PSEUDOSPECTRAL METHODS FOR THE BENJAMIN-ONO EQUATION

R.L. James and J.A.C. Weideman
College of Oceanography / Department of Mathematics
Oregon State University
Corvallis OR 97331, USA
e-mail: rodney@oce.orst.edu / weideman@math.orst.edu

Abstract. We propose two pseudospectral methods for solving the Benjamin-Ono equation. One is based on Fourier series, the other on rational functions. We discuss various issues such as the implementation of these methods and their conservation properties, and we compare them numerically.

1. Introduction. The Benjamin-Ono (BO) equation models internal waves in stratified fluids (see [2], [8]). It is defined by

\[ u_t + u u_x + \mathcal{H}(u_x^2) = 0, \quad -\infty < x < \infty, \tag{1} \]

together with initial and boundary conditions given as

\[ u(x, 0) = u_0(x), \quad \text{and} \quad u, u_x \to 0 \quad \text{as} \quad |x| \to \infty. \]

The third term in (1) involves the Hilbert transform, defined as the principal value integral

\[ \mathcal{H}(f)(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(y)}{y-x} \, dy. \tag{2} \]

The BO equation is closely related to the more famous Korteweg-de Vries (KdV) equation, in which \( u_x x \) replaces the Hilbert transform term in (1). As discussed in [1], for example, both equations may be viewed as infinite dimensional integrable Hamiltonian systems, with an infinite number of constants of motion. For (1), three of these constants are given by

\[ I_1 = \int_{-\infty}^{\infty} u \, dx, \quad I_2 = \int_{-\infty}^{\infty} u^2 \, dx, \quad \text{and} \quad I_3 = \int_{-\infty}^{\infty} (u^3 - 3u u \mathcal{H}(u)) \, dx; \tag{3} \]

the third one being the Hamiltonian. Moreover, both the BO and KdV equations are solvable by the method of inverse scattering and both admit soliton solutions. One major distinction between them is that KdV solitons have the profile of a hyperbolic secant function whereas BO solitons are rational functions in \( x \) and \( t \). For example, one- and two-soliton solutions of (1) are given respectively by

\[ u(x,t) = \frac{4c}{c^2(x-ct)^2 + 1}, \quad \text{and} \quad u(x,t) = \frac{4c_1 c_2 \left( c_1 \theta_1^2 + c_2 \theta_2^2 + (c_1 + c_2) \theta_1^{-1} \theta_2^{-1} (c_1 - c_2)^{-2} \right)}{\left( c_1 c_2 \theta_1^2 - (c_1 + c_2) \theta_2^2 (c_1 - c_2)^{-2} \right)^2 + (c_1 \theta_1 + c_2 \theta_2)^2}, \tag{4} \]

where \( \theta_1 = x - c_1 t - \phi_1, \theta_2 = x - c_2 t - \phi_2, \) and \( c, c_1, c_2, \phi_1, \phi_2 \) are arbitrary constants (see [9]).

Despite the fact that closed form solutions like these may be generated by analytical means, finding the solution for arbitrary initial data \( u_0(x) \) is no easy task. Numerical methods are therefore important for studying the dynamics of (1). In this paper we propose two such methods.

Solving (1) numerically poses several challenges. First, one needs a method for dealing with the singular integral (2). Second, the fact that BO solutions are rational functions is troublesome for the following reason. Assuming the solution is negligible for \(|x|\) sufficiently large, the infinite domain is generally truncated to a finite one for computational work. But since rational solutions decay slowly as \(|x| \to \infty\), a very wide computational domain is required. For this reason BO computations are more challenging than KdV computations, in which the hyperbolic secant solitons decay rapidly as \(|x| \to \infty\).
When deciding on appropriate methods for solving PDEs, one is invariably confronted with the choice between finite difference, finite element, and spectral methods. For the BO equation we opted for spectral methods, for the following reasons. First, in [10] it was shown that spectral methods for the KdV equation are reliable. Second, the main advantage of finite difference and finite element methods is that they are local, leading to sparse matrices and in general a less expensive computation than with spectral methods. However, for (1) the integral term is global anyway, so this advantage of finite differences and elements is lost. Third, it is not clear how to treat the Hilbert transform with finite differences or finite elements. By contrast both of our methods deal naturally with the Hilbert transform; see (7) and (14) below.

2. A Fourier method. A Fourier based spectral-iterative method for (1) was proposed in [7]. Here we introduce a related but not equivalent method. It is based on the method-of-lines strategy, in which we first discretize the space variable in (1), leaving a semi-discrete system of ODEs to be integrated in time.

