Scientific Computing

Art to Science

Shaoqiang Tang

Email: maotang@pku.edu.cn

College of Engineering, Peking University

Version 2012

Contents

Chapter 0

Introduction

0.1 References

1. J. W. Thomas. *Numerical Partial Differential Equation: Finite Difference Methods.* Springer, 1995.
2. John C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations.* Wadsworth, 1989.
3. R. Leveque. *Finite Volume Methods for Hyperbolic Problems.* Cambridge University Press, 2002.
4. D. Braess. *Finite Elements,* 2nd ed. Cambridge University Press, 1997.

0.2 Scientific Computing

- Hardware computer science and engineering
- Software supercomputing
- Physical science
- Engineering science
- Mathematics

0.2.1 Scheme/Algorithm b/t Art and Science

- Design
- Analysis: from art to science
- Experimentation

0.2.2 Goals

- Understanding scheme
- Analyzing scheme
- Appreciating scheme

This course will not be a comprehensive exposure of all aspects. Instead, we aim at providing a flavor of analysis and an overview of general issues. A standing point for your scientific career with computation.

0.3 Prerequisites

- Computer Language: Matlab, Fortran, C
- Mathematics: PDE, application and analysis, or equivalent

0.4 Syllabus

- Finite difference method: parabolic equations (4weeks)
- Finite volume method: hyperbolic equations (4–5weeks)
- Finite element method: parabolic equations (4weeks)
- Special topics: spectral method, multiscale method (2–3weeks)

Difficulties: no general quantitative PDE theory, therefore schemes are problem-dependent.

0.5 Grading

- Attendance
- Assignment: submitted every two weeks including analysis and program
- Final

Office hours: by appointment.

Chapter 1

Finite Difference Method for Parabolic Equations

1.1 Parabolic Equations—An Overview

The understanding of the underlying equations is crucial for the design and analysis for schemes. Here we briefly discuss the basic features of parabolic equations.

1.1.1 Heat Equation—A Linear Example

Let a function $u(t, x)$ be the temperature of an object. We have the heat equation

$$u_t = bu_{xx}. \tag{1.1}$$

Different boundary conditions may be imposed.

- *Cauchy problem (IVP: Initial-Value-Problem)* Consider (1.1) in $(x, t) \in \mathbb{R} \times \mathbb{R}^+$.

  $$u(0, x) = u_0(x) \tag{1.2}$$

- *IBVP (Initial-Boundary-Value-Problem)* Consider (1.1) in $(x, t) \in [a, b] \times \mathbb{R}^+$. Initial condition is

  $$u(0, x) = u_0(x). \tag{1.3}$$

  Boundary conditions are

  $$u(t, a) = u_a(t),\ u(t, b) = u_b(t) \text{ (Dirichlet boundary condition)} \tag{1.4}$$

  or

  $$u_x(t, a) = 0,\ u_x(t, b) = 0 \text{ (Neumann boundary condition)} \tag{1.5}$$

Fourier transform is an indispensable tool for studying linear partial differential equations. Regarding t as a parameter, we define

$$\mathcal{F}(u(t,x)) = \hat{u}(t,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u(t,x)e^{-i\omega x}\mathrm{d}x. \tag{1.6}$$

The function in physical space may be obtained by an inverse Fourier transform.

$$u(t,x) = \mathcal{F}^{-1}(\hat{u}(t,\omega)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \hat{u}(t,\omega)e^{i\omega x}\mathrm{d}\omega. \tag{1.7}$$

An important property of the Fourier transform is the Parseval's relation.

$$\int_{-\infty}^{+\infty} |u(t,x)|^2\mathrm{d}x = \int_{-\infty}^{+\infty} |\hat{u}(t,\omega)|^2\mathrm{d}\omega. \tag{1.8}$$

Therefore, the Fourier transform $\mathcal{F} : L^2 \longrightarrow L^2$ defines an isometry

$$\| u(t,x)^2 \|_2 = \| \hat{u}(t,\omega)^2 \|_2 . \tag{1.9}$$

The Fourier transform is a linear operator. It commutes with the time differentiation operator.

$$\mathcal{F}(u_t) = \frac{\partial}{\partial t}\hat{u}. \tag{1.10}$$

For the spatial differentiation, we compute

$$\begin{aligned}
\mathcal{F}(u_x) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u_x e^{-i\omega x}\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega x}\mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}}(ue^{-i\omega x} |_{-\infty}^{+\infty} + i\omega \int_{-\infty}^{+\infty} ue^{-i\omega x}\mathrm{d}x) \\
&= \frac{i\omega}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} ue^{-i\omega x}\mathrm{d}x \\
&= i\omega\hat{u}.
\end{aligned} \tag{1.11}$$

We should note that in equations (1.10) and (1.11), the interchange the order of differentiation and integral requires $u(t,x)$ to be “good” enough.

In the wave-number (spectral) space, the heat equation is transformed into

$$\left\{ \begin{array}{l} \dfrac{\partial}{\partial t}\hat{u} = -b\omega^2\hat{u}, \\ \hat{u}(0,\omega) = \hat{u}_0(\omega). \end{array} \right. \tag{1.12}$$

At each wave-number k, we have the solution $\hat{u}(t,\omega) = \hat{u}_0 e^{-b\omega^2 t}$, while the dispersion relation gives $\lambda = -b\omega^2$. Therefore, we compute

$$\begin{aligned} u(t,x) &= \mathcal{F}^{-1}(\hat{u}(t,\omega)) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{i\omega x} e^{-b\omega^2 t} [\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-i\omega y} u_0(y) \mathrm{d}y] \mathrm{d}\omega \\ &= \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{+\infty} (\frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{i\omega(x-y)} e^{-b\omega^2 t} \mathrm{d}\omega) u_0(y) \mathrm{d}y \\ &= \frac{1}{2\sqrt{\pi b t}} \int_{-\infty}^{+\infty} e^{-\frac{(x-y)^2}{4bt}} u_0(y) \mathrm{d}y \end{aligned} \tag{1.13}$$

In this expression, the term $e^{-\frac{(x-y)^2}{4bt}}$ in (1.13) equation is the heat kernel. This expression implies a “smoothing” effect of the heat diffusion, as well as the regularity of the solution. Actually, because the term $e^{-b\omega^2}$ dominates as $\omega \to +\infty$, we find that

$$\frac{\partial^{l+m} u(t,x)}{\partial t^l \partial x^m} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{i\omega x} (i\omega)^m (-b\omega^2)^l e^{-b\omega^2} \hat{u}_0(\omega) \mathrm{d}\omega. \tag{1.14}$$

An extension of the heat equation is an advection-diffusion equation. It reads

$$u_t + a u_x = b u_{xx}. \tag{1.15}$$

We make a change of variable.

$$w(t,y) = u(t, y + at). \tag{1.16}$$

Then we get

$$w_t = u_t + a u_x, \tag{1.17}$$

$$w_y = u_x, \tag{1.18}$$

$$w_{yy} = u_{xx}. \tag{1.19}$$

The convection-diffusion equation (1.15) is transformed into the heat equation.

$$w_t = b w_{yy}. \tag{1.20}$$

We observe that (1.15) includes two mechanisms, namely, an advection at a constant speed a (to the left when $a > 0$), and a “smoothing ”mechanism due to diffusion.

1.1.2 Nonlinear Parabolic Equation

The Burgers’ equation is an example for nonlinear parabolic equation.

$$u_t + u u_x = b u_{xx}. \tag{1.21}$$

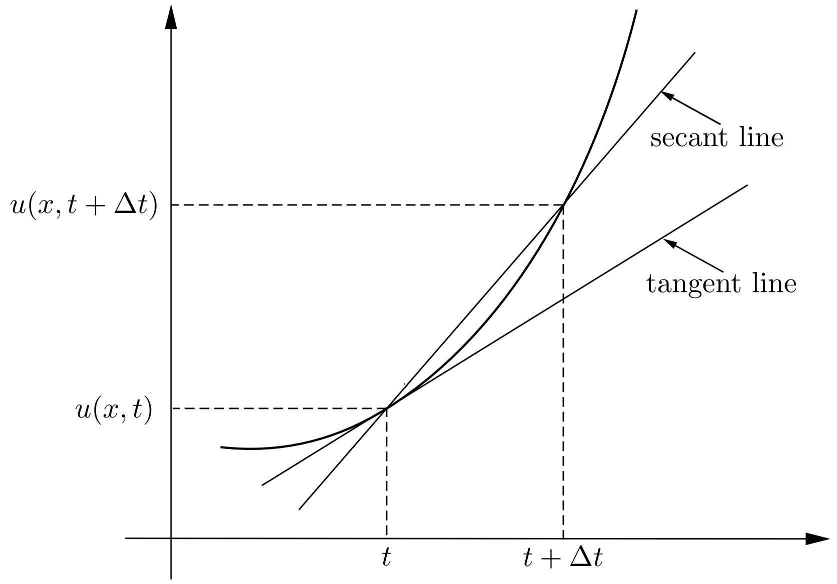


Figure 1.1: Derivative and difference

Using the Cole-Hopf transform

$$u = -2b\frac{\varphi_x}{\varphi}, \tag{1.22}$$

we obtain a heat equation

$$\varphi_t = b\varphi_{xx}. \tag{1.23}$$

1.2 Introduction to Finite Difference Method

Heuristically, the finite difference method comes naturally from the definition of derivatives as a limit of the quotient of differences, as shown in Figure 1.1

$$\frac{\partial u}{\partial t}(t,x) = \lim_{\Delta t \to 0} \frac{u(t+\Delta t,x) - u(t,x)}{\Delta t}. \tag{1.24}$$

$$\frac{\partial u}{\partial x}(t,x) = \lim_{\Delta x \to 0} \frac{u(t,x+\Delta x) - u(t,x)}{\Delta x}. \tag{1.25}$$

A finite difference method includes the following basic ingredients.

- Grid: domain discretization (see Figure 1.2).

$$\begin{cases} \text{Time:} & 0 = t_0 < t^1 < t^2 < \cdots < t^n < \cdots < t^N = T; \\ \text{Space:} & a = x_0 < x_1 < x_2 < \cdots < x_m < \cdots < x_N = b. \end{cases} \tag{1.26}$$

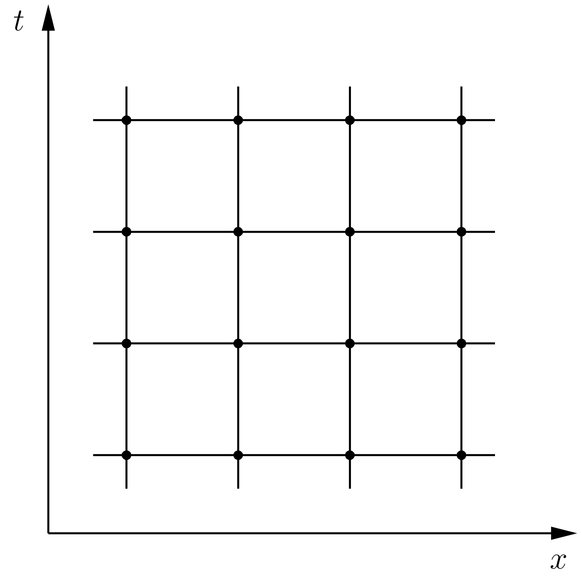


Figure 1.2: Schematic view of the grid.

In particular, for a uniform grid, we have

$$\begin{cases} t^n = n\Delta t \\ x_m = a + m\Delta x \end{cases} \tag{1.27}$$

Later on, we shall use the notations $k = \Delta t, h = \Delta x$.

- In numerical computations, we work with a finite set of numerical values. Typically, though not necessarily, we relate them with the values of the continuous function $u(x,t)$ at the grid points.

$$u_m^n \approx u(t^n, x_m). \tag{1.28}$$

- Initial data is typically assigned from the continuous initial condition.

$$u_m^0 = u(0, x_m). \tag{1.29}$$

Treatment of boundary data is more complex, and we defer the discussions later.

- The numerical solution is obtained from a scheme, that is, a formula to derive numerical values at a higher (later) time level from previous ones. The goal of the scheme is to approximate the continuous solution properly, i.e. $u_m^n \to u(t^n, x_m)$. The meaning of the limit will be precisely defined later. The major task of scientific computing or numerical analysis is to design and to analyze schemes.

The design of a scheme starts with numerical derivatives. But from a derivative, there are many different ways to approximate. For temporal derivative, we may take either of the following approximations.

$$
\begin{aligned}
&\frac{\partial u}{\partial t}(t^n, x_m)\\
\approx\ & \frac{u(t^{n+1}, x_m) - u(t^n, x_m)}{k} \rightarrow \frac{u_m^{n+1} - u_m^n}{k}\\
\text{or}\ & \frac{u(t^n, x_m) - u(t^{n-1}, x_m)}{k} \rightarrow \frac{u_m^n - u_m^{n-1}}{k}\\
\text{or}\ & \frac{u(t^{n+1}, x_m) - \frac{u(t^n, x_{m-1}) + u(t^n, x_{m+1})}{2}}{k} \rightarrow \frac{u_m^{n+1} - \frac{u_{m+1}^n + u_{m-1}^n}{2}}{k}\\
\text{or}\ & \frac{u(t^{n+1}, x_m) - u(t^{n-1}, x_m)}{2k} \rightarrow \frac{u_m^{n+1} - u_m^{n-1}}{2k}.
\end{aligned}
\tag{1.30}
$$

The spatial derivative also has several possibilities.

$$
\begin{aligned}
&\frac{\partial u}{\partial x}(t^n, x_m)\\
\approx & \frac{u(t^n, x_{m+1}) - u(t^n, x_{m-1})}{2h} \rightarrow \frac{u_{m+1}^n - u_m^{n-1}}{2h} \equiv D_c u_m^n \quad \text{central difference}\\
\text{or} & \frac{u(t^n, x_{m+1}) - u(t^n, x_m)}{h} \rightarrow \frac{u_{m+1}^n - u_m^{n-1}}{h} \equiv D_+ u_m^n \quad \text{forward difference}\\
\text{or} & \frac{u(t^n, x_m) - u(t^n, x_{m-1})}{h} \rightarrow \frac{u_m^n - u_{m-1}^n}{h} \equiv D_- u_m^n \quad \text{backward difference.}
\end{aligned}
\tag{1.31}
$$

The second order derivative may be approximated by a combination of first order derivatives, see Figure 1.3. For instance, a second order of central difference may be obtained by combining a forward difference and a backward difference.

$$
\begin{aligned}
&u_{xx}(t^n, x_m)\\
\approx & \frac{u_x(t^n, x_{m+1}) - u_x(t^n, x_m)}{h} \quad \text{forward}\\
\approx & \frac{\frac{u(t^n, x_{m+1}) - u(t^n, x_m)}{h} - \frac{u(t^n, x_m) - u(t^n, x_{m-1})}{h}}{h} \quad \text{backward}\\
= & \frac{u(t^n, x_{m+1}) - 2u(t^n, x_m) + u(t^n, x_{m-1})}{h^2}\\
\rightarrow & \frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}\\
= & D_+ D_- u_m^n.
\end{aligned}
\tag{1.32}
$$

Similarly, we may approximate the second order derivative by two for-

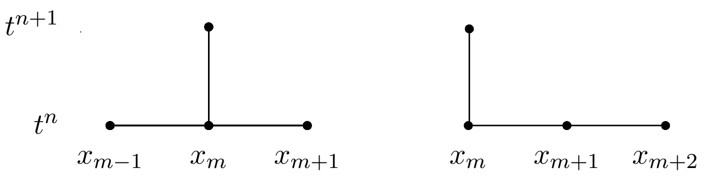


Figure 1.3: Central and one-sided difference for second order spatial derivative.

ward differences.

$$
\begin{aligned}
u_{xx}(t^n, x_m) &\approx D_+D_+u^n_m \\
&= \frac{\frac{u(t^n,x_{m+2})-u(t^n,x_{m+1})}{h} - \frac{u(t^n,x_{m+1})-u(t^n,x_m)}{h}}{h} \\
&\to \frac{u^n_{m+2} - 2u^n_{m+1} + u^n_m}{h^2}.
\end{aligned}
\tag{1.33}
$$

1.3 Basic Properties of a Finite Difference Scheme

For a difference scheme, we analyze how good it is by mainly three or four properties, namely, accuracy and consistency, stability and convergence.

1.3.1 Accuracy and Consistency: Taylor Expansion

In a finite difference scheme, we approximate continuous derivatives by differences. In turn, we may study the error of the differences by Taylor expansion. For instance, u_{xx} is approximated by (1.34) and (1.35). It may be calculated that, if expanded at (t^n, x_m),

$$
\begin{aligned}
&\frac{u(t^n,x_{m+1}) - 2u(t^n,x_m) + u(t^n,x_{m-1})}{h^2} \\
=&\frac{1}{h^2}[u(t^n,x_m) + hu_x(t^n,x_m) + \frac{h^2}{2}u_{xx}(t^n,x_m) + \frac{h^3}{6}u_{xxx}(t^n,x_m) \\
&+ \frac{h^4}{24}u_{xxxx}(t^n,x_m) - 2u(t^n,x_m) + u(t^n,x_m) - hu_x(t^n,x_m) \\
&+ \frac{h^2}{2}u_{xx}(t^n,x_m) - \frac{h^3}{6}u_{xxx}(t^n,x_m) + \frac{h^4}{24}u_{xxxx}(t^n,x_m) + O(h^5)] \\
=&u_{xx}(t^n,x_m) + \frac{h^2}{12}u_{xxxx}(t^n,x_m) + O(h^3).
\end{aligned}
\tag{1.34}
$$

In contrast, another difference form yields

$$\begin{aligned}
&\frac{u(t^n, x_{m+2}) - 2u(t^n, x_{m+1}) + u(t^n, x_m)}{h^2}\\
=&\frac{1}{h^2}[u(t^n, x_m) + 2hu_x(t^n, x_m) + \frac{4h^2}{2}u_{xx}(t^n, x_m)\\
&+ \frac{8h^3}{6}u_{xxx}(t^n, x_m) + O(h^4) - 2(u(t^n, x_m) + hu_x(t^n, x_m)\\
&+ \frac{h^2}{2}u_{xx}(t^n, x_m) + \frac{h^3}{6}u_{xxx}(t^n, x_m) + O(h^4)) + u(t^n, x_m)]\\
=&u_{xx}(t^n, x_m) + hu_{xxx}(t^n, x_m) + O(h^2).
\end{aligned} \tag{1.35}$$

Nevertheless, both difference forms approximate the continuous derivative in the limit $h \to 0$. A similar argument applies to the temporal derivative. Based on these expansions, consistency and order of accuracy are defined as follows.

Definition 1.1 (Consistency). *Given a partial differential equation $Pu = f$ and a finite difference scheme $P_{k,h}v = f$, we say that the finite difference scheme is consistent with the partial differential equation, if for any smooth function $\phi(t, x)$, it holds that*

$$P\phi - P_{k,h}\phi \to 0 \quad as \quad k, h \to 0. \tag{1.36}$$

Here the convergence is point-wise at each grid point.

Definition 1.2 (Truncation error, order of accuracy). *A scheme $P_{k,h}v = Pu = f$, which is consistent with the partial differential equation $Pu = f$, is accurate of order p in time, and of order q in space if $\forall\varphi(t, x)$ smooth, it holds that*

$$P_{k,h}\varphi - R_{k,h}f = O(k^p, h^q). \tag{1.37}$$

The above term is called truncation error. We denote this scheme accurate of order (p, q). If $k = \Lambda(h)$ is a smooth function, we say that the scheme $P_{k,h}v = R_{k,h}f$ is accurate of order r if $\forall\varphi(t, \lambda)$ smooth, it holds that

$$P_{k,h}\varphi - R_{k,h}f = O(h^r). \tag{1.38}$$

Consider the explicit one-sided difference scheme for the heat equation.

$$\frac{u_m^{n+1} - u_m^n}{k} = b\frac{u_{m-2}^n - 2u_{m-1}^n + u_m^n}{h^2}. \tag{1.39}$$

It is straightforward to find that the truncation error is

$$\frac{u_m^{n+1} - u_m^n}{k} - b\frac{u_{m-2}^n - 2u_{m-1}^n + u_m^n}{h^2} = u_t + \frac{ku_{tt}}{2} + O(k^2) - bu_{xx} + bhu_{xxx} + O(h^2). \tag{1.40}$$

Therefore, this scheme is accurate of the order $(1,1)$. Moreover, actually the numerical scheme approximates better the following equation, which is called as the modified equation for the scheme 1.55.

$$u_t + \frac{ku_{tt}}{2} - bu_{xx} + bhu_{xxx} = 0. \tag{1.41}$$

1.3.2 Basic Property of a Scheme: To Solve a Partial Differential Equation

Consistency and order of accuracy are local properties of a finite difference scheme. Other basic properties include convergence and stability, which are global properties.

Definition 1.3 (Convergence). *A "one step" finite difference scheme approximating a partial differential equation is a convergent scheme, if for any exact solution $u(t,x)$ to the partial differential equation and numerical solution u_m^n with the finite difference scheme, such that u_m^0 converges to $u_0(x)$ as $mh \to x$, then*

$$u_m^n \to u(t,x) \quad as \quad (nk, mh) \to (t,x) \quad and \quad h,k \to 0. \tag{1.42}$$

Remark 1.1. *The "convergence" will be precisely described later, usually not point-wise.*

Remark 1.2. *Convergence is usually not a simple issue. We need to define in a certain sense,*

$$\| u(nk,mh) - u_m^n \| \to 0. \tag{1.43}$$

This involves

1. *Explicit solution of the partial differential equation*
2. *Explicit solution of the FD scheme*

To fix the first issue, we define an "L_2" norm for the grid function as follows.

$$\| u(t^n,\cdot) \|_{2,h} = (h \sum_{m=-\infty}^{\infty} (u_m^n)^2)^{\frac{1}{2}}. \tag{1.44}$$

For the second issue, however, we do not have a way to treat at this point.

There is another important property for a numerical scheme, that is stability. This concern arises naturally in the following manner. Let a numerical solution $\{\widetilde{u}_m^n\}$ obtained with initial data $\{\widetilde{u}_m^0\}$, and another numerical solution $\{\bar{u}_m^n\}$ with initial data $\{\bar{u}_m^0\}$. Under the condition that $\| \widetilde{u}_m^0 - \bar{u}_m^0 \|$ is small, do we have $\| \widetilde{u}_m^n - \bar{u}_m^n \|$ small? If stability does not hold for a numerical scheme, then the scheme is useless as a small perturbation will cause totally different numerical solutions. We notice that there are uncertainties in numerical computations, such as round-off error.

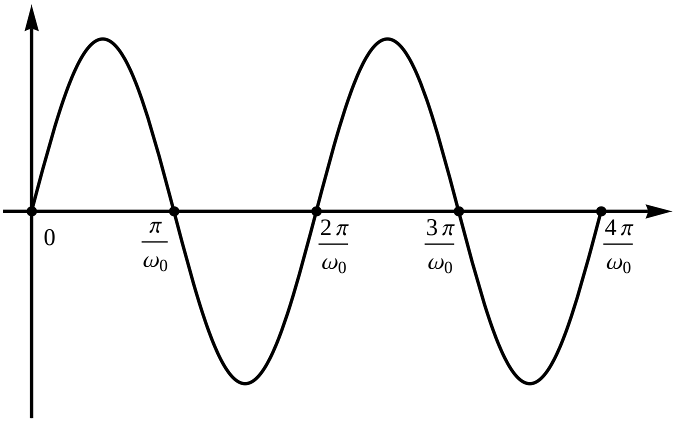


Figure 1.4: Schematic view of wave-number and gridpoint.

Definition 1.4 (Stability). *A finite difference scheme $P_{k,h}u_m^n = 0$ for a first order (in time) equation is stable if $\exists J \in \mathbb{N}$ and $h_0, k_0 > 0$, $\forall T > 0$, $\exists C_T$, such that $\forall h < h_0$, $k < k_0$, and $t < T$,*

$$\|u^n\|_h \le C_T \sum_{j=0}^{J} \|u^j\|_h. \tag{1.45}$$

Related to the stability for a numerical scheme, the stability for the partial differential equation should hold in the first place.

Definition 1.5. *A first order (in time) partial differential equation is well-posed if $\forall T > 0$, $\exists C_T$, such that for any solution $u(t,x)$, it holds that $\forall t < T$,*

$$\| u(t,\cdot) \| \le C_T \| u(0,\cdot) \| . \tag{1.46}$$

Example 1.1. *The heat equation $u_t = bu_{xx}\,(b \ge 0)$ is well-posed.*

In fact, we use the Fourier transform to deduce

$$\begin{aligned} \| u(t,\cdot) \|^2 &= \| \hat{u}(t,\cdot) \|^2 \\ &= \| \hat{u}(0,\cdot)e^{-b\omega^2 t} \|^2 \\ &\le \| \hat{u}(0,\cdot) \|^2 . \end{aligned} \tag{1.47}$$

The well-posedness is evident.

Example 1.2. *The inverse heat equation $u_t = -bu_{xx}\,(b > 0)$ is ill-posed.*

In fact, we follow the previous computations for the heat equation to obtain

$$\| u(t,\cdot) \|^2 = \| \hat{u}(0,\cdot)e^{+b\omega^2 t} \|^2 . \tag{1.48}$$

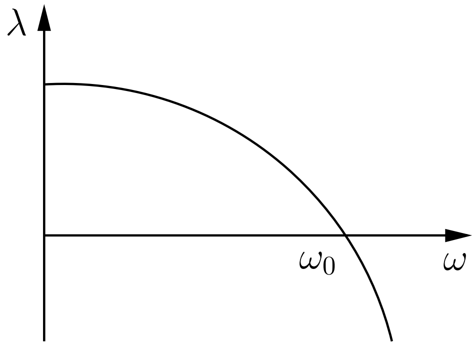


Figure 1.5: Dispersion relation for (1.50).

Then, $\forall T > 0$, $\forall C_T > 0$ and $\forall t > 0$,there always exists an ω_0 big enough, such that $e^{b\omega_0^2 t} > 2C_T$, and hence for an initial data with ω close to ω_0, and then

$$\| u(t,\cdot) \|^2 > C_T^2 \| \hat{u}(0,\cdot) \|^2 . \tag{1.49}$$

We make the following remarks.

- If the partial differential equation is ill-posed, then numerical stability is not possible in general.
- Ill-posed partial differential equation needs special function spaces to work on. The scheme should be designed in a special way as well.
- A well-posed partial differential equation does not necessarily requires negative dispersion relation for all frequencies.
- Numerically,the space grid should be fine enough to resolve those wave numbers for which the evolution is not negligible, i.e., $h < \pi/\omega_0$. Consider a solution $u(t,x) \simeq \sum_\omega \hat{u}(t,\omega)e^{\lambda t + i\omega x}$. Accordingly, it holds that $\| u(t,x) \|^2 = \sum_\omega \| \hat{u}(0,\omega)e^{\lambda t} \|^2$. If $h > \pi/\omega_0$, then $\| u_m^n(t,x) \|^2$ can not resolve the temporal increasing term for $\omega < \pi/h$, see Figure 1.4.

For instance, we consider the following advection-diffusion equation.

Example 1.3.

$$u_t = au + bu_{xx} \quad a, b > 0. \tag{1.50}$$

The dispersion relation gives $\lambda = a - b\omega^2$, as shown in Figure 1.5. For $\omega < \omega_0 = \sqrt{\frac{a}{b}}$, we may find that $\lambda > 0$. In contrast, we find that $\lambda < 0$ for $\omega > \omega_0$.

We motivate the stability analysis from the following specific scheme. A three-point explicit scheme

$$u_m^{n+1} = \alpha u_{m-1}^n + \beta u_m^n + \gamma u_{m+1}^n, \tag{1.51}$$

is stable if and only if

$$|\alpha| + |\beta| + |\gamma| \leq 1. \tag{1.52}$$

Proof. (Sufficiency) We prove the stability condition by direct calculations.

$$\begin{aligned}
&\sum_m |u_m^{n+1}|^2 \\
=&\sum_m |\alpha u_{m-1}^n + \beta u_m^n + \gamma u_{m+1}^n|^2 \\
\leq&\sum \alpha^2 (u_{m-1}^n)^2 + \beta^2 (u_m^n)^2 + \gamma^2 (u_{m+1}^n)^2 \\
&+ 2|\alpha\beta||u_{m-1}^n u_m^n| + 2|\beta\gamma||u_m^n u_{m+1}^n| + 2|\alpha\gamma||u_{m-1}^n u_{m+1}^n| \\
\leq&\sum_m \alpha^2 (u_{m-1}^n)^2 + \beta^2 (u_m^n)^2 + \gamma^2 (u_{m+1}^n)^2 \\
&+ |\alpha\beta|((u_{m-1}^n)^2 + (u_m^n)^2) + |\beta\gamma|((u_m^n)^2 + (u_{m+1}^n)^2) \\
&+ |\alpha\gamma|((u_{m-1}^n)^2 + (u_{m+1}^n)^2) \\
=&(\alpha^2 + \beta^2 + \gamma^2 + 2|\alpha\beta| + 2|\beta\gamma| + 2|\alpha\gamma|) \sum_m (u_m^n)^2 \\
=&(|\alpha| + |\beta| + |\gamma|)^2 \sum_m (u_m^n)^2 \\
\leq&(|\alpha| + |\beta| + |\gamma|)^{2n} \sum_m (u_m^0)^2.
\end{aligned} \tag{1.53}$$

□

We apply this condition to the explicit central-difference scheme

$$\frac{u_m^{n+1} - u_m^n}{k} = b\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2}. \tag{1.54}$$

Let $\mu = k/h^2$. This scheme may be recast into

$$u_m^{n+1} = b\mu u_{m-1}^n + (1 - 2b\mu)u_m^n + b\mu u_{m+1}^n. \tag{1.55}$$

The stability reads

$$|\alpha| + |\beta| + |\gamma| = 2b\mu + |1 - 2b\mu|. \tag{1.56}$$

If $1 - 2b\mu \geq 0$, i.e., $k \leq h^2/2b$, then $|\alpha| + |\beta| + |\gamma| = 1$, and the scheme is stable. It is important to note that the time step size is on the order of the space grid size squared.

1.4 Von Neumann Analysis

A systematic way to study numerical stability is the von Neumann analysis. The basic tool that facilitates the von Neumann analysis is the discrete Fourier transform.

$$\mathcal{F}\{u_m^n\} = \hat{u}^n(\xi) = \frac{1}{\sqrt{2\pi}}\sum_m e^{-imh\xi}u_m^n h, \quad \xi \in [-\frac{\pi}{h}, \frac{\pi}{h}]. \tag{1.57}$$

The inverse discrete Fourier transform is

$$u_m^n = \mathcal{F}^{-1}\{\hat{u}^n(\xi)\} = \frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh\xi}\hat{u}^n(\xi)\mathrm{d}\xi. \tag{1.58}$$

It is evident that the discrete Fourier transform may be regarded as a special case of the Fourier transform, for which the wave number is restricted to $[-\pi/h, \pi/h]$.

The Parseval's relation now reads

$$\| \hat{u}^n \|_h \equiv \sqrt{\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\hat{u}^n(\xi)|^2 \mathrm{d}\xi} = \| u^n \|_h \,. \tag{1.59}$$

We consider the explicit central difference scheme (1.55). Applying the discrete Fourier transform to both sides, we obtain

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh\xi}\hat{u}^{n+1}(\xi)\mathrm{d}\xi \\
=&\frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh\xi}[(1-2b\mu)\hat{u}^n(\xi) \\
&+ b\mu(e^{i(m-1)h\xi}\hat{u}^n(\xi) + e^{i(m+1)h\xi}\hat{u}^n(\xi))]\mathrm{d}\xi \\
=&\frac{1}{\sqrt{2\pi}}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} e^{imh\xi}[(1-2b\mu) + b\mu(e^{-ih\xi} + e^{ih\xi})]\hat{u}^n(\xi)\mathrm{d}\xi.
\end{aligned} \tag{1.60}$$

As the Fourier modes are linearly independent, we have for any $\xi \in [-\pi/h, \pi/h]$ that

$$\hat{u}^{n+1}(\xi) = [(1-2b\mu) + b\mu(e^{-ih\xi} + e^{ih\xi})]\hat{u}^n(\xi). \tag{1.61}$$

This naturally identifies an important notion in the numerical stability analysis.

