SPECTRAL METHODS:

Next we will illustrate how spectral approximations are actually constructed for the solutions to ODEs and PDEs. For this purpose we will use Burger's equation. ⟶ Simple PDE, but its discretization by spectral methods illuminates many points that occur for much more complicated problems.

The Burger's equation:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} - \nu\frac{\partial^2 u}{\partial x^2} = 0 \quad \text{in } \Omega \quad \text{for } \forall t > 0 \quad \text{(strong form)}.$$

where 'ν' is a positive constant and Ω is the spatial domain. It provides a paradigm for more complex fluid-dynamics problems such as those described by the N-S equations. It can also be written in conservation form as:

$$\frac{\partial u}{\partial t} + \frac{\partial \bar{F}(u)}{\partial x} = 0 \quad \text{in } \Omega, \quad \forall t > 0$$

where the flux $\bar{F}$ is given by $\bar{F}(u) = \frac{1}{2}u^2 - \nu\frac{\partial u}{\partial x}$.

The initial conditions are: $u(x,0) = u_0(x)$ in Ω, and appropriate BCs are also defined.

▷ The Burger's equation successfully models certain gas dynamics, acoustics, and turbulence phenomena. It became a subject of extensive studies in the 1960s to investigate in isolation the specific feature of turbulence that balances generation of smaller scales by nonlinear advection with their dissipation by diffusion.

Since the Burgers equation is one of the few nonlinear PDEs for which exact and complete solutions are known in terms of the initial values, it remains a useful model problem for evaluating numerical algorithms.

↳ Through the transformation : $u = -2\nu \frac{\phi_x}{\phi}$

the Burger's equation reduces to the heat equation for ϕ: $\frac{\partial \phi}{\partial t} - \nu \frac{\partial^2 \phi}{\partial x^2} = 0$

The solution used for non-periodic problems is: $\phi_b(x,t) = \frac{x}{t} \frac{\sqrt{a/t}\ e^{-x^2/(4\nu t)}}{1 + \sqrt{a/t}\ e^{-x^2/(4\nu t)}}$

where 'a' is a cte. For periodic problems:

$$\phi_b(x,t) = \frac{1}{\sqrt{4\pi\nu t}} \sum_{n=-\infty}^{\infty} e^{-(x-2\pi n)^2/4\nu t}$$

- Strong & Weak formulation of differential equations:

Strong form : the PDE is required to be satisfied at each point in its domain and for each time.

Weak form : is obtained by requiring that the integral of the PDE against all functions in an appropriate space X of test functions be satisfied.

↓ Precisely:

SPECTRAL METHODS:

$$\int_a^b \frac{\partial u}{\partial t} v \, dx + \int_a^b u \frac{\partial u}{\partial x} v \, dx - \int_a^b \nu \frac{\partial^2 u}{\partial x^2} v \, dx = 0 \qquad \forall v \in X, \ \forall t > 0$$

when $\Omega = (a,b)$. This is often referred to as an 'integral form of the PDE'.

The weak formulation can accomodate less regular solutions. It is worth pointing out that for time-independent problems with a symmetric spatial operator, the weak formulation is also called the variational formulation.

Boundary conditions that should be satisfied by "u" are incorporated in the boundary terms, or taken into account in the choice of test functions.

⇒ Spectral collocation methods use the strong form of the PDE, as do finite-difference methods. For spectral Galerkian and tau methods as for finite-element methods, it is preferable to use the PDE in a weak form.

- The strong form can be written compactly as:

$$u_t + G(u) + \mathcal{L}(u) = 0 \quad \text{in } \Omega, \ \forall t > 0$$

where the non-linear operator G is defined by $G(u) = u\frac{\partial u}{\partial x}$, and the linear operator $\mathcal{L}$ is just $-\nu\left(\frac{\partial^2}{\partial x^2}\right)$.

- The compact form for the weak formulation is:

$$(u_t + G(u) + \mathcal{L}(u), v) = 0 \quad \text{for } \forall v \in X, \ \forall t > 0.$$

where (u, v) denotes the inner product in X.

- Spectral approximation of the Burgers equation:

Next we will illustrate discretization processes for several spectral approximations to the Burgers equation. We will consider different treatments of the nonlinear and linear terms as well as different treatments of the boundary conditions.

