*Journal of Computational Analysis and Applications, Vol. 4, No. 1, January 2002 (© 2002)*

# Fourier–Galerkin Method for Localized Solutions of Equations with Cubic Nonlinearity

**M. A. Christou[1] and C. I. Christov[1,2]**

Using a complete orthonormal system of functions in $L^2(-\infty, \infty)$ a Fourier–Galerkin spectral technique is developed for computing of the localized solutions of equations with cubic nonlinearity. A formula expressing the triple product into series in the system is derived. Iterative algorithm implementing the spectral method is developed and tested on the soliton problem for the cubic Boussinesq equation. Solution is obtained and shown to compare quantitatively very well to the known analytical one. The issues of convergence rate and truncation error are discussed.

**KEY WORDS:** Spectral methods; Galerkin approximation; cubic Boussinesq equation; localized solutions; solitons.

## 1. INTRODUCTION

In recent years a number of physical problems have frequently led to boundary value problems in infinite domains. These are the cases when no boundary conditions are specified at certain points, but rather the solution is required to possess a summable square in the infinite domain. Then the solution is said to belong to the $L^2(-\infty, \infty)$ space. A typical example is furnished by the problem of soliton solutions of different nonlinear evolution equations or generalized wave equations. In difference or FEM numerical solutions to the problems in $L^2(-\infty, \infty)$ a lot of difficulties appears. It suffices to mention the inevitable reducing of the infinite interval to a finite one which introduces artificial eigenvalue problems the latter being irrelevant to the original infinite domain. It can happen that each of the finite-domain approximations has only a trivial solution, while the original problem possesses a nontrivial one or vice versa. Sometimes, the finite-domain problem has a solution only at some denumerable set of intervals of specific length.

[1] Department. of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504-1010.
[2] To whom all correspondence should be addressed.



0272-4324/02/0100-1063$21.50/0 

These difficulties can be surmounted if a spectral method is used with basis system of localized functions which automatically acknowledge the requirement of $L^2(-\infty, \infty)$. We make use here of a system (see, [7]) which possesses the required properties.

From the known spectral techniques we choose the Galerkin method (for other techniques, see e.g., Canuto *et al.* [6]). The Galerkin method has the advantage of simplicity in implementation in comparison with the spectral collocation method or tau-method. This turns out to be crucial for the construction of fast and efficient numerical algorithms. The only problem is that Galerkin techniques require explicit formulas expressing the products of members of the CON system into series with respect to the system. For instance, the Hermite functions and Laguerre functions do not possess that kind of explicit relation. The first systems for which a product formula does exist was proposed in [7]. A Galerkin technique based on the said system was developed in [9] and applied to KdV and KS equations with quadratic nonlinearity. It has been recently applied to 2-D problems in [8]. In a sequence of papers Boyd [3,5] showed a general way of constructing CON systems in $L^2(-\infty, \infty)$ by means of coordinate transformation to finite interval and use of Chebyshev polynomials (see [4] for references).

In order to apply a Galerkin technique to a problem with cubic nonlinearity one is to derive the formulas expressing the triple products of the basis functions into series in the system. This is the purpose of the present work.

## 2. POSING THE PROBLEM

Consider the Boussinesq equation with cubic nonlinearity

$$v_{tt} = (v - v^3 - v_{xx})_{xx}, \tag{1}$$

which arises in many different mechanical models, such as, surface water waves, elastic waves propagating over strings, beams, or rods, etc. Equation (1) is a generalized wave equation containing nonlinearity and dispersion (the fourth-order derivative). Boussinesq [1,2] showed that for this kind of equations a balance can be struck between the nonlinearity and dispersion which manifests itself through the existence of a permanent localized solution (solitary wave, soliton, etc.) which propagates with a constant phase speed in quite the same fashion as the linear waves of the hyperbolic equations do. This balance holds true also for the evolution equations (e.g., the KdV Equation [11]) which can be derived in a coordinate frame moving

with the characteristic speed of the wave equation. The solitary-wave solution is subject to the so-called asymptotic boundary conditions

$$v(t, x) \to 0, \qquad \text{for } x \to \pm\infty. \tag{2}$$

These conditions stem from the requirement for finite energy of the wave over the infinite interval, namely

$$\int_{-\infty}^{\infty} u^2(t, x)\, dx < +\infty,$$

which is nothing less than the definition of $L^2(-\infty, \infty)$ space.

