From Durran (1999) - Numerical Methods for wave Equations in Geophysical Fluid Dynamics. Springer

4

Series-Expansion Methods

Series-expansion methods that are potentially useful in geophysical fluid dynamics include the spectral method, the pseudospectral method, and the finite-element method. The spectral method plays a particularly important role in global atmospheric models, in which the horizontal structure of the numerical solution is often represented as a truncated series of spherical harmonics. Finite-element methods, on the other hand, are not commonly used in multidimensional wave propagation problems because they generally require the solution of implicit algebraic systems and are therefore not as efficient as competing explicit methods. All of these series-expansion methods share a common foundation that will be discussed in the next section.

4.1 Strategies for Minimizing the Residual

Suppose F is an operator involving spatial derivatives of ψ, and that solutions are sought to the partial differential equation

$$\frac{\partial \psi}{\partial t} + F(\psi) = 0, \tag{4.1}$$

subject to the initial condition $\psi(x, t_0) = f(x)$ and to boundary conditions at the edges of some spatial domain S. The basic idea in all series-expansion methods is to approximate the spatial dependence of ψ as a linear combination of a finite number of predetermined expansion functions. Let the general form of the series expansion be written as

$$\phi(x,t) = \sum_{k=1}^{N} a_k(t)\varphi_k(x), \tag{4.2}$$

where $\varphi_1, \ldots, \varphi_N$ are predetermined expansion functions satisfying the required boundary conditions. Then the task of solving (4.1) is transformed into a problem of calculating the unknown coefficients $a_1(t), \ldots, a_N(t)$ in a way that minimizes the error in the approximate solution. One might hope to obtain solvable expressions for the $a_k(t)$ by substituting (4.2) into the governing equation (4.1). For example, if F is a linear function of $\partial^n\psi/\partial x^n$ with constant coefficients and the expansion functions are Fourier series, direct substitution will yield a system of ordinary differential equations for the evolution of the $a_k(t)$. Unfortunately, direct substitution yields a solvable system of equations for the expansion coefficients only when the φ_k are eigenfunctions of the differential operator F—direct substitution works in precisely those special cases for which analytic solutions are available. This, of course, is a highly restrictive limitation.

In the general case where the φ_k are not eigenfunctions of F, it is impossible to specify $a_1(t), \ldots, a_N(t)$ such that an expression of the form (4.2) exactly satisfies (4.1). As an example, suppose $F(\psi) = \psi\partial\psi/\partial x$ and the expansion functions are the Fourier components $\varphi_k = e^{ikx}$, $-N \leq k \leq N$. If this Fourier series is substituted into (4.1), the nonlinear product in $F(\psi)$ introduces spatial variations at wave numbers that were not present in the initial truncated series, e.g., $F(e^{iNx}) = iNe^{i2Nx}$. A total of $4N + 1$ equations are obtained after substituting the expansion functions into (3.1) and requiring that the coefficients of each Fourier mode sum to zero. It is not possible to choose the $2N + 1$ Fourier coefficients in the original expansion to satisfy these $4N + 1$ equations simultaneously. The best one can do is to select the expansion coefficients to minimize the error.

Since the actual error in the approximate solution $\|\psi - \phi\|$ cannot be determined, the most practical way to try to minimize the error is to minimize the *residual*,

$$R(\phi) = \frac{\partial\phi}{\partial t} + F(\phi), \tag{4.3}$$

which is the amount by which the approximate solution fails to satisfy the governing equation. Three different strategies are available for constraining the size of the residual. Each strategy leads to a system of N coupled *ordinary* differential equations for the time-dependent coefficients $a_1(t), \ldots, a_N(t)$. This transformation of the partial differential equation into a system of ordinary differential equations is similar to that which occurs in grid-point methods when the spatial derivatives are replaced with finite differences.

One strategy for constraining the size of the residual is to pick the $a_k(t)$ to minimize the square of the ℓ_2-norm of the residual:

$$(\|R(\phi)\|_2)^2 = \int_S [R(\phi(x))]^2\,dx.$$

A second approach, referred to as collocation, is to require the residual to be zero at a discrete set of grid points:

$$R(\phi(j\Delta x)) = 0 \quad \text{for all } j = 1, \ldots, N.$$

The third strategy, known as the Galerkin approximation, requires the residual to be *orthogonal* to each of the expansion functions, i.e.,

$$\int_S R(\phi(x))\varphi_k(x)\,dx = 0 \quad \text{for all } k = 1, \ldots, N. \tag{4.4}$$

Different series-expansion methods rely on one or more of the preceding approaches. The collocation strategy is used in the pseudospectral method and in some finite-element formulations, but not in the spectral method. The ℓ_2-minimization and Galerkin criteria are equivalent when applied to a problem of the form (4.1), and are the basis of the spectral method. The Galerkin approximation is also used extensively in finite-element schemes.

The equivalence of the ℓ_2-minimization criterion and the Galerkin approximation can be demonstrated as follows. According to (4.3), the residual depends on both the instantaneous values of the expansion coefficients and their time tendencies. The expansion coefficients are determined at the outset from the initial conditions and are known at the beginning of any subsequent integration step. The criteria for constraining the residual are not used to obtain the instantaneous values of the expansion coefficients, but rather to determine their time evolution. If the rate of change of the kth expansion function is calculated to minimize $(\|R(\phi)\|_2)^2$, a necessary criterion for a minimum may be obtained by differentiation with respect to the quantity $da_k/dt \equiv \dot{a}_k$

$$\begin{aligned}
0 &= \frac{d}{d(\dot{a}_k)}\left\{\int_S (R(\phi))^2\,dx\right\} \\
&= \frac{d}{d(\dot{a}_k)}\left\{\int_S \left[\sum_{n=1}^{N} \dot{a}_n\varphi_n + F\left(\sum_{n=1}^{N} a_n\varphi_n\right)\right]^2 dx\right\} \\
&= 2\int_S \left[\sum_{n=1}^{N} \dot{a}_n\varphi_n + F\left(\sum_{n=1}^{N} a_n\varphi_n\right)\right]\varphi_k\,dx && (4.5) \\
&= 2\int_S R(\phi)\varphi_k\,dx. && (4.6)
\end{aligned}$$

The second derivative of $(\|R(\phi)\|_2)^2$ with respect to $\dot{a}_k$ is $2(\|\varphi_k\|_2)^2$, which is positive. Thus, the extremum condition (4.6) is associated with a true minimum of $(\|R(\phi)\|_2)^2$, and the Galerkin requirement is identical to the condition obtained by minimizing the ℓ_2-norm of the residual.

