Numerical Methods[1]

Alejandro Campos

March 20, 2025

[1]Disclaimer: most of the work in this document is not original; instead, it consists of personal notes that folow from various books, such as Lomax, Pulliam, Zing; Larsson, Thomee; Hirsch; Leveque; Blazek; Laney; Moin.

Contents

Part I

Introductory Concepts

- Analytical solution u satisfies

$$D(u) = Au - f = 0 \quad \text{in } \Omega, \tag{1}$$

where A is the differential operator.

- Discrete numerical solution U_j satisfies

$$D_h(U_j^n) = 0 \quad \text{for all } n, j, \tag{2}$$

where D_h is the discrete operator for D.

- Continuous numerical solution v satisfies

$$D_h v = 0 \quad \text{in } \Omega, \tag{3}$$

Thus, $v(t_n, x_j) = U_j^n$ for all n, j.

- Local truncation error

$$l_j^{n+1} = u(t_{n+1}, x_j) - U_j^{n+1}, \tag{4}$$

given that $U_j^n = u(t_n, x_j)$ for all j.

- Global truncation error

$$e_j^{n+1} = u(t_{n+1}, x_j) - U_j^{n+1}, \tag{5}$$

given that $U_j^0 = u(t_0, x_j)$ for all j.

- Truncation error τ: difference between the discrete and analytical equations when applied to a given continuous function g

$$\tau = D_h g - Dg. \tag{6}$$

The truncation error allows us to compute global truncation errors. Also, consider the truncation error for the analytical solution, that is, $\tau = D_h u - Du$. Using equation (1), one obtains that u is also the solution to $D_h u = \tau$. This means that the **analytical solution satisfies the numerical equation but with an additional source term equal to** τ. Conversely, consider the truncation error for the function v, $\tau = D_h v - Dv$. Using equation (3), one obtains that v is also a solution to $Dv = \tau$. This means that the **numerical solution satisfies the analytical equation but with an additional source term equal to** τ.

- Order of accuracy r defined by $\tau^n = \mathcal{O}(\Delta t^r)$ as $\Delta t \to 0$ for ODE's and $\tau^n = \mathcal{O}(\Delta x^r)$ as $\Delta x \to 0$ for PDE's, where Δt is related to Δx.

Part II

Numerical Solution of ODE's

Chapter 1

List of time integrators

Consider the IVP for $u = u(t)$

$$\frac{du}{dt} = f(t,u) \quad \text{in } (0,\infty) \tag{1.1}$$

with initial condition $u(0) = u^0$. We will discretize time into a set of finite values t^n, and express the numerical solution at a given time t^n as U^n.

1.1 Explicit

- Explicit Euler method (Forward Euler method)

$$u^{n+1} = u^n + \Delta t f(t_n, u^n) \qquad 1^{\text{st}} \text{ O.A.} \tag{1.2}$$

- Explicit Midpoint (Modified Euler, 2^{nd} order RK)

$$\begin{aligned} u^{n+1/2} &= u^n + \frac{\Delta t}{2} f(t_n, u^n) \\ u^{n+1} &= u^n + \Delta t f(t_{n+1/2}, u^{n+1/2}) \end{aligned} \tag{1.3}$$

1.2 Implicit

- Implicit Euler (Backward Euler)

$$u^{n+1} = u^n + \Delta t f(t_{n+1}, u^{n+1}) \qquad 1^{\text{st}} \text{ O.A.} \tag{1.4}$$

- Implicit Trapezoidal (Crank-Nicholson)

$$u^{n+1} = u^n + \frac{\Delta t}{2}[f(t_n, u^n) + f(t_{n+1}, u^{n+1})] \qquad 2^{\text{nd}} \text{ O.A.} \tag{1.5}$$

- Implicit Midpoint

$$u^{n+1} = u^n + \Delta t f\left(t_{n+1/2}, \frac{1}{2}\left(u^n + u^{n+1}\right)\right) \tag{1.6}$$

1.3 Predictor-Corrector

- Heun's (Explicit Trapezoidal, Improved Euler)

$$\begin{aligned} \tilde{u}^{n+1} &= u^n + \Delta t f(t_n, u^n) \\ u^{n+1} &= u^n + \frac{\Delta t}{2}[f(t_n, u^n) + f(t_{n+1}, \tilde{u}^{n+1})] \quad 2^{\text{ nd}} \text{ O.A.} \end{aligned} \tag{1.7}$$

1.4 Runge-Kutta

- 4^{th} order Runge-Kutta

$$\begin{aligned} k_1 &= f(t_n, u^n) \\ k_2 &= f(t_{n+1/2}, u^n + \frac{\Delta t}{2} k_1) \\ k_3 &= f(t_{n+1/2}, u^n + \frac{\Delta t}{2} k_2) \\ k_4 &= f(t_n, u^n + \Delta t k_3) \\ u^{n+1} &= u^n + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4) \end{aligned} \tag{1.8}$$

1.5 Multi-Step

- Leapfrog: consider the system of equations for $x = x(t)$ and $v = v(t)$

$$\frac{dx}{dt} = v \qquad \frac{dv}{dt} = f(x). \tag{1.9}$$

  The Leapfrog method is a staggered scheme defined as follows

$$v^{n+1/2} = v^{n-1/2} + f(x^n)\Delta t \tag{1.10}$$

$$x^{n+1} = x^n + v^{n+1/2}\Delta t \tag{1.11}$$

- Adams-Bashforth

Chapter 2

Solving non-linear equations

The implicit schemes result in non-linear equations that require some sort of non-linear solver. For this chapter, we'll assume the governing ODE is

$$\frac{du}{dt} = f(u) \qquad \text{in } (0, \infty), \tag{2.1}$$

or, for a system of equations,

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}) \qquad \text{in } (0, \infty). \tag{2.2}$$

2.1 Newton's method

The Euler scheme requires finding u^{n+1} so that the following is satisfied

$$u^{n+1} = u^n + \Delta t f(u^{n+1}). \tag{2.3}$$

That is, we need to find the root of the nonlinear equation

$$g(x) = x - u^n - \Delta t f(x). \tag{2.4}$$

Newton's method finds the root of a non-linear equation $g(x)$ by iterating through m as follows

$$x^{m+1} = x^m - \frac{g(x^m)}{g'(x^m)}. \tag{2.5}$$

Applying Newton's method to eq. (2.4), we have

$$x^{m+1} = x^m - \frac{x^m - u^n - \Delta t f(x^m)}{1 - \Delta t f'(x^m)}, \tag{2.6}$$

The initial value for the non-linear solver is $x^0 = u^n$. Thus, using the above, u^{n+1} is approximated by x^m as $m \to \infty$. Note that this is only one example of what is referred to as a fixed-point iteration. There are other fixed-point-iteration methods that can be used to solve eq. (2.4) that are not Newton's method.

As a side note, let's assume u^{n+1} is approximated sufficiently well by x^1, that is, only one Newton iteration is needed. Then we have

$$u^{n+1} = u^n + \frac{\Delta t f(u^n)}{1 - \Delta t f'(u^n)}, \tag{2.7}$$

which we re-write as

$$\left(\frac{1}{\Delta t} - f'(u^n)\right)\Delta u = f(u^n), \tag{2.8}$$

where $\Delta u = u^{n+1} - u^n$. The above is the approximation that gets used for pseudo-time stepping, so that the Backward Euler scheme can converge to the steady state solution as one iterates through n.

