JOURNAL OF COMPUTATIONAL PHYSICS 60, 282–305 (1985)

Numerical Experience with the Nonlinear Schrödinger Equation

B. M. HERBST

Department of Applied Mathematics, University of the Orange Free State, Bloemfontein 9300, South Africa

J. LL. MORRIS

Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

AND

A. R. MITCHELL

Department of Mathematical Sciences, The University, Dundee DD1 4HN, Scotland, United Kingdom

Received March 7, 1984; revised September 10, 1984

Increasing the magnitude of the parameter multiplying the nonlinear term of the nonlinear Schrödinger equation without changing the initial condition $u(x, 0) = \operatorname{sech} x$, leads to bound states of an increasing number of solitons. This results in very steep gradients in space and time and so provides a more severe test of numerical methods than before. In particular we find that methods which satisfy various conservation laws theoretically may now fail to do so in practice. Various analytical and numerical results relevant to this situation are discussed and illustrated by numerical examples. © 1985 Academic Press, Inc.

1. INTRODUCTION

The analytical properties of the nonlinear Schrödinger equation,

$$iu_t + u_{xx} + q\,|u|^2 u = 0, \qquad i^2 = -1 \tag{1}$$

where $u(x, t)$ is a complex-valued function defined over the whole real line and q is a real parameter, are well known. This is due to the fact that (1) is one of a relatively small number of soliton-producing equations that can be solved by the inverse scattering method, provided the initial condition $u(x, 0)$ vanishes for sufficiently large $|x|$ (see, e.g., Zakharov and Shabat [16], Strauss [13], Glassey [5], Ablowitz *et al.* [1], Miles [8]).

There also exists a growing literature on the numerical solution of (1) and recent references include Delfour *et al.* [3], Griffiths *et al.* [6], Mitchell and Morris [9],

0021-9991/85 $3.00

and Sanz-Serna and Manoranjan [12]. These investigations studied the numerical approximation of a single soliton, the interaction between two solitons, the emergence of solitons from arbitrary initial data, and the effect of dissipation on the solution. For these purposes standard numerical schemes have been modified or newer ones devised. All these investigations can be characterized as having used a relatively small value of the nonlinear coefficient q in (1).

The main purpose of the present study is to investigate the effect of increasing the value of q in (1). Increasing the value of q without changing the initial condition may lead to a bound state of more than one soliton (Miles [8]). This type of solution (see Peregrine [10] for the different types of solution allowed by the nonlinear Schrödinger equation) is not possible, for instance, for the Korteweg de Vries equation (Ablowitz *et al.* [1]) and has to our knowledge not been investigated numerically. Bound states of more than one soliton provide a more severe test for any numerical scheme and in Section 8 we show that great care must be taken for these problems.

The first part of this paper is devoted to those analytical properties of (1) which prove to be most important in obtaining and interpreting our numerical results. These are: the relationship between dispersion and nonlinearity, instability and conservation laws and finally, the existence of a bound state of more than one soliton.

The numerical solution of (1) is considered in the second part of the paper. The numerical schemes stem from two different space discretizations (17) and (21) and three different time discretizations. The methods for discretizing the time variable are:

(1) the energy-conserving, variable time step, leapfrog scheme devised by Sanz-Serna (see [11, 12]);

(2) the implicit midpoint rule, implemented in the predictor–corrector manner of Griffiths *et al.* [6]; and

(3) the scheme introduced by Delfour *et al.* [3].

The relative performances of these methods will be described for a suitably difficult test problem.

2. The Relationship between Dispersion and Nonlinearity

It is often valuable to consider the contributions from the linear and nonlinear parts of (1) separately. Accordingly we consider the linear part of (1), viz.

$$iv_t + v_{xx} = 0. \tag{2}$$

A general solution of (2) is given by

$$v(x, t) = \int_{-\infty}^{\infty} F(k) \exp i(kx - W(k)\, t)\, dk \tag{3}$$

where

$$W(k) = k^2.$$

Since

$$W''(k) \neq 0$$

the phase speed defined by

$$W(k)/k$$

depends on k. This shows our wave to be dispersive. Furthermore, an asymptotic analysis, for large values of x and t, shows that the amplitude behaves as $t^{-1/2}$, Whitham [14, Sect. 11.3].

The nonlinear terms in (1) oppose dispersion. The situation is clearly illustrated by Figs. 1(a), (b), and (c) which were obtained numerically using the same initial condition

$$u(x, 0) = \operatorname{sech} x$$

but different values of q. We note for $q = 2$ we obtain a precise balance of the dispersion and nonlinearity which allows a single soliton to be formed. A balance also occurs for $q = 8$, 18, or in general $q = 2N^2$ for integer N. These states correspond to bound states of N "solitons"—see Miles [8]. It should be pointed out that the same balance is obtained if q is fixed and a different initial condition is used. Thus a large nonlinear term $q|u|^2u$ and initial function sech x are equivalent to a term $|u|^2u$ and a large initial function $\sqrt{q}$ sech x.

Solitons are formed when a certain balance between nonlinearity and dispersion is reached. This is simply illustrated by the single soliton solution of (1) (cf. Whitham [14]),

$$u(x, t) = \left(\frac{2\alpha}{q}\right)^{1/2} \exp i\{\tfrac{1}{2}Sx - (\tfrac{1}{4}S^2 - \alpha)t\} \operatorname{sech} \alpha^{1/2}(x - St) \tag{4}$$

where S is the speed of the soliton and α a real parameter which determines its amplitude. In this situation, using as an initial condition for (1), Eq. (4) with $t = 0$, the balance of the nonlinearity and dispersion is achieved for all values of q because $q|u|^2$ is independent of q.

3. Instability and Conservation Laws

In this discussion of the stability of the nonlinear Schrödinger equation we deviate slightly from that given in Herbst *et al.* [7] and follow Whitham [14] more closely. We propose a perturbed solution to Eq. (1) of the form

$$u(x, t) = a(t) \exp(ikx) + \varepsilon_+(t) \exp(ik_+ x) + \varepsilon_-(t) \exp(ik_- x)$$

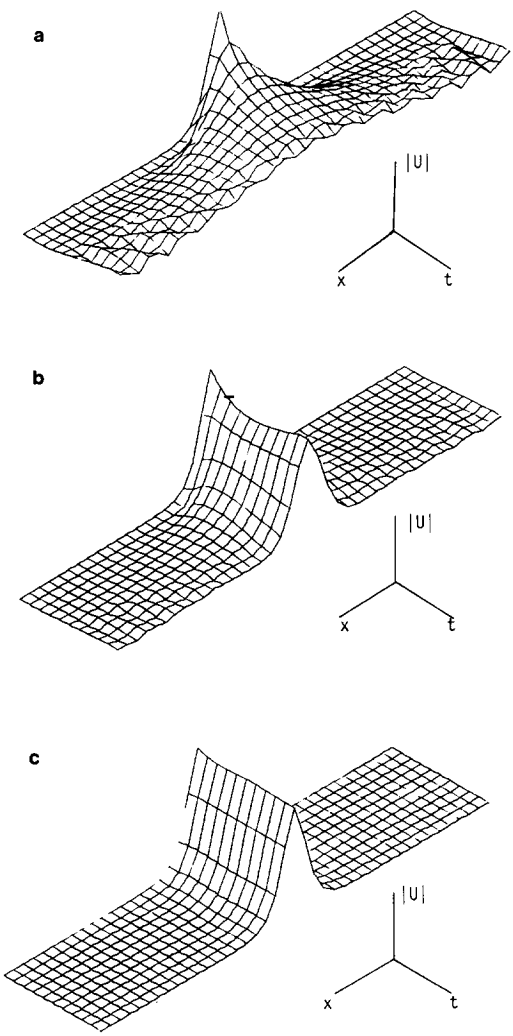

FIG 1. (a) $q=0.5$, (b) $q=1.5$, (c) $q=2.0$

where $\varepsilon_{\pm}$ are small in comparison with a. Substituting $u(x, t)$ into Eq. (1) we obtain

