JOURNAL OF COMPUTATIONAL PHYSICS 39, 94–102 (1981)

Petrov–Galerkin Methods for Nonlinear Dispersive Waves

J. M. SANZ-SERNA

Departamento de Matemáticas, Facultad de Ciencias, Valladolid, Spain

AND

I. CHRISTIE

Department of Mathematics and Statistics,
University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Received October 24, 1979

Petrov–Galerkin methods based on piecewise linear interpolants for the Korteweg–de Vries and related equations are studied. Both accuracy and stability are analysed for the linearised case. It is shown that in the nonlinear case the order of accuracy of the standard Galerkin procedure is reduced and an alternative technique is therefore proposed which retains the fourth-order accuracy in space. This is found to perform well when compared with finite difference or other finite element schemes.

1. INTRODUCTION

The Korteweg–de Vries (KdV) equation

$$u_t + uu_x + \varepsilon u_{xxx} = 0, \qquad \varepsilon > 0, \tag{1.1}$$

and several of its generalisations play a major role in the study of nonlinear dispersive waves. Examples range from water waves and lattice waves to plasma waves.

The numerical solution of (1.1) and related equations has been the subject of many papers over the last few years. Greig and Morris [4] propose a Hopscotch method and compare it with the original Zabusky–Kruskal leap frog scheme [10]. Alexander and Morris [1] study Galerkin methods based on those given by Wahlbin [9], which include the possibility of an additional dissipative term. The Korteweg–de Vries–Burgers (KdVB) equation

$$u_t + uu_x - \nu u_{xx} + \varepsilon u_{xxx} = 0, \qquad \nu, \varepsilon > 0, \tag{1.2}$$

is solved by Canosa and Gazdag [3] using finite Fourier transform techniques.

The present work concerns itself with Petrov–Galerkin finite element methods in

0021-9991/81/010094–09$02.00/0

which different trial and test functions are used (see Anderssen and Mitchell [2] for a discussion). To keep the computational effort as low as possible the interpolant is chosen to be piecewise linear; the corresponding test functions are taken to be piecewise cubic with C^1 continuity.

Following a description of the numerical procedure in Section 2, we analyse the resulting equations in the linearised case in Section 3. Section 4 shows that for the nonlinear case the accuracy of the Petrov–Galerkin procedure is reduced and to maintain a method fourth-order accurate in space an alternative technique is proposed. In the final section we present some numerical experiments which appear to show a marked improvement over other suggested methods.

Although the methods in this paper are only presented for the KdV equation the ideas can be applied also to similar equations which involve nonlinearities and dispersion.

2. Description of Method

We attempt to solve Eq. (1.1) together with the initial condition

$$u(x, 0) = f(x), \qquad -\infty < x < \infty. \tag{2.1}$$

We assume that the problem has a unique solution such that, for fixed t, $u(x, t)$, together with all its x derivatives, tends to zero as $|x| \to \infty$. Conditions on f guaranteeing existence and uniqueness are given by Lax [5] and Sjöberg [8].

Multiplying (1.1) by $v(x)$, a twice differentiable function, and integrating by parts we obtain

$$(u_t, v) + (uu_x, v) + \varepsilon(u_x, v_{xx}) = 0, \tag{2.2}$$

where $(\ ,\ )$ denotes the usual L_2 inner product

$$(f, g) = \int_{-\infty}^{\infty} f(x)\, g(x)\, dx.$$

We introduce finite elements in space in (2.2) and approximate the exact solution by

$$U(x, t) = \sum_{i=0}^{n} U_i(t)\, \phi_i(x), \tag{2.3}$$

where the trial functions $\phi_i(x)$, $i = 0, 1, \ldots, n$, have compact support. The unknown functions $U_i(t)$, $i = 0, 1, \ldots, n$, are determined from the system of ordinary differential equations

$$(U_t, \psi_j) + (UU_x, \psi_j) + \varepsilon(U_x, (\psi_j)_{xx}) = 0, \tag{2.4}$$

$j = 0, 1,..., n$, subject to appropriate initial conditions which can be obtained from (2.1). The essential feature of the Petrov–Galerkin method lies in the fact that the test functions ψ_j, $j = 0, 1,..., n$, need not be the same as the trial functions.

