Lecture 1

Series Expansion Methods Edit

Series expansion methods are the general class that encompass spectral and finite element methods. We approximate functions as a linear combination of prescribed expansion functions - we call these basis functions. For a continuous function $f(x)$, we write

$$f(x) = \sum_{j=1}^{N} a_j \phi_j(x) \tag{1.1}$$

where ϕ_j, $j = 1, \dots, N$, are the basis functions that each satisfy any boundary conditions on $f(x)$. The coefficients a_j are the unknowns and form a vector of N numbers.

Suppose that we have a partial differential equation of the arbitrary form

$$\mathcal{L}\left(f\right) = \rho(x). \tag{1.2}$$

We define the residual of equation (**??**) as

$$r(x) = \mathcal{L}\left(f\right) - \rho(x) = \mathcal{L}\left(\sum_{j=1}^{N} a_j \phi_j(x)\right) - \rho(x). \tag{1.3}$$

If $\mathcal{L}$ is linear and the basis functions, $\phi_j(x)$ are the eigen-functions of $\mathcal{L}$, then the residual can be set to zero for the whole domain and the resulting N algebraic equations can be solved for a_j. Generally speaking, $\mathcal{L}$ is non-linear so we will describe more general approaches for solving for the coefficients a_j.

We will consider three strategies: i) minimization of the of l_2-norm of the residual,

$$(||r(x)||_2)^2 = \int r(x)^2 dx, \tag{1.4}$$

ii) the collocation method where we set the residual to zero at a discrete set of positions x_k (e.g. on a regular grid $x_k = k\Delta x$),

$$r(x_k) = 0 \quad \forall \quad k = 1, \ldots, N \tag{1.5}$$

and iii) the Galerkin method which requires the residual to be orthogonal to each of the basis functions,

$$\int \phi_k r(x) dx = 0 \quad \forall \quad k = 1, \ldots, N \tag{1.6}$$

The collocation method is used in the pseudo-spectral method while the Galerkin method is used extensively in the finite element method. The spectral method is a special case where the l_2-norm and Galerkin method become equivalent.

There are many variants on these methods and we will discuss one, the Petrov-Galerkin method. Here, the residual is made orthogonal to a set of test functions, $\theta_k(x)$, which may be different to the basis set, $\phi_k(x)$:

$$\int \theta_k r(x) dx = 0 \quad \forall \quad k = 1, \ldots, N. \tag{1.7}$$

The Petrov-Galerkin method is more general than the Galerkin method because if we chose $\theta_k(x) = \phi_k(x)$ we recover equation (**??**).

1.1 The Spectral Method Edit

Spectral methods are a special case of the series expansion method; the basis functions form an orthogonal set:

$$\int \phi_i \phi_j dx = 0 \quad \forall \quad i \neq j.$$

The ability to use an orthogonal basis set is largely dictated by the domain geometry and boundary conditions. For example, it is natural to use spherical harmonics in spectral atmospheric models but difficult to model irregular coasts with the same representation.

1.1.1 Spectral method compared to the finite difference method Edit

In lecture **??** we analyzed the numerical dispersion of waves using the finite difference method applied to the linear advection problem

$$\partial_t \theta + c\partial_x \theta = 0$$

in which c was constant. Consider using a Fourier series expansion to represent $\theta(x,t)$:

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

We're implicitly assuming a periodic domain $-\pi \leq \pi$. Note, in this case, $\theta(x,t)$ is real so the coefficients in the series satisfy the property $a_k = a^*_{-k}$ where a^*_k is the complex conjugate of a_k.

Direct substitution of each component of the series, $a_k(t)e^{ikx}$, into the governing equation yields

$$\partial_t a_k + icka_k = 0 \quad \forall\ k$$

which is unusually trivial to solve. The corresponding dispersion relation shows us that all waves, including the shortest, propagate with the exact correct phase speed.