$$
\begin{aligned}
&i\dot{a}\exp(ikx)+i\dot{\varepsilon}_{+}\exp(ik_{+}x)+i\dot{\varepsilon}_{-}\exp(ik_{-}x)\\
&\quad -aw\exp(ikx)-\varepsilon_{+}w_{+}\exp(ik_{+}x)-\varepsilon_{-}w_{-}\exp(ik_{-}x)\\
&\quad +q\{|a|^{2}a\exp(ikx)+2|a|^{2}\varepsilon_{+}\exp(ik_{+}x)+2|a|^{2}\varepsilon_{-}\exp(ik_{-}x)\\
&\quad +a^{2}\varepsilon_{+}^{*}\exp(i\{2k-k_{+}\}x)+a^{2}\varepsilon_{-}^{*}\exp(i\{2k-k_{-}\}x)+O(\varepsilon_{\pm}^{2})\}=0 \qquad (5)
\end{aligned}
$$

where * denotes a complex conjugate and $w = k^2$, $w_\pm = k_\pm^2$. Thus apart from the original modes k, $k_\pm$, two additional modes $2k - k_\pm$ have been created. These new modes combine with the two original "side" modes $k_\pm$ whenever

$$2k = k_+ + k_-. \tag{6}$$

Assuming (6), compare coefficients in Eq. (5),

$$\begin{aligned} i\dot{a} - wa &= -q\,|a|^2 a \\ i\dot{\varepsilon}_+ - w_+\varepsilon_+ &= -2q\,|a|^2\varepsilon_+ - qa^2\varepsilon_-^* \\ i\dot{\varepsilon}_- - w_-\varepsilon_- &= -2q\,|a|^2\varepsilon_- - qa^2\varepsilon_+^* \end{aligned} \tag{7}$$

where we have ignored terms of $O(\varepsilon_\pm^2)$. Condition (6) will be satisfied if we choose

$$k_+ = k + \mu, \qquad k_- = k - \mu, \qquad \text{for any } \mu,$$

in which case a solution of (7) is given by

$$\begin{aligned} a(t) &= a_0 \exp i(-wt + q\,|a_0|^2\,t) \\ \varepsilon_\pm(t) &= \varepsilon_\pm(0) \exp i(q\,|a_0|^2 - w_\pm)\,t \exp i\mu^2 t \exp \pm i\sqrt{\mu^2(\mu^2 - 2q\,|a_0|^2)}\,t. \end{aligned} \tag{8}$$

The important conclusion to be drawn from (8) is that the side modes $\varepsilon_\pm$ will grow for

$$q > 0, \qquad \mu^2 < 2q\,|a_0|^2. \tag{9}$$

Of course this result only holds while $\varepsilon_\pm$ are small. The long time behaviour of these modes is determined by the conservation laws

$$\frac{d}{dt}\int_{-\infty}^{\infty} |u|^2\,dx = 0 \tag{10a}$$

$$\frac{d}{dt}\int_{-\infty}^{\infty} \left(|u_x|^2 - \frac{1}{2}q\,|u|^4\right) dx = 0 \tag{10b}$$

satisfied by solutions of (1). The conservation laws prevent the side modes from growing exponentially for an indefinite time. Although this mechanism is not perfectly understood, there exists experimental and numerical evidence that all the modes may under favourable conditions return to their initial state, and start the whole process again (Yuen and Ferguson [15]). This remarkable phenomenon is known as recurrence.

4. Bound State of More Than One Soliton

It is evident from the single soliton solution (4) that it is possible to select the speed s and amplitude α of the soliton separately. This allows the possibility that solitons with different amplitudes may move at the same speed and all the time interacting with one another. A precise result was obtained by Miles [8]. He showed that the initial condition

$$u(x, 0) = \operatorname{sech} x \tag{11}$$

will produce a bound state of N solitons if

$$q = 2N^2, \qquad N = 1, 2, \ldots. \tag{12}$$

A similar result does not exist for the Korteweg de Vries equation which does not allow independent values for the amplitudes and velocities of its solitons.

For the numerical results reported in Section 8, we solved (1) and (11) for $q = (2), 8$ and 18 which correspond to $N = (1), 2$ and 3 in (12).

5. The Finite Element Method

We approximate the solution u of (1) by U, where

$$U(x, t) = \sum_{j=0}^{m} U_j(t)\, \phi_j(x), \tag{13}$$

$U_j(t)$ are complex valued functions of time, and $\phi_j(x)$ are piecewise linear basis functions defined with respect to a uniform grid with grid spacing h.

According to the Galerkin finite element method the coefficients are determined from

$$i(\dot{U}, \phi_j) - (U_x, \phi_j') + q(|U|^2 U, \phi_j) = 0, \qquad j = 0, \ldots, m. \tag{14}$$

where a dot $\cdot$ denotes differentiation with respect to the time and a dash $'$ with respect to x. In practice we find it awkward to deal with the nonlinear term appearing in (14) and product approximation (Christie *et al.* [2]) is used instead, i.e., the nonlinear term in (14) is approximated by

$$(|U|^2 U, \phi_j) \approx \sum_{k=0}^{m} |U_k|^2\, U_k(\phi_k, \phi_j). \tag{15}$$

In addition, instead of solving a complex system, we separate real and imaginary parts

$$U_k = V_k + iW_k. \tag{16}$$

Making use of (15) and (16), Eq. (14) becomes

$$M\dot{\mathbf{U}} + \frac{1}{h^2} S\mathbf{U} + qM\mathbf{F}(\mathbf{U}) = \mathbf{0} \tag{17}$$

where

$$\mathbf{U} := (\mathbf{U}_0, \ldots, \mathbf{U}_n)^{\mathrm{T}}, \qquad \mathbf{U}_j := (V_j, W_j)^{\mathrm{T}}$$

$$M := \tfrac{1}{6} \begin{bmatrix} 2I & I & & & & \\ I & 4I & I & & O & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & O & & I & 4I & I \\ & & & & I & 2I \end{bmatrix},$$

$$S := \begin{bmatrix} -A & A & & & & \\ A & -2A & A & & O & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & O & & A & -2A & A \\ & & & & A & -A \end{bmatrix}$$

$$I := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

$$\mathbf{F} := (\mathbf{F}_0, \ldots, \mathbf{F}_n)^{\mathrm{T}}, \qquad \mathbf{F}_j := \mathbf{U}_j^{\mathrm{T}} \mathbf{U}_j A \mathbf{U}_j, \tag{18}$$

and where we have assumed natural boundary conditions (see Sect. 8). Herbst *et al.* [7] investigated the stability of (17) and, in particular, the relationship with the analytical results obtained in Section 3. They showed that small perturbations of the solution of (17) of the form

$$\boldsymbol{\phi}_j(t) = \mathbf{d}_j(t) \sum_k \exp i\alpha_k x_j$$

will grow exponentially in time whenever

$$q > 0, \qquad \frac{4}{h^2} \sin^2 \frac{\alpha h}{2} < 2\gamma q \, |a|^2 \tag{19}$$

where

$$\gamma := 1 - \frac{2}{3}\sin^2\frac{\alpha h}{2}, \qquad |a|^2 := \mathbf{U}^{\mathrm{T}}\mathbf{U}.$$

Again we require a mechanism to prevent these unstable modes from growing indefinitely. This is provided by finite element analogues of (10a) and (10b). The analogue of (10a) is obtained by multiplying (14) by U_j^*, summing over all j and taking the imaginary part; to give

$$\frac{d}{dt}\int_{-\infty}^{\infty} |U|^2\, dx = 0. \tag{20a}$$

To obtain the analogue of (10b), we multiply (14) by $\dot{U}_j^*$, sum over j and take the real part, to give

$$\frac{d}{dt}\int_{-\infty}^{\infty} \left(|U_x|^2 - \frac{1}{2}q\,|U|^4\right) dx = 0. \tag{20b}$$

