FINITE DIFFERENCE METHODS

FOR

SOLVING DIFFERENTIAL EQUATIONS

I-Liang Chern

Department of Mathematics
National Taiwan University

May 15, 2018

Contents

Chapter 1

Introduction

The goal of this course is to introduce theoretical analysis of finite difference methods for solving partial differential equations. The focuses are the stability and convergence theory. The partial differential equations to be discussed include

- parabolic equations,
- elliptic equations,
- hyperbolic conservation laws.

1.1 Finite Difference Approximation

A finite difference approximation is to approximate differential operators by finite difference operators, which is a linear combination of u on discrete points. For example,

- Forward difference: $D_+u(x) := \frac{u(x+h)-u(x)}{h}$,
- Backward difference: $D_-u(x) := \frac{u(x)-u(x-h)}{h}$,
- Centered difference: $D_0u(x) := \frac{u(x+h)-u(x-h)}{2h}$.

Here, h is called the mesh size. By Taylor expansion, we can get

- $u'(x) = D_+u(x) + O(h)$,
- $u'(x) = D_-u(x) + O(h)$,
- $u'(x) = D_0u(x) + O(h^2)$.

These formulae can be derived by performing Taylor expansion of u at x. For instance, we expand

$$u(x+h) \quad = \quad u(x) + u'(x)h + \frac{h^2}{2}u''(x) + \frac{h^3}{3!}u'''(x) + \cdots$$

$$u(x-h) \quad = \quad u(x) - u'(x)h + \frac{h^2}{2}u''(x) - \frac{h^3}{3!}u'''(x) + \cdots .$$

Subtracting these two equations yields

$$u(x+h) - u(x-h) = 2u'(x)h + \frac{2h^3}{3!}u'''(x) + \cdots .$$

This gives

$$u'(x) = D_0 u(x) - \frac{h^2}{3!}u'''(x) + \cdots = D_0 u(x) + O(h^2).$$

Thus, $u'(x)$ can be approximated by several difference operators. Indeed, we can approximate $u'(x)$ by finite difference operators which involve u on more discrete points with higher order errors. For example,

$$u'(x) = D_3 u(x) + O(h^3),$$

where

$$D_3 u(x) = \frac{1}{6h}\left(2u(x+h) + 3u(x) - 6u(x-h) + u(x-2h)\right).$$

This formula can be derived by taking Taylor expansion of $u(x+h)$, $u(x-h)$, $u(x-2h)$ about x, then making proper combination to cancel 0th, and 2nd derivatives term. That is

$$\begin{aligned}
u(x+h) &= u(x) + u'(x)h + \frac{h^2}{2}u''(x) + \frac{h^3}{3!}u'''(x) + \cdots \\
u(x-h) &= u(x) - u'(x)h + \frac{h^2}{2}u''(x) - \frac{h^3}{3!}u'''(x) + \cdots \\
u(x-2h) &= u(x) - 2u'(x)h + \frac{4h^2}{2}u''(x) - \frac{8h^3}{3!}u'''(x) + \cdots
\end{aligned}$$

Taking the combination $2u(x+h) + 3u(x) - 6u(x-h) + u(x-2h)$, we can cancel the zeroth, second derivatives and obtain $u'(x) = D_3 u(x) + O(h^3)$.

In general, suppose we are given the values of u at discrete points $\{x_j\}$. These discrete points are called grid points. We want to approximation $u^{(k)}$ at a specific point $\bar{x}$ by the values of u at some of these grid points, say $x_j, j = 0, ..., n$ with $n \geq k$. That is,

$$u^{(k)}(\bar{x}) = \sum_{j=0}^{n} c_j u(x_j) + O(h^{p-k+1})$$

Here, the mesh size h denotes $\max\{|x_i - x_j|\}$. The parameter $p \geq k$ is some positive integer. We want to design c_j so that p is as larger as possible. As we shall see later that we can choose $p = n$. To find the coefficients $c_j, j = 0, ..., n$, we make Taylor expansion of $u(x_j)$ about the point x:

$$u(x_j) = \sum_{i=0}^{p} \frac{1}{i!}(x_j - \bar{x})^i u^{(i)}(\bar{x}) + O(h^{p+1}).$$

We plug this expansion formula of $u(x_j)$ into the finite difference approximation formula for $u^{(k)}(x)$:

$$u^{(k)}(\bar{x}) = \sum_{j=0}^{n} c_j \sum_{i=0}^{p} \frac{1}{i!}(x_j - \bar{x})^i u^{(i)}(\bar{x}) + O(h^{p-k+1}).$$

Comparing both sides, we obtain

$$\sum_{j=0}^{n} \frac{(x_j - \bar{x})^i}{i!} c_j = \left\{ \begin{array}{ll} 1 & \text{if } i = k \\ 0 & \text{otherwise} \end{array} \right\}, \text{ for } i = 0, ..., p.$$

There are $p+1$ equations here, it is natural to choose $p = n$ to match the $n+1$ unknowns. This is a $n \times n$ Vandermonde system. It is nonsingular if x_i are different. The matlab code fdcoeffV(k,xbar,x) can be used to compute these coefficients. Reference: Randy LeVeque's book and his Matlab code.

In the case of uniform grid, using central finite differencing, we can get high order approximation by using less grid points. For instance, let $x_j = jh$, where $j \in \mathbb{Z}$. Let $u_j = u(x_j)$. Then

$$\begin{aligned} u'(0) &= \frac{u_1 - u_{-1}}{h} + O(h^2) \\ u''(0) &= \frac{u_1 - 2u_0 + u_{-1}}{h^2} + O(h^2) \\ u^{(3)} &= \frac{1}{2h^3}(u_2 - 2u_1 + 2u_0 - 2u_{-1} + u_{-2}) + O(h^2). \end{aligned}$$

Homeworks.

1. Consider $x_i = ih$, $i = 0, ..., n$. Let $\bar{x} = x_m$. Find the coefficients c_i for $u^{(k)}(\bar{x})$ and the coefficient of the leading truncation error for the following cases:
   - $k = 1, n = 2, 3, m = 0, 1, 2, 3.$
   - $k = 2, n = 2, m = 0, 1, 2.$

1.2 Basic Numerical Methods for Ordinary Differential Equations

The basic assumption to design numerical algorithm for ordinary differential equations is the smoothness of the solutions, which is in general valid provided the coefficients are also smooth. Basic designning techniques include numerical interpolation, numerical integration, and finite difference approximation.

Euler method

Euler method is the simplest numerical integrator for ODEs. The ODE

$$y' = f(t, y) \tag{1.1}$$

is discretized by

$$y^{n+1} = y^n + kf(t^n, y^n). \tag{1.2}$$

Here, $t^0, ..., t^n$ are the grid points of time t. $k = t^{n+1} - t^n$ is time step size of the discretization. This method is called the forward Euler method. It simply replace $dy/dt(t^n)$ by the forward finite difference $(y^{n+1} - y^n)/k$. Given a smooth solution $y(\cdot)$, by Taylor expansion, we define the local truncation error to be

$$\tau^n := y'(t^n) - \frac{y(t^{n+1}) - y(t^n)}{k} = O(k).$$

We are interested in the true error, which is defined to be $e^n := y^n - y(t^n)$. We have the following convergence theorem.

Theorem 1.1. *Assume $f \in C^1$ and suppose the solution $y' = f(t, y)$ with $y(0) = y_0$ exists on $[0, T]$. Then the Euler method converges at any $t \in [0, T]$. In fact, the true error e^n has the following estimate:*

$$|e^n| \le \frac{e^{\lambda t}}{\lambda} O(k) \to 0, \text{ as } n \to \infty. \tag{1.3}$$

Here, $\lambda = \max |\partial f/\partial y|$ and $nk = t$.

Proof. From the regularity of the solution, we have $y \in C^2[0, T]$ and

$$y(t^{n+1}) = y(t^n) + kf(t^n, y(t^n)) + k\tau^n. \tag{1.4}$$

Taking difference of (1.2) and (1.4), we obtain

$$\begin{aligned} |e^{n+1}| &\le |e^n| + k|f(t^n, y^n) - f(t^n, y(t^n))| + k|\tau^n| \\ &\le (1 + k\lambda)|e^n| + k|\tau^n|. \end{aligned}$$

where

$$|f(t, x) - f(t, y)| \le \lambda |x - y|.$$

The finite difference inequality has a fundamental solution $G^n = (1 + \lambda k)^n$, which is positive provided k is small. Multiplying above equation by $(1 + \lambda k)^{-n-1}$, we obtain

$$|e^{m+1}|G^{-m-1} \le |e^m|G^{-m} + kG^{-m-1}|\tau^m|.$$

Summing in m from $m = 0$ to $n - 1$, we get

$$\begin{aligned} |e^n| &\le \sum_{m=0}^{n-1} G^{n-m-1} k|\tau^m| \le \sum_{m=0}^{n-1} G^m O(k^2) \\ &= \frac{G^n - 1}{G - 1} O(k^2) \le \frac{G^n}{\lambda} O(k) \le \frac{e^{\lambda t}}{\lambda} O(k), \end{aligned}$$

where $t = nk$ and we have used $(1 + \lambda k)^n \le e^{\lambda t}$. □

Remarks.

1. The theorem states that the numerical method converges in $[0, T]$ as long as the solutions of the ODE exists.

2. The error is $O(k)$ if the solution is in $C^2[0, T]$.

3. The proof above relies on the existence and smoothness of the solution. However, one can also use this approach to prove the local existence theorem by showning the approximate solutions generated by the Euler method form a Cauchy sequence.

Backward Euler method

In many applications, the system is relaxed to a stable solution in a very short period of time. For instance, consider

$$y' = \frac{\bar{y} - y}{\tau}.$$

The corresponding solution $y(t) \to \bar{y}$ as $t \sim O(\tau)$. In the above forward Euler method, practically, we should require

$$1 + k\lambda \leq 1$$

in order to have G^n remain bounded. Here, λ is the Lipschitz constant. In the present case, $\lambda = 1/\tau$. If τ is very small, the the above forward Euler method will require very small k and lead to inefficient computation. In general, forward Euler method is inefficient (require small k) if

$$\max \left| \frac{\partial f(t, y)}{\partial y} \right| >> 1.$$

There are two possibilities:

- $\partial f/\partial y >> 1$: In this case, we need to choose a very small k in order to resolve details.
- $\partial f/\partial y << -1$.

For the second case, the backward Euler method is recommended:

$$y^{n+1} = y^n + kf(t^{n+1}, y^{n+1}).$$

Comparing the Taylor expansion of the exact solution at t^{n+1}:

$$y(t^{n+1}) = y(t^n) + kf(t^{n+1}, y(t^{n+1})) + O(k),$$

we get that the true error $e^n := y^n - y(t^n)$ satisfies

$$e^{n+1} = e^n + k\Delta f + O(k^2) = e^n + k \left(\frac{\partial f}{\partial y}(t, \bar{y}) \right) e^{n+1} + O(k^2).$$

$$\left(1 - k \left(\frac{\partial f}{\partial y}(t, \bar{y}) \right) \right) e^{n+1} = e^n + O(k^2).$$

Since $\frac{\partial f}{\partial y}(t,\bar{y}) << -1$, we take $\lambda = \min\left|\frac{\partial f}{\partial y}(t,\bar{y})\right|$, we get

$$\left(1 - k\left(\frac{\partial f}{\partial y}(t,\bar{y})\right)\right) \geq (1+\lambda k).$$

Thus,

$$(1+\lambda k)|e^{n+1}| \leq |e^n| + O(k^2)$$

The corresponding fundamental solution is $G^n := (1+\lambda k)^{-n}$. Notice that the error satisfies

$$\begin{aligned} |e^n| &\leq \sum_{m=0}^{n-1}(1+\lambda k)^{-m}O(k^2) \\ &\leq \frac{(1+\lambda k)^{-n+1}}{\lambda k}O(k^2) \\ &\leq \frac{e^{-\lambda T}}{\lambda}O(k). \end{aligned}$$

There is no restriction on the size of k. However, the price to pay is that we need to solve a nonlinear equation

$$y^{n+1} = y^n + kf(t^{n+1}, y^{n+1})$$

for y^{n+1} at each time step.

Leap frog method

We integrate $y' = f(t,y)$ from t^{n-1} to t^{n+1}:

$$y(t^{n+1}) - y(t^{n-1}) = \int_{t^{n-1}}^{t^{n+1}} f(\tau, y(\tau))\, d\tau.$$

We apply the midpoint rule for numerical integration, we then get

$$y(t^{n+1}) - y(t^{n-1}) = 2kf(t^n, y(t^n)) + O(k^3).$$

The midpoint method (or called leapfrog method) is

$$y^{n+1} - y^{n-1} = 2kf(t^n, y^n). \tag{1.5}$$

Homeworks.

1. Prove the convergence theorem for the leap-frog method for ODE.
   Hint: consider the system $y_1^n = y^{n-1}$ and $y_2^n = y^n$.

1.3 Runge-Kutta methods

The Runge-Kutta method (RK) is a strategy to integrate $\int_{t^n}^{t^{n+1}} f\, d\tau$ by some quadrature method. Below, RK2, RK4 are RK method with different orders.

RK2: A second order RK, denoted by RK2, is based on the trapezoidal rule of numerical integration. First, we integrate the ODE $y' = f(t, y)$ to get

$$y(t^{n+1}) - y^n = \int_{t^n}^{t^{n+1}} f(\tau, y(\tau))\, d\tau.$$

Next, this integration is approximated by

$$\int_{t^n}^{t^{n+1}} f(\tau, y(\tau))\, d\tau = \frac{k}{2}\left(f(t^n, y^n) + f(t^{n+1}, y^{n+1})\right) + O(k^3).$$

The latter term involves y^{n+1}. An explicit Runge-Kutta method approximate y^{n+1} by $y^n + kf(t^n, y^n)$. Thus, RK2 reads

$$\begin{aligned} \xi_1 &= f(t^n, y^n) \\ y^{n+1} &= y^n + \frac{k}{2}(f(t^n, y^n) + f(t^{n+1}, y^n + k\xi_1)). \end{aligned}$$

Another kind of RK2 is based on the midpoint rule of integration. It reads

$$\begin{aligned} \xi_1 &= f(t^n, y^n) \\ y^{n+1} &= y^n + kf(t^{n+1/2}, y^n + \frac{k}{2}\xi_1) \end{aligned}$$

The truncation error of RK2 is

$$y(t^{n+1}) - y(t^n) = y^{n+1} - y(t^n) + O(k^3).$$

RK4 The 4th order Runge-Kutta method uses Simpson's rule to approximate integration:

$$\int_{t^n}^{t^{n+1}} f(t, y(t))\, dt = \frac{k}{6}\left(f(t^n, y(t^n)) + 4f(t^{n+1/2}, y(t^{n+1/2})) + f(t^{n+1}, y(t^{n+1}))\right) + O(k^5).$$

The quantity $y(t^{n+1/2})$ is approximated by forward Euler method. It has the form

$$\begin{aligned} y^{n+1} &= y^n + \frac{k}{6}(\xi_1 + 2\xi_2 + 2\xi_3 + \xi_4) \\ \xi_1 &= f(t^n, y^n) \\ \xi_2 &= f(t^n + \frac{1}{2}k, y^n + \frac{k}{2}\xi_1) \\ \xi_3 &= f(t^n + \frac{1}{2}k, y^n + \frac{k}{2}\xi_2) \\ \xi_4 &= f(t^n + k, y^n + k\xi_3) \end{aligned}$$

The truncation error of RK4 is

$$y(t^{n+1}) - y(t^n) = y^{n+1} - y(t^n) + O(k^5).$$

General explicit Runge-Kutta methods The method takes the following general form

$$y^{n+1} = y^n + k\sum_{i=1}^{s} b_i\xi_i,$$

where

$$\begin{aligned}
\xi_1 &= f(t^n, y^n),\\
\xi_2 &= f(t^n + c_2k, y^n + ka_{21}\xi_1),\\
\xi_3 &= f(t^n + c_3k, y^n + ka_{31}\xi_1 + ka_{32}\xi_2),\\
&\vdots\\
\xi_s &= f(t^n + c_sk, y^n + k(a_{s1}\xi_1 + \cdots + a_{s,s-1}\xi_{s-1})).
\end{aligned}$$

We need to specify s (the number of stages), the coefficients $a_{ij}(1 \le j < i \le s)$, $b_i(i = 1, ..., s)$ and $c_i(i = 2, ..., s)$. We list them in the following Butcher table.

There are $s(s-1)/2+s+(s-1)$ unknowns to be determined for a specific scheme. We require the

0					
c_2	a_{21}				
c_3	a_{31}	a_{32}			
$\vdots$	$\vdots$		$\ddots$		
c_s	a_{s1}	a_{s2}	$\cdots$	$a_{s,s-1}$	
	b_1	b_2	$\cdots$	b_{s-1}	b_s

truncation error to be $O(k^{p+1})$. To find these coefficients, we need to expand the truncation error formula

$$y(t^{n+1}) - y^n = y^{n+1} - y^n + O(k^{p+1})$$

about (t^n, y^n) in terms of derivatives of $y(\cdot)$ at t^n. Then we can get linear equations for the coefficients.

Adaptive Runge-Kutta (Runge-Kutta-Fehlberg method, ODE45) The adaptive Runge-Kutta method is designed to be able to estimate local truncation error in each time step. From which, we can adjust time step size to have roughly uniform truncation error in each step. This is done by using two RK methods with the same sets of a_{ij} and c_i but different b_i, b_i^*. The set b_i produces RK method of order p. The auxiliary set b_i^* produces a RK method with order $p-1$. It is used to estimate the local truncation by

$$y^{n+1} - y^{n+1,*} = h\sum_{i=1}^{s}(b_i - b_i^*)k_i = O(h^p)$$

The step size h is then estimated so that the truncation error is roughly the same in each time step. Below is the Butcher table for RK5 and RK4 (b^*).

0						
$1/4$	$1/4$					
$3/8$	$3/32$	$9/32$				
$12/13$	$1932/2197$	$-7200/2197$	$7296/2197$			
1	$439/216$	-8	$3860/513$	$-845/4104$		
$1/2$	$-8/27$	2	$-3544/2565$	$1859/4104$	$-11/40$	
b	$16/135$	0	$6656/12825$	$28561/56430$	$-9/50$	$2/55$
b^*	$25/216$	0	$1408/2565$	$2197/4104$	$-1/5$	0

Convergence proof, an example Let us see the proof of the convergence of the two stage Runge-Kutta method. The scheme can be expressed as

$$y^{n+1} = y^n + k\Psi(y^n, t^n, k) \tag{1.6}$$

where

$$\Psi(y^n, t^n, k) := f(y + \frac{1}{2}kf(y)). \tag{1.7}$$

Suppose $y(\cdot)$ is a true solution, the corresponding truncation error

$$\tau^n := \frac{y(t^{n+1}) - y(t^n)}{k} - \Psi(y(t^n), t^n, k) = O(k^2)$$

Thus, the true solution satisfies

$$y(t^{n+1}) - y(t^n) = k\Psi(y(t^n), t^n, k) + k\tau^n$$

The true error $e^n := y^n - y(t^n)$ satisfies

$$e^{n+1} = e^n + k(\Psi(y^n, t^n, k) - \Psi(y(t^n), t^n, k)) - k\tau^n.$$

This implies

$$|e^{n+1}| \le |e^n| + k\lambda'|e^n| + k|\tau^n|,$$

where λ' is the Lipschitz constant of $\Psi(y, t, k)$ with respect to y. Hence, we get

$$\begin{aligned} |e^n| &\le (1 + k\lambda')^n|e^0| + k\sum_{m=0}^{n-1}(1 + k\lambda')^{n-1-m}|\tau^m| \\ &\le e^{\lambda' t}|e^0| + \frac{e^{\lambda' t}}{\lambda'}\max_m |\tau^m| \end{aligned}$$

Reference:

- Lloyd N. Trefethen, Finite Difference and Sprectral Methods for Ordinary and Partial Differential Equations,
- Randy LeVeque,
- You may also google Runge-Kutta method to get more references.

1.4 Multistep methods

The idea of multi-step method is to derive a relation between, for instance, y^{n+1}, y^n, y^{n-1}, y'^n and y'^{n-1} so that the corresponding truncation is small. The simplest multistep method is the midpoint method. Suppose y^n and y^{n-1} is given. The new state y^{n+1} is defined by

$$y^{n+1} - y^{n-1} = 2ky'^n = 2kf(t^n, y^n)$$

The truncation error is

$$\tau^n := \frac{1}{k}\left(y(t^{n+1}) - y(t^{n-1}) - 2ky'(t^n)\right) = O(k^2).$$

Thus, the method is second order.

We can also design a method which involves y^{n+1}, y^n, y^{n-1} and y'^n, y'^{n-1}. For instance,

$$y^{n+1} = y^n + \frac{k}{2}(3f(y^n) - f(y^{n-1}))$$

The truncation error

$$\tau^n := \frac{1}{k}\left(y^{n+1} - y^n + \frac{k}{2}(3f(y^n) - f(y^{n-1}))\right) = O(k^2).$$

A general r-step multistep method involves $(y^{n+1}, y^n, ..., y^{n+1-r})$ and $(y'^{n+1}, y'^n, ..., y'^{n+1-r})$. It can be written as

$$\sum_{m=0}^{r} a_m y^{n+1-r+m} = k\sum_{m=0}^{r} b_m y'^{n+1-r+m} = k\sum_{m=0}^{r} b_m f^{n+1-r+m}. \tag{1.8}$$

We will always assume $a_r \neq 0$. Because it is the coefficient corresponding to y^{n+1}, which is what we want to find. When $b_r = 0$ the method is explicit; otherwise it is implicit. For a smooth solution of (1.1), we define the truncation error τ^n to be

$$\tau^n := \frac{1}{k}\left(\sum_{m=0}^{r} a_m y(t^{n+1-r+m}) - k\sum_{m=0}^{r} b_m y'(t^{n+1-r+m})\right)$$

Definition 1.1. A multi-step method is called of order p if $\tau^n = O(k^p)$ uniformly in n. It is called consistent if $\tau^n(k) \to 0$ uniformly in n as $k \to 0$.

Remark. When f is smooth, the solution of ODE $y' = f(t, y)$ is also smooth. Then the truncation is a smooth function of k. In this case, $\tau(k) \to 0$ is equivalent to $\tau(k) = O(k)$ as $k \to 0$.

Initial setup An r-step multi-step method needs $(y0, y^1, ..., y^{r-1})^T$ to start. There is only y^0 given initially. We need to construct $y^1, ..., y^{r-1}$ by other methods. For instance RK methods. In order to maintain the order of accuracy, we should use a method of $p-1$ order. This will give initial error $y^i - y(t^i) = O(k^p)$ for $i = 0, ..., r-1$.

Derivation of multistep method of order p For notational convenience, let us extend a's and b's by setting $a_m = 0$, $b_m = 0$ for $m > r$. Taking Taylor expansion about t^{n+1-r}, we get

$$\begin{aligned}
k\tau^n &= \sum_{m=0}^{r} a_m \sum_{j=0}^{\infty} \frac{1}{j!} y^{(j)}(mk)^j - k \sum_{m=0}^{r} b_m \sum_{j=1}^{\infty} \frac{1}{(j-1)!} y^{(j)}(mk)^{j-1} \\
&= \left(\sum_{m=0}^{r} a_m \right) y^{(0)} + \sum_{j=1}^{\infty} \frac{1}{j!} \sum_{m=0}^{r} \left(m^j a_m - j m^{j-1} b_m \right) k^j y^{(j)} \\
&= \left(\sum_{m=0}^{r} a_m \right) y^{(0)} + \sum_{j=1}^{\infty} \frac{1}{j!} \sum_{m=0}^{r} m^{j-1} \left(m a_m - j b_m \right) k^j y^{(j)} \\
&= \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{m=0} m^{j-1} \left(m a_m - j b_m \right) k^j y^{(j)} \\
&= \sum_{j=0}^{\infty} C_j y^{(j)}.
\end{aligned}$$

Here, the derivatives of $y(\cdot)$ are evaluated at t^{n+1-r}. We list few equations for the coefficients a and b:

$$\begin{aligned}
C_0 &= a_0 + \cdots + a_r \\
C_1 &= (a_1 + 2a_2 + \cdots r a_r) - (b_0 + \cdots + b_r) \\
C_2 &= \frac{1}{2} \left((a_1 + 2^2 a_2 + \cdots r^2 a_r) - 2(b_1 + \cdots + r b_r) \right) \\
&\vdots \\
C_p &= \sum_{m=0}^{r} \frac{m^p}{p!} a_m - \sum_{m=1}^{r} \frac{m^{p-1}}{(p-1)!} b_m
\end{aligned}$$

To obtain a scheme of order p, we need to require

$$C_j = 0, \text{ for } j = 0, ..., p.$$

There are $2(r+1)$ unknowns for the coefficients $\{a_m\}_{m=0}^r$, $\{b_m\}_{m=0}^r$. In principle, we can choose $p = 2r + 1$ to have the same number of equations. Unfortunately, there is some limitation from stability criterion which we shall be explained in the next section. The order of accuracy p should satisfy

$$p \leq \begin{cases} r+2 \text{ if } r \text{ is even,} \\ r+1 \text{ if } r \text{ is odd,} \\ r \text{ if it is an explicit scheme.} \end{cases}$$

This is the first Dahlquist stability barrier. We shall not discuss here. See Trefethen's book, or Stiff Equation in Wiki, Let us see some concrete examples below.

Explicit Adams-Bashforth schemes When $b_r = 0$, the method is explicit. Here are some examples of the explicit schemes called Adams-Bashforth schemes, where $a_r = 1$:

- 1-step: $y^{n+1} = y^n + kf(y^n)$
- 2-step: $y^{n+1} = y^n + \frac{k}{2}(3f(y^n) - f(y^{n-1}))$
- 3 step: $y^{n+1} = y^n + \frac{k}{12}(23f(y^n) - 16f(y^{n-1}) + 5f(y^{n-2}))$

The step size is r and the order is $p = r$.

Implicit Adams-Moulton schemes Another examples are the Adams-Moulton schemes, where $b_r \neq 0$ and and the step size $r = p$

- 1-step: $y^{n+1} = y^n + \frac{k}{2}(f(y^{n+1}) + f(y^n))$
- 2-step: $y^{n+1} = y^n + \frac{k}{12}(5f(y^{n+1}) + 8f(y^n) - f(y^{n-1}))$
- 3 step: $y^{n+1} = y^n + \frac{k}{24}(9f(y^{n+1}) + 19f(y^n) - 5f(y^{n-1}) + f(y^{n-2}))$

Sometimes, we can use an explicit scheme to guess y^{n+1} as a predictor in an implicit scheme. Such a method is called a predictor-corrector method. A standard one is the following Adams-Bashforth-Moulton scheme: Its predictor part is the Adams-Bashforth scheme:

$$\hat{y}^{n+1} = y^n + \frac{k}{12}(23f(y^n) - 16f(y^{n-1}) + 5f(y^{n-2}))$$

The corrector is the Adams-Moulton scheme:

$$y^{n+1} = y^n + \frac{k}{24}(9f(\hat{y}^{n+1}) + 19f(y^n) - 5f(y^{n-1}) + f(y^{n-2}))$$

The predictor-corrector is still an explicit scheme. However, for stiff problem, we should use implicit scheme instead.

Matlab codes are available on Wikiversity with key words "Adams-Bashforth and Adams-Moulton methods."

Formal algebra Let us introduce the shift operator $Zy^n = y^{n+1}$, or in continuous sense, $Zy(t) = y(t + k)$. Let D be the differential operator. The Taylor expansion

$$y(t + k) = y(t) + ky'(t) + \frac{1}{2!}k^2D^2y(t) + \cdots$$

can be expressed formally as

$$Zy = \left(1 + (kD) + \frac{1}{2!}(kD)^2 + \cdots\right) y = e^{kD}y.$$

The multistep method can be expressed as

$$\mathcal{L}y := (a(Z) - kb(Z)D)\, y = \Big(a(e^{kD}) - kDb(e^{kD})\Big)\, y = (C_0 + C_1(kD) + \cdots)\, y.$$

Here,

$$a(Z) = \sum_{m=0}^{r} a_m Z^m, \quad b(Z) = \sum_{m=0}^{r} b_m Z^m$$

are the generating functions of $\{a_m\}$ and $\{b_m\}$. A multistep method is of order p means that

$$\Big(a(e^{kD}) - kDb(kD)\Big)\, y = O((kD)^{p+1})y.$$

We may abbreviate kD by a symbol κ. The above formula is equivalent to

$$a(e^{\kappa}) - \kappa b(e^{\kappa}) = O(\kappa^{p+1}).$$

Or equivalently,

$$\frac{a(e^{\kappa})}{b(e^{\kappa})} = \kappa + O(\kappa^{p+1}) \text{ as } \kappa \to 0. \tag{1.9}$$

We have the following theorem

Theorem 1.2. *A multistep method with $b(1) \neq 0$ is of order p if and only if*

$$\frac{a(z)}{b(z)} = \log z + O((z-1)^{p+1}) \text{ as } z \to 1.$$

It is consistent if and only if

$$a(1) = 0 \text{ and } a'(1) = b(1).$$

Proof. The first formula can be obtain from (1.9) by writing $e^{\kappa} = z$. For the second formula, we expand $\log z$ about 1. We can get

$$a(z) = b(z)\left((z-1) - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} + \cdots\right) + O((z-1)^{p+1}).$$

We also expand $a(z)$ and $b(z)$ about $z = 1$, we can get

$$a(1) + (z-1)a'(1) = b(1)(z-1) + O((z-1)^2).$$

Thus, the scheme is consistent if and only if $a(1) = 0$ and $a'(1) = b(1)$. □

Homeworks.

1. Consider the linear ODE $y' = \lambda y$, derive the finite difference equation using multistep method involving y^{n+1}, y^n, y^{n-1} and y'^n and y'^{n-1} for this linear ODE.
2. Solve the linear finite difference equations derived from previous exercise.

1.5 Linear difference equation

Second-order linear difference equation. In the linear case $y' = \lambda y$, the above difference scheme results in a linear difference equation. Let us consider general second order linear difference equation with constant coefficients:

$$ay^{n+1} + by^n + cy^{n-1} = 0, \tag{1.10}$$

where $a \neq 0$. To find its general solutions, we try the ansatz $y^n = \rho^n$ for some number ρ. Here, the n in y^n is an index, whereas the n in ρ^n is a power. Plug this ansatz into the equation, we get

$$a\rho^{n+1} + b\rho^n + c\rho^{n-1} = 0.$$

This leads to

$$a\rho^2 + b\rho + c = 0.$$

There are two solutions ρ_1 and ρ_2. In case $\rho_1 \neq \rho_2$, these two solutions are independent. Since the equation is linear, any linear combination of these two solutions is again a solution. Moreover, the general solution can only depend on two free parameters, namely, once y^0 and y^{-1} are known, then $\{y^n\}_{n\in\mathbb{Z}}$ is uniquely determined. Thus, the general solution is

$$y^n = C_1\rho_1^n + C_2\rho_2^n,$$

where C_1, C_2 are constants. In case of $\rho_1 = \rho_2$, then we can use the two solutions ρ_2^n and ρ_1^n with $\rho_2 \neq \rho_1$, but very closed, to produce another nontrivial solution:

$$\lim_{\rho_2\to\rho_1} \frac{\rho_2^n - \rho_1^n}{\rho_2 - \rho_1}$$

This yields the second solution is $n\rho_1^{n-1}$. Thus, the general solution is

$$C_1\rho_1^n + C_2 n\rho_1^{n-1}.$$

Linear finite difference equation of order r . We consider general linear finite difference equation of order r:

$$a_r y^{n+r} + \cdots + a_0 y^n = 0, \tag{1.11}$$

where $a_r \neq 0$. Since y^{n+r} can be solved in terms of $y^{n+r-1}, ..., y^n$ for all n, this equation together with initial data $y_0, ..., y_{-r+1}$ has a unique solution. The solution space is r dimensions.
To find fundamental solutions, we try the ansatz

$$y^n = \rho^n$$

for some number ρ. Plug this ansatz into equation, we get

$$a_r\rho^{n+r} + \cdots + a_0\rho^n = 0,$$

for all n. This implies

$$a(\rho) := a_r\rho^r + \cdots + a_0 = 0. \tag{1.12}$$

The polynomial $a(\rho)$ is called the characteristic polynomial of (1.11) and its roots $\rho_1, ..., \rho_r$ are called the characteristic roots.

- Simple roots (i.e. $\rho_i \neq \rho_j$, for all $i \neq j$): The fundamental solutions are ρ_i^n, $i = 1, ..., r$.
- Multiple roots: if ρ_i is a multiple root with multiplicity m_i, then the corresponding independent solutions
$$\rho_i^n, n\rho_i^{n-1}, C_2^n \rho_i^{n-2} ..., C_{m_i-1}^n \rho_i^{n-m_i+1}$$

Here, $C_k^n := n!/(k!(n-k)!)$. The solution $C_2^n \rho_i^{n-2}$ can be derived from differentiation $\frac{d}{d\rho} C_1^n \rho^{n-1}$ at ρ_i.

In the case of simple roots, we can express general solution as

$$y^n = C_1 \rho_1^n + \cdots + C_r \rho_r^n,$$

where the constants $C_1, ..., C_r$ are determined by

$$y^k = C_1 \rho_1^k + \cdots + C_r \rho_r^k, \quad k = 0, ..., r-1.$$

System of linear difference equation. The above rth order linear difference equation is equivalent to a first order linear difference system:

$$\mathbf{A}_0 \mathbf{y}^{n+1} = \mathbf{A} \mathbf{y}^n \tag{1.13}$$

where

$$\mathbf{y}^n = \begin{pmatrix} y_1^n \\ \vdots \\ y_r^n \end{pmatrix} = \begin{pmatrix} y^{n-r+1} \\ \vdots \\ y^n \end{pmatrix}$$

$$\mathbf{A}_0 = \begin{pmatrix} I_{(r-1)\times(r-1)} & 0 \\ 0 & a_r \end{pmatrix}, \ \mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{r-1} \end{pmatrix}.$$

We may divide (1.13) by $\mathbf{A}_0$ and get

$$\mathbf{y}^{n+1} = \mathbf{G} \mathbf{y}^n.$$

We call $\mathbf{G}$ the fundamental matrix of (1.13). For this homogeneous equation, the solution is

$$\mathbf{y}^n = \mathbf{G}^n \mathbf{y}^0$$

Next, we compute $\mathbf{G}^n$ in terms of eigenvalues of $\mathbf{G}$.

In the case that all eigenvalues ρ_i, $i = 1, ..., r$ of $\mathbf{G}$ are distinct, then $\mathbf{G}$ can be expressed as

$$\mathbf{G} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}, \ \mathbf{D} = \text{ diag } (\rho_1, \cdots, \rho_r),$$

and the column vectors of $\mathbf{T}$ are the corresponding eigenvectors.

When the eigenvalues of $\mathbf{G}$ have multiple roots, we can normalize it into Jordan blocks:

$$\mathbf{G} = \mathbf{T}\mathbf{J}\mathbf{T}^{-1},\ \mathbf{J} = \text{ diag } (\mathbf{J}_1, \cdots, \mathbf{J}_s),$$

where the Jordan block $\mathbf{J}_i$ corresponds to eigenvalue ρ_i with multiplicity m_i:

$$\mathbf{J}_i = \begin{pmatrix} \rho_i & 1 & 0 & \cdots & 0 \\ 0 & \rho_i & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & \rho_i \end{pmatrix}_{m_i \times m_i}.$$

and $\sum_{i=1}^{s} m_i = r$. Indeed, this form also covers the case of distinct eigenvalues.