1) Fourier Galerkin

We look for a solution that is periodic in space on the interval $(0, 2\pi)$. The trial space X_N is S_N, the set of all trigonometric polynomials of degree $\leq N/2$. The approximate function u^N is represented as the truncated Fourier series

$$u^N(x,t) = \sum_{k=-N/2}^{N/2-1} \hat{u}_k(t)\, e^{ikx}$$

In this method the fundamental unknowns are the coefficients $\hat{u}_k(t)$, with $k = -N/2, \cdots, N/2-1$. Enforcement of the weak formulation yields,

$$\int_0^{2\pi} \left(\frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - \nu \frac{\partial^2 u^N}{\partial x^2} \right) e^{-ikx}\, dx = 0 \quad , \quad k = -\frac{N}{2}, \cdots, \frac{N}{2}-1$$

Due to the orthogonality property of the test and trial functions, we obtain a set of ODEs for the $\hat{u}_k$:

SPECTRAL METHODS:

$$\frac{d\hat{u}_k}{dt} + \widehat{\left(u_N \frac{\partial u^N}{\partial x}\right)}_k + k^2 \nu \hat{u}_k = 0 \quad , \quad \text{with} \quad k = -N/2, \dots, N/2 - 1.$$

where: $$\widehat{\left(u^N \frac{\partial u^N}{\partial x}\right)}_k = \frac{1}{2\pi} \int_0^{2\pi} u^N \frac{\partial u^N}{\partial x} e^{-ikx} \, dx$$

The initial conditions are clearly: $$\hat{u}_k(0) = \frac{1}{2\pi} \int_0^{2\pi} u(x,0) e^{-ikx} \, dx.$$

* The wavenumber $k = -N/2$ appears unsymmetrically in this approximation. This can lead to a number of difficulties, and it is advisable in practice simply to enforce the condition that $\hat{u}_{-N/2}$ is zero. This could be avoided if the approximation contained an odd rather than an even number of modes. However, the most widely used FFTs require an even number of modes.

* The advection term $\widehat{\left(u^N \frac{\partial u^N}{\partial x}\right)}_k$ is a particular case of the general quadratic nonlinear term:

$$\widehat{(uv)}_k = \frac{1}{2\pi} \int_0^{2\pi} uv \, e^{-ikx} \, dx$$

where u & v denote generic trigonometric polynomials of degree $\leq N/2$. They can correspondingly be expanded and hence:

$$\widehat{(uv)}_k = \sum_{p+q=k} \hat{u}_p \hat{v}_q \quad \Longrightarrow \text{This is a convolution sum,}$$

and it requires $O(N^2)$ operations. Alternative transform methods allow this term

to be evaluated in only $O(N \log_2 N)$ operations.

2) Fourier Collocation:

Once again, we presume periodicity on $(0, 2\pi)$ and take $X_N = S_N$, but now think of the approximate solution u^N as represented by its values at the grid points $x_j = \frac{2\pi j}{N}$, with $j = 0, \cdots, N-1$.

↓

Recall that the grid-point values of u^N are related to its discrete Fourier coefficients. For the collocation method it is required that the strong form be satisfied at these points.

$$\left. \frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - \nu \frac{\partial^2 u^N}{\partial x^2} \right|_{x = x_j} = 0 \quad , \text{ with } j = 0, 1, \cdots, N-1.$$

initial conditions: $u^N(x_j, 0) = u_0(x_j)$

where recall that: $(D_N u)_j = \sum_{k=-N/2}^{N/2-1} \tilde{u}_k^{(1)} e^{2ikj\pi/N} \quad ; \; j = 0, 1, \cdots, N-1$

and $\tilde{u}_k^{(1)} = ik \tilde{u}_k = \frac{ik}{N} \sum_{l=0}^{N-1} u(x_l) e^{-2ikl\pi/N} \quad , \; k = -N/2, \cdots, N/2-1$

↘

where the derivative $\frac{\partial u^N}{\partial x}$ is most efficiently evaluated by the transform differentiation procedure.

SPECTRAL METHODS:

For the conservative form : $\frac{\partial u}{\partial t} + \frac{\partial F(u)}{\partial x} = 0$, the nonlinear operator is approximated as follows:

$$G_N(u) = \frac{1}{2} D_N\left[(u^N)^2\right]$$

Hence, the non-linear term is evaluated by first taking the pointwise square of 'u' and then differentiating.