As derived in (4.5), the Galerkin approximation and the ℓ_2-minimization of the residual both lead to a system of ordinary differential equations for the expansion coefficients of the form

$$\sum_{n=1}^{N} I_{nk}\frac{da_n}{dt} = -\int_S \left[F\left(\sum_{n=1}^{N} a_n\varphi_n\right)\varphi_k \right] dx \quad \text{for all } k = 1, \ldots, N, \tag{4.7}$$

where

$$I_{nk} = \int_S \varphi_n\varphi_k \, dx.$$

The initial conditions for the preceding system of differential equations are obtained by choosing $a_1(t_0), \ldots, a_N(t_0)$ such that $\phi(x, t_0)$ provides the "best" approximation to $f(x)$. The possible strategies for constraining the initial error are identical to those used to ensure that the residual is small. As before, the choice that minimizes the ℓ_2-norm of the initial error also satisfies the Galerkin requirement that the initial error be orthogonal to each of the expansion functions,

$$\int_S \left(\sum_{n=1}^{N} a_n(t_0)\varphi_n(x) - f(x) \right) \varphi_k(x)\, dx = 0 \quad \text{for all } k = 1, \ldots, N,$$

or, equivalently,

$$\sum_{n=1}^{N} I_{nk}a_n = \int_S f(x)\varphi_k(x)\, dx \quad \text{for all } k = 1, \ldots, N. \tag{4.8}$$

4.2 The Spectral Method

The characteristic that distinguishes the spectral method from other series-expansion methods is that the expansion functions form an orthogonal set. Since the expansion functions are orthogonal, I_{nk} is zero unless $n = k$, and the system of differential equations for the coefficients (4.7) reduces to

$$\frac{da_k}{dt} = -\frac{1}{I_{kk}}\int_S \left[F\left(\sum_{n=1}^{N} a_n\varphi_n\right)\varphi_k \right] dx \quad \text{for all } k = 1, \ldots, N. \tag{4.9}$$

This is a particularly useful simplification, since explicit algebraic equations for each $a_k(t + \Delta t)$ are obtained when the time derivatives in (4.9) are replaced with finite differences. In contrast, the finite-difference approximation of the time derivatives in the more general form (4.7) introduces a coupling between all the expansion coefficients at the new time level, and the solution of the resulting implicit system of algebraic equations may require considerable computation. The orthogonality of the expansion functions also reduces the expression for the initial value of each expansion coefficient (4.8) to

$$a_k(t_0) = \frac{1}{I_{kk}}\int_S f(x)\varphi_k(x)\, dx. \tag{4.10}$$

The choice of some particular family of orthogonal expansion functions is largely dictated by the geometry of the problem and by the boundary conditions. Fourier series are well suited to rectangular domains with periodic boundary conditions. Chebyshev polynomials are a possibility for nonperiodic domains. Associated Legendre functions are useful for representing the latitudinal dependence of a function on the spherical Earth. Since Fourier series lead to the simplest formulae, they will be used to illustrate the elementary properties of the spectral method. The special problems associated with spherical geometry will be discussed in Section 4.4.

4.2.1 *Comparison with Finite-Difference Methods*

In Chapter 2, a variety of finite-difference methods were tested on the one-dimensional advection equation

$$\frac{\partial\psi}{\partial t} + c\frac{\partial\psi}{\partial x} = 0. \tag{4.11}$$

Particular emphasis was placed on the simplest case, in which c was constant. When c is constant, it is easy to find expansion functions that are eigenfunctions of the spatial derivative term in (4.11). As a consequence, the problem of advection by a constant wind is almost too simple for the spectral method. Nevertheless, the constant-wind case reveals some of the fundamental strengths and weaknesses of the spectral method and allows a close comparison between the spectral method and finite-difference schemes.

Suppose, therefore, that c is constant and solutions are sought to (4.11) on the periodic domain $-\pi \le x \le \pi$, subject to the initial condition $\psi(x, 0) = f(x)$. A Fourier series expansion

$$\phi(x,t) = \sum_{k=-N}^{N} a_k(t)e^{ikx} \tag{4.12}$$

is the natural choice for this problem. Since individual Fourier modes are eigenfunctions of the differential operator in (4.11), evolution equations for the Fourier coefficients of the form

$$\frac{da_k}{dt} + icka_k = 0 \tag{4.13}$$

may be obtained by directly substituting (4.12) into the advection equation. In this atypically simple case, the residual is zero, and it is not necessary to adopt any particular procedure to minimize its norm. Nevertheless, (4.13) can also be obtained through the Galerkin requirement that the residual be orthogonal to each of the expansion functions. In order to apply the Galerkin formulation it is necessary to generalize the definition of orthogonality to include complex-valued functions. Two complex-valued functions $g(x)$ and $h(x)$ are *orthogonal* over the domain S if

$$\int_S g(x)h^*(x)dx = 0,$$

where $h^*(x)$ denotes the complex conjugate of $h(x)$.[1] As an example, note that for integer values of n and m,

$$\int_{-\pi}^{\pi} e^{inx}e^{-imx}\,dx = \begin{cases} \left.\dfrac{e^{i(n-m)x}}{n-m}\right|_{-\pi}^{\pi} = 0, & \text{if } m \neq n; \\ 2\pi, & \text{if } m = n, \end{cases}$$

which is just the well-known orthogonality condition for two Fourier modes. Using this orthogonality relation and setting $f(\psi) = c\partial\psi/\partial x$, with c constant, reduces (4.9) to (4.13).