In the stability analysis below, we are concerned with whether $\mathbf{G}^n$ is bounnded. It is easy to see that

$$\mathbf{G}^n = \mathbf{T}\mathbf{J}^n\mathbf{T}^{-1}, \mathbf{J}^n = \text{ diag } (\mathbf{J}_1^n, \cdots, \mathbf{J}_s^n)$$

$$\mathbf{J}_i^n = \begin{pmatrix} \rho_i^n & n\rho_i^{n-1} & C_2^n\rho^{n-2} & \cdots & C_{m_i-1}^n\rho_i^{n-m_i+1} \\ 0 & \rho_i^n & n\rho_i^{n-1} & \cdots & C_{m_i-2}^n\rho_i^{n-m_i+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & n\rho_i^{n-1} \\ 0 & 0 & 0 & \cdots & \rho_i^n \end{pmatrix}_{m_i \times m_i}.$$

where $C_k^n := \frac{n!}{k!(n-k)!}$.

Definition 1.2. The fundamental matrix $\mathbf{G}$ is called stable if $\mathbf{G}^n$ remains bounded under certain norm $\|\cdot\|$ for all n.

Theorem 1.3 (von Neumann)**.** *The fundamental matrix* $\mathbf{G}$ *is stable if and only if its eigenvalues satisfy the following condition:*

$$\begin{array}{l} \textit{either } |\rho| = 1 \textit{ and } \rho \textit{ is a simple root}, \\ \textit{or } |\rho| < 1 \end{array} \tag{1.14}$$

Proof. It is easy to see that the nth power of a Jordan form J_i^n is bounded if its eigenvalue $|\rho_i| < 1$ or if $\rho_i| = 1$ but simple. On the other hand, if $|\rho_i| > 1$ then J_i^n is unbounded; or if $\rho_i| = 1$ but not simple, then the term $n\rho_i^{n-1}$ in J_i^n will be unbounded. □

Corollary 1.1. *There exists a norm in* $\mathbb{R}^n$ *such that the above root condition for* $\mathbf{G}$ *is equivalent to* $\|\mathbf{G}\| \leq 1$ *with this norm.*

Proof. 1. First, in $\mathbb{R}^n$ (or $\mathbb{C}^n$), we define $\|\mathbf{x}\|_\infty = \max_i |x_i|$. For a linear mapping $\mathbf{G} : \mathbb{R}^n \to \mathbb{R}^n$, we define its operator norm under the $\|\cdot\|_\infty$ by

$$\|\mathbf{G}\|_\infty := \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{G}\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty}.$$

It is an easy exercise that for $\mathbf{G} = (a_{ij})_{n\times n}$, the operator norm

$$\|\mathbf{G}\|_\infty = \max_i \sum_j |a_{ij}|.$$

2. Second, a matrix $\mathbf{G}$ can be expressed as

$$\mathbf{G} = \mathbf{T}\mathbf{D}\mathbf{T}^{-1}, \quad \mathbf{D} = \text{ diag } (\mathbf{J}_1, \cdots, \mathbf{J}_s)$$

where $\mathbf{J}_i$ are Jordan blocks. For any $\epsilon_i \neq 0$, we can further transform $\mathbf{J}_i$ into

$$\mathbf{J}_i = \mathbf{S}_i \mathbf{K}_i \mathbf{S}_i^{-1}$$

where

$$\mathbf{K_i} = \begin{pmatrix} \rho_i & \epsilon & 0 & \cdots & 0 \\ 0 & \rho_i & \epsilon & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \epsilon \\ 0 & 0 & 0 & \cdots & \rho_i \end{pmatrix}_{m_i \times m_i}, \quad S_i = \text{ diag } (1, \epsilon_i, ..., \epsilon_i^{m_i - 1}).$$

Let $\mathbf{S} = \text{ diag}(\mathbf{S}_1, ..., \mathbf{S}_s)$, $\mathbf{K} = \text{ diag}(\mathbf{K}_1, ..., \mathbf{K}_s)$. We can express $\mathbf{G}$ as

$$\mathbf{G} = \mathbf{TS}\,\mathbf{K}(\mathbf{TS})^{-1}$$

We now define the new norm of $\mathbf{G}$ as

$$\|\mathbf{G}\| := \|\mathbf{K}\|_\infty$$

This means that we define the new norm $\|\cdot\|$ in $\mathbb{R}^n$ by

$$\|\mathbf{x}\| := \|(\mathbf{TS})^{-1}\mathbf{x}\|_\infty.$$

Since $\mathbf{TS}$ is invertible, this does define a norm in $\mathbb{R}^n$. With this norm, the corresponding operator norm is $\|\mathbf{K}\|_\infty$.

3. For those $\mathbf{J}_i$ with $m_i > 1$, the stability condition requires that $|\rho_i| < 1$. We choose ϵ_i such that $|\rho_i| + \epsilon_i \leq 1$. Then the corresponding $\|K_i\|_\infty \leq 1$. Thus, $\|\mathbf{G}\| \leq 1$ with the above operator norm.

□

Nonhomogeneous linear finite difference system In general, we consider the nonhomogeneous linear difference system:

$$\mathbf{y}^{n+1} = \mathbf{G}\mathbf{y}^n + \mathbf{f}^n \tag{1.15}$$

with initial data $\mathbf{y}^0$. Its solution can be expressed as

$$\begin{aligned} \mathbf{y}^n &= \mathbf{G}\mathbf{y}^{n-1} + \mathbf{f}^{n-1} \\ &= \mathbf{G}(\mathbf{G}\mathbf{y}^{n-2} + \mathbf{f}^{n-2}) + \mathbf{f}^{n-1} \\ &\vdots \\ &= \mathbf{G}^n\mathbf{y}^0 + \sum_{m=0}^{n-1} \mathbf{G}^{n-1-m}\mathbf{f}^m \end{aligned}$$

Homeworks.

1. Consider the linear ODE
$$y' = \lambda y$$
where λ considered here can be complex. Study the linear difference equation derived for this ODE by forward Euler method, backward Euler, midpoint. Find its general solutions.

2. Consider linear finite difference equation with source term
$$ay^{n+1} + by^n + cy^{n-1} = f^n$$
Given initial data $\bar{y}^0$ and $\bar{y}^1$, find its solution.

3. Find the characteristic roots for the Adams-Bashforth and Adams-Moulton schemes with steps 1-3 for the linear equation $y' = \lambda y$.

1.6 Stability analysis

There are two kinds of stability concepts.

- Zero stability: Fix $t = nk$, the computed solution y^n remains bounded as $n \to \infty$ (or equivalently, $k \to 0$).
- Absolute stability: Fix $k > 0$, the computed solution y^n remains bounded as $n \to \infty$.

1.6.1 Zero Stability

Our goal is to develop general convergence theory for multistep finite difference method for ODE: $y' = f(t, y)$ with initial condition $y(0) = y_0$. An r-step multistep finite difference scheme can be expressed as

$$\mathcal{L}y^n = \sum_{m=0}^{r} a_m y^{n+1-r+m} - k\sum_{m=0}^{r} b_m f(t^{n+1-r+m}, y^{n+1-r+m}) = 0. \tag{1.16}$$

Definition 1.3. The truncation error $\tau^n(k)$ for the above finite difference scheme is defined by

$$\tau^n(k) := \frac{1}{k}\left(\sum_{m=0}^{r} a_m y(t^{n+1-r+m}) - k\sum_{m=0}^{r} b_m f(t^{n+1-r+m}, y(t^{n+1-r+m}))\right),$$

where $y(\cdot)$ is a true solution of the ODE.

Definition 1.4. A difference scheme is called consistent if the corresponding truncation error $\tau^n(k) \to 0$ uniformly in n as the mesh size $k \to 0$. The scheme is of order p if $\tau^n(k) = O(k^p)$ uniform in n.

In the multistep method, the consistency is equivalent to $\tau(k) = O(k)$ because we assume $y(\cdot)$ is smooth and the truncation error is a smooth function in k. The consistency is $\tau(k) \to 0$ as $k \to 0$. Thus the smoothness of τ implies $\tau(k) = O(k)$.

Definition 1.5. A difference scheme is called zero stable if its solutions at time step n remain bounded as the mesh size $k \to 0$ ($nk = T$ is fixed, according $n \to \infty$).

The main theorem is the follows. We will postpone its proof at the end of this section.

Theorem 1.4 (Dahlquist equivalence theorem)**.** *For finite difference schemes for ODE $y' = f(t, y)$,*

$$\textit{consistency} \ + \ \textit{zero-stability} \ \Leftrightarrow \ \textit{convergence}$$

Stability criterion Let us investigate the condition on the coefficients a and b of an explicit multistep method for the stability

$$\mathcal{L}y^n = 0$$

to be bounded. We may assume $a_r = 1$ and $b_r = 0$. Let us write it in matrix form:

$$\mathbf{y}^{n+1} = \mathbf{A}\mathbf{y}^n + k\mathbf{B}\mathbf{f}^n$$

where

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & & & & \\ & 0 & 1 & & & \\ & & & \ddots & \ddots & \\ & & & & 0 & 1 \\ -a_0 & & \cdots & & -a_{r-2} & -a_{r-1} \end{pmatrix}, \ \mathbf{y}^n = \begin{pmatrix} y^{n-r} \\ \cdots \\ y^n \end{pmatrix},$$

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & & & & \\ & 0 & 0 & & & \\ & & & \ddots & \ddots & \\ & & & & 0 & 0 \\ b_0 & & \cdots & & b_{r-2} & b_{r-1} \end{pmatrix}, \ \mathbf{f}^n = \begin{pmatrix} f^{n-r} \\ \cdots \\ f^n \end{pmatrix},$$

In order to have solution to be bounded for a multistep scheme $\mathcal{L}y = 0$ for arbitrary f, it has at least to be valid when $f \equiv 0$. In this case, we need to invetigate the boundedness for the homogeneous equation:

$$\mathbf{y}^{n+1} = \mathbf{A}\mathbf{y}^n$$

We have seen in the last section that

Theorem 1.5. *The necessary and sufficient condition for $\|\mathbf{A}^n\|$ to be bounded is that the characteristic roots ρ_i of the characteristic equation $a(z) = 0$ satisfies:*

$$\begin{gathered} \textit{either } |\rho_i| < 1 \\ \textit{or } |\rho_i| = 1 \textit{ but simple}. \end{gathered}$$

Convergence ⇒ Stability

Proof. We only need to find an f such that the corresponding multistep is not stable implies that it does not converge. We choose $f \equiv 0$. * Since $\mathbf{A}^n$ is unbounded, which means there is an eigenvalue ρ_i with eigenvector $\mathbf{y}^i$ such that $|\rho_i| > 1$ or $|\rho_i| = 1$ but not simple. We discuss the formal case. The latter case can also be prove easily. In the former case, let $\mathbf{y}_i$ be the eigenvector of $\mathbf{A}$ corresponding to the eigenvalue ρ_i which satisfies $|\rho_i| > 1$. Let us choose y^0 and generate $\mathbf{y}^0 = (y_0^{r-1}, \cdots, y_0)^T$ by some explicit scheme starting from y^0. We can choose $\mathbf{y}^0$ such that its component on $\mathbf{y}_i$ is nonzero. Then the corresponding $\mathbf{y}^n := \mathbf{A}^n\mathbf{y}^0$ will be unbounded. Hence it cannot converge to a constant, as $k \to 0$. On the other hand, $\mathbf{y}^0$ depends on the mesh size k and $\mathbf{y}^0(k) \to (y_0, \cdots, y_0)^T$ as $k \to 0$. Thus, the method does not converge for $f \equiv 0$. □

Convergence ⇒ Consistency

Proof. From Theorem 1.2, we need to show that $a(1) = 0$ and $a'(1) = b(1)$. To show the first, we consider the ODE: $y' = 0$ with $y(0) = 1$. For the second, we consider the ODE: $y' = 1$ and $y(0) = 0$.

- Show $a(1) = 0$: We choose $\mathbf{y}^0 = (1, \cdots, 1)^T$. From $\mathbf{y}^1 = A\mathbf{y}^0$, we get

  $$y^r = -a_0 y^0 - \cdots - a_{r-1} y^{r-1} = -a_0 - \cdots - a_{r-1}.$$

  Since y^r is independent of k, and we should have $y^r \to 1$ as $k \to 0$ (by convergence), we conclude that $y^r = 1$. Thus, we get $a(1) = a_0 + \cdots + a_{r-1} + 1 = 0$.

- Show $a'(1) = b(1)$. We choose $f \equiv 1$, $y(0) = 0$. The corresponding ODE solution is $y(t) = t$. The multistep method gives

  $$a(Z)y^n - kb(Z)1 = 0. \tag{1.17}$$

  We write

  $$a(Z) = a'(1)(Z-1) + O((Z-1)^2), \quad b(Z)1 = b(1).$$

*Suppose a multistep method is convergence for every smooth f, then in particular, for $f \equiv 0$. In this case, if this multistep method is unstable, we want to show it does not converge. This is a contradiction.

Then the principal part of the above finite difference is

$$(Z-1)y - k\frac{b(1)}{a'(1)} = 0.$$

This is an arithmetic series. Its solution is $y^n = nk\frac{b(1)}{a'(1)}$. Indeed, this sequence also satisfies (1.17) provided its initial data y^n also has the form $y^n = nk\frac{b(1)}{a'(1)}$ for $0 \leq n < r$. Thus, arithmetic series $y^n = nk\frac{b(1)}{a'(1)}$ is a solution of the difference equation (1.17). Since $nk = t$, the convergence $y^n \to t$ as $n \to \infty$ enforces $\frac{b(1)}{a'(1)} = 1$.

□

Stability + Consistency ⇒ Convergence

Proof. We recall that we can express the scheme as

$$\mathbf{y}^{n+1} = \mathbf{A}\mathbf{y}^n + k\mathbf{B}\mathbf{f}^n.$$

Let Y be an exact solution, then plug it into the above scheme, we get

$$\mathbf{Y}^{n+1} = \mathbf{A}\mathbf{Y}^n + k\mathbf{B}\mathbf{F}^n + k\tau^n,$$

where $\mathbf{Y}^n := (Y(t^{n-r}), ...Y(t^n))^T$. We subtract these two and call $\mathbf{e}^n := \mathbf{Y}^n - \mathbf{y}^n$. We get

$$\mathbf{e}^{n+1} = \mathbf{A}\mathbf{e}^n + k\mathbf{B}\left(\mathbf{F}^n - \mathbf{f}^n\right) + k\tau^n.$$

The term $\mathbf{F}^n - \mathbf{f}^n$ can be repressed as

$$\begin{aligned}\mathbf{F}^n - \mathbf{f}^n &= (f(Y^{n-r}) - f(y^{n-r}), \cdots, f(Y^n) - f(y^n))^T \\ &= (L_{-r}e^{n-r}, \cdots, L_0 e^n)^T \\ &:= \mathbf{L}_n\mathbf{e}^n\end{aligned}$$

where

$$L_{-m} := \int_0^1 f'(y^{n-m} + te^{n-m})\,dt.$$

Thus, we get

$$\begin{aligned}\mathbf{e}^{n+1} &= \left(\mathbf{A} + k\mathbf{B}\mathbf{L}_n\right)\mathbf{e}^n + k\tau^n \\ &= \mathbf{G}_n(k)\mathbf{e}^n + k\tau^n\end{aligned}$$

with C independent of n and k. The reason is the follows. First, we assume that f is Lipschitz. Thus, the functions L_{-m} above are uniformly bounded (independent of n). Hence the term $\|\mathbf{BL}\|$ is uniformly bounded. Second we have a lemma

Lemma 1.1. *If $\|\mathbf{A}^n\|$ is bounded and $\|\mathbf{B}_n\|$ are uniformly bounded, then the product*

$$\|(\mathbf{A}+\frac{1}{n}\mathbf{B}_1)\cdots(\mathbf{A}+\frac{1}{n}\mathbf{B}_n)\|$$

is also uniformly bounded.

We have

$$\begin{aligned}\mathbf{e}^n &\leq \mathbf{G}_{n-1}\mathbf{e}^{n-1}+k\tau^{n-1}\\ &\leq \mathbf{G}_{n-1}\mathbf{G}_{n-2}\mathbf{e}^{n-2}+k\left(\mathbf{G}_{n-2}\tau^{n-2}+\tau^{n-1}\right)\\ &\leq \mathbf{G}_{n-1}\mathbf{G}_{n-2}\cdots\mathbf{G}_0\mathbf{e}^0\\ &\quad +k\left(\mathbf{G}_{n-2}\cdots\mathbf{G}_0\tau^0+\cdots+\mathbf{G}_{n-2}\tau^{n-2}+\tau^{n-1}\right)\end{aligned}$$

From the lemma, we get

$$\|\mathbf{e}^n\|\leq C\|\mathbf{e}^0\|+nkC\max_n\|\tau^n\|\leq C\|\mathbf{e}^0\|+O(k^p).$$

□

Proof of Lemma 1.1

Proof. 1. We have seen that $\|\mathbf{A}^n\|$ is uniformly bounded under some norm is equivalent to $\|\mathbf{A}\|\leq 1$ for some other operator norm. Thus, we may just assume $\|\mathbf{A}\|\leq 1$.

2. Since all norms in finite dimension are equivalent, we may assume $\|\mathbf{B}_i\|\leq b$ for all $i=1,...,n$.

3. We have

$$\|(\mathbf{A}+\frac{1}{n}\mathbf{B}_1)\cdots(\mathbf{A}+\frac{1}{n}\mathbf{B}_n)\|\leq(\|\mathbf{A}\|+\frac{b}{n})^n\leq(1+\frac{b}{n})^n\leq\exp(b).$$

□

Theorem 1.6 (First Dahlquist barrier)**.** *A zero-stable and linear r-step multistep method with p order of convergence should satisfies*

$$p\leq\begin{cases}r+2 \text{ if } r \text{ is even,}\\ r+1 \text{ if } r \text{ is odd,}\\ r \text{ if it is an explicit scheme.}\end{cases}$$

For proof, see pp. Hairer, Norsett, Wanner, Solving Ordinary Differential Equations, 384-387.

1.6.2 Absolute Stability

See Randall LeVeque, Finite Difference Methods for Ordinary and Partial Differential Equations, Chapter 7.

Chapter 2

Finite Difference Methods for Linear Parabolic Equations

2.1 Finite Difference Methods for the Heat Equation

2.1.1 Some discretization methods

Let us start from the simplest parabolic equation, the heat equation:

$$u_t = u_{xx}$$

Let $h = \Delta x$, $k = \Delta t$ be the spatial and temporal mesh sizes. Define $x_j = jh$, $j \in \mathbb{Z}$ and $t^n = nk$, $n \geq 0$. Let us abbreviate $u(x_j, t^n)$ by u_j^n. We shall approximate u_j^n by U_j^n, where U_j^n satisfies some finite difference equations.

Spatial discretization : The simplest one is that we use centered finite difference approximation for u_{xx}:

$$u_{xx} = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + O(h^2)$$

This results in the following systems of ODEs

$$\dot{U_j}(t) = \frac{U_{j+1}(t) - 2U_j(t) + U_{j-1}(t)}{h^2}$$

or in vector form

$$\dot{U} = \frac{1}{h^2} AU$$

where $U = (U_0, U_1, ...)^t$, $A = \text{diag } (1, -2, 1)$.

Homeworks.

1. Derive the 4th order centered finite difference approximation for u_{xx}:

$$u_{xx} = \frac{1}{h^2}(-u_{j-2} + 16u_{j-1} - 30u_j + 16u_{j+1} - u_{j+2}) + O(h^4).$$

2. Derive a 2nd order centered finite difference approximation for $(\kappa(x)u_x)_x$.

Temporal discretization We can apply numerical ODE solvers

- Forward Euler method:
$$U^{n+1} = U^n + \frac{k}{h^2}AU^n \tag{2.1}$$
- Backward Euler method:
$$U^{n+1} = U^n + \frac{k}{h^2}AU^{n+1} \tag{2.2}$$
- 2nd order Runge-Kutta (RK2):
$$U^{n+1} - U^n = \frac{k}{h^2}AU^{n+1/2},\ U^{n+1/2} = U^n + \frac{k}{2h^2}AU^n \tag{2.3}$$
- Crank-Nicolson:
$$U^{n+1} - U^n = \frac{k}{2h^2}(AU^{n+1} + AU^n). \tag{2.4}$$

These linear finite difference equations can be solved formally as

$$U^{n+1} = GU^n$$

where

- Forward Euler: $G = 1 + \frac{k}{h^2}A$,
- Backward Euler: $G = (1 - \frac{k}{h^2}A)^{-1}$,
- RK2: $G = 1 + \frac{k}{h^2}A + \frac{1}{2}\left(\frac{k}{h^2}\right)^2 A^2$
- Crank-Nicolson: $G = \frac{1+\frac{k}{2h^2}A}{1-\frac{k}{2h^2}A}$

For the Forward Euler, We may abbreviate it as

$$U_j^{n+1} = G(U_{j-1}^n, U_j^n, U_{j+1}^n), \tag{2.5}$$

where

$$G(U_{j-1}, U_j, U_{j+1}) = U_j + \frac{k}{h^2}(U_{j-1} - 2U_j + U_{j+1})$$

2.1.2 Stability and Convergence for the Forward Euler method

Our goal is to show under what condition can U_j^n converges to $u(x_j, t^n)$ as the mesh sizes $h, k \to 0$.

To see this, we first see the local error a true solution can produce. Plug a true solution $u(x,t)$ into (2.1). We get

$$u_j^{n+1} - u_j^n = \frac{k}{h^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right) + k\tau_j^n \tag{2.6}$$

where

$$\tau_j^n = D_{t,+}u_j^n - (u_t)_j^n - (D_+D_-u_j^n - (u_{xx})_j^n) = O(k) + O(h^2).$$

Let e_j^n denote for $u_j^n - U_j^n$. Then subtract (2.1) from (2.6), we get

$$e_j^{n+1} - e_j^n = \frac{k}{h^2}\left(e_{j+1}^n - 2e_j^n + e_{j-1}^n\right) + k\tau_j^n. \tag{2.7}$$

This can be expressed in operator form:

$$e^{n+1} = \mathbf{G}e^n + k\tau^n. \tag{2.8}$$

$$\begin{aligned}
\|e^n\| &\leq \|\mathbf{G}e^{n-1}\| + k\|\tau^{n-1}\| \\
&\leq \|\mathbf{G}^2e^{n-2}\| + k(\|\mathbf{G}\tau^{n-2}\| + \|\tau^{n-1}\|) \\
&\leq \|\mathbf{G}^ne^0\| + k(\|\mathbf{G}^{n-1}\tau^0\| + \cdots + \|\mathbf{G}\tau^{n-2}\| + \|\tau^{n-1}\|)
\end{aligned}$$

Suppose $\mathbf{G}$ satisfies the *stability* condition

$$\|\mathbf{G}^nU\| \leq C\|U\|$$

for some C independent of n. Then

$$\|e^n\| \leq C\|e^0\| + C\max_m|\tau^m|.$$

If the local truncation error has the estimate

$$\max_m\|\tau^m\| = O(h^2) + O(k)$$

and the initial error e^0 satisfies

$$\|e^0\| = O(h^2),$$

then so does the global true error satisfies

$$\|e^n\| = O(h^2) + O(k) \text{ for all } n.$$

The above analysis leads to the following definitions and the equivalence theorem.

Definition 2.1. A finite difference method is called consistent if its local truncation error τ satisfies

$$\|\tau_{h,k}\| \to 0 \text{ as } h, k \to 0.$$

Definition 2.2. A finite difference scheme $U^{n+1} = \mathbf{G}_{h,k}(U^n)$ is called stable in a region $(h,k) \in R$ if there exists a norm $\|\cdot\|$ such that

$$\|\mathbf{G}_{h,k}^n U\| \le C\|U\|$$

for all $n > 0$. Here, C is a constant independent of n.

Definition 2.3. A finite difference method is called convergence if the true error

$$\|e_{h,k}\| \to 0 \text{ as } h,k \to 0.$$

In the above analysis, we have seen that for forward Euler method for the heat equation,

Theorem 2.1. *The forward Euler method for the heat equation has the property:*

$$\textit{stability} \ + \ \textit{consistency} \ \Leftrightarrow \ \textit{convergence}.$$

We have already proven stability + consistency $\Rightarrow$ convergence. The proof of the other way is the same as the Dahlquist equivalent theorem 1.4.

2.2 L^2 Stability – von Neumann Analysis

We have seen from the above discussion that the convergence issue is reduced to the stability issue. In the stability analysis, we need to choose a norm to measure stability of the amplification operator $\mathbf{G}$, we will choose operator norm in L^2. For constant coefficient case, the von Neumann analysis (via Fourier method) provides a necessary and sufficient condition for stability. For more general cases such as variable coefficient cases, Kreiss' matrix theorem provides a good characterization of stability.

Below, we describe L^2 stability analysis. We introduce two equivalent methods: Fourier method and energy method, the Fourier method is also known as the von Neumann stability analysis. Given $\{U_j\}_{j\in\mathbb{Z}}$, let us define

$$\|U\|^2 = \sum_{j\in\mathbb{Z}} |U_j|^2$$

and its Fourier transform

$$\hat{U}(\xi) = \frac{1}{2\pi}\sum_{j\in\mathbb{Z}} U_j e^{-ij\xi}, \xi \in [0, 2\pi).$$

There are two advantages to analyze stability of a finite difference scheme using Fourier method.

- The shift operator is transformed to a multiplier:

  $$\widehat{TU}(\xi) = e^{i\xi}\hat{U}(\xi),$$

  where $(TU)_j := U_{j+1}$, thus the difference equation becomes an algebraic equation.

- The Parseval equality

$$\begin{aligned} \|U\|^2 &= \|\hat{U}\|^2 \\ &\equiv \int_{-\pi}^{\pi} |\hat{U}(\xi)|^2 \, d\xi \end{aligned}$$

holds, thus one can control the L^2-norm of U and $\mathbf{G}U$ in the Fourier space.

Now, let us consider a finite difference scheme of the form:

$$U_j^{n+1} = (\mathbf{G}U^n)_j = \sum_{k=-l}^{m} a_k U_{j+k}^n.$$

Taking Fourier transform, we obtain

$$\widehat{U^{n+1}} = \widehat{G}(\xi)\widehat{U^n}(\xi),$$

where

$$\widehat{G}(\xi) := \sum_{k=-l}^{m} a_k e^{ik\xi}.$$

From the Parseval equality,

$$\begin{aligned} \|U^{n+1}\|^2 &= \|\widehat{U^{n+1}}\|^2 \\ &= \int_{-\pi}^{\pi} |\widehat{G}(\xi)|^2 \, |\widehat{U^n}(\xi)|^2 \, d\xi \\ &\leq \max_{\xi} |\widehat{G}(\xi)|^2 \int_{-\pi}^{\pi} |\widehat{U^n}(\xi)|^2 \, d\xi \\ &= |\widehat{G}|_\infty^2 \|U\|^2 \end{aligned}$$

Thus a sufficient condition for stability is

$$|\widehat{G}|_\infty \leq 1. \tag{2.9}$$

Theorem 2.2. *A finite difference scheme*

$$U_j^{n+1} = \sum_{k=-l}^{m} a_k U_{j+k}^n$$

with constant coefficients is stable if

$$\widehat{G}(\xi) := \sum_{k=-l}^{m} a_k e^{ik\xi}$$

satisfies

$$\max_{-\pi \leq \xi \leq \pi} |\widehat{G}(\xi)| \leq 1. \tag{2.10}$$

For the forward Euler method for the heat equation,

$$U_j^{n+1} = G(U_{j-1}, U_j, U_{j+1}) = \alpha U_{j-1} + (1-2\alpha)U_j + \alpha U_{j+1}, \quad \alpha = \frac{k}{h^2}.$$

the corresponding

$$\widehat{G}(\xi) = \alpha(e^{i\xi} + e^{-i\xi}) + (1-2\alpha) = 1 - 4\alpha \sin^2\left(\frac{\xi}{2}\right).$$

The condition (2.10) is equivalent to

$$\alpha \leq \frac{1}{2}.$$

That is,

$$\frac{k}{h^2} \leq \frac{1}{2}.$$

Or equivalently, U_j^{n+1} is the convex combination of U_{j-1}, U_j and U_{j+1}.

Homeworks.

1. Compute the $\widehat{G}$ for the schemes: Backward Euler, RK2 and Crank-Nicolson.

2.3 Energy method

Let us write the finite difference scheme as

$$U_j^{n+1} = \alpha U_{j-1}^n + \beta U_j^n + \gamma U_{j+1}^n, \tag{2.11}$$

where

$$\alpha, \beta, \gamma \geq 0 \text{ and } \alpha + \beta + \gamma = 1.$$

We multiply (2.11) by U_j^{n+1} on both sides, apply Cauchy-Schwarz inequality, we get

$$\begin{aligned}(U_j^{n+1})^2 &= \alpha U_{j-1}^n U_j^{n+1} + \beta U_j^n U_j^{n+1} + \gamma U_{j+1}^n U_j^{n+1} \\ &\leq \frac{\alpha}{2}((U_{j-1}^n)^2 + (U_j^{n+1})^2) + \frac{\beta}{2}((U_j^n)^2 + (U_j^{n+1})^2) + \frac{\gamma}{2}((U_{j+1}^n)^2 + (U_j^{n+1})^2)\end{aligned}$$

Here, we have used $\alpha, \beta, \gamma \geq 0$. We multiply this inequality by h and sum it over $j \in \mathbb{Z}$. Denote

$$\|U\|_2 := \left(\sum_j |U_j|^2 h\right)^{1/2}.$$

We get

$$\|U^{n+1}\|^2 \leq \frac{\alpha}{2}(\|U^n\|^2 + \|U^{n+1}\|^2) + \frac{\beta}{2}(\|U^n\|^2 + \|U^{n+1}\|^2) + \frac{\gamma}{2}(\|U^n\|^2 + \|U^{n+1}\|^2)$$

$$= \frac{1}{2}(\|U^n\|^2 + \|U^{n+1}\|^2).$$

Here, $\alpha + \beta + \gamma = 1$ is applied. Thus, we get the energy estimate

$$\|U^{n+1}\|^2 \le \|U^n\|^2. \tag{2.12}$$

Homeworks.

1. Can the RK-2 method possess an energy estimate?

2.4 Stability Analysis via Entropy Estimates

Stability in the maximum norm

We notice that the action of G is a convex combination of U_{j-1}, U_j, U_{j+1}, provided

$$0 < \frac{k}{h^2} \le \frac{1}{2}. \tag{2.13}$$

Thus, we get

$$\min\{U_{j-1}^n, U_j^n, U_{j+1}^n\} \le U_j^{n+1} \le \max\{U_{j-1}^n, U_j^n, U_{j+1}^n\}.$$

This leads to

$$\min_j U_j^{n+1} \ge \min_j U_j^n,$$
$$\max_j U_j^{n+1} \le \max_j U_j^n,$$

and

$$\max_j |U_j^{n+1}| \le \max_j |U_j^n|.$$

That is, $\mathbf{G}$ is stable in $\|\cdot\|_\infty$.

Entropy estimates

The property that U^{n+1} is a convex combination (average) of U^n is very important. Given any convex function $\eta(u)$, by Jenson's inequality, we have*

$$\eta(U_j^{n+1}) \le \alpha\eta(U_{j-1}^n) + \beta\eta(U_j^n) + \gamma\eta(U_{j+1}^n). \tag{2.14}$$

---

*η is convex implies

$$\eta(\alpha U_{j-1} + (1-\alpha)V) \le \alpha\eta(U_{j-1}) + (1-\alpha)\eta(V).$$

Take $V = (\beta U_j + \gamma U_{j+1})/(1-\alpha)$. Apply the definition of convex function again, we get

$$\eta(V) \le \frac{\beta}{1-\alpha}\eta(U_j) + \frac{\gamma}{1-\alpha}\eta(U_{j+1}).$$

Combine these two inequalities, we get

$$\eta(\alpha U_{j-1} + \beta U_j + \gamma U_{j+1}) \le \alpha\eta(U_{j+1}) + \beta\eta(U_j) + \gamma\eta(U_{j+1}).$$

Summing over all j and using $\alpha + \beta + \gamma = 1$, we get

$$\sum_j \eta(U_j^{n+1}) \le \sum_j \eta(U_j^n). \tag{2.15}$$

The convex function is called *entropy* in this setting. The above inequality means that the "entropy" decreases in time. In particular, we choose

- $\eta(u) = |u|^2$, we recover the L^2 stability,
- $\eta(u) = |u|^p$, $1 \le p < \infty$, we get

  $$\sum_j |U_j^{n+1}|^p \le \sum_j |U_j^n|^p$$

  This leads to

  $$\left(\sum_j |U_j^{n+1}|^p h \right)^{1/p} \le \left(\sum_j |U_j^n|^p h \right)^{1/p},$$

  the general L^p stability. Taking $p \to \infty$, we recover L^∞ stability.
- $\eta(u) = |u - c|$ for any constant c, we obtain

  $$\sum_j |U_j^{n+1} - c| \le \sum_j |U_j^n - c|$$

  This is called Kruzkov's entropy estimate. We will see this inequality in hyperbolic theory again.

Homeworks.

1. Show that the solution of the difference equation derived from the RK2 satisfies the entropy estimate. What is the condition required on h and k for such entropy estimate?

2.5 Entropy estimate for backward Euler method

The backward Euler method for the heat equation is

$$U^{n+1} = U^n + \lambda A U^{n+1}, \quad A = \text{diag}(1, -2, 1), \quad \lambda = \frac{k}{h^2},$$

the amplification matrix is given by

$$U^{n+1} = GU^n, \quad G = (I - \lambda A)^{-1}. \tag{2.16}$$

The matrix $M := I - \lambda A$ has the following property:

Definition 2.4. A matrix $M = (m_{ij})$ is called an M-matrix if it satisfies

$$m_{ii} > 0,\ m_{ij} \le 0,\ \sum_{j \ne i} |m_{ij}| \le m_{ii} \tag{2.17}$$

For $M = I - \lambda A$ arisen from the backward Euler method, the corresponding $m_{ii} = 1 + 2\lambda$, $m_{ij} = -\lambda$ for $j \ne i$. Thus, it is an M-matrix.

Theorem 2.3. *The inverse of an M-matrix is a nonnegative matrix, i.e. all its entries are nonnegative.*

I shall not prove this general theorem. You can read Golub-von Loan's book, or consult wiki. Instead, I will find the inverse of M for the above specific M-matrix. Let us express

$$M = I - \lambda \,\text{diag}(1, -2, 1) = \frac{1 + 2\lambda}{2} \,\text{diag}\,(-a, 2, -a).$$

Here,

$$a = \frac{2\lambda}{1 + 2\lambda},\ \text{and } 0 < a < 1 \text{ if } h, k > 0.$$

The general solution of the difference equation

$$-au_{j-1} + 2u_j - au_{j+1} = 0 \tag{2.18}$$

has the form:

$$u_j = C_1\rho_1^j + C_2\rho_2^j$$

where ρ_1, ρ_2 are the characteristic roots, i.e. the roots of the polynomial equation

$$-a\rho^2 + 2\rho - a = 0.$$

Thus,

$$\rho_i = \frac{1 \pm \sqrt{1 - a^2}}{a}.$$

From the assumption of the M-matrix, $0 < a < 1$, we have $\rho_1 < 1$ and $\rho_2 > 1$.