The penalty we pay for the simplification obtained from using product approximation is that (20a) and (20b) are no longer satisfied. However, discrete analogues of (20) are satisfied if we resort to mass lumping, i.e., if we replace M in (17) by

$$\tilde{I} := \begin{bmatrix} \frac{1}{2}I & & & \\ & I & & O \\ & & \ddots & \\ & O & & I \\ & & & \frac{1}{2}I \end{bmatrix}$$

in which case (17) becomes

$$\tilde{I}\dot{\mathbf{U}} + \frac{1}{h^2}S\mathbf{U} + q\tilde{I}\mathbf{F}(\mathbf{U}) = 0. \tag{21}$$

Apart from the boundary conditions, (21) is a straightforward difference replacement of (1), see, e.g., Sanz-Serna and Manoranjan [12]. In a similar way as before, replacing integrals by appropriate summation, it can be shown that the theoretical solution of (21) satisfies

$$\frac{d}{dt}\left[h\sum_j |U_j|^2\right] = 0 \tag{22a}$$

$$\frac{d}{dt}\left[h\sum_j \left(\left|\frac{U_{j+1}-U_j}{h}\right|^2 - \frac{1}{2}q\,|U_j|^4\right)\right] = 0. \tag{22b}$$

It is therefore theoretically possible to have discrete analogues of (10) satisfied as long as the time remains continuous. Utmost care should therefore be exercised in discretizing the time variable, as this may be the main source of error in a discretized scheme.

6. Conservation of the First Quantity

Discretizing the time variable, we require (20a) and (22a) to be satisfied at all time levels, i.e.,

$$\int_{-\infty}^{\infty} |U^n|^2 \, dx = \int_{-\infty}^{\infty} |U^0|^2 \, dx, \qquad n = 1, 2, \ldots, \tag{23}$$

and

$$\sum_j |U_j^n|^2 = \sum_j |U_j^0|^2, \qquad n = 1, 2, \ldots, \tag{24}$$

where the superscript n denotes the nth time level.

We first observe that (14) and (21) may be written in the form

$$G\dot{\mathbf{U}} = \mathbf{H}(\mathbf{U}) \tag{25}$$

where $G := M$ or $G := \tilde{I}$. In both cases $\mathbf{H}(\mathbf{U})$ satisfies (cf. Sanz-Serna [11])

$$\mathbf{U}^T \mathbf{H}(\mathbf{U}) = 0. \tag{26}$$

The implicit midpoint rule for solving (25) is given by

$$G\mathbf{U}^{n+1} = G\mathbf{U}^n + \Delta t \mathbf{H}(\tfrac{1}{2}(\mathbf{U}^n + \mathbf{U}^{n+1})). \tag{27}$$

If we premultiply (27) by $\frac{1}{2}(\mathbf{U}^n + \mathbf{U}^{n+1})^T$ and make use of (26) we obtain

$$\mathbf{U}^{n+1^T} G \mathbf{U}^{n+1} = \mathbf{U}^{n^T} G \mathbf{U}^n.$$

It is not possible for (24) to be satisfied in general by solving (25) by an explicit scheme. However, Sanz-Serna [11, 12] showed that by using a variable time step the leapfrog scheme is capable of doing this. The leapfrog scheme applied to (25) gives

$$G\mathbf{U}^{n+1} = G\mathbf{U}^{n-1} + (\tau_n + \tau_{n-1}) \mathbf{H}(\mathbf{U}^n) \tag{28}$$

where

$$\tau_n := t^{n+1} - t^n. \tag{29}$$

Premultiplying (28) by $\mathbf{U}^{n+1^{\mathrm{T}}}$ and elimination of $\mathbf{U}^{n+1^{\mathrm{T}}}$ from the right hand side of the resulting expression shows that

$$\mathbf{U}^{n+1^{\mathrm{T}}} G \mathbf{U}^{n+1} = \mathbf{U}^{n-1^{\mathrm{T}}} G \mathbf{U}^{n-1}$$

provided

$$(\tau_n + \tau_{n\ 1})[2\mathbf{U}^{n-1^{\mathrm{T}}}\mathbf{H}(\mathbf{U}^n) + (\tau_n + \tau_{n-1})\,\mathbf{H}^{\mathrm{T}}(\mathbf{U}^n)\,G^{-1}\mathbf{H}(\mathbf{U}^n)] = 0. \tag{30}$$

From stability considerations Sanz-Serna and Manoranjan [12] advise the use of

$$\tau_n = 2(\mathbf{U}^n - \mathbf{U}^{n-1})^{\mathrm{T}}\mathbf{H}(\mathbf{U}^n)/\mathbf{H}^{\mathrm{T}}(\mathbf{U}^n)\,G^{-1}\mathbf{H}(\mathbf{U}^n) - \tau_{n-1} \tag{31}$$

which may be obtained from (30) and (26). Because of the appearance of G^{-1} in (31) we used the choice

$$G = \tilde{I}$$

in all our numerical experiments with the leapfrog scheme (28) and (31).

It should be pointed out that although we have derived and written the variable time step leapfrog scheme in a slightly different way than the originators (see, e.g., Sanz-Serna [11] and Sanz-Serna and Manoranjan [12]), the present scheme is in fact the same as the original. More specifically, given t^0, t^1, $\mathbf{U}^0$, and $\mathbf{U}^1$ the two schemes produce exactly the same sequences

$$t^n, \qquad \mathbf{U}^n, \qquad n = 2, 3, \ldots,$$

as a comparison between our Eqs. (28), (29), (31) and eqs. (3.1), (3.2), and (4.1) of Sanz-Serna [11] shows.

6. Conservation of the Second Quantity

The time discretized analogues of (20b) and (22b), are

$$\int_{-\infty}^{\infty} (|U_x^n|^2 - \tfrac{1}{2}q(U^n)^4)\,dx = \int_{-\infty}^{\infty} (|U_x^0|^2 - \tfrac{1}{2}q\,|U^0|^4)\,dx, \qquad n = 1, 2, \ldots, \tag{32}$$

and

$$\sum_j \left(\left| \frac{U_{j+1}^{n+1} - U_j^n}{h} \right|^2 - \frac{1}{2} q\,|U_j^n|^4 \right) = \sum_j \left(\left| \frac{U_{j+1}^0 - U_j^0}{h} \right|^2 - \frac{1}{2} q\,|U_j^0|^4 \right), \qquad n = 1, 2, \ldots, \tag{33}$$

respectively. We now return to the finite element Eqs. (14) and apply the implicit midpoint rule to obtain

$$i(U^{n+1}-U^n,\phi_j)-\tfrac{1}{2}\Delta t(U_x^{n+1}+U_x^n,\phi_j')+\tfrac{1}{8}\Delta tq(|U^{n+1}+U^n|^2(U^{n+1}+U^n,\phi_j)=0. \tag{34}$$

Multiply (34) by $U_j^{*n+1}-U_j^{*n}$, sum over j, and take the real part to obtain

$$-\tfrac{1}{2}\int_{-\infty}^{\infty}|U_x^{n+1}|^2\,dx+\tfrac{1}{8}q\int_{-\infty}^{\infty}|U^{n+1}+U^n|^2\,|U^{n+1}|^2\,dx$$

$$=-\tfrac{1}{2}\int_{-\infty}^{\infty}|U_x^n|^2\,dx+\tfrac{1}{8}q\int_{-\infty}^{\infty}|U^{n+1}+U^n|^2|U^n|^2\,dx. \tag{35}$$

Making use of the identity

$$\tfrac{1}{4}|U^{n+1}+U^n|^2=\tfrac{1}{2}(|U^{n+1}|^2+|U^n|^2)-\tfrac{1}{4}|U^{n+1}-U^n|^2, \tag{36}$$