Note, we did not use any of the methods listed in section **??**. Let us apply the Galerkin approximation (equation **??**) to see what happens. Note that two complex functions, $g(x)$ and $h(h)$, are orthogonal is the integral over the domain, S, of the product of one with the complex conjugate of the other is zero:

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

For this problem, the Galerkin approximation is

$$\int_{-\pi}^{\pi} e^{-ijx} \left[\partial_t \sum_{k=-n}^{N} a_k(t)e^{ikx} + c\partial_x \sum_{k=-n}^{N} a_k(t)e^{ikx} \right] dx = 0 \quad \forall\ j = -N, \ldots, N \tag{1.8}$$

and we need to evaluate the following integral:

$$\int_{-\pi}^{\pi} e^{-ijx} e^{ikx}\,dx = \begin{cases} \left[\frac{e^{i(k-j)x}}{k-j}\right]_{-\pi}^{\pi} = 0 & j \neq k \\ \\ 2\pi & j = k \end{cases}$$

Thus the Galerkin approximation yields

$$2\pi\partial_t a_k + 2ic\pi k a_k = 0$$

which is exactly the same results has obtained by direct substitution.

The spectral approximation does not introduce phase speed or amplitude errors, ignoring time-discretization errors.

1.1.2 Spectral Stommel model in 1-D Edit

This is a somewhat contrived use of the spectral method but allows us to make a direct comparison with finite difference method used in section **??**.

To re-state the problem, we seek solutions to the differential equation:

$$\epsilon\partial_{xx}\psi + \partial_x\psi = -1$$

with boundary conditions $\psi(0,1) = 0$.

We will express the solution, $\psi(x)$, in terms of a sin-series:

$$\psi(x) = \sum_{j=1}^{N} a_j \sin(j\pi x) \tag{1.9}$$

since the functions $\sin(j\pi x)$ all satisfy the boundary conditions. The residual of the governing equation is

$$\begin{aligned} r(x) &= \epsilon\sum_{j=1}^{N} a_j\partial_{xx}\sin(j\pi x) + \sum_{j=1}^{N} a_j\partial_x\sin(j\pi x) + 1 \\ &= -\epsilon\pi^2\sum_{j=1}^{N} a_j j^2\sin(j\pi x) + \pi\sum_{j=1}^{N} a_j j\cos(j\pi x) + 1 \qquad (1.10) \end{aligned}$$

Using the Galerkin method (**??**) we require the residual to be orthogonal to the basis functions:

$$\begin{aligned} \int_0^1 r(x)\sin(k\pi x)\,dx &= \int_0^1 \left\{ -\epsilon\pi^2\sum_{j=1}^{N} a_j j^2\sin(k\pi x)\sin(j\pi x) \right. \\ &\qquad \left. +\pi\sum_{j=1}^{N} a_j j\sin(k\pi x)\cos(j\pi x) + \sin(k\pi x) \right\} dx \\ &= 0 \quad \forall\, k = 1,\ldots,N \qquad (1.11) \end{aligned}$$

Once we evaluate the integrals and sums it will become apparent that (**??**) represents a set of N algebraic equations in the unknowns, a_k.

The first term in (**??**) involves the expression

$$\int_0^1 \sum_{j=1}^{N} a_j j^2 \sin(k\pi x) \sin(j\pi x)\, dx = \sum_{j=1}^{N} a_j j^2 \int_0^1 \sin(k\pi x) \sin(j\pi x)\, dx.$$

Evaluating the integral inside the sum we get

$$\int_0^1 \sin(k\pi x)\sin(j\pi x)\, dx = \begin{cases} \frac{1}{2} & j = k \\ \\ 0 & j \neq k \end{cases}$$

which reflects the orthogonality of the basis set. Evaluating the sum over j, the first term in (**??**) becomes

$$-\epsilon\pi^2 \sum_{j=1}^{N} a_j j^2 \int_0^1 \sin(k\pi x)\sin(j\pi x)\, dx = \frac{-\epsilon\pi^2}{2} a_k k^2.$$

The last term is similarly straight forward:

$$\int_0^1 \sin(k\pi x)\, dx = \frac{-1}{k\pi}\left[\cos(k\pi x)\right]_0^1 = \frac{1-\cos(k\pi)}{k\pi} = \frac{1-(-1)^k}{k\pi}$$

which is $2/k\pi$ when k is odd and is zero when k is even. The beta term is more complicated; the inner integral evaluates to

$$\int_0^1 \sin(k\pi x)\cos(j\pi x)\, dx = \frac{k - k(-1)^{(j+k)}}{\pi(k^2 - j^2)}.$$

When $j+k$ is even, which includes $j = k$, the numerator and integral is zero. When $j+k$ is odd the expression becomes $2k/\pi(k^2 - j^2)$.