To solve for the root of a system of non-linear equations $\mathbf{g}(\mathbf{x})$, Newton's method is as follows

$$\mathbf{x}^{m+1} = \mathbf{x}^m - \mathbf{J}^{-1}(\mathbf{x}^m)\mathbf{g}(\mathbf{x}^m). \tag{2.9}$$

In the above, $\mathbf{J}^{-1}(\mathbf{x})$ is the inverse of the Jacobian matrix $\mathbf{J}(\mathbf{x})$, which is given by

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \\ \vdots & & \ddots \end{pmatrix}. \tag{2.10}$$

For the backward Euler scheme, the equivalent of eq. (2.6) would be

$$\mathbf{x}^{m+1} = \mathbf{x}^m - \left(\mathbf{I} - \Delta t \left.\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}^m}\right)^{-1} \left(\mathbf{x}^m - \mathbf{u}^n - \Delta t \mathbf{f}(\mathbf{x}^m)\right), \tag{2.11}$$

and the equivalent of eq. (2.8) would be

$$\left(\frac{\mathbf{I}}{\Delta t} - \left.\frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{u}^n}\right)\Delta \mathbf{u} = \mathbf{f}(\mathbf{u}^n), \tag{2.12}$$

where $\Delta \mathbf{u} = \mathbf{u}^{n+1} - \mathbf{u}^n$.

Part III

Finite Difference for PDE's

Chapter 3

Introduction

3.1 Finite difference formulas

Forward	Backward	Central	Central (2nd order)
$\partial U_j = \frac{U_{j+1} - U_j}{h}$	$\bar{\partial} U_j = \frac{U_j - U_{j-1}}{h}$	$\hat{\partial} U_j = \frac{U_{j+1} - U_{j-1}}{2h}$	$\partial\bar{\partial} U_j = \frac{U_{j+1} - 2U_j + U_{j-1}}{h^2}$

3.2 Fourier analysis

This analysis is based on a test function that is periodic and thus has the following form

$$u(x) = \sum_{n=-\infty}^{n=\infty} \hat{u}_n e^{i\kappa_n x}. \tag{3.1}$$

We are interested in using the different numerical methods to compute the resulting approximate derivatives of the generic mode $e^{i\kappa x}$. We note that one can then combine the approximate derivatives of all the modes to obtain the approximate derivative of our test function u.

The analytically derivative of our generic mode $e^{i\kappa x}$ is $i\kappa e^{i\kappa x}$. We now assume that the approximate derivatives at a point x_j obtained using numerical schemes will be of the form $i\kappa^* e^{i\kappa x_j}$, where κ^* is a modified wave number. The closer κ^* is to κ the higher the accuracy of the numerical scheme. For example, for the central first order scheme

$$\delta U_j = \frac{U_{j+1} - U_{j-1}}{2h} \tag{3.2}$$

we have

$$i\kappa^* e^{i\kappa x_j} = \frac{e^{i\kappa(x_j+h)} - e^{i\kappa(x_j-h)}}{2h} \tag{3.3}$$

which leads to

$$\kappa^* = \frac{\sin \kappa h}{h}. \tag{3.4}$$

For the sixth-order compact scheme

$$\alpha\delta U_{j-1} + \delta U_j + \alpha\delta U_{j+1} = \frac{a}{2h}\left(U_{j+1} - U_{j-1}\right) + \frac{b}{4h}\left(U_{j+2} - U_{j-2}\right) \tag{3.5}$$

we have

$$\alpha i\kappa^* e^{i\kappa(x_j-h)} + i\kappa^* e^{i\kappa x_j} + \alpha i\kappa^* e^{i\kappa(x_j+h)} = \frac{a}{2h}\left[e^{i\kappa(x_j+h)} - e^{i\kappa(x_j-h)}\right] + \frac{b}{2h}\left[e^{i\kappa(x_j+2h)} - e^{i\kappa(x_j-2h)}\right] \tag{3.6}$$

which leads to

$$\kappa^* = \frac{\frac{a}{h}\sin\kappa h + \frac{b}{2h}\sin 2\kappa h}{1 + 2\alpha\cos\kappa h}. \tag{3.7}$$

A spectral method, on the other hand, will explicitly express the derivative of a mode as $i\kappa e^{i\kappa x_j}$, up to the last mode $\kappa = \frac{2\pi}{L}\frac{N}{2}$, and will not be able to capture higher modes. Thus,

$$\kappa^* = \begin{cases} \kappa & \text{for } \frac{2\pi}{L}\left(-\frac{N}{2}+1\right) \le \kappa \le \frac{2\pi}{L}\frac{N}{2} \\ 0 & \text{o.w.} \end{cases} \tag{3.8}$$

Chapter 4

Elliptic

Define the following norm: $|U|_S = \max_{x_j \in S} |U_j|$.
For $Au = -au'' + cu = f$ in $\Omega = (0,1)$, where $a(x) > 0$ and $c(x) \geq 0$, with boundary conditions $u(0) = U_0$ and $u(1) = U_m$, and $A_h = -a_j \partial\bar{\partial} U_j + c_j U_j$:

- **Lemma 4.2**
$$|U|_{\bar{\Omega}} \leq \max\{|U_0|, |U_M|\} + C|A_h U|_\Omega$$

- **Theorem 4.1** The error bound follows,
$$|U - u|_\Omega \leq Ch^2 ||u||_C^4$$

For $Au = -\Delta u = f$ in $\Omega = (0,1) \times (0,1)$, with boundary doncitions $u = 0$ in Γ, and $A_h = -\Delta_h = -\partial_1\bar{\partial}_1 U_j - \partial_2\bar{\partial}_2 U_j$:

- **Lemma 4.4**
$$|U|_{\bar{\Omega}} \leq |U|_\Gamma + C|\Delta_h U|_\Omega$$

- **Theorem 4.2** The error bound follows,
$$|U - u|_\Omega \leq Ch^2 ||u||_C^4$$

Chapter 5

Parabolic

For $u_t = u_{xx}$ in $\mathbf{R} \times \mathbf{R}_+$, with initial condition $u(\cdot, 0) = v$ in $\mathbf{R}$.

- Each scheme can be associated with its discrete solution operator E_k, defined in either of the following two ways, where U_j^n is defined only at mesh points, and $u^n(x)$ is the corresponding numerical solution defined over all space,

$$U_j^{n+1} = (E_k U^n)_j = \sum_p a_p U_{j-p}^n$$

$$u^{n+1}(x) = (E_k u^n)(x) = \sum_p a_p u^n(x - x_p)$$

- Repeated application yields $U_j^n = (E_k^n V)_j$ or $u^n(x) = (E_k^n v)(x)$, where V_i and $v(x)$ are the initial conditions, specified at mesh points and over all space, respectively.

- The **stability** of E_k (and hence the associated scheme) is defined by $||U^n||_{l_p} = ||E_k^n V||_{l_p} \leq ||V||_{l_p}$ and $||u^n||_{L_p} = ||E_k^n v||_{L_p} \leq ||v||_{L_p}$.

- The **symbol** or characteristic polynomial of E_k is defined as follows,

$$\tilde{E}(\xi) = \sum_p a_p e^{-ip\xi}.$$

  If we make use of the following Fourier series and Fourier transform,

$$\hat{V}(\xi) = h \sum_{j=-\infty}^{\infty} V_j e^{-ij\xi}$$

$$\hat{v}(\xi) = \int_{-\infty}^{\infty} v(x) e^{-ix\xi} dx$$

  then we obtain $(E_k V)\hat{}(\xi) = \tilde{E}(\xi)\hat{V}(\xi)$ and $(E_k v)\hat{}(\xi) = \tilde{E}(h\xi)\hat{v}(\xi)$, which for repeated application leads to,

$$(E_k^n V)\hat{}(\xi) = \tilde{E}^n(\xi)\hat{V}(\xi)$$
$$(E_k^n v)\hat{}(\xi) = \tilde{E}^n(h\xi)\hat{v}(\xi).$$

- Using Parseval's theorem one arrives at the **von Neumann's stability condition**, namely, $|\tilde{E}(\xi)| \leq 1|$ for all ξ is sufficient for stability in l_2 and L_2, and necessary for stability in l_∞.