Now we introduce new independent variable $\xi = x - ct$ and denote $u(\xi) = v(t, x)$. Then the following equation for $u$ is derived

$$c^2 u' = u'' - (u^3)'' - u''''. \tag{3}$$

Upon integrating twice and acknowledging the asymptotic boundary conditions (2) in the form $\xi \to \pm\infty$ $u(\xi) \to 0$, we get

$$(1 - c^2)u - u^3 - u'' = 0. \tag{4}$$

The trivial solution $u \equiv 0$ is always present, hence (4), (2) is a bifurcation problem. This requires a special care when devising an iterative algorithm and will be addressed in detail in the due place.

It is readily shown that if $u(\xi)$ is a solution to our problem, then $u(-\xi)$ is a solution too. For this reason we may consider only even functions as solutions.

For $|c| < 1$, Eq. (4) possesses analytical localized solution

$$u(\xi) = \frac{\sqrt{2(1 - c^2)}}{\cosh(\xi\sqrt{1 - c^2})}, \tag{5}$$

which is important for our purposes providing the necessary check for the consistency and accuracy of the spectral technique developed. For definiteness we select $c = 0.8$. The solution for the other values of $c$ can be obtained via re-scaling.

Before proceeding to the description of the numerical method we mention that scaling the independent variable does not change the nature of

the boundary value problem in $L^2$. Then upon introducing $\xi = \beta\eta$ we render (4) to the following boundary value problem

$$(1 - c^2)u - u^3 - \frac{1}{\beta^2}u'' = 0\,, \qquad \int_{-\infty}^{\infty} u^2(\eta)\,d\eta < +\infty, \tag{6}$$

where the primes stand for differentiation with respect to $\eta$. The additional parameter $\beta$ is crucial for the optimization of the method. It's introduction allows one to bring in concert the typical length scales of the employed system of functions and that of the support of the sought localized solution.

## 3. THE FOURIER–GALERKIN METHOD IN $L^2(-\infty, \infty)$

### 3.1. The Complete Orthonormal (CON) system

The system

$$\rho_n = \frac{1}{\sqrt{\pi}}\frac{(ix-1)^n}{(ix+1)^{n+1}}, \qquad n = 0, 1, 2, \ldots \tag{7}$$

was introduced by Wiener [12] as a Fourier transformation of the Laguerre functions. Higgins [10] defined it also for negative indices $n$ and proved its completeness and orthogonality. The significance of (7) for nonlinear problems was demonstrated in [7], where the product formula was derived

$$\rho_n\rho_k = \frac{\rho_{n+k} - \rho_{n-k}}{2\sqrt{\pi}} \tag{8}$$

and the two real-valued subsequences of odd functions $S_n$ and even functions $C_n$ were introduced, namely

$$S_n = \frac{\rho_n + \rho_{-n-1}}{i\sqrt{2}}, \qquad C_n = \frac{\rho_n - \rho_{-n-1}}{\sqrt{2}}. \tag{9}$$

Using (8) one easily shows that the products of members of the real-valued sequences are expanded in series with respect to the system as follows (see, [7]):

$$C_n C_k = \frac{1}{2\sqrt{2\pi}}[\ C_{n+k+1} - C_{n+k} - C_{n-k} + C_{n-k-1}], \tag{10}$$

$$S_n S_k = \frac{1}{2\sqrt{2\pi}}[\ C_{n+k+1} - C_{n+k} + C_{n-k} - C_{n-k-1}], \tag{11}$$

$$S_n C_k = \frac{1}{2\sqrt{2\pi}}[-S_{n+k+1} + S_{n+k} + S_{n-k} - S_{n-k-1}]. \tag{12}$$

For the second derivatives of the basis functions one has (see [7])

$$C_n'' = \sum_{m=0}^{\infty} \chi_{m,n} C_m\ , \qquad S_n'' = \sum_{m=0}^{\infty} \chi_{m,n} S_m,$$

where

$$\chi_{m,n} = -\frac{n(n-1)}{4}\delta_{m,n-2} + n^2\delta_{m,n-1} - \frac{(n+1)(n+2)}{4}\delta_{m,n+2}$$
$$-\frac{n^2 + (2n+1)^2 + (n+1)^2}{4}\delta_{m,n} + (n+1)^2\delta_{m,n+1}. \tag{13}$$