↳ (if using the Galerkin Method, the end set of equations results the same, contrary to what happens with the collocation method).

Examples: (see figures 3.1, 3.2 & 3.3).

3) Chebyshev Tau

The solution now needs to satisfy the Dirichlet boundary conditions on $(-1, 1)$.

$$u(-1, t) = u_L(t) \quad , \quad u(1, t) = u_R(t)$$

The trial space X_N consists of all the members of P_N (set of algebraic Polynomials of degree $< N$). The discrete sdution is expressed as the truncated Chebyshev series

$$u^N(x, t) = \sum_{k=0}^{N} \hat{u}_k(t)\ T_k(x),$$

with the Chebyshev coefficients comprising the fundamental representation of the approximation.

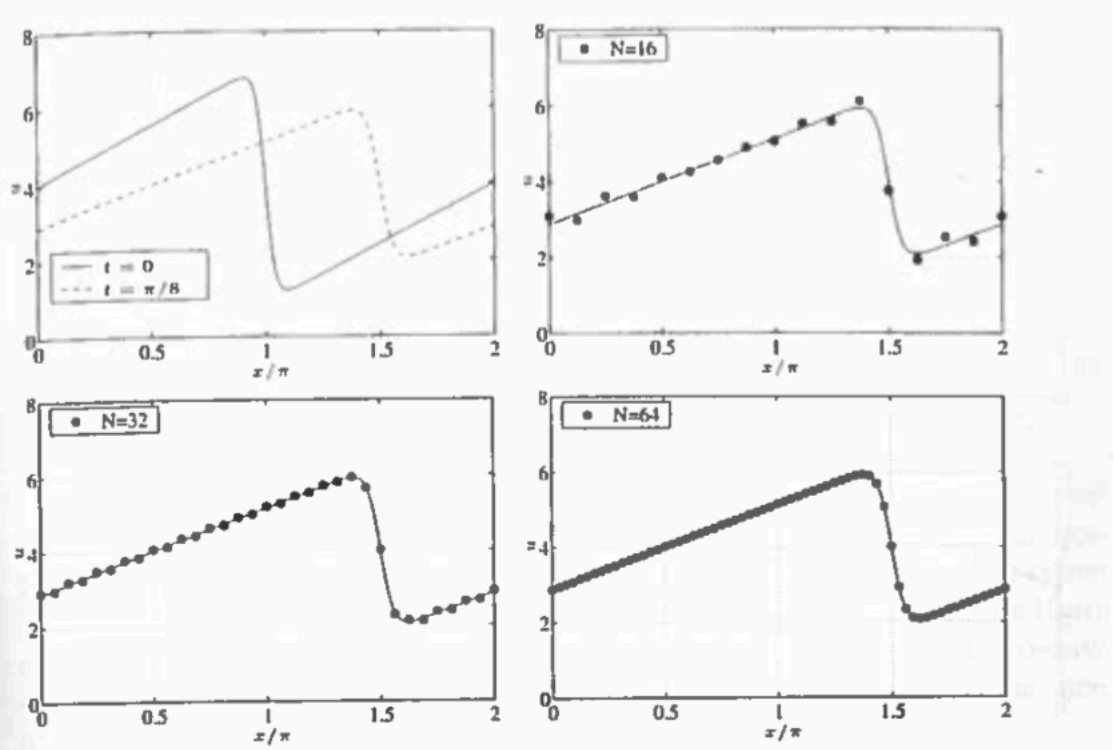

Fig. 3.1. The exact solution for the periodic Burgers equation problems (*top left*) and Fourier collocation solutions at $t = \pi/8$ for $N = 16$ (*top right*), $N = 32$ (*bottom left*), and $N = 64$ (*bottom right*)

