# 7. Galerkin Methods.

## 7.1 Introduction.

Until now, finite-difference methods for solving partial differential equations were applied. These methods specify the dependent variables at certain gridpoints in space and time and the derivatives in the equations are evaluated using Taylor Series expansions. The definitions of convergence, accumulated error, ... are based on comparing the solution $U_j^n = U(j\Delta x, n\Delta t)$ to the actual/exact solution $u(x,t)$ at grid point locations. The Galerkin procedure represents the dependent variables with a sum of functions that have a prescribed spatial structure. The coefficient associated with each function is normally a function of time. This procedure transforms a partial differential equation into a set of ordinary differential equations for the coefficients which are usually solved with finite differences in time. The two most useful Galerkin methods are the spectral method and the finite element method.

We now look at grid point values $U_j^n$ as being representative of grid-box averages of $u(x,t)$. Thus, in the case of one spatial dimension, we would compare

$$U_j^n = \frac{1}{\Delta x}\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} U_j^n \, dx \quad \text{to} \quad \frac{1}{\Delta x}\int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} u(x,t)\, dx$$

rather than $U_j^n$ to $u(x\Delta x, n\Delta t)$.

We can reformulate by defining

(1) $$\phi_j(x) = \begin{cases} 1/\Delta x & \text{for } (j-1/2)\Delta x \le x < (j+1/2)\Delta x \\ 0 & \text{elsewhere} \end{cases}$$

and the $U^n_j$ values as a vehicle for defining a piecewise constant approximation $U(x, n\Delta t)$ such that

(2) $$U(x, n\Delta t) = \sum_{j=-\infty}^{\infty} U^n_j \phi_j(x)$$

A formal series expansion for the purpose of approximating $u(x,t)$ can be carried out with an infinite variety of function $\phi_j$. The ones from (1) are representative of the mathematical grid point values associated with the standard finite-difference equations.

Example: Advection equation $\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$

By using (2), the advection equation can be rewritten

(3) $$\sum_j \frac{\partial U^n_j}{\partial t}\phi_j = -c\sum_j U^n_j \frac{\partial \phi_j}{\partial x}$$

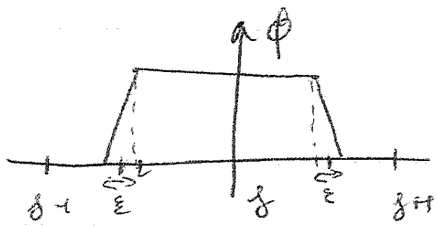
The basis function (1) are orthogonal $\Rightarrow$

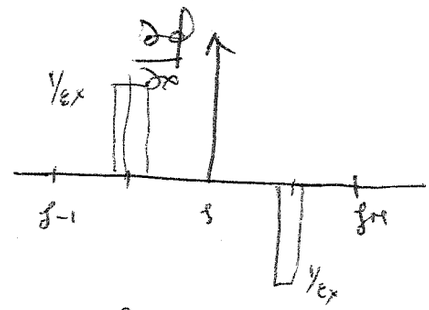
$$\int_{-\infty}^{\infty} \phi_i(x)\phi_j(x)\,dx = \begin{cases} (\Delta x) & \text{if } i=j \\ 0 & \text{elsewhere} \end{cases}$$

We can obtain an equation for $\frac{\partial U^n_i}{\partial t}$ by multiplying by $\phi_k$ and integrating over $x$

(4) $$\sum_j \frac{\partial U^n_i}{\partial t}\int_{-\infty}^{\infty}\phi_j\phi_k\,dx = -c\sum_j U^n_j \int_{-\infty}^{\infty}\frac{\partial \phi_j}{\partial x}\phi_k\,dx$$

In order to compute the RHS integral, we consider the step function to be a trapeze