Suppressing the time dependence for now, we assume that \( u(x) \) is negligible for \( |x| > L \). We truncate the infinite line to \( x \in [-L, L] \), and introduce the uniform grid

\[
x_j = jh, \quad h = L/N, \quad j = -N, \ldots, N.
\]

Next we expand the data \( u_j = u(x_j) \) as a discrete Fourier series

\[
u_j = \sum_{n=-N}^{N-1} a_n e^{i \mu_n x_j}, \quad \text{where} \quad a_n = \frac{1}{2N} \sum_{j=-N}^{N-1} u_e^{-i \mu_n x_j} \quad \text{and} \quad \mu_n = n \pi / L.
\]

This leads to approximation formulas for \( u'(x_j), u''(x_j), \) and \( \mathcal{H}(u'')(x_j) \):

\[
u'_j = i \sum_{n=-N+1}^{N-1} \mu_n a_n e^{i \mu_n x_j}, \quad u''_j = -\sum_{n=-N}^{N-1} \mu_n^2 a_n e^{i \mu_n x_j}, \quad \text{and} \quad \mathcal{H}(u'')(x_j) = -i \sum_{n=-N+1}^{N-1} \text{sgn}(n) \mu_n^2 a_n e^{i \mu_n x_j}.
\]

The third approximation is motivated by the well-known fact that

\[
\mathcal{H}(e^{i \mu x}) = i \text{sgn}(\mu) e^{i \mu x}, \quad \mu \text{ real}.
\]

Note that in the first and third formulas in (6) we set \( a_{-N} = 0 \). This prevents the introduction of a spurious imaginary component for real data, as discussed in [3] for example.

It is convenient to present the above formulas in matrix notation. If we define

\[
u = (u_{-N}, \ldots, u_{N-1})^T \quad \text{and} \quad a = (a_{-N}, \ldots, a_{N-1})^T,
\]

then (5) may be summarized as

\[
u = \mathcal{F}^{-1} a \quad \text{and} \quad a = \mathcal{F} \nu,
\]

where \( \mathcal{F} \) is the Fourier matrix, and \( \mathcal{F}^{-1} \) its inverse, with entries

\[
\mathcal{F}_{nj} = \frac{1}{2N} e^{-i \mu_n x_j}, \quad \mathcal{F}_{jn}^{-1} = e^{i \mu_n x_j}, \quad -N \leq n, j \leq N - 1.
\]

The matrices representing the first and second derivatives and the Hilbert transform in Fourier space are given by

\[
\mathcal{E}_1 = i \text{diag}(0, \mu_{-N+1}, \ldots, \mu_{N-1}), \quad \mathcal{E}_1 = -\text{diag}(\mu_{-N+1}^2, \ldots, \mu_{N-1}^2), \quad \text{and} \quad \mathcal{E}_H = i \text{diag}(0, -1, \ldots, -1, 0, 1, \ldots, 1).
\]

Formulas (6) may therefore be summarized as

\[
u' = \mathcal{F}^{-1} \mathcal{E}_1 \mathcal{F} \nu, \quad u'' = \mathcal{F}^{-1} \mathcal{E}_2 \mathcal{F} \nu, \quad \text{and} \quad h = \mathcal{F}^{-1} \mathcal{E}_H \mathcal{E}_2 \mathcal{F} \nu.
\]
The entries of the matrix $F^{-1}E_1F$ can be found in closed form in [3], and those of $F^{-1}E_HE_2F$ were obtained in [6]. In practice the matrices are not formed explicitly however, instead the Fast Fourier Transform (FFT) is used to compute the matrix-vector products involving $F$ and $F^{-1}$ efficiently.

Defining $\mathcal{H} = \text{diag}(u_{-N}, \ldots, u_{-1})$, and $0$ as the zero vector, the semi-discrete version of (1) now becomes

$$\frac{du}{dt} + \mathcal{H}(F^{-1}E_1F)u + F^{-1}E_HF^{-1}E_2F = 0,$$

which is summarized as

$$\frac{du}{dt} + f(u) = 0.$$  (10)