Substituting all these results back into **??** we obtain

$$\frac{-\epsilon\pi^2 k^2}{2} a_k + \sum_{j=1}^{N} a_j \frac{jk(1-(-1)^{(j+k)})}{k^2 - j^2} = \frac{(-1)^k - 1}{k\pi} \quad \forall\, k = 1, \ldots, N \qquad (1.12)$$

which represents N algebraic equations for the unknowns a_k. We pose this as a linear algebra problem of the form $\underline{\underline{A}}\,\underline{a} = \underline{b}$ where

$$\underline{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ \vdots \end{pmatrix}, \underline{b} = \begin{pmatrix} \frac{2}{\pi} \\ \vdots \\ \frac{(-1)^k - 1}{k\pi} \\ \vdots \end{pmatrix}$$

and $\underline{\underline{A}}$ whose elements are given by

$$A_{jk} = \frac{\epsilon \pi^2 k^2}{2} \delta_{ij} + \frac{jk(1 - (-1)^{(j+k)})}{k^2 - j^2}$$

where the symbol δ_{ij} is the Kronecka function; $\delta_{ij} = 1 \ \forall \ i = j$ and $\delta_{ij} = 0 \ \forall \ i \neq j$. This is a "full" matrix meaning it is not sparse; many elements are non-zero. Recall that the matrix problem corresponding to the second order finite difference Stommel model was a tri-diagonal matrix which is a lot easier to invert. The basis functions and solution using $N = 6$ are plotted in Fig. **??** and examples of higher truncations given in Fig. **??**. The error for a range of truncations is plotted as a function of N in Fig. **??**. Note that the slope of the error curves seems to steepen downward as N increases - this is a strong motivation for using the property of spectral methods; they have incredible convergence properties.

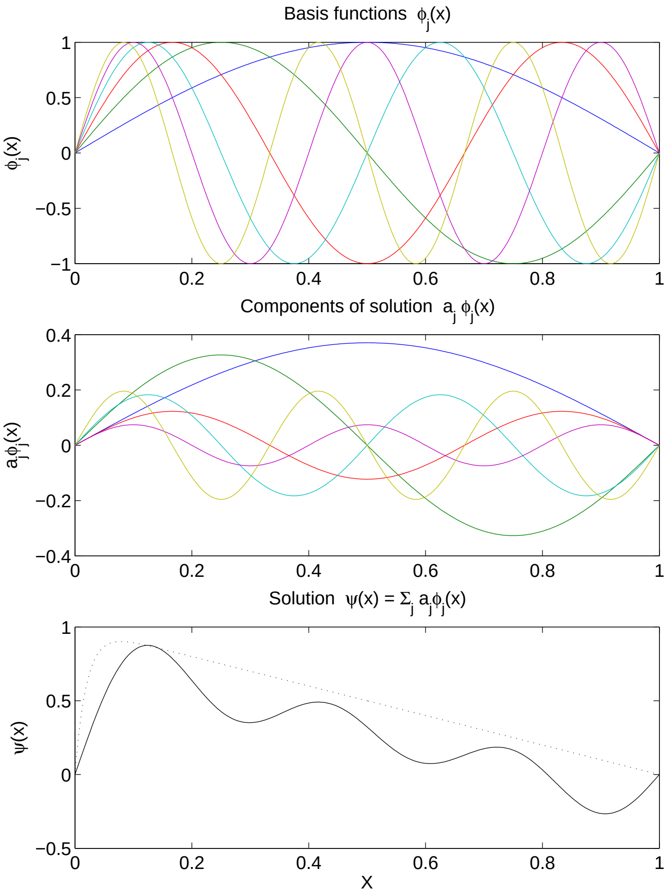

Figure 1.1: Top: The basis functions used in the spectral method solution of the Stommel model (N=6). Middle: the components of the solution due to each basis function, $a_j\phi_j(x)$. Bottom: the numerical solution and analytical solution.

