

# CME 306 Computational Methods of Applied Mathematics

Instructor: Lexing Ying

## Topics to be covered:

- ODE (2 weeks)
- PDEs:
  - ◆ Elliptic: FDM, FEM (2 weeks)
  - ◆ Parabolic: FDM (1 week)
  - ◆ Hyperbolic: FVM (2 weeks)
- SDE: Stochastic DE (1 week)
- Monte Carlo methods: MCMC (1 week)
- Wavelets & FFT: Applied harmonic analysis (1 week)

## Textbooks:

- ODE: S.H.D.
- PDEs: Larsson + Thomee, LeVeque
- SDE & MCMC: E et. al.
- Wavelets: Mallat, Chapter 7

## Week 1: Lecture 1. ODE Recap.

$$\dot{x}(t) = f(t, x(t)), \quad x(0) = x_0, \quad t > 0, \quad x(t) \in \mathbb{R}$$

Find a solution whose slope satisfies the specified vector field

- Existence and uniqueness

Ex. 1:  $\dot{x} = x^2$ ,  $x(0) = 1$ . The solution is  $x(t) = 1/(1 - t)$ , only local existence

Ex. 2:  $\dot{x} = \sqrt{x}$ ,  $x(0) = 0$ . One solution is  $x(t) \equiv 0$ . In fact, there are infinite solutions:

$$x(t) = \begin{cases} \frac{1}{4}(x-s)^2, & x \geq s \\ 0, & x < s \end{cases}$$

Consider only  $f(t, x)$  that is **Lipschitz** in  $x$ :

$$\forall t \in [0, T], \quad \forall x, y, \quad |f(t, x) - f(t, y)| \leq L|x - y|$$

Intuitively, this condition means that

$$\left| \frac{\partial f}{\partial x}(t, x) \right| \leq L$$

Ex. 1:  $f_x(t, x) = x$ , not bounded by  $L$  for  $x \rightarrow \infty$

Ex. 2:  $f_x(t, x) = 1/(2\sqrt{x})$ , not bounded by  $L$  for  $x \rightarrow 0$

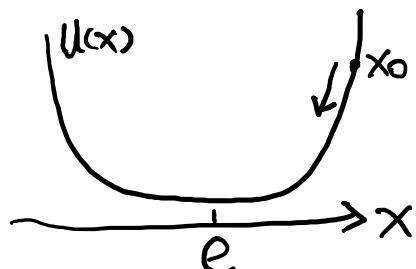
**Thm.** If  $f(t, x)$  is Lipschitz in  $x$  in  $t \in [0, T]$ , then the solution to ODE exists and is unique.

- Gradient system (related to ML)

Optimize the energy loss function  $U(x)$  by

$$\dot{x} = -\nabla_x U(x)$$

If  $e$  is stationary, then the gradient is 0.



- Hamiltonian dynamics

From the Newton's form with potential  $U(x)$

$$\ddot{x} = -\nabla_x U(x)$$

We can define the Hamiltonian dynamics in the form

$$q = x, \quad p = \dot{x}, \quad \dot{q} = p, \quad \dot{p} = -\nabla U(q)$$

In this system, the energy is conserved

$$E = \frac{p^2}{2} + U(q), \quad \dot{E}(t) = p\dot{p} + \nabla U(q)\dot{q} = -\nabla U(q)p + \nabla U(q)p = 0$$

- Damped Hamiltonian dynamics (e.g., exponential form of decay)

$$\dot{q} = p, \quad \dot{p} = -\nabla U(q) - \lambda p$$

If the damping is extremely strong ( $\lambda \gg 1$ )

$$-\nabla U(q) - \lambda p \approx 0, \quad p = -\frac{\nabla U(q)}{\lambda}$$

Putting back to the system gives the form of the gradient dynamics

$$\dot{q} = p = -\frac{\nabla U(q)}{\lambda} = -\nabla\left(\frac{U}{\lambda}\right)(q)$$

**The gradient system is the limit of a damped HD.** In ML, the momentum GD (Gradient Descent) is equal to the damped HD.

- Numerics

From Calculus (derivative = limit of difference) to Numerics (derivative  $\approx$  small difference)

To solve the following ODE:

$$\dot{y} = f(t, y), \quad y(0) = 0, \quad t \in [0, T]$$

For a step size  $h$  and the number of points  $N = T/h$

$$y_n = y(t_n), \quad w_n \approx y_n = y(t_n), \quad t_i = ih$$

An approx. equation of the exact solution is

$$\frac{y_{n+1} - y_n}{h} \approx f(t, y_n)$$

Now we set the following (exact solution for approx. ODE)

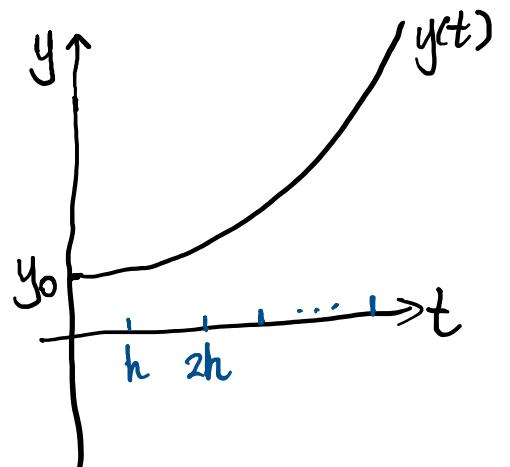
$$\frac{w_{n+1} - w_n}{h} = f(t, w_n)$$

Explicit Euler method:

$$w_{n+1} = w_n + h \cdot f(t_n, w_n)$$

Implicit Euler method:

$$w_{n+1} = w_n + h \cdot f(t_{n+1}, w_{n+1})$$



Trapezoidal method:

$$w_{n+1} = w_n + h \left[ \frac{1}{2} f(t_n, w_n) + \frac{1}{2} f(t_{n+1}, w_{n+1}) \right]$$

➤ Proof of Explicit Euler method

To estimate the error  $e_n = |y_n - w_n|$ , from Taylor expansion we have

$$y_{n+1} = y_n + h \cdot f(t_n, y_n) + \frac{1}{2} h^2 \cdot y''(\xi_n), \quad \xi_n \in [t_n, t_{n+1}]$$

The remainder can be bounded considering  $|y''(\xi_n)| \leq M$ . According to the Lipschitz condition, now the error term becomes

$$\begin{aligned} e_{n+1} &\leq e_n + h |f(t_n, y_n) - f(t_n, w_n)| + \frac{1}{2} h^2 M \leq e_n (1 + hL) + \frac{1}{2} h^2 M \\ e_{n+1} + \frac{\frac{1}{2} h^2 M}{hL} &\leq (1 + hL)e_n + \frac{1}{2} h^2 M + \frac{\frac{1}{2} h^2 M}{hL} = (1 + hL) \left[ e_n + \frac{\frac{1}{2} h^2 M}{hL} \right] \end{aligned}$$

Therefore, the error is bounded by (note that  $e_0 = 0$ )

$$e_n + \frac{\frac{1}{2} h^2 M}{hL} \leq (1 + hL)^n \cdot \left( e_0 + \frac{\frac{1}{2} h^2 M}{hL} \right) \leq e^{nhL} \cdot \frac{hM}{2L}, \quad \text{using } (1 + hL)^{\frac{1}{hL}} \rightarrow e$$

And we can obtain the bound of the global truncation error

$$e_n \leq (e^{nhL} - 1) \cdot \frac{hM}{2L} \sim e^{TL} \cdot h, \quad \forall n \in [1, N]$$

Note that the error converges:

$$\max_{1 \leq n \leq N} e_n = O(h) \rightarrow 0, \quad h \rightarrow 0$$

However, the exponential term  $e^{TL}$  can destroy all analysis for large time  $T$

**Explicit Euler converges, but the convergence is too slow, and there is an exponential error blow-up with the factor  $e^{TL}$ .**

➤ Runge-Kutta method

For previous Euler schemes

$$y(t_{n+1}) = y(t_n) + h \cdot y'(t_n) + O(h^2)$$

The remainder  $O(h^2)$  gives rise to the final  $O(h)$  truncation error.

The second-order derivative is

$$y'' = \frac{d}{dt} [f(t, y(t))] = f_t + f_y \cdot y' = f_t + f_y \cdot f$$

However, for engineering problem the function  $f$  is usually a black box, which is impossible to evaluate the analytic derivatives of  $f$ .

To obtain numerical schemes with higher order of convergence, we write

$$y(t_{n+1}) = y(t_n) + h \cdot y'(t_n) + \frac{1}{2} h^2 \cdot y''(t_n) + O(h^3)$$

The idea is to approximate  $y''$  (and  $y'$ ) with  $f$  evaluated at carefully selected locations:

$$y(t_{n+1}) = y(t_n) + h [a_1 f(t, y) + a_2 f(t + \varepsilon, y + \delta f(t, y))]$$

Now we need to choose  $a_1, a_2, \varepsilon, \delta$  by matching it with Taylor expansion

$$f + \frac{h}{2} (f_t + f_y \cdot f) = a_1 f + a_2 (f + \varepsilon f_t + f_y \cdot \delta f)$$

The above equality should be true for all possible  $f, f_t, f_y$ . We thus obtain

$$a_1 + a_2 = 1, \quad a_2 \varepsilon = \frac{h}{2}, \quad a_2 \delta = \frac{h}{2}$$

Now we can obtain different numerical schemes.

Several second-order RK methods (need 2 function evals):

$$a_1 = 0, \quad a_2 = 1, \quad \varepsilon = \delta = \frac{h}{2}$$

$$w_{n+1} = w_n + h \cdot f \left( t_n + \frac{h}{2}, w_n + \frac{h}{2} f(t_n, w_n) \right)$$

$$a_1 = \frac{1}{2}, \quad a_2 = \frac{1}{2}, \quad \varepsilon = \delta = h$$

$$w_{n+1} = w_n + h \cdot \left[ \frac{1}{2} f(t_n, w_n) + \frac{1}{2} f(t_n + h, w_n + h f(t_n, w_n)) \right]$$

These methods have  $O(h^2)$  accuracy,  $e_n \leq e^{TL} \cdot O(h^2)$

Fourth-order RK method:

$$w_{n+1} = w_n + \frac{1}{6}[k_1 + 2k_2 + 2k_3 + k_4], \quad |e_n| \leq e^{LT} \cdot O(h^4)$$
$$k_1 = h \cdot f(t_n, w_n), \quad k_2 = h \cdot f\left(t_n + \frac{h}{2}, w_n + \frac{k_1}{2}\right),$$
$$k_3 = h \cdot f\left(t_n + \frac{h}{2}, w_n + \frac{k_2}{2}\right), \quad k_4 = h \cdot f(t_n + h, w_n + k_3)$$

Topics for next lecture:

Stiffness of the gradient dynamics

Symplectic scheme of Hamiltonian dynamics:  $O(h)$  and  $O(h^2)$  schemes

## Week 1: Lecture 2. Stiffness & Symplectic schemes

- Stiffness of an ODE system

Consider a vector  $\mathbf{y}(t)$  with  $\lambda \gg 1$

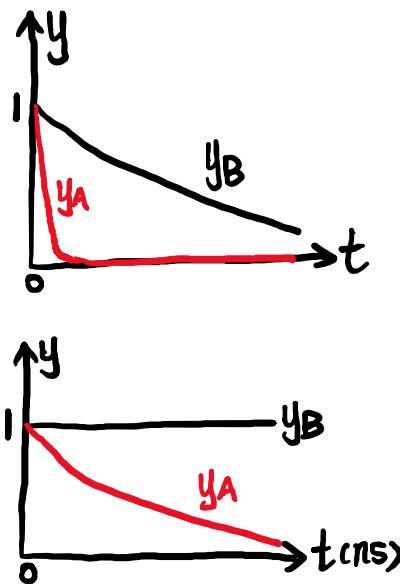
$$\begin{aligned}\dot{\mathbf{y}}^A(t) &= -\lambda \mathbf{y}^A(t), & \mathbf{y}^A(0) &= 1 \\ \dot{\mathbf{y}}^B(t) &= -\mathbf{y}^B(t), & \mathbf{y}^B(0) &= 1\end{aligned}$$

The solution is

$$\mathbf{y}^A(t) = e^{-\lambda t}, \quad \mathbf{y}^B(t) = e^{-t}$$

“Boring” or “interesting” solutions depend on the observer.

For a much faster time scale,  $\mathbf{y}^B$  is nearly constant and becomes the “boring” solution.



In terms of matrix notation

$$\dot{\mathbf{y}}(t) = -\begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix} \mathbf{y}(t)$$

In a different basis

$$M\dot{\mathbf{y}}(t) = -M \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix} M^{-1} M \mathbf{y}(t), \quad \mathbf{z}(t) = -A \mathbf{z}(t)$$

Usually our problem is given in the second form, but doing matrix diagonalization at each time step is very expensive.

Consider how to solve  $\mathbf{y}(t)$  numerically using **explicit Euler methods**

$$w_{n+1}^A = w_n^A + h \cdot (-\lambda) \cdot w_n^A = (1 - h\lambda)w_n^A, \quad w_{n+1}^B = (1 - h)w_n^B$$

Recursion gives (with  $w_0^A = w_0^B = 1$ )

$$w_n^A = (1 - h\lambda)^n, \quad w_n^B = (1 - h)^n$$

Both exact solutions decay with time, so at least we want

$$w_n \rightarrow 0, \quad n \rightarrow \infty$$

To obtain this asymptotic behavior, we need

$$|1 - h\lambda| < 1, \quad |1 - h| < 1$$

$$0 < h < \frac{2}{\lambda}, \quad 0 < h < 2$$

Because  $\lambda \gg 1$  we are forced to have  $0 < h < 2/\lambda$  due to part A, the “boring” solution.

For explicit Euler method, the “boring” part forces us to take tiny step size.

**Stiffness:** For a linear ODE system, if the exponential decaying behavior is drastically different among components, then the system is a stiff system.

On the contrary, for **implicit Euler method**

$$\begin{aligned} w_{n+1}^A &= w_n^A - h\lambda w_{n+1}^A, & w_{n+1}^B &= w_n^B - hw_{n+1}^B \\ w_{n+1}^A &= \frac{w_n^A}{1 + h\lambda}, & w_{n+1}^B &= \frac{w_n^B}{1 + h} \end{aligned}$$

Recursion also gives

$$w_n^A = \left( \frac{1}{1 + h\lambda} \right)^n, \quad w_n^B = \left( \frac{1}{1 + h} \right)^n$$

To obtain the asymptotic behavior, we need

$$|1 + h\lambda| > 1, \quad |1 + h| > 1$$

This leads to a trivial condition  $h > 0$ . Note that this only guarantees decaying solutions, but not accurate solutions. **For implicit Euler method, the choice of step size  $h$  is not restricted by  $\lambda$ , but is dictated by the desired accuracy for the “interesting” part.**

➤ Application to example stiff systems

1. **Heat equation** includes rapidly decaying high-frequency mode (“boring”) and slowly decaying low-frequency mode (“interesting”)

$$u_t = u_{xx}$$

2. **Plate equation** (very stiff system)

$$u_t = -u_{xxxx}$$

➤ Trapezoidal method for a stiff system

$$w_{n+1}^A = w_n^A + h \left[ \frac{1}{2} f(w_n^A) + \frac{1}{2} f(w_{n+1}^A) \right], \quad \left(1 + \frac{h\lambda}{2}\right) w_{n+1}^A = \left(1 - \frac{h\lambda}{2}\right) w_n^A$$

The recursive solutions are

$$w_n^A = \left( \frac{2 - h\lambda}{2 + h\lambda} \right)^n, \quad w_n^B = \left( \frac{2 - h}{2 + h} \right)^n$$

Again  $h > 0$  is enough to guarantee the desired decaying behavior.

**For stiff systems, explicit methods do not work, while implicit methods are appropriate.**

Extra notes: For Schrodinger equation, all modes are rotating around a unit circle, which is hard to determine “boring” or “interesting” modes. WKB methods are applied to solve specific modes efficiently, avoiding the usage of tiny time step size.

➤ Gradient descent system for optimization

$$\dot{x} = -\nabla_x E(x)$$

$$E(x) = \frac{\lambda x^2}{2}, \quad \nabla_x E(x) = \lambda x, \quad \dot{x} = -\lambda x$$

If the goal is only to find the minimum  $e$ , we can use different ODEs

$$\dot{x} = -M(x) \cdot \nabla_x E(x), \quad x \in \mathbb{R}^d, \quad M(x) \in \mathbb{R}^{d \times d} \text{ is sym. pos. def.}$$

The reason is that the derivative is still negative (energy still decreases)

$$\frac{d}{dt} E(x(t)) = \nabla_x E^T(x) \cdot \dot{x} = -\nabla_x E^T \cdot M \cdot \nabla_x E < 0$$

➤ Hamiltonian system & Symplectic integrator

$$\dot{q}(t) = p(t) = H_p, \quad \dot{p}(t) = -\nabla_q V(q(t)) = -H_q$$

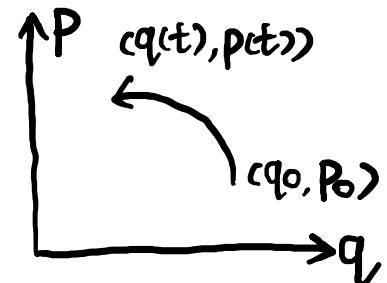
H.S. is everywhere, and the long-time error can be treated to not blow up (instead of  $e^{LT}$ )

The Hamiltonian is defined as

$$H(q, p) = \frac{p^2}{2} + V(q)$$

And it is conserved over time (energy conservation)

$$\frac{d}{dt} H = H_q \dot{q} + H_p \dot{p} = H_q H_p + H_p \cdot (-H_q) = 0$$



The symplectic form is a signed volume form

$$dq \wedge dp = -dp \wedge dq$$

The HS also preserves the above symplectic form  $dq \wedge dp$ . Consider time  $t = \varepsilon$

$$q(\varepsilon) \approx q + H_p \varepsilon + O(\varepsilon^2), \quad p(\varepsilon) \approx p - H_q \varepsilon + O(\varepsilon^2)$$

The wedge product becomes

$$dq(\varepsilon) \wedge dp(\varepsilon) = dq \wedge dp + dq \wedge d(-H_q \varepsilon) + d(H_p \varepsilon) \wedge dp + O(\varepsilon^2)$$

The chain rule gives

$$dH_p = H_{pp}dp + H_{pq}dq, \quad dH_q = H_{qp}dp + H_{qq}dq$$

Note the following properties of the symplectic form

$$dp \wedge dp = dq \wedge dq = 0 \quad (\text{"zero area"})$$

So we have

$$dq(\varepsilon) \wedge dp(\varepsilon) = dq \wedge dp + \varepsilon(-dq \wedge H_{qp}dp - dq \wedge H_{qq}dq + H_{pp}dp \wedge dp + H_{pq}dq \wedge dp) + O(\varepsilon^2)$$

$$dq(\varepsilon) \wedge dp(\varepsilon) = dq \wedge dp + O(\varepsilon^2)$$

Therefore, we have the conservation of  $dq \wedge dp$

$$\frac{dq(\varepsilon) \wedge dp(\varepsilon) - dq(0) \wedge dp(0)}{\varepsilon} = O(\varepsilon), \quad \frac{d}{dt}[dq(t) \wedge dp(t)] = 0$$

Hamiltonian  $H(q, p)$  and volume form  $dq \wedge dp$  are preserved along the Hamiltonian system.

Therefore,  $f(H)dV$  is also preserved.

Goal: Design numerical schemes for H.S. such that

$H(q, p)$  is **approx.** preserved

$dq \wedge dp$  is **exactly** preserved

These schemes are called symplectic integrators.

## 1. Euler-B method

$$\begin{aligned} q^{n+1} &= q^n + \Delta t H_p(q^n, p^{n+1}) \\ p^{n+1} &= p^n - \Delta t H_q(q^n, p^{n+1}) \end{aligned}$$

Numerically, the second equation is solved first. This is in general an implicit scheme.

However, for a decoupled Hamiltonian

$$H(q, p) = \frac{p^2}{2} + V(q), \quad H_q = V'(q)$$

This Euler-B method becomes explicit when  $H_q(q, p)$  only depends on  $q$

## 2. Euler-A method

$$\begin{aligned} q^{n+1} &= q^n + \Delta t H_p(q^{n+1}, p^n) \\ p^{n+1} &= p^n - \Delta t H_q(q^{n+1}, p^n) \end{aligned}$$

Numerically, the first equation is solved first. With the standard Hamiltonian, we then have a fully explicit scheme

$$\begin{aligned}q^{n+1} &= q^n + \Delta t \cdot p^n \\p^{n+1} &= p^n - \Delta t \cdot V'(q^{n+1})\end{aligned}$$

Applications: Celestial mechanics and Fluid dynamics

## Week 2: Lecture 3. Symplectic schemes

- Review on Hamiltonian system

Example to be working with (related to Newton's law):

$$H(q, p) = \frac{1}{2}p^2 + V(q), \quad \dot{q} = H_p = p, \quad \dot{p} = -H_q = -V'(q)$$

Two important properties:

1. **Energy**  $H(q, p)$  preserved along the flow

$$\frac{d}{dt} H(q(t), p(t)) = 0, \quad H(q(t), p(t)) = H(q(0), p(0))$$

2. **Symplectic form**  $dq \wedge dp$  preserved along the flow (constant area)

$$\frac{d}{dt} [dq(t) \wedge dp(t)] = 0$$

- Review on Euler-B method

$$q^{n+1} = q^n + \Delta t H_p(q^n, p^{n+1}), \quad \text{2nd step}$$

$$p^{n+1} = p^n - \Delta t H_q(q^n, p^{n+1}), \quad \text{1st step}$$

In general, the scheme is implicit due to the first step. But for our  $H(q, p)$  the scheme becomes fully explicit (decoupled Hamiltonian)

$$q^{n+1} = q^n + \Delta t \cdot p^{n+1}, \quad p^{n+1} = p^n - \Delta t \cdot V'(q^n)$$

- Euler-B method **exactly** preserves the symplectic form

For our decoupled Hamiltonian, the differentials are expressed as

$$dq^{n+1} = dq^n + \Delta t \cdot dp^{n+1}, \quad dp^{n+1} = dp^n - \Delta t V''(q^n) \cdot dq^n$$

Therefore, the symplectic form is

$$\begin{aligned} dq^{n+1} \wedge dp^{n+1} &= dq^n \wedge dp^{n+1} + \Delta t \cdot dp^{n+1} \wedge dp^{n+1} \\ &= dq^n \wedge dp^n - \Delta t \cdot V''(q^n) \cdot dq^n \wedge dq^n = dq^n \wedge dp^n \end{aligned}$$

For the general case, the last step uses the property of symmetry matrix

- Splitting method

For matrices  $A, B$  of  $O(\varepsilon)$ , we have

$$e^{A+B} = I + (A + B) + O(\varepsilon^2)$$

$$e^A e^B = (I + A + O(\varepsilon^2))(I + B + O(\varepsilon^2)) = I + (A + B) + O(\varepsilon^2)$$

When  $A, B$  are large matrices, calculation of  $e^A, e^B$  and  $e^A e^B$  is easy, while  $e^{A+B}$  is very hard

Example:  $A = \Delta$  and  $B = V(x)$ .  $e^{\varepsilon A} = e^{\varepsilon \Delta}$  is the kernel for heat equation,  $e^{\varepsilon B} = e^{\varepsilon V(x)}$  only involves matrix multiplication.  $u_t = \Delta u, u(\varepsilon) = e^{\varepsilon \Delta} u(0)$ .