If the ordinary differential equation (4.13) is solved analytically (in practical applications it must be computed numerically), solutions have the form $a_k(t) = \exp(-ickt)$. Thus, in the absence of time-differencing errors, the frequency of the kth Fourier mode is identical to the correct value for the continuous problem $\omega = ck$. The spectral approximation does not introduce phase speed or amplitude errors—even in the shortest wavelengths! The ability of the spectral method to correctly capture the amplitude and phase speed of the shortest resolvable waves is a significant advantage over conventional grid-point methods, in which the spatial derivative is approximated by finite differences, yet surprisingly, the spectral method is not necessarily a good technique for modeling short-wavelength disturbances. The problem lies in the fact that it is only those waves retained in the truncated series expansion that are correctly represented in the spectral solution. If the true solution has a great deal of spatial structure on the scale of the shortest wavelength in the truncated series expansion, the spectral representation will not accurately approximate the true solution.

The problems with the representation of short-wavelength features in the spectral method are illustrated in Fig. 4.1, which shows ten grid-point values forming a $2\Delta x$-wide spike against a zero background on a periodic domain with a uniformly spaced grid. Also shown is the curve defined by the truncated Fourier series passing through those ten grid-point values. The Fourier series approximation to the $2\Delta x$ spike exhibits large oscillations about the zero background state on both sides of the spike. Now suppose that the data in Fig. 4.1 represent the initial condition for a constant-wind-speed advection problem. The over- and under-shoots associated with the Fourier approximation will not be apparent at the initial time if the data are sampled only at the points on the discrete mesh. If time-differencing errors are neglected, the grid-point values will also be exact at those subsequent times at which the initial distribution has translated an integral number of grid intervals. The grid-point values will, however, reveal the oscillatory error at intermediate times. The worst errors in the grid-point values will occur at times when the solution has traveled $n + \frac{1}{2}$ grid intervals, where n is any integer. In this particular example, the error on the discrete grid does not accumulate with time; it oscillates instead, achieving a minimum when the solution has translated an integral number of grid intervals. The maximum error is limited by the error generated when the initial condition is projected onto the truncated Fourier series.

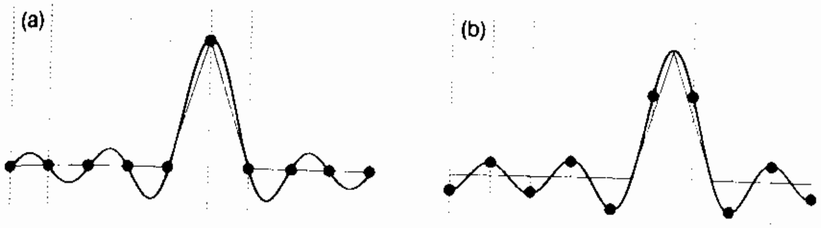

FIGURE 4.1. (a) Ten periodic grid-point values exhibiting a piecewise linear $2\Delta x$ spike, and the truncated Fourier series approximation passing through those ten points. (b) Values of the truncated Fourier series sampled at the same grid-point location after translating the curve one-half grid point to the right.

The Finite Fourier Transform

There is a simple relationship between the nine[2] independent grid-point values in Fig. 4.1 and the nine coefficients determining the truncated Fourier series passing through those points. If the Fourier-series expansion of a real-valued function is truncated at wave number N, the set of Fourier coefficients contains $2N+1$ pieces of data. Assuming that the Fourier expansion functions are periodic on the domain $0 \le x \le 2\pi$, an equivalent amount of information is contained by the $2N+1$ function values $\phi(x_j, t)$, where

$$x_j = j\left(\frac{2\pi}{2N+1}\right) \quad \text{and} \quad j = 1, \ldots, 2N+1.$$

It is obvious that the set of Fourier coefficients $a_{-N}(t), \ldots, a_N(t)$ defines the grid-point values $\phi(x_j, t)$ through the relation

$$\phi(x_j, t) = \sum_{k=-N}^{N} a_k(t)e^{ikx_j}. \tag{4.14}$$

Although it is not as self-evident, the $2N+1$ Fourier coefficients may also be determined from the $2N+1$ grid-point values. An exact algebraic expression for

[1] Multiplication by the complex conjugate ensures that if $g(x) = a(x) + ib(x)$ with a and b real, then

$$\int_S g(x)g^*(x)dx = \int_S (a^2 + b^2)dx.$$

[2] The tenth grid-point value is redundant information because the solution is periodic.