Naturally we would like our scheme to capture the dynamics of (1) accurately. One question in this direction is whether (10) succeeds in conserving the integrals (3). To investigate this, consider

$$I_1 = \sum_{n=-N}^{N-1} u_n, \quad I_2 = \sum_{n=-N}^{N-1} u_n^2$$

which are the discrete analogs of the first two integrals in (3). At this point we refer to [3, Sect. 4.5], where the conservation properties of Fourier semi-discretizations of $u_t + uu_x = 0$ are discussed. It is straightforward to include the term $\mathcal{H}(u_{xx})$ in the analysis, the important property being the fact that the matrix $F^{-1}E_HF^{-1}E_2F$ in (10) is skew-symmetric. Thus we deduce that $dI_1/dt = 0$ but in general $dI_2/dt \neq 0$. In order to achieve exact conservation of the second quantity the scheme (10) needs to be modified—the nonlinear term has to be treated as $uu_x = \frac{1}{2} uu_x + \frac{1}{2}(u^2)_x$. The corresponding semi-discretization requires one extra FFT per evaluation as compared with (10), and it sacrifices exact conservation of the first quantity. It is possible to have both $I_1$ and $I_2$ conserved at the same time, but then a Galerkin or full spectral method (as opposed to pseudospectral as in (10)) should be used. But that introduces complications such as convolution sums which are expensive to deal with. For further discussion of these issues, including the distinction between spectral and pseudospectral, we refer to [3].

As for the Hamiltonian, $I_3$ in (3), it is not clear whether (10) can be modified to achieve exact conservation of this quantity. As will be discussed in Sect. 4, however, the discretization (10) conserves all three quantities in (3) remarkably well in practice. Thus we settled on (10) as our basic Fourier semi-discretization.

It remains to pick a method for integrating (10). Many candidates are available, but the leapfrog method

$$u^{m+1} = u^{m-1} - 2\Delta t f(u^m), \quad m = 1, 2, \ldots$$  (11)

seems to do well for nonlinear dispersive problems such as these. We start it with the explicit midpoint rule

$$u^1 = u^0 - \frac{1}{2} \Delta t f(u^0), \quad u^1 = u^0 - \Delta t f(u^{1/2}).$$

The main drawback of the Fourier method described here is the following. When truncating the infinite domain to the finite interval $[-L, L]$, a Fourier series introduces artificial periodicity. If the function $u(x)$ and its derivatives do not match precisely at the boundaries $x = \pm L$, a discontinuity is introduced. When this occurs the Fourier method is not as accurate as it normally is for smooth functions on a truly periodic domain. In the present problem we are dealing with solutions that decay as $|x| \to \infty$, so one could reduce the effect of the boundary discontinuity by choosing $L$ sufficiently large. However, this means fewer points in the interior of the domain and resolution is lost. We propose to overcome this difficulty using an expansion in rational functions, which should be better suited for solutions such as (4).

3. A rational method. Consider the rational expansion

$$u(x,t) = \sum_{n=-\infty}^{\infty} a_n(t) \phi_n(x), \quad \text{where} \quad \phi_n(x) = \frac{(L + iz)^n}{(L - iz)^{n+1}}.$$  (12)

$L$ is a free parameter, to be chosen to optimize accuracy. The functions $\phi_n(x)$ are complete and orthogonal in the space $L_2(-\infty, \infty)$ (see [4]). The orthogonality property is

$$\int_{-\infty}^{\infty} \phi_n(x)\phi_m(x)^* \, dx = \begin{cases} \pi/L & n = m, \\ 0 & n \neq m, \end{cases} \quad \text{so} \quad a_n = \frac{L}{\pi} \int_{-\infty}^{\infty} u(x)\phi_n^*(x) \, dx.$$  (13)
where \( \ast \) denotes the complex conjugate. Our method is based on truncated versions of the above formulas.

Spectral methods based on (12) and (13) for solving stationary problems are described in [4], [5] and [11]. We suspect the present paper describes the first application to an evolution equation.

As partial motivation for this approach, note the following. The initial condition corresponding to the one-soliton solution in (4) can be represented by only two terms, namely

\[
\frac{4c}{c^2x^2 + 1} = 2 \phi_0(x) + 2 \phi_1(x),
\]

provided we pick \( L = 1/c \). By contrast, it requires many Fourier modes to approximate this function accurately on a truncated domain.

We start with the computation of the Hilbert transform. It is shown in [12] that

\[
\mathcal{H}(\phi_n)(x) = i \text{Sgn}(n) \phi_n(x),
\]

where we use “Sgn” to indicate that \( \text{Sgn}(0) = 1 \); normally \( \text{sgn}(0) = 0 \).

