DESCRIPTION AND PHILOSOPHY OF SPECTRAL METHODS

Philip S. Marcus

Massachusetts Institute of Technology

We describe the use of spectral methods in computational fluid dynamics. Spectral methods are generally more accurate and often faster than finite-differences. For example, the ∇^2 operator in 2 or 3 dimensions is easier to invert with spectral techniques because the spatial dependence of the operator separates in a more natural way. We warn against the use of some of the more common spectral expansions. Bessel series expansions of functions in cylindrical geometries converge poorly. However, other series expansions of the same functions converge quickly. We show how to choose basis functions that give fast convergence and outline the differences between Galerkin, tau, modal, collocation, and pseudo-spectral methods.

INTRODUCTION

After perfecting a numerical code, it is tempting to try and find every physical and astrophysical problem that one can solve with the code. However, in developing a code in the first place, one is often motivated by some particular physical or astrophysical calculation. If motivated by solar convection and rotation, the hydrodynamics appears to be easy: there is no relativity; there are no shocks; the flow is at low Mach number and is smooth. On the other hand, the numerical simulation of these flows is very difficult due to turbulence. In turbulence, many decades of length scales are excited and contribute to the dynamics. To have correct numerical results, it is necessary to resolve or model all of the physically important scales. In choosing a numerical scheme that is useful in computing turbulent flow, we shall need a method that is efficient in three dimensions (since turbulence is inherently

K.-H. A. Winkler and M. L. Norman (eds.), Astrophysical Radiation Hydrodynamics, 359–386.

three-dimensional) and provides the greatest possible spatial resolution. Spectral methods are ideal for these flows.

TWO FAMILIAR EXAMPLES OF SPECTRAL METHODS

Before giving an exact definition of a spectral method, we provide two examples to show that most readers have already used spectral techniques.

Heat Equation

The first example is an analytic calculation similar to a graduate-level electrostatics problem with boundary conditions. Consider the one-dimensional heat equation:

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} \tag{1}$$

with boundary conditions:

$$T(t,0) = T(t,1) = 0 \tag{2}$$

and initial condition

$$T(0,x) = f(x). \tag{3}$$

One way of solving the heat equation is to expand T in a sine series

$$T(t,x) = \sum_{m=1}^{\infty} a_m(t) \sin(m\pi x) \tag{4}$$

The sines are a natural choice because they are a complete set of basis functions, and each sine individually obeys the boundary conditions. Furthermore, the sines are eigenfunctions of the second derivative operator that appears on the right-hand side of equation (1). Fourier analyzing the initial data, we obtain

$$f(x) = \sum_{m=1}^{\infty} f_m \sin(m\pi x). \tag{5}$$

Substituting T(t,x) from equation (4) into the heat equation and comparing the coefficients of the sines on both sides of the equation [or to be more formal, multiplying both sides of the heat equation by $\sin(m'\pi x)$ and integrating both sides over x from zero to one], we obtain an infinite set of ordinary differential equations for the spectral coefficients $a_m(t)$:

$$\frac{\partial a_m(t)}{\partial t} = -\kappa m^2 a_m(t) \quad m = 1,2,\ldots \infty \tag{6}$$

Equation (6) can be integrated analytically using the initial condition $a_m(0) = f_m$.

$$a_m(t) = f_m \exp(-\kappa m^2 t) \quad m = 1,2,\ldots, \infty \tag{7}$$

Equation (7) is the analytic spectral solution to the heat equation. To find a numerical spectral solution to the heat equation we simply replace $T(t,x)$by the discrete approximation $T_N(t,x)$,

$$T_N(t,x) \tilde{=} T(t,x), \tag{8}$$

where we have discretized by the simple truncation

$$T_N(x,t) \equiv \sum_{m=1}^{N} a_m(t)\sin(m\pi x) \tag{9}$$

We need to know how well $T_N(t,x)$ approximates the exact solution of the heat equation. One way of measuring the error caused by the approximation is to calculate the mean-square error, L_2, which is defined:

$$L_2 \equiv \left[\int_0^1 |T(t,x) - T_N(t,x)|^{\,2}\, dx\right]^{1/2} \tag{10}$$

Using equations (7) and (9), we find that the mean-square error becomes

$$L_2 = \left[\frac{1}{2} \sum_{m=N+1}^{\infty} |f_m|^2 \, e^{-2\kappa m^2 t}\right]^{1/2} \tag{11}$$

We see immediately that the error depends solely on the spectral coefficients that we have discarded in the truncation (i.e.,those spectral coefficients a_m with $m > N$). Equation (11) shows that for $t > 0$ and $\kappa > 0$, the mean-square error decays exponentially with increasing numerical resolution, N. The hallmark of a good spectral method is that the error decreases exponentially with increasing resolution; whereas, in a second-order accurate finite-difference method,the error (by definition) decreases only as the square of the spatial resolution.

To compare directly the error due to finite-differences with the error due to spectral expansion, we use a spectral analysis to evaluate the finite-difference approximation of the second derivative operator used in the heat equation. Let the initial temperature distribution contain only one Fourier component

$$f(x) = \sin(p\pi x) \tag{12}$$

Using a centered, second-order accurate finite-difference operator, the second derivative of $\sin(p\pi x)$ at x_i is

$$\left.\frac{d^2\sin(p\pi x)}{d\,x^2}\right|_{x_i} = \frac{\sin[p\pi(x_i+\Delta x)]+\sin[p\pi(x_i-\Delta x)]-2\sin(p\pi x_i)}{(\Delta x)^2}$$

$$= \frac{2[\cos(p\pi\Delta x)-1]}{\Delta x^2}\sin(p\pi x_i) \tag{13}$$

$$\equiv \lambda_{f.-d.}\sin(p\pi x_i)$$

The exact second derivative of $\sin(p\pi x)$ is

$$\left.\frac{d^2}{dx^2}\sin(p\pi x)\right|_{x_i} = -\,p^2\pi^2\sin(p\pi x_i)$$

$$\equiv \lambda_{exact}\sin(p\,\pi x_i) \tag{14}$$

We see that the finite-difference operator gives the exact eigenfunction but an incorrect eigenvalue. If Δx is small, we can Taylor expand the finite-difference eigenvalue and thereby determine the fractional error

$$\left|\frac{\lambda_{f.-d.}-\lambda_{exact}}{\lambda_{exact}}\right| \cong \frac{\pi^2 p^2\Delta x^2}{12}\,. \tag{15}$$

If we want the eigenvalue correct to within one percent we require that

$$p\,\Delta x \lesssim 0.11 \tag{16}$$

Since there are $p/2$ wavelengths of $f(x)$ between 0 and 1 and since there are $1/\Delta x$ finite-difference grid points between 0 and 1, equation (16) tells us that we need approximately 20 grid points per wavelength in order to achieve one percent accuracy. This factor of 20 should be compared to the fact that we need only one Fourier mode per wavelength to get perfect accuracy in the numerical second derivative eigenvalue when the derivatives are computed spectrally.

