PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

# Conservative, High-Order Numerical Schemes for the Generalized Korteweg-de Vries Equation

J. L. Bona, V. A. Dougalis, O. A. Karakashian and W. R. McKinney

*Phil. Trans. R. Soc. Lond. A* 1995 **351**, 107-164
doi: 10.1098/rsta.1995.0027

# Conservative, high-order numerical schemes for the generalized Korteweg–de Vries equation†

By J. L. Bona[1], V. A. Dougalis[2], O. A. Karakashian[3] and W. R. McKinney[4]

[1]*Department of Mathematics and Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802 U.S.A.*
[2]*Mathematics Department, National Technical University, Zographou, 15780 Athens, Greece, and Institute of Applied and Computational Mathematics, F.O.R.T.H., Greece*
[3]*Department of Mathematics, University of Tennessee, Knoxville, TN 37996 U.S.A.*
[4]*Department of Mathematics, North Carolina State University, Raleigh, NC 27607 U.S.A.*

## Contents

| | PAGE |
|---|---|
| 1. Introduction | 108 |
| 2. The numerical methods | 110 |
| 3. Stability and convergence of the base scheme | 114 |
| 4. Computational considerations | 131 |
| 5. Numerical experiments: adaptive procedures, instability, and blow-up of solutions | 142 |
| 6. Conclusions and conjectures | 159 |
| References | 163 |

A class of fully discrete schemes for the numerical simulation of solutions of the periodic initial-value problem for a class of generalized Korteweg–de Vries equations is analysed, implemented and tested. These schemes may have arbitrarily high order in both the spatial and the temporal variable, but at the same time they feature weak theoretical stability limitations. The spatial discretization is effected using smooth splines of quadratic or higher degree, while the temporal discretization is a multi-stage, implicit, Runge–Kutta method. A proof is presented showing convergence of the numerical approximations to the true solution of the initial-value problem in the limit of vanishing spatial and temporal discretization. In addition, a careful analysis of the efficiency of particular versions of our schemes is given. The information thus gleaned is used in the investigation of the instability of the solitary-wave solutions of a certain class of these equations.

† Dedicated to Garrett Birkhoff on the occasion of his 83rd birthday.

*Phil. Trans. R. Soc. Lond.* A (1995) **351**, 107–164
*Printed in Great Britain*

## 1. Introduction

Considered herein are fully discrete, numerical approximation schemes for solutions of the generalized Korteweg–de Vries equation (GKdV equation henceforth) that possess high accuracy and high-order convergence rates in both the spatial and the temporal variables. This work is a continuation of the developments presented in an earlier paper (Bona *et al.* 1986). The particular contribution contained in the present paper concerns schemes with arbitrarily high rates of convergence. As will be made clear below, such schemes are of very considerable use in investigating the consequences of the balance between nonlinearity and dispersion that is the hallmark of the GKdV equations, and in making comparisons between experimental data and numerical simulations. The focus of our interest, the GKdV equation, may be written in the form

$$u_t + u^p u_x + u_{xxx} = 0, \tag{1.1a}$$

where $u = u(x,t)$ is a function of the two real variables $x$ and $t$ which correspond to space and time, respectively, and $p$ is a non-negative integer. This equation will be considered on the spatial interval $x \in [0,1]$ for $t \in [0,t^*]$, with initial data

$$u(x,0) = u_0(x) \tag{1.1b}$$

specified for $0 \leqslant x \leqslant 1$ and with $u_0$ belonging to a suitable class of periodic functions having period 1. Classical solutions whose spatial variation maintains the initially imposed periodicity will be considered and numerical approximations thereto will be proposed and investigated. The special case $p = 1$ corresponds to the Korteweg–de Vries equation itself (KdV equation, Korteweg & de Vries 1895), $p = 2$ to the modified Korteweg–de Vries equation (MKdV equation, Miura 1968), and $p = 0$ is a linear, dispersive equation whose exact solution may be found by Fourier analysis.

The particular equations exhibited in (1.1*a*) are part of a more general class which has arisen in recent years as approximate models for the unidirectional propagation of plane waves in a variety of nonlinear, dispersive media (cf. Benjamin *et al.* 1972; Saut 1975; Bona 1980, 1981*a*). For the KdV and the MKdV equations, and for certain other members of the general class to which allusion was just made, the inverse scattering transform (IST) provides a method of representing solutions from which detailed information may be extracted. However, in general, the IST does not apply and so numerical simulations come to the fore as an investigative tool. Even in cases where an IST exists, it is sometimes more convenient to use direct simulation of the partial differential equation in determining properties of solutions. An example of this arises when comparisons between experimentally obtained data and a model equation are desired. While such comparisons can be effected in a telling way using IST techniques (cf. Zabusky & Galvin 1971; Hammack 1973; Hammack & Segur 1974) there are a number of complications connected with dissipation and the imposition of boundary conditions that are obviated by more direct methods (e.g. Bona *et al.* 1981). In the situation that comes about when comparing laboratory data with numerical simulations, experience obtained in the last-cited reference shows clearly the efficacy of a scheme that is of higher order in the temporal variable. Also, in general investigations of the outcome of competition between nonlinearity and dispersion, one aspect of which is reported in § 5 of the present paper, some delicate properties




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

such as decay rates, formation of singularities, and instabilities associated with the initial-value problem require a very reliable and highly accurate numerical scheme such as those provided here.

The paper is organized in the following way. In § 2, after explaining notation and various other preliminaries, the numerical schemes are described in detail. The spatial discretizations are effected using smooth splines of quadratic or higher degree and the temporal discretizations are conservative, multi-stage, implicit Runge–Kutta methods. In fact, the schemes are written for (1.1) posed in the slightly more general form

$$u_t + \eta u_x + u^p u_x + \epsilon u_{xxx} = 0, \tag{1.2a}$$

where $\eta$ and $\epsilon$ are fixed constants, $p$ is as before, and the same initial condition

$$u(x, 0) = u_0(x) \tag{1.2b}$$

is imposed. Of course (1.2) is completely equivalent to (1.1) by way of the simple change of variables, $U(x, t) = u(\beta(x - \eta t), \beta t)$ with $\beta = \epsilon^{-1/2}$, but it is convenient both theoretically and practically to have the extra flexibility inherent in formulation (1.2). The proof of convergence of the numerical approximations to the solution of (1.2) in the case of uniform spatial meshlength and temporal step is presented in § 3. We obtain the optimal-order rate of convergence as far as the spatial discretization is concerned, and the optimal rates for the temporal discretization for the one- and two-stage time stepping procedures. For three- and higher-stage time stepping methods, the proven rate of convergence for the temporal discretization is sub-optimal. (For the special case $p = 1$ of the KdV equation, it has been shown recently by Karakashian & McKinney 1990 that the optimal temporal rate is achieved for such schemes with arbitrary number of stages.)

The implementation of the scheme and the outcome of detailed convergence studies are presented in § 4 for a two-stage time-stepping procedure and several different choices of the degree of the splines used in the spatial representation of the approximation. Although not reported here, a careful study of the relative and absolute efficiency of our schemes was also made, allowing us to compare the present scheme with other methods for the integration of KdV-type equations. Where direct comparisons are available, the present scheme appears to be superior to competing methods in terms of accuracy achieved for work expended.

Finally, in § 5, the fruits of our labor are used in an investigative mode in attempting to understand the stability and instablity of solitary-wave solutions of (1.2*a*). A recent theory by Bona *et al.* (1987) has shown these special, travelling-wave solutions of (1.2*a*) to be stable if and only if $p < 4$. However, the theory leaves completely open the manifestation of instability. The experiments reported in § 5 were carried out with a version of the algorithm described in §§ 2–4 that also performs adaptive grid refinement in space and time. They indicate that instability leads to the formation of singularities in the solution.

The last section recounts briefly the earlier accomplishments and then concentrates on formulating a specific conjecture suggested by the numerical experiments in § 5 that the singularity formation is of a similarity type. Further analysis of the numerical results is made in support of this proposition.





PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

## 2. The numerical methods

After explaining notation and reviewing certain preliminaries about splines and Runge–Kutta methods, the numerical algorithms that will hold attention thenceforth are displayed.

For the most part, the notation employed will be that which is currently standard in the numerical analysis of partial differential equations. Each of the function classes that intervenes in our analysis is a Banach space comprised of real-valued functions defined on $\mathbb{R}$ which are periodic of periodic 1. For easy reference, they are recorded here, along with the abbreviation used for their norms.

$L_q$ for $1 \leqslant q < \infty$ is the collection of periodic functions of period 1 which are $q$th-power integrable over [0,1], endowed with the norm

$$|f|_q = \left[\int_0^1 |f(x)|^q \mathrm{d}x\right]^{1/q}.$$

The usual modification applies if $q = \infty$, and the norm on $L_\infty$ is denoted $|\cdot|_\infty$.

$W_q^s$ for $s \geqslant 0$ and $1 \leqslant q \leqslant \infty$ is the Sobolev space of 1-periodic functions which, along with their first $s$ derivatives, belong to $L_q$ . The usual norm on this space is written $\|\cdot\|_{s,q}$ (cf. Adams 1975).

$H^s$ for $s \geqslant 0$ coincides with $W_2^s$ and the norm is abbreviated as simply $\|\cdot\|_s$. These spaces are Hilbert spaces, but this structure will not be used except in the case $s = 0$. These spaces are also defined for fractional or negative values of $s$ (cf. again Adams 1975), but such cases play no essential role in what follows.

$L_2$ has two abbreviations according to the above scheme. Its norm appears so frequently that it will be written unadorned as $\|\cdot\|$. The inner product in $L_2$ also appears frequently and is likewise written unadorned as $(\cdot,\cdot)$.

$C(0,T;X)$ is the space of continuous mappings $u : [0,T] \to X$ where $X$ is any Banach space. Its norm is $\max_{0\leqslant t\leqslant T} \|u(t)\|_X$.

$C^k(0,T;X)$ is the space of $X$-valued functions defined on [0,T] that are $k$-times continuously differentiable.

Before embarking upon a detailed description of the approximation techniques, it is worthwhile recalling the state of the analytical theory pertaining to the initial-value problem (1.2). Many authors have written about (1.2) or its near relatives. Perhaps the best results set in the $L_2$-based Sobolev spaces are those of Kato (1983) whose paper also contains a rather complete bibliography, and the recent work of Bourgain (1993, 1994). While Kato's results are couched in terms of the pure initial-value problem on the whole real line, with initial data having various smoothness and decay properties at infinity, many aspects of his theory go over to the periodic initial-value problems considered here (cf. Bona & Smith (1975) for remarks on the periodic problem for the case $p = 1$). The proofs for the periodic problem are sensibly the same as those already exposed in detail by Kato (1983), and so we content ourselves with a statement of the theorem which will find use here.

**Theorem 2.1.** *Let $u_0 \in H^s$ where $s \geqslant 2$ and let $p$ be a non-negative integer. Then there exists a positive time $t^* = t^*(\|u_0\|_s)$ and a unique function $u \in C(0,t^*;H^s)$ which also lies in $C^k(0,t^*;H^{s-3k})$ for $s - 3k \geqslant -2$, solving (1.2).*




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

*The solution $u$ depends continuously upon $u_0$ in the sense that the mapping associating to $u_0$ the unique solution $u$, whose existence was just asserted, is continuous from $H^s$ to*

$$\bigcap_{s-3k\geqslant -2} C^k(0,t^*;H^{s-3k}).$$

*If $p<4$, $t^*$ may be specified to be any positive number, whereas if $p\geqslant 4, t^*$ may be specified arbitrarily only if $u_0$ is sufficiently small.*

*Remarks.* It is an open question whether or not (1.2) has global smooth solutions for large, smooth initial data if $p\geqslant 4$. Numerical evidence presented in Bona *et al.* (1986) and in § 5 of the present script indicate that smooth solutions form singularities in finite time.

Bourgain's results, mentioned earlier, allow one to consider the cases $p=1,2$ in weaker spaces, but this aspect is not important for the present developments.

Turning now to the description of the numerical scheme, suppose $r$ to be an integer larger than 2 and let $S_h=S_h^r$ connote the space of 1-periodic smooth splines of order $r$ (degree $r-1$) on [0,1] with uniform mesh length $h=1/N$, where $N$ is a positive integer. The finite-dimensional spaces $S_h$ have the following approximation properties. If $v$ is a smooth, 1-periodic function, then there exists a $\chi\in S_h$ such that for any $s$ with $1\leqslant s\leqslant r$,

$$\sum_{j=0}^{s-1} h^j\|v-\chi\|_{j,\alpha}\leqslant c\,h^s\|v\|_{s,\alpha}, \tag{2.1}$$

for $\alpha=2,\infty$, where $c$ is a constant independent of $\chi$, $v$ and $h$. In addition, the spaces $S_h$ possess the following inverse properties. There exists a constant $c$, independent of $h$, such that for all $\chi\in S_h$ and for any $\alpha,\beta$ with $0\leqslant\alpha\leqslant\beta\leqslant r-1$,

$$\|\chi\|_\beta\leqslant c\,h^{-(\beta-\alpha)}\|\chi\|_\alpha,\qquad \|\chi\|_{\alpha,\infty}\leqslant c\,h^{-(\alpha+\frac{1}{2})}\|\chi\|. \tag{2.2}$$

Although it will not figure in the final analysis, the following semi-discretization is useful in motivating the design of the fully discrete schemes to be considered presently. As is customary, the semi-discretization corresponding to the initial-value problem (1.2) is a differentiable map $v_h:[0,t^*]\longrightarrow S_h$ satisfying the relation

$$(v_{ht}+{v_h}^p v_{hx}+\eta\,v_{hx},\chi)=\epsilon\,(v_{hxx},\chi_x) \tag{2.3}$$

for all $\chi\in S_h$ , and for which

$$v_h(0)=\Pi_h u_0, \tag{2.4}$$

where $\Pi_h u_0$ denotes any of a number of approximations of $u_0$ by an element of $S_h$ (e.g. an interpolant, $L_2$-projection, quasi-interpolant, etc.) such that

$$\|\Pi_h u_0-u_0\|\leqslant c\,h^r \tag{2.5}$$

for some constant $c$ which is independent of $h$. Let $P:L_2\longrightarrow S_h$ denote the orthogonal projection of the Hilbert space $L_2$ onto the finite-dimensional subspace $S_h$. Define $F:S_h\longrightarrow S_h$ by requiring that

$$(F(v),\chi)=-(v^p v_x+\eta\,v_x,\chi)+\epsilon\,(v_{xx},\ \chi_x) \tag{2.6}$$

for all $\chi\in S_h$. With this notation, the semi-discretization is a map $v_h:[0,t^*]\longrightarrow S_h$

satisfying

$$v_{ht} = F(v_h) \quad \text{for } 0 \leqslant t \leqslant t^*, \quad v_h(0) = \Pi_h u_0. \tag{2.7}$$

A proof that $v_h$ converges in $L_2$ to $u$ as $h \downarrow 0$ may be constructed along the lines developed in Baker *et al.* (1983) for the special case $p = 1$, at least over any time interval $[0, t^*]$ for which $u$ exists and is sufficiently smooth. In fact, for smooth enough initial data, one shows that

$$\max_{0 \leqslant t \leqslant t^*} \|v_h - u\| \leqslant c\, h^r,$$

where the constant $c$ is independent of $h$.

Upon choosing a basis for $S_h$ and representing $v_h$ in terms of this basis, it is evident that (2.7) may be viewed as a system of ordinary differential equations. As such, one may contemplate using any appropriate method for initial-value problems for systems of ordinary differential equations to approximate its solution $v_h$.

We shall discretize (2.7) in the temporal variable by way of a class of implicit Runge–Kutta methods. General remarks concerning Runge–Kutta-type methods along with a considerable bibliography may be found in the books by Dekker & Verwer (1984) and by Butcher (1987). For $q$ a positive integer, a $q$-stage implicit Runge–Kutta (IRK) method is defined by a tableau

$$\begin{array}{c|c} \mathrm{A} & \tau \\ \hline b^{\mathrm{T}} & \end{array},$$

where $A = (a_{ij})$ is a $q \times q$ matrix and $b = (b_i)$, $\tau = (\tau_i)$ are $q$-vectors. Of particular interest will be the $q$-stage Gauss–Legendre family, a class of IRK methods of collocation type defined as follows. For a fixed $q \geqslant 1$, let $\tau_i$, $1 \leqslant i \leqslant q$, be the zeros of the (shifted) Legendre polynomials $(\mathrm{d}/\mathrm{d}x)^q\ (x(1-x))^q$ of degree $q$ on [0,1] (cf. Dekker & Verwer 1984, p. 85). The $\tau_i$ are distinct and lie in (0,1), while the weights $b_i$ are defined so that the quadrature rule

$$\int_0^1 g(\tau)\, \mathrm{d}\tau \simeq \sum_{j=1}^{q} b_j g(\tau_j) \tag{2.8}$$

is exact for all polynomials $g$ of degree at most $q-1$. The $b_i$ are thus determined as the solution of the Vandermonde system of equations

$$\sum_{j=1}^{q} b_j (\tau_j)^\ell = \frac{1}{\ell + 1}, \quad \text{for} \quad 0 \leqslant \ell \leqslant q-1. \tag{2.9}$$

It is well known that the $b_i$, which coincide with the weights of the Gauss–Legendre quadrature rules on [0,1], are positive and that instead of (2.9) one actually obtains the superaccuracy conditions

$$\sum_{j=1}^{q} b_j (\tau_j)^\ell = \frac{1}{\ell + 1}, \quad \text{for} \quad 0 \leqslant \ell \leqslant 2q-1. \tag{2.10}$$

The $a_{ij}$ are now defined so that the quadrature rules

$$\int_0^{\tau_i} g(\tau) d\tau \simeq \sum_{j=1}^{q} a_{ij} g(\tau_j), \tag{2.11}$$

for $1 \leqslant i \leqslant q$, are exact for polynomials $g$ of degree at most $q-1$. Thus the $a_{ij}$ are obtained as solutions of the system of equations

$$\sum_{j=1}^{q} a_{ij}(\tau_j)^{\ell} = (\ell+1)^{-1}(\tau_i)^{\ell+1}, \tag{2.12}$$

for $0 \leqslant \ell \leqslant q-1$, $1 \leqslant i \leqslant q$. It is not hard to see that $A$ is invertible.

For $q = 1$ the construction just outlined yields the *midpoint method* with $a_{11} = \frac{1}{2}$, $\tau_1 = \frac{1}{2}$, $b_1 = 1$ while for $q = 2$ there results the two-stage Gauss–Legendre method corresponding to the table:

$$\begin{array}{c|cc} \frac{1}{4} & \frac{1}{4} - \frac{1}{2\sqrt{3}} & \frac{1}{2} - \frac{1}{2\sqrt{3}} \\ \frac{1}{4} + \frac{1}{2\sqrt{3}} & \frac{1}{4} & \frac{1}{2} + \frac{1}{2\sqrt{3}} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \tag{2.13}$$

One recognizes that the method represented by the tableau (2.13) is closely allied with the (2,2) Padé approximate. Indeed, in the context of homogeneous, constant-coefficient, linear systems of ordinary differential equations, the $q$-stage Gauss–Legendre method corresponds exactly to the $q$th diagonal Padé approximation $r_q(z)$ to $e^z$. As a consequence, the $q$-stage Gauss–Legendre methods are A-stable, have orders of accuracy $2q$, and are conservative ($|r_q(ix)| = 1$ for all real $x$) when used on such systems of ordinary differential equations. In the context of suitable classes of nonlinear systems of ordinary differential equations, these methods are algebraically stable, conservative, and also of order $2q$ (see Butcher 1975; Crouzeix 1979; Burrage & Butcher 1979). In the next section, some of the special properties of the Gauss–Legendre methods will enter in an important way in the proof of convergence for our fully discrete schemes, and consequently they will be explained in more detail there.

The fully discrete schemes we have in mind simply amount to using the Gauss–Legendre methods on the system of ordinary differential equations that arise from the Galerkin semi-discretization (2.7). More precisely, fix $r$, $q$, and the tableau for the $q$-stage Gauss–Legendre method. Corresponding to the initial data $u_0$, let $u$ be the solution of (1.2) defined at least for $0 \leqslant t \leqslant t^*$. Let $t_n = nk$, $n = 0, 1, \ldots, J$, where $t^* = Jk$. We seek $U^n \in S_h$ for $0 \leqslant n \leqslant J$ which approximates $u^n = u(\cdot, t_n)$ such that

$$U^0 = \Pi_h u_0, \tag{2.14}$$

where $\Pi_h$ is as previously described near (2.4). The approximation $U^{n+1}$ is constructed from $U^n$ by way of the intermediaries $U^{n,i}$ in $S_h$, $1 \leqslant i \leqslant q$, which are the solutions of the $q \times q$ system of nonlinear equations

$$U^{n,i} = U^n + k\sum_{j=1}^{q} a_{ij} F(U^{n,j}), \tag{2.15a}$$

for $1 \leqslant i \leqslant q$, using the formula

$$U^{n+1} = U^n + k\sum_{j=1}^{q} b_j F(U^{n,j}). \tag{2.15b}$$

Since $A$ is invertible, solving for the $F(U^{n,j})$ using the formulas (2.15$a$) and inserting the result in (2.15$b$) gives

$$U^{n+1} = U^n + \sum_{j=1}^{q} b_i (A^{-1})_{ij} \, (U^{n,j} - U^n), \tag{2.15$c$}$$

which is the formula actually used in practice for the computation of $U^{n+1}$.

In §§ 3–4 we study, theoretically and experimentally, various issues related to the convergence, stability and efficient implementation of the schemes defined by (2.14)–(2.15$a$,$b$). In § 5 we shall use one of them (namely the 2-stage Gauss–Legendre method coupled with cubic splines) as the basis for constructing a variable-mesh algorithm that performs adaptive grid refinement in both the spatial and temporal variables.

## 3. Stability and convergence of the base scheme

In this section we study the stability and convergence of what will be called the base scheme, defined by (2.14)–(2.15$a$,$b$). We begin by introducing notation and recounting some preliminary results that will be used in the analysis.

Consider the map $Q : S_h \times S_h \to S_h$ defined for $v, w \in S_h$ as

$$(Q(v,w), \chi) = \frac{1}{p+1}(v^p w, \chi') \quad \text{for all } \chi \in S_h. \tag{3.1}$$

Since $v, w \in S_h \subseteq H^1$, it follows that $v^p w \in H^1$. Therefore, integration by parts applied to (3.1) shows that $(Q(v,w),\chi) = -((v^p w)_x, \chi)/(p+1)$ for $\chi \in S_h$, which is to say that

$$Q(v,w) = -\frac{1}{p+1} P[(v^p w)_x], \tag{3.2}$$

where $P$ is the $L_2$-projection onto $S_h$ as before. Let $\Theta : S_h \to S_h$ be the linear operator defined for $v \in S_h$ by

$$(\Theta v, \chi) = \epsilon(v_{xx}, \chi') - \eta(v_x, \chi) \tag{3.3}$$

for all $\chi \in S_h$, and note that if $r \geqslant 4$ we may write that $\Theta v = -P[\epsilon v_{xxx} + \eta v_x]$ for $v \in S_h$. Finally, define $F : S_h \times S_h \to S_h$ by

$$F(v,w) = Q(v,w) + \Theta v \tag{3.4}$$

for $v, w \in S_h$. Abusing notation mildly, we shall also denote by $Q$ and $F$ the maps from $S_h$ into $S_h$ induced by $Q, F$, respectively, when they act on the diagonal of $S_h \times S_h$. Accordingly, for $v \in S_h$, define $Q(v),\ F(v) \in S_h$ as

$$Q(v) \equiv Q(v,v) = P(-v^p v_x), \tag{3.5}$$

$$F(v) \equiv F(v,v) = Q(v) + \Theta v. \tag{3.6}$$

The definition of $F(v)$ in (3.6) is consistent with the definition of $F$ introduced in (2.6).

The following identities and estimates concerning these mappings will find use presently. By periodicity, it follows that

$$(Q(v), v) = 0 \quad \text{and} \quad (F(v), v) = 0 \tag{3.7}$$





PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

for any $v \in S_h$. Also, by (3.5), one may write

$$Q(v+w) = Q(v) + Q(w) + R(v,w), \tag{3.8}$$

where, for $v, w \in S_h$,

$$R(v,w) = R(w,v) = -\frac{1}{p+1} P\left[\partial_x \left(\sum_{j=1}^{p} \binom{p+1}{j} v^{p+1-j} w^j\right)\right]. \tag{3.9}$$

As a consequence of (3.6) and (3.8) it follows that, for $v,\ w \in S_h$,

$$F(v+w) = F(v) + F(w) + R(v,w). \tag{3.10}$$

From (3.5) one deduces that for $v \in S_h$,

$$\|Q(v)\| \leqslant \|v^p v_x\| \leqslant |v|_\infty^p \|v_x\|, \tag{3.11}$$

while (3.9) implies that there is a constant $C_p$ depending only on $p$, such that for all $v, w \in S_h$,

$$\|R(v,w)\| \leqslant C_p \sum_{j=1}^{p} \left(|v|_\infty^{p-j} \|v_x\|\ |w|_\infty^j + |v|_\infty^{p+1-j} |w|_\infty^{j-1} \|w_x\|\right) \tag{3.12}$$

and

$$|\,(R(v,w),w)\,| \leqslant C_p \max_{1 \leqslant m \leqslant p} \|v\|_{1,\infty}^m \sum_{j=1}^{p} \int_0^1 |w|^{j+1} \mathrm{d}x. \tag{3.13}$$

Following the line of argument laid out by Baker *et al.* (1983) and by Dougalis & Karakashian (1985), approximations to $u(x,t)$ in $S_h$ will frequently be compared to a convenient auxiliary function in $S_h$, namely the *quasi-interpolant* $u_h$ of $u$ defined for $(x,t) \in [0,1] \times [0,t^*]$ by

$$u_h(x,t) = \sum_{j=1}^{N} u(jh,t) \widetilde{\Phi}_j(x),$$

where $\{\widetilde{\Phi}_j\}_{j=1}^N$ is a suitable basis of $S_h$ (cf. Baker *et al.* 1983). It is straightforward to verify that there exist constants $C_i$ independent of $h$ such that

$$\max_{0 \leqslant t \leqslant t^*} \|D_t^i (u_h - u)(\cdot,t)\|_j \leqslant C_i h^{r-j}, \quad \text{for} \quad i = 0,1,2,\ldots, \quad j = 0,1, \tag{3.14}$$

and that

$$\max_{0 \leqslant t \leqslant t^*} \|D_t^i u_h(t)\|_{1,\infty} \leqslant C_i, \quad \text{for} \quad i = 0,1,2,\ldots\ . \tag{3.15}$$

(In (3.14), (3.15) and the sequel, the notation $c$, $c_i$, $C$, etc. will be used to denote positive constants that are independent of the discretization parameters, but which may depend upon the solution in question of (1.2).