Definition 1.6 (Amplification factor)**.** *For a Fourier mode $u_m^n = U^n e^{inh\xi}$, the ratio between the amplitude at t^{n+1} and that at t^n from a numerical scheme defines an amplification factor $g(\theta, k, h)$ with $\theta = h\xi$.*

For instance, the amplification factor for the explicit central difference scheme is

$$g(h\xi) = (1 - 2b\mu) + b\mu(e^{-ih\xi} + e^{ih\xi}). \tag{1.62}$$

The meaning for the amplification factor is straightforward. If initial data consists of a single mode, numerically it evolves according to $\hat{u}(nk) = (g(\theta, h, k))^n \hat{u}(0)$.

There is a very useful theorem, which allows stability analysis through the amplification factor.

Theorem 1.1. *A one-step finite difference scheme is stable if and only if $\exists K$, independent of θ, k, h, and $h_0, k_0 > 0$, such that $|g(\theta, k, h)| \leq 1 + Kk$ for any $\theta, 0 < k \leq k_0, 0 < h \leq h_0$. If further $g(\theta, k, h)$ is independent of h, k, then the stability condition is $|g(\theta)| \leq 1$.*

Before prove this theorem, we first present the following corollary.

Corollary 1.1. *If a finite scheme is modified in a certain manner such that only an $O(k)$ modification uniformly in ξ is introduced, then the modified scheme is stable provided that the original one is so.*

Proof. Let the amplification factor for the new scheme is $g^{'} = g + O(k)$. If $|g| \leq 1 + Kk$, then $|g^{'}| = |g + O(k)| \leq 1 + Kk + Vk = 1 + K^{'}k$.

We remark that $g^{'} = g + O(k)$ is sufficient and necessary. □

Next, we prove the theorem.

Proof. (1) (Sufficiency). For the n-th time step, we know that $nk \leq T$. We estimate

$$\begin{aligned} \| u^n \|_h^2 &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |g(\theta, k, h)|^{2n} |\hat{u}^0(\xi)|^2 \mathrm{d}\xi \\ &\leq (1 + Kk)^{2n} \| u^0 \|_h^2 \\ &\leq (1 + Kk)^{2T/k} \| u^0 \|_h^2 \\ &\leq e^{2KT} \| u^0 \|_h^2 . \end{aligned} \tag{1.63}$$

Therefore, the scheme is stable.

(2) (Necessity) The necessity is shown by contradiction.

Assume that for $\forall C > 0$, $h_0 > 0$, $k_0 > 0$, $\exists \theta^* \in [0, \pi]$, and $\exists h \in (0, h_0]$, $\exists k \in (0, k_0]$ such that $|g(\theta^*, k, h)| \geq 1 + 2Ck$. Due to the continuity of $g(\xi, h, k)$, there exists an interval $[\theta_1, \theta_2]$, such that $\forall \theta \in [\theta_1, \theta_2]$, it holds that $|g(\theta, k, h)| \geq 1 + Ck$.

We construct an initial data with its Fourier transform as follows.

$$\hat{u}_0(\xi) = \begin{cases} 0 & \text{if} \quad \theta \notin [\theta_1, \theta_2], \\ \sqrt{h(\theta_2 - \theta_1)^{-1}} & \text{if} \quad \theta \in [\theta_1, \theta_2]. \end{cases} \tag{1.64}$$

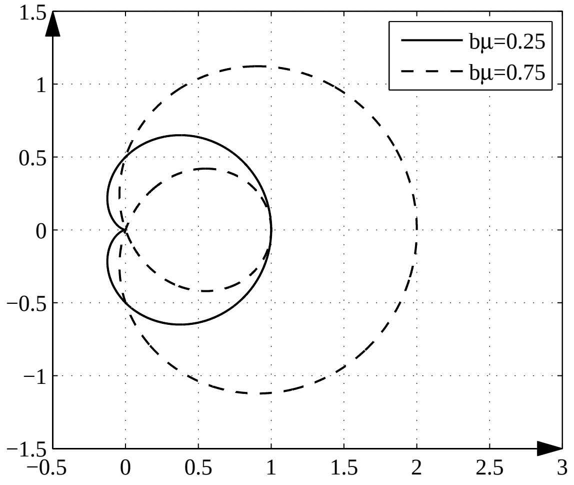


Figure 1.6: The amplification factor for the explicit central difference scheme in two different cases: $b\mu < 0.5$ (solid) and $b\mu > 0.5$ (dashed).

It is easy to find that

$$\| u^0 \|_h = 1. \tag{1.65}$$

On the other hand, we compute the norm at t^n.

$$\begin{aligned} \| u^n \|_h^2 &= \int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} (\hat{u}(\xi))^2 \mathrm{d}\xi \\ &= \int_{\theta_1}^{\theta_2} |g|^{2n} \frac{\mathrm{d}\theta}{\theta_2 - \theta_1} \\ &\geq (1 + Ck)^{2n}. \end{aligned} \tag{1.66}$$

In particular, at the time step very close to the terminal time T, we have

$$\| u^n \|_h^2 \geq \frac{1}{2} e^{2TC} \| u^0 \|_h^2 . \tag{1.67}$$

This proves the instability. □

This theorem justifies the von Neumann analysis, namely, one finds the amplification factor and check whether the stability condition above holds or not.

For instance, we apply the von Neumann analysis to the explicit central difference. In fact, from (1.62), we have (see Figure 1.6)

$$g = (1 - 2b\mu) + 2b\mu \cos\theta. \tag{1.68}$$

The stability condition may be found from

$$|g| \le 1 \Leftrightarrow 2b\mu \le 1 \Leftrightarrow \mu \le \frac{1}{2b} \Leftrightarrow k \le \frac{h^2}{2b}. \tag{1.69}$$

We observe that this is consistent with our previous result.

Next, we apply the von Neumann analysis to an implicit central difference scheme.

$$\frac{u_m^{n+1} - u_m^n}{k} = b\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{h^2}. \tag{1.70}$$

In fact, the computation of amplification factor is simply to make a correspondence g with a time forward stepping, and a correspondence $e^{il\theta}$ with u_{n+l}. For the implicit central-difference scheme, we have

$$\frac{g-1}{k} = b\frac{ge^{i\theta} - 2g + ge^{-i\theta}}{h^2}. \tag{1.71}$$

or,

$$-(1+2b\mu)g + b\mu g(e^{i\theta} + e^{-i\theta}) = -1. \tag{1.72}$$

Therefore, the amplification factor is

$$g = \frac{1}{(1+2b\mu) - 2b\mu\cos\theta}. \tag{1.73}$$

We conclude that this implicit scheme is unconditionally stable, as $0 < g \le 1, \forall \mu$. This property allows large time step size, which is very useful for many applications.

In the following, we introduce several other schemes and make von Neumann analysis.

First, the Crank-Nicolson scheme uses the average of u_{xx} at t^n and t^{n+1}.

$$\frac{u_m^{n+1} - u_m^n}{k} = \frac{b}{2}\left(\frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2} + \frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2}\right). \tag{1.74}$$

Note that here the first term on the righthand side is explicit, and the second one implicit. Moreover, if we perform Taylor expansion at $(x_m, t^{n+1/2})$, it is easy to show that the scheme has a second order accuracy in both space and time.

We may rewrite the scheme in a vector form. Let $u^n = (\cdots, u_j^n, \cdots)^T$ be the vector at time t^n. The Crank-Nicolson scheme is

$$u^{n+1} - u^n = \frac{b\mu}{2}(Au^n + Au^{n+1}). \tag{1.75}$$

Here $A = \text{tridiag}(1, -2, 1)$. Therefore, the scheme gives

$$(I - \frac{b\mu}{2}A)u^{n+1} = (I + \frac{b\mu}{2}A)u^n. \tag{1.76}$$

or,

$$u^{n+1} = (I - \frac{b\mu}{2}A)^{-1}(I + \frac{b\mu}{2}A)u^n. \tag{1.77}$$

Actually, some previous discussed schemes may be rewritten in a vector form. For instance, the explicit central-difference scheme is

$$u^{n+1} = (I + b\mu A)u^n. \tag{1.78}$$

The implicit scheme is

$$u^{n+1} = (I - b\mu A)^{-1}u^n. \tag{1.79}$$

Comparing these two schemes and the Crank-Nicolson scheme, we observe that all are an approximation (time integration) for the semi-discrete system

$$\frac{\partial}{\partial t}u = bAu. \tag{1.80}$$

The exact solution to the semi-discrete system is

$$u(t) = e^{bAt}u(0) = (1 + bAt + \frac{(bAt)^2}{2} + \cdots)u(0). \tag{1.81}$$

Besides the Crank-Nicolson scheme, another naturally devised scheme with second order accuracy in both space and time is the leapfrog scheme.

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - 2v_m^n + v_{m-1}^n}{h^2}. \tag{1.82}$$

We formally follow the von Neumann analysis and compute the amplification factor as follows.

$$\frac{g - \frac{1}{g}}{2k} = b\frac{e^{i\theta} - 2 + e^{-i\theta}}{h^2}. \tag{1.83}$$

That is,

$$g - \frac{1}{g} = -8b\mu \sin^2\frac{\theta}{2}. \tag{1.84}$$

The amplification factor g solves

$$g^2 + 8b\mu \sin^2\frac{\theta}{2}g - 1 = 0. \tag{1.85}$$

The too roots satisfy

$$g_1 g_2 = -1, \quad g_1 + g_2 = -8b\mu \sin^2\frac{\theta}{2} \tag{1.86}$$

Therefore, we have either $|g_1| > 1$ or $|g_2| > 1$. The leapfrog scheme is therefore unstable.

To rectify the stability and maintain the accuracy for the leapfrog scheme, a Dufort-Frankel scheme was proposed.

$$\frac{v_m^{n+1} - v_m^{n-1}}{2k} = b\frac{v_{m+1}^n - (v_m^{n+1} + v_m^{n-1}) + v_{m-1}^n}{h^2}. \tag{1.87}$$

In a similar way, we compute the amplification factor from

$$\frac{g - \frac{1}{g}}{2k} = \frac{b}{h^2}(e^{i\theta} - (g + \frac{1}{g}) + e^{-i\theta}). \tag{1.88}$$

This leads to

$$(1 + 2b\mu)g^2 - 4b\mu\cos\theta g - (1 - 2b\mu) = 0. \tag{1.89}$$

The roots are

$$g_\pm = \frac{2b\mu\cos\theta \pm \sqrt{1 - 4b^2\mu^2\sin^2\theta}}{1 + 2b\mu}, \tag{1.90}$$

if $1 - 4b^2\mu^2\sin^2\theta \geq 0$. In this case, we find that

$$|g_\pm| \leq \frac{2b\mu|\cos\theta| + \sqrt{1 - 4b^2\mu^2\sin^2\theta}}{1 + 2b\mu} \leq \frac{2b\mu + 1}{1 + 2b\mu} = 1. \tag{1.91}$$

The scheme is stable.

On the other hand, if $1 - 4b^2\mu^2\sin^2\theta < 0$, we find that

$$|g_\pm|^2 \leq \frac{(2b\mu\cos\theta)^2 - 1 + 4b^2\mu^2\sin^2\theta}{(1 + 2b\mu)^2} = \frac{4b^2\mu^2 - 1}{4b^2\mu^2 + 4b\mu + 1} < 1. \tag{1.92}$$

Therefore, the Dufort-Frankel scheme is unconditionally stable and possess a second order of accuracy.

1.5 Lax-Richtmyer Equivalence Theorem

As mentioned before, numerical convergence is an important property for a scheme. Because it is a global property, to prove convergence is not an easy task. On the other hand, the von Neumann analysis renders a handy tool for us to analyze the stability, which is another global property for a scheme. The Lax-Richtmyer theorem bridges these two properties, and therefore pave the way toward the convergence analysis. The main theorem is as follows.

Theorem 1.2. *For a consistent one step linear scheme for Cauchy problem of a well-posed linear partial differential equation, stability is the necessary and sufficient condition to convergence.*

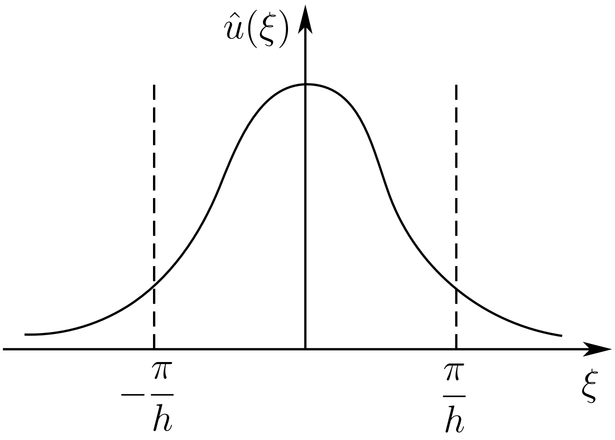


Figure 1.7: The truncation operator is a low-pass filter.

To facilitate the proof, we need to be more precise in interpreting numerical solutions. While the solution to partial differential equation lies in $L^2(\mathbb{R})$, the numerical solution lies in l^2 (also denoted as $L^2(h\mathbb{Z})$. We relate them to the same function space by a *truncation operator* and an *interpolation operator.*

The truncation operator $T : L^2(\mathbb{R}) \to l^2$ maps a continuous function $u(x)$ to a grid function as follows. Suppose that the Fourier transform of $u(x)$ is $\hat{u}(\xi)$, then grid function Tu takes m-th grid value as

$$Tu_m = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{imh\xi} \hat{u}(\xi) \mathrm{d}\xi. \tag{1.93}$$

In fact, they are related by

$$\widehat{Tu}(\xi) = \hat{u}(\xi), \quad |\xi| \le \pi/h. \tag{1.94}$$

We remark that $Tu_m \neq u(x_m)$ in general. Furthermore, the truncation operator may be regarded as a low-pass filter in terms of signal processing, as shown in Figure 1.7.

Next, we define the interpolation operator $S : l^2 \to L^2(\mathbb{R})$, which maps a grid function v to a continuous function $Sv(x)$ as follows.

Suppose that the discrete Fourier transform of v is $\hat{v}(\xi)$, then we define Sv by

$$Sv(x) = \frac{1}{\sqrt{2\pi}} \int_{-\pi/h}^{\pi/h} e^{ix\xi} \hat{v}(\xi) \mathrm{d}\xi. \tag{1.95}$$

In other words, the spectra are related by (see Figure 1.8)

$$\widehat{Sv}(\xi) = \begin{cases} \hat{v}(\xi), & |\xi| \le \pi/h, \\ 0, & |\xi| > \pi/h. \end{cases} \tag{1.96}$$

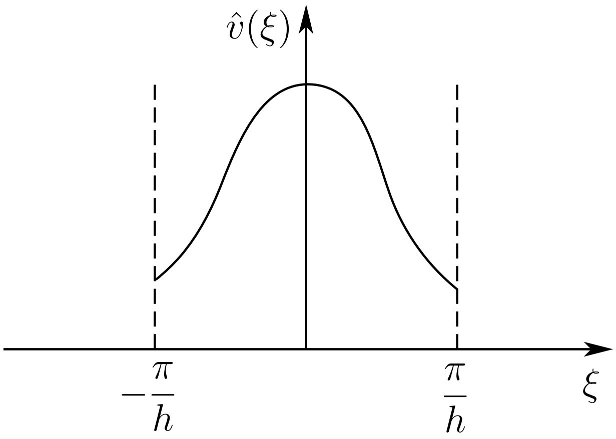


Figure 1.8: Interpolation operator.

With these operators, we have clearer definitions for some properties of a numerical scheme. In particular, a scheme is *convergent* for Cauchy problem of a partial differential equation if $\forall h, k \to 0$, $\forall u$ solution to partial differential equation, and $\forall v$ numerical solution with initial data $Sv^0 \to u(\cdot, 0)$ in L^2, it holds that $Sv^n \to u(\cdot, t^n)$ in L^2 for $t^n = nk$.

As we already know, a linear partial differential equation has a unique dispersion relation $q(\xi)$, i.e. its solution is $\hat{u}(\xi, t) = \hat{u}(\xi, 0)e^{q(\xi)t}$. Meanwhile, a scheme has a unique amplification factor $g(h\xi, h, k)$, i.e. numerical solution is $\hat{u}^{n+1}(\xi) = g\hat{u}^n(\xi)$. Therefore, a scheme is *consistent* to a partial differential equation if for $|\xi| \leq \pi/h$, it holds that

$$\frac{e^{kq(\xi)} - g}{k} = o(1), \quad \text{in } (h, k). \tag{1.97}$$

A scheme is *stable* if $\forall T > 0$, there exists $C_T > 0$, such that $\forall 0 < n < T/k$, it holds $|g^n| \leq C_T$. As we have proved in the von Neumann analysis, the stability is equivalent to the existence of a K such that $|g| \leq 1 + Kk$.

Proof. We first prove that stability implies convergence. We first assume that the numerical initial data is taken as precisely $v^0 = Tu(x, 0)$. By this choice, we have

$$\|u(\cdot, 0) - Sv^0\|^2 = \|u(\cdot, 0) - STu(\cdot, 0)\|^2 = \int_{|\xi| > \pi/h} |\hat{u}_0(\xi)|^2 \mathrm{d}\xi. \tag{1.98}$$

Because $u(\cdot, 0) \in L^2$, this term approaches to zero as $h \to 0$.

The Fourier transforms of the exact solution and numerical solution are expressed explicitly. So we have, at time $t = t^n$

$$\begin{aligned}
&\|u(\cdot,t^n) - Sv^n\|^2 \\
= &\int_{-\infty}^{+\infty} |u(x,t^n) - Sv^n(x)|^2 \mathrm{d}x \\
= &\int_{-\infty}^{+\infty} |\hat{u}(\xi,t^n) - \widehat{Sv^n}(\xi)|^2 \mathrm{d}\xi \\
= &\int_{-\pi/h}^{\pi/h} |e^{qt^n} - g^n|^2 |\hat{u}_0(\xi)|^2 \mathrm{d}\xi + \int_{|\xi|>\pi/h} |e^{qt^n}|^2 |\hat{u}_0(\xi)|^2 \mathrm{d}\xi.
\end{aligned} \tag{1.99}$$

Now we define a function

$$\Phi_h(\xi) = \left\{ \begin{array}{ll} |e^{qt^n} - g^n|^2 |\hat{u}_0(\xi)|^2, & |\xi| \le \pi/h, \\ |e^{qt^n}|^2 |\hat{u}_0(\xi)|^2, & |\xi| > \pi/h. \end{array} \right. \tag{1.100}$$

Due to the well-posedness of the partial differential equation and the stability of the numerical scheme, we know that both $|e^{qt^n} - g^n|$ and $|e^{qt^n}|$ are bounded for any $t^n \le T$ (T is a given terminate time). In addition, we have $u(\cdot, 0) \in L^2$, therefore it holds that $\Phi_h \in L^1$.

For each fixed ξ, it is observed that

$$\Phi_h(\xi) = |e^{qt^n} - g^n|^2 |\hat{u}_0(\xi)|^2, \quad \text{as } h \to 0. \tag{1.101}$$

Moreover, from the bounds of g and q, as well as the consistency, we may derive

$$\begin{aligned}
&|e^{qt^n} - g^n| \\
= &|e^{qk} - g| \sum_{j=0}^{n-1} e^{(n-j-1)qk} g^j \\
\le &|e^{qk} - g| n C_T^2 \\
\le &n C_T^2 k o(1).
\end{aligned} \tag{1.102}$$

Due to $nk \le T$, the above term is on the order of $o(1)$.

Noticing that $\Phi_h(\xi) \le (2C_T)^2 |\hat{u}_0(\xi)|^2$, by the Lebesgue dominated convergence theorem, we have

$$\lim_{h\to 0} \int_{-\infty}^{+\infty} \Phi_h \mathrm{d}\xi = \lim_{h\to 0} \int_{-\infty}^{+\infty} |\hat{u}(\xi,t^n) - \widehat{Sv^n}(\xi)|^2 \mathrm{d}\xi = 0. \tag{1.103}$$

Up to this point, the convergence is proved for the case $v^0 = Tu(\cdot, 0)$. For general numerical initial data $v^0 \ne Tu(\cdot, 0)$, we compare an corresponding solution v^n with the numerical solution w^n that takes initial data $w^0 = Tu(\cdot, 0)$. By the triangular inequality, we derive that

$$\|u(\cdot,t^n) - Sv^n\| \le \|u(\cdot,t^n) - Sw^n\| + \|Sw^n - Sv^n\|. \tag{1.104}$$

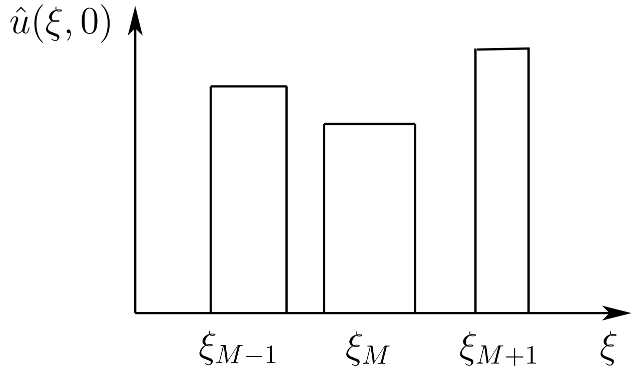


Figure 1.9: Construction of initial data in spectral space.

The first term converges to 0, and the second term converges as well. To see this, we derive

$$\begin{array}{ll} & \|Sw^n - Sv^n\| \\ = & \|S(w^n - v^n)\| \\ = & \|w^n - v^n\|_{l^2} \\ \leq & C_T\|w^0 - v^0\|_{l^2} \\ = & C_T\|Tu(\cdot,0) - v^0\|_{l^2} \\ \leq & C_T\|u(\cdot,0) - Sv^0\| \\ \to & 0. \end{array} \tag{1.105}$$

Now we turn to the second part of the theorem, namely, to prove that instability implies divergence. By instability, we know that for any $M \in I\!N$, there exist ξ_M, h_M, k_M, such that

$$|g(h_M\xi_M, h_M, k_M)| > 1 + Mk_M, \quad \text{with } |h_M\xi_M| \leq \pi. \tag{1.106}$$

Since g is continuous, there exists a neighborhood of ξ_M with radius $\eta_M > 0$, denoted as I_M, such that

$$|g(h_M\xi, h_M, k_M)| \geq 1 + \frac{M}{2}k_M, \quad \text{for } \xi \in I_M. \tag{1.107}$$

In particular, we may choose $\eta_M \leq M^{-2}, h_M \leq h_{M-1}, k_M \leq k_{M-1}$.We claim, yet without proof, that it is possible to choose these intervals I_M disjoint.

We define $\alpha_M = 1/(\sqrt{\eta_M}M)$, and take initial data $u(x,0) = \sum_M w_M(x)$, see Figure 1.9. Here w_M has the spectral representation

$$\hat{w}_M(\xi) = \left\{ \begin{array}{ll} \alpha_M, & \xi \in I_M, \\ 0, & \text{elsewhere.} \end{array} \right. \tag{1.108}$$

We check that $u(\cdot,0) \in L^2$ by

$$\|u(\cdot,0)\|^2 = \|\hat{u}(\cdot,0)\|^2 = \sum_M \|\hat{w}_M\|^2 = 2\sum_M \alpha_M^2 \eta_M = 2\sum_M \frac{1}{M^2} < +\infty. \tag{1.109}$$

At a given time T, by stability of the partial differential equation, we know that $|e^{qt}| \le C_T, \forall t \le T$. Take $M \ge 8(C_T - 1)/T$, and take time step $n \in [T/2k_M, T/k_M]$.

Let numerical solution to be v^n at time t^n. From instability, we know that for $\forall \xi \in I_M$,

$$\begin{aligned} |g(h\xi)^n - e^{q(\xi)nk}| &\ge |g(h\xi)|^n - C_T \\ &\ge (1 + \frac{1}{2}Mk_M)^n - C_T \\ &\ge 1 + \frac{n}{2}Mk_M - C_T \\ &\ge \frac{M}{2}\frac{T}{2} - \frac{MT}{8} \\ &= \frac{MT}{8}. \end{aligned} \tag{1.110}$$

Therefore we estimate the difference between the numerical solution and the exact solution as follows.

$$\begin{aligned} \|Sv^n - u(\cdot,t^n)\|^2 &\ge \int_{I_M} |g(h\xi)^n - e^{q(\xi)nk}|^2 |\hat{u}(\xi,0)|^2 \mathrm{d}\xi \\ &\ge \int_{I_M} \left(\frac{MT}{8}\right)^2 \alpha_M^2 \mathrm{d}\xi \\ &= 2\frac{M^2T^2}{64}\alpha_M^2 \eta_M \\ &\ge \frac{T^2}{32}. \end{aligned} \tag{1.111}$$

This does not converge to zero. Therefore convergence does not hold if stability does not hold. □

We remark that though we only prove that v^n does not converge to the exact solution, actually the numerical solution $\|v^n\|_{l^2} \to +\infty$.

We also remark that by the Duhamel's principle, it may be shown that the same theorem holds for general inhomogeneous linear partial differential equation.

1.6 Some Further Discussions

1.6.1 Boundary Condition: A Brief Discussion

Consider the heat equation

$$u_t = bu_{xx}, \quad x \in [\alpha, \beta]. \tag{1.112}$$

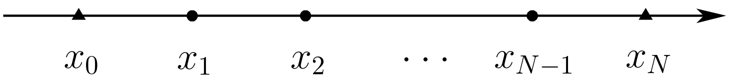


Figure 1.10: The Dirichlet boundary condition.

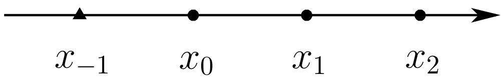


Figure 1.11: The Neumann boundary condition.

For the Dirichlet boundary condition,

$$\begin{cases} u(\alpha, t) = u_\alpha(t), \\ u(\beta, t) = u_\beta(t), \end{cases} \tag{1.113}$$

We notice that the explicit central difference scheme

$$u_m^{n+1} = (1 - 2b\mu)u_m^n + b\mu(u_{m-1}^n + u_{m+1}^n), \tag{1.114}$$

uses a three-point stencil. The numerical scheme forms a close system under the assignment for the boundary grid points with (see Figure 1.10)

$$\begin{cases} u_0^n = u_\alpha(t^n), \\ u_N^n = u_\beta(t^n). \end{cases} \tag{1.115}$$

On the other hand, a Neumann boundary condition $u_x(\alpha, t) = 0$ (and similar for the boundary condition at $x = \beta$) requires more detailed analysis.

One way to treat the Neumann boundary condition is to take

$$\frac{u_1 - u_0}{h} = 0. \tag{1.116}$$

This leads to $u_0 = u_1$, and hence the evolution is governed by

$$u_1^{n+1} = (1 - b\mu)u_1^n + b\mu u_2^n = u_1^n + b\mu(u_2^n - u_1^n). \tag{1.117}$$

The local truncation error is on the order of $O(h)$, if we compare the numerical boundary condition with the exact one at $x = \alpha$.

In general, the overall accuracy depends on both the inner scheme and the numerical boundary condition. Therefore, it is important to improve the accuracy at the boundary.

One way to improve the accuracy at the boundary is to introduce a ghost point $x_{-1} = \alpha - h$, as shown in Figure 1.11. Similar treatment applies to the other boundary $x_N = \beta$. Noticing that $\frac{u_1 - u_{-1}}{2h} = u_x + O(h^2)$, we have $u_{-1} = u_1$ which leads to

$$u_0^{n+1} = (1 - 2b\mu)u_0^n + 2b\mu u_1^n = u_0^n + 2b\mu(u_1^n - u_0^n). \tag{1.118}$$

An alternative is to design a one-sided second order boundary condition. To this end, we make Taylor expansions to the second order as follows.

$$u_1 = u_0 + u_x h + \frac{u_{xx}}{2}h^2 + O(h^3), \tag{1.119}$$

$$u_2 = u_0 + 2u_x h + \frac{u_{xx}}{2}(2h)^2 + O(h^3). \tag{1.120}$$

$$\tag{1.121}$$

This means

$$u_x = \frac{4u_1 - 3u_0 - u_2}{2h} + O(h^2) \tag{1.122}$$

Therefore, we obtain a second order accurate boundary condition

$$u_0 = \frac{4u_1 - u_2}{3}. \tag{1.123}$$

This leads to a formula for u_1 as follows.

$$\begin{aligned} u_1^{n+1} &= (1 - 2b\mu)u_1^n + b\mu(\frac{4u_1^n - u_2^n}{3} + u_2^n) \\ &= (1 - \frac{2b\mu}{3})u_1^n + \frac{2b\mu}{3}u_2^n \\ &= u_1^n + \frac{2b\mu}{3}(u_2^n - u_1^n) \end{aligned} \tag{1.124}$$

We remark that the construction of one-sided boundary condition is not unique, as more grid points may be included in the expansion.

1.6.2 An Alternative Way of Thinking: Approximate Integral Method

We now discuss an algorithm, which takes quite different way of thinking. We still consider the diffusion equation

$$u_t = bu_{xx}. \tag{1.125}$$

For the heat equation, the exact solution is

$$u(t + k, x) = \frac{1}{\sqrt{4\pi bk}} \int_{-\infty}^{+\infty} e^{-(x-y)^2/4bk} u(t, y)\mathrm{d}y. \tag{1.126}$$

Consider a simple reconstruction of the numerical data as shown in Figure 1.12

$$u(t^n, x) = u_m^n \quad \text{for} |x - mh| < \frac{h}{2}. \tag{1.127}$$

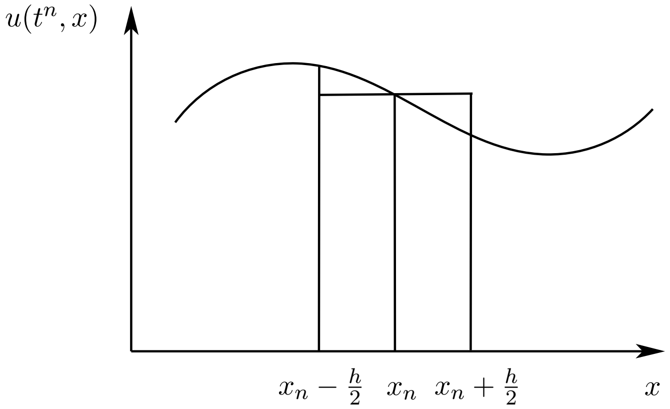


Figure 1.12: Reconstruction of the numerical data.

We obtain the exact solution at t^{n+1}

$$u_m^{n+1} = \sum_l \frac{1}{\sqrt{4\pi bk}} u_l^n \int_{x_l - \frac{h}{2}}^{x_l + \frac{h}{2}} e^{-(mh-y)^2/4bk} \mathrm{d}y. \tag{1.128}$$

Denoting the coefficients

$$K_j = \frac{1}{\sqrt{4\pi bk}} \int_{(j-\frac{1}{2})h}^{(j+\frac{1}{2})h} e^{-\epsilon^2/4bk} \mathrm{d}\epsilon = \frac{1}{2}\left(\operatorname{erf}\left(\frac{j+\frac{1}{2}}{2\sqrt{b\mu}}\right) - \operatorname{erf}\left(\frac{j-\frac{1}{2}}{2\sqrt{b\mu}}\right)\right), \tag{1.129}$$

we obtain

$$u_m^{n+1} = \sum_l K_{m-l} u_l^n = \sum_j K_j u_{m-j}^n. \tag{1.130}$$

Theoretically speaking, this algorithm takes numerical error only from the reconstruction and the average. However, in real applications, the numerical convolution $K * u^n$ is very expensive, and a truncated convolution is usually performed. That is, we use

$$u_m^{n+1} = \sum_{j=-p}^{p} K_j u_{m-j}^n. \tag{1.131}$$

To make K_p negligible, we notice that

$$K_p \approx \frac{h}{2}\frac{\mathrm{d}}{\mathrm{d}x}\operatorname{erf}\left(\frac{p}{2\sqrt{b\mu}}\right) = \frac{h}{\sqrt{4\pi bk}} e^{-p^2/4b\mu}. \tag{1.132}$$

Therefore, we choose $p \geq C_0\sqrt{\mu}$ for a certain constant C_0. We call this method the approximate integral method.