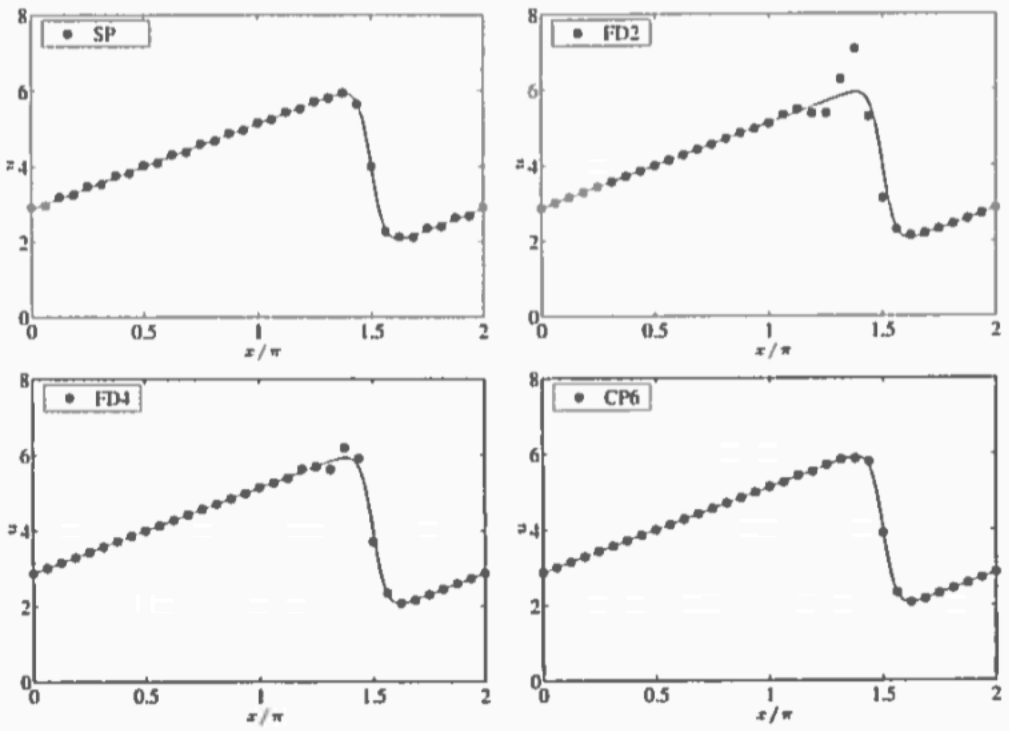

Fig. 3.2. Solutions to the periodic Burgers equation problem at $t = \pi/8$: comparison between Fourier collocation solution and finite-difference solutions of order 2, 4 and 6

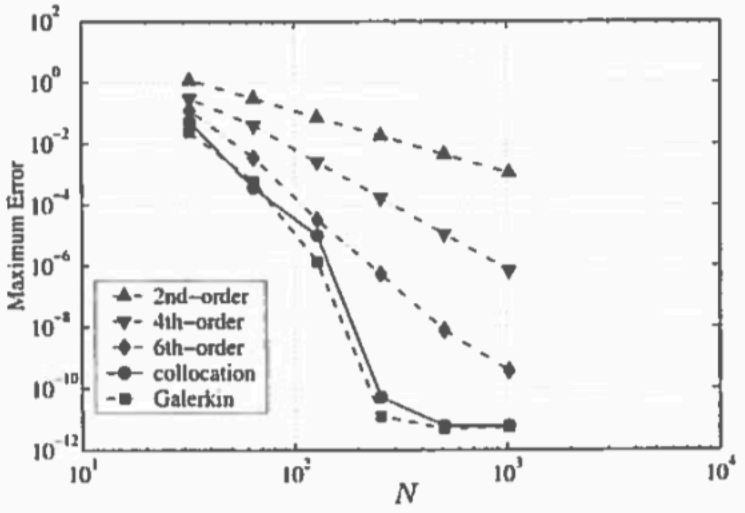

Fig. 3.3. Maximum errors for the periodic Burgers equation problem at $t = \pi/8$

hence: $$\int_{-1}^{1} \left(\frac{\partial u^N}{\partial t} + u^N \frac{\partial u^N}{\partial x} - \nu \frac{\partial^2 u^N}{\partial x^2} \right)(x)\ T_k(x)\ (1-x^2)^{-1/2}\ dx = 0$$

with $k = 0, \dots, N-2$.