Consider the Hamiltonian is decomposed into

$$H^{(1)}(p, q) = \frac{p^2}{2}, \quad H^{(2)}(p, q) = V(q)$$

**Step 1:** Apply explicit Euler on  $H^{(2)}$

$$\dot{q} = H_p^{(2)}(q, p), \quad \dot{p} = -H_q^{(2)}(q, p)$$

We obtain the intermediate results

$$\tilde{q} = q^n, \quad \tilde{p} = p^n - \Delta t \cdot V'(q^n)$$

**Step 2:** Apply explicit Euler on  $H^{(1)}$

$$\dot{q} = H_p^{(1)}(q, p), \quad \dot{p} = -H_q^{(1)}(q, p)$$

Using the intermediate results, we have

$$\begin{aligned} q^{n+1} &= \tilde{q} + \Delta t \cdot \tilde{p} = q^n + \Delta t \cdot p^{n+1} \\ p^{n+1} &= \tilde{p} = p^n - \Delta t \cdot V'(q^n) \end{aligned}$$

This is exactly the Euler-B method. If we apply  $H^{(1)}$  first, then we get the Euler-A method.

➤ Strang-splitting method

If  $A, B$  are of  $O(\varepsilon)$ , we have

$$e^{A+B} = e^{\frac{B}{2}} e^A e^{\frac{B}{2}} + O(\varepsilon^3)$$

This can be shown as below:

$$\begin{aligned} e^{\frac{B}{2}} e^A e^{\frac{B}{2}} &= \left( I + \frac{B}{2} + \frac{B^2}{8} + O(\varepsilon^3) \right) \left( I + A + \frac{A^2}{2} + O(\varepsilon^3) \right) \left( I + \frac{B}{2} + \frac{B^2}{8} + O(\varepsilon^3) \right) \\ &= I + (A + B) + \frac{A^2 + AB + BA + B^2}{2} + \tilde{O}(\varepsilon^3) = e^{A+B} + O(\varepsilon^3) \end{aligned}$$

**Remark:** We only need to implement  $e^A$  and  $e^B$  within accuracy of  $O(\varepsilon^3)$ .

$$e^{B/2} = L_{B/2} + O(\varepsilon^3), \quad e^A = L_A + O(\varepsilon^3), \quad L_{B/2}L_A L_{B/2} \approx e^{A+B} + O(\varepsilon^3)$$

➤ Strang-splitting method on Hamiltonian system

1. Implementation for  $H^{(2)}$  has  $O(\Delta t^3)$

$$\dot{q} = H_p^{(2)}(q, p) = 0, \quad \dot{p} = -H_q^{(2)}(q, p) = -V'(q), \quad \ddot{p} = -V''(q)\dot{q} = 0$$

$$\tilde{q} = q^n + \Delta t \cdot 0 + \frac{\Delta t^2}{2} \cdot 0 + O(\Delta t^3), \quad \tilde{p} = p^n - \Delta t \cdot V'(q^n) + \frac{\Delta t^2}{2} \cdot 0 + O(\Delta t^3)$$

2. Combine Euler-A and Euler-B by Strang-splitting

$$H^{(2)} \text{ for } \frac{\Delta t}{2}: \quad p^{n+1/2} = p^n - \frac{\Delta t}{2} \cdot V'(q^n)$$

$$H^{(1)} \text{ for } \Delta t: \quad q^{n+1} = q^n + \Delta t \cdot p^{n+1/2}$$

$$H^{(2)} \text{ for } \frac{\Delta t}{2}: \quad p^{n+1} = p^{n+1/2} - \frac{\Delta t}{2} \cdot V'(q^{n+1})$$

This can be combined into a better scheme on the staggered grid

$$p^{n+\frac{1}{2}} = p^{n-\frac{1}{2}} - \Delta t \cdot V'(q^n)$$

$$q^{n+1} = q^n + \Delta t \cdot p^{n+1/2}$$

This is the **Verlet integrator** with global error of  $O(\Delta t^2)$ . Since both Euler-A

and Euler-B preserve  $dq \wedge dp$ , Verlet integrator also does.

3. Hamiltonian is conserved for Verlet scheme

We use **harmonic oscillator** as a simple example to prove it

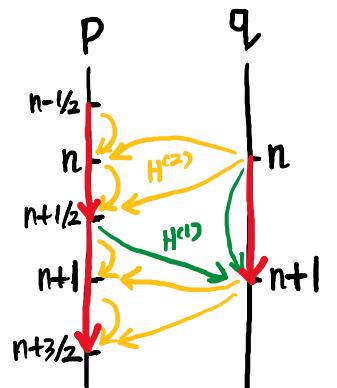
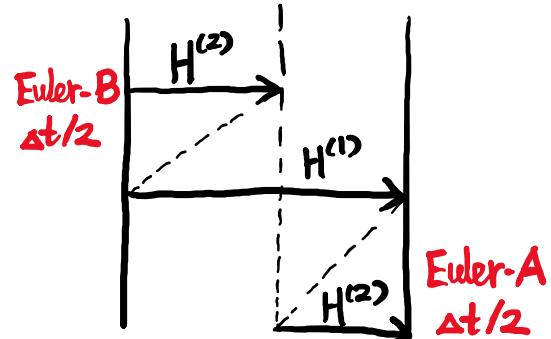
$$H(q, p) = \frac{p^2}{2} + \frac{q^2}{2}, \quad V'(q) = q$$

Start with a surrogate Hamiltonian

$$\tilde{H}_n \equiv \frac{1}{2} (p_{n-1/2}^2 + q_n^2 - \Delta t \cdot p_{n-1/2} \cdot q_n)$$

We can prove that this surrogate (“shadow”) is exactly preserved

$$\tilde{H}_{n+1} = \tilde{H}_n$$



This is shown as below, by repeatedly using the expression of Verlet scheme

$$2\tilde{H}_n = p_{n-\frac{1}{2}} \left( p_{n-\frac{1}{2}} - \Delta t \cdot q_n \right) + q_n^2 = p_{n-\frac{1}{2}} \cdot p_{n+\frac{1}{2}} + q_n^2 = p_{n+\frac{1}{2}} \left( p_{n+\frac{1}{2}} + \Delta t \cdot q_n \right) + q_n^2$$

$$2\tilde{H}_{n+1} = p_{n+\frac{1}{2}}^2 + q_{n+1} \left( q_{n+1} - \Delta t \cdot p_{n+\frac{1}{2}} \right) = p_{n+\frac{1}{2}}^2 + q_{n+1} q_n = p_{n+\frac{1}{2}}^2 + q_n \left( \Delta t \cdot p_{n+\frac{1}{2}} + q_n \right)$$

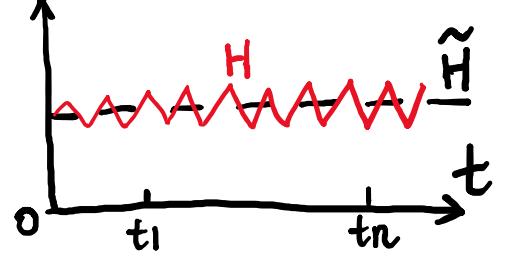
Now define the following  $H_n$  and it is approximately conserved

$$H_n \equiv H(q_n, p_{n-1/2}) = \frac{q_n^2}{2} + \frac{p_{n-1/2}^2}{2}$$

This is shown by calculating

$$\frac{2}{\Delta t} (H_n - \tilde{H}_n) = p_{n-1/2} \cdot q_n \leq \frac{p_{n-1/2}^2}{2} + \frac{q_n^2}{2} = H_n$$

$$H_n \leq \frac{2}{2 - \Delta t} \tilde{H}_n, \quad H_n - \tilde{H}_n \leq \frac{\Delta t}{2 - \Delta t} \tilde{H}_n$$



Therefore, we have

$$H_i - H_j = (H_i - \tilde{H}_i) - (H_j - \tilde{H}_j) + (\tilde{H}_i - \tilde{H}_j), \quad |H_i - H_j| \leq \frac{2\Delta t}{2 - \Delta t} \tilde{H}_0$$

For general Hamiltonian  $H(q, p)$ , the Verlet integrator guarantees

$$|H(q^n, p^{n-1/2}) - \text{Energy}| \leq O(\Delta t^{\text{order}}) \quad \text{for exp. long time } \sim e^{1/\Delta t}$$

### ➤ Verlet integrator is second order

From Taylor series, we have

$$\dot{y} = \begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} p \\ -V'(q) \end{pmatrix}, \quad \ddot{y} = \begin{pmatrix} \ddot{p} \\ -V''(q) \dot{q} \end{pmatrix} = \begin{pmatrix} -V'(q) \\ -V''(q) p \end{pmatrix}$$

$$\begin{pmatrix} q^{n+1} \\ p^{n+1} \end{pmatrix} = \begin{pmatrix} q^n \\ p^n \end{pmatrix} + \Delta t \begin{pmatrix} p^n \\ -V'(q^n) \end{pmatrix} + \frac{1}{2} \Delta t^2 \begin{pmatrix} -V'(q^n) \\ -V''(q^n) p^n \end{pmatrix} + O(\Delta t^3)$$

The analysis of Verlet scheme gives

$$q^{n+1} = q^n + \Delta t \left[ p^n - \frac{1}{2} \Delta t \cdot V'(q^n) \right] = q^n + \Delta t \cdot p^n - \frac{1}{2} \Delta t^2 \cdot V'(q^n)$$

$$p^{n+1} = p^n - \frac{1}{2} \Delta t \cdot V'(q^n) - \frac{1}{2} \Delta t \cdot V'(q^{n+1})$$

$$= p^n - \frac{1}{2} \Delta t \cdot V'(q^n) - \frac{1}{2} \Delta t [V'(q^n) + V''(q^n)(q^{n+1} - q^n)] + O(\Delta t^3)$$

$$= p^n - \Delta t \cdot V'(q^n) - \frac{1}{2} \Delta t \cdot V''(q^n) \cdot \Delta t \left[ p^n - \frac{1}{2} \Delta t \cdot V'(q^n) \right] + O(\Delta t^3)$$

$$= p^n - \Delta t \cdot V'(q^n) - \frac{1}{2} \Delta t^2 \cdot V''(q^n) p^n + O(\Delta t^3)$$

Up to the second order term, Verlet scheme is accurate. Hence its convergence is  $O(h^2)$

- Topics covered in ODE
  - ✓ Explicit and Implicit Euler, Trapezoid, RK
  - ✓ Stiffness
  - ✓ Splitting method
  - ✓ Symplectic schemes for  $H(q, p)$ , Euler-A/B and Verlet scheme

## Week 2: Lecture 4. FDM for Elliptic PDE

- Two-point boundary-value ODE problem

$$-u''(x) = f(x), \quad u(0) = u_L, \quad u(1) = u_R, \quad \text{in } \Omega = (0,1)$$

For a 2<sup>nd</sup> derivative, consider its value at point  $x$

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

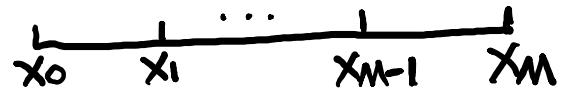
The 2<sup>nd</sup> derivative “compares the average of neighbors with itself”.

If  $u''(x) = 0$ , then the local average is equal to the value at the point. Unless  $u(x) \equiv \text{const.}$ , there exists neighbors larger than itself. Therefore, the max and min are always on the boundary. This is called the **maximal principle**.

- Discretization & Finite difference

To discretize: Domain and Derivative operator

Goal: Look for  $U_j \approx u(x_j)$  with  $U_0 = u_L, U_M = u_R$



The derivatives are approximated by finite difference with the following notations. For 1<sup>st</sup> derivatives, the forward, backward and central differences are defined as

$$\partial U_j \equiv \frac{U_{j+1} - U_j}{h}, \quad \bar{\partial} U_j \equiv \frac{U_j - U_{j-1}}{h}, \quad \hat{\partial} U_j = \frac{U_{j+1} - U_{j-1}}{2h}$$

For 2<sup>nd</sup> derivatives, we can show

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

- Accuracy of finite difference

Define the unknown exact solution as (restriction of exact solution on the grid)

$$u_j \equiv u(x_j)$$

The accuracy of derivative approximation is bounded by

$$|\hat{\partial} u_j - u'(x_j)| \leq Ch^2 |u|_{C^3}, \quad |u|_{C^3} \equiv \max_{x \in \Omega} |u^{(3)}(x)|$$

$$|\partial \bar{\partial} u_j - u''(x_j)| \leq Ch^2 |u|_{C^4}, \quad |u|_{C^4} \equiv \max_{x \in \Omega} |u^{(4)}(x)|$$

As an example, for central difference we have

$$u(x_{j+1}) = u(x_j) + hu'(x_j) + \frac{h^2}{2}u''(x_j) + \frac{h^3}{3!}u^{(3)}(\xi_j)$$

$$u_{j+1} - u_{j-1} = 2h \cdot u'(x_j) + O(h^3)|u|_{C^3}, \quad |\hat{u}_j - u'(x_j)| \leq Ch^2|u|_{C^3}$$

➤ Finite difference scheme

$$-u''(x) = f(x), \quad -\partial \bar{\partial} U_j = f_j \quad \text{for } j = 1, 2, \dots, M-1, \quad U_0 = u_L, \quad U_M = u_R$$

We have  $M-1$  unknowns and constraints. At the boundary ( $j = 1, M-1$ ), we have

$$2U_1 - U_2 = h^2 \cdot f_1 + U_0, \quad -U_{M-2} + 2U_{M-1} = h^2 \cdot f_{M-1} + U_M$$

The interior constraints give

$$-U_{j-1} + 2U_j - U_{j+1} = h^2 \cdot f_j$$

The linear system has the form of  $AU = F$ , which is  $h$ -dependent

$$\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{M-1} \end{bmatrix} = \begin{bmatrix} h^2 f_1 + U_0 \\ h^2 f_2 \\ \vdots \\ h^2 f_{M-1} + U_M \end{bmatrix}$$

The tridiagonal system can be solved by LU factorization

**Discrete**  $U_j \approx u_j \quad -U_{j-1} + 2U_j - U_{j+1} = h^2 f_j$

**Continuous (exact)**  $u(x_j) = u_j \quad -u''(x) = f(x)$

➤ Analysis of FDM: How accurate is  $U_j$ ?

General goal: How much the discretized exact solution satisfy the discrete equations (i.e., moving the continuous universe into the discrete universe)

**Discrete maximal principle:** For some vector  $V$  and the augmented linear system  $\tilde{A}$ , if the vector  $V$  satisfies the inequality in its piecewise sense

$$\tilde{A}_h V = \frac{1}{h^2} \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \\ & & & & & -1 \end{bmatrix} \begin{bmatrix} V_0 \\ V_1 \\ \vdots \\ V_{M-1} \\ V_M \end{bmatrix} \leq 0, \quad \tilde{A}_h \in \mathbb{R}^{(M-1) \times (M+1)}$$

Then the maximal value is on the boundary

$$\max_{j=0,1,\dots,M} V_j = \max(V_0, V_M)$$

**Proof.** For each interior  $j$  we have

$$-V_{j-1} + 2V_j - V_{j+1} \leq 0, \quad V_j \leq \frac{V_{j-1} + V_{j+1}}{2}$$

If  $V_j \equiv \text{const.}$ , it is obvious. If not, we can repeat the comparison until hitting the boundary. #

Now introduce the following notations

$$\Omega = (0,1), \quad \bar{\Omega} = [0,1], \quad |z|_S \equiv \max_{x_j \in S} |z_j|$$

As an example

$$|z|_\Omega = \max_{j=1,\dots,M-1} |z_j|, \quad |z|_{\bar{\Omega}} = \max_{j=0,\dots,M} |z_j|$$

**Lemma:** More generally, for any  $Z$  we have

$$\begin{aligned} \max_{j=0,1,\dots,M} |Z_j| &\leq \max(|Z_0|, |Z_M|) + C \max_{j=1,\dots,M-1} |(\tilde{A}_h Z)_j| \\ |Z|_{\bar{\Omega}} &\leq \max(|Z_0|, |Z_M|) + C \cdot |\tilde{A}_h Z|_\Omega \end{aligned}$$

**Proof.** Introduce a parabola

$$w(x) = \frac{1}{4} - \left( x - \frac{1}{2} \right)^2, \quad 0 \leq w(x) \leq \frac{1}{4}, \quad |w|_{C^4} = 0$$

Then at the mesh points we have

$$W_j = w(x_j) \leq \frac{1}{4}, \quad W_0 = W_M = 0, \quad (\tilde{A}_h W)_j = -w''(x_j) = 2$$

Define and calculate

$$\begin{aligned} V^\pm &= \pm Z - \frac{1}{2} |\tilde{A}_h Z|_\Omega W, \quad V_0^\pm = \pm Z_0 - \frac{1}{2} |\tilde{A}_h Z|_\Omega W_0 = \pm Z_0, \quad V_M^\pm = \pm Z_M \\ (\tilde{A}_h V^\pm)_j &= \pm (\tilde{A}_h Z)_j - |\tilde{A}_h Z|_\Omega \leq 0 \end{aligned}$$

Now we apply the Discrete Maximal Principle

$$\begin{aligned} V_j^\pm &= \pm Z_j - \frac{1}{2} |\tilde{A}_h Z|_\Omega W_j \leq \max(V_0^\pm, V_M^\pm) = \max(\pm Z_0, \pm Z_M) \\ Z_j &\leq \max(Z_0, Z_M) + \frac{1}{8} |\tilde{A}_h Z|_\Omega, \quad -Z_j \leq \max(-Z_0, -Z_M) + \frac{1}{8} |\tilde{A}_h Z|_\Omega \end{aligned}$$

Therefore we prove (the pointwise version)

$$|Z_j| \leq \max(|Z_0|, |Z_M|) + \frac{1}{8} |\tilde{A}_h Z|_\Omega$$

## Week 3: Lecture 5. Error analysis of FDM for Elliptic PDE

- Max-norm error analysis

**Theorem:** The error bound is

$$|U - u|_{\Omega} \leq Ch^2|u|_{C^4}$$

Remark: The method is  $O(h^2)$ , but the error depends on  $|u|_{C^4}$  (regularity) which may not exist (or regularity is not guaranteed)

**Proof.** We define

$$Z_j \equiv U_j - u_j, \quad Z_0 = Z_M = 0$$

To apply the Lemma, we need to calculate

$$\begin{aligned} (\tilde{A}_h Z)_j &= (AU)_j - (Au)_j = f_j - \left( -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} \right) \\ &= -u''(x_j) - \left( -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} \right) \end{aligned}$$

Therefore, the  $|\tilde{A}_h Z|_{\Omega}$  term is also bounded by

$$|(\tilde{A}_h Z)_j| = \left| \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} - u''(x_j) \right| \leq Ch^2|u|_{C^4}, \quad |\tilde{A}_h Z|_{\Omega} \leq Ch^2|u|_{C^4}$$

Using the previous Lemma, we have

$$|U - u|_{\Omega} \leq \max(|Z_0|, |Z_M|) + C \cdot |\tilde{A}_h Z|_{\Omega} \lesssim Ch^2|u|_{C^4}$$

This theorem is about the **infinity norm**  $|U - u|_{\Omega}$  using the maximal principle. However, for solving the linear system we usually use 2-norm  $\|\tilde{x} - x\|_{l^2}$

- $L^2$ -norm error estimate

For FDM, we have the tridiagonal linear systems for the approximated and exact solutions

$$A_h U = F, \quad A_h u = F - \tau, \quad |\tau_j| \leq O(h^2)$$

Therefore, we have

$$A_h Z \equiv A_h(U - u) = \tau, \quad \|Z\|_{l^2} \leq \|A_h^{-1}\|_{l^2 \rightarrow l^2} \cdot \|\tau\|_{l^2}$$

Now calculate the  $l^2$ -norms

$$\|\tau\|_{l^2} = \sqrt{\sum_{j=1}^{M-1} |\tau_j|^2} \lesssim \sqrt{Mh^4} \leq \sqrt{M}h^2 \quad (\sim h^{1.5})$$

For the  $A_h$  matrix, it is a **TST (Toeplitz symmetric tridiagonal) matrix**. For a general TST matrix of the form

$$A = \begin{bmatrix} \alpha & \beta & & & \\ \beta & \alpha & \beta & & \\ & \ddots & \ddots & \ddots & \\ & & \beta & \alpha & \beta \\ & & & \beta & \alpha \end{bmatrix}, \quad \alpha = \frac{2}{h^2}, \quad \beta = -\frac{1}{h^2}$$

The eigenvalue decomposition is

$$A = Q\Lambda Q^{-1}, \quad \lambda_j = \alpha + 2\beta \cos\left(\frac{\pi j}{M}\right), \quad q_{jk} = \sqrt{\frac{2}{M}} \sin\left(\frac{\pi j k}{M}\right)$$

For our  $A_h$  the eigenvalues are

$$\lambda_j = \frac{4}{h^2} \left[ \frac{1}{2} - \frac{1}{2} \cos\left(\frac{\pi j}{M}\right) \right] = \frac{4}{h^2} \sin^2\left(\frac{\pi j}{2M}\right)$$

The first and the last eigenvalues are about

$$\lambda_1 \approx \frac{\pi^2}{h^2 M^2} = \pi^2 \sim O(1), \quad \lambda_{M-1} \approx \frac{4}{h^2}$$

Similarly, for the inverse matrix  $A_h^{-1}$  we have

$$\|A_h^{-1}\|_{l^2 \rightarrow l^2} = \max \lambda_j^{-1} \sim O(1)$$

The meaning of the matrix norm is

$$\|A\|_{l^2 \rightarrow l^2} = \max_v \frac{\|Av\|_{l^2}}{\|v\|_{l^2}} = \max_{\|v\|_{l^2}=1} \|Av\|_{l^2}$$

As a summary, on the average sense we have  $|z_j|$  is about  $O(h^2)$

$$\|Z\|_{l^2} \leq \|A_h^{-1}\|_{l^2 \rightarrow l^2} \cdot \|\tau\|_{l^2} \lesssim \sqrt{M} h^2, \quad |z_j| \sim O(h^2)$$

➤ 2D PDE example: Poisson equation

$$-\Delta u = f, \quad x \in \Omega = (0,1)^2, \quad u(x) = 0 \text{ on boundary } \partial\Omega$$

Discretize the domain: Introduce  $j = (j_1, j_2)$  and define

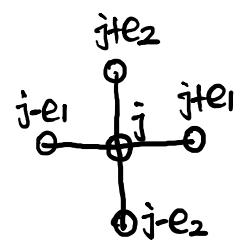
$$x_j = (j_1 h, j_2 h), \quad u_j \equiv u(x_j), \quad U_j \approx u_j, \quad e_1 = (1,0), \quad e_2 = (0,1)$$

Discretize the operator: The Laplacian is approximated with central difference

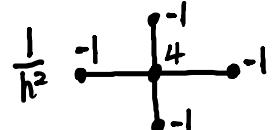
$$\partial_1 \partial_1 u \approx \bar{\partial}_1 \partial_1 U, \quad -\bar{\partial}_1 \partial_1 U_j - \bar{\partial}_2 \partial_2 U_j = f_j$$

At point  $j$  the equation becomes

$$\frac{1}{h^2} (4U_j - U_{j-e_1} - U_{j+e_1} - U_{j-e_2} - U_{j+e_2}) = f_j, \quad 1 \leq j_1, j_2 \leq M-1$$



Stencil



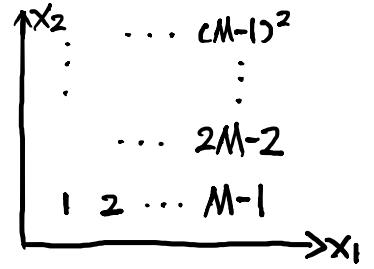
The linear system needs to be written carefully. The number of the unknowns and linear equations is  $(M-1)^2$ . The **ordering** is important.

