MATHEMATICS OF COMPUTATION
VOLUME 45, NUMBER 172
OCTOBER 1985, PAGES 329–345

On Some High-Order Accurate Fully Discrete Galerkin Methods for the Korteweg-de Vries Equation*

By Vassilios A. Dougalis and Ohannes A. Karakashian

Abstract. We construct and analyze fully discrete Galerkin (finite-element) methods of high order of accuracy for the numerical solution of the periodic initial-value problem for the Korteweg-de Vries equation. The methods are based on a standard space discretization using smooth periodic splines on a uniform mesh. For the time stepping, we use two schemes of third (resp. fourth) order of accuracy which are modifications of well-known, diagonally implicit Runge-Kutta methods and require the solution of two (resp. three) nonlinear systems of equations at each time step. These systems are solved approximately by Newton's method. Provided the initial iterates are chosen in a specific, accurate way, we show that only one Newton iteration per system is needed to preserve the stability and order of accuracy of the scheme. Under certain mild restrictions on the space mesh length and the time step we prove L^2-error estimates of optimal rate of convergence for both schemes.

1. Introduction. In this paper we shall be concerned with the numerical solution by fully discrete Galerkin methods of the periodic initial-value problem for the Korteweg-de Vries (KdV) equation. Given $0 < T < \infty$, we shall approximate a real-valued function, $u = u(x, t)$, for $(x, t) \in [0,1] \times [0, T]$, 1-periodic in x for all $t \in [0, T]$, and satisfying

$$\begin{cases} u_t + uu_x + u_{xxx} = 0, & (x,t) \in [0,1] \times (0,T], \\ u(x,0) = u^0(x), & x \in [0,1], \end{cases} \tag{1.1}$$

where u^0 is a given 1-periodic function smooth enough, cf., e.g., [3], to guarantee that (1.1) has a unique, sufficiently smooth solution so that the various convergence estimates below hold. For error estimates of other numerical methods for (1.1), cf., e.g., the references of [2].

We begin by introducing notation. For integer $s \geqslant 0$ and real $1 \leqslant p \leqslant \infty$, denote by $W_p^s = W_p^s(0,1)$ the usual real Sobolev spaces on $(0,1)$, and by $\|\cdot\|_{s,p}$ the associated norms. Let $H^s = W_2^s$ and $\|\cdot\|_s = \|\cdot\|_{s,2}$. The inner product and norm on $L^2 = L^2(0,1)$ are denoted by $(\cdot,\cdot)$ and $\|\cdot\|$, respectively, and the norm of $L^\infty(0,1)$ by $\|\cdot\|_\infty$. If v: $[0,T] \to X$ is a (strongly) measurable map with values in a Banach space $\{X, \|\cdot\|_X\}$, let

$$\|v\|_{L^p(X)} = \left[\int_0^T \|v(t)\|_X^p \, dt\right]^{1/p} \quad \text{for } 1 \leqslant p < \infty,$$

$$\|v\|_{L^\infty(X)} = \operatorname*{ess\,sup}_{0 \leqslant t \leqslant T} \|v(t)\|_X.$$

Received February 8, 1983; revised October 24, 1983 and October 5, 1984.
1980 *Mathematics Subject Classification*. Primary 65N30; Secondary 35Q20, 65M15, 65L07.
*Research supported in part by USARO Grant No. DAAG29-80-K-0056.

0025-5718/85 $1.00 + $.25 per page

For integer $r \geqslant 4$, let S_h^r be the space of 1-periodic smooth splines of order r (degree $r-1$) on $[0,1]$ with uniform mesh length $h = 1/N$ for integer $N > 0$. It is well-known that if v is 1-periodic and sufficiently smooth, then, there exists a $\chi \in S_h^r$ such that

$$\sum_{j=0}^{s-1} h^j \|v - \chi\|_j \leqslant ch^s \|v\|_s, \qquad 1 \leqslant s \leqslant r, \tag{1.2}$$

$$\sum_{j=0}^{m-1} h^j \|v - \chi\|_{j,\infty} \leqslant ch^m \|v\|_{m,\infty}, \qquad 1 \leqslant m \leqslant r. \tag{1.3}$$

Here c is a constant independent of h, v and χ. (Throughout the paper c will denote a generic constant, not necessarily the same in all instances.) In addition, S_h^r satisfies the following *inverse properties*: there exists a constant c, independent of h, such that for all $\chi \in S_h^r$,

$$\|\chi\|_\beta \leqslant ch^{-(\beta-\alpha)} \|\chi\|_\alpha, \qquad \|\chi\|_{\alpha,\infty} \leqslant ch^{-(\alpha+1/2)} \|\chi\|, \qquad 0 \leqslant \alpha \leqslant \beta \leqslant r-1. \tag{1.4}$$

In what follows we let Πu^0 denote any conveniently chosen element of S_h^r (e.g., L^2-projection, interpolant, etc.) that satisfies, for u^0 sufficiently smooth,

$$\|\Pi u^0 - u^0\| \leqslant ch^r. \tag{1.5}$$

We define the map $F\colon S_h^r \times S_h^r \to S_h^r$ for $v, w \in S_h^r$ by

$$(F(v,w), \chi) = \tfrac{1}{2}(vw, \chi_x) + (v_{xxx}, \chi) \quad \forall \chi \in S_h^r, \tag{1.6}$$

i.e., if $P\colon L^2 \to S_h^r$ is the orthogonal L^2-projection operator onto S_h^r, we let $F(v,w) = -P[\frac{1}{2}(vw)_x + v_{xxx}]$. We shall write $F(v) = F(v,v)$ for $v \in S_h^r$. Then, the (standard) Galerkin semidiscretization of (1.1) in S_h^r is a map $v_h\colon [0,T] \to S_h^r$ satisfying

$$v_{ht} = F(v_h), \qquad 0 < t \leqslant T,\ v_h(0) = \Pi u^0. \tag{1.7}$$

It is known, cf. [8], [2], that $v_h(t)$ exists uniquely for $t \in [0,T]$ and, provided u is sufficiently smooth, satisfies the optimal rate of convergence error estimate $\|u - v_h\|_{L^\infty(L^2)} \leqslant ch^r$, for some constant $c = c(u,T)$.

We shall be interested in *full* (i.e., in time also) discretizations of the system of ordinary differential equations (ODE's) represented by (1.7). In [2], a second-order accurate in time Crank-Nicolson type fully discrete scheme (coupled with Newton's method for the solution of the attendant nonlinear systems of equations at each time step) was analyzed. Here we turn to *higher-order accurate* full discretizations. In sequel, let J be a positive integer, put $k = T/J$ and, for a continuous function v defined on $[0,T]$, let $v^n = v(t^n)$, $t^n = nk$, $n = 0, 1, \ldots, J$.

As a base for our fully discrete schemes we shall use two well-known semi-implicit (diagonally implicit) Runge-Kutta (RK) methods; cf., e.g., [5], [1], and the references of [1]. A *q-stage Diagonally Implicit RK (DIRK) method* for the autonomous system of ODE's $y'(t) = f(y(t))$ is determined by a table of real constants of the form $A \mid b$, where $A = (a_{ij})_{1 \leqslant i,j \leqslant q}$ is a lower triangular $q \times q$ matrix with $a_{ii} = \beta \neq 0$ and b is a q-vector $b = [b_1, \ldots, b_q]^T$. The corresponding algorithm produces approximations y^n

to $y(t^n) = y(nk)$ for $n = 0, 1, \ldots,$ given by

$$y^{n,i} = y^n + k \sum_{j=1}^{q} a_{ij} f(y^{n,j}), \qquad 1 \leqslant i \leqslant q, \tag{1.8}$$

$$y^{n+1} = y^n + k \sum_{j=1}^{q} b_j f(y^{n,j}). \tag{1.9}$$

(1.8) is equivalent to

$$y^{n,i} = y^n + k\beta f(y^{n,i}) + \sum_{j=1}^{i-1} \mu_{ij}(y^{n,j} - y^n), \qquad 1 \leqslant i \leqslant q, \tag{1.10}$$

when μ_{ij} are the elements of the $q \times q$ strictly lower triangular matrix $M = I - \beta A^{-1}$. (In (1.10), and elsewhere, we use the convention $\sum_{j=l}^{m} = 0$ if $m < l$). We shall also frequently replace (1.9) by the following (in view of (1.8)) formula, in which $a_{ij}^{-1} \equiv (A^{-1})_{ij}$:

$$y^{n+1} = y^n + \sum_{i,j=1}^{q} b_i a_{ij}^{-1}(y^{n,j} - y^n). \tag{1.11}$$

In particular, we shall consider two such specific DIRK methods. First the two-stage method (i.e., $q = 2$) given by the tableau

$$\begin{array}{cc|l} \beta & 0 & b_1 \\ 1-2\beta & \beta & b_2 \end{array} \quad \begin{array}{l} \beta = (3+\sqrt{3})/6 = \text{the largest root of } \beta^2 - \beta + 1/6 = 0, \\ b_1 = b_2 = 1/2. \end{array} \tag{1.12}$$