1.6.3 Convection-Diffusion Equation

A linear advection-diffusion equation reads

$$u_t + au_x = bu_{xx}. \tag{1.133}$$

A natural choice is an explicit central-difference scheme.

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_{m+1}^n - u_{m-1}^n}{2h} = b\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}. \tag{1.134}$$

or, we have

$$u_m^{n+1} = (1 - 2b\mu)u_m^n + b\mu(1-\alpha)u_{m+1}^n + b\mu(1-\alpha)u_{m-1}^n. \tag{1.135}$$

Here we define $\alpha = \frac{ha}{2b}$. It is a cell Reynolds number as h, a and b are the grid length, the advection velocity and the viscosity, respectively.

From previous study, we know that this scheme is stable if an only if

$$|1 - 2b\mu| + |b\mu(1-\alpha)| + |b\mu(1+\alpha)| \leq 1. \tag{1.136}$$

This is equivalent to $2b\mu \leq 1$ and $1 - \alpha \geq 0$, i.e., $\mu \leq \frac{1}{2b}$ and $h \leq \frac{2b}{a}$. The first requirement is standard for parabolic equations. The second one confines the spatial grid size to be small enough.

A consideration comes from the study of the pure advection equation

$$u_t + au_x = 0, \quad a \geq 0. \tag{1.137}$$

As shall be explained later in the next chapter, it is recommended to use an upwind scheme, namely, to obtain u_m^{n+1} in terms of u_{m-1}^n, u_m^n.

$$u_x \sim \frac{u_m^n - u_{m-1}^n}{h}. \tag{1.138}$$

Therefore, we take an upwind scheme for the advection-diffusion equation as follows.

$$\frac{u_m^{n+1} - u_m^n}{k} + a\frac{u_m^n - u_{m-1}^n}{h} = b\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}. \tag{1.139}$$

It may be rewritten as

$$u_m^{n+1} = (1 - 2b\mu(1+\alpha))u_m^n + b\mu u_{m+1}^n + b\mu(1+2\alpha)u_{m-1}^n. \tag{1.140}$$

Therefore, the stability condition is

$$|1 - 2b\mu(1+\alpha)| + b\mu + b\mu(1+2\alpha) \leq 1 \tag{1.141}$$

This is equivalent to

$$2b\mu(1+\alpha) \leq 1, \tag{1.142}$$

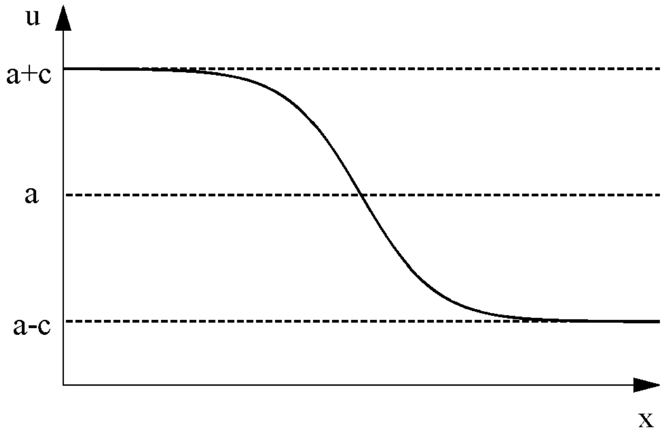


Figure 1.13: Wave profile for the Burgers' equation.

or, with the definition $\lambda = k/h$,

$$a\lambda + 2b\mu \leq 1. \tag{1.143}$$

We remark that $a\lambda \leq 1$ is the CFL (Courant-Friedrichs-Lewy) stability condition for $u_t + au_x = 0$. We shall discuss this in the next chapter. In the mean time, $2b\mu \leq 1$ is the stability condition for the diffusion equation.

We may use the schemes discussed before to treat nonlinear convection-diffusion equations. For instance, we consider the Burgers' equation

$$u_t + \frac{(u^2)_x}{2} = bu_{xx}. \tag{1.144}$$

An exact solution to this equation is (see Figure 1.13)

$$u(t,x) = a - c\tanh[\frac{c}{2b}(x - at)]. \tag{1.145}$$

Let $a > c > 0$. An explicit upwind scheme reads

$$\frac{u_m^{n+1} - u_m^n}{k} + \frac{(u_m^n)^2 - (u_{m-1}^n)^2}{2h} = b\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}. \tag{1.146}$$

It is obvious that the von Neumann analysis does not work anymore due to the nonlinearity. Even stability holds, the Lax-Richtmeyer equivalence theorem does not apply and the convergence is under question. Nevertheless, we expect that stability holds if

$$(a + c)\lambda + 2b\mu \leq 1. \tag{1.147}$$

1.6.4 Von Neumann Analysis for More General Situation

We shall briefly discuss stability for two more cases. First, for a multi-step scheme, the von Neumann stability analysis leads to a polynomial for the amplification factor. For instance, for the leap-frog scheme

$$\frac{u_m^{n+1} - u_m^{n-1}}{2k} = b\frac{u_{m+1}^n - 2u_m^n + u_{m-1}^n}{h^2}, \tag{1.148}$$

we may formally substitute u_{m+l}^{n+j} with $g^j e^{il\theta}$ to obtain

$$\frac{g - \frac{1}{g}}{2k} = b\frac{e^{i\theta} - 2 + e^{-i\theta}}{h^2}. \tag{1.149}$$

Therefore, the amplification factor is a root to

$$g^2 + 8b\mu \sin^2 \frac{\theta}{2} g - 1 = 0. \tag{1.150}$$

We observe that g is not a polynomial of θ.

In general, g is a root to the polynomial $\phi(g, \theta) = 0$. Let the ν-th branch of root be $g_\nu(\theta)$. The stability condition reads as follows.

Theorem 1.3. *Stability holds if all the roots $g_\nu(\theta)$ to $\phi(g, \theta)$ satisfy the following conditions. First, $\exists K$, s.t. $\forall \nu$, $|g_\nu(\theta)| \leq 1 + Kk$. Second, $\exists c_0, c_1 > 0$, such that for all $c_0 \leq |g_\nu(\theta)| \leq 1 + Kk$, $|g_\nu(\theta)|$ is a simple root; and for any other root $g_\mu(\theta)$, it holds that $|g_\mu(\theta) - g_\nu(\theta)| \geq c_1$ for h, k sufficiently small.*

Next, we consider the stability for a system of partial differential equations. We consider

$$\underline{u}_t = B\underline{u}_{xx}. \tag{1.151}$$

The numerical stability is again based on the well-posedness of the partial differential equations. To check the well-posedness, we perform a dispersion relation analysis with the form of solution

$$\underline{u} = \underline{U} e^{\lambda t + i\omega x}. \tag{1.152}$$

Substituting this into (1.151), we obtain

$$\lambda \underline{U} = -B\omega^2 \underline{U}. \tag{1.153}$$

Therefore, the growing rate is an eigenvalue to the linear system

$$\lambda I + \omega^2 B = 0. \tag{1.154}$$

To make Re$\lambda \leq 0$, we require B to be positive-definite. In particular, a special case is $B = diag(b_1, \cdots, b_n)$ with $b_i \geq 0$.

For a positive-definite matrix B, we consider the solution in the form of

$$\underline{u}_m^n = \underline{U}^n e^{i\omega x}. \tag{1.155}$$

Then a numerical scheme leads to an amplification factor defined by

$$\underline{U}^{n+1} = G\underline{U}^n, \tag{1.156}$$

where G is a matrix. Numerical solution

$$\underline{U}^n = G^n\underline{U}^0. \tag{1.157}$$

Obviously, stability holds if $\parallel G^n \parallel \leq C_T$. It is therefore necessary that $\rho(G) \leq 1 + Kk$. Yet it is not enough to maintain stability in general.

If the G matrix has two distinct eigenvalues, then it may diagonalized and the condition $\rho(G) \leq 1$ guarantees the stability. However, if a double eigenvalue is the case, then we notice that

$$\begin{bmatrix} \alpha & \beta \\ 0 & \beta \end{bmatrix}^n = \begin{bmatrix} \alpha^n & n\alpha^{n-1}\beta \\ 0 & \alpha^n \end{bmatrix}. \tag{1.158}$$

For example, a first order scheme for the system

$$\begin{cases} u_t^1 = u_{xx}^2, \\ u_t^2 = 0, \end{cases} \tag{1.159}$$

may be taken as follows (assume $k = h$)

$$\begin{cases} u_m^{1,n+1} = u_m^{1,n} - (u_{m+1}^{2,n} - 2u_m^{2,n} + u_{m-1}^{2,n}), \\ u_m^{2,n+1} = u_m^{2,n}. \end{cases} \tag{1.160}$$

The amplification factor matrix is

$$G = \begin{pmatrix} 1 & 4\sin^2\frac{\theta}{2} \\ 0 & 1 \end{pmatrix}. \tag{1.161}$$

The spectral radius is $\rho(G) = 1$. However, instability occurs as we have

$$G^n = \begin{pmatrix} 1 & 4n\sin^2\frac{\theta}{2} \\ 0 & 1 \end{pmatrix}. \tag{1.162}$$

For another example, we consider a two-variable system.

$$\begin{cases} v_t = b_{11}v_{xx} + b_{12}w_{xx}, \\ w_t = b_{21}v_{xx} + b_{22}w_{xx}. \end{cases} \tag{1.163}$$

Consider an explicit central difference scheme

$$\begin{cases} \dfrac{v_m^{n+1}-v_m^n}{k} = b_{11}\dfrac{v_{m-1}^n - 2v_m^n + v_{m+1}^n}{h^2} + b_{12}\dfrac{w_{m-1}^n - 2w_m^n + w_{m+1}^n}{h^2}, \\ \dfrac{w_m^{n+1}-w_m^n}{k} = b_{21}\dfrac{v_{m-1}^n - 2v_m^n + v_{m+1}^n}{h^2} + b_{22}\dfrac{w_{m-1}^n - 2w_m^n + w_{m+1}^n}{h^2}. \end{cases} \tag{1.164}$$

Substituting the form of the amplification factor expression

$$\begin{bmatrix} v^{n+1} \\ w^{n+1} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \begin{bmatrix} v^n \\ w^n \end{bmatrix}, \tag{1.165}$$

we find that

$$\begin{cases} (g_{11}-1)v^n + g_{12}w^n = b_{11}\mu(-4\sin^2\frac{\theta}{2})v^n + b_{12}\mu(-4\sin^2\frac{\theta}{2})w^n, \\ g_{21}v^n + (g_{22}-1)w^n = b_{21}\mu(-4\sin^2\frac{\theta}{2})v^n + b_{22}\mu(-4\sin^2\frac{\theta}{2})w^n. \end{cases} \tag{1.166}$$

Therefore, the amplification factor matrix is

$$G = \begin{bmatrix} 1-4b_{11}\mu\sin^2\frac{\theta}{2} & -4b_{12}\mu\sin^2\frac{\theta}{2} \\ -4b_{21}\mu\sin^2\frac{\theta}{2} & 1-4b_{22}\mu\sin^2\frac{\theta}{2} \end{bmatrix} = I - 4\mu\sin^2\frac{\theta}{2}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}. \tag{1.167}$$

1.7 Multi-dimensional Diffusion Equation

In the previous sections, we studied schemes in one-space dimensions. Now we consider the heat equation in two space dimensions.

$$u_t = b_{11}u_{xx} + 2b_{12}u_{xy} + b_{22}u_{yy}. \tag{1.168}$$

Here $b_{11}, b_{12} > 0$, and $b_{12}^2 \le b_{11}b_{22}$. After a change of coordinates, the B matrix may be diagonalized. The equation then becomes

$$u_t = \widetilde{b}_{11}u_{\widetilde{x}\widetilde{x}} + \widetilde{b}_{22}u_{\widetilde{y}\widetilde{y}}. \tag{1.169}$$

Here $\widetilde{b}_{11}, \widetilde{b}_{12} > 0$. In the following, we omit the tildes and still denote the equation as

$$u_t = b_{11}u_{xx} + b_{22}u_{yy}. \tag{1.170}$$

1.7.1 Time Splitting

In a square uniform grid with mesh size h_x, h_y in the two dimensions, respectively, the simplest explicit scheme reads as follows.

$$\frac{u_{l,m}^{n+1} - u_{l,m}^n}{k} = b_{11}\frac{u_{l-1,m}^n - 2u_{l,m}^n + u_{l+1,m}^n}{h_x^2} + b_{22}\frac{u_{l,m-1}^n - 2u_{l,m}^n + u_{l,m+1}^n}{h_y^2}. \tag{1.171}$$

Here $u_{l,m}^n \sim u(x_l, y_m, t^n)$. The scheme may be slightly modified as follows. First, we consider the equation and scheme

$$\begin{cases} u_t = b_{11} u_{xx}, \\ u^\star = P_x(k) u^n, \end{cases} \tag{1.172}$$

where P_x denotes an advancing for one time-step k. Next, we consider the equation and scheme

$$\begin{cases} u_t = b_{22} u_{yy}, \\ u^{n+1} = P_y(k) u^\star, \end{cases} \tag{1.173}$$

where P_y denotes an advancing for one time-step k.

If we take explicit centered difference schemes in both dimensions, then the resulted scheme is as follows.

$$\begin{cases} u_{l,m}^\star = (1 - 2b\mu_x) u_{l,m}^n + b\mu_x (u_{l-1,m}^n + u_{l+1,m}^n), \\ u_{l,m}^{n+1} = (1 - 2b\mu_y) u_{l,m}^\star + b\mu_y (u_{l,m-1}^\star + u_{l,m+1}^\star). \end{cases} \tag{1.174}$$

Here $\mu_x = \frac{k}{h_x^2}, \mu_y = \frac{k}{h_y^2}$. It may also be formally written as

$$u^{n+1} = P_y(k) P_x(k) u. \tag{1.175}$$

However, it is evident that this scheme differs from $u^{n+1} = P_x(k) P_y(k) u$, and both time-splitting methods introduce error on the first order $O(k)$. In the explicit scheme, the error is first order in time even for one space dimension. Therefore, this time splitting is reasonably satisfactory. However, if a second order scheme is used, e.g., the Crank-Nicolson scheme or the DuFort-Frankel scheme, then the above splitting is not compatible and reduces the accuracy order for the overall scheme. To solve this problem, a Strang-splitting technique is used, which yields a second order accuracy. More precisely, if we have second order schemes in time

$$u_{n+1} = P_x(k)(u_n, u_{n-1}), \tag{1.176}$$

for one-step update in the x-dimension, and

$$u_{n+1} = P_y(k)(u_n, u_{n-1}), \tag{1.177}$$

for one-step update in the y-dimension, then we may take

$$u^{n+1} = P_x(\frac{k}{2}) P_y(k) P_x(\frac{k}{2})(u_n, u_{n-1}). \tag{1.178}$$

1.7.2 ADI Method on a Square

Consider the Crank-Nicolson scheme in two space dimensions

$$\begin{aligned}\frac{u^{n+1}_{l,m}-u^n_{l,m}}{k} &= \frac{1}{2}b_{11}\frac{u^n_{l+1,m}-2u^n_{l,m}+u^n_{l-1,m}}{h_x^2}+\frac{1}{2}b_{12}\frac{u^{n+1}_{l+1,m}-2u^{n+1}_{l,m}+u^{n+1}_{l+1,m}}{h_x^2}\\ &+\frac{1}{2}b_{22}\frac{u^n_{l,m+1}-2u^n_{l,m}+u^n_{l,m-1}}{h_y^2}+\frac{1}{2}b_{21}\frac{u^{n+1}_{l,m+1}-2u^{n+1}_{l,m}+u^{n+1}_{l,m-1}}{h_y^2}\end{aligned} \tag{1.179}$$

Let $u^n_m = (u^n_{0,m}, \cdots, u^n_{l,m}, \cdots, u^n_{L,m})$ be the value at $y = mh, t = nk$, and $u^n = (u^n_0, \cdots, u^n_m, \cdots, u^n_M)^T$ be the grid function at $t = nk$. The Crank-Nicolson scheme may be rewritten as

$$\frac{\underline{u}^{n+1}-\underline{u}^n}{k} = \frac{1}{2}(A_1u^{n+1}+A_1u^n)+\frac{1}{2}(A_2u^{n+1}+A_2u^n)+O(k^2). \tag{1.180}$$

It is readily shown to be second order in both space and time, when the Taylor expansion is performed at $t = (n+\frac{1}{2})k, x = lh_x, y = mh_y$. Here A_1 is a tridiagonal matrix, and A_2 is a blocked tri-diagonal matrix. The inversion of the Crank-Nicolson scheme is involved.

First, we rewrite the scheme as

$$(I-\frac{k}{2}A_1-\frac{k}{2}A_2)u^{n+1} = (I+\frac{k}{2}A_1+\frac{k}{2}A_2)u^n+O(k^3). \tag{1.181}$$

Next, we purposely insert two terms on both hand sides.

$$(I-\frac{k}{2}A_1-\frac{k}{2}A_2+\frac{k^2}{4}A_1A_2)u^{n+1} = (I+\frac{k}{2}A_1+\frac{k}{2}A_2+\frac{k^2}{4}A_1A_2)u^n+O(k^3). \tag{1.182}$$

We decompose both sides to derive

$$(I-\frac{k}{2}A_1)(I-\frac{k}{2}A_2)u^{n+1} = (I+\frac{k}{2}A_1)(I-\frac{k}{2}A_2)u^n+O(k^3). \tag{1.183}$$

This means

$$u^{n+1} = (I-\frac{k}{2}A_2)^{-1}(I-\frac{k}{2}A_1)^{-1}(I+\frac{k}{2}A_1)(I+\frac{k}{2}A_2)u^n. \tag{1.184}$$

We notice that $(I-\frac{k}{2}A_1)^{-1}$ and $(I+\frac{k}{2}A_1)$ commutes. That is,

$$u^{n+1} = (I-\frac{k}{2}A_2)^{-1}(I+\frac{k}{2}A_1)(I-\frac{k}{2}A_1)^{-1}(I+\frac{k}{2}A_2)u^n. \tag{1.185}$$

This leads to the Peaceman-Rachford algorithm.

$$\begin{cases}(I-\frac{k}{2}A_1)u^{n+1/2} = (I+\frac{k}{2}A_2)u^n,\\ (I-\frac{k}{2}A_2)u^{n+1} = (I+\frac{k}{2}A_1)u^{n+1/2}.\end{cases} \tag{1.186}$$

We remark that each equation in this algorithm involve only the inversion of the tri-diagonal matrix. The computing load is greatly reduced. The whole method is called an ADI (Alternately directional implicit) method.

Remark 1.3. *We notice that $u^{n+1/2}$ may be not a good approximation to u at $t = (n + \frac{1}{2})k$ in general.*

An important issue is the numerical boundary condition. In fact, the partial differential equation takes boundary conditions, which automatically impose conditions for u^n and u^{n+1} at the boundary. However, the boundary condition for $u^{n+1/2}$ is not obvious.

Adding the two equations in (1.186)), we easily obtain $2u^{n+1/2} = (I - \frac{k}{2}A_1)u^{n+1} + (I + \frac{k}{2}A_2)u^n$. With this, we determine the boundary condition for $u^{n+1/2}$.

Finally, we make a stability analysis for the ADI algorithm. To this end, we first assume the form of solution as

$$u^n_{l,m} = U^n e^{i(l\theta+m\phi)}. \tag{1.187}$$

For the first step in the Peaceman-Rachford algorithm we assume an amplification factor $\widetilde{g}$ with which

$$\begin{aligned} u^{n+1/2}_{l,m} &= U^{n+1/2}e^{i(l\theta+m\phi)} \\ &\equiv \widetilde{g}U^n e^{i(l\theta+m\phi)}. \end{aligned} \tag{1.188}$$

For the second step, we have an amplification factor g with which

$$\begin{aligned} u^{n+1}_{l,m} &= U^{n+1}e^{i(l\theta+m\phi)} \\ &\equiv gU^{n+1/2}e^{i(l\theta+m\phi)} \\ &= g\widetilde{g}U^n e^{i(l\theta+m\phi)}. \end{aligned} \tag{1.189}$$

As we use the central difference in space, it is easy to obtain that

$$\begin{cases} (1 + 4b_1\mu_1 \sin^2\frac{\theta}{2})\widetilde{g} = 1 - 4b_2\mu_2\sin^2\frac{\phi}{2}, \\ (1 + 4b_2\mu_2 \sin^2\frac{\phi}{2})g = 1 - 4b_1\mu_1\sin^2\frac{\theta}{2}. \end{cases} \tag{1.190}$$

Hence, we obtain the amplification factor

$$g\widetilde{g} = \frac{(1 - 4b_1\mu_1\sin^2\frac{\theta}{2})(1 - 4b_2\mu_2\sin^2\frac{\phi}{2})}{(1 + 4b_1\mu_1\sin^2\frac{\theta}{2})(1 + 4b_2\mu_2\sin^2\frac{\phi}{2})}. \tag{1.191}$$

It is obvious that $|g\widetilde{g}| \leq 1$. Therefore, the ADI algorithm is unconditionally stable.

Assignments

1. Show that the Cole-Hopf transform

$$u = -2b\frac{\varphi_x}{\varphi}, \tag{1.192}$$

transforms the Burgers' equation

$$u_t + uu_{xx} = bu_{xx}. \tag{1.193}$$

into a heat equation

$$\varphi_t = b\varphi_{xx}. \tag{1.194}$$

2. Prove that $\frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{i\omega(x-y)} e^{-b\omega^2 t} \mathrm{d}\omega = \frac{1}{\sqrt{bt}} e^{-(x-y)^2/4bt}$.

3. Find the order of accuracy for the scheme

$$\frac{u_m^{n+1} - u_m^{n-1}}{2k} = \frac{u_m^n - 2u_{m-1}^n + u_{m-2}^n}{h^2}.$$

4. For the explicit central difference scheme

$$\frac{u_m^{n+1} - u_m^n}{k} = \frac{u_{m-1}^n - 2u_m^n + u_{m+1}^n}{h^2},$$

it is easy to find that the modified equation is

$$u_t + \frac{k}{2} u_{tt} - b(u_{xx} + \frac{h^2}{12} u_{xxxx}) = 0.$$

Noticing that $u_{tt} = b^2 u_{xxxx}$, the accuracy may be improved if we take $k = \dfrac{h^2}{6b}$. Verify this by numerical computations.

5. Find a sufficient condition for an implicit scheme to be stable.

$$\frac{u_m^{n+1} - u_m^n}{k} = b\frac{u_{m-1}^{n+1} - 2u_m^{n+1} + u_{m+1}^{n+1}}{h^2}. \tag{1.195}$$

6. Perform the von Neumann analysis to determine stability for the Crank-Nicolson scheme.

7. For a multi-step scheme, e.g., the leap-frog scheme, we perform the von Neumann analysis with equation (1.84), and solve it to get the amplification factor $g_\pm$. The question is: what do we mean by g_+ and g_- in this application? *(Hint: Find the eigen-function corresponding to g_+ and g_- respectively.)*

8. For

$$u_t = u_{xx}, u_0(x) = \begin{cases} 1, & x \in [-1, 1], \\ 0, & \text{elsewhere}, \end{cases} \tag{1.196}$$

compute with the explicit central-difference scheme under the following boundary conditions, respectively.

- Dirichlet boundary condition

$$u_0(t,-A) = u(t,A) = 0. \tag{1.197}$$

- Neumann boundary condition

$$u_x(t,-A) = u_x(t,A) = 0. \tag{1.198}$$

  There are two different ways to discretize this boundary condition, namely, either

$$\begin{cases} u_N(t) = u_{N-1}(t), \\ u_N(t) = u_{-N+1}(t), \end{cases} \tag{1.199}$$

  or

$$\begin{cases} u_{N+1}(t) = u_{N-1}(t), \\ u_{N+1}(t) = u_{-(N-1)}(t). \end{cases} \tag{1.200}$$

- Play with different choices of (h,k), e.g., $k_n = \frac{h_n^2}{4}$ and $h_n = h_0/2^n$.
- Error analysis (numerical convergence rate). Let the solution with the finest grid h_N be u_{exact}. Calculate error in the numerical solution $u^{(n)}$ by a coarser gird h_n as follows.

$$E_n = [\sum_m h_n(u_m^{(n)} - u_{exact}(x_m))^2]^{1/2}. \tag{1.201}$$

  Plot $\lg(E_n)$ versus $\lg(h_n)$, and find the slope. This slope is called the numerical convergence rate.

9. Compute

$$\begin{cases} u_t = & u_{xx}, \\ u(x,0) = & e^{-x^2}. \end{cases} \tag{1.202}$$

   with the approximate integral method for $p = 1, 2$.

10. Perform numerical tests to check the numerical stability for the explicit upwind scheme (1.146) for the Burgers' equation. Take initial data from the exact solution (1.145). Assign a proper boundary condition, e.g., $u = 0$. Check that the scheme is stable if the stability condition (1.147) holds; and otherwise (longer time step size) it is unstable. *(Notice that this is a nonlinear problem, our stability condition may not be as straightforward as for the linear problems. Moreover, here the stability only means that around the given exact solution (1.146).)*

11. Show that the inhomogeneous equation

$$u_t = A_1 u + A_2 u + f(x,y,t)$$

can be approximated to second order accuracy by

$$\begin{aligned}(I-\frac{k}{2}A_{1h})\widetilde{u}^{n+1} &= (I+\frac{k}{2}A_{2h})u^n+\frac{k}{2}f^{n+1/2},\\ (I-\frac{k}{2}A_{2h})u^{n+1} &= (I+\frac{k}{2}A_{1h})\widetilde{u}^{n+1/2}+\frac{k}{2}f^{n+1/2},\end{aligned}$$

where $f^{n+1/2}$ is the vector formed by $f(x_l, y_m, t^{n+1/2}$.

12. Realize the ADI algorithm in a square.

Chapter 2

Finite Volume Method for Hyperbolic Equations

Hyperbolic differential equations are of great importance in sciences and engineering, particularly when wave phenomena are under consideration. Roughly speaking, most practical physical systems are diffusive hence of parabolic type essentially. Their steady states, typically describing the asymptotic behaviors, are of elliptic type. The wave phenomena, under a proper scaling, are often better described by neglecting the diffusions and the systems become hyperbolic.

2.1 Recap of Hyperbolic Partial Differential Equations

2.1.1 Linear Wave Equation

The most well-known hyperbolic partial differential equation is the linear wave equation, which describes free wave propagation in a homogeneous medium.

$$u_{tt} - c^2 u_{xx} = 0. \tag{2.1}$$

By the d'Alembert principle, the solution may be expressed in terms of a left-going wave and a right-going wave.

$$u(x,t) = \phi(x+ct) + \psi(x-ct), \tag{2.2}$$

where the form of ϕ and ψ are determined by initial data of the Cauchy Problem. For an initial-boundary-value problem, boundary conditions need to be incorporated properly. For details, please refer to the appendix of this chapter.

In general, a boundary condition is posed only along a line with inward characteristic. For instance, for a linear advection equation

$$u_t + cu_x = 0, \tag{2.3}$$

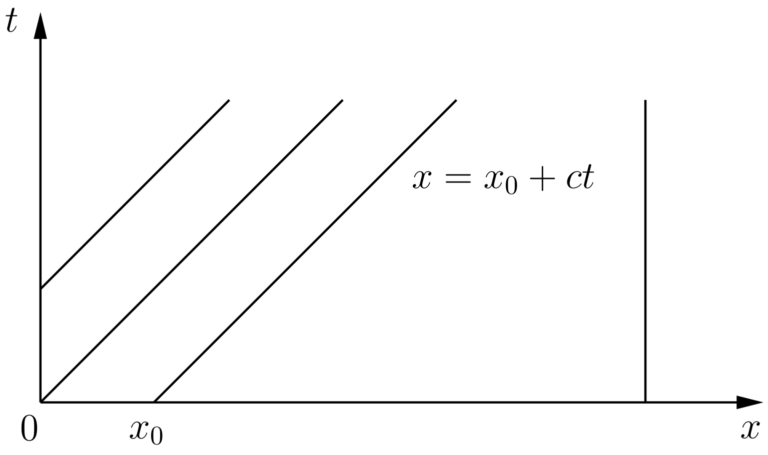


Figure 2.1: Boundary conditions for linear advection equation.

in a finite domain $(x,t) \in [0,1] \times \mathbb{R}^+$ with $c > 0$, we impose a boundary condition at $x = 0$. No boundary condition is needed at $x = 1$, see Figure 2.1.

A characteristic curve/line is one of the major tool in studying wave propagations in a hyperbolic system. For the linear advection equation, for instance, along a straight line defined by

$$x = x_0 + ct, \tag{2.4}$$

It holds that

$$\frac{\mathrm{d}u}{\mathrm{d}t}\Big|_{\frac{\mathrm{d}x}{\mathrm{d}t}=c} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = 0. \tag{2.5}$$

This implies that u maintains constant along this line, that is, $u(x,t) = u(x_0, 0)$. This line is called as a characteristic line. This line brings the information from the initial data. In more general cases, as we shall see for a nonlinear hyperbolic system later, the information may propagate in a curve, which is called as a characteristic curve.

The direct consequence of the discussion for characteristic curve is the finite propagation speed in a hyperbolic system. We recall that the heat diffusion propagate with infinite speed.

Due to the finite propagation speed, the solution $u(x,t)$ depends only on previous information within a finite range in the (x,t)-plane. This defines a *domain of dependence.*

Meanwhile, the information at (x,t) only influences a limited subdomain in the future, which defines a *range of influence.*

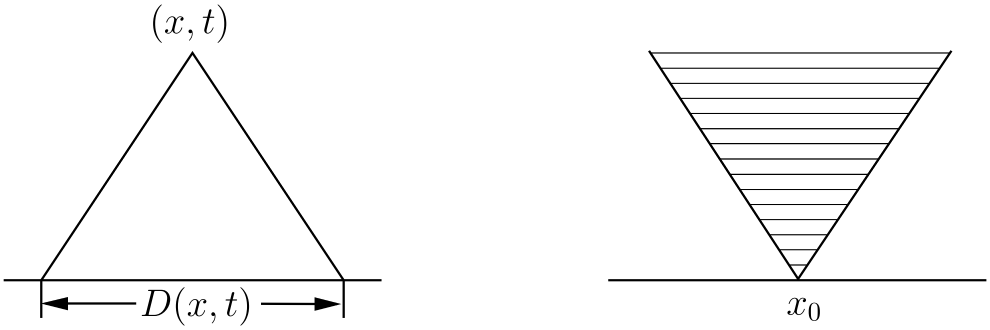


Figure 2.2: Domain of dependence and range of influence.

2.1.2 A Function Space

The most distinct feature in hyperbolic partial differential equations is discontinuity. The generation and propagation of a spatially discontinuous wave require discussions in a suitable function space, which should be "bigger" than the continuous function space such as C^1. It turns out that a total variation bounded (TVB) space is a suitable choice. We remark that for two space dimensions, the suitable function space is not discovered yet.

The total variation for a function $f(x)$ is defined as follows.

$$TV(f) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{\mathbb{R}} |f(x) - f(x - \epsilon)| \mathrm{d}x. \tag{2.6}$$

In our numerical studies, the numerical solution is usually interpreted as a piecewise constant function, that is, a grid function. For such a function, the total variation is equivalent to

$$TV(f) = \sum_i |f_i - f_{i-1}|. \tag{2.7}$$

We remark that total variation does not define a norm. In fact, $TV(f) = 0$ only leads to $f(x) = C$ where C may be non-zero.

It is straightforward to see that $TV(u(x,t))$ remains unchanged for a linear advection equation. It is very important that $TV(u(x,t))$ is non-increasing for a nonlinear equation in general.

2.1.3 Linear System

A direct generalization for a linear advection equation is a linear system. Consider for $\underline{u}$ a vector-valued function, which is governed by

$$\underline{u}_t + A\underline{u}_x = 0. \tag{2.8}$$

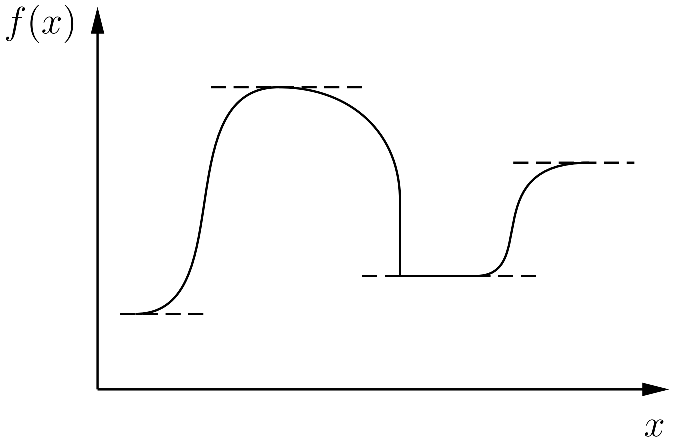


Figure 2.3: Total variation.