## Row ordering scheme

The vector  $U$  and  $F$  is ordered as

$$U = [U_1, \dots, U_{M-1}, U_M, \dots, U_{2M-2}, \dots, \dots, U_{(M-1)^2}]^T$$

$$F = [F_1, \dots, F_{M-1}, F_M, \dots, F_{2M-2}, \dots, \dots, F_{(M-1)^2}]^T$$



The linear system  $AU = F$  looks like

$$A_h = \frac{1}{h^2} \begin{bmatrix} D & -I & 0 & 0 & 0 & \cdots & 0 \\ -I & D & -I & 0 & 0 & \cdots & 0 \\ 0 & -I & D & -I & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -I & D & -I & 0 \\ 0 & \cdots & \cdots & 0 & -I & D & -I \\ 0 & \cdots & \cdots & \cdots & 0 & -I & D \end{bmatrix}, \quad \text{with } D = \begin{bmatrix} 4 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 4 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 4 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 4 & -1 & 0 \\ 0 & \cdots & \cdots & 0 & -1 & 4 & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & -1 & 4 \end{bmatrix}$$

As an intuitive example, for  $M - 1 = 3$  we have

$$A_h = \frac{1}{h^2} \begin{array}{c|ccc|ccc} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 \\ \hline -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & -1 \\ \hline 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 \end{array}$$

Now the matrix is a banded matrix of  $b = O(M)$  band. The LU cost scales as

$$Nb^2 = M^2 \cdot M^2 = M^4 = N^2$$

### Week 3: Lecture 6-1. Error analysis of 2D FDM for Elliptic PDE

➤ Error analysis for Poisson equation

**Discrete maximal principle:** Consider  $V = (V_j)$ ,  $0 \leq j_1, j_2 \leq M$ . Suppose  $(\tilde{A}_h V)_j \leq 0$ ,

then  $V$  achieves its maximal value on boundary.

**Proof:** The condition gives

$$4V_j - V_{j-e_1} - V_{j+e_1} - V_{j-e_2} - V_{j+e_2} \leq 0, \quad V_j \leq \text{Ave}(V_{j-e_1}, V_{j+e_1}, V_{j-e_2}, V_{j+e_2})$$

This means  $V_j$  is smaller than at least one of its neighbors, unless the function is constant.

Recursively, this comparison continues until hitting the boundary

**Lemma 2:** Suppose  $(\tilde{A}_h Z)_j$  is not all zero, we generally have

$$|Z|_{\bar{\Omega}} \leq \max_{j \in \partial\Omega} |Z_j| + C |\tilde{A}_h Z|_{\Omega}$$

**Proof:** Similarly construct a function

$$w(x) = \frac{1}{2} - \left| x - \left( \frac{1}{2}, \frac{1}{2} \right) \right|^2 = \frac{1}{4} - \left( x_1 - \frac{1}{2} \right)^2 + \frac{1}{4} - \left( x_2 - \frac{1}{2} \right)^2$$

Then we apply Discrete Maximal Principle to the following vector

$$Z^+ = Z - \frac{1}{4} |\tilde{A}_h Z|_{\Omega} W$$

**Max-norm error analysis:** To analyze the error, we have

$$(A_h U)_j = f_j, \quad (A_h u)_j = -\Delta u(x_j) - \tau_j = f_j - \tau_j, \quad \tau_j \leq Ch^2 |u|_{C^4}$$

$$(A_h Z)_j = [A_h(U - u)]_j = \tau_j$$

Based on Lemma 2, we obtain

$$|Z|_{\bar{\Omega}} \leq |Z|_{\partial\Omega} + C |\tau|_{\Omega} \lesssim h^2 |u|_{C^4}$$

**L2-norm error analysis:** Following 1D case, we still have

$$A_h Z \equiv A_h(U - u) = \tau, \quad \|Z\|_{l^2} \leq \|A_h^{-1}\|_{l^2 \rightarrow l^2} \cdot \|\tau\|_{l^2}$$

**Kronecker product:**

$$B \otimes C = \begin{bmatrix} b_{11}C & \cdots & b_{1n}C \\ \vdots & \ddots & \vdots \\ b_{m1}C & \cdots & b_{mn}C \end{bmatrix}$$

Therefore, the 2D FDM matrix  $A_h$  can be written from the 1D FDM matrix  $B_h$

$$A_h = B_h \otimes I + I \otimes B_h, \quad B_h, I \in \mathbb{R}^{(M-1) \times (M-1)}, \quad A \in \mathbb{R}^{(M-1)^2 \times (M-1)^2}$$

The eigenvalue decomposition becomes

$$A_h = (Q \Lambda Q^T) \otimes (Q I Q^T) + (Q I Q^T) \otimes (Q \Lambda Q^T)$$

$$\begin{aligned}
&= (Q \otimes Q)(\Lambda \otimes I)(Q \otimes Q)^T + (Q \otimes Q)(I \otimes \Lambda)(Q \otimes Q)^T \\
&= (Q \otimes Q)(\Lambda \otimes I + I \otimes \Lambda)(Q \otimes Q)^T
\end{aligned}$$

The eigenvalues of  $A_h$  are

$$\lambda_{p,q} = \lambda_p(B) + \lambda_q(B), \quad \lambda_p(B) = \frac{4}{h^2} \sin^2 \left( \frac{\pi p}{2M} \right)$$

Hence, the eigenvalue spectrum is basically the same with  $B_h$ , with an extra factor of 2

Precisely, we have the L2-norm error bound

$$\|Z\|_{l^2} \leq \|A_h^{-1}\|_{l^2 \rightarrow l^2} \cdot \|\tau\|_{l^2} \lesssim \|\tau\|_{l^2}$$

On the “average sense”, we have  $|z_j| \sim O(h^2)$ .

## Week 3: Lecture 6-2. Hilbert Space

Finite element method solves the following elliptic PDE

$$-\Delta u = f \quad \text{or} \quad -\nabla \cdot (a(x)\nabla u) = f$$

### ➤ Introduction to Hilbert space

	Linear Algebra	Hilbert Space
Vectors	$\mathbb{R}^n$	“ $\infty$ -dim” vector space $V$
Linear functional (Linear form)	$w^T v: \mathbb{R}^n \rightarrow \mathbb{R}$	$L: V \rightarrow \mathbb{R}$
Bilinear form	$u^T B v: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$	$B: V \times V \rightarrow \mathbb{R}$
Sym. Pos. Def.	$B_{ij} = B_{ji}$ $w^T B w > 0, \quad w \neq 0$	$B(u, v) = B(v, u)$ $B(u, u) > 0, \quad \forall u \in V, \quad u \neq 0$
Distance (Norm)	$\ u\ _B = \sqrt{u^T B u}$	$\ u\ _B = \sqrt{B(u, u)}$
Semi-norm	$\ u\ _{\text{semi}} \geq 0$ (e.g., $\ u\ _{\text{semi}} =  u_1 $ ) Triangle inequality	$\ u\ _{\text{semi}} \geq 0$ Triangle inequality

**Example:**  $V = L^2(\mathbb{R})$ , square-integrable function space is a Hilbert space

$$B(f, g) = \int_{-\infty}^{+\infty} f(x)g(x) dx, \quad \|f\|_{L^2} = \sqrt{\int_{-\infty}^{+\infty} f^2(x) dx}$$

**Example:**  $V = H^1(\mathbb{R})$ , an example of Sobolev space

$$B(f, g) = \int_{-\infty}^{+\infty} fg + \nabla f \cdot \nabla g dx, \quad \|f\|_{H^1} = \sqrt{\int_{-\infty}^{+\infty} f^2 + |\nabla f|^2 dx}$$

**Definition.** Hilbert space  $(V, \|\cdot\|_B)$  is a complete inner product space. This means:

1. Bilinear form  $B$  on vector space  $V$  is sym. pos. def. (i.e., inner product space)
2.  $V$  is complete (i.e., every Cauchy sequence in  $V$  is convergent)

### Equivalent norms

For linear algebra, two norms are always equivalent:

$$\forall u \in \mathbb{R}^n, \quad \exists \gamma_{AB}, \Gamma_{AB} > 0, \quad \gamma_{AB}\|u\|_B \leq \|u\|_A \leq \Gamma_{AB}\|u\|_B$$

For Hilbert space, we define that the two norms are equivalent when

$$\forall u \in V, \quad \exists \gamma_{AB}, \Gamma_{AB} > 0, \quad \gamma_{AB}\|u\|_B \leq \|u\|_A \leq \Gamma_{AB}\|u\|_B$$

## Bounded linear operator

For linear algebra, the linear operator  $M: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is always bounded.

For Hilbert space, we define that the linear operator  $L: (V, \|\cdot\|_B) \rightarrow (W, \|\cdot\|_A)$  is bounded if

$$\exists C_L \in \mathbb{R}, \quad \|Lu\|_A \leq C_L \|u\|_B, \quad \forall u \in V$$

Therefore, we can define the norm of a bounded linear operator

$$\|L\| = \sup_{u \in V \setminus \{0\}} \frac{\|Lu\|_A}{\|u\|_B}$$

**Dual space:** The set of all bounded linear functionals  $L: V \rightarrow \mathbb{R}$  on  $V$  is the dual space  $V^*$ . The (dual) norm in  $V^*$  is (norm on a dual space, same definition as above)

$$\|L\|_{V^*} = \sup_{u \in V \setminus \{0\}} \frac{|L(u)|}{\|u\|_B}, \quad L \in V^*$$

## Riesz representation theorem

For each bounded  $L: V \rightarrow \mathbb{R}$  on the Hilbert space  $(V, \|\cdot\|_B)$ , there is a unique  $x_L \in V$  such that

$$Lu = B(x_L, u), \quad \forall u \in V \quad \quad \|L\|_{B^*} = \|x_L\|_B$$

**Remark:** We can thus identify the linear functionals  $L \in V^*$  with the associated  $x_L \in V$ . Now, we obtain another Hilbert space  $(V^*, \|\cdot\|_{B^*})$

**Example:** In linear algebra, any linear functional  $Lu = w^T u$  can be represented by an SPD bilinear form (B-SPD)

$$w^T u = x_L^T B u, \quad x_L = B^{-1} w$$

The inverse  $B^{-1}$  is also SPD. Now we can define a dual way to measure  $L$  in the dual space, and it is the same with the norm of its representation  $x_L$ :

$$\|L\|_{B^*} \equiv \sqrt{w^T B^{-1} w}, \quad \|x_L\|_B = \sqrt{w^T B^{-1} w}$$

The two norms are linked through the representation theorem

## Poincare's inequality

$$\|f\|_{H^1} = \int_{[0,1]} f^2 + |\nabla f|^2 dx, \quad \|\nabla f\| = \int_{[0,1]} |\nabla f|^2 dx, \quad \|f\| = \int_{[0,1]} f^2 dx$$

Poincare's inequality gives

$$\|\nabla f\| \leq \|f\|_{H^1} \leq C \cdot \|\nabla f\|$$

This implies that the first two norms are equivalent

$$\|\nabla f\| \leftrightarrow \|f\|_{H^1}, \quad \|\nabla f\| \leftrightarrow \|f\|, \quad \|f\|_{H^1} \leftrightarrow \|f\|$$

## Week 4: Lecture 7. Introduction of Sobolev Space

➤ Review on Hilbert space

	Linear Algebra	Hilbert Space
Norm	$\ x\  = \sqrt{x^T B x}$	$\ u\ _B = \sqrt{B(u, u)}$
Linear functional	$l^T(x) = l^T x \in \mathbb{R}$	$L: V \rightarrow \mathbb{R}$
Riesz Represent.	$l^T x = u^T B x, \quad u = B^{-1} l$	$\exists u_l \in V, \quad l(x) = B(u_l, x), \quad \forall x \in V$
Dual norm	$\ l\ _{B^{-1}} = \ u\  = \sqrt{l^T B^{-1} l}$	$\ l\ _{B^*} = \ u_l\ _B$

In finite element method (FEM), for Hilbert space  $(V, \|\cdot\|_B)$ , we want another bilinear form  $A$  which is **symmetric**

$$A(u, v) = A(v, u)$$

**bounded** with respect to  $B$

$$|A(u, v)| \leq M \|u\|_B \|v\|_B, \quad \forall u, v \in V$$

and **coercive** in  $V$

$$A(u, u) \geq \alpha \|u\|_B^2, \quad \forall u \neq 0$$

The above is analogous to max. eigenvalue  $< \infty$  and min. eigenvalue  $> 0$  in linear algebra

Now we can define another space  $(V, \|\cdot\|_A)$ , and the two norms are **equivalent**, because from the properties of  $A$  we have

$$\underbrace{\alpha \|u\|_B^2 \leq A(u, u)}_{\text{coercive}} = \underbrace{\|u\|_A^2 \leq M \|u\|_B^2}_{\text{bounded}}, \quad \sqrt{\alpha} \|u\|_B \leq \|u\|_A \leq \sqrt{M} \|u\|_B$$

Suppose  $A(\cdot, \cdot)$  is symmetric, bounded, and coercive. We claim that  $(V, \|\cdot\|_A)$  is also a Hilbert space. Given a linear functional, by Riesz rep. theorem we have

$$\exists u_l \in V, \quad s.t. \quad A(u_l, v) = l(v), \quad \forall v \in V, \quad \|u_l\|_A = \|l\|_{A,*}$$

Now let  $v = u_l$ , we have

$$\underbrace{\alpha \|u_l\|_B^2 \leq A(u_l, u_l)}_{\text{coercive}} = \underbrace{l(u_l) \leq \|u_l\|_B \|l\|_{B,*}}_{\text{bounded linear functional}}$$

This leads to (energy estimate)

$$\alpha \|u_l\|_B \leq \|l\|_{B,*}, \quad \|u_l\|_B \leq \frac{1}{\alpha} \|l\|_{B,*}$$

**Remark.** Even though we use a problem-specific norm  $A$ , we can still translate the result into a ‘common’ norm  $B$ , with the only difference being a constant factor

## Projection

Suppose a subspace  $W \subset V$ , the projection  $w \in W$  of  $v \in V$  satisfies  $\|v - w\|_B^2$  is minimized, and  $v - w \perp_B$  any vector  $w' \in W$

## Connection with optimization

Let  $u \in V$  be the solution of  $A(u, v) = l(v)$ ,  $\forall v \in V$ . This is equivalent to an optimization

$$\min_u \frac{1}{2} A(u, u) - l(u)$$

As an analogy, the linear system  $Au = l$  is equivalent to an optimization problem

$$\min_u E(u), \quad E(u) = \frac{1}{2} u^T A u - l^T u, \quad \frac{\partial E}{\partial u} = A u - l = 0$$

**Example:**  $L_2$  space for a compact domain  $\Omega$

$$L_2(\Omega) = \{f \mid \int_{\Omega} f^2(x) dx < \infty\}$$

$$A(f, g) = \int_{\Omega} \overline{f(x)} \cdot g(x) dx, \quad \|f\|_{L_2(\Omega)} = \sqrt{\int_{\Omega} |f(x)|^2 dx}$$

### Weak derivatives

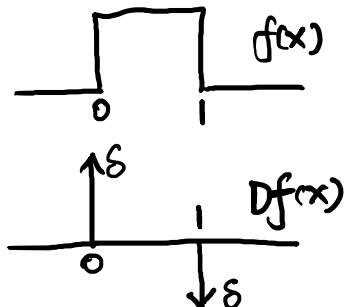
For usual derivatives, we have the integration by parts

$$\int_{\Omega} f' g dx = f g|_{\partial\Omega} - \int_{\Omega} g' f dx$$



Now consider  $f$  is not differentiable (e.g., jumps, kinks). If test function  $g$  is smooth, the LHS is bad while the RHS is still appropriate. We want to define the derivative for  $f(x)$  as a linear functional  $Df$  such that for any smooth  $g(x)$

$$(Df)(g) = \underbrace{- \int_{\Omega} g' f dx}_{\text{Linear Functional}}$$



As an example, for the boxcar function  $f(x)$

$$(Df)(g) = - \int_{\mathbb{R}} g' f dx = - \int_0^1 g'(x) dx = g(0) - g(1)$$

### Sobolev space

Suppose  $f, g \in L_2(\mathbb{R})$ , the weak derivative  $Df$  is a distribution. If  $Df$  and  $Dg$  happen to be also in  $L_2(\mathbb{R})$ , then we can define the following bilinear form

$$A(f, g) = \int_{\mathbb{R}} (fg + Df Dg) dx$$

Then we have the following space  $H^1(\mathbb{R})$

$$H^1(\mathbb{R}) \equiv \{f \mid A(f, f) < \infty\}, \quad \|f\|_{H^1} = \sqrt{\int_{\mathbb{R}} [f^2(x) + (Df)^2(x)] dx}$$

If  $Df \in L_2$  and  $D(Df) \equiv D^2f \in L_2$ , then we can define  $H^2$  space

$$\|f\|_{H^2} = \sqrt{\int_{\mathbb{R}} [f^2(x) + (Df)^2(x) + (D^2f)^2(x)] dx}$$

Similarly, we can define  $H^k$  space

$$\|f\|_{H^k} = \sqrt{\int_{\mathbb{R}} \sum_{j=1}^k |D^j f(x)|^2 dx}$$

Now we have a sequence of spaces

$$L_2 = H^0 \supset H^1 \supset \dots \supset H^k$$

The corresponding sequence of the dual spaces is (the dual space of  $L_2$  is itself)

$$L_2 = H^0 \subset H^{-1} \subset \dots \subset H^{-k}$$

The full sequence is thus

$$\dots \supset H^{-2} \supset H^{-1} \supset H^0 = L_2 \supset H^1 \supset H^2 \supset \dots$$

For a finite domain (FEM purpose), we define

$$H_0^1(\Omega) \equiv \{f \in H^1(\Omega) \text{ but also vanishes on the boundary of } \Omega\}$$

$$\|f\|_{H_0^1(\Omega)} = \sqrt{\int_{\Omega} f^2(x) + |\nabla f|^2(x) dx}$$

More rigorously,  $H_0^1(\Omega)$  is the completion of  $C_0^\infty(\Omega)$  function w.r.t. the norm  $\|f\|_{H_0^1(\Omega)}$

### ➤ Poincare's inequality (1D)

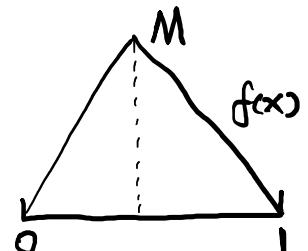
Suppose  $\Omega = (0, 1)$  and  $f \in H_0^1(\Omega)$ , we have

$$\|f\|_{L_2} \lesssim \|Df\|_{L_2}$$

If this is not true, then  $\|f\|_{L_2}/\|Df\|_{L_2}$  is unbounded. This contradicts intuition, as having a large maximum requires steep slopes, given that  $f = 0$  on the boundary.

**Proof.** We have

$$f(x) = \int_0^x f'(y) dy$$



The Cauchy-Schwartz inequality gives

$$|f(x)| = \left| \int_0^x 1 \cdot f'(y) dy \right| \leq \sqrt{\int_0^1 1^2 dy} \cdot \sqrt{\int_0^1 f'(y)^2 dy}, \quad |f(x)| \leq \|Df\|_{L_2}$$

Therefore, we have

$$\|f\|_{L_2} = \sqrt{\int_0^1 f^2(x) dx} \leq \sqrt{\int_0^1 \|Df\|_{L_2}^2 dx} = \|Df\|_{L_2}, \quad \|f\|_{L_2(\Omega)} \leq C(\Omega) \|\nabla f\|_{L_2(\Omega)}$$

The implication of Poincare's inequality is

$$\begin{aligned} \int_{\Omega} |Df|^2 dx &\leq \int_{\Omega} (|f|^2 + |Df|^2) dx \leq 2 \int_{\Omega} |Df|^2 dx \\ \|\nabla f\|_{L_2} &\leq \|f\|_{H_0^1} \leq \sqrt{2} \|\nabla f\|_{L_2} \end{aligned}$$

This says that  $\|\nabla f\|_{L_2}$  can be regarded as a norm on  $H_0^1(\Omega)$ , and in fact,  $\|\nabla f\|_{L_2}$  is equivalent w.r.t.  $H_0^1(\Omega)$  norm

**Remark.** It is very important that we consider  $f \in H_0^1(\Omega)$ . If  $f$  doesn't vanish on boundary, Poincare's inequality doesn't hold, and one counter-example is a non-zero constant function.

➤ Finite element method (FEM) for 1D problem

Recall in FDM we solve  $-u'' = f$  with zero on the boundary. The consequence is

$$\max_j |U_j - u(x_j)| \leq Ch^2 |u|_{C^4}$$

The regularity term  $|u|_{C^4}$  is bad and will be fixed for FEM.