As with similar formulas of Baker *et al.* (1983) and Dougalis & Karakashian (1985), it may be proved that $u_h$ satisfies the equations

$$(u_{ht} + u^p u_{hx} + \eta u_{hx}, \chi) - \epsilon\,(u_{hxx}, \chi') = (\psi(t), \chi) \tag{3.16}$$

for all $\chi \in S_h$ and $0 \leqslant t \leqslant t^*$, where the truncation error $\psi(t)$ satisfies

$$\max_{0 \leqslant t \leqslant t^*} \|D_t^i \psi(t)\| \leqslant c_i h^r,\ i = 0,1,2,\ldots\ . \tag{3.17}$$

Differentiating (3.16) shows that for all $\chi \in S_h$ and $0 \leqslant t \leqslant t^*$,

$$\left(D_t^i\left(u_{ht} + u_h^p u_{hx} + \eta u_{hx}\right), \chi\right) - \epsilon\left(D_t^i u_{hxx}, \chi'\right)$$
$$= \left(D_t^i\left[\psi + (u_h^p - u^p)u_{hx}\right], \chi\right), \qquad \text{for } i = 0, 1, 2, \ldots . \tag{3.18}$$

It is also straightforward to see using (3.14) and (3.15) that there exist constants $c_{i,p}$ such that $\|D_t^i\left[(u_h^p - u^p)\, u_{hx}\right]\| \leqslant c_{i,p} h^r$, $i = 0, 1, 2, \ldots$ . Hence, we conclude from (3.17), (3.18) and the triangle inequality that

$$D_t^i u_{ht} = D_t^i F(u_h) + E_0^{(i)}(t), \quad i = 0, 1, 2, \ldots, \quad 0 \leqslant t \leqslant t^*, \tag{3.19}$$

where $E_0^{(i)} : [0, t^*] \longrightarrow S_h$ satisfies

$$\max_{0 \leqslant t \leqslant t^*} \|E_0^{(i)}(t)\| \leqslant c_i h^r, \quad i = 0, 1, 2, \ldots . \tag{3.20}$$

Now we come to some preliminary results and properties of the $q$-stage Gauss–Legendre methods that will be needed in what follows. Note first that the order relations (2.10) and (2.12) may be written, as

$$b^{\mathrm{T}}\, T^\ell e = \frac{1}{\ell + 1}, \quad 0 \leqslant \ell \leqslant 2q - 1, \tag{3.21}$$

$$A\, T^\ell e = \frac{1}{\ell + 1} T^{\ell+1} e, \quad 0 \leqslant \ell \leqslant q - 1, \tag{3.22}$$

respectively, where $T = \operatorname{diag}(\tau_1, \ldots, \tau_q) \in \mathbb{R}^{q \times q}$ and $e = (1, 1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^q$. A simple recursive argument using (3.22) leads to

$$A^j e = T^j e / j!, \quad \text{for} \quad 0 \leqslant j \leqslant q. \tag{3.23}$$

Use will also be made of the following additional order condition for the Gauss–Legendre methods (known as condition (D); cf. Dekker & Verwer 1984; Butcher 1987), namely that

$$b^{\mathrm{T}} T^\ell A = b^{\mathrm{T}}(I - T^{\ell+1})/(\ell + 1), \quad \ell = 0, 1, \ldots, q - 1. \tag{3.24}$$

It is well known (see Dekker & Verwer 1984; Butcher 1987) that the Gauss–Legendre methods are *algebraically stable*, which means that the associated constants $a_{ij}$ and $b_i$ satisfy

$$\left.\begin{aligned} &b_i \geqslant 0,\ 1 \leqslant i \leqslant q, \text{ and the } q \times q \text{ matrix } (m_{ij}), \text{ where} \\ &m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j,\ 1 \leqslant i, j \leqslant q, \text{ is positive semi-definite.} \end{aligned}\right\} \tag{3.25}$$

In fact, these methods are conservative, which means that they actually satisfy

$$M = (m_{ij}) = 0. \tag{3.26}$$

In the proof of existence of our fully discrete approximations, and in several 'diagonalization' arguments in the sequel, we shall use the following additional property of the Gauss–Legendre methods (cf. Dekker & Verwer 1984, p. 157):

$$\left.\begin{aligned} &\text{For each } q \geqslant 1, \text{ there exists a diagonal } q \times q \text{ matrix } D \text{ with positive} \\ &\text{diagonal elements such that } DAD^{-1} \text{ is positive definite on } \mathbb{R}^q. \end{aligned}\right\} \tag{3.27}$$

We now embark upon showing the existence and stability of solutions of the fully

discrete scheme (2.15). For the existence result, we will refer to the following well-known variant of Brouwer's fixed point theorem.

**Lemma 3.1.** *Let $H$ be a real, finite-dimensional Hilbert space with inner product $(\cdot,\cdot)_H$ and norm $\|\cdot\|_H$. Let $g : H \longrightarrow H$ be continuous and suppose there exists $\alpha > 0$ such that $(g(z), z)_H \geqslant 0$ for all $z$ such that $\|z\|_H = \alpha$. Then, there exists $z^*$ in $H$ such that $\|z^*\|_H \leqslant \alpha$ and $g(z^*) = 0$.*

In fact, use will actually be made of the following corollary of lemma 3.1.

**Lemma 3.2.** *Let $\{S, \langle\cdot,\cdot\rangle\}$ be a real, finite-dimensional Hilbert space and let $f : S \longrightarrow S$ be a continuous map such that*

$$\langle f(\varphi), \varphi\rangle \leqslant 0, \quad \text{for all } \varphi \in S. \tag{3.28}$$

*For positive integers $q$, consider the product space $H = S^q$, equipped with the inner product $(\Phi, \Psi)_H = \sum_{i=1}^{q}\langle\varphi_i, \psi_i\rangle$, where $\Phi = (\varphi_i)$, $\Psi = (\psi_i) \in H$, and let $\|\cdot\|_H = (\cdot,\cdot)_H^{1/2}$ be the associated norm. Let $\mathcal{F} : H \longrightarrow H$ denote the diagonal map defined for $\Phi = (\varphi_i) \in H$ by*

$$(\mathcal{F}(\Phi))_i = f(\varphi_i), \quad \text{for } 1 \leqslant i \leqslant q, \tag{3.29}$$

*and let $A$ be an invertible $q\times q$ real matrix for which (3.27) holds. Given $W \in H$ and $k > 0$, consider the map $\mathcal{G} : H \longrightarrow H$ defined for $\Phi \in H$ by*

$$\mathcal{G}(\Phi) = \Phi - W - kA\mathcal{F}(\Phi). \tag{3.30}$$

*Then there exists $\Phi^* \in H$ such that $\mathcal{G}(\Phi^*) = 0$. Moreover, there exists a positive constant $c$ that depends only on $A, D$ and $q$, such that if $\Phi \in H$ is any solution of $\mathcal{G}(\Phi) = 0$, then*

$$\|\Phi\|_H \leqslant c\|W\|_H. \tag{3.31}$$

*Proof.* From (3.30) it follows that for $\Phi \in H$ and $D$ as in (3.27),

$$(D^2A^{-1}\mathcal{G}(\Phi), \Phi)_H = (D^2A^{-1}\Phi, \Phi)_H - (D^2A^{-1}W, \Phi)_H - k\,(D^2\mathcal{F}(\Phi), \Phi)_H\,. \tag{3.32}$$

By (3.28) and (3.29), if $D = \mathrm{diag}(d_i)$ and $\Phi = (\varphi_i)$, then it transpires that

$$(D^2\mathcal{F}(\Phi), \Phi)_H = \sum_{i=1}^{q} d_i^2\langle f(\varphi_i), \varphi_i\rangle \leqslant 0. \tag{3.33}$$

Of course we may write $(D^2A^{-1}\Phi, \Phi)_H$ as $(\widetilde{A}^{-1}D\Phi, D\Phi)_H$ where $\widetilde{A} = DAD^{-1}$. In view of (3.27) and the last remark, it is apparent that there are constants $c'$, $c_1 > 0$ depending only on $D$ and $A$ such that for any $\Phi \in H$,

$$(D^2A^{-1}\Phi, \Phi)_H \geqslant c'\|D\Phi\|_H^2 \geqslant c_1\|\Phi\|_H^2. \tag{3.34}$$

Finally, it is clear that for some positive constant $c_2$, depending only on $D$, $A$ and $q$, and any $\Phi \in H$,

$$|\,(D^2A^{-1}W, \Phi)_H\,| \leqslant c_2\|W\|_H\|\Phi\|_H. \tag{3.35}$$

Combining (3.32)–(3.35) yields the inequality

$$(D^2A^{-1}\mathcal{G}(\Phi), \Phi)_H \geqslant c_1\|\Phi\|_H^2 - c_2\|\Phi\|_H\|W\|_H = c_1\|\Phi\|_H\,(\|\Phi\|_H - c_2\|W\|_H/c_1)\,, \tag{3.36}$$

which holds for any $\Phi \in H$.

It is concluded that if $\alpha = 1 + c_2\|W\|_H/c_1$ and $\Phi \in H$ is such that $\|\Phi\|_H = \alpha$, then $(D^2A^{-1}\mathcal{G}(\Phi), \Phi)_H > 0$. Lemma 3.1 implies that there exists a $\Phi^* \in H$ such that $D^2A^{-1}\mathcal{G}(\Phi^*) = 0$ which in turn means that $\mathcal{G}(\Phi^*) = 0$.

Finally, if $\Phi$ is a solution of $\mathcal{G}(\Phi) = 0$, then (3.32)–(3.35) yield (3.31) with $c = c_2/c_1$. The lemma is thus established. ■

Lemma 3.2 will now be used to guarantee the existence of a solution of the nonlinear system (2.15$a$).

**Proposition 3.1.** *Given $U^n \in S_h$, there are elements $U^{n,i}, 1 \leq i \leq q$, and $U^{n+1}$ in $S_h$ satisfying (2.15$a$) and (2.15$b$), respectively.*

*Proof.* The system (2.15$a$) may be written in the form

$$\mathcal{U}^n = U^n e + k\, A\mathcal{F}(\mathcal{U}^n), \tag{3.37}$$

where $\mathcal{U}^n = (U^{n,1}, \ldots, U^{n,q})^{\mathrm{T}} \in H = (S_h)^q$, $e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^q$, and $\mathcal{F} : H \longrightarrow H$ is defined by $(\mathcal{F}(\mathcal{U}^n))_i = F(U^{n,i})$, $1 \leqslant i \leqslant q$, with $F$ defined in (2.6). For fixed $h$ the continuity of $F$ follows from the inverse assumptions (2.3). In view of (3.7), the existence in $S_h$ of $U^{n,i}$, for $1 \leqslant i \leqslant q$, follows from lemma 3.2 if we identify $\{S, \langle\cdot,\cdot\rangle\}$ with $\{S_h, (\cdot,\cdot)\}$.

The proof of the lemma is thus concluded. ■

As a consequence of (3.25), the proposed numerical method (2.15$a$,$b$) is stable. Indeed, as we now demonstrate, it is also conservative in $L_2$.

**Proposition 3.2.** *Let $\{U^n\}$, $0 \leqslant n \leqslant J$, be a solution of (2.14), (2.15$a$,$b$). Then for $0 \leqslant n \leqslant J$,*

$$\|U^n\| = \|U^0\|. \tag{3.38}$$

*Proof.* Suppose that (3.38) holds for $0 \leqslant n \leqslant J - 1$. Then (2.15$b$) implies that

$$\|U^{n+1}\|^2 = \|U^n\|^2 + 2k \sum_{i=1}^{q} b_i \left(U^n, F(U^{n,i})\right) + k^2 \sum_{i,j=1}^{q} b_i b_j \left(F(U^{n,i}), F(U^{n,j})\right).$$

In the first sum on the right-hand side, replace $U^n$ in the $i$th summand by its expression in terms of the $\{U^{n,j}\}_{j=1}^q$ from (2.15$a$) and then use (3.7) and (3.26) to deduce the formulas

$$\|U^{n+1}\|^2 = \|U^n\|^2 - k^2 \sum_{i,j=1}^{q} \left(b_i a_{ij} + b_j a_{ji} - b_i b_j\right) \left(F(U^{n,i}), F(U^{n,j})\right) = \|U^n\|^2,$$

from which the desired conclusion (3.38) follows. ■

Hence the $q$-stage Gauss–Legendre schemes conserve the discrete analog of the second invariant $I_2 = \|u(\cdot,t)\|^2$ of (1.2). (Of course, the discrete analog of the first invariant $I_1 = \int_0^1 u(x,t)\mathrm{d}x$ is easily seen to be conserved since

$$\int_0^1 U^{n+1}\mathrm{d}x = \int_0^1 U^n \mathrm{d}x$$

follows from the fact that $\int_0^1 F(v)\,\mathrm{d}x = 0$ for $v$ in $S_h$.)

Perhaps the most important step in proving convergence of solutions of the fully discrete scheme (2.14), (2.15$a$,$b$) to the solution of (1.2) is to establish the

*consistency* of the scheme. To this end, we shall presently show that the local error in $L_2$ engendered by our scheme is $O\left(k(k^{q+2}+h^r)\right)$ as $k, h \longrightarrow 0$, provided $q \geqslant 2$. (For $q = 1$, we can show the local error to be $O\left(k(k^2+h^r)\right)$, and this result may be obtained at considerably less technical expense than the higher-order cases $q \geqslant 2$. In consequence, attention will be fixed upon the high-order cases $q \geqslant 2$ in the sequel.) It follows from the local error estimate mentioned above and the stability of the method that

$$\max_{0 \leqslant n \leqslant J} \|U^n - u(\cdot, t^n)\| = O(k^{q+2} + h^r)$$

as $k, h \longrightarrow 0$. Thus the proof yields the classical, optimal temporal rate of convergence $\nu = 2q$ of the Gauss–Legendre methods if $q = 2$, but yields a non-optimal rate for $q \geqslant 3$. For the special case $p = 1$ of the KdV equation, two of us have demonstrated that the optimal temporal order $2q$ is obtained for this scheme for any $q$ (Karakashian & McKinney 1990). The techniques in the last-quoted reference differ in detail from those presented here. Indeed, for $p > 1$, a substantial portion of the effort entailed in obtaining the local error estimates goes in to avoiding stringent mesh restrictions (i.e. restrictions on the ratio of $k$ to a suitable power of $h$) that arise if a straightforward analysis is pursued.

In carrying out the consistency proof for our schemes, several auxiliary functions in $S_h$ are introduced and studied. For $0 \leqslant n \leqslant J-1$, define $V^{n,i}$ for $1 \leqslant i \leqslant q$ and $V^{n+1}$ in $S_h$ by the formulas

$$V^{n,i} = u_h^n + k \sum_{j=1}^{q} a_{ij} F(V^{n,j}), \quad 1 \leqslant i \leqslant q, \tag{3.39}$$

$$V^{n+1} = u_h^n + k \sum_{j=1}^{q} b_j F(V^{n,j}), \tag{3.40}$$

where $u_h(t)$ is the quasi-interpolant of $u(\cdot, t)$ and $u_h(t^n)$ is denoted $u_h^n$. Applying the argument used in the proof of proposition 3.1 shows that the $V^{n,i}$ and $V^{n+1}$ exist in $S_h$. Moreover, from (3.31) and the fact that

$$V^{n+1} = u_h^n + \sum_{i,j=1}^{q} b_i (A^{-1})_{ij} (V^{n,j} - u_h^n), \tag{3.41}$$

which follows from (3.39), (3.40), and (3.15), one may confirm the *a priori* estimate

$$\max_{0 \leqslant n \leqslant J-1} \left( \sum_{i=1}^{q} \|V^{n,i}\| + \|V^{n+1}\| \right) \leqslant c. \tag{3.42}$$

In proposition 3.3 below it will be shown that $\|V^{n+1} - u_h^{n+1}\| = O\left(k(k^{q+2}+h^r)\right)$. To prove this for large $p$ without requiring a stringent mesh condition on $k$ and $h$, it is useful to establish that the $V^{n,i}$ are also uniformly bounded in $L_\infty$. To this end, let

$$M = \max\left( C_0, \max_{(x,t) \in [0,1] \times [0,t^*]} |u(x,t)| \right),$$

where $u$ is the solution of (1.2) and $C_0$ is the constant occurring in (3.15). Define

the function $\widetilde{f} : \mathbb{R} \longrightarrow \mathbb{R}$ by

$$\widetilde{f}(x) = \begin{cases} (-2M)^{p+1} & \text{if} \quad x < -2M, \\ x^{p+1} & \text{if} \quad -2M \leqslant x \leqslant 2M, \\ (2M)^{p+1} & \text{if} \quad x > 2M, \end{cases}$$

as a Lipschitz continuous, bounded extension of the mapping $x \mapsto x^{p+1}$ from $[-2M, 2M]$ to $\mathbb{R}$. Now define the maps $\widetilde{Q}, \widetilde{F} : S_h \longrightarrow S_h$ by the relations

$$\begin{aligned} &(\widetilde{Q}(v), \chi) = (\widetilde{f}(v), \chi')/(p+1) \quad \text{for all } \chi \in S_h, \\ &\widetilde{F}(v) = \widetilde{Q}(v) + \Theta v, \end{aligned} \tag{3.43}$$

for $v \in S_h$, where $\Theta$ is as in (3.3). In addition, define the map $\widetilde{\mathcal{F}} : (S_h)^q \to (S_h)^q$ by $(\widetilde{\mathcal{F}}(\mathcal{V}))_i = \widetilde{F}(v_i)$ $1 \leqslant i \leqslant q$, for $\mathcal{V} = (v_1, \ldots, v_q)^{\mathrm{T}} \in (S_h)^q$.

**Lemma 3.3.** *For each $n$ with $0 \leqslant n \leqslant J-1$, there exists a*

$$\widetilde{\mathcal{V}} = (\widetilde{V}^{n,1}, \ldots, \widetilde{V}^{n,q})^{\mathrm{T}} \in (S_h)^q$$

*satisfying*

$$\widetilde{\mathcal{V}} = u_h^n e + kA\widetilde{\mathcal{F}}(\widetilde{\mathcal{V}}), \tag{3.44}$$

*where $e = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^q$. Moreover, there exists a constant $c^* > 0$ such that if $kh^{-1} < c^*$, then*

$$\max_{n,i} \|u_h(t^{n,i}) - \widetilde{V}^{n,i}\| \leqslant c(k^2 + kh^r). \tag{3.45}$$

*Proof.* We follow the notation of lemma 3.2 and proposition 3.1, letting $H$ stand for $(S_h)^q$ and $\|\cdot\|_H$ denote the $(L_2)^q$-norm on $H$. Define $\widetilde{\mathcal{G}} : H \to H$ by

$$\widetilde{\mathcal{G}}(\Phi) = \Phi - u_h^n e - kA\widetilde{\mathcal{F}}(\Phi)$$

for $\Phi \in H$. It follows that if $\Phi \in H$ and $D$ is as in (3.27), then

$$(D^2 A^{-1}\widetilde{\mathcal{G}}(\Phi), \Phi)_H = (D^2 A^{-1}\Phi, \Phi)_H - (D^2 A^{-1}(u_h^n e), \Phi)_H - k(D^2\widetilde{\mathcal{F}}(\Phi), \Phi)_H. \tag{3.46}$$

From (3.43), (3.3) and periodicity, it follows that $(D^2\widetilde{\mathcal{F}}(\Phi), \Phi)_H = 0$. Combining this relation with (3.46) and arguing as in the proofs of lemma 3.2 and proposition 3.1, we establish the existence of a solution $\Phi$ of $\widetilde{\mathcal{G}}(\Phi) = 0$, that is, of a $\widetilde{\mathcal{V}} = (\widetilde{V}^{n,1}, \ldots, \widetilde{V}^{n,q})^{\mathrm{T}} \in (S_h)^q$ satisfying (3.44).

Let $\eta^{n,i} \in S_h$, $1 \leqslant i \leqslant q$, be defined by

$$u_h(t^{n,i}) = u_h^n + k\sum_{j=1}^{q} a_{ij}\widetilde{F}\left(u_h(t^{n,j})\right) + \eta^{n,i}, \quad \text{for } 1 \leqslant i \leqslant q. \tag{3.47}$$

It follows from the definition of $M$ that $\max_{n,i} |u_h(t^{n,i})|_\infty \leqslant M$. Hence in (3.47) it is seen that $\widetilde{F}\left(u_h(t^{n,j})\right) = F\left(u_h(t^{n,j})\right)$, for $1 \leqslant j \leqslant q$, in view of (3.43), (3.5) and (3.6). Therefore, because $\sum_{j=1}^{q} a_{ij} = \tau_i$ (see (3.23)), Taylor's theorem and (3.15), (3.19) and (3.20) may be used in (3.47) to obtain that

$$\max_{n,i} \|\eta^{n,i}\| \leqslant c(k^2 + kh^r). \tag{3.48}$$

If (3.47) is rewritten in the form

$$\mathcal{U}_h^n = u_h^n e + kA\widetilde{\mathcal{F}}(\mathcal{U}_h^n) + \mathcal{H}^n, \tag{3.49}$$

where $\mathcal{U}_h^n = (u_h(t^{n,1}), \dots, u_h(t^{n,q}))^{\mathrm{T}}$, $\mathcal{H}^n = (\eta^{n.1}, \dots, \eta^{n,q})^{\mathrm{T}}$,

then one obtains from (3.49) and (3.44) that

$$\widetilde{A}^{-1}D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n) = kD(\widetilde{\mathcal{F}}(\widetilde{\mathcal{V}}) - \widetilde{\mathcal{F}}(\mathcal{U}_h^n)) - DA^{-1}\mathcal{H}^n,$$

where $\widetilde{A} = DAD^{-1}$ as before. Taking the inner product in $H$ of this identity with the element $D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n)$ gives

$$\begin{aligned}(\widetilde{A}^{-1}D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n),\ D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n))_H \\ = k(D\left(\widetilde{\mathcal{F}}(\widetilde{\mathcal{V}}) - \widetilde{\mathcal{F}}(\mathcal{U}_h^n)\right), D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n))_H - (DA^{-1}\mathcal{H}^n, D(\widetilde{\mathcal{V}} - \mathcal{U}_h^n))_H.\end{aligned}$$

Combining this with (3.43), (3.48), and the facts that $\widetilde{A}$ is positive definite and $\widetilde{f}$ is Lipschitz leads to the estimate

$$\begin{aligned}\sum_{i=1}^{q} \|\widetilde{V}^{n,i} - u_h(t^{n,i})\|^2 \leqslant ck \sum_{i=1}^{q} \|\widetilde{V}^{n,i} - u_h(t^{n,i})\| \|(\widetilde{V}^{n,i} - u_h(t^{n,i}))_x\| \\ + c(k^2 + kh^r) \sum_{i=1}^{q} \|\widetilde{V}^{n,i} - u_h(t^{n,i})\|.\end{aligned} \tag{3.50}$$

Using the $H^1 - L_2$ inverse assumption (2.2) and taking $kh^{-1}$ to be sufficiently small, the inequality (3.50) implies that

$$\sum_{i=1}^{q} \|\widetilde{V}^{n,i} - u_h(t^{n,i})\| \leqslant c(k^2 + kh^r),$$

from which (3.45) follows.

Thus, if $kh^{-1}$ is sufficiently small, we may combine (3.45) and (2.2) to conclude that

$$|\widetilde{V}^{n,i} - u_h(t^{n,i})|_\infty \leqslant ch^{-1/2}(k^2 + kh^r) \leqslant ch^{3/2}.$$

In particular, because of (3.15), for $h$ sufficiently small it follows that

$$\max_{n,i} |\widetilde{V}^{n,i}|_\infty \leqslant \tfrac{3}{2}M, \tag{3.51}$$

an important consequence of which is that $\widetilde{\mathcal{V}} = (\widetilde{V}^{n,i}, \dots, \widetilde{V}^{n,q})^{\mathrm{T}}$ actually satisfies the equation $\widetilde{\mathcal{V}} = u_h^n e + kA\mathcal{F}(\widetilde{\mathcal{V}})$. Therefore, if $kh^{-1}$ is sufficiently small, we may take it that

$$V^{n,i} = \widetilde{V}^{n,i}, \quad \text{for } 0 \leqslant n \leqslant J-1,\ 1 \leqslant i \leqslant q, \tag{3.52}$$

and hence that the *a priori* $L_\infty$-estimate (3.51) holds for $V^{n,i}$ as well. ■

To prove the main consistency estimate for $V^{n+1}$ we seem to require some additional technical results. In what follows, multi-index notation will frequently be employed. Let $\ell_i$, $1 \leqslant i \leqslant p+1$, be non-negative integers and reserve the symbols $\lambda$ and $|\lambda|$ for the multi-index $(\ell_1, \dots, \ell_{p+1})$ and the scalar $\sum_{i=1}^{p+1} \ell_i$, respectively.

For functions $v_i$, $1 \leqslant i \leqslant p+1$, in $S_h$, and also for any sufficiently smooth, 1-periodic function, extend the definition (3.2) by letting

$$Q(v_1, \ldots, v_{p+1}) = -\frac{1}{p+1} P[(v_1 v_2 \cdots v_{p+1})_x]. \tag{3.53}$$

It will also be convenient to extend the mapping $\Theta$ to act on 1-periodic functions $v$ which are smooth enough, as in (3.3) or, equivalently, as

$$\Theta v = -P(\epsilon v_{xxx} + \eta v_x).$$

**Lemma 3.4.** *Let $E_0^{(i)n} = E_0^{(i)}(t^n)$, $i = 0, 1, 2, \ldots$, be as in (3.19), and let $\alpha_{i\ell} \in S_h$, $1 \leqslant i \leqslant q$, $0 \leqslant \ell \leqslant q+1$, be defined recursively by*

$$\alpha_{i0} = u_h^n, \ 1 \leqslant i \leqslant q,$$

$$\alpha_{i,\ell+1} = \sum_{j=1}^{q} a_{ij} \left\{ \sum_{|\lambda|=\ell} Q\left(\alpha_{j\ell_1}, \ldots, \alpha_{j\ell_{p+1}}\right) + \Theta \alpha_{j\ell} + t_j^{\ell} E_0^{(\ell)n}/\ell! \right\},$$

$$\text{for } \ell = 0, \ldots, q, \ 1 \leqslant i \leqslant q. \tag{3.54}$$

*Then, denoting by $\alpha_\ell$ the vector $(\alpha_{1\ell}, \ldots, \alpha_{q\ell})^{\mathrm{T}} \in (S_h)^q$, we may write*

$$\alpha_\ell = T^\ell e D_t^\ell u_h^n/\ell!, \quad \ell = 0, 1, \ldots, q, \tag{3.55}$$

$$\alpha_{q+1} = A\, T^q e D_t^{q+1} u_h^n/q!. \tag{3.56}$$

*Now define $\widetilde{\alpha}_\ell$, $\ell = 0, 1, \ldots, q+1$, by (3.55) and (3.56) with $u^n = u(\cdot, t^n)$ replacing $u_h^n$, and suppose that $\alpha_{q+2} \in (S_h)^q$ is defined by*

$$\alpha_{i,q+2} = \sum_{j=1}^{q} a_{ij} \left\{ \sum_{|\lambda|=q+1} Q\left(\widetilde{\alpha}_{j\ell_1}, \ldots, \widetilde{\alpha}_{j\ell_{p+1}}\right) + \Theta \widetilde{\alpha}_{j,q+1} \right\}, \tag{3.57}$$

*for $1 \leqslant i \leqslant q$. It follows that*

$$\alpha_{q+2} = \frac{A\, T^{q+1} e}{(q+1)!} P\left(D_t^{q+2} u^n\right) + A\left(\frac{AT^q}{q!} - \frac{T^{q+1}}{(q+1)!}\right) e$$
$$\times \left\{ -P\left[\left(D_t^{q+1} u^n (u^n)^p\right)_x\right] + \Theta\left(D_t^{q+1} u^n\right) \right\}. \tag{3.58}$$

*Proof.* The formulas (3.55) and (3.56) follow by an induction argument using (3.22), (3.6) and (3.19).