It should be noted that the approximant (2.3) has compact support. This is no drawback since the type of problem under consideration exhibits exponential decay as $|x| \to \infty$. Other upstream/downstream conditions can be accommodated by a suitable modification of the approximant.

We introduce a grid $x_0 < x_1 < \cdots < x_n$ with uniform spacing h and define $\phi_i(x)$ to be the usual piecewise linear hat function associated with the node x_i, $i = 0, 1,..., n$ (i.e., $\phi_i(x_j) = \delta_{ij}$, the Kronecker delta). With this choice of trial functions $U(x_i, t) = U_i(t)$.

The integration by parts which had to be performed to arrive at formula (2.4) provides us with the possibility of introducing a C^0 interpolant, resulting in a much lower computational effort than that found when using splines or Hermite cubics, which would be necessary if identical trial and test functions were used. This approach is analogous to the H^{-1}–Galerkin method of Rachford and Wheeler [7].

The support of the test functions is chosen to be an interval of length $4h$, in order to give a five-point replacement for u_{xxx}. It is clear from (2.4) that C^1 continuity is necessary for the test functions. With these requirements in mind we define

$$\psi_i(x) = \psi((x - x_0)/h - i), \qquad i = 0, 1,..., n, \tag{2.5}$$

where $\psi(x)$ is a C^1 function with support $[-2, 2]$ and which reduces to a cubic polynomial in each of the intervals $[i, i + 1]$, $i = -2, -1, 0, 1$. Clearly there is a six parameter family of such piecewise cubics and an individual member of the family can be specified by the values $\alpha_i = \psi(i)$, $\beta_i = \psi'(i)$, $i = -1, 0, 1$. We can write ψ in terms of the basis functions for the Hermite cubic interpolation as follows. Define

$$\sigma(x) = \begin{cases} (|x| - 1)^2 (2|x| + 1) & \text{if } |x| \leqslant 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\rho(x) = \begin{cases} x(|x| - 1)^2 & \text{if } |x| \leqslant 1 \\ 0 & \text{otherwise,} \end{cases}$$

so that $\sigma(0) = 1$, $\sigma(-1) = \sigma(1) = 0$, $\sigma'(-1) = \sigma'(0) = \sigma'(1) = 0$, and $\rho'(0) = 1$, $\rho'(-1) = \rho'(1) = 0$, $\rho(-1) = \rho(0) = \rho(1) = 0$. Now

$$\begin{aligned} \psi(x) = {} & \alpha_{-1}\sigma(x + 1) + \alpha_0\sigma(x) + \alpha_1\sigma(x - 1) + \beta_{-1}\rho(x + 1) \\ & + \beta_0\rho(x) + \beta_1\rho(x - 1). \end{aligned} \tag{2.6}$$

With ψ_j given by (2.5) and (2.6), the system (2.4) becomes, after tedious evaluations of the inner products:

$$
\begin{aligned}
&(1/60)[(9\alpha_1 + 2\beta_1)\,\dot{U}_{i+2} + (9\alpha_0 + 42\alpha_1 + 2\beta_0)\,\dot{U}_{i+1} \\
&\quad + (42\alpha_0 + 9\alpha_1 + 9\alpha_{-1} - 2\beta_1 + 2\beta_{-1})\,\dot{U}_i \\
&\quad + (9\alpha_0 + 42\alpha_{-1} - 2\beta_0)\,\dot{U}_{i-1} + (9\alpha_{-1} - 2\beta_{-1})\,\dot{U}_{i-2}] \\
&\quad + (1/60h)[(9\alpha_1 + 2\beta_1)\,U_{i+2}^2 + (12\alpha_1 + \beta_1)\,U_{i+1}\,U_{i+2} \\
&\quad + (9\alpha_0 + 2\beta_0 - 6\beta_1)\,U_{i+1}^2 + (12\alpha_0 - 12\alpha_1 + \beta_0 + \beta_1)\,U_i U_{i+1} \\
&\quad + (9\alpha_{-1} - 9\alpha_1 - 6\beta_0 + 2\beta_{-1} + 2\beta_1)\,U_i^2 \\
&\quad - (12\alpha_0 - 12\alpha_{-1} - \beta_0 - \beta_{-1})\,U_{i-1}U_i - (9\alpha_0 - 2\beta_0 + 6\beta_{-1})\,U_{i-1}^2 \\
&\quad - (12\alpha_{-1} - \beta_{-1})\,U_{i-2}U_{i-1} - (9\alpha_{-1} - 2\beta_{-1})\,U_{i-2}^2] \\
&\quad + (\varepsilon/h^3)[-\beta_1 U_{i+2} + (2\beta_1 - \beta_0)\,U_{i+1} + (2\beta_0 - \beta_1 - \beta_{-1})\,U_i \\
&\quad + (2\beta_{-1} - \beta_0)\,U_{i-1} - \beta_{-1}U_{i-2}] = 0, \qquad (2.7)
\end{aligned}
$$

where $i = 0, 1, \ldots, n$, and we set $U_{-2} = U_{-1} = U_{n+1} = U_{n+2} = 0$. Since multiplication of $\psi(x)$ by a constant does not alter Eqs. (2.7), only five essential parameters among α_i, β_i, $i = -1, 0, 1$, remain. Furthermore, application of Taylor expansions to (2.7) show that, to give approximations consistent with (1.1), additional conditions are required. In this way we arrive at the following set of constraints:

$$
\begin{aligned}
\alpha_{-1} + \alpha_0 + \alpha_1 &= 1, \\
\beta_{-1} + \beta_0 + \beta_1 &= 0, \qquad (2.8) \\
\beta_{-1} - \beta_1 &= 1.
\end{aligned}
$$

The remaining three parameters can be used to vary the degree of asymmetry in the test functions. This enables us to introduce upwinding into our numerical schemes, which, in a convection dominated situation can be very useful (see for example Mitchell and Christie [6]). However to study travelling waves, emphasis should be placed on conservation properties and it will be clear from the following analysis that this can be achieved by using symmetric test functions. We therefore obtain the additional constraints:

$$
\begin{aligned}
\alpha_{-1} &= \alpha_1, \\
\beta_{-1} &= -\beta_1, \qquad (2.9) \\
\beta_0 &= 0.
\end{aligned}
$$

Hence, from (2.8) and (2.9) $\beta_{-1} = \frac{1}{2}$, $\beta_1 = -\frac{1}{2}$, which leaves α_1 (say) as the only free parameter. When $\alpha_1 = \frac{1}{6}$, $\psi(x)$ reduces to the well-known Schoenberg cubic spline and continuity is increased to C^2. In what follows we shall assume that conditions (2.8) and (2.9) apply and consider a family of test functions $\psi(x)$ depending on the single parameter $\alpha = \alpha_1$.

3. Linearised Equation

In this section we look at the linearised KdV equation

$$L(u) \equiv u_t + \mu u_x + \varepsilon u_{xxx} = 0, \tag{3.1}$$

where μ is a constant. With the test and trial functions defined previously, the linear analogue of Eq. (2.4) is

$$M\dot{\mathbf{U}} + S\mathbf{U} = \mathbf{0}, \tag{3.2}$$

where $\mathbf{U} = [U_0(t), U_1(t),..., U_n(t)]^T$, the dot denotes differentiation with respect to t, and M and S are five-band $(n+1) \times (n+1)$ matrices. M is usually called a mass matrix. The system (3.2) is given explicitly by