For FEM, we focus on

$$-[a(x)u'(x)]' + Cu(x) = f(x), \quad u(0) = u(1) = 0,$$

with smooth function  $a(x)$  and constant  $C$  satisfying

$$0 < \underline{a} \leq a(x) \leq \bar{a}, \quad C \geq 0$$

The above problem is in the divergence form. In the operator form, we have (set  $C = 0$ )

$$-D[a(Du)] = D^T[a(Du)] = f(x)$$

The LHS is a sym. pos. def. operator.  $-D = D^T$  is the adjoint of  $D$  in this finite domain based on integration by parts.

Now focus on  $H_0^1(\Omega)$  which is a Hilbert space, with any test function  $\varphi(x) \in H_0^1(\Omega)$

$$\int_{\Omega} -[au'(x)]'\varphi(x) dx = \int_{\Omega} f(x)\varphi(x) dx$$

From integration by parts, we obtain

$$-au'\varphi|_{\partial\Omega} + \int_{\Omega} a(x)u'(x)\varphi'(x) dx = \int_{\Omega} f(x)\varphi(x) dx, \quad \forall \varphi(x) \in H_0^1(\Omega)$$

The LHS is sym. pos. def. (SPD)

## Summary

In the integral formulation, we only need one derivative for  $u$ , and thus consider  $u \in H_0^1(\Omega)$ . Therefore,  $u$  is a **weak solution** of  $-(au')' = f$  if we have  $u \in H_0^1(\Omega)$  and for any  $\varphi$

$$\int_{\Omega} a(x)u'(x)\varphi'(x) dx = \int_{\Omega} f(x)\varphi(x) dx, \quad \forall \varphi(x) \in H_0^1(\Omega)$$

## Topic for next lecture:

The Hilbert space  $H_0^1(\Omega)$  provides the base norm  $\|\cdot\|_B \rightarrow \|\cdot\|_{H_0^1(\Omega)}$ . The problem specific bilinear form is

$$A(u, \varphi) \equiv \int_{\Omega} a(x)u'(x)\varphi'(x) dx$$

This bilinear form is symmetric, bounded, and coercive, as shown below

$$|A(u, \varphi)| \leq \bar{a} \int_{\Omega} |u'\varphi'| dx \leq C \|u\|_{H_0^1(\Omega)} \|\varphi\|_{H_0^1(\Omega)}$$

$$A(u, u) = \int_{\Omega} au'u' dx \gtrsim \int_{\Omega} |u'|^2 dx \gtrsim \|u\|_{H_0^1(\Omega)}^2$$

## Week 4: Lecture 8. FEM for 1D elliptic PDE

- Review on weak solution

In the integral formulation,  $u \in H_0^1(\Omega)$  is a **weak solution** of  $-(au')' = f$  if we have

$$\int_{\Omega} a(x)u'(x)\varphi'(x) dx = \int_{\Omega} f(x)\varphi(x) dx, \quad \forall \varphi(x) \in H_0^1(\Omega)$$

More compactly, we can write it as

$$A(u, \varphi) = (f, \varphi)_{L_2}, \quad \forall \varphi(x) \in H_0^1(\Omega)$$

Note that the RHS is just the usual  $L_2(\Omega)$  inner product. If  $f \in H_0^1(\Omega)$ , then the RHS should be strictly written as  $f(\varphi)$ , but when  $f \in L_2(\Omega)$  then it just becomes the inner product.

### Advantage

1. Only need one derivative for  $u$
2. Bilinear form, we can apply Hilbert space technique

### Existence of solution

The bilinear form  $A(u, \varphi)$  is symmetric, bounded, and coercive

$$\begin{aligned} |A(u, \varphi)| &\leq \bar{a} \int_{\Omega} |u' \varphi'| dx \leq \bar{a} \sqrt{\int_{\Omega} |u'|^2 dx} \sqrt{\int_{\Omega} |\varphi'|^2 dx} \leq \bar{a} \|u\|_{H_0^1(\Omega)} \|\varphi\|_{H_0^1(\Omega)} \\ A(u, u) &= \int_{\Omega} au'u'dx \geq \underline{a} \int_{\Omega} |u'|^2 dx \geq \underline{a} \|u\|_{H_0^1(\Omega)} \end{aligned}$$

Now we show that  $A(u, \varphi)$  is an SPD bilinear form on  $H_0^1(\Omega)$  and it is equivalent to the usual norm  $\|u\|_{H_0^1(\Omega)}$ . From Riesz rep. theorem, we can find  $u_f \in H_0^1(\Omega)$  such that

$$A(u_f, \varphi) = f(\varphi) = (f, \varphi)$$

If  $A$  is non-symmetric, then we need to use Lax-Milgram theorem

- Finite element method (FEM)

The goal is to look for  $u_h \in S_h \subsetneq H_0^1(\Omega)$  such that

$$A(u_h, \varphi_h) = (f, \varphi_h), \quad \forall \varphi_h \in S_h$$

where  $S_h$  is a finite dimension subspace of  $H_0^1(\Omega)$

**Remark.** The subspace  $S_h \subsetneq H_0^1(\Omega)$  is finite dimension, easy to work with, and captures  $u$

**Remark.** The problems are equivalent to the following optimization problems

$$\min_{u \in H_0^1(\Omega)} \frac{1}{2} A(u, u) - (f, u), \quad \min_{u_h \in S_h} \frac{1}{2} A(u_h, u_h) - (f, u_h)$$

The FEM is a finite dimension optimization, in contrast to the infinite dimension one.

The left figure indicates the infinite dimension problem, while the red line in the right figure indicates the FEM finite dimension problem. The advantage is that we **reduce dimension**, while the drawback is that there is a small **gap** between  $u_h$  and the true solution  $u$



### FEM partition

$$0 = x_0 < x_1 < \dots < x_M = 1$$

Small intervals  $K_j$  are elements with

$$K_j = [x_{j-1}, x_j], \quad h_j = x_j - x_{j-1}, \quad h = \max_j h_j$$



The discrete solution will be found in the finite dimension subspace  $S_h$  defined as

$$S_h = \{ f \in C(\bar{\Omega}) \mid \text{piecewise linear w.r.t this partition, } f(0) = f(1) = 0 \}$$

### Courant hat basis for $S_h$

The hat basis is defined as

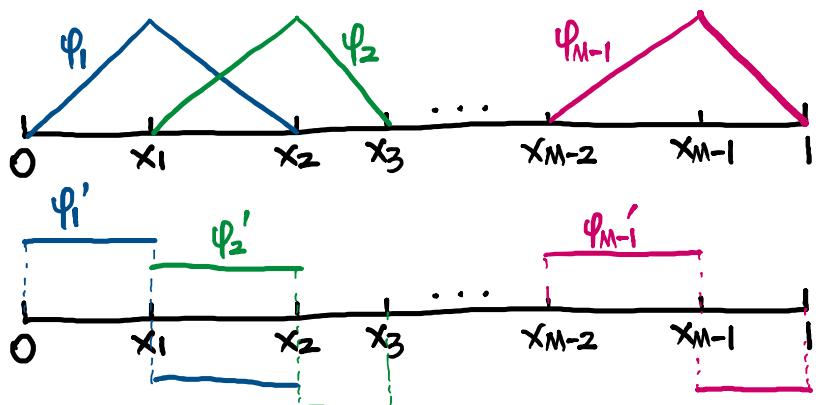
$$\varphi_i(x_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

Note that the intervals need not to be uniformly spaced

It is easy to recognize

$$S_h = \text{span}(\varphi_1, \varphi_2, \dots, \varphi_{M-1})$$

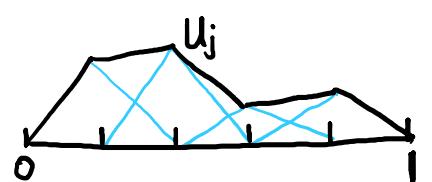
The representation by  $\{\varphi_i\}$  is uniquely determined by function values at interior points



### FEM linear system

Now write our target solution  $u_h$  as

$$u_h = \sum_{j=1}^{M-1} \varphi_j(x) U_j$$



We look for  $\{U_j\}$  to satisfy (now denote test functions as  $\chi_h$ )

$$A(u_h, \chi_h) = A\left(\sum_{j=1}^{M-1} \varphi_j U_j, \chi_h\right) = (f, \chi_h), \quad \forall \chi_h \in S_h$$

With  $\{\varphi_i\}$  being the basis, we only need to check  $\chi_h \equiv \varphi_i$

$$A\left(\sum_{j=1}^{M-1} \varphi_j U_j, \varphi_i\right) = (f, \varphi_i), \quad i = 1, 2, \dots, M-1$$

Now we need to solve

$$\sum_{j=1}^{M-1} A(\varphi_i, \varphi_j) U_j = (\varphi_i, f), \quad i = 1, 2, \dots, M-1$$

This becomes the following linear system

$$A_{ij} = A(\varphi_i, \varphi_j), \quad F_i = (\varphi_i, f), \quad AU = F, \quad A \in \mathbb{R}^{(M-1) \times (M-1)}$$

The matrix  $A$  is a “shadow” of  $A(u, v)$  on  $S_h$

1. Since  $A(u, v)$  is SPD, then the restriction of  $A(u, v)$  on  $S_h$ , the matrix  $A$ , is also SPD
2. Matrix  $A$  is tridiagonal. For example, the supports of  $\varphi_1$  and  $\varphi_3$  are disjoint
3.  $A_{ii} > 0, A_{ij} < 0$

$$A_{12} = \int a(x) \varphi'_1(x) \varphi'_2(x) dx < 0$$

**Example:** Uniform spacing  $h$  with  $a(x) \equiv 1$

$$A(\varphi_i, \varphi_i) = \int_0^1 |\varphi'_i(x)|^2 dx = \frac{2}{h}, \quad A(\varphi_i, \varphi_{i\pm 1}) = -\frac{1}{h}$$

This matrix is the same with FDM matrix. It seems that the constant factor is  $1/h$ , but the missing  $h$  is into the  $F_i = (\varphi_i, f)$ . FEM and FDM matrices happen to be the same for the uniform spacing discretization

### ➤ Error analysis of FEM

For the problem

$$-[a(x)u'(x)]' = f, \quad u(0) = u(1) = 0$$

Existence of solution indicates

$$\begin{aligned} f \in H^{-1} &\Rightarrow u \in H_0^1(\Omega) \\ f \in L_2 = H^0 &\Rightarrow u \in H^2(\Omega) \end{aligned}$$

**Claim.** Solution  $u$  is smoother than  $f$

The weak formulation is

$$A(u, \chi) = (\chi, f), \quad \forall \chi \in H_0^1(\Omega)$$

The FEM problem is

$$A(u_h, \chi_h) = (\chi_h, f), \quad \forall \chi_h \in S_h \subsetneq H_0^1(\Omega)$$

Difference between the two problems gives

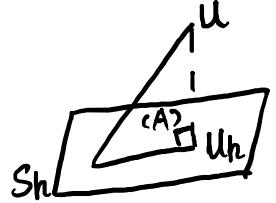
$$A(u - u_h, \chi_h) = 0, \quad \forall \chi_h \in S_h$$

Therefore, with respect to the inner product  $A(\cdot, \cdot)$  we have

$$(u - u_h) \perp_A \chi_h, \quad \forall \chi_h \in S_h$$

The geometric characterization of FEM solution  $u_h$  is

$$\|u - u_h\|_A = \min_{\chi_h} \|u - \chi_h\|_A$$



Or equivalently

$$\|u - u_h\|_A \leq \|u - \chi_h\|_A, \quad \forall \chi_h \in S_h$$

**Remark.** As  $\|\cdot\|_A$  is equivalent to  $H_0^1(\Omega)$  norm, we have

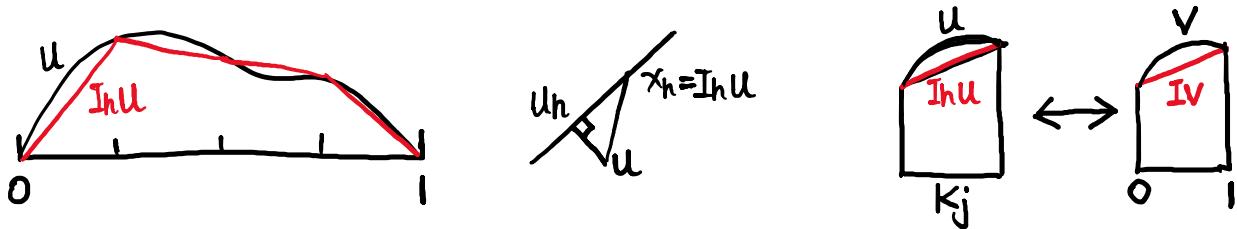
$$C_1 \|u - u_h\|_{H_0^1(\Omega)} \leq \|u - u_h\|_A \leq \|u - \chi_h\|_A \leq C_2 \|u - \chi_h\|_{H_0^1(\Omega)}, \quad \forall \chi_h \in S_h$$

$$\|u - u_h\|_{H_0^1(\Omega)} \leq C \|u - \chi_h\|_{H_0^1(\Omega)}, \quad \forall \chi_h \in S_h$$

This is not problem-specific anymore. All we need to do now is to find a good  $\chi_h \in S_h$  such that  $\|u - \chi_h\|_{H_0^1}$  is small. This  $\chi_h$  can depend on  $u$  as long as  $\chi_h \in S_h$

### Approximation of $u$ in $S_h$

Define the piecewise linear interpolation of  $u \in H_0^1(\Omega)$  as  $I_h u$ . We will use  $\chi_h \equiv I_h u$



### Error estimate of FEM

Consider an element  $K_j$  and extend the range to  $[0,1]$ . We have

$$\|v - Iv\|_{L_2(\Omega)} \lesssim \|v''\|_{L_2(\Omega)}$$

**Proof.** Poincare's inequality gives

$$\|v - Iv\|_{L_2} \lesssim \|(v - Iv)'\|_{L_2}$$

By mean value theorem,  $(v - Iv)'$  has a zero-crossing point  $x_0$ . Therefore, we can modify the proof of Poincare's inequality using this  $x_0$  and similarly obtain

$$\|(v - Iv)'\|_{L_2} \lesssim \|(v - Iv)''\|_{L_2}$$

Therefore, because  $Iv$  is piecewise linear, we have  $(Iv)'' = 0$  and thus

$$\|v - Iv\|_{L_2} \lesssim \|(v - Iv)'\|_{L_2} \lesssim \|(v - Iv)''\|_{L_2} = \|v''\|_{L_2}$$

Going back to the original  $K_j$  element, as we have

$$u(x) = v\left(\frac{x - x_{j-1}}{h}\right), \quad u' = \frac{v'}{h}, \quad u'' = \frac{v''}{h^2}$$

we can obtain

$$\|u - I_h u\|_{L_2(K_j)} \lesssim h^2 \|u''\|_{L_2(K_j)}, \quad \|(u - I_h u)'\|_{L_2(K_j)} \lesssim h \|u''\|_{L_2(K_j)}$$

For the whole domain  $\Omega = (0,1)$ , we have

$$\begin{aligned} \|u - I_h u\|_{L_2}^2 &= \sum_{j=1}^M \int_{K_j} |u - I_h u|^2 dx \lesssim h^4 \sum_{j=1}^M \int_{K_j} |u''|^2 dx = (h^2 \|u''\|_{L_2})^2 \\ \|(u - I_h u)'\|_{L_2}^2 &= \sum_{j=1}^M \int_{K_j} |(u - I_h u)'|^2 dx \lesssim h^2 \sum_{j=1}^M \int_{K_j} |u''|^2 dx = (h \|u''\|_{L_2})^2 \end{aligned}$$

These results show (scaling argument)

$$\|u - I_h u\|_{L_2} \lesssim h^2 \|u''\|_{L_2}, \quad \|(u - I_h u)'\|_{L_2} \lesssim h \|u''\|_{L_2}$$

Summarizing all together, we obtain the **Theorem\***

$$\|u - u_h\|_{H_0^1} \lesssim \underbrace{\|u - I_h u\|_{H_0^1}}_{\text{Poincare}} \lesssim \underbrace{\|(u - I_h u)'\|_{L_2}}_{\text{PDE Theory}} \lesssim h \|u''\|_{L_2} \lesssim h \|u\|_{H^2} \lesssim h \|f\|_{L_2}$$

Recall in FDM, we prove that

$$\max_j |u - u_j| \lesssim h^2 |u|_{C^4}$$

But note that  $H_0^1(\Omega)$  norm applied in FEM is a stronger one. Now the question is: can we estimate the  $L_2$  error for FEM

## Week 5: Lecture 9. FEM for 2D elliptic PDE

- $L_2$  error analysis of 1D FEM

### Theorem

$$\|u - u_h\|_{L_2(\Omega)} \lesssim h^2 \|u\|_{H^2(\Omega)}$$

**Proof.** Define the error function  $e_h = u_h - u$ , and we have

$$A(e_h, \chi_h) = 0, \quad \forall \chi_h \in S_h$$

This means that  $e_h \perp_A \chi_h$ . Now consider the following problem

$$\mathcal{A}\varphi = -(a'\varphi)' = e_h, \quad \varphi(0) = \varphi(1) = 0$$

Within  $L_2$ -norm we can write

$$\|e_h\|_{L_2}^2 = (e_h, e_h)_{L_2} = e_h^T \cdot e_h = e_h^T \cdot \mathcal{A}\varphi = e_h^T \mathcal{A}(\varphi - I_h\varphi)$$

The last step uses the perpendicular property since  $I_h\varphi \in S_h$ . Using Theorem\*

$$e_h^T \mathcal{A}(\varphi - I_h\varphi) \lesssim \|e_h\|_A \cdot \|\varphi - I_h\varphi\|_A \lesssim \|e_h\|_A \cdot h \|\varphi\|_{H^2} \lesssim \|e_h\|_A \cdot h \|e_h\|_{L_2}$$

Therefore, using Theorem\* again, we have

$$\|e_h\|_{L_2} \lesssim h \cdot \|e_h\|_A \lesssim h^2 \|f\|_{L_2}$$

- FEM for 2D elliptic PDE



$$\mathcal{A}u = -\nabla \cdot (a(x)\nabla u) = f, \quad u(x) = 0 \text{ on } \partial\Omega$$

We require that  $\Omega$  is a convex, piecewise linear boundary. This implies  $u \in H^2(\Omega)$ , which guarantees that  $u$  is continuous in 2D and we can construct  $I_h u$

### Weak solution

$$\int_{\Omega} -\nabla \cdot (a\nabla u) \varphi \, dx = - \int_{\partial\Omega} \varphi (a\nabla u) \hat{n} \, dx + \int_{\Omega} a\nabla u \cdot \nabla \varphi \, dx = \int_{\Omega} f \varphi \, dx$$

We search for solution  $u \in H_0^1(\Omega)$  such that

$$A(u, \varphi) = \int_{\Omega} a\nabla u \cdot \nabla \varphi \, dx = \int_{\Omega} f \varphi \, dx = (f, \varphi), \quad \forall \varphi \in H_0^1(\Omega)$$

Again, we consider  $a(x)$  has upper and lower limits, and it is smooth in  $\Omega$ . The bilinear form is symmetric, bounded and coercive on  $H_0^1(\Omega)$

$$\begin{aligned} |A(u, \varphi)| &\leq \bar{a} \int_{\Omega} |\nabla u| |\nabla \varphi| \, dx \leq \bar{a} \sqrt{\int_{\Omega} |\nabla u|^2 \, dx} \sqrt{\int_{\Omega} |\nabla \varphi|^2 \, dx} \leq \bar{a} \|u\|_{H_0^1(\Omega)} \|\varphi\|_{H_0^1(\Omega)} \\ A(u, u) &\geq \underline{a} \int_{\Omega} |\nabla u|^2 \, dx \gtrsim \underline{a} \|u\|_{H_0^1(\Omega)} \end{aligned}$$

Therefore,  $A(u, \varphi)$  is an SPD bilinear form on  $H_0^1(\Omega)$  and it is equivalent to  $\|u\|_{H_0^1(\Omega)}$ .