Now, we define a fundamental solution:

$$g_j = \begin{cases} \rho_1^j & \text{for } j \ge 0 \\ \rho_2^j & \text{for } j < 0 \end{cases}.$$

We can check that $g_j \to 0$ as $|j| \to \infty$. Moreover, g_j satisfies the difference equation (2.18) for $|j| \ge 1$. For $j = 0$, we have

$$-ag_{-1} + 2g_0 - ag_1 = -a\rho_2^{-1} + 2 - a\rho_1 = 2 - a(\rho_1 + \rho_2^{-1}) = d$$

We reset

$$g_j \leftarrow g_j \left(\frac{2\lambda}{(1 + 2\lambda)d}\right).$$

Then we have

$$\frac{1+2\lambda}{2}\left(-ag_{-1}+2g_0-ag_1\right)=1,$$

$$\frac{1+2\lambda}{2}\left(-ag_{j-1}+2g_j-ag_{j+1}\right)=0,\forall j\neq 0.$$

This means

$$\sum_j g_{i-j}m_{j,k}=\delta_{i,k},$$

or

$$G(I-\lambda A)=Id.$$

Thus, $M^{-1}=(g_{i-j})$ is a positive matrix (i.e. all its entries are positive). Furthermore, from

$$g_{i-j}-\lambda\left(-g_{i-j-1}+2g_{i-j}-g_{i-j+1}\right)=\delta_{ij},$$

summing over $j\in\mathbb{Z}$, we obtain

$$\sum_j g_{i-j}=1 \text{ for all } i.$$

This means that

$$U_i^{n+1}=\left(GU^n\right)_i=\sum_{j\in\mathbb{Z}} g_{i-j}U_j^n$$

is indeed average of U^n with weights g_{i-j}. With this property, we can apply Jensen's inequality to get the entropy estimates:

Theorem 2.4. *Let $\eta(u)$ be a convex function. Let U_j^n be a solution of the difference equation derived from the backward Euler method for the heat equation. Then we have*

$$\sum_j \eta(U_j^n)\leq\sum_j \eta(U_j^0). \tag{2.19}$$

Remark 1.

- From entropy estimate, we get stability estimates for $\mathbf{G}$ in all L^p norms with $1\leq p\leq\infty$.
- It is important to note that there is no restriction on the mesh sizes h and k for stability for the Backward Euler method.

Homeworks.

1. Can the Crank-Nicolson method for the heat equation satisfy the entropy estimate? What is the condition on h and k?

2.6 Existence Theory

We can prove existence theorem of PDEs through finite difference approximation. In order to do so, let us define continuous and discrete Sobolev spaces and make a connection between them.

The continuous Sobolev space is defined as

$$H^m := \{u : \mathbb{R} \to \mathbb{R} | u, u', ..., u^{(m)} \in L^2(\mathbb{R})\}.$$

The discrete Sobolev space for functions defined on grid $G_h := \{x_j := jh | j \in \mathbb{Z}\}$.

$$H_h^m := \{U : G_h \to \mathbb{R} | U, D_{x,+}U, ..., D_{x,+}^m U \in \ell^2\}.$$

Here, $(D_{x,+}U)_j^n := (U_{j+1}^n - U_j^n)/h$

For any discrete function $U_j \in H_h^m$ we can construct a function u_h in H^m defined by

$$u_h(x) := \sum_j U_j \phi_h(x - x_j) \tag{2.20}$$

where $\phi_h(x) = \text{sinc}(x/h)$. We have

$$u_h(x_j) = U_j, \text{ for all } x_j \in G_h$$

It can be shown that

$$\|D_x^m u_h\| \approx \|D_{x,+}^m U\|. \tag{2.21}$$

Here, the norm is the L^2 norm. Similarly, the space $L_k^\infty(H_h^m)$ can be embeded into $L^\infty(H^m)$ by defining

$$u_{h,k}(x,t) = \sum_{n \geq 0} \sum_j U_j^n \phi_k(t) \phi_h(x)$$

The discrete norm and the continuous norm are equivalent.

2.6.1 Existence via forward Euler method

The forward Euler method for the heat equation $u_t = u_{xx}$ reads

$$U_j^{n+1} = U_j^n + \frac{k}{h^2} \left(U_{j-1}^n - 2U_j^n + U_{j+1}^n \right).$$

Here, We have seen that we can get the energy estimate:

$$\|U^n\| \leq \|U^0\|.$$

We perform finite difference operation on the above equation, say the forward Euler equation, for instance, let $V_j^n = (D_{x,+}U)_j^n := (U_{j+1}^n - U_j^n)/h$. Then V_j^n satisfies the same finite difference equation

$$V_j^{n+1} = V_j^n + \frac{k}{h^2} \left(V_{j-1}^n - 2V_j^n + V_{j+1}^n \right).$$

Thus, it also possesses the same energy estimate. Similar estimate for $D_{x,+}^2 U$. In general, we have

$$\|D_{x,+}^m U^n\| \leq \|D_{x,+}^m U^0\|. \tag{2.22}$$

If we assume the initial data $u_0 \in H^2$, then we get $U^n \in H_h^2$ for all $n \geq 0$.

Theorem 2.5. *If the initial data $u_0 \in H^m$,$m \geq 2$ and $k/h^2 \leq 1/2$, then the solution of forward Euler equation has the estimate*

$$\|D_{x,+}^m U^n\| \leq \|D_{x,+}^m U^0\|, \|D_{t,+} U^n\| \leq \|D_{x,+}^2 U^0\| \tag{2.23}$$

Further, the corresponding smoothing function $u_{h,k}$ has the same estimate and has a subsequence converges to a solution $u(x,t)$ of the original equation.

Proof. The functions $u_{h,k}$ are unformly bounded in $W^{1,\infty}(H^2)$. Hence they have a subsequence converges to a function $u \in W^{1,\infty}(H^2)$ weakly in $W^{1,\infty}(H^2)$ and strongly in $L^\infty(H^1)$. The functions $u_{h,k}$ satisfy

$$u_{h,k}(x_j, t^{n+1}) - u_{h,k}(x_j, t^n) = \frac{k}{h^2}(u_{h,k}(x_{j-1}, t^n) - 2u_{h,k}(x_j, t^n) + u_{h,k}(x_{j+1}, t^n))$$

Multiply a test smooth function ϕ, sum over j and n, take summation by part, we can get the subsequence converges to a solution of $u_t = u_{xx}$ *weakly*. □

2.6.2 A Sharper Energy Estimate for backward Euler method

In this subsection, we will get a sharper energy estimate for solutions obtained from the backward Euler method. Recall the backward Euler method for solving the heat equation is

$$U_j^{n+1} - U_j^n = \lambda(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}) \tag{2.24}$$

where $\lambda = k/h^2$. An important technique is the summation by part:

$$\sum_j (U_j - U_{j-1})V_j = -\sum_j U_j(V_{j+1} - V_j) \tag{2.25}$$

There is no boundary term because we consider periodic condition in the present case.

We multiply both sides by U_j^{n+1}, then sum over j. We get

$$\begin{aligned}
\sum_j (U_j^{n+1})^2 - U_j^{n+1}U_j^n &= \sum_j \lambda(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1})U_j^{n+1} \\
&= \lambda\left[\sum_j (U_{j-1}^{n+1} - U_j^{n+1})U_j^{n+1} + \sum_j (U_{j+1}^{n+1} - U_j^n)U_j^{n+1}\right] \\
&= \lambda\left[\sum_j (U_j^{n+1} - U_{j+1}^{n+1})U_{j+1}^{n+1} + \sum_j (U_{j+1}^{n+1} - U_j^n)U_j^{n+1}\right]
\end{aligned}$$

$$= -\lambda \sum_j |U_{j+1}^{n+1} - U_j^{n+1}|^2.$$

The term

$$U_j^{n+1} U_j^n \le \frac{1}{2}((U_j^{n+1})^2 + (U_j^n)^2)$$

by Cauchy-Schwartz. Hence, we get

$$\frac{1}{2}\sum_j \Big((U_j^{n+1})^2 - (U_j^n)^2)\Big) \le -\lambda \sum_j |U_{j+1}^{n+1} - U_j^{n+1}|^2$$

Or

$$\frac{1}{2} D_{t,-} \|U^{n+1}\|^2 \le -\frac{h^2}{k}\frac{k}{h^2}\|D_{x,+}U^{n+1}\| = -\|D_{x,+}U^{n+1}\|^2. \tag{2.26}$$

where,

$$D_{t,-}V_j^{n+1} := \frac{V_j^{n+1} - V_j^n}{k},\ D_{x,+}U_j^{n+1} := \frac{U_{j+1}^{n+1} - U_j^{n+1}}{h},$$

We sum in n from $n = 1$ to N, we get the following theorem.

Theorem 2.6. *For the backward Euler method, we have the estimate*

$$\|U^N\|^2 + C\sum_{n=1}^{N} \|D_{x,+}U^n\|^2 \le \|U^0\|^2 \tag{2.27}$$

This gives controls not only on $\|U^n\|^2$ but also on $\|D_{x,+}U^n\|$.

Homeworks.

1. Show that the Crank-Nicolson method also has similar energy estimate.
2. Can forward Euler method have similar energy estimate?

2.7 Relaxation of errors

In this section, we want to study the evolution of an error on a periodic domain $[0, 2\pi)$. We consider

$$u_t = u_{xx}, x \in [0, 2\pi), \tag{2.28}$$

with initial data u_0. The grid points $x_j = 2\pi j/N$ and $h = 2\pi/N$. The error $e_j^n := u(x_j, t^n) - U_j^n$ satisfies

$$e_j^{n+1} = e_j^n + \lambda(e_{j-1}^n - 2e_j^n + e_{j+1}^n) + k\tau_j^n. \tag{2.29}$$

We want to know how error is relaxed to zero from an initial error e^0. We study the homogeneous finite difference quation first. That is

$$e_j^{n+1} = e_j^n + \lambda(e_{j-1}^n - 2e_j^n + e_{j+1}^n). \tag{2.30}$$

or $e^{n+1} = G(u^n)$. The matrix is a tridiagonal matrix. It can be diagonalized by Fourier method. The eigenfunctions and eigenvalues are

$$v_{k,j} = e^{2\pi ijk/N}, \rho_k = 1 - 2\lambda + 2\lambda\cos(2\pi k/N) = 1 - 4\lambda\sin^2(\pi k/N), k = 0, ..., N-1.$$

When $\lambda \leq 1/2$, all eigenvalues are negative except ρ_0:

$$1 = \rho_0 > |\rho_1| > |\rho_2| > \cdots .$$

The eigenfunction

$$v_0 \equiv 1.$$

Hence, the projection of any discrete function U onto this eigenfunction is the average: $\sum_j U_j$.

Now, we decompose the error into

$$e^n = \sum_{k=0}^{N-1} e_k^n v_k, \quad n \geq 0$$

Then

$$e_k^{n+1} = \rho_k e_k^n.$$

Thus,

$$e_k^n = \rho_k^n e_k^0.$$

Since $\rho_0 = 1$, we see that $e_0^n = e_0^0$, which is the average of e^n, does not decay, unless $e_0^0 = 0$ initially. To guarantee the average of e^0 is zero, we may choose U_j^n to be the cell average of $u(x, t^n)$ in the jth cell:

$$U_j^n = \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t^n)\, dx.$$

instead of the grid data. This implies the initial error has zero local averages, and thus so does the global average.

For the decay behaviours of errors e_k^n for $k = 1, ..., N-1$, we notice that for $1 \leq k \leq N-1$,

$$\rho_k = 1 - 4\lambda\sin^2\left(\frac{\pi k}{N}\right) \approx 1 - 4\lambda\left(\frac{\pi k}{N}\right)^2, \text{ for } N >> 1.$$

The largest values of ρs are ρ_1 and ρ_{N-1}:

$$\rho_1 = \rho_{N-1} \approx 1 - 4\lambda\left(\frac{\pi}{N}\right)^2 = 1 - 4\frac{\Delta t}{h^2}\frac{\pi^2}{N^2} = 1 - \Delta t.$$

They correspond to low frequency eigenmodes: $v_1 = (e^{2\pi ij/N})_{j=0}^{N-1}$ and $v_{N-1} = (e^{-2\pi ij/N})_{j=0}^{N-1}$. The decay rate

$$\rho_1^n \approx (1 - \Delta t)^n \approx e^{-t}.$$

Here, $t = n\Delta t$. This is the decay rate of e_1^n and e_{N-1}^n with $n\Delta t = t$. They are the slowest decay modes. For $k = N/2$, the corresponding eigenmode $v_{N/2} = ((-1)^j)_{j=0}^{N-1}$ is the highest frequency mode. The corresponding eigenvalue

$$\rho_{N/2} = 1 - 4\lambda = 1 - 4\frac{\Delta t}{h^2}.$$

The decay rate is

$$\rho_{N/2}^n = \left(1 - 4\frac{\Delta t}{h^2}\right)^n \approx e^{-\frac{4t}{h^2}}.$$

which decays very fast.

The contribution of the truncation error to the true error is:

$$e^{n+1} = \rho_k e_k^n + \Delta t \tau_k^n$$

Its solution is

$$e_k^n = \rho_k^n e_k^0 + \Delta t \sum_{m=0}^{n-1} \rho_k^{n-1-m} \tau_k^m$$

We see that the term e_0^n does not tend to zero unless $\tau_0^m = 0$. This can be achieved if we choose U_j as well as f_j to be the cell averages instead the grid data. We have seen that the truncation error is second order. That is

$$\tau_{k,\max} := \max_{0 \le m \le n-1} |\tau_k^m| = O(h^2).$$

Then for $k \ge 1$,

$$\Delta t \sum_{m=0}^{n-1} |\rho_k|^{n-1-m} \le \Delta t \sum_{m=0}^{n-1} |\rho_1|^{n-1-m} = \Delta t \frac{1-\rho_1^n}{1-\rho_1} \approx \Delta t \frac{1-e^{-t}}{1-(1-\Delta t)} = 1 - e^{-t}.$$

Thus, we obtain

$$|e^n| \le e^{-t} e^0 + (1 - e^{-t}) O(h^2)$$

with $n\Delta t = t$.

Homeworks.

1. Define $U_j := \frac{1}{h} \int_{x_{j-1/2}}^{x_{j+1/2}} u(x)\, dx$. Show that if $u(x)$ is a smooth periodic function on $[0, 1]$, then

   $$u''(x_j) = \frac{1}{h^2}(U_{j-1} - 2U_j + U_{j+1}) + \tau$$

   with $\tau = O(h^2)$.

2.8 Boundary Conditions

2.8.1 Dirichlet boundary condition

Now, we consider the initial-boundary problem:

$$u_t = u_{xx}, \quad x \in [0,1]$$

The Dirichlet boundary condition is

$$u(0) = a,\ u(1) = b. \tag{2.31}$$

The initial condition is

$$u(x,0) = u_0(x).$$

We introduce uniform grids: $x_j = j/N$, $j = 0, ..., N$. The forward Euler method can be realized on $x_1, ..., x_{N-1}$ as

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{h^2}\left(U_{j-1}^n - 2U_j^n + U_{j+1}^n\right), j = 1, ..., N-1.$$

Near the boundary point x_1, the finite difference approximation of u_{xx} at x_1 involves u at $x_0 = 0$. We plug the boundary condition:

$$u_{xx}(x_1) = \frac{U_0 - 2U_1 + U_2}{h^2} + O(h^2) = \frac{a - 2U_1 + U_2}{h^2} + O(h^2) \tag{2.32}$$

Similarly,

$$u_{xx}(x_{N-1}) = \frac{U_{N-2} - 2U_{N-1} + U_N}{h^2} + O(h^2) = \frac{U_{N-2} - 2U_{N-1} + b}{h^2} + O(h^2)$$

The unknowns are $U_1^n, ..., U_{N-1}^n$ with $N-1$ finite difference equations at $x_1, ..., x_{N-1}$. Including boundary terms, we write the equation as

$$U^{n+1} = (I + \lambda A)U^n + \lambda B, \quad \lambda = \frac{\Delta t}{h^2},$$

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 \end{pmatrix}_{(N-1)\times(N-1)}, \quad B = \begin{pmatrix} a \\ 0 \\ \vdots \\ b \end{pmatrix}_{(N-1)\times 1}. \tag{2.33}$$

The matrix A is the discrete Laplacian with zero Dirichlet boundary condition. The term B comes from the Dirichlet boundary conditions.

We can have energy estimates, entropy estimates as the case of periodic boundary condition.

Next, we examine how error is relaxed for the Euler method with zero Dirichlet boundary condition. We have seen that the error $e_j^n := u(x_j, t^n) - U_j^n$ satisfies the difference equation with truncation error as the source term:

$$e_j^{n+1} = ((I + \lambda A)e^n)_j + \lambda \delta B + \Delta t \tau_j^n, \quad j = 1, ..., N-1$$

where τ_j^n is the truncation error,

$$\delta B = \begin{pmatrix} \delta a \\ 0 \\ \vdots \\ \delta b \end{pmatrix}_{(N-1)\times 1}.$$

If there is any error from $u(0) = a$, say $U_0 = a + \delta a$, it will creates a truncation error $\delta a/h^2$ at x_1. The solution for this difference equation is

$$e^n = G^n e^0 + \sum_{m=0}^{n-1} G^{n-1-m} \left(\lambda \delta B + \Delta t \tau^m\right), \quad G = I + \lambda A.$$

From Fourier method, we can compute the eigenvector and eigenvalue of A:

$$v_k = (\sin(\pi jk/N)_{j=1}^{N-1}, \quad \lambda_k = -4\sin^2(\pi k/(2N)), \quad k = 1, ..., N-1.$$

In fact, we extend v_k to an $N+1$-vector as $v_{k,0} = v_{k,N} = 0$. Using this extended vector, we can check

$$v_{k,j-1} - 2v_{k,j} + v_{k,j+1} = (2\cos(\pi k/N) - 2)\sin(\pi jk/N) = -4\sin^2\left(\frac{\pi k}{2N}\right) v_{k,j}, \quad j = 1, ..., N-1, k = 1, ..., N-1.$$

The eigenvalues of $I + 4\lambda A$ are

$$\rho_k = 1 - 4\lambda \sin^2\left(\frac{\pi k}{2N}\right), \quad k = 1, ..., N-1.$$

In the present case, all eigenvalues

$$\rho_k < 1, k = 1, ..., N-1.$$

provided the stability condition

$$\lambda \leq 1/2.$$

In this case,

$$1 > \rho_1 > |\rho_2| > ... > |\rho_{N-1}|.$$

The lowest mode is ρ_1, which is

$$\rho_1 = 1 - 4\lambda \sin^2(\pi/2N) \approx 1 - \lambda\left(\frac{\pi}{N}\right)^2 = 1 - \frac{\Delta t}{h^2}\frac{\pi^2}{N^2} = 1 - \pi^2 \Delta t.$$

and

$$\rho_1^n \approx \left(1 - \pi^2 \Delta t\right)^n \approx e^{-\pi^2 t}$$

Thus,

$$\|G^n\| \leq e^{-\pi^2 t}, \quad n\Delta t = t.$$

The accumulation effect is

$$\sum_{m=0}^{n-1} \|G^{n-1-m}\| \leq \frac{1 - e^{-\pi^2 t}}{\Delta t}.$$

Thus, the error from the initial data is

$$\|G^n e^0\| \leq e^{-\pi^2 t} \|e^0\|$$

The error coming from truncation is

$$\sum_{m=0}^{n-1} G^{n-1-m} \left(\Delta t \tau^m\right) = (1 - e^{-\pi^2 t}) O(h^2).$$

The error due to boundary is

$$\sum_{m=0}^{n-1} G^{n-1-m} \left(\lambda \delta B\right) = (1 - e^{-\pi^2 t}) \frac{1}{h^2} \|\delta B\|.$$

2.8.2 Neumann boundary condition

The Neumann boundary condition is

$$u'(0) = \sigma_0, \; u'(1) = \sigma_1. \tag{2.34}$$

We may use the following discretization methods:

- First order:

$$\frac{U_1 - U_0}{h} = \sigma_0.$$

- Second order-I:

$$\frac{U_1 - U_0}{h} = u_x(x_{1/2}) = u_x(0) + \frac{h}{2} u_{xx}(x_0) = \sigma_0 + \frac{h}{2} f(x_0) \tag{2.35}$$

- Second order-II: we use extrapolation

$$\frac{-3U_0 + 4U_1 - U_2}{2h^2} = \sigma_0.$$

The knowns are U_j^n with $j = 0, ..., N$. In the mean time, we add two more equations at the boundaries.

Homeworks.

1. Find the eigenfunctions and eigenvalues for the discrete Laplacian with the Neumann boundary condition (consider both first order and second order approximation at boundary). Notice that there is a zero eigenvalue.

   Hint: You may use Matlab to find the eigenvalues and eigenvectors.

Here, I will provide another method. Suppose A is the discrete Laplacian with Neumann boundary condition. A is an $(N+1) \times (N+1)$ matrix. Suppose $Av = \lambda v$. Then for $j = 1, ..., N-1$, v satisfies

$$v_{j-1} - 2v_j + v_{j+1} = \lambda v_j, j = 1, ..., N-1.$$

For v_0, we have

$$-2v_0 + 2v_1 = \lambda v_0.$$

For v_N, we have

$$-2v_N + 2v_{N-1} = \lambda v_N.$$

Then this matrix has the following eigenvectors:

$$v_j^k = \cos(\pi jk/N), \quad k = 0, ..., N$$

with eigenvalue

$$\lambda^k = -2 + 2\cos(\pi k/N) = -4\sin^2\left(\frac{\pi k}{2N}\right), \quad k = 0, ..., N.$$

Notice that $\lambda^0 = 0$. The error corresponding this eigenmode does not decay.

Homeworks.

1. Complete the calculation.
2. Consider

   $$u_t = u_{xx} + f(x)$$

   on $[0, 1]$ with Neumann boundary condition $u'(0) = u'(1) = 0$. If $\int f(x)\, dx \neq 0$. What wil happen to u as $t \to \infty$?

2.9 The discrete Laplacian and its inversion

We consider the elliptic equation

$$u_{xx} - \alpha u = f(x), x \in (0, 1),$$

with the Dirichlet boundary condition

$$u(0) = a,\ u(1) = b. \tag{2.36}$$

The finite difference approximation of u_{xx} at x_1 involves u at $x_0 = 0$. We plug the boundary contion:

$$u_{xx}(x_1) = \frac{U_0 - 2U_1 + U_2}{h^2} + O(h^2) = \frac{a - 2U_1 + U_2}{h^2} + O(h^2)$$

Similarly,

$$u_{xx}(x_{N-1}) = \frac{U_{N-2} - 2U_{N-1} + U_N}{h^2} + O(h^2) = \frac{U_{N-2} - 2U_{N-1} + b}{h^2} + O(h^2)$$

The unknowns are $U_1^n, ..., U_{N-1}^n$ with $N-1$ finite difference equations at $x_1, ..., x_{N-1}$. The discrete Laplacian becomes

$$A = \begin{pmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 1 & -2 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 \end{pmatrix}_{(N-1)\times(N-1)} . \tag{2.37}$$

This is a discrete Laplacian with Dirichlet boundary condition. We have seen in the last section that A can be diagonalized by the discrete Fourier sin functions $v^k = (\sin(\pi jk/N)_{j=1^{N-1}}$. In this section, we will find explicit expression of A^{-1}. Indeed, we will find $(A-2\beta)^{-1}$ where $\beta = \alpha h^2/2$. The difference equation

$$U_{j-1} - (2 + 2\beta)U_j + U_{j+1} = 0$$

has two independent solutions ρ_1^j and ρ_2^j, where ρ_i are roots of

$$\rho^2 - (2 + 2\beta)\rho + 1 = 0.$$

That is

$$\rho = 1 + \beta \pm \sqrt{(1+\beta)^2 - 1}.$$

Our goal below is to construct fundamental solution G_{ij}, which is $G = (A - 2\beta)^{-1}$.

Case 1: $\beta = 0$ When $\beta = 0$, the two independent solutions are $U_j = 1$ and $U_j = j$. Let us construct the fundamental solution centered at i, call it G_{ij}. It has the form:

$$G_{i,j} = \begin{cases} jC_i & j \le i, \\ (N-j)C_i' & j \ge i, \end{cases} \qquad 1 \le i, j \le (N-1). \tag{2.38}$$

for some constants C_i and C_i'. From $G_{i,i-1} - 2G_{i,i} + G_{i,i+1} = 1$ and $iC_i = (N-i)C_i'$, we obtain that

$$C_i = -(N-i)/N \quad C_i' = -i/N.$$

This gives explicit formula of $G = A^{-1}$.

Case 2: $\beta > 0$ When $\beta > 0$, the two roots are $\rho_1 < 1$ and $\rho_2 > 1$. The fundamental solution G_{ij} has the following form

$$G_{ij} = \begin{cases} C_1\rho_1^j + C_2\rho_2^j & \text{for } j \le i \\ D_1\rho_1^j + D_2\rho_2^j & \text{for } j \ge i \end{cases} \qquad 1 \le i \le (N-1), \quad 0 \le j \le N.$$

Here, we extend $G_{i,j}$ with $1 \le j \le (N-1)$ to $0 \le j \le N$. The constants C_1, C_2, D_1, D_2 are determined by

$$G_{i0} = 0, \quad G_{i,N} = 0,$$
$$G_{i,i-1} - (2+2\beta)G_{i,i} + G_{i,i+1} = 1$$
$$C_1\rho_1^i + C_2\rho_2^i = D_1\rho_1^i + D_2\rho_2^i.$$

Homeworks.

1. Find the coefficients C_1, C_2, D_1, D_2 above.

Let us go back to the original equation:

$$u_{xx} - \alpha u = f(x)$$

The above study of the Green's function of the discrete Laplacian helps us to quantify the the error produced from the source term. If $Au = f$ and $A^{-1} = G$, then an error in f, say τ, will produce an error

$$e = G\tau.$$

The error from the boundary also has the same behaviour. If the off-diagonal part of G decays exponentially (i.e. $\beta > 0$), then the error is "localized," otherwise, it pollutes everywhere.

Project 2. Solve the following equation

$$u_{xx} - \alpha u + f(x) = 0, \ x \in [0,1]$$

numerically with periodic, Dirichlet and Neumann boundary condition. The equilibrium

1. A layer structure
$$f(x) = \begin{cases} -1 & 1/4 < x < 3/4 \\ 1 & \text{otherwise} \end{cases}$$
2. An impluse
$$f(x) = \begin{cases} \gamma & 1/2 - \delta < x < 1/2 + \delta \\ 0 & \text{otherwise} \end{cases}$$
3. A dipole
$$f(x) = \begin{cases} \gamma & 1/2 - \delta < x < 1/2 \\ -\gamma & 1/2 < x < 1/2 + \delta \\ 0 & \text{otherwise} \end{cases}$$

You may choose $\alpha = 0.1, 1, 10$, observe how solutions change as you vary α.

Project 3. Solve the following equation

$$-u_{xx} + f(u) = g(x),\ x \in [0, 1]$$

numerically with Neumann boundary condition. Here, $f(u) = F'(u)$ and the potential is

$$F(u) = u^4 - \gamma u^2.$$

Study the solution as a function of γ. Choose simple g, say piecewise constant, a delta function $\delta(x - x_0)$, or a dipole $\delta(x - x_0 + \epsilon) - \delta(x - x_0 - \epsilon)$.

Chapter 3

Finite Difference Methods for Linear Elliptic Equations

3.1 Discrete Laplacian in two dimensions

We will solve the Poisson equation

$$\triangle u = f$$

in a domain $\Omega \subset \mathbb{R}^2$ with Dirichlet boundary condition

$$u = g \text{ on } \partial\Omega$$

Such a problem is a core problem in many applications. We may assume $g = 0$ by substracting a suitable function from u. Thus, we limit our discussion to the case of zero boundary condition. Let h be the spatial mesh size. For simplicity, let us assume $\Omega = [0, 1] \times [0, 1]$. But many discussion below can be extended to general smooth bounded domain.

3.1.1 Discretization methods

Centered finite difference The Laplacian is approximated by

$$A = \frac{1}{h^2} \left(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{i,j} \right).$$

For the square domain, the indeces run from $1 \leq i, j \leq N - 1$ and

$$U_{0,j} = U_{N,j} = U_{i,0} = U_{i,N} = 0$$

from the boundary condition.

If we order the unknowns U by $i + j * (N-1)$ with j being outer loop index and i the inner loop index, then the matrix form of the discrete Laplacian is

$$A = \frac{1}{h^2} \begin{pmatrix} T & I & & & \\ I & T & I & & \\ & I & T & I & \\ & & \ddots & \ddots & \ddots \\ & & & I & T \end{pmatrix}$$

This is an $(N-1) \times (N-1)$ block tridiagonal matrix. The block T is an $(N-1) \times (N-1)$ matrix

$$T = \begin{pmatrix} -4 & 1 & & & \\ 1 & -4 & -1 & & \\ & 1 & -4 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -4 \end{pmatrix}$$

Since this discrete Laplacian is derived by centered finite differencing over uniform grid, it is second order accurate, the truncation error

$$\begin{aligned} \tau_{i,j} &:= \frac{1}{h^2} \left(u(x_{i-1}, y_j) + u(x_{i+1}, y_j) + u(x_i, y_{j-1}) + u(x_i, y_{j+1}) - 4u(x_i, y_j) \right) \\ &= O(h^2). \end{aligned}$$

3.1.2 The 9-point discrete Laplacian

The Laplacian is approximated by

$$\nabla_9^2 = \frac{1}{6h^2} \begin{pmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{pmatrix}$$

One can show by Taylor expansion that

$$\nabla_9^2 u = \nabla^2 u + \frac{1}{12} h^2 \nabla^4 u + O(h^4).$$

If u is a solution of $\nabla^2 u = f$, then

$$\nabla_9^2 u = f + \frac{1}{12} h^2 \nabla^2 f + O(h^4).$$

Thus, we get a 4th order method:

$$\nabla_9^2 U_{ij} = f_{ij} + \frac{h^2}{12} \nabla^2 f_{ij}$$

3.2 Stability of the discrete Laplacian

We have seen that the true solution of $\triangle u = f$ with Dirichlet boundary condition satisfies

$$Au = f + \tau,$$

where A is the discrete Laplacian and τ is the truncation error and satisfies $\tau = O(h^2)$ in maximum norm. The numerical solution U satisfies $AU = f$. Thus, the true error satisfies

$$Ae = \tau,$$

where $e = u - U$. Thus, e satisfies the same equation with right-hand side τ and with the Dirichlet boundary condition. To get the convergence result, we need an estimate of e in terms of τ. This is the stability criterion of A. We say that A is stable if there exists some norm $\|\cdot\|$ and a constant C such that

$$\|e\| \leq C\|Ae\|.$$

3.2.1 Fourier method

Since our domain $\Omega = [0,1] \times [0,1]$ and the coefficients are constant, we can apply Fourier transform. Let us see one dimensional case first. Consider the Laplacian d^2/dx^2 on domain $[0,1]$ with Dirichlet boundary condition. The discrete Laplacian is $A = \frac{1}{h^2}\text{diag}\,(1,-2,1)$, where $h = 1/N$. We can check below that the eigenvectors of A are $v_k = (\sin(\pi jkh))_{j=1}^{N-1}$, $k = 1, ..., N-1$. The corresponding eigenvalues are $-\frac{4}{h^2}\sin^2(\pi hk/2)$.

$$\begin{aligned}[Av_k]_j = [A\sin(j\pi kh)]_j &= \frac{1}{h^2}\left(\sin((j+1)\pi hk) + \sin((j-1)\pi hk) - 2\sin(j\pi hk)\right) \\ &= \left[\frac{2}{h^2}(\cos(\pi hk) - 1)\right]\sin(j\pi hk) = -\frac{4}{h^2}\sin^2(\pi hk/2)[v_k]_j.\end{aligned}$$

For two dimensional case, the eigenfunctions of the discrete Laplacian are $U^{k,\ell}$, $1 \leq, k, \ell \leq N-1$,

$$(U^{k,\ell})_{i,j} = \sin(i\pi kh)\sin(j\pi \ell h), \quad 1 \leq i, j \leq N-1.$$

The corresponding eigenvalues are

$$\begin{aligned}\lambda^{k,\ell} &= \frac{2}{h^2}(\cos(k\pi h) + \cos(\ell\pi h) - 2) \\ &= -\frac{4}{h^2}(\sin^2(k\pi h/2) + \sin^2(\ell\pi h/2)), \quad 1 \leq, k, \ell \leq N-1.\end{aligned}$$

The smallest eigenvalue (in magnitude) is

$$\lambda^{1,1} = -\frac{8}{h^2}\sin^2(\pi h/2) \approx -2\pi^2 \quad \text{for small } h.$$

To show the stability, we take Fourier transform of U and A. We then have

$$\|\widehat{A}\hat{U}\|\|\hat{U}\| \geq \left|\langle \widehat{A}\hat{U}, \hat{U}\rangle\right| \geq |\lambda^{1,1}|\|\hat{U}\|^2 \approx 2\pi^2\|\hat{U}\|^2.$$

Hence, the L^2 norm of $\widehat{A}$ has the following estimate:

$$\|\widehat{A}\hat{U}\| \geq 2\pi^2\|\hat{U}\|.$$

Thus, we get

$$\|\hat{U}\| \leq \frac{1}{2\pi^2}\|\widehat{A}\hat{U}\|.$$

From Parseval equality, we have

$$\|U\| \leq \frac{1}{2\pi^2}\|AU\|$$

Applying this stability to the formula: $Ae = \tau$, we get

$$\|e\| \leq \frac{1}{2\pi^2}\|\tau\| = O(h^2).$$

Homeworks.

1. Compute th eigenvalues and eigenfunctions of the 9-point discrete Laplacian on the domain $[0,1] \times [0,1]$ with zero boundary condition.

3.2.2 Energy method

Below, we use energy method to prove the stability result for discrete Laplacian. We shall prove it for rectangular domain. However, it can be extended to more general domain. To perform energy estimate, we rewrite the discrete Laplacian as

$$AU_{i,j} = \frac{1}{h^2}\left(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{i,j}\right) = ((D_{x+}D_{x-} + D_{y+}D_{y-})U)_{i,j}$$

where

$$(D_{x+}U)_{i,j} = \frac{U_{i+1,j} - U_{i,j}}{h}$$

the forward differencing. We multiply the discrete Laplacian by $U_{i,j}$, then sum over all i, j. By applying the summation by part, we get

$$\begin{aligned}\langle AU, U\rangle &= \langle (D_{x+}D_{x-} + D_{y+}D_{y-})U, U\rangle \\ &= -\langle D_{x-}U, D_{x-}U\rangle - \langle D_{y-}U, D_{y-}U\rangle \\ &= -\|\nabla_h U\|_h^2\end{aligned}$$

Here, the discrete L^2 norm is defined by

$$\|U\|_h^2 = \sum_{i,j} |U_{i,j}|^2 h^2.$$

The boundary term does not show up because we consider the zero Dirichlet boundary problem. Thus, the discrete Poisson equation has the estimate

$$\|\nabla_h U\|_h^2 = |\langle f, U\rangle| \le \|f\|_h \|U\|_h. \tag{3.1}$$

Next, for the zero Dirichlet boundary condition, we have the Poincaré inequality, which will be shown below. Before stating the Poincare inequality, we need to clarify the meaning of zero boundary condition in the discrete sense. We define the Sobolev space $H^1_{h,0}$ to be the completion of the restriction of all C^1_0 functions to the grid points under the discrete H^1 norm. Here, C^1_0 function is a C^1 function that is zero on the boundary; the discrete H^1 norm is

$$\|U\|_{h,1} := \|U\|_h + \|\nabla_h U\|_h.$$

Lemma 3.1. *Let Ω be a bounded domain in $\mathbb{R}^2$, then there exist a constant d_Ω, which is the diameter of the domain Ω, such that for any $U \in H^1_{h,0}$,*

$$\|U\|_h \le d_\Omega \|\nabla_h U\|_h. \tag{3.2}$$

Proof. Let us take $\Omega = [0, X] \times [0, Y]$ as an example for the proof. We assume $X = Mh, Y = Nh$. From zero boundary condition, we have

$$\begin{aligned} U_{i,j}^2 &= (\sum_{i'=1}^{i} D_{x-} U_{i',j} h)^2 \\ &\le (\sum_{i'=1}^{i} 1^2) \cdot (\sum_{i'=1}^{i} (D_{x-} U_{i',j})^2) h^2 \text{ (Hölder's inequa;ity)} \\ &\le i(\sum_{i'=1}^{M} (D_{x-} U_{i',j})^2) h^2 \end{aligned}$$

multiply both sides by h^2 then sum over all i, j, we get

$$\begin{aligned} \|U\|_h^2 &= \sum_{i,j} U_{i,j}^2 h^2 \\ &\le (\sum_{i=1}^{M} i) h^2 \sum_{i',j} (D_{x-} U_{i',j})^2 h^2 \\ &\le \frac{M^2}{2} h^2 \sum_{i',j} (D_{x-} U_{i',j})^2 h^2 \\ &= \frac{M^2}{2} h^2 \|D_{x-} U\|_h^2 \end{aligned}$$

Similarly, we have

$$\|U\|_h^2 \le \frac{N^2}{2} h^2 \|D_{y-} U\|_h^2$$

Thus,

$$\begin{aligned} \|U\|_h^2 &\leq h^2 \frac{1}{2} \max\{M^2, N^2\} \|\nabla_h U\|^2 \\ &\leq d_\Omega^2 \|\nabla_h U\|_h^2. \end{aligned}$$

□

With the Poincare inequality, we can obtain two estimates for U.