### 3.2. The Triple-Product Formula

The main purpose of the present work is to develop further the Galerkin technique with application to problems with cubic nonlinearity. For this reason we derive a formula expressing the triple product of functions of the system into a series with respect to the system. In deriving these formulas we follow [7] and make repeated use of (10), (11), and (12). For the triple product of the functions $C_n$ from the even subsequence we obtain

$$C_l C_n C_k = \frac{1}{8\pi}[C_{l+n+k+2} - 2C_{l+n+k+1} + C_{l+n+k} - 2C_{l-n-k-1} + C_{l-n-k}$$
$$- C_{n-k+l+1} - C_{l-n+k+1} + 2C_{n-k+l} + 2C_{l-n+k} - C_{l-n+k-1}$$
$$- C_{n-k+l-1} + C_{l-n-k-2}], \tag{14}$$

where the indices are allowed to assume negative values. For the purposes of the computer implementation of the method one needs the same formula expressed only through functions with positive indices. Making use of the

relations from [7]: $C_{-n} = -C_{n-1}$, $S_{-n} = S_{n-1}$, after some manipulations one gets

$$C_l C_n C_k \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \beta_{lnk,m} C_m(x),$$

$$\begin{aligned}
\beta_{lnk,m} = {} & \delta_{m,n+k+l+2} - 2\delta_{m,n+k+l+1} + \delta_{m,n+k+l} \\
& - 2\ \mathrm{sgn}(l-n-k-0.5)\delta_{m,[|l-n-k-0.5|]} \\
& - \mathrm{sgn}(l-n+k-0.5)\delta_{m,[|l-n+k-0.5|]} \\
& - \mathrm{sgn}(l+n-k-0.5)\delta_{m,[|l+n-k-0.5|]} \\
& + \mathrm{sgn}(l-n-k-1.5)\delta_{m,[|l-n-k-1.5|]} \\
& + \mathrm{sgn}(l-n-k+0.5)\delta_{m,[|l-n-k+0.5|]} \\
& + 2\ \mathrm{sgn}(l+n-k+0.5)\delta_{m,[|l+n-k+0.5|]} \\
& + 2\ \mathrm{sgn}(l-n+k+0.5)\delta_{m,[|l-n+k+0.5|]} \\
& - \mathrm{sgn}(l+n-k+1.5)\delta_{m,[|l+n-k+1.5|]} \\
& - \mathrm{sgn}(l-n+k+1.5)\delta_{m,[|l-n+k+1.5|]},
\end{aligned}$$

where $[a]$ stands for the biggest integer number smaller than $a$ and $\delta$ is the Kroneker delta-function. In a similar fashion:

$$S_l S_n S_k \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \alpha_{lnk,m} S_m(x),$$

$$\begin{aligned}
\alpha_{lnk,m} = {} & -\delta_{m,n+k+l+2} + 2\delta_{m,n+k+l+1} - \delta_{m,n+k+l} + 2\delta_{m,[|l-n-k-0.5|]} \\
& - \delta_{m,[|l-n+k-0.5|]} - \delta_{m,[|l+n-k-0.5|]} - \delta_{m,[|l-n-k-1.5|]} \\
& - \delta_{m,[|l-n-k+0.5|]} + 2\delta_{m,[|l+n-k+0.5|]} + 2\delta_{m,[|l-n+k+0.5|]} \\
& - \delta_{m,[|l+n-k+1.5|]} - \delta_{m,[|l-n+k+1.5|]},
\end{aligned}$$

$$S_l S_n C_k \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \bar{\gamma}_{lnk,m} C_m(x),$$

$$\begin{aligned}
\bar{\gamma}_{lnk,m} = & -\delta_{m,n+k+l+2} + 2\delta_{m,n+k+l+1} - \delta_{m,n+k+l} \\
& - 2\,\mathrm{sgn}(l-n-k-0.5)\delta_{m,[|l-n-k-0.5|]} \\
& - \mathrm{sgn}(l-n+k-0.5)\delta_{m,[|l-n+k-0.5|]} \\
& + \mathrm{sgn}(l+n-k-0.5)\delta_{m,[|l+n-k-0.5|]} \\
& + \mathrm{sgn}(l-n-k-1.5)\delta_{m,[|l-n-k-1.5|]} \\
& + \mathrm{sgn}(l-n-k+0.5)\delta_{m,[|l-n-k+0.5|]} \\
& - 2\,\mathrm{sgn}(l+n-k+0.5)\delta_{m,[|l+n-k+0.5|]} \\
& + 2\,\mathrm{sgn}(l-n+k+0.5)\delta_{m,[|l-n+k+0.5|]} \\
& + \mathrm{sgn}(l+n-k+1.5)\delta_{m,[|l+n-k+1.5|]} \\
& - \mathrm{sgn}(l-n+k+1.5)\delta_{m,[|l-n+k+1.5|]},
\end{aligned}$$

$$S_l C_n C_k \stackrel{\text{def}}{=} \sum_{m=0}^{\infty} \bar{\bar{\gamma}}_{lnk,m} S_m(x),$$