The suspicious reader can argue that the preceding comparison is unfair because our spectral expansion of f(x) uses basis functions that are the exact analytic eigenfunctions of the second derivative operator. However, we shall show soon that expanding f(x) in any other set of suitable basis functions require only three or four modes per wavelength to obtain one percent accuracy; whereas, second-order finite difference methods always need approximately 20 grid points per wavelength to obtain the same accuracy. Furthermore, choosing the basis functions to be the set of analytic eigenfunctions of the second derivative operator is not always wise. We show later that solving the heat equation in cylindrical coordinates with an expansion in the eigenfunctions of the second derivative operator has disastrous consequences .

Quadrature

As a second example of a spectral method, we consider numerical quadrature. One method of numerically integrating a function is to use a Newton-Cotes quadrature formula. These formulae are really just finite-difference methods. Let f(x) be a function that is tabulated at equally spaced intervals x_i, where $x_{i+1} - x_i \equiv \Delta x$. We numerically integrate f(x) from x_{i-1} to x_{i+1} by replacing the function f(x) in the interval $[x_{i-1}, x_{i+1}]$ with the parabola that passes through the three points $(x_{i-1}, f(x_{i-1}))$, $(x_i, f(x_i))$ and $(x_{i+1}, f(x_{i+1}))$. Analytic integration of the parabola gives Simpson's rule:

$$\int_{x_{i-1}}^{x_{i+1}} f(x)\, dx = \frac{\Delta x}{3} (f(x_{i+1}) + 4f(x_i) + f(x_{i-1})) + \mathcal{O}(\Delta x^5) \tag{17}$$

To see that Simpson's rule is really a second-order finite difference method, we expand f(x) in a Taylor series about $x = x_i$. Integrating the Taylor series from x_{i-1} to x_{i+1} we obtain

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx = \int_{x_i-\Delta x}^{x_i+\Delta x} (f(x_i)+xf'(x_i) + \frac{x^2}{2} f''(x_i) + \ldots)dx =$$

$$= 2\Delta x f(x_i) + \frac{1}{3}\Delta x^3 f''(x_i) + \mathcal{O}(\Delta x^5) \tag{18}$$

In order to evaluate this integral, we need to determine $f''(x_i)$. If we substitute the second-order centered finite difference

$$f''(x_i) = \frac{f(x_{i+1}) + f(x_{i-1}) - 2f(x_i)}{\Delta x^2} + \mathcal{O}(\Delta x^2) \tag{19}$$

into equation (18), then we recover Simpson's rule.

An alternative approach to quadrature is Gauss's method. In Gaussian quadrature, we abandon the constraint of equally spaced sampling points. The numerical integral of f(x) from -1 to 1 is approximated by a linear sum of weights, w_i, multiplied by f(x) evaluated at the sampling points.

$$\int_{-1}^{1} f(x)dx \cong \sum_{i=0}^{N} f(x_i)w_i \tag{20}$$

In equation (20) there are 2(N+1) unknown quantities: (N+1) x_i's and (N+1) w_i's. To determine these unknowns we impose 2(N+1) equations: equation (20) must be satisfied exactly for f(x) = 1, f(x) = x, f(s) = x^2,..., f(x) = x^{2N}, f(x) = x^{2N+1}. Equivalently, the quadrature formula in equation (20) must be exact for all polynomials of order (2N+1) or less. The well-known solution to this problem is that the x_i are roots of the Legendre polynomial of order N+1.

$$P_{N+1}(x_i) = 0 \qquad i = 0,N \tag{21}$$

What are we really doing when we use Gaussian quadrature? We are actually replacing f(x) with the spectral approximation, $f_N(x)$

$$f_N(x) \equiv \sum_{0}^{N} a_n P_n(x) \tag{22}$$

where the $P_n(x)$ are Legendre polynomials and where the coefficients a_n, are determined by the method of least squares (cf. Dahlquist, et al., 1979). The quadrature formula in equation (20) is the result of an analytic integration of $f_N(x)$.

DEFINITIONS OF SPECTRAL METHODS - SELECTION OF BASIS FUNCTIONS

Spectral methods are useful in numerical calculations because they allow us to represent a continuous function, f(x), as a discrete approximation, $f_N(x)$. The approximation is written as a finite sum of basis functions multiplied by coefficients.

$$f(x) \cong f_N(x) = \sum_{i=1}^{N} a_i \phi_i(x) \tag{23}$$

The basis functions, $\phi_i(x)$, are arbitrary and do not have to be eigenfunctions or orthogonal. The two tasks of the numericist are to: (1) select a set of basis functions $\phi_i(x)$ and (2) compute the coefficients, a_i. Both the basis functions and the method of computing the coefficients should be chosen so that the boundary and initial conditions are easily satisfied, the spectral sum converges quickly, and both the numericist and computer have a minimal amount of work to perform. The choice of basis functions and the manner of computing the coefficients determines a method's name, such as modal, Galerkin, spectral, tau, collocation, pseudospectral, or Rayleigh-Ritz.

Fourier Series and the 9 % Solution.

We first consider the task of choosing basis functions that make up the spectral series. The simplest choice of a Fourier series with $\phi_k(x) = e^{ikx}$. We remind the reader of an important theorem about Fourier series: if f(x) is a continuous, piecewise function over the domain 0 to 2π, where f(x) is of bounded variation, and if Fourier coefficients, a_k, are defined by

$$a_k \equiv \frac{1}{2\pi} \int_0^{2\pi} f(x)\ e^{-ikx} dx\ , \tag{24}$$

and if the spectral sum g(x) is defined by

$$g(x) \equiv \sum_{k=-\infty}^{\infty} a_k\ e^{ikx}\ , \tag{25}$$

then

$$g(x) = 1/2\ \left[f(x^+) - f(x^-)\right]. \tag{26}$$

This theorem means that if f(x) is a continuous function, then the Fourier series, g(x), is equal to f(x) at every point. If there is a discontinuity in the function, then g(x) is equal to the arithmetic mean of f(x) at the discontinuity. It is important to realize that for a numericist this theorem has no practical value. In numerical approximations we calculate the partial sums $f_N(x)$

$$f_N(x) \equiv \sum_{k=-N}^{N} a_k \, e^{ikx} \tag{27}$$

The preceding theorem does not guarantee that $f_N(x)$ converges uniformly to $g(x)$ or $f(x)$ as N approaches infinity. In fact, near a discontinuity, $f_N(x)$ never uniformly converges to $f(x)$ or $g(x)$. This lack of convergence is well-known to anyone who has Fourier analyzed a step function and calculated the Gibbs overshoot of approximately 9 % at the discontinuity. If more Fourier modes are included in the partial sum, the 9 % error does not decrease. If the function, $f(x)$, is itself continuous but has a discontinuity in one of its derivatives, then $f_N(x)$ may converge to $f(x)$ but the rate of convergence will be poor. One way of measuring the convergence rate is to examine the mean square error, $L_2(N)$ of the N^{th} partial sum:

$$L_2(N) \equiv \left[\int_0^{2\pi} |f(x)-f_N(x)|^2 \, dx \right]^{1/2} \tag{28}$$

$$= \left[2\pi \sum_{|k|=N+1}^{\infty} |a_k|^2 \right]^{1/2} \tag{29}$$

To evaluate the error, we must first determine how the Fourier coefficients, a_k, depend on k. Integrating equation (24) by parts we obtain

$$a_k = \left. \frac{if(x)e^{-ikx}}{2\pi k} \right]_0^{2\pi} + \frac{1}{i2\pi k} \int_0^{2\pi} f'(x)e^{-ikx} dx \tag{30}$$