$a_n(t)$ can be obtained by noting that

$$\sum_{j=1}^{2N+1} \phi(x_j,t)e^{-inx_j} = \sum_{j=1}^{2N+1}\left(\sum_{m=-N}^{N} a_m(t)e^{imx_j}\right)e^{-inx_j}$$
$$= \sum_{m=-N}^{N} a_m(t)\left(\sum_{j=1}^{2N+1} e^{imx_j}e^{-inx_j}\right). \tag{4.15}$$

Further simplification of the preceding equation is possible because the final summation in (4.15) obeys an orthogonality condition on the discrete mesh. Using the definition of x_j,

$$\sum_{j=1}^{2N+1} e^{imx_j}e^{-inx_j} = \sum_{j=1}^{2N+1}\left(e^{\frac{i2\pi(m-n)}{2N+1}}\right)^j. \tag{4.16}$$

If $m = n$, then (4.16) sums to $2N + 1$; for $m \neq n$ the formula for the sum of a finite geometric series,

$$1 + r + r^2 + \cdots + r^n = \frac{1 - r^{n+1}}{1 - r}, \tag{4.17}$$

may be used to reduce (4.16) to

$$\sum_{j=1}^{2N+1} e^{imx_j}e^{-inx_j} = \frac{e^{\frac{i2\pi(m-n)}{2N+1}}\left(1 - e^{i2\pi(m-n)}\right)}{1 - e^{\frac{i2\pi(m-n)}{2N+1}}} = 0. \tag{4.18}$$

Using these orthogonality properties, (4.15) becomes

$$a_n(t) = \frac{1}{2N+1}\sum_{j=1}^{2N+1} \phi(x_j,t)e^{-inx_j}. \tag{4.19}$$

The relations (4.19) and (4.14), known as *finite Fourier transforms*, are discretized analogues to the standard Fourier transform and its inverse. The integrals in the continuous transforms are replaced by finite sums in the discrete expressions. These formulae, or more specifically the mathematically equivalent Fast Fourier Transform (FFT) algorithms, are essential for obtaining efficient spectral solutions in many practical applications where it is advantageous to transform the solution back and forth between wave number space and physical space once during the execution of every time step.

The Equivalent Grid-Point Method.

If c is constant, the spectral solution to the advection equation (4.11) can be recast in the form of an equivalent finite-difference method. Observe that

$$\frac{d\phi(x_j,t)}{dt} = \sum_{n=-N}^{N} \frac{da_n}{dt}e^{inx_j}$$
$$= \sum_{n=-N}^{N} inca_n(t)e^{inx_j}$$
$$= -c\sum_{n=-N}^{N} in\left(\frac{1}{2N+1}\sum_{k=1}^{2N+1}\phi(x_k,t)e^{-inx_k}\right)e^{inx_j}$$
$$= -c\sum_{k=1}^{2N+1} C_{j,k}\,\phi(x_k,t),$$

where

$$C_{j,k} = \frac{1}{2N+1}\sum_{n=-N}^{N} ine^{in(x_j-x_k)}.$$

The finite-difference coefficient $C_{j,k}$ depends only on the difference between j and k, and is zero if $j = k$. If $j \neq k$, a simpler expression for $C_{j,j+\ell}$ can be obtained by defining

$$s = x_j - x_{j+\ell} = -\ell\left(\frac{2\pi}{2N+1}\right),$$

in which case

$$C_{j,j+\ell} = \frac{1}{2N+1}\frac{d}{ds}\left(\sum_{n=-N}^{N} e^{ins}\right) = \frac{1}{2N+1}\frac{d}{ds}\left(e^{-iNs}\sum_{n=0}^{2N}(e^{is})^n\right).$$

Using (4.17) to sum the finite geometric series, differentiating, and noting that $e^{i(2N+1)s} = 1$, the preceding becomes

$$C_{j,j+\ell} = \frac{(-1)^{\ell+1}}{2\sin\left(\frac{\ell\pi}{2N+1}\right)},$$

which implies that since $C_{j,j+\ell} = -C_{j,j-\ell}$, the equivalent finite-difference formula is centered in space.

Two grid-point values are used in the centered second-order finite-difference approximation to $\partial\psi/\partial x$. A fourth-order centered difference utilizes four points; the sixth-order difference requires six grid points. Every grid-point on the numerical mesh (except the central point) is involved in the spectral approximation of $\partial\psi/\partial x$. As will be shown in the next section, the use of all these grid points allows the spectral method to compute derivatives of smooth functions with very high accuracy. Merilees and Orszag (1979) have compared the weighting coefficients for the spectral method on a seventeen-point periodic grid with the weighting coefficients for second- through sixteenth-order centered finite differences. Their calculations appear in Table 4.1, which shows that the influence of remote grid points on the spectral calculation is much greater than the remote influence in any of the

Δx away from central point	*2nd order*	*4th order*	*6th order*	*16th order*	*spectral*
1	0.500	0.667	0.750	0.889	1.006
2		−0.083	−0.150	−0.311	−0.512
3			0.017	0.113	0.351
4				−0.035	−0.274
5				0.009	0.232
6				−0.001	−0.207
7				0.000	0.192
8				−0.000	−0.186

TABLE 4.1. Comparison of weight accorded each grid point as a function of its distance to the central grid point in centered finite differences and in a spectral method employing 17 expansion coefficients.

finite-difference formulas. The large degree of remote influence in the spectral method has been a source of concern, since the true domain of dependence for the constant-wind-speed advection equation is a straight line. Practical evidence suggests that this remote influence is not a problem provided that enough terms are retained in the truncated Fourier series to adequately resolve the spatial variations in the solution.

Order of Accuracy

The accuracy of a finite difference is characterized by the truncation error, which is computed by estimating a smooth function's values at a series of grid points through the use of Taylor series, and by substituting those Taylor-series expansions into the finite-difference formula. The discrepancy between the finite-difference calculation and the true derivative is the truncation error and is usually proportional to some power of the grid interval. A conceptually similar characterization of accuracy is possible for the computation of spatial derivatives via the spectral method.