- The **order of accuracy** r is defined by $\tau^n = \mathcal{O}(h^r)$ as $h \to 0$. Since $u^{n+1}(x) = E_k u^n(x) + k\tau^n(x)$, where $u^n(x)$ is now the analytical solution and k is the time step, then the order of accuracy is also obtained from $u^{n+1}(x) - E_k u^n(x) = k\mathcal{O}(h^r)$. This expression for order of accuracy is equivalent to $\tilde{E}(\xi) = e^{-\lambda\xi^2} + \mathcal{O}(|\xi|^{r+2})$, as $\xi \to 0$.

Chapter 6

Hyperbolic

For $u_t = au_x$ in $\mathbf{R} \times \mathbf{R}_+$, with initial condition $u(\cdot, 0) = v$ in $\mathbf{R}$:

- As for parabolic equations, von Neumann's condition $|\tilde{E}(\xi)| \leq 1$ for all ξ is a necessary and sufficient condition for stability in the L_2 norm.
- Order of accuracy is defined in the same manner as for the parabolic case, except that now the symbol of E_k is given by $\tilde{E}(\xi) = e^{ia\lambda\xi} + \mathcal{O}(|\xi|^{r+1})$ as $\xi \to 0$.
- **CFL condition** for stability: a necessary condition for stability is that the domain of dependence of the finite difference scheme at (x, t) contains the domain of dependence of the continuous problem.
- The Friedrichs scheme (first order accurate) follows,

$$(E_k U^n)(x) = 1/2(1 + a\lambda)U^n(x + h) + 1/2(1 - a\lambda)U^n(x - h)$$

- The Lax-Wendroff scheme (second order accurate) follows,

$$(E_k U^n)(x) = 1/2(a^2\lambda^2 + a\lambda)U^n(x + h) + (1 - a^2\lambda^2)U^n(x) + 1/2(a^2\lambda^2 - a\lambda)U^n(x - h)$$

Part IV

Finite Volume for PDE's

Chapter 7

Introduction

7.1 Some schemes for high-speed flows

- Central
  - Jameson
- Upwinding
  - Flux Vector Splitting
    - Steger-Warming
    - AUSM
  - Flux Difference Splitting
    - Godunov
    - Roe
  - Total Variation Diminishing

Chapter 8

Elliptic

Chapter 9

Parabolic

Chapter 10

Hyperbolic

10.1 One-dimensional case

Consider the hyperbolic equation

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = 0, \tag{10.1}$$

where $f = f(u)$ is the flux of u. The finite volume method consists on discretizing the spatial domain into volumes, which we will denote by the index i, and which are defined by $x \in [x_{i-1/2}, x_{i+1/2}]$, where $x_{i-1/2}$ and $x_{i+1/2}$ represent the boundaries of the finite volume i.

We now proceed by averaging the equation over each control volume, that is

$$\frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{1+1/2}} \frac{\partial u}{\partial t} dx + \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{1+1/2}} \frac{\partial f}{\partial x} dx = 0, \tag{10.2}$$

where $\Delta x_i = x_{i+1/2} - x_{i-1/2}$. Moving the time derivative and using the Divergence theorem, the above becomes

$$\frac{d}{dt}\left(\frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} u\, dx\right) + \frac{1}{\Delta x_i}(f_{i+1/2} - f_{i-1/2}) = 0, \tag{10.3}$$

where $f_{i\pm 1/2}$ is the flux evaluated at $x_{i\pm 1/2}$. The spatially-discrete numerical solution then satisfies

$$\frac{dU_i}{dt} + \frac{1}{\Delta x_i}(F_{i+1/2} - F_{i-1/2}) = 0, \tag{10.4}$$

where $U_i = U_i(t)$ is the discrete solution at some specific point within the finite volume, and is used to approximate the average of u. $F_{i\pm 1/2}$ are the so-called numerical fluxes.

Godunov Scheme:

$$F^n_{j+1/2} = \begin{cases} \min_{U_j \le u \le U_{j+1}} f(u) & \text{for} \quad U_j < U_{j+1} \\ \min_{U_{j+1} \le u \le U_j} f(u) & \text{for} \quad U_j > U_{j+1} \end{cases}$$

Lax-Friedrichs:

$$F^n_{j+1/2} = \frac{h}{2k}(U_j - U_{j+1} + \frac{1}{2}(f(U_j) + f(U_{j+1}))$$

10.2 Multi-dimensional case

We follow a similar approach for the multidimensional case. Consider a generic conservation equation

$$\frac{\partial w_i}{\partial t} + \frac{\partial f^{(c)}_{ij}}{\partial x_j} = \frac{\partial f^{(v)}_{ij}}{\partial x_j} + q_i \tag{10.5}$$

where w_i is the vector of conservative variables, $f_{ij}^{(c)}$ the convective flux tensor, $f_{ij}^{(v)}$ the viscous flux tensor, and q_i some heat source.

Averaging the equation over a generic finite volume denoted by the index I, one obtains

$$\frac{d}{dt}\left(\frac{1}{\Omega_I}\int_{\Omega_I} w_i\, dV\right) + \frac{1}{\Omega_I}\int_{\delta\Omega_I} f_i^{(c)}\, dS = \frac{1}{\Omega_I}\int_{\delta\Omega_I} f_i^{(v)}\, dS + \frac{1}{\Omega_I}\int_{\Omega_I} q_i\, dV, \tag{10.6}$$

where $f_i^{(c)} = f_{ij}^{(c)} n_j$ and $f_i^{(v)} = f_{ij}^{(v)} n_j$ are the vectors of convective and viscous fluxes, respectively. In vector notation, this is written as

$$\frac{d}{dt}\left(\frac{1}{\Omega_I}\int_{\Omega_I} \mathbf{w}\, dV\right) + \frac{1}{\Omega_I}\int_{\delta\Omega_I} \mathbf{f}^{(c)}\, dS = \frac{1}{\Omega_I}\int_{\delta\Omega_I} \mathbf{f}^{(v)}\, dS + \frac{1}{\Omega_I}\int_{\Omega_I} \mathbf{q}\, dV. \tag{10.7}$$

A specific example is the Navier-Stokes equations, for which we have

$$\mathbf{w} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho w \\ \rho E \end{bmatrix} \qquad \mathbf{f}^{(c)} = \begin{bmatrix} \rho(u_j n_j) \\ \rho u(u_j n_j) + p n_1 \\ \rho v(u_j n_j) + p n_2 \\ \rho w(u_j n_j) + p n_3 \\ \rho\left(E + \frac{p}{\rho}\right)(u_j n_j) \end{bmatrix} \qquad \mathbf{f}^{(v)} = \begin{bmatrix} 0 \\ \tau_{1j} n_j \\ \tau_{2j} n_j \\ \tau_{3j} n_j \\ u_i \tau_{ij} n_j + \kappa \frac{\partial T}{\partial x_j} n_j \end{bmatrix} \qquad \mathbf{q} = \begin{bmatrix} 0 \\ \rho g_1 \\ \rho g_2 \\ \rho g_3 \\ \rho u_i g_i \end{bmatrix}. \tag{10.8}$$