The surface term in equation (30) vanishes if $f(x)$ is continuous and periodic. If all of the derivatives of $f(x)$ are continuous, then further integration by parts produces no surface terms. Integration by parts of equation (24) m times gives

$$a_k = \frac{1}{(ik)^m 2\pi} \int_0^{2\pi} f^{(m)}(x)e^{-ikx} dx \tag{31}$$

From equation (31), we see that a_k falls off as $1/k^m$ for all functions whose first (m-1) derivatives are continuous and periodic. Equation (29 shows that the mean-square error, $L_2(N)$, decreases at least as fast as $1/N^{m-1}$. If f(x) is a c_∞ function, then the Fourier coefficients and mean-square error decrease exponentially with N. Therefore, the convergence of the partial Fourier sums of a c_∞ periodic function is exponential. This exponential rate of convergence should be compared to second-order finite difference methods where (by definition) the rate of convergence is only quadratic.

If f(x) has a discontinuity at x_o, then integration by parts of equation (24) gives

$$a_k = \frac{i}{2\pi k} e^{-ikx_o} \left[f\right] + \frac{1}{2\pi ik} \int_o^2 f'(x) e^{-ikx} dx \qquad (32)$$

when $[f]$ is the discontinuity in f. The surface term in equation (32) is not zero. No matter how many times we integrate equation (32) by parts, there will always be a contribution to the Fourier coefficient, a_k, that decreases slowly as 1/k. To see how the discontinuity affects the convergence, we write the error in the Fourier partial sum as

$$f(x) - f_N(x) = \sum_{|k|=N+1}^{\infty} a_k \, e^{ikx} \qquad (33)$$

In estimating the sum of the series in equation (33), we can argue that if all the terms have random phases (say, far away from the discontinuity), then the sum is approximately equal to the leading term in the series, a_n, which decreases as 1/N. If the terms in equation (33) are in phase (say, near the discontinuity) then the error in equation (33) is of order unity. The 9 % Gibbs overshoot in the truncated sum made from the Fourier transforms of a square wave is an example of an error of order unity near a discontinuity. Far from the discontinuity the error in the Fourier sum is much smaller and is of order 1/N.

In general, if the lowest order derivative of f(x) that has a discontinuity or non-periodicity is the mth derivative, then the convergence of the partial Fourier sums is of order $(1/N)^{m+1}$ far from the discontinuity and is of order $(1/N)^m$ near the discontinuity. To illustrate this convergence rate, consider the heat equation with an inhomogeneous source term,

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2} + 1. \tag{34}$$

Let $f(x)$ be defined over the domain $[0,1]$ with boundary conditions

$$f(0) = f(1) = 0 \tag{35}$$

Expanding both $f(x)$ and 1 in a Fourier sine series,

$$f(t,x) = \sum_{k=1}^{\infty} f_k(t) \sin(k\pi x) \tag{36}$$

$$1 = \sum_{k=1,\ k=\text{odd}}^{\infty} \frac{4}{k\pi} \sin(k\pi x) \tag{37}$$

and substituting into equation (34), we obtain an ordinary differential equation for each Fourier coefficient $f_k(t)$. Integrating these equations we find

$$\begin{aligned} f_k(t) &= f_k(0)\, e^{-k^2\pi^2 t} && \text{for } k = \text{even} \\ f_k(t) &= f_k(0) e^{-k^2\pi^2 t} + \frac{4}{\pi^3 k^3} (1-e^{-\pi^2 k^2 t}) && \text{for } k = \text{odd} \end{aligned} \tag{38}$$

We see explicitly that the spectral coefficients do not decrease exponentially with k; instead, we find that they decrease only as $(1/k)^3$. The convergence is not exponential; it is cubic. Why doesn't the series converge any faster? To understand the poor rate of convergence we examine the Fourier coefficients of $f(t,x)$ which are defined

$$f_k(t) \equiv 2 \int_0^1 f(t,x) \sin(k\pi x) dx \tag{39}$$

A first integration by parts yields

$$f_k(t) = \frac{-2f(t,x)}{k\pi} \cos(k\pi x) \Big]_0^1 + \frac{2}{k\pi} \int_0^1 f'(t,x) \cos(k\pi x) dx \tag{40}$$

The surface term in equation (40) vanishes because $f(t,x)$ vanishes at 0 and 1 due to the boundary conditions. A second integration by parts gives

$$f_k(t) = \frac{2f'(t,x)}{k^2\pi^2} \sin(k\pi x) \Big]_0^1 - \frac{2}{k^2\pi^2} \int_0^1 f''(t,x) \sin(k\pi x) dx \tag{41}$$

The surface term vanishes again, not due to any property of $f(t,x)$ or its derivatives, but because the sine vanishes at 0 and 1. A third integration by parts gives

$$f_k(t) = \frac{2f''(t,x)\cos(k\pi x)}{k^3\pi^3}\Big]_0^1 - \frac{2}{k^3\pi^3}\int_0^1 f'''(t,x)\cos(k\pi x)dx \qquad (42)$$

This time the surface term is non-zero. The surface term and $f_k(t)$ both decrease as $(1/k)^3$. This is the reason that the convergence rate is cubic and not exponential.