$$\begin{aligned}
\bar{\bar{\gamma}}_{lnk,m} = & -\delta_{m,n+k+l+2} + 2\delta_{m,n+k+l+1} - \delta_{m,n+k+l} + 2\delta_{m,[|l-n-k-0.5|]} \\
& + \delta_{m,[|l-n+k-0.5|]} + \delta_{m,[|l+n-k-0.5|]} - \delta_{m,[|l-n-k-1.5|]} \\
& - \delta_{m,[|l-n-k+0.5|]} - 2\delta_{m,[|l+n-k+0.5|]} - 2\delta_{m,[|l-n+k+0.5|]} \\
& + \delta_{m,[|l+n-k+1.5|]} + \delta_{m,[|l-n+k+1.5|]}.
\end{aligned}$$

Now we are equipped to handle problems with cubic nonlinearities.

### 3.3. Algebraic System for the Coefficients

As already mentioned above our problem admits even functions as solutions and hence we develop the sought solution $u$ into series with respect to the subsequence of functions $C_n$ only, namely

$$u(\eta) = \sum_{n=0}^{\infty} a_n C_n(\eta). \tag{15}$$

Then

$$u^3(\eta) = \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \sum_{m_3=0}^{\infty} \sum_{n=0}^{\infty} a_{m_1} a_{m_2} a_{m_3} \beta_{m_1 m_2 m_3, n} C_n(\eta), \tag{16}$$

$$u''(\eta) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} a_m \chi_{m,n} C_n(\eta). \tag{17}$$

Since for the Galerkin method the sets of trial and test functions coincide with the set $C_n$, then upon introducing (15), (16), and (17) into (4), combining the terms with the like functions $C_n$, and taking the respective coefficients to be equal to zero (due to the independence of members of subsequence $C_n$ and its completeness in the subspace of even functions in $L^2(-\infty, \infty)$), we obtain the following infinite nonlinear algebraic system for the unknown coefficients $a_n$:

$$\sum_{m=0}^{\infty} [-\chi_{m,n} + (1 - c^2)\delta_{m,n}] a_m - \sum_{m_1=0}^{\infty} \sum_{m_2=0}^{\infty} \sum_{m_3=0}^{\infty} a_{m_1} a_{m_2} a_{m_3} \beta_{m_1 m_2 m_3 n} = 0. \tag{18}$$

In the numerical calculations a truncated version of the above system is used where the infinity is replaced by $N$. The problems of the truncation error are addressed in the next section.

## 4. ALGORITHM

In order to avoid the trivial solution we consider the rescaled vector of unknowns $a_i = \sqrt{\alpha}\hat{a}_i$ (where $\alpha$ is unknown parameter) and impose a condition on the first coefficient, namely we set $\hat{a}_0 = 1$.

If the series converge the first coefficient is supposed to be the greatest or at least not much smaller than the second or third one. The coefficients with larger numbers are much smaller. This allows one to impose the condition on the first one.

After $\hat{a}_0$ is set to be equal to unity the system becomes overposed. In order to avoid this problem the equation for $\hat{a}_0$ should not be used any more. It becomes an equation for determination of $\alpha$, namely

$$\alpha = \frac{(1-c^2)\hat{a}_0 - \sum_{m=0}^{N} \hat{a}_m \chi_{m,0}}{\vartheta},$$

$$\vartheta \stackrel{\text{def}}{=} \sum_{m_1=0}^{N} \sum_{m_2=0}^{N} \sum_{m_3=0}^{N} \hat{a}_{m_1} \hat{a}_{m_2} \hat{a}_{m_3} \beta_{m_1 m_2 m_3, 0}.$$

Note that the above procedure is valid only when the expected solution has non-zero amplitude in the origin of the coordinate system. When this is not the case (i.e for solutions that are odd functions), one can impose a similar condition on one of the coefficients before one of the lower-order odd functions (say, $S_0$ or $S_1$).