Proposition 3.1. *Consider the discrete Laplacian with zero boundary condition. We have*

$$\|U\|_h \leq d_\Omega^2 \|f\|_h, \tag{3.3}$$

$$\|\nabla_h U\| \leq d_\Omega \|f\|_h. \tag{3.4}$$

Proof. From

$$\|\nabla_h U\|_h^2 \leq \|f\|_h \cdot \|U\|_h$$

We apply the Poincare inequality to the left-hand side, we obtain

$$\|U\|_h^2 \leq d_\Omega^2 \|\nabla U\|_h^2 \leq d_\Omega^2 \|f\|_h \|U\|_h$$

This yields

$$\|U\|_h \leq d_\Omega^2 \|f\|_h$$

If we apply the Poincare inequality to the right-hand side, we get

$$\|\nabla_h U\|_h^2 \leq \|f\|_h \cdot \|U\|_h \leq \|f\|_h \cdot d_\Omega \|\nabla_h U\|_h$$

Thus, we obtain

$$\|\nabla_h U\| \leq d_\Omega \|f\|_h$$

When we apply this result to $Ae = \tau$, we get

$$\begin{aligned} \|e\| &\leq d_\Omega^2 \|\tau\| = O(h^2) \\ \|\nabla_h e\| &\leq d_\Omega \|\tau\| = O(h^2). \end{aligned}$$

□

Remark The discrete Laplacian has many good properties as those of continuous Laplacian. For continuous Laplacian, we can have $\|u\|_H^{s+2}$ estimated by some $\|f\|_{H^s}$. In the case of discrete Laplacian, we have similar result. As the truncated error is of $\|\tau\|_{H^s} = O(h^2)$ in terms of the discrete norm, then we have $\|e\|_{H^{s+2}} = O(h^2)$. Using Sobolev inequality, wecan get $|e|_\infty = O(h^2)$.

Chapter 4

Finite Difference Theory For Linear Hyperbolic Equations

4.1 A review of smooth theory of linear hyperbolic equations

Hyperbolic equations appear commonly in physical world. The propagation of acoustic wave, electric-magnetic waves, etc. obey hyperbolic equations. Physical characterization of hyperbolicity is that the signal propagates at finite speed. Mathematically, it means that compact-supported initial data yield compact-supported solutions for all time. This hyperbolicity property has been characterized in terms of coefficients of the corresponding linear partial differential equations through Fourier method.

They are two techaniques for hyperbolic equations, one is based on Fourier method (Garding et al.), the other is energy method (Friedrichs' symmetric hyperbolic equations). A good reference is F. John's book. For computational purpose, we shall only study one dimensional cases. For analysis, the techniques include methods of characteristics, energy methods, Fourier methods.

4.1.1 Linear advection equation

We start from the Cauchy problem of the linear advection in one-space dimension

$$\begin{aligned} u_t + a u_x &= 0, && (4.1)\\ u(x,0) &= u_0(x). && (4.2) \end{aligned}$$

Its solution is simply a translation of u_0, namely,

$$u(x,t) = u_0(x - at).$$

More generally, we can solve the linear advection equation with variable coefficients by the *method of characteristics*. Consider

$$u_t + a(x,t) u_x = 0.$$

This equation merely says that the direction derivative of u is 0 in the direction $(1, a) \parallel (dt, dx)$. If $x(t, \xi)$ is the solution of the ODE

$$\frac{dx}{dt} = a(x, t).$$

with initial data $x(0, \xi) = \xi$, then

$$\begin{aligned} \frac{d}{dt}|_\xi u(x(t,\xi),t) &= \partial_t u + \partial_x u \frac{dx}{dt} \\ &= u_t + a u_x = 0 \end{aligned}$$

In other words, u is unchanged along the curve: $dx/dt = a$. Such a curve is called the characteristic curve. Suppose from any point $(x, t), t > 0$, we can find the characteristic curve $\xi(s, t, x)$ backward in time and $\xi(\cdot, t, x)$ can be extended to $s = 0$. Namely, $\xi(\cdot, t, x)$ solves the ODE: $d\xi/ds = a(\xi, s)$ with $\xi(t, t, x) = x$, and $\xi(\cdot, t, x)$ exists on $[0, t]$. The solution to the Cauchy problem is then given by $u(x, t) = u_0(\xi(0, t, x))$.

Note that the characteristics are the curves where signals propagate along.

Now, we consider the linear advection equation with source term:

$$u_t + a(x, t)u_x = f(x, t).$$

Let $x(t, \xi)$ be its characteristic curves. Along the characteristic curve, the equation becomes

$$\frac{d}{dt}|_\xi u(x(t,\xi),t) = u_t + a u_x = f(x(t,\xi),t).$$

We integrate this equation in t with fixed ξ. We obtain

$$u(x(t,\xi),t) = u_0(\xi) + \int_0^t f(x(s,\xi),s)\, ds.$$

This is a function in (ξ, t). The final solution is obtain by replacing ξ by $\xi(0, t, x)$.

Homeworks

1. Find the solution of

   $$u_t - (\tanh x)u_x = 0$$

   with initial data u_0. Also show that $u(x, t) \to 0$ as $t \to \infty$, provided $u_0(x) \to 0$ as $|x| \to \infty$.

2. Show that the initial value problem for

   $$u_t + (1 + x^2)u_x = 0$$

   is not well defined. (Show the characteristics issued from x-axis do not cover the entire domain: $x \in R, t \geq 0$.)

4.1.2 Linear systems of hyperbolic equations

Methods of characteristics Second-order hyperbolic equations can be expressed as hyperbolic systems. For example, the wave equation

$$u_{tt} - c^2 u_{xx} = 0$$

can be written as

$$\begin{pmatrix} u_x \\ u_t \end{pmatrix}_t - \begin{pmatrix} 0 & 1 \\ c^2 & 0 \end{pmatrix} \begin{pmatrix} u_x \\ u_t \end{pmatrix}_x = 0.$$

To solve this system of equations, we diagonalize it. The eigenvalues and eigenvectors of the matrix is

$$\lambda_1 = -c, \quad \ell_1 = (-c, 1), \quad r_1 = (-c, 1)^T,$$
$$\lambda_1 = c, \quad \ell_2 = (c, 1), \quad r_2 = (c, 1)^T.$$

We multiply the system by ℓ_1 from the left and obtain

$$(u_t - cu_x)_t + c(u_t - cu_x)_x = 0.$$

By multiplying ℓ_2, we obtain

$$(u_t + cu_x)_t - c(u_t + cu_x)_x = 0.$$

Let $v_1 = u_t - cu_x$, and $v_2 = u_t + cu_x$. Then v_1 and v_2 satisfy linear advection equations, and u satisfies linear advection equation with source term. These can be solved by previous characteristic method for linear advection equation.

In general, systems of hyperbolic equations have the following form

$$u_t + A(x,t)u_x = B(x,t)u + f(x,t).$$

Here, u is an n-vector and A, B are $n \times n$ matrices. Such a system is called hyperbolic if A is diagonalizable with real eigenvalues. That is, A has real eigenvalues

$$\lambda_1 \leq \cdots \leq \lambda_n$$

with left/right eigenvectors l_i/r_i, respectively. We normalize these eigenvectors so that $l_i r_j = \delta_{i,j}$. Let $R = (r_1, \cdots, r_n)$ and $L = (l_1, \cdots, l_n)^t$. Then

$$\begin{aligned} A &= R\Lambda L, \\ \Lambda &= \text{ diag } (\lambda_1, \cdots, \lambda_n) \\ LR &= I. \end{aligned}$$

We can use L and R to diagonalize this system. First, we introduce $v = Lu$, then multiply the equation by L from the left:

$$Lu_t + LAu_x = LBu + Lf.$$

This gives

$$v_t + \Lambda v_x = Cv + g,$$

where $C = LBR + L_t R + \Lambda L_x R$ and $g = Lf$. The i-th equation:

$$v_{i,t} + \lambda_i v_{i,x} = \sum_j c_{i,j} v_j + g_i$$

is simply an ODE in the direction $dx/dt = \lambda_i(x,t)$. As before, from a point (x,t) with $t > 0$, we draw characteristic curves $\xi_i(\cdot, t, x)$, $i = 1, \cdots, n$:

$$\begin{aligned} \frac{d\xi_i}{ds} &= \lambda_i(\xi_i, s), i = 1, \cdots, n \\ \xi_i(t,t,x) &= x \end{aligned}$$

We integrate the i-th equation along the i-th characteristics to obtain

$$v_i(x,t) = v_{0,i}(\xi_i(0,t,x)) + \int_0^t (\sum_j c_{i,j} v_j + g_i)(\xi_i(s,t,x), s)\, ds.$$

An immediate conclusion we can draw here is that the domain of dependence of (x,t) is $[\xi_n(0,t,x), \xi_1(0,t,x)]$, which, we denote by $D(x,t)$, is finite. This means that if u_0 is zero on $D(x,t)$, then $u(x,t) = 0$.

One can obtain local existence theorem from this integral equation provided v_0 and $v_{0,x}$ are bounded. Its proof is mimic to that of the local existence of ODE. We define a function space $C_b(\mathbb{R})$, the bounded continuous functions on $\mathbb{R}$, using the sup norm: $\|u\|_\infty := \sup_x |u(x)|$. Define a map

$$Tv = v_{0,i}(\xi_i(0,t,x)) + \int_0^t \left(\sum_j c_{i,j} v_j + g_i \right).$$

Then the above integral equation is equivalent to find a fixed point of T

$$v = Tv$$

in the space $C_b(\mathbb{R})$. The operator T is a contraction in $C_b(\mathbb{R})$ if the time is short enough. The contraction map T yields a fixed point. This is the solution.

The global existence follows from a priori estimates (for example, C^1-estimates) using the above integral equations. A necessary condition for global existence is that all characteristics issued from any point (x,t), $x \in R$, $t > 0$ should be traced back to initial time. A sufficient condition is that $A(x,t)$ is bounded in the upper half plane in x-t space.

A nice reference for the method of characteristics for systems of hyperbolic equations in one-dimension is John's book, PDE, Sec. 5, Chapter 2.

Energy method for symmetric hyperbolic equations Many physical systems can be written in the symmetric hyperbolic equations:

$$A_0 u_t + A(x,t)u_x = B(x,t)u + f,$$

where A_0, A are $n \times n$ symmetric matrices and A_0 is positive definite. We take inner product of this equation with u, later we integrate in x over the whole space. For simplicity, we assume A_0 and A are constant matrices temporarily. We get

$$\frac{\partial}{\partial t}\frac{1}{2}A_0 u \cdot u + \frac{\partial}{\partial x}\frac{1}{2}Au \cdot u = Bu \cdot u + f \cdot u.$$

Here we have used the symmetric properties of A_0 and A:

$$\frac{\partial}{\partial x}Au \cdot u = Au_x \cdot u + Au \cdot u_x = 2Au_x \cdot u.$$

As we integrate in x over the whole space, we get

$$\frac{d}{dt}\frac{1}{2}\langle A_0 u, u\rangle = \langle Bu, u\rangle + \langle f, u.\rangle$$

Here, the term

$$\langle Au_x, u\rangle = \langle u_x, Au\rangle = \int \frac{1}{2}(Au, u)_x \, dx = 0.$$

We have used symmetry property of A. The positivity of A_0 yields that $(A_0 u, u)$ is equivalent to $\|u\|_2^2$, namely, there are two constants C_1 and C_2 such that for any $u \in L^2(\mathbb{R})$,

$$C_1 \int |u|^2 \, dx \le \langle A_0 u, u\rangle \le C_2 \int |u|^2 \, dx.$$

If we use $\langle A_0 u, u\rangle$ as a new norm $|||u|||^2$, then we get

$$\frac{d}{dt}\frac{1}{2}|||u(t)|||^2 \le C|||u|||^2 + C'|||u||| \cdot \|f\|$$

Here, we have used the boundedness of B. Eliminating $\|u\|$, we get

$$\frac{d}{dt}|||u(t)||| \le C|||u||| + C'\|f\|$$

This yields (by Gronwell inequality)

$$|||u(t)||| \le e^{Ct}|||u(0)||| + C' \int_0^t e^{C(t-s)}\|f(s)\| \, ds$$

Thus, $|||u(t)|||$ s bounded for any finite time if $\|u(0)\|$ is bounded.

We can apply this method to the equations for derivatives of u by differentiating the equations. This will give us the boundedness of all derivatives, from which we get compactness of approximate solution and existence theorem. For general "smooth" theory for symmetric hyperbolic systems in high-dimension we refer to Chapter 6 of John's book.

4.2 Finite difference methods for linear advection equation

4.2.1 Design techniques

We shall explain some design principles for the linear advection equation:

$$u_t + au_x = 0.$$

We shall assume $a > 0$ a constant. Despite of its simplicity, the linear advection equation is a prototype equation to design numerical methods for nonlinear hyperbolic equations in multi-dimension.

First, we choose $h = \Delta x$ and $k = \Delta t$ to be the spatial and temporal mesh sizes, respectively. We discretize the $x - t$ space by the grid points (x_j, t_n), where $x_j = j\Delta x$ and $t_n = n\Delta t$. We shall use the data U_j^n to approximate $u(x_j, t_n)$. To derive finite difference schemes, we use finite differences to approximate derivatives. We demonstrate spatial discretization first, then the temporal discretization.

1. Spatial discretization. There are two important design principles here, the interpolation and upwinding.

1. Derivatives are replaced by finite differences. For instance, u_{xj} can be replaced by

$$\frac{U_j - U_{j-1}}{h}, \text{ or } \frac{U_{j+1} - U_{j-1}}{2h}, \text{ or } \frac{3U_j - 4U_{j-1} + U_{j-2}}{2h}.$$

   The first one is first-order, one-side finite differencing, the second one is the central differencing which is second order, the third one is a one-side, second-order finite differencing. This formulae can be obtained by make Taylor expansion of u_{j+k} about x_j.

2. Upwinding. We assume $a > 0$, this implies that the information comes from left. Therefore, it is reasonable to approximate u_x by "left-side finite difference":

$$\frac{U_j - U_{j-1}}{h} \text{ or } \frac{3U_j - 4U_{j-1} + U_{j-2}}{2h}.$$

2. Temporal discretization.

1. Forward Euler: We replace $u_t{}_j^n$ by $(U_j^{n+1} - U_j^n)/k$. As conbining with the upwinding spatial finite differencing, we obtain the above upwinding scheme.

2. backward Euler: We replace $u_t{}_j^{n+1}$ by $(U_j^{n+1} - U_j^n)/k$, but replace u_x by $D_x)_j^{n+1}$, where D is spatial finite difference above.

3. Leap frog: We replace $u_t{}_j^n$ by $(U_j^{n+1} - U_j^{n-1})/2k$.

4. An important trick is to replace high-order temporal derivatives by high-order spatial derivatives through the help of P.D.E.: for instance, in order to achieve high order approximation of u_t, we can expand

$$\frac{u_j^{n+1} - u_j^n}{k} = u_{t,j}^n + \frac{k}{2}u_{tt,j}^n + \cdots,$$

We can replace u_{tt} by

$$u_{tt} = -au_{xt} = a^2 u_{xx},$$

then approximate u_{xx} by central finite difference. Thus, we obtain a second order approximation of u_t:

$$u_t \leftarrow \frac{U_j^{n+1} - U_j^n}{k} - \frac{k}{2h^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

For u_x, we can use central difference approximation

$$u_x \leftarrow \frac{U_{j+1}^n - U_{j-1}}{2h}$$

The resulting scheme is

$$U_j^{n+1} - U_j^n = -\frac{ak}{2h}\left(U_{j+1}^n - U_{j-1}\right) + \frac{a^2k^2}{2h^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

This is a second order scheme in both space and time. The scheme is called Lax-Wendroff scheme.

We list some finite difference schemes below. Let $\sigma = ak/h$.

$$\begin{aligned}
\text{Upwind} \quad &: \quad U_j^{n+1} = U_j^n - \sigma(U_j^n - U_{j-1}^n) \\
\text{Lax-Friedrichs} \quad &: \quad U_j^{n+1} = \frac{U_{j+1}^n + U_{j-1}^n}{2} + \frac{\sigma}{2}(U_{j+1}^n - U_{j-1}^n) \\
\text{Backward Euler} \quad &: \quad U_j^{n+1} - U_j^n = \frac{\sigma}{2}(U_{j-1}^{n+1} - U_{j+1}^{n+1}) \\
\text{Lax-Wendroff} \quad &: \quad U_j^{n+1} = U_j^n - \frac{\sigma}{2}(U_{j+1}^n - U_{j-1}^n) + \frac{\sigma^2}{2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n) \\
\text{Beam-Warming} \quad &: \quad U_j^{n+1} = U_j^n - \frac{\sigma}{2}(3U_j^n - 4U_{j-1}^n + U_{j-2}^n) + \frac{\sigma^2}{2}(U_j^n - 2U_{j-1}^n + U_{j-2}^n)
\end{aligned}$$

In Beam-Warming, the term u_x is approximated by second order finite difference with upwinding:

$$u_x \leftarrow \frac{1}{2h}(3U_j^n - 4U_{j-1}^n + U_{j-2}^n)$$

Here the upwinding means that $a > 0$, the information comes from left, and we use U_{j-2}, U_{j-1} and U_j as our stencil. The term

$$u_t \leftarrow \frac{U_j^{n+1} - U_j^n}{k} - \frac{k}{2h^2}(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

as that in the Lax-Wendroff scheme. Thus, Beam-Warming is a second order scheme.

In general, an (explicit) finite difference scheme for the linear advection equation can be expressed as

$$U_j^{n+1} = G(U_{j-l}^n, U_{j-l+1}^n, \cdots, U_{j+m}^n) = \sum_{k=-l}^{m} a_k U_{j+k}^n$$

Remark.

1. From characteristics method, $u(x_j, t_{n+1}) = u(x_j - ak, t_n)$. We can approximate it by interpolation at neighboring grid points. For instance, a linear interpolation at x_{j-1} and x_j gives

$$u_j^{n+1} \approx \frac{ak}{h} u_{j-1}^n + (1 - \frac{ak}{h}) u_j^n.$$

   The corresponding finite difference scheme is then defined by

$$U_j^{n+1} = \frac{ak}{h} U_{j-1}^n + (1 - \frac{ak}{h}) U_j^n.$$

   This is the well-known upwind scheme. Where the spatial discretization is exactly the above one-side, first-order finite differencing.

2. The term $(u_j^{n+1} - u_j^n)/k$ in a forward Euler method introduces an anti-diffusion term $-a^2 u_{xx}$, namely,

$$\frac{u_j^{n+1} - u_j^n}{k} = u_t + \frac{k}{2} u_{tt} + O(k^2) = u_t + \frac{a^2 k}{2} u_{xx} + O(k^2).$$

   Thus, a high-order upwind differencing $\frac{\sigma}{2}(3U_j^n - 4U_{j-1}^n + U_{j-2}^n)$ for au_x and first-order difference in time will be unstable. We will see this in the modified equation later.

Homeworks

1. Use the trick $u_{tt} = a^2 u_{xx}$ and central finite difference to derive Lax-Wendroff scheme by yourself.

2. Derive a finite difference using method of characteristics and a quadratic interpolation at x_{j-2}, x_{j-1} and x_j. Is this scheme identical to the Beam-Warming scheme?

3. Do the same thing with cubic interpolation at $x_{j-2}, \cdots, x_{j+1}$?

4. Write a computer program using the above listed schemes to the linear advection equation. Use periodic boundary condition. The initial condition are

   (a) square wave,

   (b) hat function

   (c) Gaussian

   (d) $e^{-x^2/D} \sin mx$

   Refine the mesh by a factor of 2 to check the convergence rates.

4.2.2 Courant-Friedrichs-Levy condition

For a finite difference scheme:

$$U_j^{n+1} = G(U_{j-\ell}^n, \cdots, U_{j+m}^n),$$

We can define numerical domain of dependence of (x_j, t_n) to be $[x_{j-n\ell}, x_{j+nm}]$ (denoted by $D(j,n)$). For instance, the numerical domain of upwind method is $[x_{j-n}, x_j]$. If $U_k^0 = 0$ on $D(j,n)$, then $U_j^n = 0$. In order to have our finite difference schemes physically meaningful, a natural condition is

$$\text{physical domain of dependence } \subset \text{ numerical domain of dependence.}$$

This gives a constraint on the ratio of h and k. Such a condition is called the Courant-Friedrichs-Levy (C-F-L) condition. For the linear advection equation with $a > 0$, the condition is

$$\{x_j - ak\} \subset [x_{j-\ell}, x_{j+m}].$$

This leads to

$$0 \leq \frac{ak}{\ell h} \leq 1$$

If this condition is violated, we can easily construct an initial condition which is zero on numerical domain of dependence of (x,t), yet $u(x,t) \neq 0$. The finite difference scheme will produce 0 at (x,t). Thus, its limit is also 0. But the true solution $u(x,t)$ is not zero.

Below, we shall fix the ratio h/k during the analysis and take $h \to 0$ in the approximation procedure.

4.2.3 Consistency and Truncation Errors

Let us express our difference scheme in the form:

$$U^{n+1} = GU^n.$$

Given a smooth solution $u(x,t)$ to the PDE. Let us denote $u(jh, nk)$ by u_j^n. Plug u^n into this finite difference equation, then make Taylor expansion about (jh, nk). For instance, we plug a smooth function u into a upwind scheme:

$$\frac{1}{k}(u_j^{n+1} - u_j^n) + \frac{1}{h}(u_j^n - u_{j-1}^n) = (u_t + au_x) + k(u_{tt} - \sigma u_{xx}) + O(h^2 + k^2)$$

Thus, we define the truncation error as

$$\tau^n(h,k) = \frac{u^{n+1} - Gu^n}{k}.$$

A finite difference scheme is called consistent if $\tau(h,k) \to 0$ as $h, k \to 0$. Naturally, this is a minimal requirement of a finite difference scheme. If the scheme is expressed as

$$U_j^{n+1} = \sum_{k=-l}^{m} a_k U_{j+k}^n,$$

then a necessary and sufficient condition for consistency is

$$\sum_{k=-l}^{m} a_k = 1.$$

This is easy to see because the constant is a solution.

If $\tau = O(k^r)$, then the scheme is called of order r. We can easily check that $\tau = O(k)$ for the upwind method by Taylor expansion:

$$\begin{aligned}
\tau &= \frac{1}{k}\left(u_j^{n+1} - u_j^n + \sigma(u_j^n - u_{j-1}^n)\right) \\
&= \frac{1}{k}\left(u_t k + \frac{1}{2}u_{tt}k^2 + \frac{ak}{h}(-u_x h + \frac{1}{2}u_{xx}h^2\right) + HOT \\
&= (u_t + au_x) + \frac{k}{2}\left(u_{tt} + \frac{ah}{k}u_{xx}\right) + HOT \\
&= (u_t + au_x) - \frac{h^2}{2k}\sigma(1-\sigma)u_{xx} + HOT
\end{aligned}$$

The term $\frac{h^2}{2k}\sigma(1-\sigma)u_{xx}$ is $O(h)$ if we keep $\sigma = ak/h$ fixed. Thus, the upwind scheme is first order.

Homework Find the truncation error of the schemes listed above.

4.2.4 Lax's equivalence theorem

Suppose U^n is generated from a finite difference scheme: $U^{n+1} = G(U^n)$, we wish the solution remain bounded under certain norm as we let the mesh size $\Delta t \to 0$. This is equivalent to let the time step number $n \to \infty$. A scheme is called stable if $\|U^n\|$ remains bounded under certain norm $\|\cdot\|$ for all n.

Let u be an exact solution of some linear hyperbolic P.D.E. and U be the solution of a corresponding finite difference equation, We want to estimate the true error $e_j^n = u_j^n - U_j^n$.

First we estimate how much errors accumulate in one time step.

$$e^{n+1} := u^{n+1} - U^{n+1} = ke^n + Gu^n - GU^n = ke^n + Ge^n.$$

If we can have an estimate (called stability condition) like

$$\|GU\| \le \|U\| \tag{4.3}$$

under certain norm $\|\cdot\|$, then we obtain

$$\|u^n - U^n\| \le \|u^0 - U^0\| + k(\tau^{n-1} + \cdots + \tau^1).$$

From the consistency, we obtain $\|e^n\| \to 0$ as $k \to 0$. If the scheme is of order r, then we obtain

$$\|e^n\| \le \|u^0 - U^0\| + O(k^r).$$

We have the following theorems.

Theorem 4.1 (Lax equivalence theorem)**.** *Given a linear hyperbolic partial differential equation. Then a consistent finite difference scheme is stable if and only if is is convergent.*

We have proved stability $\Rightarrow$ convergence. We shall prove the other part in the next section.

4.2.5 Stability analysis

Since we only deal with smooth solutions in this section, the L^2-norm or the Sobolev norm is a proper norm to our stability analysis. For constant coefficient and scalar case, the von Neumann analysis (via Fourier method) provides a necessary and sufficient condition for stability. For system with constant coefficients, the von Neumann analysis gives a necessary condition for statbility. For systems with variable coefficients, the Kreiss' matrix theorem provides characterizations of stability condition. We describe the von Neumann analysis below.

Given $\{U_j\}_{j\in\mathbb{Z}}$, we define

$$\|U\|^2 = \sum_j |U_j|^2$$

and its Fourier transform

$$\hat{U}(\xi) = \frac{1}{2\pi}\sum U_j e^{-ij\xi}.$$

The advantages of Fourier method for analyzing finite difference scheme are

- the shift operator is transformed to a multiplier:

$$\widehat{TU}(\xi) = e^{i\xi}\hat{U}(\xi),$$

  where $(TU)_j := U_{j+1}$;

- the Parseval equility

$$\begin{aligned}\|U\|^2 &= \|\hat{U}\|^2 \\ &\equiv \int_{-\pi}^{\pi} |\hat{U}(\xi)|^2\, d\xi.\end{aligned}$$

If a finite difference scheme is expressed as

$$U_j^{n+1} = (GU^n)_j = \sum_{i=-l}^{m} a_i (T^i U^n)_j,$$

then

$$\widehat{U^{n+1}} = \widehat{G}(\xi)\widehat{U^n}(\xi).$$

From the Parseval equality,

$$\|U^{n+1}\|^2 = \|\widehat{U^{n+1}}\|^2$$

$$
\begin{aligned}
&= \int_{-\pi}^{\pi} |\widehat{G}(\xi)|^2 \, |\widehat{U^n}(\xi)|^2 \, d\xi \\
&\leq \max_{\xi} |\widehat{G}(\xi)|^2 \int_{-\pi}^{\pi} \widehat{U^n}(\xi)|^2 \, d\xi \\
&= |\widehat{G}|_\infty^2 \|U\|^2
\end{aligned}
$$

Thus a necessary condition for stability is

$$
|\widehat{G}|_\infty \leq 1. \tag{4.4}
$$

Conversely, Suppose $|\widehat{G}(\xi_0)| > 1$, from$\widehat{G}$ being a smooth function in ξ, we can find ϵ and δ such that

$$
|\widehat{G}(\xi)| \geq 1 + \epsilon \text{ for all } |\xi - \xi_0| < \delta.
$$

Let us choose an initial data U_0 in ℓ^2 such that $\widehat{U^0}(\xi) = 1$ for $|\xi - \xi_0| \leq \delta$. Then

$$
\begin{aligned}
\|\widehat{U^n}\|^2 &= \int |\widehat{G}|^{2n}(\xi) |\widehat{U^0}|^2 \\
&\geq \int_{|\xi - \xi_0| \leq \delta} |\widehat{G}|^{2n}(\xi) |\widehat{U^0}|^2 \\
&\geq (1+\epsilon)^{2n} \delta \to \infty \text{ as } n \to \infty
\end{aligned}
$$

The operator G^n is unbounded in $\|\cdot\|_2$ operator norm. It is a fact that it will not be bounded by any equivalent norm, which involves more analysis and will be omit here. Thus, the scheme can not be stable. We conclude the above discussion by the following theorem.

Theorem 4.2. *A finite difference scheme*

$$
U_j^{n+1} = \sum_{k=-l}^{m} a_k U_{j+k}^n
$$

with constant coefficients is stable if and only if

$$
\widehat{G}(\xi) := \sum_{k=-l}^{m} a_k e^{-ik\xi}
$$

satisfies

$$
\max_{-\pi \leq \xi \leq \pi} |\widehat{G}(\xi)| \leq 1. \tag{4.5}
$$

As a simple example, we show that the scheme:

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

is unstable. The operator $G = 1 + \frac{\sigma}{2}(T - T^{-1})$. The corresponding $\widehat{G}(\xi) = 1 + i\sigma \sin \xi$, which cannot be bounded by 1 in magnitude. One the other hand, the Lax-Friedrichs scheme replaces U_j^n

in the above scheme by the average $(U^n_{j-1} + U^n_{j+1})/2$. The corresponding $\widehat{G}(\xi) = \cos\xi + i\sigma \sin\xi$, which is bounded by 1 in magnitude provided $|\sigma| \leq 1$. The above replacement is equivalent to add a term $(U^n_{j-1} - 2U^n_j + U^n_{j+1})/2$ to the right hand side of the above unstable finite difference. It then stabilizes the scheme. This quantity is called a numerical viscosity. We see the discussion in the next section.

Homeworks.

1. Compute the $\widehat{G}$ for the schemes: Lax-Friedrichs, Lax-Wendroff, Leap-Frog, Beam-Warming, and Backward Euler.

4.2.6 Modified equation

We shall study the performance of a finite difference scheme to a linear hyperbolic equation. Consider the upwind scheme for the linear advection equation. Let $u(x,t)$ be a smooth function. Expand u in Taylor series, we obtain

$$u^{n+1}_j - G(u^n)_j = (u_t + au_x)\Delta t - \frac{(\Delta x)^2}{2}(\sigma - \sigma^2)u_{xx} + O((\Delta t)^3).$$

The truncation error for the upwind method is $O(\Delta t)$ if u satisfies the linear advection scheme. However, if we fix Δx and Δt, then the error is $O(\Delta t^3)$ if u satisfies

$$u_t + au_x - \nu u_{xx} = 0,$$

where

$$\nu = \frac{(\Delta x)^2}{2\Delta t}(\sigma - \sigma^2).$$

This equation is called modified equation. *The solution of the finite difference equation is closer to the solution of this modified equation than the original equation.* The role of νu_{xx} is a dissipation term in the scheme. The term $-(\Delta x)^2/(\Delta t)\sigma^2 u_{xx}$ comes from the forward Euler approximation to u_t. It is an anti-diffusion. The term $(\Delta x)^2/(\Delta t)\sigma u_{xx}$ comes from the upwind discretization for au_x. It is a diffusion. The effect diffusion is νu_{xx}. The constant ν is called numerical viscosity. We observe that $\nu \geq 0$ if and only if $0 \leq \sigma \leq 1$, which is exactly the (C-F-L as well as von Neumann) stability condition. This is consistent to the well-postedness of diffusion equations (i.e. $\nu \geq 0$).

The effect of numerical viscosity is that it will make solution smoother, and will smear out discontinuities. To see this, let us solve the Cauchy problem:

$$\begin{aligned} u_t + au_x &= \nu u_{xx} \\ u(x,0) &= H(x) := \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \end{aligned}$$

The function H is called the Heaviside function. The corresponding solution is given by

$$u(x,t) = \frac{1}{\sqrt{4\pi\nu t}} \int_{-\infty}^{\infty} e^{-\frac{(x-at-y)^2}{4\nu t}} u(y,0)\, dy$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{4\pi\nu t}} \int_0^\infty e^{-\frac{(x-at-y)^2}{4\nu t}} dy \\
&= \text{erf}((x-at)/\sqrt{4\nu t}),
\end{aligned}
$$

where

$$
\text{erf}(x) := \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-z^2}\, dz.
$$

Let $u_e(x,t)$ be the exact solution of $u_t + au_x = 0$ with $u(x,0) = H(x)$. Then

$$
|u_e(y+at,t) - u(y+at,t)| = \text{erf}(-|y|/\sqrt{4\nu t}).
$$

Hence,

$$
\begin{aligned}
\|u_e(\cdot,t) - u(\cdot,t)\|_{L^1} &= 2\int_{-\infty}^{0} \text{erf}(\frac{y}{\sqrt{4\nu t}})\, dy \\
&= C\sqrt{\nu t}
\end{aligned}
$$

Since $\nu = O(\Delta t)$, we see that

$$
\|u_e^n - u^n\| = O(\sqrt{\Delta t}).
$$

On the other hand, if U is the solution of the finite difference equation, then we expect that $\|U^n - u^n\|_{L^1} = O(\Delta t)$, because it is first order. Indeed, it is only $O(\sqrt{\Delta t})$ and

$$
\|U^n - u_e^n\|_{L^1} = O(\sqrt{\Delta t}).
$$

Thus, a first order scheme is only of half order for "linear discontinuities."

One can also observe the smearing (averaging) of discontinuities from the finite difference directly. In upwind scheme, U_j^{n+1} may be viewed as weighted averages of U_j^n and U_{j-1}^n:

$$
U_j^{n+1} = (1-\sigma)U_j^n + \sigma U_{j-1}^n.
$$

If $U_{j-1}^n = 0$ and $U_j^n = 1$, then U_j^{n+1} is a value between 0 and 1. This is a smearing process (averaging process). The smearing process will spread out. Its width is $(\sqrt{n}\Delta x) = O(\sqrt{\Delta t})$ from the estimate of binomial distribution.

It should be noticed that the magnititute of the numerical viscosity of the upwind method is smaller than that of the Lax-Friedrichs method. The upwind method uses the information of charateristic speed whereas the Lax-Friedrichs does not use this information.

Homeworks.