(35) becomes

$$\int_{-\infty}^{\infty}(|U_x^{n+1}|^2-\tfrac{1}{2}q\,|U^{n+1}|^4)\,dx=\int_{-\infty}^{\infty}(|U_x^n|^2-\tfrac{1}{2}q\,|U^n|^4)\,dx$$

$$+\tfrac{1}{2}\int_{-\infty}^{\infty}|U^{n+1}-U^n|^2(|U^{n+1}|-|U^n|)(|U^{n+1}|+|U^n|)\,dx.$$

Thus, apart from an $O(\Delta t^3)$ term the second quantity (32) is conserved by the implicit midpoint rule. This quantity may clearly be conserved exactly if we ignore the $O(\Delta t^2)$ on the right-hand side of (36), a result first used by Delfour, Fortin, and Payre [3]. Thus, instead of (34) we use

$$i(U^{n+1}-U^n,\phi_j)-\tfrac{1}{2}\Delta t(U_x^{n+1}+U_x^n,\phi_j')$$

$$+\tfrac{1}{4}\Delta tq((|U^{n+1}|^2+|U^n|^2)(U^{n+1}+U^n),\phi_j)=0. \tag{37}$$

Similarly (33) is satisfied theoretically if we discretize (21) using the implicit midpoint rule but instead of using

$$\tfrac{1}{8}(\mathbf{U}_j^n+\mathbf{U}_j^{n+1})^T(\mathbf{U}_j^n+\mathbf{U}_j^{n+1})\,A(\mathbf{U}_j^n+\mathbf{U}_j^{n+1}) \tag{38}$$

for the nonlinear terms, we use

$$\tfrac{1}{4}(\mathbf{U}_j^{nT}\mathbf{U}_j^n+\mathbf{U}_j^{n+1T}\mathbf{U}_j^{n+1})\,A(\mathbf{U}_j^n+\mathbf{U}_j^{n+1}). \tag{39}$$

In the next section we discuss the results of numerical experiments based on the discretized schemes described in this paper.

8. Numerical Results

The numerical results reported in this section are obtained by solving (1) for various values of q, using the initial condition (11). The following methods of solution are used:

I. (17), together with the implicit midpoint rule (38).

II. (21), together with the implicit midpoint rule (38).

III. (17), together with the Delfour *et al.* modification (39).

IV. (21), together with the Delfour *et al.* modification (39).

V. (21), together with the variable time-step, leapfrog scheme (28) and (31).

The first four schemes are implicit schemes which require a nonlinear system of equations to be solved at each time level. These equations may be written in the form

$$G\mathbf{U}^{n+1} = \mathbf{P}(\mathbf{U}^n, \mathbf{U}^{n+1}) \tag{40}$$

where G is given by

$$G := M + rS \qquad \text{or} \qquad G := \tilde{I} + rS, \qquad r := \Delta t/h^2,$$

and $\mathbf{P}$ is determined by the specific method in use. The nonlinear system (40) was solved by a predictor–corrector procedure. First determine $\mathbf{U}^{(1)}$ by using Euler's method. Successive improvements are calculated from

$$G\mathbf{U}^{(k)} = \mathbf{P}(\mathbf{U}^n, \mathbf{U}^{(k-1)}), \qquad k = 2, 3, \ldots. \tag{41}$$

The main advantage of (41) lies in its low computational cost. Since G is time independent it needs to be factorized once only. After its factorization at the beginning only one forward and one backward substitution are required for each iteration. The computational cost of more sophisticated iteration procedures may become prohibitive. For instance, if a Newton type procedure is employed the Jacobian needs to be updated at very short time intervals due to the large temporal gradients we wish to compute. In addition it will be shown that (41) is adequate at least for methods I and III. In order to use the leapfrog scheme, method V, the first time step τ_0 and solution at the first time level must be provided. Following the suggestion of Sanz-Serna and Manoranjan [12], we used values of τ_0 in the vicinity of the linearized stability limit

$$\tau_0 = \tfrac{1}{4}h^2. \tag{42}$$

The starting solution at $t = \tau_0$ will be provided by using Euler's method or alternatively the implicit midpoint rule.

We had already observed that schemes I and III do not conserve (exactly) either of the two quantities implicit in (20a) and (20b). Schemes II and V conserve only the first quantity and only scheme IV conserves both quantities. Due to round-off error and the computational cost of iterating (41) a large number of times, the best any method can do in practice is to conserve the quantities to a fixed number of decimal places. On the other hand the importance of the conservation laws was argued in Sections 3 and 5 and by Herbst *et al.* [7]. It is therefore of considerable interest to investigate the ability of the methods to conserve the quantities in practice.

Except when otherwise stated we adopt the following strategy in our numerical procedures. At each time level a certain maximum number of corrector iterations (cf. (41)) is specified. After each iteration the quantities

$$C_{1_n} := \sum_j |U_j^n|^2 \tag{43a}$$

$$C_{2_n} := \sum_j \left(\left| \frac{U_{j+1}^n - U_j^n}{h} \right|^2 - \frac{1}{2} q |U_j^n|^4 \right) \tag{43b}$$

are calculated and compared with the quantities C_{1_0} and C_{2_0} obtained from the initial condition. In order to save on computation cost, the iteration terminated whenever

$$|C_{1_n} - C_{1_0}| < 2 \times 10^{-4}. \tag{44}$$

Because of our inability to compute over the whole real axis, we imposed natural boundary conditions at $x = -20$ and $x = 20$. In this manner our imposed boundary conditions had no influence on the solution.

The numerical approximation of the single soliton solution of (1) with $q = 2.0$ and initial condition (11) has already been thoroughly investigated without any problems. For instance Griffiths *et al.* [6] did not require to iterate the corrector as in (41) and Sanz-Serna and Manoranjan [12] obtained reasonable results using Euler's method to provide the missing starting value required at $t = \tau_0$. Even the case $q = 8$ (bound state of two solitons according to Sect. 4) does not present serious problems. Figures 2 and 3 show the results obtained with methods II and IV. In all our surface plots we show $|U|$ as a function of time and x. In this case a fixed number of 10 iterations per time step was used. As was predicted theoretically, method IV conserved both quantities and method II the first quantity, at least to four decimal places. Even though method II does not conserve the second quantity even to two decimal places, there is no apparent difference between the two figures.

The situation changes dramatically when we increase the value of q to 18 (bound state of three solitons). Again using $h = 0.125$, the solution from IV became unbounded after 14 and 18 time steps for values of $\Delta t = 0.0125$ and 0.01, respectively. A further decrease in Δt to 0.005 did not lead to an unbounded solution (at least for the first 200 time steps); however, the iteration procedure did start having

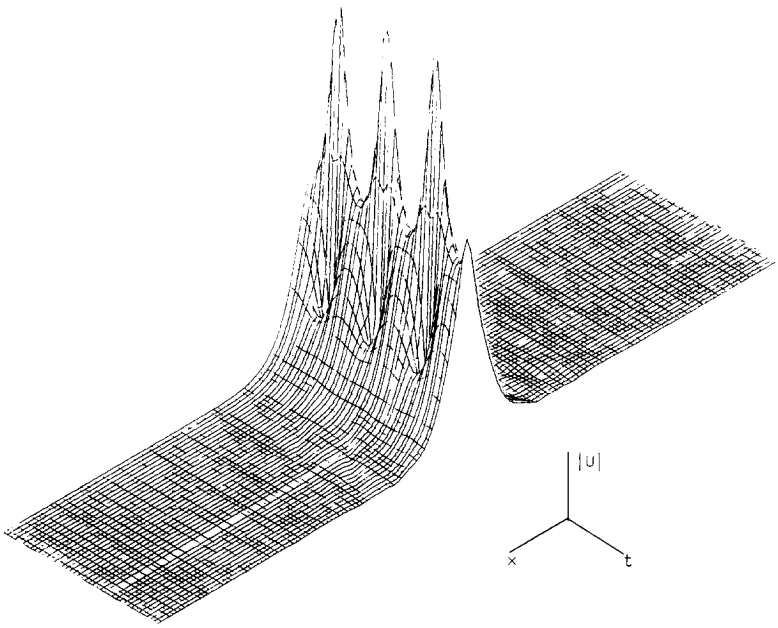

FIG. 2. Method II, $q = 8.0$, $\Delta t = 0.0125$, $h = 0.125$, printed every 5th step for 200 steps.