The nonlinear system is solved by means of so-called "simple iteration". When the nonlinear term is evaluated the coefficients are treated as known quantities from the previous iteration, namely $\hat{a}_i^n$ and then the linear system with pentadiagonal matrix is solved for the "new" iteration designated by $\tilde{a}_i$, $\tilde{\alpha}$:

$$\sum_{m=1}^{N} [-\chi_{m,n} + (1-c^2)\delta_{m,n}]\tilde{a}_m - \alpha^n \sum_{m_1=0}^{N} \sum_{m_2=0}^{N} \sum_{m_3=0}^{N} \hat{a}^n_{m_1} \hat{a}^n_{m_2} \hat{a}^n_{m_3} \beta_{m_1 m_2 m_3, k} = 0, \tag{19}$$

$$\tilde{\alpha} = \frac{1}{\vartheta}\left[(1-c^2)\hat{a}^n_0 - \sum_{m=0}^{N} \hat{a}^n_m \chi_{m,0}\right],$$

$$\vartheta = \sum_{m_1=0}^{N} \sum_{m_2=0}^{N} \sum_{m_3=0}^{N} \hat{a}^n_{m_1} \hat{a}^n_{m_2} \hat{a}^n_{m_3} \beta_{m_1 m_2 m_3, 0}. \tag{20}$$

In order to control the rate of convergence and to avoid divergence when the process is started from a very rough initial approximation a relaxation is applied according to the ubiquitous formulas

$$\hat{a}_i^{n+1} = \omega \tilde{a}_i + (1-\omega)\hat{a}_i^n, \qquad \alpha^{n+1} = \omega \tilde{\alpha} + (1-\omega),$$

**Table I.** Dependence of the Solution on $N$ for Scaling Parameter $\beta = 1$

| $N$ | 0 | 9 | 19 | 29 | 39 |
|---|---|---|---|---|---|
| $a_0$ | 0.14701e + 01 | 0.14549e + 01 | 0.14545e + 01 | 0.14545e + 01 | 0.14545e + 01 |
| $a_1$ | | 0.50280e + 00 | 0.50458e + 00 | 0.50457e + 00 | 0.50457e + 00 |
| $a_2$ | | 0.11926e + 00 | 0.12202e + 00 | 0.11999e + 00 | 0.11999e + 00 |
| $a_3$ | | −0.20383e − 01 | −0.22042e − 01 | −0.22107e − 01 | −0.22107e − 01 |
| $a_4$ | | −0.58713e − 01 | −0.63500e − 01 | −0.63590e − 01 | −0.63592e − 01 |
| $a_5$ | | −0.57651e − 01 | −0.65849e − 01 | −0.65949e − 01 | −0.65954e − 01 |
| $a_6$ | | −0.43642e − 01 | −0.55118e − 01 | −0.55207e − 01 | −0.55215e − 01 |
| $a_7$ | | −0.27684e − 01 | −0.41831e − 01 | −0.41884e − 01 | −0.41985e − 01 |
| $a_8$ | | −0.14176e − 01 | −0.29851e − 01 | −0.29838e − 01 | −0.29851e − 01 |
| $a_9$ | | −0.47649e − 02 | −0.20269e − 01 | −0.20157e − 01 | −0.20171e − 01 |
| $a_{10}$ | | | −0.13094e − 01 | −0.12849e − 01 | −0.12864e − 01 |
| $a_{11}$ | | | −0.79698e − 02 | −0.75614e − 02 | −0.75744e − 02 |
| $a_{12}$ | | | −0.44667e − 02 | −0.38665e − 02 | −0.38758e − 02 |
| $a_{13}$ | | | −0.21894e − 02 | −0.13770e − 02 | −0.13799e − 02 |
| $a_{14}$ | | | −0.80827e − 03 | 0.22662e − 03 | 0.23300e − 03 |
| $a_{15}$ | | | −0.60053e − 04 | 0.11943e − 02 | 0.12132e − 02 |
| $a_{16}$ | | | 0.26120e − 03 | 0.17157e − 02 | 0.17503e − 02 |
| $a_{17}$ | | | 0.31565e − 03 | 0.19316e − 02 | 0.19853e − 02 |
| $a_{18}$ | | | 0.22816e − 03 | 0.19451e − 02 | 0.20208e − 02 |
| $a_{19}$ | | | 0.96939e − 04 | 0.18303e − 02 | 0.19308e − 02 |
| $a_{20}$ | | | | 0.16402e − 02 | 0.17675e − 02 |
| $a_{21}$ | | | | 0.14119e − 02 | 0.15671e − 02 |
| $a_{22}$ | | | | 0.11707e − 02 | 0.13540e − 02 |
| $a_{23}$ | | | | 0.93418e − 03 | 0.11441e − 02 |
| $a_{24}$ | | | | 0.71365e − 03 | 0.94742e − 03 |
| $a_{25}$ | | | | 0.51662e − 03 | 0.76939e − 03 |
| $a_{26}$ | | | | 0.34784e − 03 | 0.61272e − 03 |
| $a_{27}$ | | | | 0.21027e − 03 | 0.47807e − 03 |
| $a_{28}$ | | | | 0.10574e − 03 | 0.36481e − 03 |
| $a_{29}$ | | | | 0.35410e − 04 | 0.27152e − 03 |
| $a_{30}$ | | | | | 0.19629e − 03 |
| $a_{31}$ | | | | | 0.13705e − 03 |
| $a_{32}$ | | | | | 0.91636e − 04 |
| $a_{33}$ | | | | | 0.57935e − 04 |
| $a_{34}$ | | | | | 0.33946e − 04 |
| $a_{35}$ | | | | | 0.17803e − 04 |
| $a_{36}$ | | | | | 0.77946e − 05 |
| $a_{37}$ | | | | | 0.23577e − 05 |
| $a_{38}$ | | | | | 0.78492e − 07 |
| $a_{39}$ | | | | | −0.32213e − 06 |