To prove (3.58), simply note that by the definition of the $\widetilde{\alpha}_{j\ell}$ for $\ell \leqslant q+1$,

$$\sum_{|\lambda|=q+1} \widetilde{\alpha}_{j\ell_1} \ldots \widetilde{\alpha}_{j\ell_{p+1}} = (p+1)\widetilde{\alpha}_{j,q+1}(u^n)^p + \sum_{\substack{\lambda:\ell_i \leqslant q \\ |\lambda|=q+1}} \widetilde{\alpha}_{j\ell_1} \ldots \widetilde{\alpha}_{j\ell_{p+1}}$$

$$= \frac{(p+1)}{q!} \left( \sum_{m=1}^{q} a_{jm} \tau_m^q \right) (D_t^{q+1} u^n)(u^n)^p$$

$$+ \frac{\tau_j^{q+1}}{(q+1)!} \sum_{\substack{\lambda:\ell_i \leqslant q \\ |\lambda|=q+1}} \frac{(q+1)!}{\ell_1! \ldots \ell_{p+1}!} D_t^{\ell_1} u^n \ldots D_t^{\ell_{p+1}} u^n$$

$$= (p+1)\left[\frac{1}{q!}\sum_{m=1}^{q} a_{jm}\tau_m^q - \frac{1}{(q+1)!}\tau_j^{q+1}\right]\left(D_t^{q+1}u^n\right)(u^n)^p + \frac{\tau_j^{q+1}}{(q+1)!}D_t^{q+1}[(u^n)^{p+1}],$$

from which (3.58) follows if we define $\alpha_{i,q+2}$ by (3.57) and use the equation obtained by differentiating (1.2*a*) $q+1$ times with respect to $t$. ■

**Corollary 3.1.** *With the notation above, we have that*

$$b^{\mathrm{T}}A^{-1}\alpha_\ell = \frac{1}{\ell!}D_t^\ell u_h^n,\ \ell = 1,\dots,q+1, \tag{3.59}$$

*and*

$$b^{\mathrm{T}}A^{-1}\alpha_{q+2} = \frac{1}{(q+2)!}P\left(D_t^{q+2}u^n\right),\quad \text{if } q \geqslant 2. \tag{3.60}$$

*Proof.* Using (3.22) and (3.21), we have by (3.55) that,

$$b^{\mathrm{T}}A^{-1}\alpha_\ell = b^{\mathrm{T}}A^{-1}T^\ell eD_t^\ell u_h^n/\ell! = b^{\mathrm{T}}A^{-1}AT^{\ell-1}e\ D_t^\ell u_h^n/(\ell-1)! = D_t^\ell u_h^n/\ell!,$$

for $\ell = 1,\dots,q$, and by (3.56),

$$b^{\mathrm{T}}A^{-1}\alpha_{q+1} = b^{\mathrm{T}}T^q eD_t^{q+1}u_h^n/q! = D_t^{q+1}u_h^n/(q+1)!.$$

In view of (3.58) and (3.21), it suffices to show that

$$b^{\mathrm{T}}\left(\frac{AT^q}{q!} - \frac{T^{q+1}}{(q+1)!}\right)e = 0,$$

for $q \geqslant 2$ in order that (3.60) be accounted valid. Using (3.24) for $\ell = 0$ and (3.21), one concludes that $b^{\mathrm{T}}AT^q e = b^{\mathrm{T}}(1-T)T^q e = (q+1)^{-1} - (q+2)^{-1} = (q+1)^{-1}(q+2)^{-1}$ since $q \geqslant 2$. Hence by (3.21) again, it follows that

$$b^{\mathrm{T}}\left(\frac{AT^q}{q!} - \frac{T^{q+1}}{(q+1)!}\right)e = \frac{1}{(q+2)!} - \frac{1}{(q+2)!} = 0.$$

The corollary is thus established. ■

Attention is now given to the proof of the main consistency result for the schemes under discussion.

**Proposition 3.3.** *Let $V^{n,i}, V^{n+1}$ be defined by (3.39)–(3.40), where $1 \leqslant i \leqslant q$. Let $k$ be sufficiently small if $p = 1$, $kh^{-1/2}$ be sufficiently small if $p = 2$ and $kh^{-1}$ be sufficiently small if $p \geqslant 3$. Then*

$$\max_{0\leqslant n\leqslant J-1}\|V^{n+1} - u_h^{n+1}\| \leqslant ck(k^2 + h^r),\quad \text{if } q = 1, \tag{3.61}$$

$$\max_{0\leqslant n\leqslant J-1}\|V^{n+1} - u_h^{n+1}\| \leqslant ck(k^{q+2} + h^r),\quad \text{if } q \geqslant 2. \tag{3.62}$$

*Proof.* The proof of (3.61) when $q = 1$ follows from a straightforward modification of the proof for the cases $q \geqslant 2$ and is therefore omitted. Let $q \geqslant 2$ and

$0 \leqslant n \leqslant J-1$. Define $e^{n,i} \in S_h$, for $1 \leqslant i \leqslant q$ by the equations

$$V^{n,i} = \sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell} + e^{n,i}, \quad \text{for } 1 \leqslant i \leqslant q. \tag{3.63}$$

Substituting this expression in (3.39) and performing some straightforward calculations using the definitions of $F$, $Q$ and $\Theta$ leads to

$$\sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell} + e^{n,i} = u_h^n + k \sum_{j=1}^{q} a_{ij} \left\{ \sum_{\ell=0}^{q+1} k^\ell \sum_{|\lambda|=\ell} Q\left(\alpha_{j\ell_1}, \ldots, \alpha_{j\ell_{p+1}}\right) + I_j + \Theta\left( \sum_{\ell=0}^{q+2} k^\ell \alpha_{j\ell} + e^{n,j} \right) \right\}$$

for $1 \leqslant i \leqslant q$, where for $j = 1, \ldots, q$,

$$I_j = -(p+1)^{-1} P \left[ \partial_x \left\{ \sum_{\ell=q+2}^{(q+2)(p+1)} k^\ell \Pi_\ell \left(\alpha_{j0}, \ldots, \alpha_{j,q+2}\right) + \sum_{m=0}^{p} \binom{p+1}{m} \left( \sum_{\ell=0}^{q+2} k^\ell \alpha_{j\ell} \right)^m \left(e^{n,j}\right)^{p+1-m} \right\} \right] \tag{3.64}$$

and $\Pi_\ell$ is a polynomial of degree $p+1$ in $q+3$ independent variables. It follows from (3.54) and (3.57) that for $1 \leqslant i \leqslant q$,

$$e^{n,i} = k \sum_{j=1}^{q} a_{ij} \left\{ k^{q+1} \sum_{|\lambda|=q+1} \left(Q(\alpha_{j\ell_1}, \ldots, \alpha_{j\ell_{p+1}}) - Q(\tilde{\alpha}_{j\ell_1}, \ldots, \tilde{\alpha}_{j\ell_{p+1}})\right) + I_j \right.$$
$$\left. + k^{q+1} \Theta(\alpha_{j,q+1} - \tilde{\alpha}_{j,q+1}) + k^{q+2} \Theta \alpha_{j,q+2} - \sum_{\ell=0}^{q} \frac{k^\ell \tau_j^\ell}{\ell!} E_0^{(\ell)n} + \Theta e^{n,j} \right\}. \tag{3.65}$$

The $L_2$-norm of the $e^{n,i}$ is now estimated from the expression (3.65). To this end, a diagonalization argument as in the proofs of lemmas 3.2 and 3.3 is used. Let again $\tilde{A}^{-1} = DA^{-1}D^{-1}$, where $D = \operatorname{diag}(d_1, \ldots, d_q)$. Viewing (3.65) as a vector equation in $(S_h)^q$ and multiplying both sides by the $q \times q$ matrix $DA^{-1}$ gives the equation

$$\sum_{j=1}^{q} (\tilde{A}^{-1})_{ij} d_j e^{n,j} = k d_i (r_i + I_i + \beta_i + \gamma_i + \Theta e^{n,i}), \tag{3.66}$$

for $1 \leqslant i \leqslant q$, where the $r_i, \beta_i, \gamma_i \in S_h$ are defined by

$$r_i = k^{q+1} \sum_{|\lambda|=q+1} \left(Q\left(\alpha_{i\ell_1}, \ldots, \alpha_{i\ell_{p+1}}\right) - Q\left(\tilde{\alpha}_{i\ell_1}, \ldots, \tilde{\alpha}_{i\ell_{p+1}}\right)\right),$$
$$\beta_i = k^{q+1} \Theta\left(\alpha_{i,q+1} - \tilde{\alpha}_{i,q+1}\right) + k^{q+2} \Theta \alpha_{i,q+2},$$
$$\gamma_i = -\sum_{\ell=0}^{q} \frac{k^\ell \tau_i^\ell}{\ell!} E_0^{(\ell)n}.$$

Taking the $L_2$-inner product of the $i$th equation in (3.66) with $d_i e^{n,i}(x)$ and

summing with respect to $i$ from 1 to $q$ leads to the equation

$$\sum_{i,j=1}^{q} (\widetilde{A}^{-1})_{ij} \left(d_i e^{n,i}, d_j e^{n,j}\right) = k \sum_{i=1}^{q} \left[(r_i, d_i^2 e^{n,i}) + (I_i, d_i^2 e^{n,i}) + (\beta_i, d_i^2 e^{n,i}) + (\gamma_i, d_i^2 e^{n,i})\right]. \tag{3.67}$$

Since $\widetilde{A}^{-1}$ is positive definite, it is easily deduced that there is a constant $c \geqslant 0$ such that

$$\sum_{i,j=1}^{q} (\widetilde{A}^{-1})_{ij} \left(d_i e^{n,i}, d_j e^{n,j}\right) \geqslant c \sum_{i=1}^{q} \|e^{n,i}\|^2. \tag{3.68}$$

But for periodic $\varphi$ in $W_\infty^1$, it is always the case that

$$\left(\partial_x \left(\varphi (e^{n,i})^m\right), e^{n,i}\right) = (1+m)^{-1} \left(\varphi_x, (e^{n,i})^{m+1}\right).$$

Hence one obtains from (3.64) the inequality

$$\begin{aligned}
&\left|(I_i, d_i^2 e^{n,i})\right| \\
&\quad \leqslant (p+1)^{-1} \sum_{\ell=q+2}^{(q+2)(p+1)} k^\ell \left|\left([\Pi_\ell(\alpha_{i0}, \ldots, \alpha_{i,q+2})]_x, d_i^2 e^{n,i}\right)\right| \\
&\qquad + (p+1)^{-1} \sum_{m=1}^{p} \binom{p+1}{m} \frac{1}{(p+2-m)} \left|\left(\left(\left[\sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell}\right]^m\right)_x, d_i^2 (e^{n,i})^{p+2-m}\right)\right| \\
&\quad \leqslant c k^{q+2} \sum_{\ell=q+2}^{(q+2)(p+1)} \left\|\left(\Pi_\ell(\alpha_{i0}, \ldots, \alpha_{i,q+2})\right)_x\right\| \left\|e^{n,i}\right\| \\
&\qquad + c \sum_{m=1}^{p} \left\|\left(\sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell}\right)^m\right\|_{1,\infty} \left|e^{n,i}\right|_\infty^{p-m} \left\|e^{n,i}\right\|^2,
\end{aligned} \tag{3.69}$$

which is valid for $1 \leqslant i \leqslant q$. Applying the Cauchy–Schwarz inequality in $(L_2)^q$ and then the arithmetic-geometric mean inequality to the right-hand side of (3.67), one obtains from (3.68) and (3.69) that

$$\begin{aligned}
\sum_{i=1}^{q} \|e^{n,i}\|^2 \leqslant\ & c k^{2(q+2)} \sum_{i=1}^{q} \sum_{|\lambda|=q+1} \left\| \left[\alpha_{i\ell_1} \cdots \alpha_{i\ell_{p+1}} - \widetilde{\alpha}_{i\ell_1} \cdots \widetilde{\alpha}_{i\ell_{p+1}}\right]_x \right\|^2 \\
& + c k^{2(q+3)} \sum_{i=1}^{q} \sum_{\ell=q+2}^{(q+2)(p+1)} \left\| \left(\Pi_\ell(\alpha_{i0}, \ldots, \alpha_{i,q+2})\right)_x \right\|^2 \\
& + c k \sum_{i=1}^{q} \sum_{m=1}^{p} \left\| \left(\sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell}\right)^m \right\|_{1,\infty} |e^{n,i}|_\infty^{p-m} \|e^{n,i}\|^2 \\
& + c k^{2(q+2)} \sum_{i=1}^{q} \|\Theta(\alpha_{i,q+1} - \widetilde{\alpha}_{i,q+1})\|^2 + c k^{2(q+3)} \sum_{i=1}^{q} \|\Theta \alpha_{i,q+2}\|^2 \\
& + c k^2 \sum_{i=1}^{q} \|\gamma_i\|^2.
\end{aligned} \tag{3.70}$$

To suitably bound the various terms on the right-hand side of (3.70), note that for $\ell_j \leqslant q+1$, the quantity $\left(\alpha_{i\ell_1}\cdots\alpha_{i\ell_{p+1}} - \widetilde{\alpha}_{i\ell_1}\cdots\widetilde{\alpha}_{i\ell_{p+1}}\right)_x$ can be expressed as a sum of terms of the form $f(\alpha_{i\ell}-\widetilde{\alpha}_{i\ell})_x g$ and $f(\alpha_{i\ell}-\widetilde{\alpha}_{i\ell})g$ where $f$ and $g$ are polynomial expressions in $\alpha_{i\ell}, \alpha_{i\ell x}, \widetilde{\alpha}_{i\ell}$ and $\widetilde{\alpha}_{i\ell x}$, for $\ell \leqslant q+1$, which are bounded in $W^1_\infty$ by virtue of (3.55), (3.56) and (3.15). In view of the case $j=1$ in (3.14), one therefore obtains the estimate

$$\| \left[\alpha_{i\ell_1}\cdots\alpha_{i\ell_{p+1}} - \widetilde{\alpha}_{i\ell_1}\cdots\widetilde{\alpha}_{i\ell_{p+1}}\right]_x \| \leqslant ch^{r-1}. \tag{3.71}$$

For a sufficiently smooth, periodic function $v$, the approximation and inverse properties (2.1) and (2.2) appertaining to the spaces $S_h$ imply that $\|Pv\|_{1,\infty} \leqslant c\|v\|_{1,\infty}$ for some constant $c$ which is independent of $v$ and $h$. Also, for sufficiently smooth, periodic $v$, it is straightforward to show that the inequality $\|\Theta(Pv)\| \leqslant c\|v\|_3$ holds, where $c$ is a constant that is independent of $v$ and $h$. From these observations, and using (3.58), one obtains, on the one hand that $\|\alpha_{i,q+2}\|_{1,\infty} \leqslant c$, from which

$$\sum_{\ell=q+2}^{(q+2)(p+1)} \| \left(\Pi_\ell(\alpha_{i0},\ldots,\alpha_{i,q+2})\right)_x \|^2 + \sum_{m=1}^{q} \left\| \left(\sum_{\ell=0}^{q+2} k^\ell \alpha_{i\ell}\right)^m \right\|_{1,\infty} \leqslant c \tag{3.72}$$

follows for $1 \leqslant i \leqslant q$, and on the other that

$$\|\Theta\alpha_{i,q+2}\| \leqslant c \tag{3.73}$$

for $1 \leqslant i \leqslant q$. Note that for $1 \leqslant i \leqslant q$, (3.56), (3.19), (1.2$a$) yield the relations

$$\begin{aligned}
&\Theta\left(\alpha_{i,q+1} - \widetilde{\alpha}_{i,q+1}\right) \\
&\quad = \frac{(A\,T^q e)_i}{q!}\,\Theta\left(D_t^{q+1}(u_h^n - u^n)\right) \\
&\quad = \frac{(A\,T^q e)_i}{q!}\left\{D_t^{q+1}\left(u_{ht}^n - Q(u_h^n)\right)\right) - E_0^{(q+1)n} - P\left[D_t^{q+1}\left(u_t^n + (u^n)^p u_x^n\right)\right]\right\} \\
&\quad = \frac{(A\,T^q e)_i}{q!}\left\{P\,D_t^{q+2}(u_h^n - u^n) + PD_t^{q+1}\left[(u_h^n)^p\,u_{hx}^n - (u^n)^p\,u_x^n\right] - E_0^{(q+1)n}\right\}.
\end{aligned}$$

Furthermore, since

$$\begin{aligned}
&\|D_t^{q+1}\left[(u_h^n)^p\,u_{hx}^n - (u^n)^p\,u_x^n\right]\| \\
&\qquad \leqslant \|D_t^{q+1}\left[\left((u_h^n)^p - (u^n)^p\right)u_{hx}^n\right]\| + \|D_t^{q+1}\left((u^n)^p(u_{hx}^n - u_x^n)\right)\| \leqslant ch^{r-1},
\end{aligned}$$

as is easily seen using (3.14) and (3.15), it follows using (3.14) and (3.20) that

$$\|\Theta\left(\alpha_{i,q+1} - \widetilde{\alpha}_{i,q+1}\right)\| \leqslant ch^{r-1} \tag{3.74}$$

for $1 \leqslant i \leqslant q$. Now observe that by applying (3.63), the definition of the $\alpha_{i\ell}$, $0 \leqslant \ell \leqslant q+2$, and (3.42), it is inferred that $\max_{n,i}\|e^{n,i}\| \leqslant c$. Hence, it is concluded by (2.2) that for $m = 1,\ldots,p$,

$$|e^{n,i}|_\infty^{p-m} \leqslant ch^{-(p-m)/2}\|e^{n,i}\|^{p-m} \leqslant ch^{-(p-1)/2}.$$

Thus it transpires that for $1 \leqslant m \leqslant p$,

$$\max_{n,i}|e^{n,i}|_\infty^{p-m} \leqslant \begin{cases} c & \text{if } p=1, \\ ch^{-1/2} & \text{if } p=2. \end{cases} \tag{3.75$a$}$$

Suppose now that $p \geqslant 3$ and that $kh^{-1}$ is sufficiently small as assumed in lemma 3.3. Then, by (3.52), (3.51), (3.63) and the fact that the $\{\alpha_{i\ell}\}$, $1 \leqslant i \leqslant q$, $0 \leqslant \ell \leqslant q+2$ are uniformly bounded in $L_\infty$ (see (3.72)), it follows that

$$\max_{n,i} |e^{n,i}|_\infty \leqslant c. \tag{3.75b}$$

Using (3.20), (3.70)–(3.74), (3.75*a*) for $p = 1$ or 2 and (3.75*b*) if $p \geqslant 3$, it is deduced under the hypotheses on $k$ and $h$ stated in the proposition for the various values of $p$ that

$$\begin{aligned}\max_n \sum_{i=1}^{q} \|e^{n,i}\|^2 &\leqslant ck^2 \left\{k^{2(q+1)}h^{2(r-1)} + k^{2(q+2)} + h^{2r}\right\} \\ &\leqslant ck^2 \left\{k^{2(q+2)} + h^{2r}\right\},\end{aligned}$$

which yields

$$\max_{n,i} \|e^{n,i}\| \leqslant ck\left(k^{q+2} + h^r\right). \tag{3.76}$$

Finally, since (3.41), (3.63), (3.59), and (3.60) with $e^n = (e^{n,1}, \ldots, e^{n,q})^{\mathrm{T}}$ imply that

$$\begin{aligned}V^{n+1} &= u_h^n + \sum_{\ell=1}^{q+2} k^\ell b^{\mathrm{T}} A^{-1} \alpha_\ell + b^{\mathrm{T}} A^{-1} e^n \\ &= u_h^n + \sum_{\ell=1}^{q+2} \frac{k^\ell}{\ell!} D_t^\ell u_h^n + \frac{k^{q+2}}{(q+2)!} D_t^{q+2}\left[Pu^n - u_h^n\right] + b^{\mathrm{T}} A^{-1} e^n,\end{aligned}$$

we conclude by (3.14), (3.76) that (3.62) holds. ■

We are now in position to prove the main result of this section, which is the following convergence theorem.

**Theorem 3.1.** *Suppose that, as* $h \to 0$,
*if* $p = 1$, $k = O\left(h^{3/2(q+2)}\right)$ *for* $q \geqslant 2$ *or* $k = O\left(h^{3/4}\right)$ *for* $q = 1$,
*if* $p = 2$, $kh^{-1/2}$ *is sufficiently small for* $q \geqslant 2$ *or* $k = O(h^{3/4})$ *for* $q = 1$,
*if* $p \geqslant 3$, $kh^{-1}$ *is sufficiently small for all* $q \geqslant 1$.
*Then for* $h$ *sufficiently small, there exists a unique solution* $U^n$ *of (2.14)–(2.15a,b) such that*

$$\max_{0 \leqslant n \leqslant J} \|U^n - u^n\| \leqslant c(k^2 + h^r) \quad \text{for} \quad q = 1, \tag{3.77}$$

$$\max_{0 \leqslant n \leqslant J} \|U^n - u^n\| \leqslant c(k^{q+2} + h^r) \quad \text{for} \quad q \geqslant 2. \tag{3.78}$$

*Remark.* We will refer to the hypotheses (i), (ii) and (iii) collectively as the *mesh conditions* or *mesh hypotheses.*

*Proof.* The proof of (3.77) is a straightforward modification of that used to establish (3.78), and so it will be omitted. Suppose $q \geqslant 2$ and that $V^{n,i}, V^{n+1}$ are defined by (3.39) and (3.40) for $1 \leqslant i \leqslant q$. Set $\zeta^n = U^n - u_h^n$ and make the induction hypothesis that for some $0 \leqslant n \leqslant J-1$

$$\|\zeta^n\| \leqslant \sigma\, e^{\sigma t_n}\left(k^{q+2} + h^r\right), \tag{3.79}$$

where $\sigma \geqslant 1$ is some constant to be specified below which is independent of $k$, $h$,

and $n$ ($\sigma$ will depend only on the solution and data of (1.2) and the constants pertaining to the numerical method). Obviously, (3.79) holds for $n = 0$ in view of (2.14) provided $\sigma \geqslant c_0$ for some appropriate constant $c_0$. Make the definition $\epsilon^{n,i} = U^{n,i} - V^{n,i}$. As a first step in the proof, it is shown that

$$\|\epsilon^{n,i}\| \leqslant c_1 \|\zeta^n\|, \tag{3.80}$$

for $1 \leqslant i \leqslant q$, where $c_1$ depends only on the solution of (1.2), the constants of the Gauss–Legendre method, and the constants occurring in the approximation and inverse properties of $S_h$. In particular $c_1$ does not depend on $k$, $h$, $n$ or $\sigma$. First consider the case where $p = 1$ or 2. Note that (3.31), (3.38), (2.5) and (3.42) yield

$$\|\epsilon^{n,i}\| \leqslant \|U^{n,i}\| + \|V^{n,i}\| \leqslant c\|U^0\| + \|V^{n,i}\| \leqslant c, \tag{3.81}$$

for $1 \leqslant i \leqslant q$. Subtracting (3.39) from (2.15$a$) and using (3.10) leads to

$$\epsilon^{n,i} = \zeta^n + k \sum_{j=1}^{q} a_{ij} \left( F(\epsilon^{n,j}) + R(V^{n,j}, \epsilon^{n,j}) \right) \tag{3.82}$$

for $1 \leqslant i \leqslant q$. Using (3.63), (3.76), (2.2) and the fact that $\|\alpha_{i,\ell}\|_{1,\infty} \leqslant c$ for $1 \leqslant i \leqslant q$, $0 \leqslant \ell \leqslant q+2$, $0 \leqslant n \leqslant J-1$, as shown in the course of the proof of proposition 3.3, it is concluded using the mesh hypotheses that

$$\max_{i,n} \|V^{n,i}\|_{1,\infty} \leqslant c + ch^{-3/2} \left( k^{q+3} + kh^r \right) \leqslant c. \tag{3.83}$$

It follows by (3.13), (2.2), (3.83) and (3.81) that for $1 \leqslant j \leqslant q$,

$$| \left( R(V^{n,j}, \epsilon^{n,j}), \epsilon^{n,j} \right) | \leqslant c \sum_{i=1}^{p} \int_0^1 |\epsilon^{n,j}|^{i+1} \mathrm{d}x \leqslant ch^{-\frac{1}{2}(p-1)} \|\epsilon^{n,j}\|^2.$$

Since $p = 1$ or 2, we obtain by (3.82), the above equation, a diagonalization argument similar to the one already used in previous proofs, and our mesh hypotheses that (3.80) holds. Tracing through the various constants arising in the proof confirms the claim made about the nature of $c_1$.

Attention is now turned to the case where $p \geqslant 3$. Here, use will be made of the mapping $\widetilde{\mathcal{F}}$ introduced in lemma 3.3. Let $\widetilde{\mathcal{U}}^n = (\widetilde{U}^{n,1}, \ldots, \widetilde{U}^{n,q})^{\mathrm{T}} \in (S_h)^q$ satisfy the equations

$$\widetilde{\mathcal{U}}^n = U^n e + k\, A\, \widetilde{\mathcal{F}}(\widetilde{\mathcal{U}}^n). \tag{3.84}$$

Such a $\widetilde{\mathcal{U}}^n$ exists by virtue of the argument made in the beginning of the proof of lemma 3.3. Recall now, e.g. from (3.52), that $\mathcal{V} = (V^{n,1}, \ldots, V^{n,q})^{\mathrm{T}}$ satisfies the equation

$$\mathcal{V} = u_h^n e + k\, A\, \widetilde{\mathcal{F}}(\mathcal{V}). \tag{3.85}$$

Forming the difference of (3.85) and (3.84), we obtain

$$\widetilde{\mathcal{U}}^n - \mathcal{V} = \zeta^n e + k\, A \left( \widetilde{\mathcal{F}}(\widetilde{\mathcal{U}}^n) - \widetilde{\mathcal{F}}(\mathcal{V}) \right).$$

Hence, a diagonalization argument as in the second half of the proof of lemma 3.3 with computations analogous to those leading to (3.50) yields the inequality

$$\sum_{i=1}^{q} \|\widetilde{U}^{n,i} - V^{n,i}\|^2 \leqslant c\|\zeta^n\| \sum_{i=1}^{q} \|\widetilde{U}^{n,i} - V^{n,i}\| + ck \sum_{i=1}^{q} \|\widetilde{U}^{n,i} - V^{n,i}\| \, \|(\widetilde{U}^{n,i} - V^{n,i})_x\|.$$

It is concluded from this result and (2.2) that, under the imposed mesh condition wherein $kh^{-1}$ is taken to be sufficiently small,

$$\|\widetilde{U}^{n,i} - V^{n,i}\| \leqslant c\|\zeta^n\| \tag{3.86}$$

for $1 \leqslant i \leqslant q$. Hence, the induction hypothesis (3.79) coupled with (2.2) yield

$$|\widetilde{U}^{n,i} - V^{n,i}|_\infty \leqslant c\sigma e^{\sigma t_n} h^{-1/2} \left(k^{q+2} + h^r\right) \leqslant \min(1, \tfrac{1}{2}M), \tag{3.87}$$

where $M$ was defined before lemma 3.3 and the last inequality is valid due to the mesh hypotheses. (After choosing $\sigma$, take $k^{q+2}h^{-1/2}$ small enough to achieve $c\sigma e^{\sigma t^*} h^{-1/2} \left(k^{q+2} + h^r\right) \leqslant \min(1, \frac{1}{2}M)$.) Now, by (3.87) and (3.51)–(3.52), we have

$$|\widetilde{U}^{n,i}|_\infty \leqslant |V^{n,i}|_\infty + |\widetilde{U}^{n,i} - V^{n,i}|_\infty \leqslant 2M$$

for all $n, i$. Thus, the vector $\widetilde{\mathcal{U}}^n$ satisfies (3.84) with $\widetilde{\mathcal{F}}$ replaced by $\mathcal{F}$. Hence, we may as well take $U^{n,i}$ to be $\widetilde{U}^{n,i}$ and then (3.80) follows from (3.86). Tracing the constants appearing in this argument reveals again that $c_1$ depends as advertised after (3.80).