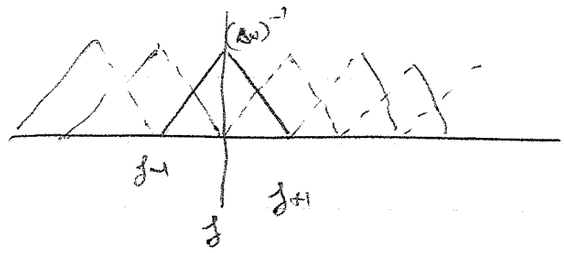




$$\Rightarrow \int_{-\infty}^{\infty} \frac{\partial \phi_j}{\partial x}\,\phi_k\,dx = \begin{cases} -1/2\Delta x^2 & j = k-1 \\ +1/2\Delta x^2 & j = k+1 \\ 0 & \text{elsewhere} \end{cases}$$

(independent of the slope) (valid for $\varepsilon \to 0$)

(4) then becomes $\dfrac{\partial U_k^n}{\partial t} = -c\,\dfrac{U_{k+1}^n - U_{k-1}^n}{2\Delta x}$ Centered in space

* If we use a different $\phi$ such as a piece wise linear representation




They are not orthogonal $\displaystyle\int_{-\infty}^{\infty} \phi_j\,\phi_k\,dx = \begin{cases} \frac{2}{3\Delta x} & j = k \\ \frac{1}{6\Delta x} & |j-k| = 1 \\ 0 & \text{elsewhere} \end{cases}$

and result in

(5) $$\frac{1}{6}\left(\frac{\partial U_{k-1}^n}{\partial t} + 4\,\frac{\partial U_k^n}{\partial t} + \frac{\partial U_{k+1}^n}{\partial t}\right) = -c\,\frac{U_{k+1}^n - U_{k-1}^n}{2\Delta x}$$

* We can generalize these ideas into a formal definition of the Galerkin method

Given a differential equation $\mathcal{L}(u) = f(x)$ where $\mathcal{L}$ is a differential operator, $u$ the dependent variable and $f(x)$ a specified forcing function.

in the domain R (x may be multidimensional). The Galerkin approximation is defined by

$$U(x,t) = \sum_{j=1}^{N} A_j(t)\,\phi_j(x) \qquad (6)$$

where the coefficients $A_j(t)$ are determined by requiring that the error

$$e_N = \mathcal{L}(U(x,t) - f(x) = \mathcal{L}\left(\sum_{j=1}^{N} A_j(t)\,\phi_j(x)\right) - f(x) \qquad (7)$$

be orthogonal to each basis function.

$$\underbrace{\int_R}_{\text{domain}} e_N\,\phi_j(x)\,dx = 0 \qquad j = 1, \ldots N \qquad (8)$$

The final form is

$$\int_R \phi_k \left(\sum_{j=1}^{N} A_j(t)\,\phi_j(x)\right) dx - \int_R \phi_k\, f(x)\,dx = 0 \qquad (9)$$

$$k = 1, \ldots N.$$

This reduces to the problem of N algebraic equations that relate the unknown coefficients $A_j(t)$ to the "transform" of the forcing function. They are normally solved by finite differences in time.

There are various ways to interpret (8).

(1) The residual error is orthogonal to orthogonal to $\phi_j$, i.e. the error should have no components in the space spanned by the $\phi_j$.

(2) The coefficients $A_j$ should be chosen to minimize the integral $\int_R e^2(x,t)\,dx$

Straightforward when $\mathcal{L}$ is a linear operator. In more complicated cases, not necessarily valid.

(3) $\mathcal{L}(u) = f(x)$ is approximated by $\mathcal{L}(U) = f(x)$ as in the Introduction.

Schemes employing basis functions defined in terms of periodic functions are referred to as "spectral". The ones using more "localized" basis functions are "finite elements" schemes.

