Numerical Partial Differential Equations: Advanced Methods and Analysis

Jeongbin Jo

Department of Physics, Yonsei University, jeongbin033@yonsei.ac.kr

March 2, 2026

Abstract

This document outlines the fundamental and advanced mathematical frameworks for solving Partial Differential Equations (PDEs) computationally. It synthesizes finite difference, finite volume, and finite element approaches, incorporating stability analysis and modern discontinuous Galerkin techniques.

Contents

---

1 Finite Difference Method (FDM)

1.1 Basic Concepts of Finite Difference Method

The core idea of the Finite Difference Method (FDM) is to approximate continuous partial differential equations (PDEs) into a discrete form that can be computationally solved [9]. Consider the continuity equation with a diffusion term:

$$\frac{\partial u}{\partial t} - \nu \nabla^2 u + u \cdot \nabla u + \nabla p = f \tag{1}$$

To convert this continuous form into a discrete form for computer calculations, let $u(x)$ be a sufficiently smooth function. At a specific node x_0 with a uniform grid spacing h, we define the following basic difference approximations:

- **Forward Difference**:
$$D_+u(x_0) = \frac{u(x_0+h) - u(x_0)}{h} \tag{2}$$
- **Backward Difference**:
$$D_-u(x_0) = \frac{u(x_0) - u(x_0-h)}{h} \tag{3}$$
- **Centered Difference**: This intuitively combines the forward and backward differences, providing a more accurate approximation.
$$D_0u(x_0) = \frac{u(x_0+h) - u(x_0-h)}{2h} = \frac{1}{2}(D_+u(x_0) + D_-u(x_0)) \tag{4}$$

As $h \to 0$, all these discrete operators mathematically converge to the exact continuous derivative $u'(x_0)$.

1.2 Truncation Error Analysis

To rigorously evaluate the mathematical accuracy of these approximations, we utilize the Taylor series expansion [9]. For the forward difference, expanding $u(x+h)$ yields:

$$u(x+h) = u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \dots \tag{5}$$

Rearranging this equation to solve for $u'(x)$, we find the local truncation error (LTE):

$$u'(x) = \frac{u(x+h) - u(x)}{h} - \frac{h}{2}u''(x) + \mathcal{O}(h^2) \tag{6}$$

This theoretically indicates that both forward and backward differences exhibit a first-order truncation error, strictly bounded by $\mathcal{O}(h)$.

However, for the centered difference, the first-order error terms mathematically cancel out when combining the expansions of $u(x+h)$ and $u(x-h)$:

$$u'(x) = \frac{D_+u(x) + D_-u(x)}{2} + \mathcal{O}(h^2) \tag{7}$$

Thus, the centered difference systematically achieves second-order accuracy, $\mathcal{O}(h^2)$, making it significantly superior for overall numerical stability.

1.3 Deriving Arbitrary Finite Difference Approximations

We can systematically derive difference approximations for arbitrarily spaced nodes using the method of undetermined coefficients. Suppose we want to approximate $u'(x)$ using a one-sided stencil with the nodes x, $x-h$, and $x-2h$:

$$Du(x) = au(x) + bu(x-h) + cu(x-2h) \tag{8}$$

Applying Taylor expansion to $u(x-h)$ and $u(x-2h)$ and substituting them back into the equation yields:

$$Du(x) = (a+b+c)u(x) - (b+2c)hu'(x) + \frac{1}{2}(b+4c)h^2u''(x) - \frac{1}{6}(b+8c)h^3u'''(x) + \dots \tag{9}$$

To eliminate the dominant error terms and perfectly isolate $u'(x)$, we construct the following linear algebraic system:

$$a + b + c = 0 \quad \text{(to eliminate } u(x)\text{)} \tag{10}$$

$$-(b + 2c)h = 1 \implies b + 2c = -\frac{1}{h} \quad \text{(to isolate } u'(x)\text{)} \tag{11}$$

$$b + 4c = 0 \quad \text{(to eliminate } u''(x)\text{)} \tag{12}$$

Solving this system yields the unique coefficients $a = \frac{3}{2h}$, $b = -\frac{2}{h}$, and $c = \frac{1}{2h}$. Substituting these back, we obtain a strictly second-order accurate asymmetric (backward) approximation formula, which is particularly highly useful for maintaining accuracy when implementing Neumann boundary conditions [9]:

$$Du(x) = \frac{1}{2h}(3u(x) - 4u(x - h) + u(x - 2h)) + \mathcal{O}(h^2) \tag{13}$$

1.4 Higher Order Derivatives and Multi-Dimensional Extension

The standard centered approximation for the second derivative $u''(x)$ is elegantly derived by applying the first derivative operator twice, $D_+D_-u(x)$:

$$D^2u(x) = \frac{1}{h^2}[u(x - h) - 2u(x) + u(x + h)] \tag{14}$$

This approximation strictly maintains second-order accuracy, $\mathcal{O}(h^2)$, and forms the foundation for solving 1D Poisson's and Heat equations.

