AMATH 442 (Fall 2014 - 1149)

Numerical Solutions of Partial Differential Equations

Prof. L. Krivodonona
University of Waterloo

LaTeXer: W. KONG
http://wwkong.github.io
Last Revision: December 7, 2014

Table of Contents

These notes are currently a work in progress, and as such may be incomplete or contain errors.

ACKNOWLEDGMENTS:

Special thanks to *Michael Baker* and his LaTeX formatted notes. They were the inspiration for the structure of these notes.

Abstract

The purpose of these notes is to provide the reader with a secondary reference to the material covered in AMATH 442. The formal prerequisite to this course is either AMATH 351 or AMATH 350.

Errata

6-7 Assignments Biweekly

25% Assignments, 25% Midterm, 50% Final Exam

Office hours: W, Th @ 2-3pm

Midterm: Oct. 21st @ 4-5:30pm

1 Introduction

We begin with a quick review of the theoretical bases of **partial differential equations**.

1.1 Classification of 2nd Order Linear PDEs

There are 3 types of (linear) PDEs:

1. Parabolic PDEs (e.g. heat equation, diffusion equation)
   (a) Has the form $u_t = \sigma u_{xx}$ or in the multivariate case, $u_t = \sigma(u_{xx} + u_{yy} + u_{zz})$
2. Elliptic PDEs (e.g. Laplace's equation, Poisson equation)
   (a) Has the form $u_{xx} + u_{yy} = f(x, y)$ or $\Delta u = f(x, y)$
3. Hyperbolic PDEs (e.g. 1st, 2nd order wave equations)
   (a) Has the form $u_{tt} - c^2 u_{xx} = 0$ or $u_t + a u_x = 0$

There are also **non-linear PDEs**:

- Burger's equation: $u_t + u u_x = 0$
- Non-linear heat equation: $u_t = (\sigma(u) u_x)_x$
- Higher-order PDEs: $u_t + u u_x = \sigma u_{xxx}$
- Mixed types

1.2 Examples of Linear PDEs

(1) Let's begin by looking at the classic **linear advection (a.k.a. wave) equation**. The basic form is

$$u_t + a u_x = 0, a \in \mathbb{R}, (x, t) \in \mathbb{R} \times \mathbb{R}$$

Claim 1.1. Any $\phi(x - at)$ is a solution.

Proof. Substitution and chain rule:

$$u = \phi(x - at) \implies u_t = \phi'(x - at)(-a), u_x = \phi'(x - at) \implies -a\phi' + a\phi' = 0$$

Therefore ϕ is a solution. □

With the initial condition $u(x,0) = u_0(x)$, the solution is $u = u_0(x - at)$. The PDE with the aforementioned initial condition is called the **Cauchy problem**. We can interpret the parameter a as a speed parameter.

Now suppose that we introduce a finite domain $\Omega = [\alpha, \beta]$ and **boundary conditions**. Saying $u(\beta, t) = b_{right}(t)$ might lead to contradiction since $u_0(x-at) \neq b_{right}(\beta, t)$ or $u_0 = b_{right}$ (no new information is given). Instead, we provide $u(\alpha, t) = b_{left}(t)$ if $a > 0$ and we provide $u(\beta, t) = b_{right}(t)$ if $a < 0$.[1]

Conclusion 1. Here are some conclusions regarding the above wave equation:

[1] The solution of (1) does not grow or decay over time.

[2] New extrema can be introduced only through boundary conditions.

(2) Moving on, we have the **diffusion (heat) equation**. The basic form is

$$u_t = \sigma u_{xx}, \sigma \in \mathbb{R}$$

Assume that the initial conditions (I.C.) and boundary conditions (B.C.) are such that

$$u(x,t) = \hat{u}(k,t) \sin kx$$

is a solution, with k fixed. By substitution,

$$\begin{aligned} u_t &= \hat{u}_t \sin kx \\ u_x &= k\hat{u} \cos kx \\ u_{xx} &= -k^2 \hat{u} \sin kx \end{aligned}$$

and so

$$\hat{u_t} \sin kx = -\sigma k^2 \hat{u} \sin kx \implies \hat{u}_t = -\sigma k^2 \hat{u} \implies \hat{u}(k,t) = ce^{-\sigma k^2 t} \implies u(x,t) = ce^{-\sigma k^2 t} \sin kx$$

If we set $c = 1$ then the I.C. should be

$$u(x,0) = e^{-\sigma k^2 \cdot 0} \sin kx = \sin kx$$

If the domain is $\Omega = [-1, 1]$ and $k = \pi$ then the B.C. is

$$\begin{cases} u(-1,t) &= 0 \\ u(1,t) &= 0 \end{cases}$$

Remark 1.1. Here are some remarks about the solution:

[1] If $\sigma > 0$ then $u(x,t)$ decays with time (proper heat equation) and if $\sigma < 0$ then $u(x,t)$ grows with time (inverse or backwards heat equation). For this course, we always assume that $\sigma > 0$.

[2] The larger the σ, the faster the decay with respect to time. We call σ the **diffusion coefficient**.

[3] The larger the k, the faster the decay $\implies$ high frequencies decay faster.

2 Finite Difference Methods

Recall that

$$\begin{aligned} u_x &:= \lim_{\Delta x \to 0} \frac{u(x + \Delta x, t) - u(x,t)}{\Delta x} \\ \Delta^+ u &:= \frac{u(x + \Delta x, t) - u(x,t)}{\Delta x} \\ \Delta^- u &:= \frac{u(x,t) - u(x - \Delta x, t)}{\Delta x} \end{aligned}$$

[1] $x = \alpha$ is called inflow while $x = \beta$ is called outflow.

where we call the last two the **1st forward difference** and **1st backward difference** respectively. By convention, $\Delta x > 0$ and $\Delta t > 0$. Note that Δx is finite which is where the name "finite difference" comes from. u_x will be approximated by $\Delta^+ u$ or $\Delta^- u$. Similarly,

$$u_t \approx \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t}$$

We then introduce a discretization of space where

$$\begin{aligned} \Delta x_j &= x_{j+1} - x_j \\ \Delta t_n &= t_{n+1} - t_n \end{aligned}$$

For simplicity, assume uniform discretization. That is, $\Delta x_j = \Delta x, \Delta t_n = \Delta t$ for all j and t. In general $\Delta x \neq \Delta t$. We will use the notation

$$\begin{aligned} (x_j, t_n) &\equiv (j, n) \\ u_j^n &\equiv u(x_j, t_n) \end{aligned}$$

Finally, we denote the numerical solution as $U_j^n \approx u_j^n$. Now recall the Taylor series expansion of u about (x_j, t_n) in x:

$$\begin{aligned} u(x_j + \Delta x, t_n) &= u(x_j, t_n) + \Delta x u_x(x_j, t_n) + \frac{\Delta x^2}{2} u_{xx}(x_j, t_n) + ... \\ u(x_j - \Delta x, t_n) &= u(x_j, t_n) - \Delta x u_x(x_j, t_n) + \frac{\Delta x^2}{2} u_{xx}(x_j, t_n) - ... \end{aligned}$$

or more compactly,

$$\begin{aligned} u_{j+1}^n &= u_j^n + \Delta x \, (u_x)_j^n + \frac{\Delta x^2}{2} (u_{xx})_j^n + ... \\ u_{j-1}^n &= u_j^n - \Delta x \, (u_x)_j^n + \frac{\Delta x^2}{2} (u_{xx})_j^n - ... \end{aligned}$$

If we solve for $(u_x)_j^n$ in the first equation, then we get

$$(u_x)_j^n = \frac{u_{j+1}^n - u_j^n}{\Delta x} - \frac{\Delta x}{2} (u_{xx})_{j+\xi}^n, 0 < \xi < 1$$

by the mean value theorem. We call $\tau_j^n = -\frac{\Delta x}{2} (u_{xx})_{j+\xi}^n$ the **discretization (truncation) error**. Similarly from the second equation,

$$(u_x)_j^n = \frac{u_{j+1}^n - u_j^n}{\Delta x} \underbrace{+ \frac{\Delta x}{2} (u_{xx})_{j-\xi}^n}_{\tau_j^n}, 0 < \xi < 1$$

If we subtract the two equations together, then

$$(u_x)_j^n = \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} \underbrace{- \frac{1}{6} (u_{xxx})_{j+\xi}^n \Delta x^2}_{\tau_j^n}, 0 < \xi < 1$$

We call the first term on the right side the **1st central difference**. Central difference is more accurate than forward and backward difference. More accuracy is achievable with more points x_{j+2}, x_{j+3}. Adding the two equations will give us

$$(u_{xx})_j^n = \frac{u_{j+1} - 2u_j^n + u_{j-1}^n}{\Delta x^2} - \frac{\Delta x^2}{12} (u_{xxxx})_{j+\eta}^n, 0 < \eta < 1$$

In general, higher derivatives and more accurate approximations require more points (i.e. larger **stencil**).

Using **big-O notation**, we can write:

$$\begin{aligned}(u_x)_j^n &= \frac{u_{j+1}^n - u_j^n}{\Delta x} + O(\Delta x)\\ (u_t)_j^n &= \frac{u_j^{n+1} - u_j^n}{\Delta t} + O(\Delta t)\\ (u_{xx})_j^n &= \frac{u_{j+1} - 2u_j^n + u_{j-1}^n}{\Delta x^2} + O(\Delta x^2)\end{aligned}$$

Example 2.1. Let's construct a finite difference (FD) scheme for the heat equation:

$$\begin{aligned}u_t &= \sigma u_{xx}, -\infty < x < \infty\\ u(x,0) &= \phi(x)\end{aligned}$$

We have

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \sigma \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} \implies u_j^{n+1} = ru_{j-1}^n + (1-2r)u_j^n + ru_{j+1}^n$$

where $r = \sigma \Delta t / \Delta x^2$. If we know u_j^n for all j then we can compute u_j^{n+1} for all j. We need u_j^0 so we set $u_j^0 = u(x_j, 0) = \phi(x_j)$ for all j.

Let's plug in some values. Suppose that $\sigma = 1$ and choose the I.C. such that $u_0^0 = 1, u_j^0 = 0, \forall j \neq 0$ and $\Delta x = 1, \Delta t = 1/4 \implies r = 4$. This gives us

$$u_j^{n+1} = 4u_{j-1}^n - 7u_j^n + 4u_{j+1}^n$$

From stability analysis (CS 476), you will see that:

1. u_j^n grows

2. u_j^n oscillates (+ve, -ve, +ve, -ve, ...)

Instead, let's try: $\Delta x = 1/4, \Delta t = 1/64 \implies r = 1/4$ with:

$$u_j^{n+1} = \frac{1}{4}u_{j-1}^n + \frac{1}{2}u_j^n + \frac{1}{4}u_{j+1}^n$$

This will provide reasonable results. In general, we want u_0^n to be a good approximation of u_j^n.

Definition 2.1. A scheme is **convergent** on $0 < t \leq T$ if

$$\|u^n - U^n\| \to 0$$

as $\Delta x \to 0, \Delta t \to 0, n \to \infty, n\Delta t \leq T$. Here, $\|\cdot\|$ is some norm with u^n as a vector of all the (u_j^n)'s. A scheme is **convergent of order** k if

$$\|u^n - U^n\| = O(\Delta x^k)$$

Fact 2.1. *Convergence is difficult to prove directly. Instead, we look at:*

- Stability
- Consistency

Going back to our last example, consider $\|u\|_\infty = \max_j |u_j|$. From the general equation

$$\begin{aligned}|u_j^{n+1}| &\leq |r||u_{j-1}^n| + |1-2r||u_j^n| + |r||u_{j+1}^n|\\ &\leq (|r| + |1-2r| + |r|)\|u^n\|_\infty\end{aligned}$$

If $0 < r < \frac{1}{2}$ then $|u_j^{n+1}| \leq \|u^n\|_\infty, \forall j \implies \|u^{n+1}\| \leq \|u^n\|_\infty$. If $r > \frac{1}{2}$, then

$$|r| + |1-2r| + |r| = 2r - 1 + 2r = 4r - 1 \geq 1$$

and hence

$$|u_j^{n+1}| \leq (4r-1)\|u^n\|_\infty$$

Definition 2.2. A scheme is **stable** if $\exists C > 0$ **independent** of $\Delta x, \Delta t, u^0$ such that

$$\|u^n\| \leq C\|u^0\|, \forall n \in \mathbb{N}, \Delta x \leq \overline{\Delta x}, \Delta t \leq \overline{\Delta t}, n\Delta t \leq T$$

Note 1. (1) We allow some growth in the solution. Don't confuse this definition of stability with stability in ODE theory.

(2) Scheme is usually stable only for fixed values of some parameters. For example, Δt as a function of Δx or r.

In our example above, we showed that it was a stable scheme for the heat equation when $r < \frac{1}{2}$.

Definition 2.3. Alternatively, if u^n, v^n are solutions with $u^0 = \phi, v^0 = \psi$ (same problem, different I.C.), then a scheme is stable if $\exists C > 0$ independent of $\Delta x, \Delta t, u^0$ such that

$$\|u^n - v^n\| \leq C\|u^0 - v^0\|, \forall n \in \mathbb{N}, \Delta x \leq \overline{\Delta x}, \Delta t \leq \overline{\Delta t}, n\Delta t \leq T$$

Example 2.2. Going back to heat equation, suppose we choose I.C.

$$u^0 = (..., -1, 1, -1, 1, ...) \implies u_j^0 = (-1)^j$$

and hence

$$\begin{aligned} u_j^1 &= 2r(-1)^{j+1} + (1-2r)(-1)^j \\ &= (-1)^j(-2r+1-2r) \\ &= -(4r-1)(-1)^j \\ u_j^n &= (-1)^{j+1}(4r-1)^n \end{aligned}$$

Taking norms, we have

$$\|u^n\|_\infty = (4r-1)^n\|u^0\|_\infty = (4r-1)^n$$

We call this **exponential growth** in the case of $r > \frac{1}{2}$. As $\Delta x, \Delta t \to 0$ with fixed T and $n \to \infty$, we have

$$\|u^n\|_\infty \to \infty$$

So with $r > \frac{1}{2}$, the results are **unstable**.

Remark 2.1. Stability for numerical methods is equivalent to **well-posedness** for PDEs:

- Solution exists given suitable I.C. and B.C.
- Solution is unique
- Solution is continuously dependent on initial data

2.1 Consistency

We now change our notation so that U^n is the finite difference estimate and u^n is the exact solution. We want to know how much $u(x,t)$ satisfies the below equation

$$(2)\ \frac{U_j^{n+1} - U_n^n}{\Delta t} = \sigma\frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}$$

which is a discretization of the heat equation. Note that $u(x,t)$ only exactly solves

$$(1)\ u_t = \sigma u_{xx}$$

Define

$$P(v) = \frac{v_j^{n+1} - v_n^n}{\Delta t} - \sigma\frac{v_{j+1}^n - 2v_j^n + v_{j-1}^n}{\Delta x^2}$$

We have $P(U^n) = 0$. Define $\tau_j^n \equiv P(u)$ where $u = u(x,t)$. We call τ_j^n the **truncation or discretization error**. Plug $u(x,t)$ into (2) to get

$$\begin{aligned}\tau_j^n &= \frac{u_j^n + \Delta t(u_t)_j^n + \frac{\Delta t^2}{2}(u_{tt})_j^n + O(\Delta t^3) - u_j^n}{\Delta t} - \\ &\sigma\frac{u_j^n + \Delta x(u_x)_j^n + \frac{\Delta x^2}{2}(u_{xx})_j^n + \frac{\Delta x^3}{6}(u_{xxx})_j^n + \frac{\Delta x^4}{24}(u_{xxxx})_j^n + O(\Delta x^5)}{\Delta x^2} - \\ &\sigma\frac{-2u_j^n + \left(u_j^n - \Delta x(u_x)_j + \frac{\Delta x^2}{2}(u_{xx})_j^n - \frac{\Delta x^3}{6}(u_{xxx})_j^n + \frac{\Delta x^4}{24}(u_{xxxx})_j^n + O(\Delta x^5)\right)}{\Delta x^2}\end{aligned}$$

This can be reduced to

$$\begin{aligned}\tau_j^n &= \underbrace{(u_t)_j^n - \sigma(u_{xx})_j^n}_{=0} + \frac{\Delta t}{2}(u_{tt})_j^n - \sigma\frac{\Delta x^2}{12}(u_{xxxx})_j^n + O(\Delta t^2) \\ &= \frac{\Delta t}{2}(u_{tt})_{j+\xi}^n - \sigma\frac{\Delta x^2}{12}(u_{xxxx})_{j+\eta}^n\end{aligned}$$

Suppose that the function $u(x,t)$ is smooth enough such that there exists M with the property $|u_{tt}|, |u_{xxxx}| \leq M$. We then get

$$|\tau_j^n| \leq M\left(\frac{\Delta t}{2} + \frac{\Delta x^2}{12}\right)$$

and hence $(\tau_j^n) = O(\Delta t, \Delta x^2)$. Since we need $r < \frac{1}{2}$ for stability, we have

$$\sigma\frac{\Delta t}{\Delta x^2} < \frac{1}{2} \implies \Delta t < \frac{\Delta x^2}{2\sigma} \implies (\tau_j^n) = O(\Delta x^2) \text{ if } r < \frac{1}{2}$$