1. Find the modified equations for the following schemes:

$$
\begin{aligned}
\text{Lax-Friedrichs} &: \quad u_t + au_x = \frac{(\Delta x)^2}{2\Delta t}(1-\sigma^2)u_{xx} \\
\text{Lax-Wendroff} &: \quad u_t + au_x = \frac{(\Delta x)^2}{6} a(\sigma^2-1)u_{xxx} \\
\text{Beam-Warming} &: \quad u_t + au_x = \frac{(\Delta x)^2}{6} a(2-3\sigma+\sigma^2)u_{xxx}
\end{aligned}
$$

2. Expand u up to u_{xxxx}, find the modified equation with the term u_{xxxx} for the Lax-Wendroff scheme and Beam-Warming. That is

$$u_t + au_x = \mu u_{xxx} + \kappa u_{xxxx}.$$

Show that the coefficient $\kappa < 0$ for both scheme if and only if the C-F-L stability condition.

3. Find the solution U_j^n of the upwind scheme with initial data $U_j^0 = \delta_{j0}$. (Hint: a binomial distribution.) Now, condider the Heaviside function as our initial data. Using the above solution formula, superposition principle and the Stirling formula, show that $\sum_j |u_j^n - U_j^n|\Delta x = O(\sqrt{n}\Delta x) = O(\sqrt{\Delta t})$.

Next, we study second-order scheme for solutions with discontinuities. We use Fourier method to study the solution of the modified equation:

$$u_t + au_x = \mu u_{xxx}.$$

By taking Fourier transform in x:

$$\hat{u}(\xi, t) := \int u(x,t)e^{-ix\xi}\, dx,$$

we find

$$\hat{u}_t = (-ia\xi - i\mu\xi^3)\hat{u} = -i\omega(\xi)\hat{u}$$

Hence

$$u(x,t) = \int e^{i(x\xi - \omega(\xi)t)}\hat{u}(\xi, 0)\, d\xi.$$

The initial data we consider here is the Heaviside function $H(x)$. However, in the discrete domain, its Fourier expansion is truncated. The corresponding inversion has oscillation on both side of the discontinuity, called Gibb's phenomena. The width is $O(\Delta x)$, the height is $O(1)$. We propagate such an initial data by the equation $u_t + au_x = \mu u_{xxx}$. The superposition of waves in different wave number ξ cause interference of waves. Eventually, it forms a wave package: a high frequency wave modulated by a low frequency wave. By the method of stationary phase, we see that the major contribution of the integral is on the set when

$$\frac{d}{d\xi}(x\xi - \omega(\xi)t) = 0.$$

The correspond wave $e^{i(x-\omega'(\xi)t)}$ is the modulated wave. Its speed $\omega'(\xi)$ is called the group velocity v_p. For the Lax-Wendroff scheme, we see that the group speed is

$$v_p = a + 3\mu\xi^2.$$

For the Beam-Warming, $v_p = a + 3\mu\xi^2$. Since $\mu < 0$ for the Lax-Wendroff, while $\mu > 0$ for the Beam-Warming, we observe that the wave package leaves behind (ahead) the discontinuity in the Lax-Wendroff (Beam-Warming).

One can also observe this oscillation phenomena directly from the scheme. In Beam-Warming, we know that U_j^{n+1} is a quadratic interpolation of U_{j-2}^n, U_{j-1}^n and U_j^n. If $U_{j-2}^n = 0$, and $U_{j-1}^n = U_j^n = 1$, then the quadratic interpolation gives an overshoot at U_j^{n+1} (that is, $U_j^{n+1} > 1$). Similarly, in the Lax-Wendroff scheme, U_j^{n+1} is a quadratic interpolation of U_{j-1}^n, U_j^n and U_{j+1}^n. If $U_{j-1}^n = U_j^n = 0$, and $U_{j+1}^n = 1$, then $U_j^{n+1} < 0$ (an undershoot).

Homeworks.

1. Measure the width of the oscillation as a function of number of time steps n.

4.3 Finite difference schemes for linear hyperbolic system with constant coefficients

4.3.1 Some design techniques

We consider the system

$$u_t + Au_x = 0$$

with A being a constant $n \times n$ matrix. The first designing principle is to diagonal the system. Using the left/right eigenvectors, we decompose

$$\begin{aligned} A &= R\Lambda L \\ &= R(\Lambda^+ - \Lambda^-)L \\ &= A^+ - A^- \end{aligned}$$

Here, $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ and $\Lambda^\pm$ are the positive/negative parts of Λ.

With this decomposition, we can define the upwind scheme:

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x} A^+ (U_{j-1}^n - U_j^n) - \frac{\Delta t}{\Delta x} A^- (U_{j+1}^n - U_j^n).$$

The Lax-Friedrichs is still

$$\begin{aligned} U_j^{n+1} &= \frac{U_{j-1}^n + U_{j+1}^n}{2} + \frac{\Delta t}{2\Delta x} A(U_{j-1}^n - U_{j+1}^n) \\ &= U_j^n + \frac{\Delta t}{2\Delta x} A(U_{j-1}^n - U_{j+1}^n) + \frac{U_{j-1}^n - 2U_j^n + U_{j+1}^n}{2} \end{aligned}$$

We see the last term is a dissipation term. In general, we can design modified L-F scheme as

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x} A(U_{j-1}^n - U_{j+1}^n) + D\frac{U_{j-1}^n - 2U_j^n + U_{j+1}^n}{2}$$

where D is a positive constant. D is chosen so that the scheme is stable by the von-Neumann analysis.

The Lax-Wendroff scheme is given by

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{2\Delta x} A(U_{j-1}^n - U_{j+1}^n) + \frac{(\Delta t)^2}{2(\Delta x)^2} A^2(U_{j+1}^n - 2U_j^n + U_{j-1}^n).$$

The C-F-L condition for upwind, L-F, L-W are

$$\max_i |\lambda_i| \frac{\Delta t}{\Delta x} \leq 1.$$

Homeworks.

1. Find the modified equation for the above schemes.
2. What is the stability condition on D for the modified L-F scheme.
3. Write a compute program to compute the solution of the wave equation:

$$\begin{aligned} u_t &= v_x \\ v_t &= c^2 u_x \end{aligned}$$

   using upwind, modified L-F, L-W schemes. The initial data is chosen as those for the linear advection equation. Use the periodic boundary condition.

4.3.2 Stability analysis

The definition of L^2-stability is that the L^2-norm of the solution of finite difference scheme

$$\sum_j |U_j^n|^2 \Delta x$$

is uniformly bounded.

This L^2-theory for smooth solutions was well developed in the 60s. First, Lax's equivalence theorem was originally proved for well-posed linear systems even in multi-dimension. Thus, the essential issue for finite difference scheme is still the stability problem.

Let us suppose the system is expressed as

$$u_t = \sum_i A_i u_{x_i} + Bu + f$$

Here, A_i, B are constant matrices. We assume that the system is hyperbolic. This means that $\sum_i \xi A_i$ is diagonal with real eigenvalues. Suppose the corresponding finite difference scheme is expressed as

$$U^{n+1} = GU^n = \sum a_\alpha T^\alpha U^n.$$

Here, $\alpha = (\alpha_1, \cdots, \alpha_n)$ is multi-index, a_α are matrices. Consider the Fourier transform of G:

$$\widehat{G}(k) = \sum_\alpha a_\alpha e^{i \sum_m \alpha_m k_m \Delta x_m}$$

If we take Δx_m as a function of Δt, then $\widehat{G}$ is a function of $(k, \Delta t)$. Using $\widehat{G}$, we have

$$\widehat{U^n} = \widehat{G}^n \widehat{U^0}.$$

From the Parseval equality: $\|U\|^2 = \int |\hat{U}|^2$, we obtain that the stability of a scheme $U^{n+1} = GU^n$ is equivalent to $\|\widehat{G}^n\|$ is uniformly bounded. Von Neumann gave a necessary condition for stability for system case.

Theorem 4.3. *A necessary condition for stability is that all eigenvalues of $\widehat{G}(k, \Delta t)$ satisfies*

$$|\lambda_i(k, \Delta t)| \leq 1 + O(\Delta t), \forall k, \forall \Delta t \leq \tau.$$

Proof. The spectral radius of $\widehat{G}(k, \Delta t)$ is the maximum value of the absolute values of the its eigenvalues. That is,

$$\rho(\widehat{G}) := \max_i |\lambda_i|$$

Since there is an eigenvector v such that $|\widehat{G}v| = \rho|v|$, we have that

$$\rho \leq \|\widehat{G}\| := \max_u \frac{|\widehat{G}u|}{|u|}.$$

Also, the eigenvalues of $\widehat{G}^n$ are λ_i^n. Hence we have

$$\rho(\widehat{G}^n) = \rho(\widehat{G})^n.$$

Combine the above two, we obtain

$$\rho(\widehat{G})^n \leq \|\widehat{G}^n\|.$$

Now, if $\|\widehat{G}^n\|$ is uniformly bounded, say by a constant C depends on $t := n\Delta t$, then

$$\begin{aligned} \rho &\leq C^{1/n} \\ &\leq 1 + O(\Delta t). \end{aligned}$$

□

For single equation, we have seen that von Neumann condition is also a sufficient condition for stability.

In general, Kreiss provided characterization of matrices which are stable.

Definition 4.1. A family of matrices $\{\mathrm{A}\}$ is stable if there exists a constant C such that for all $A \in \{A\}$ and all positive integer n,

$$\|A^n\| \leq C.$$

Theorem 4.4 (Kreiss matrix theorem)**.** *The stability of $\{A\}$ is equivalent to each of the following statements:*

(i) There exists a constant C such that for all $A \in \{A\}$ and $z \in C, |z| > 1$, $(A - zI)^{-1}$ exists and satisfies

$$\|(A - zI)^{-1}\| \leq \frac{C}{|z| - 1}.$$

(ii) There exist constants C_1 and C_2 such that for all $A \in \{A\}$, there exists nonsingular matrix S such that (1) $\|S\|, \|S^{-1}\| \leq C_1$, and (2) $B = SAS^{-1}$ is upper triangular and its off-diagonal elements satisfy

$$|B_{ij}| \leq C_2 \min\{1 - |\kappa_i|, 1 - |\kappa_j|\}$$

where κ_i are the diagonal elements of B.

(iii) There exists a constant $C > 0$ such that for all $A \in \{A\}$, there exists a positive definite matrix H such that

$$C^{-1}I \leq H \leq CI$$

$$A^* H A \leq H$$

Remarks.

1. In the first statement, the spectral radius of A is bounded by 1.
2. In the second statement, it is necessary that all $|\kappa_i| \leq 1$.
3. The meaning of the last statement means that we should use the norm $\sum |U_j|^2 = \sum_j (HU_j, U_j)$ instead of the Euclidean norm. Then A^n is nonincreasing under this norm.

4.4 Finite difference methods for linear systems with variable coefficients

Again, the essential issue is stability because Lax's equivalence theorem.

Kreiss showed by an example that the local stability (i.e. the stability for the frozen coefficients) is neither necessary nor sufficient for overall stability of linear variable systems. However, if the system $u_t = Au$ with A being first order, Strang showed that the overall stability does imply the local stability. So, for linear first-order systems with variable coefficients, the von Neumann condition is also a necessary condition for the overall stability.

For sufficient condition, we need some numerical dissipation to damp the high frequency component from spatial inhomogeneity. To illustrate this, let us consider the following scalar equation:

$$u_t + a(x)u_x = 0,$$

and a finite difference scheme

$$U^{n+1}(x) = A(x)U^n(x - \Delta x) + B(x)U^n(x) + C(x)U^n(x + \Delta x).$$

For consistency, we need to require

$$\begin{aligned} A(x) + B(x) + C(x) &= 1 \\ A(x) - C(x) &= a(x) \end{aligned}$$

Now, we impose another condition for local stability:

$$0 \le A(x), B(x), C(x) \le 1.$$

We show stability result. Multiply the difference equation by $U^{n+1}(x)$, use Cauchy-Schwartz inequality, we obtain

$$\begin{aligned} (U^{n+1}(x))^2 &= A(x)U^n(x-\Delta x)U^{n+1}(x) + B(x)U^n(x)U^{n+1}(x) + C(x)U^n(x+\Delta x)U^{n+1}(x) \\ &\le \frac{A(x)}{2}((U^n(x-\Delta x))^2 + (U^{n+1}(x))^2) + \frac{B(x)}{2}((U^n(x))^2 + (U^{n+1}(x))^2) \\ &\quad + \frac{C(x)}{2}((U^n(x+\Delta x))^2 + (U^{n+1}(x))^2) \\ &= \frac{A(x)}{2}(U^n(x-\Delta x))^2 + \frac{B(x)}{2}(U^n(x))^2 + \frac{C(x)}{2}(U^n(x+\Delta x))^2 + \frac{1}{2}(U^{n+1}(x))^2 \end{aligned}$$

This implies

$$\begin{aligned} (U^{n+1}(x))^2 &\le A(x)(U^n(x-\Delta x))^2 + B(x)(U^n(x))^2 + C(x)(U^n(x+\Delta x))^2 \\ &= A(x-\Delta x)(U^n(x-\Delta x))^2 + B(x)(U^n(x))^2 + C(x+\Delta x)(U^n(x+\Delta x))^2 \\ &\quad + (A(x) - A(x-\Delta x))(U^n(x-\Delta x))^2 + (C(x) - C(x+\Delta x))(U^n(x+\Delta x))^2 \end{aligned}$$

Now, we sum over $x = x_j$ for $j \in Z$. This yields

$$\|U^{n+1}\|^2 \le \|U^n\|^2 + O(\Delta t)\|U^n\|^2$$

Hence,

$$\|U^n\|^2 \le (1 + O(\Delta t))^n \|U^0\|^2 \le e^{Kt}\|U^0\|^2.$$

The above analysis show that monotone schemes are stable in L^2. Indeed, the scheme has some dissipation to damp the errors from the variation of coefficient (i.e. the term like $(A(x) - A(x - \Delta x))$).

For higher order scheme, we need to estimate higher order finite difference ΔU, this will involves $|\Delta a|\|\Delta U\|$, or their higher order finite differences. We need some dissipation to damp the growth of this high frequency modes. That is, the eigenvalues of the amplification matrix should satisfies

$$|\lambda_i| \le 1 - \delta|k\Delta x|^{2r}, \text{ when } |k\Delta x| \le \pi$$

for some $\delta > 0$.

To be more precisely, we consider first-order hyperbolic system in high-space dimension:

$$u_t + \sum_{i=1}^{d} a_i(x) u_{x_i} = 0,$$

where $u \in \mathbb{R}^N$, a_i, $i = 1, ..., d$, are $N \times N$ matrices. Consider a finite difference approximation:

$$U^{n+1}(x) = \sum_{\alpha} A_\alpha(x) T^\alpha U^n(x)$$

Here $\alpha = (\alpha_1, \cdots, \alpha_d)$ is a multi-index.

Let $\widehat{G}(x, \Delta t, \xi) = \sum_\alpha A_\alpha e^{i\alpha\cdot\xi}$ be the Fourier transform of the frozen finite difference operator.

Definition 4.2. A finite difference scheme with amplification matrix $\widehat{G}(x, \Delta t, \xi)$ is called dissipative of order $2r$ if there exists a constant $\delta > 0$ such that all eigenvalues of $\widehat{G}$ satisfy

$$|\lambda_i(x, \Delta t, \xi)| \leq 1 - \delta|\xi|^{2r}$$

for all $\max_i |\xi_i| \leq \pi$, all x, and all $\Delta t < \tau$ for some constant τ.

An important theorem due to Kreiss is the following stability theorem.

Theorem 4.5. *Suppose the system is symmetric hyperbolic, i.e. the matrices a_i are symmetric. Suppose the coefficient matrices A_α are also symmetric. Assume all coefficients are uniformly bounded. If the scheme is of order $2r - 1$ and is dissipative of order r, then the scheme is stable.*

Chapter 5

Scalar Conservation Laws

5.1 Physical models

Many partial differential equations are derived from physical conservation laws such as conservation of mass, momentum, energy, charges, etc. This class of PDEs is called conservation laws. The scalar conservation law is a single conservation law.

5.1.1 Traffic flow model

An interesting model is the following traffic flow model on a high way. We use macroscopic model, which means that $\Delta x \approx 100m$. Let ρ be the car density, u be the average car velocity. The car flux at a point x is the number of car passing through x per unit time. In a time period Δt, the car which can pass x must be in the region $u(x,t)\Delta t$. Thus, the flux at x is $(\rho(x,t)u(x,t)\Delta t)/(\Delta t) = \rho(x,t)u(x,t)$. Now, consider an arbitrary region (a,b), we have

$$\text{the change of number of cars in } (a,b) = \text{ the car flux at } a- \text{ the car flux at } b.$$

In mathematical formula:

$$\begin{aligned} \frac{d}{dt}\int_a^b \rho(x,t)\,dx &= \rho(a,t)u(a,t) - \rho(b,t)u(b,t) \\ &= -\int_a^b (\rho u)_x\,dx \end{aligned}$$

This holds for any (a,b). Hence, we have

$$\rho_t + (\rho u)_x = 0. \tag{5.1}$$

This equation is usually called the continuity equation in continuum mechanics. It is not closed because it involves two knowns ρ and u. Empirically, u can be teated as a function of ρ which satisfies $u \to 0$ as $\rho \to \rho_{\max}$. For instance,

$$u(\rho) = u_{\max}(1 - \frac{\rho}{\rho_{\max}}),$$

if there is a upper velocity limit, or

$$u(\rho) = a \log(\rho_{\max}/\rho)$$

if there is no restriction of velocity. We can model u to depend on ρ_x also. For instance,

$$u = u(\rho) - \nu \frac{\rho_x}{\rho}.$$

The quantity $\rho_x/\rho = -V_x/V$ is the negative expansion rate, where V is called the specific length, the length of a car (i.e., $V = 1/\rho$). If the expansion rate is positive, then the car train is rarefied. Thus, if the car number becomes denser (resp. rarefied), then the speed is reduced (resp. increased). Here, ν is the diffusion coefficient (viscosity) which is a positive number. Thus, the final equation is

$$\rho_t + f(\rho)_x = 0, \tag{5.2}$$

or

$$\rho_t + f(\rho)_x = \nu \rho_{xx}, \tag{5.3}$$

where $f(\rho) = \rho u(\rho)$.

5.1.2 Burgers' equation

The Burgers equation is

$$u_t + \frac{1}{2}(u^2)_x = \epsilon u_{xx}. \tag{5.4}$$

When $\epsilon = 0$, this equation is called inviscid Burgers equation. This equation is a prototype equation to study conservation laws.

Homeworks.

1. The Burgers equation can be linearized by the following nonlinear transform: let

   $$v = e^{-\frac{2}{\epsilon}\int^x u(\xi,t)\,d\xi},$$

   show that v satisfies the heat equation:

   $$v_t = \epsilon v_{xx}$$

2. Show that the Cauchy problem of the Burgers equation with initial data u_0 has an explicit solution:

   $$\begin{aligned} u(x,t) &= -\frac{\epsilon}{2}\frac{v_x}{v} \\ &= \int_{-\infty}^{\infty} \left(\frac{x-y}{t}\right) p_\epsilon(x,t,y)\,dy, \end{aligned}$$

where

$$\begin{aligned} p_\epsilon(x,t,y) &= \frac{e^{-\frac{1}{2\epsilon}I(x,t,y)}}{\int_{-\infty}^{\infty} e^{-\frac{1}{2\epsilon}I(x,t,y)}\,dy}, \\ I(x,t,y) &= \frac{(x-y)^2}{2t} + \int_0^y u_0(\xi)\,d\xi. \end{aligned}$$

The Burgers equation is a prototype equation for conservation laws. It reads

$$u_t + uu_x = \frac{\epsilon}{2}u_{xx} \tag{5.5}$$

A famous transformation called the Hopf-Cole transform can transform this equation into heat equation:

$$\phi(x,t) := \int_{-\infty}^{x} u(y,t)\,dy, \quad v = e^{-\frac{1}{\epsilon}\phi}.$$

Then ϕ satisfies the Hamilton-Jacobi equation

$$\phi_t + \frac{\phi_x^2}{2} = \frac{\epsilon}{2}\phi_{xx}$$

and v satisfies heat equation:

$$v_t = -\frac{1}{\epsilon}v, \quad v_x = -\frac{1}{\epsilon}\phi_x v,$$

$$v_{xx} = -\frac{1}{\epsilon}\phi_{xx}v + \left(\frac{1}{\epsilon}\phi_x\right)^2 v.$$

Thus,

$$v_t = \frac{\epsilon}{2}v_{xx} \quad \Leftrightarrow \quad \phi_t + \frac{\phi_x^2}{2} = \frac{\epsilon}{2}\phi_{xx}.$$

The solution to the heat equation can be expressed as

$$v(x,t) = \frac{1}{\sqrt{2\pi\epsilon t}}\int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{2\epsilon t}} v(y,0)\,dy$$

From

$$\phi = -\epsilon \ln v, \quad u = \phi_x = -\epsilon\frac{v_x}{v},$$

we can obtain an explicit form of the solution u as

$$\begin{aligned} u(x,t) &= -\epsilon\frac{1}{v(x,t)}\frac{1}{\sqrt{2\pi\epsilon t}}\int_{-\infty}^{\infty}\left(-\frac{x-y}{t\epsilon}\right)e^{-\frac{(x-y)^2}{2t\epsilon}}e^{-\frac{1}{\epsilon}\phi(y,0)}\,dy \\ &= \int_{-\infty}^{\infty}\left(\frac{x-y}{t}\right)p_\epsilon(x,y,t)\,dy \end{aligned}$$

where

$$p_\epsilon(x,y,t) = \frac{e^{-\frac{1}{\epsilon}I(x,y,t)}}{\int_{-\infty}^{\infty} e^{-\frac{1}{\epsilon}I(x,y,t)}\,dy}, \quad I(x,y,t) = \frac{(x-y)^2}{2t} + \phi(y,0).$$

Taking $\epsilon \to 0+$, we obtain

$$u(x,t) = \frac{x - y(x,t)}{t},$$

where

$$y(x,t) := \text{arg min}_x I(x,y,t).$$

5.1.3 Two phase flow

The Buckley-Leverett equation models how oil and water move in a reservoir. The unknown u is the saturation of water, $0 \le u \le 1$. The equation is

$$u_t + f(u)_x = 0$$

where

$$f(u) = \frac{u^2}{u^2 + a(1-u^2)^2}.$$

Unlike previous examples, the flux f here is a non-convex function.

5.2 Basic theory

Let consider scalar conservation law

$$u_t + f(u)_x = 0. \tag{5.6}$$

The equation can be viewed as a directional derivative $\partial_t + f'(u)\partial_x$ of u is zero. That implies u is constant along the characteristic curve

$$\frac{dx}{dt} = f'(u(x,t)).$$

This yields that the characteristic curve is indeed a straight line. Using this we can solve the Cauchy problem of (5.6) with initial data u_0 implicitly:

$$u = u_0(x - ut).$$

For instance, for inviscid Burgers' equation with $u_0(x) = x$, the solution u is given by $u = x - ut$, or $u = x/(1+t)$.

Homeworks.

1. If f is convex and u_0 is increasing, then the Cauchy problem for equation (5.6) has global solution.

2. If f is convex and $u_0' < 0$ at some point, then $u_x \to -\infty$ at finite time.

The solution may blow up (i.e. $|u_x| \to \infty$) in finite time due to the intersection of characteristic curves. A shock wave (discontinuity) is formed. We have to extend our solution class to to include these discontinuous solutions. We can view (5.6) in "weak sense." That is, for every smooth test function ϕ with compact support in $R \times [0, \infty)$,

$$\int_0^\infty \int_{-\infty}^\infty \phi[u_t + f(u)_x]\, dx\, dt = 0$$

Integrate by part, we obtain

$$\int_0^\infty \int_{-\infty}^\infty [\phi_t u + \phi_x f(u)]\, dx\, dt + \int_{-\infty}^\infty \phi(x,0)u(x,0)\, dx = 0, \tag{5.7}$$

In this formulation, it allows u to be discontinuous.

Definition 5.1. A function u is called a weak solution of (5.6) if it satisfies (5.7) for all smooth test function ϕ with compact support in $R \times [0, \infty)$.

Lemma 5.1. *Suppose u is a weak solution with discontinuity across a curve $x(t)$. Suppose u is smooth on the two sides of $x(t)$. Then u satisfies the following jump condition across $x(t)$:*

$$\frac{dx}{dt}[u] = [f(u)] \tag{5.8}$$

where $[u] := u(x(t)+, t) - u(x(t)-, t)$.

Homeworks. Work out this by yourself.

5.2.1 Riemann problem

The Riemann problem is a Cauchy problem of (5.6) with the following initial data

$$u(x,0) = \begin{cases} u_\ell & \text{for } x < 0 \\ u_r & \text{for } x > 0 \end{cases} \tag{5.9}$$

The reasons why Riemann problem is important are:

(i) Discontinuities are generic, therefore Riemann problem is generic locally.

(ii) In physical problems, the far field states are usually two constant states. Because of the hyperbolicity, at large time, we expect the solution is a perturbation of solution to the Riemann problem. Therefore, Riemann problem is also generic globally.

(iii) Both the equation (5.6) and the Riemann data (5.9) are invariant under the Galilean transform: $x \to \lambda x, t \to \lambda t$ for all $\lambda > 0$. If the uniqueness is true, the solution to the Riemann problem is self-similar. That is, $u = u(x/t)$. The PDE problem is then reduced to an ODE problem.

When $f'' \neq 0$, say, $f'' > 0$, here are two important solutions.

1. shock wave: $u_\ell \geq u_r$

$$u(x,t) = \begin{cases} u_\ell & \text{for } x < \sigma t \\ u_r & \text{for } x > \sigma t \end{cases} \tag{5.10}$$

   where $\sigma = (f(u_r) - f(u_\ell))/(u_r - u_\ell)$.

2. rarefaction wave: $u_\ell < u_r$

$$u(x,t) = \begin{cases} u_\ell & \text{for } x < \lambda_\ell t \\ u & \text{for } \lambda_\ell < \lambda(u) = \frac{x}{t} < \lambda_r \\ u_r & \text{for } x > \lambda_r t \end{cases} \tag{5.11}$$

   where $\lambda(u) = f'(u)$ is an increasing function.

These two solution are of fundamental importance. We shall denote them by (u_ℓ, u_r).

The weak solution is not unique. For instance, in the case of $u_\ell < u_r$, both (5.11) and (5.10) are weak solutions. Indeed, there are infinite many weak solutions to such a Riemann problem. Therefore, additional condition is needed to guarantee uniqueness. Such a condition is called an entropy condition.

5.2.2 Entropy conditions

To find a suitable entropy condition for general hyperbolic conservation laws, let us go back to study the gas dynamic problems. The hyperbolic conservation laws are simplified equations. The original physical equations usually contain a viscous term νu_{xx}, as that in the Navier-Stokes equation. We assume the viscous equation has uniqueness property. Therefore let us make the following definition.

Definition 5.2. A weak solution is called admissible if it is the limit of

$$u^\epsilon_t + f(u^\epsilon)_x = \epsilon u^\epsilon_{xx}, \tag{5.12}$$

as $\epsilon \to 0+$.

We shall label this condition by (A). In gas dynamics, the viscosity causes the physical entropy increases as gas particles passing through a shock front. One can show that such a condition is equivalent to the admissibility condition. Notice that this entropy increasing condition does not involve viscosity explicitly. Rather, it is a limiting condition as $\epsilon \to 0+$. This kind of conditions is what we are looking for. For general hyperbolic conservation laws, there are many of them. We list some of them below.

(L) Lax's entropy condition: across a shock (u_ℓ, u_r) with speed σ, the Lax's entropy condition is

$$\lambda_\ell > \sigma > \lambda_r \tag{5.13}$$

where λ_ℓ (λ_r) is the left (right) characteristic speed of the shock.
The meaning of this condition is that the information can only enter into a shock, then disappear. It is not allowed to have information coming out of a shock. Thus, if we draw characteristic curve from any point (x, t) backward in time, we can always meet the initial axis. It can not stop at a shock in the middle of time because it would violate the entropy condition. In other words, all information can be traced back to initial time. This is a causality property. It is also time irreversible, which is consistent to the second law of thermodynamics. However, Lax's entropy is only suitable for flux f with $f'' \neq 0$.

(OL) Oleinik-Liu's entropy condition: Let

$$\sigma(u, v) := \frac{f(u) - f(v)}{u - v}.$$

The Oleinik-Liu's entropy condition is that, across a shock

$$\sigma(u_\ell, v) \geq \sigma(u_\ell, u_r) \tag{5.14}$$

for all v between u_ℓ and u_r. This condition is applicable to nonconvex fluxes.

(GL) The above two conditions are conditions across a shock. Lax proposed another global entropy condition. First, he define entropy-entropy flux: a pair of function $(\eta(u), q(u))$ is called an entropy-entropy flux for equation (5.6) A weak solution $u(x, t)$ is said to satisfy entropy condition if for any entropy-entropy flux pair (η, q), $u(x, t)$ satisfies

$$\eta(u(x, t))_t + q(u(x, t))_x \leq 0 \tag{5.15}$$

in weak sense.

(K) Another global entropy proposed by Kruzkov is for any constant c,

$$\int_0^\infty \int_{-\infty}^\infty [|u - c|\phi_t + \text{sign}(u - c)(f(u) - f(c))\phi_x]\, dx \geq 0 \tag{5.16}$$

for all positive smooth ϕ with compact support in $\mathbb{R} \times (0, \infty)$. **(GL) $\Rightarrow$ (K):**
For any c, we choose $\eta(u) = |u - c|$, which is a convex function. One can check the corresponding $q(u) = \text{sign}(u - c)(f(u) - f(c))$. Thus, (K) is a special case of (GL). We may remark here that we can choose even simplier entropy-entropy flux:

$$\eta(u) = u \vee c,\ q(u) = f(u \vee c),$$

where $u \vee c := \max\{u, c\}$.

When the flux is convex, each of the above conditions is equivalent to the admissibility condition. When f is not convex, each but the Lax's entropy condition is equivalent to the admissibility condition.

We shall not provide general proof here. Rather, we study special case: the weak solution is only a single shock (u_ℓ, u_r) with speed σ.

Theorem 5.1. *Consider the scalar conservation law (5.6) with convex flux f. Let (u_ℓ, u_r) be its shock with speed σ. Then the above entropy conditions are all equivalent.*

Proof. **(L) $\Leftrightarrow$ (OL);**
We need to assume f to be convex. This part is easy. It follows from the convexity of f. We leave the proof to the reader.
(A) $\Leftrightarrow$ (OL):
We also need to assume f to be convex. Suppose (u_ℓ, u_r) is a shock. Its speed

$$\sigma = \frac{f(u_r) - f(u_\ell)}{u_r - u_\ell}.$$

We shall find a solution of (5.12) such that its zero viscosity limit is (u_ℓ, u_r). Consider a solution haing the form $\phi((x - \sigma t)/\epsilon)$. In order to have $\phi \to (u_\ell, u_r)$, we need to require far field condition:

$$\phi(\xi) \to \begin{cases} u_\ell & \xi \to -\infty \\ u_r & \xi \to \infty \end{cases} \tag{5.17}$$

Plug $\phi((x - \sigma t)/\epsilon)$ into (5.12), integrate in ξ once, we obtain

$$\phi' = F(\phi). \tag{5.18}$$

where $F(u) = f(u) - f(u_\ell) - \sigma(u - u_\ell)$. We find $F(u_\ell) = F(u_r) = 0$. This equation with far-field condition (5.17) if and only if, for all u between u_ℓ and u_r, (i) $F'(u) > 0$ when $u_\ell < u_r$, or (ii) $F'(u) < 0$ when $u_\ell > u_r$. One can check that (i) or (ii) is equivalent to (OL).

Next, we study global entropy conditions.
(A) $\Rightarrow$ (GL)
If u is an admissible solution. This means that it is the limit of u^ϵ which satisfy the viscous conservation law (5.12). Let (η, q) be a pair of entropy-entropy flux. Multiply (5.12) by $\eta'(u^\epsilon)$, we obtain

$$\begin{aligned} \eta(u^\epsilon)_t + q(u^\epsilon)_x &= \epsilon\eta'(u^\epsilon)u^\epsilon_{xx} \\ &= \epsilon\eta(u^\epsilon)_{xx} - \epsilon\eta''(u^\epsilon_x)^2 \\ &\le \epsilon\eta(u^\epsilon)_{xx} \end{aligned}$$

We multiply this equation by any positive smooth test function ϕ with compact support in $R \times (0, \infty)$, then integrate by part, and take $\epsilon \to 0$, we obtain

$$\int_0^\infty \int_{-\infty}^\infty [\eta(u)\phi_t + q(u)\phi_x]\, dx\, dt \ge 0$$

This means that $\eta(u)_t + q(u)_x \le 0$ in weak sense.

(K) $\Rightarrow$ (OL) for single shock:

Suppose (u_ℓ, u_r) is a shock. Suppose it satisfies (K). We want to show it satisfies (OL). The condition (GL), as applied to a single shock (u_ℓ, u_r), is read as

$$-\sigma[\eta] + [q] \le 0.$$

Here, we choose $\eta = |u - c|$. The condition becomes

$$-\sigma(|u_r - c| - |u_\ell - c|) + \text{sign}(u_r - c)(f(u_r) - f(c)) - \text{sign}(u_\ell - c)(f(u_\ell) - f(c)) \le 0$$

Or

$$-\sigma(u_\ell, u_r)(|u_r - c| - |u_\ell - c|) + |u_r - c|\sigma(u_r, c) - |u_\ell - c|\sigma(u_\ell, c) \le 0 \tag{5.19}$$

We claim that this condition is equivalent to (OL). First, if c lies outside of u_ℓ and u_r, then the left-hand side of (5.19) is zero. So (5.19) is always true in this case. Next, if c lies betrween u_ℓ and u_r, one can easily check it is equivalent to (OL). □

5.2.3 Rieman problem for nonconvex fluxes

The Oleinik-Liu's entropy condition can be interpreted as the follows graphically. Suppose (u_ℓ, u_r) is a shock, then the condition (OL) is equivalent to one of the follows. Either $u_\ell > u_r$ and the graph of f between u_ℓ, u_r lies below the secant $(u_r, f(u_r)), (u_\ell, f(u_\ell))$. Or $u_\ell < u_r$ and the graph of f between u_ℓ, u_r lies above the secant $((u_\ell, f(u_\ell)), (u_r, f(u_r)))$. With this, we can construct the solution to the Riemann problem for nonconvex flux as the follows.

If $u_\ell > u_r$, then we connect $(u_\ell, f(u_\ell))$ and $(u_r, f(u_r))$ by a convex envelope of f (i.e. the largest convex function below f). The straight line of this envelope corresponds to an entropy shock. In curved part, $f'(u)$ increases, and this portion corresponds to a centered rarefaction wave. Thus, the solution is a composition of rarefaction waves and shocks. It is called a copmposite wave.

If $u_\ell < u_r$, we simply replace convex envelope by concave envelope.

Example. Consider the cubic flux: $f(u) = \frac{1}{3}u^3$. If $u_\ell < 0$, $u_r > 0$ From u_ℓ, we can draw a line tangent to the graph of f at $u_\ell^* = -u_\ell/2$. If $u_r > u_\ell^*$, then the wave structure is a shock (u_ℓ, u_ℓ^*) follows by a rarefaction wave (u_ℓ^*, u_r). If $u_r \le u_\ell^*$, then the wave is a single shock. Notice that in a composite wave, the shock may contact to a rarefaction wave. Such a shock is called a contact shock.

Homeworks.

1. For the flux $f(u) = u^3/3$, construct the general solution to the Riemann problem for general left/right states u_ℓ andu_r.