For computing derivatives we need to relate the coefficients in the formula

\[
\sum_{n=-\infty}^{\infty} a_n \phi_n^{(k)}(x) = \sum_{n=-\infty}^{\infty} a_n^{(k)} \phi_n(x), \quad \text{for} \quad k = 1, 2.
\]

Direct differentiation of \( \phi_n(x) \), and using the identity \( \phi_n(x) + \phi_{n-1}(x) = 2L(\mathcal{L} + ix)^{n-1}/(\mathcal{L} - ix)^{n+1} \), yields

\[
a_n^{(1)} = \frac{i}{2L} \left[ n a_{n-1} + (2n+1)a_n + (n+1)a_{n+1} \right],
\]

\[
a_n^{(2)} = -\frac{1}{4L^2} \left[ n(n-1)a_{n-2} + 4n^2a_{n-1} + (5n^2 + 6n + 2)a_n + 4(n+1)^2a_{n+1} + (n+2)(n+1)a_{n+2} \right].
\]

In practice we truncate (12) to \(-N \leq n \leq N-1 \); this means setting \( a_n = 0 \) whenever \( n < -N \) or \( n > N-1 \) in (16) and (17). Thus we summarize

\[
a_n^{(1)} = C_1 a_n, \quad a_n^{(2)} = C_2 a_n, \quad \text{where} \quad a = (a_{-N}, \ldots, a_{N-1})^T.
\]

The matrix \( C_1 \) is skew-hermitian and tridiagonal, and \( C_2 \) is hermitian and pentadiagonal. They are the first and second derivative analogues of \( \mathcal{E}_1 \) and \( \mathcal{E}_2 \) in (9). Further properties of \( C_1 \) and \( C_2 \), including their eigenvalues, are given in [11]. Similarly, from (14) we conclude that the Hilbert transform analog of \( \mathcal{E}_H \) in (9) is

\[
C_H = i \text{diag}(-1, \ldots, -1, 1, \ldots, 1).
\]

The remaining question is the computation of the expansion coefficients \( a_n \) in terms of the function values and vice versa. This can be done efficiently with the FFT if one exploits the fact that (12) is essentially a transformed Fourier series. To see this, define \( \theta \) through

\[
e^{i\theta} = \frac{L + ix}{L - ix}, \quad \text{or} \quad x = L \tan \frac{1}{2} \theta, \quad -\pi \leq \theta \leq \pi,
\]

and introduce the uniform grid

\[
\theta_j = jh, \quad h = \pi/N, \quad j = -N, \ldots, N.
\]

This translates into the non-uniform grid in the original variable: \( z_j = L \tan \frac{1}{2} \theta_j \). Note that \( x = \pm \infty \) is included as a grid-point where we set the function to zero; no domain truncation is required with this approach. On this grid, the finite version of (12), namely

\[
u(x) = \sum_{n=-N}^{N-1} a_n \phi_n(x), \quad \text{becomes} \quad u_j(L - iz_j) = \sum_{n=-N}^{N-1} a_n e^{in\theta_j}.
\]
This formula can be evaluated and inverted with the FFT, summarized as
\[ u = \mathcal{F}^{-1} \mathcal{F}^{-1} a, \quad a = \mathcal{F} \mathcal{F} u. \]

The Fourier matrices \( \mathcal{F}, \mathcal{F}^{-1} \) are defined in (8), with \( n \theta_j \) replacing \( \mu_n z_j \). The weighting matrix is defined by
\[ \mathcal{P} = \text{diag}(L - i x_{-N}, \ldots, L - i x_{N-1}). \]

It has the undefined quantity \( L + i \infty \) in its first position, but when \( \mathcal{P} \) multiplies \( u \) the first entry is set to zero due to the boundary condition \( u \to 0 \) as \( |x| \to \infty \).

The above formulas may now be used to define the approximations to \( u'(x_j), u''(x_j), \) and \( \mathcal{H}(u'')(x_j) \):
\[ u' = \mathcal{P}^{-1} \mathcal{F}^{-1} \mathcal{C}_1 \mathcal{F} \mathcal{P} u, \quad u'' = \mathcal{P}^{-1} \mathcal{F}^{-1} \mathcal{C}_2 \mathcal{F} \mathcal{P} u, \quad \text{and} \quad h = \mathcal{P}^{-1} \mathcal{F}^{-1} \mathcal{C}_H \mathcal{C}_2 \mathcal{F} \mathcal{P} u. \]

Note that with this method we do not set \( a_{-N} = 0 \) in any of the formulas, as we did in the Fourier case. In the series (18) all terms appear as conjugate pairs; no spurious imaginary component will be introduced for real data.

Defining again \( \mathcal{U} = \text{diag}(a_{-N}, \ldots, a_{N-1}) \), the rational semi-discretization of (1) now reads
\[ \frac{d u}{d t} + \mathcal{U} \mathcal{P}^{-1} \mathcal{F}^{-1} \mathcal{C}_1 \mathcal{F} \mathcal{P} u + \mathcal{P}^{-1} \mathcal{F}^{-1} \mathcal{C}_H \mathcal{C}_2 \mathcal{F} \mathcal{P} u = 0, \]
which we integrate with the leapfrog method (11). Although formidable at first sight, this discretization is only marginally more expensive than the corresponding Fourier discretization (10). The added expense comes from carrying along the weighting matrix \( \mathcal{P} \), and the fact that the differentiation matrices \( \mathcal{C}_1 \) and \( \mathcal{C}_2 \) in (19) are banded whereas \( \mathcal{E}_1 \) and \( \mathcal{E}_2 \) in (10) are diagonal. In computing both (10) and (19), the bulk of the work consists of the two FFTs.