Definition 2.4. A scheme is called **consistent** if $\tau_j^n \to 0$ as $\Delta x \to 0, \Delta t \to 0$. A scheme is called **consistent of order** k in Δx and m in Δt if

$$\tau_j^n = O(\Delta x^k, \Delta t^m)$$

Remark 2.2. Regarding the truncation error:

1. τ_j^n measures how far (2) is from (1)
2. τ_j^n is a purely analytical tool. Don't try to find it in your code!
3. τ_j^n is easy to compute $\implies$ the reason it is used
4. For many schemes (all ours) if $\tau_j^n = O(\Delta x^k, \Delta t^m)$ then

$$\|e^n\| = \|u^n - U^n\| = O(\Delta x^k, \Delta t^m)$$

Note 2. We note that

$$\begin{cases} U_j^{n+1} &= rU_{j-1}^n + (1-2r)U_j^n + rU_{j+1}^n \\ u_j^{n+1} &= ru_{j-1}^n + (1-2r)u_j^n + ru_{j+1}^n + \Delta\tau_j^n\Delta t \end{cases}$$

and hence

$$\begin{aligned} e_j^{n+1} &= re_{j-1}^n + (1-2r)e_j^n + re_{j+1}^n + \tau_j^n\Delta t \\ |e_j^{n+1}| &\leq (|r| + |1-2r| + |r|)\|e^n\| + \Delta t\|\tau^n\| \end{aligned}$$

and $r < \frac{1}{2}$ gives us

$$\begin{aligned}\|e^{n+1}\| &\leq \|e^n\| + \Delta t\|\tau^n\| \\ &\leq \|e^{n-1}\| + \Delta t(\|\tau^n\| + \|\tau^{n-1}\|) \\ &\leq \|e^0\| + \Delta t\sum_{k=1}^{n}\|\tau^k\|\end{aligned}$$

Since $\|e^0\| = 0$ because $U_j^0 = u_j^0$ if we let $\tau = \max \|\tau^k\|$, then

$$\|e^{n+1}\| \leq \Delta t \sum_{k=1}^{n} \tau = \Delta t \cdot n \cdot \tau = t_n \cdot \tau$$

and hence

$$\|e^{n+1}\| = \|u^n - U^n\| \leq \underbrace{t_n}_{\text{finite}} \cdot C(\Delta x^2 + \Delta t) \leq \bar{C}\Delta x^2$$

for some constants $C, \bar{C}$. We should expect quadratic convergence on smooth solutions of (1) using (2).

Remark 2.3. If $\tau_j^n = 0$ then $e_j^n = 0$ and hence if $u(x,t)$ is linear in time and cubic in space, then (2) solves (1) exactly.

Recall

1. Stability doesn't grow with time uncontrollably
2. Consistency gives the convergence (and its rate)
3. Convergence is *good*

Theorem 2.1. *(Lax Equivalence Theorem) We have*

$$\textit{Stability+Consistency} \iff \textit{Convergence}$$

The forward direction is easy to prove, while the reverse direction is difficult to prove. This is true for most (and of all of our) methods.

2.2 Convergence in Practice and Error Estimation

From the previous section, we saw that

$$\|e_j^n\| \sim C\Delta x^2 \implies \|e_j^n\| = O(\Delta x^2)$$

Suppose we have two meshes with Δx and $\frac{\Delta x}{2}$ and we know that in general $e_j^n = O(\Delta x^k)$. We then have

$$\begin{cases} \|e_{\Delta x}^n\| & \sim C_1 \Delta x^k \\ \|e_{\frac{\Delta x}{2}}^n\| & \sim C_2 \left(\frac{\Delta x}{2}\right)^k \end{cases} \implies \frac{\|e_{\Delta x}^n\|}{\|e_{\frac{\Delta x}{2}}^n\|} \sim \frac{C_1 \Delta x^k}{C_2 \left(\frac{\Delta x}{2}\right)^k} \implies \log_2 \frac{\|e_{\Delta x}^n\|}{\|e_{\frac{\Delta x}{2}}^n\|} \sim k, \ (C_1 \approx C_2)$$

Convergence Tests when $u(x,t)$ is not known

Suppose that we have three solutions $\{U_{\Delta x}, U_{\Delta x/2}, U_{\Delta x/4}\}$ which are methods of order k.

1. Pick a very fine mesh, say $\Delta x/64$ (arbitrary) and view if as an exact solution.
2. Consider

$$\log_2 \frac{\|U_{\Delta x} - U_{\Delta x/2}\|}{\|U_{\Delta x/2} - U_{\Delta x/4}\|} \leq \log_2 \frac{\|U_{\Delta x} - u\| + \|u - U_{\Delta x/2}\|}{\|U_{\Delta x/2} - u\| + \|u - U_{\Delta x/4}\|} \sim \log_2 \frac{C_1 \Delta x^k + C_2 \left(\frac{\Delta x}{2}\right)^k}{C_2 \left(\frac{\Delta x}{2}\right)^k + C_3 \left(\frac{\Delta x}{4}\right)^k}$$

   and simplifying with $C = C_1 \sim C_2 \sim C_3$, we we get

$$\log_2 \frac{C_1 \Delta x^k + C_2 \left(\frac{\Delta x}{2}\right)^k}{C_2 \left(\frac{\Delta x}{2}\right)^k + C_3 \left(\frac{\Delta x}{4}\right)^k} \sim \log_2 2^k = k$$

Richardson Extrapolation (Error Estimation)

Suppose we have two solutions $U_{\Delta x}^n, U_{\Delta x/2}^n$. Then,

$$U_{\Delta x}^n - U_{\Delta x/2}^n = (U_{\Delta x}^n - u^n) + (u_n - U_{\Delta x/2}^n) \approx c_1 \Delta x^k - c_2 \left(\frac{\Delta x}{2}\right)^k \approx C\left(1 - \frac{1}{2^k}\right)\Delta x^k$$

The error of the Δx grid is $e^n_{\Delta x} \approx C\Delta x^k$ and hence

$$e^n_{\Delta x} \sim C\Delta x^k \approx \frac{U^n_{\Delta x} - U^n_{\Delta x/2}}{1 - \frac{1}{2^k}}$$

Similarly for the $\Delta x/2$ grid, we have $e^n_{\Delta x/2} \approx C\left(\frac{\Delta x}{2}\right)^k$ and hence

$$\frac{U^n_{\Delta x/2} - U^n_{\Delta x/4}}{2^k\left(1 - \frac{1}{2^k}\right)} = \frac{U^n_{\Delta x/2} - U^n_{\Delta x/4}}{2^k - 1} \sim e^n_{\Delta x/2}$$

So $e_{\Delta x}$ is more reliable than $e_{\Delta x/2}$ but $e_{\Delta x/2}$ is an estimate for a better solution.

2.3 Von-Neumann Stability Analysis

This is a general tool applicable to schemes other than finite difference methods.

Review. Recall **Euler's formula**

$$e^{\beta i} = \cos\beta + i\sin\beta \implies \cos\beta = \frac{1}{2}(e^{\beta i} - e^{-\beta i}), \sin\beta = \frac{i}{2}(e^{-\beta i} - e^{\beta i})$$

Given

$$g(t) = e^{(\alpha + i\beta)} = e^{\alpha t}(\cos\beta t + i\sin\beta t)$$

we have that α is responsible for the growth in $g(t)$ w.r.t. (with respect to) time and β is the frequency.

Review. Suppose that $f(x)$ is on $[-\pi, \pi]$. Then the **Fourier series** (F.S.) of $f(x)$ is

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k\cos kx + b_k\sin kx)$$

Theorem 2.2. *If $f(x)$ is periodic and C^1 then the Fourier series of $f(x)$ converges to $f(x)$ in the infinity and L_2 norms.*

Consider the exponential for the F.S. using Euler's formula as a substitution:

$$\begin{aligned} f(x) &\sim \frac{a_0}{2} + \sum_{k=1}^{\infty}\frac{a_k}{2}(e^{kxi} + e^{-kxi}) + \sum_{k=1}^{\infty}\frac{ib_k}{2}(e^{-kxi} - e^{kxi}) \\ &= \frac{a_0}{2} + \sum_{k=1}^{\infty}\underbrace{\frac{1}{2}(a_k - ib_k)}_{c_k}e^{kxi} + \sum_{k=1}^{\infty}\underbrace{\frac{1}{2}(a_k + ib_k)}_{c_{-k}}e^{-kxi} \\ &= \sum_{k=-\infty}^{\infty} c_k e^{kxi} \end{aligned}$$

It is easy to show that $c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-kxi}$. In the discrete version, we first choose a function $[-\pi, \pi] \mapsto [0, J]$ using

$$x(\xi) = \frac{2\pi}{J}\xi - \pi$$

Then

$$e^{kxi} = e^{\frac{2\pi\xi k}{J}i}e^{-\pi ki} = (-1)^k e^{\frac{2\pi k}{J}\xi i}$$

If we substitute this into the exponential form, then we get the F.S. on $[0, J]$:

$$f(\xi) \sim \sum_{k=-\infty}^{\infty} c_k(-1)^k e^{\frac{2\pi k}{J}\xi i} = \sum_{k=-\infty}^{\infty}\hat{c}_k e^{\frac{2\pi k}{J}\xi i}, \hat{c}_k = c_k(-1)^k$$

Now U_j^n is a discrete function defined at $x = x_j, j \in [0, J], \xi = \delta$. We claim that

$$(*)\ U_j = \sum_{k=0}^{J-1} A_k e^{\frac{2\pi k}{J} ji}$$

where we will call A_k the **discrete Fourier coefficients**. For the justification of (*), remark that:

1. The summation should be finite (stops at $k = J - 1$) because

   $$e^{\frac{2\pi J}{J} ji} = e^{2\pi ji} = e^{0ji} = 1$$

   Similar reasoning can be applied for any $k = J + s, 0 < s < J$.

2. If we rewrite (*) for U_j^n, then

   $$(**)\ U_j^n = \sum_{k=0}^{J-1} A_k^n w_j^k, w_j^k = e^{\frac{2\pi i}{J} kj}$$

   where A_k^n is time-indexed with a superscript and w_j^k is of degree k (power k).

3. (Orthogonality relation) Note that

   $$\sum_{j=0}^{J-1} w_j^k \bar{w}_j^m = \begin{cases} J & k \equiv m \mod J \\ 0 & \text{otherwise} \end{cases}$$

   Multiply $(**)$ by $\bar{w}_j^m$ and sum over j (m is fixed) to get

   $$\sum_{j=0}^{J-1} U_j^n \bar{w}_j^m = \sum_{j=0}^{J-1} \sum_{k=0}^{J-1} A_k^n w_j^k \bar{w}_j^m = \sum_{j=0}^{J-1} A_k^n \sum_{k=0}^{J-1} w_j^k \bar{w}_j^m = J A_m^n$$

   and hence

   $$A_m^n = \frac{1}{J} \sum_{j=0}^{J-1} U_j^n \bar{w}_j^m$$

4. (Discrete Parseval's Relation) It follows from above that

   $$\|U^n\|_2^2 = J \|A^n\|_2^2$$

   which follows from orthogonality. Compare this with the continuous case (very similar).

Remark 2.4. For the general heat equation $u_t = \sigma u_{xx} + f(x)$, if $u(x, t)$ tends to the $\bar{u}(x)$, called the **steady state**, then

$$\frac{\partial}{\partial t} u(x, t) = \frac{\partial}{\partial t} \bar{u}(x) = 0$$

and $\sigma u_{xx} = -f(x)$ which is an elliptic equation. Elliptic equations can be viewed as a steady state of parabolic equations.

Example. Here is the Von Neumann analysis applied to

$$(1)\ u_t + a u_x = 0$$

with periodic boundary conditions. In this problem, we want to find the stability condition for (1) (if any). Recall that

$$U_j^n = \sum_{k=0}^{J-1} A_k^n w_j^k, 0 \le j \le J$$

Note that we require periodic boundary conditions to allow the Fourier series to converge. One of the many FDMs for (1) is

$$(2)\ \frac{U_j^{n+1} - U_j^n}{\Delta t} + a \frac{U_j^n - U_{j-1}^n}{\Delta x} = 0, U_0^n = U_j^n$$

We call this scheme is FTBS. Rewriting, we have

$$U_j^{n+1} = (1-\alpha)U_j^n + \alpha U_{j-1}^n, \alpha = \frac{a\Delta t}{\Delta x}$$

Plug in the F.S. expansion to get

$$\sum_{k=0}^{J-1} A_k^{n+1} w_j^k = \sum_{k=0}^{J-1} \left((1-\alpha)A_k^n w_j^k + \alpha A_k^n w_{j-1}^k\right)$$

Collect terms with w_j^k with the fact that

$$w_{j-1}^k = e^{\frac{2\pi i}{J}kj} e^{-\frac{2\pi i}{J}k} = w_j^k e^{-\frac{2\pi i}{J}k}$$

to get

$$\sum_{k=0}^{J-1} \left(A_k^{n+1} - \left[(1-\alpha)A_k^n + \alpha A_k^n e^{-\frac{2\pi i}{J}k}\right]\right) w_j^k = 0 \implies A_k^{n+1} = \underbrace{\left[(1-\alpha) + \alpha e^{-\frac{2\pi i}{J}k}\right]}_{M_k} A_k^n$$

by linear independence.[2] By recurrence,

$$A_k^{n+1} = (M_k)^{n+1} A_k^0 \implies U_j^n = \sum_{k=0}^{J-1} (M_k)^n A_k^0 w_j^k$$

and hence by Parseval's identity

$$\|U^n\|_2^2 = J \sum_{k=0}^{J-1} |M_k|^{2n} |A_k^0|^2$$

If $|M_k| \leq 1$ for all k then

$$\|U^n\|_2^2 \leq J \sum_{k=0}^{J-1} |A_k^0| = \|U^0\|_2^2$$

So U^n is stable in 2-norm with $C = 1$. Since the exact solution of (1) doesn't grow in time, it is reasonable to require the same from U_j^n, i.e. $C = 1$.

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

Note 3. Say we have $k = \hat{k}$ such that $M_{\hat{k}} > 1$ and $M_k \leq 1, \forall k \neq \hat{k}$. Then the corresponding wave will grow in amplitude and dominate over the other smaller waves.

So now we want to find α such that $|M_k| \leq 1$. Instead, look for $|M_k|^2 \leq 1$ with

$$\begin{aligned}
|M_k|^2 = M_k \bar{M}_k &= (1 - \alpha + \alpha\cos\theta)^2 + (\alpha \sin\theta)^2, \theta = \frac{2\pi k}{J} \\
&= 1 - 2\alpha + \alpha^2 + 2\alpha(1-\alpha)\cos\theta + \alpha^2\cos^2\theta + \alpha^2 \sin\theta \\
&= 1 - 2\alpha + 2\alpha^2 + 2\alpha(1-\alpha)\cos\theta \\
&= 1 - 2\alpha(1-\alpha) + 2\alpha(1-\alpha)\cos\theta \\
&= 1 - 2\alpha(1-\alpha)(1-\cos\theta)
\end{aligned}$$

So

$$0 \leq |M_k|^2 = 1 - 4\alpha(1-\alpha)\sin^2\frac{\theta}{2} \leq 1$$

Since $\alpha = \frac{a\Delta t}{\Delta x}$ then our stability condition is

$$(*)\ \Delta t \leq \frac{\alpha \Delta x}{a}, 0 < \alpha \leq 1$$

for (1)-(2) if $a > 0$. If $a < 0$, then (2) is unstable for all $\Delta x, \Delta t$. We call (*) the **CFL** (Courant-Friedrichs-Lewy) **condition** which is a stability restriction on time step size Δt for hyperbolic problems.

Remark 2.5. Consider $\alpha = 1$ where $|M_k|^2 = 1 \implies$ no amplitude loss and and exact propagation of the initial profile. That

[2] This is due to the fact that the $w_j's$ form a linear independent basis.

is, $U_j^{n+1} = U_{j-1}^n$. If $a < 0$, we can show that

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + a\frac{U_{j+1}^n - U_j^n}{\Delta x} = 0$$

is stable with (*) under $\Delta t \leq \frac{\alpha \Delta x}{|a|}$ with $0 \leq \alpha \leq 1$.