We shall also consider in Section 4 the three-stage method ($q = 3$):

$$\begin{array}{ccc|l} \beta & 0 & 0 & b_1 \\ \frac{1}{2}-\beta & \beta & 0 & b_2 \\ 2\beta & 1-4\beta & \beta & b_3 \end{array} \quad \begin{array}{l} \beta = \dfrac{1}{\sqrt{3}}\cos\dfrac{\pi}{18} + \dfrac{1}{2} = \text{the largest root of} \\ \beta^3 - \dfrac{3}{2}\beta^2 + \dfrac{1}{2}\beta - \dfrac{1}{24} = 0, \\ b_1 = b_3 = 1/6(2\beta-1)^2, \quad b_2 = 1 - 2b_1. \end{array} \tag{1.13}$$

It is well-known, cf., e.g., [5], [1], [4], that the methods (1.12), respectively (1.13), have orders of accuracy 3, respectively 4, and good stability properties for a wide class of nonlinear problems. Usually a q-stage RK method of order of accuracy p is called a (q, p) RK method. We shall accordingly refer to (1.12), respectively (1.13), as the (2, 3), respectively (3, 4), DIRK method under consideration.

Using, e.g., the formulation (1.10)–(1.11) we obtain now the following full discretization of (1.7): seek U^n, $0 \leqslant n \leqslant J$ and $U^{n,i}$, $1 \leqslant i \leqslant q$, $0 \leqslant n \leqslant J-1$ in S_h^r satisfying

$$U^0 = \Pi u^0, \tag{1.14}$$

and for $n = 0, 1, \ldots, J-1$,

$$U^{n,i} = U^n + k\beta F(U^{n,i}) + \sum_{j=1}^{i-1} \mu_{ij}(U^{n,j} - U^n), \qquad 1 \leqslant i \leqslant q, \tag{1.15}$$

$$U^{n+1} = U^n + \sum_{i,j=1}^{q} b_i a_{ij}^{-1}(U^{n,j} - U^n). \tag{1.16}$$

In the specific cases of the DIRK methods (1.12) and (1.13) under consideration, it may be shown that if u, the solution of (1.1), is sufficiently smooth, and if the discretization parameters k and h satisfy certain conditions, then, the scheme (1.14)–(1.16) has a unique solution, is consistent with (1.1), and stable. Moreover, numerical experiments that we have performed, indicate, as expected, that the approximate solution of (1.14)–(1.16), obtained using Newton's method for approximating $U^{n,i}$, the solutions of the q $N \times N$ nonlinear systems (1.15), with one Newton iteration per stage using appropriate starting values, is indeed, for every n, $O(k^p + h^r)$ close to u^n in the L^2-norm. However, we were unable to prove rigorously that the scheme has a local error of $O(k^{p+1} + kh^r)$ in L^2 and, consequently, we could not infer that its global L^2-error has the optimal rate of convergence bound of $O(k^p + h^r)$.

We were able, however, to prove that *modified* versions of (1.14)–(1.16), obtained by perturbing (1.15) by "small" terms (that still permit solving q $N \times N$ nonlinear systems for the intermediate stages) yield schemes whose implementation via Newton's method is almost as efficient as that of (1.14)–(1.16) and which are stable and convergent with a global L^2-error of $O(k^p + h^r)$ (for $p = 3$ and $p = 4$). We present the modified scheme in the case $p = 3$ below and summarize our main convergence results. The detailed motivation of the perturbation terms and the local error analysis for this scheme will be presented in Section 2; in Section 3 we prove our optimal rate of convergence L^2-error estimate for the scheme coupled with Newton's method for the solution of the associated nonlinear systems. In Section 4 we state a modified scheme corresponding to (1.13) ($p = 4$) and the relevant convergence result without proof. Details of omitted proofs and numerical experiments may be found in a technical report available from the authors.

The modified fully discrete scheme corresponding to the (2, 3) DIRK method (1.12) (henceforth referred to as the "modified (2, 3) scheme") is defined as follows. Let the map Q: $S_h^r \times S_h^r \to S_h^r$ be given, for $v, w \in S_h^r$, by

$$(Q(v, w), \chi) = \tfrac{1}{2}(vw, \chi_x) \quad \forall \chi \in S_h^r, \tag{1.17}$$

i.e., by $Q(v, w) = P[-\frac{1}{2}(vw)_x]$ and denote $Q(v) = Q(v, v)$. Then, seek U^n, $0 \leqslant n \leqslant J$, $U^{n,i}$, $i = 1, 2$, $0 \leqslant n \leqslant J - 1$ in S_h^r, such that

$$U^0 = \Pi u^0, \tag{1.18}$$

and for $n = 0, 1, \ldots, J - 1$,

$$\begin{aligned} U^{n,i} = U^n + k\beta\big[F(U^{n,i}) + p_i Q(U^{n,i} - U^n)\big] \\ + \sum_{j=1}^{i-1} \mu_{ij}(U^{n,j} - U^n), \qquad i = 1, 2, \end{aligned} \tag{1.19}$$

$$U^{n+1} = U^n + \sum_{i,j=1}^{2} b_i a_{ij}^{-1}(U^{n,j} - U^n). \tag{1.20}$$

Here a_{ij}, b_j, β are given by (1.12), $\mu_{21} = (1 - 2\beta)/\beta$ (all other $\mu_{ij} = 0$) and the perturbation parameters p_i by

$$p_1 = 1, \qquad p_2 = -\beta^2/(1 - \beta)^2. \tag{1.21}$$

It may be proved, using the estimation techniques of Sections 2 and 3, that for each n, the solution U^n of (1.18)–(1.20) exists uniquely and satisfies $\|U^n - u^n\| \leqslant c(k^3 + h^r)$, if u is sufficiently smooth and k and h are sufficiently small and satisfy $k \leqslant \alpha h$ for some $\alpha > 0$. However, we shall not be interested in the solution of (1.18)–(1.20) per se, but we shall approximate $U^{n,i}$ by Newton's method. Let j_0, $j_1, \ldots, j_J$ be a collection of nonnegative integers to be specified below; j_{n+1} will be the number of Newton iterations performed at each one of the stages $i = 1, 2$ in (1.19). For each n, $0 \leqslant n \leqslant J$, we denote by $U_{j_n}^n$ in S_h^r the approximation to U^n, i.e., the final output of the fully discrete scheme at each step. For $0 \leqslant n \leqslant J-1$, given $U_{j_n}^n$ and appropriate starting values $U_0^{n,1}$, $U_0^{n,2}$ in S_h^r, we construct iteratively sequences $U_j^{n,1}$, $U_j^{n,2}$, $j = 1, 2, \ldots, j_{n+1}$ in S_h^r by applying j_{n+1} steps of Newton's method to (1.19) first for $i = 1$ and then for $i = 2$. It is straightforward to see that given $U_j^{n,i}$, $U_{j+1}^{n,i}$ satisfies the following linear system of equations

$$\begin{aligned} &U_{j+1}^{n,i} - k\beta F\big(U_{j+1}^{n,i}, U_j^{n,i}\big) - k\beta Q\big(U_{j+1}^{n,i}, [1+2p_i]U_j^{n,i} - 2p_i U_{j_n}^n\big) \\ &\quad = U_{j_n}^n + \sum_{j=1}^{i-1} \mu_{ij}\big(U_{j_{n+1}}^{n,j} - U_{j_n}^n\big) - k\beta\big[(1+p_i)Q\big(U_j^{n,i}\big) - p_i Q\big(U_{j_n}^n\big)\big], \\ &\qquad 1 \leqslant j \leqslant j_{n+1} - 1,\ i = 1, 2. \end{aligned} \tag{1.22}$$

$U_{j_{n+1}}^n$ is then defined by replacing U^n, $U^{n,i}$ in (1.20) by their final approximations, i.e., by

$$U_{j_{n+1}}^{n+1} = U_{j_n}^n + \sum_{i,j=1}^{2} b_i a_{ij}^{-1}\big(U_{j_{n+1}}^{n,j} - U_{j_n}^n\big). \tag{1.23}$$

"Good" starting values $U_0^{n,i}$ must be supplied for each n and i so that the convergence of the Newton iterates $U_j^{n,i}$ to $U^{n,i}$ is as fast as possible, i.e., so that we may be able by performing just *one* iteration per stage (i.e., by taking $j_{n+1} = 1$) to preserve the stability and global order of accuracy of the "exact" scheme (1.18)–(1.20). It turns out that this is possible if we perform one additional iteration for $n = 0, 1$. We first take

$$j_0 = 0, \qquad U_{j_0}^0 = U_0^0 = \Pi u^0. \tag{1.24}$$

Then we compute an intermediate value $U_*^1 \in S_h^r$ by

$$\big(U_*^1 - U_{j_0}^0, \chi\big) + k\big(U_{j_0}^0[U_*^1]_x, \chi\big) + k\big([U_*^1]_{xxx}, \chi\big) = 0 \quad \forall \chi \in S_h^r. \tag{1.25}$$