The weight function, $w(x) = (1-x^2)^{-1/2}$, appropriate for the Chebyschev polynomials is used in the orthogonality condition.

hence:

$$\frac{\partial \hat{u}_k}{\partial t} + \widehat{\left(u^N \frac{\partial u^N}{\partial x} \right)}_k - \nu\, \hat{u}_k^{(2)} = 0 \qquad \text{with} \quad k = 0, 1, \dots, N-2.$$

where $$\hat{u}_k^{(2)} = \frac{1}{c_k} \sum_{\substack{p=k+2 \\ p+k \text{ even}}}^{\infty} p\left(p^2 - k^2\right) \hat{u}_p \,, \quad k \geq 0$$ (see references for details).

and:

$$\widehat{\left(u^N \frac{\partial u^N}{\partial x} \right)}_k = \frac{2}{\pi c_k} \int_{-1}^{1} \left(u_N \frac{\partial u^N}{\partial x} \right)(x)\ T_k(x)\ (1-x^2)^{-1/2}\ dx$$

where $$c_k = \begin{cases} 2, & k = 0 \\ 1, & k \geq 1. \end{cases}$$

In terms of the Chebyshev coefficients the boundary conditions become:

$$\sum_{k=0}^{N} \hat{u}_k = u_R \qquad \sum_{k=0}^{N} (-1)^k \hat{u}_k = u_L$$

SPECTRAL METHODS

The initial conditions are: $\hat{u}_k(0) = \frac{2}{\pi c_k} \int_{-1}^{1} u_0(x)\, T_k(x)\, (1-x^2)^{-1/2}\, dx, \quad k = 0, 1, \cdots N$

* The expression for $\widehat{\left(u^N \frac{\partial u^N}{\partial x}\right)}_k$ is a special case of

$$\widehat{(uv)}_k = \frac{2}{\pi c_k} \int_{-1}^{1} u(x)\, v(x)\, T_k(x)\, (1-x^2)^{-1/2}\, dx$$

which is equal to the following expression involving the convolution sums:

$$\widehat{(uv)}_k = \frac{1}{2} \sum_{p+q=k} \hat{u}_p \hat{v}_q + \sum_{|p-q|=k} \hat{u}_p \hat{v}_p$$

* if the BCs are of Neuman type, $u'(-1,t) = 0$, $u'(1,t) = 0$, then conditions are replaced by:

$$\sum_{k=1}^{N} k^2 \hat{u}_k = 0, \qquad \sum_{k=1}^{N} (-1)^k k^2 \hat{u}_k = 0$$

Nonperiodic Numerical Example The nonperiodic exact solution u corresponding to (3.1.8), (3.1.7) and (3.1.5) for $\nu = 0.01$, $c = 1$, $a = 16$ and $t_0 = 1$ is shown in Fig. 3.4 (left) at $t = 0$ and $t = 1$. The Burgers equation is solved on the interval $(-1, 1)$ with initial and boundary data taken from this exact solution.

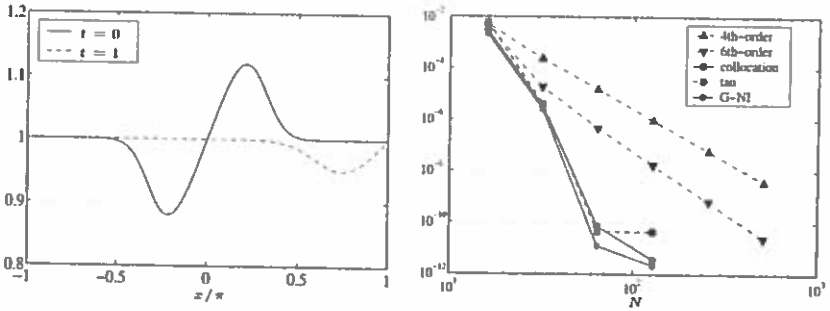

Fig. 3.4. The exact solution for the nonperiodic Burgers equation problems (*left*) and computed maximum errors at $t = 1$ (*right*)

Figure 3.4 (right) illustrates the errors from the Chebyshev tau, Chebyshev collocation and G-NI numerical schemes on this problem, integrated in time with the RK4 method (see (D.2.17)). Also included for comparison are solutions for fourth-order and sixth-order compact differences. Compact-difference approximations to the first and second derivatives require special one-sided stencils for the points at and adjacent to the boundaries. For the fourth-order scheme, the stencils used here are taken from Lele (1992); they are third-order accurate at the boundaries and fourth-order accurate for all the interior points. The asymptotic decay rate of the fourth-order solutions shown in Fig. 3.4 is fourth order. The stencils for the sixth-order scheme are third order at the boundary points, fourth order at the points adjacent to the boundaries and sixth order everywhere else. (See Sect. 3.7 and CHQZ3, Sect. 4.2 for further discussion of the challenges of appropriate boundary stencils for compact schemes. As illustrated in CHQZ3, Fig. 4.2, higher order stencils near the boundaries for this class of sixth-order schemes are temporally unstable.) The asymptotic decay rate of the sixth-order results is less than fifth order. All the spectral results decay faster than algebraically without requiring any special treatment at the boundaries.