Definition 2.5. A FDM (finite difference method) satisfies the **Von-Neumann condition** if $\exists C > 0$ independent of $\Delta x, \Delta t, k$ such that

$$|M_k| \leq 1 + C\Delta t, \forall \Delta t \leq \overline{\Delta t}, \Delta x \leq \overline{\Delta x}$$

Theorem 2.3. *A constant coefficient scalar one-level FDM is stable in the 2-norm iff it satisfies the Von Neumann conditions.*

Proof. ($\Longleftarrow$) Suppose U^n satisfies the Von Neumann conditions. Then,

$$\begin{aligned} \|U^n\|_2^2 &= J\sum_{k=0}^{J-1} |M_k|^{2n} |A_k^0|^2 \\ &\leq (1 + c\Delta t)^{2n} \|U^0\|_2^2 \end{aligned}$$

Now recall that

$$e^x = 1 + x + \frac{x^2}{2} + ... \implies 1 + x \leq e^x$$

and hence

$$(1 + c\Delta t)^{2n} \|U^0\|_2^2 \leq e^{2c\underbrace{\Delta t n}_{\Delta t_n}} \|U^0\|_2^2 \leq e^{2cT} \|U^0\|_2^2 \leq \bar{C}\|U^0\|_2^2$$

where $0 \leq t_n \leq T$ where T is the final time. □

($\Longrightarrow$) Suppose the scheme is stable and $\exists k = k^*$ such that $|M_{k^*}| > (1 + c\Delta t), \forall c$. Choose I.C. such that $A_{k^*}^0 \neq 0, A_k^0 = 0, k \neq k^*$. Then $U_j^0 = A_{k^*}^0 w_k^{k^*}$ and

$$U_j^n = (M_{k^*})^n A_{k^*}^0 w_j^{k^*} = (M_{k^*})^n U_j^0$$

Hence,

$$\|U^n\|_2^2 = |M_{k^*}|^{2n} \|U^0\|_2^2 > (1 + c\Delta t)^{2n} \|U^0\|_2^2$$

which implies that $\|U^n\|_2^2$ cannot be bounded by $\bar{C}\|U^0\|_2^2$ and hence is unstable. This is impossible and thus the scheme must satisfy the Von Neumann condition.

2.4 Implicit Methods

Recall the stability condition for the discretized heat equation

$$(1)\ u_t = \sigma u_{xx}$$

which was

$$(2)\ \frac{U_j^{n+1} - U_j^n}{\Delta t} = \sigma \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}, r = \frac{\sigma \Delta t}{\Delta x^2} \leq \frac{1}{2} \implies \Delta t \leq \frac{\Delta x^2}{2\sigma}$$

This is very restrictive and seldom used in practice. For example, if $\Delta x = 10^{-3}$ then $\Delta t \approx 10^{-6}$ and if $T = 1$ then $N = 10^6$. Consider

$$(3)\ \frac{U_j^{n+1} - U_j^n}{\Delta t} = \sigma \frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{\Delta x^2}$$

where

$$(4)\ U_j^n = -rU_{j+1}^{n+1} + (1 + 2r)U_j^{n+1} - rU_{j-1}^{n+1}, \tau_j^{n+1} = O(\Delta x^2, \Delta t)$$

Substitute $U_j^n = \sum_{k=0}^{J-1} A_k^n w_j^k$ into (4) to get

$$\sum_{k=0}^{J-1} A_k^n w_j^k = \sum_{k=0}^{J-1} \left(-rA_k^{n+1} w_{j+1}^k + (1 + 2r)A_k^{n+1} w_j^k - rA_k^{n+1} w_{j-1}^k\right)$$

and factoring out w_j^k (and matching coefficients) we get

$$A_k^n = \underbrace{\left(-re^{\frac{2\pi k}{J}i} + (1+2r) - re^{-\frac{2\pi k}{J}i}\right)}_{\equiv M_k^{-1}} A_k^{n+1} \implies A_k^{n+1} = M_k A_k^n$$

where

$$\begin{aligned} M_k^{-1} &= (1+2r) - 2r\cos\frac{2\pi k}{J} \\ &= 1 + 2r\left(1 - \cos\frac{2\pi k}{J}\right) = 1 + 4r\left(\sin\frac{\pi}{J}\right)^2 \end{aligned}$$

So $M_k^{-1} \geq 1, \forall r > 0, r = \frac{\sigma\Delta t}{\Delta x^2} \implies M_k \leq 1$ and (3) will be unconditionally stable. In practice, Δt is taken to be $O(\Delta x)$ for unconditionally stable schemes. However, (2) is $O(\Delta x^2)$ accurate ($r < 1/2$) and (3) is only $O(\Delta x)$ with $\Delta t \approx \Delta x$.

2.5 Crank-Nicolson Method

The **Crank-Nicholson** (CN) method is defined as

$$(5)\ \frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{\sigma}{2}\left[\frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{\Delta x^2} + \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}\right]$$

CN is unconditionally stable (see notes). It can be shown that $\tau_j^n = O(\Delta x^2, \Delta t^2)$:

$$\begin{aligned} \tau_j^n &= (u_t)_j^n + \frac{\Delta t}{2}(u_{tt})_j^n + O(\Delta t^2) - \frac{\sigma}{2}\left[(u_{xx})_j^{n+1} + O(\Delta x^2) + (u_{xx})_j^n + O(\Delta x^2)\right] \\ &= (u_t)_j^n + \frac{\Delta t}{2}(u_{tt})_j^n - \frac{\sigma}{2}\left[(u_{xx})_j^n + \Delta t(u_{xxt})_j^n + (u_{xx})_j^n + O(\Delta x^2) + O(\Delta t^2)\right] + O(\Delta t^2) \end{aligned}$$

Now $u_t = \sigma u_{xx} \implies u_{tt} = \sigma u_{xxt}$ and the result follows.

Solution Algorithm for Crank-Nicholson (CN) Method

Consider the heat equation with the following conditions:

$$\begin{aligned} (1)\ u_t = \sigma u_{xx} \quad & \text{on } x \in (\alpha, \beta) \\ u(x,0) = u_0(x) \quad & \text{I.C} \\ \begin{cases} u(\alpha,t) = f_l(t) \\ u(\beta,t) = f_r(t) \end{cases} \quad & \text{Diriclet B.C} \end{aligned}$$

and the scheme

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} = \frac{\sigma}{2}\left[\frac{U_{j+1}^{n+1} - 2U_j^{n+1} + U_{j-1}^{n+1}}{\Delta x^2} + \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{\Delta x^2}\right]$$

Rewrite CN as

$$(2)\ U_j^{n+1} - \frac{r}{2}\left(U_{j-1}^{n+1} - 2U_j^{n+1} + U_{j+1}^{n+1}\right) = U_j^n + \frac{r}{2}\left(U_{j-1}^n - U_j^n + U_{j+1}^n\right), 1 \leq j \leq J-1$$

$$U_0^n = f_l(t_n) = f_l^n; U_J^n = f_r(t_n) = f_r^n$$

Rewrite (2) as a matrix where (2) is a system of $J-1$ linear equations. Let

$$C_{ij} = M_{ij} = \begin{cases} 1 & |i-j| = 1 \\ -2 & i = j \\ 0 & \text{otherwise} \end{cases}, f^n = \begin{cases} f_l^n & i = 1 \\ f_r^n & i = J \\ 0 & \text{otherwise} \end{cases}$$

Then the system (2) can be rewritten as

$$\begin{aligned} U^{n+1} - \frac{r}{2}CU^{n+1} - \frac{r}{2}f^{n+1} &= U^n + \frac{r}{2}CU^n + \frac{r}{2}f^n \\ \implies \underbrace{\left(I - \frac{r}{2}C\right)}_{A} U^{n+1} &= \underbrace{\left(I + \frac{r}{2}C\right)U^n + \frac{r}{2}\left(f^n + f^{n+1}\right)}_{F} \\ \implies AU^{n+1} &= F \end{aligned}$$

and the last equation is solvable using linear algebra methods. Remark that A is a sparse tridiagonal matrix with $\sim 3(J-1)$ non-zero elements. This will make A^{-1} dense with $(J-1)^2$ non-zero elements.

Tridiagonal Algorithm

Suppose we have $AX = F$ with $A \in \mathbb{R}^{N\times N}$ being tridiagonal. Consider the LU decomposition

$$(1)\ L\underbrace{UX}_{y} = F, A = LU \implies (2)\ Ly = F,\ (3)\ UX = y$$

Suppose that

$$A_{ij} = \begin{cases} b_{ij} & j-i=1 \\ a_{ij} & i=j \\ c_{ij} & i-j=1 \\ 0 & \text{otherwise} \end{cases} \implies L_{ij} = \begin{cases} l_{ij} & j-i=1 \\ 1 & i=j \\ 0 & \text{otherwise} \end{cases}, U_{ij} = \begin{cases} u_{ij} & i=j \\ v_{ij} & i-j=1 \\ 0 & \text{otherwise} \end{cases}$$

We need to find u_j, v_j, l_j. The first row and first column, which we denote by $R_1 \cdot C_1$, of the LU decomposition implies

$$u_1 \cdot 1 = a_1$$

Similarly,

$$R_2 \cdot C_1 \implies l_2 u_1 = b_2 \implies l_2 = b_2/u_1$$

and in general,

$$R_j \cdot C_1 \implies l_j = b_j/u_{j-1}$$

With the first row and second column, using the same logic,

$$R_1 \cdot C_2 \implies v_1 \cdot 1 = c_1$$

$$R_2 \cdot C_j \implies l_2 v_1 + u_2 = a_2 \implies u_2 = a_2 - l_2 v_1$$

$$R_j \cdot C_2 \implies u_j = a_j - l_j v_{j-1}$$

This is LU factorization that takes into account the sparsity of A.

Note 4. Pivoting is usually not necessary for matrices arising in FD & FE (**finite element**) discretizations. For the number of operations, we have

$$\begin{aligned} 3 \text{ arithmetic} \times N &\sim 3N \\ (3 \text{ assignments of values} &\sim 3N) \end{aligned}$$

So it is $O(N)$ in the number of operations. Forward substitution for the heat equation gives us

$$y_j = f_j - l_j y_{j-1}$$

and we can use backward substitution to solve $UX = y$.

Summary 1. We have the following description of our discretizations:

	Explicit	Implicit	CN
Accuracy	$O(\Delta x^2, \Delta t)$	$O(\Delta x^2, \Delta t)$	$O(\Delta x^2, \Delta t^2)$
Stability	$\frac{\sigma \Delta t}{\Delta x^2} < \frac{1}{2}$	Unconditionally Stable	Unconditionally Stable
Work Per Time Step	$O(J)$	$O(J)$	$O(J)$
Total Work	$O(J \times N) = O(J^3)$	$O(J \times N) = O(J^2)$	$O(J \times N) = O(J^2)$
Reason (above)	$r < \frac{1}{2} \implies \Delta t \sim \Delta x^2$	since$\Delta t \sim \Delta x$	since$\Delta t \sim \Delta x$

Note 5. Even though CN is unconditionally stable, Δt should be about Δx for:

1) Accuracy

2) Convergence of iterative solvers for $AX = F$

Conclusion 2. Implicit schemes are more efficient per time step[3], but they might be more efficient (more than explicit schemes) overall if the number of time steps is smaller.

Boundary Conditions

1. Dirichlet B.C. are B.C. on $u(x,t)$
2. Neumann B.C. are B.C. on u_x
   (a) $u_x(-1,t) = u_x(1,t) = 0$ imply that the ends are insulated and no heat enters or leaves
   (b) Consider $u_x(\alpha,t) = g_l(t)$ for $u_t = \sigma u_{xx}$
   (c) Method 1:
      i. $(U_x)_0^n = g_l^n$, $\frac{U_1^n - U_0^n}{\Delta x} = g_l^n \implies U_0^n = U_1^n - \Delta x g_l^n$
      ii. At $j = 1$,

$$\begin{aligned} U_1^{n+1} = U_0^n - 2U_1^n + U_2^n &= U_1^n - \Delta x g_l^n - 2U_1^n + U_2^n \\ &= -U_1^n + U_2^n - \Delta g_l^n \end{aligned}$$

      iii. The first line in C is $AU + f$ where

$$A_{ij} = \begin{cases} -1 & (i,j) = (1,1) \\ 1 & |i-j| = 1 \\ -2 & i = j \end{cases}, f_i = \Delta x g_l^n$$

   (d) Method 2:
      i. Use higher order approximation for u_x where

$$(U_x)_0^n = \frac{-3U_0^n + 4U_1^n - U_2^n}{2\Delta x} = g_l^n \implies U_0^n = \frac{-(2\Delta x g_l^n - 4U_1^n + U_2^n)}{3}$$

      Eliminating U_0^n from the approximation of $(*)$ u_{xx} as in $(*)$
   (e) Method 3 (**Ghost Cell Approach**):
      i. Create [imaginary] cells (grid points) $j = -1, j = J+1$ and use

$$(U_x)_0^n = g_l(t_n) \approx \underbrace{\frac{U_1^n - U_{-1}^n}{2\Delta x}}_{O(x^2)} \implies U_{-1}^n = U_1^n 2 - \Delta x g_l^n$$

      and

$$(U_{xx})_0^n \approx \frac{U_{-1}^n - 2U_0^n + U_1^n}{\Delta x^2} = \frac{U_1^n - 2\Delta x g_l^n - 2U_0^n + U_1^n}{\Delta x^2}$$

      ii. Modify the first row of C. Also, note that we have $J+1$ unknowns: $U_0^n, ..., U_J^n$
3. Robin B.C. is in the form $\alpha u(1,t) + \beta u_x(1,t) = f(t)$
4. Mixed B.C. is when you have one half Dirichlet and one half Neumann

[3]This is due to the fact that in implicit schemes, you need to create a system of equations and solve it per time step. You will need to code this and it WILL take quite a bit of time. So efficiency here refers to amount of time invested.

2.6 Higher Dimensions

Consider the two dimensional heat equation

$$u_t = \sigma(u_{xx} + u_{yy}), U^n_{j,k} \approx u(x_j, y_k, t_n)$$

The basic explicit discretization, which is 2nd order in space, and 1st order in time is

$$\frac{U^{n+1}_{j,k} - U^n_{j,k}}{\Delta t} = \sigma \frac{U^n_{j+1,k} - 2U^n_{j,k} + U^n_{j-1,k}}{\Delta x^2} + \sigma \frac{U^n_{j,k+1} - 2U^n_{j,k} + U^n_{j,k+1}}{\Delta x^2}$$

This scheme is stable if

$$r_x + r_y < \frac{1}{2}, r_x = \frac{\sigma \Delta t}{\Delta x^2}, r_y = \frac{\sigma \Delta t}{\Delta y^2}$$

Assuming that $\Delta x = \Delta y$, then

$$\frac{\sigma \Delta t}{\Delta x^2} < \frac{1}{4}$$

Compare work in 1D and 2D:

- 1D: J points in space $\times$ N layers in time $= JN$
- 2D: J^2 points in space $\times$ N layers in time $= J^2 N$

In 3D this becomes $J^3 N$.

2.7 Crank-Nicolson and ADI Methods

This is using the definition of the operator

$$\delta^2_x U^n_{j,k} := U^n_{j-1,k} - 2U^n_{j,k} + U^n_{j+1,k}$$

with the scheme

$$\frac{U^{n+1}_{j,k} - U^n_{j,k}}{\Delta t} = \sigma \left(\frac{\delta^2_x (U^n_{j,k} + U^{n+1}_{j,k})}{2\Delta x^2} + \frac{\delta^2_y (U^n_{j,k} + U^{n+1}_{j,k})}{2\Delta y^2} \right)$$

which simplifies to

$$\left(1 - \frac{r_x \delta^2_x + r_y \delta^2_y}{2}\right) U^{n+1}_{j,k} = \left(1 + \frac{r_x \delta^2_x + r_y \delta^2_y}{2}\right) U^n_{j,k}, r_x = \frac{\Delta t \sigma}{\Delta x^2}$$

How do we organize $U^n_{j,k}$ into a vector U^n?

1. By row $U_{11}, U_{12}, ..., U_{21}, U_{22}, ...$
2. By column $U_{11}, U_{12}, ..., U_{12}, U_{22}, ...$
3. Any way you want!