The starting values $U_0^{n,i}$, $i = 1, 2$, $n = 0, 1$, are given by

$$U_0^{n,1} = (1 - n - \beta)U_{j_0}^0 + (n + \beta)U_*^1, \qquad n = 0, 1, \tag{1.26}$$

$$U_0^{n,2} = (\beta - n)U_{j_0}^0 + (1 + n - \beta)U_*^1, \qquad n = 0, 1. \tag{1.27}$$

For $n \geqslant 2$, the starting values that we shall use are

$$U_0^{n,i} = d_{i0}U_{j_n}^n + d_{i1}U_{j_{n-1}}^{n-1} + d_{i2}U_{j_{n-2}}^{n-2}, \qquad i = 1, 2, \tag{1.28}$$

where

$$\begin{aligned} &d_{10} = \beta^2 + 3\beta/2 + 1, \quad d_{11} = -2\beta(1+\beta), \quad d_{12} = \beta(1+2\beta)/2, \\ &d_{20} = (6 - 5\beta)/2, \quad d_{21} = 4\beta - 3, \quad d_{22} = (2 - 3\beta)/2. \end{aligned} \tag{1.29}$$

In Theorem 3.1, we prove that, if u is sufficiently smooth, if k and h are sufficiently small and satisfy $k \leqslant \alpha h$ for some $\alpha > 0$, and if we take $j_1 = j_2 = 2$ and $j_n = 1$ for $3 \leqslant n \leqslant J$, then, all intermediate approximations defined by (1.22)–(1.29) exist uniquely. Moreover, there exists a constant $c = c(u, T, \alpha)$ such that

$$\max_{0 \leqslant n \leqslant J} \|U_{j_n}^n - u^n\| \leqslant c(k^3 + h^r).$$

Hence, by solving, for $n \geqslant 3$, two $N \times N$ linear systems of equations per step (one per stage) we may achieve an L^2-bound for the error $U_{j_n}^n - u^n$ of optimal rate of convergence in space and time. It should be noted that the matrices of the linear systems (1.22) (i.e., the Jacobians of the nonlinear systems (1.19)), change from step to step and from stage to stage. However, their sparsity structure is the same as that of, e.g., the Gram matrix associated with the usual B-spline basis of S_h^r. Hence, these matrices are "cyclically banded" due to the periodic boundary conditions and, under the hypotheses, e.g., of Theorem 3.1, are (nonsymmetric) positive definite. Such linear systems can be easily solved by direct methods and updating their elements is not expensive in our one-dimensional situation. Note that the unmodified scheme ($p_1 = p_2 = 0$ in (1.22)) is not significantly less expensive to implement than the modified one.

To perform the error estimations in S_h^r we shall, cf. [2], compare the solutions of the various fully discrete schemes with a certain quasi-interpolant u_h: $[0, T] \to S_h$ of u, which is defined, [2], [7], by

$$u_h(x, t) = \sum_{j=1}^{N} u(jh, t)\tilde{\Phi}_j(x), \qquad (x, t) \in [0,1] \times [0, T], \tag{1.30}$$

where $\{\tilde{\Phi}_j\}_{1 \leqslant j \leqslant N}$ is a suitably chosen basis of S_h^r, cf. [7, Lemma 2.4], so that

$$\|u_h(t) - u(t)\| \leqslant ch^r \left\| \frac{\partial^r u}{\partial x^r}(t) \right\|, \qquad 0 \leqslant t \leqslant T, \tag{1.31}$$

holds. Following the proofs of Lemmas 2.2 and 2.4 of [7], we obtain

$$(u_{ht} + uu_{hx} + u_{hxxx}, \chi) = (\psi(t), \chi) \quad \forall \chi \in S_h^r, 0 \leqslant t \leqslant T, \tag{1.32}$$

where the "truncation error" $\psi(t)$ satisfies, for u sufficiently smooth, the estimates

$$\|D_t^i \psi\|_{L^\infty(L^2)} \leqslant c_i h^r, \qquad i = 0, 1, 2, \ldots. \tag{1.33}$$

Here $D_t^i = \partial^i / \partial t^i$ and the c_i are positive constants depending on u and T only. (We shall henceforth generally omit mentioning that such constants may depend on u and T unless there is a specific reason for doing so. The symbols c_i will also denote generic positive constants not necessarily the same in any two places.) Now, since the quasi-interpolation operator commutes with time differentiation, (1.31) gives

$$\|D_t^i u_h - D_t^i u\|_{L^\infty(L^2)} \leqslant c_i h^r, \qquad i = 0, 1, 2, \ldots. \tag{1.34}$$

It is straightforward to check that (1.34) and (1.2)–(1.4) imply that

$$\|D_t^i u_h\|_{L^\infty(W_\infty^1)} \leqslant c_i, \qquad i = 0, 1, 2, \ldots. \tag{1.35}$$

We finally mention for further reference that (1.32) implies, for each $i = 0, 1, 2, \ldots$, and for all $\chi \in S_h^r$,

$$\left(D_t^i [u_{ht} + u_h u_{hx} + u_{hxxx}](t), \chi\right) = \left(D_t^i [\psi + (u_h - u)u_{hx}](t), \chi\right). \tag{1.36}$$

Note also that it follows from (1.31) and (1.33)–(1.35) that

$$\|D_t^i[\psi + (u_h - u)u_{hx}]\|_{L^\infty(L^2)} \leqslant c_i h^r, \qquad i = 0, 1, 2, \ldots. \tag{1.37}$$

2. Analysis of the Local Error. To study the local error of the scheme (1.19)–(1.20) and also to motivate the choice of the perturbation terms, we first make some remarks on the local errors of the (2, 3) DIRK method in the context of the scalar ODE $D_t y = f(y)$. With f smooth and $y(t^n)$ replacing y^n in (1.8)–(1.9), series expansions in powers of k give

$$y^{n,1} = y(t^n) + k\beta f_n + k^2\beta^2 f_n f_n' + k^3\beta^3\left[(f_n)^2 f_n''/2 + (f_n')^2 f_n\right] + O(k^4), \tag{2.1}$$

and an analogous expression for $y^{n,2}$, where $f_n = f(y(t^n))$, $f_n' = f'(y(t^n))$ etc. Define $\tilde{e}^{n,i}$ to be the residuals after 4 terms of the Taylor expansions of $y^{n,i}$ about t^n; it may be easily seen that

$$y^{n,i} = \sum_{j=0}^{3} \tau_{ij} k^j D_t^j y(t^n) + \tilde{e}^{n,i}, \qquad i = 1, 2, \tag{2.2}$$

where

$$\tau_{ij} = \sum_{l=1}^{2} a_{il}\tau_{l,j-1}, \qquad i = 1, 2, \quad 1 \leqslant j \leqslant 3, \qquad \tau_{i0} = 1, \qquad i = 1, 2. \tag{2.3}$$

It follows, by comparing (2.1) with (2.2), that

$$\tilde{e}^{n,1} = -\beta^3 k^3 (f_n)^2 f_n''/2 + O(k^4). \tag{2.4}$$

In an entirely analogous manner, we obtain

$$\tilde{e}^{n,2} = \mu_{21}\tilde{e}^{n,1} + \beta^3 k^3 (f_n)^2 f_n''/2 + O(k^4). \tag{2.5}$$

Finally, using the order relations of the method, i.e.

$$b^T A^{j-1} e = 1/j!, \qquad 1 \leqslant j \leqslant 3, b = (b_1, b_2)^T, e = (1,1)^T, \tag{2.6}$$

we obtain by (1.11), (2.2), (2.3) that

$$y^{n+1} = y(t^{n+1}) + (1/2\beta)\left[\tilde{e}^{n,1} + \left(\tilde{e}^{n,2} - \mu_{21}\tilde{e}^{n,1}\right)\right] + O(k^4).$$

Hence, (2.4) and (2.5) yield finally that $y^{n+1} = y(t^{n+1}) + O(k^4)$, i.e., that the local error is indeed of $O(k^4)$. The example confirms a well-known property of many RK methods, namely the fact that although some type of intermediate residuals (e.g., the $\tilde{e}^{n,i}$ here) may be of lower order of accuracy (here, third), nevertheless, the large errors cancel and the correct order of the local error emerges when y^{n+1} is finally computed in terms of y^n and $y^{n,i}$. The local error estimate can easily be rigorously justified, cf. [5], for, say, sufficiently smooth f with bounded appropriate higher derivatives. In case the ODE system in question is stiff and represents the semidiscretization of a PDE, one should rigorously justify the computations by setting up equations for the local errors and estimating them using the properties of the particular partial differential operator without imposing severe limitations on k as a function of h that are not dictated by stability requirements. We attempt to do this in our present case by studying the local error of the (2, 3) DIRK method when applied to (1.7).