Now that we understand why the rate of convergence is slow we can accelerate it. The surface term from the first integration by parts in equation (40) vanishes because $f(x)$ is zero at the boundaries. The surface term vanishes after the second integration by parts because $\sin(k\pi x)$ vanishes at 0 and 1. The surface term from the third integration by parts does not vanish because $f''(0)$ and $f''(1)$ are not equal to zero. In fact, from the inhomogeneous heat equation (34) we see that

$$f''(0) = f''(1) = -1 \qquad (43)$$

To improve our rate of convergence we forgo expanding $f(x)$ in a sine series and follow the procedure of Gottlieb and Orszag (1977). Fourier expand a new function $g(t,x)$, where

$$g(t,x) \equiv f(t,x) + \frac{x(x-1)}{2} \qquad (44)$$

We have defined $g(t,x)$ so that the boundary conditions on g and g'' are homogeneous:

$$g(t,0) = g(t,1) = g''(t,0) = g''(t,1) = 0 \qquad (45)$$

In fact, by using equations (34) and (44) we find

$$g(t,0)^{(2n)} = g(t,1)^{(2n)} = 0 \text{ for } n=0,1,2,\ldots,\infty \qquad (46)$$

When we compute the Fourier coefficients of $g(t,x)$ by integration by parts, we find that all of the surface terms vanish. The Fourier sine series expansion of $g(x)$ converges exponentially. Equivalently, the partial sums

$$f_N(t,x) \equiv -\frac{x(x-1)}{2} + \sum_{k=1}^{N} f_k(t)\sin(k\pi x) \qquad (47)$$

converge exponentially to the exact solution of the inhomogeneous heat equation which is:

$$f(t,x) = -\frac{x(x-1)}{2} + \sum_{k=1}^{\infty} \hat{f}_k(0)\sin(k\pi x)e^{-\pi^2 k^2 t} \qquad (48)$$

where

$$\hat{f}_k(0) \equiv 2\int_0^{\pi}\left[f(t=0,x) + \frac{x(x-1)}{2}\right]\sin(k\pi x)dx \qquad (49)$$

The leading term of the error, $f(x,t)-f_N(t,x)$, is $\hat{f}_{N+1}(0)\sin[(N+1)\pi x]e^{-\pi^2(N+1)^2 t}$, which decays exponentially in N for $t > 0$.

Polynomial Basis Functions

In approximating a non-periodic function with a Fourier sum, we often find that including an appropriate polynomial in the spectral sum improves the rate of convergence. The method of selecting an appropriate polynomial depends on the boundary conditions of the function that is being approximated. We would like to discuss a more general method of finding polynomials to use in spectral approximations so that the rate of convergence does not explicitly depend on boundary conditions. We abandon sines and cosines as basis functions; instead we use the eigenfunction $\{\phi_n(x)\}$, of a Sturm-Liouville operator:

$$\frac{d}{dx}\left[p(x)\frac{d\phi_n}{dx}\right] + \left[\lambda_n w(x) - q(x)\right]\phi_n(x) = 0 \qquad (50)$$

where

$$p(x) \geq 0 \qquad (51)$$

$$w(x) \geq 0 \qquad (52)$$

$$q(x) \geq 0 \qquad (53)$$

$$a \leq x \leq b \qquad (54)$$

Here, the λ_n are the eigenvalues associated with $\phi_n(x)$ and $w(x)$ is the weighting function that is used to define the inner product. The eigenfunctions are complete and orthonormal with respect to these weighting functions

$$\int_a^b w(x)\phi_n(x)\phi_m(x)\, dx = \delta_{nm} \qquad (55)$$

We can express $f(x)$ as an infinite series in $\phi_n(x)$ and approximate $f(x)$ with the partial sum $f_N(x)$. The mean square error of the approximation with respect to the weighting function $w(x)$ is $L_2(N)$.

$$f(x) = \sum_{n=0}^{\infty} a_n \, \phi(x) \tag{56}$$

$$f_N(x) = \sum_{n=0}^{N} a_n \phi(x) \tag{57}$$

$$L_2(N) = \left[\sum_{N+1}^{\infty} |a_n|^2 \right]^{1/2} \tag{58}$$

where

$$a_n = \int_a^b f(x)\phi_n(x)w(x)dx \tag{59}$$

The error is the square of the truncated spectral coefficients. To show how $L_2(N)$ and a_N decrease with N we follow the procedure of Lanczos (1956), and integrate equation (59) twice by parts to obtain

$$a_n = \frac{1}{\lambda_n} p(x) \left[(\phi_n \frac{df}{dx} - f \frac{d\phi_n}{dx}) \right]_a^b - \frac{1}{\lambda_n} \int_a^b \phi_n(x)h(x)w(x)dx \tag{60}$$

where $h(x)$ is defined

$$h(x) \equiv \frac{1}{w(x)} \frac{d}{dx} (p \frac{df}{dx}) - \frac{q(x)f(x)}{w(x)} \tag{61}$$

The surface term in equation (60) vanishes regardless of the values of $f(x)$ and its derivatives at the boundaries (as long as they remain nonsingular) if $p(a) = p(b) = 0$. When $p(a) = p(b) = 0$ we call the Sturm-Liouville operator singular. For singular operators, each time equation (60) is integrated by parts the surface terms vanish as long as the higher derivatives of $f(x)$ remain bounded. Integrating equation (60) by parts 2p times, shows that the spectral coefficient a_n is of order $(1/\lambda_n)^p$.

As $p \to \infty$, the spectral series, $f_N(x)$ converges exponentially.

As an example of a singular operator we consider the Chebyshev polynomials, $T_n(x)$. The Sturm-Liouville equation that generates the Chebyshev polynomials has

$$p(x) = \sqrt{1-x^2} \tag{62}$$

$$w(x) = 1/\sqrt{1-x^2} \tag{63}$$

$$q(x) = 0 \tag{64}$$

$$\lambda_n = n^2 \tag{65}$$

$$-1 \leq x \leq 1 \tag{66}$$

One way to convince ourselves that the Chebyshev polynomials really do have exponential convergence is to approximate a sine function with a Chebyshev series (cf. Gottlieb and Orszag, 1977). Computing the spectral coefficients from equation (59) we find

$$\sin(m\pi x) = 2 \sum_{\substack{n=1 \\ n=\text{odd}}}^{\infty} \left[J_n(m\pi)(-1)^{(n-1)/2}\right] T_n(x) \tag{67}$$

Only the odd Chebyshev polynomials enter the sum in equation (67) because $\sin(m\pi x)$ is an odd function of x. The coefficients of the Chebyshev series are proportional to Bessel functions, $J_n(m\pi)$. Bessel functions of low order behave like sines, but when their order becomes greater than their argument they exponentially decay. Therefore, for $n > m\pi$ the Chebyshev spectral coefficients decay exponentially, and the convergence of the partial sums is exponential. Between -1 and 1, $\sin(m\pi x)$ has m wavelengths. Since $m\pi$ Chebyshev polynomials are required for exponential convergence, we conclude that approximately π Chebyshev polynomials are needed per wavelength of the function being approximated. This factor of π should be compared to the requirements of second-order finite difference methods where we found that approximately 20 grid points per wavelength are needed to obtain 1% accuracy.