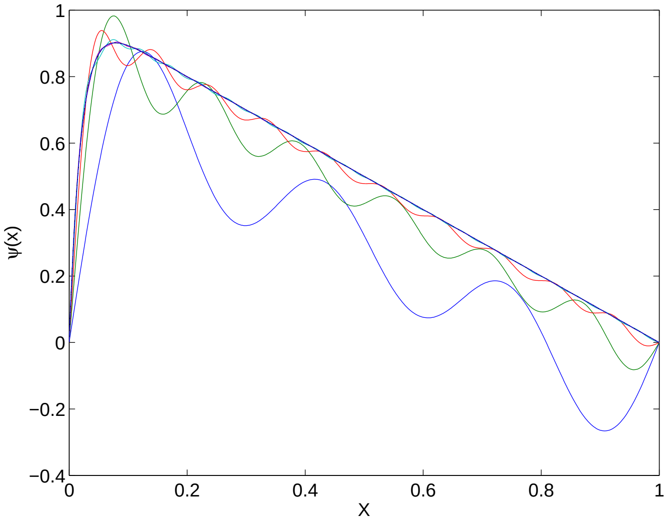

Figure 1.2: Spectral solutions to the Stommel problem using different length Fourier series, $N = 6, 12, 20, 40, 60, 100, 200, 400$.

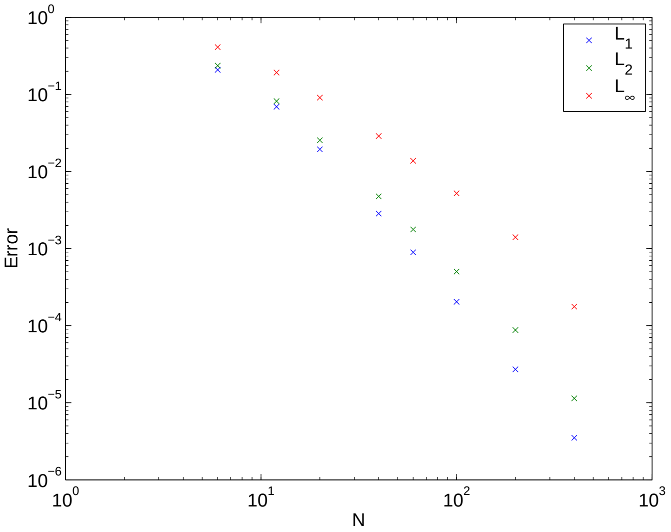

Figure 1.3: Convergence of the spectral method. The error, measured by the l_1, l_2 and l_∞ norms, is plotted as a function of N. Note the increasing rate of convergence as N is increased.

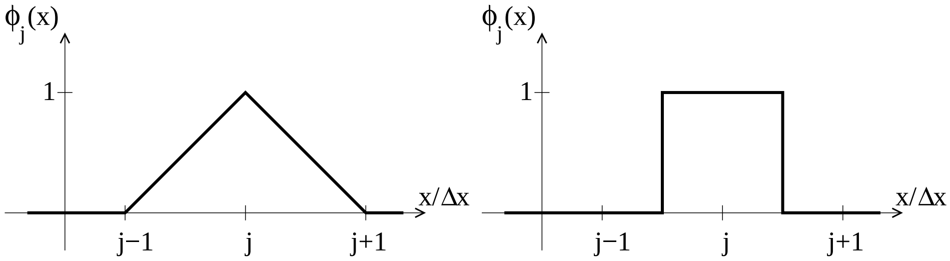

Figure 1.4: a) The chapeau function and b) the top hat function

1.2 Finite elements using Chapeau functions

Edit

As an example, consider the problem of constant advection in one-dimension:

$$\partial_t u + c\partial_x u + \nu\partial_{xx} u = 0$$

We'll describe the dependent variable, u, using the chapeau basis functions:

$$\phi_j(x) = \begin{cases} 0 & \forall\ |x - x_j| > \Delta x \\ 1 - \frac{|x-x_j|}{\Delta x} & \forall\ |x - x_j| \leq \Delta x \end{cases} \tag{1.13}$$

which is plotted in Fig. **??**a. Using the Galerkin approximation (**??**), we have:

$$\sum_{j=1}^{N} \int_{-\infty}^{\infty} \phi_i \left(\partial_t u_j \phi_j + c u_j \partial_x \phi_j - \nu u_j \partial_x x \phi_j\right) dx = 0 \quad \forall\ i = 1, \ldots, N \tag{1.14}$$

In this expression we see products of basis functions and derivatives of basis functions. We can evaluate the integrals of these expressions:

$$\begin{aligned} \int_{-\infty}^{\infty} \phi_i \phi_i dx &= \frac{4}{6}\Delta x \\ \int_{-\infty}^{\infty} \phi_i \phi_{i\pm 1} dx &= \frac{1}{6}\Delta x \\ \int_{-\infty}^{\infty} \phi_i \partial_x \phi_i dx &= 0 \\ \int_{-\infty}^{\infty} \phi_i \partial_x \phi_{i\pm 1} dx &= \pm\frac{1}{2} \end{aligned}$$

$$\begin{aligned}\int_{-\infty}^{\infty} \phi_i \partial_{xx} \phi_i dx &= \frac{-2}{\Delta x} \\ \int_{-\infty}^{\infty} \phi_i \partial_{xx} \phi_{i\pm1} dx &= \frac{\pm1}{\Delta x}\end{aligned}$$

where the last two integrals were carried out by parts. Equation **??** then becomes

$$\frac{1}{6}(\partial_t u_{i-1} + 4\partial_t u_i + \partial_t u_{i+1}) + \frac{c}{2\Delta x}(u_{i+1} - u_{i-1}) - \frac{\nu}{\Delta x^2}(u_{i+1} - 2u_i + u_{i-1}) = 0$$

or, returning the finite difference notation,

$$\mathcal{A}^x \partial_t u + \frac{c}{\Delta x}\delta_i \overline{u}^i - \frac{\nu}{\Delta x^2}\delta_{ii} u = 0$$

where the averaging operator, $\mathcal{A}^x$ is defined:

$$\mathcal{A}^x u = u + \frac{1}{6}\delta_{ii} u = \frac{1}{6}(u_{i-1} + 4u_i + u_{i+1})$$

The finite element method gives fourth order spatial accuracy in this problem. For comparison, the second order finite difference approximation is

$$\partial_t u + \frac{c}{\Delta x}\delta_i \overline{u}^i - \frac{\nu}{\Delta x^2}\delta_{ii} u = 0$$

and the fourth order finite difference approximation:

$$\partial_t u + \frac{c}{\Delta x}\delta_i \overline{u - \frac{1}{6}\delta_{ii} u}^i - \frac{\nu}{\Delta x^2}\delta_{ii}(u - \frac{1}{12}\delta_{ii} u) = 0$$

which uses a five point stencil. The dispersion relation for the undamped waves ($\nu = 0$) in these approximations are plotted in Fig. **??**. Note that although the stencil of the finite elements method is only three points (as for the second order finite difference approximation), and that the formal truncation is $O(\Delta x^4)$, the dispersion relation is far more accurate than even the sixth order finite difference approximation.

Finite element Stommel model

Note that we can trivially see what the finite element approximation to the Stommel problem would be by simply dropping the time-derivative in the above problem. The resulting discretization is exactly the same as the second-order finite difference discretization.