To this effect, define now, with u_h as in (1.30), $\tilde{V}^{n,i}$ and $\tilde{V}^{n+1}$ in S_h^r by

$$\tilde{V}^{n,i} = u_h^n + k\beta F(\tilde{V}^{n,i}) + \sum_{j=1}^{i-1} \mu_{ij}(\tilde{V}^{n,j} - u_h^n), \qquad i = 1, 2, \tag{2.7}$$

$$\tilde{V}^{n+1} = u_h^n + \sum_{i,j=1}^{2} b_i a_{ij}^{-1}(\tilde{V}^{n,j} - u_h^n). \tag{2.8}$$

(The existence of $\tilde{V}^{n,i}$ follows, e.g., from Lemma 2.1 below.) Put

$$\Lambda_i u_h^n \equiv \sum_{j=0}^{3} \tau_{ij} k^j D_t^j u_h^n, \qquad i = 1, 2, \tag{2.9}$$

and define $\tilde{\varepsilon}^{n,i} \in S_h^r$ by

$$\tilde{V}^{n,i} = \Lambda_i u_h^n + \tilde{\varepsilon}^{n,i}, \qquad i = 1, 2. \tag{2.10}$$

Inserting (2.10) in (2.7) and using (2.9), (1.6), (1.17) and (2.3) yields after a rather lengthy but straightforward computation

$$\begin{aligned} &\tilde{\varepsilon}^{n,i} - 2k\beta Q(\Lambda_i u_h^n, \tilde{\varepsilon}^{n,i}) - k\beta F(\tilde{\varepsilon}^{n,i}) - \sum_{j=1}^{i-1} \mu_{ij} \tilde{\varepsilon}^{n,j} \\ &\quad = \beta k^3 \left[(\tau_{i1})^2 - 2\tau_{i2}\right] Q(D_t u_h^n) + \tilde{E}^{n,i}, \qquad i = 1, 2, \end{aligned} \tag{2.11}$$

where, using (1.35)–(1.37) it can be seen that

$$\max_{n,i} \|\tilde{E}^{n,i}\| \leqslant c(k^4 + kh^r). \tag{2.12}$$

The equations (2.11) are the analogs of (2.4) and (2.5) of the scalar case. Note that the coefficients of the term $Q(D_t u_h^n)$ in (2.11) are equal to $-\beta^3 k^3$, respectively $\beta^3 k^3$, if $i = 1$, respectively 2. Hence, using, e.g., the estimation technique of Proposition 2.1 below, we may infer that we cannot achieve more than $O(k^3)$ temporal accuracy for each $\tilde{\varepsilon}^{n,i}$. Proceeding now to the final phase at step n and substituting (2.9) and (2.10) in (2.8), using (2.6), (2.3), Taylor's theorem and (1.35), we obtain as in the scalar case that

$$\tilde{V}^{n+1} - u_h^{n+1} = E^{n+1} + (1/2\beta)\left[\tilde{\varepsilon}^{n,1} + (\tilde{\varepsilon}^{n,2} - \mu_{21}\tilde{\varepsilon}^{n,1})\right], \tag{2.13}$$

where E^{n+1} is of optimal order, i.e.

$$\max_n \|E^{n+1}\| \leqslant ck^4. \tag{2.14}$$

Now, by (2.11), (2.12) it is seen (in L^2) that

$$\tilde{V}^{n+1} - u_h^{n+1} = (k/2) \sum_{i=1}^{2} \left[2Q(\Lambda_i u_h^n, \tilde{\varepsilon}^{n,i}) + F(\tilde{\varepsilon}^{n,i})\right] + O(k^4 + kh^r).$$

Hence, if, e.g., the nonlinear terms $2Q(\Lambda_i u_h^n, \tilde{\varepsilon}^{n,i}) + F(\tilde{\varepsilon}^{n,i})$ were bounded above in L^2 by a term of $O(k^3 + h^r)$—something that we were unable to show—then, an optimal rate of convergence $O(k^4 + kh^r)$ bound would follow for $\tilde{V}^{n+1} - u_h^{n+1}$.

We now shift our attention to a different strategy: if, by modifying the intermediate stages of the RK method, we could cancel the $O(k^3)$ terms in the right-hand side of (2.11), then (2.11)–(2.14) and the triangle inequality would certainly give the desired $O(k^4 + kh^r)$ bound for $\|\tilde{V}^{n+1} - u_h^{n+1}\|$. One way to do this is by modifying

the scheme as done in (1.19). To study the local error of the modified method, define $V^{n,i}$, $i = 1, 2$ and V^{n+1} in S_h^r (for their existence, cf. Lemma 2.1) by

$$V^{n,i} = u_h^n + k\beta\left[F(V^{n,i}) + p_i Q(V_{n,i} - u_h^n)\right] + \sum_{j=1}^{i-1} \mu_{ij}(V_{n,j} - u_h^n), \tag{2.15}$$

$$V^{n+1} = u_h^n + \sum_{i,j=1}^{2} b_i a_{ij}^{-1}(V^{n,j} - u_h^n). \tag{2.16}$$

With $\Lambda_i u_h^n$ as in (2.9), introduce the residuals $e^{n,i} \in S_h^r$ by

$$V^{n,i} = \Lambda_i u_h^n + e^{n,i}, \qquad i = 1, 2. \tag{2.17}$$

Inserting now (2.9) and (2.17) into (2.15) and proceeding with similar calculations to those that led to (2.11), it may be seen that the effect of perturbation terms such as the ones introduced here is to cancel precisely the $O(k^3)$ term in the right-hand side of (2.11). The new error equations are

$$\begin{aligned} e^{n,i} - k\beta\left[2Q\left((1 + p_i)\Lambda_i u_h^n - p_i u_h^n, e^{n,i}\right) + p_i Q(e^{n,i}) + F(e^{n,i})\right] \\ - \sum_{j=1}^{i-1} \mu_{ij} e^{n,j} = E^{n,i}, \qquad i = 1, 2 \end{aligned} \tag{2.18}$$

where

$$\max_{n,i} \|E^{n,i}\| \leqslant c(k^4 + kh^r). \tag{2.19}$$

We can now formally state a result about the local error of the modified (2, 3) scheme. First we need a preliminary result.

LEMMA 2.1. *Given w, $v \in S_h^r$ and λ, μ real numbers, let G: $S_h^r \to S_h^r$ be given, for $\phi \in S_h^r$, by*

$$G(\phi) = \phi - w - k\left[\lambda F(\phi) + \mu Q(\phi - v)\right]. \tag{2.20}$$

Then, if $k|\mu|\,\|v_x\|_\infty < 2$, the equation $G(\phi) = 0$ has a solution ϕ that satisfies

$$\|\phi\| \leqslant \left(2 - k|\mu|\,\|v_x\|_\infty\right)^{-1}\left(2\|w\| + k|\mu|\,\|(v^2)_x\|\right). \tag{2.21}$$

Proof. By (1.4) it is seen that, for each $h > 0$, G is a continuous map in $\{S_h^r, \|\cdot\|\}$. Integration by parts and the Cauchy-Schwarz inequality now yield for $\phi \in S_h^r$ that

$$(G(\phi), \phi) \geqslant \|\phi\|\left[\left(1 - \frac{k}{2}|\mu|\,\|v_x\|_\infty\right)\|\phi\| - \left(\|w\| + \frac{k}{2}|\mu|\,\|(v^2)_x\|\right)\right].$$

It follows by our hypothesis that, for $\|\phi\|$ sufficiently large, $(G(\phi), \phi) > 0$. Using a well-known variant of Brouwer's fixed point theorem (Lemma 3.3 in [2]), we conclude that $G(\phi) = 0$ has a solution; (2.21) then follows from the previous estimate. □

The main result of this section is the proposition that follows.