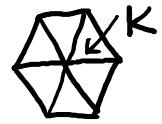
## FEM solution

We search for  $u_h \in S_h \subsetneq H_0^1(\Omega)$  such that

$$A(u_h, \varphi_h) = (f, \varphi_h)$$

The domain is decomposed into a triangular mesh that satisfies

1. Small maximum radius  $h = \max h_K$
2. No triangles with angle close to  $0^\circ$  or  $180^\circ$
3. No bad decomposition like this

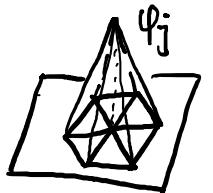


Now we can define  $S_h$  as

$$S_h = \{f \in C(\bar{\Omega}) \mid \text{piecewise linear w.r.t triangulation, } f(\partial\Omega) = 0\}$$

The basis function is similar to the hat function

$$S_h \equiv \text{span}\{\varphi_j\}_{\text{interior}} \quad \dim(S_h) = \# \text{ of interior vertices}$$



Same with 1D case, now we need to solve

$$\sum_{j=1}^{\#\text{int}} A(\varphi_i, \varphi_j) U_j = (\varphi_i, f), \quad \forall i = \text{interior point}$$

This becomes the following linear system

$$A_{ij} = A(\varphi_i, \varphi_j), \quad F_i = (\varphi_i, f), \quad AU = F, \quad A \in \mathbb{R}^{\#\text{int} \times \#\text{int}}$$

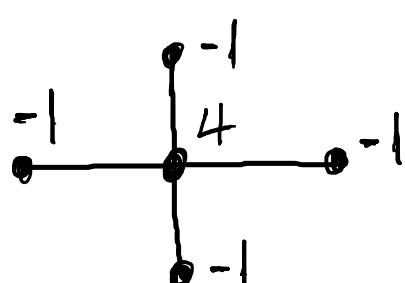
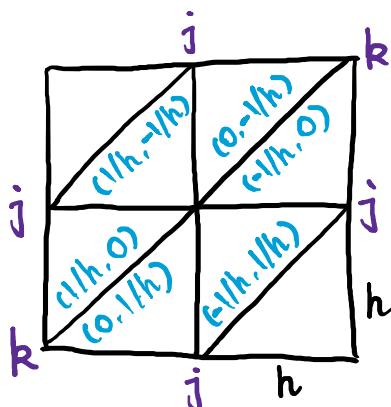
The matrix  $A$  is SPD. It is also very sparse, as we need  $i$  and  $j$  only 1 step away

**Example:**  $a(x) \equiv 1, -\nabla \cdot (\nabla u) = f$

$$A_{ii} = \int_{\Omega} |\nabla \varphi_i|^2 \, dx = \frac{h^2}{2} \cdot 2 \left( \frac{2}{h^2} + \frac{1}{h^2} + \frac{1}{h^2} \right) = 4$$

$$A_{ij} = \frac{h^2}{2} \left( -\frac{1}{h^2} - \frac{1}{h^2} \right) = -1, \quad A_{ik} = 0$$

The stencil is the same with 2D finite difference



## Error analysis

Repeat the 1D analysis, we obtain the same picture

$$A(e_h, \varphi_h) = (f, \varphi_h), \quad e_h \perp_A \varphi_h, \quad \forall \varphi_h \in S_h$$

$$\|u - u_h\|_A = \min_{\chi_h \in S_h} \|u - \chi_h\|_A, \quad \|u - u_h\|_{H_0^1(\Omega)} \lesssim \|u - \chi_h\|_{H_0^1(\Omega)}$$

We similarly do piecewise linear interpolation  $\chi_h = I_h u$ , but require other conditions. This is because in 2D  $u \in H_0^1(\Omega)$  does not imply  $u$  is continuous. We need  $u \in H^2(\Omega)$ , which requires the domain  $\Omega$  to be convex (polygonal) and  $f \in L_2(\Omega)$ . Therefore, we can interpolate  $\chi_h = I_h u$

By **scaling argument**, we similarly have

$$\|\nu - I\nu\|_{L_2} \lesssim \|\nabla^2 \nu\|_{L_2}, \quad \|\nabla(\nu - I\nu)\| \lesssim \|\nabla^2 \nu\|_{L_2}$$

$$\|u - I_h u\|_{L_2} \lesssim h^2 \|\nabla^2 u\|_{L_2}, \quad \|\nabla(u - I_h u)\|_{L_2} \lesssim h \|\nabla^2 u\|_{L_2}$$

$$\|u - I_h u\|_{H_0^1} \lesssim h \|u\|_{H^2}$$

Also with the requirement that the triangles are regular. Otherwise, the Jacobian matrix will be bad when mapping the element into an equilateral triangle. Therefore, we obtain

$$\|u - I_h u\|_{H_0^1} \leq h \|u\|_{H^2} \leq h \|f\|_{L_2}, \quad \|u - u_h\|_{H_0^1} \lesssim h \|f\|_{L_2}$$

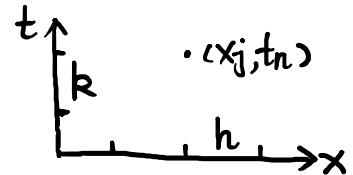
**Theorem.**  $L_2$ -error estimate

$$\|u - u_h\|_{L_2} \lesssim h^2 \|u\|_{H^2} \lesssim h^2 \|f\|_{L_2}$$

## Week 6: Lecture 10. FDM for parabolic PDE

➤ 1D heat equation

$$u_t = u_{xx}, \quad u(x, 0) = v(x), \quad x \in \mathbb{R}, \quad t > 0$$



### Discretization

$$x_j = jh, \quad t_n = nk, \quad U_j^n \approx u(x_j, t_n)$$

At grid point  $(x_j, t_n)$  we have **forward difference in time** and **central difference in space**

$$u_t \approx \partial_t U_j^n \equiv \frac{U_j^{n+1} - U_j^n}{k}, \quad u_{xx} \approx \partial_x \bar{\partial}_x U_j^n \equiv \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}$$

The equation becomes

$$\partial_t U_j^n = \partial_x \bar{\partial}_x U_j^n, \quad \frac{U_j^{n+1} - U_j^n}{k} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2}$$

The usual way to write the system is

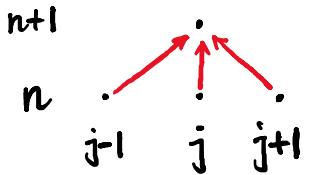
$$U_j^{n+1} = \lambda U_{j+1}^n + (1 - 2\lambda) U_j^n + \lambda U_{j-1}^n, \quad \lambda = \frac{k}{h^2}$$

This is an **explicit method** since it does not involve solving a linear system

### Time marching

Define the column vector at time  $t_n$  as  $U^n$ , we have the following convolution

$$U^{n+1} = \begin{bmatrix} 1 - 2\lambda & \lambda & & & \\ \lambda & 1 - 2\lambda & \lambda & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda & 1 - 2\lambda & \lambda \\ & & & \lambda & 1 - 2\lambda \end{bmatrix} U^n = E_{kh} U^n$$



### Error analysis

To simplify the analysis, we assume that  $U(x)$  is defined everywhere, not only on the grid points. We now consider  $U$  is continuous and the numerical scheme is

$$U^{n+1}(x) = \lambda U^n(x - h) + (1 - 2\lambda) U^n(x) + \lambda U^n(x + h)$$

The operator becomes **discrete convolution**

$$(E_{kh} U^n)(x) = [(\lambda \delta_{-h} + (1 - 2\lambda) \delta_0 + \lambda \delta_h) * U^n](x)$$

For the approximate solution  $U$

$$\partial_t U = \partial_x \bar{\partial}_x U, \quad U^{n+1} = E_{kh} U^n$$

For the exact solution  $u$  (exact solution applied to FD), we denote the **local error**  $\tau^n$

$$\frac{u^{n+1}(x) - u^n(x)}{k} = \frac{u^n(x + h) - 2u^n(x) + u^n(x - h)}{h^2} + \tau^n(x), \quad u^{n+1} = E_{kh} u^n + k\tau^n$$

Define the error  $Z^n = U^n - u^n$

$$Z^n = E_{kh} Z^{n-1} + (-k\tau^{n-1}) = E_{kh}^n Z^0 - k[E_{kh}^{n-1}\tau^0 + E_{kh}^{n-2}\tau^1 + \dots + E_{kh}^0\tau^{n-1}]$$

Then we have

$$\|Z^n\| \leq k[\|E_{kh}^{n-1}\|\|\tau^0\| + \dots + \|E_{kh}^0\|\|\tau^{n-1}\|] \leq n \cdot \max_p \|E_{kh}^p\| \cdot \max_p \|k\tau^p\|$$

To ensure  $\|Z_n\|$  is small, we need the following conditions

$$\underbrace{\|E_{kh}^p\| \leq C}_{\textcircled{1}}, \quad \underbrace{\|\tau^p\| \text{ small}}_{\textcircled{2}}, \quad \forall p$$

In order to have small  $\|E_{kh}^p\|$ , note that it is a convolution operator. We analyze it with the Fourier transform (i.e., a rotation)

$$f(x) \in L_2(\mathbb{R}) \rightarrow \mathcal{F}\{f\}(\xi) = \hat{f}(\xi) \in L_2(\mathbb{R})$$

The convention we use is

$$\hat{f}(\xi) = \int_{\mathbb{R}} e^{-ix\xi} f(x) dx, \quad \check{g}(x) = \int_{\mathbb{R}} e^{ix\xi} g(\xi) d\left(\frac{\xi}{2\pi}\right)$$

The Fourier transform has the following properties

$$\check{\hat{f}} \equiv f, \quad \forall f \in L_2, \quad \|f\|_{L_2(dx)} = \|\hat{f}\|_{L_2\left(\frac{d\xi}{2\pi}\right)}, \quad \mathcal{F}\{f * g\}(\xi) = \hat{f}(\xi)\hat{g}(\xi)$$

For the convolution property, we can show it as

$$\begin{aligned} \mathcal{F}\{f * g\}(\xi) &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-ix\xi} [f(y)g(x-y) dy] dx \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-iy\xi} e^{-i(x-y)\xi} f(y)g(x-y) dy dx = \hat{f}(\xi)\hat{g}(\xi) \end{aligned}$$

Now we analyze the FDM scheme in the Fourier domain

$$\begin{aligned} U^{n+1} &= E_{kh} U^n = [\lambda\delta_{-h} + (1-2\lambda)\delta_0 + \lambda\delta_h] * U^n \\ \widehat{U}^{n+1}(\xi) &= [1 - 2\lambda + 2\lambda \cos(h\xi)] \cdot \widehat{U}^n(\xi) \end{aligned}$$

The condition  $\textcircled{1}$

$$\|E_{kh}^p\| \leq C, \quad \forall p \leftrightarrow \|E_{kh}\| \leq 1$$

is now equivalent to

$$\max_{\xi} |1 - 2\lambda + 2\lambda \cos(h\xi)| \leq 1$$

When  $\xi \approx 0$ , the above condition is met. Focus on  $\xi \approx \pi/h$ , we can obtain

$$|1 - 4\lambda| \leq 1, \quad 0 \leq \lambda \leq \frac{1}{2}$$

**Conclusion.** When  $\lambda > 1/2$  the solution blows up

## Interpretation

The high frequency mode (large  $\xi$ ), which immediately disappears (“boring mode”), enforces a tiny time step ( $k \leq h^2/2$ ). This corresponds to **stiffness**. In Fourier domain, we have

$$\frac{d}{dt} \hat{u}(\xi, t) = -\xi^2 \hat{u}(\xi, t)$$

From ODE analysis, we can get the same interpretation.

## Local error

For condition ②, the local error is

$$k\tau^n = u^{n+1} - E_{kh}u^n = (e^{k\Delta} - E_{kh})u^n$$

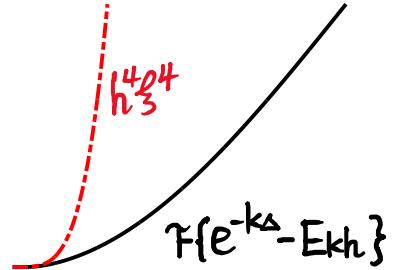
Fourier transform gives

$$k\hat{\tau}^n = \mathcal{F}\{e^{k\Delta} - E_{kh}\} \cdot \hat{u}^n(\xi) = \{e^{-k\xi^2} - [1 - 2\lambda + 2\lambda \cos(h\xi)]\} \cdot \hat{u}^n(\xi)$$

$$\mathcal{F}\{e^{k\Delta} - E_{kh}\} = \left(\frac{k^2}{2} - \frac{kh^2}{12}\right)\xi^4 + \dots = C \cdot h^4\xi^4 + \dots$$

Therefore, we have the following bound

$$|k\hat{\tau}^n(\xi)| \lesssim \underbrace{h^4\xi^4|\hat{u}^n(\xi)|}_{\text{PDE: } v \text{ is initial cond.}} \leq h^4\xi^4|\hat{v}(\xi)|$$



The  $L_2$ -error estimate becomes

$$\|k\tau^n\|_{L_2}^2 = \int_{\mathbb{R}} |k\hat{\tau}^n(\xi)|^2 d\xi \lesssim h^8 \int_{\mathbb{R}} |\xi^4 \hat{v}(\xi)|^2 d\xi \leq h^8 \|v\|_{H^4(\mathbb{R})}^2$$

The last step applies Fourier isometry ( $\|v^{(4)}(x)\| \sim \|\xi^4 \hat{v}(\xi)\|$ ), and thus we have

$$\|k\tau^n\|_{L_2} \lesssim h^4 \|v\|_{H^4}$$

Finally, we obtain

$$\|Z^n\| \leq n \cdot \max_p \|E_{kh}^p\| \cdot \max_p \|k\tau^p\| \lesssim \underbrace{nh^4 \|v\|_{H^4}}_{t_n=nk, k \sim h^2} \lesssim t_n k \|v\|_{H^4}$$

**Summary:** **Explicit** FDM for heat equation is

1. Convergent when  $\lambda = k/h^2 \leq 0.5$  (stiffness)
2. 1<sup>st</sup> order in time step  $k$
3. Depends on  $\|v\|_{H^4}$  (similar to FDM for elliptic PDE)

## Week 6: Lecture 11. Implicit FDM for parabolic PDE

### ➤ Implicit FDM scheme

$$\frac{U^{n+1}(x) - U^n(x)}{k} = \frac{U^{n+1}(x+h) - 2U^{n+1}(x) + U^{n+1}(x-h)}{h^2}$$

$$(1 + 2\lambda)U^{n+1}(x) - \lambda U^{n+1}(x-h) - \lambda U^{n+1}(x+h) = U^n(x)$$

The linear system becomes

$$A_{kh}U^{n+1} = U^n, \quad U^{n+1} = A_{kh}^{-1}U^n = E_{kh}U^n$$

In the Fourier domain, we have

$$[1 + 2\lambda - 2\lambda \cos(h\xi)] \cdot \hat{U}^{n+1}(\xi) = \hat{U}^n(\xi), \quad \hat{E}_{kh}(\xi) = \frac{1}{1 + 2\lambda - 2\lambda \cos(h\xi)}$$



### Error analysis

Following the same procedure, we have

$$Z^{n+1} = E_{kh}Z^n + (-k\tau^n), \quad \|Z^n\| \leq n \cdot \max_p \|E_{kh}^p\| \cdot \max_p \|k\tau^p\|$$

For the operator part, we always have

$$\left| \frac{1}{1 + 2\lambda - 2\lambda \cos(h\xi)} \right| \leq 1$$

Therefore, we can arbitrarily choose  $\lambda = k/h^2$  to satisfy this condition

For the local error part

$$k\hat{\tau}^n = \mathcal{F}\{e^{k\Delta} - E_{kh}\} \cdot \hat{u}^n(\xi) = \left[ e^{-k\xi^2} - \frac{1}{1 + 2\lambda - 2\lambda \cos(h\xi)} \right] \cdot \hat{u}^n(\xi)$$

Based on Taylor expansion, we obtain

$$\mathcal{F}\{e^{k\Delta} - E_{kh}\} \lesssim k^2\xi^4 + kh^2\xi^4, \quad \|k\hat{\tau}^n(\xi)\|_{L_2} \lesssim \|(k^2\xi^4 + kh^2\xi^4)\hat{v}(\xi)\|_{L_2}$$

The Fourier isometry gives

$$\|k\tau^n\|_{L_2} \lesssim (k^2 + kh^2) \|v\|_{H^4}$$

Therefore, the error bound is

$$\|Z^n\| \lesssim n(k^2 + kh^2) \|v\|_{H^4} \leq t_n(k + h^2) \|v\|_{H^4}$$

**Summary:** Implicit FDM for heat equation is

1. No longer stiff
2. 1<sup>st</sup> order in time
3. Depends on  $\|v\|_{H^4}$

➤ Trapezoidal rule FDM scheme

$$(1 + \lambda)U^{n+1}(x) - \frac{\lambda}{2}U^{n+1}(x-h) - \frac{\lambda}{2}U^{n+1}(x+h) = (1 - \lambda)U^n(x) + \frac{\lambda}{2}U^n(x-h) + \frac{\lambda}{2}U^n(x+h)$$

The linear system becomes

$$AU^{n+1} = BU^n, \quad U^{n+1} = A^{-1}BU^n = E_{kh}U^n$$

In the Fourier domain, we have

$$\hat{E}_{kh}(\xi) = \frac{1 - \lambda + \lambda \cos(h\xi)}{1 + \lambda - \lambda \cos(h\xi)}$$

### Error analysis

For the operator part, we always have

$$\left| \frac{1 - \lambda[1 - \cos(h\xi)]}{1 + \lambda[1 - \cos(h\xi)]} \right| \leq 1$$

Therefore, we can still arbitrarily choose  $\lambda = k/h^2$  to satisfy this condition

For the local error part

$$k\hat{\tau}^n = \mathcal{F}\{e^{k\Delta} - E_{kh}\} \cdot \hat{u}^n(\xi) = \left[ e^{-k\xi^2} - \frac{1 - \lambda + \lambda \cos(h\xi)}{1 + \lambda - \lambda \cos(h\xi)} \right] \cdot \hat{u}^n(\xi)$$

Based on Taylor expansion, we have

$$\mathcal{F}\{e^{k\Delta} - E_{kh}\} \lesssim (k^3 \xi^6 + kh^2 \xi^4)$$

Following the same procedure gives

$$\|k\tau^n\|_{L_2} \lesssim k^3 \|v\|_{H^6} + kh^2 \|v\|_{H^4}$$

Therefore, the error bound is

$$\|Z^n\| \lesssim t_n(k^2 \|v\|_{H^6} + h^2 \|v\|_{H^4})$$

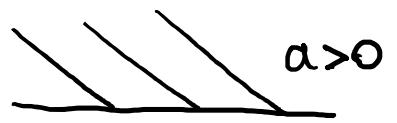
**Summary:** **Trapezoidal rule** FDM for heat equation is

1. No longer stiff
2. 2<sup>nd</sup> order in time
3. Depends on  $\|v\|_{H^6}$

## Week 7: Lecture 12. FDM for hyperbolic PDE

- 1D transport (advection) equation

$$\frac{\partial u}{\partial t} = a \frac{\partial u}{\partial x}$$



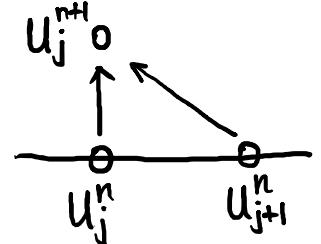
When  $a > 0$ , the function is transported in  $-x$  direction. This implies that our scheme should be **upwind** to use the information from the previous time step

### Discretization

$$x_j = jh, \quad t_n = nk, \quad U_j^n \approx u(x_j, t_n)$$

At grid point  $(x_j, t_n)$  we use **forward difference in time and space**

$$\begin{aligned} \frac{U_j^{n+1} - U_j^n}{k} &= a \frac{U_{j+1}^n - U_j^n}{h} \\ U_j^{n+1} &= a\lambda U_{j+1}^n + (1 - a\lambda) U_j^n, \quad \lambda = \frac{k}{h} \end{aligned}$$



The backward difference in space is bad, as it is not upwind.

### Analysis

Again, we assume  $U^n(x)$  as a continuous function in space and obtain the **lifted equation**

$$U^{n+1}(x) = a\lambda U^n(x + h) + (1 - a\lambda) U^n(x)$$

In the operator form

$$U^{n+1} = E_{kh} U^n, \quad E_{kh} = (1 - a\lambda) \delta_0(x) + a\lambda \delta_{-h}(x)$$

Now apply FD scheme to the exact solution

$$u^{n+1} = E_{kh} u^n + k\tau^n, \quad Z^{n+1} = E_{kh} Z^n + (-k\tau^n)$$

Telescoping gives

$$\begin{aligned} Z^n &= E_{kh}^{n-1}(-k\tau^0) + \dots + E_{kh}^0(-k\tau^{n-1}) \\ \|Z^n\| &\leq n \cdot \max_p \|E_{kh}^p\| \cdot \max_p \|k\tau^p\| \end{aligned}$$

In the Fourier domain

$$\widehat{U}^{n+1}(\xi) = \widehat{E}_{kh}(\xi) \widehat{U}^n(\xi), \quad \widehat{E}_{kh}(\xi) = (1 - a\lambda) + a\lambda e^{i\xi h}$$

The operator norm should be bounded by 1

$$\|\widehat{E}_{kh}\| = \max_{\xi} |(1 - a\lambda) + a\lambda e^{i\xi h}| \leq 1$$

We only need to focus on the situations where  $e^{i\xi h} = \pm 1$ , which leads to

$$a\lambda \leq 1, \quad k \leq \frac{h}{a}$$

**Remark.** When  $a$  is large, the speed is fast, and the characteristics are ‘flat’. At this case, the time step  $k$  should be small

The local error term is analyzed as

$$k\tau^n \equiv u^{n+1} - E_{kh}u^n = (e^{k\cdot a\partial_x} - E_{kh})u^n, \quad \|k\tau^n\| \leq \|(e^{ka\partial_x} - E_{kh})u^n\|$$

Using Fourier isometry, we have

$$\|k\hat{\tau}^n\| \leq \|e^{ika\xi} - (1 - a\lambda) - a\lambda e^{ih\xi}\| \cdot \|\hat{u}^n\|$$

By Taylor expansion

$$\begin{aligned} e^{ika\xi} - (1 - a\lambda) - a\lambda e^{ih\xi} &= ika\xi - \frac{1}{2}k^2a^2\xi^2 - a\lambda\left(ih\xi - \frac{1}{2}h^2\xi^2\right) + O(h^3, k^3) \\ &= ika\xi - a\lambda ih\xi + C(h^2 + k^2)\xi^2 + O(h^3, k^3) \\ &= Ch^2\xi^2 + O(h^3) \end{aligned}$$

Therefore, we have the error bound

$$|k\hat{\tau}^n| \leq Ch^2\xi^2|\hat{u}^n(\xi)| \leq Ch^2\xi^2|\hat{v}(\xi)|$$

$$\|k\tau^n\|_{L_2}^2 \leq \int_{\xi} (h^2\xi^2)^2 |\hat{v}(\xi)|^2 d\xi = h^4 \|v\|_{H^2}^2$$

$$\|Z^n\| \leq n \cdot \max_p \|E_{kh}^p\| \cdot \max_p \|k\tau^p\| \leq n \cdot h^2 \|v\|_{H^2} \leq Th \|v\|_{H^2}$$

**Theorem.** If  $\lambda = k/h < 1/a$ , then we have

$$\|E_{kh}\| \lesssim 1, \quad \|k\tau^n\| \leq Ch^2 \|v\|_{H^2}, \quad \|Z^n\| \leq t_n h \|v\|_{H^2}$$

### Meta-theorem: Lax-equivalence theorem

1. Stability: Time step control
2. Consistency: Plugging in PDE solution in FD scheme gives small error

With these two conditions, then the FD scheme is convergent

➤ Influence of stencil on transport equation

For the backward difference scheme, we have

$$U_j^{n+1} - U_j^n = a\lambda(U_j^n - U_{j-1}^n), \quad U_j^{n+1} = (1 + a\lambda)U_j^n - a\lambda U_{j-1}^n$$

This scheme is not stable for  $a > 0$ , as the operator is not bounded

$$|\hat{E}_{kh}(\xi)| = |1 + a\lambda - a\lambda e^{-i\xi h}| \geq 1, \quad \text{when } a > 0$$

However, when  $a < 0$  we need backward difference in space, which is upwinding.