For two-dimensional physical problems, the continuous Laplacian operator $\nabla^2 u = u_{xx} + u_{yy}$ is universally approximated using a 5-point stencil on a uniform Cartesian grid with spacing h:

$$\nabla_h^2 u(x, y) = \frac{u(x - h, y) + u(x + h, y) + u(x, y - h) + u(x, y + h) - 4u(x, y)}{h^2} \tag{15}$$

This 2D extension is indispensable for modeling practical computational physics phenomena and inherently preserves $\mathcal{O}(h^2)$ spatial accuracy across the domain [9].

2 Steady States and Boundary Value Problems (BVP)

2.1 Heat Equation and Boundary Conditions

The governing equation for heat conduction in a spatial domain Ω is defined as:

$$\partial_t u(x, t) = \nabla \cdot (\kappa(x)\nabla u(x, t)) + f(x, t) \tag{16}$$

If the material properties are isotropic and strictly homogeneous (κ is constant), the equation simplifies to $\partial_t u = \kappa\nabla^2 u + f$. To uniquely solve this partial differential equation system, necessary initial and boundary conditions must be appropriately defined [9]:

- **Initial Condition**: $u(x, 0) = u^0(x)$ defined for all $x \in \Omega$.
- **Dirichlet Boundary Condition**: The value of the solution is explicitly prescribed on the boundary $\partial\Omega$ ($u(x, t) = h(x, t)$).
- **Neumann Boundary Condition**: The normal derivative (representing flux) is prescribed on the boundary ($\frac{\partial u}{\partial n}(x, t) = g(x, t)$). If $g = 0$, it mathematically represents a perfectly insulated boundary.

2.2 FDM for Steady-State Problems (Poisson Equation)

For a steady-state thermal problem where the state variables are completely independent of time ($\partial_t = 0$), the governing equation reduces to the classical Poisson Equation: $-\kappa\nabla^2 u = f(x)$. Consider a 1D problem $-u''(x) = f(x)$ on the domain $0 < x < 1$ with Dirichlet conditions $u(0) = \alpha$ and $u(1) = \beta$. We systematically construct a uniform spatial mesh $x_j = j \cdot h$ with the grid spacing $h = \frac{1}{n+1}$ for interior nodes $j = 1, \ldots, n$.

Using the strictly second-order centered difference approximation derived in Section 1, the discrete system for the interior nodes is formulated as:

$$\frac{1}{h^2}(-U_{j-1} + 2U_j - U_{j+1}) = f(x_j) \tag{17}$$

This naturally translates into a global linear algebraic matrix equation $\mathbf{AU} = \mathbf{F}$. The resulting coefficient matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a tridiagonal, highly sparse, symmetric, and positive-definite (SPD) matrix [9]. The SPD property strictly guarantees that iterative numerical solvers (e.g., Conjugate Gradient method) will converge unconditionally and efficiently:

$$\mathbf{A} = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & -1 & 2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} \tag{18}$$

2.3 Handling Neumann Boundary Conditions

Implementing Neumann conditions, such as $u'(0) = \sigma$, requires special mathematical care to preserve the global accuracy of the underlying scheme. There are three primary numerical approaches to resolve this boundary constraint [9]:

2.3.1 First Approach (First-Order Forward Difference)

Using a simple forward difference at the left boundary x_0:

$$u'(0) \approx \frac{U_1 - U_0}{h} = \sigma \tag{19}$$

This directly adds the equation $\frac{1}{h}(-U_0 + U_1) = \sigma$ to the discrete system. While mathematically natural and easy to implement, its local truncation error (LTE) is tightly bounded by $\mathcal{O}(h)$, which tragically degrades the entire global system to first-order accuracy.

2.3.2 Second Approach (Ghost Node Method)

To robustly maintain second-order accuracy, we introduce a fictitious "ghost node" U_{-1} located outside the physical domain and apply a centered difference approximation at the boundary:

$$\frac{U_1 - U_{-1}}{2h} = \sigma \implies U_{-1} = U_1 - 2h\sigma \tag{20}$$

Substituting this relation into the standard D^2U_0 difference equation elegantly yields a modified boundary row for the global matrix:

$$\frac{1}{h}(-U_0 + U_1) = \sigma + \frac{h}{2}f(x_0) \tag{21}$$

This ghost node approach successfully preserves the strict $\mathcal{O}(h^2)$ global accuracy of the numerical scheme while preserving matrix sparsity.