The basic idea is to examine the difference between the actual derivative of a smooth function and the approximate derivative computed from the spectral representation of the same function. Suppose that a function $\psi(x)$ is periodic on the domain $-\pi \le x \le \pi$ and that the first few derivatives of ψ are continuous. Then ψ and its first derivative can be represented by the convergent Fourier series

$$\psi(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx}, \tag{4.20}$$

and

$$\frac{\partial \psi}{\partial x} = \sum_{k=-\infty}^{\infty} ika_k e^{ikx}.$$

If ψ is represented by a spectral approximation, the series will be truncated at some wave number N, but the Fourier coefficients for all $|k| \le N$ will be identical to those in the infinite series (4.20).[3] Thus, the error in the spectral representation of $\partial\psi/\partial x$ is

$$E = \sum_{|k|>N} ika_k e^{ikx}.$$

If the pth derivative of ψ is piecewise continuous, and all lower-order derivatives are continuous, the Fourier coefficients satisfy the inequality

$$|a_k| \le \frac{C}{|k|^p}, \tag{4.21}$$

where C is a positive constant (see Problem 10). Thus,

$$|E| \le 2 \sum_{k=N+1}^{\infty} \frac{C}{|k|^{p-1}} \le 2C \int_N^{\infty} \frac{ds}{s^{p-1}} = \frac{2C}{p-2}\left(\frac{1}{N^{p-2}}\right).$$

As demonstrated in the preceding section, a $2N + 1$ mode spectral representation of the derivative is equivalent to some finite-difference formula involving $2N+1$ grid points equally distributed throughout the domain. The spectral computation is therefore equivalent to a finite-difference computation with grid spacing $\Delta x_e = 2\pi/(2N + 1)$. Thus $\Delta x_e \propto N^{-1}$ and

$$|E| \le \tilde{C}(\Delta x_e)^{p-2}, \tag{4.22}$$

where $\tilde{C}$ is another constant. It follows that the effective order of accuracy of the spectral method is determined by the smoothness of ψ. If ψ is infinitely differentiable, the truncation error in the spectral approximation goes to zero faster than any finite power of Δx_e. In this sense, spatial derivatives are represented with infinite-order accuracy by the spectral method.

The preceding error analysis suggests that if a Fourier series approximation to $\psi(x)$ (as opposed to $d\psi/dx$) is truncated at wave number N, the error will be $O(1/N^{p-1})$. This error estimate is actually too pessimistic. As noted by Gottlieb and Orszag (1977, p. 26), (4.21) can be tightened, because if the pth derivative of ψ is piecewise continuous and all lower-order derivatives are continuous,

$$|a_k| \ll \frac{1}{|k|^p} \qquad \text{as} \qquad k \to \pm\infty. \tag{4.23}$$

[3] In order to ensure that the a_k are identical in both the infinite and truncated Fourier series, it is necessary to compute the integral in the Fourier transform,

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi(x)e^{-ikx}\,dx,$$

with sufficient accuracy to avoid aliasing error.

Away from the points where $d^p\psi/dx^p$ is discontinuous, the maximum-norm error in the truncated Fourier series decays at a rate similar to the magnitude of the first few neglected Fourier coefficients, and according to (4.23) this rate is faster than $O(1/N^p)$. In practice, the error is $O(1/N^{p+1})$ away from the points where $d^p\psi/dx^p$ is discontinuous and $O(1/N^p)$ near the discontinuities (Fornberg 1996, p. 13).

Time-Differencing

The spectral representation of the spatial derivatives reduces the original partial differential equation to the system of ordinary differential equations (4.9). In most practical applications, this system must be solved numerically. The time-differencing schemes discussed in Section 2.3 provide a number of possibilities, among which the leapfrog and Adams–Bashforth methods are the most common choices. Integrations performed using the spectral method typically require smaller time steps than those used when spatial derivatives are computed with low-order finite differences. This decrease in the maximum allowable time step is a natural consequence of the spectral method's ability to correctly resolve the spatial gradient in a $2\Delta x$ wave.

In order to better examine the source of this time-step restriction, consider the case of advection by a constant wind speed c. When c is constant, the time dependence of the kth Fourier component is governed by (4.13), which is just the oscillation equation (2.30) with $\kappa = ck$. The maximum value of $|k|$ is $N = \pi/\Delta x_e - \frac{1}{2} \approx \pi/\Delta x_e$, where Δx_e is the equivalent mesh size introduced in connection with (4.22) and $\pi/\Delta x_e \gg \frac{1}{2}$ if the total domain $[-\pi, \pi]$ is divided into at least ten grid intervals. If the oscillation equation is integrated using the leapfrog scheme, the stability requirement is $|\kappa\Delta t| \leq 1$. Thus, the time step in a stable leapfrog spectral solution must satisfy $|c\Delta t/\Delta x_e| \leq 1/\pi$.

Now suppose the advection equation (4.11) is approximated using a centered second-order spatial difference. The time evolution of the approximate solution at the jth grid point is, once again, governed by the oscillation equation. In this case, however, $\kappa = -c\sin(k\Delta x)/\Delta x$. The misrepresentation of the shorter wavelengths by the finite difference reduces the maximum value of $|\kappa|$ to $c/\Delta x$, and the leapfrog stability criterion relaxes to $|c\Delta t/\Delta x| \leq 1$. If higher-order finite differences are used, the error in the shorter wavelengths is reduced, and the maximum allowable value of $|c\Delta t/\Delta x|$ decreases to 0.73 for a fourth-order difference, and to 0.63 for a sixth-order difference. Machenhauer (1979) notes that as the order of a centered finite-difference approximation approaches infinity, the maximum value of $|c\Delta t/\Delta x|$ for which the scheme is stable approaches $1/\pi$. The maximum stable time step for the leapfrog spectral method is therefore consistent with the interpretation of the spectral method as an infinite-order finite-difference scheme.