There is no way to create U^n so that C is tridiagonal. Modify CN with

$$\left(1 - \frac{1}{2} r_x \delta^2_x\right)\left(1 - \frac{1}{2} r_y \delta^2_y\right) U^{n+1} = \left(1 + \frac{1}{2} r_x \delta^2_x\right)\left(1 + \frac{1}{2} r_y \delta^2_y\right) U^n$$

Introduce intermediate $U^{n+\frac{1}{2}}$ with

$$\left(1 - \frac{1}{2} r_x \delta^2_x\right) U^{n+\frac{1}{2}} = \left(1 + \frac{1}{2} r_y \delta^2_y\right) U^n - O(J)$$

and

$$\left(1 - \frac{1}{2} r_y \delta^2_y\right) U^{n+1} = \left(1 + \frac{1}{2} r_x \delta^2_x\right) U^{n+\frac{1}{2}} - O(J)$$

The accuracy is then

$$\left(1-\frac{1}{2}r_x\delta_x^2\right)\left(1-\frac{1}{2}r_y\delta_y^2\right)=\underbrace{1+\frac{1}{2}r_x\delta_x^2+\frac{1}{2}r_y\delta_y^2}_{CN}+\frac{1}{4}r_xr_y\delta_x^2\delta_y^2$$

Summary 2. Here is a summary of the all the finite difference methods for the linear advection equation

$$u_t+au_x=0, u(x,0)=u_0(x)$$

that we've learned (and an extra one):

- Upwind Method:
$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_j^n-u_{j-1}^n}{\Delta x}=0, a>0$$
has error $O(\Delta t,\Delta x)$
- Central Method
$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_{j+1}^n-u_{j-1}^n}{\Delta 2x}=0, a>0$$
has error $O(\Delta t,\Delta x^2)$ but the solution grows instead of decays (very bad)
- Lax-Friedrichs Method:
$$\frac{u_j^{n+1}-\frac{1}{2}\left(u_{j-1}^n+u_{j+1}^n\right)}{\Delta t}+a\frac{u_{j+1}^n-u_{j-1}^n}{\Delta 2x}=0$$
has error $O(\Delta t,\Delta x^2,\Delta x^2/\Delta t)$ and doesn't depend on sgn(α) which makes it better than the upwind method in some sense
- Leapfrog Method:
$$\frac{u_j^{n+1}-u_j^{n-1}}{2\Delta t}+a\frac{u_{j+1}^n-u_{j-1}^n}{\Delta 2x}=0$$
has error $O(\Delta t^2,\Delta x^2)$ and is known as a multilevel or multistep scheme
  - This the most accurate so far but uses data at n and $n-1$ which requires two starting values U^0 and U^1 and hence needs 2x more memory
  - If you analyze leapfrog, you'll realize why weather predictions are so bad
- Lax-Wendroff (NEW! And still alive!):
$$u_j^{n+1}\approx u_j^n-\Delta ta\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}+\frac{\Delta t^2}{2}a^2\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}$$
where this is derived using the Taylor series expansion of $u(x,t+\Delta t)$
  - If $\alpha=a\Delta t/\Delta x$ then the above can be written as
$$u_j^{n+1}=u_j^n-\frac{\alpha}{2}(u_{j+1}^n-u_{j-1}^n)+\frac{\alpha^2}{2}(u_{j+1}^n-2u_j^n+u_{j-1}^n)$$

All of these schemes are stable when $|a|\Delta t/\Delta x\leq 1$ with $\tau_j^n=(\Delta x^2,\Delta t^2)$.

2.8 Dissipation and Dispersion Error

Definition 2.6. We define **dissipation** as the dying of signal over time and **dispersion** as when a wave travels at the wrong speed. We use these definition as tools to compare schemes.

Example 2.3. Consider the upwind method

$$(*)\ \frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_j^n-u_{j-1}^n}{\Delta x}=0, a>0$$

and rewrite the second term as

$$\frac{u_j^n-u_{j-1}^n}{\Delta x}\pm\frac{u_{j+1}^n}{2\Delta x}=\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}-\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}\cdot\frac{\Delta x}{2}$$

and hence

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}=\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}\cdot\frac{\Delta x}{2}$$

is an approximation to $u_t+au_x=\frac{\Delta x}{2}u_{xx}$ where $(*)$ has a "hidden" diffusion term. We call this "numerical (artificial) diffusion" or "dissipation". As $\Delta x\to 0$, there is less diffusion. Through more work, we can find the u_{xxx} term.

In general, we can write our schemes like this:

$$(2)\ u_j^{n+1}=u_j^n-\frac{\alpha}{2}(u_{j+1}^n-u_{j-1}^n)+\frac{\beta}{2}(u_{j+1}^n-2u_j^n+u_{j-1}^n)$$

for

$$(1)\ u_t+au_x=0$$

where $\alpha=\frac{a\Delta t}{\Delta x}$ and the β values are based on the scheme and are summarized below:

Scheme	β
Central	$\beta=0$
Upwind	$\beta=\lvert\alpha\rvert$
Downwind	$\beta=-\lvert\alpha\rvert$
Lax-Friedrichs	$\beta=1$
Lax-Wendroff	$\beta=\alpha^2$

Assuming periodic boundary conditions, we can use Von Neumann analysis as follows:

$$\begin{aligned}
A_k^{n+1} &= A_k^n\left(1-\frac{\alpha}{2}(w^k-w^{-k})+\frac{\beta}{2}(w^k-2+w^{-k})\right)\\
&= A_k^n\left((1-\beta)+\frac{1}{2}(\beta-\alpha)w^k+\frac{1}{2}(\beta+\alpha)w^{-k}\right)\\
&= A_k^n\left(1-\beta-\alpha i\sin\left(\frac{2\pi k}{J}\right)+\beta\cos\left(\frac{2\pi k}{J}\right)\right)\\
&= A_k^n\left(1-2\beta\sin^2\left(\frac{\pi k}{J}\right)+\alpha i\sin\left(\frac{2\pi k}{J}\right)\right)
\end{aligned}$$

So the real part is the diffusion and the imaginary part the is the drift. Note that

$$|A_k^{n+1}|^2=|A_k^n|\left(1-2\beta\sin^2\left(\frac{\pi k}{J}\right)+4\beta\sin^4\left(\frac{\pi k}{J}\right)+\alpha^2\sin^2\left(\frac{2\pi k}{J}\right)\right)$$

where

$$|M_k|^2=1-2\beta\sin^2\left(\frac{\pi k}{J}\right)+4\beta\sin^4\left(\frac{\pi k}{J}\right)+\alpha^2\sin^2\left(\frac{2\pi k}{J}\right)$$

To see the above remark, let $\theta=\frac{k\pi}{J}$ where

$$M_k=1-2\beta\sin^2\theta-i\alpha\sin 2\theta$$

Here, $\Re(M_k)$ is responsible for the change in amplitude and $\Im(M_k)$ for phase shift. Note for the exact solution $\Re(M_k)=1$. To simplify analysis, assume I.C. are such that $U_j^0=w_j^k$ for some fixed k and numerical I.C. for (2). Assuming $x\in[0,1]$ then

$u(x,0) = e^{2\pi kxi}$ is the exact I.C. for (1). Now as above,

$$\begin{aligned} |M_k|^2 &= 1 - 4\beta \sin^2\theta + 4\beta \sin^4\theta + \alpha^2 \sin^2 2\theta \\ &= 1 - 4\beta \sin^2\theta + 4\beta^2(1-\cos^2\theta)\sin^2\theta + \alpha^2 \sin^2 2\theta \\ &= 1 + (\alpha^2 - \beta^2)\sin^2 2\theta - 4\beta(1-\beta)\sin^2\theta \end{aligned}$$

Definition 2.7. We say a scheme is **dissipative of order** $2r$ if

$$|M_k(\theta)| \leq 1 - C|\theta|^{2r}, 0 \leq \theta \leq \frac{\pi}{2}$$

where $C > 0$ and independent of Δx and Δt. Note that an exact solution is not dissipative.

Example 2.4. (Lax-Wendroff) For this scheme $\beta = \alpha^2$ and

$$|M_k|^2 = 1 + 4\alpha^2(1-\alpha^2)\sin^2\theta(\cos^2\theta - 1) = 1 - 4\alpha^2(1-\alpha^2)\sin^4\theta$$

So Lax-Wendroff is dissipative of order 4. This is because $x > \sin x$ for $x \in [0, \frac{\pi}{2}]$.

(Upwind) Here $\beta = |\alpha|$ and

$$|M_k|^2 = 1 - 4|\alpha|(1-|\alpha|)\sin^2\theta$$

So Upwind is dissipative of order 2. You will see that the other schemes are also order 2.

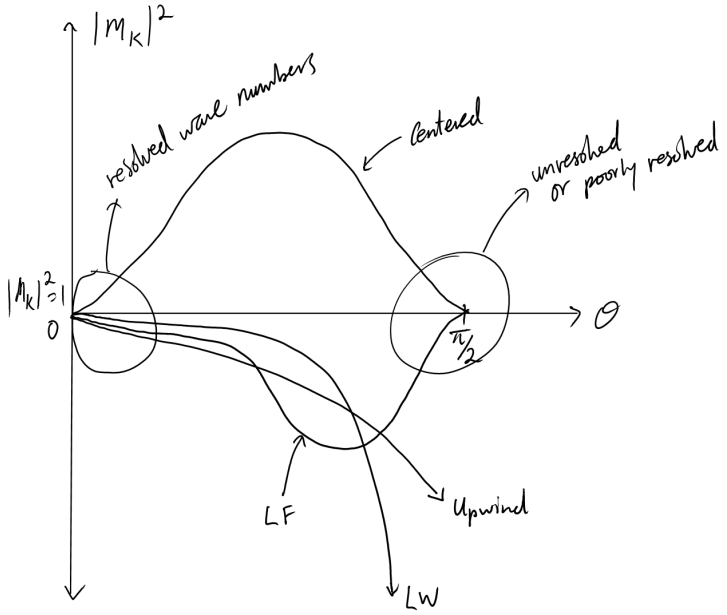
Conclusion 3. Lax-Wendroff (LW) is the least dissipative of the 4 schemes and hence is the "best" in some sense.

Note 6. For centered, $\beta = 0$ and $|M_k|^2 = 1 + \alpha^2 \sin^2\theta$, and for Lax-Friedrichs (LF), $\beta = 1$ with $|M_k|^2 = 1 + (\alpha^2 - 1)\sin^2 2\theta$.

Remark 2.6. Recall that $\theta = \frac{\pi k}{J}$ with k fixed. A small θ means $J \gg k \implies$ the grid is fine relative to k. For example, consider the I.C.'s

$$\begin{cases} u(x,0) = \sin \pi x & J = 10 \implies \theta = \frac{\pi}{10} \\ u(x,0) = \sin 10\pi x & J = 10 \implies \theta = \frac{10\pi}{10} \end{cases}$$

We call the first example a **resolved wave number**. For resolved wave numbers, we have have $|M_k| \approx 1$ for all schemes (we approximate them well). For large θ, say about $\frac{\pi}{2} \implies J \approx 2k$ since $\frac{\pi}{2} \approx \theta = \frac{\pi k}{J}$, we call these unresolved waves. LF and LW dissipate high frequency waves w.r.t the mesh waves (i.e. dumps them):



Remark 2.7. This is good, because they are largely numerical noise.

Note 7. For $\alpha = 1$ we have no amplitude loss and we have an exact translation.

Remark 2.8. Fix Δx. A smaller Δt is not necessarily better since this makes α very tiny.

Dissipation Error

Rewrite $M_k = |M_k|e^{i\phi_k}$ and where $-\alpha = \phi_k$. Note that

$$(1)\ \tan\phi = -\tan\alpha = \frac{\alpha\sin\theta}{1-2\beta\sin^2\theta}$$

Then

$$\begin{aligned}(*)\ U_j^n &= |M_k|^n e^{-in\phi_k} e^{2\pi k\frac{j}{J}i} \\ &= |M_k| e^{(2\pi k\frac{j}{J} - n\phi_k)i} \\ &= M_k e^{2\pi k(x_j - a_k t_n)i}\end{aligned}$$

where $t_n = n\Delta t$ and $a_k = \phi_k/(2\pi k\Delta t)$. In the linear advection equation, compare this with the exact solution $u(x,t) = e^{2\pi k(x-at)i}$ at (x_j, t_n) which is

$$u(x,t) = e^{2\pi k(x_j - at_n)i}$$

We see that a_k is the numerical wave speed. Usually $a_k \neq a$. Expanding (1) for ϕ_k, with the assumption that θ is small and

$$\begin{aligned}\tan^{-1} z &= z - \frac{z^3}{3} + O(z^5) \\ \sin x &= x - \frac{x^3}{6} + O(x^5) \\ \frac{1}{1-y} &= 1 + y + O(y^2)\end{aligned}$$

gives us

$$\begin{aligned}\phi_k &= \frac{\alpha\sin 2\theta}{1-2\beta\sin^2\theta} - \frac{1}{3}\left(\frac{\alpha\sin 2\theta}{1-2\beta\sin^2\theta}\right)^3 + ... \\ &= \underbrace{\alpha\left(2\theta - \frac{8\theta^3}{6} + ...\right)\cdot\left(1 + 2\beta\theta^2 + ...\right)}_{z} - \frac{z^3}{3} \\ &\approx 2\alpha\theta\left(1 - \frac{2}{3}\theta^2\right)\left(1 + 2\beta\theta^2\right) - \frac{8\alpha^3\theta^3}{3} \\ &\approx 2\alpha\theta\left(1 - \frac{2}{3}\theta^2 + 2\beta\theta^2 - \frac{4\alpha^2\theta^2}{3}\right) \\ &= \underbrace{2\alpha\theta\left(1 - \frac{2}{3}\theta^2\left[1 + 2\alpha^2 - 3\beta\right]\right)}_{Q(\theta)}\end{aligned}$$

where the third and fourth equations are keeping up to cubic terms. Now

$$a_k = \frac{\phi_k}{2\pi k\Delta t} \approx \frac{2\underbrace{\frac{\pi k}{J}}_{\theta}\cdot\underbrace{\frac{a\Delta}{\Delta x}}_{\alpha}Q(\theta)}{2\pi k\Delta} = aQ(\theta) = a\left(1 - \frac{2}{3}\theta^2\left[1 + 2\alpha^2 - 3\beta\right]\right)$$

So $a_k - a = O(\theta^2) = O(\Delta x^2)$. Therefore, a_k is an approximation of a.

Example 2.5. Consider LW with $\beta = \alpha^2$ and

$$a_k \approx a\left(1 - \frac{2\theta^2}{3}\left[\underbrace{1-\alpha^2}_{>0}\right]\right)$$

Then $a_k \leq a$. The numerical solution moves slower than $u(x,t)$. With LF where $\beta = 1$, we have

$$\begin{aligned} a_k &= a\left(1 - \frac{2\theta}{3}(-2 + 2\alpha^2)\right) \\ &= a\left(1 + \frac{4\theta^2}{3}(1 - \alpha^2)\right) \end{aligned}$$

and hence the numerical solution moves faster than the exact one.

3 Finite Volume Methods

Finite volume methods (FVM) are largely applied to nonlinear hyperbolic problems of the form

$$(1)\ u_t + f(u)_x = 0$$

where we call $f(u)$ the flux function. Assume that $f(u)$ is differentiable. (1) is called a conservative form while a non-conservative form looks like

$$(2)\ u_t + f_u u_x = 0$$

Compare this to the linear wave equation

$$u_t + au_x = 0 \implies u_t + (au)_x = 0 \implies f_u = a$$

In conclusion, f_u is the nonlinear wave speed where it is a usually a function of u but can also depend on x and t.

Example 3.1. (Burgers' Equation) Consider

$$(3)\ u_t + \frac{1}{2}(u^2)_x = u_t + uu_x = 0$$

where a modified equation has the form

$$u_t + \frac{1}{2}(u^2)_x = \varepsilon u_{xx}, \varepsilon \approx 0, \varepsilon \to 0 \text{ (viscous)}$$

There is no solution to (3) with arbitrary I.C. that can be expressed with basic functions.

3.1 Method of Characteristics

Start with $u_t + au_x = 0, -\infty < x < \infty$ with $u(x,0) = u_0(x)$. Consider $u(x,t)$ restricted to some curve $x(t)$ in the $x - t$ plane. We have $u(x(t),t)$ and differentiating w.r.t. t gives us

$$\frac{d}{dt}u(x(t),t) = \frac{\partial u}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial u}{\partial t} = u_t + \frac{dx}{dt}u_x$$

If $x(t)$ is such that $\frac{dx}{dt} = a$ then $\frac{d}{dt}(u(x(t),t)) = 0$ which implies that u is constant along this curve. The curve is called a **characteristic**. $\frac{dx}{dt} = a$ should be a line. We can back along a characteristic to find a solution using the initial condition. Since the equation of a characteristic passing through (x_1, t_1) is

$$x(t) = at + x_0$$

where $x_0 = x_1 - at_1$. Then, $u(x,t) = u(x_0, 0) = u_0(x - at)$ which is a function of our initial condition.

Example 3.2. Going back to Burgers' equation, suppose $x = x(t)$ for some curve. Then $u(x,t)$ is constant[4] on the curves with slope $\frac{dx}{dt} = u(x,t)$. But u is constant so $\frac{dx}{dt}$ is constant so we have lines again. Consider the I.C.

$$u_0(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}, -\infty < x < \infty$$

[4]This is because $\frac{du}{dt} = 0$ from the previous example.