2.3.3 Third Approach (One-Sided Second-Order Difference)

Alternatively, utilizing the 3-point asymmetric formula derived earlier, we can enforce:

$$\frac{1}{2h}(-3U_0 + 4U_1 - U_2) = \sigma \tag{22}$$

This methodology also rigorously achieves $\mathcal{O}(h^2)$ accuracy at the boundary. However, it fundamentally destroys the strictly symmetric tridiagonal structure of the coefficient matrix $\mathbf{A}$, complicating the selection of efficient linear solvers.

2.4 Uniqueness and Matrix Singularity

If a BVP is defined with purely Neumann boundary conditions on all enclosing boundaries (e.g., a completely insulated thermodynamic system where $\frac{\partial u}{\partial n} = 0$ everywhere), the assembled system matrix $\mathbf{A}$ inevitably becomes singular ($\det \mathbf{A} = 0$). Algebraically, the constant vector $\mathbf{v} = [1, 1, \ldots, 1]^T$ resides perfectly within the null space of $\mathbf{A}$ (i.e., $\mathbf{Av} = \mathbf{0}$). A singular matrix explicitly implies no unique inverse ($\mathbf{A}^{-1}$ does not exist). Physically and mathematically, if $\hat{u}$ is a valid solution to the system, any arbitrary linear combination $\hat{u} + c$ (where c is a constant) is mathematically also a valid solution. Thus, the continuous problem is ill-posed and intrinsically lacks uniqueness unless a specific anchoring condition (e.g., explicitly fixing the solution value at a single interior node) is strictly provided to eliminate the null space.

3 Initial Value Problems for Ordinary Differential Equations (ODEs)

3.1 Linear ODEs and Duhamel's Principle

Consider a fundamental system of linear ODEs:

$$u'(t) = A(t)u(t) + g(t) \tag{23}$$

where $A(t) \in \mathbb{R}^{s \times s}$ is the coefficient matrix and $g(t) \in \mathbb{R}^s$ is the external forcing term. If the coefficient matrix A is constant, the exact analytical solution can be expressed using the matrix exponential and Duhamel's principle [5]:

$$u(t) = e^{A(t-t_0)}u(t_0) + \int_{t_0}^{t} e^{A(t-s)}g(s)ds \tag{24}$$

The first exponential term represents the homogeneous solution (the transient response strictly based on initial conditions), while the integral term encapsulates the non-homogeneous solution driven entirely by the external forcing $g(t)$.

3.2 Lipschitz Continuity and Well-posedness

For highly nonlinear initial value problems of the general form $u'(t) = f(u, t)$, we typically lack explicit analytical formulas. To mathematically guarantee stability, the function f is required to

be **Lipschitz continuous** with respect to u over a specified domain $D = \{(u,t) : |u - u(t_0)| \leq a, |t - t_0| \leq b\}$ [5]. This condition is strictly satisfied if there exists a positive constant L such that:

$$|f(u,t) - f(u^*,t)| \leq L|u - u^*| \quad \text{for all } (u,t),(u^*,t) \in D \tag{25}$$

If f is continuously differentiable with respect to u, the Lipschitz constant L can be rigorously bounded by the maximum of its partial derivative norm over the domain: $L = \max_{(u,t)\in D} \left|\frac{\partial f}{\partial u}\right|$.

3.2.1 Picard-Lindelöf Theorem (Existence and Uniqueness)

The rigorous mathematical foundation of ODE solutions relies heavily on the globally recognized **Picard-Lindelöf Theorem** [5]:

- **Theorem of Existence**: If f is continuous and strictly bounded in D (i.e., $|f(u,t)| \leq M$), the ODE is guaranteed to possess at least one valid solution up to a specific time constraint $T = \min(t_0 + b, t_0 + \frac{a}{M})$.
- **Theorem of Uniqueness**: If f is not only continuous but also satisfies the Lipschitz continuity condition, the solution is mathematically proven to be strictly unique within that interval.