5.3 Uniqueness and Existence

Theorem 5.2 (Kruzkov)**.** *Assume f is Lipschitz continuous and the initial data u_0 is in $L^1 \cap BV$. Then there exists a global entropy solution (satisfying condition (K)) to the Cauchy problem for*

(5.6). Furthermore, the solution operator is contractive in L^1, that is, if u, v are two entropy solutions, then

$$\|u(t) - v(t)\|_{L^1} \leq \|u(0) - v(0)\|_{L^1} \tag{5.20}$$

As a consequence, we have uniqueness theorem and the total variation diminishing property:

$$T.V.u(\cdot, t) \leq T.V.u(\cdot, 0) \tag{5.21}$$

Proof. The part of total variation diminishing is easy. We prove it here. The total variation of u is defined by

$$T.V.u(\cdot, t) = \text{Sup}_{h>0} \int \frac{|u(x+h,t) - u(x,t)|}{h}\, dx$$

We notice that if $u(x,t)$ is an entropy solution, so is $u(x+h,t)$. Apply the contraction estimate for $u(\cdot,t)$ and $v = u(\cdot + h, t)$. We obtain the total variation diminishing property.

To prove the L^1-contraction property, we claim that the constant c in the Kruzhkov entropy condition **(K)** can be replaced by any other entropy solution $v(t,x)$. That is

$$\int\int [|u(t,x) - v(t,x)|\psi_t + \text{sign}(u(t,x) - v(t,x))(f(u(t,x)) - f(v(t,x)))\psi_x]\, dx\, dt \geq 0$$

for all positive smooth ψ with compact support in $\mathbb{R} \times [0, \infty)$. To see this, we choose a test function $\phi(s,x,t,y)$, the entropy conditions for u and v are

$$\int\int [|u(s,x) - k|\phi_s(s,x,t,y) + \text{sign}(u(s,x) - k)(f(u(s,x)) - f(k))\phi_x(s,x,t,y)]\, dx\, ds \geq 0$$

$$\int\int [|v(t,y) - k'|\phi_t(s,x,t,y) + \text{sign}(v(t,y) - k')(f(v(t,y)) - f(k'))\phi_y(s,x,t,y)]\, dx\, ds \geq 0$$

Set $k = v(t,y)$ in the first equation and $k' = u(s,x)$ in the second equation. Integrate the rest variables and add them together. We get

$$\int\int\int\int \{|u(s,x) - v(t,y)|(\phi_s + \phi_t) + \text{sign}(u(s,x) - v(t,y)) \cdot [f(u(s,x)) - f(v(t,y))] \cdot (\phi_x + \phi_y)\}\, dx\, ds\, dy\, dt \geq$$

Now we choose $\phi(s,x,t,y)$ such that it concentrates at the diagonal $s = t$ and $x = y$. To do so, let $\rho_h(x) = h^{-1}\rho(x/h)$ be an approximation of the Dirac mass measure. Let $\psi(T,X)$ be a non-negative test function on $(0,\infty) \times \mathbb{R}$. Choosing

$$\phi(s,x,t,y) = \psi\left(\frac{s+t}{2}, \frac{x+y}{2}\right) \rho_h\left(\frac{s-t}{2}\right) \rho_h\left(\frac{x-y}{2}\right),$$

we get

$$\int\int\int\int \rho_h\left(\frac{s-t}{2}\right) \rho_h\left(\frac{x-y}{2}\right) \left\{|u(s,x) - v(t,y)|\psi_T\left(\frac{s+t}{2}, \frac{x+y}{2}\right)\right.$$
$$\left. + \text{sign}(u(s,x) - v(t,y)) \cdot [f(u(s,x)) - f(u(v(t,y)))] \cdot \psi_X\left(\frac{s+t}{2}, \frac{x+y}{2}\right)\right\} dx\, dy\, ds\, dt \geq 0.$$

Now taking limit $h \to 0$, we can get the desired inequality.

Next, we choose

$$\psi(t,x) = [\alpha_h(t) - \alpha_h(t-\tau)] \cdot [1 - \alpha_h(|x| - R + L(\tau - t))],$$

where $\alpha_h(z) = \int_{-\infty}^{z} \rho_h(s)\, ds$. We can get the desired L^1 contraction estimate. □

The existence theorem mainly based on the same proof of the uniqueness theorem. Suppose the initial data is in $L^1 \cap L^\infty \cap BV$, we can construct a sequence of approximate solutions which satisfy entropy conditions. They can be construncted by finite difference methods (see the next section), or by viscosity methods, or by wave tracking methods (by approximate the flux function by piecewise linear functions). Let us suppose the approximate solutions are constructed via viscosity method, namely, u^ε are solutions of

$$u_t^\varepsilon + f(u^\varepsilon)_x = \varepsilon u_{xx}^\varepsilon.$$

Following the same proof for $(GL) \Rightarrow (K)$, we can get that the total variation norms of the approximate solutions u^ε are bounded by $T.V.u_0$. This gives the compactness in L^1 and a convergent subsequence leads to an entropy solution.

Remark. The general existence theorem can allow only initial data $u_0 \in L^1 \cap L^\infty$. Even the initial data is not in BV, the solution immediately has finite total variation at any $t > 0$.

Chapter 6

Finite Difference Schemes For Scalar Conservation Laws

6.1 Major problems

First of all, we should keep in mind that local stability is necessary in designing finite difference schemes for hyperbolic conservation laws. Namely, the scheme has to be stable for hyperbolic conservation laws with frozen coefficients, see Chapter 1. In addition, there are new things that we should be careful for nonlinear equations. The main issue is how to compute discontinuities correctly. We list common problems on this issue.

- Spurious oscillation appears around discontinuities in every high order schemes. The reason is that the solution of finite difference scheme is closer to a PDE with higher order derivatives. The corresponding dispersion formula demonstrates that oscillation should occur. Also, one may view that it is incorrect to approximate weak derivative at discontinuity by higher order finite differences. The detail spurious structure can be analyzed by the study of the discrete traveling wave corresponding to a finite difference scheme.

  To cure this problem, we have to lower down the order of approximation near discontinuities to avoid oscillation. We shall devote to this issue later.

- The approximate solution may converge to a function which is not a weak solution. For instance, let us apply the Courant-Isaacson-Rees (C-I-R) method to compute a single shock for the inviscid Burgers equation. The C-I-R method is based on characteristic method. Suppose we want to update the state U_j^{n+1}. We draw a characteristic curve back to time t_n. However, the slope of the characteristic curve is not known yet. So, let us approximate it by U_j^n. Then we apply upwind method:

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{\Delta x} U_j^n (U_{j-1}^n - U_j^n) \quad \text{if } U_j^n \geq 0$$
$$U_j^{n+1} - U_j^n = \frac{\Delta t}{\Delta x} U_j^n (U_j^n - U_{j+1}^n) \quad \text{if } U_j^n < 0$$

Now, we take the following initial data:

$$U_j^0 = \begin{cases} 1 & \text{for } j < 0 \\ 0 & \text{for } j \geq 0 \end{cases}$$

It is easy to see that $U_j^n = U_j^0$. This is a wrong solution. The reason is that we use wrong characteristic speed U_j^n when there is a discontinuity passing x_j from t_n to t_{n+1}.

To resolve this problem, it is advised that one should use a conservative scheme. We shall discuss this issue in the next section.

- Even the approximate solutions converge to a weak solution, it may not be an entropy solution. For instance, consider the invisid Burgers equation $u_t + uu_x = 0$ with the initial data:

$$U_j^0 = \begin{cases} -1 & \text{for } j < 0 \\ 1 & \text{for } j \geq 0 \end{cases}$$

We define the scheme by

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(F(U_{j-1}^n, U_j^n) - F(U_j^n, U_{j+1}^n))$$

where

$$F(U,V) = \begin{cases} f(U) & \text{if } U + V \geq 0 \\ f(V) & \text{if } U + V < 0 \end{cases}$$

We find that $F(U_j^n, U_{j+1}^n) = F(U_{j-1}^n, U_j^n)$. Thus, the solution is $U_j^n = U_j^0$. This is a nonentropy solution.

6.2 Conservative schemes

A finite difference scheme is called conservative if it can be written as

$$\boxed{U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(F_{j-1/2}^{n+1/2} - F_{j+1/2}^{n+1/2})} \tag{6.1}$$

where $F_{j+1/2}^{n+1/2}$ is a function of U^n and possibly U^{n+1}. The advantage of this formulation is that the total mass is conservative:

$$\sum_j U_j^n = \sum_j U_j^{n+1} \tag{6.2}$$

There is a nice interpretation of F if we view U_j^n as an approximation of the cell-average of the solution u over the cell $(x_{j-1/2}, x_{j+1/2})$ at time step n. Let us integrate the conservation law $u_t + f(u)_x = 0$ over the box: $(x_{j-1/2}, x_{j+1/2}) \times (t_n, t_{n+1})$. Using divergence theorem, we obtain

$$\bar{u}_j^{n+1} = \bar{u}_j^n + \frac{\Delta t}{\Delta x}(\bar{f}_{j-1/2}^{n+1/2} - \bar{f}_{j+1/2}^{n+1/2}) \tag{6.3}$$

where

$$\begin{aligned} \bar{u}_j^n &= \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t_n)\,dx \\ \bar{f}_{j+1/2}^{n+1/2} &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+1/2},t))\,dt \end{aligned}$$

Thus, in a conservative scheme (6.1), we may view U_j^n as an approximation of the cell average $\bar{u}_j^n$ and $F_{j+1/2}^{n+1/2}$ as an approximation of the flux average $\bar{f}_{j+1/2}^{n+1/2}$. This formulation is closer to the original integral formulation of a conservation, and it does not involve derivatives of the unknown quantity u.

A conservative scheme is consistent if $F_{j+1/2}(U,U) = f(u)$, where U is a vector with $U_j = u$. For explicit scheme, $F_{j+1/2}$ is a function of U^n only and it only depends on $U_{j-\ell+1}^n, \cdots, U_{j+m}^n$. That is

$$F_{j+1/2} = F(U_{j-\ell+1}^n, \cdots, U_{j+m}^n).$$

We usually assume that the function is a Lipschitz function.

The most important advantage of conservative schemes is the following Lax-Wendroff theorem. Which says that its approximate solutions, if converge, must to a weak solution.

Theorem 6.1 (Lax-Wendroff)**.** *Suppose $\{U_j^n\}$ be the solution of a conservative scheme (6.1). The Define $u_{\Delta x} := U_j^n$ for $[x_{j-1/2}, x_{j+1/2}) \times [t_n, t_{n+1})$. Suppose $u_{\Delta x}$ is uniformly bounded and converges to u almost everywhere. Then u is a weak solution of (5.6).*

Proof. Let ϕ be a smooth test function with compact support on $R \times [0,\infty)$. We multiply (6.1) by ϕ_j^n and sum over j and n to obtain

$$\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty} \phi_j^n (U_j^{n+1} - U_j^n) = \frac{\Delta t}{\Delta x}\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty} \phi_j^n [F_{j-1/2}(U^n) - F_{j+1/2}(U^n)]$$

Using summation by part, we obtain

$$\sum_{j=-\infty}^{\infty} \phi_j^0 U_j^0 + \sum_{n=1}^{\infty}\sum_{j=-\infty}^{\infty} (\phi_j^n - \phi_j^{n-1}) U_j^n + \sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty} (\phi_{j+1}^n - \phi_j^n) F_{j+1/2}(U^n) = 0$$

Since ϕ is of compact support and $u_{\Delta x}$, hence $F(U^n)$, are uniformly bounded, we obtain the convergence in the above equation is uniformly in j and n. If $(x_j, t_n) \to (x,t)$, then from the consistency condition, $F_{j+1/2}(U^n) \to f(u(x,t))$. We obtain that u is a weak solution. □

Below, we show that many scheme can be written in conservation form. We may view $F_{j+1/2}^{n+1/2}$ as a numerical flux at $x_{j+1/2}$ between t_n and t_{n+1}.

1. Lax-Friedrichs:

$$F_{j+1/2}^{n+1/2} = F(U_j, U_{j+1}) = \frac{1}{2}(f(U_{j+1}) + f(U_j)) + \frac{\Delta t}{2\Delta x}(U_j - U_{j+1}). \tag{6.4}$$

The second term is a numerical dissipation.

2. Two-step Lax-Wendroff:

$$\begin{aligned} F_{j+1/2}^{n+1/2} &= f(U_{j+1/2}^{n+1/2}) \\ U_{j+1/2}^{n+1/2} &= \frac{U_j^n + U_{j+1}^n}{2} + \frac{\Delta t}{2\Delta x}\left[f(U_j^n) - f(U_{j+1}^n)\right] \end{aligned}$$

Homeworks. Construct an example to show that the Lax-Wendroff scheme may produce nonentropy solution.

6.3 Entropy and Monotone schemes

Definition 6.1. A scheme expressed as

$$U_j^{n+1} = G(U_{j-\ell}^n, \cdots, U_{j+m}^n) \tag{6.5}$$

is called a monotone scheme if

$$\frac{\partial G}{\partial U_{j+k}} \geq 0, k = -\ell, \cdots, m \tag{6.6}$$

In the case of linear equation, the monotone scheme is

$$U_j^{n+1} = \sum_{k=-\ell}^{m} a_k U_{j+k}^n$$

with $a_k \geq 0$. The consistency condition gives $\sum_k a_k = 1$. Thus, a monotone scheme in linear cases means that U_j^{n+1} is an average of $U_{j-\ell}^n, \cdots, U_{j+m}^n$. In the nonlinear case, this is more or less "true." For instance, the sup norm is nonincreasing, the solution operator is ℓ^1-contraction, and the total variation is dimishing. To be precise, let us define the norms for $U = \{U_j\}$:

$$\begin{aligned} |U|_\infty &= \sup_j |U_j| \\ \|U\|_1 &= \sum_j |U_j| \Delta x \\ T.V.(U) &= \sum_j |U_{j+1} - U_j| \end{aligned}$$

We have the following theorem.

Theorem 6.2. *For a monotone scheme (6.5), we have*

(i) ℓ^∞- bound:

$$|U^{n+1}|_\infty \leq |U^n|_\infty$$

(ii) ℓ^1-contraction: if U, V are two solutions of (6.1), then

$$\|U^{n+1} - V^{n+1}\|_1 \leq \|U^n - V^n\|_1 \tag{6.7}$$

(iii) total variation diminishing:

$$T.V._x(U^{n+1}) \leq T.V._x(U^n) \tag{6.8}$$

(iv) boundedness of total variation: there exists a constant C such that

$$T.V._{x,t}(U) \leq C \tag{6.9}$$

Proof. 1.

$$\begin{aligned} U_j^{n+1} &= G(U_{j-\ell}^n, \cdots, U_{j+m}^n) \\ &\leq G(\max U^n, \cdots, \max U^n) \\ &= \max U^n \end{aligned}$$

Hence, we have $\max U^{n+1} \leq \max U^n$. Similarly, we also have $\min U^{n+1} \geq \min U^n$.

2. Let us denote the vector (U_j^n) by U^n, the scheme (6.5) by an operator $U^{n+1} = G(U^n)$. $U \leq V$ means that $U_j \leq V_j$ for each j. Denote by $U \vee V$ for the vector $(\max\{U_j, V_j\})$. The monotonicity reads

$$G(U) \leq G(V) \text{ if } U \leq V.$$

We have $G(U \vee V) \geq G(V)$. Hence,

$$(G(U) - G(V))^+ \leq ((G(U \vee V) - G(V))^+ = G(U \vee V) - G(V).$$

We take summation in j, and use conservative property of G, namely, $\sum_j (G(U))_j = \sum_j U_j$, we obtain

$$\sum_j (G(U) - G(V))_j^+ \leq \sum_j ((U \vee V) - V)_j = \sum_j (U - V)_j^+.$$

Similarly, we have

$$\sum_j (G(V) - G(U))_j^+ \leq \sum_j (V - U)_j^+.$$

Adding these two, we obtain the ℓ^1-contraction:

$$\sum_j |G(U)_j - G(V)_j| \leq \sum_j |U_j - V_j|.$$

3. Suppose U_j^n is a solution of (6.5). We take V_j^n to be U_{j+1}^n. Then V_j^n also satisfies (6.5). From the ℓ^1-contraction property, we have

$$\sum_j |U_{j+1}^{n+1} - U_j^{n+1}| \le \sum_j |U_{j+1}^n - U_j^n|$$

This shows the total variation dimishing property of (6.5).

4. The total variation of U in x, t with $0 \le t \le T$ is defined by

$$\begin{aligned} T.V._{x,t}(U) &= \sum_{n=0}^{N} \sum_{j=-\infty}^{\infty} \left[\frac{|U_{j+1}^n - U_j^n|}{\Delta x} + \frac{|U_j^{n+1} - U_j^n|}{\Delta t} \right] \Delta x \, \Delta t \\ &= \sum_{n=0}^{N} \left[T.V._x U^n \Delta t + \|U^{n+1} - U^n\|_{L^1} \right] \\ &= T.V._x U^n T + \sum_{n=0}^{N} \|U^{n+1} - U^n\|_{L^1}. \end{aligned}$$

Here $N\Delta t = T$. We claim that $\|U^{n+1} - U^n\|_{L^1} \le O(\Delta t)$. If so, then we obtain the result with $C \le T + NO(\Delta t) \le T + KT$ for some constant K. Now, we prove this claim:

$$\begin{aligned} \sum_j |U_j^{n+1} - U_j^n| &= \sum_j |G(U_{j-\ell}^n, \cdots, U_{j+m}^n) - G(U_j^n, \cdots, U_j^n)| \\ &\le \sum_j L(|U_{j-\ell}^n - U_j^n| + \cdots + |U_{j+m}^n - U_j^n|) \\ &\le L(\ell + m)^2 T.V._x(U^n). \end{aligned}$$

Here, we have used that G is Lipschitz continuous. Hence, we conclude

$$\sum_j |U_j^{n+1} - U_j^n| \Delta t \le O(\Delta t).$$

□

The boundedness of total variation of U in (x, t) implies that we can substract a subsequence $u_{\Delta x}$ which converges in L^1. Below, we show that its limit indeed satisfies entropy condition.

Theorem 6.3. *The limiting function of the approximate solutions constructed from a monotone scheme satisfies Kruzkov's entropy condition.*

Proof. We choose $\eta = (u - c)^+ = u \vee c - c$. The corresponding entropy flux is $q(u) = f(u \vee c) - f(c)$. It is natural to choose the numerical entropy flux to be $Q(U_{j-\ell+1}, \cdots, U_{j+m}) = F(U_{j-\ell+1} \vee c, \cdots, U_{j+m} \vee c) - F(c, \cdots, c)$. We have

$$(U^{n+1} \vee c) = G(U_{j-\ell}^n, \cdots, U_{j+m}^n) \vee G(c, \cdots, c)$$

$$
\begin{aligned}
&\leq G(U^n_{j-\ell} \vee c, \cdots, U^n_{j+m} \vee c) \\
&= U^n_j \vee c + \frac{\Delta t}{\Delta x}\left[F(U^n_{j-\ell} \vee c, \cdots, U^n_{j+m-1} \vee c) - F(U^n_{j-\ell+1} \vee c, \cdots, U^n_{j+m} \vee c)\right] \\
&= U^n_j \vee c + \frac{\Delta t}{\Delta x}\left[Q(U^n_{j-\ell}, \cdots, U^n_{j+m-1}) - Q(U^n_{j-\ell+1}, \cdots, U^n_{j+m})\right]
\end{aligned}
$$

Multiply this inequality by ϕ^n_j, sum over j and n, and apply "summation-by-part", then take limit $\Delta t, \Delta x \to 0$. We obtain that u is an entropy solution. □

Theorem 6.4 (Harten-Hyman-Lax)**.** *A monotone scheme (6.5) is at most first order.*

Proof. We claim that the modified equation corresponding to a monotone scheme has the following form

$$
u_t + f(u)_x = \Delta t[\beta(u, \lambda)u_x]_x \tag{6.10}
$$

where $\lambda = \Delta t/\Delta x$,

$$
\beta = \frac{1}{2\lambda^2}\sum_{k=-\ell}^{m} k^2 G_k(u, \cdots, u) - \frac{1}{2}f'(u)^2, \quad G_k := \frac{\partial G}{\partial u_k}, \tag{6.11}
$$

and $\beta > 0$ except for some exceptional cases. Since for smooth solution, the solution of finite difference equation is closer to the modified equation, we see that the scheme is at most first order.

To show (6.10), we take Taylor expansion of G about $(u_0, \cdots, u_0)$:

$$
\begin{aligned}
G(u_{-\ell}, \cdots, u_m) &= G(u_0, \cdots, u_0) \\
&\quad + \sum_{k=-\ell}^{m} G_k(u_k - u_0) \\
&\quad + \frac{1}{2}\sum_{j,k=-\ell}^{m} G_{j,k}(u_j - u_0)\,(u_k - u_0) + O(\Delta x)^3 \\
&= u_0 + \Delta x u_x \sum_{k=-\ell}^{m} kG_k + \frac{1}{2}(\Delta x)^2 u_{xx}\sum_{k=-\ell}^{m} k^2 G_k \\
&\quad + \sum_{j,k}\frac{1}{2}(\Delta x)^2 u_x^2 jkG_{j,k} + O(\Delta x)^3 \\
&= u_0 + \Delta x u_x \sum_{k=-\ell}^{m} kG_k + \frac{1}{2}(\Delta x)^2\left(\sum_{k=-\ell}^{m} k^2 G_k u_x\right)_x \\
&\quad + \sum_{j,k}\frac{1}{2}(\Delta x)^2 u_x^2 (jk - k^2)G_{j,k} + O(\Delta x)^3
\end{aligned}
$$

On the other hand,

$$
G(u_{-\ell}, \cdots, u_m) = u_0 + \lambda(F(\bar{u}) - F(T\bar{u}))
$$

where $\bar{u} = (u_{-\ell}, \cdot, u_{m-1})$, $T\bar{u} = (u_{-\ell+1}, \cdots, u_m)$. We differentiate this equation to obtain

$$\begin{aligned} G_k &= \delta_{0,k} + \lambda[F_k(\bar{u}) - F_{k-1}(T\bar{u})] \\ G_{j,k} &= \lambda[F_{j,k}(\bar{u}) - F_{j-1,k-1}(T\bar{u})] \end{aligned}$$

We differentiate the consistency condition $F(u_0, \cdots, u_0) = f(u_0)$ to obtain

$$\sum_{-\ell}^{m-1} F_k(u_0, \cdots, u_0) = f'(u_0).$$

Therefore,

$$\begin{aligned} \sum_{k=-\ell}^{m} G_k &= 1 \\ \sum_{k=-\ell}^{m} kG_k &= \lambda \sum (F_k - F_{k-1})k = -\lambda f'(u_0) \\ \sum_{j,k} (j-k)^2 G_{j,k} &= \lambda \sum (j-k)^2 [G_{j-1,k-1} - G_{j,k}] = 0 \end{aligned}$$

Using this and the symmetry $G_{j,k} = G_{k,j}$, we obtain

$$\sum_{j,k} G_{j,k}(jk - k^2) = -\frac{1}{2} \sum G_{j,k}(j-k)^2 = 0.$$

Hence we obtain

$$G(u_{-\ell}, \cdots, u_m) = u_0 - \Delta x \lambda f'(u) u_x + (\frac{1}{2}\Delta x)^2 u_{xx} \sum_k k^2 G_k + O(\Delta x)^3$$

Now, from the Taylor expansion:

$$\begin{aligned} u_0^1 &= u_0 + \Delta t u_t + \frac{1}{2}(\Delta t)^2 u_{tt} + O(\Delta t)^3 \\ &= u_0 - \Delta t f(u)_x + (\frac{1}{2}\Delta t)^2 [f'(u)^2 u_x]_x + O(\Delta t)^3 \end{aligned}$$

Combine these two, we obtain that smooth solution of the finite difference equation satisfy the modified equation up to a truncation error $(\Delta t)^2$.

To show $\beta \geq 0$, from the monotonicity $G_k \geq 0$. Hence

$$\begin{aligned} \lambda^2 f'(u)^2 &= \left(\sum_k kG_k \right)^2 = \left(\sum k\sqrt{G_k}\sqrt{G_k} \right)^2 \\ &\leq \sum k^2 G_k \cdot \sum G_k = \sum_k k^2 G_k \end{aligned}$$

The equality holds only when $G_k(u, \cdots, u) = 0$ for all k except 1. This means that $G(u_\ell, \cdots, u_m) = u_1$. This is a trivial case.

□

Chapter 7

Finite Difference Methods for Hyperbolic Conservation Laws

Roughly speaking, modern schemes for hyperbolic conservation laws can be classified into the following two categories.

1) flux-splitting methods

2) high-order Godunov methods

1) is more algebraic construction while 2) is more geometrical construction.
Among 1), there are

- artificial viscosity methods,
- flux correction transport (FCT),
- total variation diminishing (TVD),
- total variation bounded (TVB),
- central scheme,
- relaxation schemes,
- relaxed scheme.

Among 2), there are

- High order Godunov methods,
- MUSCL,
- piecewise parabolic method (PPM),
- essential nonoscillatory. (ENO)

In 1) we describe total variation diminishing method while in 2) we show the high order Godunov methods.

7.1 Flux splitting methods

The basic thinking for these methods is to add a switch such that the scheme becomes first order near discontinuity and remains high order in the smooth region.

Suppose we are given

$$F^L \text{ a lower order numerical flux}$$
$$F^H \text{ a higher order numerical flux}$$

Define

$$\begin{aligned} F_{j+\frac{1}{2}} &= F^L_{j+\frac{1}{2}} + \phi_{j+\frac{1}{2}}(F^H_{j+\frac{1}{2}} - F^L_{j+\frac{1}{2}}) \\ &= F^H_{j+\frac{1}{2}} + (1-\phi_{j+\frac{1}{2}})(F^L_{j+\frac{1}{2}} - F^H_{j+\frac{1}{2}}). \end{aligned}$$

Here, $\phi_{j+\frac{1}{2}}$ is a switch or a limiter. We require
$\phi_{j+\frac{1}{2}} \sim 0$, i.e. $F_{j+\frac{1}{2}} \sim F^L_{j+\frac{1}{2}}$, near a discontinuity,
$\phi_{j+\frac{1}{2}} \sim 1$, i.e. $F_{j+\frac{1}{2}} \sim F^H_{j+\frac{1}{2}}$, in smooth region.
In FCT, ϕ is chosen so that $\max U_j^{n+1} \le \max(U^n_{j-1}, U^n_j, U^n_{j+1})$ and $\min U_j^{n+1} \ge \min(U^n_{j-1}, U^n_j, U^n_{j+1})$.

Design Criterion for $\phi_{j+\frac{1}{2}}$

7.1.1 Total Variation Diminishing (TVD)

Consider the linear advection equation

$$u_t + au_x = 0, \quad a > 0.$$

We show the ideas by choosing

$$\begin{aligned} F^L_{j+\frac{1}{2}} &= aU_j && \text{be upwind's flux, and} \\ F^H_{j+\frac{1}{2}} &= aU_j + \tfrac{1}{2}a(1-\tfrac{a\Delta t}{\Delta x})(U_{j+1}-U_j) && \text{be Lax-Wendroff's flux.} \end{aligned}$$

Then the numerical flux is

$$F_{j+\frac{1}{2}} = aU_j + \phi_{j+\frac{1}{2}}(\frac{1}{2}a(1-\frac{a\Delta t}{\Delta x})(U_{j+1}-U_j)). \tag{7.1}$$

Here

$$\phi_{j+\frac{1}{2}} = \phi(\theta_{j+\frac{1}{2}}),$$
$$\theta_{j+\frac{1}{2}} := \frac{U_j - U_{j-1}}{U_{j+1}-U_j}.$$

Theorem 7.1. *1. If ϕ is bounded, then the scheme is consistent with the partial differential equation.*

2. If $\phi(1) = 1$, and ϕ is Lipschitz continuous(or C^1) at $\theta = 1$, then the scheme is second order in smooth monoton region.(i.e., u is smooth and $u_x \neq 0$)

3. If $0 \leq \frac{\phi(\theta)}{\theta} \leq 2$ and $0 \leq \phi(\theta) \leq 2$, then the scheme is TVD.

Proof. 1. $F_{j+\frac{1}{2}}(u,u) = f(u) = au$.

2. Hint: Apply truncation error analysis.

3. From (7.1), the next time step U_j^{n+1} is

$$U_j^{n+1} = U_j^n - c_{j-1}^n(U_j^n - U_{j-1}^n),$$

where $c_{j-1}^n = \nu + \frac{1}{2}\nu(1-\nu)(\frac{\phi_{j+\frac{1}{2}}(U_{j+1}^n - U_j^n) - \phi_{j-\frac{1}{2}}(U_j^n - U_{j-1}^n)}{U_j^n - U_{j-1}^n})$, $\nu = \frac{a\Delta t}{\Delta x}$.
In other words, U_j^{n+1} is the average of U_j^n and U_{j-1}^n with weights $(1 - c_{j-1}^n)$ and c_{j-1}^n.

$$\begin{aligned} U_{j+1}^{n+1} - U_j^{n+1} &= (U_{j+1}^n - c_j^n(U_{j+1}^n - U_j^n)) - (U_j^n - c_{j-1}^n(U_j^n - U_{j-1}^n)) \\ &= (1 - c_j^n)(U_{j+1}^n - U_j^n) + c_{j-1}^n(U_j^n - U_{j-1}^n) \end{aligned}$$

Suppose $1 \geq c_j^n \geq 0 \ \ \forall j, n$

$$|U_{j+1}^{n+1} - U_j^{n+1}| \leq (1 - c_j^n)|U_{j+1}^n - U_j^n| + c_{j-1}^n|U_j^n - U_{j-1}^n|$$

$$\begin{aligned} \sum_j |U_{j+1}^{n+1} - U_j^{n+1}| &\leq \sum_j (1 - c_j^n)|U_{j+1}^n - U_j^n| + \sum c_{j-1}^n|U_j^n - U_{j-1}^n| \\ &= \sum_j (1 - c_j^n)|U_{j+1}^n - U_j^n| + \sum c_j^n|U_{j+1}^n - U_j^n| \\ &= \sum_j |U_{j+1}^n - U_j^n|, \end{aligned}$$

then the computed solution is total variation diminishing.
Next, we need to find ϕ such that $0 \leq c_j^n \leq 1, \ \forall j, n$. Consider

$$\frac{\phi_{j+\frac{1}{2}}(U_{j+1} - U_j) - \phi_{j-\frac{1}{2}}(U_j - U_{j-1})}{U_j - U_{j-1}} = \frac{\phi(\theta_{j+\frac{1}{2}})}{\theta_{j+\frac{1}{2}}} - \phi(\theta_{j-\frac{1}{2}}),$$

$$\Longrightarrow c_{j-1}^n = \nu + \frac{1}{2}\nu(1-\nu)(\frac{\phi(\theta_{j+\frac{1}{2}})}{\theta_{j+\frac{1}{2}}} - \phi(\theta_{j-\frac{1}{2}})) \quad 0 \leq \nu \leq 1$$

A sufficient condition for $0 \leq c_{j-1}^n \leq 1 \ \ \forall j$ is

$$|\frac{\phi(\theta_{j+\frac{1}{2}})}{\theta_{j+\frac{1}{2}}} - \phi(\theta_{j-\frac{1}{2}})| \leq 2. \tag{7.2}$$

If $\theta_{j+\frac{1}{2}} < 0$, $\phi(\theta_{j+\frac{1}{2}}) = 0$.
If $0 \leq \frac{\phi(\theta)}{\theta} \leq 2$, $0 \leq \phi(\theta) \leq 2$, then (7.2) is valid.

□

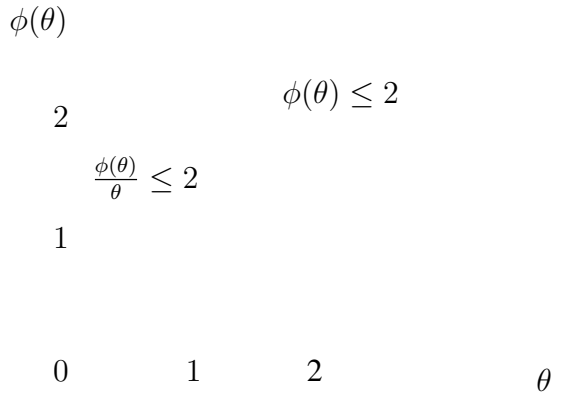


Figure 7.1: The region in which $\phi(\theta)$ should lie so that the scheme will be TVD.

7.1.2 Other Examples for $\phi(\theta)$

1. $\phi(\theta) = 1$. This is the Lax-Wendroff scheme.
2. $\phi(\theta) = \theta$. This is Beam-Warming.
3. Any ϕ between ϕ_{B-W} and ϕ_{L-W} with $0 \leq \phi \leq 2$, $0 \leq \frac{\phi(\theta)}{\theta} \leq 2$ is second order.
4. Van Leer's minmod
$$\phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|}.$$
It is a smooth limiter with $\phi(1) = 1$.
5. Roe's superbee
$$\phi(\theta) = \max(0, \min(1, 2\theta), \min(\theta, 2))$$

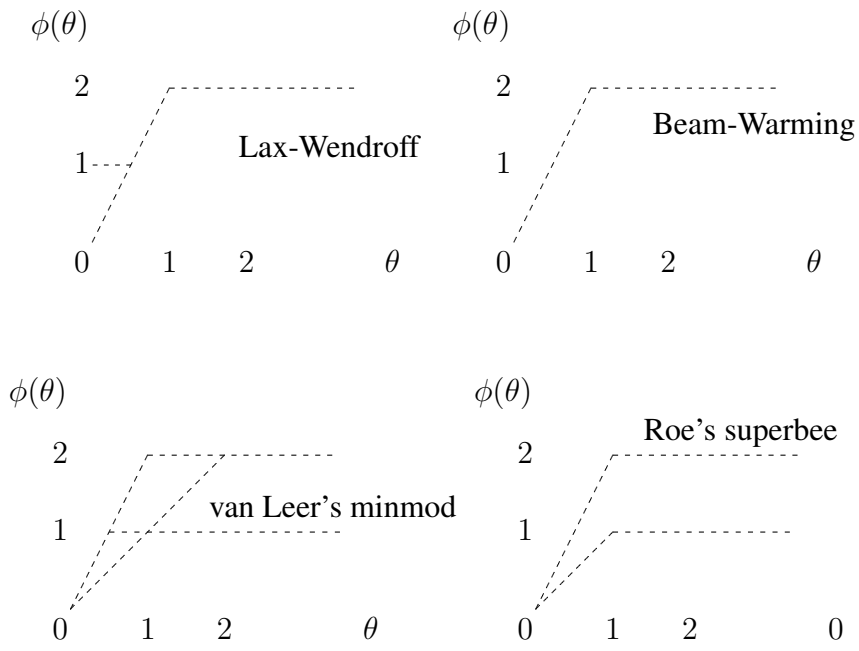


Figure 7.2: Several limiters

7.1.3 Extensions

There are two kinds of extensions. One is the $a < 0$ case, and the other is the linear system case.

For $a < 0$, we let

$$\begin{aligned} F^L_{j+\frac{1}{2}} &= \frac{1}{2}a(U_j + U_{j+1}) - \frac{1}{2}|a|(U_{j+1} - U_j) \\ &= \begin{cases} aU_j & \text{if } a > 0 \\ aU_{j+1} & \text{if } a < 0 \end{cases} \\ F^H_{j+\frac{1}{2}} &= \frac{1}{2}a(U_j + U_{j+1}) - \frac{1}{2}\nu a(U_{j+1} - U_j) \quad \nu = \frac{a\Delta t}{\Delta x} \end{aligned}$$

Then

$$\begin{aligned} F_{j+\frac{1}{2}} &= F^L_{j+\frac{1}{2}} + \phi_{j+\frac{1}{2}}(F^H_{j+\frac{1}{2}} - F^L_{j+\frac{1}{2}}) \\ &= F^L_{j+\frac{1}{2}} + \phi_{j+\frac{1}{2}}\frac{1}{2}(\text{sign}(\nu) - \nu)a(U_{j+1} - U_j). \end{aligned}$$

Where $\phi_{j+\frac{1}{2}} = \phi(\theta_{j+\frac{1}{2}})$, $\theta_{j+\frac{1}{2}} = \frac{U_{j'+1} - U_{j'}}{U_{j+1} - U_j}$, and $j' = j - \text{sign}(\nu) = j \pm 1$.