4.2.2 Improving Efficiency Using the Transform Method

The computational effort required to obtain spectral solutions to the advection equation ceases to be trivial if there are spatial variations in the wind speed. In

such circumstances, the Galerkin requirement (4.9) becomes

$$\frac{da_k}{dt} = -\frac{i}{2\pi}\sum_{n=-N}^{N} na_n \int_{-\pi}^{\pi} c(x,t)e^{i(n-k)x}\,dx. \tag{4.24}$$

Although it may be possible to evaluate the integrals in (4.24) exactly for certain special flows, in most instances the computation must be done by numerical quadrature. In a typical practical application, $c(x,t)$ would be available at the same spatial resolution as $\psi(x,t)$; indeed, many models might include equations that simultaneously predict c and ψ. Suppose, therefore, that c is given by the Fourier series

$$c(x,t) = \sum_{m=-N}^{N} c_m(t)e^{imx}. \tag{4.25}$$

Substitution of this series expansion into (4.24) gives

$$\frac{da_k}{dt} = -\frac{i}{2\pi}\sum_{n=-N}^{N}\sum_{m=-N}^{N} nc_m a_n \int_{-\pi}^{\pi} e^{i(n+m-k)x}\,dx,$$

which reduces, by the orthogonality of the Fourier modes, to

$$\frac{da_k}{dt} = -\sum_{\substack{m+n=k \\ |m|,\,|n|\leq N}} inc_m a_n. \tag{4.26}$$

The notation below the summation indicates that the sum should be performed for all indices n and m such that $|n| \leq N$, $|m| \leq N$, and $n + m = k$.

Although the expression (4.26) is relatively simple, it is not suitable for implementation in large, high-resolution numerical models. The number of arithmetic operations required to evaluate the time derivative of the kth Fourier coefficient via (4.26) is proportional to the total number of Fourier coefficients, $M \equiv 2N+1$. The total number of operations required to advance the solution one time step is therefore $O(M^2)$. On the other hand, the number of calculations required to evaluate a finite-difference formula at an individual grid point is independent of the total number of grid points. Thus, assuming that there are M points on the numerical grid, a finite-difference solution may be advanced one time step with just $O(M)$ arithmetic operations. Spectral models are therefore less efficient than finite-difference models when the approximate solution is represented by a large number of grid points—or, equivalently, a large number of Fourier modes. Moreover, the relative difference in computational effort increases rapidly with increases in M. As a consequence, spectral models were limited to just a few Fourier modes until the development of the transform method by Orszag (1970) and Eliasen et al. (1970).

The key to the transform method is the efficiency with which fast Fourier transforms can be used to swap the solution between wave-number space and physical space. Only $O(M \log M)$ operations are needed to convert a set of M Fourier coefficients, representing the Fourier transform of $\phi(x)$, into the M grid-point values

$\phi(x_j)$.[4] The basic idea behind the transform method is to compute product terms like $c\partial\psi/\partial x$ by transforming c and $\partial\psi/\partial x$ from wave number space to physical space (which takes $O(M \log M)$ operations), then multiplying c and $\partial\psi/\partial x$ at each grid point (requiring $O(M)$ operations), and finally transforming the product back to wave-number space (which again uses $O(M \log M)$ operations). The total number of operations required to evaluate $c\partial\psi/\partial x$ via the transform technique is therefore $O(M \log M)$, and when the number of Fourier components is large, it is far more efficient to perform these $O(M \log M)$ operations than the $O(M^2)$ operations necessary for the direct computation of (4.26) in wave-number space. In order to appreciate the degree to which the transform method can improve efficiency, suppose the spectral method is used in a two-dimensional problem in which the spatial dependence along each coordinate is represented by 128 Fourier modes; then an order-of-magnitude estimate of the increase in speed allowed by the transform method is

$$O\left(\frac{128 \times 128}{\log_2(128 \times 128)}\right) = O(1000).$$

The transform method is implemented as follows. Suppose that one wishes to determine the Fourier coefficients of the product of $\phi(x)$ and $\chi(x)$ such that

$$\phi(x)\chi(x) = \sum_{k=-K}^{K} p_k e^{ikx},$$

where ϕ and χ are periodic on the interval $0 \le x \le 2\pi$ and

$$\phi(x) = \sum_{m=-K}^{K} a_m e^{imx}, \quad \chi(x) = \sum_{n=-K}^{K} b_n e^{inx} \tag{4.27}$$

As just discussed, it is more efficient to transform ϕ and χ to physical space, compute their product in physical space, and to transform the result back to wave-number space than to compute p_k from the "convolution sum"

$$p_k = \sum_{\substack{m+n=k \\ |m|,\,|n| \le K}} a_m b_n.$$

The values of p_k obtained with the transform technique will be identical to those computed by the preceding summation formula, provided that there is sufficient spatial resolution to avoid aliasing error[5] during the computation of the product terms on the physical-space mesh. Suppose that the physical-space mesh is defined such that

$$x_j = \frac{2\pi j}{2N+1}, \quad \text{where} \quad j = 1, \ldots, 2N+1. \tag{4.28}$$

It might appear natural to chose $N = K$, thereby equating the number of gridpoints on the physical mesh with the number of Fourier modes. It is, however, necessary to choose $N > K$ in order to avoid aliasing error.

The amount by which N must exceed K may be most easily determined using the graphical diagram shown in Fig. 4.2, which is similar to Fig. 3.9 of Section 3.5.1. The wave number is plotted along the horizontal axis; without loss of generality, only positive wave numbers will be considered. The cutoff wave number in the original expansion is K, and the cutoff wave number on the high-resolution physical mesh is $\pi/\Delta x_e$. Any aliasing error that results from the multiplication of waves with wave numbers k_1 and k_2 will appear at wave number $\tilde{k} = k_1 + k_2 - 2\pi/\Delta x_e$. The goal is to choose a sufficiently large value for $\pi/\Delta x_e$ to guarantee that no finite-amplitude signal is aliased into those wave numbers retained in the original Fourier expansion, which lie in the interval $-K \le k \le K$. The highest wave number that will have nonzero amplitude after computing the binary product on the physical mesh is $2K$. Thus, there will be no aliasing error if

$$K < \left| 2K - \frac{2\pi}{\Delta x_e} \right| = \frac{2\pi}{\Delta x_e} - 2K.$$

Using the definition of Δx_e implied by (4.28), the criteria for the elimination of aliasing error reduces to $N > (3K-1)/2$.