3.3 Numerical Integration Methods for ODEs

3.3.1 Runge-Kutta Methods

To achieve high-order accuracy in time-stepping without requiring higher derivatives of f, the classical explicit fourth-order, four-stage Runge-Kutta (RK4) method is widely adopted [9]. It strategically evaluates the right-hand function at multiple intermediate stages to achieve an impressive global accuracy of $\mathcal{O}(\Delta t^4)$:

$$Y_1 = U^n \tag{26}$$

$$Y_2 = U^n + \frac{\Delta t}{2} f(Y_1, t_n) \tag{27}$$

$$Y_3 = U^n + \frac{\Delta t}{2} f(Y_2, t_n + \frac{\Delta t}{2}) \tag{28}$$

$$Y_4 = U^n + \Delta t f(Y_3, t_n + \frac{\Delta t}{2}) \tag{29}$$

$$U^{n+1} = U^n + \frac{\Delta t}{6}\left[f(Y_1,t_n) + 2f(Y_2, t_n + \frac{\Delta t}{2}) + 2f(Y_3, t_n + \frac{\Delta t}{2}) + f(Y_4, t_n + \Delta t)\right] \tag{30}$$

3.3.2 Linear Multistep Methods (LMM)

Unlike single-step Runge-Kutta methods, linear multistep methods reuse prior function evaluations from previous time steps, significantly improving computational efficiency. The general architectural form is [5]:

$$\sum_{j=0}^{r} \alpha_j U^{n+j} = \Delta t \sum_{j=0}^{r} \beta_j f(U^{n+j}, t_{n+j}) \tag{31}$$

According to the **Dahlquist Equivalence Theorem**, for an LMM to be theoretically convergent, it must be both consistent and zero-stable [9]. If the leading coefficient $\beta_r = 0$, the method is

explicitly defined (**Explicit Method**, e.g., Adams-Bashforth). If $\beta_r \neq 0$, the method is implicitly defined (**Implicit Method**, e.g., Adams-Moulton). Implicit methods mathematically require solving a nonlinear algebraic equation at each step but offer vastly superior numerical stability properties (such as A-stability) [5].

Examples of Explicit Methods:

- **Forward Euler (1-step)**: $U^{n+1} = U^n + \Delta t f(U^n, t_n)$ yields an accuracy of $\mathcal{O}(\Delta t)$.
- **Adams-Bashforth (2-step)**: $U^{n+2} = U^{n+1} + \frac{\Delta t}{2}[-f(U^n) + 3f(U^{n+1})]$ yields an accuracy of $\mathcal{O}(\Delta t^2)$.

Examples of Implicit Methods:

- **Backward Euler (1-step)**: $U^{n+1} = U^n + \Delta t f(U^{n+1}, t_{n+1})$ yields an unconditionally stable scheme of $\mathcal{O}(\Delta t)$.
- **Trapezoidal Rule (1-step)**: $U^{n+1} = U^n + \frac{\Delta t}{2}[f(U^n) + f(U^{n+1})]$ yields an accuracy of $\mathcal{O}(\Delta t^2)$.
- **Adams-Moulton (2-step)**: $U^{n+2} = U^{n+1} + \frac{\Delta t}{12}[-f(U^n) + 8f(U^{n+1}) + 5f(U^{n+2})]$ yields an accuracy of $\mathcal{O}(\Delta t^3)$.

4 Diffusion Equations and Parabolic Problems

4.1 The Heat Equation and Discretization

The one-dimensional heat equation, a classic example of a parabolic partial differential equation, is defined as:

$$u_t = \kappa u_{xx} \quad \text{for } x \in (0,1),\ t > 0 \tag{32}$$

where $\kappa > 0$ is the constant diffusion coefficient. We establish a discrete computational grid with spatial step $h = \Delta x$ and time step $k = \Delta t$. The numerical approximation of the exact continuous solution $u(x_j, t_n)$ is denoted as U_j^n [9].

4.2 Crank-Nicolson Method

While the fully explicit Forward Euler method is strictly bounded by a severe stability restriction ($k \leq h^2/2\kappa$), the Crank-Nicolson method elegantly overcomes this limitation by averaging the explicit (Forward Euler) and implicit (Backward Euler) spatial approximations [9].