The spatially-discrete numerical solution then satisfies

$$\frac{d\mathbf{W}_I}{dt} + \frac{1}{\Omega_I}\sum_{K\in N(I)} \mathbf{F}_K^{(c)}\, \Delta S_K = \frac{1}{\Omega_I}\sum_{K\in N(I)} \mathbf{F}_K^{(v)}\, \Delta S_K + \mathbf{Q}_I \tag{10.9}$$

where $\mathbf{W}_I = \mathbf{W}_I(t)$ is the numerical solution at some specific point within the finite volume, and is used to approximate the average of the vector $\mathbf{w}$. The faces of the finite volumes are indexed, and the set $N(I)$ consists of the indices of the faces of the finite volume I. The variables $\mathbf{F}_K^{(c)} = \mathbf{F}_K^{(c)}(\mathbf{W}_1, \mathbf{W}_2, \dots)$ and $\mathbf{F}_K^{(v)} = \mathbf{F}_K^{(v)}(\mathbf{W}_1, \mathbf{W}_2, \dots)$ are the numerical fluxes, and ΔS_K the surface area, for face K. The averaged source term $\mathbf{Q}_I = \mathbf{Q}_I(\mathbf{W}_I)$ is approximated as the source vector $\mathbf{q}$ evaluated using $\mathbf{W}_I$. The above is typically rewritten as

$$\frac{d\mathbf{W}_I}{dt} = -\frac{1}{\Omega_I}\mathbf{R}_I, \tag{10.10}$$

where the residual $\mathbf{R}_I = \mathbf{R}_I(\mathbf{W}_1, \mathbf{W}_2, \dots)$ is given by

$$\mathbf{R}_I = \sum_{K\in N(I)} \mathbf{F}_K^{(c)}\, \Delta S_K - \sum_{K\in N(I)} \mathbf{F}_K^{(v)}\, \Delta S_K - \mathbf{Q}_I\, \Omega_I \tag{10.11}$$

The Jacobian is then

$$\frac{\partial \mathbf{R}_I}{\partial \mathbf{W}_J} = \sum_{K\in N(I)} \frac{\partial \mathbf{F}_K^{(c)}}{\partial \mathbf{W}_J}\, \Delta S_K - \sum_{K\in N(I)} \frac{\partial \mathbf{F}_K^{(v)}}{\partial \mathbf{W}_J}\, \Delta S_K - \frac{\partial \mathbf{Q}_I}{\partial \mathbf{W}_J}\, \Omega_I \tag{10.12}$$

Note the the Jacobian is only non zero when $J \in M(I)$, where $M(I)$ represents the set of all finite volumes that the numerical fluxes of finite volume I depend on.

10.2.1 An example from turbulence modeling

Consider the SST turbulence model. The two transport equations solved by the model are

$$\frac{\partial \rho k}{\partial t} + \frac{\partial \rho k u_j}{\partial x_j} = P - \beta^* \rho w k + \frac{\partial}{\partial x_j}\left[(\mu + \sigma_k \mu_t)\frac{\partial k}{\partial x_j}\right] \tag{10.13}$$

$$\frac{\partial \rho w}{\partial t} + \frac{\partial \rho w u_j}{\partial x_j} = \frac{\gamma}{\nu_t} P - \beta \rho w^2 + \frac{\partial}{\partial x_j}\left[(\mu + \sigma_w \mu_t)\frac{\partial w}{\partial x_j}\right] + 2(1 - F_1)\frac{\rho \sigma_{w2}}{w}\frac{\partial k}{\partial x_j}\frac{\partial w}{\partial x_j} \tag{10.14}$$

The vector of conservative variables is $\mathbf{w} = [\rho k, \rho w]^T$. The convective and viscous flux vectors are

$$\mathbf{f}^{(c)} = \begin{bmatrix} \rho k (u_j n_j) \\ \rho w (u_j n_j) \end{bmatrix} \qquad \mathbf{f}^{(v)} = \begin{bmatrix} (\mu + \sigma_k \mu_t)\frac{\partial k}{\partial x_j} n_j \\ (\mu + \sigma_w \mu_t)\frac{\partial w}{\partial x_j} n_j \end{bmatrix}. \tag{10.15}$$

The source $\mathbf{q}$ is

$$\mathbf{q} = \begin{bmatrix} P - \beta^* \rho w k \\ \frac{\gamma}{\nu_t} P - \beta \rho w^2 + 2(1 - F_1)\frac{\rho \sigma_{w2}}{w}\frac{\partial k}{\partial x_j}\frac{\partial w}{\partial x_j} \end{bmatrix} \tag{10.16}$$

Convective numerical fluxes: A simple upwinding scheme is as follows

$$\mathbf{F}_K^{(c)} = \begin{bmatrix} (\rho k)_l a_0 + (\rho k)_r a_1 \\ (\rho w)_l a_0 + (\rho w)_r a_1 \end{bmatrix} \tag{10.17}$$

where subscripts l and r denote values at the center nodes of the control volumes to the left and right of the K^{th} control surface, respectively. The variables a_0 and a_1 are given by

$$a_0 = \begin{cases} q_{lr} & \text{if } q_{lr} > 0 \\ 0 & \text{o.w.} \end{cases} \qquad a_1 = \begin{cases} 0 & \text{if } q_{lr} > 0 \\ q_{lr} & \text{o.w.} \end{cases} \tag{10.18}$$

where

$$q_{lr} = \frac{(\mathbf{u}_l \cdot \mathbf{n}) + (\mathbf{u}_r \cdot \mathbf{n})}{2}. \tag{10.19}$$

The corresponding Jacobian is

$$\frac{\partial \mathbf{F}_K^{(c)}}{\partial \mathbf{W}_J} = \begin{bmatrix} \frac{\partial \mathbf{F}_K^{(c)}(1)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{F}_K^{(c)}(1)}{\partial \mathbf{W}_J(2)} \\ \frac{\partial \mathbf{F}_K^{(c)}(2)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{F}_K^{(c)}(2)}{\partial \mathbf{W}_J(2)} \end{bmatrix} = \begin{bmatrix} a_0 \delta_{lJ} + a_1 \delta_{rJ} & 0 \\ 0 & a_0 \delta_{lJ} + a_1 \delta_{rJ} \end{bmatrix}. \tag{10.20}$$

Viscous numerical fluxes: the viscous fluxes can be discretized as follows

$$\mathbf{F}_K^{(v)} = \begin{bmatrix} (\mu + \sigma_k \mu_t)_{avg}\left[\frac{k_r - k_l}{l_{lr}}\mathbf{t}_{lr} + \nabla k_{avg} - (\nabla k_{avg} \cdot \mathbf{t}_{lr})\,\mathbf{t}_{lr}\right] \cdot \mathbf{n} \\ (\mu + \sigma_\omega \mu_t)_{avg}\left[\frac{\omega_r - \omega_l}{l_{lr}}\mathbf{t}_{lr} + \nabla \omega_{avg} - (\nabla \omega_{avg} \cdot \mathbf{t}_{lr})\,\mathbf{t}_{lr}\right] \cdot \mathbf{n} \end{bmatrix}. \tag{10.21}$$

where the subscript $_{avg}$ means the left and right values have been averaged, $\mathbf{t}_{lr}$ is the unit vector that points from the center of the l^{th} control volume (the one to the left of the face) to the center of the r^{th} control volume (the one to the right of the face), and l_{lr} is the length between these two centers. The corresponding Jacobian is approximated as follows

$$\frac{\partial \mathbf{F}_K^{(v)}}{\partial \mathbf{W}_J} = \begin{bmatrix} \frac{\partial \mathbf{F}_K^{(v)}(1)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{F}_K^{(v)}(1)}{\partial \mathbf{W}_J(2)} \\ \frac{\partial \mathbf{F}_K^{(v)}(2)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{F}_K^{(v)}(2)}{\partial \mathbf{W}_J(2)} \end{bmatrix}, \tag{10.22}$$

where

$$\frac{\partial \mathbf{F}_K^{(v)}(1)}{\partial \mathbf{W}_J(1)} = (\mu + \sigma_k \mu_t)_{avg} \frac{1}{l_{lr}} \mathbf{t}_{lr} \cdot \mathbf{n} \frac{\delta_{rJ}}{\rho_r} - (\mu + \sigma_k \mu_t)_{avg} \frac{1}{l_{lr}} \mathbf{t}_{lr} \cdot \mathbf{n} \frac{\delta_{lJ}}{\rho_l}, \tag{10.23}$$

$$\frac{\partial \mathbf{F}_K^{(v)}(2)}{\partial \mathbf{W}_J(2)} = (\mu + \sigma_\omega \mu_t)_{avg} \frac{1}{l_{lr}} \mathbf{t}_{lr} \cdot \mathbf{n} \frac{\delta_{rJ}}{\rho_r} - (\mu + \sigma_\omega \mu_t)_{avg} \frac{1}{l_{lr}} \mathbf{t}_{lr} \cdot \mathbf{n} \frac{\delta_{lJ}}{\rho_l} \tag{10.24}$$

and $\frac{\partial \mathbf{F}_K^{(v)}(1)}{\partial \mathbf{W}_J(2)} = \frac{\partial \mathbf{F}_K^{(v)}(2)}{\partial \mathbf{W}_J(1)} = 0$.