Now let $\epsilon^{n+1} = U^{n+1} - V^{n+1}$. The second step in the proof is to show the *stability estimate*

$$\|\epsilon^{n+1}\| \leqslant (1 + c_2 k)\|\zeta^n\|, \tag{3.88}$$

where $c_2$ is a positive constant that only depends on the same quantities as does $c_1$. From the equations that the $\epsilon^{n,i}$ and $\epsilon^{n+1}$ satisfy, namely, for $1 \leqslant i \leqslant q$,

$$\epsilon^{n,i} = \zeta^n + k \sum_{j=1}^{q} a_{ij} \left(F(U^{n,j}) - F(V^{n,j})\right),$$

and

$$\epsilon^{n+1} = \zeta^n + k \sum_{j=1}^{q} b_j \left(F(U^{n,j}) - F(V^{n,j})\right),$$

one obtains from the algebraic stability of the Gauss–Legendre methods (much as in the proof of proposition 3.2) and formula (3.10) that

$$\begin{aligned} \|\epsilon^{n+1}\|^2 &= \|\zeta^n\|^2 + 2k \sum_{i=1}^{q} b_i \left(\epsilon^{n,i},\ F(U^{n,i}) - F(V^{n,i})\right) \\ &= \|\zeta^n\|^2 + 2k \sum_{i=1}^{q} b_i \left(R(V^{n,i}, \epsilon^{n,i}), \epsilon^{n,i}\right). \end{aligned} \tag{3.89}$$

Consider again the case wherein $p = 1$ or 2. Using the induction hypotheses, (2.2) and (3.80), it is determined that

$$|\epsilon^{n,i}|_\infty \leqslant ch^{-1/2}\|\zeta^n\| \leqslant c\sigma e^{\sigma t_n} h^{-1/2}(k^{q+2} + h^r) \leqslant 1,$$

for $1 \leqslant i \leqslant q$, provided we arrange as before that $c\sigma e^{\sigma t^*} h^{-1/2}(k^{q+2} + h^r) \leqslant 1$ which is again possible because of the mesh conditions. Then, using (3.83) with (3.89) and (3.13) gives

$$\|\epsilon^{n+1}\|^2 \leqslant \|\zeta^n\|^2 + ck \sum_{i=1}^{q} \|\epsilon^{n,i}\|^2,$$

from which (3.88) follows in view of (3.80). If $p \geqslant 3$, use (3.87) and the remark that $U^{n,i}$ can be identified with $\widetilde{U}^{n,i}$ to conclude that $\|\epsilon^{n,i}\| \leqslant 1$ for all $i$. In addition, by (3.51), (3.45), (3.15), and (2.2) we obtain that

$$\|V^{n,i}\|_{1,\infty} \leqslant \|u_h(t^{n,i})\|_{1,\infty} + \|u_h(t^{n,i}) - \widetilde{V}^{n,i}\|_{1,\infty} \leqslant c + ch^{-3/2}(k^2 + kh^r) \leqslant c$$

because of the mesh conditions. It is concluded therefore, by (3.89), (3.13), (3.80) as before, that (3.88) holds again. Finally, (3.88), (3.62) and (3.79) yield

$$\begin{aligned}\|\zeta^{n+1}\| = \|U^{n+1} - u_h^{n+1}\| &\leqslant \|U^{n+1} - V^{n+1}\| + \|V^{n+1} - u_h^{n+1}\| \\ &\leqslant (1 + c_2 k)\|\zeta^n\| + c_3 k(k^{q+2} + h^r) \\ &\leqslant (1 + c_2 k)\sigma e^{\sigma t_n}(k^{q+2} + h^r) + c_3 k(k^{q+2} + h^r) \\ &\leqslant (1 + (c_2 + c_3)k)\,\sigma e^{\sigma t_n}(k^{q+2} + h^r), \end{aligned} \tag{3.90}$$

since $\sigma \geqslant 1$, where $c_3$ denotes the constant $c$ in (3.62). Choose $\sigma = \max(1,\ c_0,\ c_2 + c_3)$. Then $(1 + (c_2 + c_3)k) \leqslant e^{\sigma k}$ and (3.90) shows that

$$\|\zeta^{n+1}\| \leqslant \sigma e^{\sigma t_{n+1}}(k^{q+2} + h^r),$$

i.e. that (3.79) holds for $n + 1$. The inductive step is complete and (3.78) follows from (3.79) and (3.14).

For $n$ in the range $[0, J - 1]$, let $\{U^{n,i}\}$ and $\{W^{n,i}\}$ be two solutions of (2.15$a$) corresponding to the same $U^n$, and let $Y^{n,i} = U^{n,i} - W^{n,i}$. Suppose $p = 1$ or 2. Since

$$Y^{n,i} = k\sum_{j=1}^{q} a_{ij}(F(Y^{n,j}) + R(U^{n,j}, Y^{n,j})),$$

the familiar diagonalization argument yields

$$\sum_{j=1}^{q} \|Y^{n,j}\|^2 \leqslant ck \sum_{j=1}^{q} |\,(R(U^{n,j}, Y^{n,j}), Y^{n,j})\,|. \tag{3.91}$$

Because of the mesh hypotheses, if follows by (2.2), (3.80), (3.79), (3.83) that

$$\|U^{n,j}\|_{1,\infty} \leqslant \|U^{n,j} - V^{n,j}\|_{1,\infty} + \|V^{n,j}\|_{1,\infty} \leqslant ch^{-3/2}(k^{q+2} + h^r) + c \leqslant c$$

for all $n, j$. In addition, (3.31) and (3.38) imply that

$$\|Y^{n,j}\| \leqslant \|W^{n,j}\| + \|U^{n,j}\| \leqslant c\|U^n\| \leqslant c,$$

for all $n, j$. In the usual manner, then, (3.91) yields

$$\sum_{j=1}^{q} \|Y^{n,j}\|^2 \leqslant c\ k\ h^{-(p-1)/2} \sum_{j=1}^{q} \|Y^{n,j}\|^2,$$

which in turn implies that $Y^{n,j} = 0$ (since $p = 1$ or 2) under our mesh conditions. Uniqueness of $U^{n+1}$ follows. If $p \geqslant 3$, let $\mathcal{U}^n = \{U^{n,i}\}$, $\mathcal{W}^n = \{W^{n,i}\}$ and assume that $\max_{n,i} |U^{n,i}|_\infty, \max_{n,i} |W^{n,i}|_\infty \leqslant 2M$. Then both $\mathcal{U}^n$, $\mathcal{W}^n$ satisfy the equations $\mathcal{V} = U^n e + kA\widetilde{\mathcal{F}}(\mathcal{V})$. In consequence, we have

$$\mathcal{U}^n - \mathcal{W}^n = kA\left(\widetilde{\mathcal{F}}(\mathcal{U}^n) - \widetilde{\mathcal{F}}(\mathcal{W}^n)\right),$$

from which, by the familiar diagonalization argument and inverse assumptions,

one deduces that

$$\|\mathcal{U}^n - \mathcal{W}^n\|_H \leqslant ckh^{-1}\|\mathcal{U}^n - \mathcal{W}^n\|_H,$$

whence $\mathcal{U}^n = \mathcal{W}^n$ because of the mesh conditions. Therefore, solutions $\mathcal{U}^n = \{U^{n,i}\}$ of (2.15$a$) are unique if $kh^{-1}$ is sufficiently small and $|U^{n,i}|_\infty \leqslant 2M$. But the latter inequality holds for the accurate solutions $U^{n,i}$ of the first part of the proof since $|U^{n,i}|_\infty \leqslant |U^{n,i} - V^{n,i}|_\infty + |V^{n,i}|_\infty \leqslant ch^{-1/2}(k^{q+2} + h^r) + \frac{3}{2}M \leqslant 2M$ by (3.80), (3.79), (3.51), and (3.52).

The proof of the theorem is now complete. ■

## 4. Computational considerations

From now on, attention will be restricted to the full discretization (2.15$a$,$c$) by means of the two-stage Gauss–Legendre method with constants $a_{ij}$, $b_i$, $\tau_i$, $1 \leqslant i \leqslant j \leqslant 2$, given by the tableau (2.13). In the present section, consideration is given to implementing this method and to reporting on various aspects of it's accuracy. In particular, a summary is made of the outcome of numerical experiments that were performed on the periodic initial-value problem (1.2$a$,$b$) using the time-stepping procedure detailed above together with splines of order $r = 3, 4$ and 6 to represent the spatial structure of solutions. All the calculations were run in VS Fortran on the IBM 3090 at the University of Tennessee, Knoxville, using a code that implements in double precision the scheme described in § 2 and the first part of this section. These computations are used to ascertain the accuracy, stability and computational efficiency of the proposed numerical schemes. Recourse will frequently be made to comparisons of the computer-generated approximations with exact solutions of the partial differential equations. The results of this section may be used to compare the efficiency of the schemes put forth here with that of other numerical methods. It transpires that the schemes examined are the best currently available in terms of accuracy achieved for effort expended, and for the exploratory studies to be reported in § 5.

At each time step we solve the nonlinear system represented by (2.15$a$) using Newton's method as follows. Given $n \geqslant 0$, let $U_0^{n,i} \in S_h$, $i = 1, 2$ be an accurate enough (see below) initial guess for $U^{n,i}$, the solution of (2.15$a$). Then the iterates of Newton's method for (2.15$a$) (called the *outer* iterates for reasons that will become clear presently) $U_j^{n,i}$, $j = 1, 2, \ldots$ ($U_j^{n,i}$ approximates $U^{n,i}$) satisfy the $2 \times 2$ block linear system in $S_h \times S_h$,

$$\begin{bmatrix} I + ka_{11}J(U_j^{n,1}) & ka_{12}J(U_j^{n,2}) \\ ka_{21}J(U_j^{n,1}) & I + ka_{22}J(U_j^{n,2}) \end{bmatrix} \begin{bmatrix} U_{j+1}^{n,1} \\ U_{j+1}^{n,2} \end{bmatrix} = \begin{bmatrix} U^n \\ U^n \end{bmatrix} - kp \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} Q(U_j^{n,1}) \\ Q(U_j^{n,2}) \end{bmatrix} \tag{4.1}$$

for $j = 0, 1, \ldots$, where, for a given $\varphi \in S_h$, the linear mapping $J(\varphi) : S_h \to S_h$ is defined by

$$J(\varphi)\psi = -(p+1)Q(\varphi, \psi) - \Theta\psi, \tag{4.2}$$

and the mappings $Q(\cdot, \cdot)$, $Q(\cdot)$ and $\Theta$ were introduced in (3.1), (3.5) and (3.3), respectively. Upon choosing a basis for $S_h$, it becomes apparent that (4.1) represents a $2N \times 2N$ linear system for the unknown coefficients of the new Newton iterates $U_{j+1}^{n,i}$, $i = 1, 2$, for each $j$. The following device was used to uncouple the two operator equations in (4.1). Evaluating all four entries of the matrix on the

left-hand side of (4.1) at a point $U^* \in S_h$ defined by

$$U^* = \tfrac{1}{2}(U_0^{n,1} + U_0^{n,2}), \tag{4.3}$$

(which makes the operators in the entries of this matrix independent of $j$ and allows them to commute with each other), we may then write (4.1) equivalently as

$$\begin{bmatrix} I + ka_{11}J(U^*) & ka_{12}J(U^*) \\ ka_{21}J(U^*) & I + ka_{22}J(U^*) \end{bmatrix} \begin{bmatrix} U_{j+1}^{n,1} \\ U_{j+1}^{n,2} \end{bmatrix}$$
$$= \begin{bmatrix} U^n \\ U^n \end{bmatrix} - kp \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} Q(U_j^{n,1}) \\ Q(U_j^{n,2}) \end{bmatrix}$$
$$+k \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} J(U^*) - J(U_j^{n,1}) & 0 \\ 0 & J(U^*) - J(U_j^{n,2}) \end{bmatrix} \begin{bmatrix} U_{j+1}^{n.1} \\ U_{j+1}^{n.2} \end{bmatrix}$$

for $j \geqslant 0$, a form that immediately suggests an iterative scheme for approximating $U_{j+1}^{n,i}$, $i = 1, 2$. This scheme generates *inner* iterates denoted by $U_{j+1}^{n,i,\ell}$ for given $n, i, j$ and $\ell = 0, 1, 2, \ldots$ ($U_{j+1}^{n,i,\ell}$ approximates $U_{j+1}^{n,i}$) that are found recursively from the equations

$$\begin{bmatrix} I + ka_{11}J(U^*) & ka_{12}J(U^*) \\ ka_{21}J(U^*) & I + ka_{22}J(U^*) \end{bmatrix} \begin{bmatrix} U_{j+1}^{n,1,\ell+1} \\ U_{j+1}^{n,2,\ell+1} \end{bmatrix} = \begin{bmatrix} r_{j+1}^{n,1,\ell} \\ r_{j+1}^{n,2,\ell} \end{bmatrix} \tag{4.4}$$

for $\ell \geqslant 0$, where

$$r_{j+1}^{n,i,\ell} = U^n - kp \sum_{m=1}^{2} a_{im} Q(U_j^{n,m}) + k \sum_{m=1}^{2} a_{im} \left( J(U^*) - J(U_j^{n,m}) \right) U_{j+1}^{n,m,\ell}.$$

The linear system (4.4) can be solved efficiently as follows: since $a_{12}a_{21} < 0$, it is possible, upon scaling the matrix on the left-hand side of (4.4) by a diagonal similarity transformation, to write it as

$$\begin{bmatrix} I + \frac{1}{4}kJ(U^*) & -kJ(U^*)/4\sqrt{3} \\ kJ(U^*)/4\sqrt{3} & I + \frac{1}{4}kJ(U^*) \end{bmatrix} \begin{bmatrix} U_{j+1}^{n,1,\ell+1} \\ \mu U_{j+1}^{n,2,\ell+1} \end{bmatrix} = \begin{bmatrix} r_{j+1}^{n,1,\ell} \\ \mu r_{j+1}^{n,2,\ell} \end{bmatrix}, \tag{4.5}$$

where $\mu = 2 - \sqrt{3}$. The system (4.5) is equivalent to the single, complex $N \times N$ system

$$(I + k\beta\, J(U^*))\, Z = R, \tag{4.6}$$

where $\beta = \frac{1}{4} + \mathrm{i}/4\sqrt{3}$, and where $Z$ and $R$ are complex-valued functions with real and imaginary parts in $S_h$ which depend upon $n, \ell$ and $j$ and are given by

$$Z = U_{j+1}^{n,1,\ell+1} + i\,\mu\, U_{j+1}^{n,2.\ell+1}, \quad R = r_{j+1}^{n,1,\ell} + i\,\mu\, r_{j+1}^{n,2,\ell}. \tag{4.7}$$

The complexification (4.6) of (4.5) may be regarded as the analog in the nonlinear case of the idea used in the context of homogeneous, linear, time-independent-coefficient parabolic partial differential equations discretized in time by the (2,2) Padé scheme, in Baker *et al.* (1977) and Fairweather (1978).

In practice, only a finite number of outer and inner iterates are computed at each time step. Specifically, for $i = 1, 2$, $n \geqslant 0$, we compute approximations to the outer iterates $U_j^{n,i}$ for $j = 1, \ldots, J_{\text{out}}$, for some small positive integer $J_{\text{out}}$. For




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

each $j$, $0 \leqslant j \leqslant J_{\text{out}} - 1$, $U_{j+1}^{n,i}$ is approximated by the last inner iterate $U_{j+1}^{n,i,J_{\text{inn}}}$ of the sequence of inner iterates $U_{j+1}^{n,i,\ell}$, $0 \leqslant \ell \leqslant J_{\text{inn}}$ that satisfy linear systems of the form (4.6); consideration of initiating values is provided below. In practice, it was observed that taking $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ was sufficient to conserve the accuracy and stability properties of our schemes in almost all cases that arose, provided suitable starting values were used. The relevant numerical experiments will be described in the next section.

Given $U^n$, the required starting values $U_0^{n,i}$ for the outer (Newton) iteration were computed by extrapolation from previous values as

$$U_0^{n,i} = \alpha_{0,i} U^n + \alpha_{1,i} U^{n-1} + \alpha_{2,i} U^{n-2} + \alpha_{3,i} U^{n-3}, \tag{4.8}$$

for $i = 1, 2$, where the coefficients $\alpha_{j,i}$ are such that $U_0^{n,i}$ is the value at $t = t^{n,i}$ of the Lagrange interpolating polynomial of degree at most 3 in $t$ that interpolates to the data $U^{n-j}$ at the four points $t^{n-j}$, $0 \leqslant j \leqslant 3$. (If $0 \leqslant n \leqslant 2$, we use the same linear combination, putting $U^j = U^0$ if $j < 0$, and compensate for the reduced accuracy by increasing the number of iterations to $J_{\text{out}} = J_{\text{inn}} = 3$. Here, the function $U^0$ is taken to be the $L_2$-projection of $u_0$ onto $S_h$).

The complete algorithm for computing one step of the method, that is, determining $U^{n+1}$ given $U^n$, is then as follows.

(i) Compute $U_0^{n,i}$, $i = 1, 2$, by (4.8).
(ii) Set $U^* = \frac{1}{2}(U_0^{n,1} + U_0^{n,2})$.
(iii) For $j = 0, 1, \ldots, J_{\text{out}} - 1$:
  Initialize $U_{j+1}^{n,i,0} = U_j^{n,i}$, $i = 1, 2$
  For $\ell = 0, 1, \ldots, J_{\text{inn}} - 1$:
    Compute $U_{j+1}^{n,i,\ell+1}$, $i = 1, 2$ solving the linear system (4.6)
  Set $U_{j+1}^{n,i} = U_{j+1}^{n,i,J_{\text{inn}}}$, $i = 1, 2$
(iv) Set $U^{n,i} = U_{J_{\text{out}}}^{n,i}$, $i = 1, 2$
(v) Compute $U^{n+1}$ from $U^n$, $U^{n,i}$, $i = 1, 2$ via (2.15*c*). (4.9)

It is clear from the outline (4.9) that the heart of the computation is forming the right-hand side $R$ of the linear system (4.6) for each $j$ and $\ell$ and then solving for $Z$; note that the operator $I + k\beta J(U^*)$ is independent of the inner and outer iteration indices and is hence formed and decomposed once at each time step. (In practice, if the solution is not changing rapidly with time, the same $U^*$ may be safely used for, say, 10 to 20 time steps, without increasing $J_{\text{inn}}$ or $J_{\text{out}}$).

It is outside the scope of this paper to analyse rigorously the convergence of the doubly iterative scheme (4.9) to the solution of (2.15*a*). Such an analysis can be made along the lines of the analogous proof in Dougalis & Karakashian (1985). Later in this section it is verified *experimentally* that, for $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$, the resulting overall time-stepping procedure is stable and has an $L_2$-error bound of $O(k^4 + h^r)$ in the cases of current practical interest $r = 3, 4, 6$. We content ourselves here with pointing out the crux of the matter, namely that for $k$ sufficiently small, the operator in the linear system (4.6) is invertible if the scheme is stable and the previous values $U^{n-j}$, $0 \leqslant j \leqslant 3$ are sufficiently accurate.

The implementation of the algorithm (4.9) as a computer program follows the general plan laid out in the case $p = 1$ in § 3 of Bona *et al.* (1986). The only exception is that general nonlinear terms of the form $(\psi^p \phi, \chi)$ where $\psi, \phi, \chi \in S_h$




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 1. *CPU time (seconds) per mesh interval per time step*

| | $J_{\text{out}}=1,\ J_{\text{inn}}=2$ | | | $J_{\text{out}}=2,\ J_{\text{inn}}=2$ | | |
|---|---|---|---|---|---|---|
| $p$ | $r=3$ | $r=4$ | $r=6$ | $r=3$ | $r=4$ | $r=6$ |
| 1 | 2.56 (−4) | 3.94 (−4) | 7.57 (−3) | 3.94 (−4) | 6.02 (−4) | 1.16 (−3) |
| 2 | 4.29 (−4) | 6.87 (−4) | 1.43 (−3) | 6.45 (−4) | 1.02 (−3) | 2.09 (−3) |
| 3 | 4.76 (−4) | 8.05 (−4) | 1.67 (−3) | 7.22 (−4) | 1.22 (−3) | 2.51 (−3) |
| 4 | 5.28 (−4) | 8.75 (−4) | 1.85 (−3) | 8.12 (−4) | 1.34 (−3) | 2.83 (−3) |
| 5 | 5.76 (−4) | 9.92 (−4) | 2.12 (−3) | 9.06 (−4) | 1.56 (−3) | 3.27 (−3) |

are now evaluated using Gaussian quadrature with a sufficient number of nodes that the quadrature is exact.

We omit reporting the detailed operation counts and timings of various parts of the complete algorithm, but some of this data is summarized in table 1. Recorded there is the CPU time per time step per spatial mesh used by the algorithm for the methods corresponding to the two cases $\{J_{\text{out}}=1, J_{\text{inn}}=2\}$ and $\{J_{\text{out}}=2, J_{\text{inn}}=2\}$ for $r=3,4,6$ and $p=1,2,3,4,5$ when run in double precision on an IBM 3090 at the University of Tennessee, Knoxville. While these aspects are not labored here, it deserves remark that interesting issues arise under this computer-science aegis. The authors stand ready to provide the interested reader with further details.

Attention is now turned to the order of accuracy, stability and related aspects of the methods when they are applied to the generalized KdV equation. In order to build confidence in the methods in view of the more challenging numerical experiments to be described in § 5, we investigate briefly their accuracy in some well-controlled experiments, namely in approximating solitary-wave solutions. The initial-value problem for solitary waves comprises equation (1.2$a$), reproduced here for convenience, with a solitary wave as initial data,

$$\left.\begin{aligned} u_t + \eta u_x + u^p u_x + \epsilon u_{xxx} &= 0, \\ u(x,0) &= A\,\mathrm{sech}^{2/p}[K(x-x^0)], \end{aligned}\right\} \tag{4.10$a$}$$

whose solution is given by

$$u(x,t) = A\,\mathrm{sech}^{2/p}[K(x-x^0)-\omega t], \tag{4.10$b$}$$

for $x,t\in\mathbb{R}$, where

$$K = p\,(A^p/2\epsilon(p+1)(p+2))^{1/2}$$

and

$$\omega = K\,[\eta + 2A^p/(p+1)(p+2)]\,.$$

Unless otherwise specified, the waves were initially centred at $x^0=\frac{1}{2}$. These solutions are stable if $1\leqslant p<4$ and unstable for $p\geqslant 4$ (see § 5). However, in the order-of-accuracy experiments for $p\geq 5$, small enough amplitudes $A$ and time intervals $[0,T]$ were chosen so that the instability did not manifest itself during the calculations. Although (4.10$b$) is a solution to the pure initial-value problem (4.10$a$) on the whole real line, if $K$ is sufficiently large it may be considered as a good approximation to a solution corresponding to periodic boundary conditions

Table 2. *Errors $E(t)$ and spatial rates of convergence for quadratic splines ($r = 3$)*

| | | $t = 0.1$ | | $t = 0.5$ | | $t = 1.0$ | |
|---|---|---|---|---|---|---|---|
| $h^{-1}$ | $k^{-1}$ | $E(t)$ | rate | $E(t)$ | rate | $E(t)$ | rate |
| 96 | 1000 | 0.9516 (−3) | 3.42 | 0.1170 (−2) | 3.44 | 0.1363 (−2) | 3.45 |
| 144 | 1000 | 0.2375 (−3) | 3.43 | 0.2900 (−3) | 3.55 | 0.3368 (−3) | 3.62 |
| 192 | 1000 | 0.8853 (−4) | 3.33 | 0.1044 (−3) | 3.48 | 0.1189 (−3) | 3.58 |
| 256 | 1000 | 0.3393 (−4) | 3.24 | 0.3831 (−4) | 3.39 | 0.4243 (−4) | 3.50 |
| 320 | 1000 | 0.1646 (−4) | 3.15 | 0.1796 (−4) | 3.28 | 0.1942 (−4) | 3.39 |
| 512 | 1000 | 0.3743 (−5) | 3.08 | 0.3844 (−5) | 3.14 | 0.3941 (−5) | 3.20 |
| 768 | 1000 | 0.1075 (−5) | 3.01 | 0.1074 (−5) | 3.01 | 0.1076 (−5) | 3.01 |
| 1024 | 1500 | 0.4519 (−6) | | 0.4521 (−6) | | 0.4522 (−6) | |

since the tails of the solitary wave decay exponentially and are zero to machine accuracy within the period (see Bona 1981*b*).

One result emerging from the experiments is that the rates of convergence for $p > 1$ are essentially the same as those for $p = 1$. Thus in the following we restrict ourselves to reporting results for the KdV equation $p = 1$. As in Bona *et al.* (1986, § 4), we took $\eta = 1$ and used the parameters $\epsilon = 0.2058 \times 10^{-4}$, $x^0 = \frac{1}{2}$ and $A = 0.2275$.

First, the convergence rates of the scheme in both the spatial and temporal variables with $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ were investigated for $r = 3, 4, 6$. The measure of error used was the normalized $L_2$-norm given by

$$E(t) = \|U^n - u(\cdot, t)\|/\|u_0\| \tag{4.11}$$

if $t = nk$, $n = 1, 2, \ldots$, whereas for other values of $t$, $E$ is defined by linear interpolation. To determine experimentally the spatial convergence rate, the approximate solution was determined for $0 \leqslant t \leqslant 1$ using values of $N = h^{-1}$ ranging from 96 to 1024 (from 96 to 768 when $r = 6$). For these runs, very small time steps were taken to render the temporal error negligible. The observed error as defined in (4.11) was recorded at $t = 0.1$, 0.5 and 1. The convergence rate corresponding to two different runs with spatial meshes $h_1$ and $h_2$ and corresponding errors $E_1$ and $E_2$ is defined to be $\log(E_1/E_2)/\log(h_1/h_2)$, as usual. The convergence rates derived from the runs mentioned above are presented along with the associated errors in tables 2, 3 and 4 which correspond to the values $r = 3, 4$ and 6, respectively. A rate shown in the tables at a given value of $N$ is computed using the values of $E$ and $h$ for that value of $N$ and those corresponding to the value of $N$ following it in the table.

From these tables it is safe to conclude that the convergence rates for the spatial $L_2$-error between the exact solution and the approximation produced by our computer code are indeed 3, 4 and 6 for quadratic, cubic and quintic splines, respectively.