As a second example of polynomial expansions, we use the Legendre polynomials as a set of basis functions. The Legendre polynomials are eigenfunctions of a Sturm-Liouville equation with:

$$p(x) = (1-x^2) \tag{68}$$

$$w(x) = 1 \tag{69}$$

$$q(x) = 0 \tag{70}$$

$$\lambda_n = n(n+1) \tag{71}$$

$$-1 \leq x \leq 1 \tag{72}$$

Again, expressing $\sin(m\pi x)$ as a spectral sum we obtain

$$\sin(m\pi x) = \sum_{\substack{n=1\\n=\text{odd}}}^{\infty} \left[\frac{(2n+1)(-1)^{(n-1)/2}}{(2m)^{1/2}} J_{n+1/2}(m\pi)\right] P_n(x) \tag{73}$$

The coefficients multiplying the Legendre polynomials are Bessel functions of half integral order; they decrease exponentially when their order is greater than their argument. Again, we need on the order of π polynomials per wavelength to obtain exponential convergence.

One way to heuristically see the resolution properties of a spectral sum is to examine the spacings between zero-crossings of the basis functions. Fourier series have equally spaced zeroes which make them well-suited for approximating periodic functions. A truncated Fourier series resolves boundary layers poorly. Legendre and Chebyshev polynomials have more zero crossings near the boundaries at -1 and 1 than they do at the interior of their domain near 0. Near the interior $P_N(x)$ and $T_N(x)$ both have an average separation between zero crossings of π/N. Near the boundaries, the average spacing reduces to $\pi^2/2N^2$. Partial sums of $P_N(x)$ or $T_n(x)$ are well suited for approximation function with boundary-layers. If a boundary layer has thickness of order δ, then approximately $(1/\delta)^{1/2}$ terms are needed in the spectral sum to resolve the boundary layer.

As our last example of choosing basis functions, we consider the one-dimensional axisymmetric heat equation in cylindrical coordinates with an inhomogeneous forcing term:

$$\frac{\partial T}{\partial t} = \kappa\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)T + 1 \tag{74}$$

with boundary conditions $T(t,r=1) = 0$, $T(t,r=0)$ finite, and initial conditions $T(t=0,r) = f(r)$. An apparently obvious choice of basis functions are the Bessel functions of index zero

$$\phi_n(r) \equiv \sqrt{2}\ \frac{J_0(j_{0n}r)}{J_1(j_{0n})} \tag{75}$$

where j_{0n} is the n^{th} root of J_0. The eigenfunctions are normalized to obey equation (55) with weighting function $w(r) = r$. The $J_0(j_{0n}r)$ Bessel function is an eigenfunction of $\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r}\right)$ which is a Sturm-Liouville operator with $p(r) = r$ and eigenvalue jon. Because the $\phi_n(r)$ are complete, obey the same homogeneous boundary conditions as $T(t,x)$, and are eigenfunctions of the differential operator on the right-hand-side of equation (74), they are the natural basis functions for this problem,

$$T(t,x) = \sum_{n=1}^{\infty} a_n(t)\phi_n(x) \tag{76}$$

The exact solution to equation (74) is

$$T(t,x) = \sum_{n=1}^{\infty} \left[f_n\ e^{-j_{on}\kappa t} - \frac{\sqrt{2}}{\kappa(j_{0n})^2} \right] \phi_n(r) \tag{77}$$

where

$$f_n = \int_0^1 \phi_n(r)f(r)r\ dr \tag{78}$$

The solution expanded in terms of the Bessel functions does not converge exponentially. From the asymptotic behavior of j_{0n},

$$j_{0n} \sim \pi\ (n - 1/4), \tag{79}$$

we see that the exact solution converges as $1/n^2$ far from the boundary and as $1/n$ near the boundary. The slow convergence is due to the fact that $p(r) \neq 0$ at the boundary and the Sturm-Liouville operator is not singular. For this problem, second-order accurate finite differences are more efficient than a Bessel expansion. Spectral methods can still be used, but Chebyshev or Legendre polynomials should be employed. Although Chebyshev or Legendre functions do not appear to be "natural" choices for cylindrical geometry, they converge rapidly and work well (Marcus, 1983).

IMPLEMENTATION OF SPECTRAL METHODS

Galerkin's Method

Once the spectral numericist has decided on a set of basis functions, his second task is to compute efficiently the spectral coefficients. Generally, this task requires solving a large set of coupled ordinary differential equations. The most straight-forward of all spectral techniques is a modal or Galerkin method. In Galerkin's method each basis function obeys the same boundary conditions as the function that is being approximated. Consider the combined initial value, boundary-value problem

$$\frac{\partial u}{\partial t} = \mathcal{L}(u) \tag{80}$$

where $\mathcal{L}$ is some arbitrary spatial operator. We impose homogeneous boundary conditions on both u and the basis functions, $\phi_i(x)$

$$u(t,x = -1) = u(t,x=1) = 0 \tag{81}$$

$$\phi_n(-1) = \phi_n(1) = 0 \quad \text{for all } n. \tag{82}$$

As usual, the partial sum u_N is defined

$$u_N(t,x) = \sum_{n=1}^{N} a_n(t)\phi_n(x), \tag{83}$$

The easiest case that can be solved with Galerkin's method occurs when $\mathcal{L}$ is a linear differential operator with constant coefficients and where the $\phi_n(x)$ are eigenfunctions of $\mathcal{L}$. In this case, substitution of equation (83) into equation (80) results in a set of N, non-coupled ordinary differential equations which can be solved easily for the $a_n(t)$.

In a slightly more complicated case, we still restrict $\mathcal{L}$ to be a linear differential operator with constant coefficients, but we no longer require that the $\phi_n(x)$ be eigenfunctions of $\mathcal{L}$. For example, we consider the wave equation with $\mathcal{L} = \frac{d}{dx}$

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x}, \tag{84}$$

with $-1 \le x \le 1$, and with boundary condition,

$$u(1) = 0. \tag{85}$$

Using the basis functions,

$$\phi_n(x) \equiv [P_n(x) - 1] \tag{86}$$

to expand $u(t,x)$, substituting the series in equation (84), and taking the inner product of both sides of the resulting equation with respect to $\phi_n(x)$, we obtain a set of coupled ordinary differential equations for the spectral coefficients $a_n(t)$. Let $\underline{a}$ be the column vector whose n^{th} element is $a_n(t)$. The equations for the spectral coefficients can be written in matrix form:

$$\underline{\underline{P}}\,\dot{\underline{a}} = \underline{\underline{M}}\,\underline{a} \tag{87}$$

The dot above the vector $\underline{a}$ imples that each element is differentiated with respect to time . The elements of $\underline{\underline{M}}$ and $\underline{\underline{P}}$ are easily written in terms of inner products.

$$P_{mn} = (\phi_m,\phi_n) \tag{88}$$

$$M_{mn} = (\phi_m,\mathcal{L}(\phi_n)) \tag{89}$$

In practice, we would never directly use the matrix equation (87) to solve for $a_n(t)$. Matrix arithmetic (multiplication or inversion) requires of order N^2 or N^3 operations (where N is the number of terms in the spectral sum). Usually $\underline{\underline{P}}$ and $\underline{\underline{M}}$ are sparse or have a special form due to the fact that $\mathcal{L}$ implies a recursion relationship among the spectral coefficients. By exploiting the sparsity or symmetry of $\underline{\underline{M}}$,equation (87) can often be solved in only N operations per timestep.

