Delta y = 2\pi/52$. Now, by integrating with the maximal stable integration step for this method, $\Delta t = 1/7$, we achieve that the same amount of work is required - to

reach the end point t = T - as in the case of the RK methods. By this we mean that the total number of evaluations of an expression of the form $\zeta \cdot \frac{W_+ - W_-}{\Delta}$ is the same for both types of methods. Note that, using this working-unit, one time step with the Lax-Wendroff method (which is second order in space) is four times as cheap as one step with the RK method (fourth order in space).

To be able to apply the exponentially fitted method we need the values of the fitting parameters ν_0 and μ_0. We will have a minimal phase error (minimal dispersion) if we solve the minimax problem

$$\min_{(\mu_0,\nu_0)\in R} \max_{(\mu_0,\nu_0)\in R} |\phi(\nu,\mu)|, \quad R := [\underline{\mu},\bar{\mu}]\times[\underline{\nu},\bar{\nu}] \tag{6.8}$$

for given intervals $[\underline{\nu},\bar{\nu}]$ and $[\underline{\mu},\bar{\mu}]$. For this two-dimensional minimization problem we used the routine E04JAF from the NAG library [8].

To measure the obtained accuracy we define

$$cd := -{}^{10}\log(\|\text{global error at the endpoint } t = T\|_\infty), \tag{6.9}$$

denoting the number of correct digits in the numerical approximation at the end point.

First, as a reference, we give the cd-values of the various methods obtained at the end point T = 100.

cd-value of the classical RK method	cd-value of the exponentially fitted RK method	cd-value of the Lax-Wendroff method
2.31	11.02	1.66

As the exponentially fitted method was given the exact values for the space- and time frequencies, its integration process is "exact" (relative to the machine-precision).

Next we vary the frequency intervals to test their influence on the accuracy of the exponentially fitted methods. Such an interval can be considered as an uncertain estimate of the exact value of the corresponding

frequency. We start with varying the length of the ν-interval, centered around the exact value $0.8\,\Delta t$, whereas the μ-parameter is assumed to be estimated correctly, i.e. $\underline{\mu} = \bar{\mu} = \Delta x$. The results can be found in Table 6.1. Further, the rôles of ν and μ have been interchanged, yielding the cd-values as listed in Table 6.2.

Table 6.1. cd-values for the exponentially fitted method for several ν-intervals; $\underline{\mu} = \bar{\mu} = \Delta x = \pi/20$, $\Delta t = 1/3$ and $T = 100$.

$[\underline{\nu},\bar{\nu}]$:	$[.75,.85]\Delta t$	$[.7,.9]\Delta t$	$[.6,1.1]\Delta t$	$[.5,1.2]\Delta t$
cd :	4.03	3.71	3.37	3.16

Table 6.2. cd-values for the exponentially fitted method for several μ-intervals; $\underline{\nu} = \bar{\nu} = 0.8\,\Delta t = 4/15$, $T = 100$ and $\Delta x = \pi/20$

$[\underline{\mu},\bar{\mu}]$:	$[.95,1.05]\Delta x$	$[.9,1.1]\Delta x$	$[.8,1.2]\Delta x$	$[.5,1.5]\Delta x$
cd :	3.88	3.56	3.24	2.76

These results clearly show that the exponentially fitted method can take profit from a good estimate of the frequencies of the (dominant) solution component. However, even if this estimate is rather poor there still remains a substantial gain in accuracy when compared with the classical method (note that this method yields cd = 2.31 for this problem). Especially if we realize that the additional effort to adapt the Runge-Kutta scheme to the behaviour of the solution is quite negligible.

It should be observed that the errors are mainly due to phase errors (dispersion) and not to dissipation. Hence, we have linear accumulation of the phase errors made in each integration step. This observation is confirmed by an experiment in which we set the end point of the integration interval to $T = 10$; the cd-values obtained were approximately 1.0 larger than those given in the Tables of results, i.e. we have a 10 times smaller global error. This feature places great emphasis on the necessity of having a small phase error in long term integration processes.