The experimental determination of the temporal accuracy is a somewhat more delicate matter because long runs with very small values of $h$ are prohibitively expensive, both in terms of run time and storage. We took three values of $h$, namely $h^{-1} = 192$, 384 and 480, and computed solutions to the periodic initial-




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 3. *Errors $E(t)$ and spatial rates of convergence for cubic splines ($r = 4$)*

| $h^{-1}$ | $k^{-1}$ | $t = 0.1$ $E(t)$ | rate | $t = 0.5$ $E(t)$ | rate | $t = 1.0$ $E(t)$ | rate |
|---|---|---|---|---|---|---|---|
| 96 | 5000 | 0.1777 (−3) | 4.79 | 0.1806 (−3) | 4.81 | 0.1847 (−3) | 4.84 |
| 144 | 5000 | 0.2547 (−4) | 4.42 | 0.2566 (−4) | 4.43 | 0.2593 (−4) | 4.45 |
| 192 | 5000 | 0.7145 (−5) | 4.23 | 0.7168 (−5) | 4.24 | 0.7199 (−5) | 4.25 |
| 256 | 5000 | 0.2113 (−5) | 4.14 | 0.2115 (−5) | 4.14 | 0.2118 (−5) | 4.15 |
| 320 | 5000 | 0.8387 (−6) | 4.07 | 0.8391 (−6) | 4.07 | 0.8397 (−6) | 4.07 |
| 512 | 5000 | 0.1237 (−6) | 4.03 | 0.1237 (−6) | 4.03 | 0.1238 (−6) | 4.03 |
| 768 | 5000 | 0.2415 (−7) | 4.01 | 0.2415 (−7) | 4.01 | 0.2416 (−7) | 4.01 |
| 1024 | 7500 | 0.7611 (−8) | | 0.7611 (−8) | | 0.7611 (−8) | |

Table 4. *Errors $E(t)$ and spatial rates of convergence for quintic splines ($r = 6$)*

| $h^{-1}$ | $k^{-1}$ | $t = 0.1$ $E(t)$ | rate | $t = 0.5$ $E(t)$ | rate | $t = 1.0$ $E(t)$ | rate |
|---|---|---|---|---|---|---|---|
| 96 | 2000 | 0.2131 (−4) | 8.16 | 0.2126 (−4) | 8.16 | 0.2143 (−4) | 8.18 |
| 144 | 3000 | 0.7777 (−6) | 7.12 | 0.7778 (−6) | 7.12 | 0.7781 (−6) | 7.12 |
| 192 | 5000 | 0.1002 (−6) | 6.70 | 0.1002 (−6) | 6.70 | 0.1002 (−6) | 6.70 |
| 256 | 8000 | 0.1460 (−7) | 6.45 | 0.1460 (−7) | 6.45 | 0.1460 (−7) | 6.45 |
| 320 | 10,000 | 0.3459 (−8) | 6.25 | 0.3459 (−8) | 6.25 | 0.3460 (−8) | 6.25 |
| 512 | 25,000 | 0.1833 (−9) | 6.11 | 0.1833 (−9) | 6.04 | 0.1833 (−9) | 6.07 |
| 768 | 22,500 | 0.1539 (−10) | | 0.1581 (−10) | | 0.1567 (−10) | |

value problem with solitary-wave initial data up to $T = 1$ for various values of $k$. It was found that as the value of $k$ decreases, the $L_2$-error $E(T)$ ceases to decrease at a certain point because the temporal error becomes much smaller than the spatial error at which point their combined effect cannot be distinguished from the spatial error. It is thus hard to see the asymptotic rate of the temporal error. A way around this problem is now explained. For a fixed value of $h$, we made a reference calculation with a small value $k = k_{\text{ref}}$. We took $k_{\text{ref}} = \frac{1}{20}h$, a value well below the threshold of about $\frac{1}{3}h$ to $\frac{1}{6}h$ where $E(T)$ stabilized as a function of $k$. The approximate solution $U^m = U^m(h, k_{\text{ref}})$ determined by the reference simulation differs from the exact solution by an error that is almost purely from the spatial discretization. For the same values of $h$, we then define a modified error associated to values of $k$ that are larger than $k_{\text{ref}}$, namely

$$E^*(t) = \|U^n(h,k) - U^m(h,k_{\text{ref}})\|/\|u_0\| \tag{4.12}$$

where $t = nk = mk_{\text{ref}}$. It transpires that for small values of $k$ which are nevertheless considerably larger than $k_{\text{ref}}$, the expected temporal rate of convergence is visible because subtracting $U^m(h, k_{\text{ref}})$ from $U^n(h,k)$ essentially cancels the spatial error inherent in the latter approximation. The results of these comparisons




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

are shown in table 5 which refer to splines of order 3, 4 and 6, respectively. For each simulation, the tables show $E(T)$, $E^*(T)$ and the error $E(T)$ associated with the reference approximation $U^m(h, k_{\text{ref}})$ by means of which $E^*(T)$ is computed. The expected temporal rate of convergence $\sigma = 4$ emerges clearly from these experiments for all values of $h$ and $r$ tested. In addition, it does not appear that an upper bound on some quantity such as $kh^{-1}$ is needed to insure stability of the scheme.

The next set of experiments reported here feature computing the approximate solution of (4.10) up to $T = 5$ in order to study various kinds of errors pertinent to the numerical approximation of waves and, also, to assess the effect of the number of outer and inner iterations on the accuracy of the method over longer temporal intervals.

To begin with, based on experiments not reported here, we concluded that the combination $\{J_{\text{out}} = 1,\ J_{\text{inn}} = 1\}$ was unstable. It also became evident that it is unnecessary to perform more than two outer iterations since no extra accuracy seems to be gained by further pursuing this aspect of the algorithm. Moreover, in either of the cases $J_{\text{out}} = 1$ or $J_{\text{out}} = 2$, the experiments with $J_{\text{inn}} = 2$ and $J_{\text{inn}} = 3$ gave essentially identical results. Hence, in what follows, attention will be restricted to the two cases $\{J_{\text{out}} = 1,\ J_{\text{inn}} = 2\}$ and $\{J_{\text{out}} = 2,\ J_{\text{inn}} = 2\}$. We studied the various errors associated with these two combinations in three runs with discretization parameters suitably chosen to yield errors $E(t)$ of magnitudes on the order of $10^{-1}$, $10^{-3}$ and $10^{-5}$, respectively.

In table 6, errors are shown at times $t_i = i,\ 1 \leqslant i \leqslant 5$ for the methods corresponding to the two aforementioned combinations of $J_{\text{out}}$ and $J_{\text{inn}}$, both run with parameters $r = 3,\ N = 44$ and $J = 500$ time steps on the temporal interval [0,5]. These values were calculated to yield errors $E(t)$ of order $10^{-1}$ for $0 \leqslant t \leqslant 5$. For these runs, various measures of error are computed for the approximations to the solitary wave. In addition to the normalized $L_2$-error, we also provide the $L_2$-based *shape error*, the *phase error* and the *amplitude error*, quantities which are now defined. The shape error $S^n$ is defined for each time step $n = 0, 1, \ldots,\ J$ as follows. Fix $n$ and consider the quantity

$$\xi^2(\tau) = \int_0^1 |u(x,\tau) - U^n(x)|^2\,\mathrm{d}x \Big/ \int_0^1 u^2(x,0)\,\mathrm{d}x$$

where $u(x,t)$ is given in (4.10*b*) and $U^n$ is the computed solution at the $n$th time step. Let $\tau^* = \tau^*(n)$ denote the value of $\tau$ near $nk$ where $\xi^2(\tau)$ takes its minimum value. If $U^n$ resembles a solitary wave in shape, it follows that $\tau^*$ is well defined. Then $S^n = \xi^2(\tau^*)$ measures by how far the computed solution differs from the original solitary wave as regards its shape, as measured by the normalized $L_2$-norm. The phase error $P^n$ at any time step $n$ with $0 \leqslant n \leqslant J$ is defined to be $nk - \tau^*(n)$. This quantity measures the error in the position at which the wave is located. The amplitude error $A^n$ is defined as $(A - U^n_{\max})/A$ where $A$ is the amplitude parameter in (4.10*b*) and $U^n_{\max}$ is the maximum value of $U^n(x)$.

In figure 1 we show $E(t)$ and the shape error as functions of time for the data in table 6. Table 7 and figure 2 show the analogous data for $r = 4,\ N = 96,\ J = 900$, parameters designed to yield an $E(t)$ of order $10^{-3}$ for $0 \leqslant t \leqslant 5$. Finally, table 8 and figure 3 correspond to $r = 6,\ N = 128,\ J = 2600$, parameters that yield an $E(t)$ whose order of magnitude is about $10^{-5}$. Tables 6 and 7 and figures 1

Table 5. *Temporal rates of convergence at* $T = 1$ *for* $h^{-1} = 480$ *and* $r = 3, 4, 6$

quadratic splines

| $k^{-1}$ | $kh^{-1}$ | $E(T)$ | $E^*(T)$ | rate |
|---|---|---|---|---|
| 240 | 2 | 0.1348(−3) | 0.1371(−3) | 4.08 |
| 480 | 1 | 0.7250(−5) | 0.8089(−5) | 4.01 |
| 640 | $\frac{3}{4}$ | 0.4480(−5) | 0.2554(−5) | 4.00 |
| 720 | $\frac{2}{3}$ | 0.4547(−5) | 0.1594(−5) | 4.00 |
| 960 | $\frac{1}{2}$ | 0.4857(−5) | 0.5041(−6) | 4.00 |
| 1440 | $\frac{1}{3}$ | 0.5028(−5) | 0.9954(−7) | 4.00 |
| REF | | | | |
| 9600 | $\frac{1}{20}$ | 0.5074(−5) | | |

cubic splines

| $k^{-1}$ | $kh^{-1}$ | $E(T)$ | $E^*(T)$ | rate |
|---|---|---|---|---|
| 240 | 2 | 0.1371(−3) | 0.1371(−3) | 4.08 |
| 480 | 1 | 0.8086(−5) | 0.8089(−5) | 4.01 |
| 640 | $\frac{3}{4}$ | 0.2554(−5) | 0.2554(−5) | 4.00 |
| 720 | $\frac{2}{3}$ | 0.1597(−5) | 0.1594(−5) | 4.00 |
| 960 | $\frac{1}{2}$ | 0.5250(−6) | 0.5041(−6) | 4.00 |
| 1440 | $\frac{1}{3}$ | 0.1868(−6) | 0.9954(−7) | 4.00 |
| REF | | | | |
| 9600 | $\frac{1}{20}$ | 0.1607(−6) | | |

quintic splines

| $k^{-1}$ | $kh^{-1}$ | $E(T)$ | $E^*(T)$ | rate |
|---|---|---|---|---|
| 240 | 2 | 0.1371(−3) | 0.1371(−3) | 4.08 |
| 480 | 1 | 0.8089(−5) | 0.8089(−5) | 4.01 |
| 640 | $\frac{3}{4}$ | 0.2553(−5) | 0.2554(−5) | 4.00 |
| 720 | $\frac{2}{3}$ | 0.1594(−5) | 0.1594(−5) | 4.00 |
| 960 | $\frac{1}{2}$ | 0.5041(−6) | 0.5041(−6) | 4.00 |
| 1440 | $\frac{1}{4}$ | 0.9959(−7) | 0.9954(−7) | 4.00 |
| REF | | | | |
| 9600 | $\frac{1}{20}$ | 0.2770(−9) | | |

and 2 indicate that for the accuracy levels $10^{-1}$ and $10^{-3}$ there is a significant difference in the growth of $E(t)$ with $t$ between the methods corresponding to the two different combinations of inner and outer iterations. Specifically, for $T = 5$ and errors at the $10^{-1}$ level, the experiments run with two outer iterations gave





PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 6. *Long-time errors,* $r = 3$, $N = 44$, $J = 500$, $T = 5$, $E$ *of order* $10^{-1}$

| | $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ | | | | $J_{\text{out}} = 2$, $J_{\text{inn}} = 2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $t$ | $E(t)$ | shape error | phase error | amplitude error | $E(t)$ | shape error | phase error | amplitude error |
| 1 | 0.592(−1) | 0.403(−1) | 0.164(−2) | 0.352(−1) | 0.577(−1) | 0.403(−1) | 0.156(−2) | 0.338(−1) |
| 2 | 0.872(−1) | 0.408(−1) | 0.292(−2) | 0.281(−1) | 0.817(−1) | 0.406(−1) | 0.268(−2) | 0.259(−1) |
| 3 | 1.213(−1) | 0.440(−1) | 0.430(−2) | 0.426(−1) | 1.098(−1) | 0.437(−1) | 0.383(−2) | 0.413(−2) |
| 4 | 1.529(−1) | 0.434(−1) | 0.558(−2) | 0.475(−1) | 1.334(−1) | 0.424(−1) | 0.479(−2) | 0.418(−2) |
| 5 | 1.906(−1) | 0.446(−1) | 0.707(−2) | 0.374(−1) | 1.611(−1) | 0.432(−1) | 0.590(−2) | 0.308(−2) |

Table 7. *Long-time errors,* $r = 4$, $N = 96$ $J = 900$, $T = 5$, $E$ *of order* $10^{-3}$

| | $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ | | | | $J_{\text{out}} = 2$, $J_{\text{inn}} = 2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $t$ | $E(t)$ | shape error | phase error | amplitude error | $E(t)$ | shape error | phase error | amplitude error |
| 1 | 0.463(−3) | 0.258(−3) | 0.142(−4) | 0.362(−3) | 0.385(−3) | 0.241(−3) | 0.110(−4) | 0.316(−3) |
| 2 | 0.813(−3) | 0.277(−3) | 0.282(−4) | 0.541(−3) | 0.583(−3) | 0.248(−3) | 0.194(−4) | 0.465(−3) |
| 3 | 1.246(−3) | 0.296(−3) | 0.446(−4) | 0.534(−4) | 0.794(−3) | 0.252(−3) | 0.277(−4) | 0.641(−4) |
| 4 | 1.744(−3) | 0.314(−3) | 0.632(−4) | 0.226(−3) | 0.999(−3) | 0.255(−3) | 0.356(−4) | 0.897(−4) |
| 5 | 2.324(−3) | 0.338(−3) | 0.847(−4) | 0.559(−3) | 1.216(−3) | 0.261(−3) | 0.438(−4) | 0.392(−3) |

an error that was about 85% of the error of the run with $J_{\text{out}} = 1$, while at the $10^{-3}$ level it was about 52%. Another interesting fact derived from figure 2 is that, for the problem considered, it seems that the scheme with $J_{\text{out}} = 2$ gives an almost perfectly linear growth of $E(t)$ while the scheme with $J_{\text{out}} = 1$ leads to a somewhat faster growth. It should be kept in mind (cf. table 1) that the cost of the $J_{\text{out}} = 1$ method is about 65% of the cost of the $J_{\text{out}} = 2$ scheme at both the error levels $10^{-1}$ and $10^{-3}$. At the higher accuracy level (table 8 and figure 3) there is no appreciable difference between the two methods and $E(t)$ is apparently increasing linearly. The phase error curves, which are not shown in the figures, behave qualitatively exactly like the $E(t)$ curves.

After a transitory initial time interval, the shape error for both methods in all cases seems to remain practically constant. This raises the interesting possibility that the conservative Gauss–Legendre methods, coupled with splines, may possess exact, discrete, travelling-wave solutions that propagate with a shape and a phase speed that are very close to those of the given, solitary-wave solution of the KdV equation. They apparently share this latter property with the Crank–Nicolson method (which effectively coincides with the 1-stage member of the Gauss–Legendre family). On the other hand the dissipative Calahan method does not enjoy this property (cf. Bona *et al.* 1986, figure 1). Finally, it is worth note that the amplitude error remained small and, as $t$ increased, it fluctuated about a mean value that was approximately the same for both methods.

As mentioned before, the generalized KdV equation has three invariant func-

Table 8. *Long-time errors,* $r = 6$, $N = 128$, $J = 2600$, $T = 5$, $E$ *of order* $10^{-5}$

| | $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ | | | | $J_{\text{out}} = 2$, $J_{\text{inn}} = 2$ | | | |
|---|---|---|---|---|---|---|---|---|
| $t$ | $E(t)$ | shape error | phase error | amplitude error | $E(t)$ | shape error | phase error | amplitude error |
| 1 | 0.614(−5) | 0.357(−5) | 0.184(−6) | 0.108(−3) | 0.613(−5) | 0.356(−5) | 0.184(−6) | 0.108(−3) |
| 2 | 0.955(−5) | 0.370(−5) | 0.324(−6) | 0.874(−4) | 0.951(−5) | 0.370(−5) | 0.323(−6) | 0.873(−4) |
| 3 | 1.312(−5) | 0.378(−5) | 0.463(−6) | 0.923(−6) | 1.305(−5) | 0.378(−5) | 0.460(−6) | 0.923(−6) |
| 4 | 1.657(−5) | 0.387(−5) | 0.594(−6) | 0.132(−3) | 1.647(−5) | 0.387(−5) | 0.590(−6) | 0.132(−3) |
| 5 | 2.021(−5) | 0.398(−5) | 0.730(−6) | 0.683(−4) | 2.007(−5) | 0.398(−5) | 0.724(−6) | 0.683(−4) |

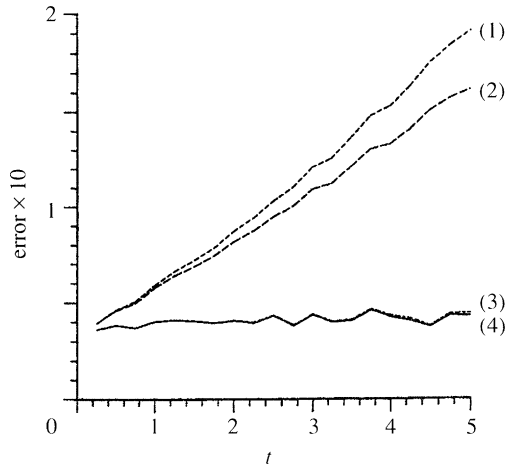


Figure 1. $E(t)$ and shape error for the data of table 6. Labelling of curves: $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$ : (1) = $E(t)$, (3) = shape error. $J_{\text{out}} = 2$, $J_{\text{inn}} = 2$ : (2) = $E(t)$, (4) = shape error.

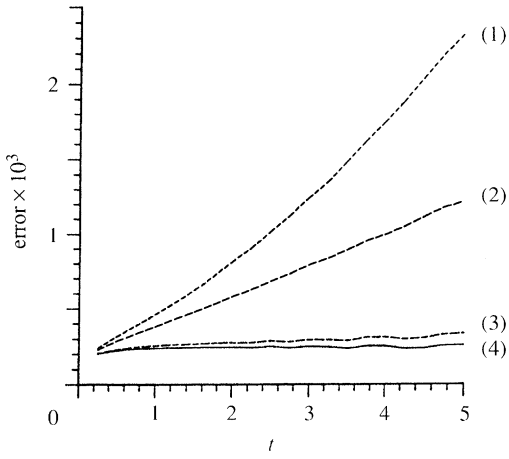


Figure 2. $E(t)$ and shape error for the data of table 7. Labelling of curves: $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$: (1) = $E(t)$, (3) = shape error. $J_{\text{out}} = 2$, $J_{\text{inn}} = 2$: (2) = $E(t)$, (4) = shape error.




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

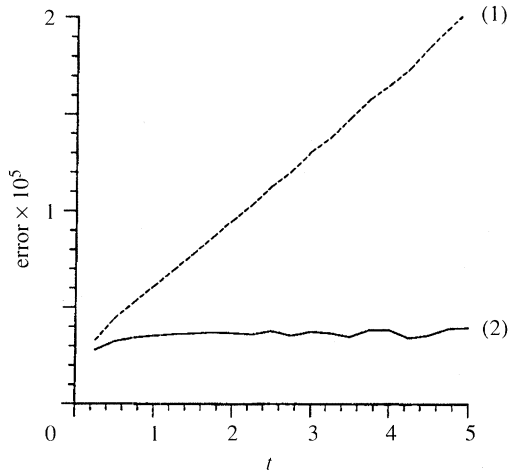


Figure 3. $E(t)$ and shape errors for the data of table 8. Labelling of curves: (1) = $E(t)$ for ($J_{\text{out}} = 1$, $J_{\text{inn}} = 2$) or ($J_{\text{out}} = 2$, $J_{\text{inn}} = 2$). (2) = shape error for ($J_{\text{out}} = 1$, $J_{\text{inn}} = 2$) or ($J_{\text{out}} = 2$, $J_{\text{inn}} = 2$).

tionals, namely

$$\left.\begin{aligned} I_1 &= \int_0^1 u(x,t)\,\mathrm{d}x, \qquad I_2 = \int_0^1 u^2(x,t)\,\mathrm{d}x, \\ I_3 &= \int_0^1 \left(u^{p+2} - \tfrac{1}{2}(p+1)(p+2)\epsilon u_x^2\right)\mathrm{d}x. \end{aligned}\right\} \tag{4.13}$$

For reasonably smooth, 1-periodic solutions of the generalized KdV equation, these functionals are independent of $t$ and therefore determined for all time by the initial data. It is interesting to inquire how these functionals vary as a function of $t$ when evaluated on the approximate solutions produced by our two schemes. It turns out that $I_1$ is conserved up to round-off error by our schemes, as one checks without difficulty. Hence we will concentrate on the variation of $I_2$ and $I_3$. From the data used to create tables 6, 7 and 8, three sets of values of $I_2$ and $I_3$ evaluated on the discrete approximation to the solution of the KdV equation ($p = 1$) were obtained. At $t = 0$, $I_2$ and $I_3$ are computed using $U^0$, the $L_2$-projection of $u_0$ onto $S_h$. The outcome of these calculations is reported in table 9, where it appears that even when the error level is set for order $10^{-5}$, the method with $J_{\text{out}} = 2$ is superior to that with $J_{\text{out}} = 1$.

Despite this outcome, it is our view that the scheme using $\{J_{\text{out}} = 1,\ J_{\text{inn}} = 2\}$ is more effective overall if one takes account of accuracy versus cost. Moreover, in the more interesting numerical experiments described in § 5 wherein $p > 1$, the values of $k$ and $h$ needed for accurate numerical computation are much smaller than the ones used to generate the data of table 8. In this circumstance, the difference in accuracy between the $J_{\text{out}} = 1$ and $J_{\text{out}} = 2$ schemes is not significant.

Although not considered herein, the scheme using $\{J_{\text{out}} = 2,\ J_{\text{inn}} = 1\}$ was found to be highly competitive with errors and timings similar to the case $\{J_{\text{out}} = 1,\ J_{\text{inn}} = 2\}$, and is also suitable for the experiments in § 5.

Table 9. *Invariants $I_2$ and $I_3$ for numerical schemes for the KdV equation corresponding to the parameters of tables 6–8*

Data from table 6.

| $I_i$ | $t$ | $J_{out} = 1,\ J_{inn} = 2$ | $J_{out} = 2,\ J_{inn} = 2$ |
|---|---|---|---|
| $I_2$ | 0.0 | 0.227276952065 | 0.227276952065 |
| $(\times 10^{-2})$ | 5.0 | 0.225441094323 | 0.227276951121 |
| $I_3$ | 0.0 | 0.30877046040 | 0.30877046040 |
| $(\times 10^{-3})$ | 5.0 | 0.30408278723 | 0.30822810524 |

Data from table 7.

| | | | |
|---|---|---|---|
| $I_2$ | 0.0 | 0.227365272650 | 0.227365272650 |
| $(\times 10^{-2})$ | 5.0 | 0.227306893470 | 0.227365273577 |
| $I_3$ | 0.0 | 0.31035264712 | 0.31035262165 |
| $(\times 10^{-3})$ | 5.0 | 0.30408278723 | 0.30822810524 |

Data from table 8.

| | | | |
|---|---|---|---|
| $I_2$ | 0.0 | 0.227365280110 | 0.227365280110 |
| $(\times 10^{-2})$ | 5.0 | 0.227365275602 | 0.227365280110 |
| $I_3$ | 0.0 | 0.31035360715 | 0.31035360715 |
| $(\times 10^{-3})$ | 5.0 | 0.31035359689 | 0.31035360714 |

## 5. Numerical experiments: adaptive procedures, instability, and blow-up of solutions

In this section, one of the numerical schemes analysed and tested in §§ 3–4 is used as a tool to investigate some interesting aspects of the GKdV equations. We begin with a short discussion of the state of the theory that provides impetus for the present study.

The KdV equation itself is solvable by the inverse-scattering transform (cf. Ablowitz & Segur 1981) and consequently we understand a great deal about the solutions of the pure initial-value problems (1.1) or (1.2) in case $p = 1$. This is also true of the case $p = 2$, but for $p > 2$ the GKdV equation is apparently not integrable by an inverse-scattering transform (McLeod & Olver 1983). There are many, interesting lessons to be gleaned from the completely integrable cases $p = 1, 2$. For our purposes, the major point of interest is that the solitary-wave solutions introduced earlier play a distinguished role in the solution of the pure initial-value problems (1.1) or (1.2). In fact, for $p = 1$ or 2, a large class of initial data has the property that the solutions of (1.1) emanating therefrom resolve themselves into a finite sequence of independently propagating solitary waves and very little else. The residue after the solitary waves are accounted is termed a dispersive tail, a waveform composed of relatively high frequencies that slowly




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

spreads and decays in amplitude. Another, surprising property emerging from the inverse-scattering theory is the exact interaction of solitary waves, so leading to the term soliton for such special solitary waves. In this aspect, a solitary wave overtaking a smaller-amplitude solitary wave on account of its greater phase speed, emerges unscathed and leaves the smaller solitary wave likewise unsullied after the nonlinear interaction between the two. Although this property of exact interaction does not generally carry over to non-integrable equations, the resolution into solitary waves and the general importance of solitary waves continue to be guiding features of the long-time behavior of solutions of the initial-value problem for the GKdV equations. Indeed, the presentation of evidence in favour of this last assertion is one of the goals in sight here.

Because it is interesting in its own right, and because of the conviction enunciated above about the general importance of solitary waves, there has been a considerable theory developed concerned with the stability of these waveforms as solutions of the initial-value problem. The theory began with the paper of Benjamin (1972) which in turn spawned many refinements and extensions (e.g. Bona 1975; Bennett *et al.* 1983; Weinstein 1986, 1987, Albert *et al.* 1987; Grillakis *et al.* 1987; Bona *et al.* 1987; Bona & Sachs 1988; Souganidis & Strauss 1990; Albert & Bona 1991; Albert 1992; Pego & Weinstein 1992). The upshot of this theory as it applies to the GKdV equation is that, whatever the value of the phase speed of the solitary wave in question, it is stable if and only if $p < 4$.

The notion of stability to which the last statement refers is the following. Let $u_0 \in H^k$, where $k \geqslant 2$, be an initial datum and let $u$ be the associated solution of (1.1). Suppose $u_0$ is near in $L_2$ to a particular solitary-wave solution $\phi$ of (1.1). Then $\phi$ is called *stable* if to any $\epsilon > 0$ there corresponds a $\delta > 0$ such that if $\|u_0 - \phi\| \leqslant \delta$, then

$$\inf_{y \in \mathbb{R}} \|u(\cdot, t) - \phi(\cdot + y)\| \leqslant \epsilon \tag{5.1}$$

for all $t \geqslant 0$. In the current context, this is just the usual notion of orbital stability. Notice that the quantity on the left-hand side of (5.1) is, up to a normalization, the continuous version of what in § 4 was called the shape error.

An issue that appears to be intimately related to the question of stability of solitary waves is that of global existence of solutions of the initial-value problem. As stated earlier (see theorem 2.1), the initial-value problem (1.1) is known to always possess global solutions corresponding to Sobolev-classes of initial data exactly when $p < 4$.