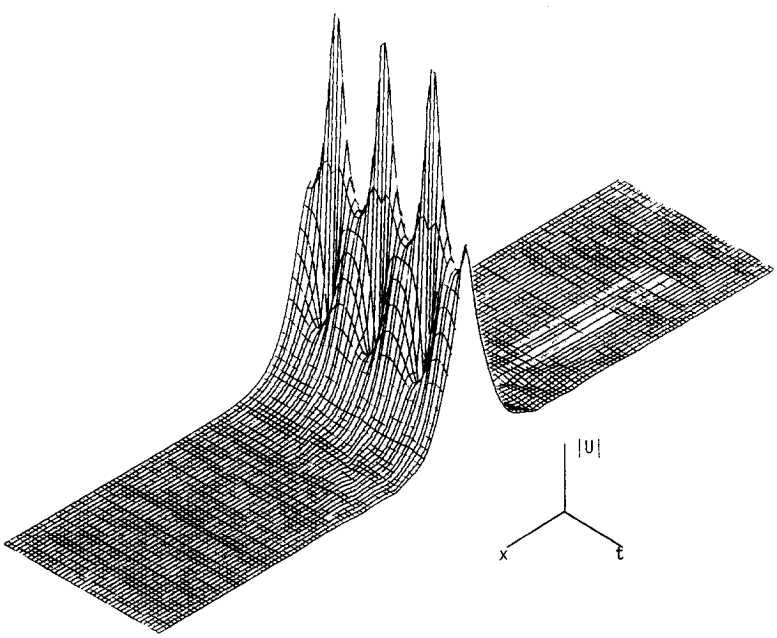

FIG. 3. Method IV, $q = 8.0$, $\Delta t = 0.0125$, $h = 0.125$, printed every 5th step for 200 steps.

convergence problems around 36 time steps. For a value of $h = 0.1$ the same pattern emerged. For values of $\Delta t = 0.0125$ and 0.01 the solution became unbounded after 15 and 47 time steps, respectively. For $\Delta t = 0.01$ the iteration procedure had difficulty in converging after 17 time steps. For $\Delta t = 0.005$ the solution remained bounded although some difficulty to converge was noticed around 38 and after 86 time steps. This solution is shown in Fig. 4. The difficulties mentioned above were also reflected in the behaviour of the conserved quantities (43). For instance, for $h = 0.1$ and $\Delta t = 0.01$, C_{1_n} changed from 2.0000 to 1.9417 and C_{2_n} from -5.6748 to -3.8983 after 17 to 21 time steps. After this time these values remained the same until the solution became unbounded after 47 time steps. Also, for $h = 0.1$ and $\Delta t = 0.005$, C_{1_n} and C_{2_n} changed from 1.9999 to 1.9175 and from -5.6251 to -1.7746 respectively, after 80 to 90 time steps. After these changes the values remained unchanged for the remainder of our calculations.

All the difficulties mentioned above coincided with the formation of one of the spikes seen in Figs. 8 and 9. At these spikes the temporal gradient of the solution is very large. The improved behaviour obtained when decreasing Δt suggests that the source of the trouble is our iteration scheme which has difficulty in converging from a bad prediction by Euler's method using too large values of Δt.

Method II also had difficulty in giving a reasonable representation of the analytical solution. The numerical solution using $h = 0.125$ and $\Delta t = 0.0125$ is shown in Fig. 5. A comparison with Figs. 8 and 9 shows that this solution does not

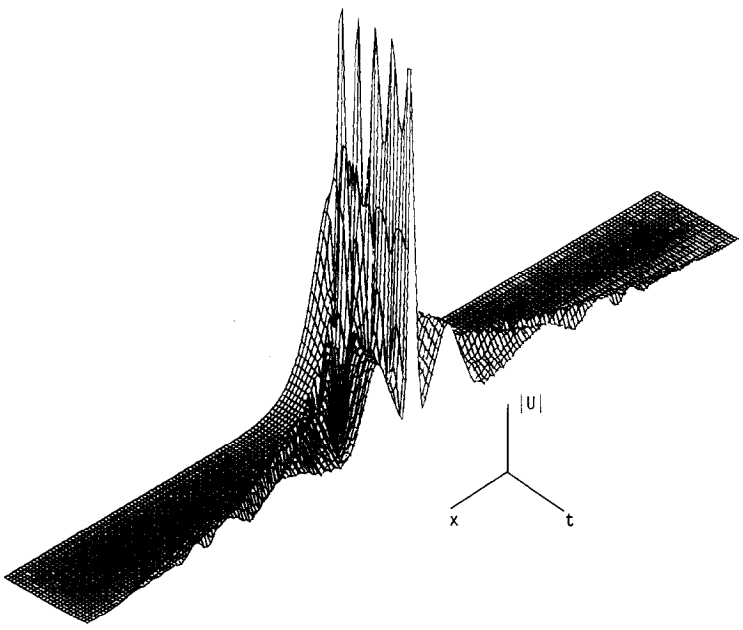

FIG. 4. Method IV, $q = 18$, $\Delta t = 0.005$, $h = 0.1$, printed every 10th step for 250 steps.

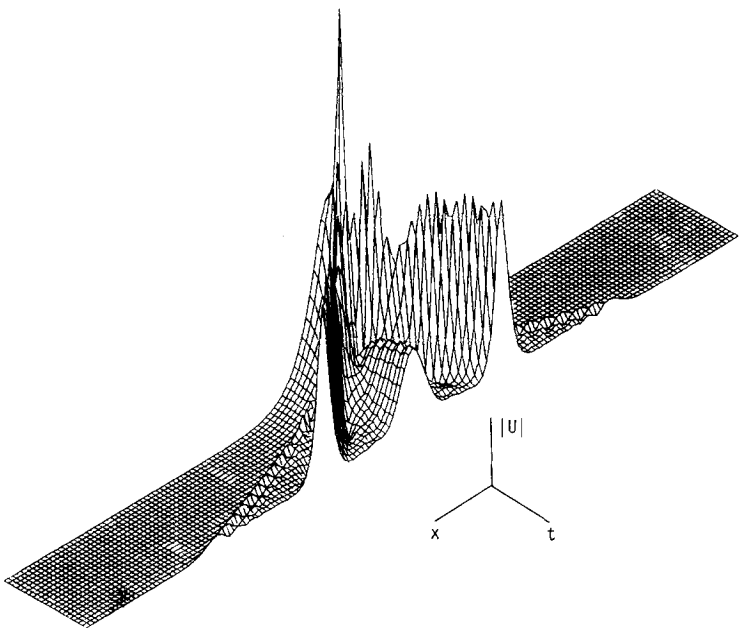

FIG. 5. Method II, $q = 18$, $\Delta t = 0.0125$, $h = 0.125$, printed every 5th step for 100 steps.

resemble the analytical solution. The difficulties are the same as with method IV. For instance, between 10 and 15 time steps, which coincide with the first spike, the values of C_{1_n} and C_{2_n} changed from 2.0000 to 1.5402 and from -5.7394 to 2.6824, respectively. Again an improvement is obtained if the values of h and Δt are reduced. Figure 6 shows the solution using $h = 0.1$ and $\Delta t = 0.005$. We observed that the method had difficulty in converging between 85 and 115 time steps, where up to 20 iterations (the maximum number allowed) were required. For the remainder of the calculations the method had no difficulty in converging. Also, no significant