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{\kappa}{2}\left(D_x^2 U_j^n + D_x^2 U_j^{n+1}\right) \tag{33}$$

where the standard centered difference operator is $D_x^2 U_j = \frac{1}{h^2}(U_{j-1} - 2U_j + U_{j+1})$. By defining the parabolic Courant number (or diffusion number) as $r = \frac{\kappa k}{2h^2}$, the scheme can be algebraically rearranged into a tightly coupled tridiagonal linear system:

$$-rU_{j-1}^{n+1} + (1+2r)U_j^{n+1} - rU_{j+1}^{n+1} = rU_{j-1}^n + (1-2r)U_j^n + rU_{j+1}^n \tag{34}$$

4.2.1 Truncation Error and Accuracy

To rigorously derive the Local Truncation Error (LTE), we substitute the exact smooth solution $u(x,t)$ into the difference scheme and expand it using a multivariate Taylor series centered precisely around the midpoint $(x_j, t_{n+1/2})$ [9]:

$$\tau_j^n = \left[u_t + \frac{k^2}{24}u_{ttt} + \mathcal{O}(k^4)\right] - \kappa\left[u_{xx} + \frac{h^2}{12}u_{xxxx} + \mathcal{O}(h^4)\right] \tag{35}$$

Since the physical exact solution is strictly governed by $u_t = \kappa u_{xx}$, the leading error terms mathematically demonstrate that the Crank-Nicolson method yields a highly accurate, symmetric scheme of $\mathcal{O}(k^2 + h^2)$.

4.3 Von Neumann Stability Analysis

To rigorously assess numerical stability for linear PDEs, we utilize Von Neumann analysis by substituting a single Fourier mode $U_j^n = g^n e^{i\xi jh}$ into the discrete equation, where $g(\xi)$ represents the amplification factor [3, 9]. For the Crank-Nicolson scheme, this algebraic substitution yields:

$$g(\xi) = \frac{1 - 2r(1 - \cos(\xi h))}{1 + 2r(1 - \cos(\xi h))} = \frac{1 - 4r\sin^2(\xi h/2)}{1 + 4r\sin^2(\xi h/2)} \tag{36}$$

Because $r > 0$ and $\sin^2(\xi h/2) \geq 0$, it strictly holds that $|g(\xi)| \leq 1$ for all mathematically possible values of spatial and temporal step sizes. Therefore, the Crank-Nicolson method is mathematically proven to be **unconditionally stable**.

However, from an advanced computational perspective, it is worth noting that as $r \to \infty$, the amplification factor $g(\xi) \to -1$. This implies that taking excessively large time steps, while mathematically stable, can introduce non-physical, slowly decaying oscillatory behavior in the transient solution due to a lack of strict L-stability [5].

5 Advection Equations and Hyperbolic Systems

5.1 The Linear Advection Equation

The fundamental one-dimensional linear advection equation, representing the transport of a scalar quantity u at a constant velocity a, is given by:

$$u_t + au_x = 0 \tag{37}$$

Given an initial condition $u(x,0) = \eta(x)$, the analytical exact solution is simply a rigid translation of the initial profile: $u(x,t) = \eta(x - at)$. Despite its simplicity, this equation serves as the primary test case for analyzing numerical stability and phase errors in hyperbolic systems [8].

5.2 Classical Finite Difference Schemes

Let the advective Courant (CFL) number be defined as $\nu = \frac{ak}{h}$, where $k = \Delta t$ and $h = \Delta x$. For numerical stability, the CFL condition requires $|\nu| \leq 1$, ensuring the mathematical domain of dependence is contained within the numerical stencil [3].

5.2.1 Upwind Method

Assuming $a > 0$, the physical information propagates strictly from left to right. The first-order **Upwind Method** respects this characteristic direction:

$$U_j^{n+1} = U_j^n - \nu(U_j^n - U_{j-1}^n) \tag{38}$$

This method is first-order accurate, denoted as $\mathcal{O}(k+h)$. Through *modified equation analysis*, it can be shown that the upwind scheme actually solves $u_t + au_x = \frac{ah}{2}(1-\nu)u_{xx}$. The leading-order error term behaves like a diffusion term, which artificially smears out sharp gradients—a phenomenon known as **numerical diffusion** [9].

5.2.2 Lax-Wendroff Method

To achieve second-order accuracy, we employ a Taylor series expansion in time: $u(x, t+k) \approx u + ku_t + \frac{1}{2}k^2u_{tt}$. By substituting the temporal derivatives with spatial derivatives using the original PDE ($u_t = -au_x$ and $u_{tt} = a^2u_{xx}$), we obtain the **Lax-Wendroff Method**:

$$U_j^{n+1} = U_j^n - \frac{\nu}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{\nu^2}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \tag{39}$$

While this scheme achieves $\mathcal{O}(k^2+h^2)$ accuracy, its modified equation reveals a third-order derivative error term proportional to u_{xxx}. This leads to **phase errors** where different Fourier modes travel at different speeds, resulting in spurious, non-physical oscillations (Gibbs-like phenomena) near discontinuities [8].