This is called a **Riemann problem** (infinite domain, I.C. are two constant states). We have

$$\frac{dt}{dx} = \begin{cases} \infty & x < 0 \\ 1 & x \geq 0 \end{cases}$$

Compare this to what happens when we set

$$u(x,0) = \begin{cases} 1 & x \geq \gamma \\ \frac{x}{\gamma} & 0 < x < \gamma \\ 0 & x \leq 0 \end{cases}$$

Then we should expect

$$\frac{dt}{dx} = \begin{cases} 1 & x < 1 \\ \frac{\gamma}{x} & 0 < x < \gamma \\ \infty & x \geq \gamma \end{cases}$$

That is, we expect the slopes to continuously change from ∞ to 1. If we take $\gamma \to 0$ then

$$u(x,t) = \begin{cases} 1 & \frac{x}{t} \geq 1 \\ \frac{x}{t} = \xi & 0 < \xi < 1 \\ 0 & x \leq 0 \end{cases}$$

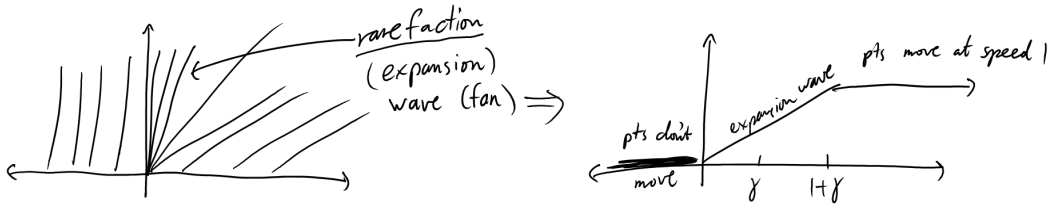
and the graph will be like a linear interpolation of the slopes as $x \to 0$. We call this phenomenon **rarefaction** (expansion) wave (fan). Next, consider

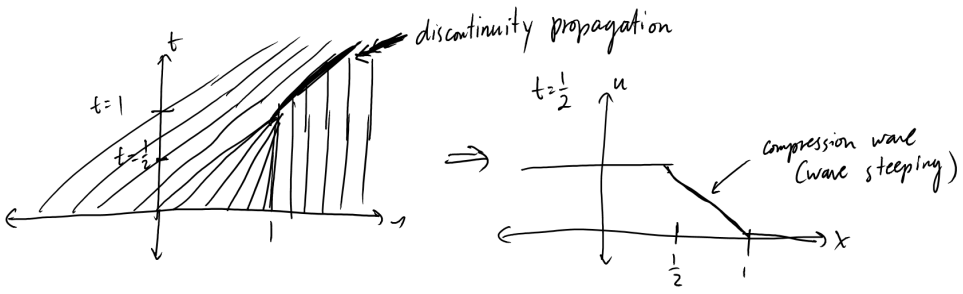
$$u(x,0) = \begin{cases} 1 & x < 0 \\ 1 - x & 0 \leq x \leq 1 \\ 0 & x > 0 \end{cases}$$

Then we have

$$\frac{dt}{dx} = \lim_{\gamma \to 0} \begin{cases} 1 & x < 1 \\ \frac{\gamma}{1-x} & 0 < x < \gamma \\ \infty & x \geq \gamma \end{cases}$$

This will create what is called a **compression wave** (wave steeping) with a shock / discontinuity at the limit of the compression. Here are some pictures for illustrative purposes:





Question. What is the slope of the intersecting characteristic that is formed. That is, with what speed does the discontinuity propagate?

3.2 Rankine-Hugoniot Condition

Consider a nonlinear conservation law

$$\begin{aligned} (1)\ u_t + f(u)_x &= 0 \\ f(u) &= \text{flux} \end{aligned}$$

For Burgers' equation, $f(u) = u^2/2$. Let $\xi(t)$ be the location of the shock (unknown) at time t where we have some continuous curves on the left and right. Let $\dot{\xi} = \frac{d\xi}{dt}$. We then integrate (1) on $[\alpha, \beta]$ to get

$$\begin{aligned} \int_\alpha^\beta u_t\, dx + \int_\alpha^\beta f(u)_x\, dx = 0 \quad &\implies \quad \frac{d}{dt}\left(\int_\alpha^\beta u\, dx\right) = -f(u)\Big|_\alpha^\beta \\ &\implies \quad \frac{d}{dt}\left(\int_\alpha^{\xi^-(t)} u\, dx + \int_{\xi^+(t)}^\beta u\, dx\right) = -f(u)\Big|_\alpha^\beta \\ &\implies \quad \left[\frac{d\xi^-}{dt}\right] u^- + \int_\alpha^{\xi^-(t)} u_t\, dx - \left[\frac{d\xi^+}{dt}\right] u^+ + \int_{\xi^+(t)}^\beta u_t\, dx = -f(u)\Big|_\alpha^\beta \end{aligned}$$

where $u^- = \lim_{x\to\xi(t)^-} u(x,t)$ and $u^+ = \lim_{x\to\xi(t)^+} u(x,t)$. The first Now we expect

$$\frac{d\xi^-}{dt} = \frac{d\xi^+}{dt} = \frac{d\xi}{dt} = \dot{\xi}$$

That is, the left shock and right shock move at the same speed. Take $\alpha \to \xi^-$ and $\beta \to \xi^+$ to get

$$-\dot{\xi}(u^+ - u^-) + \underbrace{\int_\alpha^{\xi^-(t)} u_t\, dx + \int_{\xi^+(t)}^\beta u_t}_{\to 0} = \underbrace{f(u)\Big|_\alpha - f(u)\Big|_\beta}_{f(u^-)-f(u^+)}$$

and hence

$$\dot{\xi} = \frac{f(u^+) - f(u^-)}{u^+ - u^-}$$

which we call the **Rankine-Hugoniot condition**. Here we can see that shock speed depends on values of u on the left and right of it as well as the flux.

Example 3.3. In Burgers' equation,

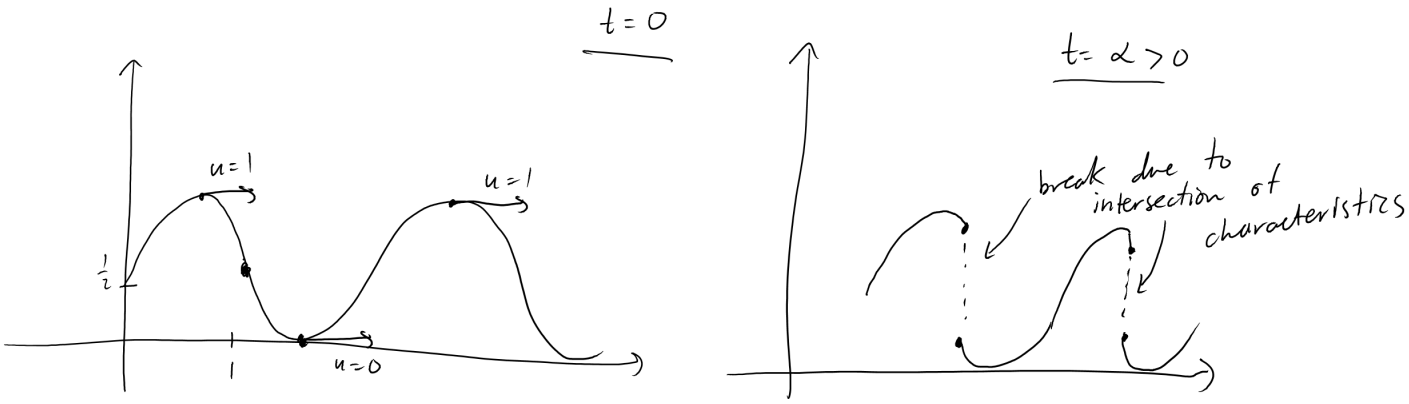
$$f(u) = \frac{u^2}{2} \implies \dot{\xi} = \frac{(u^+)^2/2 - (u^-)^2/2}{u^+ - u^-} = \frac{1}{2}(u^+ + u^-)$$

Note 8. The shock moves with its own speed $\neq$ speed to the left or right of it.

Example 3.4. Consider Burgers' equation with the I.C.

$$u(x,0) = \frac{1}{2}(1 + \sin \pi x)$$

The peaks move at speed 1 while the troughs stay in place (compression):



3.3 System of Hyperbolic Equations

Consider the vectorized equation

$$(2)\ u_t + f(u)_x = 0, u \in \mathbb{R}^n, f : \mathbb{R}^n \mapsto \mathbb{R}^n$$

Some examples include Euler equations, Maxwell's equation, and shallow water. Consider the special case

$$u_t + f_u u_x = 0$$

where $f_u = \nabla f$, the Jacobian matrix. (2) is hyperbolic if f_u has n real eigenvalues and a full set of eigenvectors. Consider the linear case

$$(3)\ u_t + Aux = 0$$

where $AV = V\Lambda$, V is a matrix of eigenvectors and $\Lambda = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ and $V^{-1}AV = \Lambda$. Multiply (3) by V^{-1} to get

$$V^{-1}u_t + V^{-1}AVV^{-1}u_x = 0$$

Let $w = V^{-1}u \implies w_t = V^{-1}u_t$ since V^{-1} is constant. So

$$w_t + \Lambda w_x = 0$$

Therefore, hyperbolic equations are really a combination of n scalar waves.

Example 3.5. (2nd Order Wave Equation) Consider

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

and let $u_1 = u_t, u_2 = cu_x$. Then

$$u_{1t} = u_{tt} = c^2 u_{xx} = c(cu_x)_x = cu_{2x}$$

$$u_{2t} = cu_{xt} = c(u_t)_x = cu_{1x}$$

$$\implies \begin{cases} u_{1t} - cu_{2x} &= 0 \\ u_{2t} - cu_{1x} &= 0 \end{cases} \implies \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \underbrace{\begin{pmatrix} 0 & -c \\ -c & 0 \end{pmatrix}}_{=A} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x = 0\ (*)$$

This system is hyperbolic if A is diagonalizable. Since

$$\begin{vmatrix} -\lambda & -c \\ -c & -\lambda \end{vmatrix} = \lambda^2 - c^2 = 0 \implies \lambda_{12} = \pm c \implies (*) \text{ is hyperbolic}$$

Example 3.6. (Shallow water equations) Consider the system:

$$\begin{cases} h_t + hu &= 0 \\ (hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x &= 0 \end{cases}$$

where we interpret $h(x,t)$ as the height of the wave at the point x at time t, and $u(x,t)$ is the velocity. The variable $g > 0$ is

the gravitational constant. Let $q_1 = h, q_2 = hu$ and remark that

$$\begin{cases} q_{1,t} + q_{2,x} & = 0 \\ q_{2,t} + \left(\frac{q_2^2}{q_1} + \frac{1}{2}gq_1^2\right)_x & = 0 \end{cases} \implies \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}_t + \underbrace{\begin{pmatrix} q_2 \\ \frac{q_2^2}{q_1} + \frac{1}{2}gq_1^2 \end{pmatrix}_x}_{\text{flux function for } f} = 0$$

Now

$$f_q = \frac{\partial(f_1, f_2)}{\partial(q_1, q_2)} = \begin{pmatrix} 0 & 1 \\ -\frac{q_2^2}{q_1^2} + gq_1 & \frac{2q_2}{q_1} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -u^2 + gh & 2u \end{pmatrix}$$

and the eigenvalues for f_q can be calculated as follows.

$$\begin{vmatrix} -\lambda & 1 \\ -u^2 + gh & 2u - \lambda \end{vmatrix} = \lambda^2 - 2u\lambda + u^2 + gh = 0 \implies \lambda_{12} = u \pm \sqrt{gh}, g > 0, h > 0$$

So the system is hyperbolic.

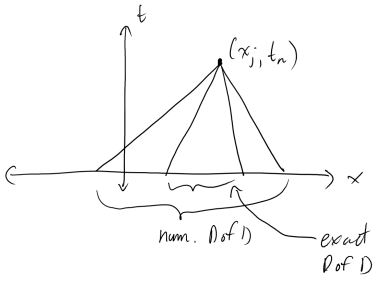
3.4 Domain of Dependence

Domain of Dependence

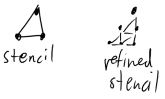
Theorem 3.1. *The Exact D of D ⊂ Numerical D of D for consistency of a numerical scheme (original CFL condition).*

Proof. Assume the condition is not true:

Proof of Exact D of D ⊂ Numerical D of D :



- Choose special I.C. s.t.
  $u(x_j, 0) = 0$ in Num. D of D
  $u(x_j, 0) = 1$ otherwise
- Then the num. sol. at (x_j, t_n) is 0 as $\Delta x \to 0$ (assuming fixed α)
  * This is because refinement doesn't change num. D of D :



Note that $u = 0$ at (x_j, t_n) and we have no convergence, which is impossible. □

From this diagram, we should have

$$\Delta x/\Delta t \leq |\lambda_i|, i = 1, 2, ..., n$$

or alternatively

$$\Delta x/\Delta t \leq |\lambda_I|, I = \text{argmax}_i|\lambda_i|$$

3.5 Discontinuous and Weak Solutions

Example 3.7. Consider

$$u_t + au_x = 0, t > 0$$

$$u(x, 0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases}$$

According to the method of characteristics, $u(x, t) = u_0(x - at)$. Should we accept the discontinuous solution as a solution? For practical purposes, yes. A physical example would be a shockwave near or at a wing of an airplane. We still need to find a mathematical way of dealing with them.

Weak Solutions

Consider the PDE

$$(1)\ u_t + f(u)_x = 0$$

Multiply by a smooth function $v(x,t)$ and integrate on $[x_1,x_2]\times[t_1,t_2]\subset[\alpha,\beta]\times[0,1]$, where $[\alpha,\beta]$ is the domain of (1), to get

$$(2)\ \int_{t_1}^{t_2}\int_{x_1}^{x_2}(u_t+f(u)_x)v\ dx\ dt = 0$$

or more generally,

$$(3)\ \int_0^\infty\int_{-\infty}^{\infty}(u_t+f(u)_x)v\ dx\ dt, u=0 \text{ outside of } [\alpha,\beta]$$

Integrate by parts to obtain

$$-\int_0^\infty\int_{-\infty}^{\infty}uv_t\ dx\ dt+\int_{-\infty}^{\infty}uv\ dx\Big|_{t=0}^{t=\infty}-\int_0^\infty\int_{-\infty}^{\infty}f(u)v_x\ dx\ dt+\int_0^\infty uv\ dt\Big|_{x=-\infty}^{x=\infty}=0$$

Let's require v to decay at infinity. That is, $v(\pm\infty,t)=0$ and $v(x,\infty)=0$. This makes the last integral equal to 0 and

$$\int_{-\infty}^{\infty}uv\ dx\Big|_{t=0}^{t=\infty}=-\int_{-\infty}^{\infty}u(x,0)v(x,0)\ dx$$

Hence,

$$(4)\ \int_0^\infty\int_{-\infty}^{\infty}(uv_t+f(u)v_x)\ dx\ dt=-\int_{-\infty}^{\infty}u(x,0)v(x,0)\ dx$$

A function $u(x,t)$ that satisfies (4) is called a **weak solution** of (1). We call $v(x)$ a **test function** and it should be differentiable. Note that (4) should be satisfied with the properties $v(\pm\infty,t)=v(x,\infty)=0, v\in\mathcal{C}^1$ and hence it is valid for discontinuous u. The whole point was to shift derivative from $\underline{u}$ to v.

Now, $u(x,t)$ that solves (1) is called a **strong solution** and a strong solution is also a weak solution, but not the converse in general.

3.6 Godunov Schemes

We want to model the problem

$$(1)\ u_t + f(u)_x = 0$$

using the ideas of weak solutions and the method of characteristics (Riemann problems). Remark that this is equivalent to

$$u_t + f_u u_x = 0$$

Before, we had the pointwise approximation $U_j^n\approx u(x_j,t_n)$, but now we will denote $U_j^n\approx$ the average of $u(x,t)$ at $t=t_n$ on an interval $I_j=(x_{j-\frac{1}{2}},x_{j+\frac{1}{2}})$, call it $\bar{u}_j^n$ where

$$\bar{u}_j^n=\frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}u(x,t_n)\ dx, \Delta x_j := x_{j+\frac{1}{2}}-x_{j-\frac{1}{2}}$$

So U^n is a piecewise constant approximation to $\bar{u}^n$. Let's integrate (1) on $[x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}]\times[t_n,t_{n+1}]$ to get

$$\begin{aligned}0=\int_{t_n}^{t_{n+1}}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}(u_t+f(u)_x)\ dx\ dt &= \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}u\Big|_{t_n}^{t_{n+1}}\ dx+\int_{t_n}^{t_{n+1}}f(u)\Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}\ dt\\ &= \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}\left[u(x,t_{n+1})-u(x,t_n)\right]\ dx+\int_{t_n}^{t_{n+1}}\Big[f(u(x_{j+\frac{1}{2}},t))-f(u(x_{j-\frac{1}{2}},t))\Big]\ dt\\ &= (\bar{u}_j^{n+1}-\bar{u}_j^n)\Delta x_j+\Delta t_n\left(f(u_{j+\frac{1}{2}}^*)-f(u_{j-\frac{1}{2}}^*)\right)\end{aligned}$$

where the (*) values still need to be determined. This gives us the numerical scheme

$$(2)\ U_j^{n+1} = U_j^n - \frac{\Delta t_n}{\Delta x_j}\left(F^n_{j+\frac{1}{2}} - F^n_{j-\frac{1}{2}}\right), F^n_{j+\frac{1}{2}} \approx \frac{1}{\Delta t_n}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\, dt$$

where we call $F^n_{j+\frac{1}{2}}$ the numerical flux. What we need now is the values of $f(U)$ on the line $x = x_{j+\frac{1}{2}}$ and $t_n \le t \le t_{n+1}$. We call $x = x_{j+\frac{1}{2}}$ an **interface** between cells I_j and I_{j+1}. We get this from solving a Riemann problem. In particular, we examine

$$u(x,0) = \begin{cases} U_j^n & x < 0 \\ U_{j+1}^n & x \ge 0 \end{cases}$$

This is a mental shift of $x_{j+\frac{1}{2}} \to 0$, $t_n \to 0$. We use this value to find $F_{j+\frac{1}{2}}$. Let's assume that Δt is small enough so that solutions from $x_{j+\frac{1}{2}}$ and $x_{j-\frac{1}{2}}$ don't intersect. That is we want only U_{j-1} and U_j to determine solutions along $x = x_{j-\frac{1}{2}}$ with no input from U_{j+1}. For classical **Godunov schemes**,

$$\Delta t \le \frac{\Delta x}{2\lambda},\ \lambda = \max_{1\le i\le m, j} |\lambda_i|$$

where we have m equations j refers to the cells. From modern schemes,

$$\Delta t \le \frac{\Delta x}{\lambda}$$

is usually good enough. $F_{j+\frac{1}{2}}$ is equation (or f) specific. For all common equations, $F_{j+\frac{1}{2}}$ has been derived.