$$\begin{aligned} L_h(U_i) \equiv\ & (1/60)[(9\alpha - 1)\,\dot{U}_{i+2} + (9 + 24\alpha)\,\dot{U}_{i+1} + (44 - 66\alpha)\,\dot{U}_i \\ & + (9 + 24\alpha)\,\dot{U}_{i+1} + (9\alpha - 1)\,\dot{U}_{i-2}] \\ & + (\mu/24h)[(12\alpha - 1)\,U_{i+2} + (14 - 24\alpha)\,U_{i+1} \\ & - (14 - 24\alpha)\,U_{i-1} - (12\alpha - 1)\,U_{i-2}] \\ & + (\varepsilon/2h^3)[U_{i+2} - 2U_{i+1} + 2U_{i-1} - U_{i-2}] = 0, \end{aligned} \tag{3.3}$$

where $i = 0, 1,..., n$ and we set $U_{-2} = U_{-1} = U_{n+1} = U_{n+2} = 0$.

For general α, Taylor expansion gives

$$L_h(u) = L(u) + O(h^2), \tag{3.4}$$

which is second-order accurate. However when $\alpha = 1/6$ (Schoenberg spline case) the truncation error is

$$L_h(u) - L(u) = (h^2/4)(L(u))_{xx} + O(h^4) \tag{3.5}$$

which, upon using (3.1), is seen to be fourth-order accurate.

Next we consider the conservation properties of the scheme. Multiplication of (3.2) by $\mathbf{U}^T$ yields

$$\mathbf{U}^T M\dot{\mathbf{U}} + \mathbf{U}^T S\mathbf{U} = 0. \tag{3.6}$$

The skew symmetry of S implies that $\mathbf{U}^T S\mathbf{U} = 0$ and, due to the symmetry of M, (3.6) becomes

$$\frac{d}{dt}(\mathbf{U}^T M\mathbf{U}) = 0. \tag{3.7}$$

Furthermore, the matrix M can be shown to be positive definite and diagonally dominant provided $1/9 < \alpha < 7/33$, in which case $(\mathbf{U}^T M\mathbf{U})^{1/2}$ is a norm for $\mathbf{U}$.

If the system of ordinary differential equations (3.2) is solved by means of the

trapezoidal rule (Crank–Nicholson), application of the familiar von Neumann stability test reveals that the amplification factors have modulus unity, thus guaranteeing unconditional stability and conservation of the amplitude of the Fourier modes.

Note that (3.2) is an *implicit* system of ODE's and hence any method for the discretization in time will result in a implicit scheme, unless the mass matrix M is lumped.

4. Nonlinear Case

Returning to the KdV equation (1.1), the system (2.4) can be written as

$$M\dot{\mathbf{U}} + \mathbf{S}(\mathbf{U}) = \mathbf{0}, \tag{4.1}$$

where M and $\mathbf{U}$ are as defined in the linear case and $\mathbf{S}(\mathbf{U})$ is a nonlinear vector function. The ith component of Eq. (4.1) is of the form (3.3) except for the fact that the term involving μ is replaced by

$$\begin{aligned}(1/120h)[&(18\alpha - 2)\, U_{i+2}^2 + (24\alpha - 1)\, U_{i+2}U_{i+1} + (24 - 36\alpha)\, U_{i+1}^2 \\ &+ (23 - 72\alpha)\, U_{i+1}U_i - (23 - 72\alpha)\, U_iU_{i-1} - (24 - 36\alpha)\, U_{i-1}^2 \\ &- (24\alpha - 1)\, U_{i-1}U_{i-2} - (18\alpha - 2)\, U_{i-2}^2].\end{aligned} \tag{4.2}$$