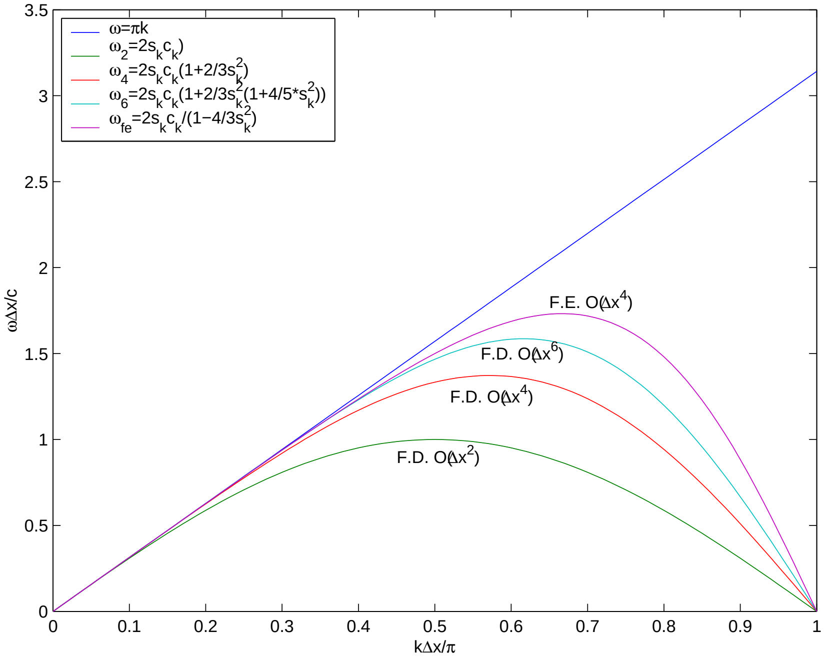

Figure 1.5: Dispersion relations for the $O(\Delta x^2)$, $O(\Delta x^4)$ and $O(\Delta x^6)$ finite difference methods and the $O(\Delta x^4)$ finite elements method (using the chapeau basis functions).

1.3 Note on Finite Fourier series Edit

Different texts use different apparent representations of finite Fourier series which are in fact equivalent. For a *real* valued function, $\phi(x)$, one series representation may be written

$$\phi(x) = a_0 + \sum_{k=1}^{N} a_k \cos(kx) + \sum_{k=1}^{N} b_k \sin(kx)$$

which has $2N+1$ degrees of freedom and where all coefficients are *real*. This representation has the advantage that it is immediately obvious that all terms are real. A more succinct series representation is

$$\phi(x) = \sum_{k=-N}^{N} c_k e^{ikx}$$

which also has $2N+1$ coefficients but where the coefficients c_k are *complex*. A complex number has two components, real and imaginary, and one may wonder if there are in fact $4N+2$ degrees of freedom. If $\phi(x)$ is real, then the coefficients must satisfy a certain constraint as we derive now.

Let us write each coefficient $c_k = c_k^R + ic_k^I$ where c_k^R and c_k^I are real numbers. Then each term in the series is

$$\begin{aligned} c_k e^{ikx} &= (c_k^R + ic_k^I)\left[\cos(kx) + i\sin(kx)\right] \\ &= \left[c_k^R \cos(kx) - c_k^I \sin(kx)\right] + i\left[c_k^R \sin(kx) + c_k^I \cos(kx)\right]. \end{aligned}$$

It is then obvious that the sum of contributions from modes k and $-k$ can be written in term of $\cos(kx)$ and $\sin(kx)$ alone:

$$\begin{aligned} c_k e^{ikx} + c_{-k} e^{-ikx} &= \left[(c_k^R + c_{-k}^R)\cos(kx) + (c_{-k}^I - c_k^I)\sin(kx)\right] \\ &\quad + i\left[(c_k^R + c_{-k}^R)\sin(kx) + (c_k^I + c_{-k}^I)\cos(kx)\right]. \end{aligned}$$

To ensure that the function has no imaginary component we need to ensure

$$c_k^R - c_{-k}^R = 0 \quad \text{and} \quad c_k^I + c_{-k}^I = 0$$

or more simply

$$c_k = c_{-k}^*.$$

We can also associate each term in the two forms of Fourier series:

$$\begin{aligned} a_0 &= c_0 \qquad \text{(which is real)} \\ a_k &= c_k^R + c_{-k}^R = 2c_k^R \\ b_k &= c_{-k}^I - c_k^I = -2c_k^I. \end{aligned}$$