If $\mathcal{L}$ is a nonlinear differential operator or linear differential operator with non-constant coefficients, then Galerkin's method becomes unwieldy. As an example, consider the nonlinear wave equation

$$\frac{\partial u}{\partial t} = - u \frac{\partial}{\partial x} u \tag{90}$$

over the domain $0 \leq x \leq \pi$ and with boundary condition

$$u(t,x=0) = 0 \tag{91}$$

Writing $u(t,x)$ as a Fourier sine series (which is complete over $0 \leq x \leq \pi$) equation (90) becomes

$$\sum_{n=1}^{N} \dot{a}_n(t)\sin(n\pi x) = -\left[\sum_{k=1}^{N} a_k(t)\sin(k\pi x)\right]\frac{d}{dx}\left[\sum_{m=1}^{N} a_m(t)\sin(m\pi x)\right] \tag{92}$$

$$= -\frac{\pi}{2}\sum_{n=1}^{N-1}\sum_{m=1}^{N-n} m a_m a_{n+m}\sin(n\pi x) - \frac{\pi}{2}\sum_{n=1}^{2N'}\sum_{\substack{m=n-N\\ m>0}}^{N} \operatorname{sgn}(n-m) m a_m a_{|m-n|}\sin(n\pi x) \tag{93}$$

where sgn(n-m) is the sign of (n-m) and is equal to zero if (n-m)=0. The right-hand side of equation (93) is a convolution product. The convolution leads to three problems: first, it couples all of the ordinary differential equations for $a_n(t)$; second, the convolution product is expensive to compute - it requires (on the order of) N^2 multiplications; thirdly, the convolution product of two partial sums of spectral modes generates spectral modes that are not included in or spanned by the original finite set of basis functions. The convolution of a sine series with N modes and its derivative produces a product containing 2N sine modes. In previous cases we found that when a linear differential operator with constant coefficients operates on a finite sum of basis functions, the output is itself spanned by this same finite set of basis functions; in these cases we say that the finite set of basis functions is closed with respect to that operator. In general, a finite set of basis functions is not closed with respect to a nonlinear operator. The lack of closure requires us to project the output of $\mathcal{N}$ back onto the initial set of basis functions. In the example in equation (93), projection is done by setting all of the spectral coefficients a_n with $n > N$ equal to zero. Galerkin's method uses truncation to project the output of $\mathcal{N}$. Other spectral methods use other types of projection. From equation (93), we see that the set of all sine functions is closed under $\mathcal{N}$; no cosines are produced. On the other hand, the set of cosines are not closed under $\mathcal{N}$; $\mathcal{N}$ operating on a cosine series produces a sine series. The closure or lack of closure of an infinite set of sines or cosines is due to the spatial symmetry properties of $\mathcal{N}$. $\mathcal{N}$ is anti-reflection symmetric about $x = 0$ and so are the sines; therefore, the set of sines is closed. The cosines do not have the anti-reflection symmetry of $\mathcal{N}$ and therefore they are not closed. In general, for every incomplete or sparse set of basis functions that is closed with respect to an operator, there is a physical symmetry common to the operator and the sparse set. It is to the numericist's advantage to exploit the fact that some equations together with their boundary and initial conditions have a symmetry. The clever numericist will choose a sparse set of basis functions that shares the symmetry. For example, if we use Galerkin's method to compute thermal convection in a star and if we are interested only in convective cells with duodecahedral symmetry (i.e., the convective cells tesselate the stellar surface in a soccer ball pattern), then using a set of basis functions with duodecahedral symmetry is more efficient than using the spherical harmonics as basis functions. A calculation of duodecahedral convection that uses a spectral sum with all spherical harmonics, $Y_{\ell,m}$ with $|m|<\ell$ and $0 \leq \ell \leq 32$ (1089 modes) can also be done with a much smaller series containing only 19 duodecahedral harmonics to obtain the identical spatial resolution. Each convolution with the smaller set of duodecahedral harmonics is approximately 10,000 times faster than the convolution with the larger set of spherical harmonics.

Tau Method

Galerkin's method cannot be used to approximate a function f(x) if the basis functions do not obey the same homogeneous boundary conditions as f(x) or if f(x) has inhomogeneous boundary conditions. For these boundary-value problems we use the tau method, developed by Lanczos (1956). Again, consider the heat equation (equation 1) over the domain $-1 \leq x \leq 1$ with boundary conditions

$$T(t,-1) = \alpha \tag{94}$$

$$T(t,L) = \beta \tag{95}$$

This time, we write T(t,x) as a spectral sum of Legendre polynomials

$$T(t,x) = \sum_{\ell=0}^{N} a_\ell(t)\, P_\ell(x) \tag{96}$$

The Legendre polynomials do not obey the inhomogeneous boundary conditions of T(t,x); instead, they obey

$$P_\ell(-1) = (-1)^\ell \tag{97}$$

$$P_\ell(1) = 1 \tag{98}$$

Substituting equation (96) into equation (1), multiplying both sides of equation (1) by $P_n(x)$, and integrating the resulting equation from -1 to 1, we obtain (N+1) equations for the (N+1) unknown spectral coefficients, a_n, n=0,1,...N. However, to satisfy the two boundary equations, (94) - (95), we also require that

$$\sum_{\ell=0}^{N} a_\ell(t)P_\ell(-1) = \sum_{\ell=0}^{N} a_\ell(t)(-1)^\ell = \alpha \tag{99}$$

and

$$\sum_{\ell=0}^{N} a_\ell(t)P_\ell(1) = \sum_{\ell=0}^{N} a_\ell(t) = \beta \tag{100}$$

This presents the dilemma that we have (N+3) equations and only (N+1) unknowns. To solve this problem, we must first examine the second derivative operator in more detail. Writing the second derivative of T in terms of the Legendre sum in equation (96) we obtain

$$\frac{d^2}{dx^2} T(x,t) \equiv \sum_{n=0}^{N} b_n(t)P_n(x) = \sum_{n=0}^{N} a_n \frac{d^2}{dx^2} P_n(x) \tag{101}$$