Sources: The discretized source terms are computed following eq. (10.16). The Jacobian is expressed as

$$\frac{\partial \mathbf{Q}_I}{\partial \mathbf{W}_J} = \begin{bmatrix} \frac{\partial \mathbf{Q}_I(1)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{Q}_I(1)}{\partial \mathbf{W}_J(2)} \\ \frac{\partial \mathbf{Q}_I(2)}{\partial \mathbf{W}_J(1)} & \frac{\partial \mathbf{Q}_I(2)}{\partial \mathbf{W}_J(2)} \end{bmatrix}. \tag{10.25}$$

According to Wilcox p. 413, the elements of the matrix above are approximated as follows

$$\frac{\partial \mathbf{Q}_I(1)}{\partial \mathbf{W}_J(1)} \approx -\beta^* w_I \delta_{IJ} \qquad \frac{\partial \mathbf{Q}_I(2)}{\partial \mathbf{W}_J(2)} \approx -2\beta w_I \delta_{IJ} \qquad \frac{\partial \mathbf{Q}_I(1)}{\partial \mathbf{W}_J(2)} = \frac{\partial \mathbf{Q}_I(2)}{\partial \mathbf{W}_J(1)} \approx 0. \tag{10.26}$$

10.3 Implicit time integration

We combine eq. (10.10) for all I into a single equation as follows

$$\frac{d}{dt} \begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \end{pmatrix} = \begin{pmatrix} -\frac{1}{\Omega_1} \mathbf{R}_1 \\ -\frac{1}{\Omega_2} \mathbf{R}_2 \\ \vdots \end{pmatrix}. \tag{10.27}$$

Note that the above is of the same form as eq. (2.2). Thus, the corresponding form of eq. (2.12) would be

$$\left[\frac{1}{\Delta t} \begin{pmatrix} \mathbf{I} & 0 & \dots \\ 0 & \mathbf{I} & \\ \vdots & & \ddots \end{pmatrix} - \begin{pmatrix} -\frac{1}{\Omega_1} \frac{\partial \mathbf{R}_1(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_1} & -\frac{1}{\Omega_1} \frac{\partial \mathbf{R}_1(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_2} & \dots \\ -\frac{1}{\Omega_2} \frac{\partial \mathbf{R}_2(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_1} & -\frac{1}{\Omega_2} \frac{\partial \mathbf{R}_2(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_1} & \\ \vdots & & \ddots \end{pmatrix}_{\mathbf{x}_\alpha = \mathbf{W}_\alpha^n} \right] \begin{pmatrix} \Delta \mathbf{W}_1 \\ \Delta \mathbf{W}_2 \\ \vdots \end{pmatrix}$$
$$= \begin{pmatrix} -\frac{1}{\Omega_1} \mathbf{R}_1(\mathbf{W}_1^n, \mathbf{W}_2^n, \dots) \\ -\frac{1}{\Omega_2} \mathbf{R}_2(\mathbf{W}_1^n, \mathbf{W}_2^n, \dots) \\ \vdots \end{pmatrix}. \tag{10.28}$$

Multiplying each of the major rows above one obtains

$$\left[\frac{1}{\Delta t} \begin{pmatrix} \Omega_1 \mathbf{I} & 0 & \dots \\ 0 & \Omega_2 \mathbf{I} & \\ \vdots & & \ddots \end{pmatrix} + \begin{pmatrix} \frac{\partial \mathbf{R}_1(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{R}_1(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_2} & \dots \\ \frac{\partial \mathbf{R}_2(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{R}_2(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_2} & \\ \vdots & & \ddots \end{pmatrix}_{\mathbf{x}_\alpha = \mathbf{W}_\alpha^n} \right] \begin{pmatrix} \Delta \mathbf{W}_1 \\ \Delta \mathbf{W}_2 \\ \vdots \end{pmatrix}$$
$$= \begin{pmatrix} -\mathbf{R}_1(\mathbf{W}_1^n, \mathbf{W}_2^n, \dots) \\ -\mathbf{R}_2(\mathbf{W}_1^n, \mathbf{W}_2^n, \dots) \\ \vdots \end{pmatrix}. \tag{10.29}$$

As previously noted, the only Jacobians $\partial \mathbf{R}_I / \partial \mathbf{W}_J$ that are non-zero are those for which $J \in M(I)$. Thus, the second submatrix of the above is mostly sparse. The Ith major row can be written as

$$\sum_{J=M(I)} \left[\frac{\Omega_I}{\Delta t} \delta_{IJ} + \left. \frac{\partial \mathbf{R}_I(\mathbf{x}_1, \mathbf{x}_2, \dots)}{\partial \mathbf{x}_J} \right|_{\mathbf{x}_\alpha = \mathbf{W}_\alpha^n} \right] \Delta \mathbf{W}_J = -\mathbf{R}_I(\mathbf{W}_1^n, \mathbf{W}_2^n, \dots). \tag{10.30}$$

Part V

Finite Element for PDE's

Chapter 11

Introduction

11.1 The finite-element method

The finite element method is a technique for solving partial differential equations in which the solution is represented as a sum of predefined functions, which are referred to as basis functions. This is not unlike the spectral method, but the key difference is that the basis functions of the finite element method are localized, that is, they are non-zero only among some specific regions of the domain, whereas the basis functions of the spectral method are defined over the entire domain.

The finite element method typically relies on the Galerkin method to obtain a solution. This is an additional difference with the spectral method, which can be based on either the Galerkin method or other approaches such as the collocation and tau methods. For completeness, the collocation method is briefly described in appendix A.

11.2 The Galerkin method

11.2.1 Weak form or variational formulation

Consider the problem in which a scalar $u = u(\mathbf{x}, t)$ is sought as the solution to the following

$$\begin{aligned} &\frac{\partial u}{\partial t} + Au = f && \text{in } \Omega \times \mathbf{R}_+ \\ &u = 0 && \text{on } \partial\Omega \times \mathbf{R}_+ \\ &u(\cdot, 0) = u_0 && \text{in } \Omega \end{aligned} \tag{11.1}$$

where A is any given linear operator, Ω is the spatial domain, $\partial\Omega$ its boundary, and $\mathbf{R}_+$ the set of positive real numbers. We now introduce $\mathcal{U}$ as the space of trial functions and $\mathcal{V}$ as the space of test functions. The weak formulation of the equation above is then to find the weak solution $u \in \mathcal{U}$ such that

$$\left(\frac{\partial u}{\partial t}, v\right) + (Au, v) = (f, v) \qquad \forall v \in \mathcal{V}, \tag{11.2}$$

where $(a, b) = \int_\Omega av\, dV$ is the L_2 inner product. Using the product rule and the divergence theorem, some of the derivatives in (Au, v) can be moved from operating on u to operating on v. We label the end result as $a(u, v)$, which is a bilinear form. We also label (f, v) as $L(v)$, which is a linear form. Thus, the above is rewritten as

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u, v) = L(v) \qquad \forall v \in \mathcal{V}. \tag{11.3}$$