PROPOSITION 2.1. *If k is sufficiently small, then the $V^{n,i}$, V^{n+1}, $e^{n,i}$, defined by* (2.15)–(2.17) *exist and satisfy, for some constant c,*

$$\max_n \left(\|V^{n,1}\| + \|V^{n,2}\| + \|V^{n+1}\|\right) \leqslant c, \tag{2.22}$$

$$\max_{n,i} \|e^{n,i}\| \leqslant ck(k^3 + h^r), \tag{2.23}$$

$$\max_n \left\|V^{n+1} - u_h^{n+1}\right\| \leqslant ck(k^3 + h^r). \tag{2.24}$$

Proof. The existence of $V^{n,i}$ follows immediately by applying Lemma 2.1 to (2.15) and taking into account (1.35). (2.22) then follows from (2.21), (1.35) and (2.16). Taking the L^2-inner product of $e^{n,i}$ with itself in (2.18) yields

$$\|e^{n,i}\|^2 - \sum_{j=1}^{i-1} \mu_{ij}(e^{n,j}, e^{n,i}) + (k\beta/2)\left(\left[\tilde{\Lambda}_i u_h^n\right]_x, (e^{n,i})^2\right) = (E^{n,i}, e^{n,i}),$$

where $\tilde{\Lambda}_i u_h^n \equiv (1 + p_i)\Lambda_i u_h^n - p_i u_h^n$. There follows that

$$\|e^{n,i}\|\left[1 - (k\beta/2)\left\|\left[\tilde{\Lambda} u_h^n\right]_x\right\|_\infty\right] \leqslant \sum_{j=1}^{i-1} |\mu_{ij}|\, \|e^{n,j}\| + \|E^{n,i}\|.$$

For k sufficiently small, taking into account (1.35) and (2.19), use of the above for $i = 1$ and 2 gives (2.23); (2.24) now follows from (2.23) and the triangle inequality applied to (2.13), which, of course, still holds (and (2.14) also) if we replace $\tilde{V}^{n+1}$, $\tilde{\varepsilon}^{n,i}$ in it by V^{n+1}, $e^{n,i}$, respectively. □

We emphasize that it is in the case of special nonlinearities, like the quadratic $F(u)$ in the KdV case, that such perturbations (which cancel terms involving higher derivatives of f) have simple expressions, say, for third- or fourth-order accurate RK methods. Let us also point out that in the *unmodified case* we can obtain immediately from (2.11)–(2.14), in the manner of Proposition 2.1, the suboptimal in time estimate

$$\max_n \left\|\tilde{V}^{n+1} - u_h^{n+1}\right\| \leqslant ck(k^2 + h^r).$$

3. Convergence of the Modified (2, 3) Method. In this section we derive optimal L^2-error estimates for the scheme (1.22)–(1.23) with the initial conditions (1.24)–(1.29). We note first an identity for later reference: given v, w, η, $\theta \in S_h^r$ and β, σ real numbers, let ϕ, χ satisfy (cf. Lemma 2.1) the equations

$$\phi = v + k\beta\left[F(\phi) + \sigma Q(\phi - \eta)\right], \qquad \chi = w + k\beta\left[F(\chi) + \sigma Q(\chi - \theta)\right].$$

Then, if $\phi - \chi = \varepsilon$, $\eta - \theta = \zeta$, we have

$$\varepsilon = v - w + k\beta\left[F(\varepsilon) + \sigma Q(\varepsilon) + \sigma Q(\zeta) + Q(\varepsilon, \delta) + Q(\zeta, \nu)\right], \tag{3.1}$$

where $\delta = 2(1 + \sigma)\chi - 2\sigma(\zeta + \theta)$ and $\nu = 2\sigma(\theta - \chi)$.

THEOREM 3.1. *Let k, h be sufficiently small and suppose that*

$$\text{there exists } \alpha > 0 \text{ such that } k \leqslant \alpha h. \tag{3.2}$$

Let $j_1 = j_2 = 2$ and $j_n = 1$ for $3 \leqslant n \leqslant J$. Then, $U_j^{n,i}$, $U_{j_n}^n$, defined by (1.22)–(1.29) *exist uniquely. Moreover, the following holds*:

$$\max_{0 \leqslant n \leqslant J} \left\|U_{j_n}^n - u^n\right\| \leqslant c(k^3 + h^r). \tag{3.3}$$

Proof. In what follows, some constants, depending at most on u, T and α will play a distinguished role; we reserve for them the symbols c_n^*, $-3 \leqslant n \leqslant J$, $\bar{c}_n$, $0 \leqslant n \leqslant J$, c^* and C, whereas by c we denote as usual "uninteresting" constants. By (1.24), (1.5) and (1.31) it follows that $\|U_{j_0}^0 - u_h^0\| \leqslant c_0 h^r$. Hence, if we choose $c_{-1}^* \geqslant c_0$, any nonnegative constants c_{-2}^*, c_{-3}^* and compute c_0^* by (I.c) below for $i = 0$, then (I.a,b,c) hold for $i = 0$ and any $C \geqslant 0$. Hence, given n, $0 \leqslant n \leqslant J - 1$, we make the following

INDUCTION HYPOTHESIS I (on n).

$$\text{(I)} \quad \begin{cases} \text{(a)} & U_{j_i}^i \text{ exists uniquely}, \qquad 0 \leqslant i \leqslant n, \\ \text{(b)} & \left\| U_{j_i}^i - u_h^i \right\| \leqslant c_i^* (k^3 + h^r), \qquad 0 \leqslant i \leqslant n, \\ \text{(c)} & c_i^* = Ck + (1 + Ck) c_{i-1}^* + Ck(c_{i-2}^* + c_{i-3}^*), \qquad 0 \leqslant i \leqslant n, \end{cases}$$

where C is a positive constant to be chosen later, independent of n, h, k. We shall show that (I) holds for $i = n + 1$ and choose C in the process. Note that (I.c) implies that the c_i^*, $0 \leqslant i \leqslant n$, are bounded above by a positive constant c^* depending only on C and T, i.e., at most on u, T, α, upon eventual choice of C. We subdivide the proof of the inductive step (I) into six parts.

1. *Existence of* $\tilde{U}^{n,i}$, $i = 1, 2$, $\tilde{U}^{n+1}$. Let $\tilde{U}^{n,i}$, $i = 1, 2$, $\tilde{U}^{n+1}$ be defined by

$$\tilde{U}^{n,i} = U_{j_n}^n + k\beta \left[F(\tilde{U}^{n,i}) + p_i Q\left(\tilde{U}^{n,i} - U_{j_n}^n\right) \right] + \sum_{j=1}^{i-1} \mu_{ij} \left(\tilde{U}^{n,j} - U_{j_n}^n\right), \tag{3.4}$$

$$\tilde{U}^{n+1} = U_{j_n}^n + \sum_{i,j=1}^{2} b_i a_{ij}^{-1} \left(\tilde{U}^{n,j} - U_{j_n}^n\right). \tag{3.5}$$

The existence of $\tilde{U}^{n,i}$ then follows from applying Lemma 2.1 for $i = 1$ and then for $i = 2$ to (3.4) provided we require that

$$k\beta \left\| \left(U_{j_n}^n\right)_x \right\|_\infty < 2. \tag{3.6}$$

Note that (1.35), (1.4) and (I.b) imply

$$\left\| \left(U_{j_n}^n\right)_x \right\|_\infty \leqslant \|u_{hx}^n\|_\infty + \left\| \left(u_h^n - U_{j_n}^n\right)_x \right\|_\infty \leqslant c + cc_n^* (k^3 h^{-3/2} + h^{r-3/2}). \tag{3.7}$$

Hence, (3.6) holds if, e.g., we take k sufficiently small so that $ck < 1$ and also, in view of (3.2), require that $cc_n^*(\alpha^4 h^{5/2} + \alpha h^{r-1/2}) < 1$. The latter can always be guaranteed by eventually taking (i.e., when the choice of C is made) h sufficiently small (independently of n), since $c_n^* \leqslant c^*$. Conditions like this or, more generally, of the form

$$cc_n^*(k^\lambda + h^\mu) < 1, \qquad \lambda, \mu > 0, \tag{3.8}$$

will be frequently assumed in sequel and follow (upon eventual choice of C) by taking k, h sufficiently small so that $cc^*(k^\lambda + h^\mu) < 1$. For brevity's sake they will be referred to as "conditions of (3.8) type."