Example 3.8. Consider the numerical flux, a.k.a., a Riemann solver for the Burgers equation:

$$F^n_{j+\frac{1}{2}} = F(U_j^n, U_{j+1}^n) = F(U_L, U_R) = f(U^*)$$

where U^* is to be determined and $U_L = U_j^n, U_R = U_{j+1}^n$. Breaking this down into 5 cases, we have

$$U^* = \begin{cases} U_L & \text{if } U_L > 0, U_R > 0 \\ U_R & \text{if } U_L < 0, U_R < 0 \\ 0 & \text{if } U_L < 0, U_R > 0 \\ U_L & \text{if } U_L > 0, U_R < 0, \frac{U_L+U_R}{2} > 0 \\ U_R & \text{if } U_L > 0, U_R < 0, \frac{U_L+U_R}{2} < 0 \end{cases}$$

where $U_L = U_R$ are the slopes of the characteristics for Burgers' equation. Note that the last two cases are because

$$\dot{\xi} = \frac{U_L + U_R}{2}$$

and hence $F_{j+\frac{1}{2}} \approx \int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))\, dt = \int_{t_n}^{t_{n+1}} f(u^*)\, dt = \Delta t_n f(u^*)$. Note that U^* is the specific U in (U_j^n, U_{j+1}^n) in the integral mean value theorem seen previously. The results above follow especially because we choose Δt so information propagates only at most one of half a space step.

Definition 3.1. Rewrite (2) in the form of (1) to get

$$(3)\ \frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{1}{\Delta x}\left(F^n_{j+\frac{1}{2}} - F^n_{j-\frac{1}{2}}\right) = 0$$

To simplify discussion, assume that $f_u > 0$ (at least locally, near (j,n)). Then, assume

$$\begin{cases} F_{j+\frac{1}{2}} = F(U_j) & \Longrightarrow \text{Upwind} \\ F_{j-\frac{1}{2}} = F(U_{j-1}) \\ u_t + \underbrace{f_u}_{>0} u_x = 0 \end{cases}$$

Assume $u(x,t)$ is smooth (no discontinuities). Near discontinuities we do truncation analysis. Plug $u(x,t) \to (3)$ and get

$$\frac{\bar{u}_j^{n+1} - \bar{u}_j^n}{\Delta t} + \frac{1}{\Delta x}\left(f(\bar{u}_j^n) - f(\bar{u}_{j-1}^n)\right) = 0$$

Expand $u(x,t_n)$ into Taylor series about x_j to get

$$u(x,t_n) = u_j^n + (u_x)_j^n(x-x_j) + \frac{1}{2}(u_{xx})_j^n(x-x_j)^2 + ...$$

and

$$\begin{aligned}
\bar{u}\Big|_{I_j} = \bar{u}_j^n &= \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_n)\,dx \\
&= \frac{1}{\Delta x}u_j^n\left(x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}\right) + \underbrace{\frac{1}{\Delta x}(u_x)_j^n\frac{(x-x_j)^2}{2}\Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}}_{=0} + \frac{1}{\Delta x}(u_{xx})_j^n\frac{(x-x_j)^6}{6}\Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} + ...] \\
&= u_j^n + O(\Delta x^2)
\end{aligned}$$

By the same logic,

$$(v)\ \bar{u}_j^{n+1} = u_j^{n+1} + O(\Delta x^2) = u_j^n + \Delta t(u_t)_j^n + O(\Delta x^2, \Delta t^2) \implies \frac{U_j^{n+1} - U_j^n}{\Delta t} = (u_t)_j^n + O(\Delta x^2, \Delta t)$$

Next, examining the flux term via Taylor expansions,

$$\begin{aligned}
f(\bar{u}_j^n) - f(\bar{u}_{j-1}^n) &= f(\bar{u}_j^n) - \left(f(\bar{u}_j^n) - (\bar{u}_j^n - \bar{u}_{j-1}^n)f_u(\bar{u}_j^n) + \frac{1}{2}(\bar{u}_j^n - \bar{u}_{j-1}^n)^2 f_{uu}(\bar{u}_j^n) + ...\right) \\
&= (\bar{u}_j - \bar{u}_{j-1})f_u\Big|_{\bar{u}_j^n} + O(\Delta\bar{u}_j^2) \\
&= \underbrace{(u_j - u_{j-1})}_{=u_x\Delta x + O(\Delta x^2)} f_u\Big|_{\bar{u}_j^n} + O(\Delta u^2, \Delta x^2) \\
&= \Delta x f_u\Big|_{\bar{u}_j^n} u_x + O(\Delta x^2) + O(\Delta x^2, \Delta u^2)
\end{aligned}$$

Further expansion of Taylor series (T.S.) gives us

$$(vv)\ f_u(\bar{u}_j^n) = f(u_j^n) + O(\Delta x^2) \implies f(\bar{u}_j^n) - f(\bar{u}_{j-1}^n) = f(u_j^n)(u_x)_j^n\Delta x + O(\Delta x^2)$$

Combining (v) and (vv) gives us $\tau_j^n = O(\Delta x, \Delta t)$.

Conclusion 4. The scheme (2) is only first order accurate (convergent of order one). (3) is very similar to finite difference methods (FDM).

Question. Why not use the form

$$u_t + f_u u_x = 0$$

and the FDM

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + f_u(u_j^n)\left[\frac{u_j^n - u_{j-1}^n}{\Delta x}\right] = 0$$

Answer. Several reasons:

1. Can't assume that f_u has a constant sign (i.e. always positive or negative) or for some systems we might have $\lambda_1 > 0, \lambda_2 < 0 \implies$ no clear upwind
2. These formulations are not equivalent on discontinuous solutions

On the Importance of Conservation

Consider

$$(1)\ u_t + f(u)_x = 0$$

and integrate assuming u has compact support (so u decays to 0 at $-\infty$ and ∞) to get

$$\frac{d}{dt}\int_{-\infty}^{\infty} u\ dx + f(u)\Big|_{-\infty}^{\infty} = 0 \implies \frac{d}{dt}\int_{-\infty}^{\infty} u\ dx = 0$$

The total of u does not change with time (i.e. conserved). Hence the name, conservation law (1). (1) is an equation ins a conserved form. The below equation is not a conservative form.

$$u_t + f_u u_x = 0$$

Each (1) can be written in a non-conservative form. However, not all

$$u_t + a(u)u_x = 0$$

can be written in form (1) for some $f(u)$ such that $f_u = a(u)$.

Remark 3.1. *If (1) is on $[\alpha, \beta]$ we have*

$$\frac{d}{dt}\int_{\alpha}^{\beta} u\ dx = -\left(f(\beta) - f(\alpha)\right)$$

Total change in u comes from stuff entering from left boundary and leaving through right boundary (assumed $f(\alpha) > 0$ and $f(\beta) > 0$). That is, total mass is conserved or changes due to influx or outflux at the boundaries. FVM have the conservation property. To see this, remark:

$$U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}(F^n_{j+\frac{1}{2}} - F_{j-\frac{1}{2}})$$

$$\int_{\alpha}^{\beta} U\ dx \approx \sum_j U_j^n \Delta x$$

$$\begin{aligned}\sum_j U_j^{n+1}\Delta x &= \sum_j \left[U_j^n \Delta x - \Delta t\left(F_{j+\frac{1}{2}} - F_{j+\frac{1}{2}}\right)\right] \\ &\overset{(1)}{=} \left(\sum_j U_j^n \Delta x\right) - \underbrace{\Delta t\left(F^n_{J+\frac{1}{2}} - F^n_{\frac{1}{2}}\right)}_{(2)}\end{aligned}$$

where (1) is by telescoping and (2) causes the changes in the total mass.

Note 9. All good schemes for hyperbolic conservation laws are conservative.

Example 3.9. Consider Burgers' equation with the forms

$$\begin{aligned}(1)\ u_t + f(u)_x &= 0 \\ (2)\ u_t + f_u u_x &= 0\end{aligned}$$

We can pick special I.C. to show that discretizing (2) is not always correct (because of the loss of conservation). Consider

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

Assume $u \geq 0$ and use upwind flux in FVM:

$$(3)\ U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\left(\frac{(U_j^n)^2}{2} - \frac{(U_{j-1}^n)^2}{2}\right)$$

For the non-conservative form, we have

$$(4)\ \frac{U_j^{n+1}-U_j^n}{\Delta t}+U_j^n\left(\frac{U_j^n-U_{j-1}^n}{\Delta x}\right)=0 \implies U_j^{n+1}=U_j^n-\frac{\Delta t\cdot U_j^n}{\Delta x}(U_j^n-U_{j-1}^n)$$

Consider the I.C.

$$U_j^0=\begin{cases}1 & j\le 0\\ 0 & j>0\end{cases}$$

If we compute U^1 using (4), then

$$U_0^1=U_0^0-\frac{\Delta t\cdot U_0^0}{\Delta x}(U_0^0-U_1^0)=1-\frac{\Delta t}{\Delta x}(1-1)=1$$

and similarly for $U_j^1=1, j<0$. Next,

$$U_1^1-\frac{\Delta t\cdot U_1^0}{\Delta x}(U_1^0-U_0^0)=0-\frac{\Delta t\cdot 0}{\Delta x}(0-1)=0$$

and similarly for $U_j^1=0, j>1$. Iterating, it is clear that $U_j^n=U_j^0$. The solution is a stationary shock (one that does not move). The exact solution, however, is a shock that moves with speed $\dot\xi=\frac{U_L+U_R}{2}=\frac{1}{2}$. Let's compute U^1 using (3):

$$\begin{aligned}U_0^1 &= U_0^0-\frac{\Delta t}{\Delta x}\left(\frac{\left(U_0^0\right)^2}{2}-\frac{(U_{-1}^0)^2}{2}\right)=1\\ U_1^1 &= U_1^0-\frac{\Delta t}{\Delta x}\left(\frac{\left(U_1^0\right)^2}{2}-\frac{(U_0^0)^2}{2}\right)=0-\frac{\Delta t}{\Delta x}\left(-\frac{1}{2}\right)=\frac{\Delta t}{2\Delta x}\\ U_2^1 &= 0\end{aligned}$$

In general,

$$U_j^1=\begin{cases}1 & j\le 0\\ \frac{\Delta t}{2\Delta x} & j=1\\ 0 & j>1\end{cases}$$

The phenomenon of having a value between 1 and 0 and the $j=1$ node is called **smearing of the shock** due to numerical diffusion. Note that rearranging (4), we can write

$$U_j^{n+1}=U_j^n-\frac{\Delta t}{\Delta x}\left(\frac{1}{2}(U_j^n)^2-\frac{1}{2}(U_{j-1}^n)^2\right)-\underbrace{\frac{\Delta t\Delta x}{2}\left(\frac{U_j^n-U_{j-1}^n}{\Delta x}\right)^2}_{(5)}$$

where (5) is an approximation of $(u_x)^2$. On smooth solutions, the last term is $O(\Delta x^2)$ so the difference between (3) and (4) is the order of discretization. However, if u has a discontinuity, $\frac{U_j-U_{j-1}}{\Delta x}$ is not necessarily small and hence (3) and (4) are not equivalent.

Conclusion 5. Non-conservative schemes might result in a wrong shock speed that persist under mesh refinement (strictly speaking this means non-convergent). Conservation is important for problems with discontinuities.

3.7 Boundary Conditions

Usually imposed weakly via **ghost states** (not ghost cells).

1. Dirichlet B.C. (Inflow or Outflow)

(a) Say the domain is $[\alpha, \beta]$ and we have $u(\alpha, t) = \phi(t)$, $U_0^n = \phi(t_n) = \phi^n$. Our discretizations are

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n\right) \\ U_1^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}\left(F_{\frac{3}{2}}^n - F_{\frac{1}{2}}^n\right) \\ F_{\frac{1}{2}}^n &= F^n(U_1^n, U_0^n) = F^n(U_1^n, \phi^n) \end{aligned}$$

and "weakly" means here via the flux and not the value of U at $x_{\frac{1}{2}}$.

2. Solid Wall or Reflecting B.C. a.k.a. no flow through the wall

(a) Impose $u = 0$ (weakly), as velocity, and a ghost vector at the boundary (actually a little outside the boundary) such that

$$\vec{u}_{ghost} = -\vec{u} \text{ (in 1D)}$$

(b) Numerically, this means that

$$U_0^n = -U_1^n$$

(c) For example, for the shallow water equations, where there are two or more unknowns, $\binom{h}{hu}_t$, at the wall, we have

$$\begin{aligned} h^g &= h_1^n = h_0^n \\ u^g &= -u_1^n = -u_0^n \end{aligned}$$

3. Reflecting B.C.

(a) See OneNote

3.8 Lax-Friedrichs in FVM

Consider the standard hyperbolic problem with the FVM

$$\begin{aligned} &(1) \qquad u_t + f(u)_x = 0 \\ &(2) \qquad U_j^{n+1} = U_j^n - \frac{\Delta t}{\Delta x}\left(F_{j+\frac{1}{2}}^n - F_{j-\frac{1}{2}}^n\right) \end{aligned}$$

Recall LF for the linear wave equation $u_t + au_x = 0$ or $u_t + (au)_x = 0$ which was

$$U_j^{n+1} = \frac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \frac{\Delta t}{2\Delta x} a\left(U_{j+1}^n - U_{j-1}^n\right)$$

This motivates LF for (1):2

$$(3)\ U_j^{n+1} = \frac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \frac{\Delta t}{2\Delta x}\left(f(U_{j+1}^n) - f(U_{j-1}^n)\right)$$

Claim 3.1. (3) is a FVM with

$$(4)\ F_{j+\frac{1}{2}}^n = F^n(U_j^n, U_{j+1}^n) = \frac{1}{2}\left(f(U_{j+1}^n) + f(U_j^n)\right) - \frac{\Delta x}{2\Delta t}\left(U_{j+1}^n - U_j^n\right)$$

Proof. We check by substituting (4) into (2) to recover (3):

$$\begin{aligned} U_j^{n+1} &= U_j^n - \frac{\Delta t}{\Delta x}\left[\frac{1}{2}\left(f(U_{j+1}^n) + f(U_j^n)\right) - \frac{1}{2}\left(f(U_j^n) + f(U_{j-1}^n)\right)\right] - \frac{\Delta t}{\Delta x}\frac{\Delta x}{2\Delta t}\left(U_{j+1}^n - U_j^n - (U_j^n - U_{j-1}^n)\right) \\ &= \frac{1}{2}\left(U_{j-1}^n + U_{j+1}^n\right) - \frac{\Delta t}{2\Delta x}\left(f(U_{j+1}^n) - f(U_{j-1}^n)\right) \end{aligned}$$

Hence this approximation works. □

Remark 3.2. (4) can be viewed as a flux for

$$u_t + f(u)_x = \frac{\Delta x^2}{2\Delta t} u_{xx}$$

and

$$\begin{aligned} F^n_{j+\frac{1}{2}} = F(U^n_{j+1}, U^n_j) &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(f(u)_x - \frac{\Delta x^2}{2\Delta t} u_{xx} \right) dx\, dt \\ &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} \left[f(u)_x \Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} - \frac{\Delta x^2}{2\Delta t} u_x \Big|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \right] dt \end{aligned}$$

where

$$F^n_{j+\frac{1}{2}} = \frac{1}{2}\left(f(U^n_{j+1}) + f(U^n_j)\right) - \frac{\Delta x}{2\Delta t}\left(U^n_{j+1} - U^n_j\right)$$

is an approximation for the upper component $\left(\Big|^{x+\frac{1}{2}} \right)$. The first term is an average and the second term is an approximation for $u_x \frac{\Delta x^2}{2\Delta t}$. So LF is very diffusive. (4) gives rise to the popular **LF flux**:

$$F(U_L, U_R) = \frac{1}{2}\left[f(U_L) + f(U_R)\right] - \frac{|\lambda|}{2}(U_R - U_L)$$

where $\Delta x/\Delta t \geq |\lambda|$.