Thus two questions emerge naturally from the current state of theory regarding (1.1) or (1.2). First, what happens to an unstable solitary wave? Second, in case $p \geqslant 4$, is (1.1) or (1.2) globally well posed or not? It is proposed to cast light on these two questions by use of our numerical scheme. Regarding the instability of solitary waves for $p \geqslant 4$, we are aided by an appreciation of some of the details in the work of Bona *et al.* (1987) which gives an indication of the direction in function space that produces instability. As for the issue of global existence, it was shown in Albert *et al.* (1988) that a solution $u$ of (1.1) or (1.2) is global if and only if the solution remains bounded on bounded time intervals. (Indeed, even if $u$ remains bounded in $L_q$ on bounded time intervals for some $q > p - 2$, the same arguments will show $u$ to be global.) Thus, evidence in favour of a global solution corresponding to a particular initial value is that it settles down to some sort of

Table 10. *Errors, $U_{\max}$ and $I_3$ for the run of figure 4*

| $t$ | 6 | 12 | 18 |
|---|---|---|---|
| $E(t)$ | 0.244(−2) | 0.779(−2) | 0.373(−1) |
| shape error | 0.212(−2) | 0.417(−2) | 0.957(−2) |
| phase error | 0.130(−3) | −0.705(−3) | −0.384(−2) |
| $U_{\max}$ | 0.800 | 0.803 | 0.806 |
| $I_3$ | −0.274508236(−2) | −0.274508234(−2) | −0.274508231(−2) |

bounded state, whereas if a solution is not to be global, then it must necessarily become unbounded.

As hinted earlier, our investigations indicate these two questions are related. It seems that the theoretically predicted instability manifests itself in the solitary wave giving way to a similarity solution that forms a singularity in finite time. This in turn yields a negative answer to the question of global existence. Indeed, indications are that even for initial data far removed from the branch of solitary waves, the solution derived therefrom resolves itself into the same similarity solution and ceases to exist in finite time.

In the process of examining the issues just discussed, it was found to be extremely useful to extend the work described in earlier sections by allowing for local spatial and temporal refinement in our numerical scheme.

Attention is now turned to a detailed description of our adaptive scheme and an interpretation of the outcome of our numerical experiments, including several concrete conjectures generated by our observations. In all cases the parameter $\eta$ in (1.2*a*) was taken to be zero. The vast majority of the numerical experiments reported in this section were performed on a SUN Sparcstation 1 with a double precision, Fortran, variable-grid realization of the numerical scheme described earlier. In all the experiments reported in this section, we used as a base scheme the 2-stage Gauss–Legendre method ($q = 2$) with cubic splines ($r = 4$) and iteration numbers $J_{\text{out}} = 1$, $J_{\text{inn}} = 2$. All initial wave profiles were organized to be symmetric about the point $x^0 = \frac{1}{2}$.

Described first are experiments on the solitary-wave solution of (4.10*a*) given by (4.10*b*). The numerical scheme with constant spatial mesh length and constant time-step (the situation analysed in the preceding sections) seems to be adequate for describing solitary-wave solutions of (1.2) having small amplitude A, at least over time scales for which such solutions remain stable. For example, figure 4 shows the temporal evolution corresponding to solitary-wave initial data with parameters $p = 5$, $A = 0.8$, $\epsilon = 10^{-4}$, obtained with the uniform discretization parameters $h = 1/384$ and $k = 10^{-2}$. The profiles are plotted at $t = 0$, 6, 12 and 18.

In addition, in table 10 we show the normalized $L_2$-error $E(t)$, the shape and phase errors (all defined in § 4) as well as the amplitude $U_{\max}$ and the invariant $I_3$ of the discrete solution for this run at the three times $t = 6$, 12 and 18. (The value of the approximation to $I_3$ at $t = 0$ was $-0.274508237 \times 10^{-2}$.)

It is evident that the uniform-mesh scheme was able, with these discretization

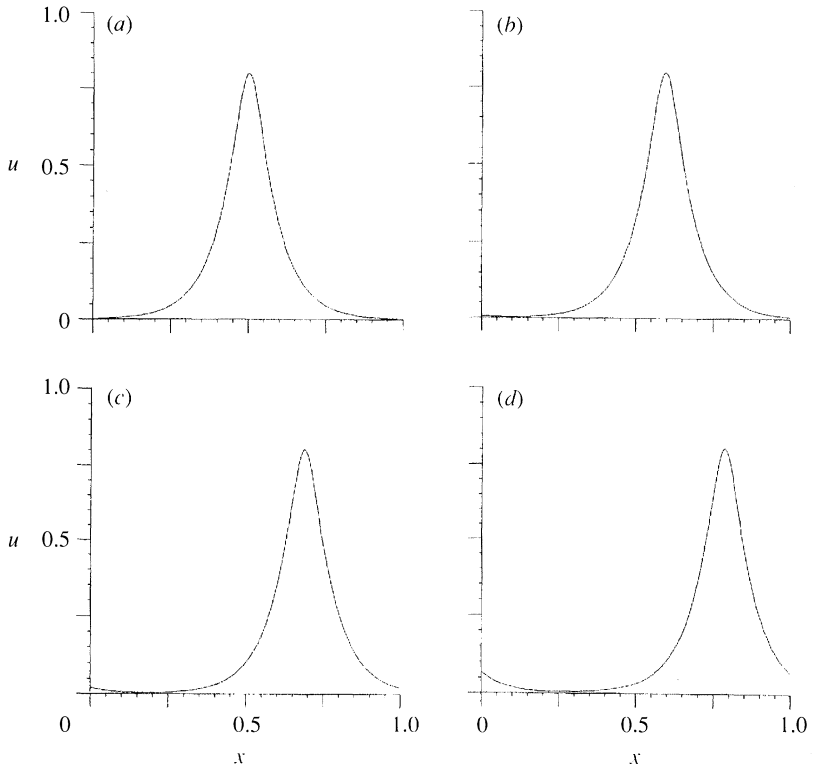


Figure 4. Numerical simulation of a solitary-wave solution, $p = 5$, $A = 0.8$, $\epsilon = 10^{-4}$ using a uniform mesh with $h = 1/384$, $k = 10^{-2}$. (*a*) $t = 0$, $u_{\max} = 0.800$; (*b*) $t = 6$, $u_{\max} = 0.799$; (*c*) $t = 12$, $u_{\max} = 0.802$; (*d*) $t = 18$, $u_{\max} = 0.806$.

parameters, to approximate the travelling wave quite satisfactorily. Moreover, there was no hint of a developing instability during this time period.

The situation was different when, with $p = 5$, the amplitude was increased to $A = 2$ and $\epsilon$ was taken to be $5 \times 10^{-4}$. The temporal step size was reduced to $2 \times 10^{-4}$ while the spatial meshlength remained $1/384$. Graphical output from this run is depicted in figure 5 at the instants $t = 0,\ 0.04,\ 0.08$ and $0.1$.

The solitary wave rapidly lost stability, exhibiting a significant change in amplitude. The growth in amplitude in turn triggered what appeared to be a numerical instability, wherein the approximate solution began to break down by developing small oscillations due to dispersive pollution. For this particular experiment, the oscillations disappear if the spatial mesh length and the temporal step size are reduced. However, they reappear at a later time as the amplitude of the sharp peak increases. It became evident that tracking an unstable solution which quickly develops into a large-amplitude disturbance is difficult with a fixed space-time grid. But this experiment and others like it provided the first hint that the instability of the solitary-wave solutions of (1.2) for $p \geqslant 4$ may result in the formation of a singularity where the solution becomes unbounded at a single point. To better track the development of the instability and to increase confidence in the supposition that the solution becomes unbounded in finite time, we turned to an algorithm based on the 2-stage Gauss–Legendre method and cubic splines

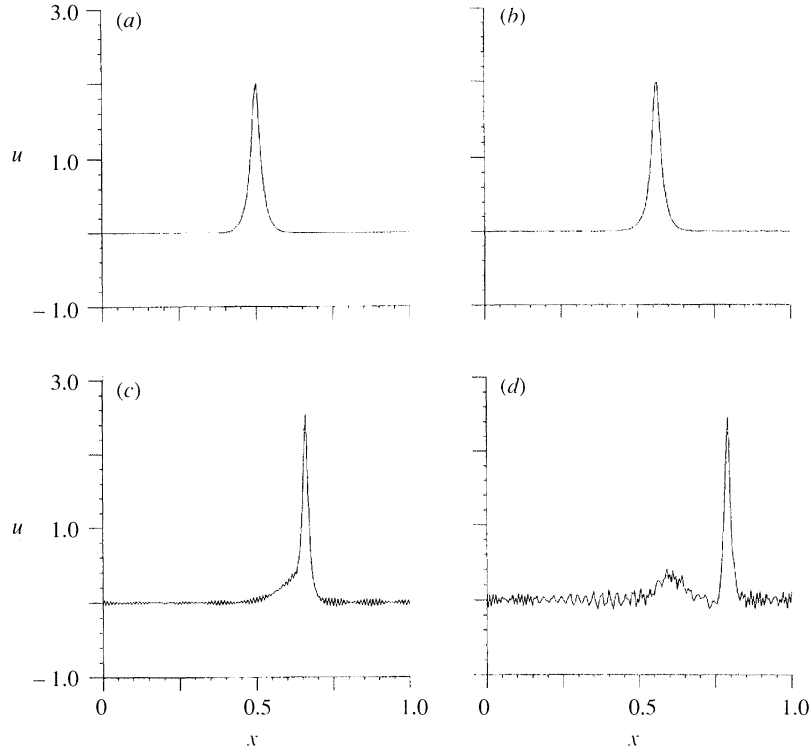


Figure 5. Numerical instability in the simulation of an unstable solitary-wave solution, $p = 5$, $A = 2$, $\epsilon = 5 \times 10^{-4}$, using a uniform mesh with $h = 1/384$, $k = 2 \times 10^{-4}$. (*a*) $t = 0.00$, $u_{\max} = 2.00$; (*b*) $t = 0.04$, $u_{\max} = 1.98$; (*c*) $t = 0.08$, $u_{\max} = 2.53$; (*d*) $t = 0.10$, $u_{\max} = 2.45$.

which can perform automatic grid refinement in both the spatial and temporal variables.

Note that the error analysis of § 3 does not preclude changing the temporal stepsize $k$ since all time-stepping schemes in the class considered are single-step. (Care should be exercised in interpolating starting values for the new time step and, of course, the various constraints appearing for example in theorem 3.1 should be taken into account.) However, the error estimates obtained in § 3 do depend on the assumption of a uniform spatial mesh because of reliance on the approximation properties of the quasi-interpolant in the space of periodic splines. Based on our numerical experience, it appeared evident that this assumption would need to be relaxed if the problem of approximating well the apparent blow-up of solutions is to be tractable. Indeed, from the discrete conservation law (3.38) and the $L_\infty - L_2$ inverse inequality in (2.2), it follows that for $n \geqslant 0$,

$$|U^n|_\infty \leqslant ch^{-1/2}\|U^n\| = ch^{-1/2}\|U^0\|. \tag{5.2}$$

Hence the approximations $\{U^n(x)\}_{n\geqslant 0}$ cannot develop an arbitrarily high peak without $h$ becoming arbitrarily small. For instance, if $\|U^0\|$ and $c$ in (5.2) are of order one, say, then a modest peak of amplitude $10^3$ cannot be attained unless $h$ is about $10^{-6}$. Both in terms of computational time and storage, it is not currently possible for us to sustain calculations on a uniform spatial grid with a million points.

Attention is now given to the aforementioned adaptive algorithm developed in response to the difficulties documented above. This algorithm and the corresponding computer code are geared towards approximating solutions of the initial-value problem that develop a single, infinite singularity at a point $(x^*, t^*)$ for some $x^* \in [0, 1]$ at some finite time $t^* > 0$. The adaptive mechanism of the code consists of three main parts:

(i) local refinement of the spatial grid,

(ii) temporal step-size reduction, and

(iii) spatial translation of the peak to a region with a finer grid.

Our spatial, local grid refinement is effected by occasional additions of even numbers of new nodes that are distributed evenly about the midpoint $x = 0.5$, but within successively smaller intervals. Since the task in view is to approximate wave profiles that travel to the right and apparently develop a single, very high peak, the solution is occasionally translated to the left and the peak centred near $x = 0.5$ to keep it in the region of highest density of nodes and away from regions of coarser mesh. Thus, in effect, we translate the solution to conform to the grid, which is being refined locally around a fixed point, rather than design a grid that moves with the peak and condenses around it. This technique proved to be far simpler to program than actually moving the nodes, but it lacks the generality of a fully spatially adaptive scheme. Nevertheless, this idea was quite effective when applied to the problems considered here.

In describing the adaptive procedure of spatial refinement, it is convenient to introduce some notation. At $t = 0$, the temporal integration is initiated with a uniform partition of the spatial interval $[0, 1]$ consisting of $N$ mesh intervals of length $h_0 = 1/N$. At some later time, spatial refinement starts according to a criterion to be specified below. At a refinement stage, let NSPLIT stand for the number of times additional nodes have been introduced thus far and $h_* = h_{\mathrm{NSPLIT}} = 2^{-\mathrm{NSPLIT}}/N$ be the current, finest meshlength. Each time new nodes are added, NSPLIT increases by one and $h_*$ is cut in half. At each refinement stage, a fixed, even number NADD of nodes is added, symmetrically and contiguously about the point $x^0 = \frac{1}{2}$. (After some experimentation, NADD was set to 36 except at the first refinement where it was given the value 72.) Then the interval $\Omega_*$ given by

$$\Omega_* = [0.5 - \mathrm{NADD} \times h_* \,,\; 0.5 + \mathrm{NADD} \times h_*]$$

is the current neighborhood of the midpoint $x^0 = 0.5$ to which the most recent new nodes have been added symmetrically about $x = 0.5$, and therefore it is the region with the finest grid (having a uniform meshlength $h_*$). Each time nodes are introduced, $\Omega_*$ is thereby cut in half. With this kind of system of spatial refinement, it is clear that the effect of the spatial translations in step (iii) above should be to insure that the peak of the numerical approximation lies in $\Omega_*$. Keeping this notation in mind, we turn to the details of the steps (i)–(iii) of the adaptive refinement.

(i) *Local refinement of the spatial grid.* This is based on the $L_\infty - L_2$ inverse inequality (5.2). Let $U^n(x)$ be the current, fully discrete approximation, and compute

$$Z_2 = \left( \int_{\Omega_*} [U^n(x)]^2 \mathrm{d}x \right)^{1/2}.$$

Estimate the max-norm of $U^n$ by $Z_\infty = \max_{x\in Q} |U^n(x)|$, where $Q$ denotes the set of all Gauss-quadrature abscissae in [0,1] which are used to evaluate integrals. (A set of $n_q$ Gauss points is used in every spatial mesh interval to ensure exact computation of integrals; e.g. for $p = 5$ and $r = 4$, $n_q = 11$.) Then, refine the mesh locally, as described previously, if

$$\frac{Z_\infty h_*^{1/2}}{c_* Z_2} > \mathrm{TOL1}, \tag{5.3}$$

where $0 < \mathrm{TOL1} < 1$ is an empirically chosen tolerance, usually taken to be equal to 0.2 or 0.1 in practice, and where the constant $c_*$ is an estimate of the constant $c$ appearing in (5.2). A reasonable approximation of $c$ is the constant occurring in the $L_\infty - L_2$ inverse equality $|\chi|_\infty = c_* h^{-1/2} \|\chi\|$, where $\chi$ is the bell-shaped, cubic spline basis function with support on [0,4h]. Thus in (5.3) we took

$$c_* = \frac{\max_{[0,4h]}\ \chi(x)}{\left(\int_0^{4h} \chi^2(x)\,\mathrm{d}x\right)^{1/2}} h^{1/2} = \left(\frac{151}{140}\right)^{1/2} \simeq 1.04.$$

(The exact value of $c$ is greater than $c_*$ and has an easily calculated, coarse upper bound of about 4.2. Hence a value of $\mathrm{TOL1} \leqslant 0.2$ represents a conservative choice.)

If the inequality (5.3) holds at a certain stage in the calculations, then at that point the spatial mesh is locally refined by cutting $h_*$ and $\Omega_*$ in half to give the numerical scheme a chance to approximate the peak better; in effect, allowing the right-hand side of (5.2) to grow so as not to inhibit the growth of the left-hand side. Note that since new nodes are added without shifting the old ones, the subspace $S_{h_j}$ is imbedded in $S_{h_{j+1}}$, for all $j$, and the coefficients of a function $v \in S_{h_j}$ are simply recomputed to correspond to the new basis.

(ii) *Temporal stepsize reduction.* The temporal stepsize $k$ is adjusted in an attempt to preserve accuracy in the fully discrete approximation of the third invariant (see (4.13))

$$I_3(v) = \int_0^1 [v^{p+2} - \tfrac{1}{2}(p+1)(p+2)\epsilon\, v_x^2]\,\mathrm{d}x.$$

Given $U^n$, the fully discrete approximation of $u(\cdot, t^n)$, an estimate of $U^{n+1}$ is computed by our single-step scheme using the previously available value of $k$. This estimate is accepted if

$$\frac{|I_3(U^{n+1}) - I_3(U^n)|}{\int_0^1 (U_x^{n+1})^2 \mathrm{d}x} < \mathrm{TOL2}, \tag{5.4}$$

where TOL2 is a small parameter, typically taken to be in the range $10^{-4}$ to $10^{-5}$. If (5.3) is not satisfied, the time step is cut in half and the approximation to the solution $u(\cdot, t^{n+1})$ recomputed as $U^{n+1}$. The denominator in (5.4), one of the two terms comprising $I_3$ , plays a normalizing role. In fact, it was found that taking the denominator to be $|I_3(U^n)|$, say, thus imposing a maximum relative error tolerance on $I_3$, was too stringent a restriction that typically resulted in cutting the time step too soon and so drastically that the numerical process could not even approach the blow-up point. Considerable experimentation indicated that normalizing the numerator in (5.4) by the square of the $L_2$ norm of $U_x^{n+1}$

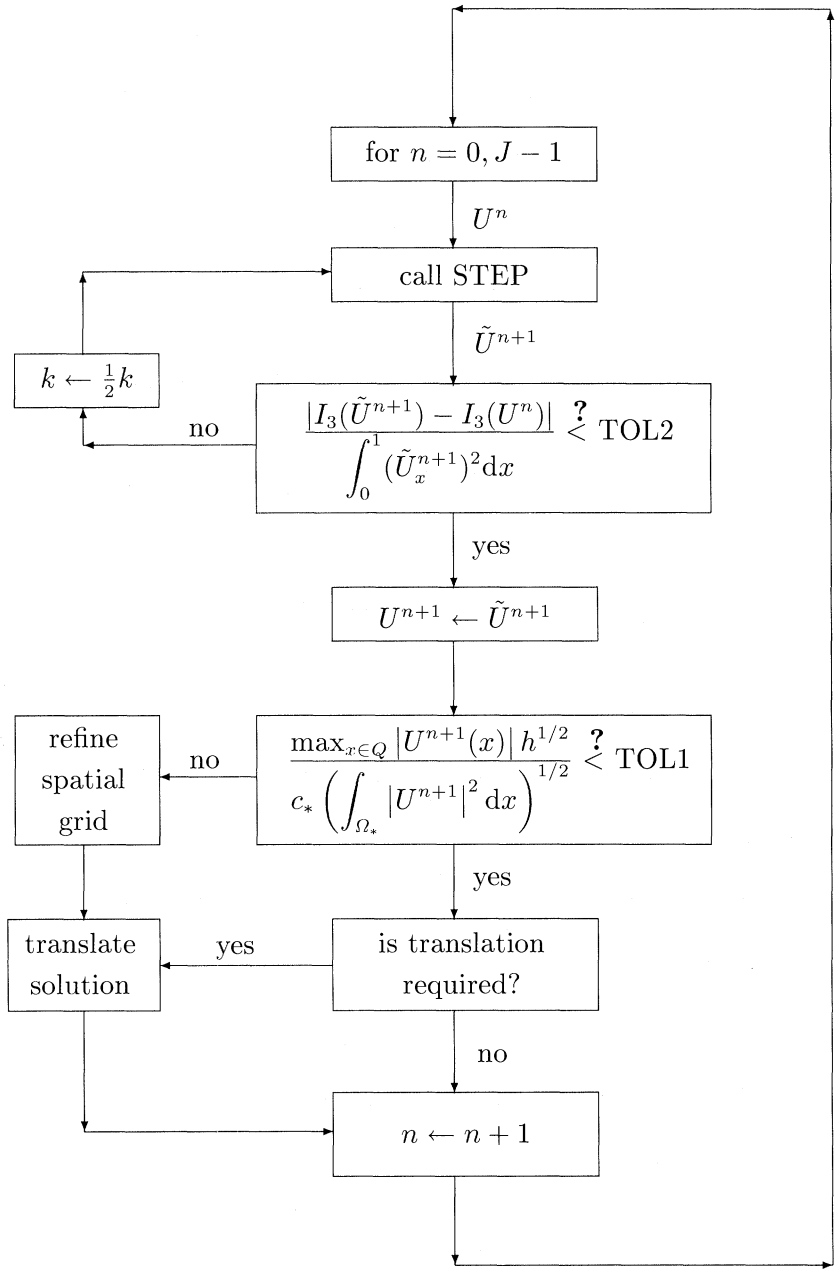


Diagram 1. Subprogram STEP

was an effective choice. This normalization coupled with a suitable choice of TOL2 generates a rule for cutting the time step at a rate which is sufficient for a satisfactory approximation of the solution even when the amplitude becomes large.

It should be noted that keeping $I_3$ under control is difficult in actual computations since it is the difference of two terms $\int_0^1 u^{p+2}$ and $\int_0^1 u_x^2$, both of which must

become large if the solution is to form a singularity. Our experience showed that refining the time step by imposing condition (5.4) on the fully discrete version of $I_3$ was helpful in avoiding deterioration in numerical accuracy as the solution approached what appears to be a blow-up time.

(iii) *Spatial translation of the peak to a finer grid region.* As mentioned previously, part of the idea of the adaptive scheme discussed here is to occasionally shift the numerical solution with the purpose of keeping the large function- and derivative-values within $\Omega_*$, the region of the highest node density. This is accomplished in the following manner. Evaluate $\max_{x\in Q}|U^n(x)|$ and let $y$ be the right-most point in $Q$ (the set of Gauss abscissae) where the maximum is attained. If $y > \frac{1}{2}(1+\mathrm{NADD}\times h_*)$, translate $U^n$ to the left by a distance $s = x_r - \frac{1}{2}$, where $[x_\ell, x_r]$ is the mesh interval containing $y$. In general the translated function $U^n(x-s)$ will not lie in $S_h$. Hence the translation $U^n(\cdot - s)$ is projected onto an element $\hat{U}^n$ in $S_h$ by requiring

$$(\hat{U}^n, \chi) = (U^n(\cdot - s), \chi) \quad \text{for all } \chi \in S_h.$$

The flowchart (diagram 1) summarizes the three steps of the adaptive mechanism and indicates the sequence in which they are implemented. In the chart, one full temporal step is performed by the subprogram STEP.

Having constructed a code with this adaptive mechanism in place, it is interesting to repeat numerical experiments such as the one whose results are depicted in figure 5 in which a relatively large-amplitude solitary wave was propagated. A collection of such experiments will be described below, which suggest that solitary-wave solutions are indeed unstable if $p \geqslant 4$, the instability probably being precipitated by truncation and roundoff errors in the representation of the initial data and the solution. To hasten the onset of instability it was convenient to use as initial data functions of the form

$$u_0(x) = \lambda A \operatorname{sech}^{2/p}[K(x - \tfrac{1}{2})], \tag{5.5}$$

where $K$ is defined after (4.10), (with $\eta = 0$, $x^0 = \frac{1}{2}$) and where $\lambda$ is a perturbation factor, usually taken to be either 1.05 or 1.01.

In a preliminary experiment we took $A = 2$, $p = 5$, $\epsilon = 5\times 10^{-4}$ (the same parameters as those present in the run corresponding to figure 5) and $\lambda = 1.01$. Starting with initial, uniform-mesh parameters $h_0 = 1/192$, $k_0 = 10^{-3}$ and using TOL1 $= 0.2$, TOL2 $= 10^{-5}$ as tolerances in the adaptive procedure, the temporal evolution was that depicted in figure 6. It is worth note that the small-scale numerical oscillations are no longer in evidence. In this figure, the four plots are taken at four times when additional spatial refinement was called for, specifically when NSPLIT became 2, 4, 6 and 9, respectively. As the peak value $U_{\max}$ increases, the vertical axis is rescaled so that the entire profile is shown. This has the effect of making the solution appear to be quite small everywhere except near the peak and the shelf trailing immediately behind it. Because of the translations that are part of our adaptive scheme, the peak is always kept near the midpoint of the interval. The tolerances chosen in the adaptive procedure enabled us to continue this run to the point where $U_{\max}$ was about $2\times 10^5$, all the time maintaining a smooth, small-oscillation-free profile.

A closer look at the structure around the solution's peak as it continues to grow may be obtained by rescaling the horizontal axis as well as the vertical




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

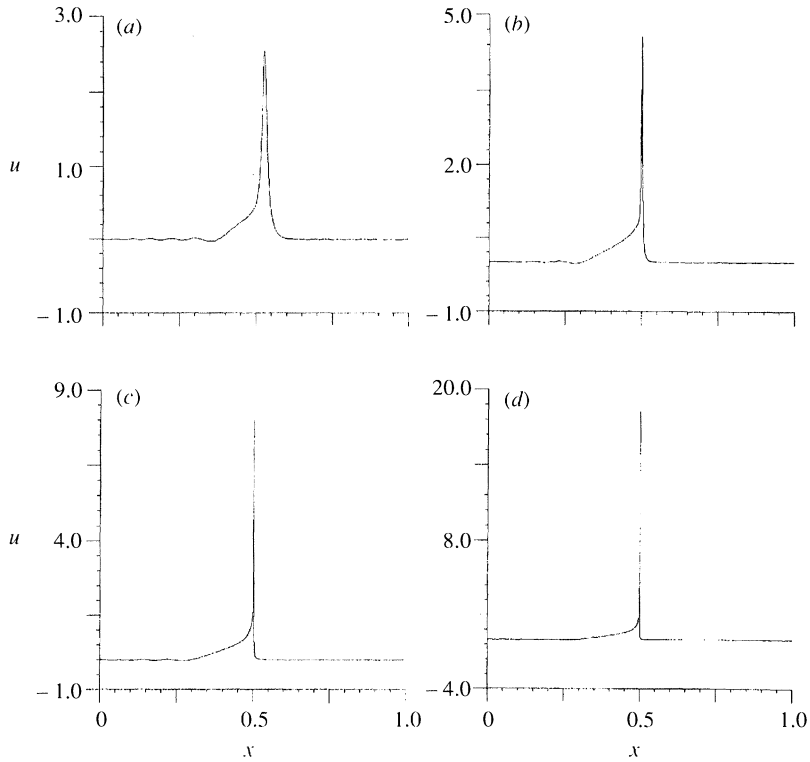


Figure 6. Numerical simulation of the instability of a solitary-wave solution, $p = 5$, $A = 2$, $\epsilon = 5 \times 10^{-4}$, $\lambda = 1.01$. Variable grid with $h_0 = 1/192$, $k_0 = 10^{-3}$, TOL1 $= 0.2$, TOL2 $= 10^{-5}$. (*a*) $t = 0.01975$, $u_{\max} = 2.55$; (*b*) $t = 0.02251$, $u_{\max} = 4.57$; (*c*) $t = 0.02254$, $u_{\max} = 8.02$; (*d*) $t = 0.02254$, $u_{\max} = 18.3$.

axis. This is done in figure 7, where, at four different times the solution has been translated by $-\frac{1}{2}$ so that the peak lies near zero, and then only that portion is plotted corresponding to an interval centred at zero and having the same length as the then current, fully refined interval $\Omega_*$. These four plots were taken at times when NSPLIT $= 10$, 21, 29 and 40, respectively. By the final plot the local spatial and temporal mesh sizes had decreased to the point where $h \simeq 10^{-14}$ and $k \simeq 10^{-38}$, respectively. The corresponding amplitudes $U_{\max}$ are indicated on the legends. This experiment provides strong evidence supporting the conjecture that, not only is the solitary-wave solution unstable, but the instability manifests itself as a single-point blow-up in finite time. Moreover, as the graphs in figure 7 illustrate, the blow-up appears to be of a similarity type. We shall return to this point presently.