**Table II.** Dependence of the Solution on $N$ for Scaling Parameter $\beta = 4$

| $N$ | 0 | 4 | 9 | 19 | 29 |
|---|---|---|---|---|---|
| $a_0$ | 0.99162e + 00 | 0.71111e + 00 | 0.71111e + 00 | 0.71111e + 00 | 0.71111e + 00 |
| $a_1$ | | −0.30480e + 00 | −0.30484e + 00 | −0.30484e + 00 | −0.30484e + 00 |
| $a_2$ | | 0.36126e − 01 | 0.36111e − 01 | 0.36112e − 01 | 0.36112e − 01 |
| $a_3$ | | −0.70547e − 02 | −0.67024e − 02 | −0.67002e − 02 | −0.67002e − 02 |
| $a_4$ | | 0.45251e − 02 | 0.54480e − 02 | 0.54518e − 02 | 0.54518e − 02 |
| $a_5$ | | | 0.14133e − 02 | 0.14162e − 02 | 0.14162e − 02 |
| $a_6$ | | | 0.79325e − 03 | 0.78845e − 03 | 0.78845e − 03 |
| $a_7$ | | | 0.14238e − 03 | 0.12030e − 03 | 0.12030e − 03 |
| $a_8$ | | | −0.14838e − 04 | −0.63290e − 04 | −0.63304e − 04 |
| $a_9$ | | | −0.32585e − 04 | −0.10855e − 03 | −0.10858e − 03 |
| $a_{10}$ | | | | −0.88158e − 04 | −0.88208e − 04 |
| $a_{11}$ | | | | −0.56755e − 04 | −0.56817e − 04 |
| $a_{12}$ | | | | −0.30265e − 04 | −0.30319e − 04 |
| $a_{13}$ | | | | −0.12746e − 04 | −0.12751e − 04 |
| $a_{14}$ | | | | −0.29156e − 05 | −0.28101e − 05 |
| $a_{15}$ | | | | 0.15989e − 05 | 0.18900e − 05 |
| $a_{16}$ | | | | 0.29328e − 05 | 0.34857e − 05 |
| $a_{17}$ | | | | 0.26195e − 05 | 0.34859e − 05 |
| $a_{18}$ | | | | 0.16519e − 05 | 0.28220e − 05 |
| $a_{19}$ | | | | 0.64720e − 06 | 0.20037e − 05 |
| $a_{20}$ | | | | | 0.12708e − 05 |
| $a_{21}$ | | | | | 0.70858e − 06 |
| $a_{22}$ | | | | | 0.32270e − 06 |
| $a_{23}$ | | | | | 0.85139e − 07 |
| $a_{24}$ | | | | | −0.41622e − 07 |
| $a_{25}$ | | | | | −0.92850e − 07 |
| $a_{26}$ | | | | | −0.97255e − 07 |
| $a_{27}$ | | | | | −0.76460e − 07 |
| $a_{28}$ | | | | | −0.46159e − 07 |
| $a_{29}$ | | | | | −0.17779e − 07 |

with $0 < \omega < 1$. For the different cases treated in the recent work we have attained convergence for $0.1 < \omega < 1$.

## 5. RESULTS AND VERIFICATIONS

We consider here two different values of the scaling parameter, namely $\beta = 1$ and $\beta = 4$. The former corresponds, in fact, to absence of any scaling. The latter is roughly proportional to the ratio between the supports of the analytical solution (5), and of the function $C_0$.

### 5.1. Convergence with the Number of Terms $N$

When no *a priori* information is available concerning the length scale of the support of the sought localized solution it is only natural to begin with

$\beta = 1$. We have solved the algebraic problem in this case with different number of terms in the truncated series. Results are presented in Table I. The Fourier–Galerkin method demonstrates an efficiency well beyond the expectations. Even the solution with only one term ($N = 0$) is within 1% of the solution for $a_0$ with 40 terms ($N = 39$). It means that a substantial part of the energy of the solitary wave can be predicted with a truncated series containing merely one term.