6.2. The shallow water equations

Consider the basic, linearized form of the shallow water equations:

$$(6.10)\qquad \frac{\partial \vec{u}}{\partial t} = -g\nabla h,\ \frac{\partial h}{\partial t} = -h_0\nabla\vec{u},\quad t \in [0,T],(x,y) \in \mathbb{R}^2$$

where $\vec{u}$ is the depth-averaged velocity, h is the depth below the moving water surface, h_0 is the depth when the water is in rest, and g is the acceleration of gravity. Initial and boundary conditions on the square $0 \le x,y \le L$ were taken from the exact solution

$$(6.11)\qquad (\vec{u},h-h_0) = \tfrac{1}{4}\sin((-\sqrt{2gh_0}\,t+x+y)\,\frac{2\pi}{L}).(1,1,\sqrt{2h_0/g})$$

In our experiments we used $h_0 = 80$, $g = 10$, $L = 600000$, $\Delta x = \Delta y = L/24$ and $T = 18000$; the integration step Δt was chosen maximal with respect to stability and was found to be ≈ 818.2, which resulted in 22 steps. Furthermore, $D_x = D_y$ were defined according to (6.5).

When written in the form (1.1) the eigenvalues of the matrix $\omega_x A+\omega_y B = 2\pi(A+B)/L$ are given by

$$(6.12)\qquad \alpha_{\pm} = \pm\sqrt{gh_0(\omega_x^2+\omega_y^2)} = \pm\frac{2\pi}{L}\sqrt{2gh_0}.$$

In the exponentially fitted method the fitting point (α_0,ω_0) was chosen at $(\alpha_+,2\pi/L)$ so that $(\nu_0,\delta_0) = (\Delta t\alpha_+,\delta(2\pi\Delta x/L))$. We observe that fitting at $(\alpha_+,2\pi/L)$ automatically implies fitting at $(\alpha_-,2\pi/L)$.

In addition to the linear system (6.10) we also integrated the *non-linear* modifications which are closer to the actual shallow water equations:

$$(6.13)\qquad \frac{\partial \vec{u}}{\partial t} = -g\nabla h - (\vec{u}.\nabla)\vec{u},\ \frac{\partial h}{\partial t} = -h_0\nabla\vec{u},$$

$$(6.14)\qquad \frac{\partial \vec{u}}{\partial t} = -g\nabla h,\ \frac{\partial h}{\partial t} = -\nabla(h\vec{u}),$$

$$(6.15)\qquad \frac{\partial \vec{u}}{\partial t} = -g\nabla h - (\vec{u}.\nabla)\vec{u},\ \frac{\partial h}{\partial t} = -\nabla(h\vec{u}).$$

In Table 6.3 the number of correct digits, defined according to (6.9), are listed; here, the error is given by $h-h_0$.

Table 6.3. cd-values for the conventional and exponentailly fitted methods: $\Delta x = \Delta y = 25000$, $\Delta t = 818.2$, $T = 18000$

	(6.10)	(6.13)	(6.14)	(6.15)
Conventional method	2.72	2.58	2.41	2.28
Exponentially fitted method	11.64	3.35	2.91	2.64

These results clearly show the effect of deviating from the model problem situation (eq.(6.10)). However, in spite of the considerable drop in accuracy when nonlinear terms are introduced, we still obtain an error that is smaller by a factor 6 to 3, with insignificant additional effort (in this connection we remark that in the case of the equations (6.14) and (6.15) we replaced h_0 by h in (6.12) and computed the Runge-Kutta parameters β_3 and β_4 by exponential fitting in every grid point).

REFERENCES

[1] GAUTSCHI, W., *Numerical integration of ordinary differential equations based on trigonometric polynomials*, Numer. Math., 3 (1961), pp.381-397.

[2] HOUWEN, P.J. VAN DER, *Construction of integration formulas for initial-value problems*, North-Holland, Amsterdam, 1977.

[3] HOUWEN, P.J. VAN DER, B.P. SOMMEIJER and F.W. WUBS, *Reduction of dispersion in hyperbolic difference schemes by adapting the space discretization* (in preparation).