Different linear operators A lead to different bilinear forms. Take as an example the operator $\nabla \cdot (\boldsymbol{\lambda} u)$, where $\boldsymbol{\lambda}$ is a vector. For this case, the product rule and the divergence theorem lead to

$$\begin{aligned}(\nabla \cdot (\boldsymbol{\lambda} u), v) &= \int_\Omega \nabla \cdot (\boldsymbol{\lambda} u) v \, dV \\ &= \int_\Omega \nabla \cdot (\boldsymbol{\lambda} u v) \, dV - \int_\Omega u \boldsymbol{\lambda} \cdot \nabla v \, dV \\ &= \int_{\partial\Omega} u v \boldsymbol{\lambda} \cdot \mathbf{n} \, dS - \int_\Omega u \boldsymbol{\lambda} \cdot \nabla v \, dV. \end{aligned} \tag{11.4}$$

We assume that either u, v, or $\boldsymbol{\lambda}$ is zero on $\partial\Omega$, and thus we get

$$(\nabla \cdot (\boldsymbol{\lambda} u), v) = -(u\boldsymbol{\lambda}, \nabla v). \tag{11.5}$$

Some examples of operators and the corresponding bilinear forms are shown in the table below

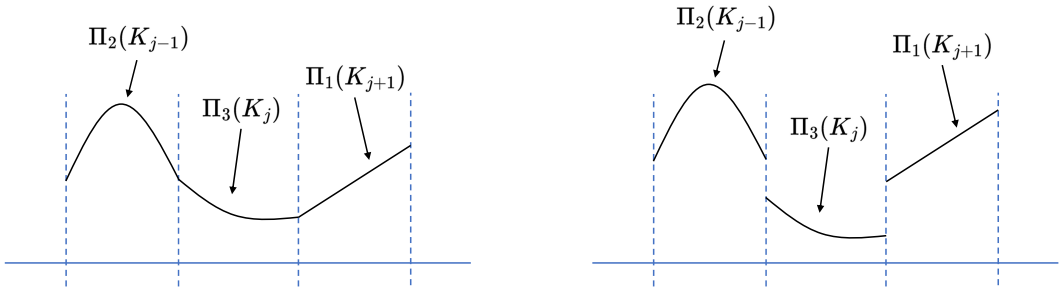
Operator	Definition	Bilinear Form
Mass	λu	$(\lambda u, v)$
Divergence	$-\nabla \cdot (\boldsymbol{\lambda} u)$	$(u\boldsymbol{\lambda}, \nabla v)$
Diffusion	$-\nabla \cdot (\lambda \nabla u)$	$(\lambda \nabla u, \nabla v)$
Convection	$\boldsymbol{\lambda} \cdot \nabla u$	$(\boldsymbol{\lambda} \cdot \nabla u, v)$

11.2.2 Discretization into finite elements

The domain Ω is now subdivided into smaller finite elements K_j, with $j = 1, ..., m$. We introduce the following function space

$$\mathcal{S}_{h,k} = \{v \in C\left(\bar{\Omega}\right) : v(K_j) \in \Pi_k(K_j) \text{ for } j = 1, ..., m, \text{ and } v = 0 \text{ on } \partial\Omega \times \mathbf{R}_+\}. \tag{11.6}$$

In the above, $\Pi_k(K_j)$ denotes the space of polynomials of degree $\leq k$ defined within the finite element K_j. In other words, $\mathcal{S}_{h,k}$ is the space of piecewise-polynomial continuous functions. See for example fig. 11.1a. for a pictorial representation of these functions. For finite k, $S_{h,k}$ is a finite dimensional space: any function in $S_{h,k}$ can be constructed as a linear combination of basis polynomial functions.



(a) Sample function for CG.

(b) Sample function for DG.

Figure 11.1

We now introduce $\mathcal{U}_h \subset \mathcal{U}$ and $\mathcal{V}_h \subset \mathcal{V}$ as functional spaces that belong to $\mathcal{S}_{h,k}$. The aim is to find $u_h \in \mathcal{U}_h$ such that

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) = L(v_h) \qquad \forall v_h \in \mathcal{V}_h. \tag{11.7}$$

Label the finite basis of $\mathcal{U}_h$ as $\{\Phi_i\}_{i=1}^n$, where n is the size of the basis (also referred to as degrees of freedom), and $\Phi_i = \Phi_i(\mathbf{x})$. For the sake of simplicity, lets use $\mathcal{V}_h = \mathcal{U}_h$ in this description.

Any $v_h \in \mathcal{V}_h$ can be written as $v_h = \sum_{i=1}^n V_i \Phi_i$, where $V_i = V_i(t)$. Equation (11.7) is then automatically satisfied if the following is satisfied

$$\left(\frac{\partial u_h}{\partial t}, \Phi_i\right) + a(u_h, \Phi_i) = L(\Phi_i) \qquad \text{for } i = 1, ..., n. \tag{11.8}$$

Since $u_h \in \mathcal{U}_h$, we can write $u_h = \sum_{j=1}^n U_j \Phi_j$, where $U_j = U_j(t)$. Thus, the above becomes

$$\sum_{j=1}^n (\Phi_j, \Phi_i) \frac{dU_j}{dt} + \sum_{j=1}^n U_j a(\Phi_j, \Phi_i) = L(\Phi_i) \qquad \text{for } i = 1, ..., n. \tag{11.9}$$

Define the vectors $\mathbf{U}$ and $\mathbf{b}$, as those whose components are U_j and $L(\Phi_i)$, respectively. Define the matrix $\mathbf{A}$ as that whose components are $a_{ij} = a(\Phi_j, \Phi_i)$. Similarily, define the mass matrix $\mathbf{M}$ as that whose components are $m_{i,j} = (\Phi_j, \Phi_i)$. Then, the equation above can be written as

$$\mathbf{M}\frac{d\mathbf{U}}{dt} + \mathbf{A}\mathbf{U} = \mathbf{b}. \tag{11.10}$$

11.3 The discontinuous Galerkin method

The dicsontinuous Galerkin (DG) method uses a space of trial functions $\mathcal{U}$ and a space of test functions $\mathcal{V}$ that consist of discontinuous functions. The discontinuities in functions lead to an alteration of the divergence theorem. In eq. (11.4) we used the divergence theorem, which is defined below

$$\int_\Omega \nabla \cdot (\boldsymbol{\lambda} u)\, dV = \int_{\partial\Omega} u\boldsymbol{\lambda} \cdot \mathbf{n}\, dS. \tag{11.11}$$

This holds true only when u is a continuous function. An alternate formulation holds for cases where u is discontinuous (note that we still assume $\boldsymbol{\lambda}$ is continuous). Label ω_j, with $j = 1, ...$ as the subdomains in Ω over which the functions are continuous. The set of all ω_j is τ. Thus, we now have

$$\int_\Omega \nabla \cdot (\boldsymbol{\lambda} u)\, dV = \sum_{w_j \in \tau} \int_{w_j} \nabla \cdot (\boldsymbol{\lambda} u)\, dV = \sum_{w_j \in \tau} \int_{\partial w_j} u\boldsymbol{\lambda} \cdot \mathbf{n}_{w_j}\, dS, \tag{11.12}$$

where $\mathbf{n}_{w_j}$ is the unit normal for the subdomain w_j. We introduce the notation

$$[\![u]\!] = u^+ - u^-, \tag{11.13}$$

where u^+ is the value of u on one side of the subdomain boundary and u^- is the value on the other side. Additionally, introduce Γ as the skeleton of τ, that is, the inner boundaries of all the subdomains w_j in the domain Ω_j. We can now write

$$\int_\Omega \nabla \cdot (\boldsymbol{\lambda} u)\, dV = \int_{\partial\Omega} u\boldsymbol{\lambda} \cdot \mathbf{n}\, dS + \int_\Gamma [\![u]\!]\boldsymbol{\lambda} \cdot \mathbf{n}_\Gamma\, dS, \tag{11.14}$$

where $\mathbf{n}_\Gamma$ is a unit normal for each point in the skeleton Γ (the orientation of $\mathbf{n}_\Gamma$ is arbitrary). Equation (11.14) is the divergence theorem for discontinuous functions u. We introduce the notation

$$\{u\} = \frac{1}{2}\left(u^+ + u^-\right). \tag{11.15}$$

From its definitions, it can be shown that

$$[\![uv]\!] = \{u\}[\![v]\!] + [\![u]\!]\{v\}. \tag{11.16}$$

Thus, the divergence theorem for two discontinuous functions u and v can be written as

$$\int_\Omega \nabla \cdot (\boldsymbol{\lambda} uv)\, d\mathbf{x} = \int_{\partial\Omega} uv\boldsymbol{\lambda} \cdot \mathbf{n}\, dS + \int_\Gamma (\{u\}[\![v]\!] + [\![u]\!]\{v\})\, \boldsymbol{\lambda} \cdot \mathbf{n}_{\mathbf{n}_\Gamma}\, dS. \tag{11.17}$$