The dependence of the higher-order coefficients on the total number of terms $N + 1$ is qualitatively similar to $\hat{a}_0$. Naturally, the smaller the modulus of a coefficient is, the larger the relative error. Yet, for $a_3$ which is 8 times smaller than $a_0$ the difference between the solutions with 10 and 40 terms is only 7.5%.

The results for larger $N$ as shown in Table I testify to an exponential decay with the increase of $i$. A fairly good fit is given by the relation $|a_i| \approx 0.2 \exp(-0.23i)$. For $\beta = 1$ the influence of the truncation is "felt" approximately 7–8 terms before the last one, i.e., for $N = 39$ the coefficients deviate appreciably from the above exponential law for $i \geq 32$.

As shown in Table II the convergence for scale factor $\beta = 4$ is much faster and the adverse effects of the truncation are "felt" only 4–5 terms before the last one. In addition these terms are of much smaller magnitude so that they virtually do not affect the solution more than on 0.01%. The trade-off is that the one-term solution is farther from the accurate one (some 18%). But this is in fact more natural since nobody expects a good quantitative result for a nonlinear problem with just one term in the Fourier–Galerkin series. Now, a best fit for the dependence of the coefficient on its number is $|a_i| \approx 0.006 \exp(-0.41i)$ which shows much faster convergence (almost twice as large an exponent than that for $\beta = 1$).

### 5.2. Truncation Error

The most important characteristics of a spectral method is the truncation error which allows one to estimate the accuracy with which the solution can be obtained. When an analytical reference solution is available (which is the case of the present paper) then the truncation error is calculated as the difference between the truncated spectral solution and the analytical one. When an analytical solution is not available one has to calculate the solution with very large number of terms and to use these calculations as a reference.

In Fig. 1 we present the pointwise truncation error for the case $\beta = 1$. One sees that the error is of order of $10^{-5}$. This outlines the accuracy of the truncated solution with number of terms $N + 1 = 40$. For better accuracy one is to use twice as many terms. But then another obstacle arises

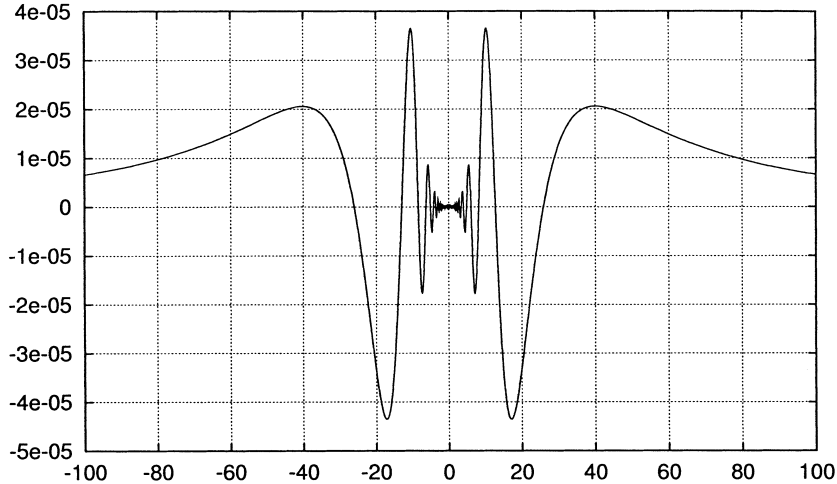


**Fig. 1.** Pointwise truncation error for scale factor $\beta = 1$ and $N = 39$.

connected with the accumulation of round-off error when the function $u(\eta)$ is evaluated from the series (15). This means that finding the optimal $\beta$ has not only theoretical significance, but a profound impact on the practical results as well.

The pointwise truncation error for $\beta = 4$ as presented in Fig. 2 is of three orders of magnitudes smaller than the error for $\beta = 1$. One sees in the figure that, in fact, the truncation error is smaller than the round-off error which is the cause of the spurious behavior in the interval $\eta \in [-20, 20]$. It is interesting to mention here that the round-off error is of order of $10^{-8}$ although the calculations with double precision are supposed to be accurate at least down to $10^{-12}$. As mentioned in the above the larger round-off error for $u(\eta)$ is due to the fact that at each point the latter is calculated as a sum of $N + 1$ functional values each of which can have an error of order of $10^{-12}$ and even larger since $C_n$ are calculated through summation of power series. Thus we have reached the limit for accuracy imposed by the finite representation of the numbers in the computer.