[4] JELTSCH, R. and O. NEVANLINNA, *Stability of explicit time discretizations for solving initial-value problems*, Num. Math. 37 (1981), pp.61-91.

[5] LAMBERT, J.D., *Computational methods in ordinary differential equations*, Wiley, London, 1973.

[6] LINIGER, W. and R.A. WILLOUGHBY, *Efficient integration methods for stiff systems of ordinary differential equations*, SIAM J. Numer. Anal., 7 (1970), pp.47-66.

[7] MITCHELL, A.R. and D.F. GRIFFITHS, *The finite difference method in partial differential equations*, Wiley, Chichester, 1980.

[8] NAG, *Numerical Algorithms Group*, FORTRAN Library Manual Mark 10, Oxford-Illinois, 1982.

APPENDIX

PROOF OF THEOREM 3.1. Let us substitute the expressions (3.1) into the semi-discrete equation (2.2) to obtain the so-called *modified equation*

$$(A.1)\qquad \frac{\partial \vec{w}}{\partial t} = \left[A(\vec{w})\,\frac{\partial}{\partial x}\,X(\Delta x\,\frac{\partial}{\partial x}\,,\Delta y\,\frac{\partial}{\partial y}\,) + B(\vec{w})\,\frac{\partial}{\partial y}\,X(\Delta y\,\frac{\partial}{\partial y}\,,\Delta x\,\frac{\partial}{\partial x}\,)\right]\vec{w}$$

$$=: \left[A(\vec{w})\,\frac{\partial}{\partial x} + B(\vec{w})\,\frac{\partial}{\partial y}\right]\vec{w} + E_\Delta \vec{w}(t,x,y),$$

where the error term $E_\Delta \vec{w}$ vanishes as $\Delta x \to 0$. Thus, when solving (2.2) we are not solving the equation (1.1), but a slightly perturbed equation which differs from the original equation (1.1) by the term $E_\Delta \vec{w}$. From (3.2) it is immediate that the operator E_Δ is given by

$$(A.2)\qquad E_\Delta = A\,\frac{\partial}{\partial x}[X(\Delta x\,\frac{\partial}{\partial x},\ \Delta y\,\frac{\partial}{\partial y})-1] + B\,\frac{\partial}{\partial y}[X(\Delta y\,\frac{\partial}{\partial y},\ \Delta x\,\frac{\partial}{\partial x})-1] = O(\Delta),$$

where $\Delta = \max\{\Delta x,\Delta y\}$.

Using Runge-Kutta theory we are now able to derive the truncation error of the difference scheme {(2.1),(2.4)}. Let us assume that the approximation $\vec{W}^n$ in (2.4) is exact, i.e. $\vec{W}^n = \vec{W}(t_n) = (\vec{w}(t_n, j\Delta x, \ell\Delta y))$, where $\vec{W}(t)$ denotes the exact solution of (2.3) through the point $(t_n,\vec{W}(t_n))$.

We compare the numerical solution with the exact solution $\vec{w}$ of the original equation (1.1). On Ω_Δ this solution satisfies the system of ODEs

$$(A.3)\qquad \frac{d\tilde{\vec{W}}}{dt} = \vec{F}_\Delta(\tilde{\vec{W}}) - [E_\Delta \vec{w}]_{\Omega_\Delta}, \qquad \tilde{\vec{W}} = [\vec{w}]_{\Omega_\Delta},$$

where $[\cdot]_{\Omega_\Delta}$ denotes the restriction of a continuous function onto the grid Ω_Δ. We write the local truncation error as

$$(A.4)\qquad \vec{W}^{n+1} - \tilde{\vec{W}}(t_{n+1}) = \vec{W}^{n+1} - \vec{W}(t_{n+1}) + \vec{W}(t_{n+1}) - \tilde{\vec{W}}(t_{n+1}).$$