2. *A key stability result.* We next establish the *stability* estimate

$$\|\tilde{U}^{n+1} - V^{n+1}\| \leqslant (1 + ck) \left\| U_{j_n}^n - u_h^n \right\|, \tag{3.9}$$

where V^{n+1} has been defined by (2.16). Recalling (2.15), put $\varepsilon^{n,i} = \tilde{U}^{n,i} - V^{n,i}$, $i = 1, 2$, $\varepsilon^{n+1} = \tilde{U}^{n+1} - V^{n+1}$, $\zeta^n = U^n_{j_n} - u^n_h$. Subtracting (2.15) from (3.4) and using (3.1) yields

$$\begin{aligned}\varepsilon^{n,i} &= \zeta^n + \sum_{j=1}^{i-1} \mu_{ij}\left(\varepsilon^{n,j} - \zeta^n\right) \\ &\quad + k\beta\left[F(\varepsilon^{n,i}) + p_i Q(\varepsilon^{n,i}) + Q(\varepsilon^{n,i}, \delta^{n,i}) + p_i Q(\zeta^n) + Q(\zeta^n, \nu^{n,i})\right], \\ &\qquad\qquad i = 1, 2,\end{aligned} \tag{3.10}$$

where

$$\delta^{n,i} = 2(1 + p_i)V_{n,i} - 2p_i(\zeta^n + u^n_h), \qquad \nu^{n,i} = -2p_i\left(V^{n,i} - u^n_h\right). \tag{3.11}$$

Taking L^2-inner products with $\varepsilon^{n,i}$ in (3.10) and using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left[1 - (k\beta/4)\|\delta^{n,i}_x\|_\infty\right]\|\varepsilon^{n,i}\| &\leqslant \|\zeta^n\| + \sum_{j=1}^{i-1} |\mu_{ij}|\,\|\varepsilon^{n,j} - \zeta^n\| \\ &\quad + k\beta|p_i|\,\left\|\zeta^n\zeta^n_x - \left[\zeta^n\left(V^{n,i} - u^n_h\right)\right]_x\right\|, \qquad i = 1, 2.\end{aligned} \tag{3.12}$$

Now, from (2.9), (2.10), (1.35), (1.4), (2.23) and (3.2) there follows that $\|V^{n,i}_x\|_\infty \leqslant c$. Also, using (1.4), (I.b) and (3.2) yields, under a condition of (3.8) type, that $\|\zeta^n_x\|_\infty < 1$. As a consequence, (3.11) and (1.35) give that $\|\delta^{n,i}_x\|_\infty \leqslant c$. It follows now by (1.4), (1.35), (2.9), (2.10), (2.23) and (3.2) that

$$\begin{aligned}&\left\|\zeta^n\zeta^n_x - \left[\zeta^n\left(V^{n,i} - u^n_h\right)\right]_x\right\| \\ &\quad \leqslant \|\zeta^n\|\left(\|\zeta^n_x\|_\infty + \left\|\left(V^{n,i} - u^n_h\right)_x\right\|_\infty + ch^{-1}\|V^{n,i} - u^n_h\|_\infty\right) \\ &\quad \leqslant \|\zeta^n\|\left[c + ch^{-1}k + ch^{-5/2}(k^4 + kh^r)\right] \leqslant c\|\zeta^n\|.\end{aligned}$$

Therefore, (3.12) gives, for k sufficiently small, for $i = 1, 2$,

$$\|\varepsilon^{n,i}\| \leqslant (1 - ck)^{-1}\left[(1 + ck)\|\zeta^n\| + \sum_{j=1}^{i-1} |\mu_{ij}|\,\|\varepsilon^{n,j} - \zeta^n\|\right], \tag{3.13}$$

from which, for k sufficiently small, there follow

$$\|\varepsilon^{n,1}\| \leqslant (1 + ck)\|\zeta^n\|, \qquad \|\varepsilon^{n,2}\| \leqslant c\|\zeta^n\|. \tag{3.14}$$

Note that the constants c in (3.14) are independent of C.

We now introduce some notation. For $\phi, \chi \in S^r_h$ define, for $i = 1, 2$, $\Phi_i(\phi, \chi) = F(\phi) + p_i Q(\phi - \chi)$ and put

$$v^{n,i} = \Phi_i\left(\tilde{U}^{n,i}, U^n_{j_n}\right), \qquad w^{n,i} = \Phi_i\left(V^{n,i}, u^n_h\right). \tag{3.15}$$

It follows that (2.15) and (2.16) become—in the form of (1.8), (1.9)—

$$V^{n,i} = u^n_h + k\sum_{j=1}^{2} a_{ij}w^{n,j}, \qquad V^{n+1} = u^n_h + k\sum_{i=1}^{2} b_i w^{n,i}. \tag{3.16}$$

Also, by definition, we have

$$\tilde{U}^{n,i} = U^n_{j_n} + k\sum_{j=1}^{2} a_{ij}V^{n,j}, \qquad \tilde{U}^{n+1} = U^n_{j_n} + k\sum_{i=1}^{2} b_i v^{n,i}. \tag{3.17}$$

Subtracting (3.16) from (3.17) gives

$$\varepsilon^{n,i} = \zeta^n + k\sum_{j=1}^{2} a_{ij}(v^{n,j} - w^{n,j}), \qquad \varepsilon^{n+1} = \zeta^n + k\sum_{i=1}^{2} b_i(v^{n,i} - w^{n,i}). \tag{3.18}$$

The next step of the proof uses the *algebraic stability* or *B-stability* properties of the (2, 3) DIRK scheme (1.12); cf., [4], [6]. As in [4, Theorem 2.2] or [6, Theorem 1], taking the L^2-inner product of ε^{n+1} in (3.18) with itself, yields

$$\begin{aligned}\|\varepsilon^{n+1}\|^2 = \|\zeta^n\|^2 &+ 2k\sum_{i=1}^{2} b_i(\varepsilon^{n,i}, v^{n,i} - w^{n,i}) \\ &- k^2\sum_{i,j=1}^{2} m_{ij}(v^{n,i} - w^{n,i}, v^{n,j} - w^{n,j}),\end{aligned} \tag{3.19}$$

where the matrix $m_{ij} = b_i a_{ij} + b_j a_{ji} - b_i b_j$ is nonnegative definite, cf., [4], [6]. We easily conclude that

$$\|\varepsilon^{n+1}\|^2 \leqslant \|\zeta^n\|^2 + 2k\sum_{i=1}^{2} b_i(\varepsilon^{n,i}, v^{n,i} - w^{n,i}). \tag{3.20}$$

To estimate the last term of (3.20), note that (3.10) and (3.18) give

$$k\beta(\varepsilon^{n,i}, v^{n,i} - w^{n,i}) = -(k\beta/4)\big(\delta_x^{n,i}, [\varepsilon^{n,i}]^2\big) - (k\beta p_i/2)\big([\zeta^n]_x^2, \varepsilon^{n,i}\big),$$

from which, with the aid of the same type of estimates that were used in deriving (3.14) from (3.12), we obtain from (3.14) for k sufficiently small, that

$$k(\varepsilon^{n,i}, v^{n,i} - w^{n,i}) \leqslant ck\|\zeta^n\|^2$$

holds. (3.9) follows now from (3.20).

3. *Uniqueness of* $\tilde{U}^{n,i}$, $\tilde{U}^{n+1}$. We may now show that $\tilde{U}^{n,i}$, $i = 1, 2$, are unique. In addition to $\tilde{U}^{n,i}$, let $\tilde{W}^{n,i} \in S_h^r$ satisfy (3.4). Then if $\tilde{Y}^{n,i} = \tilde{W}^{n,i} - \tilde{U}^{n,i}$, (3.1) gives

$$\tilde{Y}^{n,i} = \sum_{j=1}^{i-1} \mu_{ij}\tilde{Y}^{n,j} + k\beta\Big[F(\tilde{Y}^{n,i}) + p_i Q(\tilde{Y}^{n,i}) + 2Q\big(\tilde{Y}^{n,i}, (1+p_i)\tilde{U}^{n,i} - p_i U_{j_n}^n\big)\Big],$$

from which, taking the L^2-inner product of $\tilde{Y}^{n,i}$ with itself, we obtain

$$\Big[1 - (k\beta/2)\Big\|\big[(1+p_i)\tilde{U}^{n,i} - p_i U_{j_n}^n\big]_x\Big\|_\infty\Big]\|\tilde{Y}^{n,i}\| \leqslant \sum_{j=1}^{i-1} |\mu_{ij}|\,\|\tilde{Y}^{n,j}\|.$$

Now, (1.4), (3.14), (I.b), (3.2), (1.35), a condition of (3.8) type and the estimates following (3.12) show that $\|\tilde{U}_x^{n,i}\|_\infty \leqslant c$, $\|(U_{j_n}^n)_x\|_\infty \leqslant c$, which, substituted in the above give, for k sufficiently small, that $\tilde{Y}^{n,i} = 0$, i.e., that $\tilde{U}^{n,i}$ (and $\tilde{U}^{n+1}$ also) are unique.