## 7.2 Energy conservation

If we consider the simplified equation

$$\frac{\partial u}{\partial t} + \mathcal{L}(u) = 0 \tag{10}$$

then the Galerkin form is

$$\sum_{j=1}^{N} \frac{dA_j}{dt} \int_R \phi_k \phi_j \,dx + \int_R \phi_k \mathcal{L}\left(\sum_{j=1}^{N} A_j \phi_j\right) dx = 0 \tag{11}$$

$$k = 1, \ldots N$$

This process gives $N$ coupled ordinary differential equations in the coefficients $A_j(t)$. This can be solved by introducing finite differences in time.

We already discussed the importance of energy conserving schemes. The Galerkin method leads naturally to energy conservation in equations with quadratic energy invariants.

For an energy conserving system

$$(12)\qquad \int_R \frac{\partial (u^2/2)}{\partial t} = -\int_R u\, \mathcal{L}(u)\, dx$$

, the operator $\mathcal{L}$ must satisfy the condition

$\int_R \psi\, \mathcal{L}(\psi)\, dx = 0$ where $\psi$ is any reasonable function that satisfies the boundary conditions. Then (12) satisfies

$$(13)\qquad \frac{d}{dt}\int_R u^2/2\, dx = 0$$

which shows the energy conservation for the exact equation. We need to demonstrate that it holds for the finite sum.

We multiply the $k^{th}$ equation (11) by $A_k$ and sum from $k=1$ to $N$

$$(14)\qquad \int_R \left(\sum_{k=1}^N A_k \phi_k\right) \frac{\partial}{\partial t}\left(\sum_{j=1}^N A_j \phi_j\right) dx =$$

$$- \int_R \left(\sum_{k=1}^N A_k \phi_k\right) \mathcal{L}\left(\sum_{j=1}^N A_j \phi_j\right) dx$$

The integral on the right side vanishes if we set $\psi = \sum_{j=1}^N A_j \phi_j = \sum_{k=1}^N A_k \phi_k$

and (14) can be rewritten

$$\frac{1}{2}\int_R \frac{\partial}{\partial t}\left(\sum_{k=1}^N A_k \phi_k\right)^2 dx = 0$$

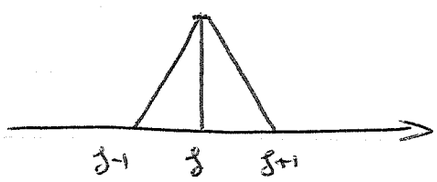
21 st March 1991

**Energy conservation for the Galerkin Approximation**

## 7.3 The Advection equation with Finite Elements

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0$$

We use again piece wise linear elents such as




$$\phi_j(x) = \begin{cases} 0 & x < (j-1)\Delta x \\ (x-(j-1)\Delta x)/\Delta x & (j-1)\Delta x < x < j\Delta x \\ ((j+1)\Delta x - x)/\Delta x & j\Delta x < x < (j+1)\Delta x \\ 0 & x > (j+1)\Delta x \end{cases}$$

The Galerkin equation is obtained by setting

$$L = c\frac{\partial}{\partial x}, \quad f(x) = 0$$

$$(5) \quad \sum_{j=1}^{N} \frac{\partial A_j}{\partial t} \int_R \phi_k \phi_j \, dx + c \sum_{j=1}^{N} A_j \int_R \phi_k \frac{\partial \phi_j}{\partial x} = 0$$

$$k = 1 \ldots N$$

The resulting equation is the one derived in the introduction

$$(16) \quad \frac{1}{6}\left(\frac{\partial A_{k+1}}{\partial t} + 4\frac{\partial A_k}{\partial t} + \frac{\partial A_{k-1}}{\partial t}\right) = -c\frac{A_{k+1} - A_{k-1}}{2\Delta x}$$

The advection term is the same as if obtained from centered differencing, but the time derivatives appears as a weighted average over three points. This greatly increases the accuracy of the solution.