4 Convolution Sums

The principal algorithmic component of efficient Galerkin methods for nonlinear or variable-coefficient problems is the evaluation of convolution sums.

Consider, however, the Fourier Galerkin treatment of the product

$$s(x) = u(x)v(x)\,. \tag{3.4.1}$$

In the case of an infinite series expansion, we have the familiar convolution sum

$$\hat{s}_k = \sum_{m+n=k} \hat{u}_m \hat{v}_n\,, \tag{3.4.2}$$

where

$$u(x) = \sum_{m=-\infty}^{\infty} \hat{u}_m e^{imx}, \qquad v(x) = \sum_{n=-\infty}^{\infty} \hat{v}_n e^{inx}\,, \tag{3.4.3}$$

and

$$\hat{s}_k = \frac{1}{2\pi}\int_0^{2\pi} s(x)e^{-ikx}\,dx\,. \tag{3.4.4}$$

In the present context u and v are finite Fourier series of degree $\le N/2$, i.e., trigonometric polynomials belonging to S_N, whereas $s \in S_{2N}$. The values of $\hat{s}_k$, though, are only of interest for $|k| \le N/2$. So, we truncate the product (3.4.1) at degree $N/2$ (i.e., taking $P_N(uv)$). Then (3.3.2) becomes

$$\hat{s}_k = \sum_{\substack{m+n=k \\ |m|,|n|\le N/2}} \hat{u}_m \hat{v}_n\,, \qquad |k| \le N/2\,, \tag{3.4.5}$$

which amounts to requiring (3.4.4) for $|k| \le N/2$. The direct summation implied by (3.4.5) takes $O(N^2)$ operations. (In three dimensions, the cost is $O(N^4)$, provided, as discussed in Orszag (1980), that one utilizes the tensor-product nature of multidimensional spectral approximations.) This is prohibitively expensive, especially when one considers that for a nonlinear term a finite-difference algorithm takes $O(N)$ operations in one dimension (and $O(N^3)$ in three). However, the use of transform methods enables (3.4.5) to be evaluated in $O(N \log_2 N)$ operations (and the three-dimensional generalization in $O(N^3 \log_2 N)$ operations). This technique was developed independently by Orszag (1969, 1970) and Eliasen, Machenhauer and Rasmussen (1970). It was the single most important development that made spectral Galerkin methods practical for large-scale computations.

3.4.1 Transform Methods and Pseudospectral Methods

The approach taken in the transform method for evaluating (3.4.5) for u, v in S_N is to use the inverse discrete Fourier transform (DFT) to transform $\hat{u}_m$ and $\hat{v}_n$ to physical space, to perform there a multiplication similar to (3.4.1), and then to use the DFT to determine $\hat{s}_k$. This must be done carefully, however. To illustrate the subtle point involved, we introduce the discrete transforms (Sect. 2.1.2):

$$u_j = \sum_{k=-N/2}^{N/2-1} \hat{u}_k e^{ikx_j} ,$$

$$j = 0, 1, \ldots, N-1 , \qquad (3.4.6)$$

$$v_j = \sum_{k=-N/2}^{N/2-1} \hat{v}_k e^{ikx_j} ,$$

and define

$$s_j = u_j v_j , \qquad j = 0, 1, \ldots, N-1 , \qquad (3.4.7)$$

and

$$\tilde{s}_k = \frac{1}{N} \sum_{j=0}^{N-1} s_j e^{-ikx_j} , \qquad k = -\frac{N}{2}, \ldots, \frac{N}{2} - 1 , \qquad (3.4.8)$$

where

$$x_j = 2\pi j/N .$$

Note that the $\tilde{s}_k$ are the discrete Fourier coefficients of the function s (see 2.1.25)). Use of the discrete transform orthogonality relation (2.1.26) leads o

$$\tilde{s}_k = \sum_{m+n=k} \hat{u}_m \hat{v}_n + \sum_{m+n=k\pm N} \hat{u}_m \hat{v}_n = \hat{s}_k + \sum_{m+n=k\pm N} \hat{u}_m \hat{v}_n . \qquad (3.4.9)$$