The second derivative of $P_n(x)$ is a linear combination of Legendre polynomials whose order is less than or equal to (n-2). The second derivative operator is a lowering operator; it maps the set of basis functions $\{P_n(x) : n=0,\ldots,N\}$ into the set $\{P_n(x) : n=0,\ldots, N-2\}$. Furthermore, since the operator $\frac{d^2}{dx^2}$ preserves parity, the second derivative of a Legendre polynomial of even (or odd) order is a sum of Legendre polynomials of even (or odd) order. We can write down a relationship between the b_n and a_n of equation (101) as the vector equation

$$\left|\begin{array}{c} b_0 \\ b_1 \\ b_2 \\ b_3 \\ \cdot \\ \cdot \\ \cdot \\ b_{N-3} \\ b_{N-2} \\ b_{N-1} \\ b_N \end{array}\right| = \left|\begin{array}{cccccccccccc} 0 & 0 & x & 0 & x & 0 & . & . & . & x & 0 & x \\ 0 & 0 & 0 & x & 0 & x & . & . & . & 0 & x & 0 \\ 0 & 0 & 0 & 0 & x & 0 & . & . & . & x & 0 & x \\ 0 & 0 & 0 & 0 & 0 & x & . & . & . & 0 & x & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & . & . & . & 0 & x & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & . & . & . & 0 & 0 & x \\ 0 & 0 & 0 & 0 & 0 & 0 & . & . & . & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & . & . & . & 0 & 0 & 0 \end{array}\right| \left|\begin{array}{c} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ a_{N-3} \\ a_{N-2} \\ a_{N-1} \\ a_N \end{array}\right| \tag{102}$$

or equivalently

$$\underline{b} = \underline{\underline{D}}^2 \ \underline{a} \tag{103}$$

where x stands for any non-zero matrix element. Note that $\underline{\underline{D}}^2$ is singular because the last two rows are zero. The fact that $\underline{\underline{D}}^2$ is non-invertible is due to the fact that $\frac{d^2}{dx^2}$ cannot be inverted unless two boundary conditions are supplied.

We can try to solve the spectral heat equation $\dot{\underline{a}} = \kappa \underline{\underline{D}}^2 \ \underline{a}$ with a forward Euler integration in time, or

$$a(t+\Delta t) = \underline{a}(t) + \kappa(\Delta t) \ \underline{\underline{D}}^2 \ \underline{a}(t) \tag{104}$$

In solving the heat equation explicitly,we avoid having to invert the singular matrix $\underline{\underline{D}}^2$. Unfortunately, in the explicit equation (104), the last two rows of $\underline{\underline{D}}^2$ make $a_N(t) = a_N(0)$ and $a_{N-1}(t) = a_{N-1}(0)$ for all time. This is undesirable. Furthermore, after one timestep, the $a_n(t)$ no longer satisfy the boundary conditions. The remedy to all of these problems is to modify the matrix $\underline{D}^2$ in the heat equation so that the bottom two rows are replaced by the boundary conditions of equations (99) - (100).

$$\left|\begin{array}{c} \dot{a}_0 \\ \dot{a}_1 \\ \dot{a}_2 \\ \dot{a}_3 \\ \cdot \\ \cdot \\ \cdot \\ \dot{a}_{N-3} \\ \dot{a}_{N-2} \\ \beta \\ \alpha \end{array}\right| = \left|\begin{array}{cccccccccccc} 0 & 0 & x & 0 & 0 & 0 & \cdot & \cdot & \cdot & x & 0 & x \\ 0 & 0 & 0 & 0 & 0 & x & \cdot & \cdot & \cdot & 0 & x & 0 \\ 0 & 0 & 0 & 0 & x & 0 & \cdot & \cdot & \cdot & x & 0 & x \\ 0 & 0 & 0 & 0 & 0 & x & \cdot & \cdot & \cdot & 0 & x & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 & x & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 & 0 & x \\ 1 & 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & \cdot & \cdot & \cdot & 1 & -1 & 1 \end{array}\right| \left|\begin{array}{c} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \cdot \\ \cdot \\ \cdot \\ a_{N-3} \\ a_{N-2} \\ a_{N-1} \\ a_N \end{array}\right| \tag{105}$$

Notice that the matrix on the right-hand side of equation (105) is invertible; if we integrate $\dot{a}_i(t)$ forward in time with a stable implicit method, then the matrix can be inverted at each timestep. Using a backwards Euler method to solve equation (105), we obtain

$$(\underline{\underline{\hat{I}}} - \kappa\Delta t\underline{\underline{\hat{D}}})\ \underline{a}(t+\Delta t) = \underline{\underline{\hat{I}}}\ \underline{a}(t) + \underline{x}(\beta,\alpha) \tag{106}$$

where $\underline{\underline{\hat{D}}}$ is the matrix on the right-hand side of equation (105), where $\underline{\underline{\hat{I}}}$ is the identity with the last two rows replaced by zeroes and where $\underline{x}(\alpha,\beta)$ is a column vector whose elements are all zero except for the last two which are β and α respectively.

Using equation (106) , the $a_n(t)$ exactly satisfy the boundary conditions for all time, but do they still satisfy the heat equation? We evaluate the mean-square error that arises from approximating a function $f(x)$ with boundary conditions $f(1) = \beta$ and $f(-1) = \alpha$ by using the tau method. In a Legendre-tau method we approximate $f(x)$ with $f_N(x)$,

$$f_N(x) \cong f(x) \tag{107}$$

where

$$f_N(x) = \sum_{n=0}^{N} a_n P_n(x). \tag{108}$$

Writing $f(x)$ exactly as an infinite series

$$f(x) = \sum_{n=0}^{\infty} c_n P_n(x) \tag{109}$$

The mean-square error is

$$L_2 = \left[\int_{-1}^{1} |f(x)-f_N(x)|^2 \, dx \right]^{1/2} = \left[\sum_{n=0}^{N} |c_n-a_n|^2 + \sum_{n=N+1}^{\infty} |c_n|^2 \right]^{1/2} \tag{110}$$

With the tau method we evaluate the first (N-2) spectral coefficients of a_n by taking inner products of $f(x)$ with $P_n(x)$ or

$$a_n = \int_{-1}^{1} f(x) P_n(x) dx = c_n \quad \text{for} \quad 0 \leq n \leq N-2 \tag{111}$$

The last two spectral coefficients are not determined from inner products; instead, they come from the boundary conditions (99)-(100). The mean-square error becomes

$$L_2 = \left[|a_{N-1}-c_{N-1}|^2 + |a_N-c_N|^2 + \sum_{n=N+1}^{\infty} |c_n|^2 \right]^{1/2} \tag{112}$$

The last term on the right-hand side of equation (112) is the same error that arises with Galerkin's method and is exponentially small if $f(x)$ is sufficiently smooth. We must show that the additional two terms $|a_N - c_N|^2$ and $|a_{N-1} - c_{N-1}|^2$ are also exponentially small. The exact function $f(x)$ obeys the exact boundary conditions

$$f(-1) = \sum_{n=0}^{\infty} c_n (-1)^N = \alpha \tag{113}$$

$$f(1) = \sum_{n=0}^{\infty} c_n = \beta \tag{114}$$

Comparing equation (99) with equation (113) and equation (100) with equation (114) , we see that

$$(-1)^{N-1} a_{N-1} + (-1)^N a_N = \sum_{n=N-1}^{\infty} (-1)^n c_n \tag{115}$$

and

$$a_{N-1} + a_N = \sum_{n=N-1}^{\infty} c_n . \tag{116}$$

The solution to equations (115) - (116) is

$$a_N = \sum_{k=0}^{\infty} c_{N+2k} \tag{117}$$

$$a_{N-1} = \sum_{k=0}^{\infty} c_{N-1+2k} . \tag{118}$$

Equations (117)-(118) show that the second contribution to L_2 in equation (112) is exponentially small:

$$|a_{N-1} - c_{N-1}|^2 + |a_N - c_N|^2 \leq$$

$$2 \left| \sum_{k=0}^{\infty} c_{N-1+2k} \right|^2 + 2 \left| \sum_{k=0}^{\infty} c_{N+2k} \right|^2 + 2|c_{N-1}|^2 + 2|c_N|^2 \tag{119}$$

The right-hand side of equation (119) decreases exponentially with N.