We now apply the leap-frog time differencing scheme

(17) $$\frac{1}{12\Delta t}\left( A_{k+1}^{n+1} - A_{k+1}^{n-1} + 4\left(A_k^{n+1} - A_k^{n-1}\right) + A_{k-1}^{n+1} - A_{k-1}^{n-1} \right) = -c\,\frac{A_{k+1}^{n} - A_{k-1}^{n}}{2\Delta x}$$

The stability and phase error of this scheme can be investigated by substituting $A_k^n = A\, e^{i(\mu\Delta x n k + \alpha n \Delta t)}$ (Variation of the Fourier Transform).

Substitution into (17) leads to

(18) $$\sin \alpha\Delta t = -\frac{c\Delta t}{\Delta x}\left( \left( 3\sin\mu\Delta x \right)/2 + \cos \mu\Delta x \right)$$

( For the leap-frog $\sin \alpha\Delta t = -\frac{c\Delta t}{\Delta x}$ )

The solution is stable (neutral solution with no damping or amplification) if $\alpha$ is real or $|\sin \alpha\Delta t| \leq 1$. To ensure stability for all wavelengths, it is necessary to find the maximum magnitude of the RHS of (18)

Maximum when $\mu\Delta x = 120° \Rightarrow$

$$|c\,\Delta t/\Delta x| \leq 1/\sqrt{3}$$

which is more restrictive than the leapfrog FD scheme. However it gives even better phase speed than the fourth-order leapfrog scheme

Finite elements are an interesting alternative to classic FD methods. They offer a high level of flexibility offered for the use of grids of

variable size, shape and flexibility and are attractive despite a higher cost in computer time. They are popular in the engineering field and central domain. For a review, Le Provost (1985) in O'Brien

## 7.4 The Spectral and Pseudo-Spectral method applied to the non-linear advection equation (Burger)

Before the advent of the fast Fourier Transform (FFT) spectral methods played only a minor role in fluid dynamics because they were far less economical than grid points methods.

We want to solve $\mathcal{L}(u) = \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x}$

Let assume a cyclic boundary condition on the domain $-1 \leq x \leq 1$

We use for the basis function, the trigonometric

$$\phi_j(x) = e^{i\pi j x} \qquad (j = -J, J) \tag{19}$$

The Galerkin approximation is then

$$\left\langle \phi_m, \sum_j A'_j \phi_j \right\rangle + \left\langle \phi_m, \left(\sum A_j \phi_j\right)\left(\sum A_k \phi'_k\right) \right\rangle = 0 \tag{20}$$

$$m = -J, \dots J$$

and $\langle \phi_i, \phi_j \rangle = \int_R \phi_i \, \overline{\phi_j}$ ) conjugate (in complex numbers)

The general form of this set of equation is

$$\sum_j a_{mj} A'_j + \sum_j \sum_k b_{mjk} A_j A_k = 0 \qquad m = -J \to J \tag{21}$$

To advance the solution by one step, one needs $(2\sigma+1)^3$ multiplications for the "interaction" term (RHS of (21)) and an inversion of the matrix of coefficients $a_{m_s}$. The basis functions are orthogonal $\Longrightarrow$

$$\langle \phi_{m'}, \phi_s \rangle = \begin{cases} 2 & m = s \\ 0 & \text{otherwise} \end{cases}$$

$$\langle \phi_m, \phi_s \phi'_k \rangle = \begin{cases} 2i\pi k & m = s+k \\ 0 & \text{otherwise} \end{cases}$$

(21) then reduces to

$$A'_m + \sum_{s+k=m} i\pi k A_s A_k = 0 \tag{22}$$

Thus, in practice, the inversion problem does not arise and the operation count for the interaction terms is only $\sigma^2$. The operation count for grid points methods on the other is proportional to to the number of grid points $(2\sigma+1)$. The difference is significant and acted as a major deterrent in the past

the FFT can improve the speed of the spectral method. There is not apparent need to perform Fourier transforms except at $t=0$ when the initial condition have to be transformed from physical to phase space and vice-versa at the end the speed of these transforms does not affect the overall efficiency of the method.