The preceding result may be verified algebraically by considering the formula for $\tilde{p}_k$, the kth component of the finite Fourier transform computed from the gridpoint values of $\phi\chi$ on the physical mesh,

$$\tilde{p}_k \equiv \frac{1}{M} \sum_{j=1}^{M} \phi(x_j)\chi(x_j)e^{-ikx_j}. \tag{4.29}$$

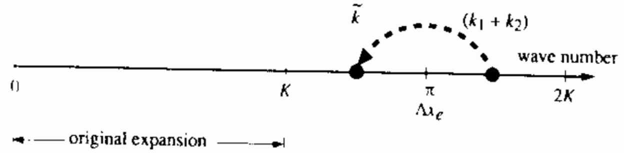

FIGURE 4.2. Aliasing of $k_1 + k_2$ into $\tilde{k}$ such that $|\tilde{k}|$ appears as the symmetric reflection of $k_1 + k_2$ about the cutoff wave number on the high-resolution physical mesh.

[4] To be specific, if M is a power of two, a transform can be computed in $2M \log_2 M$ operations using the FFT algorithm.

[5] Aliasing error occurs when a short-wavelength fluctuation is sampled at discrete intervals and misinterpreted as a longer-wavelength oscillation. As discussed in Section 3.5.1, aliasing error can be generated in attempting to evaluate the product of two poorly resolved waves on a numerical mesh.

Here $M = 2N + 1$ is the total number of grid points on the physical mesh. Let the values of $\phi(x_j)$ and $\chi(x_j)$ appearing in the preceding formula be expressed in the form

$$\phi(x_j) = \sum_{m=-N}^{N} a_m e^{imx_j}, \qquad \chi(x_j) = \sum_{n=-N}^{N} b_n e^{inx_j},$$

where those values of a_n and b_m that were not included in the original series expansions (4.27) are zero, i.e.,

$$a_\ell = b_\ell = 0 \quad \text{for} \quad K < |\ell| \leq N. \tag{4.30}$$

Substituting these expressions for $\phi(x_j)$ and $\chi(x_j)$ into (4.29), one obtains

$$\tilde{p}_k = \sum_{m=-N}^{N} \sum_{n=-N}^{N} a_m b_n \left(\frac{1}{M} \sum_{j=1}^{M} e^{i(m+n-k)x_j} \right). \tag{4.31}$$

Since $x_j = 2\pi j/M$, each term in the inner summation in (4.31) is unity when $m + n - k$ is 0, M, or $-M$. The inner summation is zero for all other values of m and n by the discrete orthogonality condition (4.18). Thus, (4.31) may be written

$$\tilde{p}_k = \sum_{\substack{m+n=k \\ |m|,\,|n| \leq N}} a_m b_n + \sum_{\substack{m+n=k+M \\ |m|,\,|n| \leq N}} a_m b_n + \sum_{\substack{m+n=k-M \\ |m|,\,|n| \leq N}} a_m b_n, \tag{4.32}$$

where the last two terms represent aliasing error, only one of which can be nonzero for a given value of k. The goal is to determine the minimum resolution required on the physical mesh (or, equivalently, the smallest M) that will prevent aliasing errors from influencing the value of p_k^* associated with any wave number retained in the original Fourier expansion. Any aliasing into a negative wave number will arise through the summation

$$\sum_{\substack{m+n=k+M \\ |m|,\,|n| \leq N}} a_m b_n.$$

If follows from (4.30) that $a_m b_n = 0$ if $m + n > 2K$, so for a given wave number k all the terms in the preceding summation will be zero if $m + n = k + M > 2K$. Thus, there will be no aliasing error in p_k^* for those wave numbers retained in the original expansion if M satisfies $-K + M > 2K$. An equivalent condition expressed in terms of the wave number N is $N > (3K - 1)/2$, which is the same result obtained using Fig. 4.2. A similar argument may be used to show that this same condition also prevents the third term in (4.32) from generating aliasing error in the retained wave numbers. The choice $N = 3K/2$ is therefore sufficient to ensure that $p_k = \tilde{p}_k$ for all $|k| \leq K$ and guarantee that the transform method yields the same algebraic result as the convolution sum in wave-number space. In order to maximize the efficiency of the fast Fourier transforms used in practical applications, N is often chosen to be the smallest integer exceeding $(3k - 1)/2$ that contains no prime factor larger than five.

The procedure used to implement the transform method may be summarized as follows. In order to be concrete, suppose that a solution to the variable-wind-speed advection equation is sought on the periodic domain $0 \leq x \leq 2\pi$ and that $c(x,t)$ is being simultaneously predicted by integrating a second unspecified equation. Let both ϕ and c be approximated by Fourier series expansions of the form (4.14) and (4.25) with cutoff wave numbers $N = K$.

1. Pad the coefficients in the Fourier expansions of c and ϕ with zeros by defining $a_k = c_k = 0$, for $K < |k| \leq 3K/2$.
2. Multiply each a_k by ik to compute the derivative of ϕ in wave-number space.
3. Perform two inverse FFTs to obtain $c(x_j)$ and $\partial\phi(x_j)/\partial x$ on the physical-space grid, whose nodal points are located at $x_j = 2\pi j/(3K + 1)$.
4. Compute the product $c(x_j)\partial\phi(x_j)/\partial x$ on the physical-space grid.
5. (If terms representing additional forcing are present in the governing equation, and if those terms are more easily evaluated on the physical mesh than in wave-number space, evaluate those terms now and add the result to to $c(x_j)\partial\phi(x_j)/\partial x$.)
6. Fast-Fourier transform $c(x_j)\partial\phi(x_j)/\partial x$ to obtain the total forcing at each wave number, i.e., to get the right-hand-side of (4.26). Discard the forcing at wave numbers for which $|k| > K$.
7. Step the Fourier coefficients forward to the next time level using an appropriate time-differencing scheme.