Consider the derivation carried out to obtain eq. (11.5). If both u and v are discontinuous, then eq. (11.17) is used rather than the traditional divergence theorem and the result becomes

$$(\nabla \cdot (\boldsymbol{\lambda} u), v) = -(u\boldsymbol{\lambda}, \nabla v) + \int_\Gamma (\{u\}[\![v]\!] + [\![u]\!]\{v\})\, \boldsymbol{\lambda} \cdot \mathbf{n}_\Gamma\, dS. \tag{11.18}$$

Thus, for DG, an additional term appears for every bilinear form. For example, eq. (11.3) for the weak formulation given the specific operator $\nabla \cdot (\boldsymbol{\lambda} u)$ is now

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u, v) + \int_\Gamma (\{u\}[\![v]\!] + [\![u]\!]\{v\})\, \boldsymbol{\lambda} \cdot \mathbf{n}_\Gamma\, dS = L(v) \qquad \forall v \in \mathcal{V}. \tag{11.19}$$

Unlike the finite dimensional space shown in eq. (11.6), we now use the following for DG

$$\mathcal{S}_{h,k} = \{v \in L_2\left(\bar{\Omega}\right) : v(K_j) \in \Pi_k(K_j) \text{ for } j = 1, ..., m\}, \tag{11.20}$$

where $\Pi_k(K_j)$ can be the same space as in the continuous Galerkin case. The main difference, however, is that the functions $v \in S_{h,k}$ can now be discontinuous from element to element (they do not belong to $C(\bar{\Omega})$). See fig. 11.1b for a pictorial representation of these functions. We note that the functions are discontinuous only across elements, and not within elements. Thus, we can now make the correspondence $\omega_j \to K_j$. The set of all K_j is τ_h, and the skeleton is Γ_h. The corresponding form of eq. (11.7) is thus

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) + \int_{\Gamma_h} (\{u_h\}[\![v_h]\!] + [\![u_h]\!]\{v_h\})\, \boldsymbol{\lambda} \cdot \mathbf{n}_{\Gamma_h}\, dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h. \tag{11.21}$$

Note that we can break up the integrals above into sums of integrals over each finite element. That is,

$$\sum_{K_j \in \tau_h} \left(\frac{\partial u_h}{\partial t}, v_h\right)_{K_j} + \sum_{K_j \in \tau_h} a(u_h, v_h)_{K_j} + \sum_{K_j \in \tau_h} \int_{\partial K_j} u_h v_h \boldsymbol{\lambda} \cdot \mathbf{n}_{K_j}\, dS = \sum_{K_j \in \tau_h} L(v_h)_{K_j} \qquad \forall v \in \mathcal{V}. \tag{11.22}$$

To satisfy the above, we can require

$$\left(\frac{\partial u_h}{\partial t}, v_h\right)_{K_j} + a(u_h, v_h)_{K_j} + \int_{\partial K_j} u_h v_h \boldsymbol{\lambda} \cdot \mathbf{n}_{K_j}\, dS = L(v_h)_{K_j} \qquad \forall v_h \in \mathcal{V}_h. \tag{11.23}$$

For many schemes, the flux term $f(u_h) = \boldsymbol{\lambda} u_h$ in the surface integrals of eqs. (11.21) and (11.23) is replaced by a numerical flux $\hat{f}(u_h)$. In this case, we'll focus on the numerical flux $\hat{f}(u_h) = \boldsymbol{\lambda}\hat{u}_h$, where

$$\hat{u}_h = \{u_h\} + \delta \frac{|\boldsymbol{\lambda} \cdot \mathbf{n}_{\Gamma_h}|}{\boldsymbol{\lambda} \cdot \mathbf{n}_{\Gamma_h}} [\![u_h]\!]. \tag{11.24}$$

Since $\hat{u}_h$ is now continuous, we have $[\![\hat{u}_h]\!] = 0$ and $\{\hat{u}_h\} = \hat{u}_h$. Thus, eq. (11.21) becomes

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) + \int_{\Gamma_h} \hat{u}_h [\![v_h]\!] \boldsymbol{\lambda} \cdot \mathbf{n}_{\Gamma_h}\, dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h, \tag{11.25}$$

or, after expanding the definition of $\hat{u}_h$

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) + \int_{\Gamma_h} \boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}\{u_h\}[\![v_h]\!]\,dS + \delta\int_{\Gamma_h} |\boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}|\,[\![u_h]\!][\![v_h]\!]\,dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h. \tag{11.26}$$

Given that $a(u_h, v_h) = -(u_h\boldsymbol{\lambda}, \nabla v_h)$ for the operator $\nabla\cdot(\boldsymbol{\lambda}u)$, we can rewrite eq. (11.26) as

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) - (u_h\boldsymbol{\lambda}, \nabla v_h) + \int_{\Gamma_h} \boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}\{u_h\}[\![v_h]\!]\,dS + \delta\int_{\Gamma_h} |\boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}|\,[\![u_h]\!][\![v_h]\!]\,dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h. \tag{11.27}$$

Using eq. (11.18), we get

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + (\nabla\cdot(\boldsymbol{\lambda}u_h), v_h) - \int_{\Gamma_h} \left(\{u_h\}[\![v_h]\!] + [\![u_h]\!]\{v_h\}\right)\boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}\,dS + \int_{\Gamma_h} \boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}\{u_h\}[\![v_h]\!]\,dS + \delta\int_{\Gamma_h} |\boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}|\,[\![u_h]\!][\![v_h]\!]\,dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h, \tag{11.28}$$

which, upon simplifying gives

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + (\nabla\cdot(\boldsymbol{\lambda}u_h), v_h) - \int_{\Gamma_h} \boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}[\![u_h]\!]\{v_h\}\,dS + \delta\int_{\Gamma_h} |\boldsymbol{\lambda}\cdot\mathbf{n}_{\Gamma_h}|\,[\![u_h]\!][\![v_h]\!]\,dS = L(v_h) \qquad \forall v_h \in \mathcal{V}_h. \tag{11.29}$$

Equation (11.26) is the weak form of the equation, whereas eq. (11.29) is the strong form.