5.2.3 Godunov's Theorem and Monotonicity

The contrast between the diffusive Upwind method and the oscillatory Lax-Wendroff method illustrates **Godunov's Theorem**: no linear numerical scheme of second-order or higher can be monotonicity-preserving [8]. This fundamental constraint necessitates the development of nonlinear high-resolution schemes such as TVD limiters.

6 Finite Volume Method (FVM) and Conservation Laws

6.1 Integral Formulation and Numerical Fluxes

Unlike FDM which approximates pointwise derivatives, the Finite Volume Method (FVM) tracks the evolution of integral averages across specific control volumes (cells) $C_j = [x_{j-1/2}, x_{j+1/2}]$ [8]. Let Q_j^n represent the cell average of the physical quantity:

$$Q_j^n \approx \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} q(x, t_n)\, dx \tag{40}$$

For a general conservation law $q_t + f(q)_x = 0$, integrating over the cell C_j and the time step $[t_n, t_{n+1}]$ yields the exact conservative update formula:

$$Q_j^{n+1} = Q_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+1/2}^n - F_{j-1/2}^n\right) \tag{41}$$

Here, $F_{j\pm1/2}^n$ denotes the **Numerical Flux**, which approximates the time-averaged physical flux acting across the cell interfaces. In this framework, the interface values are typically determined by solving a local **Riemann Problem** at each cell boundary [8].

6.2 High-Resolution Schemes and Flux Limiters

Godunov's theorem mathematically states that no linear scheme can be both second-order accurate and perfectly monotone (free of spurious oscillations). To bypass this fundamental limitation, we employ non-linear **High-Resolution Methods** utilizing flux limiters [6].

The total numerical flux is formulated as a precise non-linear blend of a low-order, strictly monotone flux (F_L, e.g., Upwind) and a high-order, oscillatory flux (F_H, e.g., Lax-Wendroff):

$$F^n_{j-1/2} = F_L + \Phi(\theta_{j-1/2})\,(F_H - F_L) \tag{42}$$

The parameter $\theta_{j-1/2}$ measures the local smoothness of the data by taking the ratio of consecutive spatial gradients:

$$\theta_{j-1/2} = \frac{Q^n_j - Q^n_{j-1}}{Q^n_{j+1} - Q^n_j} \quad \text{(assuming velocity } a > 0) \tag{43}$$

6.2.1 Total Variation Diminishing (TVD) Limiters

To strictly ensure that numerical oscillations do not develop, the scheme must satisfy the **Total Variation Diminishing (TVD)** property, where $TV(Q^{n+1}) \leq TV(Q^n)$ [6]. For this to hold, the limiter function $\Phi(\theta)$ must reside within the **Sweby Region** [10].

Standard, highly effective limiter functions include:

- **Minmod Limiter**: The most conservative limiter, selecting the smallest available slope to strictly prevent any overshoots.

$$\Phi(\theta) = \max(0, \min(1, \theta)) \tag{44}$$

- **Superbee Limiter**: An aggressive limiter designed to maintain maximum sharpness at shock fronts.

$$\Phi(\theta) = \max(0, \min(1, 2\theta), \min(2, \theta)) \tag{45}$$

- **Monotonized Central (MC) Limiter**: A robust compromise between accuracy and stability, widely used in practical CFD [10].

$$\Phi(\theta) = \max\left(0, \min\left(\frac{1+\theta}{2}, 2, 2\theta\right)\right) \tag{46}$$

- **Van Leer Limiter**: A smoothly differentiable limiter providing excellent convergence properties for steady-state problems.

$$\Phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|} \tag{47}$$

7 Mathematical Foundations of the Finite Element Method

7.1 Weak Derivatives and Sobolev Spaces

Classical (strong) solutions to differential equations require a high degree of smoothness, such as $u \in C^2(\Omega)$ for a second-order PDE. However, this requirement is often too restrictive for physical problems involving shocks, material interfaces, or singular forcing terms. To provide a robust mathematical framework for such cases, we utilize the concept of the **Weak Derivative** [4].

Let $u, v \in L^1_{loc}(\Omega)$ be locally integrable functions. The function v is defined as the α-th weak partial derivative of u, denoted by $D^\alpha u = v$, if it satisfies the following generalized integration by parts formula for all test functions $\phi \in C_c^\infty(\Omega)$ (infinitely differentiable functions with compact support in Ω):

$$\int_\Omega u D^\alpha \phi \, dx = (-1)^{|\alpha|} \int_\Omega v \phi \, dx \tag{48}$$

This formulation effectively shifts the requirement of differentiability from the unknown function u to the smooth, infinitely differentiable test function ϕ.