To better understand the blow-up instability of solitary waves, a series of numerical experiments was performed aimed at estimating the rates with which various norms and semi-norms of the solution tend to infinity as $t$ approaches the blow-up time $t^*$. Let $M(t)$ be a quantity of interest associated with a solution $u$ and let $\rho > 0$ be its blow-up rate at $t^*$, which is to say it is presumed that $M(t) \sim c(t^* - t)^{-\rho}$ as $t \uparrow t^*$ for some nonzero constant $c$. If this supposition is valid, then in principle if the values of $M$ are known at two distinct instants $\tau_1$

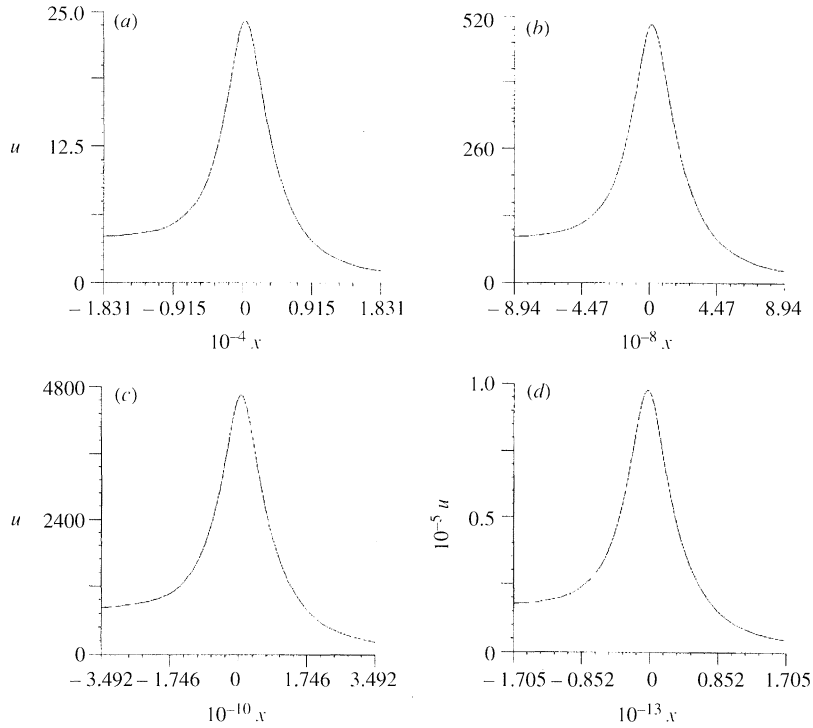


Figure 7. The solution of figure 6 near its peak, with both axes scaled and the $x$-axis translated. ($a$) $t = 0.02254$, $u_{\max} = 2.4 \times 10$; ($b$) $t = 0.02254$, $u_{\max} = 5.0 \times 10^2$; ($c$) $t = 0.02254$, $u_{\max} = 4.6 \times 10^3$; ($d$) $t = 0.02254$, $u_{\max} = 9.8 \times 10^4$.

and $\tau_2$ near to, but less than $t^*$, the quantity $\rho$ may be estimated as the ratio

$$-\frac{\log\left(M(\tau_1)/M(\tau_2)\right)}{\log\left((t^*-\tau_1)/(t^*-\tau_2)\right)}. \tag{5.6}$$

Of course, in practice $t^*$ is not known and must be estimated numerically. In addition, for an accurate determination of $\rho$ one needs to compute all the quantities in (5.6) carefully to avoid loss of precision due to cancellation.

The following procedure appeared to be successful in determining the values of $\rho$ corresponding to a number of interesting quantities. As the peak of the solution steepens, the code starts refining in space by locally cutting $h$ as outlined previously. Let $\tau_i$, $i = 1, 2, \ldots, f$, be the time at which the $i$th spatial refinement occurs and let $\Delta\tau_i = \tau_i - \tau_{i-1}$, $2 \leqslant i \leqslant f$. We approximate the actual blow-up time $t^*$ as the time of the final spatial refinement $\tau_f$. The code itself terminates either when the maximum amplitude of the peak exceeds a specified ceiling or when the differences $\Delta\tau_i = \tau_i - \tau_{i-1}$ fall below a certain floor. Define the quantity $s_i$ by $s_i = \tau_f - \tau_i$, $1 \leqslant i \leqslant f-1$. Then, the rate $\rho$ in (5.6) is approximated by the sequence of rates $\rho_i$ given by

$$\rho_i = -\frac{\log\left(M(\tau_i)/M(\tau_{i+1})\right)}{\log(s_i/s_{i+1})}, \quad i = 1, 2, \ldots, f-2. \tag{5.7}$$

In the experiments reported here, the times $\tau_i$ are exceedingly close to $\tau_f$ for larger $i$. To avoid loss of accuracy due to subtractive cancellation in forming

Table 11. *Blow-up rates. Solitary wave,* $p = 5$, $\epsilon = 5 \times 10^{-4}$, $\tau_f = 0.22543 \times 10^{-1}$, $f = 42$, $x^* = 0.61333$, $U_{\max} = 224,766$, $k_{\min} = 0.23 \times 10^{-40}$, $\Delta\tau_f = 0.16 \times 10^{-38}$, TOL1 $= 0.2$, TOL2 $= 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.5029(−1) | 0.6683(−1) | 0.7795(−1) | 0.8590(−1) | 0.1336 | 0.3008 | 0.4657 |
| 10 | 0.5047(−1) | 0.6729(−1) | 0.7853(−1) | 0.8657(−1) | 0.1348 | 0.3028 | 0.4731 |
| 15 | 0.4983(−1) | 0.6647(−1) | 0.7759(−1) | 0.8554(−1) | 0.1334 | 0.2992 | 0.4618 |
| 20 | 0.4989(−1) | 0.6658(−1) | 0.7773(−1) | 0.8572(−1) | 0.1338 | 0.2999 | 0.4690 |
| 25 | 0.5044(−1) | 0.6728(−1) | 0.7851(−1) | 0.8654(−1) | 0.1347 | 0.3029 | 0.4747 |
| 30 | 0.4974(−1) | 0.6633(−1) | 0.7741(−1) | 0.8534(−1) | 0.1329 | 0.2985 | 0.4685 |
| 35 | 0.5001(−1) | 0.6672(−1) | 0.7786(−1) | 0.8583(−1) | 0.1336 | 0.3004 | 0.4654 |

Table 12. *Blow-up rates. Solitary wave,* $p = 5$, $\epsilon = 5 \times 10^{-4}$, $\tau_f = 0.22618 \times 10^{-1}$, $f = 44$, $x^* = 0.61383$, $U_{\max} = 307,834$, $k_{\min} = 0.15 \times 10^{-41}$, $\Delta\tau_f = 0.16 \times 10^{-39}$, TOL1 $= 0.15$, TOL2 $= 2 \times 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.5035(−1) | 0.6675(−1) | 0.7779(−1) | 0.8567(−1) | 0.1328 | 0.2998 | 0.4681 |
| 10 | 0.5003(−1) | 0.6662(−1) | 0.7768(−1) | 0.8558(−1) | 0.1329 | 0.2995 | 0.4677 |
| 15 | 0.5000(−1) | 0.6672(−1) | 0.7789(−1) | 0.8587(−1) | 0.1339 | 0.3007 | 0.4698 |
| 20 | 0.5043(−1) | 0.6727(−1) | 0.7850(−1) | 0.8654(−1) | 0.1347 | 0.3027 | 0.4724 |
| 25 | 0.4982(−1) | 0.6641(−1) | 0.7745(−1) | 0.8531(−1) | 0.1324 | 0.2989 | 0.4605 |
| 30 | 0.5006(−1) | 0.6681(−1) | 0.7799(−1) | 0.8600(−1) | 0.1342 | 0.3011 | 0.4743 |
| 35 | 0.5027(−1) | 0.6706(−1) | 0.7828(−1) | 0.8630(−1) | 0.1342 | 0.3019 | 0.4655 |
| 40 | 0.5006(−1) | 0.6679(−1) | 0.7795(−1) | 0.8593(−1) | 0.1342 | 0.3014 | 0.4770 |

$\tau_f - \tau_i$, we compute $s_{i+1}$ as the sum $s_{i+1} = \sum_{j=i+2}^{f} \Delta\tau_j$. (The quantities $\Delta\tau_j$ are easily accumulated as sums of a few successive values of the current time step. Typically, near a singularity the code appears on the average to cut about three times in time for every cut in space.) The denominator in (5.7) is then evaluated as $\log\left((s_{i+1} + \Delta\tau_{i+1})/s_{i+1}\right)$.

Described now are the experiments and the blow-up rates that were calculated. For $p = 4, 5, 6$, and 7 we recorded the blow-up rates $\rho_i$ of the $L_m$ norms of the approximate solution $U$ for $m = p - 1,\ p,\ p + 1,\ p + 2$ and $\infty$ and also the $L_2$ and $L_\infty$ norms of $U_x$ (shown in the tables under columns $L_{2,D}$ and $L_{\infty,D}$) at the times $\tau_i$, usually every few $i$. Positive rates were obtained for all these quantities. Thus the experiments suggest that they blow up as $t \to t^*$.

In tables 11–14 are shown the blow-up rates that were obtained for solitary-wave solutions for $p = 5$, initial amplitude $A = 2$ and perturbation factor $\lambda = 1.01$. In tables 11–13 we took $\epsilon = 5 \times 10^{-4}$ and changed the parameters TOL1 and TOL2 occurring in the spatial and temporal refinement procedures, respectively. In the legends are recorded the final time $\tau_f$, used as an approximation to the exact blow-up time, the index $f$ corresponding to $\tau_f$, the approximation to the




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Table 13. *Blow-up rates. Solitary wave,* $p = 5$, $\epsilon = 5 \times 10^{-4}$, $\tau = 0.22617 \times 10^{-1}$, $f = 41$, $x^* = 0.61384$, $U_{\max} = 171,382$, $k_{\min} = 0.94 \times 10^{-40}$, $\Delta\tau_f = 0.11 \times 10^{-37}$, TOL1 $= 0.2$, TOL2 $= 2 \times 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.5061(−1) | 0.6727(−1) | 0.7849(−1) | 0.8650(−1) | 0.1348 | 0.3028 | 0.4682 |
| 10 | 0.4985(−1) | 0.6643(−1) | 0.7750(−1) | 0.8541(−1) | 0.1331 | 0.2989 | 0.4631 |
| 15 | 0.5016(−1) | 0.6691(−1) | 0.7809(−1) | 0.8608(−1) | 0.1338 | 0.3011 | 0.4698 |
| 20 | 0.4998(−1) | 0.6660(−1) | 0.7766(−1) | 0.8556(−1) | 0.1329 | 0.2993 | 0.4614 |
| 25 | 0.4933(−1) | 0.6571(−1) | 0.7660(−1) | 0.8436(−1) | 0.1308 | 0.2955 | 0.4534 |
| 30 | 0.4960(−1) | 0.6608(−1) | 0.7706(−1) | 0.8489(−1) | 0.1319 | 0.2970 | 0.4610 |
| 35 | 0.5009(−1) | 0.6687(−1) | 0.7808(−1) | 0.8610(−1) | 0.1345 | 0.3014 | 0.4673 |

Table 14. *Blow-up rates. Solitary wave,* $p = 5$, $\epsilon = 2 \times 10^{-4}$, $\tau_f = 0.14605 \times 10^{-1}$, $f = 41$, $x^* = 0.57355$, $U_{\max} = 248,639$, $k_{\min} = 0.59 \times 10^{-41}$, $\Delta\tau_f = 0.47 \times 10^{-39}$, TOL1 $= 0.2$, TOL2 $= 2 \times 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.5073(−1) | 0.6736(−1) | 0.7859(−1) | 0.8663(−1) | 0.1350 | 0.3028 | 0.4707 |
| 10 | 0.5085(−1) | 0.6773(−1) | 0.7901(−1) | 0.8707(−1) | 0.1355 | 0.3046 | 0.4753 |
| 15 | 0.4941(−1) | 0.6583(−1) | 0.7676(−1) | 0.8457(−1) | 0.1316 | 0.2961 | 0.4629 |
| 20 | 0.4992(−1) | 0.6657(−1) | 0.7766(−1) | 0.8558(−1) | 0.1330 | 0.2997 | 0.4579 |
| 25 | 0.4955(−1) | 0.6606(−1) | 0.7705(−1) | 0.8490(−1) | 0.1318 | 0.2975 | 0.4627 |
| 30 | 0.4917(−1) | 0.6548(−1) | 0.7635(−1) | 0.8412(−1) | 0.1306 | 0.2940 | 0.4559 |
| 35 | 0.5116(−1) | 0.6824(−1) | 0.7964(−1) | 0.8779(−1) | 0.1368 | 0.3072 | 0.4781 |

spatial blow-up point $x^*$ (obtained as the sum of the midpoint $x^0 = \frac{1}{2}$ plus the accumulated total shift resulting from all translations up to time $\tau_f$), the amplitude of the peak $U_{\max}$ at $\tau_f$, as well as $\Delta\tau_f$ and the smallest time step reached, $k_{\min}$. (In the experiments of tables 11 to 15 we took initially $h_0 = 1/192$ and $k_0 = 10^{-3}$. For tables 16 and 17, we took $h_0 = 1/384$, $k_0 = 10^{-3}$, while tables 18 and 19 have $h_0 = 1/384$, $k_0 = 0.5 \times 10^{-3}$.) As one would expect, although $\tau_f$, $U_{\max}$ and $x^*$ vary somewhat as TOL1 and TOL2, (and f) vary, the rates are quite robust and remain (to two digits) independent of the refinement parameters and indeed practically constant as $i$, the number of spatial refinements, increases. Hence we are rather confident that the quantities $M(t)$ grow due to the behavior of the solution and not as a result of numerical instability or bias in the refinement procedure. Tables 13 and 14 differ only in the value of $\epsilon$. Changing $\epsilon$ results in the blow-up occuring at different points $(x^*, t^*)$; however, the rates remain sensibly the same and the ratio $\tau_f/(x^* - x^0)$ also remains constant due to the scaling of the independent variables that gives the equivalence of (1.1*a*) with (1.2*a*).

Tables 15 and 16 show the computed blow-up rates and the associated parameters for solitary waves (both with $\epsilon = 5 \times 10^{-4}$) for $p = 6$ and 7, respectively. The computations are more difficult now due to the higher values of $p$ (note the re-

Table 15. *Blow-up rates. Solitary wave,* $p = 6$, $\epsilon = 5 \times 10^{-4}$, $\tau_f = 0.51541 \times 10^{-2}$, $f = 41$, $x^* = .54732$, $U_{\max} = 26,099$, $k_{\min} = 0.47 \times 10^{-40}$, $\Delta\tau_f = 0.26 \times 10^{-38}$, TOL1 $= 0.1$, TOL2 $= 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.4479(−1) | 0.5578(−1) | 0.6372(−1) | 0.6969(−1) | 0.1116 | 0.2786 | 0.4465 |
| 10 | 0.4440(−1) | 0.5548(−1) | 0.6340(−1) | 0.6935(−1) | 0.1110 | 0.2773 | 0.4419 |
| 15 | 0.4438(−1) | 0.5548(−1) | 0.6341(−1) | 0.6936(−1) | 0.1109 | 0.2774 | 0.4460 |
| 20 | 0.4442(−1) | 0.5552(−1) | 0.6345(−1) | 0.6941(−1) | 0.1111 | 0.2775 | 0.4410 |
| 25 | 0.4423(−1) | 0.5528(−1) | 0.6317(−1) | 0.6909(−1) | 0.1106 | 0.2764 | 0.4396 |
| 30 | 0.4473(−1) | 0.5560(−1) | 0.6354(−1) | 0.6950(−1) | 0.1113 | 0.2781 | 0.4431 |
| 35 | 0.4452(−1) | 0.5565(−1) | 0.6359(−1) | 0.6955(−1) | 0.1112 | 0.2781 | 0.4455 |

Table 16. *Blow-up rates. Solitary wave,* $p = 7$, $\epsilon = 5 \times 10^{-4}$, $\tau_f = 0.13676 \times 10^{-2}$, $f = 35$, $x^* = 0.52458$, $U_{\max} = 1191$, $k_{\min} = 0.13 \times 10^{-32}$, $\Delta\tau_f = 0.61 \times 10^{-31}$, TOL1 $= 0.1$, TOL2 $= 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.4172(−1) | 0.4935(−1) | 0.5526(−1) | 0.5993(−1) | 0.9760(−1) | 0.2618 | 0.4360 |
| 10 | 0.3986(−1) | 0.4779(−1) | 0.5376(−1) | 0.5841(−1) | 0.9559(−1) | 0.2628 | 0.4320 |
| 15 | 0.3972(−1) | 0.4766(−1) | 0.5362(−1) | 0.5825(−1) | 0.9534(−1) | 0.2622 | 0.4278 |
| 20 | 0.3974(−1) | 0.4768(−1) | 0.5364(−1) | 0.5827(−1) | 0.9536(−1) | 0.2622 | 0.4301 |
| 25 | 0.3961(−1) | 0.4753(−1) | 0.5348(−1) | 0.5810(−1) | 0.9511(−1) | 0.2614 | 0.4267 |
| 30 | 0.3950(−1) | 0.4740(−1) | 0.5331(−1) | 0.5792(−1) | 0.9471(−1) | 0.2605 | 0.4219 |

duction of TOL1 and TOL2) and the calculations terminate when $U_{\max} = 26\,099$ and 1191, respectively. However, the blow-up rates still seem to be quite stable to two significant digits as $i$ increases.

The outcome of numerical simulations performed for the borderline case $p = 4$ shows more subtle aspects than do those relating to the cases where $p \geqslant 5$. The simple amplitude perturbation in which $\lambda$ is taken to be slightly larger than 1 in (5.5) did not seem to produce a solution that blows up in finite time. However, when the solitary wave was perturbed in a way that corresponds to the instability theory of Bona *et al.* (1987), a solution formed that appears to blow up in finite time (see figure 8). The special initial datum that generated the evolution depicted in figure 8 was

$$u_0(x) = 1.01\{A\ \operatorname{sech}^{2/p}(v(x)) + 0.02[1 - w(x)\tanh(v(x))]\}, \tag{5.8}$$

where $v(x) = C_1(x - \frac{1}{2})$, $C_1 = \frac{1}{2}p(2A^p/\epsilon(p+1)(p+2))^{1/2}$ and $w$ is a cutoff function defined by

$$w(x) = \begin{cases} -\frac{1}{4}C_1 & \text{if } 0 \leqslant x \leqslant \frac{1}{4}, \\ v(x) & \text{if } \frac{1}{4} < x < \frac{3}{4}, \\ \frac{1}{4}C_1 & \text{if } \frac{3}{4} \leqslant x \leqslant 1. \end{cases}$$

Table 17. *Blow-up rates. Perturbed solitary wave,* $p = 4$, $\epsilon = 5 \times 10^{-4}$, $\tau_f = 0.59411 \times 10^{-1}$, $f = 10$, $x^* = 0.74118$, $U_{\max} = 26.56$, $k_{\min} = 0.76 \times 10^{-9}$, $\Delta\tau_f = 0.25 \times 10^{-6}$, TOL1 $= 0.1$, TOL2 $= 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 2 | 0.4990(−1) | 0.8523(−1) | 0.1048 | 0.1170 | 0.1652 | 0.3479 | 0.5501 |
| 3 | 0.5278(−1) | 0.9052(−1) | 0.1115 | 0.1247 | 0.1889 | 0.3714 | 0.5068 |
| 4 | 0.5435(−1) | 0.9196(−1) | 0.1124 | 0.1252 | 0.1870 | 0.3746 | 0.5701 |
| 5 | 0.5483(−1) | 0.9069(−1) | 0.1101 | 0.1225 | 0.1835 | 0.3673 | 0.5491 |
| 6 | 0.5958(−1) | 0.9612(−1) | 0.1161 | 0.1291 | 0.1939 | 0.3869 | 0.5796 |
| 7 | 0.6259(−1) | 0.9918(−1) | 0.1195 | 0.1328 | 0.1994 | 0.3990 | 0.6026 |
| 8 | 0.6407(−1) | 0.1001 | 0.1204 | 0.1339 | 0.2014 | 0.4023 | 0.5981 |

We took $A = 2$, $p = 4$ and $\epsilon = 5 \times 10^{-4}$ in the calculations whose outcome is depicted in figure 8. Note that $u_0(x)$ is continuous but not differentiable at $x = \frac{1}{4}$ and at $x = \frac{3}{4}$; this aspect did not seem to have any significant effect on the computation. The initial value $u_0$ in (5.8) should not be thought of as a small perturbation to the solitary wave for $p = 4$ in the same sense as were the previous initial profiles (for $p \geqslant 5$). Rather, it is a datum obtained from the solitary wave by first adding a constant value which results in non-zero asymptotic values at $x = 0$ and $x = 1$, and then perturbing the result in a very special direction which in turn is cut off to preserve periodicity. Despite the more substantive nature of the perturbation than those seen heretofore, the solution that develops from $u_0$ appears to be dominated by travelling-wave behaviour for small time. However, as the perturbation is felt, the solution steepens and develops a thin spike like those seen for the cases $p \geqslant 5$. Following this leading peak is an almost horizontal shelf which develops more complex structure as the evolution continues. At about $t = 0.05941$ ($f = 10$) the leading peak had risen to a maximum value of about 26.56. Our numerical simulation appeared to lose accuracy thereafter. The rates of the supposed blow-up are recorded in table 17. They are not as convincing as those obtained earlier for the cases where $p > 4$, perhaps reflecting the difficulty the numerical procedures were experiencing with this borderline case.

It is worth noting that we tried a considerable range of pertubations of the solitary wave in the case $p = 4$ that seemed not to lead to blow-up. For example, we tried perturbing the solitary wave in the direction $1 - v\tanh(v)$ mollified by a weight function that brought the perturbed initial data rapidly and smoothly to zero at $x = 0$ and $x = 1$, but the resulting solution did not appear to form a singularity in finite time.

In a second set of experiments we computed the solution of the GKdV equation starting from an initial Gaussian profile

$$u_0(x) = \exp\left(-100(x - \tfrac{1}{2})^2\right), \tag{5.9}$$

taking at first $p = 5, 6$, with $\epsilon = 1.21 \times 10^{-4}$. (The calculations started with $h_0 = 1/384$, and $k_0 = 0.5 \times 10^{-3}$.) For $p = 5$ it was observed that the solution produced a solitary-type wave travelling to the right and a 'hump' that followed. After emerging, the solitary wave became unstable and rapidly formed a singularity

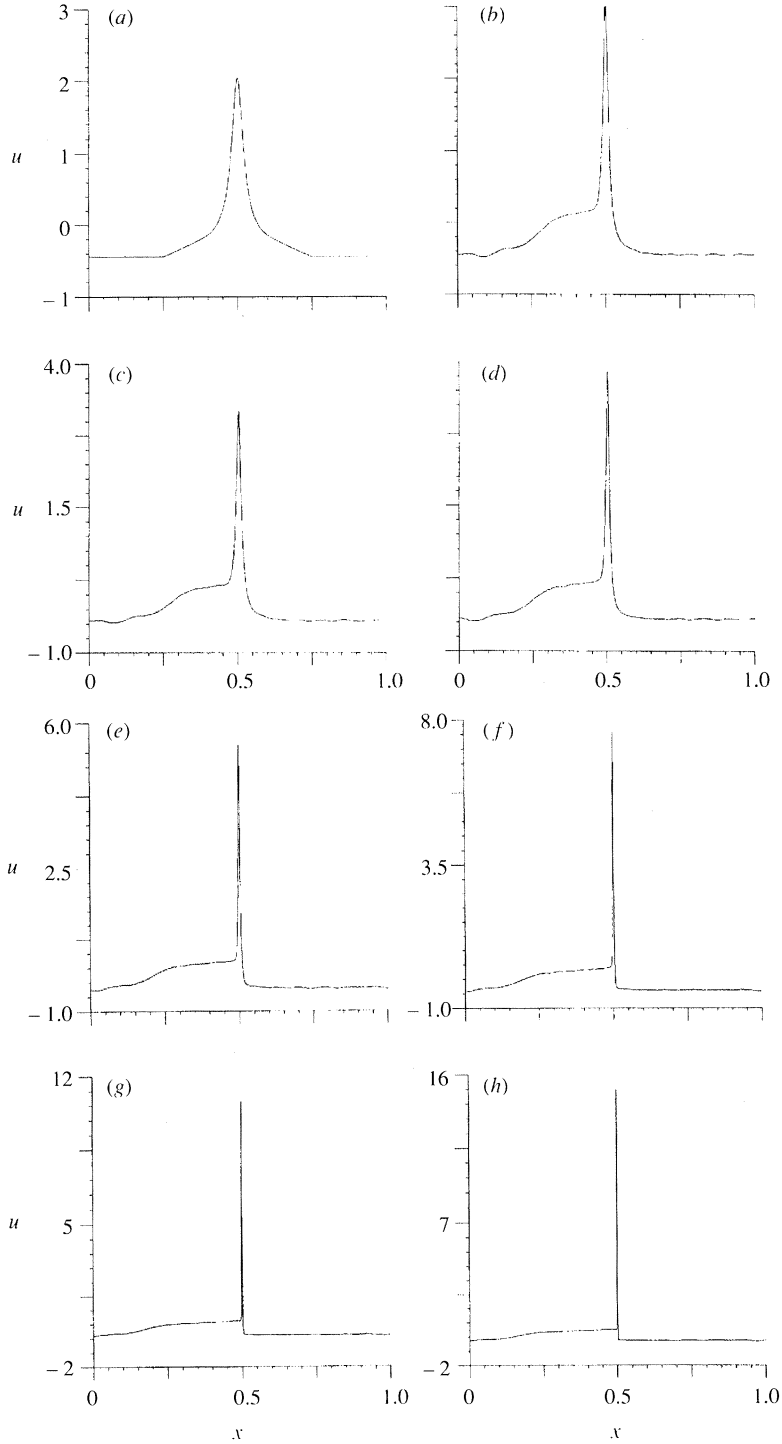


Figure 8. Blow-up of a perturbed solitary wave, $p = 4$. (Data corresponding to the values listed in table 17.) (*a*) $t = 0.0000$, $u_{\max} = 2.0$; (*b*) $t = 0.0542$, $u_{\max} = 3.1$; (*c*) $t = 0.0550$, $u_{\max} = 3.2$; (*d*) $t = 0.0578$, $u_{\max} = 3.8$; (*e*) $t = 0.0592$, $u_{\max} = 5.5$; (*f*) $t = 0.0594$, $u_{\max} = 7.7$; (*g*) $t = 0.0594$, $u_{\max} = 11$; (*h*) $t = 0.0594$, $u_{\max} = 15$.