Note that the transform method allows processes that are difficult to describe mathematically in wave-number space to be conveniently evaluated during the portion of the integration cycle when the solution is available on the physical mesh. For example, if ϕ represents the concentration of water vapor, any change in ϕ produced by the condensation or evaporation of water depends on the degree to which the vapor pressure at a given grid point exceeds the saturation vapor pressure. The degree of supersaturation is easy to determine in physical space but very difficult to assess in wave-number space.

4.2.3 Conservation and the Galerkin Approximation

The mathematical equations describing non-dissipative physical systems often conserve domain averages of quantities like energy or momentum. When spectral methods are used to approximate such systems, the numerical solution replicates some of the important conservation properties of the true solution. In order to examine the conservation properties of the spectral method for a relatively general class of problems consider those partial differential equations of the form (4.1) for which the forcing has the property that $\overline{vF(v)} = 0$, where the overbar denotes the integral over the spatial domain and v is any sufficiently smooth function that satisfies the boundary conditions.

An example of this type of problem is the simulation of passive tracer transport by nondivergent flow in a periodic spatial domain, which is governed by the equation

$$\frac{\partial \psi}{\partial t} + \mathbf{v} \cdot \nabla \psi = 0.$$

In this case $F(v) = \mathbf{v} \cdot \nabla v$. One can verify that $\overline{vF(v)} = 0$ if v is any periodic function with continuous first derivatives by noting that

$$\int_D v(\mathbf{v} \cdot \nabla v)\, dV = \frac{1}{2}\int_D \nabla \cdot (v^2 \mathbf{v}) - v^2(\nabla \cdot \mathbf{v})\, dV = 0,$$

where the second equality follows from periodicity and the nondivergence of the velocity field.

If ϕ is an approximate spectral solution to (4.1) in which the time dependence is not discretized, then

$$\frac{\partial \phi}{\partial t} + F(\phi) = R(\phi), \tag{4.33}$$

where $R(\phi)$ denotes the residual. Suppose that the partial differential equation being approximated is a conservative system for which $\overline{\phi F(\phi)} = 0$, then multiplying (4.33) by ϕ and integrating over the spatial domain yields

$$\frac{1}{2}\overline{\frac{\partial \phi^2}{\partial t}} = \overline{\phi R(\phi)}.$$

The right side of the preceding is zero because ϕ is a linear combination of the expansion functions and $R(\phi)$ is orthogonal to each individual expansion function. As a consequence,

$$\frac{d}{dt}\|\phi\|_2 = 0, \tag{4.34}$$

implying that spectral approximations to conservative systems are not subject to nonlinear instability because (4.34) holds independent of the linear or nonlinear structure of $F(\psi)$. The only potential source of numerical instability is in the discretization of the time derivative.

Neglecting time-differencing errors, spectral methods will also conserve $\overline{\phi}$ provided that $\overline{F(v)} = 0$, where once again v is any sufficiently smooth function that satisfies the boundary conditions. The conservation of $\overline{\phi}$ can be demonstrated by integrating (4.33) over the domain to obtain

$$\frac{\partial \overline{\phi}}{\partial t} = \overline{R(\phi)} \propto \overline{R(\phi)\varphi_0} = 0,$$

where φ_0 is the lowest-wave-number orthogonal expansion function, which is a constant.

4.3 The Pseudospectral Method

The spectral method uses orthogonal expansion functions to represent the numerical solution and constrains the residual error to be orthogonal to each of the expansion functions. As discussed in Section 3.1, there are alternative strategies for constraining the size of the residual. The pseudospectral method utilizes one of these alternative strategies: the collocation approximation, which requires the residual to be zero at every point on some fixed mesh. Spectral and pseudospectral methods might both represent the solution with the same orthogonal expansion functions; however, as a consequence of the collocation approximation, the pseudospectral method is basically a grid-point scheme—series expansion functions are used only to compute derivatives.

In order to illustrate the pseudospectral procedure, suppose that solutions are sought to the advection equation (4.11) on the periodic domain $0 \le x \le 2\pi$ and that the approximate solution ϕ and the spatially varying wind speed $c(x)$ are represented by Fourier series truncated at wave number K:

$$\phi(x,t) = \sum_{n=-K}^{K} a_n e^{inx}, \quad c(x,t) = \sum_{m=-K}^{K} c_m e^{imx}.$$

The collocation requirement at grid point j is

$$\sum_{n=-K}^{K} \frac{da_n}{dt} e^{inx_j} + \sum_{m=-K}^{K} c_m e^{imx_j} \sum_{n=-K}^{K} ina_n e^{inx_j} = 0. \tag{4.35}$$

Enforcing $R(\phi(x_j)) = 0$ at $2K+1$ points on the physical-space grid leads to a solvable linear system for the time derivatives of the $2K+1$ Fourier coefficients. In the case of the Fourier spectral method, the most efficient choice for the location of these points is the equally spaced mesh

$$x_j = j\left(\frac{2\pi}{2K+1}\right), \quad j = 1, 2, \ldots, 2K+1. \tag{4.36}$$

There is no need actually to solve the linear system for the da_k/dt. It is more efficient to write (4.35) in the equivalent form

$$\frac{d\phi}{dt}(x_j) + c(x_j)\frac{\partial \phi}{\partial x}(x_j) = 0, \tag{4.37}$$

where

$$\frac{\partial \phi}{\partial x}(x_j) = \sum_{n=-K}^{K} ina_n e^{inx_j}. \tag{4.38}$$