The so-called "pseudo-spectral" method produces similar results to the spectral method by transforming the variables back and forth between grid point and phase space every few steps

The transform (Pseudo-Spectral) method uses the series at certain spatial grid points and these fields are multiplied together at each point to form the new linear terms. Then these terms are transformed back to spectral space. The usefulness of this method is enhanced by the existence of efficient transform methods such as the FFTs. This method is essentially a grid point method which uses spectral decomposition techniques to eliminate the problem of FD in space, namely the phase retardation of short waves by computing the spatial derivatives by differentiating the individual Fourier components.

We define the grid points by $x_j = j/J$ for $-J < j < J$. With $\phi_k(x_j) = e^{i\pi k x_j}$ as basis functions we have

$$U_j(t) \equiv U(x_j, t) = \sum_{k=-J}^{J-1} A_k(t)\, e^{i\pi k x_j} \tag{23}$$

Note that the number of Fourier components matches the number of grid points (Cyclic). Also, if $k = J$, then we have two identical basis functions.

The orthogonality relation is then

$$\langle \phi_p, \phi_q \rangle = \frac{1}{J} \sum_{j=-J}^{J-1} e^{i\pi(p-q)x_j} = \begin{cases} 2 & p-q \overset{\text{Modulo}}{=} (2J) \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

The inverse transform of (23) is then

$$A_k(t) = \frac{1}{2J} \sum_{j=-J}^{J-1} U_j(t)\, e^{-i\pi k x_j} \tag{25}$$

Starting with grid point values $U_j$, the

the coefficients $A_k$ can be computed from (25)
The derivative $\partial U_j / \partial x$ is obtained by a second transform

$$\frac{\partial U_j(t)}{\partial x} = \sum_{k=-J}^{J-1} i\pi k A_k(t) e^{i\pi k x_j} \tag{26}$$

Time integration is normally done by finite-differencing. An alternative method is to perform the time integration in phase space, but then transform the dependent variables back and forth to grid point space for the evaluation of the non linear terms.

The final equation which approximate the advection equation is then

$$\frac{\partial U_j(t)}{\partial t} + \underbrace{U_j(t)}_{\text{Burger}} \underbrace{c}_{\text{advection}} \sum_{k=-J}^{J-1} i\pi k A_k(t) e^{i\pi k x_j} = 0 \tag{27}$$

Let's now compare the two methods (Spectral (S) and Pseudo Spectral (PS)).

We first transform (27) (PS) into an equation for the $A_k$

$$(2J) A'_k(t) + \sum_p \sum_q \sum_j i\pi q A_p(t) A_q(t) e^{i\pi(p+q-k)x_j} = 0 \tag{28}$$

Using the orthogonality of the basis functions, we can rewrite (28) as

(29) $$A'_k(t) + \sum_{p+q=k} i\pi q\, A_p A_q + \sum_{p+q=k+2J} i\pi q\, A_p A_q$$

$$+ \sum_{p+q=k-2J} i\pi q\, A_p A_q = 0$$

(since the orthogonality is modulo $2\sigma$)

The last two sums show a significant difference between (P) and (PS). They are referred to as "aliased" terms brought by the finite sampling interval of the discrete Fourier decomposition. This method is clearly much faster since the number of operations is $(2\sigma+1)\log_2(2\sigma+1)$ versus $(2\sigma+1)^2$ for the interaction method (P)

There are two basic techniques for removing the aliasing error introduced in (29)

(1) Aliasing removal by Padding or Truncation. The key is to use a discrete transform with $M$ rather than $N$ points where $M \geq \frac{3N}{2}$

(2) Aliasing removal by phase shifts

Both methods can be extended to two and three dimensions. For a complete description of spectral methods

"Spectral Methods in Fluid Dynamics"
by Canuto
Hussaini
Quarteroni
Zang
Springer-Verlag
1988