4. *Accuracy of the initial Newton iterates.* We now prove that the starting values $U_0^{n,i}$ required in (1.22) and defined by (1.24)–(1.29) are close to $\tilde{U}^{n,i}$. First, note that it was proved in [2] that U_*^1 defined by (1.25) exists uniquely in S_h^r and satisfies $\|U_*^1 - u_h^1\| \leqslant c(k^2 + h^r)$. Now, (1.26), (1.24), (2.9) and (2.17) give

$$\begin{aligned}U_0^{0,1} - \tilde{U}^{0,1} = \beta\big(U_*^1 - u_h^1\big) - \varepsilon^{0,1} + \big[(1-\beta)\Pi u^0 + \beta u_h^1 - \big(u_h^0 + \beta k u_{ht}^0\big)\big] \\ - \left(\sum_{j=2}^{3} k^j \tau_{1j} D_t^j u_h^0 + e^{0,1}\right).\end{aligned}$$

Hence, it may be easily checked, using Taylor's theorem, (1.5), (1.31), (1.35), (3.14), (I.b), (2.23) and the triangle inequality that $\|U_0^{0,1} - \tilde{U}^{0,1}\| = O(k^2 + h^4)$. Taking into account (1.27) for $n = 0$, we may also prove the same estimate for the second stage. We can write, in fact,

$$\|U_0^{0,i} - \tilde{U}^{0,i}\| \leqslant c(1 + c_0^*)(k^2 + h^r), \qquad i = 1, 2. \tag{3.21}$$

If $n = 1$, a similar analysis at $t = t^1$ and (1.26), (1.27) for $n = 1$ give

$$\|U_0^{1,i} - \tilde{U}^{1,i}\| \leqslant c(1 + c_1^*)(k^2 + h^r), \qquad i = 1, 2. \tag{3.22}$$

If now $n \geqslant 2$, it may be seen that (1.28), (2.9), (2.17) and Taylor's theorem imply, in view of (I.b), (3.14), (1.29), (1.35) and (2.23), that

$$\|U_0^{n,i} - \tilde{U}^{n,i}\| \leqslant \bar{c}_n(k^3 + h^r), \qquad i = 1, 2, \tag{3.23}$$

where $\bar{c}_n = c(1 + c_n^* + c_{n-1}^* + c_{n-2}^*)$. We summarize now (3.21)–(3.23) by

$$\|U_0^{n,i} - \tilde{U}^{n,i}\| \leqslant \bar{c}_n(k^{2\theta(n)} + h^r), \qquad i = 1, 2, \tag{3.24}$$

where $\theta(n) \equiv 1$ if $n = 0$ or 1 and $\theta(n) \equiv 3/2$ if $n \geqslant 2$, and where

$$\bar{c}_n = c(1 + c_n^* + c_{n-1}^* + c_{n-2}^*). \tag{3.25}$$

Note that the constants c in this subsection (and in particular the constant c in (3.25)) are independent of C.

5. *Convergence of the Newton iterates $U_j^{n,i}$ to $\tilde{U}^{n,i}$.* Next, we prove that the $U_j^{n,i}$, $1 \leqslant j \leqslant j_{n+1}$, given by (1.22) exist uniquely and approach "quadratically" $\tilde{U}^{n,i}$ as j increases. We achieve this inductively by an "internal" to (I) loop (II) (on the index j). First note from (3.24) that $\|U_0^{n,i} - \tilde{U}^{n,i}\| < \bar{c}_n(k^{\theta(n)} + h^{r/2})^2$. Hence, we make for some $0 \leqslant j \leqslant j_{n+1} - 1$ the

INDUCTION HYPOTHESIS II (on j).

$$\text{(II)} \quad \begin{cases} \text{(a)} & U_m^{n,i} \text{ exists uniquely for } i = 1, 2,\ 0 \leqslant m \leqslant j, \\ \text{(b)} & \|U_m^{n,i} - \tilde{U}^{n,i}\| \leqslant \bar{c}_n(k^{\theta(n)} + h^{r/2})^{2^{m+1}}, \qquad i = 1, 2,\ 0 \leqslant m \leqslant j. \end{cases}$$

We shall show that (II) holds for $m = j + 1$. Note that $U_{j+1}^{n,i}$ satisfies an equation of the form $LU_{j+1}^{n,i} = W$, for some $W = W(n, i, j) \in S_h^r$, where the linear map $L = L(i, j, n)$: $S_h^r \to S_h^r$ is defined, for $\phi \in S_h^r$, as

$$L\phi = \phi - k\beta F\big(\phi, U_j^{n,i}\big) - k\beta Q\big(\phi, (1 + 2p_i)U_j^{n,i} - 2p_i U_{j_n}^n\big).$$

Hence

$$(L\phi, \phi) \geqslant \|\phi\|^2\Big\{1 - (k\beta/2)\Big\|\big[(1 + p_i)U_j^{n,i} - p_i U_{j_n}^n\big]_x\Big\|_\infty\Big\}.$$

Using now (II.b), (3.25) and similar estimates to the ones in part 3 above, we obtain, under conditions of type (3.8), that $\|(U_j^{n,i})_x\|_\infty \leqslant c$, $\|(U_{j_n}^n)_x\|_\infty \leqslant c$. It follows, for k sufficiently small, that L is positive-definite and that $U_{j+1}^{n,i}$ exist uniquely, i.e., that (II.a) is true for $m = j + 1$. To prove (II.b) for $m = j + 1$, subtracting (3.4) from (1.22), taking L^2-inner products with $U_{j+1}^{n,i} - \tilde{U}^{n,i}$ and using (1.4), we obtain, for $i = 1, 2$,

$$\begin{aligned} &\|U_{j+1}^{n,i} - \tilde{U}^{n,i}\|\Big\{1 - (k\beta/2)\Big\|\big[(1 + p_i)U_j^{n,i} - p_i U_{j_n}^n\big]_x\Big\|_\infty\Big\} \\ &\qquad \leqslant ckh^{-3/2}\|U_j^{n,i} - \tilde{U}^{n,i}\|^2 + \sum_{j=1}^{i-1} |\mu_{ij}|\,\|U_{j_{n+1}}^{n,j} - \tilde{U}^{n,j}\|. \end{aligned} \tag{3.26}$$

If $i = 1$, (3.26) and similar considerations to the ones already used yield, for k sufficiently small, that $\|U_{j+1}^{n,i} - \tilde{U}^{n,1}\| \leqslant ckh^{-3/2}\|U_j^{n,1} - U^{n,1}\|^2$. Hence, by (II.b)

$$\left\|U_{j+1}^{n,1} - \tilde{U}^{n,1}\right\| \leqslant \lambda \bar{c}_n \left(k^{\theta(n)} + h^{r/2}\right)^{2^{j+1}+1}, \tag{3.27}$$

where $\lambda = ck^{-3/2}\bar{c}_n(k^{\theta(n)} + h^{r/2}) \leqslant c\bar{c}_n(k^2h^{-3/2} + kh^{(r-3)/2})$. Hence, (3.2), (3.25) and a condition of type (3.8) allow making $\lambda < 1$. (3.27) implies then that (II.b) holds for $m = j + 1$ and $i = 1$. For $i = 2$, (3.26) gives

$$\sigma\left\|U_{j+1}^{n,2} - \tilde{U}^{n,2}\right\| \leqslant ckh^{-3/2}\left\|U_j^{n,2} - \tilde{U}^{n,2}\right\|^2 + |\mu_{21}|\left\|U_{j_{n+1}}^{n,1} - \tilde{U}^{n,1}\right\|,$$

where $\sigma = 1 - (k\beta/2)\|[(1 + p_2)U_j^{n,2} - p_2U_{j_n}^n]_x\|_\infty$. Hence, (II.b) and (3.27) ($\lambda < 1$), give, since $j + 1 \leqslant j_{n+1}$,

$$\begin{aligned}\left\|U_{j+1}^{n,2} - \tilde{U}^{n,2}\right\| &\leqslant \sigma^{-1}ckh^{-3/2}\left\|U_j^{n,2} - \tilde{U}^{n,2}\right\|^2 + \sigma^{-1}|\mu_{21}|\left\|U_{j_{n+1}}^{n,1} - \tilde{U}^{n,1}\right\| \\ &\leqslant \lambda^*\bar{c}_n\left(k^{\theta(n)} + h^{r/2}\right)^{2^{j+1}+1},\end{aligned} \tag{3.28}$$

where $\lambda^* = \sigma^{-1}[|\mu_{21}| + ckh^{-3/2}\bar{c}_n(k^{\theta(n)} + h^{r/2})]$. Since $|\mu_{21}| = (2\beta - 1)/\beta < 1$ and σ may take any value in $(0, 1)$ if k is sufficiently small, we may assume that $|\mu_{21}| < \sigma < 1$. Hence, a condition of type (3.8) yields that $\lambda^* < 1$ and (II.b), $i = 2$, $m = j + 1$, follows from (3.28); the inductive step II is now complete.