If A is diagonalizable with complete eigen-vectors, then we collect all the left eigenvectors to form a matrix P, namely, $PAP^T = \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_N)$. Let $\underline{w} = P\underline{u}$, we find that

$$\underline{w}_t + \Lambda \underline{w}_x = 0. \tag{2.9}$$

Therefore, we obtain a decoupled system. For each $w_i(x,t)$, the wave propagation is governed by a linear advection equation

$$\partial_t w_i + \lambda_i \partial_x w_i = 0. \tag{2.10}$$

We remark that a special case which guarantees A to be diagonalizable is when all the eigenvalues of A are distinct.

2.1.4 Nonlinear Conservation Laws: The Difficulties

For a scalar nonlinear conservation law

$$q_t + f(q)_x = 0, \tag{2.11}$$

we further require a convex flux, that is

$$f^{''}(q) > 0. \tag{2.12}$$

A representative model is the inviscid Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0. \tag{2.13}$$

For a classical solution $u(x,t) \in C^1$, it may be rewritten as

$$u_t + uu_x = 0. \tag{2.14}$$

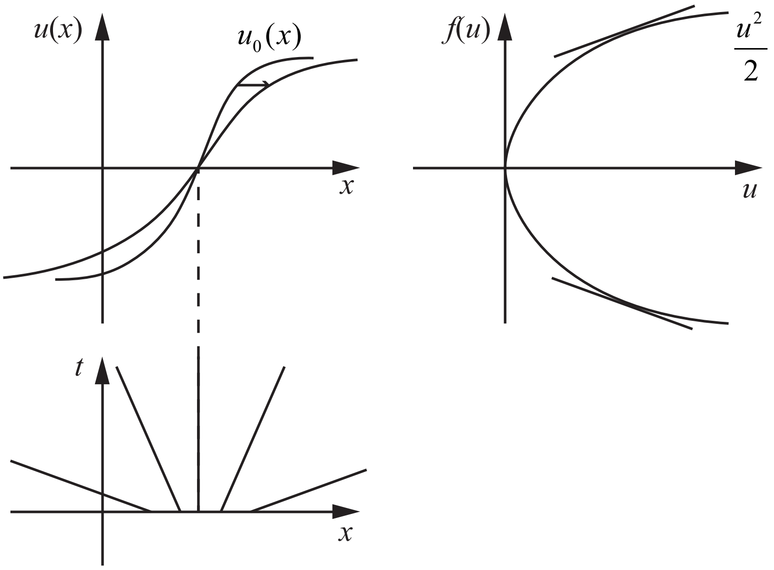


Figure 2.4: Monotone increasing initial data: rarefaction.

In this form, we may apply the characteristic approach. Consider a characteristic curve defined implicitly by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = u(x,t), \quad x(0) = x_0. \tag{2.15}$$

Along this curve, again we may compute

$$\frac{\mathrm{d}u}{\mathrm{d}t}|_{\frac{\mathrm{d}x}{\mathrm{d}t}=u} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = 0. \tag{2.16}$$

This means that if a solution lies in C^1, then u keeps constant along each characteristic curve. Moreover, as u is a constant, the characteristic curve is actually a straight line. Then the study of the equation becomes: for any (x,t), does there exist a unique characteristic line going back to the x-axis. If this is true, we simply take the corresponding initial data to get the value for $u(x,t)$.

This leads to a geometric approach. We plot simultaneously three plots, namely, the flux function in the $(q, f(q))$-plane, the solution in (x,u)-plane, and the characteristic lines in (x,t)-plane. Consider monotone initial data. There are two possibilities.

- If the initial data is monotone increasing (and C^1), then exist unique solution in C^1.

- If the initial data is monotone decreasing (and C^1), then the characteristic lines intersect in finite time. At this time, the solution becomes

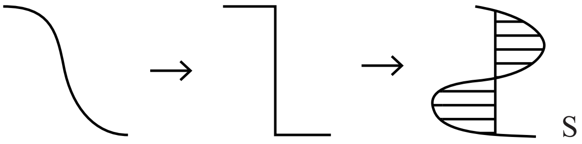


Figure 2.5: Monotone decreasing initial data: shock.

discontinuous as $u_x \to +\infty$. If one keeps going with the characteristic approach beyond this time, the solution flips over and becomes multi-valued.

For the first case, the wave profile gets flattened, see Figure 2.4.

For the second case, the discontinuity propagates at a speed determined by u across the shock front, see Figure 2.5.

For general initial data, both monotone increasing and monotone decreasing branches exist, and discontinuity develops for whatever smooth initial data.

In studying nonlinear hyperbolic equations, there are two special cases, which serve as the elementary waves. For the inviscid Burgers' equation, they are explicitly expressed as follows.

- Centered refraction wave $u_- < u_+$:

$$u(x,t) = \begin{cases} u_-, & \text{if } x < u_- t; \\ x/t, & \text{if } u_- t < x < u_+ t; \\ u_+, & \text{if } x > u_+ t. \end{cases} \tag{2.17}$$

- Shock wave $u_- > u_+$:

  1. The Rankine-Hugoniot relation determines the propagation speed. It is actually an integral form of the equation. It may be obtained formally by substituting ∂t by $-s$ where s is the shock front propagation speed, and then put a jump sign defined by $[\![f]\!] = f_+ - f_-$. That is, we have

  $$-s[\![u]\!] + [\![\frac{u^2}{2}]\!] = 0. \tag{2.18}$$

  From this we derive

  $$s = \frac{u_- + u_+}{2}. \tag{2.19}$$

  2. The Lax entropy condition $u_- > u_+$ arises from the intersection of the two characteristics across the shock front. In particular, if we draw the two characteristic lines with an arrow pointing along the time evolving direction, we have the so-called "2-in-1-out" situation. That is, both characteristic lines point toward the wave front.

The solution therefore reads as follows.

$$u(x,t) = \begin{cases} u_-, & \text{if } x < st; \\ u_+, & \text{if } x > st. \end{cases} \tag{2.20}$$

2.1.5 Nonlinear System

For a nonlinear system it is more complicated. For instance, for the polytropic gas in the Lagrangian coordinates, we have

$$\begin{cases} v_t - u_x = 0, \\ u_t + p(v)_x = 0. \end{cases} \tag{2.21}$$

The Riemann problem is solved with left-going waves R_-, S_-, and right-going waves R_+, S_+. Here R_-, R_+ are central refraction, and S_-, S_+ are shocks.

In the phase space, the left-going wave and the right-going wave give

$$\begin{cases} u_\star = u_- + f_-(v_\star - v_-), \\ u_+ = u_\star + f_+(v_+ - v_\star). \end{cases} \tag{2.22}$$

Combining these two facts, we obtain $v_\star$ from

$$u_+ = u_- + f_+(v_+ - v_\star) + f_-(v_\star - v_-). \tag{2.23}$$

As $u_\star$ is obtained accordingly, the two waves are identified. This solves the Riemann problem.

2.2 Finite Difference Methods for Linear Advection Equation

Consider the linear advection equation

$$u_t + cu_x = 0. \tag{2.24}$$

The following finite difference schemes are readily obtained, among many other possible designs. All schemes listed below are explicit ones. We shall explain why implicit ones are not under consideration later.

- Central difference scheme:

$$\frac{u_m^{n+1} - u_m^n}{k} + c\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0. \tag{2.25}$$

- Forward (in space) scheme:

$$\frac{u_m^{n+1} - u_m^n}{k} + c\frac{u_{m+1}^n - u_m^n}{h} = 0. \tag{2.26}$$

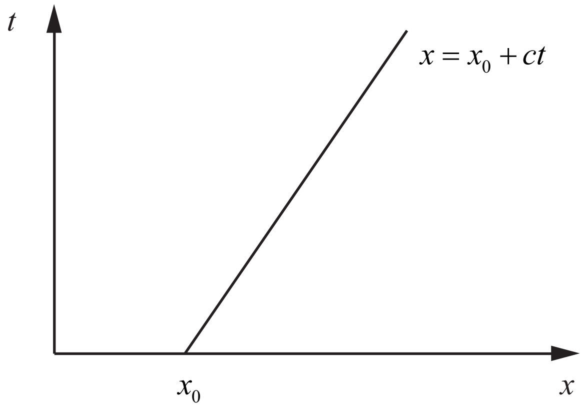


Figure 2.6: Solution for the linear advection equation.

- Backward (in space) scheme:

$$\frac{u_m^{n+1} - u_m^n}{k} + c\frac{u_m^n - u_{m-1}^n}{h} = 0. \tag{2.27}$$

- Lax-Friedrichs scheme:

$$\frac{u_m^{n+1} - \frac{u_{m+1}^n + u_{m-1}^n}{2}}{k} + c\frac{u_{m+1}^n - u_{m-1}^n}{2h} = 0. \tag{2.28}$$

If $c > 0$, the exact solution is $u(x,t) = u_0(x - ct)$, where $u_0(x)$ is the initial data, see Figure 2.6. From this exact solution, we notice that the forward scheme has no chance to be correct, because it takes information from the wrong side, and the solution is likely to contain discontinuities. This differs from the situation for the heat equation, where we are quite free in choosing the stencil to reproduce the derivatives because the solution is smooth in general.

We demonstrate the instability by a very simple argument. Let us think about initial data of a Heaviside function

$$u_0(x) = H(x) = \begin{cases} 0, & \text{for } x < 0, \\ 1, & \text{for } x > 0. \end{cases} \tag{2.29}$$

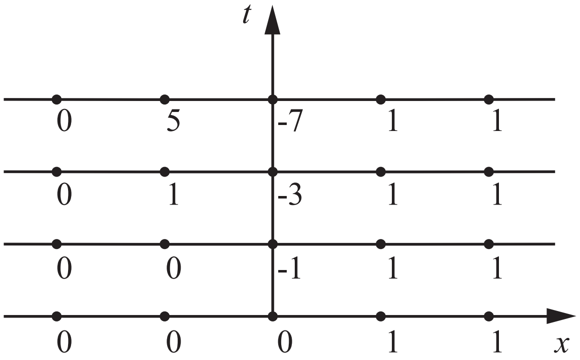
Moreover, let $ck/h = 1$. This reduces the forward scheme into

$$u_m^{n+1} = 2u_m^n - u_{m+1}^n. \tag{2.30}$$

It is evident that $u_m^n = 1$ for all $m \geq 1$. This gives, in turn, that

$$u_0^n = -(2^n - 1), \quad \text{for } n > 1. \tag{2.31}$$

Instability appears also for other grid points to the left. See Figure 2.7 for an illustration.



0	5	-7	1	1
0	1	-3	1	1
0	0	-1	1	1
0	0	0	1	1

Figure 2.7: Instability for the forward scheme.

In the meantime, the backward scheme with $ck/h = 1$ becomes

$$u_m^{n+1} = u_{m-1}^n. \tag{2.32}$$

This actually agrees with the exact solution. However, this does not apply to nonlinear equation in general.

If $c < 0$ instead, the stabilities for the forward and backward schemes change. Therefore, we define a downwind scheme and an upwind scheme instead. If the stencil only consists of grid points from where the information comes from, the upwind direction, we call the scheme upwind. If the stencil only consists of grid points from the other side, we call the scheme downwind. Therefore, the upwind scheme is a forward one if $c < 0$, and a backward one if $c > 0$.

The upwind scheme is stable, under certain restrictions of the time step size. We prove this by another type of stability analysis, namely, through the modified equation.

As we shall see the reason, it is assumed that $k \sim h$. Taylor expansions lead to the following form for the upwind scheme (the backward scheme with $c > 0$).

$$u_t + \frac{k}{2}u_{tt} + c(u_x - \frac{h}{2}u_{xx}) + O(h^2) = 0. \tag{2.33}$$

Noticing that this implies that

$$u_t = -cu_x + O(k), \tag{2.34}$$

and

$$u_{tt} = c^2 u_{xx} + O(k). \tag{2.35}$$

Therefore, the numerical scheme actually solves, to the order of $O(h)$, the following equation with $\lambda = ck/h$.

$$\begin{aligned} u_t + cu_x &= -\frac{1}{2}(kc^2 - ch)u_{xx} + O(h^2) \\ &= \frac{ch}{2}(1-\lambda)u_{xx} + O(h^2). \end{aligned} \tag{2.36}$$

We call the resulted partial differential equation as the modified equation.

$$u_t + cu_x = \frac{ch}{2}(1-\lambda)u_{xx}. \tag{2.37}$$

As a positive diffusion requires $\lambda \leq 1$, this gives the stability condition. Fail to satisfy this condition leads to a negative viscosity, and numerically, a blow up. This restriction on time step size is the famous CFL (Courant-Friedrichs-Lewy) condition.

We may investigate the stability for these schemes by the von Neumann analysis. For the central difference scheme, we compute

$$g = 1 - i\lambda \sin\theta. \tag{2.38}$$

It is always unstable.

For the upwind scheme, we compute

$$g = 1 - \lambda + \lambda e^{-i\theta}. \tag{2.39}$$

Again, we find that it is stable only if $\lambda \leq 1$.

We may compute for the implicit scheme

$$\frac{u_m^{n+1} - u_m^n}{k} + c\frac{u_{m+1}^{n+1} - u_m^{n+1}}{h} = 0. \tag{2.40}$$

The amplification factor reads

$$g = \frac{1}{1 - \lambda + \lambda e^{i\theta}} \tag{2.41}$$

It is unstable even when $\lambda < 1$.

On the other hand, we compute for another implicit scheme

$$\frac{u_m^{n+1} - u_m^n}{k} + c\frac{u_m^{n+1} - u_{m-1}^{n+1}}{h} = 0. \tag{2.42}$$

The amplification factor is

$$g = \frac{1}{1 + \lambda - \lambda e^{i\theta}}. \tag{2.43}$$

This scheme is unconditionally stable, namely, for any λ. Unfortunately, this does not apply to nonlinear problems.

Chapter 3

Finite Volume Method for Scalar Equations

3.1 Direction of Time

Though a scalar conservation law

$$q_t + f(q)_x = 0, \tag{3.1}$$

remains unchanged under the transform $(x,t) \to (-x,-t)$, the more complete physical system includes a diffusion term.

$$q_t + f(q)_x = \varepsilon q_{xx}. \tag{3.2}$$

Symmetry in (x,t) does not hold for this system. As we described before, the hyperbolic partial differential equation is obtained by dropping out the diffusion terms to better capture the wave phenomena. The time direction may be readily shown by a discussion on the entropy pairs.

Consider a classical solution with finite energy to the Burgers' equation

$$u_t + \left(\frac{u^2}{2}\right)_x = \epsilon u_{xx}. \tag{3.3}$$

We multiply the equation by $2u$.

$$(u^2)_t + \left(\frac{2u^3}{3}\right)_x = 2\epsilon u u_{xx}. \tag{3.4}$$

We integrate over $\mathbb{R}$. Noticing that

$$\lim_{x\to\pm\infty} u(x,t) = 0, \tag{3.5}$$

due to the finite total energy, we integrate by part and reach at

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} u^2 \mathrm{d}x = -2\epsilon \int_{\mathbb{R}} u_x^2 \mathrm{d}x. \tag{3.6}$$

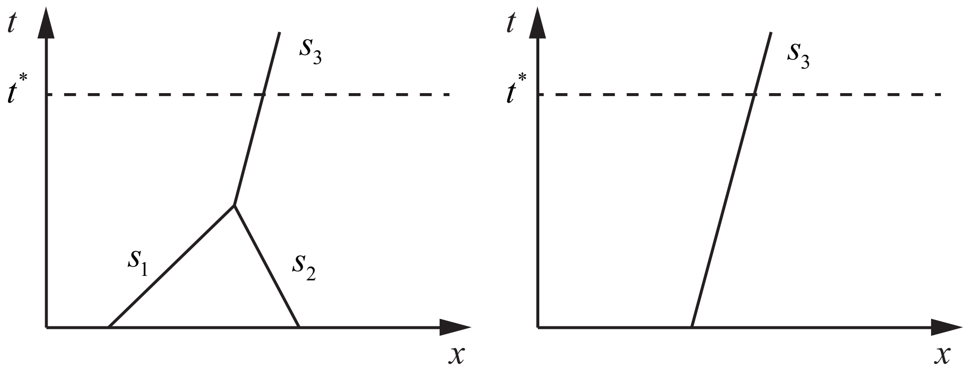


Figure 3.1: Shock generated by the collision of two shocks (left); or by the propagation of one shock (right).

The time direction therefore may be identified by requiring that the integral of the entropy function u^2 decreases. Moreover, we call the corresponding flux function $\dfrac{2u^3}{3}$ as the entropy flux. These two form an entropy pair. While it holds for a classical solution in the limit $\epsilon \to 0+$ that

$$(u^2)_t + \left(\frac{2u^3}{3}\right)_x = 0, \tag{3.7}$$

for a non-classical solution (e.g. a shock solution) we have in a integral sense that

$$(u^2)_t + \left(\frac{2u^3}{3}\right)_x \leq 0. \tag{3.8}$$

Accordingly the time direction is set as the direction for decreasing entropy function, when we drop out the diffusion terms to get the hyperbolic equation (3.1).

In general, hyperbolic problems have a fixed time direction and the structure of solution plays a key role in numerical approximations. For example, in the two subplots of Figure 3.1, we have either a shock generated from the collision of two shocks, or simply propagation of a single shock. If we look back from time $t^* = 2$, we have no idea where this shock structure comes from. This also explains why implicit schemes are not adopted for calculating hyperbolic conservation laws.

We remark that the entropy pairs are not unique for a scalar conservation law with a convex flux function. They are equivalent to each other in the sense of selecting the same discontinuities, and also equivalent to the Lax entropy condition.

3.2 Godunov Method

Godunov first proposed a numerical method capable of capturing shocks correctly in the 1960's. This method was then generalized to the finite volume method. Godunov method is designed with two important basic ideas.

First, a cell averaged view for the grid function is taken. Instead of regard the grid function as the value at a grid point, we define a cell around a grid point x_m as $C_m = [x_{m-1/2}, x_{m+1/2}]$. Then we regard the value u_m^n at (x_m, t^n) as the cell average

$$q_m^n = \frac{1}{h} \int_{C_m} q(x, t^n)\, dx. \tag{3.9}$$

We further define a numerical flux

$$F_{m-1/2}^n = \frac{1}{k} \int_{t^n}^{t^{n+1}} f(q(x_{m-1/2}, t))\, dt. \tag{3.10}$$

Integrating the equation over the domain $C_m \times [t^n, t^{n+1}]$,we obtain

$$h(q_m^{n+1} - q_m^n) + k(F_{m+1/2}^n - F_{m-1/2}^n) = 0. \tag{3.11}$$

The cell average is updated with

$$q_m^{n+1} = q_m^n - \frac{k}{h} \left(F_{m+1/2}^n - F_{m-1/2}^n \right). \tag{3.12}$$

This expression is exact, in contrast to the approximation one makes for a finite difference scheme. The numerical algorithm then transforms into the design of the numerical flux $F_{m-1/2}^n$.

The other basic idea in the Godunov method is the choice of the Riemann solver as building blocks.

This starts with the assumption that the data at t^n is taken as piecewise constant, namely, $u(x, t^n) = u_m^n$ for $x \in C_m$. Due to the finite propagation speed, the exact solution within C_m is determined by $u_{m-1}^n, u_m^n, u_{m+1}^n$, provided that the time step size is chosen to be small enough. Roughly speaking, if the maximal characteristic curve slope $\max |f'(q)| \le C$, the time step size is chosen as $k \le h/2C$. This is more stringent than the CFL condition, and may be relaxed after a modified discussion in general.

It greatly reduces the coding cost as we only need the value at cell-interface to get numerical flux. The detailed Riemann solutions need not be computed.

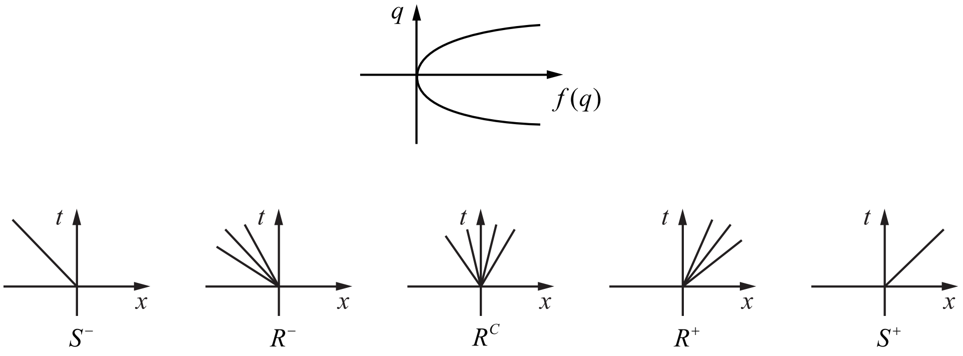


Figure 3.2: Riemann solution to $q_t + f(q)_x = 0$.

For a convex flux function $f(q)$ with $f^{''}(q) \geq 0$, we denote q_s for the unique stagnation value, namely $f'(q_s) = 0$. The numerical flux is

$$
\begin{aligned}
F^n_{m-1/2} = f(q(x_{m-1/2}, t)) &= \begin{cases} f(q^n_{m-1}) & \text{if } q^n_{m-1} > q_s, s > 0, \\ f(q^n_m) & \text{if } q^n_m < q_s, s < 0, \\ f(q_s) & \text{if } q^n_{m-1} < q_s < q^n_m. \end{cases} \\
&= \begin{cases} \min\limits_{q^n_{m-1} \leq q \leq q^n_m} f(q) & \text{if } q^n_{m-1} \leq q^n_m \\ \max\limits_{q^n_m \leq q \leq q^n_{m-1}} f(q) & \text{if } q^n_{m-1} \geq q^n_m \end{cases}
\end{aligned} \tag{3.13}
$$

Here $s = (f(q^n_m) - f(q^n_{m-1}))/(q^n_m - q^n_{m-1})$. We remark that the latter expression applies to non-convex flux function as well. We also observe that if $f'(q)$ does not change sign, then the Godunov scheme reduces to the upwind scheme.

The Godunov method can be generalized to an REA algorithm, which will be adopted extensively in our discussions.

- Reconstruct a piecewise polynomial function $\tilde{q}(x, t_n)$ from the cell averages q^n_m. High resolution methods use polynomials, whereas the Godnov's cell average is a 0-th order polynomial.

- Evolve the solution for a time step, and find either exact or approximate solution at t^{n+1}. This usually means a Riemann solver.

- Average the solution at t^{n+1} in each cell, that is,

$$
q^{n+1}_m = \frac{1}{h} \int_{C_m} q(x, t^{n+1}) \mathrm{d}x. \tag{3.14}
$$

3.3 Conservative Form and Convergence

In an REA approach, the algorithm takes a conservative form, that is, the increment of a cell average is obtained with the numerical fluxes across the cell boundary. We call such a scheme as a conservative scheme. A non-conservative scheme usually lead to wrong propagation speed for a shock wave.

We explain the shock speed by the inviscid Burgers' equation with initial data

$$u(x,0) = \begin{cases} 2, & \text{if } x < 0, \\ 1, & \text{if } x > 0. \end{cases} \tag{3.15}$$

The exact solution is a shock wave.

$$u(x,t) = \begin{cases} 2, & \text{if } x < 3t/2, \\ 1, & \text{if } x > 3t/2. \end{cases} \tag{3.16}$$

We recall that an upwind scheme is in Conservative form.

$$u_m^{n+1} = u_m^n - \frac{k}{h}\left[\frac{(u_m^n)^2}{2} - \frac{(u_{m-1}^n)^2}{2}\right]. \tag{3.17}$$

Its modified equation is

$$u_t + cu_x = \frac{ch}{2}(1-\lambda)u_{xx}. \tag{3.18}$$

It gives a traveling profile for a shock wave with the correct propagation speed, that is, the speed determined by the Burgers' equation. The transition layer has a thickness on the order of $ch(1-\lambda)/2$.

In the meantime, recasting the inviscid Burgers' equation into its primitive form

$$u_t + uu_x = 0, \tag{3.19}$$

it is natural to take a non-conservative scheme as follows.

$$u_m^{n+1} = u_m^n - \frac{k}{h}u_m^n(u_m^n - u_{m-1}^n). \tag{3.20}$$

The difference between these two schemes is clearly seen from the following form.

$$u_m^{n+1} = u_m^n - \frac{k}{h}\left[\frac{(u_m^n)^2}{2} - \frac{(u_{m-1}^n)^2}{2}\right] + \frac{k}{2h}(u_m^n - u_{m-1}^n)^2. \tag{3.21}$$

The modified equation for the non-conservative scheme is

$$u_t + cu_x = \frac{ch}{2}(1-\lambda)u_{xx} + \frac{1}{2}h(u_x)^2. \tag{3.22}$$

As a different traveling wave equation is obtained, even for $h \to 0$, different propagation speed is obtained for a shock profile.

The following theorem relates a conservative scheme with correct shock speed.

Theorem 3.1. *(Lax-Wendroff) Consider a sequence of grids* $\{^{(j)}k, {}^{(j)}h\}$, *with* $\lim\limits_{j\to+\infty} {}^{(j)}k = \lim\limits_{j\to+\infty} {}^{(j)}h = 0$. *Let the numerical solution be denoted as* ${}^{(j)}q = ({}^{(j)}q_m^n)$. *If the scheme is consistent (Lipschitz continuous), conservative, stable and* $\mathcal{S}^{(j)}q \to q$. *Then* q *is a weak solution.*

Remark 3.1. *We note that besides the consistency and convergence, stability is taken as a premise. The schemes under consideration here are nonlinear, for which the Lax-Richtmeyer theorem of equivalence does not apply.*

Before prove the theorem, we first clarify the notions appeared in this theorem.

- By consistency, we require a Lipschitz continuity of the flux function, that is, there exists a constant L such that

$$|F(q_{m-1}, q_m) - f(\bar{q})| \leq L \max(|q_m - \bar{q}|, |q_{m-1} - \bar{q}|). \tag{3.23}$$

- By stability, we assume that for each T, there is an $R > 0$ such that

$$\mathrm{TV}({}^{(j)}q(\cdot, t)) < R, \quad \forall\, 0 \leq t \leq T, \quad j = 1, 2, \cdots . \tag{3.24}$$

  We remark that this stability definition excludes the case when infinite many oscillations occur with finite amplitude. Instability may appear as overflow ($|u| \to +\infty$) for parabolic equations, and oscillation for hyperbolic equations.

- Here $\mathcal{S}^{(j)}q(x,t) = S\{{}^{(j)}q_m^n\}$ denotes a piecewise constant function that takes the value q_m^n on the space-time mesh cell $(x_{m-1/2}, x_{m+1/2}) \times [t^n, t^{n+1})$. It is indexed by j corresponding to the particular mesh used, with ${}^{(j)}h$ and ${}^{(j)}k$ both approaching zero as $j \to \infty$.

- The convergence $S^i(q) \to q$ for the function sequence ${}^{(j)}q(x,t)$ to $q(x,t)$ in the sense that over every bounded set $\Omega = [a,b] \times [0,T]$ in the (x,t)-space, it holds that

$$\int_0^T \int_a^b |\mathcal{S}^{(j)}q(x,t) - q(x,t)|\, dx\, dt \to 0, \quad \text{as } j \to \infty. \tag{3.25}$$

  This is actually the 1-norm over the set Ω, so we can simply write

$$\|\mathcal{S}^{(j)}q - q\|_{1,\Omega} \to 0, \quad \text{as } j \to \infty. \tag{3.26}$$

- The introduction of a weak solution arises as follows. For a classical solution $q(x,t)$ to the equation

$$q_t + f(q)_x = 0, \tag{3.27}$$

we find that for any $\phi(x,t) \in C_0^\infty$ (smooth function with compact support), it holds that

$$\int_{t_1}^{t_2}\int_{-\infty}^{+\infty}[q\phi_t + f(q)\phi_x]\mathrm{d}x\mathrm{d}t = 0. \tag{3.28}$$

We integrate by parts to get

$$\int_{0}^{\infty}\int_{-\infty}^{+\infty}[q\phi_t + f(q)\phi_x]\mathrm{d}x\mathrm{d}t = -\int_{-\infty}^{\infty} q(x,0)\phi(x,0)\mathrm{d}x. \tag{3.29}$$

Because continuous solution breaks down in a nonlinear hyperbolic conservation law in general, we consider a broader sense of solution. That is, if (3.29) holds for any $\phi(x,t) \in C_0^\infty$, then we call q a weak solution.

Proof. We will show that the limit function $q(x,t)$ satisfies the weak form.

Let ϕ be a C_0^∞ test function. On the j-th grid, we define its discrete version $\Phi^{(j)}$ by ${}^{(j)}\phi_m^n = \phi({}^{(j)}x_m, {}^{(j)}t_n)$, where $({}^{(j)}x_m, {}^{(j)}t_n)$ is a grid point on this grid. Similarly, ${}^{(j)}q_m^n$ denotes the numerical approximation on this grid. To simplify notation, we will drop the superscript (j) below and simply use ϕ_m^n and q_m^n, but remember that (j) implicitly presents, and in the end we must take the limits as $j \to \infty$.

Multiply the conservative scheme

$$q_m^{n+1} = q_m^n - \frac{k}{h}(F_{m+1/2}^n - F_{m-1/2}^n) \tag{3.30}$$

by ϕ_m^n to obtain

$$\phi_m^n q_m^{n+1} = \phi_m^n q_m^n - \frac{k}{h}\phi_m^n(F_{m+1/2}^n - F_{m-1/2}^n). \tag{3.31}$$

This is true for all m and n on each grid j. If we sum (3.31) over all m and $n \geq 0$, we obtain

$$\sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty}\phi_m^n(q_m^{n+1} - q_m^n) = -\frac{k}{h}\sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty}\phi_m^n(F_{m+1/2}^n - F_{m-1/2}^n). \tag{3.32}$$

We now sum by parts, which just amounts to recombining the terms in each sum.

$$\sum_{i=1}^{m} a_i(b_i - b_{i-1}) = a_m b_m - a_0 b_0 - \sum_{i=0}^{m-1}(a_{i+1} - a_i)b_i. \tag{3.33}$$

Note that the original sum involved the product of a_m with differences of b's, whereas the final sum involves the product of b_m with differences of a's. This is completely analogous to integration by parts, where the derivatives is moved from one function to the other. Just as in integration by parts, there arise boundary terms $a_m b_m - a_0 b_0$.

We apply this Abel's formula on both sides of (3.32) (for the n-sum on the left and for the m-sum on the right). By our assumption that ϕ has a compact support, then $\phi_m^n = 0$ for $|m|$ or n sufficiently large. Hence the boundary terms at $m = \pm\infty$, $n = \infty$ all drop out. The only boundary term that remains is at $n = 0$ for $t_0 = 0$. This gives

$$-\sum_{m=-\infty}^{\infty} \phi_m^0 q_m^0 - \sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty} (\phi_m^n - \phi_m^{n-1}) q_m^n = \frac{k}{h}\sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty} (\phi_{m+1}^n - \phi_m^n) F_{m-1/2}^n. \tag{3.34}$$

Note that each of these sums is in fact a finite sum, since ϕ has compact support. Multiplying by h and rearranging this equation gives

$$\begin{aligned} hk \Bigg[\sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty} \left(\frac{\phi_m^n - \phi_m^{n-1}}{k}\right) q_m^n \\ + \sum_{n=0}^{\infty}\sum_{m=-\infty}^{\infty} \left(\frac{\phi_{m+1}^n - \phi_m^n}{h}\right) F_{m-1/2}^n \Bigg] = -h \sum_{m=-\infty}^{\infty} \phi_m^0 q_m^0. \end{aligned} \tag{3.35}$$

Now let $j \to \infty$, so that ${}^{(j)}k$, ${}^{(j)}h \to 0$ in (3.35). It is reasonably straightforward, using the 1-norm convergence of ${}^{(j)}q$ to q and the smoothness of ϕ, to show that the term on the top line of (3.35) converge to $\int_0^\infty \int_{-\infty}^\infty \phi_t(x,t) q(x,t)\, dx$ as $j \to \infty$. If we define initial data q_m^0 by taking cell averages of the data $q_0(x)$, for example, then the right-hand side converges to $-\int_{-\infty}^{\infty} \phi(x,0) q(x,0)\, dx$ as well.