Taylor expansions again reveal that, for any α, the method is second-order accurate. However for $\alpha = 1/6$ the $O(h^2)$ term can no longer be cancelled and the higher order of the linearised problem cannot be attained. A means of recovering fourth-order accuracy in space is now described. We rewrite (1.1) in the form

$$u_t + (u^2/2)_x + \varepsilon u_{xxx} = 0 \tag{4.3}$$

and note from (3.3) that, in the linear case with $\alpha = 1/6$, $u_x(x_i)$ is replaced by

$$(1/24h)[U_{i+2} + 10U_{i+1} - 10U_{i-1} - U_{i-2}]. \tag{4.4}$$

By analogy we approximate $(u^2/2)_x|_{x_i}$ by

$$(1/48h)[U_{i+2}^2 + 10U_{i+1}^2 - 10U_{i-1}^2 - U_{i-2}^2]. \tag{4.5}$$

Of course a subsequent Taylor expansion reveals that the truncation error of the resulting method is now $O(h^4)$ as in the linear case. (In fact, all one has to do is to replace u by u^2 in the expansion corresponding to the linear case.)

Expression (4.5) can be generated by the described Petrov–Galerkin process provided the approximation

$$U^2(x, t) = \sum_{i=0}^{n} U_i^2(t)\, \phi_i(x) \tag{4.6}$$

is used in the term $(\frac{1}{2}u^2)_x$.

This technique is clearly capable of handling any nonlinear term of the form $(F(u))_x$.

5. Numerical Results

In order to compare the method outlined in this text with some others available, we consider the initial value problem given by (1.1) and the initial condition

$$f(x) = 3c \operatorname{sech}^2(kx + d) \tag{5.1}$$

with $c = 0.3$, $\varepsilon = 0.000484$, $k = (c/4\varepsilon)^{1/2}$, and $d = -k$. This problem was used in [1, 4] and has the theoretical solution

$$u(x, t) = 3c \operatorname{sech}^2(kx - kct + d) \tag{5.2}$$

representing a single soliton with amplitude 0.9.

We further take $x_0 = 0$, and $x_n = 2$ since, outside this region, the solution is negligibly small over the range of time used ($0 \leqslant t \leqslant 1$). Equation (4.1) was solved with $\alpha = 1/6$, first without modification (Petrov–Galerkin method) and then with (4.2) replaced by (4.5) to obtain the fourth-order accurate method. The trapezoidal rule with step length τ was employed to advance the solution in time and the resulting system of nonlinear equations solved by Newton iteration. The required initial condition $\mathbf{U}(0)$ was obtained interpolating $f(x)$ at x_i, $i = 0, 1, \ldots, n$. Numerical results are displayed in Table I, where those given by [4] are also included for comparison.

First of all we note that as h gets smaller, the error in the modified scheme decays considerably quicker than in the standard Petrov–Galerkin case, demonstrating its higher order of accuracy in space.

At $h = 0.01$ the errors associated with the modified scheme are negligible compared with both finite difference methods, even though the time step is greater by a factor of 10. Indeed, comparable accuracy is obtained between finite differences with $h = 0.01$, $\tau = 0.0005$ and the modified scheme with $h = 0.033$, $\tau = 0.01$. This balances the fact that one finite difference step is cheaper to execute than one finite element step, since the latter requires Newton iteration.

Alexander and Morris [1] use the Galerkin method with cubic splines as trial and test functions and include the possibility of an additional dissipation term. With $h = 0.05$, $t = 0.39$, and exact time integration, they report a maximum error ranging between 0.025 and 0.059, according to the chosen value of the dissipation parameter; for $h = 0.033$ and $t = 0.46$ the error presented is of the order 0.015. We observe from Table I that the modified scheme improves on these errors. It should be pointed out that the band widths in the matrices are five for the present scheme and seven for Galerkin methods with splines. However, the large computation times quoted in [1] are largely attributable to the use of a computer package for the time integration.

Our schemes were also tested satisfactorily in cases of soliton interaction. No spurious oscillations were found to appear.

Finally, the KdVB equation (1.2) was solved with the initial condition (5.1). When $\varepsilon = 0.000484$ and $\nu = 0.01$ the solution behaves like a travelling wave in which the amplitude is damped. Dissipation dominates over convection when ν is increased to 0.1 and the solution evolves in time in a manner similar to that for the heat equation.