The second term on the right-hand side is the aliasing error. If the convolution sums are evaluated as described above, then the differential equation is not approximated by a true spectral Galerkin method. Orszag (1971a) termed the resulting scheme a *pseudospectral method*. The convolution sum (3.4.5) in the pseudospectral method is evaluated at the cost of 3 FFTs and N multiplications. The total operation count is $(15/2)N \log_2 N$ multiplications. The generalization of the pseudospectral evaluation of convolution sums to more than one dimension is straightforward.

There are two basic techniques for removing the aliasing error from (3.4.9). They are discussed in the following two subsections.

4.2 Aliasing Removal by Padding or Truncation

The key to this *de-aliasing* technique is the use of a discrete transform with rather than N points, where $M \geq 3N/2$. Let

$$y_j = 2\pi j/M, \quad \breve{u}_j = \sum_{k=-M/2}^{M/2-1} \breve{u}_k e^{iky_j}, \quad \breve{v}_j = \sum_{k=-M/2}^{M/2-1} \breve{v}_k e^{iky_j}, \qquad (3.4.10)$$

$$\breve{s}_j = u_j v_j , \qquad (3.4.11)$$

for $j = 0, 1, \ldots, M-1$, where

$$\breve{u}_k = \begin{cases} \hat{u}_k , & |k| \lesssim N/2 \\ 0 & \text{otherwise} \end{cases} . \qquad (3.4.12)$$

(Note that the $\bar{u}_j$ (and $\bar{v}_j$ and $\bar{s}_j$) are the values of u at $y_j = 2\pi j/M$, whereas the u_j defined in the previous section are the values of u at $x_j = 2\pi j/N$.) Thus, the $\breve{u}_k$ coefficients are the $\hat{u}_k$ coefficients padded with zeros for the additional wavenumbers. Similarly, let

$$\breve{s}_k = \frac{1}{M} \sum_{j=0}^{M-1} \breve{s}_j e^{-iky_j} , \qquad k = -\frac{M}{2}, \ldots, \frac{M}{2} - 1 . \qquad (3.4.13)$$

Then

$$\breve{s}_k = \sum_{m+n=k} \breve{u}_m \breve{v}_n + \sum_{m+n=k\pm M} \breve{u}_m \breve{v}_n . \qquad (3.4.14)$$

We are only interested in $\breve{s}_k$ for $|k| \leq N/2$, and choose M so that the second term on the right-hand side vanishes for these k. Since $\breve{u}_m$ and $\breve{v}_m$ are zero for $|m| > N/2$, the worst-case condition is

$$-\frac{N}{2} - \frac{N}{2} \leq \frac{N}{2} - 1 - M ,$$

or

$$M \geq \frac{3N}{2} - 1 . \qquad (3.4.15)$$

With M so chosen we have obtained the de-aliased coefficients

$$\hat{s}_k = \breve{s}_k , \qquad k = -\frac{N}{2}, \ldots, \frac{N}{2} - 1 . \qquad (3.4.16)$$

The operation count for this transform method is $(45/4)N \log_2(\frac{3}{2}N)$, which is roughly 50% larger than the simpler, but aliased, method discussed earlier. For obvious reasons this technique is sometimes referred to as the *3/2-rule*. As described here it requires an FFT that can handle prime factors of 3. If only a prime factor 2 FFT is available, then this de-aliasing technique can be implemented by choosing M as the smallest power of 2 that satisfies (3.4.15). This de-aliasing technique is also termed truncation and is sometimes referred to as the *2/3-rule*.

3.4.3 Aliasing Removal by Phase Shifts

A second method to remove the aliasing terms, due to Patterson and Orszag (1971), employs phase shifts. In this case (3.4.6) is replaced with

$$u_j^{\Delta} = \sum_{k=-N/2}^{N/2-1} \hat{u}_k e^{ik(x_j+\Delta)}, \quad v_j^{\Delta} = \sum_{k=-N/2}^{N/2-1} \hat{v}_k e^{ik(x_j+\Delta)} , \qquad (3.4.17)$$

$$j = 0, 1, \ldots, N-1 ,$$