## 6. CONCLUSIONS

In the present paper Fourier–Galerkin spectral technique is developed for calculating localized solutions of equations with cubic nonlinearity. A complete orthonormal (CON) basis system earlier proposed in the authors

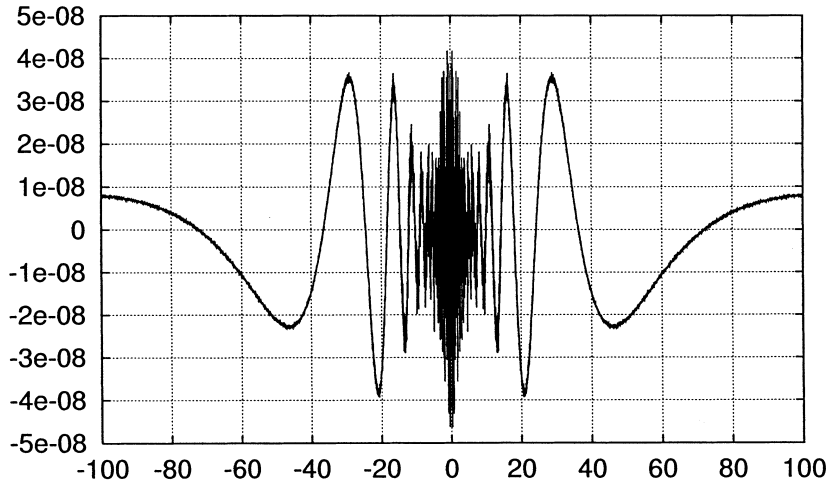


**Fig. 2.** Pointwise truncation error for scale factor $\beta = 4$ and $N = 29$.

works is used as a basis system. New formulas expressing the triple products of the basis functions are derived. These formulas are crucial for the Galerkin approach. A scaling parameter is introduced which allows fine-tuning and optimization of the technique proposed.

Iterative algorithm is devised for solving the truncated version of the algebraic system for the coefficients of the Galerkin series. Numerical experiments with different number of terms are conducted for two different values of the scaling parameter. These experiments establish the practical convergence of the method and indicate an exponential decay of the coefficients with the increase of their number (exponential convergence).

The global truncation error is assessed via comparison to the known soliton solution of the cubic Boussinesq equation. It turns out that the error is of order of $10^{-8}$ when the series are truncated after the 30th which shows the efficiency of the proposed technique. It is shown that a reasonable approximation can be obtained even with as little as 5–10 terms in the truncated series.

## ACKNOWLEDGMENT

This work is supported by Grant LEQSF(1999–2002)-RD-A-49 from the Louisiana Board of Regents.

## REFERENCES


1. J. V. Boussinesq, Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire, *Comp. Rend. Hebd. des Seances de l'Acad. des Sci.* 72, 755–759 (1871).
2. J. V. Boussinesq, Théorie des ondes et des remous qui se propagent le long d'un canal rectangulaire horizontal, en communiquant au liquide contenu dans ce canal des vitesses sensiblement pareilles de la surface au fond, *Journal de Mathématiques Pures et Appliquées* 17, 55–108 (1872).
3. J. P. Boyd, Spectral methods using rational basis on an infinite interval. *J. Comp. Phys.* 69, 112–142 (1987).
4. J. P. Boyd, *Spectral Methods*, Springer-Verlag, New York, 1989.
5. J. P. Boyd, The orthogonal rational functions of Higgins and Christov and algebraically mapped Chebyshev polynomials. *J. Approx. Theory* 61, 98–105 (1990).
6. C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer-Verlag, New York, 1988.
7. C. I. Christov, A complete orthonormal sequence of functions in $L^2(-\infty, \infty)$ space, *SIAM J. Appl. Math.* 42, 1337–1344 (1982).
8. C. I. Christov, Fourier–Galerkin algorithm for 2-D localized solutions. *Annuaire de l'Univ. Sof., Fac. de Mathématiques et Informatique*, 95 (lb.2-Mathematiques Appliquée et Informatique), 169–179 (1995).
9. C. I. Christov and K. L. Bekyarov, A Fourier-series method for solving soliton problems, *SIAM J. Sci. Stat. Comp.* 11, 631–647 (1990).
10. J. R. Higgins, *Completeness and basis properties of sets of special functions,* Cambridge University Press, London, 1977.
11. D. J. Korteweg and G. de Vries, On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves, *Phil. Mag.* ser.5, 39, 422–443 (1895).
12. N. Wiener, *Extrapolation, Interpolation and smoothing of stationary time series,* Technology Press MIT and John Wiley, New York, 1949.