TABLE I

Error $\times 10^3$

		Zabusky–Kruskal		Hopscotch		Petrov–Galerkin		Modified P–G	
	t	L_2	L_∞	L_2	L_∞	L_2	L_∞	L_2	L_∞
$h = 0.05$		$\tau = 0.025$		$\tau = 0.025$		$\tau = 0.025$		$\tau = 0.025$	
	0.25	34.64	19.4	61.21	32.7	81.39	42.18	52.15	30.22
	0.50	122.68	63.5	122.41	67.4	102.54	51.85	64.90	22.85
	0.75	210.44	122.4	181.35	99.3	125.84	87.60	89.01	35.86
	1.00	298.19	161.4	228.10	141.6	150.57	100.41	107.20	39.39
$h = 0.033$						$\tau = 0.01$		$\tau = 0.01$	
	0.25					31.18	14.27	5.94	2.80
	0.50					43.35	21.65	7.56	4.53
	0.75					56.21	29.78	8.70	4.85
	1.00					74.08	39.37	9.49	5.85
$h = 0.01$		$\tau = 0.0005$		$\tau = 0.0005$		$\tau = 0.005$		$\tau = 0.005$	
	0.25	5.94	2.05	3.79	1.11	4.46	1.21	0.21	0.07
	0.50	13.17	4.22	9.28	2.14	7.01	2.15	0.38	0.11
	0.75	21.08	6.36	14.14	3.54	10.08	3.09	0.57	0.17
	1.00	28.66	8.13	18.72	4.91	13.26	3.83	0.74	0.21

6. Conclusions

A scheme for the numerical solution of equations governing nonlinear dispersive waves has been proposed. The new method gives a significant improvement, both from the point of view of accuracy and efficiency, over others which are currently available.

The Petrov–Galerkin approach enables us to use a C^0 interpolant, resulting in a much lower computational effort than that associated with the standard Galerkin method based on splines or Hermite cubics. Also the possibility of introducing asymmetric test functions (upwinding) has been presented.

The order of accuracy of approximations to linearised equations is generally higher than that for the corresponding nonlinear equation and a technique has been developed for improving the latter.

Acknowledgments

The authors want to express their gratitude to the Department of Mathematics, University of Dundee, for the help given and facilities offered to them while carrying out his work.

References

1. M. E. Alexander and J. Ll. Morris, *J. Computational Physics* **30** (1979), 428–451.
2. R. S. Anderssen and A. R. Mitchell, *Math. Methods Appl. Sci.* **1** (1979), 3–15.
3. J. Canosa and J. Gazdag, *J. Computational Physics* **23** (1977), 393–403.
4. I. S. Greig and J. Ll. Morris, *J. Computational Physics* **20** (1976), 67–80.
5. P. D. Lax, *Comm. Pure Appl. Math.* **21** (1968), 467–490.
6. A. R. Mitchell and I. Christie, Finite diference/finite element methods at the parabolic-hyperbolic interface, *in* "Numerical Analysis of Singular Perturbation Problems" (P. W. Hemker and J. J. H. Miller, Eds.) Academic Press, London, 1979.
7. H. H. Rachford and M. F. Wheeler, An H^{-1} Galerkin procedure, *in* "Mathematical Aspects of Finite Elements" (C. de Boor, Ed.), Academic Press, New York, 1974.
8. A. Sjöberg, *J. Math. Anal. Appl.* **29** (1970), 569–579.
9. L. Wahlbin, A dissipative Galerkin method for the numerical solution of first order hyperbolic equations, *in* "Mathematical Aspects of Finite Elements" (C. de Boor, Ed.), Academic Press, New York, 1974.
10. N. J. Zabusky and M. D. Kruskal, *Phys. Rev. Lett.* **15** (1965), 240–243.