The remaining term in (3.35), involving $F_{m-1/2}^n$, is more subtle and requires the additional assumptions on F and q that we have imposed. For a three-point method (such as Godunov's method), we have

$$^{(j)}F_{m-1/2}^n = \mathcal{F}(^{(j)}q_{m-1}^n, {}^{(j)}q_m^n). \tag{3.36}$$

and the consistency condition (3.23), with the choice $\bar{q} = {}^{(j)}q_m^n$, gives

$$|^{(j)}F_{m-1/2}^n - f(^{(j)}q_m^n)| \le L|^{(j)}q_m^n - {}^{(j)}q_{m-1}^n|, \tag{3.37}$$

where L is the Lipschitz constant for the numerical flux function. Since $q^{(j)n}$ has bounded total variation, uniformly in j, it must be that

$$|^{(j)}F_{m-1/2}^n - f(^{(j)}q_m^n)| \to 0, \quad \text{as } j \to \infty. \tag{3.38}$$

for almost all values of m. Using this and the fact that ${}^{(j)}q^n$ converges to q, it can be shown that

$$hk \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} \left(\frac{\phi_{m+1}^n - \phi_m^n}{h} \right) F_{m-1/2}^n \to \int_0^\infty \int_{-\infty}^\infty \phi_x(x,t) f(q(x,t))\, dxdt, \tag{3.39}$$

as $j \to \infty$. This completes the demonstration that (3.35) converges to the weak form (3.29). Since this is true for any test function $\phi \in C_0^1$, we have proved that q is in fact a weak solution. □

For simplicity we assumed the numerical flux $F_{m-1/2}^n$ depends only on the two neighboring values q_{m-1}^n and q_m^n. However, the proof is readily extended to schemes with a wider stencil, under a consistency condition that the flux function is uniformly Lipschitz-continuous in all values on which it depends.

Finally, we make several remarks.

- There usually exist more than one weak solutions. For the inviscid Burger's equation with initial data

$$u(x,0) = \left\{ \begin{array}{ll} 1, & x < 0, \\ 2, & x > 0, \end{array} \right. \tag{3.40}$$

  the following two weak solutions exist.

$$u(x,t) = \left\{ \begin{array}{ll} 1, & x < t, \\ x/t, & t < x < 2t \\ 2, & x > 2t; \end{array} \right. \tag{3.41}$$

$$u(x,t) = \left\{ \begin{array}{ll} 1, & x < \frac{3}{2}t, \\ 2, & x > \frac{3}{2}t. \end{array} \right. \tag{3.42}$$

- The theorem guarantees the correct propagation speed when a numerical shock profile is obtained. However, this does not mean that the numerical shock profile gives a correct solution. The entropy condition needs verification.

- Numerically, we do not really take $j \to \infty$. Instead, we just double grid.

Here we give two examples of conservative schemes. For an upwind scheme of the inviscid Burgers' equation, the numerical flux is

$$F_{m+1/2} = \left\{ \begin{array}{ll} \frac{u_m^2}{2}, & u_m, u_{m+1} > 0, \\ 0, & u_m \cdot u_{m+1} \le 0, \\ \frac{u_{m+1}^2}{2}, & u_m, u_{m+1} < 0. \end{array} \right. \tag{3.43}$$

For a central difference scheme, we have

$$F_{m+1/2} = \frac{u_m^2 + u_{m+1}^2}{2}. \tag{3.44}$$

3.4 Entropy Condition

As discussed in the previous section, the Lax-Wendroff theorem does not guarantee a weak solution to satisfy the entropy condition. From the theory of hyperbolic conservation laws, we know that a "correct" solution should satisfy an entropy condition. In a sense, the limiting behavior for a more complete diffusive system

$$q_t + f(q)_x = \epsilon q_{xx}, \tag{3.45}$$

as $\epsilon \to 0+$ is described by the combination of the weak form of equation

$$q_t + f(q)_x = 0, \tag{3.46}$$

and an entropy condition.

Three kinds of entropy conditions are commonly used.

- **Lax entropy condition.** For a convex scalar conservation law, a discontinuity propagating at speed s (given by the Rankine-Hugoniot relation) satisfies
$$f'(q_l) > s > f'(q_r). \tag{3.47}$$

- **Oleinik entropy condition.** There exists a constant $E > 0$ such that for all $a > 0$, $t > 0$, and $x \in \mathbb{R}$, it holds that
$$\frac{q(x+a,t) - q(x,t)}{a} < \frac{E}{t}. \tag{3.48}$$

- **Entropy pair.** The last approach defines an *entropy function* $\eta(q)$, motivated by thermodynamic considerations in gas dynamics. The entropy function $\eta(q)$ is convex in q, i.e., $\eta^{''}(q) > 0$. Along with an entropy function $\eta(q)$, we define an entropy flux $\Psi(q) = \displaystyle\int \eta'(q) f'(q)\, \mathrm{d}q$. For a classical solution, it is easy to check that
$$\eta_t + \Psi_x = 0. \tag{3.49}$$

A weak solution q satisfies the weak form of the entropy inequality for all $\phi \in C_0^1(\mathbb{R} \times \mathbb{R}^+)$ with $\phi(x,t) \geq 0$,

$$\int_0^\infty \int_{-\infty}^\infty [\phi_t \eta(q) + \phi_x \Psi(q)]\, dxdt + \int_{-\infty}^\infty \phi(x,0)\eta(q(x,0))\, dx \geq 0. \tag{3.50}$$

The entropy inequality (3.50) is often written formally as

$$\eta(q)_t + \Psi(q)_x \leq 0. \tag{3.51}$$

In case of a scalar conservation law in one space dimension with a convex flux, it may be shown that the three entropy condition are equivalent. Although the Lax entropy condition seems to be most straightforward, it is usually more convenient to adopt the entropy pair approach for theorem proof of a general Cauchy problem. In fact, the Lax entropy condition may be used for each shock. There is a general methodology of shock tracking for solving hyperbolic conservation laws. In that approach, one detects and follows each shock, using the Rankine-Hugoniot relation and the entropy condition. The whole problem is then transformed to the calculations of a smooth solution away from the shock, and the calculation of how the smooth solution subdomains evolve or how the shock fronts evolve. The drawback of this approach becomes obvious when many, possibly small or weak, shock fronts emerges, which is typically the case in applications. Different from the shock tracking method, it is more commonly used that one solves the equations by a carefully designed scheme that does not explicitly identify the shock front. Instead, a shock front appears in terms of a sharp numerical gradient. This gives a shock capturing approach. The programming is much simpler for such a scheme. As a numerical scheme always contains artificial viscosity to stabilize the computation around a shock, such a scheme may overlook a shock if the numerical dissipation smooths out a relatively flat gradient. Finite volume method falls into this shock capturing approach.

In the following, we shall prove that the Godunov scheme yields an entropic solution in the limit of $j \to 0$. As a matter of fact, the Godunov scheme is in a conservative form, and the Riemann solver implies a local entropy inequality across each cell. We shall show that the entropy inequality is obtained over the whole domain by the Godunov method.

In numerical computations, the Godunov scheme considers locally each pair of neighboring cells as a Riemann problem. So far as the CFL condition is satisfied, the two Riemann problems around each cell do not interact with each other.

We first divide the domain of interest into uniform grids, as shown in Fig **??**. For each cell, we solve two Riemann problems, for which the entropy equality (3.51) holds. We define two functions in $[x_{m-1/2}, x_{m+1/2}]\times[t^n, t^{n+1}]$. First, we take

$$\phi_1(x,t) \sim \begin{cases} 1, & (x,t) \in [x_{m-1/2}, x_m] \times [t^n, t^{n+1}], \\ 0, & \text{elsewhere.} \end{cases} \tag{3.52}$$

Here we use $\sim$ to denote that ϕ_1 is actually a smoothed function, e.g. by a standard mollifier in exponential form.

Similarly, we define

$$\phi_2 \sim \begin{cases} 1, & (x,t) \in [x_m, x_{m+1/2}] \times [t^n, t^{n+1}], \\ 0, & \text{elsewhere.} \end{cases} \tag{3.53}$$

For the first Riemann problem, we know that an entropy inequality holds.

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}}(\phi_{1t}\eta+\phi_{1x}\psi)\mathrm{d}x\mathrm{d}t+\int_{\mathbb{R}}\phi_1(x,0)\eta(x,0)\mathrm{d}x\geq 0. \tag{3.54}$$

Noticing that $\phi_{1t}\sim\delta(t-t^n)-\delta(t-t^{n+1})$ and $\phi_{1x}\sim\delta(x-x_{m-1/2})-\delta(x-x_m)$, we have

$$\begin{aligned}&\int_{x_{m-1/2}}^{x_m}(\eta(q(x,t^n))-\eta(q(x,t^{n+1}))\mathrm{d}x\\ &+\int_{t^n}^{t^{n+1}}\Psi(q(x_{m-1/2},t))-\Psi(q(x_m,t))\mathrm{d}t+\int_{\mathbb{R}}\phi_1(x,0)\eta(q(x,0))\mathrm{d}x\geq 0.\end{aligned} \tag{3.55}$$

The last term drops out unless we discuss the first time step.

Similarly, we have

$$\begin{aligned}&\int_{x_m}^{x_{m+1/2}}(\eta(q(x,t^n))-\eta(q(x,t^{n+1}))\mathrm{d}x\\ &+\int_{t^n}^{t^{n+1}}\Psi(q(x_m,t))-\Psi(q(x_{m+1/2},t))\mathrm{d}t+\int_{\mathbb{R}}\phi_2(x,0)\eta(q(x,0))\mathrm{d}x\geq 0.\end{aligned} \tag{3.56}$$

Summing the ϕ_1 and ϕ_2 inequalities together, we obtain

$$\begin{aligned}\int_{x_{m-1/2}}^{x_{m+1/2}}\eta(q(x,t^{n+1}))\,dx\leq&\int_{x_{m-1/2}}^{x_{m+1/2}}\eta(q(x,t^n))\,dx\\ +\int_{t^n}^{t^{n+1}}\Psi(q(x_{m-1/2},t))\,dt-&\int_{t^n}^{t^{n+1}}\Psi(q(x_{m+1/2},t))\,dt.\end{aligned} \tag{3.57}$$

If $n=0$, an additional initial term should be included. Here $q(x,t)$ solves the equation with a piecewise initial data

$$q(x,t^n)=q_m^n,\ x_{m-1/2}<x<x_{m+1/2}. \tag{3.58}$$

Because $q(x,t^n)$ is constant in the m-th cell, we observe

$$\int_{x_{m-1/2}}^{x_{m+1/2}}\eta(q(x,t^n))\,dx=h\eta_m^n. \tag{3.59}$$

On the other hand, $\int_{x_{m-1/2}}^{x_{m+1/2}}\eta(q(x,t^{n+1}))\,dx$ does not equal to $h\eta_m^{n+1}$ because we average $q(x,t^{n+1})$ to get q_m^{n+1}. However, because $\eta''(q)\geq 0$, the Jenssen's inequality gives

$$\begin{aligned}&\eta_m^{n+1}\equiv\eta(q_m^{n+1})=\eta\left(\frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}}q(x,t^{n+1})\,dx\right)\\ \leq&\frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}}\eta\left(q(x,t^{n+1})\right)\,dx\leq\eta_m^n-\frac{k}{h}[\Psi_{m+1/2}-\Psi_{m-1/2}],\end{aligned} \tag{3.60}$$

where $\Psi_{m+1/2} \equiv \Psi(q_{m+1/2})$. Using the same argument for the Lax-Wendroff theorem and taking the limit $j \to \infty$, we may prove that

$$\int_{\mathbb{R}^+} \int_{\mathbb{R}} (\phi_t \eta + \phi_x \Psi) \mathrm{d}x \mathrm{d}t + \int_{\mathbb{R}} \phi(x,0) \eta(q(x,0)) \mathrm{d}x \geq 0. \tag{3.61}$$

In fact, the four conditions in the Lax-Wendroff theorem may be checked as follows. First, the consistency holds with $\bar{\phi}_{m+1/2}(\bar{q}) = \phi(q)$. Secondly, the formulation (inequality) for (η, Ψ) is in conservative form. Thirdly, it is stable as $TV(\eta) \leq C \times TV(q)$ because η is a C^2 function in q. Finally, the convergence holds again due to $\eta \in C^2$ which yields $\mathcal{S}\eta(^{(j)}q) \to \eta(q)$ provided $\mathcal{S}^{(j)}q \to q$.

From Equation (3.61) we know that in the weak sense

$$\frac{\partial}{\partial t} \eta(q(x,t)) + \frac{\partial}{\partial x} \Psi(q(x,t)) \leq 0. \tag{3.62}$$

We conclude that the Godunov method satisfying a local entropy condition (for each cell Riemann problem) guarantees the global entropy condition.

3.5 Nonlinear Stability and Convergence

For a Linear advection equation

$$u_t + c u_x = 0, \qquad (c > 0) \tag{3.63}$$

we consider a general linear scheme

$$u_m^{n+1} = \sum_j \alpha_j u_{m+j}^n. \tag{3.64}$$

It may be shown that if $\alpha_j \geq 0$ $(\forall j)$, then this scheme is at most of the first order accuracy except for the special case of $u_m^{n+1} = u_{m-l}^n$ with $ck = lh$.

This scheme has some nice properties. First, because $\alpha_j \geq 0$, it is a positive scheme. Consequently, it preserves monotonicity, and is hence called as a monotone scheme. That is, if the profile of u^n is monotone, so is u^{n+1}. In contrast, a non-monotone scheme usually generates new oscillations.

Secondly, the positive scheme leads to a contractive operator. More generally, if we have a scheme that gives a solution with

$$u_m^{n+1} = N(u_{m-1}^n, u_m^n, u_{m+1}^n). \tag{3.65}$$

Suppose that it gives another numerical solution

$$\tilde{u}_m^{n+1} = N(\tilde{u}_{m-1}^n, \tilde{u}_m^n, \tilde{u}_{m+1}^n). \tag{3.66}$$

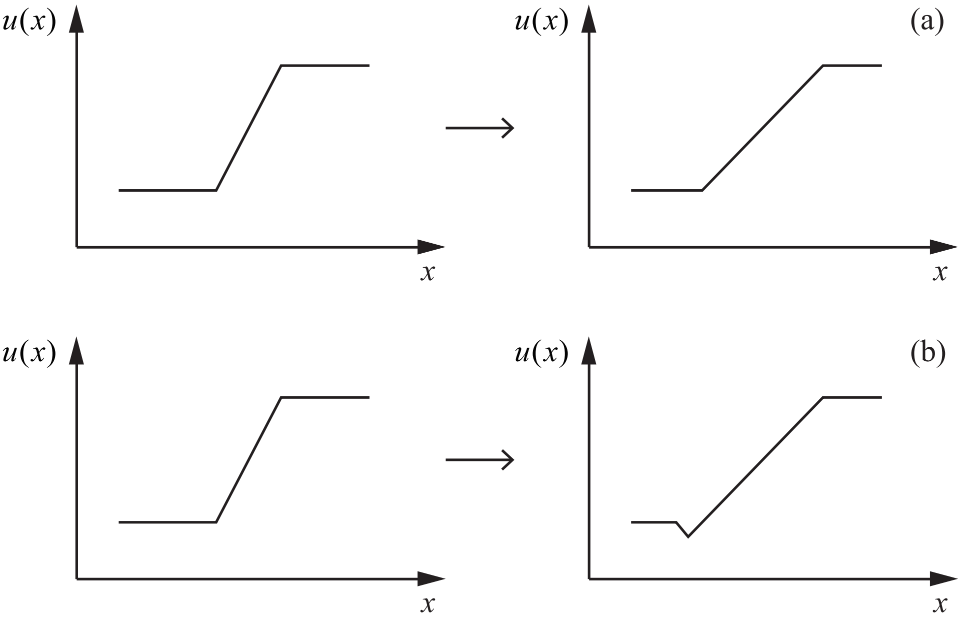


Figure 3.3: (a) Positive scheme preserves monotonicity. (b) Non-monotone scheme generates new oscillations.

We obtain a contractive operator if it holds that

$$\| u^{n+1} - \tilde{u}^{n+1} \| \leq \| u^n - \tilde{u}^n \| \,. \tag{3.67}$$

A positive scheme always leads to a contractive operator.

$$\begin{aligned}
\| u_m^{n+1} - \tilde{u}_m^{n+1} \| &= \sum_m \sum_j |\alpha_j (u_{m+j}^n - \tilde{u}_{m+j}^n)| \\
&\leq \sum_m \sum_j \alpha_j |(u_{m+j}^n - \tilde{u}_{m+j}^n)| \\
&\leq \sum_{l=m+j} \sum_j \alpha_j |(u_l^n - \tilde{u}_l^n)| \\
&\leq \sum_l |(u_l^n - \tilde{u}_l^n)| = \| u^n - \tilde{u}^n \| \,.
\end{aligned}$$

Here we have used the fact $\sum_j \alpha_j \leq 1$, which is a consistency and stability requirement.

Finally, the previous positive scheme is stable. More generally, for the nonlinear equation

$$q_t + f(q)_x = 0, \tag{3.68}$$

a nonlinear scheme

$$q_m^{n+1} = N(q_{m-I}^n, \cdots, q_{m+I}^n), \tag{3.69}$$

is a monotone scheme if $\partial N/\partial q_j \geq 0$, $\forall j$. There appears no new oscillation in numerical computations with such a scheme. However, it is only of first order accuracy.

We remark that monotonicity of a scheme is too restrictive. Therefore, we consider the TV-stability (total variation stability) instead.

We first define $L^{1,T}$ for time $T > 0$ with norm

$$\|v\|_{1,T} = \int_0^T \|v(,t)\|_1 \mathrm{d}t = \int_0^T \int_{-\infty}^{+\infty} |v(x,t)| \mathrm{d}x\mathrm{d}t. \tag{3.70}$$

Then we define a total variation for time T by

$$TV_T(q) = \lim_{\epsilon\to 0} \sup \frac{1}{\epsilon} \int_0^T \int_{-\infty}^{+\infty} |q(x+\epsilon,t) - q(x,t)| + |q(x,t+\epsilon) - q(x,t)| \mathrm{d}x\mathrm{d}t. \tag{3.71}$$

It is actually the sum of integrals of the total variations in time and in space. That is,

$$TV_T(q) = \int_0^T TV(q(\cdot,t))\mathrm{d}t + \int_{-\infty}^{+\infty} TV(q(x,\cdot))\mathrm{d}x. \tag{3.72}$$

In particular, for a discrete solution, we have

$$\begin{aligned} TV_T(q^{(k)}) &= \sum_{n=0}^{T/k} \sum_{m=-\infty}^{+\infty} \left[k|q_{m+1}^n - q_m^n| + h|q_m^{n+1} - q_m^n|\right] \\ &= \sum_{n=0}^{T/k} [kTV(q^n) + \|q^{n+1} - q^n\|_{L^1}]. \end{aligned} \tag{3.73}$$

It may be shown that

$$\mathcal{K} = \left\{q \in L^{1,T} : TV_T(q) \leq R, \mathrm{supp}(q(\cdot,t)) \subseteq [-M,M], \forall t \in [0,T]\right\} \tag{3.74}$$

is compact, that is, any bounded sequence in this set has a convergent subsequence in $L^{1,T}$.

We call a numerical scheme TV-stable if $\exists \delta > 0$, such that $\forall^{(j)}k < \delta$, numerical solution ${}^{(j)}q$ lies in $\mathcal{K}$ for certain R and M. Here R and M may depend on the initial data, T and $f(q)$, but not on ${}^{(j)}k$.

We note that the existence of M is guaranteed if initial data has compact support, due to the finite propagative speed.

Unlike in the linear case, the relation among stability and convergence is more complicated. By convergence, we usually mean $q_m^n \to q(x,t)$ as $h, k \to 0$, as discussed before. We notice that there may exist more than one weak solution q. So we define a solution set $\mathcal{W} = \{q | q \text{ is a weak solution}\}$. In this section, by convergence we mean

$$\mathrm{dist}(\mathcal{S}q, \mathcal{W}) \equiv \inf_{w\in\mathcal{W}} \|\mathcal{S}q - w\|_{1,T} \to 0 \quad \text{as } k \to 0. \tag{3.75}$$

Moreover, for simplicity, we let $\lambda = \frac{k}{h}$ be fixed.

Theorem 3.2. *A conservative method with Lipschitz continuous numerical flux is TV-stable, if* $\forall q^0, \exists \delta, R > 0$*, such that* $TV(q^n) \leq R, \quad \forall k \leq \delta, nk < T$.

Proof. We first justify that $\exists \alpha > 0$ such that $\|q^{n+1} - q^n\|_{L^1} \leq \alpha k$, $\forall k < k_0$, $nk \leq T$.

We employ a conservative method

$$q_m^{n+1} - q_m^n = -\frac{k}{h}(F_{m+1/2}^n - F_{m-1/2}^n). \tag{3.76}$$

Therefore we have

$$\|q^{n+1} - q^n\|_{L^1} = k \sum |F_{m+1/2} - F_{m-1/2}|. \tag{3.77}$$

Since q^n has a compact support and $TV(q^n) \leq R$, we have $|q_m^n| \leq \frac{R}{2}$. Therefore, by the Lipschitz continuity, it holds for certain K that

$$|F_{m+1/2} - F_{m-1/2}| \leq K \sum_j |q_{m+j}^n - q_{m+j-1}^n|. \tag{3.78}$$

Here j is the index for the neighboring cells in the stencil. Then we estimate the L^1 norm of q^n and compute the total variation of ${}^{(j)}q$.

$$\begin{aligned}
\|q^{n+1} - q^n\|_{L^1} &\leq k \cdot K \sum_{m=-\infty}^{+\infty} \sum_j |q_{m+j}^n - q_{m+j-1}^n| \\
&\leq k \cdot K \sum_j TV(q^n) \\
&\leq k \cdot K \cdot (\text{number of } j\ ) \cdot R \\
&\equiv \alpha k.
\end{aligned} \tag{3.79}$$

$$\begin{aligned}
TV({}^{(j)}q) &= \sum_{n=0}^{T/k} [kTV(q^n) + \|q^{n+1} - q^n\|_{L^1}] \\
&\leq \sum_{n=0}^{T/k} [k \cdot R + \alpha k] \\
&\leq k \cdot (R + \alpha) \cdot T/k = (R + \alpha) \cdot T.
\end{aligned} \tag{3.80}$$

Finally, we arrive at the conclusion that ${}^{(j)}q$ is TV-stable. □

Theorem 3.3. *let* ${}^{(j)}q$ *be obtained by a conservative and consistent scheme with a Lipschitz continuous numerical flux. If the method is TV-stable, then it is convergent. That is,* $dist({}^{(j)}q, \mathcal{W}) \to 0$ *as* $j \to \infty$.

Proof. Assume if it is not true, then there exists $\varepsilon > 0$ and sequences $\{^{(1)}q, ^{(2)}q, \cdots\}$ with $^{(j)}k \to 0$. where

$$\mathrm{dist}(q^{k_j}, \mathcal{W}) > \varepsilon, \quad \forall j.$$

Therefore, we can extract a subsequence in the compact set $\mathcal{K}$, which converges to $v \in \mathcal{K}$.
Then it holds that

$$\|^{(j)}q - v\|_{1,T} < \varepsilon \quad \text{as } j \to \infty.$$

By the Lax-Wendroff theorem, $v \in \mathcal{W}$. This contradicts with the assumption. Therefore, the theorem holds. □

Chapter 4

Finite Volume Method for Nonlinear Systems

4.1 General Setting

In this chapter, we design numerical schemes for either a linear system

$$q_t + Aq_x = 0, \quad q \in \mathbb{R}^d, \tag{4.1}$$

or a nonlinear system

$$q_t + f(q)_x = 0, \quad q \in \mathbb{R}^d. \tag{4.2}$$

Same as before, a finite volume method starts with the cell average

$$Q_m^n \equiv \frac{1}{h}\int_{x_{m-1/2}}^{x_{m+1/2}} q(x, t^n)\mathrm{d}x, \tag{4.3}$$

and a flux function

$$F_{m-1/2}^n \equiv \frac{1}{k}\int_{t^n}^{t^{n+1}} f(q(x_{m-1/2}, t^n))\mathrm{d}t. \tag{4.4}$$

We remark that Q is used here to emphasize that a vector function is under consideration here.

The exact formula for updating is

$$Q_m^{n+1} = Q_m^n - \frac{k}{h}(F_{m+1/2}^n - F_{m-1/2}^n). \tag{4.5}$$

The key issue is then the approximation of the exact flux by a numerical flux. Typically, we require the CFL condition $|\lambda|_{\max}\frac{k}{h} < \frac{1}{2}$. Usually this may be relaxed to $|\lambda|_{\max}\frac{k}{h} < 1$.

The approximate flux is usually designed as a function of several cell averages. The resulted scheme the depends on corresponding cells, which form a stencil. In a simplest form, two cell averages are used.

$$F^n_{m-1/2} \equiv \mathcal{F}(Q^n_{m-1}, Q^n_m) \equiv f(Q^\star_{m-1/2}(Q^n_{m-1}, Q^n_m)). \tag{4.6}$$

We define fluctuations

$$\begin{cases} \mathcal{A}^-\Delta Q^n_{m-1/2} = f(Q^\star_{m-1/2}) - f(Q^n_{m-1}), \\ \mathcal{A}^+\Delta Q^n_{m-1/2} = f(Q^n_m) - f(Q^\star_{m-1/2}). \end{cases} \tag{4.7}$$

The scheme then reads

$$Q^{n+1}_m = Q^n_m - \frac{k}{h}(\mathcal{A}^+\Delta Q^n_{m-1/2} + \mathcal{A}^-\Delta Q^n_{m+1/2}). \tag{4.8}$$

We observe that the increment of Q comes solely from the right-going waves from the left cell boundary, and the left-going waves from the right cell boundary.

4.2 Godunov Method for Linear Systems

We develop some basic notions for a system through the Godunov method for the linear system (4.1). Formally, the finite volume method gives

$$\frac{Q^{n+1}_m - Q^n_m}{k} + \frac{AQ^n_{m+1/2} - AQ^n_{m-1/2}}{h} = 0. \tag{4.9}$$

By hyperbolicity, we know that there is a transform matrix P composed of row eigenvectors for A, such that

$$PAP^T = \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d). \tag{4.10}$$

Under this transform, the system becomes decoupled. Each of the state variable is governed by a linear advection equation. Numerically, we define

$$PQ^n_{m-1/2} = \tilde{Q}^n_{m-1/2}, \quad PQ^n_{m-1} = \tilde{Q}^n_{m-1}, \quad PQ^n_m = \tilde{Q}^n_m. \tag{4.11}$$

By the Godunov method for each decoupled equation, we obtain

$$\tilde{Q}^{n+1}_m = \tilde{Q}^n_m - \frac{k}{h}\Lambda(\tilde{Q}^n_{m+1/2} - \tilde{Q}^n_{m-1/2}). \tag{4.12}$$

Here, the intermediate state is defined by

$$\begin{aligned} \tilde{Q}^n_{m-1/2} &= \begin{cases} \text{p-th entry in } \tilde{Q}^n_{m-1}, & \text{for}\lambda_p > 0 \\ \text{p-th entry in } \tilde{Q}^n_m, & \text{for } \lambda_p < 0 \end{cases} \\ &= \tilde{Q}^n_{m-1} + \{\text{p-th entry in } \tilde{Q}^n_m - \tilde{Q}^n_{m-1} \text{ for } \lambda_p < 0\} \\ &= \tilde{Q}^n_m - \{\text{p-th entry in } \tilde{Q}^n_m - \tilde{Q}^n_{m-1} \text{ for } \lambda_p > 0\}. \end{aligned} \tag{4.13}$$

To simplify the notations, we omit the superscript n for the discussions on numerical flux.

We define

$$\lambda_p^- = \min(0, \lambda_p), \quad \lambda_p^+ = \max(0, \lambda_p), \tag{4.14}$$

$$\Lambda^\pm = \mathrm{diag}(\lambda_1^\pm, \cdots, \lambda_d^\pm), \quad |\Lambda| = \mathrm{diag}(|\lambda_1|, \cdots, |\lambda_d|), \tag{4.15}$$

and accordingly

$$A^\pm = P^T \Lambda^\pm P, \quad |A| = P^T |\Lambda| P. \tag{4.16}$$

The Godunov flux for the original variable may be computed as follows.

$$\begin{aligned} AQ_{m-1/2} &= AQ_{m-1} + AP^T(\tilde{Q}_{m-1/2} - \tilde{Q}_{m-1}) \\ &= AQ_{m-1} + P^T PAP^T(\tilde{Q}_{m-1/2} - \tilde{Q}_{m-1}) \\ &= AQ_{m-1} + P^T \Lambda(\tilde{Q}_{m-1/2} - \tilde{Q}_{m-1}) \\ &= AQ_{m-1} + \sum_p \lambda_p (P^T(\tilde{Q}_{m-1/2} - \tilde{Q}_{m-1}))_{\text{p-th entry}} \\ &= AQ_{m-1} + \sum_p \lambda_p^- w_{m-1/2}^{(p)}. \end{aligned} \tag{4.17}$$

Here $Q_m - Q_{m-1} = \sum w_{m-1/2}^{(p)}$ is the normal mode decomposition. That is, $w_{m-1/2}^{(p)}$ is an eigenvector of A corresponding to the eigenvalue λ_p.

Similarly, we may show that

$$AQ_{m-1/2} = AQ_m - \sum_p \lambda_p^+ w_{m-1/2}^{(p)}. \tag{4.18}$$

In summary, we can rewrite the Godunov numerical flux as follows.

$$F_{m-1/2}^n = AQ_{m-1} + \sum_{p=1}^{d} \lambda_p^- w_{m-1/2}^{(p)}, \tag{4.19}$$

or,

$$F_{m-1/2}^n = AQ_m - \sum_{p=d}^{m} \lambda_p^+ w_{m-1/2}^{(p)}. \tag{4.20}$$

Furthermore, the average of the above two expressions gives

$$F_{m-1/2} = \frac{1}{2}(AQ_{m-1} + AQ_m) - \frac{1}{2}|A|(Q_m - Q_{m-1}). \tag{4.21}$$

This can be viewed as the arithmetic average plus a correction term that stabilizes the method. For the constant-coefficient linear problem this is simply another way to rewrite the Godunov or upwind flux, but this form is often seen in extensions to nonlinear problems based on approximate

Riemann solvers, as discussed in the next section. This formulation is also useful in studying the numerical dissipation of the upwind method.

Using the flux above we can get the following updating formula.

$$Q_m^{n+1} = Q_m - \frac{1}{2}\frac{k}{h}A(Q_{m+1} - Q_{m-1}) + \frac{1}{2}\frac{k}{h}\sum_{p=1}^{d}(|\lambda_p|w_{m+1/2}^{(p)} - |\lambda_p|w_{m-1/2}^{(p)}). \tag{4.22}$$

This is equivalent to

$$Q_m^{n+1} = Q_m - \frac{1}{2}\frac{k}{h}A(Q_{m+1} - Q_{m-1}) + \frac{k}{2h}|A|(Q_{m-1} - 2Q_m + Q_{m+1}). \tag{4.23}$$

The second term is a central difference, whereas the last term stands for the viscosity. Notice that the central difference possesses second-order accuracy, yet lacks of stability. The additional viscous term stabilizes the scheme and help capturing the entropic shock wave, yet the order of accuracy is thence reduced. The main goal for developing high resolution schemes is then to balance the two contradictory demands of stability and accuracy.