4. Numerical Experiments. We now compare the methods (10) and (19). As a first test, we consider simulating the one-soliton solution (4) with speed \( c = 1/5 \) (a rather arbitrary choice.) The initial condition for the numerical schemes is picked from the exact solution. As for choosing the parameter \( L \), we decided to pick it optimally for the initial condition, realizing that it will become non-optimal as soon as the soliton moves.

For the rational method we pick \( L = 1/c = 5 \) as discussed below (13). For the Fourier method it is not so easy to determine the optimal \( L \). We did it numerically by minimizing the residual (truncation error) at \( t = 0 \) in (10). For \( N = 64 \) we thus settled on \( L = 100 \).

Fig. 1 shows the results of this simulation. In the first picture the errors (defined as the maximum absolute error at the gridpoints) are shown as a function of time. In the second and third pictures the actual soliton is shown as it moves to the right. Initially the rational method is about six orders of magnitude better than the Fourier method. However, around \((x, t) = (10, 50)\) the error in the rational method starts increasing rapidly, and soon after that the numerical soliton loses its coherence. These effects may be attributed to the non-uniform grid \( x_j = L \tan \frac{\theta_j}{2} \). Near the origin the grid spacing is roughly \( h = 0.12 \), which increases to \( h = 0.64 \) near \( x = 10 \), and then unboundedly as \( x \to \infty \). The Fourier method, although not dramatically accurate, manages a useful approximation for a much longer period of time. In fact, the true solution is superimposed as a dot-dash curve in Fig. 1 (b) and (c), but it is hardly visible in the Fourier picture. Only much later, when the influence of the right hand boundary becomes a factor, does the Fourier solution break down.

One can improve the performance of the rational method by increasing the value of \( L \), thereby sacrificing accuracy initially in exchange for improved results over a longer period of time. Or one could try and devise methods for moving the grid and/or adapting \( L \) dynamically. We will not pursue this further, however.
Figure 1. One-soliton simulation as described in the text. Here $c = 0.2$, $\Delta t = 0.001$, $N = 64$, and $L = 100$, $L = 5$ for the Fourier and rational methods, respectively.

Figure 2. Two-soliton Fourier simulation as described in the text. Here $c_1 = 0.3$, $c_2 = 0.6$, $\phi_1 = -30$, $\phi_2 = -55$, $\Delta t = 0.001$, $N = 128$, $L = 100$. 376
Having established that the Fourier method is more robust in long-time calculations, we disregard the rational method for the more challenging two-soliton collision shown in Fig. 2. Again we superimposed the true solution, but this is invisible. The success of the simulation is evident. The solitons appear unscathed from the collision, as they should. Also, as observed in [9], they do not undergo a phase shift during collision, as do solitons of KdV type. A mild oscillation can be observed in the $t = 180$ picture, this being due to the presence of the artificial boundary on the right. It can be removed by increasing $N$, with a corresponding increase in $L$.

We also computed the trapezoidal rule approximations of the invariant integrals in (3) at various time levels. These quantities are conserved to a remarkable degree—over the interval $t \in [0,180]$ the values of $I_1$, $I_2$, and $I_3$ did not change by more than $6 \times 10^{-13}$, $2 \times 10^{-4}$, and $3 \times 10^{-8}$ percent, respectively. Considering that the calculation involved $1.8 \times 10^8$ time steps, these results reflect well on the ability of the Fourier method to capture the significant dynamics of the BO equation in a stable manner.

In summary, we have proposed two spectral methods for the BO equation. Although we have stopped short of proving convergence of the methods and deriving error bounds, numerical experiments confirm that both methods are useful and worth studying further. The rational method is capable of spectacular accuracy, provided the solution does not wander too far from the origin. In long time calculations, however, the Fourier method is decidedly superior.

Acknowledgments. The authors acknowledge useful discussions with Bob Higdon, Enrique Thomann, and Nick Trefethen. The second author would also like to acknowledge the hospitality of the Department of Computer Science, Cornell University, where part of this research was performed, as well as financial support from the Center for Applied Mathematics and the Advanced Computing Research Institute at Cornell.

References