For the central difference scheme

$$U_j^{n+1} - U_j^n = \frac{a\lambda}{2}(U_{j+1}^n - U_{j-1}^n), \quad U_j^{n+1} = U_j^n + \frac{a\lambda}{2}(U_{j+1}^n - U_{j-1}^n)$$

This scheme is not stable, as the operator is not bounded

$$|\hat{E}_{kh}(\xi)| = |1 + ia\lambda \sin(h\xi)| \geq 1$$

**Comments.** We want to improve the following aspects:

1. We don't always know  $\text{sign}(a)$
2. The error  $\|Z^n\|$  is first order in  $h$  and depends on  $\|v\|_{H^2}$

### ➤ Friedrichs method

Replace the value at  $U_j^n$  by the average of the neighbors

$$U_j^{n+1} - \frac{1}{2}(U_{j-1}^n + U_{j+1}^n) = \frac{a\lambda}{2}(U_{j+1}^n - U_{j-1}^n), \quad E_{kh} = \frac{1+a\lambda}{2}\delta_{-h}(x) + \frac{1-a\lambda}{2}\delta_h(x)$$

For stability condition, we have

$$|\hat{E}_{kh}(\xi)| = |\cos(h\xi) + ia\lambda \sin(h\xi)| \leq 1, \quad |a\lambda| \leq 1$$

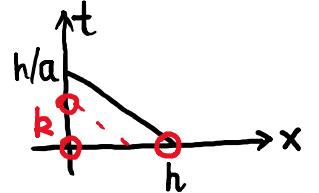
Now the sign of wave speed does not matter. However, this method is still first order in  $h$

### ➤ CFL condition (Courant-Friedrichs-Lowy)

The numerical stencil support should include the characteristics (or domain of dependence)

$$\lambda = \frac{k}{h} < \frac{1}{a}, \quad k < \frac{h}{a}$$

However, this is not sufficient for stability, only a **necessary condition**.



Similarly for heat equation, the explicit scheme requires

$$\lambda = \frac{k}{h^2} \leq \frac{1}{2}$$

This can be understood as the CFL condition in the limiting sense. For the implicit scheme, its domain of dependence includes every grid points on the previous time step, since it solves the linear system

### ➤ Lax-Wendroff method

$$\begin{aligned} u(x, t+k) &= u(x, t) + ku_t(x, t) + \frac{k^2}{2}u_{tt}(x, t) + O(k^3) \\ &= u(x, t) + kau_x(x, t) + \frac{k^2}{2}a^2u_{xx}(x, t) + O(k^3) \end{aligned}$$

Now we consider the following scheme

$$U^{n+1}(x) = U^n(x) + ak \cdot \frac{U^n(x+h) - U^n(x-h)}{2h} + \frac{k^2 a^2}{2} \cdot \frac{U^n(x+h) - 2U^n(x) + U^n(x-h)}{h^2}$$

We apply the central difference for spatial derivatives. We thus obtain

$$U^{n+1}(x) = (1 - a^2 \lambda^2)U^n(x) + \frac{a\lambda + a^2 \lambda^2}{2}U^n(x+h) + \frac{-a\lambda + a^2 \lambda^2}{2}U^n(x-h)$$

The operator norm becomes

$$|\hat{E}_{kh}(\xi)| = |1 - a^2 \lambda^2 + a^2 \lambda^2 \cos(h\xi) + ia\lambda \sin(h\xi)|$$

The stability condition again indicates

$$|a\lambda| \leq 1, \quad \frac{k}{h} \leq \frac{1}{|a|}$$

It can be shown that Lax-Wendroff method is second order in  $h$

## Week 7: Lecture 13.1. FDM for wave equations

- 1D multi-component system

$$\frac{\partial \mathbf{u}}{\partial t} = A \frac{\partial \mathbf{u}}{\partial x}, \quad A \text{ is symmetric}$$

When  $A$  is symmetric, this is a system of several advection equations. (If not, then this is the Cauchy-Riemann equation.)

For the Friedrichs and Lax-Wendroff schemes, we have

$$\begin{aligned}\mathbf{U}^{n+1}(x) &= \frac{1}{2}(I + A\lambda)\mathbf{U}^n(x+h) + \frac{1}{2}(I - A\lambda)\mathbf{U}^n(x-h) \\ \mathbf{U}^{n+1}(x) &= (I - A^2\lambda^2)\mathbf{U}^n(x) + \frac{1}{2}(A\lambda + A^2\lambda^2)\mathbf{U}^n(x+h) + \frac{1}{2}(-A\lambda + A^2\lambda^2)\mathbf{U}^n(x-h)\end{aligned}$$

### Stability condition

In the Fourier domain, now  $\hat{E}_{kh}(\xi)$  is a matrix, and we need

$$\hat{\mathbf{U}}^{n+1}(\xi) = \underbrace{\hat{E}_{kh}(\xi)}_{N \times N} \hat{\mathbf{U}}^n(\xi), \quad \|\hat{E}_{kh}(\xi)\|_{l_2 \rightarrow l_2} \leq 1$$

For Friedrichs method we have

$$\hat{E}_{kh}(\xi) = I \cos(h\xi) + i\lambda A \sin(h\xi)$$

Based on diagonalization, we have

$$I \cos(h\xi) + i\lambda A \sin(h\xi) = Q[I \cos(h\xi) + i\lambda A_D \sin(h\xi)]Q^T$$

So we can only require

$$\|\cos(h\xi) + i\lambda A_D \sin(h\xi)\|_{l_2 \rightarrow l_2} \leq 1, \quad \max_{\xi,j} |\cos(h\xi) + i\lambda a_j \sin(h\xi)| \leq 1$$

The stability condition is

$$\lambda = \frac{k}{h} \leq \frac{1}{\max_j |a_j|} = \frac{1}{\|A\|_{l_2 \rightarrow l_2}}$$

For Lax-Wendroff method, we have the same result.

- 1D wave equation

$$w_{tt} = w_{xx}, \quad \text{given } w(0, x) \text{ and } w_t(0, x)$$

We write it into a system and then apply either Friedrichs or Lax-Wendroff method.

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} aw_x \\ w_t \end{bmatrix}$$

Therefore, each component satisfies

$$\frac{\partial u_1}{\partial t} = aw_{xt} = a \frac{\partial u_2}{\partial x}, \quad \frac{\partial u_2}{\partial t} = w_{tt} = a^2 w_{xx} = a \frac{\partial u_1}{\partial x}$$

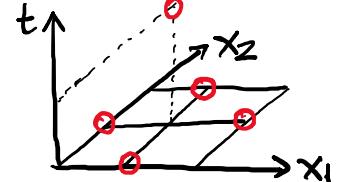
The system can be written as

$$\frac{\partial}{\partial t} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \cdot \frac{\partial}{\partial x} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad \frac{\partial \mathbf{u}}{\partial t} = A \frac{\partial \mathbf{u}}{\partial x}, \quad \mathbf{u}(0, x) = \begin{bmatrix} a \frac{\partial w(0, x)}{\partial x} \\ w_t(0, x) \end{bmatrix}$$

➤ Multi-dimension, multi-component system

Now consider  $\mathbf{u}(t, x_1, x_2, \dots, x_d)$

$$\frac{\partial \mathbf{u}}{\partial t} = \sum_{j=1}^d A_j \frac{\partial \mathbf{u}}{\partial x_j}, \quad A_j \text{ is symmetric}$$



Friedrichs method gives

$$\begin{aligned} \mathbf{U}^{n+1}(x) - \frac{1}{2d} \begin{bmatrix} \mathbf{U}^n(x - e_1 h) + \mathbf{U}^n(x + e_1 h) \\ \vdots \\ \mathbf{U}^n(x - e_d h) + \mathbf{U}^n(x + e_d h) \end{bmatrix} &= \sum_{j=1}^d A_j \frac{\mathbf{U}^n(x + e_j h) - \mathbf{U}^n(x - e_j h)}{2h} \\ \mathbf{U}^{n+1}(x) &= \frac{1}{2} \sum_{j=1}^d \left[ \left( \frac{I}{d} + \lambda A_j \right) \mathbf{U}^n(x + e_j h) + \left( \frac{I}{d} - \lambda A_j \right) \mathbf{U}^n(x - e_j h) \right] \end{aligned}$$

The Fourier transform in  $d$ -dimension is

$$\hat{E}_{kh}(\xi) = \sum_{j=1}^d \left[ \frac{I}{d} \cos(h\xi_j) + i\lambda A_j \sin(h\xi_j) \right]$$

### Stability condition

To ensure the operator is bounded, we can choose  $\lambda$  such that

$$\begin{aligned} \max_{\xi_j} \left\| \frac{I}{d} \cos(h\xi_j) + i\lambda A_j \sin(h\xi_j) \right\|_{l_2 \rightarrow l_2} &\leq \frac{1}{d} \quad \forall j = 1, 2, \dots, d \\ \lambda \leq \frac{1}{d \|A_j\|} \quad \forall j = 1, 2, \dots, d &\Rightarrow \lambda \leq \frac{1}{d \cdot \max \|A_j\|} \end{aligned}$$

This is due to that  $A_j$  does not necessarily commute with each other. Therefore, we make sure each one of them is bounded.

➤ 2D wave equation

$$w_{tt} = w_{xx} + w_{yy}$$

Similarly, we write the wave equation into a system

$$\mathbf{u} = (u_1, u_2, u_3)^T, \quad u_1 = w_x, \quad u_2 = w_y, \quad u_3 = w_t$$

Therefore, each component satisfies

$$\frac{\partial u_1}{\partial t} = w_{xt} = \frac{\partial u_3}{\partial x}, \quad \frac{\partial u_2}{\partial t} = w_{yt} = \frac{\partial u_3}{\partial y}, \quad \frac{\partial u_3}{\partial t} = w_{tt} = w_{xx} + w_{yy} = \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y}$$

The system can be written as

$$\frac{\partial}{\partial t} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \cdot \frac{\partial}{\partial x} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \cdot \frac{\partial}{\partial y} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \quad \frac{\partial \mathbf{u}}{\partial t} = A_1 \frac{\partial \mathbf{u}}{\partial x} + A_2 \frac{\partial \mathbf{u}}{\partial y}$$

## Week 7: Lecture 13.2. Conservation law

- Nonlinear wave equation (**Conservation Law**)

$$u_t + [f(u)]_x = 0$$

The function  $f(u)$  is convex or concave. Integrating over space gives

$$\int_a^b (u_t + [f(u)]_x) dx = 0, \quad \frac{d}{dt} \int_a^b u dx + f[u(b,t)] - f[u(a,t)] = 0$$

The physics interpretation is the conservation law: The rate of total mass change is equal to the net flux. In this simple model, **the flux is a function of local density**.

### Examples

1. Transport equation:  $f(u) = au$
2. Burgers' equation:  $f(u) = u^2/2$

$$u_t + \frac{\partial}{\partial x} \left( \frac{u^2}{2} \right) = 0, \quad u_t + uu_x = 0$$

3. Consider  $u \in [0,1]$  is the car density. One simple model for speed is  $v(u) = 1 - u$  and the flux is  $f(u) = u \cdot v(u) = u(1 - u)$ . The **traffic flow equation** is

$$u_t + [u(1 - u)]_x = 0$$



### Characteristics

The characteristic  $x(t)$  satisfies

$$\frac{dx(t)}{dt} = f'[u(x(t), t)]$$

Along  $x(t)$  we can show that  $u$  is conserved:

$$\frac{d}{dt} [u(x(t), t)] = u_t + u_x \frac{dx}{dt} = u_t + f'(u) \cdot u_x = u_t + [f(u)]_x = 0$$

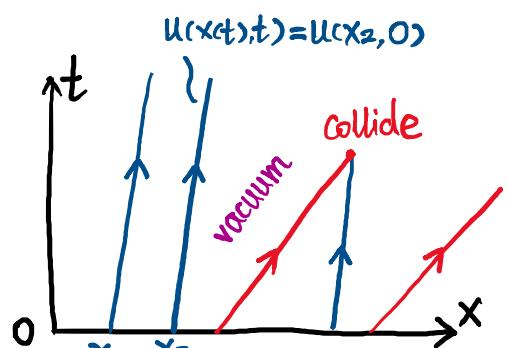
For Burgers' equation, the characteristics are straight lines:

$$\frac{dx(t)}{dt} = f'[u(x(t), t)] = \underbrace{u(x(t), t)}_{\text{conserved along } x(t)} \equiv \text{const.}$$

The slope will depend on initial conditions.

Two issues:

1. If two solutions (characteristics) **collide**, the above analysis fails
2. There is **vacuum** between two lines



## Shock formation

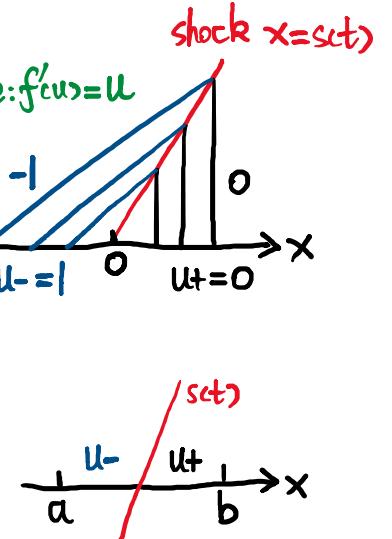
When two characteristics collide, a **shock** is formed. Assume that the initial condition is step-like with  $u_-$  and  $u_+$  (Riemann Problem)

Integral form of the conservation law is

$$0 = \frac{d}{dt} \int_a^b u(x, t) dx + f[u(b, t)] - f[u(a, t)]$$

The integral can be decomposed into

$$0 = \frac{d}{dt} \left[ \int_a^{s(t)} u(x, t) dx + \int_{s(t)}^b u(x, t) dx \right] + f[u(b, t)] - f[u(a, t)]$$



From the Leibniz rule we have

$$\begin{aligned} \frac{d}{dt} \int_a^{s(t)} u(x, t) dx &= \int_a^{s(t)} u_t dx + u_-(s(t), t) \cdot s'(t) \\ \frac{d}{dt} \int_{s(t)}^b u(x, t) dx &= \int_{s(t)}^b u_t dx - u_+(s(t), t) \cdot s'(t) \end{aligned}$$

Adding and subtracting terms  $f[u_-(s(t), t)]$  and  $f[u_+(s(t), t)]$ , we can use the conservation law on both sides

$$0 = \left\{ \int_a^{s(t)} u_t dx + f[u_-(s(t), t)] - f(u(a, t)) \right\} + \left\{ \int_{s(t)}^b u_t dx + f(u(b, t)) - f[u_+(s(t), t)] \right\} + f[u_+(s(t), t)] - f[u_-(s(t), t)] - u_+(s(t), t) \cdot s'(t) + u_-(s(t), t) \cdot s'(t)$$

This leads to the **Rankine-Hugoniot jump condition**:

$$s'(t) = \frac{f(u_+) - f(u_-)}{u_+ - u_-}$$

For Burger's equation, we have

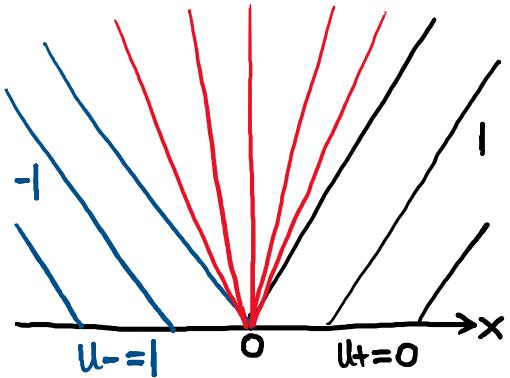
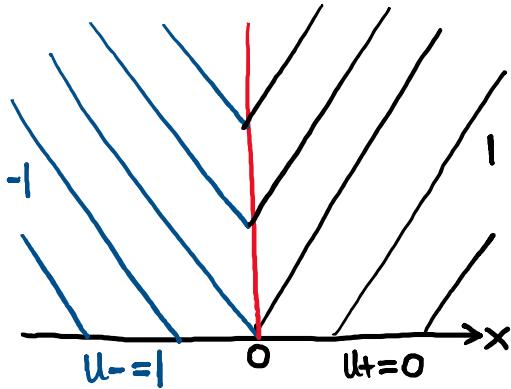
$$f(u) = \frac{u^2}{2}, \quad s'(t) = \frac{u_+ + u_-}{2}$$

## Rarefaction wave

For traffic flow equation, the characteristics are also straight lines with slope

$$f(u) = u(1-u), \quad \frac{dx(t)}{dt} = f'[u(x(t), t)] = 1 - 2u$$

For the same initial condition  $u_- = 1, u_+ = 0$ , now we have  $f'(u_-) < f'(u_+)$  and there is a vacuum. We have two ways to construct a weak solution.



The shock wave solution (**left**) becomes non-physical and entropy-violating. Another solution is the rarefaction wave (**right**), which is physical and fills the vacuum.

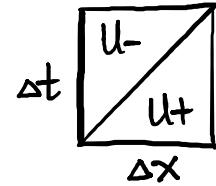
## Week 8: Lecture 14. Finite volume method for conservation law

➤ Review on conservation law

### Physics derivation of Rankine-Hugoniot jump condition

The weak form of conservation law for any  $[a, b]$  is

$$\frac{d}{dt} \int_a^b u(x, t) dx = f[u(a, t)] - f[u(b, t)]$$



Locally at a shock, we have

$$\frac{u_- \Delta x - u_+ \Delta x}{\Delta t} = f(u_-) - f(u_+), \quad s = \frac{\Delta x}{\Delta t} = \frac{f(u_-) - f(u_+)}{u_- - u_+} \equiv \frac{\llbracket f(u) \rrbracket}{\llbracket u \rrbracket}$$

### Shock & Rarefaction wave

Consider Burger's equation with initial condition  $u_- = -1$  and  $u_+ = 1$ . One weak solution is the artificial shock

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

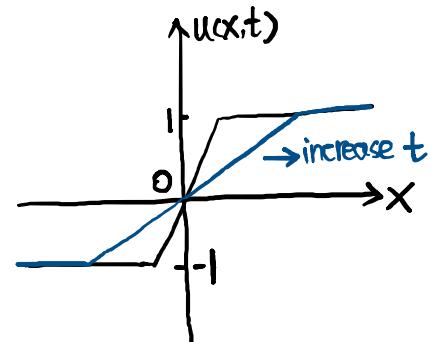
**Remark.** At a shock,  $u$  is a weak solution if and only if  $u$  satisfies the RH condition.

Another weak solution is the rarefaction wave

$$u(x, t) = \begin{cases} -1, & x \leq -t \\ x/t, & -t < x < t \\ 1, & x \geq t \end{cases}$$

This is a continuous solution, and we can directly check it by

$$\left( \frac{x}{t} \right)_t + \left( \frac{x^2}{2t^2} \right)_x = -\frac{x}{t^2} + \frac{2x}{2t^2} = 0$$



The derivative is not defined at the kinks, and thus we call it a weak solution

**Remark.** For the artificial shock, we can switch to the rarefaction wave at any later time  $t$ . However, rarefaction wave cannot go back to the artificial shock. Therefore, macroscopically we only observe the rarefaction wave, which is stable in terms of entropy.

**Principle.** The characteristic can only enter the shock and never leave from the shock.

### Artificial viscosity

$$u_t^\varepsilon + [f(u^\varepsilon)]_x = \varepsilon u_{xx}^\varepsilon$$

The viscosity solution  $u^\varepsilon$  is smooth and never has a shock. In the limit  $\varepsilon \rightarrow 0$ , viscosity solution  $u^\varepsilon$  converges to the rarefaction wave in  $L_1$  norm.

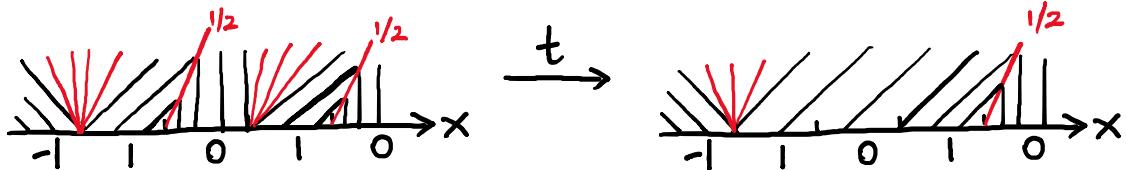
**Theorem.** Consider the two equations

$$u_t + [f(u)]_x = 0, \quad w_t + [f(w)]_x = 0$$

We have the  **$L_1$ -contractivity**

$$\|u(t, \cdot) - w(t, \cdot)\|_{L_1(\mathbb{R})} \leq \|u(0, \cdot) - w(0, \cdot)\|_{L_1(\mathbb{R})}$$

This indicates that all the details will be smeared out eventually at long enough  $t$ .



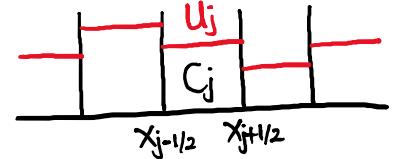
### ➤ Finite volume method (FVM) for conservation law

Start from the weak form of conservation law

$$\frac{d}{dt} \int_a^b u(x, t) dx = f[u(a, t)] - f[u(b, t)]$$

The domain is decomposed into cells, and the solution is approximated with piecewise constant function (cell averages)

$$\begin{aligned} \frac{d}{dt} (\Delta x \cdot U_j) &= f(u(x_{j-1/2}, t)) - f(u(x_{j+1/2}, t)) \\ \frac{1}{\Delta t} [\Delta x \cdot U_j^{n+1} - \Delta x \cdot U_j^n] &= f(u_{j-1/2}(t)) - f(u_{j+1/2}(t)) \end{aligned}$$



Therefore, we obtain

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{\Delta x} [f(u_{j-1/2}(t)) - f(u_{j+1/2}(t))]$$

However, we still need to approximate the fluxes at grid points.

$$f(u_{j-1/2}(t)) \approx \hat{f}(U_{j-1}^n, U_j^n), \quad f(u_{j+1/2}(t)) \approx \hat{f}(U_j^n, U_{j+1}^n)$$

Our numerical scheme now becomes

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{\Delta x} [\hat{f}(U_{j-1}^n, U_j^n) - \hat{f}(U_j^n, U_{j+1}^n)]$$

This is the **conservative schemes** or the **finite volume method**. Different schemes construct the **numerical flux  $\hat{f}$**  in different ways to evaluate the flux at the interface.

### ➤ Godunov scheme

- When  $\Delta t$  is sufficiently small, we can neglect the information from nearby cells

Because the wave speed is bounded by initial  $\max|f'(u_0)|$ , we have the **CFL condition**

$$\Delta t \leq \frac{\Delta x}{\max_{u(t=0)} |f'(u)|}$$

2. Now the local problem becomes the **Riemann's problem**

Consider  $f(u)$  is convex, i.e. the derivative  $f'(u)$  is monotone. We have in total two categories (rarefaction wave & shock) and five cases.