4.3 Approximate Riemann Solvers

To apply the Godunov method on a system of equations, we essentially only need to determine $q^{\downarrow}(q_l, q_r)$, the state along $x/t = 0$ based on the Riemann data q_l and q_r. We do not need the entire structure, but to compute $q^{\downarrow}$ we must typically determine something about the full wave structure and the wave speeds in order to determine where $q^{\downarrow}$ lies in state space. The process of solving the Riemann problem is thus often quite expensive, even in the end we use very little information from this solution to define the flux. A wide variety of approximate Riemann solvers have been proposed that can be applied much more cheaply than the exact Riemann solver and yet give results that in many cases are equally good when used in Godunov or high-resolution methods.

For given data Q_{m-1} and Q_m, an approximate Riemann solution might define a function $\hat{Q}_{m-1/2}(x/t)$ that approximates the true similarity solution to Riemann problem with data Q_{m-1} and Q_m. This function typically consists of a set of d waves. A p-th family wave corresponds to an increment vector $w_{m-1/2}^{(p)}$, propagating with a speed $s_{m-1/2}^{(p)}$. That is,

$$Q_m - Q_{m-1} = \sum_{p=1}^{d} w_{m-1/2}^{(p)}. \tag{4.24}$$

To generalize the Godunov method using this function, one may either approximate the state, or approximate directly the flux. For the first choice, we define the numerical flux by $F_{m-1/2} = f(\hat{Q}^{\downarrow}_{m-1/2})$, where

$$\hat{Q}^{\downarrow}_{m-1/2} = Q_{m-1} + \sum_{s^p_{m-1/2}<0} w^p_{m-1/2}. \tag{4.25}$$

This $\hat{Q}^{\downarrow}_{m-1/2}$ is the intermediate state along the cell interface. Then we set

$$\mathcal{A}^{-}\triangle Q_{m-1/2} = f(\hat{Q}^{\downarrow}_{m-1/2}) - f(Q_{m-1}), \mathcal{A}^{+}\triangle Q_{m-1/2} = f(Q_m) - f(\hat{Q}^{\downarrow}_{m-1/2}). \tag{4.26}$$

Alternatively, we may directly use the waves and speeds from the approximate Riemann solution to define

$$\mathcal{A}^{-}\triangle Q_{m-1/2} = \sum_{p=1}^{d}(s^{(p)}_{m-1/2})^{-} w^{(p)}_{m-1/2}, \mathcal{A}^{+}\triangle Q_{m-1/2} = \sum_{p=1}^{d}(s^{(p)}_{m-1/2})^{+} w^{(p)}_{m-1/2}. \tag{4.27}$$

With either definition of the fluxes, we then adopt the updating formula (4.8). In case of a linear system, the above two approaches are equivalent. But for a nonlinear system, they are different in general.

4.3.1 Linearized Riemann Solvers

To approximate Riemann solutions, one may replace the nonlinear problem by a linear one

$$\hat{q}_t + \hat{A}_{m-1/2}\hat{q}_x = 0. \tag{4.28}$$

The matrix $\hat{A}_{m-1/2}$ is chosen to be some approximation to $f'(q)$ valid in a neighborhood of the data Q_{m-1} and Q_m. There are two requirements on the matrix $\hat{A}_{m-1/2}$. First, to make the resulted linear system hyperbolic, $\hat{A}_{m-1/2}$ should be diagonalizable with real eigenvalues. Furthermore, to make the linearize system compatible with the nonlinear one, we require

$$\hat{A}_{m-1/2} \to f'(\bar{q}) \quad \text{as } Q_{m-1}, Q_m \to \bar{q}. \tag{4.29}$$

A natural way to define $\hat{A}_{m-1/2}$ is to take the average of $f'(Q_{m-1})$ and $f'(Q_m)$. But the resulted matrix is usually not diagonalizable, even if $f'(Q_{m-1})$ and $f'(Q_m)$ have real eigenvalues. The other choice is to take $\hat{A}_{m-1/2} = f'(\hat{Q}_{m-1/2})$, where $\hat{Q}_{m-1/2}$ is a certain average of Q_{m-1} and Q_m. For instance, one may take $\hat{Q}_{m-1/2} = \frac{1}{2}(Q_{m-1}+Q_m)$. But this simple choice does not work properly. If a cell Riemann solution contains only left-going waves, the correct choice should be $\hat{Q}_{m-1/2} = Q_m$. A wave view of the Riemann solution is more suitable in developing linear approximate solvers.

4.3.2 Roe Linearization

The situation discussed in the previous section is quite complicated. Mathematically speaking, the general Riemann solution for a cell problem includes all d waves. However, in applications, Roe noticed that the cell Riemann problems typically have a large jump at most in one wave family. This greatly ease the problem and motivates the Roe linearization.

Let us consider a shock wave for the p-th family connecting the states Q_{m-1} and Q_m. This means $||w^{(j)}_{m+1/2}|| = O(\triangle x)$ for all other $j \neq p$. For other types of waves, we shall make a discussion later on. To capture precisely this shock, the linearized system should admit the eigenvector $w^{(p)}_{m-1/2}$ with eigenvalue $s^{(p)}_{m-1/2}$. Together with the Rankine-Hugoniot condition, we obtain that

$$\hat{A}_{m-1/2}(Q_m - Q_{m-1}) = s^{(p)}_{m-1/2}(Q_m - Q_{m-1}) = f(Q_m) - f(Q_{m-1}). \tag{4.30}$$

It should hold for arbitrary wave family and data that

$$\hat{A}_{m-1/2}(Q_m - Q_{m-1}) = f(Q_m) - f(Q_{m-1}). \tag{4.31}$$

Before the construction of the approximate matrix $\hat{A}_{m-1/2}$, we remark that the corresponding scheme is conservative. Moreover, this also allows a consistent numerical flux, as it produces $F_{m-1/2}(\bar{Q}, \bar{Q}) = 0 = f(\bar{Q}) - f(\bar{Q})$. Consider the straight-line path parameterized by

$$q(\xi) = Q_{m-1} + (Q_m - Q_{m-1})\xi, \quad 0 \leq \xi \leq 1. \tag{4.32}$$

Then $f(Q_m) - f(Q_{m-1})$ can be written as the line integral.

$$f(Q_m) - f(Q_{m-1}) = \left[\int_0^1 \nabla_q f(q(\xi)) \mathrm{d}\xi\right] (Q_m - Q_{m-1}). \tag{4.33}$$

This suggests us to take

$$\hat{A}_{m-1/2} = \int_0^1 \nabla_q f(q(\xi)) \mathrm{d}\xi. \tag{4.34}$$

This is usually not easy to compute. Roe introduced a parameter vector $z \in \mathbb{R}^d$, which is actually a change of variables, and ease the calculation for integrals. The inverse transform of $z(q)$ is denoted as $q(z)$. For z, we integrate along the path

$$z(\xi) = Z_{m-1} + (Z_m - Z_{m-1})\xi, \tag{4.35}$$

where $Z_m = z(Q_m)$ and $Z_{m-1} = z(Q_{m-1})$. Then we get

$$f(Q_m) - f(Q_{m-1}) = \left[\int_0^1 \nabla_z f(q(z(\xi))) \mathrm{d}\xi\right] (Z_m - Z_{m-1}). \tag{4.36}$$

Take $q(z)$ as the inverse transform of the variables, we have

$$Q_m - Q_{m-1} = \left[\int_0^1 \nabla_z q(z(\xi)) \mathrm{d}\xi \right] (Z_m - Z_{m-1}). \tag{4.37}$$

Let

$$\hat{B}_{m-1/2} = \int_0^1 \nabla_z q(z(\xi)) \mathrm{d}\xi, \quad \hat{C}_{m-1/2} = \int_0^1 \nabla_z f(q(z(\xi))) \mathrm{d}\xi. \tag{4.38}$$

Then we have

$$\hat{A}_{m-1/2} = \hat{C}_{m-1/2} \hat{B}^{-1}_{m-1/2}. \tag{4.39}$$

To better illustrate this procedure, we derive the Roe matrix for the shallow water equations.

$$h_t + (hu)_x = 0, \quad (hu)_t + (hu^2 + \frac{1}{2} g h^2)_x = 0. \tag{4.40}$$

For the shallow water equations we have

$$q = \begin{pmatrix} h \\ hu \end{pmatrix} = \begin{pmatrix} q^1 \\ q^2 \end{pmatrix}, \tag{4.41}$$

$$f(q) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2} g h^2 \end{pmatrix} = \begin{pmatrix} q^2 \\ q^2/q^1 + \frac{1}{2} g (q^1)^2 \end{pmatrix}, \tag{4.42}$$

and

$$\nabla_q f(q) = \begin{pmatrix} 0 & 1 \\ -(q^2/q^1)^2 + g q^1 & 2q^2/q^1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -u^2 + gh & 2u \end{pmatrix}. \tag{4.43}$$

As a parameter vector we choose $z = h^{-1/2} q$, so that

$$\begin{pmatrix} z^1 \\ z^2 \end{pmatrix} = \begin{pmatrix} h^{1/2} \\ h^{1/2} u \end{pmatrix}. \tag{4.44}$$

The inverse transform reads

$$q(z) = \begin{pmatrix} (z^1)^2 \\ z^1 z^2 \end{pmatrix}. \tag{4.45}$$

We find that

$$\nabla_z q = \begin{pmatrix} 2z^1 & 0 \\ z^2 & z^1 \end{pmatrix} \tag{4.46}$$

and

$$f(q(z)) = \begin{pmatrix} z^1 z^2 \\ (z^2)^2 + \frac{1}{2} g (z^1)^4 \end{pmatrix}. \tag{4.47}$$

Therefore, the Jacobian matrix is

$$\nabla_z f = \begin{pmatrix} z^2 & z^1 \\ 2g(z^1)^3 & 2z^2 \end{pmatrix}. \tag{4.48}$$

We now set $z^p = Z^p_{m-1} + (Z^p_m - Z^p_{m-1})\xi$, for $p = 1, 2$. By direct integration, we get

$$\int_0^1 z^p(\xi)\mathrm{d}\xi = \frac{1}{2}(Z^p_{m-1} + Z^p_m) \equiv \bar{Z}^p, \tag{4.49}$$

and

$$\begin{aligned}\int_0^1 (z^1(\xi))^3\mathrm{d}\xi &= \frac{(Z^1_m)^4 - (Z^1_{m-1})^4}{4(Z^1_m - Z^1_{m-1})} \\ &= \frac{(Z^1_m + Z^1_{m-1})}{2} \cdot \frac{(Z^1_m)^2 + (Z^1_{m-1})^2}{2} \\ &= \bar{Z}^1_m \cdot \bar{h},\end{aligned} \tag{4.50}$$

where

$$\bar{h} = \frac{1}{2}(h_{m-1} + h_m). \tag{4.51}$$

Hence we obtain

$$\hat{B}_{m-1/2} = \begin{pmatrix} 2\bar{Z}^1 & 0 \\ \bar{Z}^2 & \bar{Z}^1 \end{pmatrix},\ \hat{C}_{m-1/2} = \begin{pmatrix} \bar{Z}^2 & \bar{Z}^1 \\ 2g\bar{Z}^1\bar{h} & 2\bar{Z}^2 \end{pmatrix}. \tag{4.52}$$

So we get

$$\begin{aligned}\hat{A}_{m-1/2} &= \begin{pmatrix} 0 & 1 \\ -(\bar{Z}^2/\bar{Z}^1)^2 + g\bar{h} & 2\bar{Z}^2/\bar{Z}^1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ -(\hat{u})^2 + g\bar{h} & 2\hat{u} \end{pmatrix}.\end{aligned} \tag{4.53}$$

Here $\hat{u}$ is the Roe average

$$\hat{u} = \frac{\bar{Z}^2}{\bar{Z}^1} = \frac{h^{1/2}_{m-1}u_{m-1} + h^{1/2}_m u_m}{h^{1/2}_{m-1} + h^{1/2}_m}. \tag{4.54}$$

The resulted Roe matrix is analogous to the Jacobian matrix to the original shallow water equations. We remark that the linearized problem has a linear Jacobian matrix, which varies from cell problem to cell problem, as it is determined by Q_{m-1} and Q_m.

Moreover, the eigen-structure is also similar to the nonlinear problem. In fact, we have eigenvalues

$$\hat{\lambda}^1 = \hat{u} - \hat{c},\ \hat{\lambda}^2 = \hat{u} + \hat{c}, \tag{4.55}$$

and eigen-vectors

$$\hat{r}^1 = \left(\begin{array}{c} 1 \\ \hat{u} - \hat{c} \end{array} \right), \ \hat{r}^2 = \left(\begin{array}{c} 1 \\ \hat{u} + \hat{c}. \end{array} \right) \tag{4.56}$$

Here the sound speed $\hat{c} = (g\bar{h})^{1/2}$. We find a decomposition for $\Delta Q_{m-1/2} = \left(\begin{array}{c} \Delta h \\ \Delta m \end{array} \right)$ as follows.

$$Q^n_m - Q^n_{m-1} = \alpha^1_{m-1/2}\hat{r}_1 + \alpha^2_{m-1/2}\hat{r}_2 \equiv w^1_{m-1/2} + w^2_{m-1/2}. \tag{4.57}$$

The coefficients are

$$\alpha^1_{m-1/2} = \frac{(\hat{u} + \hat{c})\Delta h - \Delta m}{2\hat{c}}, \tag{4.58}$$

$$\alpha^2_{m-1/2} = \frac{-(\hat{u} - \hat{c})\Delta h + \Delta m}{2\hat{c}}. \tag{4.59}$$

Same as for the linear system, the Roe linearized problem yields a numerical flux

$$\begin{aligned} F_{m-1/2} &= f(Q_{m-1}) + \hat{A}^-_{m-1/2}(Q_m - Q_{m-1}) \\ &= f(Q_m) - \hat{A}^+_{m-1/2}(Q_m - Q_{m-1}) \\ &= \frac{1}{2}(f(Q_{m-1}) + f(Q_m)) - \frac{1}{2}|\hat{A}_{m-1/2}|(Q_m - Q_{m-1}). \end{aligned} \tag{4.60}$$

The finite volume scheme reads

$$\begin{aligned} \hat{Q}^{n+1}_m &= \hat{Q}^n_m - \frac{k}{h}(F_{m+1/2} - F_{m-1/2}) \\ &= Q^n_m - \frac{k}{h}\{[\frac{1}{2}(f(Q_m) + f(Q_{m+1})) - \frac{1}{2}|\hat{A}_{m+1/2}|(Q_{m+1} - Q_m)] \\ &\quad - [\frac{1}{2}(f(Q_m) + f(Q_{m-1})) - \frac{1}{2}|\hat{A}_{m-1/2}|(Q_m - Q_{m-1})]\} \\ &= \hat{Q}^n_m - \frac{k}{2h}[f(Q_{m+1}) - f(Q_{m-1})] \\ &\quad + \frac{k}{2h}[|\hat{A}_{m+1/2}|(Q_{m+1} - Q_m) - |\hat{A}_{m-1/2}|(Q_m - Q_{m-1})]. \end{aligned} \tag{4.61}$$

The last term is a numerical viscosity term $\sim |\hat{A}|(Q_{m+1} - 2Q_m + Q_{m-1})$.

Now we consider the application of the Roe solver to the nonlinear problem. Under the assumption of one wave per cell problem, the Roe linearization actually gives a correct flux if the Rankine-Hugoniot relation holds. That means, it is correct for a shock or a contact discontinuity.

For a rarefaction wave, we need a careful exploration. Suppose the cell problem is solved by a rarefaction wave of the p-th family. If the rarefaction wave is purely left-going, that is, $\lambda_p(Q_{m-1}) < \lambda_p(Q_m) < 0$, then $\hat{A}^+_{m-1/2} = 0$ and the numerical flux $F_{m-1/2} = f(Q_m) - \hat{A}^+_{m-1/2}(Q_m - Q_{m-1}) = f(Q_m)$. So, the linearized system produces the correct flux. Similarly, if $0 < \lambda_p(Q_{m-1}) < \lambda_p(Q_m)$, then $\hat{A}^-_{m-1/2} = 0$ and the numerical flux $F_{m-1/2} = f(Q_{m-1}) + \hat{A}^-_{m-1/2}(Q_m - Q_{m-1}) = f(Q_{m-1})$. It is again correct.

However, if we have a transonic rarefaction, namely, $\lambda_p(Q_{m-1}) < 0 < \lambda_p(Q_m)$, the exact solution is a stagnation state $Q^\star$ that makes $\lambda^p(Q^\star) = 0$. But with the linearized problem, we have $f(Q_m) - f(Q_{m-1}) = s(Q_m - Q_{m-1})$, with which the intermediate state is Q_m if $s < 0$, and Q_{m-1} if $s > 0$.

Another aspect to see the difficulty is from the point of view of the numerical viscosity. In a transonic case, $\lambda_p \sim 0$ which makes $|\hat{A}|(q_{m+1} - 2q_m + q_{m-1})$ a too small viscosity.

E. Harten suggested the following entropy fix. In the previous updating formula, we replace $|\hat{A}_{m-1/2}|(Q_m - Q_{m-1}) = \sum_{p=1}^{d} |\hat{A}_{m-1/2}|w^{(p)}_{m-1/2} = \sum_{p=1}^{d} |\lambda_p| w^{(p)}_{m-1/2}$ by $\sum_{p=1}^{d} \phi_\delta(\lambda_p) w^{(p)}_{m-1/2}$, with

$$\phi_\delta(\lambda) = \begin{cases} |\lambda|, & \text{if } |\lambda| > \delta, \\ \frac{\lambda^2+\delta^2}{2\delta}, & \text{if } |\lambda| < \delta. \end{cases} \tag{4.62}$$

Here δ is a small positive parameter. In other equivalent forms of the numerical flux, one may replace $\lambda^\pm$ by

$$\begin{cases} \lambda^- = \frac{1}{2}(\lambda - \phi_\delta(\lambda)), \\ \lambda^+ = \frac{1}{2}(\lambda + \phi_\delta(\lambda)). \end{cases} \tag{4.63}$$

4.4 High Resolution Methods

In the Godunov method, one makes reconstruction of data at each t^n by piecewise constants. This results in the first order of accuracy. To improve the accuracy order, we consider better reconstructions.

4.4.1 Limiters for the Linear Advection Equation

We consider a linear reconstruction for a scalar variable.

$$\widetilde{q}(x, t^n) = q_m + \sigma_m(x - x_m), \quad x \in [x_{m-1/2}, x_{m+1/2}]. \tag{4.64}$$

Noticing that this preserves the cell average q_m, we are free to choose the slope σ_m.

For the linear advection equation

$$q_t + cq_x = 0, (c > 0), \tag{4.65}$$

by straight-forward computations, we find that

$$\begin{aligned} q_m^{n+1} &= \frac{ck}{h}\left(q_{m-1} + \frac{1}{2}(h-ck)\sigma_{m-1}\right) + \left(1 - \frac{ck}{h}\right)\left(q_m - \frac{ck}{2}\sigma_m\right) \\ &= q_m - \frac{ck}{h}(q_m - q_{m-1}) - \frac{ck}{2h}(h-ck)(\sigma_m - \sigma_{m-1}). \end{aligned} \tag{4.66}$$

This corresponds to a numerical flux

$$F_{m-1/2} = cq_{m-1} + \frac{c}{2}(h-ck)\sigma_{m-1}. \tag{4.67}$$

There are various ways to define the slope σ_m. When the nearest neighboring cells are taken into account, one may adopt either of the following choices.

$$\begin{aligned} \text{Lax-Wendroff :} \quad \sigma_m &= \frac{q_{m+1} - q_m}{h}, \\ \text{Beam-Waming :} \quad \sigma_m &= \frac{q_m - q_{m-1}}{h}, \\ \text{Fromm :} \quad \sigma_m &= \frac{q_{m+1} - q_{m-1}}{2h}. \end{aligned}$$

We make some further discussions on the Lax-Wendroff scheme. With the above choice, the scheme reads

$$q_m^{n+1} = q_m - \frac{ck}{h}(q_m - q_{m-1}) - \frac{ck}{2h}(1 - \frac{ck}{h})(q_{m+1} - 2q_m + q_{m-1}). \tag{4.68}$$

Under the CFL condition, an interpretation for the Lax-Wendroff scheme is that it introduces anti-dissipation to the upwind scheme. The upwind scheme possesses too strong viscosity.

In the mean time, we may rewrite the scheme as

$$q_m^{n+1} = q_m - \frac{ck}{2h}(q_{m+1} - q_{m-1}) + \frac{c^2k^2}{2h^2}(q_{m+1} - 2q_m + q_{m-1}). \tag{4.69}$$

The Lax-Wendroff scheme therefore may also be viewed as the centered difference scheme with additional viscosity. In fact, the more natural way to derive this scheme is by Taylor expansion, not the linear reconstruction. We remark that the Taylor expansion derivation is from a finite difference point of view, whereas the linear reconstruction is from a finite volume point of view.

We start with the central difference scheme.

$$\frac{q_m^{n+1} - q_m}{k} + c\frac{q_{m+1} - q_{m-1}}{2h} = 0. \tag{4.70}$$

The temporal term may be expanded as $q_t + \frac{1}{2}kq_{tt} + O(k^2)$. The spatial term is $c(q_x + h^2\frac{q_{xxx}}{6}) + o(h^2)$. Therefore, as $k \sim h$ in the computations, the order of accuracy is lowered due to the temporal term. From the linear advection equation, we know that $q_{tt} = c^2 q_{xx}$. Therefore, we include an additional term to correct $\frac{1}{2}kq_{tt} = \frac{1}{2}c^2 kq_{xx}$. That is, we design (4.69).

Though the Lax-Wendroff method has second order accuracy, there are also some disadvantages. If a numerical solution takes speed α for a sinusoidal wave $e^{i\omega x}$, then a wave package $e^{i\omega x}p(x)$ with long wave envelop $p(x)$ propagates at group velocity smaller than α. This causes phase error. A more severe problem for the Lax-Wendroff scheme is overshooting. It is not a TVD (total-variation-diminishing) scheme.

The three above choices for slope are linear, one may also take nonlinear ones. The motivation for developing nonlinear slopes comes from the compromise between accuracy and stability. To maintain the stability, upwind or Godunov type of schemes are more appealing, as they produce stable numerical results, particularly the entropic shock waves. However, they are only of first order accuracy. The Lax-Wendroff type of schemes are of high accuracy order. The idea is then to blend these two types of schemes. More precisely, in the region where the solution is smooth, we adopt the Lax-Wendroff flux. On the other hand, in the region where the solution likely has discontinuity, we use the upwind flux. In another word, we introduce a switch that turns on and off according to the solution profile. The solution profile is actually identified by discrete gradient. When this is done, we call a scheme is high resolution method, noticing the difference from the high order of accuracy.

We may realize the idea of switch by adopting a minmod limiter. In the reconstruction of data, the three above choices do not make much difference in a smooth region. On the other hand, near a sharp gradient, it is appealing to use the information from a more smooth neighboring cell to form a linear profile in a cell. This leads to

$$\sigma_m = \text{minmod}\left(\frac{q_m - q_{m-1}}{h}, \frac{q_{m+1} - q_m}{h}\right), \tag{4.71}$$

where the function

$$\text{minmod}(a,b) = \begin{cases} a, & ab > 0 \text{ and } |a| \le |b|, \\ b, & ab > 0 \text{ and } |a| > |b|, \\ 0, & ab < 0. \end{cases} \tag{4.72}$$

One may use an MC (monotonized center-difference) limiter as well.

$$\sigma_m = \text{minmod}\left(\frac{q_{m+1} - q_{m-1}}{2h}, \frac{q_m - q_{m-1}}{h}, \frac{q_{m+1} - q_m}{h}\right). \tag{4.73}$$

With these nonlinear slope limiters, the reconstructed data becomes smoother and preserves monotonicity.

The numerical flux formulation is determined for a given slope limiter. Consider the linear advection equation for both $c \geq 0$ and $c < 0$, we may find the flux function with linear reconstruction.

$$F_{m-1/2} = \begin{cases} cq_{m-1} + \frac{c}{2}(h - ck)\sigma_{m-1}, & c \geq 0, \\ cq_m - \frac{c}{2}(h - ck)\sigma_m, & c < 0. \end{cases} \tag{4.74}$$

With $\Delta q_{m-1/2} = q_m - q_{m-1}$, we rewrite the flux function as

$$F_{m-1/2} = c^- q_m + c^+ q_{m-1} + \frac{|c|}{2}\left(1 - \frac{ck}{h}\right)\delta_{m-1/2}. \tag{4.75}$$

where

$$\begin{aligned} \delta_{m-1/2} &= \begin{cases} h\sigma_{m-1}, & c \geq 0, \\ h\sigma_m, & c < 0, \end{cases} \\ &\equiv \phi(\theta_{m-1/2})\Delta q_{m-1/2}, \end{aligned} \tag{4.76}$$

$$\theta_{m-1/2} = \frac{\Delta q_{I-1/2}}{\Delta q_{m-1/2}}, \quad I = \begin{cases} m-1, & c \geq 0, \\ m+1, & c < 0. \end{cases} \tag{4.77}$$

We call $\phi(\theta)$ a flux limiter. Same as the slope limiter, the flux limiter also has many different choices, both linear and nonlinear.

Linear flux limiters include

$$\text{upwind:} \quad \phi(\theta) = 0, \tag{4.78}$$

$$\text{Lax-Wendroff:} \quad \phi(\theta) = 1, \tag{4.79}$$

$$\text{Beam-Warming:} \quad \phi(\theta) = \theta, \tag{4.80}$$

$$\text{Fromm:} \quad \phi(\theta) = \frac{1+\theta}{2}. \tag{4.81}$$

Nonlinear flux limiters include

$$\text{minmod:} \quad \phi(\theta) = \text{minmod}(1, \theta), \tag{4.82}$$

$$\text{MC:} \quad \phi(\theta) = \max\left(0, \min\left(\frac{1+\theta}{2}, 2, 2\theta\right)\right). \tag{4.83}$$

To select a flux limiter, we usually require it to be TVD. The following Harten's theorem gives a sufficient condition.

Theorem 4.1. *(Harten) A scheme*

$$q_m^{n+1} = q_m - C_{m-1}(q_m - q_{m-1}) + D_m(q_{m+1} - q_m), \tag{4.84}$$

is TVD, if

$$C_m \geq 0, \quad D_m \geq 0, \quad C_m + D_m \leq 1, \quad \forall m. \tag{4.85}$$

Proof. From the scheme, we have

$$\begin{aligned} q_{m+1}^{n+1} - q_m^{n+1} =& (1 - C_m - D_m)(q_{m+1} - q_m) \\ &+ D_{m+1}(q_{m+2} - q_{m+1}) + C_{m-1}(q_m - q_{m-1}). \end{aligned} \tag{4.86}$$

Then the total variation may be estimated as follows.

$$\begin{aligned} \frac{1}{h}TV(q^{n+1}) &= \sum_m |q_{m+1}^{n+1} - q_m^{n+1}| \\ &\leq \sum_m (1 - C_m - D_m)|q_{m+1} - q_m| \\ &\quad + D_{m+1}|q_{m+2} - q_{m+1}| + C_{m-1}|q_m - q_{m-1}| \\ &= \sum_m [(1 - C_m - D_m) + C_m + D_m]|q_{m+1} - q_m| \\ &= \frac{1}{h}TV(q^n). \end{aligned} \tag{4.87}$$

□

Consider $c > 0$, and let $\gamma = ck/h$. The scheme is

$$\begin{aligned} q_m^{n+1} =& q_m - \gamma(q_m - q_{m-1}) - \\ & \frac{\gamma(1-\gamma)}{2}[\phi(\theta_{m+1/2})(q_{m+1} - q_m) - \phi(\theta_{m-1/2})(q_m - q_{m-1})]. \end{aligned} \tag{4.88}$$

Under this circumstance, we have

$$C_{m-1} = \gamma + \frac{\gamma(1-\gamma)}{2}\left[\frac{\phi(\theta_{m+1/2})}{\theta_{m+1/2}} - \phi(\theta_{m-1/2})\right], \quad D_m^n = 0. \tag{4.89}$$

According to the Harten's theorem, we take $0 \leq C_{m-1} \leq 1$. Due to the CFL condition $0 \leq \gamma \leq 1$, it becomes

$$\left|\frac{\phi(\theta_1)}{\theta_1} - \phi(\theta_2)\right| \leq 2. \tag{4.90}$$

It is then enough to have

$$0 \leq \frac{\phi(\theta)}{\theta} \leq \text{minmod}\left(2, \frac{2}{\theta}\right). \tag{4.91}$$

This region for $\phi(\theta)$ is called the TVD region, see Figure 4.1. The upwind scheme $\phi(\theta) = 0$ lies in this region. The Lax-Wendroff scheme, the Beam-Warming scheme, and the Fromm's scheme are not completely contained in this region, therefore not TVD. The minmod limiter lies in the TVD region. We further remark that Sweby suggested that the flux limiter should be a convex combination of $\phi(\theta) = 1$ and $\phi(\theta) = \theta$.

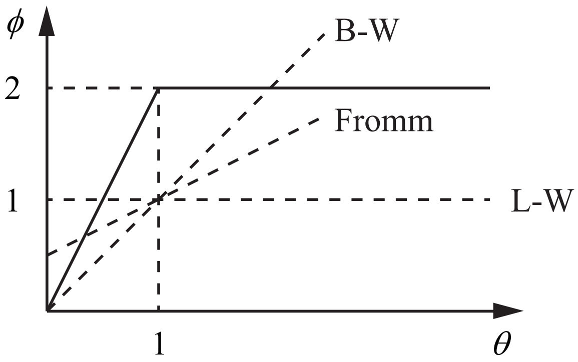


Figure 4.1: TVD choices for flux limiters.

4.5 Limiters for Systems

Similar to the scalar equations, a piecewise linear reconstruction with downwind slope for a linear system leads to the Lax-Wendroff scheme.

Consider a linear system

$$q_t + Aq_x = 0, \quad q \in \mathbb{R}^d. \tag{4.92}$$

The Lax-Wendroff scheme reads

$$Q_m^{n+1} = Q_m - \frac{k}{2h}A\left(Q_{m+1} - Q_{m-1}\right) + \frac{1}{2}\left(\frac{k}{h}\right)^2 A^2\left(Q_{m+1} - 2Q_m + Q_{m-1}\right). \tag{4.93}$$

The numerical flux function is

$$F_{m-1/2} = \frac{1}{2}A\left(Q_m + Q_{m-1}\right) - \frac{k}{2h}A^2\left(Q_m - Q_{m-1}\right). \tag{4.94}$$

The flux function can be decomposed into two parts, i.e., the flux function of upwind scheme (low order), and a high order correction.

$$\begin{aligned} F_{m-1/2} &= \frac{1}{2}A\left(Q_m + Q_{m-1}\right) - \frac{k}{2h}A^2\left(Q_m - Q_{m-1}\right) \\ &= A^-Q_m + A^+Q_{m-1} + \frac{1}{2}|A|\left(I - \frac{k}{h}|A|\right)\left(Q_m - Q_{m-1}\right). \end{aligned} \tag{4.95}$$

The first term is the upwind flux function, and the second one the high order correction.

The general form for a high resolution method using a flux limiter $\varphi_{m-1/2}$ may be expressed as follows.

$$F_{m-1/2} = F_L(q_{m-1}, q_m) + \varphi_{m-1/2}\left[F_H(q_{m-1}, q_m) - F_L(q_{m-1}, q_m)\right]. \tag{4.96}$$

Here F_L represents flux function with low order of accuracy, and F_H is the flux function with high accuracy. In a smooth region, essentially we have

$\varphi_{m-1/2} \approx 1$, and hence $F_{m-1/2} \approx F_H$. In a region with sharp numerical gradient, we have $\varphi_{m-1/2} \approx 0$, which leads to $F^n_{m-1/2} \approx F_L$.