First, we estimate the difference of the solutions of (2.3) and (A.3) by Taylor expansion:

$$\vec{W}(t_{n+1})-\tilde{\vec{W}}(t_{n+1}) = \{\vec{W}(t_n)+\Delta t\vec{F}_\Delta(\vec{W}(t_n))+\tfrac{1}{2}\Delta^2 t\left(\frac{\partial\vec{F}_\Delta}{\partial\vec{W}}\vec{F}_\Delta\right)(\vec{W}(t_n))$$

$$+\ldots\}-\{\tilde{\vec{W}}(t_n)+\Delta t(\vec{F}_\Delta(\tilde{\vec{W}}(t_n))-[E_\Delta\vec{w}]_{\Omega_\Delta,t_n})$$

$$+\tfrac{1}{2}\Delta^2 t\left(\frac{\partial}{\partial\tilde{\vec{W}}}\left(\vec{F}_\Delta(\tilde{\vec{W}}(t_n))-[E_\Delta\vec{w}]_{\Omega_\Delta,t_n}\right)\right)\cdot(\vec{F}_\Delta(\tilde{\vec{W}}(t_n))-[E_\Delta\vec{w}]_{\Omega_\Delta,t_n})+\ldots\}.$$

Since, by assumption, $\vec{W}(t_n) = \tilde{\vec{W}}(t_n)$, we find

$$\text{(A.5a)}\qquad \vec{W}(t_{n+1})-\tilde{\vec{W}}(t_{n+1}) = \Delta t[E_\Delta\vec{w}]_{\Omega_\Delta,t_n}(1+0(\Delta t)).$$

From (A.2) it follows that $E_\Delta\vec{w} = 0(\Delta^4 t)$ if

$$X(\Delta x\frac{\partial}{\partial y},\Delta y\frac{\partial}{\partial y}) = 1+0(\Delta^4 t).$$

Since

$$X(x,y) = 2\gamma_1 + \frac{1}{3}\gamma_{21}x^2 + \gamma_{22}y^2 + 0(x^4+y^4+x^2y^2),$$

where $\gamma_1,\gamma_{21},\gamma_{22}$ are the expressions in terms of the weights $\xi_j^{(\ell)}$ given in the theorem, it follows that the contribution of the spatial discretization to the truncation error is $0(\Delta^5 t)$ provided that the $\gamma_1,\gamma_{21},\gamma_{22}$ satisfy the conditions of the theorem.

The contribution of the Runge-Kutta discretization to the local truncation error is given by (cf. e.g. [5])

$$\text{(A.5b)}\qquad \vec{W}^{n+1} - \vec{W}(t_{n+1}) = (\beta_1-1)\Delta t\vec{D}_1 + (\beta_2-\tfrac{1}{2})\Delta^2 t\vec{D}_2 + (\beta_3-\tfrac{1}{6})\Delta^3 t\vec{D}_3$$

$$+ (\beta_{3,1}-\tfrac{1}{3})\Delta^3 t\vec{D}_{3,1} + (\beta_4-\tfrac{1}{24})\Delta^4 t\vec{D}_4$$

$$+ (\beta_{4,1}-\tfrac{1}{12})\Delta^4 t\vec{D}_{4,1}+(\beta_{4,2}-\tfrac{1}{8})\Delta^4 t\vec{D}_{4,2}+(\beta_{4,3}-\tfrac{1}{4})\Delta^4 t\vec{D}_{4,3}$$

$$+ 0(\Delta^5 t),\quad \Delta^j t := (\Delta t)^j.$$

Here the $\vec{D}_j,\vec{D}_{j,\ell}$ are expressions in the right-hand side function $\vec{F}_\Delta$, for example

$$\vec{D}_1 = \vec{F}_\Delta(\vec{W}(t_n)),\ \vec{D}_2 = \frac{\partial \vec{F}_\Delta}{\partial \vec{W}}(\vec{W}(t_n))\vec{F}_\Delta(\vec{W}(t_n)).$$

The coefficients β_j and $\beta_{j\ell}$ are expressions in terms of the Runge-Kutta parameters a_{iq} and b_q as given in the theorem. Evidently, the conditions of the theorem imply that (A5.b) is of order $\Delta^5 t$ as $\Delta t \to 0$.

The full truncation error defined by (A.5a) and (A.5b) satisfies the order equation

$$\vec{W}^{n+1} - \tilde{\vec{W}}(t_{n+1}) = 0(\Delta^5 t)$$

which proves the theorem. □