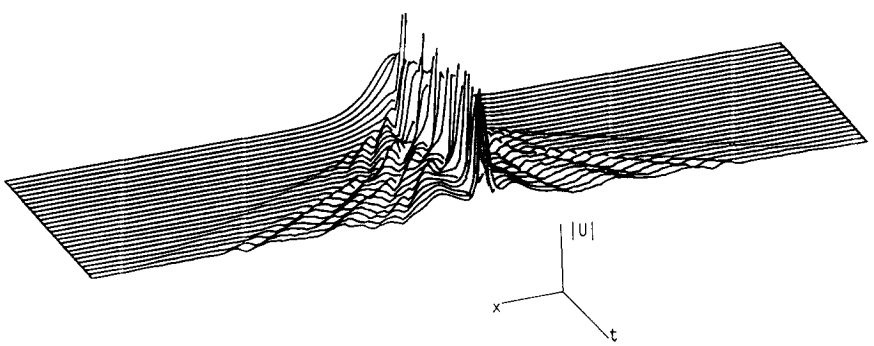

FIG. 6. Method II, $q = 18$, $\Delta t = 0.005$, $h = 0.1$, printed every 10th step for 250 steps.

change in the conserved quantities was observed. However, Fig. 6 shows the pronounced downstream oscillations which started just before the first convergence difficulties were encountered. Since the total energy (43a) remained within the limit imposed by (44), the oscillations provide a means of "leaking" energy from the central spine with noticable adverse effects on the quality of the solution (cf. Figs. 9 and 10).

Figures 7 and 8 show the solution obtained from methods I and III using $h = 0.125$ and $\Delta t = 0.0125$ and although the quality of the approximation may not be completely acceptable the solutions do resemble the analytical solution. Thus, even though neither quantity is conserved theoretically, somewhat surprisingly, these methods do better in practice than methods II and IV.

A further improvement may be obtained by reducing the values of h and Δt to 0.1 and 0.005, respectively. The results for methods I and III are shown in Figs. 9 and 10. For these values it was possible to satisfy (44) using an average of approximately 3 or 4 iterations per time step. Thus, not only was the quality of the approximation improved but the solutions were obtained at lower computational cost. Again there is little difference between the two figures, even though neither method conserved the second quantity even to one significant figure. A further reduction in the values of h and Δt to 0.067 and 0.0025 did not have any significant effect on the solutions.

Figure 11 shows the solution obtained from method I using $h = 0.5$, a con-

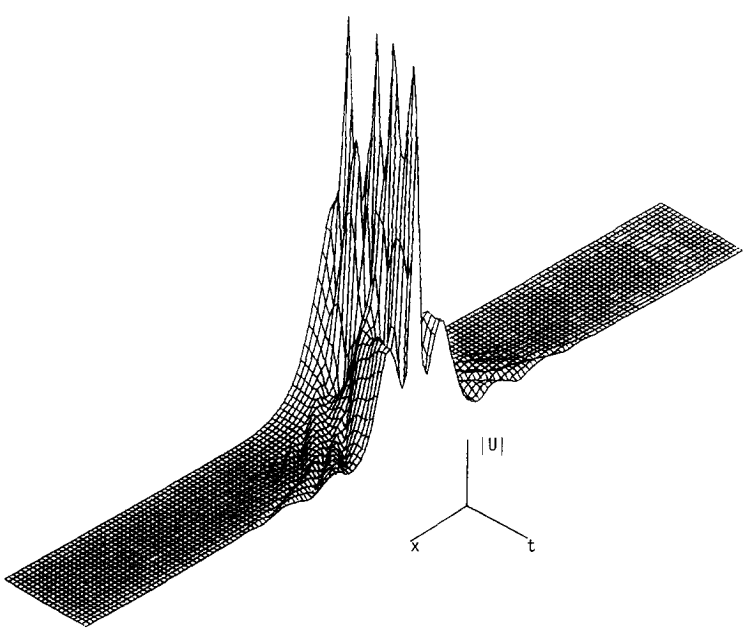

FIG. 7. Method I, $q = 18$, $\Delta t = 0.0125$, $h = 0.125$, printed every 5th step for 100 steps.

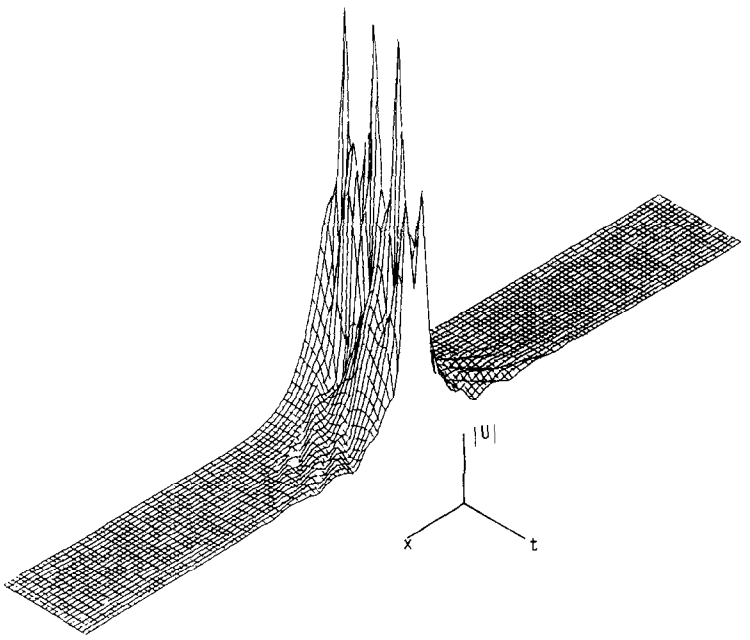

FIG. 8. Method III, $q = 18$, $\Delta t = 0.0125$, $h = 0.125$, printed every 5th step for 100 steps.

siderable increase in the size of the space step. The interesting feature of this solution is that, although (44) was satisfied using just over 3 iterations per time step, the solutions do not resemble the solutions in Figs. 9 and 10. This may be explained in the terms of Section 2. A large grid spacing h does not allow an accurate approximation of the linear, dispersive part of (1). This leads to an imbalance between dispersion and nonlinearity in the numerical scheme which may easily result in a different solution from the one expected.

Finally we used the leapfrog scheme, method V. Since this method is devised to conserve the first quantity very accurately and hopefully guarantee a good solution, our main interest lies in the behaviour of the time step. Figure 12 shows τ_n as a function of n using Euler's method to provide the extra starting values at $t = \tau_0$. According to (42), τ_0 is chosen as 0.0025 when $h = 0.1$ (smaller and bigger values, $\tau_0 = 0.005$ and $\tau_0 = 0.001$, give less satisfactory results). From Fig. 12 it is clear that there is a split in the behaviour at odd and even numbers of time steps. The reason for this behaviour lies in the fact that Euler's method is used to provide the extra starting values. Since Euler's method does not conserve the energy, the time step is adjusted in such a way that different quantities are conserved at odd and even time levels.

Use of the implicit midpoint rule to provide the starting values improves matters dramatically. Now the same quantity is conserved at all time levels and we have no distinction between even and odd time levels, see Fig. 13. From the figure two facts

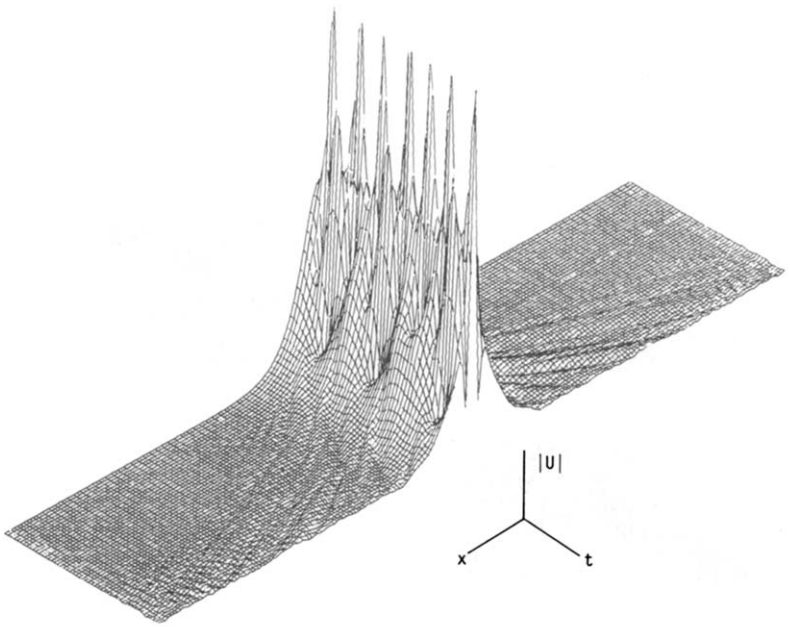

FIG. 9. Method I, $q = 18$, $\Delta t = 0.005$, $h = 0.1$, printed every 10th step for 500 steps.