7.2 Sobolev Spaces and Hilbert Framework

Building upon the definition of weak derivatives, we define the **Sobolev Space** $W^{k,p}(\Omega)$. This space consists of all functions in $L^p(\Omega)$ whose weak derivatives up to order k also reside in $L^p(\Omega)$ [4].

In the specific and highly important case where $p = 2$, the Sobolev space becomes a complete inner-product space (a **Hilbert Space**), denoted as $H^k(\Omega) = W^{k,2}(\Omega)$. For most second-order elliptic problems, the space $H^1(\Omega)$ is central. Its norm is rigorously defined as:

$$\|u\|_{H^1(\Omega)} = \left(\int_\Omega (|u|^2 + |\nabla u|^2) \, dx \right)^{\frac{1}{2}} \tag{49}$$

Furthermore, we define $H^1_0(\Omega)$ as the subspace of $H^1(\Omega)$ containing functions that vanish on the boundary $\partial\Omega$ in the sense of the **Trace Operator**. This space is essential for enforcing Dirichlet boundary conditions within the variational framework [1].

7.3 The Role of Hilbert Spaces in FEM

The Hilbert space structure of $H^k(\Omega)$ is fundamental to the Finite Element Method. It allows the use of inner products to define orthogonality and projections, which are the mathematical pillars of the **Galerkin method**. By working within these spaces, we can guarantee that the numerical approximation is the "best" possible solution within a finite-dimensional subspace, measured in terms of the energy norm [1].

8 The Finite Element Method (FEM)

8.1 Variational (Weak) Formulation

Consider a general second-order elliptic boundary value problem (BVP):

$$-\nabla \cdot (A \nabla u) + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega \tag{50}$$

To construct the weak formulation, we multiply the PDE by an arbitrary test function $v \in H^1_0(\Omega)$ and integrate over the domain Ω. Applying the divergence theorem (Green's First Identity) transfers one derivative from the unknown solution u to the smooth test function v, naturally incorporating the Dirichlet boundary condition into the function space [4]:

$$\int_\Omega (A \nabla u \cdot \nabla v + b \cdot \nabla u \, v + cuv) \, dx = \int_\Omega f v \, dx \tag{51}$$

This is expressed as the standard Variational Problem: Find $u \in H_0^1(\Omega)$ such that:

$$a(u, v) = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega) \tag{52}$$

where $a(u, v)$ is a continuous bilinear form, and $\langle f, v \rangle$ is a bounded linear functional in the dual space $H^{-1}(\Omega)$.

8.2 Lax-Milgram Lemma and Energy Minimization

The existence and uniqueness of the weak solution are strictly guaranteed by the **Lax-Milgram Lemma** [1]. The bilinear form $a(u, v)$ must satisfy two critical conditions:

- **Continuity (Boundedness)**: $|a(u, v)| \leq \alpha \|u\|_{H^1} \|v\|_{H^1}$ for some $\alpha < \infty$.
- **Coercivity (Ellipticity)**: $a(u, u) \geq \beta \|u\|_{H^1}^2$ for some strictly positive constant $\beta > 0$.

Furthermore, if the operator A is symmetric (leading to $a(u, v) = a(v, u)$), solving the variational problem is equivalent to finding the unique minimizer of the **Energy Functional**:

$$J(v) = \frac{1}{2} a(v, v) - \langle f, v \rangle \tag{53}$$

This minimization principle links the PDE solution to the physical state of minimum potential energy [1].

8.3 Galerkin Approximation and Basis Functions

In FEM, we project the infinite-dimensional Hilbert space problem onto a finite-dimensional subspace $V_h \subset H_0^1(\Omega)$. According to **Cea's Lemma**, the Galerkin approximation u_h is the best possible approximation of u in the V_h subspace with respect to the energy norm:

$$\|u - u_h\|_{H^1} \leq \frac{\alpha}{\beta} \inf_{v_h \in V_h} \|u - v_h\|_{H^1} \tag{54}$$

The domain Ω is partitioned into discrete elements (e.g., triangles or quadrilaterals). Any approximate solution $u_h \in V_h$ is expressed as a linear combination of localized, piecewise polynomial basis functions $\phi_j(x)$:

$$u_h(x) = \sum_{j=1}^{N} U_j \phi_j(x) \tag{55}$$

Substituting this into the weak formulation yields a sparse linear algebraic system $\mathbf{KU} = \mathbf{F}$, where $\mathbf{K}$ is the **Stiffness Matrix** ($K_{ij} = a(\phi_j, \phi_i)$) and $\mathbf{F}$ is the **Load Vector** ($F_i = \langle f, \phi_i \rangle$). Because the basis functions have compact support, $\mathbf{K}$ is a sparse, Symmetric Positive Definite (SPD) matrix, allowing for efficient computational inversion [1].