In the linear system case, our equation is

$$u_t + Au_x = 0. \tag{7.3}$$

We can decompose A so that $A = R\Lambda R^{-1}$ with $\Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$ constituting by A's eigenvalues and $R = [r_1, \cdots, r_n]$ being right eigenvectors.That is, $Ar_i = \lambda_i r_i$. We know that $U_{j+1} - U_j = \sum_{k=1}^{n} \alpha_{j,k} r_k$, let

$$\begin{aligned} \nu_k &= \lambda_k \frac{\Delta t}{\Delta x} \\ \theta_{j,k} &= \frac{\alpha_{j',k}}{\alpha_{j,k}} \quad j' = j - \text{sign}(\nu_k). \end{aligned}$$

Therefore,

$$\begin{aligned} F^L &= \frac{1}{2}A(U_j + U_{j+1}) - \frac{1}{2}|A|(U_{j+1} - U_j) \\ F^H &= \frac{1}{2}A(U_j + U_{j+1}) - \frac{1}{2}\frac{\Delta t}{\Delta x}A^2(U_{j+1} - U_j) \end{aligned}$$

where $|A| = R|\Lambda|R^{-1}$. The numerical flux is

$$F_{j+\frac{1}{2}} = F^L_{j+\frac{1}{2}} + \frac{1}{2}\sum_k \phi(\theta_{j,k})(\text{sign}(\nu_k) - \nu_k)\lambda_k \alpha_{j,k} r_k.$$

7.2 High Order Godunov Methods

Algorithm

1. Reconstruction: start from cell averages $\{U_j^n\}$, we reconstruct a piecewise polynomial function $\tilde{u}(x, t_n)$.
2. "Exact" solver for $u(x,t)$, $t_n < t < t_{n+1}$. It is a Riemann problem with initial data $\tilde{u}(x, t_n)$.
3. Define
$$U_j^{n+1} = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}(x, t_{n+1})\, dx.$$

If 2. is an exact solver, using

$$\int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_t + f(u)_x \, dxdt = 0$$

we have

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(\tilde{f}_{j-\frac{1}{2}} - \tilde{f}_{j+\frac{1}{2}}),$$

where $\tilde{f}_{j+\frac{1}{2}} = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(\tilde{u}(x_{j+\frac{1}{2}}, t))\, dt$ is the average flux. Thus 2. and 3. can be replaced by

2'. an 'Exact solver" for u at $x_{j+\frac{1}{2}}, t_n < t < t_{n+1}$ to compute averaged flux $\tilde{f}_{j+\frac{1}{2}}$.

3'. $U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(\tilde{f}_{j-\frac{1}{2}} - \tilde{f}_{j+\frac{1}{2}})$

1. Reconstruction: We want to construct a polynomial in each cell. The main criterions are

(1) high order in regions where u is smooth and $u_x \neq 0$

(2) total variation no increasing.

In other words, suppose we are given a function $u(x)$, let

$$U_j = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\, dx$$

From $\{U_j\}$, we can use some reconstruct algorithm to construct a function $\tilde{u}(x)$. We want the reconstruction algorithm to satisfy

(1) $|\tilde{u}(x) - u(x)| = O(\Delta x)^r$, where u is smooth in $I_j = (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ and $u_x \neq 0$ near I_j.

(2) $T.V.\tilde{u}(x) \leq T.V.u(x)(1 + O(\Delta x))$

7.2.1 Piecewise-constant reconstruction

Our equation is

$$u_t + f(u)_x = 0 \tag{7.4}$$

Following the algorithm, we have

(1) approximate $u(t, x)$ by piecewise constant function, i.e., $\{U_j^n\}$ represents the cell average of $u(x, t_n)$ over $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$.

$$\Delta x$$

$$x_{j-\frac{1}{2}} \qquad\qquad x_{j+\frac{1}{2}}$$

(2) solve Riemann problem
(u_j, u_{j+1}) on the edge $x_{j+\frac{1}{2}}$, its solution $\tilde{u}(x_{j+\frac{1}{2}}, t)$,$t_n < t < t_{n+1}$ can be found, which is a constant.

(3) integrate the equation (7.4) over $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times (t_n, t_{n+1})$

$$\begin{aligned}
\Longrightarrow U_j^{n+1} &= \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}(x, t_{n+1})\, dx \\
&= U_j^n + \frac{\Delta t}{\Delta x}\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \Big(f(\tilde{u}(x_{j-\frac{1}{2}}, t)) - f(\tilde{u}(x_{j+\frac{1}{2}}, t)\Big)\, dt \\
&= u_j^n + \frac{\Delta t}{\Delta x}[f(\tilde{u}(x_{j-\frac{1}{2}}, t_{n+\frac{1}{2}})) - f(\tilde{u}(x_{j+\frac{1}{2}}, t_{n+\frac{1}{2}}))]
\end{aligned}$$

Example 1 $f(u) = au \quad a > 0$

Riemann problem gives $\tilde{u}(x,t) = \begin{cases} u_j & \text{if } x - x_{j+\frac{1}{2}} < at, t_n < t < t_{n+1} \\ u_{j+1} & \text{if } x - x_{j+\frac{1}{2}} > at, t_n < t < t_{n+1} \end{cases}$

$$\begin{aligned}
u_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= \tilde{u}(x_{j+\frac{1}{2}}, t_{n+\frac{1}{2}}) = u_j \\
F_{j+\frac{1}{2}} &= aU_{j+\frac{1}{2}}^{n+1} = aU_j \\
\therefore U_j^{n+1} &= U_j^n + \frac{\Delta t}{\Delta x}(aU_{j-1}^n - aU_j^n)
\end{aligned}$$

This is precisely the upwind scheme.

Example 2 Linear system

$$u_t + Au_x = 0$$

Let $R^{-1}AR = \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_n)$. We need to solve Riemann problem with initial data (U_j, U_{j+1}). Let $L = (\ell_1, \cdots, \ell_n) = R^{-1}, \ell_i A = \lambda_i \ell_i, i = 1, \ldots, n$ be the left eigenvectors. Project initial data onto $r_1, \ldots, r_n$

$$u(x, t_n) = \begin{cases} U_j & x < x_{j+\frac{1}{2}} \\ U_{j+1} & x > x_{j+\frac{1}{2}} \end{cases}$$

by $\ell_i r_j = \delta_{ij}$, $\sum(\ell_i u(x, t_n))r_i = u(x, t_n)$.

$$\begin{aligned}
& \ell_i(u_t + Au_x) = 0 \\
\Longrightarrow \quad & \alpha_{it} + \lambda_i \alpha_{ix} = 0
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \quad \alpha_i(x,t) &= \alpha_i(x - \lambda_i(t - t_n), t_n) \\
&= \ell_i u(x - \lambda_i(t - t_n), t_n)
\end{aligned}$$

$$\tilde{u}(x,t) = \sum_i (\ell_i \tilde{u}(x - \lambda_i(t - t_n), t_n))r_i$$

$$\tilde{u}(x_{j+\frac{1}{2}}, t) = \sum_{\lambda_i \geq 0} (\ell_i \tilde{u}(x - \lambda_i(t - t_n), t_n))r_i + \sum_{\lambda_i < 0} (\ell_i \tilde{u}(x - \lambda_i(t - t_n), t_n))r_i$$

$$\tilde{U}^{n+\frac{1}{2}}_{j+\frac{1}{2}} = \sum_{i,\lambda_i \geq 0} \ell_i U_j r_i + \sum_{i,\lambda_i<0} \ell_i U_{j+1} r_i$$

$$\begin{aligned} F_{j+\frac{1}{2}} &= A\tilde{U}^{n+\frac{1}{2}}_{j+\frac{1}{2}} \\ &= \sum_{i,\lambda_i \geq 0} \lambda_i \ell_i U_j r_i + \sum_{i,\lambda_i<0} \lambda_i \ell_i U_{j+1} r_i \end{aligned}$$

solve $\tilde{u}(x,t)$ for $\frac{x}{t} = \lambda$

$$\tilde{u}(x,t) = \sum_{\lambda_i \geq \lambda} \lambda_i \ell_i U_j r_i + \sum_{\lambda_i<\lambda} \lambda_i \ell_i U_{j+1} r_i$$

consider the following cases

(1) $\lambda < \lambda_1 < \cdots < \lambda_n$

$$\tilde{u}(x,t) = \sum_{\lambda_i \geq \lambda} \ell_i U_j r_i = U_j$$

(2) $\lambda_1 < \lambda < \lambda_2 < \cdots < \lambda_n$

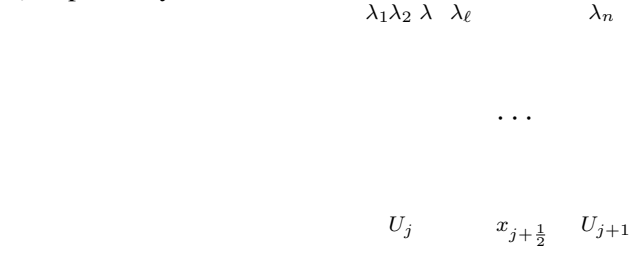
$$\begin{aligned} \tilde{u}(x,t) &= \sum_{i=2}^{n} \ell_i U_j r_i + \ell_1 U_{j+1} r_1 \\ &= \sum_{i=1}^{n} \ell_i U_j r_i + \ell_1 U_{j+1} r_1 - \lambda_1 \ell_1 U_j r_1 \\ &= U_j + \ell_1 (U_{j+1} - U_j) r_1 \end{aligned}$$

There is a jump $\ell_1(U_{j+1} - U_j)r_1$

(3) $\lambda_1 < \lambda_2 < \lambda < \lambda_3 < \cdots < \lambda_n$

$$\tilde{u}(x,t) = U_j + \ell_1 (U_{j+1} - U_j) r_1 + \ell_2 (U_{j+1} - U_j) r_2$$

Therefore the structure of the solution of Riemann problem is composed of n waves $\ell_1(U_{j+1} - U_j)r_1, \cdots, \ell_n(U_{j+1} - U_j)r_n$ with left state U_j and right state U_{j+1}. Each wave propagate at speed λ_i respectively.



7.2.2 piecewise-linear reconstruction

(1) Reconstruction
Given cell average $\{U_j\}$, we want to reconstruct a polynimial $\tilde{u}(x,t_n)$ in each cell $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ under following criterions

a) high order approximation in smooth regions.

b) TVD or TVB or ENO

(2) Riemann solver
solve equation "exactly" for (t_n, t_{n+1}).

Once we have these two, define $U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(\tilde{u}(x_{j-\frac{1}{2}},t)) - f(\tilde{u}(x_{j+\frac{1}{2}},t))\,dt$. For second order temporal discretization,

$$\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(\tilde{u}(x_{j+\frac{1}{2}},t))\,dt \approx f(\tilde{u}(x_{j+\frac{1}{2}},t_{n+\frac{1}{2}})),$$

$$U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}[f(\tilde{u}(x_{j-\frac{1}{2}},t_{n+\frac{1}{2}})) - f(\tilde{u}(x_{j+\frac{1}{2}},t_{n+\frac{1}{2}}))].$$

For Scalar Case

(1) Reconstruction
Suppose $\tilde{u}(x,t_n) = a + b(x-x_j) + c(x-x_j)^2$, want to find a,b,c such that the average of $\tilde{u} = U_j$.

$$\begin{aligned}
\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{u}(x,t_n)\,dx &= U_j \\
\frac{1}{\Delta x}\int_{x_{j-\frac{3}{2}}}^{x_{j-\frac{1}{2}}} \tilde{u}(x,t_n)\,dx &= U_{j-1} \\
\frac{1}{\Delta x}\int_{x_{j+\frac{1}{2}}}^{x_{j+\frac{3}{2}}} \tilde{u}(x,t_n)\,dx &= U_{j+1}
\end{aligned}$$

$$\Longrightarrow a = U_j,\ b = \frac{U_{j+1}-U_{j-1}}{2\Delta x},\ c = 0$$

Lemma 7.1. *Given a smooth function $u(x)$, let $U_j = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\,dx$, and let $\tilde{u}(x) = U_j + \delta U_j \frac{x-x_j}{\Delta x}$ $\delta U_j = (U_{j+1}-U_{j-1})/2$, then $|\tilde{u}(x)-u(x)| = O(\Delta x)^3$ for $x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$*

When u has discontinuities or u_x changes sign, we need to put a "limiter" to avoid oscillation of $\tilde{u}$.

Example of limiters

(a)

$$\begin{aligned}\delta U_j &= \text{minmod}(U_{j+1}-U_j, U_j-U_{j-1}) \\ &= \begin{cases} \text{sign}(U_{j+1}-U_j)\min\{|U_{j+1}-U_j|, |U_j-U_{j-1}|\} & \text{if } U_{j+1}-U_j \text{ and } U_j-U_{j-1} \text{ have the same sign} \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

(b) $\delta U_j = \text{minmod}(\frac{U_{j+1}-U_{j-1}}{2}, 2(U_j-U_{j-1}), 2(U_{j+1}-U_j))$

(2) Exact solver for small time step
Consider the linear advection equation

$$u_t + au_x = 0.$$

with precise linear data

$$\tilde{u}(x,t_n) = \begin{cases} U_j + \delta U_j \frac{x-x_j}{\Delta x} & x < x_{j+\frac{1}{2}} \\ U_{j+1} + \delta U_{j+1}\frac{x-x_{j+1}}{\Delta x} & x > x_{j+\frac{1}{2}} \end{cases}$$

Then

$$\begin{aligned}\tilde{u}_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= \tilde{u}(x_{j+\frac{1}{2}} - a(t-t_n), t_n) \quad (a>0) \\ &= U_j + \delta U_j(x_{j+\frac{1}{2}} - a(t_{n+\frac{1}{2}}-t_n) - x_j)/\Delta x \\ &= U_j + \delta U_j(\frac{1}{2} - \frac{a\Delta t}{2\Delta x}) \quad \text{let } \nu = \frac{a\Delta t}{\Delta x} \\ F_{j+\frac{1}{2}} &= a\tilde{U}_{j+\frac{1}{2}}^{n+\frac{1}{2}} = a(U_j + \delta U_j(\frac{1}{2} - \frac{\nu}{2}))\end{aligned}$$

To compare with the TVD scheme, let $\delta U_j = \text{minmod}(U_{j+1}-U_j, U_j-U_{j-1})$

$$\begin{aligned}F_{j+\frac{1}{2}} &= aU_j + (\frac{1}{2} - \frac{\nu}{2})a(U_{j+1}-U_j)\cdot\phi_{j+\frac{1}{2}} \\ \phi_{j+\frac{1}{2}} &= \frac{\text{minmod}(U_{j+1}-U_j, U_j-U_{j-1})}{U_{j+1}-U_j}\end{aligned}$$

$$\phi(\theta) = \begin{cases} 0 & \theta \le 0 \\ \theta & 0 \le \theta \le 1 \\ 1 & \theta \ge 1 \end{cases} \qquad \theta = \frac{U_j - U_{j-1}}{U_{j+1}-U_j}$$

Its graph is shown in Fig.(7.3).

If $a < 0$, then

$$\begin{aligned}\tilde{u}_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= U_{j+1} + \delta U_{j+1}(-\frac{1}{2} - \frac{a\Delta t}{2\Delta x}) \quad |\frac{a\Delta t}{\Delta x}| \le 1 \\ F_{j+\frac{1}{2}} &= a(U_{j+1} + \delta U_{j+1}(-\frac{1}{2} - \frac{\nu}{2}))\end{aligned}$$

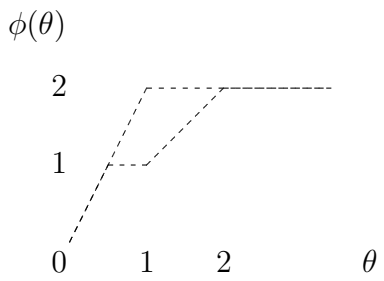


Figure 7.3: The limiter of second order Godunov method

For System Case

$$u_t + Au_x = 0 \tag{7.5}$$

(1) Reconstruction
Construct $\tilde{u}(x, t_n)$ to be a piecewise linear function.

$$\tilde{u}(x, t_n) = U_j^n + \delta U_j^n(\frac{x - x_j}{\Delta x})$$

The slope is found by $\delta U_j^n = \text{minmod}(U_j - U_{j-1}, U_{j+1} - U_j)$. We can write it characteristic-wisely: let

$$\begin{aligned} \alpha_{j,k}^L &= \ell_k(U_j - U_{j-1}), \\ \alpha_{j,k}^R &= \ell_k(U_{j+1} - U_j), \\ \alpha_{j,k} &= \text{minmod}(\alpha_{j,k}^L, \alpha_{j,k}^R). \end{aligned}$$

Then $\delta U_j = \sum \alpha_{j,k} r_k$.

(2) Exactly solver
We trace back along the characteristic curve to get u in half time step.

$$\begin{aligned} u_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= \sum_k \ell_k \tilde{u}(x_{j+\frac{1}{2}} - \lambda_k(t_{n+\frac{1}{2}} - t_n), t_n) r_k \\ &= \sum_{\lambda_k \geq 0} \ell_k(U_j + \delta U_j(\frac{1}{2} - \frac{\nu_k}{2}))r_k + \sum_{\lambda_k < 0} \ell_k(U_{j+1} + \delta U_{j+1}(-\frac{1}{2} - \frac{\nu_k}{2}))r_k \\ &= \text{initial state of Riemann data } (U_j, U_{j+1}) + \\ &\quad \sum_{\lambda_k \geq 0} (\ell_k(\frac{1}{2} - \frac{\nu_k}{2})r_k)\delta U_j + \sum_{\lambda_k < 0} (\ell_k(-\frac{1}{2}) - \frac{\nu_k}{2})r_k)\delta U_{j+1}. \end{aligned}$$

In another viewpoint, let $u_{j+\frac{1}{2},L}^{n+\frac{1}{2}}$ be the solution of (7.5) with initial data = $\begin{cases} \tilde{u}(x, t_n) & x \in (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \\ 0 & \text{otherwise} \end{cases}$.

$$u_{j+\frac{1}{2},L}^{n+\frac{1}{2}} = u_j^n + \sum_{\lambda_k \geq 0} \ell_k \delta U_j^n (\frac{x_{j+\frac{1}{2}} - \lambda_k \frac{\Delta t}{2} - x_j}{\Delta x}) r_k$$

$$= \quad u_j^n + \sum_{\lambda_k \geq 0} \ell_k \delta U_j^n (\frac{1}{2} - \frac{\nu_k}{2}) r_k$$

where ℓ_k, r_k are left / right eigenvector, λ_k is eigenvalue and $\nu_k = \frac{\lambda_k \Delta t}{\Delta x}$.
Similarly,

$$\begin{aligned} u_{j+\frac{1}{2},R}^{n+\frac{1}{2}} &= u_{j+1}^n - \sum_{\lambda_k < 0} \ell_k \delta U_{j+1}^n (\frac{x_{j+\frac{1}{2}} - \lambda_k \frac{\Delta t}{2} - x_{j+1}}{\Delta x}) r_k \\ &= u_{j+1}^n - \sum_{\lambda_k < 0} \ell_k \delta U_{j+1}^n (-\frac{1}{2} - \frac{\nu_k}{2}) r_k \end{aligned}$$

Then we solve (7.5) with$(u_{j+\frac{1}{2},L}^{n+\frac{1}{2}}, u_{j+\frac{1}{2},R}^{n+\frac{1}{2}})$ as the Riemann data. This gives $u_{j+\frac{1}{2}}^{n+\frac{1}{2}}$. Therefore

$$\begin{aligned} u_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= u_{j+\frac{1}{2},L}^{n+\frac{1}{2}} + \sum_{\lambda_k \geq 0} \ell_k \delta U_{j+\frac{1}{2}} (-\frac{\lambda_k \frac{\Delta t}{2}}{\Delta x}) r_k \\ &= u_{j+\frac{1}{2},L}^{n+\frac{1}{2}} + \sum_{\lambda_k \geq 0} \ell_k \delta U_{j+\frac{1}{2}} (-\frac{\nu_k}{2}) r_k \\ \text{or } u_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= u_{j+\frac{1}{2},R}^{n+\frac{1}{2}} - \sum_{\lambda_k \leq 0} \ell_k \delta U_{j+\frac{1}{2}} (-\frac{\nu_k}{2}) r_k \\ \text{or } u_{j+\frac{1}{2}}^{n+\frac{1}{2}} &= \frac{U_{j+\frac{1}{2},L}^{n+\frac{1}{2}} + U_{j+\frac{1}{2},R}^{n+\frac{1}{2}}}{2} - \frac{1}{2} \sum \text{sign}(\nu_k) \ell_k \delta U_{j+\frac{1}{2}} \frac{\nu_k}{2} r_k \end{aligned}$$

where $\delta U_{j+\frac{1}{2}} = U_{j+\frac{1}{2},R}^{n+\frac{1}{2}} - U_{j+\frac{1}{2},L}^{n+\frac{1}{2}}$.

(3) $U_j^{n+1} = U_j^n + \frac{\Delta t}{\Delta x}(f(U_{j-\frac{1}{2}}^{n+\frac{1}{2}}) - f(U_{j+\frac{1}{2}}^{n+\frac{1}{2}}))$.

7.3 Multidimension

There are two kinds of methods.

1. Splitting method.

2. Unsplitting method.

We consider two-dimensional case.

7.3.1 Splitting Method

We start from

$$u_t + Au_x + Bu_y = 0. \tag{7.6}$$

This equation can be viewed as

$$u_t = (-A\partial_x - B\partial_y)u.$$

Then the solution operator is:

$$e^{-t(A\partial_x + B\partial_y)},$$

which can be approximate by $e^{-tA\partial_x}e^{-tB\partial_y}$ for small t. Let $\mathcal{A} = -A\partial_x, \mathcal{B} = -B\partial_y$, we have

$$u = e^{t(\mathcal{A}+\mathcal{B})}u_0.$$

Consider $e^{t(\mathcal{A}+\mathcal{B})}$,

$$\begin{aligned}
e^{t(\mathcal{A}+\mathcal{B})} &= 1 + t(\mathcal{A}+\mathcal{B}) + \frac{t^2}{2}(\mathcal{A}^2 + \mathcal{B}^2 + \mathcal{A}\mathcal{B} + \mathcal{B}\mathcal{A}) + \cdots \\
e^{t\mathcal{B}} \cdot e^{t\mathcal{A}} &= (1 + t\mathcal{B} + \frac{t^2}{2}\mathcal{B}^2 + \cdots)(1 + t\mathcal{A} + \frac{t^2}{2}\mathcal{A}^2 + \cdots) \\
&= 1 + t(\mathcal{A}+\mathcal{B}) + \frac{t^2}{2}(\mathcal{A}^2 + \mathcal{B}^2) + t^2\mathcal{B}\mathcal{A} + \cdots \\
&\therefore e^{t(\mathcal{A}+\mathcal{B})} - e^{t\mathcal{B}} \cdot e^{t\mathcal{A}} = t^2(\frac{\mathcal{A}\mathcal{B} - \mathcal{B}\mathcal{A}}{2}) + \mathcal{O}(t^3).
\end{aligned}$$

Now we can design splitting method as:
Given $\{U^n_{i,j}\}$,

1. For each j, solve $u_t + Au_x = 0$ with data $\{U^n_j\}$ for Δt step. This gives $\bar{U}^n_{i,j}$.

$$\bar{U}^n_{i,j} = U^n_{i,j} + \frac{\Delta t}{\Delta x}(F(U^n_{i-1,j}, U^n_{i,j}) - F(U^n_{i,j}, U^n_{i+1,j}))$$

   where $F(U, V)$ is the numerical flux for $u_t + Au_x = 0$.

2. For each i, solve $u_t + Bu_y = 0$ for Δt step with data $\{\bar{U}^n_{i,j}\}$. This gives $U^{n+1}_{i,j}$.

$$U^{n+1}_{i,j} = \bar{U}^n_{i,j} + \frac{\Delta t}{\Delta y}(G(\bar{U}^n_{i,j-1}, \bar{U}^n_{i,j}) - G(\bar{U}^n_{i,j}, \bar{U}^n_{i,j+1}))$$

The error is first order in time $n(\Delta t)^2 = O(\Delta t)$.

To reach higher order time splitting, we may approximate $e^{t(\mathcal{A}+\mathcal{B})}$ by polynomials $P(e^{t\mathcal{A}}, e^{t\mathcal{B}})$ or rationals $R(e^{t\mathcal{A}}, e^{t\mathcal{B}})$. For example, the Trotter product (or strang splitting) is given by

$$e^{t(\mathcal{A}+\mathcal{B})} = e^{\frac{1}{2}t\mathcal{A}}e^{t\mathcal{B}}e^{\frac{1}{2}t\mathcal{A}} + O(t^3).$$

For $t = n\Delta t$,

$$\begin{aligned}
e^{t(\mathcal{A}+\mathcal{B})}u_0 &= (e^{\frac{1}{2}\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\frac{1}{2}\Delta t\mathcal{A}}) \cdots (e^{\frac{1}{2}\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\frac{1}{2}\Delta t\mathcal{A}})(e^{\frac{1}{2}\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\frac{1}{2}\Delta t\mathcal{A}})u_0 \\
&= e^{\frac{1}{2}\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\Delta t\mathcal{A}} \cdots e^{\Delta t\mathcal{A}}e^{\Delta t\mathcal{B}}e^{\frac{1}{2}\Delta t\mathcal{A}}u_0
\end{aligned}$$

Trotter product is second order.

7.3.2 Unsplitting Methods

The PDE is

$$u_t + f(u)_x + g(u)_y = 0 \tag{7.7}$$

Integrate this equation over $(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}) \times (t_n, t_{n+1})$. We have

$$U_{i,j}^{n+1} = U_{i,j}^n + \frac{\Delta t}{\Delta x}(\bar{f}_{i-\frac{1}{2},j}^{n+\frac{1}{2}} - \bar{f}_{i+\frac{1}{2},j}^{n+\frac{1}{2}}) + \frac{\Delta t}{\Delta y}(\bar{g}_{i,j-\frac{1}{2}}^{n+\frac{1}{2}} - \bar{g}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}})$$

where

$$\begin{aligned} \bar{f}_{i+\frac{1}{2},j}^{n+\frac{1}{2}} &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{i+\frac{1}{2}}, y_j, t))\, dt \\ \bar{g}_{i,j+\frac{1}{2}}^{n+\frac{1}{2}} &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} g(u(x_i, y_{j+\frac{1}{2}}, t))\, dt. \end{aligned}$$

Looking for numerical approximations $F(U_{i,j+k}^n, U_{i+1,j+k}^n), G(U_{i+\ell,j}^n, U_{i+\ell,j+1}^n)$ for $\bar{f}_{i+\frac{1}{2},j+k}^{n+\frac{1}{2}}, \bar{g}_{i+\ell,j+\frac{1}{2}}^{n+\frac{1}{2}}$. We consider Godunov type method.

1. Reconstruction

   $$\tilde{u}(x, y, t_n) = u_{i,j}^n + \delta_x U_{i,j}(\frac{x - x_i}{\Delta x}) + \delta_y U_{i,j}(\frac{y - y_j}{\Delta y}) \quad \text{in } I = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$$

   For example, $\delta_x U_{i,j} = \text{minmod}(U_{i,j} - U_{i+1,j}, U_{i+1,j} - U_{i,j})$.

2. We need to solve

   $$u_t + Au_x + Bu_y = 0 \text{ with data } \begin{cases} \tilde{u}(x, y, t_n) & \text{for } (x, y) \in I \\ 0 & \text{otherwise} \end{cases}$$

   $$\begin{aligned} \tilde{u}(x_{j+\frac{1}{2}}, y_j, \frac{\Delta t}{2}) &= U_{i,j}^n + \sum_{a>0} \delta_x U_{i,j}(\frac{x_{i+\frac{1}{2}} - \frac{a\Delta t}{2} - x_i}{\Delta x}) + \delta_y U_{i,j}(\frac{y_j - \frac{b\Delta t}{2} - y_j}{\Delta y}) \\ &= U_{i,j}^n + \sum_{a>0} (\delta_x U_{i,j}^n) \cdot (\frac{1}{2} - \frac{\nu_x}{2}) + (\delta_y U_{i,j}^n)(-\frac{\nu_y}{2}) \end{aligned}$$

   where $\nu_x = \frac{a\Delta t}{\Delta x}, \nu_y = \frac{b\Delta t}{\Delta y}$. For system case, λ_k^x, λ_k^y are eigenvalues of A and B.

   $$U_{i+\frac{1}{2},L,j}^{n+\frac{1}{2}} = U_{i,j}^n + \sum_{\lambda^x \geq 0} (\frac{1}{2} - \frac{\nu_k^x}{2})(\ell_k^x \cdot \delta_x U_{i,j}) r_k^x + \sum_k (-\frac{\nu_k^y}{2})(\ell_k^y \cdot \delta_y U_{i,j}) r_k^y$$

   similarly,

   $$U_{i+\frac{1}{2},R,j}^{n+\frac{1}{2}} = U_{i+1,j}^n + \sum_{\lambda_k^x < 0} (-\frac{1}{2} - \frac{\nu_k^x}{2})(\ell_k^x \cdot \delta_x U_{i,j}) r_k^x + \sum_k (-\frac{\nu_k^y}{2})(\ell_k^y \cdot \delta_y U_{i+1,j}) r_k^y$$

Finally, solve Riemann problem $u_t + Au_x = 0$ with data $\begin{cases} U^{n+\frac{1}{2}}_{i+\frac{1}{2},L,j} \\ U^{n+\frac{1}{2}}_{i+\frac{1}{2},R,j} \end{cases}$

$$\therefore U^{n+\frac{1}{2}}_{i+\frac{1}{2},j} = U^{n+\frac{1}{2}}_{i+\frac{1}{2},L,j} + \sum_{\lambda^x_k \geq 0} \ell_k \cdot \delta U_{i+\frac{1}{2},j} r_k$$

Chapter 8

Systems of Hyperbolic Conservation Laws

8.1 General Theory

We consider

$$u_t + f(u)_x = 0,\ u = \begin{pmatrix} u_1 \\ u_2 \\ \ddots \\ u_n \end{pmatrix}\ f : \mathbb{R}^n \to \mathbb{R}^n \text{the flux} \tag{8.1}$$

The system (8.1) is called *hyperbolic* if $\forall u$, the $n \times n$ matrix $f'(u)$ is diagonalizable with real eigenvalues $\lambda_1(u) \leq \lambda_2(u) \leq \cdots \leq \lambda_n(u)$. Let us denote its left/right eigenvectors by $\ell_i(u)/r_i(u)$, respectively.
It is important to notice that the system is Galilean invariant, that is , the equation is unchanged under the transform:

$$t \longrightarrow \lambda t, \quad x \longrightarrow \lambda x, \quad \forall \lambda > 0.$$

This suggests we can look for special solution of the form $u(\frac{x}{t})$.

We plug $u(\frac{x}{t})$ into (8.1) to yield

$$u' \cdot (-\frac{x}{t^2}) + f'(u)u' \cdot \frac{1}{t} = 0$$

$$\Longrightarrow \quad f'(u)u' = \frac{x}{t}u'$$

This implies that there exists i such that $u' = r_i(u)$ and $\frac{x}{t} = \lambda_i(u(\frac{x}{t}))$. To find such a solution, we first construct the integral curve of $r_i(u)$: $u' = r_i(u)$. Let $R_i(u_0, s)$ be the integral curve of $r_i(u)$ passing through u_0, and parameterized by its arclength. Along R_i, the speed λ_i has the variation:

$$\frac{d}{ds}\lambda_i(R_i(u_0, s)) = \nabla\lambda_i \cdot R_i' = \nabla\lambda_i \cdot r_i.$$

We have the following definition.

Definition 8.1. The i-th characteristic field is called

1. genuinely nonlinear if $\nabla\lambda_i(u) \cdot r_i(u) \neq 0 \forall u$.
2. linearly degenerate if $\nabla\lambda_i(u) \cdot r_i(u) \equiv 0$
3. nongenuinely nonlinear if $\nabla\lambda_i(u) \cdot r_i(u) = 0$ on isolated hypersurface in $\mathbb{R}^n$.

For scalar equation, the genuine nonlinearity is equivalent to the convexity(or concavity) of the flux f, linear degeneracy is $f(u) = au$, while nongenuine nonlinearity is nonconvexity of f.

8.1.1 Rarefaction Waves

When the i-th field is genuiely nonlinear, we define

$$R_i^+(u_0) = \{u \in R_i(u_0) | \lambda_i(u) \geq \lambda_i(u_0)\}.$$

Now suppose $u_1 \in R_i^+(u_0)$, we construct the centered rarefaction wave, denoted by (u_0, u_1):

$$(u_0, u_1)(\frac{x}{t}) = \begin{cases} u_0 & \text{if } \frac{x}{t} \leq \lambda_i(u_0) \\ u_1 & \text{if } \frac{x}{t} \geq \lambda_i(u_1) \\ u & \text{if } \lambda_i(u_0) \leq \frac{x}{t} \leq \lambda_i(u_1) \text{and} \lambda_i(u) = \frac{x}{t} \end{cases}$$

It is easy to check this is a solution. We call (u_0, u_1) an *i-rarefaction wave*.

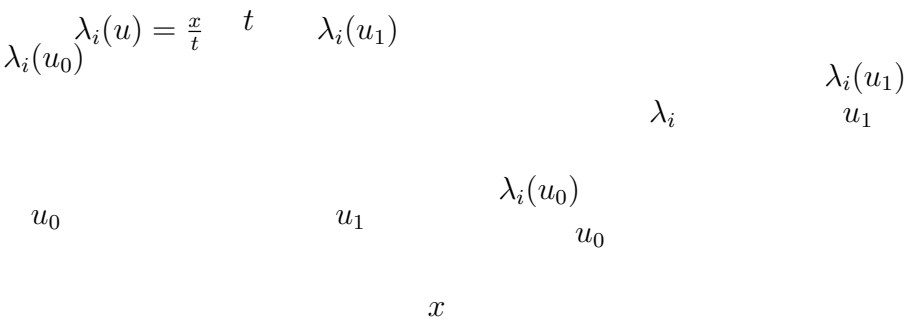


Figure 8.1: The integral curve of $u' = r_i(u)$ and the rarefaction wave.