To develop flux limiters for the linear system, we diagonalize the Jacobian matrix by $R = (r^{(1)}, \cdots, r^{(d)}$, to get $R^T A R = \Lambda$. Here $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_d)$. The new variable $\tilde{q} = R^T q$ satisfies the decoupled linear equation.

$$\tilde{q}_t + \Lambda \tilde{q}_x = 0. \tag{4.97}$$

For $\tilde{q}$, the flux function is

$$\tilde{F}^n_{m-1/2} = \Lambda^- \tilde{Q}_m + \Lambda^+ \tilde{Q}_{m-1} + \frac{1}{2}|\Lambda| \left(1 - \frac{k}{h}|\Lambda|\right) \phi\left(\theta_{m-1/2}\right) \Delta\tilde{Q}_{m-1/2}. \tag{4.98}$$

Here the p-th entry of the vector $\Delta\tilde{Q}_{m-1/2} = \tilde{Q}_m - \tilde{Q}_{m-1} = R^T \Delta Q_{m-1/2}$ is $\tilde{\alpha}^{(p)}_{m-1/2}$. The limiter $\phi\left(\theta_{m-1/2}\right)$ denotes a diagonal matrix, composed by slope limiters $\phi\left(\dfrac{\tilde{\alpha}^{(p)}_{I-1/2}}{\tilde{\alpha}^{(p)}_{m-1/2}}\right)$. We let $\alpha^{(p)}_{m-1/2} = \tilde{\alpha}^{(p)}_{m-1/2}\phi\left(\dfrac{\tilde{\alpha}^{(p)}_{I-1/2}}{\tilde{\alpha}^{(p)}_{m-1/2}}\right)$.

Then back to the original variable q, the flux function is

$$\begin{aligned}
F_{m-1/2} &= R\tilde{F}_{m-1/2} \\
&= A^- Q_m + A^+ Q_{m-1} + \frac{1}{2} R|\Lambda| \left(1 - \frac{k}{h}|\Lambda|\right) \phi\left(\theta_{m-1/2}\right) \Delta\tilde{Q}_{m-1/2} \\
&= A^- Q_m + A^+ Q_{m-1} + \frac{1}{2} R|\Lambda| \left(1 - \frac{k}{h}|\Lambda|\right) \begin{pmatrix} \vdots \\ \alpha^{(p)}_{m-1/2} \vdots \end{pmatrix} \\
&= A^- Q_m + A^+ Q_{m-1} + \frac{1}{2} \sum_{p=1}^{d} |\lambda_p| \left(1 - \frac{k}{h}|\lambda_p|\right) \alpha^{(p)}_{m-1/2} r^{(p)} \\
&= A^- Q_m + A^+ Q_{m-1} + \frac{1}{2} \sum_{p=1}^{d} |\lambda_p| \left(1 - \frac{k}{h}|\lambda_p|\right) w^{(p)}_{m-1/2}.
\end{aligned} \tag{4.99}$$

Here the wave decomposition for $\Delta Q_{m-1/2}$ gives $w^{(p)}_{m-1/2} = \alpha^{(p)}_{m-1/2} r^{(p)}$.

Next, we consider a nonlinear system

$$q_t + f(q)_x = 0. \tag{4.100}$$

For each cell problem, we define $\hat{A}_{m-1/2}$ by Roe linearization. We then apply the limiters for the approximate linear equations.

$$q_t + \hat{A}_{m-1/2} q_x = 0. \tag{4.101}$$

We decompose the flux difference into left-going and right-going waves. That is,

$$\begin{aligned} f(Q_m) - f(Q_{m-1}) &= A_{m-1/2}\Delta Q_{m-1/2} \\ &= \mathcal{A}^-_{m-1/2}\Delta Q_{m-1/2} + \mathcal{A}^+_{m-1/2}\Delta Q_{m-1/2}. \end{aligned} \tag{4.102}$$

The first order upwind scheme reads

$$Q_m^{n+1} = Q_m - \frac{k}{h}(\mathcal{A}^-\Delta Q_{m+1/2} + \mathcal{A}^+\Delta Q_{m-1/2}), \tag{4.103}$$

$$\mathcal{A}^-\Delta Q_{m+1/2} = \sum_p (s^{(p)}_{m+1/2})^- w^{(p)}_{m+1/2}, \tag{4.104}$$

$$\mathcal{A}^+\Delta Q_{m-1/2} = \sum_p (s^{(p)}_{m-1/2})^+ w^{(p)}_{m-1/2}. \tag{4.105}$$

We include further a correction term to obtain a high resolution scheme.

$$Q_m^{n+1} = Q_m - \frac{k}{h}(\mathcal{A}^-\Delta Q_{m+1/2} + \mathcal{A}^+\Delta Q_{m-1/2}) - \frac{k}{h}(F^c_{m+1/2} - F^c_{m-1/2}). \tag{4.106}$$

Here we take the correction

$$F^c_{m-1/2} = \frac{1}{2}\sum_p |s^{(p)}_{m-1/2}|(1 - \frac{k}{h}|s^{(p)}_{m-1/2}|)w^{(p)}_{m-1/2}. \tag{4.107}$$

Here $s^{(p)}_{m-1/2}$ may be chosen as the p-th eigenvalue obtained from the Roe linearization, possibly with an entropy fix to capture the correct rarefaction. The wave limiter $w^{(p)}_{m-1/2}$ is computed in the following way. First, from the decomposition of $\Delta Q_{m-1/2}$ we obtain the p-th component $\tilde{w}^{(p)}_{m-1/2} = \tilde{\alpha}^{(p)}_{m-1/2} r^{(p)}_{m-1/2}$. Next, we compute

$$\theta^{(p)}_{m-1/2} \equiv \frac{\tilde{w}^{(p)}_{I-1/2} \cdot \tilde{w}^{(p)}_{m-1/2}}{\tilde{w}^{(p)}_{m-1/2} \cdot \tilde{w}^{(p)}_{m-1/2}}. \tag{4.108}$$

Then we define

$$w^{(p)}_{m-1/2} = \tilde{\alpha}^{(p)}_{m-1/2} r^{(p)} \varphi(\theta^{(p)}_{m-1/2}). \tag{4.109}$$

4.6 Some Other Approaches: Relaxation Method and the Glimm Scheme

During the development of numerical algorithms for hyperbolic conservation laws, there are some other approaches besides the eventually dominant finite volume method with Roe linearization and flux limiter. Here we briefly describe two of them, the relaxation method and the Glimm scheme.

4.6.1 Relaxation Method

We illustrate the method with a scalar equation

$$u_t + f(u)_x = 0. \tag{4.110}$$

Consider a system with a small parameter $\epsilon > 0$, and a artificial wave speed λ.

$$\begin{cases} u^\epsilon_t + v^\epsilon_x = 0, \\ v^\epsilon_t + \lambda^2 u^\epsilon_x = \frac{1}{\epsilon}(f(u^\epsilon) - v^\epsilon). \end{cases} \tag{4.111}$$

When $\epsilon \to 0$, we expect that the source term in the second equation should not be singular, hence $f(u^\epsilon) - v^\epsilon \to 0$. Substituting this into the first equation, we recover the original equation.

This may be explained by a formal theory of Chapman-Enskog expansion.

$$\begin{aligned} 0 &= u^\epsilon_t + v^\epsilon_x \\ &= u^\epsilon_t + \left(f(u^\epsilon) - \epsilon(v^\epsilon_t + \lambda^2 u^\epsilon_x)\right)_x \\ &= u^\epsilon_t + f(u^\epsilon)_x - \epsilon\left(f(u^\epsilon)_t + \lambda^2 u^\epsilon_x\right)_x + O(\epsilon^2) \\ &= u^\epsilon_t + f(u^\epsilon)_x - \epsilon\left(f'(u^\epsilon)u^\epsilon_t + \lambda^2 u^\epsilon_x\right)_x + O(\epsilon^2) \\ &= u^\epsilon_t + f(u^\epsilon)_x - \epsilon\left(f'(u^\epsilon)(-v^\epsilon_x) + \lambda^2 u^\epsilon_x\right)_x + O(\epsilon^2) \\ &= u^\epsilon_t + f(u^\epsilon)_x - \epsilon\left(f'(u^\epsilon)(-f(u^\epsilon)_x) + \lambda^2 u^\epsilon_x\right)_x + O(\epsilon^2) \\ &= u^\epsilon_t + f(u^\epsilon)_x - \epsilon\left[\left(\lambda^2 - f'(u^\epsilon)^2\right) u^\epsilon_x\right]_x + O(\epsilon^2). \end{aligned} \tag{4.112}$$

From this expansion, we observed that under a subcharacteristic condition $\lambda > |f'(u)|$, the relaxation method yields a viscous term on the order of ϵ. As a matter of fact, this is a over simplified version for the derivation of the macroscopic Euler equations or the Navier-Stokes equations from the microscopic Boltzmann equation.

To solve this model numerically, we adopt the splitting technique. That is, for a cell problem, we first compute a wave propagation step

$$\begin{cases} \widetilde{u}^\epsilon_t + \widetilde{v}^\epsilon_x = 0, \\ \widetilde{v}^\epsilon_t + \lambda^2 \widetilde{u}^\epsilon_x = 0, \end{cases} \tag{4.113}$$
$$\widetilde{u}^\epsilon(x, t^n) = u^n_m, \quad \widetilde{v}^\epsilon(x, t^n) = v^n_m.$$

We obtain $\widetilde{u}^\epsilon(x, t^{n+1}), \widetilde{v}^\epsilon(x, t^{n+1})$. Then we make a relaxation step

$$\begin{cases} \bar{u}^\epsilon_t = 0, \\ \bar{v}^\epsilon_t = \frac{1}{\epsilon}\left(f(\bar{u}^\epsilon) - \bar{v}^\epsilon\right), \end{cases} \tag{4.114}$$
$$\bar{u}^\epsilon(x, t^n) = \widetilde{u}^\epsilon(x, t^{n+1}), \quad \bar{v}^\epsilon(x, t^n) = \widetilde{v}^\epsilon(x, t^{n+1}).$$

Then we define

$$u^\epsilon(x, t^{n+1}) = \bar{u}^\epsilon(x, t^{n+1}), \quad v^\epsilon(x, t^{n+1}) = \bar{v}^\epsilon(x, t^{n+1}). \tag{4.115}$$

In the second step, $\epsilon \to 0$ leads to a set of stiff ordinary differential equations. Due to the special structure, it can be exactly solved. That is,

$$\bar{u}^{\epsilon}(x,t) = \widetilde{u}^{\epsilon}(x,t^{n+1}), \tag{4.116}$$

and

$$\bar{v}^{\epsilon}(x,t) = (1-e^{-t/\epsilon})f(\bar{u}^{\epsilon}) + e^{-t/\epsilon}\bar{v}^{\epsilon}(t^n). \tag{4.117}$$

In particular, we may even take $\epsilon \to 0^+$ in this step to get $\bar{v}^{\epsilon}(x,t^{n+1}) \to f(\bar{u}^{\epsilon}(x,t^n))$. This leads to a relaxed scheme. Notice that the relaxed scheme still possesses viscosity due to the splitting.

4.6.2 Glimm Scheme

Consider again the scalar equation as an example with $f''(u) > 0$. We use a staggered grid to solve the problem. Suppose that we start with a uniformly distributed grid points x_m. We regard $[x_{m-1}, x_{m+1}]$ with odd m as a cell, and the initial data is a constant in this cell. With a time step size k that satisfies the stability condition, we have the exact Riemann solution $u(x,t)$ ($t \in [t^0, t^1)$) for each pair of neighboring cells. Moreover, the waves are confined within $[x_m, x_{m+2}]$ for the cell pair $[x_{m-1}, x_{m+1}]$ and $[x_{m+1}, x_{m+3}]$. Choose a random number $a_1 \in [0,1]$. We set $y_m = x_m + 2a_1 h \in [x_m, x_{m+2}]$ for each odd m. We further let $u(x,t^1) = u(y_m, t^0)$ for $x \in [x_m, x_{m+2}]$ with odd m. This forms a piecewise constant data at t^1, and the cell for this layer is $[x_{m-1}, x_{m+1}]$ with even m. Taking the same procedure with another random number $a_2 \in [0,1]$, we may solve for the time step $[t^1, t^2]$. Noticing that actually the solution u is bounded by the initial data, the time step size may be chosen as a uniform one. Therefore, we may repeat the procedure and obtain a numerical solution.

This random choice of data produces a convergent numerical solution. The scheme is introduced by J. Glimm. Proposed as a numerical scheme, this method is not popular for numerical computations. However, it is important for proving the existence of solution to hyperbolic conservation laws. Actually it gave one of the first general proof. Furthermore, partly based on this idea, A. Bressan developed one of the most general theory for nonlinear hyperbolic systems.

4.7 Multidimensional Hyperbolic Problems

Most important applications require numerical computations in multiple dimensions. Fortunately and unfortunately, the theory for multidimensional hyperbolic problems is not complete. This brings challenges as well as opportunities to numerical studies.

4.7.1 Hyperbolicity

We start with a definition of hyperbolic problem in two space dimensions. It may be readily generalized to multiple dimensions.

Definition 4.1. *The system $q_t + f(q)_x + g(q)_y = 0$ is strongly hyperbolic if $\forall \vec{n}$, the Jacobian matrix $(\nabla_q f(q), \nabla_q g(q)) \cdot \vec{n}$ is diagnalizable with real eigenvalues. That is, the system is hyperbolic in every direction.*

The p-system in two space dimensions is hyperbolic.

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2 + p)_x + (\rho uv)_y = 0, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2 + p)_y = 0. \end{cases} \tag{4.118}$$

In any direction $\vec{n}$, we have

$$\begin{cases} \rho_t + \frac{\partial \rho(\vec{u}\cdot\vec{n})}{\partial((x,y)\cdot\vec{n})} = 0, \\ (\rho\vec{u}\cdot\vec{n})_t + \frac{\partial(\rho(\vec{u}\cdot\vec{n})^2+p)}{\partial((x,y)\cdot\vec{n})} = 0. \end{cases} \tag{4.119}$$

This reflects the fact that the physical laws of conservation are coordinate independent.

The directional version of p-system leads to

$$\begin{aligned} q = \begin{pmatrix} \rho \\ \rho u \\ \rho v \end{pmatrix}, \quad f(q) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \end{pmatrix} = \begin{pmatrix} q^2 \\ (q^2)^2/q^1 + p(q^1) \\ q^2 q^3/q^1 \end{pmatrix}, \\ g(q) = \begin{pmatrix} q^3 \\ q^2 q^3/q' \\ (q^3)^2/q^1 + p(q^1) \end{pmatrix}. \end{aligned} \tag{4.120}$$

The Jacobian matrices are

$$\nabla_q f(q) = \begin{bmatrix} 0 & 1 & 0 \\ -u^2 + p'(\rho) & 2u & 0 \\ -uv & v & u \end{bmatrix}, \quad \nabla_q g(q) = \begin{bmatrix} 0 & 0 & 1 \\ -uv & v & u \\ -v^2 + p'(\rho) & 0 & 2v \end{bmatrix}. \tag{4.121}$$

Then we obtain the Jacobian matrix along $\vec{n}$ is

$$\begin{aligned} &(\nabla_q f(q), \nabla_q g(q)) \cdot \vec{n} \\ &= \begin{bmatrix} 0 & n_x & n_y \\ n_x(-u^2 + p'(\rho)) - n_y uv & 2un_x + vn_y & un_y \\ -n_x uv + n_y(-v^2 + p'(\rho)) & vn_x & un_x + 2vn_y \end{bmatrix}. \end{aligned} \tag{4.122}$$

It has three distinct eigenvalues, $un_x + vn_y \pm c_0, un_x + vn_y$, where $c_0 = \sqrt{p'(\rho_0)}$ is the sound speed.

4.7.2 Numerical Methods

A most straightforward way to solve a multidimensional problem is by dimension splitting. That is, for one step computation of the equation $q_t + f(q)_x + g(q)_y = 0$, we perform two sub-steps.

$$\begin{cases} q_t + f(q)_x = 0 & x\text{-sweep}, \\ q_t + g(q)_y = 0 & y\text{-sweep}. \end{cases} \tag{4.123}$$

More precisely, we take

$$q^{\star}_{mj} = q^n_{mj} - \frac{k}{h}(F^n_{m+1/2,j} - F^n_{m-1/2,j}), \tag{4.124}$$

$$q^{n+1}_{mj} = q^{\star}_{mj} - \frac{k}{h}(G^{\star}_{m,j+1/2} - G^{\star}_{m,j-1/2}). \tag{4.125}$$

One may adopt the Strang splitting to reach second order splitting accuracy.

$$q^{\star}_{mj} = q^n_{mj} - \frac{k}{2h}(F^n_{m+1/2,j} - F^n_{m-1/2,j}), \tag{4.126}$$

$$q^{\star\star}_{mj} = q^{\star}_{mj} - \frac{k}{h}(G^{\star}_{m,j+1/2} - G^{\star}_{m,j-1/2}), \tag{4.127}$$

$$q^{n+1}_{mj} = q^{\star\star}_{mj} - \frac{k}{2h}(F^{\star\star}_{m+1/2,j} - F^{\star\star}_{m-1/2,j}). \tag{4.128}$$

There is a second approach. Consider a semi-discrete system

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} q_{mj}(t) = &- \frac{1}{\Delta x}(F_{m+1/2,j}(q) - F_{m-1/2,j}(q)) \\ &- \frac{1}{\Delta y}(G_{m,j+1/2}(q) - G_{m,j-1/2}(q)). \end{aligned} \tag{4.129}$$

This is then solved by the Runge-Kutta method.

A third approach starts with the finite volume methodology. It is a fully discrete flux difference method. First, we have the exact formula after integration by parts.

$$\begin{aligned} &\iint_{\Omega_{mj}} \frac{\mathrm{d}}{\mathrm{d}t} q_{mj}(t) \\ =& \int_{y_{j-1/2}}^{y_{j+1/2}} f(q(x_{m-1/2}, y, t))\mathrm{d}y - \int_{y_{j-1/2}}^{y_{j+1/2}} f(q(x_{m+1/2}, y, t))\mathrm{d}y \\ +& \int_{x_{m-1/2}}^{x_{m+1/2}} g(q(x, y_{j-1/2}, t))\mathrm{d}x - \int_{x_{m-1/2}}^{x_{m+1/2}} g(q(x, y_{j+1/2}, t))\mathrm{d}x. \end{aligned} \tag{4.130}$$

Finite volume scheme reads

$$q^{n+1}_{mj} = q^n_{mj} - \frac{k}{h_x}(F^n_{m+1/2,j} - F^n_{m-1/2,j}) - \frac{k}{h_y}(G^n_{m,j+1/2} - G^n_{m,j-1/2}), \tag{4.131}$$

with numerical fluxes

$$F^n_{m-1/2,j} \approx \frac{1}{kh_y}\int_{t^n}^{t^{n+1}}\int_{y_{j-1/2}}^{y_{j+1/2}} f(q(x_{m-1/2},y,t))\mathrm{d}y\mathrm{d}t, \tag{4.132}$$

$$G^n_{m,j-1/2} \approx \frac{1}{kh_x}\int_{t^n}^{t^{n+1}}\int_{x_{m-1/2}}^{x_{m+1/2}} g(q(x,y_{j-1/2},t))\mathrm{d}x\mathrm{d}t. \tag{4.133}$$

The design of numerical fluxes is similar to that in one dimension.

Assignments

1. Perform the von Neumann analysis to the Lax-Friedrichs scheme to study its stability.

2. Find the modified equation for the Lax-Friedrichs scheme to study its stability.

3. Perform numerical experiments with the schemes (3.17) and (3.20), respectively for initial data

$$u(x,0) = \begin{cases} 2, & x<0, \\ 1, & x>0, \end{cases} \tag{4.134}$$

Compare the numerical results. Also compare the numerical results for initial data

$$u(x,0) = \begin{cases} 1, & x<0, \\ 2, & x>0. \end{cases} \tag{4.135}$$

4. Prove the Jensen's inequality

$$\eta^{''}(q) \geq 0 \Rightarrow \eta(\frac{1}{h}\int_{C_m}\tilde{q}^n(x)\mathrm{d}x) \leq \frac{1}{h}\int_{C_m}\eta(\tilde{q}^n(x))\mathrm{d}x. \tag{4.136}$$

Construct an example to show the inequality does not hold if $\eta^{''}(q) \geq 0$ does not hold.

5. Consider a scheme for the linear advection equation

$$u^{n+1}_m = \sum_j \alpha_j u^n_{m+j}. \tag{4.137}$$

Show that if $\alpha_j \geq 0$ $(\forall j)$, then this scheme is at most of the first order accuracy except for the special case of $u^{n+1}_m = u^n_{m-l}$ with $ck = lh$.

6. Prove

$$AQ_{m-1/2} = AQ_m - \sum_p \lambda^+_p w^{(p)}_{m-1/2}. \tag{4.138}$$

7. Compute with the Roe linearization for the inviscid Burgers' equation $u_t + (\frac{u^2}{2})_x = 0$, with initial data $u(x,0) = \begin{cases} -1, & x < 0, \\ 2, & x > 0. \end{cases}$ Compare the results with and without the Harten's entropy fix.

8. Derive and compute with the Roe linearization for the polytropic Euler equations, with an initial data $(\rho, u)(x,0) = (1, \exp(-25x^2))$.

9. For the shallow water equations, derive the numerical flux with minmod limiter.

10. Compute for the shallow water equations with initial data

$$(h,u)(x,0) = \begin{cases} (1,2), & x < 0, \\ (5,0), & x > 0. \end{cases} \tag{4.139}$$

Chapter 5

Introduction to Finite Element Method

5.1 Sobolev Spaces

As we have learned for the hyperbolic equations, classical solutions may not exist in general. This holds true for another type of equations, namely, elliptic partial differential equations. To investigate this type of equations, the most appropriate space of functions falls in to the category of Sobolev spaces. We shall confine ourselves to a special sub-category of spaces, namely, the Hilbert spaces.

We consider $\Omega \in \mathbb{R}^n$ a bounded open set with piecewise smooth boundary. Moreover, a cone condition is assumed in most applications to elliptic partial differential equations. That is, at any point of $\partial\Omega$, a cone with positive inner angle is locally contained within Ω.

A Hilbert space is a complete space with inner product. Depending on the inner product defined, we have a sequence of Hilbert spaces for functions.

We start with $H^0(\Omega)$, which is actually $L_2(\Omega)$. We define an inner product

$$(u, v)_0 = \int_\Omega u(x)v(x)\mathrm{d}x. \tag{5.1}$$

This leads to an L^2 norm $\parallel u \parallel_0 = \sqrt{(u, u)_0} = \int_\Omega u(x)^2\mathrm{d}x$. The $L^2(\Omega)$ is the collection of all functions with finite L^2-norm.

We notice that a function in $L^2(\Omega)$ may not be differentiable in general. So, we define a weak derivative instead. For this purpose, we define $C^\infty(\Omega)$ the smooth function space, and $C_0^\infty(\Omega)$ the subspace with each element taking a compact support. Because Ω is open, and a compact set in $\mathbb{R}$ must be bounded closed, we know that a function in $C_0^\infty(\Omega)$ must vanish at boundary.

For $u \in L^2$, we define its weak derivative $v = \partial^\alpha u \in L^2$, where α is a multiple index, if

$$(\phi, v)_0 = (-1)^{|\alpha|}(\partial^\alpha \phi, u)_0, \tag{5.2}$$

$\forall \phi \in \mathbb{C}_0^\infty(\Omega)$. We further define an inner product and corresponding norm as follows.

$$(u, v)_m = \int_\Omega \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v) \mathrm{d}x, \quad \| u \|_m = \sqrt{(u, u)_m}. \tag{5.3}$$

A semi-norm may be defined by

$$|u|_m = \sqrt{\sum_{|\alpha|=m} \| \partial^\alpha u \|_0^2}. \tag{5.4}$$

The Hilbert space is defined as $H^m(\Omega) = \{u \in L^2(\Omega) | \ \| u \|_m < +\infty\}$. The completeness may be verified easily. It is possible to show that $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$, and $H^m(\Omega)$ is the completion of $C^\infty(\Omega) \cap H^m(\Omega)$ under the H^m norm.

In a similar way, we define $H_0^m(\Omega)$ the completion of $C_0^\infty(\Omega)$. There are two sequences of inclusion.

$$L_2(\Omega) = H^0(\Omega) \supset H^1(\Omega) \supset H^2(\Omega) \supset \cdots, \tag{5.5}$$

$$H_0^0(\Omega) \supset H_0^1(\Omega) \supset H_0^2(\Omega) \supset \cdots. \tag{5.6}$$

In particular, norm and the seminorm $||_m$ are equivalent in $H_0^m(\Omega)$ by the following result based on the Poincare-Friedrichs inequality.

$$|v|_m \leq \| v \|_m \leq (1+s)^m |v|_m, \forall v \in H_0^m(\Omega). \tag{5.7}$$

Here we assume that Ω is contained in a cube with side with length s.

A key fact in Sobolev spaces is the compact imbedding. For $m > 0$ and Ω a Lipschitiz domain with cone condition, $H^{m+1}(\Omega) \hookrightarrow H^m(\Omega)$ is a compact imbedding, namely, a subset which is bounded in H^{m+1} is relatively compact in H^m. The compact imbedding facilitates theoretical studies of elliptic and parabolic partial differential equations, such as the existence and regularity of the solutions.

5.2 Variational Formulation of Second-order Elliptic Equations

A simple example of elliptic partial differential equation is the Laplace equation

$$\Delta u = u_{xx} + u_{yy} = 0, \quad (x, y) \in \Omega. \tag{5.8}$$

Notice that a certain boundary condition is necessary.

More generally, we may consider an elliptic operator

$$Lu := \sum_{i,k=1}^{d} a_{ik}(x) u_{x_i x_k}, \tag{5.9}$$

where the coefficient square matrix $\{a_{ik}\}$ is positive definite. Moreover, if $\exists \alpha > 0, \forall x \in \Omega$, it holds that $\min \lambda \geq \alpha$, we call L uniformly elliptic. For such an operator, some features are as follows.

- **Minimum principle.**
  If $Lu = f \leq 0$, then u attains its minimum on $\partial\Omega$.

- **Comparison principle.**
  If two classical solutions $u, v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ satisfy $Lu \leq Lv$ in Ω, and $u \geq v$ on $\partial\Omega$, then it holds that $u \geq v$ in Ω.

- **Continuous dependency on the boundary data.**
  For two solutions of $Lu_i = f$ $(i = 1, 2)$ with different boundary data, it holds that $\sup_{x \in \Omega} |u_1(x) - u_2(x)| = \sup_{z \in \partial\Omega} |u_1(z) - u_2(z)|$.

- **Continuous dependency on the righthand side.**
  $\forall u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, it holds that $|u(x)| \leq \sup_{\partial\Omega} |u(z)| + C \sup_{\partial\Omega} |Lu(z)|$.

- **Elliptic operator with Helmholtz term.**
  For operator $Lu := -\sum a_{ik}(x) u_{x_i x_k} + c(x)u$ with $c(x) \geq 0$, if $Lu \leq 0$, then $\sup_{x \in \Omega} u(x) \leq \max\{0, \sup_{\partial\Omega} u(z)\}$

- **Elliptic operator in divergence form.**
  For operator $Lu = -\sum_{i,k} \partial_i (a_{ik} \partial_k u) + a_0 u$, with $(a_{ik}), a_0(x) \geq 0$, we may take an associated bilinear form as follows.

$$a(u, v) = \int_\Omega (\sum_{i,k} a_{ik} \partial_i u \partial_k v + a_0 uv) \mathrm{d}x. \tag{5.10}$$

As a classical solution does not exist in general, we consider a weak solution instead.

Definition 5.1. *$u \in H_0^1(\Omega)$ is a weak solution of $Lu = f$ in Ω and $u = 0$ on $\partial\Omega$ if*

$$a(u, v) = (f, v)_0, \forall v \in H_0^1(\Omega). \tag{5.11}$$

The finite element method starts with transforming the partial differential equation into a minimization problem. We have the following theorem that relates an equation with a minimization for a bilinear form. In a sense, it is similar to the Fermat's theorem in calculus, which states that an extremum point must be a critical point.

Theorem 5.1. *(Characterization theorem) Let V be a linear space, $a : V \times V \to \mathbb{R}$ be a symmetric positive bilinear form, and $l : V \to \mathbb{R}$ be a linear functional denoted as $\langle l, v\rangle = l(v)$. Then*

$$J(v) = \frac{1}{2}a(v,v) - < l, v > \tag{5.12}$$

attains its minimum over V at u if and only if

$$a(u,v) =< l, v >, \forall v \in V. \tag{5.13}$$

Moreover, there is at most one solution.

Proof. Take $u, v \in V, t \in \mathbb{R}$, we compute

$$\begin{aligned} J(u+tv) &= \frac{1}{2}a(u+tv, u+tv) - < l, u+tv > \\ &= J(u) + t[a(u,v) - < l, v >] + \frac{1}{2}t^2 a(v,v). \end{aligned} \tag{5.14}$$

On one hand, if $a(u,v) =< l, v >, \forall v \in V$, then we have

$$J(u+v) = J(u) + \frac{1}{2}t^2 a(v,v) > J(u), \ \forall v \in V \text{ and } v \neq 0. \tag{5.15}$$

On the other hand, if J has a minimum at u, then $\forall v$, we consider the function $t \longmapsto J(u+tv)$. By the Fermat's theorem we have

$$\frac{\mathrm{d}}{\mathrm{d}t}J(u+tv)|_{t=0} = a(u,v) - < l, v >= 0. \tag{5.16}$$

Finally, we prove the uniqueness. If two solutions satisfy $a(u_1, v) =< l, v >= a(u_2, v)$, then $J(u_1)$ and $J(u_2)$ are both minimum. So $J(u_1 + (u_2 - u_1)) > J(u_1)$ leads to a contradiction. □

The theorem implies that every classical solution of the boundary values problem

$$\begin{cases} -\sum_{i,k} \partial_i(a_{ik}\partial_k u) + a_0 u = f, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \tag{5.17}$$

is a solution of the variational problem $v \in C^2(\Omega) \cap C^0(\bar{\Omega})$ with $v|_{\partial\Omega} = 0$.

$$J(v) = \int_\Omega \left[\frac{1}{2}\sum a_{ik}\partial_i v \partial_k v + \frac{1}{2}a_0 v^2 - fv\right] \mathrm{d}x \to \min!. \tag{5.18}$$

The following Lax-Milgram theorem is key to the understanding of elliptic partial differential equations.

Theorem 5.2. *(Lax-Milgram) Let V be a closed convex set in a Hilbert space H, and $a : H \times H \to \mathbb{R}$ be an elliptic bilinear form. Then $\forall l \in H'$ (dual space of H), the variational problem*

$$J(v) = \frac{1}{2}a(v,v) - < l, v > \to \min! \tag{5.19}$$

has a unique solution in V.

Proof. We claim that J is bounded from below. As a matter of fact, due to the ellipticity and the definition of dual space, we have

$$J(v) \geq \frac{1}{2}\alpha \parallel v \parallel^2 - \parallel l \parallel \parallel v \parallel = \frac{1}{2\alpha}(\alpha \parallel v \parallel - \parallel l \parallel)^2 - \frac{\parallel l \parallel^2}{2\alpha} \geq -\frac{\parallel l \parallel^2}{2\alpha}. \tag{5.20}$$

Therefore, there exists $\inf J(v) = C_1$. Let (v_n) be a minimizing sequence. We derive

$$\begin{aligned} \alpha \parallel v_n - v_m \parallel^2 &\leq a(v_n - v_m, v_n - v_m) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m) \\ &= 4J(v_n) + 4J(v_m) - 8J(\frac{v_n + v_m}{2}) \\ &\leq 4J(v_n) + 4J(v_m) - 8C_1. \end{aligned} \tag{5.21}$$

Here we have made usage of the convexity of V to derive that $\frac{v_n+v_m}{2} \in V$. The above term tends to 0 as $n, m \to \infty$. Now combining the facts that (v_n) is A Cauchy sequence, H is complete, and V is closed, we conclude that

$$u = \lim_{n\to\infty} v_n \in V. \tag{5.22}$$

Furthermore, as J is continuous, $J(u) = \lim\limits_{n\to\infty} J(v_n) = C_1 = \inf\limits_{v\in V} J(v)$.

Next, we prove the uniqueness. To this end, assuming that u_1, u_2 are both solutions, we may construct a sequence $(u_1, u_2, u_1, u_2, \cdots)$. It is obviously a minimizing sequence. It then must be a Cauchy sequence, and hence $u_1 = u_2$. □

We remark that the difference between the characterization theorem and the Lax-Milgram lies in the difference of the space. In the previous one, the whole Hilbert space is adopted. On the other hand, the Lax-Milgram theorem uses only a convex closed subset.