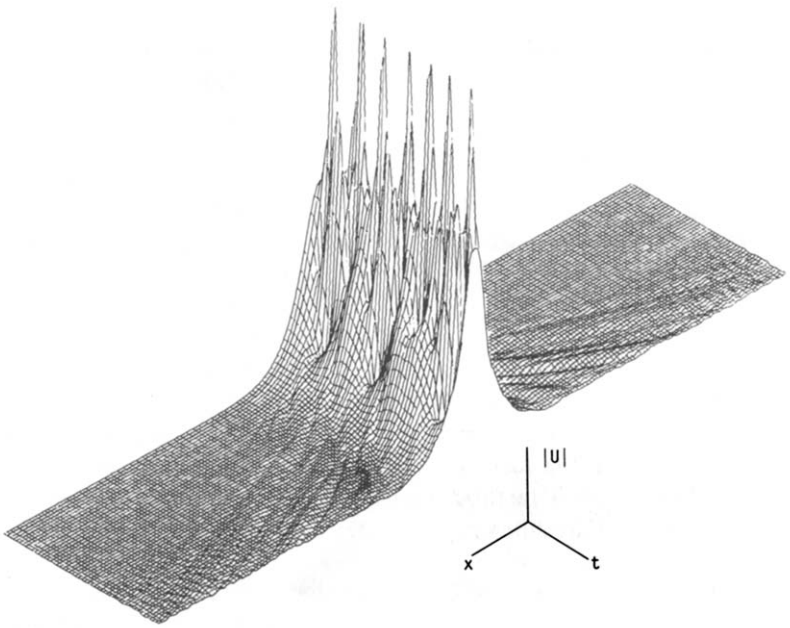

FIG. 10. Method III, $q = 18$, $\Delta t = 0.005$, $h = 0.1$, printed every 10th step for 500 steps.

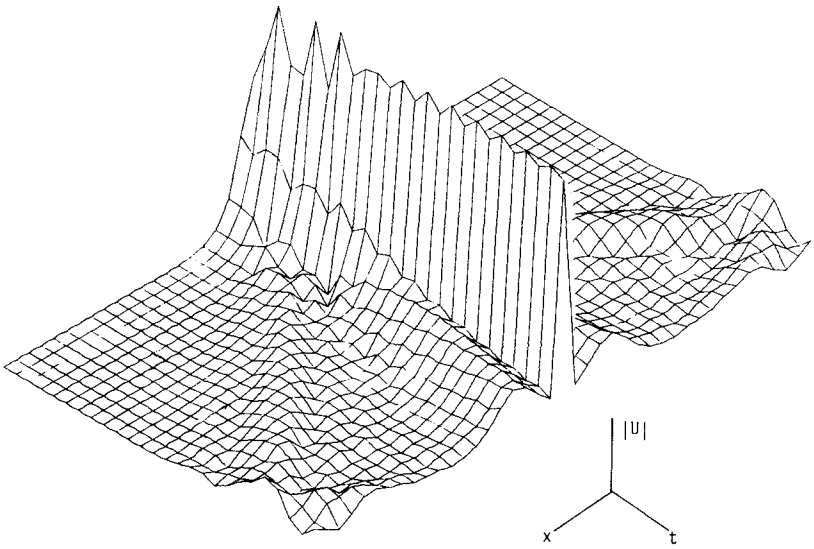

FIG. 11. Method I, $q = 18$, $\Delta t = 0.005$, $h = 0.5$, printed every 20th step for 500 steps.

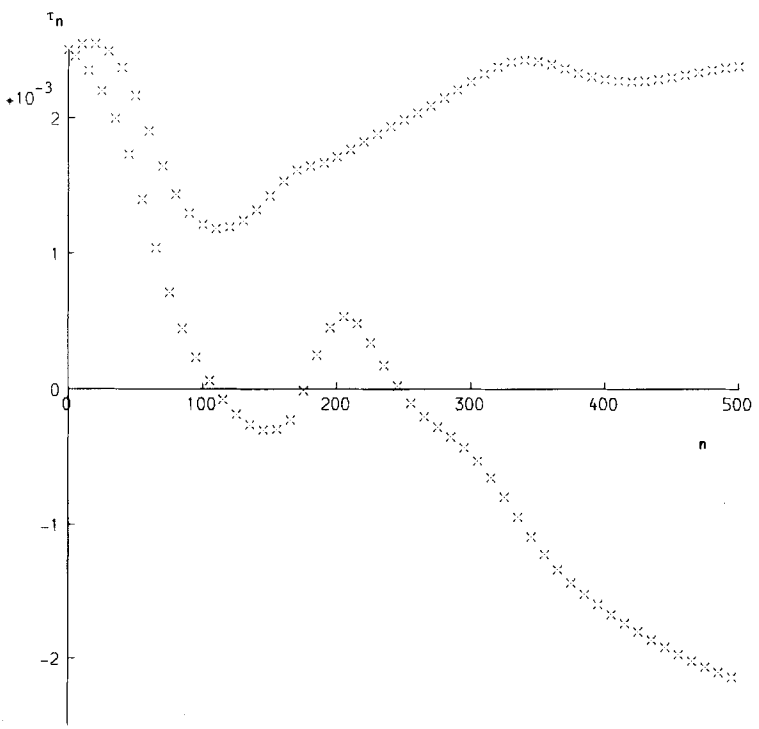

FIG. 12. Method V, $q = 18$, $h = 0.1$, $\tau_0 = 0.0025$, missing starting value by Euler's method.

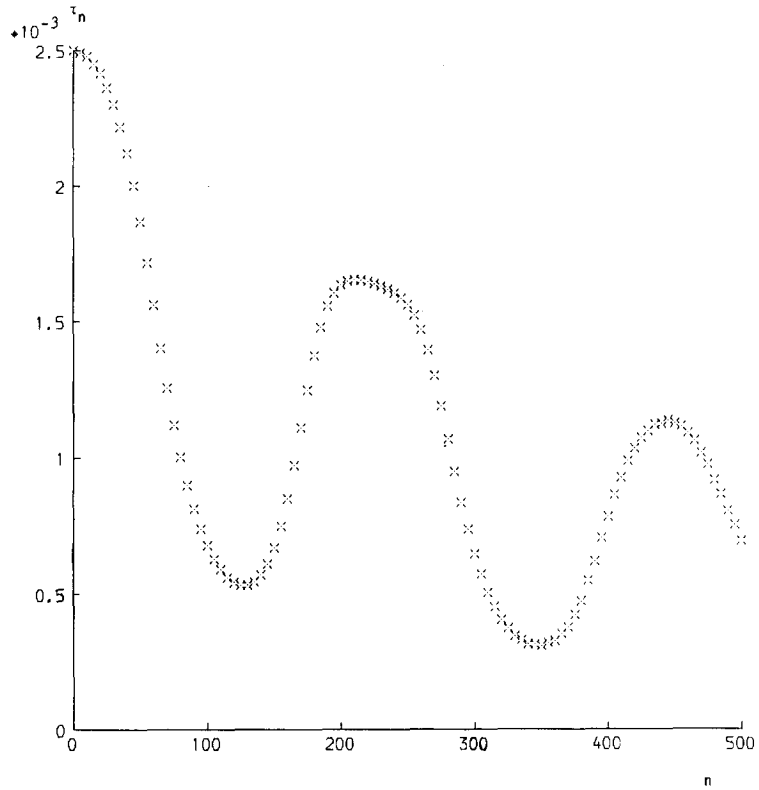

FIG. 13. Method V, $q = 18$, $h = 0.1$, $\tau_0 = 0.0025$, missing starting value by the midpoint rule.