3.9 Higher Order Conservation Laws

From the previous sections:

- FVM is only 1st order accurate
- Difficulties only occur near discontinuities
- LF can be viewed as a FVM

Example 3.10. Consider LW applied to $u_t + au_x = 0$ where

$$U^{n+1}_j = U^n_j - \frac{\alpha}{2}\left(U^n_{j+1} - U^n_{j-1}\right) + \frac{\alpha^2}{2}\left(U^n_{j+1} - 2U^n_j + U^n_{j-1}\right)$$

Note that LW for the linear case can be viewed as a FVM (not true for nonlinear; left as an exercise). Consider the special I.C.

$$u(x,0) = \begin{cases} 1 & x \leq 0 \\ 0 & x > 0 \end{cases} \implies U^0_j = \begin{cases} 1 & j \leq 0 \\ 0 & j > 0 \end{cases}$$

We then have

$$\begin{aligned} U^1_0 &= 1 + \frac{\alpha}{2} - \frac{\alpha^2}{2} \geq 1 \\ U^1_1 &= \frac{\alpha}{2} + \frac{\alpha^2}{2} \end{aligned}$$

This creates oscillations near the discontinuity. Oscillations are stable for linear problems and unstable for non-linear problems.

Example 3.11. Consider upwind for the same equation and same I.C. with

$$\begin{aligned} U^{n+1}_j &= (1-\alpha)U^n_j + \alpha U^n_{j-1} \\ U^1_0 &= 1 \\ U^1_1 &= \alpha \\ U^1_2 &= 0 \end{aligned}$$

and in the next timestep,

$$\begin{aligned} U_1^2 &= -\alpha^2 + 2\alpha \\ U_2^2 &= \alpha^2 \end{aligned}$$

So we have smoothing (or **smearing**) of discontinuities for this scheme.

Theorem 3.2. *(Godunov) An oscillation-free method is only first-order accurate.*

Note 10. True for all methods for hyperbolic equations. Only for methods that can be written as a formula. That is, they do not depend on the solution.

Gibbs Phenomenon

Given a discontinuous function $u(x,0)$, such as the one in our previous two examples, the **Fourier Series** (F.S.) of $u(x,0)$ converges to $u(x,0)$ in the L_2 norm and non-uniformly pointwise. At $x=0$, F.S. converges to

$$\frac{1}{2}\left[u(0^+,0)+u(0^-,0)\right]$$

As $n \to 0$ oscillations get closer to 0 but never disappear. This is the reason for numerical oscillations as well.

Conclusion 6. Two main conclusions:

- High-order methods create oscillations, but more accurate, i.e. desirable
- Low-order methods don't create oscillations, i.e. are stable for nonlinear problems, but they are not accurate enough for practical applications

State of the Art

We know how to construct 2nd order stable (oscillation free) methods. 3D or higher, we have tons of techniques that are adhoc & not always robust. Here is a higher order reconstruction for FVM (**Godunov type**):

In FVM we have averages on each spatial point j and we use these to reconstruct slopes. That is,

$$\tilde{U}_j^n = U_j^n + \sigma(x - x_j)$$

then

$$F_{j+\frac{1}{2}} = F\left(\tilde{U}_j^n(x_{j+\frac{1}{2}}), \tilde{U}_{j+1}^n(x_{j+\frac{1}{2}})\right)$$

We can prove that a lot of sigmas would work for 2nd order accuracy:

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

For stability, we require σ such that $\tilde{U}_j^n(x_{j+1})$ does not exceed U_{j+1}^n and $\tilde{U}_j^n(x_{j-\frac{1}{2}})$ doesn't exceed U_{j-1}^n, i.e. the reconstructed solution on cell j should lie between U_{j-1}^n and U_j. This way we don't create new extrema in the solution $\implies$ there will be no overshoots (and undershoots).

High-Order (Oscillation Free) Reconstruction

Last time, we saw that a low-order method was too diffusive while a high-order method had oscillations. The goal is to get a high-order method, with no spurious oscillation. How do we detect oscillations? Currently, *there is no known robust way*. Instead, we require no new extrema.

To do this, we use **total variation** (TV), or in other words we require the **total variation diminishing** (TVD) property. For continuous TV,

$$TV(u) := \int_\alpha^\beta |u_x|\, dx$$

and for discrete TV,

$$TV(U^n) = \sum_j |U_j^n - U_{j-1}^n|$$

The exact solution satisfies the TVD property if $TV(u)$ does not increase with respect to time. That is, we require

$$(*)\ TV(U^{n+1}) \leq TV(U^n)$$

Definition 3.2. A scheme satisfying (*) is called TVD.

Theorem 3.3. *(Harten) Consider a method of the form*

$$(1)\ U_j^{n+1} = U_j^n + \left[D_j^n(U_{j+1}^n - U_j^n) - C_{j-1}(U_j^n - U_{j-1})\right]$$

If $D_j^n, C_j^n \geq 0$ and $C_j^n + D_j^n \leq 1$ then $TV(U^{n+1}) \leq TV(U^n)$. Note that this is only a sufficient condition.

Proof. Substitute (1) into the definition of TV and expand terms. □

Example 3.12. Consider $u_t + au_x = 0$ with the upwind FV:

$$U_j^{n+1} = U_j^n - \frac{a\Delta t}{\Delta x}(U_j^n - U_{j-1}^n)$$

Here, $C_{j-1}^n = a\Delta t/\Delta x > 0$, $D_j^n = 0$. $C_{j-1}^n + D_j^n = a\Delta t/\Delta x \leq 1$ recovers the CFL condition.

We can also use the TVD condition to compute slope σ in $\tilde{U}_j^n = U_j^n + \sigma(x - x_j)$. σ should be such that there is no overshoots (no new extrema) in the reconstructed solution of $\tilde{U}_j^n(x)$. The steepest allowed slope is

$$(2)\ \sigma = \min\left(\frac{2|U_{j+1}^n - U_j^n|}{\Delta x}, \frac{2|U_j^n - U_{j-1}^n|}{\Delta x}\right)$$

if $(U_{j+1}^n - U_j^n)$ and $(U_j^n - U_{j-1}^n)$ are of the same sign. For any σ, we will create a new extrema so we need

$$(3)\ \sigma = 0$$

Numerically, we will lose accuracy at smooth extrema as slope $= 0$ is only 1st order accurate. (2) and (3) is called the **superbee reconstruction** (or **superbee limiter**). There are others where σ in (2) is less than the steepest possible.

4 Finite Element Methods

Largely applied to elliptic and parabolic equations. The model problem is of the form

$$(1)\ -(p(x)u')' + q(x)u = f(x)$$

This is an elliptic problem (will add u_t later). We assume:

1. $p(x) \geq 0, q(x) \geq 0$
2. p, q, f are smooth enough
3. $u(0) = u(1) = 0$ (for now)
4. $[0, 1]$ domain

Weak formulation

1. Multiply (1) by $v \in H_0^1([0,1])$ where H^p is a **Sobolev space** with norm

$$(*)\ \|u\|_p = \left(\int_0^1 (u^2 + (u')^2 + ... + (u^{(p)})^2)\, dx\right)^{1/2}$$

and $u \in H^p$ if (*) exists. Also, $u \in H_0^p$ if $u \in H^p$ and $u(0) = u(1) = 0$.

2. Integrate on $[0,1]$ and integrate by parts:

$$\int_0^1 \left[-v(p(x)u')' + quv - fv\right]\, dx = \int_0^1 \left[pu'v' + quv - fv\right]\, dx + \underbrace{\left[v(-pu')\right]_0^1}_{=0} = 0$$

since $v(0) = v(1) = 0$, then

$$(2)\ \int_0^1 \left[pu'v' + quv\right]\, dx = \int_0^1 fv\, dx$$

which we call the **weak** or **Galerkin form**.

3. We define $A(u,v) = (f,v)$ where

$$\begin{aligned} A(u,v) &= \int_0^1 \left[pu'v' + quv\right]\, dx \\ (f,v) &= \int_0^1 fv\, dx \end{aligned}$$

The first is a bilinear form and the second is an inner product in L^2 on $[0,1]$.

We call a solution of (1) a strong solution and (2) a weak or Galerkin solution. The idea of **finite element methods** (FEM) is to solve (2) with U, V belonging to a subspace of H_0^1. That is,

$$\begin{aligned} (3)\ \int_0^1 \left(pU'V' + qUV\right)\, dx &= \int_0^1 fV\, dx \\ U, V &\in S^P \subseteq H^1 \end{aligned}$$

To be specific, S^p is the space of piecewise polynomials of degree up to p. For U', V' to exist, we need to require that U, V to be at least continuous. Define U_j as the restriction of U to $[x_{j-1}, x_j]$. So $U_j(x)$ is the polynomial piece on $[x_{j-1}, x_j]$ and $U = \sum_j U_j(x)$. That is, $U_j(x) = 0$ if $x \notin [x_{j-1}, x_j]$.

Start with a linear approximation:

$$(**)\ U = \sum_{j=1}^{N} c_j \phi_j(x)$$

where $\phi_j(x)$ are basis functions called **hat** (or **tent**) **functions**. Divide $[0,1]$ into cells or **elements** $[x_{j-1}, x_j]$. Define or require

$$\phi_j(x) = \begin{cases} 1 & \text{if } x = x_j \\ 0 & \text{if } x = x_k, k \neq j \end{cases}$$

On $[x_{j-1}, x_j]$ we have contributions from $c_j\phi_j(x)$ from $c_j\phi_j(x)$ and $c_{j-1}\phi_{j-1}(x)$ and

$$U_j(x) = c_{j-1}\phi_{j-1}(x) + c_j\phi_j(x)$$

which is a linear polynomial. Note that $U(x_k) = x_k$ is a unique value $\implies$ U is continuous. (**) is a piecewise linear polynomial function. If we want higher order accuracy, we'll need higher-order polynomial approximations. One way to add higher polynomials is via **bubble functions** ϕ_j^2 which are quadratic. These functions ϕ_j^2 are defined only only on $[x_{j-1}, x_j]$ (i.e. zero otherwise). We also restrict $\phi_j^2(x_{j-1}) = \phi_j^2(x_j) = 0$ to preserve continuity.

Conclusion 7. Linear hat functions are non-zero on two adjacent elements give basic 2nd order accuracy and ensure continuity of U. Higher order bubble functions are non-zero on one element only. They can be included into U for higher accuracy. Here,

$$U(x) = \sum_{j=1,\ldots,N, k=2,\ldots,p} c_j\phi_j(x) + c_j^k\phi_j^k(x)$$

We plug this into the FE formula to find c_j and c_j^k. The terms with ϕ_j are the linear basis functions while the terms with the ϕ_j^k are the higher-order bubble functions.

Idea of FEM

Plug $U(x)$ above into (2) and choose $N-1$ suitable $V(x)$ to obtain a linear system for c. Solve the system to obtain the solution.

Assembling FEM

1. Observe that we need to define basis functions on all elements
2. We need to compute the integrals on all I_j
   (a) To save time and space, we map all $I_j = [x_{j-1}, x_j]$ onto a standard "computational" element $[-1,1]$ and do integration there.
   (b) This is done by mapping x to ξ via
   $$x = \frac{1-\xi}{2}\cdot x_{j-1} + \frac{1+\xi}{2}\cdot x_j$$
   (c) Map the basis functions from I_j to $[-1,1]$. That is $\phi_{j-1}(x) \mapsto N_{-1}(\xi)$ and $\phi_{j+1}(x) \mapsto N_1(\xi)$ or
   $$\begin{cases}\phi_j(x_{j-1}) & =0\\ \phi_j(x_j) & =1\end{cases} \mapsto N_1(\xi) = \begin{cases}0 & \xi=-1\\ 1 & \xi=1\end{cases}, \begin{cases}\phi_{j-1}(x_{j-1}) & =1\\ \phi_{j-1}(x_j) & =0\end{cases} \mapsto N_{-1}(\xi) = \begin{cases}1 & \xi=-1\\ 0 & \xi=1\end{cases}$$
      i. Higher order functions will be added later
   (d) Map the integrals (change of variables) via $\frac{dx}{d\xi} = \frac{x_j - x_{j-1}}{2} = \frac{h_j}{2}$. The first integral is
   $$\begin{aligned} A_j^S(v,u) \equiv \int_{x_{j-1}}^{x_j} pv'u'\,dx &= \int_{-1}^{1} p(x(\xi))\frac{dv}{d\xi}\cdot\frac{d\xi}{dx}\cdot\frac{du}{d\xi}\cdot\frac{d\xi}{dx}\cdot\frac{dx}{d\xi}\cdot d\xi\\ &= \int_{-1}^{1} p\frac{dv}{d\xi}\cdot\frac{du}{d\xi}\cdot\frac{2}{h_j}\cdot d\xi\\ &= \frac{2}{h_j}\int_{-1}^{1} pv'u'\,d\xi\end{aligned}$$
   The second integral is
   $$\begin{aligned} A_j^m(v,u) \equiv \int_{x_{j-1}}^{x_j} qvu\,dx &= \int_{-1}^{1} q(x(\xi))v(x(\xi))u(x(\xi))\frac{dx}{d\xi}\cdot d\xi\\ &= \frac{h_j}{2}\int_{-1}^{1} qvu\,d\xi\end{aligned}$$
   The final integral is
   $$\int_{x_{j-1}}^{x_j} vf\,dx = \frac{h_j}{2}\int_{-1}^{1} fv\,d\xi$$
3. How do we test against all $v \in S_N^p$ (N partitions or knots)?
   (a) Set $v = \sum_i d_i\phi_i$ where $d_i \in \mathbb{R}$ for all i. This completely models S_N^p since $\{\phi_i\}$ is a basis. From linearity, we only need to consider (3) on $v \in \{\phi_i\}$
   (b) Using all ϕ_i creates a 1-1 (onto) relation of coefficients and basis functions. Also, no equations are linearly dependent so we have an invertible square system (which is REALLY good)!
   (c) This set-up gives the system
   $$\int_0^1 pU'V'\,dx = \sum_j \frac{2}{h_j}\int_{-1}^{1} p_j(\xi)U_j'(\xi)V_j'(\xi)d\xi$$
   with
   $$\begin{aligned} U_j(\xi) &= c_{j-1}N_{-1}(\xi) + c_jN_1(\xi)\\ V_j(\xi) &= d_{j-1}N_{-1}(\xi) + d_jN_1(\xi)\end{aligned}$$

$$U_j = c_j^T N, V_j = d_j^T N$$

So we can rewrite $A_j^S(V,U)$ as

$$A_j^S(V,U) = \frac{2}{h_j}\int_{-1}^{1} p_j d_j^T \frac{\partial}{\partial\xi} N \frac{\partial}{\partial\xi} N^T c_j \; d\xi$$

and similarly

$$\int_0^1 qUV \; dx = \sum_j \frac{h_j}{2}\int_{-1}^{1} q_j(\xi) U_j(\xi) V_j(\xi) d\xi$$

which gives

$$A_j^M = \frac{h_j}{2}\int_{-1}^{1} q_j d_j^T N N^T c_j \; d\xi$$

(d) If we assume $p_j = p$ is constant then

$$A_j^S(V,U) = \frac{2p}{h_j}\int_{-1}^{1} d_j^T \frac{\partial}{\partial\xi} N \frac{\partial}{\partial\xi} N^T c_j \; d\xi$$

and $N'_{-1} = -1/2, N'_1 = 1/2$ by construction (linear case). So

$$\frac{\partial}{\partial\xi} N \frac{\partial}{\partial\xi} N^T = \begin{pmatrix} -1/2 \\ 1/2 \end{pmatrix} \begin{pmatrix} -1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} 1/4 & -1/4 \\ -1/4 & 1/4 \end{pmatrix}$$

and we can write

$$\begin{aligned} A_j^S &= \frac{p}{h_j} d_j^T \begin{pmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{pmatrix} c_j \\ &= \frac{p}{h_j}(d_j c_j - d_j c_{j-1} - d_{j-1} c_j + d_{j-1} c_{j-1}) \\ &= \frac{p}{h_j}(d_j - d_{j-1})(c_j - c_{j-1}) \\ &= d_j^T K c_j \end{aligned}$$

(e) If we assume $q_j = q$ is constant then

$$\begin{aligned} A_j^M &= \frac{h_j q}{2} d_j \int_{-1}^{1} \begin{pmatrix} N_1 N_1 & N_1 N_{-1} \\ N_{-1} N_1 & N_{-1} N_{-1} \end{pmatrix} d\xi \; c_j^T \\ &= \frac{h_j q}{2} d_j \int_{-1}^{1} \begin{pmatrix} \frac{(1+\xi)^2}{4} & \frac{1-\xi^2}{4} \\ \frac{1-\xi^2}{4} & \frac{(1-\xi)^2}{4} \end{pmatrix} d\xi \; c_j^T \\ &= \frac{h_j q}{2} d_j \begin{pmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{pmatrix} c_j^T \\ &= \frac{h_j q}{6} d_j \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} c_j^T \\ &= d_j^T M_j c_j \end{aligned}$$

and we call M the **local mass matrix**.