8.1.2 Shock Waves

The shock wave is expressed by:

$$u(\frac{x}{t}) = \begin{cases} u_0 & \text{for } \frac{x}{t} < \sigma \\ u_1 & \text{for } \frac{x}{t} > \sigma \end{cases}$$

Then (u_0, u_1, σ) need to satisfy the jump condition:

$$f(u_1) - f(u_0) = \sigma(u_1 - u_0). \tag{8.2}$$

Lemma 8.1. *(Local structure of shock waves)*

1. *The solution of (8.2) for (u, σ) consists of n algebraic curves passing through u_0 locally, named them by $S_i(u_0), i = 1, \cdots, n$.*

2. *$S_i(u_0)$ is tangent to $R_i(u_0)$ up to second order. i.e., $S_i^{(k)}(u_0) = R_i^{(k)}(u_0), k = 0, 1, 2$, here the derivatives are arclength derivatives.*

3. *$\sigma_i(u_0, u) \to \lambda_i(u_0)$ as $u \to u_0$, and $\sigma_i'(u_0, u_0) = \frac{1}{2}\lambda_i'(u_0)$*

Proof. 1. Let $S(u_0) = \{u | f(u) - f(u_0) = \sigma(u - u_0)$ for some $\sigma \in \mathbb{R}\}$. We claim that $S(u_0) = \bigcup_{i=1}^{n} S_i(u_0)$, where $S_i(u_0)$ is a smooth curve passing through u_0 with tangent $r_i(u_0)$ at u_0. When u is on $S(u_0)$, rewrite the jump condition as

$$\begin{aligned} f(u) - f(u_0) &= \left[\int_0^1 f'(u_0 + t(u - u_0)\, dt\right](u - u_0) \\ &= \tilde{A}(u_0, u)(u - u_0) \\ &= \sigma(u - u_0) \end{aligned}$$

$$\therefore \quad u \in S(u_0) \Longleftrightarrow (u - u_0) \text{ is an eigenvector of } \tilde{A}(u_0, u).$$

Assume $A(u) = f'(u)$ has real and distinct eigenvalues $\lambda_1(u) < \cdots \lambda_n(u)$, $\tilde{A}(u_0, u)$ also has real and distinct eigenvalues $\tilde{\lambda}_1(u_0, u) < \cdots < \tilde{\lambda}_n(u_0, u)$, with left/right eigenvectors $\tilde{\ell}_i(u_0, u)$ and $\tilde{r}_i(u_0, u)$, respectively, and they converge to $\lambda_i(u_0), \ell_i(u_0), r_i(u_0)$ as $u \to u_0$ respectively. Normalize the eigenvectors: $\|\tilde{r}_i\| = 1, \tilde{\ell}_i \tilde{r}_j = \delta_{ij}$. The vector which is parallel to r_i can be determined by

$$\tilde{\ell}_k(u_0, u)(u - u_0) = 0 \text{ for } k \neq i, k = 1, \cdots, n.$$

Now we define

$$S_i(u_0) = \{u | \tilde{\ell}_k(u_0, u)(u - u_0) = 0, k \neq i, k = 1, \cdots, n\}$$

We claim this is a smooth curve passing through u_0. Choose coordinate system $r_1(u_0), \cdots, r_n(u_0)$. Differentiate this equation $\tilde{\ell}_k(u_0, u)(u - u_0) = 0$ at $u = u_0$ in $r_j(u_0)$ direction:

$$\left.\frac{\partial}{\partial r_j}\right|_{u=u_0} (\tilde{\ell}_k(u_0, u)(u - u_0)) = \tilde{\ell}_k.(u_0, u_0) \cdot r_j(u_0) = \delta_{jk},$$

Thus, this is the Jacobian matrix of the map: $\tilde{\ell}_k(u_0, u)(u - u_0)$ at u_0. It is an $(n - 1) \times n$ full rank matrix. By the implicit function theorem, the set $S_i(u_0)$ is a smooth curve passing through u_0.

2,3. $R_i(u_0) = u_0 = S_i(u_0)$

$$f(u) - f(u_0) = \sigma_i(u_0, u)(u - u_0) \quad \forall u \in S_i(u_0)$$

Take arclength derivative along $S_i(u_0)$

$$f'(u)u' = \sigma_i'(u - u_0) + \sigma_i u' \text{ and } u' = S_i'.$$

When $u \to u_0$

$$f'(u_0)S_i'(u_0) = \sigma_i(u_0, u_0)S_i'(u_0)$$
$$\implies \quad S_i'(u_0) = r_i(u_0) \text{ and } \sigma_i(u_0, u_0) = \lambda_i(u_0).$$

Consider the second derivative.

$$(f''(u)u', u') + f'(u)u'' = \sigma_i''(u - u_0) + 2\sigma_i' \cdot u' + \sigma_i u''$$

At $u = u_0$, $u' = S_i'(u_0) = R_i'(u_0) = r_i(u_0)$ and $u'' = S_i''(u_0)$,

$$\implies \quad (f''r_i, r_i) + f'S_i'' = 2\sigma_i' r_i + \sigma_i S_i''$$

On the other hand, we take derivative of $f'(u)r_i(u) = \lambda_i(u)r_i(u)$ along $R_i(u_0)$, then evaluate at $u = u_0$.

$$(f''r_i, r_i) + f'(\nabla r_i \cdot r_i) = \lambda_i' r_i + \lambda_i \nabla r_i \cdot r_i,$$

where $\nabla r_i \cdot r_i = R_i''$.

$$\implies \quad (f' - \lambda_i)(S_i'' - R_i'') = (2\sigma_i' - \lambda_i')r_i$$

Taking inner product with ℓ_i leads to

$$2\sigma_i' = \lambda_i'.$$

Let $S_i'' - R_i'' = \sum_k \alpha_k r_k(u_0)$. Taking inner product with ℓ_k leads to

$$\sum_{k \neq i} (\lambda_k - \lambda_i)\alpha_k r_k = 0 \implies \alpha_k = 0 \; \forall k \neq i$$

On the other hand, from $(R_i', R_i') = 1$ and $(S_i', S_i') = 1$, we get $(R_i'', R_i') = 0$ and $(S_i'', S_i') = 0$. Since $R_i' = S_i' = r_i$, we then get

$$(S_i'' - R_i'', r_i) = 0.$$

Hence $S_i'' = R_i''$ at u_0. Hence $R_i'' = S_i''$ at u_0.

□

Suppose the i-th characteristic field is genuinely nonlinear. The Lax entropy condition reads

$$\lambda_i(u_0) > \sigma_i(u_0, u_1) > \lambda_i(u_1) \tag{8.3}$$

Let us define $S_i^-(u_0)$ to be the branch of $S_i(u_0)$ which satisfies entropy condition:

$$S_i(u_0) := \{u \in S_i(u_0) | \lambda_i(u) < \lambda_i(u_0)\}$$

Then for $u_1 \in S_i^-(u_0)$, and $u_1 \sim u_0$, (8.3) is always valid. This follows easily from $\lambda_i = 2\sigma_i'$ and $\sigma_i(u_0, u_0) = \lambda_i(u_0)$. For $u_1 \in S_i^-(u_0)$, we call the solution (u_0, u_1) an *i-shock* or *Lax-shock*.

8.1.3 Contact Discontinuity (Linear Wave)

If $\nabla\lambda_i(u)\cdot r_i(u)\equiv 0$, we call the i-th characteristic field *linearly degenerate (ℓ. dg.)*. In the case of scalar equation, this correspond $f''=0$. We claim

$$R_i(u_0)=S_i(u_0) \text{ and } \sigma_i(u_0,u)=\lambda_i(u_0) \text{ for } u\in S_i(u_0) \text{ or } R_i(u_0).$$

Indeed, along $R_i(u_0)$, we have

$$f'(u)u'=\lambda_i(u)u'.$$

and $\lambda_i(u)$ is a constant $\lambda_i(u_0)$ from the linear degeneracy. We integrate the above equation from u_0 to u along $R_i(u_0)$, we get

$$f(u)-f(u_0)=\lambda_i(u_0)(u-u_0).$$

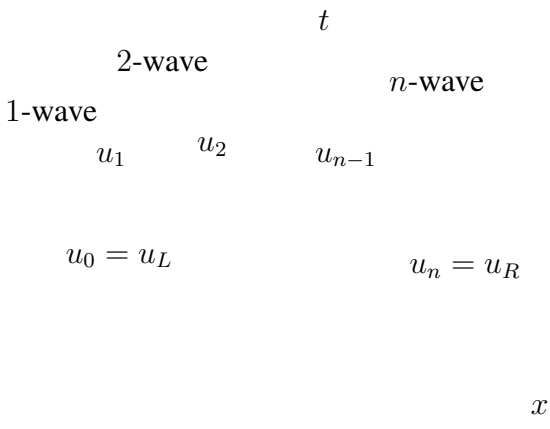
This gives the shock condition. Thus, $S_i(u_0)\equiv R_i(u_0)$ and $\sigma(u,u_0)\equiv\lambda_i(u_0)$.

Homeworks.

$$(u_0,u_1)=\begin{cases} u_0 & \frac{x}{t}<\sigma_i(u_0,u_1)\\ u_1 & \frac{x}{t}>\sigma_i(u_0,u_1)\end{cases}$$

Let $T_i(u_0)=R_i^+(u_0)\cup S_i^-(u_0)$ be called the *i-th wave curve*. For $u_1\in T_i(u_0)$, (u_0,u_1) is either a rarefaction wave, a shock, or a contact discontinuity.

Theorem 8.1. *(Lax) For strictly hyperbolic system (8.1), if each field is either genuinely nonlinear or linear degenerate, then for $u_L\sim u_R$, the Riemann problem with two end states (u_L,u_R) has unique self-similar solution which consists of n elementary waves. Namely, there exist $u_0=u_L,\cdots,u_n=u_R$ such that (u_{i-1},u_i) is an i-wave.*



Proof. Given $(\alpha_1,\cdots,\alpha_n)\in\mathbb{R}^n$, we define u_i successively as the follows. First we define $u_0=u_L$. Then we follow T_1 curve from u_0 with length α_1. This gives $u_1\in T_1(u_0)$ and (u_0,u_1) forms a 1-wave with strength α_1 (measured by the arc length α_1 on $T_1(u_0)$. From u_1, we follow $T_2(u_1)$ with length α_2 to u_2. This gives (u_1,u_2) a 2-wave with strength α_2. We continue this process until

$u_n := f(u_L, \alpha_1, ..., \alpha_n)$. This gives a map from strengths $(\alpha_1, ..., \alpha_n)$ to the final state u_n with $f(u_L, 0, ..., 0) = u_L$. The mapping is C^2 because the curves $T_i \in C^2$. Now, we are given the final state u_R. We solve the inverse problem

$$u_R = f(u_L, \alpha_1, ..., \alpha_n).$$

This mapping is locally invertible because the Jacobian

$$\frac{\partial f}{\partial \alpha_k}(u_L, 0, ..., 0) = r_k(u_L), \quad k = 1, ..., n$$

is invertible at u_L. By the inverse function theorem, when $u_R \sim u_L$, there exists a unique $\alpha_1, ..., \alpha_n)$ such that $u_R = f(u_L, \alpha_1, ..., \alpha_n)$. □

8.2 Physical Examples

8.2.1 Gas dynamics

The equations of gas dynamics are derived based on conservation of mass, momentum and energy. Before we derive these equations, let us review some thermodynamics. First, the basic thermo variables are pressure (p), specific volume (τ), called state variables. The internal energy (e) is a function of p and τ. Such a relation is called a constitutive equation. The basic assumption are

$$\left.\frac{\partial e}{\partial p}\right|_\tau > 0, \ \left.\frac{\partial e}{\partial \tau}\right|_p > 0$$

Sometimes, it is convinient to express p as a function of (τ, e).

In an adiabetic process (no heat enters or losses), the first law of thermodynamics (conservation of energy) reads

$$de + pd\tau = 0. \tag{8.4}$$

This is called a Pfaffian equation mathematically. A function $\sigma(e, \tau)$ is called an integral of (8.4) if there exists a function $\mu(e, \tau)$ such that

$$d\sigma = \mu \cdot (de + pd\tau).$$

Thus, $\sigma =$ constant represents a specific adiabetic process. For Pfaffian equation with only two independent variables, one can always find its integral. First, one can derive equation for μ: from

$$\sigma_e = \mu \text{ and } \sigma_\tau = \mu p$$

and using $\sigma_{e\tau} = \sigma_{\tau e}$, we obtain the equation for μ:

$$\mu_\tau = (\mu p)_e.$$

This is a linear first-order equation for μ. It can be solved by the method of characteristics in the region $\tau > 0$ and $e > 0$. The solutions of μ and σ are not unique. If σ is a solution, so

does $\bar{\sigma}$ with $d\bar{\sigma} = \nu(\sigma)d\sigma$ for any function $\nu(\sigma)$. We can choose μ such that if two systems are in thermo-equilibrium, then they have the same value μ. In other words, μ is only a function of emperical temperature. We shall denote it by $1/T$. Such T is called the absolute temperature. The corresponding σ is called the physical entropy S. The relation $d\sigma = \mu(de + pd\tau)$ is re-expressed as

$$de = TdS - pd\tau. \tag{8.5}$$

For ideal gas, which satisfies the laws of Boyle and Gay-Lussac:

$$p\tau = RT, \tag{8.6}$$

where R is the universal gas constant. From this and (8.5), treating S and τ as independent variables, one obtains

$$Re_S(S,\tau) + \tau e_\tau(S,\tau) = 0.$$

We can solve this linear first-order equation by the method of characteristics. We rewrite this equation as a directional differentiation:

$$\left(R\frac{\partial}{\partial S} + \tau\frac{\partial}{\partial \tau}\right) e = 0.$$

This means that e is constant along the characteristic curves

$$R\frac{d\tau}{dS} = \tau.$$

These characteristics can be integrated as

$$\tau e^{-S/R} = \phi.$$

Here ϕ is a positive constant. The energy $e(\tau, S)$ is constant when $\tau e^{-S/R}$ is a constant. That is, $e = h(\phi)$ for some function h. We notice that $h' < 0$ because $p = -(\frac{\partial e}{\partial \tau})_S = -e^{-S/R}h'(\tau H) > 0$. From $T = (\frac{\partial e}{\partial S})_\tau = -\frac{1}{R}h'(\phi)\cdot\phi$, we see that T is a function of ϕ. In most cases, T is a decreasing function of ϕ. We shall make this as an assumption. With this, we can invert the relation between T and ϕ and treat ϕ as a decreasing function of T. Thus, we can also view e as a function of T, say $e(T)$, and $e(T)$ is now an increasing function. Now, we have five thermo variables p, τ, e, S, T, and three relations:

$$\begin{aligned} p\tau &= RT \\ e &= e(T) \\ de &= TdS - pd\tau \end{aligned}$$

Hence, we can choose two of as independent thermo variables and treat the rest three as dependent variables.

For instance, e is a linear function of T, i.e. $e = c_v T$, where c_v is a constant called specfic heat at constant volume. Such a gas is called polytropic gas. We can obtain

$$p\tau = RT \text{ and } e = c_v T = \frac{p\tau}{\gamma - 1} \tag{8.7}$$

or in terms of entropy,

$$\begin{aligned} p &= A(S)\tau^{-\gamma} \\ T &= \frac{A(S)}{R}\tau^{-\gamma+1} \\ e &= \frac{c_v A(S)}{R}\tau^{-\gamma+1} \end{aligned}$$

where

$$\begin{aligned} A(S) &= (\gamma-1)\exp((S-S_0)/c_v) \\ \gamma &= 1 + R/c_v \end{aligned}$$

If we define $dQ = TdS$, it is easy to see that c_v and c_p are the specific heat at constant volume and constant pressure, respectively.

$$\begin{aligned} c_v &= \left(\frac{\partial Q}{\partial T}\right)_\tau = \left(\frac{\partial e}{\partial T}\right)_\tau, \\ c_p &:= \left(\frac{\partial Q}{\partial T}\right)_p = ((\frac{\partial e}{\partial \tau})_p + p)/(\frac{\partial T}{\partial \tau})_p \\ &= \left(\frac{\partial e}{\partial T}\right)_p + p\left(\frac{\partial \tau}{\partial T}\right)_p \end{aligned}$$

In general, $c_p > c_v$. Because c_p is the amount of heat added to a system per unit mass at constant pressure. In order to maintain constant pressure, the volume has to expand (otherwise, pressure will increase), the extra amount of work due to expansion is supplied by the extra amount of heat $c_p - c_v$.

Next, we derive the equation of gas dynamics. Let us consider an arbitrary domain $\Omega \subset R^3$. The mass flux from outside to inside per unit time per unit area dS is $-\rho v\cdot$, where n is the outer normal of Ω. Thus, the conservation of mass can be read as

$$\begin{aligned} \frac{d}{dt}\int_\Omega \rho\, dx &= \int_{\partial\Omega}[-\rho v\cdot n]dS \\ &= -\int_\Omega \text{div}\,(\rho\, v)\, dx \end{aligned}$$

This holds for arbitrary Ω, hence we have

$$\rho_t + \text{ div}(\rho\, v) = 0. \tag{8.8}$$

This is called the continuity equation.

Now, we derive momentum equation. Let us suppose the only surface force is from pressure (no viscous force). Then the momentum change in Ω is due to (i) the momentum carried in through

boundary, (ii) the pressure force exerted on the surface, (iii) the body force. The first term is $-\rho v v\cdot n$, the second term is $-pn$. Thus, we have

$$\begin{aligned}\frac{d}{dt}\int_\Omega \rho v\,dx &= \int_{\partial\Omega} -[\rho v v\cdot n + pn]\,dS + \int F\,dx \\ &= \int_\Omega \text{div}[-\rho v\otimes v - pI] + F\,dx\end{aligned}$$

This yields

$$(\rho v)_t + \text{ div}(\rho\, v\otimes v) + \nabla p = F \tag{8.9}$$

Here, the notation $\nabla\cdot \rho v\otimes v$ stands for a vector whoes ith component is $\sum_j \partial_j(\rho v^i v^j)$. The energy per unit volume is $E = \frac{1}{2}\rho\, v^2 + \rho e$. The energy change in Ω per unit time is due to (i) the energy carried in through boundary (ii) the work done by the pressure from boundary, and (iii) the work done by the body force. The first term is $-Ev\cdot n$. The second term is $-pv\cdot n$. The third term is $F\cdot v$. The conservation of energy can be read as

$$\frac{d}{dt}\int_\Omega E\,dx = \int_{\partial\Omega}[-Ev\cdot n - pv\cdot n]\,dS + \int_\Omega F\cdot v\,dx$$

By applying divergence theorem, we obtain the energy equation:

$$E_t + \text{div}[(E+p)v] = \rho F\cdot v. \tag{8.10}$$

In one dimension, the equations are (without body force)

$$\begin{aligned}\rho_t + (\rho u)_x &= 0 \\ (\rho u)_t + (\rho u^2 + p)_x &= 0 \\ (\frac{1}{2}\rho u^2 + e)_t + [(\frac{1}{2}\rho u^2 + e + p)u]_x &= 0.\end{aligned}$$

Here, the unknowns are two thermo variable ρ and e, and one kinetic variable u. Other thermo variable p is given by the constitutive equation $p(\rho, e)$.

8.2.2 Riemann Problem of Gas Dynamics

We use (ρ, u, S) as our variables.

$$\begin{pmatrix}\rho\\ u\\ S\end{pmatrix}_t + \begin{pmatrix}u & \rho & 0\\ \frac{c^2}{\rho} & u & \frac{P_S}{\rho}\\ 0 & 0 & u\end{pmatrix}\begin{pmatrix}\rho\\ u\\ S\end{pmatrix}_x = 0$$

Where $p(\rho, S) = A(S)\rho^\gamma, \gamma > 1$ and $c^2 = \left.\frac{\partial P}{\partial\rho}\right|_S$. The eigenvalues and corresponding eigenvectors are

$$\begin{array}{ccc}\lambda_1 = u-c & \lambda_2 = u & \lambda_3 = u+c\\ r_1 = \begin{pmatrix}\rho\\ -c\\ 0\end{pmatrix} & r_2 = \begin{pmatrix}-P_S\\ 0\\ c^2\end{pmatrix} & r_3 = \begin{pmatrix}\rho\\ c\\ 0\end{pmatrix}\\ \ell_1 = (c, -\rho, \frac{P_S}{c}) & \ell_2 = (0,0,1) & \ell_3 = (c, \rho, \frac{P_S}{c})\end{array}$$

Note that

$$\begin{aligned}
&\nabla\lambda_1 \cdot r_1 = \tfrac{1}{c}(\tfrac{1}{2}\rho P_{\rho\rho} + c^2) > 0\\
&\nabla\lambda_3 \cdot r_3 = \tfrac{1}{c}(\tfrac{1}{2}\rho P_{\rho\rho} + c^2) > 0\\
&\nabla\lambda_2 \cdot r_2 \equiv 0.
\end{aligned}$$

R_1 is the integral curve of $(d\rho, du, dS) \parallel r_1$ and $(d\rho, du, dS) \perp \ell_2$ and ℓ_3. Therefore on R_1,

$$\begin{aligned}
&\left\{ \begin{array}{l} (d\rho, du, dS)\cdot(0,0,1) = 0 \\ (d\rho, du, dS)\cdot(c, \rho, \frac{P_S}{c}) = 0. \end{array} \right.\\
\Longrightarrow \quad &\left\{ \begin{array}{l} dS = 0 \text{ along } R_1 \\ cd\rho + \rho du + \frac{P_S}{c} dS = 0 \end{array} \right.\\
\Longrightarrow \quad &\left\{ \begin{array}{l} cd\rho + \rho du = 0 \\ c^2 d\rho + P_S dS + c\rho du = 0 \end{array} \right. \quad \Longrightarrow \quad dP + c\rho du = 0
\end{aligned}$$

On R_2, $(d\rho, du, dS) \perp \ell_1, \ell_3$

$$\begin{aligned}
\Longrightarrow \quad &\left\{ \begin{array}{l} c^2 d\rho + c\rho du + P_S dS = 0 \\ c^2 d\rho - c\rho du + P_S dS = 0 \end{array} \right.\\
\Longrightarrow \quad &\left\{ \begin{array}{l} dP + c\rho du = 0 \\ dP - c\rho du = 0 \end{array} \right.\\
\Longrightarrow \quad &\left\{ \begin{array}{l} dP = 0 \\ du = 0 \end{array} \right. \quad \rho \neq 0
\end{aligned}$$

On R_3, $(d\rho, du, dS) \perp \ell_1, \ell_2$

$$\left\{ \begin{array}{l} dS = 0 \\ cd\rho - \rho du = 0 \end{array} \right.$$

Let $\ell = \int \frac{c(\rho,S)}{\rho}\, d\rho$. From $c = \sqrt{P_\rho} = \sqrt{A(S)\gamma\rho^{\gamma-1}}$, $\ell(P,s) = \sqrt{\gamma A(S)}\frac{2}{\gamma-1}\rho^{\frac{\gamma-1}{2}}$. Then on R_3',

$$\begin{aligned}
u - u_0 &= \mp \int_{\rho_0}^{\rho} \frac{c}{\rho}\, d\rho = \mp(\ell - \ell_0)\\
\ell &= \sqrt{\gamma A(S)}\frac{2}{\gamma-1}\rho^{\gamma-1}2 = \frac{2}{\gamma-1}\sqrt{\frac{\gamma P}{\rho}}\\
P\rho^{-\gamma} &= A(S) = A(S_0) = P_0\rho_0^{-\gamma}.
\end{aligned}$$

Express ρ interms of P, P_0, ρ_0, then plug it into ℓ,

$$\begin{aligned}
\ell - \ell_0 &= \psi(P)\\
&= \frac{2}{\gamma-1}(\sqrt{\gamma P(\frac{P_0}{P})^{\frac{1}{\gamma}}\rho_0^{-1}} - \sqrt{\frac{\gamma P_0}{\rho_0}})\\
&= \frac{2\sqrt{\gamma}}{\gamma-1}\rho_0^{-\frac{1}{2}}P_0^{\frac{1}{2\gamma}}(P^{\frac{\gamma-1}{2\gamma}} - P_0^{\frac{\gamma-1}{2\gamma}})
\end{aligned}$$

$$\therefore \quad \begin{array}{ll} R_1 & u = u_0 - \psi_0(P) \\ R_3 & u = u_0 + \psi_0(P) \end{array}$$

P

R_3^+

(ℓ)

R_1^+

u

Figure 8.2: The integral curve of the first and the third field on the (u, P) phase plane.

On R_2, which is a contact discontinuity, $du = 0, dP = 0$. Therefore $u = u_0, P = P_0$.
For S_1, S_3

$$\begin{cases} \rho_t + (\rho u)_x = 0 \\ (\rho u)_t + (\rho u^2 + P)_x = 0 \\ (\frac{1}{2}\rho u^2 + \rho e)_t + ((\frac{1}{2}\rho u^2 + \rho e + P)u)_x = 0 \end{cases}$$

Suppose the shock is along $x - \sigma t$. Let $v = u - \sigma$ (standing shock)

$$\begin{cases} [\rho v] = 0 \\ [\rho v^2 + P] = 0 \\ [(\frac{1}{2}\rho v^2 + \rho e + P)v] = 0 \end{cases}$$

Let

$$m = \rho_0 v_0 = \rho v$$

which is from the first jump condition. The second jump condition says that

$$\begin{aligned} \rho_0 v_0^2 + P_0 &= \rho v^2 + P \\ m v_0 + P_0 &= m v + P \end{aligned}$$

$$\begin{aligned} m &= -\frac{P - P_0}{v - v_0} \\ &= -\frac{P - P_0}{m\tau - m\tau_0} \text{ where } \tau = \tfrac{1}{\rho} \text{ is the specific volume.} \end{aligned}$$

$$\begin{aligned} \therefore \quad m^2 &= -\tfrac{P - P_0}{\tau - \tau_0} \\ v - v_0 &= -\tfrac{P - P_0}{m} \\ (u - u_0)^2 = (v - v_0)^2 &= -(P - P_0)(\tau - \tau_0) \end{aligned}$$

The third one is

$$\begin{aligned}(\frac{1}{2}\rho_0 v_0^2 + \rho_0 e_0 + P_0)v_0 &= (\frac{1}{2}\rho v^2 + \rho e + P)v \\ \Longrightarrow \qquad \frac{1}{2}v_0^2 + e_0 + P_0\tau_0 &= \frac{1}{2}v^2 + e + P\tau\end{aligned}$$

By $v_0^2 = m^2\tau_0^2, v^2 = m^2\tau^2, m^2 = -\frac{P-P_0}{\tau-\tau_0}$,

$$\Longrightarrow \qquad H(P,\tau) = e - e_0 + \frac{P+P_0}{2}(\tau - \tau_0) = 0$$

Recall $e = \frac{P\tau}{\gamma-1}$. From $H(P,\tau) = 0$,

$$\frac{P\tau}{\gamma-1} - \frac{P_0\tau_0}{\gamma-1} + (\frac{P+P_0}{2})(\tau-\tau_0) = 0.$$

Solve fot τ in terms of P, P_0, τ_0, then plug into

$$(u-u_0)^2 = -(P-P_0)(\tau-\tau_0)$$

Set $\phi(P) = (P-P_0)\sqrt{\frac{\frac{2}{\gamma+1}\tau_0}{P + \frac{\gamma-1}{\gamma+1}P_0}}$ Then

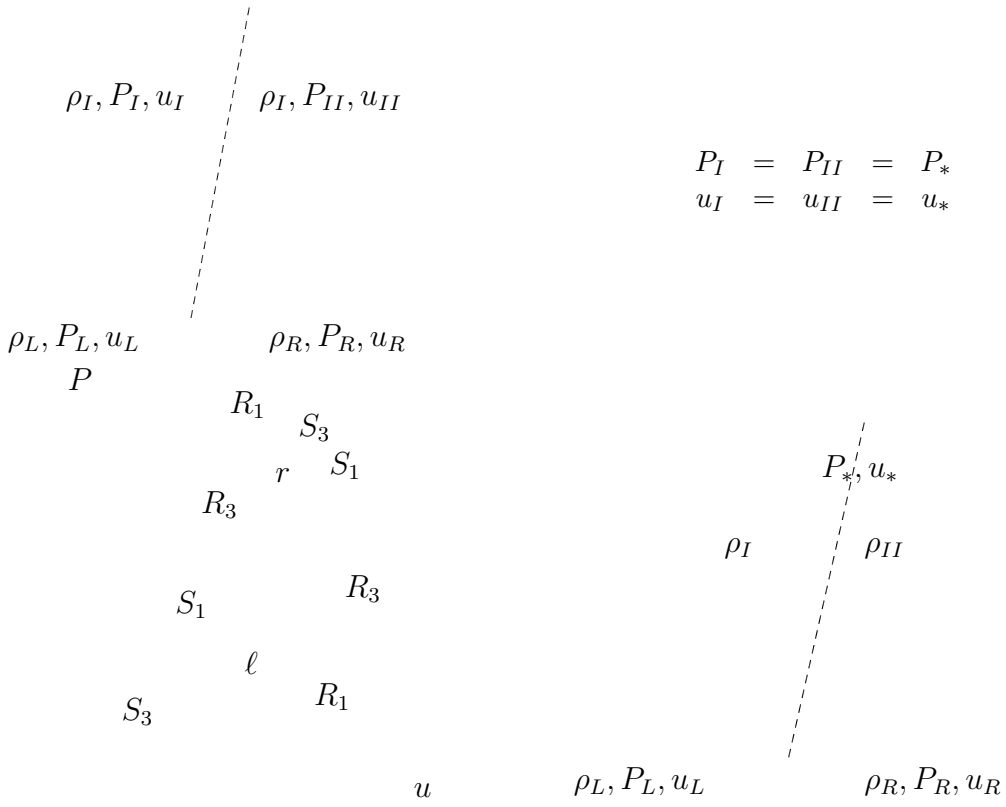
$$\begin{aligned}S_1: \quad u &= u_0 - \phi_0(P) \\ S_3: \quad u &= u_0 + \phi_0(P)\end{aligned}$$

Therefore,

$$\begin{aligned}T_1^{(\ell)}: \quad u &= \begin{cases} u_0 - \psi_0(P) & P < P_0 \\ u_0 - \phi_0(P) & P \geq P_0 \end{cases} \\ T_3^{(\ell)}: \quad u &= \begin{cases} u_0 + \psi_0(P) & P > P_0 \\ u_0 + \phi_0(P) & P \leq P_0 \end{cases}\end{aligned}$$

$$\begin{aligned}T_1^{(r)}: \quad u &= \begin{cases} u_0 - \psi_0(P) & P > P_0 \\ u_0 - \phi_0(P) & P \leq P_0 \end{cases} \\ T_3^{(r)}: \quad u &= \begin{cases} u_0 + \psi_0(P) & P < P_0 \\ u_0 + \phi_0(P) & P \geq P_0 \end{cases}\end{aligned}$$

Now we are ready to solve Riemann Problem with initial states (ρ_L, P_L, u_L) and (ρ_R, P_R, u_R). Recall that in the second field, $[P] = [u] = 0$.



The vaccum state

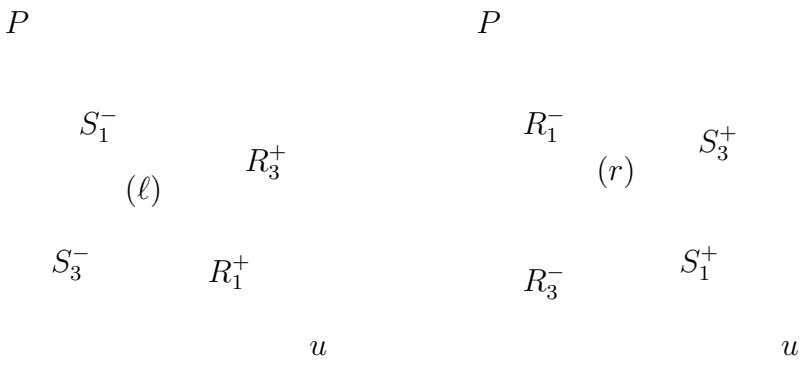


Figure 8.3: The rarefaction waves ans shocks of 1,3 field on (u, P) phase plane at left/right state.

P

r

ℓ

Solution must satisfy $P > 0$. If $u_\ell + \ell_\ell$ is less than $u_r - \ell_r$, there is no solution.

u

Finding middle states Given $U_L := (p_L, u_L, S_L)$ and $U_R := (p_R, u_R, S_R)$, we want to find two middle states U_I and U_{II} such that (U_L, U_I) forms a 1-wave, (u_{II}, U_R) forms a 3-wave and (U_I, U_{II}) forms a 2-wave. From jump condition of the 2-wave, we have $U_I = (p_*, u_*, S_I)$ and $U_{II} = (p_*, u_*, S_{II})$. With this, then S_I and S_{II} can be determined the equation on $T_1(U_L)$ and $T_3(U_R)$, respectively. The key step is to find (p_*, u_*).

Godunov gives a procedure to find the middle states (u_*, P_*). The algorithm to find P_* is to solve

$$\begin{aligned} u_\ell - f_\ell(P) &= u_I = u_{II} = u_r + f_r(P) \\ f_0(P) &= \begin{cases} \psi_0(P) & P < P_0 \\ \phi_0(P) & P \geq P_0 \end{cases} \end{aligned}$$

This is equivalent to

$$\begin{cases} Z_R(u_* - u_R) &= P_* - P_R \\ -Z_L(u_* - u_L) &= P_* - P_L. \end{cases}$$

Where

$$\begin{aligned} Z_R &= \sqrt{\frac{P_R}{\tau_R}}\Phi(\frac{P_*}{P_R}) \\ Z_L &= \sqrt{\frac{P_L}{\tau_L}}\Phi(\frac{P_*}{P_L}) \end{aligned}$$

and

$$\Phi(w) = \begin{cases} \sqrt{\frac{\gamma+1}{2}w + \frac{\gamma-1}{2}} & w > 1 \\ \frac{\gamma-1}{2\sqrt{\gamma}} \frac{1-w}{1-w^{\frac{\gamma-1}{2\gamma}}} & w \leq 1 \end{cases}$$

This is an equation for (u_*, P_*). It can be solved by Newton's method.

Approximate Riemann Solver Consider the Riemann data (u_L, u_R). We look for middle states $u_0 = u_L, u_1, ..., u_n = u_R$. Suppose $u_L \sim u_R$, the original equation can be replaced by

$$u_t + A(\bar{u})u_x = 0,$$

where $\bar{u} = \dfrac{u_L + u_R}{2}$. We will solve this linear hyperbolic equation with Riemann data (u_L, u_R). Let λ_i, ℓ_i, r_i be eigenvalues and eigenvectors of $A(\bar{u})$. Then the solution of the Riemann problem is self-similar and has the form

$$u(\frac{x}{t}) = u_L + \sum_{\lambda_i < \frac{x}{t}} (\ell_i \cdot (u_R - u_L)) \cdot r_i.$$

One severe error in this approximate Riemann solver is that rarefaction waves are approximated by discontinuities. This will produce non-entropy shocks. This is particularly serious in Godunov method which uses Riemann solution at $x/t = 0$. To cure this problem, we expand such a linear discontinuity by a linear fan. Precisely, suppose $\lambda_i(u_{i-1}) < 0, \lambda_i(u_i) > 0$, this suggests that there exists rarefaction fan crossing $\frac{x}{t} = 0$. We then expand this discontinuity by a linear fan. At $x/t = 0$, we thus choose

$$\begin{aligned} u_m &= (1-\alpha)u_{i-1} + \alpha u_i, \\ \alpha &= \frac{-\lambda_i(u_{i-1})}{\lambda_i(u_i) - \lambda_i(u_{i-1})}. \end{aligned}$$

A final remark, the above $\bar{u}$ can be replaced by Roe's state in the application of gas dynamics.

Chapter 9

Kinetic Theory and Kinetic Schemes

9.1 Kinetic Theory of Gases

9.2 Kinetic scheme

Assume the equilibrium distribution is $g_0(\xi)$. It should satisfy (i) momentum conditions, (ii) equation of states (or flux condition), (iii) positivity. That is, the moment condition:

$$\int g_0 \psi_\alpha(\xi)\, d\xi = U_\alpha$$

and flux condition:

$$\int g_0 \xi \psi_\alpha(\xi)\, d\xi = F_\alpha(U).$$

For non-convex case f,