9 Advanced Methods: Discontinuous Galerkin (DG-FEM)

9.1 Motivation and Conceptual Framework

The classical Finite Element Method demands global continuity across element interfaces (C^0 continuity), which heavily restricts its capability to efficiently solve purely hyperbolic conservation laws ($q_t + \nabla \cdot f(q) = 0$) where shocks or sharp discontinuities frequently arise.

The **Discontinuous Galerkin (DG) Method** brilliantly synthesizes the foundational strengths of the Finite Volume Method (FVM) and FEM [7]. In DG-FEM, the solution is approximated using high-order polynomials within each local element D^k, but the functions are completely permitted to be discontinuous across the element boundaries. This localized approach allows for superior handling of transport-dominated problems.

9.2 DG-FEM Formulation and Numerical Fluxes

By defining a local test function $\phi_i^k(x)$ exclusively supported on element $D^k = [x^k, x^{k+1}]$, the weak formulation is localized:

$$\int_{D^k} \left(\frac{\partial u_h^k}{\partial t} \phi_i^k - f_h^k \frac{\partial \phi_i^k}{\partial x} - g\phi_i^k \right) dx = - \left[f_h^k \phi_i^k \right]_{x^k}^{x^{k+1}} \tag{56}$$

Because the piecewise solution is discontinuous at the interface nodes x^k, the boundary term on the right-hand side is mathematically ambiguous. To resolve this, DG-FEM borrows the core concept of the **Numerical Flux** $F_{j\pm1/2}$ from FVM [2] to couple adjacent elements dynamically:

$$\left[f_h^k \phi_i^k \right]_{x^k}^{x^{k+1}} = F(u_h(x^{k+1,-}), u_h(x^{k+1,+}))\phi_i^k(x^{k+1}) - F(u_h(x^{k,-}), u_h(x^{k,+}))\phi_i^k(x^k) \tag{57}$$

This framework yields a block-diagonal mass matrix that can be trivially inverted locally, making DG-FEM explicitly solvable without large global matrix inversions.

9.3 Comparison of Numerical Methods

The choice of numerical framework profoundly depends on the governing physics and geometric complexity:

- **FDM (Finite Difference)**: Computationally intuitive and fast, but severely ill-suited for complex arbitrary geometries. It strictly relies on pointwise Taylor expansions.
- **FVM (Finite Volume)**: Specifically engineered for fluid dynamics and shock waves due to its exact adherence to integral conservation laws. Extending FVM to high-order spatial accuracy on unstructured meshes is mathematically cumbersome.
- **FEM (Finite Element)**: Extremely powerful for elliptic/parabolic PDEs (structural mechanics, heat transfer) and handles complex geometries effortlessly via unstructured meshing. However, it struggles natively with strong hyperbolic advection.
- **DG-FEM (Discontinuous Galerkin)**: Achieves arbitrarily high-order accuracy (hp-adaptivity), perfectly satisfies local conservation laws, handles complex geometries, and supports highly parallelizable explicit time-stepping. The primary drawback is a significant increase in the total degrees of freedom (computational cost).

References

[1] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York, NY, 3rd edition, 2008.

[2] Bernardo Cockburn, San-Yih Lin, and Chi-Wang Shu. Tvb runge-kutta local projection discontinuous galerkin finite element method for conservation laws. iii. one-dimensional systems. *Journal of Computational Physics*, 84(1):90–113, 1989. The critical paper that successfully coupled RK time-stepping with the spatial DG-FEM method.

[3] Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische Annalen*, 100(1):32–74, 1928. The original paper establishing the fundamental CFL condition for numerical stability.

[4] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society (AMS), Providence, RI, 2nd edition, 2010. Mathematical foundation for Weak Derivatives, Sobolev Spaces, and the Lax-Milgram Lemma.

[5] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin, Heidelberg, 2nd edition, 1993.

[6] Amiram Harten. High resolution schemes for hyperbolic conservation laws. *Journal of Computational Physics*, 49(3):357–393, 1983. A landmark paper that introduced the mathematical framework for High-Resolution TVD schemes.

[7] Jan S. Hesthaven and Tim Warburton. *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications*. Springer, New York, NY, 2007.

[8] Randall J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, UK, 2002.

[9] Randall J. LeVeque. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[10] Peter K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM journal on numerical analysis*, 21(5):995–1011, 1984. The foundational paper defining the TVD region and standard flux limiters like Minmod and Superbee.