(f) We still need $(v,f) = \frac{h_i}{2}\int_{x_{j-1}}^{x_j} v_j f(\xi) \; d\xi$

i. We could use numerical quadrature to evaluate approximately
ii. We could project f into S_N^p
iii. One can show that (i) is equivalent to (ii) by AMATH 242
iv. In the case of S_N^p linear functions we use linear interpolation to get

$$f(x) \approx f_{j-1}\phi_{j-1} + f_j \phi_j, \forall x \in [x_{j-1}, x_j] \implies f(\xi) \approx f_{j-1} N_{j-1} + f_j N_j$$

and hence

$$\begin{aligned}(v,f) &= \frac{h_j}{2}\int_{-1}^{1} d_j^T N_j N_j^T f_j \, d\xi \\ &= \frac{h_j q}{6} d_j \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} f_j^T\end{aligned}$$

Global Matrix Assembly

Recall

$$\int_{x_{j-1}}^{x_j} pU_j^{'}V_j^{'}dx = \frac{p}{2h_j}(d_{j-1},d_j)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} c_{j-1} \\ c_j \end{pmatrix} = d_j^T K_j c_j$$

Let

$$c = \begin{pmatrix} \vdots \\ c_{j-1} \\ c_j \\ c_{j+1} \\ \vdots \end{pmatrix}, d = \begin{pmatrix} \vdots \\ d_{j-1} \\ d_j \\ d_{j+1} \\ \vdots \end{pmatrix}$$

Combine all contributions into a global matrix. This is done by extending K_j into a global matrix by filling it with zeros:

$$d^T \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{K_j} c + d^T \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{K_{j+1}} c = \begin{pmatrix} 1 & -1 & & & & \\ \ddots & \ddots & \ddots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ & & & & -1 & 1 \end{pmatrix}$$

where explicitly

$$(K_j)_{st} = \begin{cases} 1 & s = t = j \text{ or } s = t = j+1 \\ -1 & s = j+1, t = j \text{ or } s = j, t = j+1 \end{cases}$$

Recall that

$$p\int_0^1 U^{'}V^{'}dx = \sum p\int_{x_{j-1}}^{x_j} U^{'}V^{'}dx$$

If we continue, we will get

$$p\int_0^1 U'V'dx = d^T Kc$$

where

$$K = \frac{p}{2h}\begin{pmatrix} 1 & -1 & & & & \\ \ddots & \ddots & \ddots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ & & & & -1 & 1 \end{pmatrix}$$

assuming a uniform mesh. Similarly

$$\int_{x_{j-1}}^{x_j} qU_j^{'}V_j^{'}dx = \frac{qh_j}{6}d_j^T \underbrace{\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}}_{M_j} c_j \implies \int_0^1 qU_j^{'}V_j^{'}dx = d^T Mc$$

where M is called the **global mass matrix** and

$$M = \frac{qh}{6}\begin{pmatrix} 2 & 1 & & & & \\ \ddots & \ddots & \ddots & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 & 0 \\ 0 & 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & \ddots & \ddots & \ddots \\ & & & & 1 & 2 \end{pmatrix}$$

Finally,

$$\int_{x_{j-1}}^{x_j} fV\ dx \approx d_j^T \frac{h}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} f_{j-1} \\ f_j \end{pmatrix} = d_j^T \underbrace{\frac{h}{6}\begin{pmatrix} 2f_{j-1}+f_j \\ f_{j-1}+2f_j \end{pmatrix}}_{L_j}$$

and

$$\int_0^1 fV\ dx \approx d^T L,\ L = \frac{h}{6}\begin{pmatrix} \vdots \\ f_{j-1}+4f_j+f_{j+1} \\ f_j+4f_{j+1}+f_{j+2} \\ \vdots \end{pmatrix}, f_j = f(x_j)$$

Combining all contributions,

$$d^T Kc + d^T Mc = d^T L \implies d^T\left((K+M)c - L\right) = 0$$

This should be true for all d so

$$(K+M)c = L$$

So we can solve for c using linear program solvers to obtain $U = \sum_j c_j\phi_j$.

Trivial Extension to Parabolic Equations

Consider $u_t = \sigma u_{xx}$. Multiply by $v \in H_0^1$ to get

$$\frac{d}{dt}\int_0^1 uv\ dx = \sigma\int_0^1 u'v'\ dx$$

Seek for FE solution $U(x,t) = \sum_j c_j(t)\phi_j(x), V(x) = \sum_j d_j\phi_j(x)$. This satisfies

$$\int_0^1 \sum \frac{d}{dt}c_j(t)\phi_j(x)\sum d_j\phi(x)\ dx = \sigma\int_0^1 \sum c_j(t)\phi'(x)\sum d_j(t)\phi_j'(x)\ dx$$

or

$$d^T M\frac{d}{dt}c = \sigma d^T Kc$$

where M and K are as before with $p=1, q=1$. This gives:

$$(*)\ M\frac{dc}{dt} = \sigma Kc$$

Now (*) is an ODE system that can be solved using an ODE solver (e.g. implicit Runge-Kutta method or backward Euler).

4.1 Optimality of Finite Element Solutions

Model problem in weak formulation:

$$(1)\ \int_0^1 (pu'v' + quv)\ dx = \int_0^1 fv, u \in H_0^1, \forall v \in H_0^1$$

or

$$A(v,u) = (v,f), \forall v \in H_0^1$$

The FE problem is

$$(2)\ \int_0^1 (pU'V' + qUV)\ dx = \int_0^1 fV\ dx, U \in S_N^p, \forall V \in S_N^p$$

Recall A is a symmetric bilinear form:

$$\begin{aligned} A(u,v) &= A(v,u) \\ A(u,v+w) &= A(u,v) + A(u,w) \end{aligned}$$

Note that (2) is valid $\forall v \in H_0^1$ and $S_N^p \subset H_0^1$. We have

$$A(V,u) = (V,f), A(V,U) = (V,f) \implies (*)\ A(V,u-U) = 0$$

So if $e \equiv u - U \implies A(V,e) = 0, \forall V \in S_N^p$. Take $V = U$ to get

$$(**)\ A(U,u-U) = 0$$

in (*). Consider $A(u-U,u-U)$. This can be expanded to get

$$\begin{aligned} A(u-U,u-U) &= \underbrace{A(u,u) - A(u,U) - A(U,u) + A(U,U)}_{\text{expanded LHS}} + \underbrace{2A(U,u-U)}_{=0 \text{ by (**)}} \\ &= A(u,u) - A(U,U) \\ &= A(u,u) - A(U,U) + A(V,V) - A(V,V) - \underbrace{2A(u-U,V)}_{=0 \text{ by (*)}} \\ &= [A(u,u) - 2A(u,V) + A(V,V)] - [A(U,U) - 2A(U,V) + A(V,V)] \\ &= A(u-V,u-V) - A(U-V,U-V) \end{aligned}$$

Remark that

$$A(U-V,U-V) \geq 0$$

since by assumption $p \geq 0, q \geq 0$ for all $x \in [0,1]$. We have thus showed that

$$(\$)\ A(u-U,u-U) \leq A(u-V,u-V), \forall v \in S_N^p$$

U is a FE solution and V is any piecewise polynomial function. Now, let's introduce the following norm, which we call the **energy norm**:

$$\|w\|_A := (A(w,w))^{1/2} = \left(\int_0^1 (p(w')^2 + qw^2)\ dx\right)^{1/2}$$

This is kind of like a weighted Sobolev norm. This is a norm because:

1. $\|\alpha w\|_A = |\alpha| \|w\|_A$
2. $\|w_1 + w_2\|_A \leq \|w_1\|_A + \|w_2\|_A$
3. $\|w\|_A = 0$ iff $w(x) = 0, \forall x$ $(p \geq 0, q \geq 0)$

Using our norm above, ($) can be rewritten as

$$\|u-U\|_A^2 \leq \|u-V\|_A^2, \forall v \in S_N^p$$

or

$$(***)\ \|e\|_A = \|u-U\|_A^2 = \min_{v \in S_N^p} \|u - A\|_A$$

Conclusion 8. Out of infinitely many possible approximations for $u(x)$ in S_N^p, the finite element solution is the best in the sense that it has the smallest error in the energy norm.

Back to $A(u-U,V)=0$. Viewing A as a inner product, we conclude that the error $e=u-U$ is orthogonal to all function in S_N^p. We can get the order of convergence from (***) by using standard interpolation results:

$$\|u-U_j\|_s \le C\cdot h^{\min(q,p+1)-s}|u|_{q;[x_{j-1},x_j]}, C\in\mathbb{R}$$

where q indicates the norm in H^s,

$$q: u\in H^q, p: U_j\in S_N^p$$

and

$$|u|_q=\left(\int\left(\frac{\partial^q u}{\partial x^q}\right)^2 dx\right)^{1/2}$$

the **Sobolev semi-norm**. Assuming u is smooth (i.e. $q\ge p+1$). we have the bound in the L^2 norm ($s=0$) as

$$\|u-U_j\|_2\le \bar{c}h^{p+1}$$

or on the whole domain:

$$\|u-U\|_{[0,1]}\le \hat{c}h^{p+1}$$

with a $p+1$ convergence rate.

4.2 Discontinuous Galerkin Methods

Cross between FE and FV. Consider the equation

$$(1)\ u_t+f(u)_x=0$$

Divide the domain into cells. Multiply (1) by a test function $v\in H^1([x_{j-1},x_j])$ and integrate on I_j. This gives

$$\int_{x_{j-1}}^{x_j} u_t v\ dx+\int_{x_{j-1}}^{x_j} f(u)_x v\ dx=0$$

By parts,

$$\int_{x_{j-1}}^{x_j} u_t v\ dx+f(u)v\Big|_{x_{j-1}}^{x_j}-\int_{x_{j-1}}^{x_j} f(u)v'dx=0,\forall v\in H'$$

We assume $u\Big|_{I_j}\approx U_j, U_j=\sum_{i=1}^P c_{ij}(t)\phi_{ij}(x)$ where $\phi_{ij}(x)$ are basis functions on I_j and $c_{ij}(t)$ are solution coefficients on I_j. Since $\int_{x_{j-1}}^{x_j} f(u)v'\ dx$ is defined for piecewise continuous functions $v=\phi_{ij}(x)$, we do not need to enforce continuity across cells $\Longrightarrow$ hence the name **discontinuous Galerkin**.

Substitute (3) into (2) and choose numerical test functions $V=\phi_{ij}, 0\le i\le p$ to get

$$(2)\ \frac{d}{dt}\int_{x_{j-1}}^{x_j} U_jV\ dx+f(U_j)V\Big|_{x_{j-1}}^{x_j}-\int_{x_{j-1}}^{x_j} f(U_j)V'\ dx=0, 0\le i\le p,\forall V\in\bar{S}_N^p$$

where $\bar{S}_N^p$ is the space of piecewise polynomial functions. With test functions, this is

$$\frac{d}{dt}\int_{x_{j-1}}^{x_j} U_j\phi_{ij}\ dx+f(U_j)\phi_{ij}\Big|_{x_{j-1}}^{x_j}-\int_{x_{j-1}}^{x_j} f(U_j)\phi_{ij}'\ dx=0, 0\le i\le p,\forall V\in\bar{S}_N^p$$

If there is a discontinuity at x_j for U, then $f(U_j(x_j))$ is not defined because of the discontinuity of $U(x)$. To remedy this, we use numerical flux:

$$\begin{aligned}\frac{d}{dt}\int_{x_{j-1}}^{x_j} U_j\phi_{ij}\ dx+\left(F(U_j(x_j),U_{j+1}(x_j))\right)\phi_{ij}(x_j)&\\ -\left(F(U_{j-1}(x_{j-1}),U_j(x_{j-1}))\right)\phi_{ij}(x_{j-1})&\\ -\int_{x_{j-1}}^{x_j} f(U_j)\phi_{ij}'\ dx &= 0\end{aligned}$$

F is a numerical flux (borrowed from FVM) e.g. Lax-Friedrichs. Now map $[x_{j-1}, x_j]$ into $[-1, 1]$ and define test functions once on $[-1, 1]$ instead of on each I_j:

$$\begin{aligned}\frac{d}{dt}\int_{-1}^{1}\frac{h_j}{2}U_j(\xi)\phi_i(\xi)\,d\xi + \left(F(U_j(1), U_{j+1}(-1))\right)\phi_i(1) \\ - \left(F(U_{j-1}(1), U_j(-1))\right)\phi_i(-1) \\ -\int_{-1}^{1} f(U_j)\phi_i'\,d\xi &= 0\end{aligned}$$

$\{\phi_i\}_{i=0}^{p}$ can be anything, but note that:

1. Polynomials are the most convenient
2. Some polynomials are better than others

The naive basis is the monomials $1, \xi, \xi^2, \xi^3, ...$ but the the good basis is $P_0(\xi), P_1(\xi), P_2(\xi), ...$ where $P_k(\xi)$ is the k^{th} **Legendre polynomial**. Note that

$$\int_{-1}^{1} P_i(\xi)P_k(\xi)\,d\xi = c\delta_{ij}$$

and so $\{P_i\}$ are an orthogonal set and c depends on the normalization. We require $P_i(1) = 1$ for all i. Then $P_i(-1) = (-1)^i$ (fact; can be derived) and $C = \frac{2}{2k+1}$. Explicitly,

$$P_0 = 1, P_1 = \xi, P_2 = \frac{3\xi^2 - 1}{2}, P_3 = \frac{5\xi^2 - 3\xi}{2}, ...$$

The Legendre basis is better because

$$\int_{-1}^{1}\sum_{i=0}^{P} c_{ij}(t)\phi_i(\xi)\cdot\phi_k(\xi)\,d\xi = \sum_{i=0}^{P} c_{ij}(t)\int_{-1}^{1}\phi_i(\xi)\phi_k(\xi)\,d\xi, k = 0, ..., p$$

The first terms gives raise to a matrix form:

$$\underbrace{\frac{h_j}{2}\begin{pmatrix}\int_{-1}^{1}\phi_0\phi_0\,d\xi & \int_{-1}^{1}\phi_0\phi_1\,d\xi & \cdots & \int_{-1}^{1}\phi_0\phi_1\,d\xi \\ \vdots & \vdots & & \vdots \\ \int_{-1}^{1}\phi_p\phi_0\,d\xi & \int_{-1}^{1}\phi_p\phi_1\,d\xi & \cdots & \int_{-1}^{1}\phi_p\phi_p\,d\xi\end{pmatrix}}_{M}\frac{d}{dt}c_j(t), c_j = \begin{pmatrix}c_{0j}\\ c_{1j}\\ \vdots \\ c_{pj}\end{pmatrix}$$

With the Legendre basis, this matrix M is diagonal and with others it is not. In the general case,

$$M\cdot\frac{d}{dt}c_j(t) = L(c)$$

where $L(c)$ combines the rest of the terms and $L(c)$ is a vector of length $(p+1)$. The i^{th} component is

$$L_i(c) = -\left(F(U_j(1), U_{j+1}(-1))\right)V_i(1) + \left(F(U_{j-1}(1), U_j(-1))\right)V_i(-1) + \int_{-1}^{1} f(U_j)\phi_i'\,d\xi$$

We need to invert M to get

$$\frac{d}{dt}c_j(t) = M^{-1}L(c)$$

and this can be solved with an ODE solver such as Runge-Kutha or Adams-Bashforth. If $\{\phi_i\}$ is orthogonal we have M is diagonal. In the case of Legendre,

$$\int_{-1}^{1}\phi_k^2\,d\xi = \frac{2}{2k+1} \implies M_{kk} = \frac{2}{2k+1}$$

This makes (2) uncouple and we get

$$\frac{h_j}{2}\cdot\frac{2}{2i+1}\cdot\frac{d}{dt}c_{ij}+F(1)\cdot 1-F(-1)\cdot(-1)^i-\int_{-1}^{1}f(U_j)P_i'\,d\xi=0$$

using $P_i(1)=1$ and $P_i(-1)=(-1)^i$

$$\frac{h_j}{2i+1}\cdot\frac{d}{dt}c_{ij}=-\left(F(U_j(1),U_{j+1}(-1))-F(U_{j-1}(-1),U_j(-1))\cdot(-1)^i\right)+\int_{-1}^{1}f(U_j)P_i'\,d\xi$$

This is the **discontinuous Galerkin formulation**. The reasons for an orthogonal basis are:

1. Computational efficiency: Multiplication by M^{-1} is costly
   (a) In 3D, the number of basis functions is $\frac{(p+1)(p+2)(p+3)}{6}$
      i. e.g. if $p=4$ then $M\in\mathbb{R}^{35\times 35}$
2. Condition number for M
   (a) e.g. with the monomial basis,
   $$M=\begin{pmatrix}\int 1 & \int x & \cdots & \int x^p\\ \vdots & \vdots & & \vdots\\ \int x^p & \int x^{p+1} & \cdots & \int x^{2p}\end{pmatrix}$$
   which is the Vandermonder matrix, an ill-conditioned matrix.
3. In 3D, this method does not depend on having a square mesh
4. Getting higher order accuracy is easier compared to finite volume methods (small stencil)

Aside. If we want a basis which is 0 on the boundaries, consider the **Lobatto polynomials** defined by $N_i=P_i-P_{i-2}$ since $N_i(1)=N_i(-1)=0$.

Index

Appendix

How to Check your Code

1. Manufacture a problem for which you know the exact solution and which you should be able to solve exactly.
   (a) In the heat equation, we could try $u(x,t) = 1$ with I.C. $u(x,0) = 1$ and B.C. $u(1,t) = 1, u(0,t) = 1$.