Rarefaction wave: $u_L < u_R \Rightarrow f'(u_L) < f'(u_R)$		
$f'(u_L) < f'(u_R) < 0$	$0 < f'(u_L) < f'(u_R)$	$f'(u_L) < 0 < f'(u_R)$
$\hat{f} = f(u_R)$	$\hat{f} = f(u_L)$	$\hat{f} = f(u_m)$
Summary: $\hat{f} \equiv \min_{u \in [u_L, u_R]} f(u)$		

Shock: $u_L > u_R \Rightarrow f'(u_L) > f'(u_R)$	
$f(u_L) < f(u_R)$	$f(u_L) > f(u_R)$
$s = \ f(u)\ /\ u\  < 0$	$s = \ f(u)\ /\ u\  > 0$
$\hat{f} = f(u_R)$	$\hat{f} = f(u_L)$
Summary: $\hat{f} \equiv \max_{u \in [u_R, u_L]} f(u)$	

Therefore, solution of the Riemann's problem gives

$$\hat{f} = \begin{cases} \min_{u \in [u_L, u_R]} f(u), & u_L < u_R \\ \max_{u \in [u_R, u_L]} f(u), & u_L > u_R \end{cases}$$

3. Godunov scheme

$$U_j^{n+1} - U_j^n = \frac{\Delta t}{\Delta x} (\hat{f}_{j-1/2}^n - \hat{f}_{j+1/2}^n)$$

➤ Lax-Friedrichs scheme

$$\hat{f}_{j+1/2} = \frac{1}{2} [f(U_j) + f(U_{j+1}) - \alpha(U_{j+1} - U_j)], \quad \alpha = \max_u |f'(u)|$$

On the other hand, the localized version chooses  $\alpha$  by

$$\alpha = \max_{u \in (U_j, U_{j+1})} |f'(u)|$$

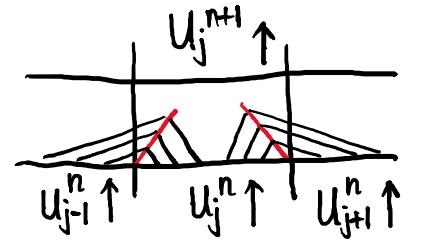
## Week 8: Lecture 15.1. Analysis of finite volume method

➤ Analysis of monotone schemes

$$U_{j+1}^n = U_j^n + \frac{\Delta t}{\Delta x} [\hat{f}(U_{j-1}^n, U_j^n) - \hat{f}(U_j^n, U_{j+1}^n)]$$

We can consider this scheme as a nonlinear operator (stencil)

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



The entropy solution indicates that

$$U_{j-1}^n \uparrow \Rightarrow U_j^{n+1} \uparrow, \quad U_{j+1}^n \uparrow \Rightarrow U_j^{n+1} \uparrow, \quad U_j^n \uparrow \Rightarrow U_j^{n+1} \uparrow$$

**Definition.** The scheme is called monotone if  $G(\uparrow, \uparrow, \uparrow)$ . This is also the stability condition.

Monotonicity requires the numerical flux to satisfy all the conditions below

$$\frac{\partial G}{\partial U_{j-1}^n} = \frac{\Delta t}{\Delta x} \hat{f}_1(U_{j-1}^n, U_j^n) \geq 0, \quad \frac{\partial G}{\partial U_{j+1}^n} = -\frac{\Delta t}{\Delta x} \hat{f}_2(U_j^n, U_{j+1}^n) \geq 0$$

$$\frac{\partial G}{\partial U_j^n} = 1 + \frac{\Delta t}{\Delta x} [\hat{f}_2(U_{j-1}^n, U_j^n) - \hat{f}_1(U_j^n, U_{j+1}^n)] \geq 0$$

Therefore, for any  $u, v, p \in \mathbb{R}$ , we require

$$\hat{f}_1(u, v) \geq 0, \quad \hat{f}_2(u, v) \leq 0, \quad 1 + \frac{\Delta t}{\Delta x} [\hat{f}_2(p, u) - \hat{f}_1(u, q)] \geq 0$$

### Lax-Friedrichs scheme

$$\hat{f}(p, q) = \frac{1}{2} [f(p) + f(q) - \alpha(q - p)], \quad \alpha = \max_{u \text{ init.}} |f'(u)|$$

The first two conditions are automatically satisfied based on the chosen  $\alpha$

$$\hat{f}_1 \equiv \partial_1 \hat{f} = \frac{1}{2} [f'(p) + \alpha] \geq 0, \quad \hat{f}_2 \equiv \partial_2 \hat{f} = \frac{1}{2} [f'(q) - \alpha] \leq 0$$

The final condition gives the constraint on time step

$$1 + \frac{\Delta t}{\Delta x} [\hat{f}_2(p, u) - \hat{f}_1(u, q)] = 1 - \alpha \frac{\Delta t}{\Delta x} \geq 0, \quad \Delta t \leq \frac{\Delta x}{\alpha} = \frac{\Delta x}{\max_{u \text{ init.}} |f'(u)|}$$

Now we prove that Lax-Friedrichs scheme is monotone

### Convergence of monotone scheme

Now suppose the numerical solution and we define its discrete  $L_1$ -norm as

$$\|U\|_{L_1} = \Delta x \sum_i |U_i|$$

**Theorem.** If  $G$  is monotone, then consider two numerical schemes

$$U_j^{n+1} \equiv G(U_{j-1}^n, U_j^n, U_{j+1}^n), \quad V_j^{n+1} \equiv G(V_{j-1}^n, V_j^n, V_{j+1}^n)$$

We similarly have the discrete  $L_1$ -contractivity

$$\|U^n - V^n\|_{L_1} \leq \|U^0 - V^0\|_{L_1}$$

**Theorem.** [Crandall-Majda] If  $G$  is monotone (stable), then  $\{U_j^n\}$  converges to the entropic solution as  $\Delta t, \Delta x \rightarrow 0$ . The entropic solution is the limit of viscosity solution with  $\varepsilon \rightarrow 0$ .

**Theorem.** Any monotone scheme has only  $O(h)$  accuracy.

**Proof scheme.** The local error is  $O(h^2)$  and thus the total error is  $O(h^2) \cdot T/h \rightarrow O(h)$

## Week 8: Lecture 15.2. Monte Carlo method

Computational methods for stochastic system and for probability-related problems

- Example problems

### Numerical integration

$$\int_a^b f(x) dx = (b - a) \int_a^b \frac{f(x)}{b - a} dx$$

With the uniform distribution  $p(x)$  over interval  $[a, b]$ , we have

$$\int_a^b f(x) dx = (b - a) \int_a^b f(x)p(x) dx$$

Now if we can sample  $X \sim p(x)dx$ , the integral becomes the expectation

$$\int_a^b f(x) dx = (b - a) \cdot \mathbb{E}_{X \sim p(x)} f(X)$$

We can thus calculate the integral by sampling  $X_1, X_2, \dots, X_N \sim p(x)$

$$I(f) \approx \hat{I}_N(f) = \left[ \frac{1}{N} \sum_{i=1}^N f(X_i) \right] (b - a)$$

According to Law of Large Number (Central Limit), this approximates the expectation

### Estimation of $\pi$

We can estimate  $\pi$  by randomly sample points within a square with area 4, and count how many samples fall into the circle with radius 1.

### Optimization and sampling

Given the energy function  $E(x)$ , the Boltzmann distribution is ( $\beta$  is inverse temperature)

$$p(x) = \frac{1}{Z} e^{-\beta E(x)}, \quad Z \equiv \int_0^{+\infty} e^{-\beta E(y)} dy, \quad \beta = \frac{1}{T}$$

To sample from this distribution  $p(x)$ , the usual method is Metropolis-Hastings, which is an example of Markov Chain Monte Carlo (MCMC) method.

Sampling problems can be naturally linked to optimization problems, as the densely sampled region corresponds to where the energy is low

## ➤ Monte Carlo method (MC)

Now assume  $[a, b] = [0, 1]$ . The expectation of the numerical integral is

$$\mathbb{E}[\hat{I}_N(f)] = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N f(X_i)\right] = \frac{1}{N} \sum_{i=1}^N E[f(X_i)] = \underbrace{E[f(X)]}_{\text{i.i.d.}} = I(f)$$

This shows that our estimation is unbiased. However, we also need to study the variance.

$$\text{var}(Z) = \mathbb{E}[(Z - \mathbb{E}(Z))^2] = \mathbb{E}(Z^2) - [\mathbb{E}(Z)]^2$$

For our numerical integral, we have

$$\begin{aligned} \text{var}[\hat{I}_N(f)] &= \mathbb{E}[\hat{I}_N(f) - I(f)]^2 = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^N (f(X_i) - I(f))\right]^2 \\ &= \frac{1}{N^2} \cdot \mathbb{E} \sum_{i,j=1}^N [f(X_i) - I(f)][f(X_j) - I(f)] \\ &= \frac{1}{N^2} \cdot N \cdot \mathbb{E}[f(X) - I(f)]^2 = \frac{1}{N} \text{var}[f(x)] \end{aligned}$$

Therefore, the convergence rate is

$$|\hat{I}_N(f) - I(f)| \propto \frac{1}{\sqrt{N}}$$

**Remark.** This method works for any dimension as long as  $\text{var}[f(x)]$  is not large in terms of dimension. With  $N$  samples, the convergence rate  $1/\sqrt{N}$  is independent of dimension. This is the foundation of success for machine learning techniques.

However, for Riemann sum when there are  $N$  cubes in  $d$ -dimension, we have  $N^{1/d}$  samples for each dimension. The quadrature error scales as

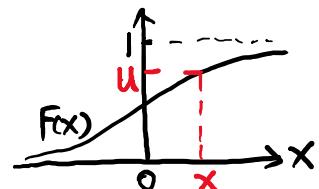
$$\text{Integration error} \propto \left(\frac{1}{N^{1/d}}\right)^2 = \frac{1}{N^{2/d}}$$

When dimension  $d \gg 1$ , the Monte-Carlo method is always better.

## ➤ Sampling from 1-D distribution

We first construct the CDF  $F(x)$ . The sampling procedure is

1. Sample  $U$  uniformly in  $[0, 1]$
2. Calculate  $X = F^{-1}(U)$
3. Return  $X \sim p(x)$



For efficient sampling of Gaussian distribution, we consider a pair  $x_1, x_2 \in N(0,1)$

$$p(x_1)p(x_2) = \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}} dx_1 dx_2 = \frac{1}{2\pi} e^{-\frac{r^2}{2}} r dr d\theta = e^{-\frac{r^2}{2}} d\left(\frac{r^2}{2}\right) \cdot \frac{1}{2\pi} d\theta$$

Therefore, the sampling procedure for a pair of Gaussian is

1.  $X \sim [0, 1], G = -\ln(1 - X)$
2.  $R = \sqrt{2G}$
3.  $\theta \sim [0, 2\pi]$
4.  $X_1 = R \cos \theta, X_2 = R \sin \theta$

➤ Analysis of Monte Carlo method

$$\hat{I}_N(f) \equiv \frac{1}{N} \sum_{i=1}^N f(X_i) \approx \int_0^1 f(x) dx$$

The variance of our estimation is

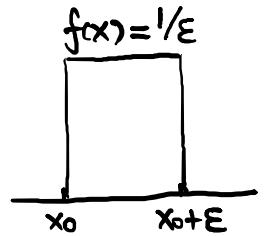
$$\text{var}[\hat{I}_N(f)] = \frac{1}{N} \text{var}[f(x)], \quad \text{var}[f(x)] = \mathbb{E}[f^2(x)] - (\mathbb{E}[f(x)])^2$$

We hope the variance of  $f(X)$  is small. However, consider the following  $f(x)$

$$f(x) = \frac{1}{\varepsilon}, \quad x \in [x_0, x_0 + \varepsilon]$$

The expectation and variance of  $f(x)$  are

$$\mathbb{E}[f(X)] = 1, \quad \mathbb{E}[f^2(X)] = \frac{1}{\varepsilon}, \quad \text{var}[f(x)] = \frac{1}{\varepsilon} - 1 \rightarrow +\infty$$



For a narrow function (which is often the case in high dimension), the variance of estimation can be large and requires to be improved for real applications.

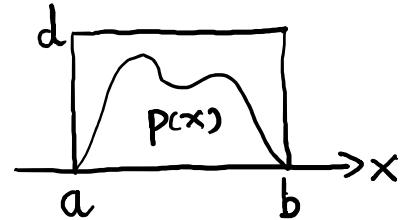
## Week 9: Lecture 16. Rejection sampling & Markov chain Monte Carlo

### ➤ Acceptance-Rejection sampling

If the CDF or its inverse is very difficult to compute, we use alternative methods. Consider our PDF  $p(x)$  is supported in  $[a, b]$  with  $0 \leq p(x) \leq d$ . The idea is to sample uniformly in the box  $[a, b] \times [0, d]$ . If the sample is under the curve, accept it. Otherwise, repeat.

The algorithm is written as:

1.  $X \in \text{Unif}[a, b], Y \in \text{Unif}[0, d]$
2. If  $Y < P(X)$ , accept and go to 4
3. Otherwise, reject and go to 1
4. Return  $X \sim p(x)$



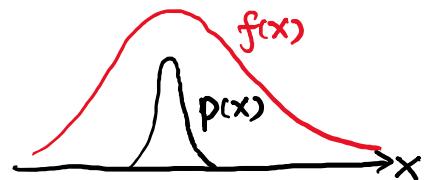
### Bad examples:

- Narrow distribution: Rejection rate is too high
- Distribution defined on  $\mathbb{R}$ : Not possible for an infinite support

### Modified rejection sampling

Require  $f(x) = Cg(x)$  with the distribution  $g(x)$  easy to sample.

1. Sample  $X \sim g(x)$  and sample  $Y$  uniformly between  $[0, f(X)]$ . This allows us to sample uniformly under the curve  $f(x)$
2. If  $Y \leq p(X)$ , accept and go to 4
3. Otherwise, reject and go to 1
4. Return  $X \sim p(x)$



**Remark.** We need  $g(x)$  which is easy to sample, and  $f(x)$  does not waste too much

### ➤ Importance sampling for integration

$$I(f) \approx \hat{I}_N(f) = \frac{1}{N} \sum_{i=1}^N f(X_i), \quad \mathbb{E}[\hat{I}_N(f)] = I(f), \quad \text{var}[\hat{I}_N(f)] = \frac{1}{N} \text{var}[f(x)]$$

For a narrow function, the variance is very large. We can use a normalization as follows

$$I(f) = \int_{\mathbb{R}} f(x) dx = \int_{\mathbb{R}} \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_{X \sim g(x)} \left[ \frac{f(X)}{g(X)} \right]$$

We want  $g(x)$  to look like  $f(x)$  up to a scaling factor. Now we sample  $X_1, X_2, \dots, X_N \sim g(x)$

$$\hat{I}_N(f) = \frac{1}{N} \sum_{i=1}^N \frac{f(X_i)}{g(X_i)}$$

The variance of  $f/g$  is calculated as

$$\text{var}_{X \sim g(x)} \left[ \frac{f(X)}{g(X)} \right] = \mathbb{E} \left[ \left( \frac{f(X)}{g(X)} \right)^2 \right] - \left( \mathbb{E} \left[ \frac{f(X)}{g(X)} \right] \right)^2 = C^2 \int_{\mathbb{R}} \frac{\hat{f}^2(x)}{g(x)} dx - [I(f)]^2$$

We now obtain an optimization problem (with  $\hat{f}$  the normalized version of  $f$ )

$$\min_g \int_{\mathbb{R}} \frac{\hat{f}^2(x)}{g(x)} dx, \quad \int_{\mathbb{R}} g(x) dx = 1, \quad \int_{\mathbb{R}} \hat{f}(x) dx = 1$$

According to Cauchy-Schwartz inequality

$$\int_{\mathbb{R}} \frac{\hat{f}^2(x)}{g(x)} dx \cdot \int_{\mathbb{R}} g(x) dx \geq \left( \int_{\mathbb{R}} \frac{\hat{f}(x)}{\sqrt{g(x)}} \sqrt{g(x)} dx \right)^2 = 1$$

The minimum is obtained when  $g(x) = \hat{f}(x)$

**Remark.** We need to choose  $g(x)$  to be close to  $f(x)$ , up to a normalization factor

➤ Variance reduction techniques for Monte Carlo method

### Control variates

If  $f(x)$  is hard to work with, we choose  $h(x)$  that is close to  $f(x)$  and is easy to compute. For a given distribution  $\pi$ , we can calculate the expectation as

$$\mathbb{E}_{x \sim \pi}(f) = \mathbb{E}_{x \sim \pi}(f - h) + \mathbb{E}_{x \sim \pi}(h)$$

The first part can be applied with MC method, since the variance now becomes much smaller

$$\text{var}_{x \sim \pi}(f - h) \ll \text{var}_{x \sim \pi}(f)$$

### Antithetic variates

If we know  $f(x)$  is (approximately) symmetric, we can sample symmetrically. For example, after sampling  $X_1, \dots, X_N$ , include  $-X_1, \dots, -X_N$ .

➤ **Metropolis algorithm**

We usually want to sample  $\pi(x)$  based on some energy function  $H(x)$

$$\pi(x) = \frac{1}{Z} e^{-\beta H(x)}$$

However, we have no access to the renormalization constant  $Z$ , i.e., we don't know how "tall" our target distribution is.

### Markov chain Monte Carlo (MCMC)

For a state space  $S$ , the transition matrix  $P_{ij}$  with  $i, j \in S$  denotes the probability of going to  $j$  if currently at state  $i$

$$P_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$$

The complicated path is hard to describe, but we can study the equilibrium distribution  $\pi_i$

$$P_i(T) = \mathbb{P}(X_T = i) \rightarrow \pi_i = \frac{1}{Z} e^{-\beta H_i}$$

MCMC designs  $P_{ij}$  such that the limit distribution is  $\pi_i$  (up to a constant factor  $Z$ )

The probability flux is written as

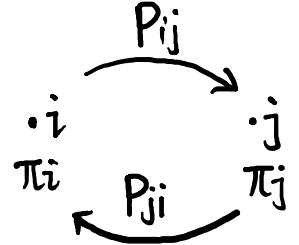
$$F_{ij} = \pi_i P_{ij}, \quad F_{ji} = \pi_j P_{ji}$$

One way to make  $\pi_i$  stationary is **detailed balance**

$$F_{ij} = \pi_i P_{ij} = \pi_j P_{ji} = F_{ji}$$

Therefore, the **renormalization constant does not matter** anymore

$$\frac{1}{Z} e^{-\beta H_i} \cdot P_{ij} = \frac{1}{Z} e^{-\beta H_j} \cdot P_{ji}, \quad \frac{P_{ij}}{P_{ji}} = e^{-\beta(H_j - H_i)}$$



### Metropolis-Hastings algorithm

Now we design the transition matrix  $P_{ij}$ . There are two conditions to be satisfied

$$\sum_k P_{ik} = 1, \quad \frac{P_{ij}}{P_{ji}} = e^{-\beta(H_j - H_i)}$$

We also want the implementation of  $P_{ij}$  to be easy. Define a simple matrix  $Q_{ij}$ , which is the proposal distribution. An example is the symmetric one

$$Q_{ij} \approx \frac{1}{N}, \quad N = \text{degree of freedom}$$

With the acceptance rate  $0 \leq A_{ij} \leq 1$ , we design the following transition matrix  $P_{ij}$

$$P_{ij} = \begin{cases} Q_{ij} A_{ij}, & i \neq j \\ 1 - \sum_{j \neq i} P_{ij}, & i = j \end{cases}$$

To satisfy the second condition (detailed balance, equilibrium  $\pi_i$ ), we require

$$\frac{A_{ij}}{A_{ji}} = e^{-\beta(H_j - H_i)} \cdot \frac{Q_{ji}}{Q_{ij}}$$

Now consider  $Q_{ij} = Q_{ji}$ . For the Metropolis strategy, the acceptance rate is defined as

$$A_{ij} = \min\left\{1, e^{-\beta(H_j - H_i)}\right\}$$

As an example, we have

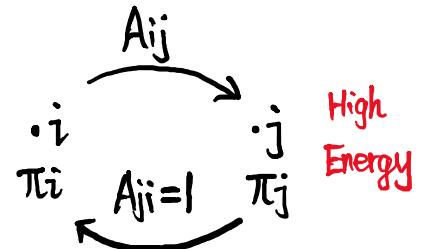
$$A_{ij} = e^{-\beta(H_j - H_i)}, \quad A_{ji} = 1, \quad \text{when } H_j > H_i$$

$$A_{ij} = 1, \quad A_{ji} = e^{-\beta(H_i - H_j)}, \quad \text{when } H_j < H_i$$

In this case, we have the detailed balance

$$\pi_i A_{ij} = \left( \frac{1}{Z} e^{-\beta H_i} \right) \cdot e^{-\beta(H_j - H_i)} = \frac{1}{Z} e^{-\beta H_j} = \pi_j$$

$H_j > H_i$	State $i$	State $j$
Energy	Low	High
Probability	High	Low



If  $Q_{ij} \neq Q_{ji}$ , then we need to adjust  $A_{ij}$  slightly. For the Metropolis strategy, we have

$$A_{ij} = \min \left\{ 1, \frac{Q_{ji} e^{-\beta H_j}}{Q_{ij} e^{-\beta H_i}} \right\}$$

This choice of  $A_{ij}$  ensures detailed balance, and thus the Metropolis algorithm converges to the equilibrium distribution

### Glauber dynamics

The acceptance rate is defined similar to a sigmoid function

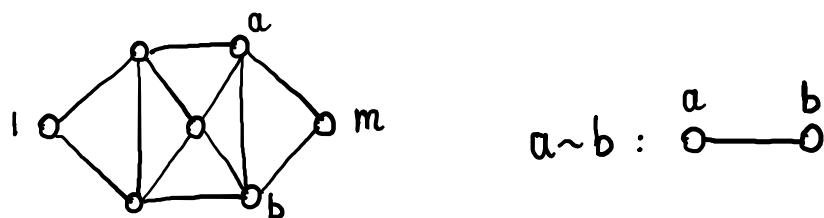
$$A_{ij} = \frac{1}{1 + e^{\beta(H_j - H_i)}} < 1$$

The detailed balance is also satisfied

$$\pi_i A_{ij} = \frac{1}{Z} e^{-\beta H_i} \cdot \frac{1}{1 + e^{\beta(H_j - H_i)}} = \frac{1}{Z} e^{-\beta H_j} \cdot \frac{1}{1 + e^{\beta(H_i - H_j)}} = \pi_j A_{ji}$$

➤ Example application of Metropolis-Hastings algorithm

Consider the following graph  $G$  of electron spins.



The state space is defined as  $S = \{(x_1, x_2, \dots, x_m), x_i = \pm 1\}$ , which is the set of all possible binary strings of length  $m$ . Then we have  $|S| = 2^m$ , which is a very large space.