Theorem 5.3. *(Existence) Let L be a second order uniformly elliptic operator, with $a_0, a_{ij} \in L_\infty(\Omega), a_0 \geq 0, f \in L_2(\Omega)$. The boundary value problem*

$$\left\{ \begin{array}{ll} Lu = f, & in\, \Omega, \\ u = 0, & on\, \partial\Omega, \end{array} \right. \tag{5.23}$$

admits a weak solution in $H_0^1(\Omega)$. It is a minimum of the variational problem

$$\frac{1}{2}a(v,v) - (f,v)_0 \to \min! \quad \textit{over} \quad H_0^1(\Omega). \tag{5.24}$$

Proof. It is possible to show that

$$\left|\sum_{i,k}\int a_{ik}\partial_i u \partial_k v \mathrm{d}x\right| \leq C\sum_{i,k}\int |\partial_i u \partial_k v|\mathrm{d}x \leq C|u|_1|v|_1. \tag{5.25}$$

Furthermore, we have

$$\left|\int a_0 uv\mathrm{d}x\right| \leq C \parallel u \parallel_0 \parallel v \parallel_0 . \tag{5.26}$$

These lead to

$$a(u,v) \leq C \parallel u \parallel_1 \parallel v \parallel_1 . \tag{5.27}$$

For any $v \in C^1$, it holds that

$$\sum a_{ik}\partial_i v \partial_k v \geq \alpha \sum (\partial_i v)^2. \tag{5.28}$$

Therefore, for any $v \in H^1$, it holds that

$$a(v,v) \geq \alpha |v|_1^2. \tag{5.29}$$

From the Poincare-Friedrichs inequality, we know that

$$|v|_1 \sim \parallel v \parallel_1 . \tag{5.30}$$

Combining (5.27)-(5.30), we find that a is an elliptic bilinear form on $H_0^1(\Omega)$. Noticing that $f \in L^2 \subset H'$, we conclude the existence and uniqueness of the weak solution from the Lax-Milgram theorem. □

We remark that the above results may be extended to non-homogeneous Dirichlet boundary value problem. Consider

$$\begin{cases} Lu = f, & \text{in } \Omega, \\ u = g, & \text{on } \partial\Omega. \end{cases} \tag{5.31}$$

We assume $\exists u_0$, such that Lu_0 exists, and $u_0|_{\partial\Omega} = g$. Now let $w = u - u_0$, then w solves a homogeneous boundary value problem

$$\begin{cases} Lw = f - Lu_0 \equiv f_1, & \text{in } \Omega, \\ w = 0, & \text{on } \partial\Omega. \end{cases} \tag{5.32}$$

5.3 Neumann Boundary Value Problem

Suppose that $\nu = (\nu_i)$ is the out-normal on the boundary $\Gamma = \partial\Omega$. The Neumann boundary value problem refers to the following setting.

$$\begin{cases} Lu = f, & in\ \Omega, \\ \sum_{i,k} \nu_i a_{ik}\partial_k u = g, & on\ \Gamma. \end{cases} \tag{5.33}$$

It is obvious that the solution is not in $H_0^1(\Omega)$ in general. In fact, if u is a solution, so is $u + C$ with C a constant. The suitable function space is $H^1(\Omega)$.

Ellipticity in $H^1(\Omega)$ requires $a_{ik} \geq \alpha \geq 0$ and $a_0 \geq \alpha$ in Ω. Consequently, we see $\forall v \in H^1(\Omega)$,

$$a(v,v) = \int_\Omega \left[\sum a_{ik}\partial_i u \partial_k v + a_0 v^2\right] \mathrm{d}x \geq \alpha |v|_1^2 + \alpha \parallel v \parallel^2 = \alpha \parallel v \parallel_1^2 . \tag{5.34}$$

Using a certain trace theorem, it may be proved that $< l, v >= \int_\Omega f v \mathrm{d}x + \int_\Gamma g v \mathrm{d}s$ defines a bounded linear functional with $f, g \in L_2(\Gamma)$. Again, the boundary value problem may be transformed to a variational problem.

Theorem 5.4. *Suppose that Ω is a bounded domain with piecewise smooth boundary, and satisfying the cone condition, then the variational problem*

$$\frac{1}{2}a(v,v) - (f,v)_{0,\Omega} - (g,v)_{0,\Gamma} \to \min! \tag{5.35}$$

has a unique solution $u \in H^1(\Omega)$. Moreover,$u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ if classical solution exists for

$$\begin{cases} Lu = f, & in\ \Omega, \\ \sum_{i,k} \gamma_i a_{ik}\partial_k u = g, & on\ \Gamma. \end{cases} \tag{5.36}$$

As an example, the Poisson equation

$$\begin{cases} -\Delta u = f, & \text{in } \Omega, \\ \frac{\partial u}{\partial \gamma} = g, & \text{on } \Gamma, \end{cases} \tag{5.37}$$

admits a unique solution up to a constant. If we restrict the solution to $V = \{v \in H^1(\Omega), \int_\Omega v \mathrm{d}x = 0\}$, then uniqueness is obtained. In fact, the Poincare-Friedrichs inequality implies that $a(u,v) = \int_\Omega \nabla u \cdot \nabla v \mathrm{d}x$ is elliptic in V. We remark that a compatibility condition is required due to the Gauss theorem, namely, $\int_\Omega f \mathrm{d}x + \int_\Gamma g \mathrm{d}s = 0$.

We may also consider a mixed boundary value problem.

$$\begin{cases} -\Delta u = 0, & \text{in } \Omega, \\ u = g, & \text{on}\Gamma_D, \\ \frac{\partial u}{\partial \gamma} = 0, & \text{on } \Gamma_N. \end{cases} \tag{5.38}$$

The suitable function space is then $\bar{W}$ with $W = \{u \in C^\infty(\Omega) \cap H^1(\Omega),\ u$ vanishes in a hold of $\Gamma_D\}$, which is between $H_0^1(\Omega)$ and $H^1(\Omega)$.

5.4 The Ritz-Galerkin Method

In the previous sections, we relate the weak solution in function space H of a boundary value problem with the minimization of a functional J over H. A natural idea for approximation by the minimization of J over a subspace S_h, called a finite element space. Here h is a characteristic length scale of the grid size. This gives the Ritz method.

Consider

$$J(v) = \frac{1}{2}a(v,v) - < l, v > \rightarrow \min_{S_h}!. \tag{5.39}$$

It is easy to know that the solution u_h satisfies

$$a(u_h, v) =< l, v >, \forall v \in S_h. \tag{5.40}$$

Now let $\{\phi_1, \cdots, \phi_N\}$ be a basis of S_h. We find that

$$a(u_n, \phi_i) =< l, \phi_i >, i = 1, \cdots, N. \tag{5.41}$$

If we expand the numerical solution also in terms of the basis $u_h = \sum_k z_k \phi_k$, we find that

$$\sum_k a(\phi_k, \phi_i) z_k =< l, \phi_i > . \tag{5.42}$$

This forms an algebraic system

$$Az = b, \tag{5.43}$$

where the stiffness matrix is $A = a(\phi_k, \phi_i)$ and the source term is $b_i =< l, \phi_i >$.

Because a is elliptic, A is positive definite and therefore the above system has a unique solution.

$$\begin{aligned} z' Az &= \sum_{i,k} z_i A_{ik} z_k \\ &= a(\sum z_k \phi_k, \sum z_i \phi_i) \\ &= a(u_h, u_h) \\ &\geq \alpha \parallel u_h \parallel_m^2 . \end{aligned} \tag{5.44}$$

We remark that the boundary value determines which function space V to use. The functional is required to be V-elliptic, that is,

$$a(v,v) \geq \alpha \parallel v \parallel_m^2; |a(u,v)| \leq C \parallel u \parallel_m \parallel v \parallel_m, \ \forall u, v \in V. \tag{5.45}$$

There are other ways to formulate the approximation. For instance, one may directly do the minimization and obtain

$$\frac{\partial}{\partial z_i} J(\sum_k z_k \phi_k) = 0. \tag{5.46}$$

The distinct feature of the finite element method is the systematic theoretical results. For instance, global stability is readily obtained as follows. In fact, from

$$\alpha \parallel u_h \parallel_m^2 \leq a(u_h, u_h) =< l, u_h > \leq \parallel l \parallel \parallel u_h \parallel_m, \tag{5.47}$$

we obtain immediately that

$$\parallel u_n \parallel_m \leq \alpha^{-1} \parallel l \parallel . \tag{5.48}$$

Furthermore, the error bound is also straightforward to prove. This also implies convergence of the finite element method.

Theorem 5.5. *(Cea's Lemma) Consider a V-elliptic bilinear form $a(u, v)$ over a certain function space $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. Let $S_h \subset V$, then*

$$\parallel u - u_h \parallel_m \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \parallel u - v_h \parallel_m . \tag{5.49}$$

Proof. From

$$a(u, v) =< l, v >, \ \forall v \in V, \tag{5.50}$$

we know that

$$a(u_h, v) =< l, v >, \ \forall v \in S_h. \tag{5.51}$$

This implies the Galerkin orthogonality

$$a(u - u_h, v) = 0, \ \forall v \in S_h. \tag{5.52}$$

Let $v_h \in S_h$, and take $v = v_h - u_h \in S_h$. We find

$$\begin{aligned}\alpha \parallel u - u_h \parallel_m^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a\underbrace{(u - u_h, v_h - u_h)}_{=0} \\ &\leq C \parallel u - u_h \parallel_m \parallel v - v_h \parallel_m .\end{aligned} \tag{5.53}$$

We conclude

$$\parallel u - u_h \parallel_m \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \parallel u - v_h \parallel_m . \tag{5.54}$$

□

From the theorem, we notice that the accuracy is up to the approximation error of u in S_h. Piecewise polynomials is adopted in general for u.

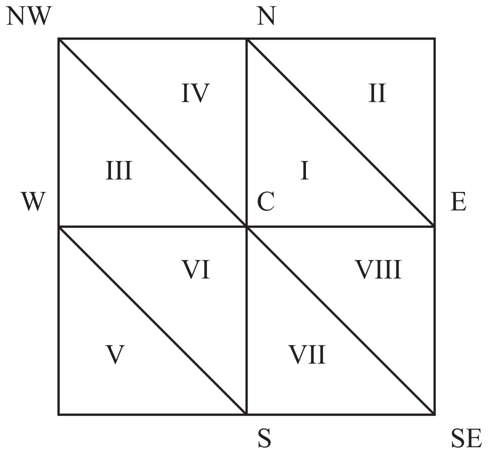


Figure 5.1: Schematic view of triangulation.

5.5 A Simple Example

To explain better the finite element method, we work out explicitly a model problem. Consider

$$\begin{cases} -\Delta u = f, & in\ \Omega = (0,1)^2, \\ u = 0 & on\ \partial\Omega. \end{cases} \tag{5.55}$$

We take the finite element space

$$S_h = \{v \in C(\bar{\Omega}) : v \text{ is a linear function in every triangle}, v|_{\partial\Omega} = 0\}. \tag{5.56}$$

The basis function in this space is denoted as $\{\phi_i\}_{i=1}^N$, where N is the number of mesh points inside the domain, and $\phi_i(x_j, y_j) = \delta_{ij}$ at every nodal points (x_j, y_j). These are tent functions. Consider a nodal point, denoted as C. Then ϕ_C is nonzero in regions $I, III, IV, VI, VII, VIII$. We refer to its neighboring vertices as E, W, S, N, NW, SE. It is obvious that $a(\phi_C, \cdot)$ is zero except for ϕ's corresponding to these neighboring vertices.

We first compute

$$\begin{aligned} a(\phi_C, \phi_C) &= 2\int_{I+III+IV} [(\partial_1\phi_C)^2 + (\partial_2\phi_C)^2]\mathrm{d}x\mathrm{d}y \\ &= 2\int_{I+III} (\partial_1\phi_C)^2\mathrm{d}x\mathrm{d}y + 2\int_{I+IV} (\partial_2\phi_C)^2\mathrm{d}x\mathrm{d}y \\ &= 2h^{-2}\int_{I+III} \mathrm{d}x\mathrm{d}y + 2h^{-2}\int_{I+IV} \mathrm{d}x\mathrm{d}y \\ &= 4. \end{aligned} \tag{5.57}$$

Next we compute

$$
\begin{aligned}
a(\phi_C, \phi_N) &= \int_{I+IV} \nabla\phi_C \cdot \nabla\phi_N \mathrm{d}x\mathrm{d}y \\
&= \int_{I+IV} \partial_y\phi_C \partial_y\phi_N \mathrm{d}x\mathrm{d}y \\
&= \int_{I+IV} (-h^{-1})h^{-1} \mathrm{d}x\mathrm{d}y \\
&= -1.
\end{aligned} \tag{5.58}
$$

Here we have used the fact that $\nabla\phi_C = (-h^{-1}, -h^{-1})$ in I, and $\nabla\phi_C = (0, -h^{-1})$ in IV; and $\nabla\phi_N = (0, h^{-1})$ in I, and $\nabla\phi_C = (h^{-1}, h^{-1})$ in IV.

By symmetry, $a(\phi_C, \phi_\alpha) = -1$ for $\alpha = S, E, W$.

Similarly, we may compute

$$
a(\phi_C, \phi_{NW}) = \int_{III+IV} [\partial_x\phi_C\partial_x\phi_{NW} + \partial_y\phi_C\partial_y\phi_{NW}]\mathrm{d}x\mathrm{d}y = 0. \tag{5.59}
$$

In summary, the stiffness matrix locally reads $\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$. This is the same as central difference scheme.

5.6 Basic Settings

In two space dimensions, the basic settings for a finite element approximation include the following issues.

A partition refers to splitting Ω into subdomains, each subdomain called as an element. An element may either a triangle or a quadrilateral. Other polygons are also used in certain special applications. The partition is regular if all elements are congruent.

We call $\mathcal{T} = \{T_1, \cdots, T_m\}$ an admissible partition if the following properties hold.

- $\bar{\Omega} = \cup_{i=1}^{M} T_i$
- If $T_i \cap T_j = \{A\}$, then the point A is a common vertex of T_i and T_j.
- If $i \neq j, T_i \cap T_j$ consists of more than one point, then $T_i \cap T_j$ is a common edge.

We some times write the partition as $\mathcal{T}_h$, if every element has diameter no greater than $2h$.

We call a partition $\mathcal{T}_h$ K-uniform, if $\exists K > 0$, such that $\forall T \in T_h$, T contains a circle of radium $\rho_T \geq h/K$.

With a partition, we next set up approximations for each element. We denote the set of polynomials whose degree is no more than t as $P_t = \{u(x,y) = \sum_{i+k\leq t} C_{ik}x^iy^k\}$. When a polynomial approximation is restricted on an edge, it reduces to one variable only, and usually the degree decreases to no greater than $t-1$.

By piecing the local polynomials together to form an approximation, there arises naturally a consideration about the regularity (smoothness) for the approximation in the whole domain Ω. In fact, we call a finite element approximation a C^k element, if the approximate functions are in $C^k(\Omega)$. While the appropriate function space for the elliptic partial differential e-quation is the Hilbert space, the following theorem provides a sharp relation between the differentiability and the Hilbert space.

Theorem 5.6. *Over a bounded domain Ω, a piecewise infinitely differentiable function $v : \bar{\Omega} \to \mathbb{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\bar{\Omega})$ for $k \geq 1$.*

As an example, if we want to solve a problem in the space $H^1(\Omega)$, it is enough to form a finite approximation with $C^0(\bar{\Omega})$ element. Similarly, if we want to solve in $H^m(\Omega)$, continuity on the order of $(m-1)$ should be enforced across the edges.

Finally, we remark that as the finite element approximation space $S_h \subset V$, we call this a conformal element.

5.7 Triangular Elements With Complete Polynomials

For a triangle element, we may define a reference triangle T with vertices at $(0,0), (1,0)$ and $(0,1)$. Then with an affine linear transformation

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} \tilde{x}_0 \\ \tilde{y}_0 \end{pmatrix}, \quad det(A) \neq 0, \tag{5.60}$$

a polynomial $p(x,y)$ over the reference triangle leads to another polynomial $\tilde{p}(\tilde{x},\tilde{y})$, which is of the same order. Therefore, we shall solely discuss approximations on the reference triangle.

For $t \geq 0$, suppose that $s = (t+1)(t+2)/2$ points $z_1, \cdots, z_s$ in T lying on $(t+1)$ lines. Then $\forall f \in C(T)$, there exists a unique polynomial p with degree no greater than t, such that $p(z_i) = f(z_i)$. This is obvious for $t = 0$. Assume the statement holds for $t-1$. Then for t, consider the edge along the x axis. There is a polynomial $p_0(x)$, such that $f(z_i) = p_0(x(z_i))$ holds for the $(t+1)$ points on this axis. Using the assumption for $(t-1)$, we may find a function $q(x,y)$ to handle the rest points. Then we construct $p(x,y) = p_0(x) + yq(x,y)$, which solves the problem for t.

This leads to the construction of some nodal bases, which refers to $\{\psi_i\}_{i=1}^s$ satisfying $\psi_i(z_j) = \delta_{i,j}$. For this choice, a set of points $z_1, \cdots, z_s$ uniquely determine a function in S_h.

Now we construct some C^0-elements on Ω with polynomials on the degree of $t \geq 1$.

After making triangulation, we select for each triangle T $s = (t+1)(t+2)/2$ points and form a polynomial $p(x, y)$ of the degree t. When restricted on an edge, the polynomial reduces to a single variable polynomial with degree no greater than t. It is uniquely determined by $(t + 1)$ points at this edge. Consider an adjacent triangle $\tilde{T}$. The polynomial on $\tilde{T}$ is also determined by the same $(t+1)$ points at this common edge when restricted to the edge. As the single variable polynomial uniquely determined by the $(t + 1)$ points at edge, these two single variable polynomials must be the same. We conclude that the approximation is C^0 across edges.

We describe several C^0 triangle elements as follows.

First, a conforming P_1 element (Courant element) $\mathcal{M}_0^1$. Over the reference triangle, we define

$$\psi_1|_T = 1 - (x + y), \quad \psi_2|_T = x, \quad \psi_3|_T = y. \tag{5.61}$$

The resulted approximation function space $\Pi_{ref} = P_1$, with $\dim \Pi_{ref} = 3$. In general, $\forall \phi|_T \in P_1$, we may expand

$$\phi = \phi_1\psi_1 + \phi_2\psi_2 + \phi_3\psi_3, \tag{5.62}$$

where $\phi_1 = \phi(0, 0)$, $\phi_2 = \phi(1, 0)$, and $\phi_3 = \phi(0, 1)$.

Next, we construct a quadratic triangle element $\mathcal{M}_0^2$ by defining

$$\begin{aligned}
\psi_1 &= (1 - (x + y))(1 - 2(x + y)), && (5.63)\\
\psi_2 &= 4x(1 - (x + y)), && (5.64)\\
\psi_3 &= 2x(x - \frac{1}{2}), && (5.65)\\
\psi_4 &= 4y(1 - (x + y)), && (5.66)\\
\psi_5 &= 4xy, && (5.67)\\
\psi_6 &= 2y(y - \frac{1}{2}). && (5.68)
\end{aligned}$$

These basis functions are constructed by selecting a particular point, and drawing two lines to cover the rest points. For this element, we have $\Pi_{ref} = P_2, \dim \Pi_{ref} = 6$.

We illustrate C^1 element by an example of the Argyris element. We take a fifth degree polynomial on the reference triangle. The number of coefficients is then $6 \times 7/2 = 21$. For a function ϕ to be approximated, we take its values and derivatives at the three vertices up to the second order, namely, $\phi, \phi_x, \phi_y, \phi_{xx}, \phi_{yy}, \phi_{xy}$. We further take the normal derivative at each edge center. Altogether, we have $6 \times 3 + 3 = 21$ conditions.

We demonstrate that the Argyris is C^1. As a matter of fact, consider along a common edge $y = 0$. The polynomials restricted to this edge is a single variable one of order 5, which requires six coefficients to fix. In fact, these six coefficients for $x^5, x^4, x^3, x^2, x, 1$ are determined uniquely by ϕ, ϕ_x, ϕ_{xx} at the two vertices. Therefore, the element is C^0. Accordingly, the x-derivative is continuous across the edge $y = 0$.

Next, we consider the normal derivative ϕ_y at $y = 0$. The normal derivative is a single variable polynomial of the order 4. There are five coefficients to be determined for $x^4, x^3, x^2, x, 1$. This is determined uniquely by ϕ_y, ϕ_{xy} at the two vertices, plus the normal derivative on this edge. This ends the proof for C^1 continuity.

Finally, we also mention a bilinear quadrilateral element (Q_1-element). Over a reference cube $[0, 1] \times [0, 1]$, the approximation is defined as $\phi(x, y) = a+bx+cy+dxy$. This may be uniquely determined by ϕ at the four vertices. We notice that $\Pi_{ref} \subset P_2$, $\dim \Pi_{ref} = 4$, and this gives C^0-element.

5.8 Finite Element and Affine Families

Definition 5.2. *A finite element is a triple (T, Π, Σ):*

- *T is a polygon in $\mathbb{R}^d$;*
- *Π is a subspace of $C(T)$ with finite dimension s;*
- *Σ is a set of s linearly independent functions in Π, such that $\forall p \in \Pi$is uniquely defined by the generalized interpolation condition, namely by fixing the values of s functionals in Σ.*

A few more notions are as follows. Each part of ∂T is called as a faces. The basis of Π are formed by shape functions. The number s is called as the local degree of freedom, or the local dimension.

Definition 5.3. *A family of finite element spaces S_h for partition $\mathcal{T}_h$ of $\Omega \subset \mathbb{R}^d$ is called as an affine family provided that $\exists$ a finite element $(T_{ref}, \Pi_{ref}, \Sigma)$, such that $\forall T_j \in \mathcal{T}_h$, there exists an affine mapping $F_j : T_{ref} \to T_j$, such that $\forall v \in S_h$, it holds that*

$$v(x)|_{T_j} = p(F_j^{-1}x), \quad p \in \pi_{ref}. \tag{5.69}$$

As an example, M_0^k is an affine family. On the other hand, an element with normal derivatives is not an affine family, e.g., the Argyris element.

5.9 Approximation Properties

As mentioned before, the Cea's lemma asserts that $\| u - u_h \|_m \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \| u - v_h \|_m, H_0^m \subset V \subset H^m(\Omega)$. This means, the convergence follows from the approximation property of the finite element method.

Because $C^0 \not\subset H^m(\Omega)$ for $m > 1$, this does apply to C^0 elements when a high order estimate is aimed at. To this end, we shall define mesh dependent norms, and confine ourselves to affine families.

First, for a partition $\mathcal{T}_h = \{T_1, \cdots, T_M\}$ and $m \geq 1$, we define

$$\| v \|_{m,h} \equiv \sqrt{\sum_{T_j \in \mathcal{T}_h} \| v \|_{m,T_j}^2}. \tag{5.70}$$

Theorem 5.7. *(Bramble-Hilbert lemma) Let $\Omega \in \mathbb{R}^2$, with Lipschitz continuous boundary, and $t \geq 2$. Assume that L is a bounded linear mapping from $H^t(\Omega)$ to a normed linear space Y. If $P_{t-1} \subset KerL$, then $\exists C = C(\Omega) \| L \| \geq 0$, it holds that*

$$\| Lv \| \leq C|v|_t, \ \forall v \in H^t(\Omega) \tag{5.71}$$

Using the Bramble-Hilbert lemma, we consider C^0 triangle elements with complete polynomials P_{t-1} $(t \geq 2)$. We take the associated affine family $S_h = \mathcal{M}_0^{t-1}(\mathcal{T}_h)$ for a shape regular triangulation $\mathcal{T}_h$, and define an interpolation $I_h : H^t(\Omega) \to S_h$. It may be prove that $\exists C = C(\Omega, k, t)$, such that

$$\| \phi - I_h\phi \|_{m,h} \leq ch^{t-m}|\phi|_{t,\Omega}, \forall \phi \in H^t(\Omega), \ 0 \leq m \leq t \tag{5.72}$$

This tells us that the convergent rate is $(t - m)$ if ϕ has a regularity up to t-th order.

$$\| \phi - I_h\phi \|_{m,h} \leq Ch^{t-m}|\phi|_{t,\Omega} \leq Ch^{t-m} \| \phi \|_{t,\Omega}, \quad m \leq t. \tag{5.73}$$

Furthermore, there is an inverse estimate for an affine family with K uniform partition. Let S_h be an affine family of FE's consisting of piecewise polynomials of degree s, then $\exists c = c(K, s, t)$, such that $\forall 0 \leq m \leq t$, it holds that

$$\| v_h \|_{t,h} \leq Ch^{m-t} \| v_h \|_{m,h}, \quad \forall v_h \in S_h. \tag{5.74}$$

The inverse estimate shows that the approximation estimate is optimal. We list the possible values of t for some elements as follows.

element	linear triangle	quadratic	cubic	bilinear	Argyris
t	2	2,3	2,3,4	2	3,4,5,6

5.10 Error Bounds

Definition 5.4. *Consider a V-elliptic bilinear form $a(\cdot, \cdot)$ with $m \geq 1$, $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. The bilinear form is called H^s-regular if $\exists C = C(\Omega, a, s)$, such that $\forall f \in H^{s-2m}(\Omega)$, the solution $u \in H^s(\Omega)$ exists for equation $a(u, v) = (f, v)_0, \forall v \in V$ with*

$$\| u \|_s \leq c \| f \|_{s-2m} \cdot \tag{5.75}$$

It may be shown that for H_0^1-elliptic linear form with sufficiently smooth coefficient functions, if further Ω is convex, then the Dirichlet problem is H^2-regular. Moreover, if Ω has a C^s boundary with $s \geq 2$, then the Dirichlet problem is H^1-regular. We remark that the Neumann problem is more complicated.

Theorem 5.8. *Let $\mathcal{T}_h$ be a family of shape-regular triangulation of a convex polygonal domain Ω, then $u_h \in S_h = \mathcal{M}_0^k, (k \geq 1)$ satisfies*

$$\| u - u_h \|_1 \leq Ch \| u \|_2 \leq Ch \| f \|_0 . \tag{5.76}$$

Proof. Because Ω is convex, therefore the problem is H^2 regular. This means, $\| u \|_2 \leq C_1 \| f \|_0$. Using the approximation property, we find that $v_h = I_h u \in S_h$ satisfies

$$\| u - v_h \|_{1,\Omega} = \| u - v_h \|_{1,h} \leq Ch \| u \|_{2,\Omega} . \tag{5.77}$$

By the Cea's lemma and the regularity, we derive

$$\| u - u_h \|_1 \leq Ch \| u \|_2 \leq Ch \| f \|_0 . \tag{5.78}$$

□

There are better estimates but we omit them.

5.11 Solving the Algebraic Equations

Finite element method usually leads to huge algebraic equations. Efficient algorithm is crucial for the application of finite element method.

In classical iterative algorithms, for an equation $Ax = b$, one decompose $A = M - N$ and derive $Mx = Nx + b$. The iterative method then reads

$$Mx^{k+1} = Nx^k + b, \tag{5.79}$$

or,

$$\begin{aligned} x^{k+1} &= x^k + M^{-1}(b - Ax^k) \\ &\equiv Gx^k + d. \end{aligned} \tag{5.80}$$

Here $G = I - M^{-1}A$ and $d = M^{-1}b$.

By the fixed point theorem, it may be shown that solution the convergence requirement reads

$$\lim_{k\to\infty} error^k = 0 \Leftrightarrow \rho(G) = \max_i |\lambda_i| < 1. \tag{5.81}$$

Different decompositions then give different algorithms. In the Jacobi method, one takes the diagonal part as D, and the off-diagonal parts as

$-L$ and $-U$, respectively. This means, $A = D - L - U$. Then we obtain $G_J = D^{-1}(L + U)$. Componentwise, we have

$$x_i^{k+1} = a_{ii}^{-1}(-\sum_{i \neq j} a_{ij} x_j^k + b_i). \tag{5.82}$$

In the Gauss-Seidel method, we take $M = D + L$, and $G_{GS} = (D + L)^{-1}U$. It is solved explicitly by

$$a_{ii} x_i^{k+1} = -\sum_{j<i} a_{ij} x_j^{k+1} - \sum_{j>i} a_{ij} x_j^{k+1} + b_i. \tag{5.83}$$

We remark that the convergence essentially requires the stiffnes matrix diagonally dominant and irreducible.

Another category of algorithms is relaxation methods. In an over-relaxation method, we compute

$$Dx^{k+1} = \omega[Lx^{k+1} + Ux^k + b] - (\omega - 1)Dx^k. \tag{5.84}$$

Convergence holds for $0 < \omega < 2$, provided that A is symmetric and positive definite.

There is a further category of algorithms, which fits very well the finite element method. These are the so-called gradient methods. The algebraic system $Ax = b$ with A a symmetric positive definite matrix, it suffices to minimize the function

$$f(x) = \frac{1}{2} x^{'} Ax - b^{'} x. \tag{5.85}$$

The standard gradient method for a general function $f(x)$ defined on $M \subset \mathbb{R}^n$ starts with an initial guess $x_0 \in M$. Then for $k = 0, 1, 2, \cdots$, we perform the following iteration.

1. Determine the direction $d_k = -\nabla f(x_k)$.

2. Line search: find $t = \alpha_k$ along the line $\{x_k + td_k : t \geq 0\} \cap M$ to search local minimal point $x_{k+1} = x_k + \alpha_k d_k$.

For the aforementioned special case of $f(x) = \frac{1}{2} x^{'} Ax - b^{'} x$, we have $d_k = b - Ax_k$ and $\alpha_k = \frac{d_k^{'} d_k}{d_k^{'} A d_k}$. Therefore, it holds that

$$\begin{aligned} f(x_{k+1}) &= f(x_k + \alpha_k d_k) \\ &= \frac{1}{2}(x_k + \alpha_k d_k)^{'} A(x_k + \alpha_k d_k) - b^{'}(x_k + \alpha_k d_k) \\ &= f(x_k) - \frac{1}{2} \frac{(d_k^{'} d_k)^2}{d_k^{'} A d_k}. \end{aligned} \tag{5.86}$$

Suppose the exact solution is $x^\star$, and define the energy norm $\| x \|_A = \sqrt{x' A x}$. For a symmetric positive definite A, we have the Kantorovitch inequality

$$\frac{(x^{'}Ax)(x^{'}A^{-1}x)}{(x'x)^2} \le (\frac{1}{2}\sqrt{k} + \frac{1}{2}k^{-1/2})^2, \quad k = dim\, A. \tag{5.87}$$

From this we derive a convergence rate as follows.

$$\| x_{k+1} - x^\star \|_A^2 = \| x_k - x^\star \|_A^2 \, [1 - \frac{(d_k^{'}d_k)^2}{d_k^{'}Ad_kd_k^{'}A^{-1}d_k}] \le (\frac{k-1}{k+1})^k \, \| x_0 - x^\star \|_A^2 \, . \tag{5.88}$$

Assignments

1. Construct the cubic triangle element M_0^3, for which $\Pi_{ref} = P_3, \dim \Pi_{ref} = 10$.

   Solution:

$$\begin{aligned}
\psi_1 &= (1-(x+y))(1-3(x+y))(1-\frac{3}{2}(x+y)), && (5.89)\\
\psi_2 &= 9x(1-\frac{3}{2}(x+y))(1-(x+y)), && (5.90)\\
\psi_3 &= \frac{27}{2}x(x-\frac{1}{3})(1-(x+y)), && (5.91)\\
\psi_4 &= \frac{9}{2}x(x-\frac{1}{3})(x-\frac{2}{3}), && (5.92)\\
\psi_5 &= 9y(1-\frac{3}{2}(x+y))(1-(x+y)), && (5.93)\\
\psi_6 &= 27xy(1-(x+y)), && (5.94)\\
\psi_7 &= \frac{27}{2}xy(x-\frac{1}{3}), && (5.95)\\
\psi_8 &= \frac{27}{2}y(y-\frac{1}{3})(1-(x+y)), && (5.96)\\
\psi_9 &= \frac{27}{2}xy(y-\frac{1}{3}), && (5.97)\\
\psi_{10} &= \frac{9}{2}y(y-\frac{1}{3})(y-\frac{2}{3}). && (5.98)
\end{aligned}$$