6. *Completion of inductive step* I. There only remains to prove that (I) holds for $i = n + 1$ and choose C. With $U_{j_{n+1}}^{n+1}$ given by (1.23) we have

$$\left\|U_{j_{n+1}}^{n+1} - u_h^{n+1}\right\| \leqslant \left\|U_{j_{n+1}}^{n+1} - \tilde{U}^{n+1}\right\| + \left\|\tilde{U}^{n+1} - V^{n+1}\right\| + \left\|V^{n+1} - u_h^{n+1}\right\|.$$

Hence, (1.23), (3.5), (3.9), (I.b) and (2.24) give

$$\begin{aligned}\left\|U_{j_{n+1}}^{n+1} - u_h^{n+1}\right\| &\leqslant c\sum_{i=1}^{2}\left\|U_{j_{n+1}}^{n,i} - \tilde{U}^{n,i}\right\| + (1 + ck)c_n^*(k^3 + h^r) \\ &\quad + ck(k^3 + h^r),\end{aligned} \tag{3.29}$$

where the constant c is independent of C. We now distinguish two cases. First, suppose that $h^{r/2} \leqslant k^{\theta(n)}$. Then, from (3.29), (II.b) and (3.25), it may be seen that

$$\left\|U_{j_{n+1}}^{n+1} - u_h^{n+1}\right\| \leqslant \left[ck + (1 + ck)c_{n-1}^* + ckc_{n-2}^*\right](k^3 + h^r),$$

where the constant c does not depend on C. Now choose C to be equal to this c and define c_{n+1}^* by $c_{n+1}^* = Ck + (1 + Ck)c_n^* + Ck(c_{n-1}^* + c_{n-2}^*)$. It follows that (I.b) and (I.c) hold for $i = n + 1$. Now if it is the case that $k^{\theta(n)} < h^{r/2}$, it is not hard to see from (3.26), (II.b), the fact that $r \geqslant 4$ and conditions of type (3.8) that $\|U_{j+1}^{n,1} - \tilde{U}^{n,1}\| \leqslant k\bar{c}_n(k^{\theta(n)} + h^{r/2})^{2^{j+1}+1}$. Arguing now as in the derivation of (3.28), we can infer the estimate $\|U_{j+1}^{n,2} - \tilde{U}^{n,2}\| \leqslant k\bar{c}_n(k^{\theta(n)} + h^{r/2})^{2^{j+1}+1}$. As a consequence, (3.29) gives the estimate

$$\left\|U_{j_{n+1}}^{n+1} - u_h^{n+1}\right\| \leqslant \left[ck\bar{c}_n + (1 + ck)c_n^* + ck\right](k^3 + h^r).$$

Hence, the choice of C and the completion of the inductive step (I) proceeds analogously.

Theorem 3.1 can now be proved: we have just argued inductively that

$$\max_n \left\|U_{j_n}^n - u_h^n\right\| \leqslant c^*(k^3 + h^r),$$

which, in conjunction with (1.31), yields (3.3). □

4. A Modified (3, 4) Method. In this section, we present a modified version of the fully discrete Galerkin method corresponding to the (3, 4) DIRK scheme (1.13). We seek U^n, $0 \leqslant n \leqslant J$, $U^{n,i}$, $1 \leqslant i \leqslant 3$, $0 \leqslant n \leqslant J-1$, in S_h^r that satisfy

$$\begin{aligned} U^{n,i} = U^n + k\beta\big[F(U^{n,i}) + q_{1i}Q(U^{n,i} - U^n) \\ + q_{2i}Q\big(U^{n,i} - U^n, U^{n,i} - \gamma_{1i}U^n - \gamma_{2i}U^{n-1}\big)\big] \\ + \sum_{j=1}^{i-1} \mu_{ij}(U^{n,j} - U^n), \end{aligned} \tag{4.1}$$

$$U^{n+1} = U^n + \sum_{i,j=1}^{3} b_i a_{ij}^{-1}(U^{n,j} - U^n). \tag{4.2}$$

Here, the perturbation terms are chosen so that the local error of the method is of $O(k^5 + kh^r)$. The constants q_{1i}, q_{2i}, γ_{1i}, γ_{2i} are given by the 3-vectors

$$(q_{1i}) = \big[1, -8\beta^2 + 8\beta - 1, \beta^2/(1-\beta)^2\big]^T,$$

$$(q_{2i}) = \big[4\beta/(2\beta+1), 2\beta^2(8\beta^2 - 4\beta - 1)/(4\beta^2 - 4\beta - 1),$$
$$(-92\beta^4 + 172\beta^3 - 106\beta^2 + 26\beta - 2)/(1-\beta)^2(2\beta^2 - 3\beta + 2)\big]^T,$$

$$(\gamma_{1i}) = [\beta + 1, 3/2, -\beta + 2]^T, \qquad (\gamma_{2i}) = [-\beta, -1/2, \beta - 1]^T,$$

with β as in (1.13). The computation of $U^{n,i}$ now requires the values U^n, U^{n-1} of the two previous steps; hence, two initial values U^0, U^1 must now be provided. Again, we solve the three nonlinear systems represented by (4.1) by Newton's method. With notation already introduced in the context of the (2, 3) scheme, this requires, given $U_{j_n}^n$, $U_{j_{n-1}}^{n-1}$ and starting values $U_0^{n,i}$, solving linear systems to find the iterates $U_j^{n,i}$, $0 < j \leqslant j_{n+1}$; $U_{j_{n+1}}^{n+1}$ is then computed by the analog of (1.23). The required initial values that we use are as follows: we take $j_0 = 0$ and $U_{j_0}^0 = U_0^0 = \Pi u^0$ as before. As $U_{j_1}^1$ we use the one obtained by the modified (2, 3) method. The starting values $U_0^{n,i}$ are given for $n = 1, 2$ by the equations

$$U_0^{1,i} = -\tau_{i1}U_{j_0}^0 + (1 + \tau_{i1})U_{j_1}^1, \qquad U_0^{2,i} = -(1 + \tau_{i1})U_{j_0}^0 + (2 + \tau_{i1})U_{j_1}^1.$$

For $n \geqslant 3$, we define $U_0^{n,i} = \Sigma_{l=0}^3 \lambda_{il} U_{j_{n-l}}^{n-l}$, where the λ_{ij} are solutions of the linear systems $\Sigma_{j=0}^3 \lambda_{ij}(-j)^m = m!\tau_{im}$, $1 \leqslant i \leqslant 3$, $0 \leqslant m \leqslant 3$ ($0^0 = 1$). Here, the constants τ_{ij}, $1 \leqslant i \leqslant 3$, $0 \leqslant j \leqslant 3$, are defined by the analog of (2.3) in the (3, 4) case. We can prove the following result:

THEOREM 4.1. *Let k, h be sufficiently small and satisfy $k \leqslant \alpha h$ for some $\alpha > 0$. Let $j_2 = j_3 = 2$ and $j_n = 1$ for $4 \leqslant n \leqslant J$. Then, the $U_j^{n,i}$, $U_{j_n}^n$ that are computed as outlined above by the modified* (3, 4) *scheme coupled with Newton's method, exist uniquely; the following estimate holds:*

$$\max_{0 \leqslant n \leqslant J} \big\| U_{j_n}^n - u^n \big\| \leqslant c(k^4 + h^r). \quad \square$$

Acknowledgment. The authors are grateful to the referee for his patience in reading a long and technical manuscript, and for his comments that led to a marked improvement in the structure of the paper.

Department of Mathematics
University of Tennessee
Knoxville, Tennessee 37916 and

Department of Mathematics
University of Crete
Iraklion, Crete, Greece

Department of Mathematics
University of Tennessee
Knoxville, Tennessee 37916

1. R. ALEXANDER, "Diagonally implicit Runge-Kutta methods for stiff O.D.E.'s," *SIAM J. Numer. Anal.*, v. 14, 1976, pp. 1006–1021.

2. G. A. BAKER, V. A. DOUGALIS & O. A. KARAKASHIAN, "Convergence of Galerkin approximations for the Korteweg-de Vries equation," *Math. Comp.*, v. 40, 1983, pp. 419–433.

→ J. L. BONA & R. SMITH, "The initial-value problem for the Korteweg-de Vries equation," *Philos. Trans. Roy. Soc. London Ser. A*, v. 278, 1975, pp. 555–604.

4. K. BURRAGE & J. C. BUTCHER, "Stability criteria for implicit Runge-Kutta methods," *SIAM J. Numer. Anal.*, v. 16, 1979, pp. 46–57.

5. M. CROUZEIX, *Sur l'Approximation des Équations Différentielles Opérationnelles Linéaires par des Méthodes de Runge-Kutta*, Thèse, Université Paris VI, 1975.

→ M. CROUZEIX, "Sur la *B*-stabilité des méthodes de Runge-Kutta," *Numer. Math.*, v. 32, 1979, pp. 75–82.

→ V. THOMÉE & B. WENDROFF," Convergence estimates for Galerkin methods for variable coefficient initial value problems," *SIAM J. Numer. Anal.*, v. 11, 1974, pp. 1059–1068.

8. L. B. WAHLBIN, "A dissipative Galerkin method for the numerical solution of first order hyperbolic equations," in *Mathematical Aspects of Finite Elements in Partial Differential Equations* (C. de Boor, ed.), Academic Press, New York, 1974, pp. 147–169.

9. R. WINTHER, "A conservative finite element method for the Korteweg-de Vries equation," *Math. Comp.*, v. 34, 1980, pp. 23–43.