The Hamiltonian is described as (with  $a \sim b$  denotes an edge in the graph)

$$H(\mathbf{x}) = - \sum_{a \sim b} x_a x_b$$

The goal is to sample  $\mathbf{x} \in S$  with probability

$$p(\mathbf{x}) = \frac{1}{Z} e^{-\beta H(\mathbf{x})}$$

Metropolis-Hastings algorithm uses the following acceptance rate

$$A_{\mathbf{xy}} = \min \left\{ 1, \frac{Q_{yx} e^{-\beta H(y)}}{Q_{xy} e^{-\beta H(x)}} \right\}$$

To make it efficient, the proposal distribution  $Q_{xy}$  is selected as symmetric

$$Q_{xy} = \begin{cases} m^{-1} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ differ by 1 slot} \\ 0 & \text{otherwise} \end{cases}, \quad Q_{xy} = Q_{yx}$$

Therefore, we have

$$A_{\mathbf{xy}} = \min \{1, e^{-\beta [H(\mathbf{y}) - H(\mathbf{x})]}\}$$

Suppose that  $\mathbf{x}$  and  $\mathbf{y}$  differ at slot  $p$

$$\begin{aligned} H(\mathbf{y}) &= \sum_{\substack{a \sim b \\ a, b \neq p}} y_a y_b + \sum_{\substack{a \sim b \\ a \text{ or } b = p}} y_a y_b \\ H(\mathbf{x}) &= \sum_{\substack{a \sim b \\ a, b \neq p}} x_a x_b + \sum_{\substack{a \sim b \\ a \text{ or } b = p}} x_a x_b \end{aligned}$$

The first term in the above expressions is the same. Since we only care about the energy difference, we can just calculate

$$H(\mathbf{y}) - H(\mathbf{x}) = \sum_{\substack{a \sim b \\ a \text{ or } b = p}} (y_a y_b - x_a x_b)$$

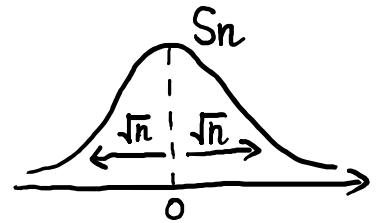
This is much more efficient than evaluating  $H(\mathbf{x})$  and  $H(\mathbf{y})$  separately, which contain many pairs in the summation.

## Week 10: Lecture 17. Introduction to stochastic differential equations

- Wiener process (Brownian motion, BM)

Consider the random walk  $\xi_i = \{1, -1\}$  with equal probability

$$S_n = \sum_{k=1}^n \xi_k \sim N(0, \sigma^2)$$



The variance is given as

$$\sigma^2 = \text{var}(S_n) = \mathbb{E}\left[\left(\sum_{k=1}^n \xi_k\right)\left(\sum_{l=1}^n \xi_l\right)\right] = \mathbb{E}\left[\sum_{k=1}^n \xi_k^2\right] = n \cdot \mathbb{E}[\xi_i^2] = n$$

Now consider the scaled version  $\tilde{\xi}_i = \{\sqrt{\Delta t}, -\sqrt{\Delta t}\}$  with time step  $\Delta t$ .

Within time  $T$ , there is a total of  $T/\Delta t$  steps

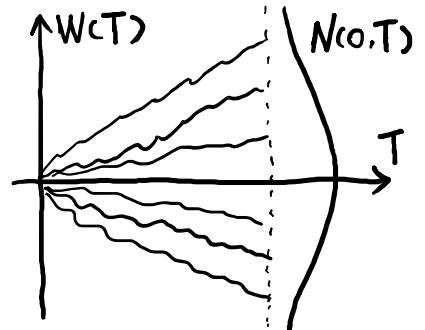
$$W_T = \sum_{k=1}^{T/\Delta t} \tilde{\xi}_k, \quad \text{var}(W_T) = \frac{T}{\Delta t} \cdot \mathbb{E}[\tilde{\xi}_i^2] = T$$

Therefore, we obtain a random variable

$$W_T \sim N(0, \sigma^2 = T)$$

Another prospective is to consider different trajectories. The BM can be viewed as a “path-wise distribution” such that

$$\mathbb{E}[W_t] = 0, \quad \mathbb{E}[W_t^2] = t$$



The local increment is

$$\mathbb{E}[dW_t] = 0, \quad E[(dW_t)^2] = dt$$

A typical path of BM is fractal, and we should be careful when taking derivatives

- Stochastic differential equations (SDEs)

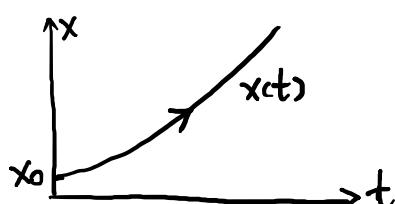
Two perspectives on SDEs:

SDE is a map from BM paths to another set of paths

SDE is a change of probability measure over the paths

**ODE: Deterministic**

$$dx = v(x)dt, \quad x(0) = x_0$$



**SDE: System with noise using BM**

$$dx = v(x)dt + dW, \quad x(0) = x_0$$



## Ito's formula

For usual derivative, the chain rule is

$$df(x) = f'(x)dx = f'(x)v(x)dt$$

However, for SDEs we have

$$(dx)^2 = [v(x)dt + dW]^2 = v^2(x)(dx)^2 + 2v(x)dtdW + (dW)^2$$

The leading term is  $(dW)^2 = dt$ . Therefore, we need to analyze the second order term in the Taylor's expansion

$$df(x) = f'(x)dx + \frac{1}{2}f''(x)(dx)^2 = f'(x)v(x)dt + f'(x)dW + \frac{1}{2}f''(x)dt$$

**Example:** Consider

$$x = W, \quad dx = dW, \quad f(x) = x^2$$

Newton's calculus gives

$$df(x) = f'(x)dx = 2xdx = 2WdW$$

Ito's calculus gives

$$df(x) = f'(x)dx + \frac{1}{2}f''(x)(dx)^2 = 2xdx + (dx)^2 = 2WdW + dt$$

**Example:** Consider

$$dx = -xdt, \quad x(0) = 1 \implies x(t) = e^{-t}$$

For SDE, we have

$$dx = -xdt + dW, \quad e^t dx + e^t xdt = e^t dW, \quad d(e^t x) = e^t dW$$

Integrating both sides gives the analytic solution

$$e^t x(t) - x(0) = \int_0^t e^s dW(s), \quad x(t) = e^{-t} + \int_0^t e^{-(t-s)} dW(s)$$

There are very few SDEs that can be analytically solved. In most cases, we need to solve the SDEs using numerical methods.

## ➤ Euler-Maruyama method for SDE

Discretize time into small intervals of  $\Delta t$ . The approximate solution  $X_n$  is given as

$$X_n \approx X(n\Delta t), \quad X_{n+1} = X_n + \Delta t \cdot v(X_n) + W_{(n+1)\Delta t} - W_{n\Delta t}$$

The BM part follows the normal distribution

$$Z_n = W_{(n+1)\Delta t} - W_{n\Delta t}, \quad Z_n \sim N(0, \Delta t)$$

Therefore, the Euler-Maruyama (E-M) scheme gives

$$X_{n+1} = X_n + \Delta t \cdot v(X_n) + Z_n, \quad Z_n \sim N(0, \Delta t)$$

## Accuracy analysis

Order  $\alpha$  strong error:

$$\mathbb{E}[|X_n - X(n\Delta t)|] \lesssim \Delta t^\alpha$$

Order  $\beta$  weak error (based on a smooth function):

$$\mathbb{E}[|f(X_n) - f(X(n\Delta t))|] \lesssim C_f \Delta t^\beta$$

E-M method has the following accuracy orders

$$\alpha = \frac{1}{2}, \quad \beta = 1$$

## Week 10: Lecture 18. Introduction to wavelets

### ➤ Wavelets

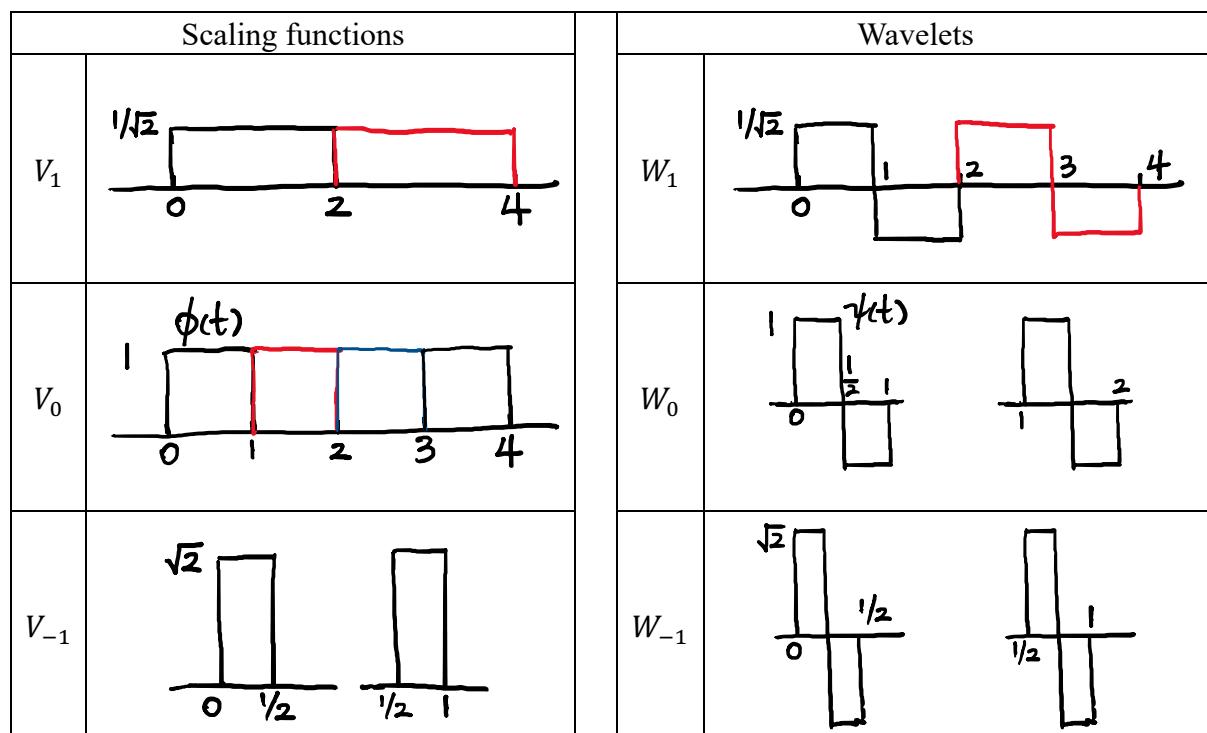
Consider functions in  $L^2(\mathbb{R})$ , the pseudo-basis can be selected as

1. Delta functions: Spikes, good for signals sparse in time domain
2. Fourier basis: Plane waves, good for signals sparse in frequency domain

However, they are not  $L^2(\mathbb{R})$  functions, so they are called pseudo-basis.

We want to construct wavelets, which are good for signals sparse in both time and frequency domains. We also require the wavelets to be **self-similar**, as it is for spikes and plane waves

### ➤ Haar wavelets



Each subspace is the span of the functions. The subspaces of scaling functions have the following properties

$$V_{j+1} \subseteq V_j, \quad \lim_{j \rightarrow +\infty} V_j = \{0\}, \quad \lim_{j \rightarrow -\infty} V_j = L^2(\mathbb{R})$$

The wavelet subspaces  $W_j$  satisfy

$$V_j \perp W_j, \quad V_{j-1} = V_j \oplus W_j$$

Recursively doing the direct sum, we have

$$V_0 = W_1 \oplus W_2 \oplus \dots$$

The Haar wavelet is **discrete self-similar**. It has both time and frequency localization

**Theorem.** The function space  $L^2(\mathbb{R})$  is the direct sum of all subspaces  $W_j$

$$L^2(\mathbb{R}) = \cdots \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus \cdots$$

➤ Generalization of wavelets

The Haar wavelet is not smooth, which leads to bad approximation property. To construct smoother wavelet  $\psi(t)$ , we follow the procedure below.

**Choose scaling function  $\phi(t)$**

We want  $\phi(t)$  and its translation to form an orthonormal basis  $V_0$ . Given a function  $\theta(t)$ , we hope to obtain an orthonormal copy of it.

We claim that  $\phi(t)$  can be represented by a linear combination of  $\theta(t - n)$

$$\phi(t) = \sum_{n \in \mathbb{Z}} a_n \theta(t - n) \Leftrightarrow \hat{\phi}(\omega) = \hat{a}(\omega) \hat{\theta}(\omega)$$

This corresponds to a convolution. Because coefficients  $a_n$  are discrete,  $\hat{a}(\omega)$  is **2π periodic**.

The orthonormality of  $\{\phi(t - k)\}_{k \in \mathbb{Z}}$  requires

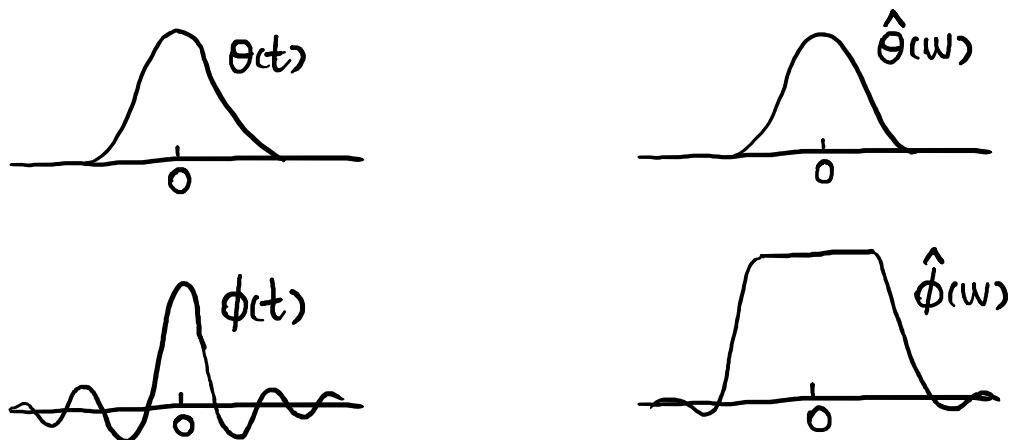
$$\langle \phi(t), \phi(t - k) \rangle = \int_{-\infty}^{+\infty} \phi(t) \phi^*(t - k) dt = \delta_k$$

In the Fourier domain, this condition becomes

$$\sum_{k \in \mathbb{Z}} |\hat{\phi}(\omega + 2\pi k)|^2 = 1$$

Now we construct  $\hat{\phi}(\omega)$  as the following to satisfy the above condition

$$\hat{\phi}(\omega) = \frac{\hat{\theta}(\omega)}{\left( \sum_{k \in \mathbb{Z}} |\hat{\theta}(\omega + 2\pi k)|^2 \right)^{1/2}}$$



## Scaling equation

The multiresolution causality  $V_{j+1} \subseteq V_j$  requires

$$\frac{1}{\sqrt{2}}\phi\left(\frac{t}{2}\right) = \sum_{n \in \mathbb{Z}} h_n \phi(t - n)$$

Again taking the Fourier transform and using recursion gives

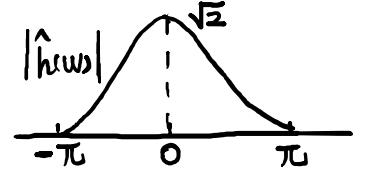
$$\hat{\phi}(2\omega) = \frac{\hat{h}(\omega)}{\sqrt{2}} \hat{\phi}(\omega), \quad \hat{\phi}(\omega) = \left[ \prod_{p=1}^{\infty} \frac{\hat{h}\left(\frac{\omega}{2^p}\right)}{\sqrt{2}} \right] \hat{\phi}(0)$$

This states that any scaling function  $\phi(t)$  is specified by a discrete filter  $h_n$ . Note that  $\hat{h}(\omega)$  is also  $2\pi$  periodic.

**Theorem.** The orthonormality of  $\{\phi(t - k)\}_{k \in \mathbb{Z}}$  is equivalent to

$$|\hat{h}(\omega)|^2 + |\hat{h}(\omega + \pi)|^2 = 2, \quad \hat{h}(0) = \sqrt{2}$$

**Comment.**  $h_n$  having finite support is a non-trivial property.



## Construct wavelets $\psi(t)$

The multiresolution causality  $W_{j+1} \subseteq V_j$  requires

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_{n \in \mathbb{Z}} g_n \phi(t - n), \quad \hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}\left(\frac{\omega}{2}\right)\hat{\phi}\left(\frac{\omega}{2}\right)$$

The orthonormality of  $\{\psi(t - k)\}_{k \in \mathbb{Z}}$  requires

$$\sum_{k=-\infty}^{+\infty} |\hat{\psi}(\omega + 2\pi k)|^2 = 1 \iff |\hat{g}(\omega)|^2 + |\hat{g}(\omega + \pi)|^2 = 2$$

The orthogonality condition  $V_j \perp W_j$  requires

$$\sum_{k=-\infty}^{+\infty} \hat{\psi}(\omega + 2\pi k) \hat{\phi}^*(\omega + 2\pi k) = 0 \iff \hat{g}(\omega)\hat{h}^*(\omega) + \hat{g}(\omega + \pi)\hat{h}^*(\omega + \pi) = 0$$

We thus have the following solution

$$\hat{g}(\omega) = e^{-i\omega} \hat{h}^*(\omega + \pi)$$

---

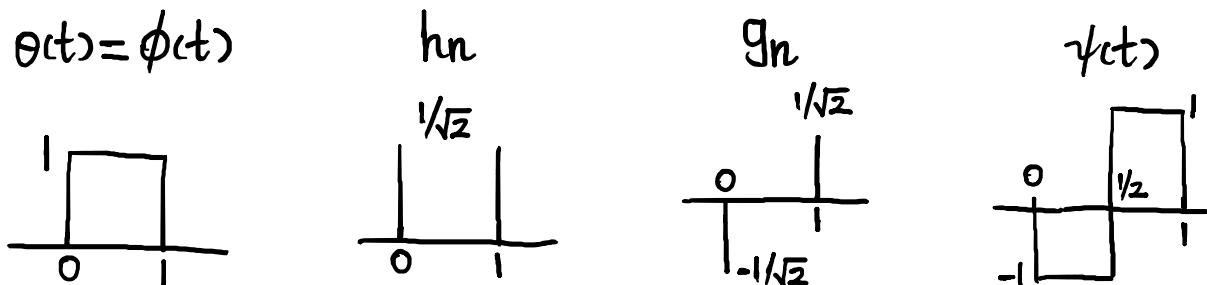
<b>Wavelets</b>	$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \sum_{n \in \mathbb{Z}} \textcolor{red}{g}_n \phi(t - n)$	$\hat{\psi}(\omega) = \frac{1}{\sqrt{2}}\hat{g}\left(\frac{\omega}{2}\right)\hat{\phi}\left(\frac{\omega}{2}\right)$
-----------------	---	--

---

<b>Steps to construct</b>	$\theta(t)$	$\phi(t)$	$h_n$	$\textcolor{red}{g}_n$			
	$\hat{\theta}(\omega)$	$\rightarrow$	$\phi(\omega)$	$\leftrightarrow$	$h(\omega)$	$\rightarrow$	$\hat{g}(\omega)$

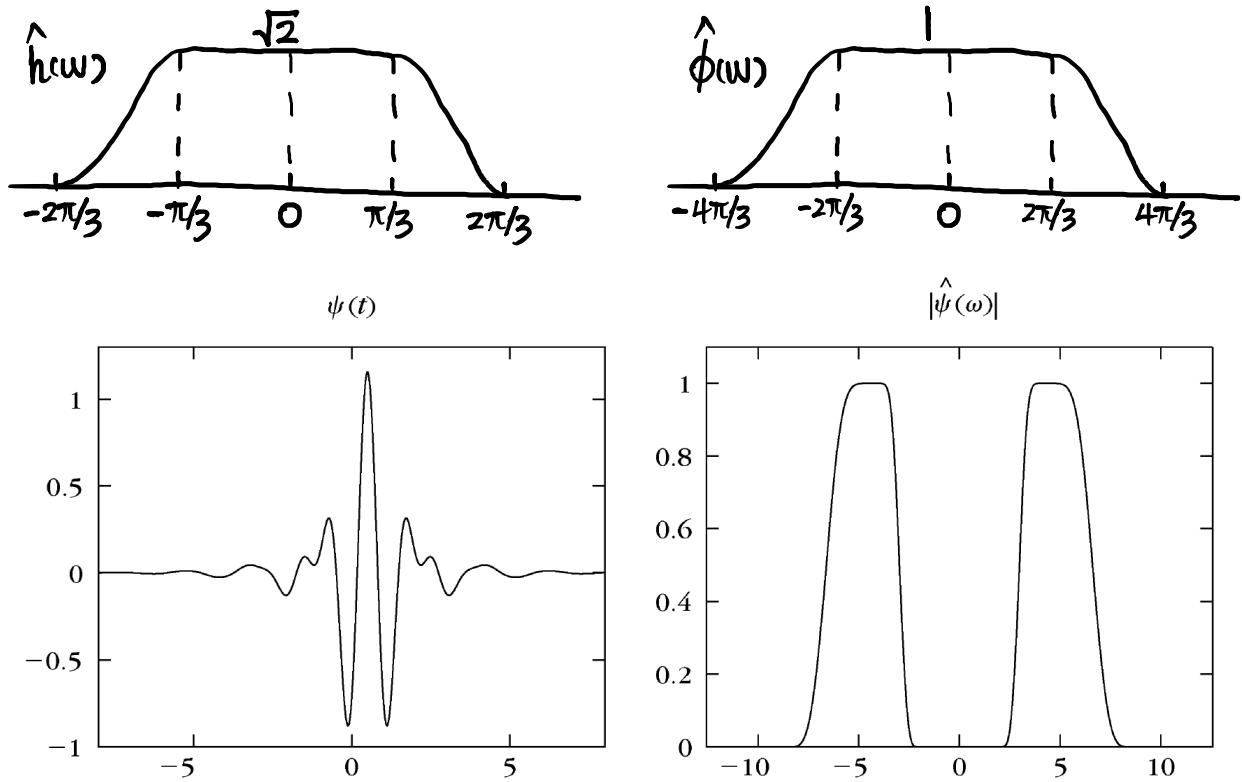
**Example:** Haar wavelets

$$\hat{h}(\omega) = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}e^{-i\omega}, \quad \hat{g}(\omega) = e^{-i\omega}\hat{h}^*(\omega + \pi) = -\frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}}e^{-i\omega}$$



**Example:** Meyer wavelets

Meyer wavelet is very localized in the frequency domain, but its convolution kernel is infinite as  $h_n \neq 0$  for all  $n$



**Example:** Daubechies compactly supported wavelets

For any given number  $p$  of vanishing moments, Daubechies wavelets  $\psi(t)$  have a support of minimum size  $[-p + 1, p]$ . The scaling function  $\phi(t)$  has a support of  $[0, 2p - 1]$ .