11.4 Multi-dimensional finite elements

To describe the application of finite elements to cases where the unknowns are vectors (e.g. $\mathbf{u} = [u_x, u_y, u_z]$), we'll use the following elliptic PDE

$$\sum_{\beta=x,y,z} -\frac{\partial}{\partial x_\beta}\left(\frac{\partial u_\alpha}{\partial x_\beta} + \frac{\partial u_\beta}{\partial x_\alpha}\right) = f_\alpha \qquad \text{in } \Omega^3 \text{ for } \alpha = x, y, z. \tag{11.30}$$

For the above, $\mathbf{u} = 0$ at boundaries. As for the 1D case, we introduce trial and test function spaces, but these are now three dimensional and are labeled as $\mathcal{U}^3$ and $\mathcal{V}^3$. As for the 1D case again, we'll assume for simplicity that $\mathcal{U}^3$ and $\mathcal{V}^3$ are the same function space. The objective is now to find the weak solution $\mathbf{u} \in \mathcal{U}^3$ such that

$$\int_{\Omega^3} \sum_{\alpha=x,y,z}\sum_{\beta=x,y,z} -\frac{\partial}{\partial x_\beta}\left(\frac{\partial u_\alpha}{\partial x_\beta} + \frac{\partial u_\beta}{\partial x_\alpha}\right) v_\alpha\,dV = \int_{\Omega^3}\sum_{\alpha=x,y,z} f_\alpha v_\alpha\,dV \qquad \forall \mathbf{v} \in \mathcal{V}^3. \tag{11.31}$$

We now use the product rule and divergence theorem to obtain

$$\int_{\Omega^3} \sum_{\alpha=x,y,z}\sum_{\beta=x,y,z} \left(\frac{\partial u_\alpha}{\partial x_\beta} + \frac{\partial u_\beta}{\partial x_\alpha}\right)\frac{\partial v_\alpha}{\partial x_\beta}\,dV = \int_{\Omega^3}\sum_{\alpha=x,y,z} f_\alpha v_\alpha\,dV \qquad \forall \mathbf{v} \in \mathcal{V}^3. \tag{11.32}$$

As before, we introduce the discrete function spaces $\mathcal{U}_h^3 \subset \mathcal{U}^3$ and $\mathcal{V}_h^3 \subset \mathcal{V}^3$ that belong to $\mathcal{S}_{h,k}^3$, which will be defined at a later point. The aim is now to find $\mathbf{u}_h \in \mathcal{U}_h^3$ such that

$$\int_{\Omega^3} \sum_{\alpha=x,y,z}\sum_{\beta=x,y,z} \left(\frac{\partial u_{h,\alpha}}{\partial x_\beta} + \frac{\partial u_{h,\beta}}{\partial x_\alpha}\right)\frac{\partial v_{h,\alpha}}{\partial x_\beta}\,dV = \int_{\Omega^3}\sum_{\alpha=x,y,z} f_\alpha v_{h,\alpha}\,dV \qquad \forall \mathbf{v}_h \in \mathcal{V}_h^3. \tag{11.33}$$

We expand $\mathbf{v}_h$ as follows

$$\mathbf{v}_h = \sum_i^n \mathbf{V}_i \Phi_i, \tag{11.34}$$

where $\mathbf{V}_i = [V_{i,x}, V_{i,y}, V_{i,z}]$ is now a vector but Φ_i is still a scalar. We thus have

$$\sum_{\alpha=x,y,z} \sum_i^n V_{i,\alpha} \int_{\Omega^3} \sum_{\beta=x,y,z} \left(\frac{\partial u_{h,\alpha}}{\partial x_\beta} + \frac{\partial u_{h,\beta}}{\partial x_\alpha} \right) \frac{\partial \Phi_i}{\partial x_\beta} \, dV =$$

$$\sum_{\alpha=x,y,z} \sum_i^n V_{i,\alpha} \int_{\Omega^3} f_\alpha \Phi_i \, dV \qquad \forall \mathbf{v}_h \in \mathcal{V}_h^3. \tag{11.35}$$

The above is satisfied if we require that the following holds

$$\int_{\Omega^3} \sum_{\beta=x,y,z} \left(\frac{\partial u_{h,\alpha}}{\partial x_\beta} + \frac{\partial u_{h,\beta}}{\partial x_\alpha} \right) \frac{\partial \Phi_i}{\partial x_\beta} \, dV = \int_{\Omega^3} f_\alpha \Phi_i \, dV \qquad \text{for } i = 1, ..., n \text{ and } \alpha = x, y, z. \tag{11.36}$$

The equation above would be the analogue of eq. (11.8) for 1D. That is, it is not necessary to test the PDE against every function in $\mathcal{V}_h^3$, but only against the basis functions of that space. We'll re-group the result as follows

$$\int_{\Omega^3} \sum_{\beta=x,y,z} \frac{\partial u_{h,\alpha}}{\partial x_\beta} \frac{\partial \Phi_i}{\partial x_\beta} \, dV + \int_{\Omega^3} \sum_{\beta=x,y,z} \frac{\partial u_{h,\beta}}{\partial x_\alpha} \frac{\partial \Phi_i}{\partial x_\beta} \, dV$$

$$= \int_{\Omega^3} f_\alpha \Phi_i \, dV \qquad \text{for } i = 1, ..., n \text{ and } \alpha = x, y, z. \tag{11.37}$$

The next step is to introduce

$$\mathbf{u}_h = \sum_j^n \mathbf{U}_j \Phi_j, \tag{11.38}$$

where again $\mathbf{U}_j = [U_{i,x}, U_{i,y}, U_{i,z}]$. Plugging this in the PDE we have

$$\sum_j^n U_{j,\alpha} \int_{\Omega^3} \sum_{\beta=x,y,z} \frac{\partial \Phi_j}{\partial x_\beta} \frac{\partial \Phi_i}{\partial x_\beta} \, dV + \sum_{\beta=x,y,z} \sum_j^n U_{j,\beta} \int_{\Omega^3} \frac{\partial \Phi_j}{\partial x_\alpha} \frac{\partial \Phi_i}{\partial x_\beta} \, dV$$

$$= \int_{\Omega^3} f_\alpha \Phi_i \, dV \qquad \text{for } i = 1, ..., n \text{ and } \alpha = x, y, z. \tag{11.39}$$

If we now introduce the bilinear forms

$$a_{ij} = \int_{\Omega^3} \sum_{\beta=x,y,z} \frac{\partial \Phi_j}{\partial x_\beta} \frac{\partial \Phi_i}{\partial x_\beta} \, dV, \tag{11.40}$$

$$b_{ij}^{\alpha\beta} = \int_{\Omega^3} \frac{\partial \Phi_j}{\partial x_\alpha} \frac{\partial \Phi_i}{\partial x_\beta} \, dV, \tag{11.41}$$

and

$$L_i^\alpha = \int_{\Omega^3} f_\alpha \Phi_i \, dV, \tag{11.42}$$

we can write the linear system of equations as

$$\sum_j^n U_{j,\alpha} a_{ij} + \sum_{\beta=x,y,z} \sum_j^n U_{j,\beta} b_{ij}^{\alpha\beta} = L_i^\alpha \qquad \text{for } i = 1, ..., n \text{ and } \alpha = x, y, z. \tag{11.43}$$

11.5 The Petrov Galerkin method

Chapter 12

Elliptic

Chapter 13

Parabolic

Chapter 14

Hyperbolic

Appendix A

The collocation method

Finite-element methods rely on the Galerkin method. Spectral methods can rely on the Galerkin method, but also on others such as the collocation and tau method. In this appendix we describe the collocation method.

As with the Galerkin method, we aim to solve eq. (11.1) by assuming the solution is $u_h \in \mathcal{U}_h$. This assumption allows us to write $u_h = \sum_{j=1}^{n} U_j \Phi_j$ and thus the governing PDE can be written as

$$\sum_{j=1}^{n} \frac{dU_j}{dt} \Phi_j + \sum_{j=1}^{n} U_j A \Phi_j = f. \tag{A.1}$$

The above equation is evaluated at a discrete set of n points, termed the collocation points. Thus we have

$$\sum_{j=1}^{n} \frac{dU_j}{dt} \Phi_j(x_i) + \sum_{j=1}^{n} U_j (A\Phi_j)(x_i) = f(x_i) \qquad \text{for } i = 1, ..., n. \tag{A.2}$$

The above can be written as

$$\tilde{\mathbf{M}} \frac{d\mathbf{U}}{dt} + \tilde{\mathbf{A}}\mathbf{U} = \tilde{\mathbf{b}}, \tag{A.3}$$

where $\tilde{\mathbf{M}}$ is a matrix whose components are $\tilde{m}_{ij} = \Phi_j(x_i)$, $\tilde{\mathbf{A}}$ is a matrix whose components are $a_{ij} = (A\Phi_j)(x_i)$, and $\tilde{\mathbf{b}}$ is a vector whose components are $f(x_i)$.