become evident. The time step displays a periodic oscillation and, on average, is decreasing all the time. In order to demonstrate that this periodic behaviour is linked to the periodic nature of the solution as shown in Figs. 9 and 10 we show the value of $|U|$ at $x = 0$ at the same time levels as τ_n in Fig. 14. The fact that τ_n is small when $|U|$ is large is explained by the equidistributing principle (cf. Sanz-Serna and Manoranjan [12])

$$\|\mathbf{U}^{n+1} - \mathbf{U}^n\| = \|\mathbf{U}^n - \mathbf{U}^{n-1}\|$$

which is satisfied by the method. This means that the time step is small when the temporal gradient is large. The temporal gradient is large in the vicinity of the spikes, hence the small τ_n when $|U|$ is large.

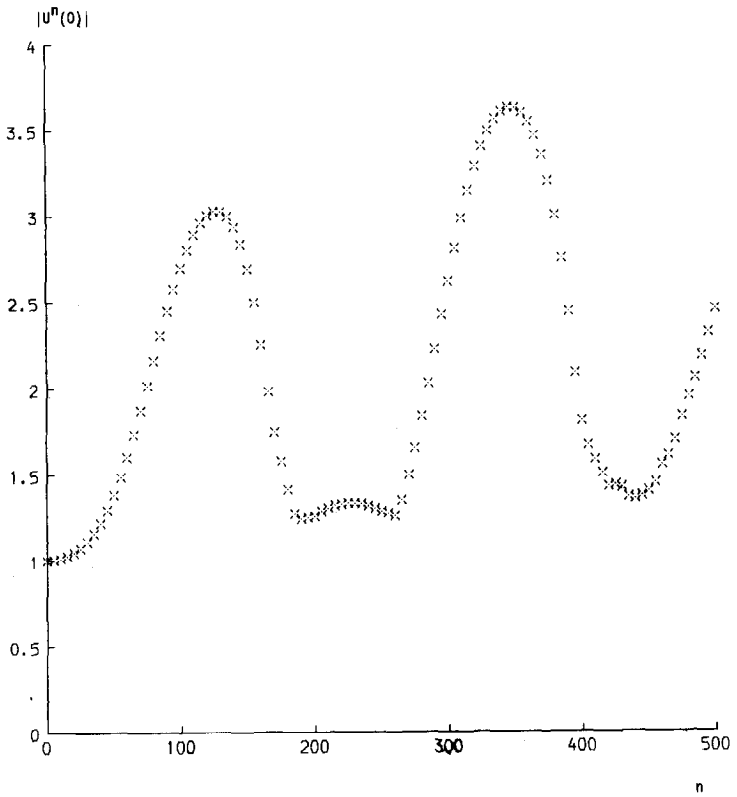

FIG. 14. Method V, $|U^n(0)|$ against n.

9. CONCLUSIONS

By increasing the nonlinear coefficient q of the NLS, while keeping the initial condition (11) unchanged, a bound state of an increasing number of solitons is obtained. Thus very steep spatial and temporal gradients are developed which provide a severe test for the various numerical schemes. Our numerical experiments under these more stringent conditions clearly show the importance of the smoothing provided by the mass matrix, which confirms the earlier conclusion of Griffiths *et al.* [6]. Without the mass matrix the iterative procedures used with both methods (implicit midpoint rule and the Delfour scheme) had difficulty in converging. Even when this did not cause the solution to become unbounded it influenced the conservation properties of the methods and caused a loss in accuracy.

Although the presence of the mass matrix does not allow the quantities (43a) and (43b) to be conserved theoretically, it did considerably improve the convergence properties of the iteration procedures. By using reasonably but not excessively small

space and time steps the improved convergence properties enabled the methods to conserve the two quantities fairly accurately and satisfactory numerical results were obtained.

We also pointed out the dangers of using too large space steps. Even if both quantities are conserved with high accuracy the numerical solution may still not resemble the analytical solution.

Because of the rapid change in the solution with time any efficient variable time step integrator should be particularly suitable for the solution of this problem. We found that the behaviour of the time step of method V did indeed follow the variation of the solution in time, provided an energy-conserving method was used to calculate the missing starting value. In addition, method V has the advantage of being explicit which means that it does not require a nonlinear system of algebraic equations to be solved at each time level. However, we found that the time step decreased to such an extent that most if not all of these advantages were lost.

In summary, the smoothing provided by the mass-matrix significantly improved the performance of our numerical methods, despite the fact that it does not allow the two quantities (43) to be conserved theoretically. It is essential that sufficiently small space and time steps should be used in order to resolve the steep spatial and temporal gradients. Neglecting to do this resulted not only in poorer approximations but in inaccurate and even unbounded solutions. Under the more stringent conditions of our numerical tests method V held no clear advantages over the implicit methods due to an excessively small time step. However, because of the potential advantages of explicit, energy-conserving variable time-step methods, more research in this direction needs to be done.

ACKNOWLEDGMENTS

This work was carried out while the first author was on sabbatical leave, at the University of Dundee, Scotland and Waterloo, Ontario, Canada. These visits were made possible by grants from the University of Orange Free State, CSIR, Pretoria, and NATO. The work of the second author was also partially supported by NSERC Research Grant A-3597. We are grateful to the reviewers for several valuable suggestions.

REFERENCES

1. M. J. ABLOWITZ, D. J. KAUP, A. C. NEWELL AND, S. SEGUR, *Stud. Appl. Math.* **53** (1974), 249.
2. I. CHRISTIE, D. F. GRIFFITHS, A. R. MITCHELL, AND J. M. SANZ-SERNA, *IMA J. Numer. Anal.* **1** (1981), 253.
3. M. DELFOUR, M. FORTIN, AND G. PAYRE, *J. Comput. Phys.* **44** (1981), 277.
4. R. K. DODD, J. C. EILBECK, J. D. GIBBON, AND H. C. MORRIS, "Solitons and Nonlinear Wave Equations," Academic Press, New York/London, 1982.
5. R. T. GLASSEY, *J. Math. Phys.* **18** (1977), 1794.
6. D. F. GRIFFITHS, A. R. MITCHELL, AND J. LL. MORRIS, *Comput. Meth. Appl. Mech. Eng.* **45** (1984), 177.

7. B. M. HERBST, A. R. MITCHELL, AND J. A. C. WEIDEMAN, *J. Comput. Phys.* **60** (1985).
8. J. W. MILES, *SIAM J. Appl. Math.* **41** (1981), 227.
9. A. R. MITCHELL AND J. LL. MORRIS, *Arab Gulf Sci. Res.* **1** (1983), 461.
10. D. H. PEREGRINE, *J. Aust. Math. Soc. Ser. B* **25** (1983), 16.
11. J. M. SANZ-SERNA, *J. Comput. Phys.* **47** (1982), 199.
12. J. M. SANZ-SERNA AND V. S. MANORANJAN, *J. Comput. Phys.* **52** (1983), 273.
13. W. A. STRAUSS, *in* "Contemporary Developments in Continuum Mechanics and Partial Differential Equations" (G. M. de la Penha and L. A. Medeiros, Eds.), North-Holland, New York, 1978.
14. G. B. WHITHAM, "Linear and Nonlinear Waves," Wiley, New York, 1974.
15. H. C. YUEN AND W. E. FERGUSON, *Phys. Fluids* **21** (1978), 1275.
16. V. E. ZAKHAROV AND A. B. SHABAT, *Sov. Phys. JETP* (*Engl. Trans.*) **34** (1972), 62.