Pseudo-Spectral Methods

Our last example of a technique for computing the spectral coefficients is known as the pseudo-spectral method, the collocation method or the method of selected points. It is used in nonlinear equations or linear equations with non-constant coefficients. Its purpose is to avoid a convolution product. The method exploits the fact that spectral differentiation is more accurate than finite-differences, but that multiplication of two functions which are tabulated at a set of selected points (or collocation points) is faster than spectral convolution. In the pseudo-spectral methods, all differentiation and quadrature is done with spectral approximations; all multiplication and division are done on a grid of points. The representation of the function goes back and forth between spectral and physical space by use of discrete (and fast) transforms. When transforming a function with N spectral coefficients from spectral space to physical space, the user should sample the function at $M = N$ grid points. Over-sampling with $M > N$ is wasteful; under-sampling with $M < N$ loses information and prohibits reconstruction of the spectral coefficients from the M sampled points. Problems arise in the pseudo-spectral method due to accidental under-sampling. The user can inadvertently under-sample a function if there are nonlinear or non-constant coefficient terms.

To see how under-sampling might arise, consider again the nonlinear wave equation

$$\frac{\partial u}{\partial t} = - u \frac{\partial u}{\partial x} \tag{120}$$

with periodic boundary conditions

$$u(t,x=0) = u(t,x=1) \tag{121}$$

Initially, u and $\frac{\partial u}{\partial x}$ are represented in spectral space as sine series (sines are complete over the interval):

$$u_N(t,x) = \sum_{n=0}^{N} a_n(t) \sin \pi nx \tag{122}$$

$$\frac{\partial u_N}{\partial x}(t,x) = \pi \sum_{n=0}^{N} a_n(t) \cos \pi nx \tag{123}$$

Fourier transforming u_N and $\frac{\partial u_N}{\partial x}$ into physical space, we obtain tabulations of the two functions $u_N(t,x_i)$ and $\frac{\partial u_N}{\partial x}(t,x_i)$ at the collocation points $x_i \equiv \frac{(i-1)}{N}$, $i=1,N$. At this step in the pseudo-spectral method we have enough information to reconstruct the spectral coefficients a_n and $(n\pi a_n)$ from the tabulated functions. The product $-u \frac{\partial u}{\partial x}$ is tabulated at the collocation point in the obvious way :

$$-u_N \frac{\partial u_N}{\partial x} (t,x_i) \equiv - u_N(t,x_i) \frac{\partial u_N}{\partial x} (t,x_i) \tag{124}$$

Now the product $\frac{\partial u_N}{\partial x}$ is tabulated at N points, but if we were to represent it spectrally (by a convolution product) it would require 2N terms (see equation 93):

$$-u_N \frac{\partial u_N}{\partial x} = \sum_{n=1}^{2N} b_n \sin(n\pi x) \tag{125}$$

We have accidentally under-sampled because the function $- u_N \frac{\partial u_N}{\partial x}$ has 2N spectral coefficients but we have only tabulated $-u_N \frac{\partial u_N}{\partial x}$ at N points. If we had used Galerkin's method to compute the convolution product, we would project the 2N spectral coefficients of the product back onto an N-dimensional spectral space by explicitly setting the $b_n = 0$ for all $n > N$ and keeping the coefficients b_n unchanged for all $0 \leq n \leq N$. With the pseudo-spectral method the under-sampled product, $-u_N \frac{\partial u_N}{\partial x}$, is also projected back onto an N-dimensional spectral space when we naively inverse transform the tabulated product back into Fourier space. The discrete inverse transform not only sets the coefficients b_n equal to zero for all $n > N$, but also mixes the spectral coefficient b_n with $0 < n \leq N$ with the spectral coefficients with $n > N$. The mixing is called aliasing. It is due to the fact that the discrete Fourier transform with N points is unable to tell the difference between Fourier modes of wavenumber n and wavenumber 2N-n. For example, $\sin(n\pi x)$ evaluated at the grid points $x_i = \frac{(i-1)}{N}$ $i=1,N$ is indistinguishable from $-\sin[(2N-n)\pi x]$ evaluated at the same grid points. It should not be surprising that the inverse transform contaminates the

n^{th} spectral coefficient with the $(2N-n)^{th}$ coefficient. There is never contamination or aliasing error if the number of sampling points of a function is greater than or equal to the number of spectral components of the function.

The aliasing error can be avoided or reduced by several methods. One way of removing the alias in equation (121) is to evaluate u_n, $\frac{\partial u_N}{\partial x}$, and the product at $\frac{3}{2}$ N collocation points. The discrete inverse transform of the product will have no aliasing errors in the first N spectral coefficients. Often, aliasing errors can be ignored with no harmful effects. The reason is that aliasing contaminates the exact solution with the spectral coefficients b_n with $n > N$. If N is chosen sufficiently large, then the spectral coefficients with $n > N$ are exponentially small and the aliasing error is exponentially small. For example, aliasing errors in the Navier-Stokes equation can usually be ignored if the equation is solved in a way such that the aliasing error does not violate energy or momentum conservation.

DISCUSSION

There are more types of spectral methods than those discussed in this paper. Most of the other techniques are hybrids of those outlined here. All good spectral methods share the property that their convergence is exponential and are therefore more economical than finite differences. Spectral methods require approximately 7 times fewer degrees of freedom (grid points or modes) per spatial dimension than do second-order finite-differences. Therefore, in 2 or 3 dimensional calculations, spectral methods are often the only practical way of obtaining adequate spatial resolution. In the future, as astrophysicists increasingly want to extend their numerical calculations to 2 and 3 dimensions, spectral methods will increasingly become part of the astrophysicist's standard numerical tools.

REFERENCES

Dahlquist, G. and Bjorke, A.: 1974, Numerical Methods, Prentice-Hall.

Gottlieb, D. and Orszag, S.A.: 1977, Numerical Analysis of Spectral Methods, SIAM.

Lanczos, C. 1956, Applied Analysis, Prentice-Hall.

Marcus, P.S. 1983, submitted to the Journal of Fluid Mechanics.