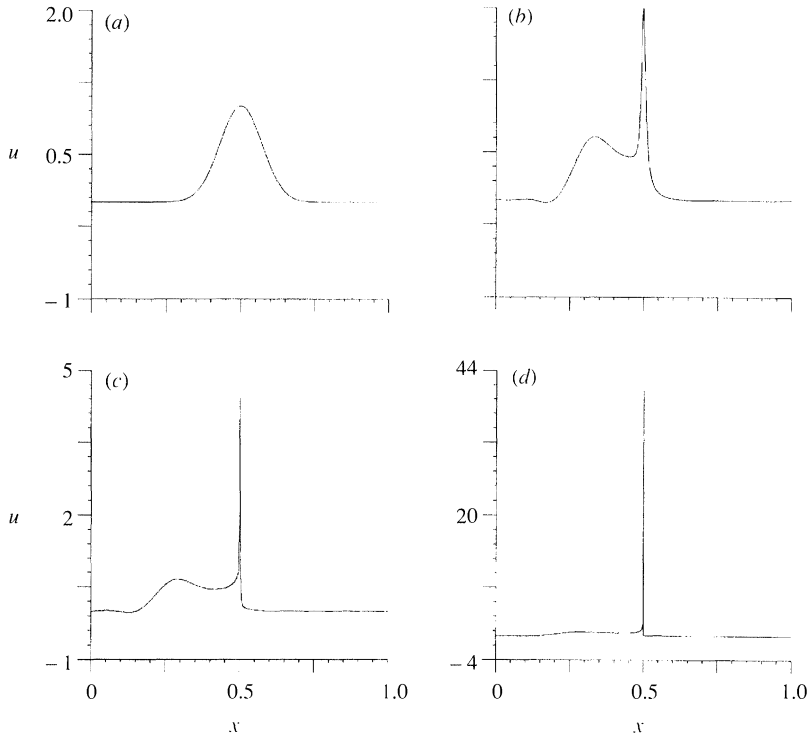


Figure 9. Gaussian: resolution into solitary wave and blow-up, $p = 5$. (Data corresponds to values listed in table 18.) (*a*) $t = 0.000$, $u_{\max} = 1.0$; (*b*) $t = 0.227$, $u_{\max} = 2.0$; (*c*) $t = 0.234$, $u_{\max} = 4.442$; (*d*) $t = 0.234$, $u_{\max} = 41$.

Table 18. *Gaussian initial data*, $p = 5$, $\epsilon = 1.21 \times 10^{-4}$, $\tau_f = 0.23429$, $f = 29$, $x^* = 0.67596$, $U_{max} = 2585$, $k_{\min} = 0.33 \times 10^{-26}$, $\Delta\tau_f = .31 \times 10^{-24}$, TOL1 = 0.1, TOL2 = $10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.4744(−1) | 0.6579(−1) | 0.7740(−1) | 0.8545(−1) | 0.1330 | 0.3002 | 0.4695 |
| 10 | 0.4970(−1) | 0.6672(−1) | 0.7789(−1) | 0.8587(−1) | 0.1342 | 0.3004 | 0.4646 |
| 15 | 0.5049(−1) | 0.6744(−1) | 0.7873(−1) | 0.8682(−1) | 0.1353 | 0.3037 | 0.4689 |
| 20 | 0.5025(−1) | 0.6701(−1) | 0.7818(−1) | 0.8617(−1) | 0.1341 | 0.3015 | 0.4734 |
| 25 | 0.5003(−1) | 0.6673(−1) | 0.7787(−1) | 0.8584(−1) | 0.1337 | 0.3004 | 0.4669 |

(see figure 9). The same behavior occurs for $p = 6$. The computed blow-up rates and the other blow-up data are shown in tables 18 and 19. It is interesting to note that the rates appear to be well-determined and are practically the same as the blow-up rates associated with the perturbed solitary-wave solutions recorded in tables 14 and 15, respectively. This leads to an interesting speculation that the solution begins its evolution by resolving itself into solitary waves and these in turn become unstable due to the ambient environment in which they find themselves.

In the case $p = 4$ (figure 10), an experiment was run with initial datum given

Table 19. *Gaussian initial data,* $p = 6$, $\epsilon = 1.21 \times 10^{-4}$, $\tau_f = 0.18864$, $x^* = 0.62763$, $f = 32$, $U_{\max} = 1503$, $k_{\min} = 0.32 \times 10^{-29}$, $\Delta\tau_f = 0.17 \times 10^{-27}$, TOL1 $= 0.1$, TOL2 $= 10^{-5}$

| $i$ | $L_{p-1}$ | $L_p$ | $L_{p+1}$ | $L_{p+2}$ | $L_\infty$ | $L_{2,D}$ | $L_{\infty,D}$ |
|---|---|---|---|---|---|---|---|
| 5 | 0.4390(−1) | 0.5576(−1) | 0.6401(−1) | 0.7009(−1) | 0.1122 | 0.2805 | 0.4455 |
| 10 | 0.4482(−1) | 0.5623(−1) | 0.6436(−1) | 0.7047(−1) | 0.1135 | 0.2818 | 0.4586 |
| 15 | 0.4431(−1) | 0.5540(−1) | 0.6331(−1) | 0.6924(−1) | 0.1109 | 0.2769 | 0.4450 |
| 20 | 0.4434(−1) | 0.5543(−1) | 0.6335(−1) | 0.6929(−1) | 0.1110 | 0.2771 | 0.4440 |
| 25 | 0.4432(−1) | 0.5539(−1) | 0.6328(−1) | 0.6920(−1) | 0.1105 | 0.2769 | 0.4481 |

by the function $u_0(x) = 3\exp(-100(x - \frac{1}{2})^2) - 1$, which is a shifted Gaussian profile with initial amplitude equal to 2 with tails decaying to $-1$, in order to speed up the growth of the incipient instability. This initial datum evolved in time under the auspices of the fully discrete, adaptive scheme using the starting parameter values $h_0 = 1/768$, $k_0 = 5 \times 10^{-4}$, $\epsilon = .121 \times 10^{-3}$, TOL1 $= 0.1$, and TOL2 $= 10^{-5}$. It was observed that the initial profile rapidly resolved itself into a series of pulses, three of which had clearly formed by the 'blow-up' time. The first pulse rose to a height of 13.24 at $t = .7398 \times 10^{-2}$. By that time, the code had cut the time step down to $0.95 \times 10^{-9}$ and the spatial mesh length locally to a minimum value of about $5 \times 10^{-6}$. Figure 10 suggests that the leading solitary-wave is on its way to blow-up. Of course, we cannot be categorical about this conclusion: the adaptive mechanism in our code is limited by our spatial refinement and translation technique which, as explained previously, is geared toward resolving a single, large-amplitude peak. In this case, a good part of the solution lies outside the region $\Omega_*$ of finest spatial refinement. It is expected that an improved version of our code will be capable of resolving this point more convincingly.

In summary, the suggestion that emerges from the set of experiments with Gaussian initial data is just the following. It seems that solutions with arbitrary initial conditions $u_0(x)$ evolve into a series of solitary-wave pulses followed by a dispersive tail. If $p \geqslant 5$ the leading solitary wave is able to form clearly and separate from the rest of the solution. Once it has emerged as an independently propagating entity, this solitary wave proceeds to become unstable and rapidly forms a singularity before the remainder of the solution has sufficient time to evolve. If $p$ is equal to 4 several such pulses have been formed by the time the first one apparently blows up.

## 6. Conclusions and conjectures

In § 2 of this paper a numerical scheme involving approximation by smooth splines in the spatial variable and an implicit Runge–Kutta discretization in the temporal variable was proposed for the periodic initial-value problem for the GKdV equations. This scheme, which has high formal order in both variables, was proved to converge at high rates in § 3, at least in the presence of a weak stability limitation. A discussion of the implementation of the scheme as a computer program and a detailed study of the accuracy of this code was provided in § 4. It was found that the suggested methods were markedly superior in several

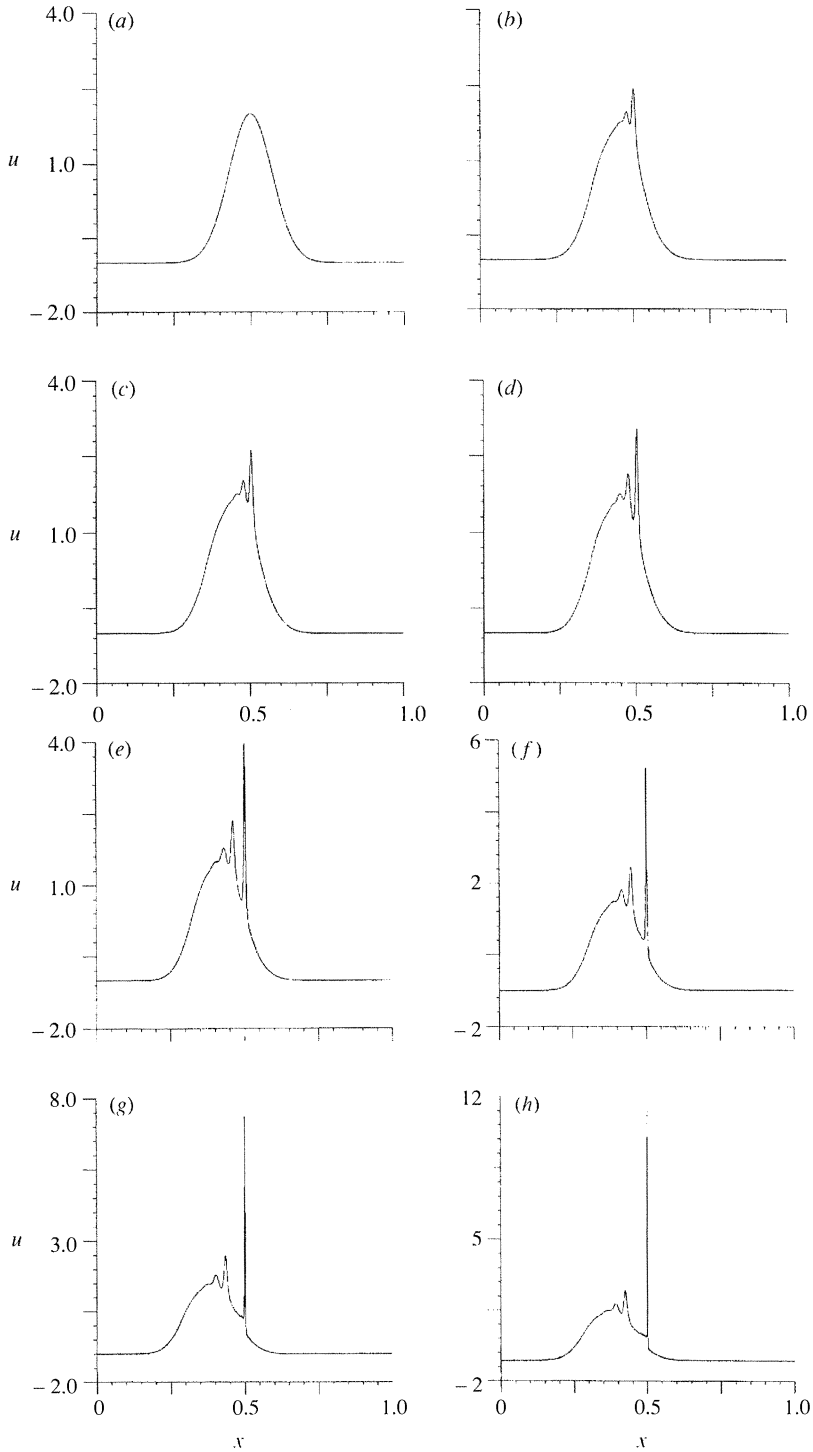


Figure 10. Evolution of Gaussian initial profile, $p = 4$. (*a*) $t = 0.00$, $u_{\max} = 2.0$; (*b*) $t = 3.75 \times 10^{-3}$, $u_{\max} = 2.4$; (*c*) $t = 4.25 \times 10^{-3}$, $u_{\max} = 2.6$; (*d*) $t = 5.25 \times 10^{-3}$, $u_{\max} = 3.0$; (*e*) $t = 6.69 \times 10^{-3}$, $u_{\max} = 3.9$; (*f*) $t = 7.22 \times 10^{-3}$, $u_{\max} = 5.2$; (*g*) $t = 7.37 \times 10^{-3}$, $u_{\max} = 7.4$; (*h*) $t = 7.40 \times 10^{-3}$, $u_{\max} = 11.5$.




PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

ways to those used previously. As reported in § 5, this code was used to study instability and singularity formation in the GKdV equation. It was observed that a uniform-mesh scheme could not resolve the structures developing out of the instability of solitary waves, and consequently an adaptive, variable-mesh version of the scheme was proposed and implemented. A computer code derived from this scheme provided detailed information about the structure of the instability, which apparently includes the formation of a singularity. Both growth properties of the putative singularity and graphs of the entire solution near the singularity were presented.

An appraisal of the numerical results reported in § 5 leads to an interesting conjecture. First, as remarked previously, it is evident from figure 7 that the essential part of the blow-up is of a similarity form. Letting $t^*$ again denote the time at which a singularity forms, one is led to conjecture that the solution $u$ has the form

$$u = \frac{1}{(t^* - t)^\alpha} \chi \left( \frac{x^* - x}{(t^* - t)^\beta} \right) + \text{ bounded term.} \tag{6.1}$$

The presumption (6.1) in turn implies that if $q > \beta/\alpha$ the $L_q$-norm of the solution $u$ will have the form

$$|u(\cdot, t)|_q \sim \frac{c}{(t^* - t)^{\alpha - \beta/q}}, \tag{6.2}$$

or that its blow-up rate $\rho$ is given by

$$\rho(|u|_q) = \alpha - \beta/q \tag{6.3}$$

as $t$ approaches $t^*$ from below. By formally differentiating (6.1), one is led to a blow-up rate for the $W_q^1$-semi-norm, namely

$$\rho(|u_x|_q) = \alpha + \beta(q-1)/q, \tag{6.4}$$

as $t$ approaches $t^*$.

Assuming blow-up occurs in the form (6.1), interesting additional information can be deduced as is now demonstrated. The temporal invariance of $I_3$ (see (4.13)) implies that $\|u_x\|^2$ and $|u|_{p+2}^{p+2}$ have the same rate of blow-up. Using (6.3) and (6.4), this observation yields that

$$2(\alpha + \tfrac{1}{2}\beta) = (p+2)\,(\alpha - \beta/(p+2))\,,$$

or what is the same,

$$\alpha = 2\beta/p. \tag{6.5}$$

Thus the $L_q$-norm is expected to blow up if $q > \frac{1}{2}p$. The invariance of the $L_2$-norm (see again (4.13)) implies that $\alpha \leqslant \frac{1}{2}\beta$, and this combined with (6.5) means that $p \geqslant 4$ for blow-up to occur, a prediction in compliance with the theory reviewed in § 2. If the relations (6.5) and (6.3)–(6.4) are coupled, it is determined that if $p \geqslant 4$, then

$$\rho\,(|u|_q) = \alpha\,(1 - p/2q) \tag{6.6a}$$

for $q > \frac{1}{2}p$ and

$$\rho\,(|u_x|_q) = \alpha\,(1 + p(q-1)/2q)\,. \tag{6.6b}$$

As soon as a good estimate of the blow-up rate for $|u(\cdot, t)|_q$ is available for some value of $q$, then the values of $\alpha$ and $\beta$ are formally determined by (6.5) and (6.6).

Table 20. *Predicted rates of blow-up for various norms for $p = 4,\ 5,\ 6,\ 7$*

| norm\p | 4 | 5 | 6 | 7 |
|---|---|---|---|---|
| $L^{p-1}$ | $\frac{1}{18} = 0.555(-1)$ | $\frac{1}{20} = 0.5(-1)$ | $\frac{2}{45} = 0.444(-1)$ | $\frac{5}{126} = 0.397(-1)$ |
| $L^{p}$ | $\frac{1}{12} = 0.833(-1)$ | $\frac{1}{15} = 0.667(-1)$ | $\frac{1}{18} = 0.555(-1)$ | $\frac{1}{21} = 0.476(-1)$ |
| $L^{p+1}$ | $\frac{1}{10} = 0.1$ | $\frac{7}{90} = 0.778(-1)$ | $\frac{4}{63} = 0.635(-1)$ | $\frac{3}{56} = 0.536(-1)$ |
| $L^{p+2}$ | $\frac{1}{9} = 0.111$ | $\frac{3}{35} = 0.857(-1)$ | $\frac{5}{72} = 0.694(-1)$ | $\frac{11}{189} = 0.582(-1)$ |
| $L^{\infty}$ | $\frac{1}{6} = 0.167$ | $\frac{2}{15} = 0.133$ | $\frac{1}{9} = 0.111$ | $\frac{2}{21} = 0.952(-1)$ |
| $L^{2}_{D}$ | $\frac{1}{3} = 0.333$ | $\frac{3}{10} = 0.3$ | $\frac{5}{18} = 0.278$ | $\frac{11}{42} = 0.262$ |
| $L^{\infty}_{D}$ | $\frac{1}{2} = 0.5$ | $\frac{7}{15} = 0.467$ | $\frac{4}{9} = 0.444$ | $\frac{3}{7} = 0.429$ |

Consideration of the blow-up rates for the $L_\infty$-norm observed in the numerical simulations for $p = 4, 5, 6$ and 7 leads to the conjecture that $\rho(|u|_\infty) = 2/3p$. From this and (6.5) it is concluded that

$$\alpha = 2/3p \quad \text{and} \quad \beta = \tfrac{1}{3}. \tag{6.7}$$

On the basis of these predictions, the relations in (6.6) become

$$\left.\begin{aligned} \rho(|u|_q) &= \frac{1}{3}\left(\frac{2}{p} - \frac{1}{q}\right) \\ \text{and} \qquad \rho(|u_x|_q) &= \frac{1}{3}\left(\frac{2}{p} + 1 - \frac{1}{q}\right). \end{aligned}\right\} \tag{6.8}$$

These blow-up rates are tabulated for the reader's convenience in table 20 for $p = 4, 5, 6, 7$ and $q = p-1,\ p,\ p+1,\ p+2,\ \infty$ for $|u|_q$ and $q = 2,\ \infty$ for $|u_x|_q$. Comparing the entries in table 20 with the experimentally obtained results reported in table 11 ($p = 5$), table 15 ($p = 6$), table 16 ($p = 7$) and table 17 ($p = 4$) lends considerable credibility to the presumption (6.1) with $\alpha$ and $\beta$ as in (6.7). Indeed, the agreement between the predictions in table 20 with the experimental results is exact to two decimal places in most cases. The agreement is not as good in the critical exponent case $p = 4$.

In conclusion, it is worth note that if (6.1) and (6.7) are valid, then the function $\chi$, which depends on $p$ of course, would satisfy a simple ordinary differential equation, namely

$$-\tfrac{2}{3}\chi(y)/p - \tfrac{1}{3}y\chi'(y) + \chi^p(y)\chi'(y) + \epsilon\chi'''(y) = 0. \tag{6.9}$$

This equation corresponds to the scaling law

$$u \mapsto \lambda^{2p}u, \qquad x \mapsto \lambda x, \qquad t \mapsto \lambda^3 t \tag{6.10}$$

for equation (1.1), as noted by F. B. Weissler (1990, personal communication). In fact, the assumption (6.1) is not quite correct; the peak does in fact propagate to the right as it blows up. A more elaborate analysis indicates that the point $X(t)$ where the spike achieves its maximum value at time $t < t^*$ evolves according to the law $X(t) = x^* - c(t^* - t)^{1/3}$ for some constant $c$. Progress has been made on the analysis of equation (6.9) which will be reported separately. It is worth note that the large-$y$ asymptotics of a putative solution of (6.9) precludes membership in the $L_2$-based function classes in which the initial-value problem is typically

posed. Hence if (6.9) governs the similarity forms found in § 5, then a question of matching solutions of (6.9) to appropriately decaying tails arises. These issues are currently under study.

Work supported in part by the Institute of Applied and Computational Mathematics of the Research Center of Crete, Iraklion, Greece, the National Science Foundation, U.S.A., the Keck Foundation, U.S.A., the Air Force Office for Scientific Research under grant AFOSR-88-0019, and the Science Alliance, University of Tennessee.

## References

Ablowitz, M. J. & Segur, H. 1981 *Solitons and the inverse scattering transform*. Philadelphia: SIAM.

Adams, R. A. 1975 *Sobolev spaces*. New York: Academic Press.

Albert, J. P. 1992 Positivity properties and stability of solitary-wave solutions of model equations for long waves. *Commun. in PDE.* **17**, 1–22.

Albert, J. P. & Bona, J. L. 1991 Total positivity and the stability of solitary waves in stratified fluids of finite depth. *IMA J. Appl. Math.* **46**, 1–19.

Albert, J. P., Bona, J. L. & Felland, M. 1988 A criterion for the formation of singularities for the generalized Korteweg–de Vries equation. *Mat. Aplic. e Comp.* **17**, 3–11.

Albert, J. P., Bona, J. L. & Henry, D. 1987 Sufficient conditions for stability of solitary-wave solutions of model equations for long waves. *Physica D* **24**, 343–366.

Baker, G. A., Bramble, J. H. & Thomée, V. 1977 Single step Galerkin approximations to parabolic problems. *Math. Comp.* **31**, 818–847.

Baker, G. A., Dougalis, V. A. & Karakashian, O. A. 1983 Convergence of Galerkin approximations for the Korteweg–de Vries equation. *Math. Comp.* **40**, 419–433.

Benjamin, T. B. 1972 The stability of solitary waves. *Proc. R. Soc. Lond.* A **328**, 153–183.

Benjamin, T. B., Bona, J. L. & Mahony, J. J. 1972 Model equations for long waves in nonlinear, dispersive systems. *Phil. Trans. R. Soc. Lond.* A **272**, 47–78.

Bennett, D. P., Brown, S. E., Stansfield, J. D., Stroughair, J. D. & Bona, J. L. 1983 The stability of internal solitary waves. *Proc. Camb. phil. Soc.* **94**, 351–379.

Bona, J. L. 1975 On the stability theory of solitary waves. *Proc. R. Soc. Lond.* A **344**, 363–374.

Bona, J. L. 1980 Model equations for waves in nonlinear, dispersive systems. In *Proc. Int. Congress of Mathematicians, Helsinki 1978*, vol. 2, 887–894. Hungary: Academia Scientiarum Fennica.

Bona, J. L. 1981*a* On solitary waves and their role in the evolution of long waves. In *Applications of nonlinear analysis in the physical sciences* (ed. H. Amann, N. Bazley & K. Kirchgässner), 183–205. London: Pitman.

Bona, J. L. 1981*b* Convergence of periodic wavetrains in the limit of large wavelength. *Appl. Scientific Res.* **37**, 21–30.

Bona, J. L., Dougalis, V. A. & Karakashian, O. A. 1986 Fully discrete Galerkin methods for the Korteweg–de Vries equation. *Comp. Math. Applic.* A **12**, 859–884.

Bona, J. L., Pritchard, W. G. & Scott, L. R. 1981 An evaluation of a model equation for water waves. *Phil. Trans. R. Soc. Lond.* A **302**, 457–510.

Bona, J. L. & Sachs, R. 1988 Global existence of smooth solutions and stability of solitary waves for a generalized Boussinesq equation. *Commun. Math. Phys.* **118**, 15–29.

Bona, J. L. & Smith, R. 1975 The initial-value problem for the Korteweg–de Vries equation. *Phil. Trans. R. Soc. Lond.* A **278**, 555–604.

Bona, J. L., Souganidis, P. E. & Strauss, W. A. 1987 Stability and instability of solitary waves of KdV type. *Proc. R. Soc. Lond.* A **411**, 395–412.

Bourgain, J. 1993 Fourier restriction phenomena for certain lattice subsets and applications to nonlinear evolution equations. II. The KdV equations. *Geom. funct. Analysis* **3**, 209–262.

Bourgain, J. 1994 On the Cauchy problem for the periodic KdV equations. IHES. (Preprint.)





PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Burrage, K. & Butcher, J. C. 1979 Stability criteria for implicit Runge–Kutta methods. *SIAM J. Numer. Analysis* **16**, 46–57.

Butcher, J. C. 1975 A stability property of implicit Runge–Kutta methods. *BIT* **15**, 357–361.

Butcher, J. C. 1975 *The numerical analysis of ordinary differential equations; Runge–Kutta methods and general linear methods.* Chichester: John Wiley.

Crouzeix, M. 1979 Sur la B-stabilité des méthodes de Runge–Kutta. *Numer. Math.* **32**, 75–82.

Dekker, K. & Verwer, J. G. 1984 *Stability of Runge–Kutta methods for stiff nonlinear differential equations.* Amsterdam: North-Holland.

Dougalis, V. A. & Karakashian, O. A. 1985 On some high order accurate fully discrete Galerkin methods for the Korteweg–de Vries equation. *Math. Comp.* **45**, 329–345.

Fairweather, G. 1978 A note on the efficient implementation of certain Padé methods for linear parabolic problems. *BIT* **18**, 101–109.

Grillakis, M., Shatah, J. & Strauss, W. A. 1987 Stability theory of solitary waves in the presence of symmetry. *J. Funct. Analysis* **74**, 160–197.

Hammack, J. L. 1973 A note on tsunamis: their generation and propagation in an ocean of finite depth. *J. Fluid Mech.* **60**, 769–799.

Hammack, J. L. & Segur H. 1974 The Korteweg–de Vries equation and water waves. Part 2: comparison with experiment. *J. Fluid Mech.* **65**, 289–314.

Karakashian, O. A. & McKinney, W. R. 1990 On optimal high order in time approximations for the Korteweg–de Vries equation. *Math. Comp.* **55**, 473–496.

Kato, T. 1983 On the Cauchy problem for the (generalized) Korteweg–de Vries equation. In *Studies in Appl. Math. Advances in Math. Suppl. Studies*, vol. 8, 93–130. New York: Academic Press.

Korteweg, D. J. & de Vries, G. 1895 On the change of form of long waves advancing in a rectangular canal and on a new type of long stationary wave. *Phil. Mag.* **39**, 422–443.

McLeod, J. B. & Olver, P. J. 1983 The connection between partial differential equations solvable by inverse scattering and ordinary differential equations of Painlevé type. *SIAM J. Math. Analysis* **14**, 56–75.

Miura, R. M. 1968 A remarkable explicit nonlinear transformation. *J. Math. Phys.* **9** 1202–1205.

Pego, R. L. & Weinstein, M. I. 1992 Eigenvalues and instabilities of solitary waves. *Phil. Trans. R. Soc. Lond.* A **340**, 47–94.

Saut, J.-C. 1975 Applications de l'interpolation nonlinéaire à des problèmes d'évolution non linéaires. *J. Math. pures appl.* **13**, 27–52.

Souganidis, P. E. & Strauss, W. A. 1990 Instability of a class of dispersive solitary waves. *Proc. R. Soc. Edinb.* A **114**, 195–212.

Weinstein, M. I. 1986 Lyapunov stability of ground states of nonlinear dispersive evolution equations. *Communs. pure appl. Math.* **39**, 51–68.

Weinstein, M. I. 1987 Existence and dynamic stability of solitary wave solutions of equations arising in long wave propagation. *Communs. partial diffl. Equat.* **12**, 1133–1173.

Zabusky, N. J. & Galvin, C. J. 1971 Shallow-water waves, the Korteweg–de Vries equation and solitons. *J. Fluid Mech.* **47**, 811–824.

*Received 24 August 1992; revised 2 June 1994; accepted 26 July 1994*





PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES