Emory University

MATH 352 PDE's in Action

Learning Notes

Jiuru Lyu

May 1, 2025

Contents

1 Numerical Approximation of IVPs

1.1 Euler's Method

Example 1.1.1 Problem Set-Up

Suppose y_{t^n} represents the population at t^n. Suppose population grow with a parameter λ. Then, we form the following equation

$$y_{t^n+\Delta t} = y_{t^n} + \Delta t \lambda y_{t^n}.$$

Then,

$$\lim_{\Delta t \to 0} \frac{y_{t^n+\Delta t} - y_{t^n}}{\Delta t} = \lambda y_{t^n}.$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y, \quad y(0) = y_0 \qquad \text{(Cauchy Problem)}$$

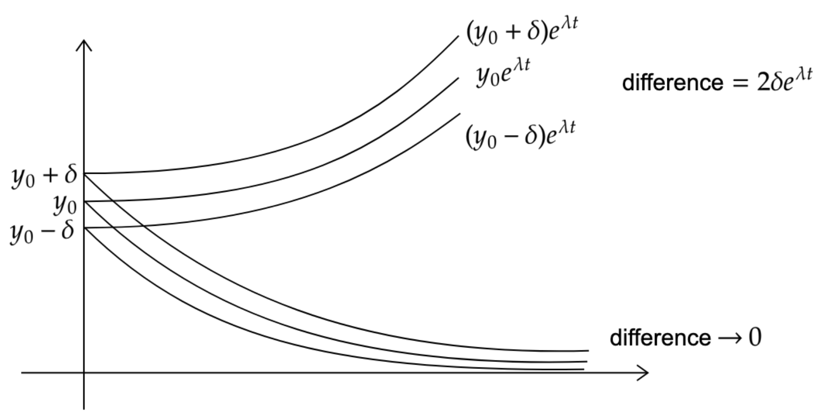
1. Solution: Separation of Variables.

$$\boxed{y(t) = y_0 e^{\lambda t}}$$

2. Evolution of Solution (Asymptotic Behavior):

   - $\lambda > 0$: $y \to \infty$ as $t \to \infty$
   - $\lambda < 0$: $y \to 0$ as $t \to 0$.
   - $\lambda = 0$: $y = y_0 \quad \forall\, t$.

3. Stability of Solution:



- When $\lambda > 0$, no matter how close our perturbation were, we will get very different asymptotic behavior $\implies$ unstable.
- When $\lambda < 0$, with perturbation, we are certain the asymptotic behavior of solution is to approach 0. So, $y = 0$ is an asymptotically stable solution.

Remark. Though we can find the exact solution in this example, it is not always the case. So, we need numerical approximation.

1.1.2 Solving the (Cauchy Problem) Numerically.

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y \implies \lim_{\Delta t \to 0} \frac{y(t+\Delta t) - y(t)}{\Delta t} = \lambda y(t).$$

1. Explicit Euler's Method: Collocate the problem at $t_1, t_2, t_3, \ldots$, where $t_{i+1} = t_i + \Delta t$.

$$\begin{aligned} \frac{y(t_0+\Delta t) - y(t_0)}{\Delta t} &= \lambda y(t_0) & \textit{Denote } u_1 &= y(t_0+\Delta t) = y(t_1) \\ \frac{u_1 - y_0}{\Delta t} &= \lambda y_0 & \implies u_1 &= y_0(1+\Delta t\lambda) \\ \frac{u_2 - u_1}{\Delta t} &= \lambda u_1 & \implies u_2 &= u_1(1+\Delta t\lambda) \\ \implies u_j &= u_{j-1}(1+\Delta t\lambda) & &= \cdots = y_0(1+\Delta t\lambda)^j \end{aligned}$$

Question: Given $\lambda < 0$. If $t \to \infty$, $j \to \infty$, does $u_j = y_0(1+\Delta t\lambda)^j \to 0$?

Short Answer: No. We need $|1+\Delta t\lambda| < 1$. So, the convergence depends on Δt.

2. Implicit Euler's Method:

Note that we can rewrite the derivative using

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{y(t) - y(t-\Delta t)}{\Delta t} = \lambda y(t).$$

$$\begin{aligned} \frac{y(t) - y(t-\Delta t)}{\Delta t} &= \lambda y(t) & \textit{Denote } u_1 &= y(t_1) \\ \frac{u_1 - y_0}{\Delta t} &= \lambda u_1 & \implies u_1 &= \frac{y_0}{1-\lambda\Delta t} \\ \frac{u_2 - u_1}{\Delta t} &= \lambda u_2 & \implies u_2 &= \frac{u_1}{1-\lambda\Delta t} = \frac{y_0}{(1-\lambda\Delta t)^2} \\ \implies u_j &= \frac{u_{j-1}}{1-\lambda\Delta t} = \frac{y_0}{(1-\lambda\Delta t)^j} \end{aligned}$$

Same question: Given $\lambda < 0$. If $t \to \infty$, $j \to \infty$, does $u_j \to 0$?

1.1.3 General Cauchy Problem.

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \\ y(0) = y_0 \end{cases} \tag{GCP}$$

Theorem 1.1.4 Existence and Uniqueness of Solution

Suppose f is continuous for $t \in I$. If f is such that $\exists$ positive constant L *s.t.* $|f(\cdot, y_1) - f(\cdot, y_2)| \leq L|y_1 - y_2|$ (*Lipschitz continuity*)

- for $y_1, y_2 \in R \subset \mathbb{R}$, $\exists$ a local unique solution to (GCP).
- $\forall\, y_1, y_2 \in \mathbb{R}$, $\exists$ a global unique solution to (GCP).

Algorithm 1: Explicit Euler (EE)

$\dfrac{u_1 - y_0}{\Delta t} = f(t_0, y_0)$;

$u_1 = y_0 + \Delta t f(t_0, y_0)$;

$u_2 = u_1 + \Delta t f(t_1, u_1)$;

$\implies u_j = u_{j-1} + \Delta t \cdot f(t_{j-1}, u_{j-1})$.

Algorithm 2: Implicit Euler (IE)

$\dfrac{u_1 - y_0}{\Delta t} = f(t_1, u_1)$ // implicit as u_1 is unknown. This is a root finding problem

$\dfrac{u_2 - y_0}{\Delta t} = f(t_2, u_2)$;

$\vdots$

1.1.5 Analysis of Explicit Euler's Method.

Definition 1.1.6 (Convergence). Let u_k be our numerical solution and y be the true solution. From EE, we know $u_k \approx y(t_k)$. Then, EE is *convergent* if

$$\lim_{\Delta t \to 0} u_k = y(t_k).$$

Theorem 1.1.7

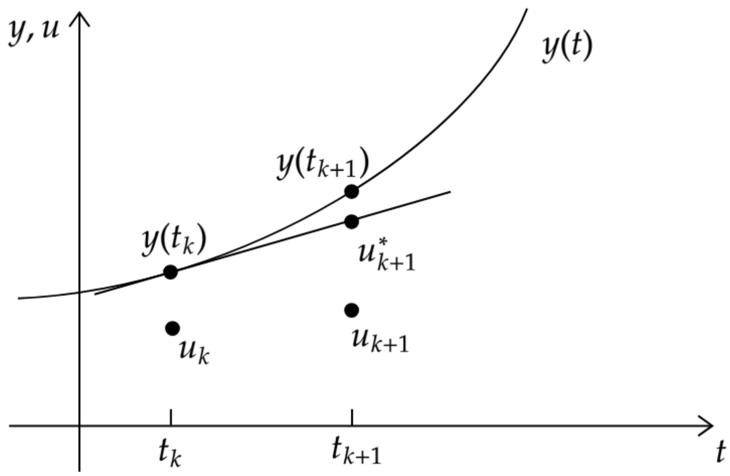
EE is convergent.

Proof 1. Define error $e_k = y(t_k) - u_k$. So, $e_{k+1} = y(t_{k+1}) - u_{k+1}$. Define the linear approximation of u_{k+1} as

$$u_{k+1}^* = y(t_k) + \Delta t f(t_k, y(t_k)).$$

Then, we can rewrite e_{k+1} into two parts:

$$e_{k+1} = y(t_{k+1}) - u_{k+1} = \underbrace{y(t_{k+1}) - u^*_{k+1}}_{\text{local}} + \underbrace{u^*_{k+1} - u_{k+1}}_{\text{Roll over}}$$



- Focus on the local part:

$$\frac{u^*_{k+1} - y(t_k)}{\Delta t} = f(t_k, y(t_k)).$$

But in general,

$$\frac{y(t_{k+1}) - y(t_k)}{\Delta t} \neq f(t_k, y(t_k)).$$

Using Taylor's expansion, we have

$$y(t_{k+1}) = y(t_k) + \frac{\mathrm{d}y}{\mathrm{d}t}(t_k)\Delta t + \frac{1}{2}\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t_k)\Delta t^2 + \cdots .$$

So,

$$\frac{y(t_{k+1}) - y(t_k)}{\Delta t} = f(t_k, y(t_k)) + \underbrace{\frac{1}{2}\frac{\mathrm{d}^2 y}{\mathrm{d}t^2}(t_k)\Delta t}_{\text{local truncation error}} .$$

Therefore,

$$e^*_{k+1} = y(t_{k+1}) - u^*_{k+1} \implies \frac{e^*_{k+1}}{\Delta t} = \frac{1}{2}c_k \Delta t, \quad \text{the local truncation error}.$$

Note that

$$\lim_{\Delta t \to 0} \frac{e^*_{k+1}}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{2}c_k \Delta t = 0 \implies \text{consistency}.$$

- The rolling over part:

$$\begin{aligned} u_{k+1}^* - u_{k+1} &= \underbrace{y(t_k)}\, + \Delta t f(t_k, y(t_k)) \underbrace{-u_k}\, - \Delta t f(t_k, u_k) \\ &= e_k + \Delta t f(t_k, y(t_k)) - \Delta t f(t_k, u_k) \end{aligned}$$

By Lipschitz continuity, we have

$$|f(t, u_A) - f(t, u_B)| \leq L \cdot |u_A - u_B|.$$

So, by triangle inequality,

$$|e_{k+1}| \leq \underbrace{|e_{k+1}^*|}_{\to 0 \text{ as } \Delta t \to 0} + \underbrace{|1 + \Delta t L||e_n|}_{\substack{\text{as } \Delta t \to 0 \text{, accumulates,} \\ \text{but bdd w.r.t } \Delta t \implies \text{stability}}}$$

So, the rate of convergence:

$$|e_k| \leq c \Delta t$$

is in the first order.

∎

Definition 1.1.8 (Absolute Stability). A numerical solution is *absolutely stable* when for $y(t) \to 0$, $t \to +\infty$, $u_i \to$ as $i \to +\infty$.

Example 1.1.9

Consider the ODE

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y; \; y(0) = y_0; \; \lambda < 0.$$

- With EE,

$$\frac{u_{i+1} - u_i}{\Delta t} = \lambda u_i \implies u_{i+1} = u_i(1 + \Delta t \lambda) = y_0(1 + \Delta t \lambda)^{i+1}.$$

  When $i \to \infty$,

$$|u_{i+1}| = \left| y_0 (1 + \Delta t \lambda)^{i+1} \right| \to 0$$

  when $|1 + \Delta t \lambda| < 1$. ($1 + \Delta t \lambda$ *is called a damping factor*)

  So, we have

$$-1 < 1 + \Delta t \lambda < 1.$$

  As $\Delta t > 0$ and $\lambda < 0$, we have

$$-1 < 1 - \Delta t |\lambda| < 1 \implies \Delta t < \frac{2}{|\lambda|}.$$

So, EE is *conditionally absolutely stable.* However, this condition is bad, especially for large λ.

- With IE,

$$\frac{u_i - u_{i-1}}{\Delta t} = \lambda u_i \implies u_i = \frac{u_{i-1}}{1 - \Delta t \lambda} = \frac{y_0}{(1 - \Delta t \lambda)^i}.$$

To have $u_i \to 0$ as $i \to +\infty$, we need

$$\frac{1}{1 - \Delta t \lambda} < 1.$$

As $\lambda < 0$, it s equivalent as

$$\frac{1}{1 + \Delta|\lambda|} < 1.$$

This is true $\forall\, \Delta t$. So IE is *(unconditionally) absolutely stable.*

1.2 Crank-Nicolson Method

Consider the Cauchy problem

$$\begin{cases} \dfrac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \\ y(0) = y_0. \end{cases}$$

One can compute $y(t)$ by

$$y(t) = y_0 + \int_0^t f(\tau, y(\tau))\, \mathrm{d}\tau.$$

So, if we discretize the problem, we have

$$y(t_1) = y_0 + \int_0^{t_1} f(\tau, y(\tau))\, \mathrm{d}\tau.$$

If we use the trapezoid rule to approximate the integral, we get the numerical solutions:

$$\begin{aligned} u_1 &= y_0 + \frac{\Delta t}{2}(f(t_0, y_0) + f(t_1, u_1)) \\ u_2 &= u_1 + \frac{\Delta t}{2}(f(t_1, y_1) + f(t_2, u_2)) \end{aligned}$$

Generalize, we have

$$u_{i+1} = u_i + \frac{\Delta t}{2}(f_i + f_{i+1}), \quad \text{where } f_i = f(t_i, u_i). \tag{CN}$$

This is an *implicit method* because u_{i+1} appears on both sides of the formula.

As the error of Trapezoid Rule is $\sim \mathcal{O}((b-a)^2)$, the error of Crank-Nicolson method is also $\sim \mathcal{O}(\Delta t^2)$.

1.3 Heun Method

Recall (CN):

$$u_{i+1} = u_i + \frac{\Delta t}{2}(f(t_i, u_i) + f(t_{i+1}, u_{i+1})) \tag{CN; Corrector}$$

is an implicit method. We can integrate it with EE:

$$u_{i+1} = u_i + \Delta t f(t_i, u_i) =: u_{i+1}^* \tag{EE; Predcitor}$$

Then, we form the Heun method as follows

$$\begin{aligned} u_{i+1} &= u_i + \frac{\Delta t}{2}(f(t_i, u_i) + f(t_{i+1}, u_{i+1})) \\ &= u_i + \frac{\Delta t}{2}(f(t_i, u_i) + f(t_{i+1}, u_i + \Delta t f(t_i, u_i))) \\ &= u_i + \frac{\Delta t}{2}\big(f(t_i, u_i) + f(t_{i+1}, u_{i+1}^*)\big) \end{aligned} \tag{H}$$

Heun is also a second order method, and it is explicit.

In Heun, u_{i+1}^* uis called a *predictor*, and CN is called a *corrector*.

Theorem 1.3.1
Crank-Nicolson is unconditionally stable.

Proof 1.

$$u_{i+1} = u_i + \frac{\Delta t}{2}(-\lambda u_i - \lambda u_{i+1}).$$

$$u_{i+1} = \frac{1 - \frac{\Delta}{2}\lambda}{1 + \frac{\Delta t}{2}\lambda} u_i \implies u_{i+1} = \left| \frac{1 - \frac{\Delta t}{2}\lambda}{1 + \frac{\Delta t}{2}\lambda} \right|^{i+1} y_0.$$

Since $\Delta t, \lambda > 0$, $1 - \frac{\Delta t}{2}\lambda < 1 + \frac{\Delta t}{2}\lambda$. Hence,

$$\left| \frac{1 - \frac{\Delta t}{2}\lambda}{1 + \frac{\Delta t}{2}\lambda} \right| < 1 \quad \forall\, \Delta t > 0.$$

So, $u_{i+1} \to 0$ when $i \to \infty$. Then, CN is unconsidtionally stable. ■

Summary: ODE Methods

Table 1: Summary of Numerical ODE Methods

Method	Order	Absolute Stability	Implicit/Explicit
Explicit Euler	1	Conditional	Explicit
Implicit Euler	1	Unconditional	Implicit
Crank-Nicolson	2	Unconditional	Implicit
Heun	2	Conditional	Explicit

- The stability condition of Heun method is the same as that of Explicit Euler.
- All explicit methods are conditionally stable.
- But implicit methods may be both conditionally or unconditionally stable. There is a trade-off: more accuracy $\implies$ less stability.
- So, it is a case-by-case decision for which method(s) to use.

1.4 From Model to General Problems

If we use λ to denote the characteristic of the problem that determines the stability of the problem, what are λ's in general problems?

(1)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \qquad \text{(General ODE)}$$

Note that

$$f(t, y) \approx f(t_0, y_0) + \frac{\partial f}{\partial y}(y - y_0) \approx \lambda y + f_0 - y_0,$$

where $f_0 = f(t_0, y_0)$, we see that $\lambda \approx \dfrac{\partial f}{\partial y}$.

(2)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = Ay \qquad \text{(System of ODEs)}$$

Let's apply EE to the system:

$$\frac{u_{i+1} - u_i}{\Delta t} = Au_i$$
$$u_{i+1} = u_i + \Delta t A u_i = (I + \Delta t A)u_i.$$

On the other hand, if we apply IE for the system,

$$(I - \Delta t A)u_{i+1} = u_i.$$

We, therefore, need to solve the following linear system:

$$Bu_{i+1} = u_i, \quad \text{where } B = I - \Delta t A.$$

Hence, IE converges as long as $I - \Delta t A$ is nonsingular.

From the two examples of applying EE and IE, we see that eigenvalues determines the stability of the system. Hence, we choose $\lambda = \max|\text{eig}(A)|$, the *spectral radius*. Meanwhile, the system is *asymptotically stable* if $\text{Re}(\text{eig}(A)) < 0$.

(3)=(1)+(2)

$$\frac{\mathrm{d}y}{\mathrm{d}t} = F(t, y),$$

where $F = (f_1, f_2, \ldots, f_m) : \mathbb{R}^m \to \mathbb{R}^n$ and $y = (y_1, y_2, \ldots, y_n)$. Then, we can form the Jacobian of F:

$$J = \left[\frac{\partial f_i}{\partial y_j}\right]_{(i,j)},$$

and thus the quantity of interest is

$$\lambda = \max|\text{eig}(J)|.$$

1.5 Multistep Methods

1.5.1 Midpoint Method (Two-Step Method)

Let's approximate the derivative in the following fashion:

$$\begin{aligned} \left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t_i} &\approx \frac{y_{i+1} - y_{i-1}}{2\Delta t} \\ f(t_i, y_i) = \left.\frac{\mathrm{d}y}{\mathrm{d}t}\right|_{t_i} &\approx \frac{u_{i+1} - u_{i-1}}{2\Delta t} \\ \implies u_{i+1} &= u_{i-1} + 2\Delta t f(t_i, y_i) \end{aligned} \tag{Midpoint}$$

- Initial Condition:

$$u_2 = y_0 + 2\Delta t f(t_1, u_1),$$

where $u_1 = y_0 + \Delta t(ft_0, y_0)$ from EE. However, this approach is bad since its error only $\sim \mathcal{O}(\Delta t)$. Another approach to consider is to use Heun to compute u_1. This approach is relatively good since its error is $\sim \mathcal{O}(\Delta t^2)$.

Remark. How to build the initial condition(s) is one key for multistep problems.

- This method is unconditionally unstable.

  Proof 1. Consider the Cauchy Problem

$$\begin{cases} \frac{\mathrm{d}y}{\mathrm{d}t} = -\lambda y, \quad \lambda > 0 \\ y(0) = y_0. \end{cases}$$

Using the (Midpoint), we have

$$u_{i+1} = u_{i-1} - 2\Delta t\lambda u_i \implies u_{i+1} + 2\Delta t\lambda u_i - u_{i-1} = 0. \quad \text{(2}^{\text{nd}}\text{ Order Difference Equation)}$$

To solve it, let's guess

$$u_i = c\rho^i, \quad c \neq 0$$

is a solution. Then, plut it in to the difference equation, we get

$$\begin{aligned} c\rho^{i+1} + 2\Delta t\lambda c\rho^i - c\rho^{i-1} &= 0, \quad c \neq 0 \\ \rho^2 + 2\Delta t\lambda\rho - 1 &= 0 \qquad \left[\text{Divide by } c\rho^{i-1}\right] \end{aligned}$$

Suppose ρ_0 and ρ_1 are two solutions. Then,

$$(\rho - \rho_0)(\rho - \rho_1) = 0 \implies \rho^2 - (\rho_0 + \rho_1)\rho + \rho_0\rho_1 = 0.$$

So, it must be that

$$|\rho_0\rho_1| = 1.$$

WLOG, suppose $\rho_0 < 1$, then $\rho_1 > 1$. Then,

$$u_i = c_0\rho_0^i + c_1\rho_1^i, \quad \text{for some } c_0, c_1.$$

Then, we know $u_1 \not\to 0$ when $i \to +\infty$ in all cases. So, this method is unconditionally unstable. ■

1.5.2 Design a Better Method: Backward Differentiation Formula (BDF)

Since (Midpoint) is unconditionally unstable, we should not use it at any cost. However, a multistep method adds more accuracy to the numerical solution. Our job now is to find a design such that the error can be of order p, where p is of the user's choice (i.e. error $\sim \mathcal{O}(\Delta t^p)$).

Taking inspiration from IE:

$$\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} = \frac{u_i - u_{i-1}}{\Delta t}.$$

So, to design a two-step method, we consider the Taylor's expansion:

$$\begin{aligned}
u_{i-1} &= u_i - \left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} \Delta t + \left.\frac{\mathrm{d}^2u}{\mathrm{d}t^2}\right|_{t_i} \frac{\Delta t^2}{2} - \left.\frac{\mathrm{d}^3u}{\mathrm{d}t^3}\right|_{t_i} \frac{\Delta t^3}{6} + \cdots \\
u_{i-2} &= u_i - \left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} 2\Delta t + \left.\frac{\mathrm{d}^2u}{\mathrm{d}t^2}\right|_{t_i} \frac{4\Delta t^2}{2} - \left.\frac{\mathrm{d}^3u}{\mathrm{d}t^3}\right|_{t_i} \frac{8\Delta t^3}{6} + \cdots
\end{aligned}$$

We want $\alpha u_{i-1} + \beta u_{i-2}$ to contain only up to the $\frac{\mathrm{d}u}{\mathrm{d}t}\Delta t$ term. So, we want

$$\begin{cases}
-\alpha - 2\beta = 1 & \text{so that the } \frac{\mathrm{d}u}{\mathrm{d}t} \text{ term has coefficient of } 1 \\
\\
\alpha + 4\beta = 0 & \text{so that the } \frac{\mathrm{d}^2u}{\mathrm{d}t^2} \text{ term has coefficient of } 0
\end{cases}.$$

Remark. Coefficients are chosen according to coefficients in the Taylor's expansion.

Solving the system, we get

$$\begin{cases}
\alpha = -2 \\
\beta = \frac{1}{2}.
\end{cases}$$

Let's test that this method really works:

$$\begin{aligned}
-2u_{i-1} &= -2u_i + 2\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} \Delta t - \left.\frac{\mathrm{d}^2u}{\mathrm{d}t^2}\right|_{t_i} \Delta t^2 + \mathcal{O}\big(\Delta t^3\big) \\
\frac{1}{2}u_{i-2} &= \frac{1}{2}u_i - \left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} \Delta t + \left.\frac{\mathrm{d}^2u}{\mathrm{d}t^2}\right|_{t_i} \Delta t^2 + \mathcal{O}\big(\Delta t^3\big) \\
-2u_{i-1} + \frac{1}{2}u_{i-2} &= -2u_i + \frac{1}{2}u_i + \left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} \Delta t + \mathcal{O}\big(\Delta t^3\big).
\end{aligned}$$

Then,

$$\begin{aligned}
\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} \Delta t &= \frac{1}{2}u_{i-2} - 2u_{i-1} - \frac{3}{2}u_i + \mathcal{O}\big(\Delta t^3\big) \\
\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t_i} &= \frac{u_{i-2} - 4u_{i-1} - 3u_i}{2\Delta t} + \mathcal{O}\big(\Delta t^3\big).
\end{aligned}$$

Thus, we have successfully built an **implicit order** 2 method.

Extension 1.1 (Higher Order Method) *If we want to build a 4-th order method, we can consider the Taylor expansion for $u_{i-1}, u_{i-2}, u_{i-3}, u_{i-4}$. Then, we choose coefficients $\alpha, \beta, \gamma, \delta$ such that $\alpha u_{i-1} + \beta u_{i-2} + \gamma u_{i-3} + \delta u_{i-4}$ only contain up to $\frac{\mathrm{d}u}{\mathrm{d}t}$ term.*

Remark 2. (Partical Considerations).

- When building such a method, we need to consider the differentiability of the function when deciding the order.
- Theoretically, we can go as many orders as we want, but we need to be careful when getting too high orders. Generally, higher order, more accuracy, but less stability.

1.6 Higher Order Methods

Definition 1.6.1 (Linear Multistep Methods).

$$u_{n+1} = \sum_{j=0}^{p} a_j u_{n-j} + \Delta t \sum_{j=0}^{p} b_j f(t_{n-j}, u_{n-j}) + \Delta t b_{-1} f(t_{n+1}, u_{n+1})$$

- This method is implicit if $b_{-1} \neq 0$.
- We can use a polynomial to represent the method:

$$\pi(\rho) = \rho^{p+1} - \sum_{j=1}^{p} a_j \rho^{p-j}.$$

Example 1.6.2 BDF Methods

Given that $\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t=t_n} \approx f(t_{n+1}, u_{n+1})$, we have

$$\frac{u_{n+1} - \sum_{j=0}^{p} a_j u_{n-j}}{\Delta t} \approx f(t_{n+1}, u_{n+1}),$$

where

$$a_j = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \quad b_j = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{ for } j = 0, 1, \ldots, p, \quad \text{and } b_{-1} \neq 0.$$

Specifically, BDF2 gives us

$$u_{n+1} = \frac{4}{3}u_n - \frac{1}{3}u_{n-1} + \frac{2}{3}\Delta t f(t_{n+1}, u_{n+1}).$$

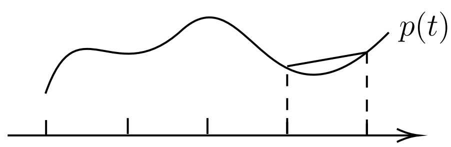
So, $\pi_{\text{BDF2}}(\rho) = \rho^2 - \frac{4}{3}\rho + \frac{1}{3}$.

Definition 1.6.3 (Adams). We know that

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(\tau, y(\tau))\,\mathrm{d}\tau.$$

We can interpolate points $\{t_i, y(t_i)\}_{i=0}^n$ using polynomial $p(t)$. Then, we have

$$y(t_{n+1}) \approx y(t_n) + \int_{t_n}^{t_{n+1}} p(t)\,\mathrm{d}t.$$



Example 1.6.4 Examples of Adams Method

- Adams-Bashforth:

$$u_{n+1} = u_n + \frac{\Delta t}{12}(23f_n - 16f_{n-1} + 5f_{n-2}) \tag{AB3}$$

Here, $b_{-1} = 0, b_1 = \frac{23}{12}, b_1 = -\frac{16}{12}, b_2 = \frac{5}{12}$, and $a_0 = 1, a_1 = 0, a_2 = 0$. Meanwhile,

$$\pi_{\text{AB3}}(\rho) = \rho^4 - \rho^2.$$

- Adams-Moulton:

$$u_{n+1} = u_n + \frac{\Delta t}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}). \tag{AM4}$$

Here, $a_0 = 1, a_1 = 0, a_2 = 0$, and $b_{-1} = \frac{9}{24}, b_0 = \frac{19}{24}, b_1 = \frac{-5}{24}, b_2 = \frac{1}{24}$.

Theorem 1.6.5 Consistency and Convergence

- If $\sum_{j=0}^{p} a_j = 1$ and $-\sum_{j=0}^{p} ja_j + \sum_{j=0}^{p} b_j + b_{-1} = 1$, then the method is consistent.
- Suppose r is the root of $\pi(\rho) = 0$. If $\forall\, r_j$, either:
  1. $|r_j| < 1$, or
  2. $|r_j| = 1$ and $\pi'(r_j) \neq 0$,

  then the method is convergent.

Example 1.6.6 BDF2 is Consistent

Recall BDF2:

$$u_{n+1} = \frac{4}{3}u_n - \frac{1}{3}u_{n-1} + \frac{2}{3}\Delta t f(t_{n+1}, u_{n+1}).$$

Then, $a_0 = \frac{4}{3}$, $a_1 = -\frac{1}{3}$, $b_{-1} = \frac{2}{3}$. So,

$$\sum_{j=0}^{1} a_j = \frac{4}{3} - \frac{1}{3} = 1$$

and

$$-\sum_{j=0}^{1} ja_j + \sum_{j=0}^{1} b_j + b_{-1} = \left(-0 \cdot \frac{4}{3} + 1\left(-\frac{1}{3}\right)\right) + 0 + 0 + \frac{1}{2} = \frac{1}{3} + \frac{2}{3} = 1.$$

So, the method is consistent. Further, the polynomial representation of BDF2 is

$$\pi_{\text{BDF2}}(\rho) = \rho^2 - \frac{4}{3}\rho + \frac{1}{3}.$$

Then, the roots are $r_1 = 1$, $r_2 = \frac{1}{3}$. Note that $|r_1| = 1$ and $|r_2| = \left|\frac{1}{3}\right| < 1$. Further, $\pi'(1) \neq 0$. So, the method is convergent.

Definition 1.6.7 (Runge-Kutta Method). $u_{n+1} = u_n + \Delta t \sum_{i=1}^{s} b_i K_i$, where s is the number of stages, and $K_i = f(t_n + c_i\Delta t, u_n + \Delta t \sum_{j=1}^{s} a_{ij}K_j)$. The quantity of c, A, and $b^\top$ will be represented using a *Butcher array*.

1.7 Systems

Consider

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t,y), \quad \text{where } f, y \text{ are vectors, and } y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_n(t) \end{bmatrix}$$

1.7.1 Stability. We can regard the system as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t,y) = Ay.$$

Then, we can diagonalize A as $A = T^{-1}DT$. Hence,

$$\begin{aligned} \frac{\mathrm{d}y}{\mathrm{d}y} &= Ay = \big(T^{-1}DT\big)y \\ T\frac{\mathrm{d}y}{\mathrm{d}t} &= T\big(T^{-1}DT\big)y \\ \frac{\mathrm{d}(Ty)}{\mathrm{d}t} &= D(Ty) && \text{Denote } w = Ty \\ \frac{\mathrm{d}w}{\mathrm{d}t} &= Dw. \end{aligned}$$

Suppose we apply EE to the system, we get

$$\begin{aligned} \frac{1}{\Delta t}(u_{n+1} - u_n) &= Au_n \\ u_{n+1} &= (I + \Delta t A)u_n. \end{aligned}$$

Then, for stability, we require

$$\Delta t < \frac{2}{|\lambda_i|} \leq \frac{2}{\max|\lambda_i|}, \quad \text{where } \max|\lambda_i| \text{is the Spectral Radius}.$$

So, EE is conditionally stable.

However, if we apply Crank-Nicolson, we get

$$\begin{aligned} \frac{u_{n+1} - u_n}{\Delta t} &= \frac{1}{2}(f(t_{n+1}, u_{n+1}) + f(t_n, u_n)). \\ \frac{1}{\Delta t}(u_{n+1} - u_n) &= \frac{1}{2}Au_n + \frac{1}{2}Au_{n+1} \\ \left(I - \frac{\Delta t}{2}A\right)u_{n+1} &= \left(I + \frac{\Delta t}{2}A\right)u_n. \end{aligned}$$

Denote $-\dfrac{\Delta t}{2}A = B$. Then, $\operatorname{eig}\left(I - \dfrac{\Delta t}{2}A\right) = \operatorname{eig}(I+B) = 1+\operatorname{eig}(B) > 0$. Therefore, the system will always be solvable, and thus CN is unconditionally stable.

1.8 Terminology Clarification

Definition 1.8.1 (Consistency). Given

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y).$$

An algorithm is *consistent* if

$$\lim_{\Delta t \to 0} \frac{y_{i+1} - y_i}{\Delta t} = f(t_{i+1}, y_{i+1}).$$

Example 1.8.2

Consider $\dfrac{\mathrm{d}y}{\mathrm{d}t} = -\lambda y$ with $y(0) = 1$. Then, $y_{\text{exact}} = e^{-\lambda t}$.

$$\begin{aligned}
\frac{y(t_{i+1}) - y(t_i)}{\Delta t} &\neq -\lambda y(t_{i+1}) \\
\frac{e^{-(t_i+\Delta t)} - e^{-\lambda t_i}}{\Delta t} &\neq -\lambda e^{-\lambda(t_i+\Delta t)}.
\end{aligned}$$

We want to investigate the quantity

$$\begin{aligned}
\frac{e^{-(t_i+\Delta t)} - e^{-\lambda t_i}}{\Delta t} - \lambda e^{-\lambda(t_i+\Delta t)} &= \frac{e^{-\lambda t_i} e^{-\lambda \Delta t} - e^{-\lambda t_i}}{\Delta t} + \lambda e^{-\lambda t_i} e^{-\lambda \Delta t} \\
&= e^{-\lambda t_i}\left(\frac{e^{-\lambda \Delta t} - 1}{\Delta t} + \lambda e^{-\lambda \Delta t}\right).
\end{aligned}$$

Consider Taylor's expansion:

$$\begin{aligned}
e^{-\lambda \Delta t} &= 1 - \lambda \Delta t + \frac{\lambda^2}{2}\Delta t^2 - \frac{\lambda^3}{3}\Delta t^3 + \cdots \\
e^{-\lambda \Delta t} - 1 &= -\lambda \Delta t + \frac{\lambda^2}{2}\Delta t^2 - \frac{\lambda^3}{3}\Delta t^3 + \cdots \\
\frac{e^{-\lambda \Delta t} - 1}{\Delta t} &= -\lambda + \frac{\lambda^2}{2}\Delta t - \frac{\lambda^3}{3}\Delta t^2 + \cdots \\
\lambda e^{-\lambda \Delta t} &= \lambda - \lambda^2 \Delta t + \frac{\lambda^3}{2}\Delta t^2 - \frac{\lambda^4}{3}\Delta t^3 + \cdots
\end{aligned}$$

So,

$$\frac{e^{-\lambda \Delta t} - 1}{\Delta t} + \lambda e^{-\lambda \Delta t} = -\frac{\lambda^2}{2}\Delta t - \frac{\lambda^3}{6}\Delta t^2 + \cdots \sim \mathcal{O}(\Delta t) = C\Delta t.$$

Then,

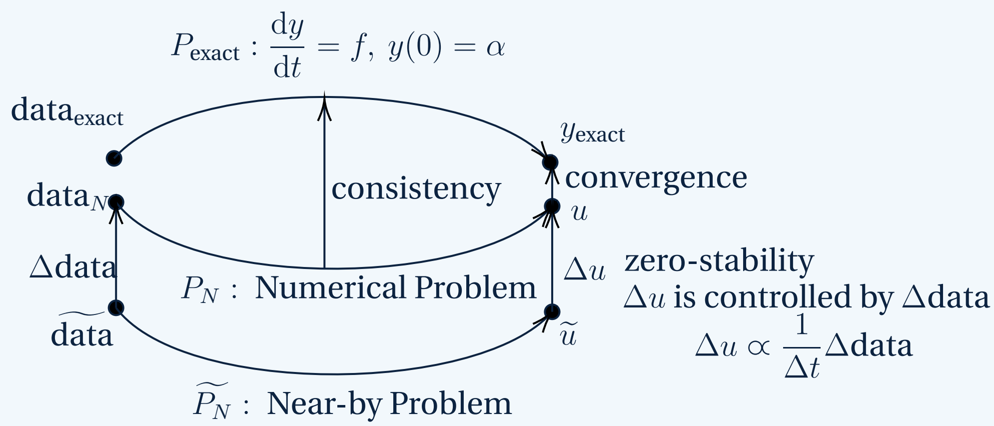
$$e^{-\lambda t_i}\left(\frac{e^{-\lambda \Delta t} - 1}{\Delta t} + \lambda e^{-\lambda \Delta t}\right) = C\Delta t e^{-\lambda t_i}.$$

When $\Delta \to 0$,

$$e^{-\lambda t_i}\left(\frac{e^{-\lambda \Delta t} - 1}{\Delta t} + \lambda e^{-\lambda \Delta t}\right) = C\Delta t e^{-\lambda t_i} \to 0.$$

So, this method is consistent.

Definition 1.8.3 (Zero Stability and Convergence).



Example 1.8.4

Consider the linear system $Au = r(\Delta t)$ with $\|r\| \to 0$ as $\Delta t \to 0$. Then,

$$u = A^{-1}r.$$

One have $\|u\| \leq \|A^{-1}\| \cdot \|r\|$. When $\Delta t \to 0$, though $\|r\| \to 0$, $\|A^{-1}\|$ can be still huge, leading to unstable u.

Definition 1.8.5 (Absolute Stability). Asympototic behavior of the method when $t \to \infty$.

2 Iterative Methods

Problem: $Ax = b$.

2.1 Introduction and Definitions

- Direct methods: Gauss-Elimination:

$$A = LU,$$

where L is lower triangular and U is upper triangular.

To solve, $Ax = LUx = b$. We solve two systems: $Ly = b$ and $Ux = y$.

(+) Cost $\mathcal{O}(n^3)$ for $A \in \mathbb{R}^{n \times n}$

(+) Finite number of steps to solution

(-) If A is sparse (# non-zero entries $\ll$ total # of entries), in general, L and U are full. Therefore, computing LU factorization will consume huge memory.

- Iterative Methods General Expression:

$$x^{(k+1)} = Bx^{(k)} + g \tag{Iter}$$

Cost: $\mathcal{O}(n^2 \cdot M)$, where M is the number of iterations. So if $n^2 \cdot M \ll n^3$ (that is, $M \ll n$), we win.

Example 2.1.1 Iterative Methods

Consider $2I_d x = b$ with exact solution $x_{\text{ex}} = \frac{1}{2}b$.
We know $x + x = b$. So,

$$x = -x + b.$$

Then, our iterative update will be

$$x^{(k+1)} = -I_d x^{(k)} + b, \qquad \text{where } B = -I_d,\ g = b$$

- If $x^{(k)} = x_{\text{ex}} = \frac{1}{2}$, do we say at x_{ex}?

$$x^{(k+1)} = -I_d \cdot \left(\frac{1}{2}b\right) + b = \frac{1}{2}b = x_{\text{ex}}.$$

So, yes. The method is therefore *consistent*.

- If $x^{(k)} = 0$, then we have

$$x^{(k+1)} = 0 + b = b, \quad x^{(k+1)} = -I_d \cdot b + b = 0, \quad x^{(k+3)} = 0 + b = b, \cdots$$

The iterates oscillates between 0 and b. BAD initial guess.

What if we change a method? Note that

$$2I_d x = \alpha I_d x + (2-\alpha) I_d x = b.$$

Then, the update rule can be

$$x^{(k+1)} = \frac{\alpha - 2}{\alpha} I_d x^{(k)} + \frac{1}{\alpha} b, \quad \text{where } B = \frac{\alpha-2}{\alpha} I_d,\ g = \frac{1}{\alpha} b.$$

Let our initial guess to be $x^{(0)} = 0$.

- If $\alpha = 2$, then the solution converge to $x_{\text{ex}} = \frac{1}{2} b$ in 1 step.
- If $\alpha = \frac{3}{2}$, then $x^{(0)} = 0$, $x^{(1)} = -\frac{1}{3} b + \frac{2}{3} b = \frac{1}{3} b$, $x^{(2)} = -\frac{5}{9} b, \ldots$. We do converge in this case, but we need a lot of steps.
- If $\alpha = \frac{1}{2}$, we have $x^{(0)} = 0$, $x^{(1)} = 2b$, $x^{(2)} = -b$. and $x^{(3)} = 5b$. In fact, we don't converge with this choice of α.

Theorem 2.1.2 Convergence of an Iterative Method
Let $\rho(B)$ be the spectrum radius of B. i.e., $\rho(B) = \max_i |\lambda_i|$.

- the iterative method converges $x^{(k)} \to \overline{x}$ as $k \to \infty \iff \rho(B) < 1$.
- $\overline{x} = x_{\text{ex}}$ (i.e., $\overline{x}$ is the exact solution for $Ax = b$) $\iff \overline{x} = B\overline{x} + g$ (i.e., $\overline{x}$ is a fixed point of the iterative method).
- The smaller $\rho(B)$, the faster convergence.

Therefore, since $B = \frac{\alpha-2}{\alpha} I_d$, we know that $\rho(B) = \left|\frac{\alpha-2}{\alpha}\right|$.

- Optimal convergence: $\rho(B) = 0$: $\frac{\alpha-2}{\alpha} = 0 \implies \alpha^* = 2$.
- When $\alpha = \frac{1}{2}$, $\rho(B) = \left|\frac{1/2 - 2}{1/2}\right| = 3 > 1 \implies$ no convergence.

Definition 2.1.3 (Consistency). An iterative method (Iter) is *consistent* with the linear system $Ax = b$ when x_{ex} is a stationary point of (Iter) (i.e., fixed point):

$$Bx_{\text{ex}} + g = x_{\text{ex}}$$

Definition 2.1.4 (Convergence of an Iterative Method). The iterative method (Iter) is convergent to the solution x_{ex} of the linear system $Ax = b$ when

$$\lim_{k\to\infty} \left\| e^{(k)} \right\| = 0,$$

where $e^{(k)} = x^{(k)} - x_{\text{ex}}$.
If $\exists C = \rho(B) < 1 \; s.t. \left\| e^{(k+1)} \right\| \leq C \cdot \left\| e^{(k)} \right\| \quad \forall\, k \geq 0$, then we guarantee convergence regardless of the initial guess $x^{(0)}$.

2.2 Richardson Method

$$\begin{aligned} Ax &= b \\ x - x &= \alpha(b - Ax) = 0 \\ x&x - \alpha Ax + \alpha b \\ x^{(k+1)} &= (I - \alpha A)x^{(k)} + \alpha b, \end{aligned}$$

where $B = I - \alpha A$, $g = \alpha b$

- We converge $\iff \rho(I - \alpha A) < 1$.
- If A is SPD (all eigenvalues are real and $x^\top Ax > 0$), then if

  $$0 < \alpha < \frac{2}{\lambda_{\max}},$$

  we converge. The optimal convergence rate attains when

  $$\alpha^* = \frac{2}{\lambda_{\min} + \lambda_{\max}}.$$

- Conditioning: $\kappa(A) = \dfrac{\lambda_{\max}}{\lambda_{\min}} \geq 1$.
  If $\kappa(A)$ is high, slow convergence. If $\kappa(A)$ is slow, fast convergence. Specially, if $\kappa(A) = 1$, then A is unitary matrix such that $A^*A = AA^* = I_d$.

- Stopping Criteria:
  - Residual: $r^{(k)} = b - Ax^{(k)}$: $\left\|r^{(k)}\right\| \leq \texttt{tol}$

    Problem: If $\kappa(A)$ is high, BAD.
  - Consecutive iterations: $\left\|x^{(k+1)} - x^{(k)}\right\| \leq \texttt{tol}$

    Why it work?

$$\underbrace{x^{(k)} - x_{\text{ex}}}_{e^{(k)}} = x^{(k)} - x^{(k+1)} + \underbrace{x^{(k+1)} - x_{\text{ex}}}_{e^{(k+1)}}$$

So,

$$\left\|e^{(k)}\right\| \leq \left\|e^{(k)} - x^{(k+1)}\right\| + \left\|e^{(k+1)}\right\|.$$

If the method is convergent, $\left\|e^{(k+1)}\right\| \leq \rho(B)\left\|e^{(k)}\right\|$. So,

$$\begin{aligned}\left\|e^{(k)}\right\| &\leq \left\|x^{(k)} - x^{(k+1)}\right\| + \left\|e^{(k+1)}\right\| \\ &\leq \left\|x^{(k)} - x^{(k+1)}\right\| + \rho(B) \cdot \left\|e^{(k)}\right\| \\ \left\|e^{(k)}\right\| &\leq \frac{1}{1-\rho(B)}\left\|x^{(k)} - x^{(k+1)}\right\|.\end{aligned}$$

2.3 Preconditioning

Definition 2.3.1 (Preconditioner). A preconditioner P is an invertible matrix (i.e., $\det(P) \neq 0$) such that $P^{-1}Ax = P^{-1}b$ with reduced $\kappa(P^{-1}A)$.

Remark. In other words, we require $P^{-1}A \approx I$. So, P needs to be close to A and be easy to solve at hte same time. However, these two requirements are exactly the opposite.

Example 2.3.2 How to come up with a P?

In Richardson method, we have

$$\begin{aligned}P\underbrace{\left(x^{(k+1)} - x^{(k)}\right)}_{\delta} &= -\alpha A x^{(k)} + \alpha b \\ &= \alpha r^{(k)}, \quad \text{where } r^{(k)} = b - Ax^{(k)} \text{ is the residual.}\end{aligned}$$

Note

$$\delta = x^{(k+1)} - x^{(k)} \implies x^{(k+1)} = x^{(k)} + \delta = -\alpha P^{-1}Ax^{(k)} + \alpha P^{-1}b.$$

So, we want $\kappa(P^{-1}A) \ll \kappa(P^{-1}b)$.

Theorem 2.3.3 Convergence

For A SPD,

$$\alpha^* = \frac{2}{\lambda_{\min} + \lambda_{\max}},$$

the following convergence estimate holds:

$$\left\|e^{(k)}\right\|_A \leq \left(\frac{\kappa(P^{-1}A) - 1}{\kappa(P^{-1}A) + 1}\right)^k \left\|e^{(0)}\right\|_A,$$

where $\|\cdot\|_A$ is the *energy norm* defined as

$$\|v\|_A = \sqrt{v^\top A v} \quad \text{for } A \text{ real, SPD}.$$

Theorem 2.3.4 Common Choices of P

- $P = \mathrm{diag}(A)$: Jacobi method.
- $P = \mathrm{lower}(A)$: Gauss-Seidel method.
- $P = \widetilde{L}\widetilde{U}$, incomplete LU factorization.

3 Finite Different for BVPs

3.1 Introduction to BVPs

Problem Set up: Suppose we have a string with fixed endpoints. There is a force adding on the string. One can write

$$\begin{cases} -\dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f(x), \quad x \in (0,1) \\ u(0) = \alpha, \dfrac{\mathrm{d}u}{\mathrm{d}x} = \beta \end{cases}$$

From ODE, we can denote $w = \frac{\mathrm{d}u}{\mathrm{d}x}$. Then, $\frac{\mathrm{d}w}{\mathrm{d}x} = f(x)$. The above problem can be written into an ODE system:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix}$$

Definition 3.1.1 (Bondary Value Problem (BVP). A *boundary-value problem (BVP)* is given by

$$\begin{cases} -\mu \dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f(x), \quad x \in (0,1),\ \mu > 0 \\ u(0) = \alpha, \quad u(1) = \beta. \end{cases} \tag{BVP}$$

Example 3.1.2 Poisson Equation

$$\begin{cases} -\left(\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2}\right) = f(x,y), \quad (x,y) \in \Omega \\ u(\text{boundary of } \Omega) = 0 \end{cases} \tag{Poisson}$$

One can further write

$$\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = \Delta u,$$

where $\Delta u = \boldsymbol{\nabla}^2 u = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$, and Δ is called the *Laplace operator*, the divergence of gradient.

3.1.3 Derive the BVP from String. Note that the energy of the string is given by

$$J(u) = \frac{1}{2} \int_0^1 \mu \left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2 \mathrm{d}x - \int_0^1 f \cdot u \, \mathrm{d}x.$$

J is called a *functional* (function of a function). The boundary condition is given by $u(0) = u(1) = 0$. In nature, things tend to minimize energy, so we want to $\min J(u)$. Let's take the gradient: suppose $\varepsilon \in \mathbb{R}$, then

$$\lim_{\varepsilon \to 0} \frac{J(u+\varepsilon v) - J(u)}{\varepsilon} = 0,$$

where v is an arbitrary function such that $v(0) = v(1) = 0$. Note that

$$\begin{aligned}
\text{Numerator} &= \frac{1}{2}\int_0^1 \mu\left(\frac{\mathrm{d}u}{\mathrm{d}x} + \varepsilon\frac{\mathrm{d}v}{\mathrm{d}x}\right)^2 \mathrm{d}x - \int_0^1 f\cdot(u+\varepsilon v)\,\mathrm{d}x - \frac{1}{2}\int_0^1 \mu\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2 \mathrm{d}x - \int_0^1 f\cdot u\,\mathrm{d}x \\
&= \cancel{\frac{1}{2}\int_0^1 \mu\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2} \mathrm{d}x + \frac{1}{2}2\varepsilon\int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x + \frac{1}{2}\varepsilon^2\int_0^1 \mu\left(\frac{\mathrm{d}v}{\mathrm{d}x}\right)^2 \mathrm{d}x \\
&\quad - \cancel{\int_0^1 f}\cdot u\,\mathrm{d}x - \varepsilon\int_0^1 f\cdot v\,\mathrm{d}x - \cancel{\frac{1}{2}\int_0^1 \mu\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)^2} \mathrm{d}x - \cancel{\int_0^1 f}\cdot u\,\mathrm{d}x \\
&= \varepsilon\int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x + \frac{1}{2}\varepsilon^2\int_0^1 \mu\left(\frac{\mathrm{d}v}{\mathrm{d}x}\right)^2 \mathrm{d}x - \varepsilon\int_0^1 f\cdot v\,\mathrm{d}x.
\end{aligned}$$

Then,

$$\frac{J(u+\varepsilon v) - J(u)}{\varepsilon} = \int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x + \frac{1}{2}\varepsilon\int_0^1 \mu\left(\frac{\mathrm{d}v}{\mathrm{d}x}\right)^2 \mathrm{d}x - \int_0^1 f\cdot v\,\mathrm{d}x.$$

So, the limit is given by

$$\lim_{\varepsilon \to 0} \frac{J(u+\varepsilon v) - J(u)}{\varepsilon} = \int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x - \int_0^1 f\cdot v\,\mathrm{d}x = 0.$$

This gives us an equilibrium solution, and

$$\int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x - \int_0^1 f\cdot v\,\mathrm{d}x = 0$$

is called *variational / weak* (we get the solution from a perturbed system).

Now, use integration by parts:

$$\int Fg = [FG] - \int fG.$$

Denote

$$\frac{\mathrm{d}u}{\mathrm{d}x} = F \quad \text{and} \quad \frac{\mathrm{d}v}{\mathrm{d}x} = g \implies \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right) = \frac{\mathrm{d}^2u}{\mathrm{d}x^2} \quad \text{and} \quad \int \frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = v.$$

So,

$$\int_0^1 \mu\frac{\mathrm{d}u}{\mathrm{d}x}\cdot\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x = \mu \underbrace{\left[\frac{\mathrm{d}u}{\mathrm{d}x}v\right]_0^1}_{=0 \text{ as } v(1)=v(0)=0} - \mu\int_0^1 \frac{\mathrm{d}^2u}{\mathrm{d}x^2}v\,\mathrm{d}x = -u\int_0^1 \frac{\mathrm{d}^2u}{\mathrm{d}x^2}v\,\mathrm{d}x.$$

So, the variational becomes

$$
\begin{aligned}
-\mu \int_0^1 \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} v \,\mathrm{d}x - \int_0^1 f \cdot v \,\mathrm{d}x = 0 \\
-\int_0^1 \left(\mu \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + f\right) \cdot v \,\mathrm{d}x = 0.
\end{aligned}
$$

We want the equation to be true $\forall\, v$, so it must be

$$
\mu \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + f = 0.
$$

That is,

$$
\begin{cases} -\mu \dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f \\ u(0) = u(1) = 0. \end{cases} \tag{BVP}
$$

Assumption: u is twice differentiable.

3.1.4 Two ways to formula a BVP.

- Find $u\ s.t.\ \forall\, v$ with $v(0) = v(1) = 0$,

  $$
  \int_0^1 \mu \frac{\mathrm{d}u}{\mathrm{d}x} \cdot \frac{\mathrm{d}v}{\mathrm{d}x} \,\mathrm{d}x = \int_0^1 f \cdot v \,\mathrm{d}x
  $$

  In this formulation, we only require u to be once differentiable. This formulation is used in *Finite Elements*

- Find $u\ s.t.$

  $$
  \begin{cases} -\mu \dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f, \quad x \in (0,1) \\ u(0) = u(1) = 0. \end{cases}
  $$

  This formulation requires u to be twice differentiable. This formulation is used for *Finite Difference*

3.2 Finite Difference

Let's use Taylor's formula to approximate $u(x_{i+1})$ and $u(x_{i-1})$:

$$
\begin{aligned}
u(x_{i+1}) = u(x_i) + \frac{\mathrm{d}u}{\mathrm{d}x}(x_i)\Delta x + \frac{1}{2}\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i)\Delta x^2 + \cdots \\
u(x_{i-1}) = u(x_i) - \frac{\mathrm{d}u}{\mathrm{d}x}(x_i)\Delta x + \frac{1}{2}\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i)\Delta x^2 + \cdots
\end{aligned}
$$

Then,

$$\begin{aligned}
u(x_{i+1}) + u(x_{i-1}) &= 2u(x_i) + \frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i)\Delta x^2 + \frac{1}{12}\frac{\mathrm{d}^4 u}{\mathrm{d}x^4}(x_i)\Delta x^4 + \mathcal{O}\big(\|\Delta x\|^4\big) \\
\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i)\Delta x^2 &= u(x_{i+1}) + u(x_{i-1}) - 2u(x_i) - \frac{1}{12}\frac{\mathrm{d}^4 u}{\mathrm{d}x^4}(x_i)\Delta x^4 + \mathcal{O}\big(\|\Delta x\|^4\big) \\
\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i) &= \frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{\Delta x^2} - \frac{1}{12}\frac{\mathrm{d}^4 u}{\mathrm{d}x^4}(x_i)\Delta x^4 + \mathcal{O}(\|\Delta x\|^2).
\end{aligned}$$

So, second order derivative approximation is

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i) \approx \frac{u(x_{i+1}) + u(x_{i-1}) - 2u(x_i)}{\Delta x^2}$$

Denote $u_i = u(x_i)$ and $f_i = f(x_i)$. Then,

$$-\mu\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_i) = -\mu\frac{u_{i+1} + u_{i-1} - 2u_i}{\Delta x^2} = f_i$$

Then, we form a linear system $Au = f$, where A is given byu

$$A = \frac{\mu}{\Delta x}\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}.$$

Claim 3.1

- $Au = f$ is solvable because A is positive definite ($x^\top Ax > 0 \quad \forall\, x \neq 0$.)
- Since A is symmetric, all eigenvalues of A is real. Further since A is positive definite, all eigenvalues are positive. So, A is nonsingular.
- $\dfrac{\lambda_{\min}}{\lambda_{\max}} \perp\!\!\!\perp \Delta x$.

> **Theorem 3.2.2 Consistency and Convergence**
> FD is consistent and convergent.

Proof 1. Note that $Au = f$ is the system we want to solve. Consider u_{ex}, the exact solution to the BVP. Then, we know, in general, $Au_{\text{ex}} \neq f$. Instead,

$$Au_{\text{ex}} = \left[\frac{\partial^2 u}{\partial x^2}(x_i)\right] + \tau_i,$$

where $\tau_i = C(x_i)\Delta x^2$. *From previously noted,*

$$C(x_i) = c\frac{\partial^4 u}{\partial x^4}(x_i).$$

So, one can write $Au_{\text{ex}} = f + \tau$.

Define $e = u_{\text{ex}} - u$. Then, $Ae = \tau \implies e = A^{-1}\tau$. So,

$$\|e\| \leq \left\|A^{-1}\tau\right\| \leq \left\|A^{-1}\right\| \cdot \|\tau\|.$$

So, to have convergence, we need

$$\left\|A^{-1}\right\| < \infty \quad \text{and} \quad \|\tau\| \to 0 \quad \text{as} \quad \Delta x \to 0.$$

As claimed before, $\dfrac{\lambda_{\text{min}}}{\lambda_{\text{max}}} \perp\!\!\!\perp \Delta x$, we know $\|A^{-1}\|$ is bounded regardless of Δx. Since $\|\tau\| \sim \Delta x^2$, $\|\tau\| \to 0$ as $\Delta x \to 0$. Then, the method is *consistent*.

Further, we have that

$$\|e\| \to 0 \quad \text{as} \quad \Delta x \to 0.$$

So, this method is *convergent*. ■

3.3 Advection-Diffussion Equation

The problem:

$$\begin{cases} \underbrace{-\mu\dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2}}_{\text{diffusion}} + \underbrace{\beta\dfrac{\mathrm{d}u}{\mathrm{d}x}}_{\text{advection}} = f \\ u(0) = u_L \\ u(1) = u_R. \end{cases} \tag{Advection-Diffusion}$$

One can think of this equation to model a particle's random walk. Based on the Guassian distribution, the particle has 50% chance to move to the left or to the right at each time point.

3.3.1 Discretization. By Taylor's Expansion:

$$u(x_{j+1}) = u(x_j) + \frac{\mathrm{d}u}{\mathrm{d}x}(x_j)\Delta x + \frac{1}{2}\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_j)\Delta x^2 - \frac{1}{6}\frac{\mathrm{d}^3 u}{\mathrm{d}x^3}(x_j)\Delta x^3 + \frac{1}{12}\frac{\mathrm{d}^4 u}{\mathrm{d}x^4}(x_j)\Delta x^4 + \mathcal{O}\left(\|\Delta x\|^4\right) \tag{1}$$

$$\begin{aligned} \frac{\mathrm{d}u}{\mathrm{d}x}(x_j)\Delta x &= u(x_{j+1}) - u(x_j) + \frac{1}{2}\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_j)\Delta x^2 \\ \frac{\mathrm{d}u}{\mathrm{d}x}(x_j) &= \frac{u_{j+1} - u_j}{\Delta x} + \frac{1}{2}\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_j)\Delta x^2 \end{aligned}$$

Can we achieve a better discretization?

$$u(x_{j-1}) = u(x_j) - \frac{\mathrm{d}u}{\mathrm{d}x}(x_j)\Delta x + \frac{1}{2}\frac{\mathrm{d}^2u}{\mathrm{d}x^2}(x_j)\Delta x^2 - \frac{1}{6}\frac{\mathrm{d}^3u}{\mathrm{d}2^3}(x_j)\Delta x^3 + \frac{1}{12}\frac{\mathrm{d}u}{\mathrm{d}x}(x_j)\Delta x^4 + \mathcal{O}\left(\|\Delta x\|^4\right) \quad (2)$$

Consider (1) − (2):

$$u(x_{j+1}) - u(x_{j-1}) = 2\frac{\mathrm{d}u}{\mathrm{d}x}(x_j)\Delta x + \frac{1}{3}\frac{\mathrm{d}^3u}{\mathrm{d}x^3}(x_j)\Delta x^3 + \mathcal{O}(\|\Delta x\|^3).$$

Then,

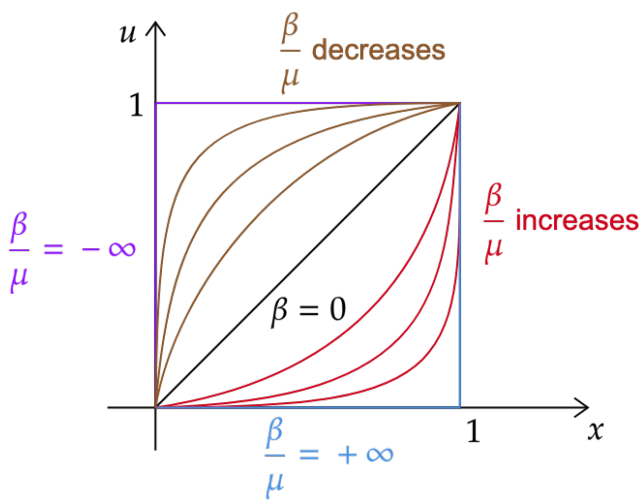
$$\frac{\mathrm{d}u}{\mathrm{d}x}(x_j) = \frac{u(x_{j+1}) - u(x_{j-1})}{2\Delta x} - \frac{1}{6}\frac{\mathrm{d}^3u}{\mathrm{d}x^3}(x_j)\Delta x^2 + \mathcal{O}\left(\frac{\|x\|^2}{2}\right).$$

So, the final numerical solution is given by

$$-\mu\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} + \beta\frac{u_{j+1} - u_{j-1}}{2\Delta x} = f_j \sim \mathcal{O}(\Delta x^2).$$

Example 3.3.2 A Specific Example

$$\begin{cases} -\mu\dfrac{\mathrm{d}^2u}{\mathrm{d}x^2} + \beta\dfrac{\mathrm{d}u}{\mathrm{d}x} = 0 \\ u(0) = 0 \\ u(1) = 1. \end{cases}$$



$$u_{\text{ex}} = \frac{e^{\frac{\beta}{\mu}x} - 1}{e^{\frac{\beta}{\mu}} - 1}.$$

If we have $\dfrac{|\beta|}{\mu} \gg 1$: convection dominated problem.

Numerical experiment shows that when $|\beta|$ is large, the numerical solution will not be consistent anymore. What's wrong?

- Mathematical explanation:

$$\mu\frac{u_{j+1}-2u_j+u_{j-1}}{\Delta x^2}+\beta\frac{u_{j+1}-u_{j-1}}{2\Delta x}=0$$
$$\left(-\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)u_{j+1}+\frac{2\mu}{\Delta x^2}u_j-\left(\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)u_{j-1}=0$$

  This is a difference equation: guess a solution $u_j=c\rho^j$. Then,

$$\left(-\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)c\rho_{j+1}+\left(\frac{2\mu}{\Delta x^2}\right)c\rho^j-\left(\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)c\rho^{j-1}=0$$
$$\left(-\frac{\mu}{\Delta x^2}+\frac{\beta}{\Delta x}\right)\rho^2+\left(\frac{2\mu}{\Delta x^2}\right)\rho-\left(\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)=0$$

  We can find ρ_1 and ρ_2 from this equation. Then,

$$u_j=c_1\rho_1+c_2\rho_2,\quad \text{a linear combination.}$$

  Note that ρ_1 and ρ_2 are solutions, so

$$\rho_1\rho_2=\frac{-\left(\frac{\mu}{\Delta x^2}+\frac{\beta}{2\Delta x}\right)}{\left(-\frac{\mu}{\Delta x^2}+\frac{\beta}{\Delta x}\right)}=\frac{1+\frac{\beta\Delta x}{2\mu}}{1-\frac{\beta\Delta x}{2\mu}}.$$

- Péclet$=\mathbb{P}_e=\dfrac{|\beta|\Delta}{2\mu}$
- If $\dfrac{|\beta|\Delta}{2\mu}>1$, $\rho_1\rho_2<0$, and then we have oscillating solutions.

3.3.3 Another Method: Upwind Method. Our previous computation relies on symmetry. However, there is a clear physical information flow. So, this problem is asymmetric in reality. We don't want as fancy as $\sim\mathcal{O}(\Delta x2^2)$ solutions, but we can use a $\sim\mathcal{O}(\Delta x)$ method:

$$\beta\frac{\partial u}{\partial x}\approx\beta\frac{u_i-u_{i-1}}{\Delta x}\qquad\text{(upwind)}$$

- Now, let's show (upwind) is *stable*:

$$\begin{aligned}\beta\frac{u_i - u_{i-1}}{\Delta x} &= \beta\frac{u_{i+1} - u_{i-1}}{2\Delta x} - \beta\frac{u_{i+1}}{2\Delta x} + \beta\frac{2u_i}{2\Delta x} \\ &= \beta\underbrace{\frac{u_{i+1} - u_{i-1}}{2\Delta x}}_{\text{central mean}} - \frac{\beta\Delta x}{2}\underbrace{\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2}}_{\text{approx. of 2nd derivative}}\end{aligned}$$

So, we can consider the equation:

$$-\underbrace{\left(\mu + \frac{|\beta|\Delta x}{2}\right)}_{\mu(1+\mathbb{P}_e)}\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial u}{\partial x} = 0.$$

Apply a central approximation:

$$-\mu\frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \beta\frac{u_{i+1} - u_{i-1}}{2\Delta x} = 0.$$

Then, upwind solution of the original problem is the central approximation of a perturbed system:

$$\text{Central (Perturbed)} = \text{Upwind (Original)}$$

Recall Péclet:

$$\mathbb{P}_e = \frac{|\beta|\Delta x}{2\mu}.$$

Then, $\mu^* = \mu(1 + \mathbb{P}_e)$. So, the Péclet of the perturbed system is

$$\mathbb{P}_e^* = \frac{|\beta|\Delta x}{2\mu^*} = \frac{|\beta|\Delta x}{2\mu(1 + \mathbb{P}_e)} = \frac{\mathbb{P}_e}{1 + \mathbb{P}_e} < 1 \quad \forall\, |\beta| \text{ and } \Delta x.$$

So, this upwind method is always stable.

- *Consistency*: when $\Delta x \to 0$, $\mu^* \to \mu$.
- *Order*: for the perturbed system, we have a 2^{nd} order approach, but with the original problem, it is only a 1^{st} order method.

3.3.4 Design a Better Method.

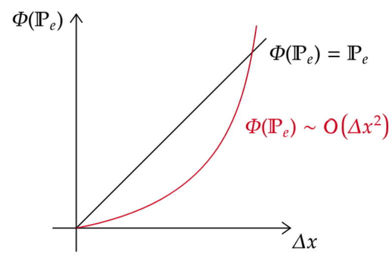
$$\mu^{\text{smart}} = \mu(1 + \Phi(\mathbb{P}_e)) \quad \text{such that}$$

- $\Phi(\mathbb{P}_e) \to 0$ as $\Delta x \to 0$.
- $\mathbb{P}_e^{\text{smart}} = \dfrac{|\beta|\Delta x}{2\mu^{\text{smart}}} < 1$.

Our upwind method takes $\Phi(\mathbb{P}_e) = \mathbb{P}_e \sim \mathcal{O}(\Delta x)$. But can we take some $\Phi(\mathbb{P}_e) \sim \mathcal{O}(\Delta x^2)$?

- We consider the *Scharfetter-Gummel Method*:

$$\Phi(\mathbb{P}_e) = \mathbb{P}_e - 1 + \underbrace{\frac{2\mathbb{P}_e}{e^{2\mathbb{P}_e} - 1}}_{\text{Bernoulli function}}$$

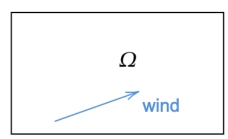


- The worst case order of Scharfetter-Gummel is $\sim \mathcal{O}(\Delta x^2)$.
- Scharfetter-Gummel is also a special $\Phi(\mathbb{P}_e)$ choice that produces exact solutions.

3.4 2-D Problem

Consider

$$\begin{cases} -\mu \Delta u + \beta \cdot \nabla u = f \\ u(\partial\Omega) = \text{data}, \end{cases}$$



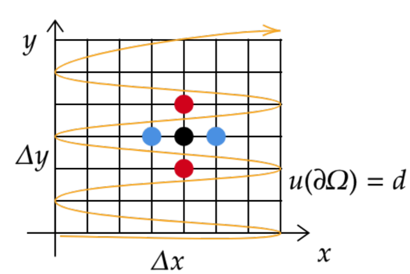
where $\partial\Omega$ is the boundary of Ω.

Write this problem out:

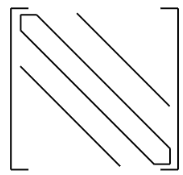
$$\begin{cases} \underbrace{-\mu\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)}_{\text{diffusion}} + \underbrace{\beta_x \frac{\partial u}{\partial x} + \beta_y \frac{\partial u}{\partial y}}_{\text{wind}} = f(x, y) \\ u(\partial\Omega) = d \end{cases}$$

3.4.1 Only consider Diffusion.

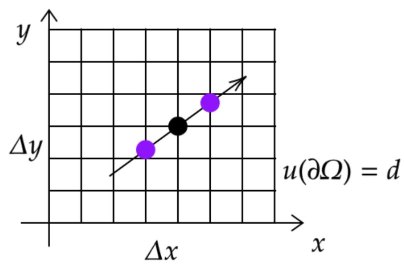
$$-\mu \frac{u_{i+i,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2} - \mu \frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2} = f(x_i, y_j)$$



To solve, we form a system: $(i,j) \to f$ such that $Au = b$, where A is SPD and takes the form of:



3.4.2 Turn on the wind.



We see that the points are not good points.

3.5 Parabolic Problems

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mu \dfrac{\partial^2 u}{\partial x^2} = f, & x \in (0,1) \text{ and } 0 < t < T \\ u(0,t) = u_L(t), \quad u(1,t) = u_R(t) \\ u(x, t=0) = u_0(x). \end{cases}$$

Discretization along x (semidiscritization): $u_j(t) = u(x_j, t)$. The equation becomes

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} - \mu \frac{u_{j+1}(t) - 2u_j(t) + u_{j-1}(t)}{\Delta x^2} = f_j(t) = f(x_j, t).$$

So, we form a system $Au = f$:

$$A = \frac{\mu}{\Delta x^2}\,\mathrm{Triad}(-1,2,1), \quad u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix}, \quad f(T) = \begin{bmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{bmatrix}.$$

Then, we have a system of ODE to solve:

$$\frac{\mathrm{d}u}{\mathrm{d}t} - Au = f.$$

We can now do time discretization and use ODE methods.

- EE/FE: $u^n = u(t^n)$. Then,

$$\begin{aligned}\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t^n} \approx \frac{u^{n+1}-u^n}{\Delta t} &= f^n + Au^n \\ u^{n+1} &= u^n + \Delta t A u^n + \Delta t f^n \\ &= (I+\Delta t A)u^n + \Delta t f^n \\ &= (I+\Delta t A)^n u_0 + \Delta t f^n.\end{aligned}$$

- IE/BE:

$$\begin{aligned}\left.\frac{\mathrm{d}u}{\mathrm{d}t}\right|_{t^n} = \frac{u^n - u^{n-1}}{\Delta t} &= f^n + Au^n \\ u^n - u^{n-1} &= \Delta t f^n + \Delta t A u^n \\ u^n - \Delta t A u^n &= \Delta t f^n + u^{n-1} \\ (I - \Delta t A)u^n &= u^{n-1} + \Delta t f^n \qquad \leftarrow \text{a linear system to solve}\end{aligned}$$

  $I - \Delta t A$ is SPD and A is time-independent. So, we may favor direct method over iterative method (as we can store $A = LU$ and reuse it).

Now, let's discuss the stability by setting $f = 0$.

- EE is conditionally stable:

  Let λ_i be eigenvalues of A. Then, we need

$$\Delta t < \frac{2}{|\lambda_i|} \quad \text{for stability.}$$

  Further, $A = \dfrac{\mu}{\Delta x^2}\,\mathrm{Triad}(1,-2,1)$, so $\rho(A) \sim \dfrac{c}{\Delta x^2}$. Then,

$$\Delta t < \frac{2}{|\lambda_i|} \leq \frac{2}{\rho(A)} = \frac{2}{c}\Delta x^2.$$

  So, if we decrease Δx by 2, to have stability,

$$\Delta t_{\text{new}} < \frac{2}{c}\left(\frac{\Delta x}{2}\right)^2 = \frac{\Delta t_{\text{old}}}{4} \implies \text{we need finer intervals for time}$$

- IE is unconditionally stable.

Definition 3.5.1 (θ Methods).

$$\frac{u^{n+1}-u^n}{\Delta t} = \theta A u^{n+1} + (1-\theta)Au^n + \theta f^{n+1} + (1-\theta)f^n, \quad \theta \in [0,1]$$

- EE: $\theta = 0$, $\sim \mathcal{O}(\Delta t)$, explicit, conditional stability
- IE: $\theta = 1$, $\sim \mathcal{O}(\Delta t)$, implicit, unconditional stability
- CN: $\theta = \frac{1}{2}$, $\sim \mathcal{O}(\Delta t^2)$, implicit, unconditional stability

To numerically solve θ methods, suppose $f = 0$. Then,

$$\begin{aligned} \frac{u^{n+1}-u^n}{\Delta t} &= \theta A u^{n+1} + (1-\theta)Au^n \\ u^{n+1} - u^n &= \Delta t \theta A u^{n+1} + \Delta t(1-\theta)Au^n \\ (I - \Delta t \theta A)u^{n+1} &= (I + \Delta t(1-\theta)A)u^n \end{aligned}$$

We essentially solve a linear system in each iteration.

Theorem 3.5.2 Stability and Order of θ Methods

- θ methods are unconditionally stable for $\theta \geq 1$. Otherwise, it is conditionally stable for $\theta < \frac{1}{2}$, and the stability condition for parabolic problem is $\Delta t < c\Delta x^2$.
- Meanwhile, the method is order 1 for $\theta \neq \frac{1}{2}$ and order 2 for $\theta = \frac{1}{2}$.

- Although the θ method is 2^{nd} order is space, the order of error is dominant and determined by the order in time.
- CN is the most vulnerable to lack of regularity and sensitive to non-smoothness.

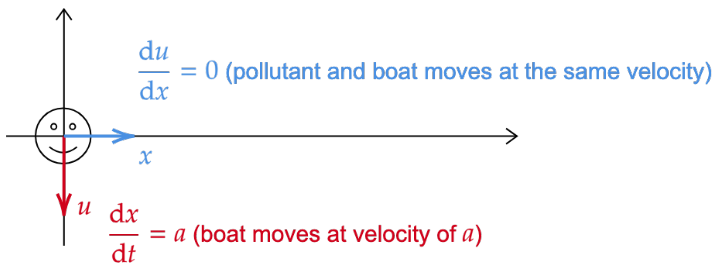
3.6 Hyperbolic Problems

$$\begin{cases} \dfrac{\partial u}{\partial t} + \alpha \dfrac{\partial u}{\partial x} = 0, & \alpha > 0 \text{ constant} \\ u(x,0) = u_0(x) \end{cases}$$

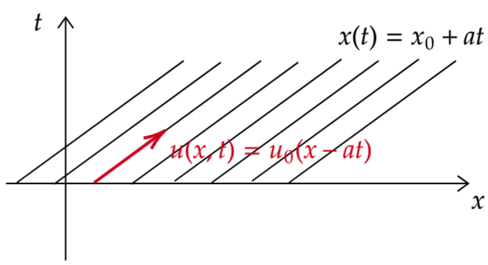
Exact solution: $u(x,t) = u_0(x - \alpha t)$.

Example 3.6.1 Modeling Density of Pollutant

u: pollutant, x: displacement of boat, t: time.



Consider the solution to $\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = a \\ x(0) = x_0. \end{cases}$ We have $x(t) = x_0 + at$. With different initial value x_0, we form different characteristic curves.



Consider $u(x(t), t)$:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} \cdot \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0.$$

3.6.2 Similar Problems.

- Conservation Law:

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

where $q(u) = v(u) \cdot u$ with $v = v_{\max}\left(1 - \frac{u}{u_{\max}}\right)$.

$$\implies \frac{\partial u}{\partial t} + \underbrace{v_{\max}\left(1 - \frac{u}{u_{\max}}\right)}_{=\text{“}a\text{”}} \frac{\partial u}{\partial x} = 0 \quad \leftarrow \text{models the density of traffic}$$

Here, a is no longer a constant.

- Heat Equation:

$$\frac{\partial^2 u}{\partial t^2} - \gamma^2 \frac{\partial^2 u}{\partial x^2} = f.$$

Define $w_1 = \dfrac{\partial u}{\partial x}$ and $w_2 = \dfrac{\partial u}{\partial t}$:

$$\begin{cases} \dfrac{\partial w_1}{\partial t} - \gamma^2 \dfrac{\partial w_2}{\partial x} = f \\ \\ \dfrac{\partial w_2}{\partial t} - \dfrac{\partial w_1}{\partial x} = 0 \qquad \left[\dfrac{\partial^2 u}{\partial x \partial t} = \dfrac{\partial^2 u}{\partial t \partial x}\right]. \end{cases}$$

Define $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ and $A = \begin{bmatrix} 0 & -\gamma^2 \\ -1 & 0 \end{bmatrix}$. Then, the original equation becomes a system

$$\frac{\partial w}{\partial t} + A \frac{\partial w}{\partial x} = 0.$$

The eigenvalues of A: $\lambda_{1,2} = \pm\gamma \implies$ Diagonalizable.

3.6.3 Find the Numerical Solution.

$$\left.\frac{\partial u}{\partial t}\right|_{t^{n+1}, u_j} = \frac{u_j^{n+1} - u_j^n}{\Delta t} \quad \text{and} \quad a \left.\frac{\partial u}{\partial x}\right|_{t^{n+1}, u_j} = \frac{a}{2} \cdot \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{\Delta t}$$

- With Backward-Euler Centered (BE-C):

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{a}{2} \cdot \frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{\Delta t} = 0$$

$$\implies \begin{bmatrix} \dfrac{1}{\Delta t} & \dfrac{a}{2\Delta t} & 0 & 0 & \cdots \\ \\ -\dfrac{a}{2\Delta t} & \dfrac{1}{\Delta t} & \dfrac{a}{2\Delta t} & 0 & \cdots \\ & & & & \ddots \end{bmatrix}.$$

- With Forward-Euler Centered (FE-C): Unconditionally unstable. NEVER USE IT!

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{a}{2} \cdot \frac{u_{j+1}^{n} - u_{j-1}^{n}}{\Delta t} = 0$$

$$\implies u_j^{n+1} = u_j^n + \frac{a\Delta t}{2\Delta t}(u_{j+1}^n - u_{j-1}^n).$$

- With Forward-Euler Upwind (FE-Upwind):

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_j^n-u_{j-1}^n}{\Delta x}=0 \quad a>0$$
$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+a\frac{u_{j+1}^n-u_j^n}{\Delta x}=0 \quad a<0$$

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{a}{2}\frac{u_{j+1}^n-u_{j-1}^n}{\Delta x}-\underbrace{\frac{|a|\Delta t}{2}\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}}_{\text{diffusion}}=0$$

- With Lax Wendroff (LW): FE-Upwind with modified coefficient

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+\frac{a}{2}\frac{u_{j+1}^n-u_{j-1}^n}{\Delta x}-\frac{a^2\Delta t}{2}\cdot\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}=0.$$

Proof 1.

$$u(x_j,t^{n+1})=u(x_j,t^n)+\left.\frac{\partial u}{\partial t}\right|_{t^n,x_j}(t^{n+1}-t^n)+\frac{1}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{t^n,x_j}(t^{n+1}-t^n)^2+\mathcal{O}\left(\left\|t^{n+1}-t^n\right\|^2\right)$$

Note that
$$\frac{\partial u}{\partial t}=-a\frac{\partial u}{\partial x},\quad \frac{\partial^2 u}{\partial x\partial y}=-a\frac{\partial^2 u}{\partial x^2},\quad \frac{\partial^2 u}{\partial x^2}=-a\frac{\partial^2 u}{\partial x\partial t}=a^2\frac{\partial^2 u}{\partial x^2}.$$

Substitute:

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

■

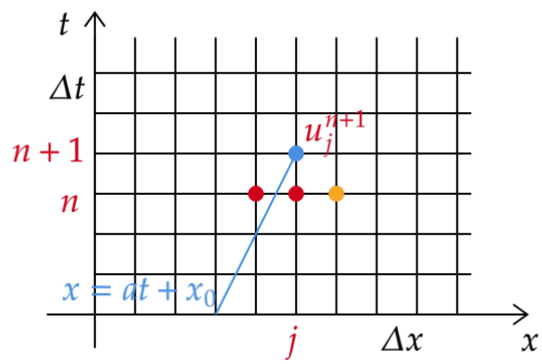
3.6.4 Consistency of Numerical Methods. τ: truncation error

- $\tau_{\text{BE-C}}\sim\mathcal{O}(\Delta t+\Delta x^2)$
- $\tau_{\text{FE-UPW}}\sim\mathcal{O}(\Delta t+\Delta x)$
- $\tau_{\text{LW}}\sim\mathcal{O}(\Delta t^2+\Delta x^2+\Delta t\Delta x)$

Theorem 3.6.5 Necessary Condition for Stability

$$\left|\frac{a\Delta t}{\Delta x}\right|=\frac{|a|\Delta t}{\Delta x}\leq 1 \qquad \text{(CFL Condition)}$$

Remark. This is also a sufficient condition for FE-UPW and LW.



- FE-UPW:

$$u_j^{n+1} = u_j^n + \frac{a}{\Delta t}(u_j^n - u_{j-1}^n)$$

- LW: u_j^{n+1} depend on u_j^n, u_{j-1}^n, and u_{j+1}^n
- Unit analysis:

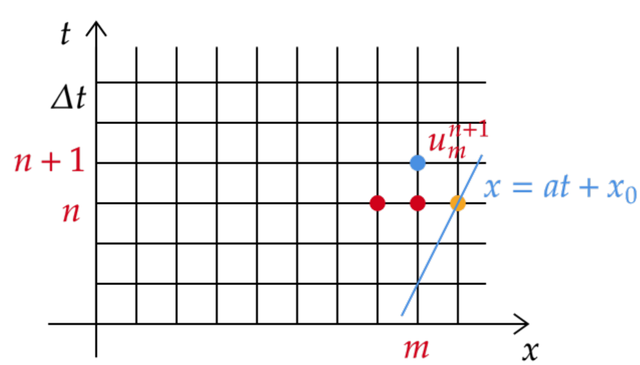
$$\frac{[u]}{[t]} = \left[[a] \cdot \frac{[u]}{[x]}\right] \implies [a] = \frac{[x]}{[t]}$$

$$\implies a \text{ is the velocity of exact solution.}$$

$$\frac{\Delta x}{\Delta t} : \text{ velocity of numerical solution}$$

  So, CFL condition: $v_{\text{exact}} \leq v_{\text{numerical}}$

- Boundary of LW: At boundary of x, we require u_{m-1}^n, u_m^n, and u_{m+1}^n to find u_m^{n+1}. However, u_{m+1}^n is out of region of interest.



What to do? We use the characteristic curves:

$$u_{m+1}^n = u_m^n + \frac{\Delta t}{\Delta x} a(u_m^n - u_{m-1}^n)$$

3.6.6 Wave/Heat Equation.

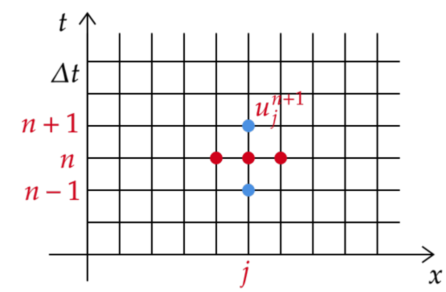
$$\frac{\partial^2 u}{\partial t^2} - \gamma^2 \frac{\partial^2 u}{\partial x^2} = 0.$$

- Form a linear system and solve using tools for conservation laws:

$$\frac{\partial w}{\partial t} + A\frac{\partial w}{\partial x} = 0.$$

$$\left(\text{Define } w_1 = \frac{\partial u}{\partial x} \quad \text{and} \quad w_2 = \frac{\partial u}{\partial t}.\right)$$

- System of first order equations: apply relevant tools.
- Wave equation Specific methods: Leapfrog Method



$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \gamma^2 \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f(x_j, t^n)$$

$$u_j^{n+1} = \Delta t^2 f_j^n + 2u_j^n - u_j^{n-1} + \frac{\gamma^2 \Delta t^2}{\Delta x^2}\left(u_{j+1}^n - 2u_j^n + u_{j-1}^n\right)$$

  - Explicit
  - Second order in time and space: $\tau \sim \mathcal{O}(\Delta t^2 + \Delta x^2)$
  - Stable under CFL condition:

$$\frac{|\gamma|\Delta t}{\Delta x} \leq 1.$$

4 Finite Elements

Motivation: Consider

$$J(u) = \frac{1}{2}\mu \int (u')^2 - \int fu, \tag{Energy}$$

where $u(0) = u(1) = 1$.

- FE: Find u $(u(0) = u(1) = 0)$ such that

$$u \int_0^1 u'v' - \int_0^1 fv = 0 \quad \forall\, v\ (v(0) = v(1) = 0),$$

  Weak as $u \in \mathcal{C}^1$ is enough.

- FD: Discretize approximation: $-\mu u'' = 0$.

  Strong and requires $u \in \mathcal{C}^2$.

4.1 Elementary Functional Analysis

Definition 4.1.1 (Space of Functions). Suppose $\mathcal{S}$ is a set of functions. $\mathcal{S}$ is a *space* of function if

- Closed under addition: $f_1, f_2 \in \mathcal{S} \implies f_1 + f_2 \in \mathcal{S}$.
- Closed under scalar multiplication: $f_1 \in \mathcal{S}$ and $\lambda \in \mathbb{R} \implies \lambda f \in \mathcal{S}$.

Definition 4.1.2 (Convergence of Functions).

- $f_n \to f \iff \lim_{n\to\infty} d(f_n, f) = 0$.
- $d(f_n, f) \to 0$ and $d(f_m, f) \to 0$ as $n, m \to \infty \implies d(f_n, f_m) \to 0$ as $n, m \to 0$.
- Cauchy sequence:

$$d(f_n, f_m) \to 0 \quad \text{as } n, m \to 0 \implies d(f_n, f) \to 0.$$

Definition 4.1.3 (Complete Space). A metric space (have distance defined) is *complete* if all sequences are Cauchy.

Definition 4.1.4 (Banach Space). A complete space with a norm defined is a *Banach space.*

Definition 4.1.5 (Hilbert Space). A Banach space with a scalar dot product defined is a *Hilbert space*.

Theorem 4.1.6 Banach Space / $\mathcal{L}^p$ / Hilbert Space
Collect all the functions on $(0,1)$ $s.t.$

$$\left|\int_0^1 f^p \, \mathrm{d}x\right| < +\infty.$$

We form a Banach space. The norm is defined as

$$\|f\|_{\mathcal{L}^p} := \left(\int_0^1 f^p \, \mathrm{d}x\right)^{1/p}.$$

This Banach space is called a $\mathcal{L}^p(0,1)$ space.
More specifically, if $p = 2$, $\mathcal{L}^2(0,1)$ is a Hilbert space. The scalar dot product is defined as

$$\langle f, g\rangle_{\mathcal{L}^2} := \int_0^1 f \cdot g \, \mathrm{d}x \implies \|f\|_{\mathcal{L}^2} = \sqrt{\int_0^1 f^2 \, \mathrm{d}x}.$$

Definition 4.1.7 (Distributional Derivative). Suppose $v \in \mathcal{C}^\infty(\mathbb{R})$ and vanishes out of an interval. Say we want to find the derivative of f, denoted as f'. Consider $f' \cdot v$:

$$\begin{aligned}\int_{\mathbb{R}} f'v \, \mathrm{d}x &= \lim_{\overline{x}\to+\infty} \int_{-\overline{x}}^{\overline{x}} f'v \, \mathrm{d}x = \lim_{\overline{x}\to+\infty} \underbrace{[f(\overline{x})v(\overline{x}) - f(-\overline{x})v(-\overline{x})]}_{=0 \text{ since } v \text{ vanishes}} - \int_{-\overline{x}}^{\overline{x}} fv' \, \mathrm{d}x \\ &= -\int_{\mathbb{R}} fv' \, \mathrm{d}x.\end{aligned}$$

So,

$$\int_{\mathbb{R}} f'v \, \mathrm{d}x = -\int_{\mathbb{R}} fv' \, \mathrm{d}x = -\int_\alpha^\beta v' \, \mathrm{d}x = -v(\beta) + v(\alpha).$$

Therefore, we define the distributional derivative as

$$f' := \int_{\mathbb{R}} f'v \, \mathrm{d}x = -v(\beta) + v(\alpha).$$

Definition 4.1.8 (Dirac-δ). The *dirac* function is defined as

$$\int_{\mathbb{R}} \delta v = v(0), \quad \text{where } v \text{ is regular enough.}$$

Meanwhile,

$$\int_{\mathbb{R}} \delta_\alpha v = v(\alpha).$$

So,

$$f' = -v(B) + v(\alpha) = -\delta_\beta + \delta_\alpha.$$

Definition 4.1.9 ($\mathcal{H}^1(0,1)$ Space). Suppose $f \in \mathcal{L}^2(0,1)$ can be differentiated using the distributional derivative. Then, the collection of f forms a space named $\mathcal{H}^1(0,1)$. $\mathcal{H}^1(0,1)$ is a Hilbert space, with

$$\begin{aligned}\langle f, g\rangle_{\mathcal{H}^1} &= \langle f, g\rangle_{\mathcal{L}^2} + \langle f', g'\rangle_{\mathcal{L}^2} \\ &= \int_0^1 fg \, \mathrm{d}x + \int_0^1 f'g' \, \mathrm{d}x.\end{aligned}$$

$\mathcal{H}^k$ space is the space of $\mathcal{L}^2$ functions with k derivatives in $\mathcal{L}^2(0,1)$.

Definition 4.1.10 ($\mathcal{H}_0^1(0,1)$). We define

$$\mathcal{H}_0^1(0,1) = \left\{ f \in \mathcal{H}^1(0,1) \mid f(0) = f(1) = 0 \right\}.$$

Remark. $\mathcal{H}_1^1(0,1)$ does not form a space.
Proof. Suppose $\mathcal{H}_1^1(0,1) = \{f \in \mathcal{H}^1(0,1) \mid f(0) = f(1) = 1\}$. Let $f, g \in \mathcal{H}_1^1(0,1)$. Then,

$$(f+g)(0) = (f+g)(1) = 2.$$

So, $f + g \notin \mathcal{H}_1^1(0,1)$, implying $\mathcal{H}_1^1$ is not a space. □

Theorem 4.1.11 Poincaré Inequality

$$\|f\|_{\mathcal{H}^1}^2 = \langle f, f\rangle_{\mathcal{H}^1} = \|f\|_{\mathcal{L}^2}^2 + \|f'\|_{\mathcal{L}^2}^2 \geq \|f\|_{\mathcal{L}^2}^2.$$

Specifrically, in $\mathcal{H}_0^1(0,1)$, $\exists$ constant $C_p > 0$ $s.t.$

$$\|f\|_{\mathcal{L}^2}^2 \leq \|f\|_{\mathcal{H}^1}^2 \leq C_p \|f'\|_{\mathcal{L}^2}^2.$$

With all the terminologies, we can rewrite (Energy) as: For

$$J = \frac{1}{2}\int_0^1 u^2 - \int fu,$$

find $u \in \mathcal{H}_0^1(0,1)$ $s.t.$

$$\int_0^1 u'v' \,\mathrm{d}x = \int_0^1 fv \,\mathrm{d}x, \quad \forall\, v \in \mathcal{H}_0^1(0,1).$$

where $f \in \mathcal{L}^2(0,1)$.

4.2 Introduction to Finite Element

Notation 4.1.

- $V := \mathcal{H}_0^1(0,1)$ is a Hilbert space.
- $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ $s.t.$ $\forall\, f, g, u, v \in V$ and $\forall\, \lambda, \mu \in \mathbb{R}$:
  - $a(\lambda f + \mu g, v) = \lambda a(f,v) + \mu a(g,v)$, and
  - $a(u, \lambda f + \mu g) = \lambda a(u,f) + \mu a(u,g)$.
- $\mathcal{F}$: a linear function on V: $\forall v_1, v_2 \in V$ and $\forall\, \lambda, \mu \in \mathbb{R}$,

$$\mathcal{F}(\lambda v_1 + \mu v_2) = \lambda \mathcal{F}(v_1) + \mu \mathcal{F}(v_2).$$

► ***General Problem for FE***

Find $u \in V$ $s.t.$

$$a(u,v) = \mathcal{F}(v) \quad \forall\, v \in V \tag{P}$$

Theorem 4.2.2 Lax-Milgram Lemma

Suppse

- $a(u,v)$ is continuous: $\forall\, u, v \in V,\ \exists\, \gamma > 0$ $s.t.$ $|a(u,v)| \le \gamma \|u\| \|v\|$,
- $\mathcal{F}(v)$ is continuous: $\forall\, v \in V,\ \exists\, M > 0$ $s.t.$ $|\mathcal{F}(v)| \le M\|v\|$, and
- $a(\cdot,\cdot)$ is coercive: $\forall\, u \in V,\ \exists\, \alpha > 0$ $s.t.$ $a(u,u) \ge \alpha \|u\|^2$.

Then, (P) is well posed. i.e., (P) is solvable and the solution is unique.

Remark.

- $|a(u,v)| \le \mu \|u'\|_{\mathcal{L}^2}\|v\|_{\mathcal{L}^2} \le \underbrace{\mu}_{=\gamma} \|u\|_{\mathcal{H}^1}\|v\|_{\mathcal{H}}$.
- $|\mathcal{F}(v)| \le \|f\|_{\mathcal{L}^2}\|v\|_{\mathcal{L}^2} \le \underbrace{\|f\|_{\mathcal{L}^2}}_{=M} \|v\|_{\mathcal{H}^1}$.
- $a(u,u) = \mu \int_0^1 (u')^2 = \mu\|u'\|^2_{\mathcal{L}^2} \ge \underbrace{\frac{\mu}{C_p}}_{\alpha} \|u\|^2_{\mathcal{H}^1}$, where $\|u\|^2_{\mathcal{H}^1} \le C_p\|u'\|^2_{\mathcal{L}^2}$.

Claim 4.3 The problem

$$\begin{cases} \mu u'' + \beta u' + \sigma u & = f \quad \sigma > 0 \\ -\mu u'' & = f \quad x \in (0,1) \\ u(0) = u(1) = 0 \end{cases}$$

can be written as

$$\underbrace{-\int_0^1 \mu u'' v + \int_0^1 \beta u' v + \int_0^1 \sigma u v}_{a(u,v)} = \underbrace{\int_0^1 f v}_{\mathcal{F}(v)} .$$

This problem satisfies Lax-Milgram conditon.

Proof 1.

- $a(u,v)$ is continuous:

$$\left|\beta \int_0^1 u'v\right| \le |\beta| \|u'\|_{\mathcal{L}^2}\|v\|_{\mathcal{L}^2} \le |\beta| \|u'\|_{\mathcal{H}^1}\|v\|_{\mathcal{H}^1}.$$

$$\beta \int_0^1 u'u = \frac{\beta}{2}\int_0^1 \frac{\mathrm{d}u^2}{\mathrm{d}x} = \frac{\beta}{2}\big(u^2(1) - u^2(0)\big) = 0.$$

$$\sigma \int u^2 = \sigma \|u\|^2_{\mathcal{L}^2}.$$

- $\mathcal{F}(v)$ is continuous.
- $a(u,u)$ is coercive:

$$a(u,u) \ge \mu C_p \|u\|^2_{\mathcal{H}^1} + \sigma\|u\|^2_{\mathcal{L}^2} \ge \mu C_p \|u\|^2_{\mathcal{H}^1}.$$

■

4.3 Galerkin Method

Find $u \in V \; s.t. \; a(u, v) = \mathcal{F}(u) \quad \forall \, v \in V$. We write the numerical problem as

$$P_N : \text{ Find } v_N \in V_N \; s.t. \; a(u_N, v_N) = \mathcal{F}(v_N) \quad \forall \, v_N \in V_N \subset V.$$

- P_N satisfies Lax-Milgram condition, and thus is well-posed.
- If u is the exact solution to the original problem, then u is also an exact solution for P_N:

$$a(u, v_N) = \mathcal{F}(v_N) \quad \forall \, v \in V_N.$$

  In other words, P_N is *strongly consistent* and truncation error $\tau = 0$.

- Convergence: Suppose

$$a(u_N, v_N) = \mathcal{F}(v_N) \quad \text{and} \quad a(u, v_N) = \mathcal{F}(v_N).$$

  What is $\|u - u_N\|_{\mathcal{H}^1}$ as $N \to \infty$?

$$\begin{aligned} \alpha \|u - u_N\|_{\mathcal{H}^1}^2 &\leq a(u - u_N, u - u_N) \\ &= a(u - u_N, u - w_N + w_N - u_N) \\ &= a(u - u_N, u - w_N) + a(u - u_N, w_N - u_N) && \text{[Bilinearity]} \end{aligned}$$

  Since u and u_N are exact for v_N. So, by strong consistency,

$$a(u, v_N) = \mathcal{F}(v_N) \quad \text{and} \quad a(u_N, v_N) = \mathcal{F}(v_N).$$

  Therefore,

$$\begin{aligned} a(u - u_N, v_N) &= a(u, v_N) - a(u_N, v_N) \\ &= \mathcal{F}(v_N) - \mathcal{F}(v_N) \\ &= 0. \end{aligned}$$

  Then,

$$\begin{aligned} a(u - u_N, u - u_N) &= a(u - u_N, u - w_N) + \underbrace{a(u - u_N, w_N - u_N)}_{=0} \\ &= a(u - u_N, u - w_N) \\ &\leq \gamma \|u - u_N\|_{\mathcal{H}^1} \cdot \|u - w_N\|_{\mathcal{H}^1}. \end{aligned}$$

We have

$$\alpha \|u - u_N\|_{\mathcal{H}^1}^{\cancel{2}} \leq \gamma \cancel{\|u - u_N\|_{\mathcal{H}^1}} \cdot \|u - w_N\|_{\mathcal{H}^1}$$
$$\|u - u_N\|_{\mathcal{H}^1} \leq \frac{\gamma}{\alpha} \|u - w_N\|_{\mathcal{H}^1}.$$

Lemma 4.1 Cea Lemma: We have

$$\|u - u_N\|_{\mathcal{H}^1} \leq \frac{\gamma}{\alpha} \inf_{w_N \in V_N} \|u - w_N\|_{\mathcal{H}^1}.$$

When $N \to \infty$, we have $\inf\limits_{w_N \in V_N} \|u - w_N\|_{\mathcal{H}^1} \to 0$. Then,

$$\|u - u_N\|_{\mathcal{H}^1} \to 0 \quad \text{as well.}$$

Remark 1. (Implication of Cea Lemma). The Galerkin solution u_N might not be the best solution w_N. However, it converges to exact solution u at the same rate as w_N.

- How to find u_N? *Interpolation with Piecewise Polynomials*

$$V_N \equiv \left\{ \text{functions} \mid {\scriptstyle \begin{array}{c} \text{continuous on a set of given intervals} \\ \text{polynomial of order 1 (linear functions)} \end{array}} \right\}.$$

We use *Lagrange polynomials*: piecewise linear polynomials $\varphi_j(x)$ $s.t.$

$$\varphi_j(x_i) = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases}$$

and

$$v_N(x) = \sum_j c_j \varphi_j(x_i) \quad \text{where } c_j = v_j.$$

So, the numerical solution is

$$u_N = \sum_j u_j \varphi_j(x).$$

Plug-in $a(u_N, v_N) = \mathcal{F}(v_N)$:

$$\sum_{j=1}^{N} u_j a(\varphi_j, v_N) = \mathcal{F}(v_N).$$

What is v_N? Try φ_i's:

$$v_N = \sum_i c_i \varphi_i.$$

Then,

$$\sum_{i=1}^{N} c_i \sum_{j=1}^{N} \underbrace{u_j}_{u_j} \underbrace{A(\varphi_j, \varphi_i)}_{A_{i,j}} = \underbrace{\mathcal{F}(\varphi_i)}_{b_i}.$$

So, we can form a linear system to solve: $\boxed{Au = b}$.

Example 4.3.2 Poisson Problem

$$u \int_0^1 u'v' = \int_0^1 fv$$

$$a(\varphi_j, \varphi_i) = \mu \int_0^1 \varphi_j' \varphi_i'$$

Note: we don't need to integrate for every combinations of i and j. For example, when $\text{support}(\varphi_2) \cap \text{support}(\varphi_7) = \varnothing \implies$ no need to compute the integral.

Therefore, the matrix A is *tridiagonal.*

4.3.1 Nonhomogenous Condition

$$\begin{cases} -\mu u'' + \beta u' + \sigma u & = f \\ x \in (0,1). \end{cases}$$

- Under non-homogeneous condition, FE will not work because

$$\mathcal{H}^1_{\text{non-hom}} = \left\{ f \in \mathcal{H}^1(0,1) : u(0) = 1,\ u(1) = 2 \right\}$$

does not form a space.

- What to do instead?

$$u(x) = \mathring{u}(x) + \ell(x), \quad \ell(0) = 1 \text{ and } \ell(1) = 2.$$

where $\ell(x)$ is a lifting function. Then, we need to find $\mathring{u} \in \mathcal{H}^1_0(0,1)$ $s.t.$

$$\mu \int_0^1 \mathring{u}' v' + \beta \int_0^1 \mathring{u}' v + \sigma \int_0^1 \mathring{u} v = \underbrace{\int_0^1 fv - \mu \int_0^1 \ell' v' - \beta \int_0^1 \ell' v - \sigma \int_0^1 \ell v}_{\mathcal{F}(v)}$$

- Another example: $u(0) = 0$ and $u'(1) = 0$. Define

$$V = \left\{ f \in \mathcal{H}^1(0,1)\ s.t.\ f(0) = 0 \right\} \equiv \mathcal{H}^1_D(0,1).$$

With FE:

$$-\mu \int_0^1 u''v + \beta \int_0^1 u'v + \sigma \int_0^1 uv = \int_0^1 fv.$$

Apply integration by parts:

$$\underbrace{\mu\Big[u'v\Big]_0^1}_{=-\mu(u'(1)v(1)-u'(0)v(0)} + \mu \int_0^1 u'v' + \beta \int_0^1 u'v + \sigma \int_0^1 uv = \int_0^1 fv$$

$$\mu \int_0^1 u'v' + \beta \int_0^1 u'v + \sigma \int_0^1 uv = \int_0^1 fv.$$

So, the problem looks the same, and the only difference is the space we search.

- $u(0) = 0$ and $u'(1) = d$. Then,

$$\mu \int_0^1 u'v' + \beta \int_0^1 u'v + \sigma \int_0^1 uv = \underbrace{\int_0^1 fv + \mu v(1)d}_{\text{New } \mathcal{F}(v)}$$

- $u(0) = 0$ and $u'(1) + u(1) = d$.

$$\underbrace{\mu\Big[u'v\Big]_0^1} + \mu \int_0^1 u'v' + \beta \int_0^1 u'v + \sigma \int_0^1 uv = \int_0^1 fv.$$

Note that

$$-\mu(u'(1)v(1) - u'(0)v(0)) = \mu d v(1) + \mu u(1)v(1) \qquad [\text{plug in } u'(1) = d - u(1)]$$

So,

$$\underbrace{\mu \int_0^1 u'v' + \beta \int_0^1 u'v + \sigma \int_0^1 uv + \mu u(1)v(1)}_{\text{New } a(u,v)} = \underbrace{\int_0^1 fv + \mu d v(1)}_{\text{New } \mathcal{F}(v)}.$$

4.3.2 Notes on Code Implementation

- Node-wise (Physical Element):

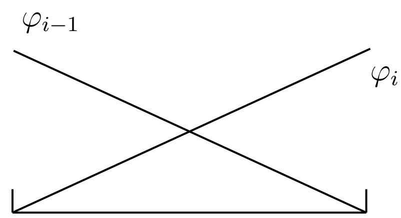
  For each note, we compute:

$$\int_{x_{i-1}}^{x_i} \varphi'_{i-1}\varphi_i$$
$$\int_{x_{i-1}}^{x_{i+1}} (\varphi''_i)^2 = \int_{x_{i-1}}^{x_i} (\varphi''_i)^2 - \int_{x_i}^{x_{i+1}} (\varphi''_i)^2$$
$$\int_{x_i}^{x_{i+1}} \varphi'_{i+1}\varphi_i$$

- Element wise (Reference Element):

  On one sub-interval:

$$\begin{bmatrix} a(\varphi_{i-1},\varphi_{i-1}) & a(\varphi_{i-1},\varphi_i) \\ a(\varphi_i,\varphi_{i-1}) & a(\varphi_i,\varphi_i) \end{bmatrix}$$



  We can further map the interval $[x_i, x_{i+1}]$ to $[0,1]$ by setting $\xi = \dfrac{x - x_i}{x_{i+1} - x_i}$. Then,

$$\widehat{\varphi}_0(\xi) = 1 - \xi \quad \text{and} \quad \widehat{\varphi}_1(\xi) = \xi.$$

  Meanwhile, we have $x = x_i + \xi(x_{i+1} - x_i)$, so we can move back-and-forth.

- Computing integral: quadrature rule:

$$\int_a^b f \approx \sum_j w_j f(x_j)$$

- φ_j can be other types of functions. For example, piecewise quadratic. Then, on each interval, we need 3 points to interpolate a quadratic function.

$$u(x) = \sum_j u_j \varphi_j(x),$$

  where $\varphi_j(x)$ is composed of midpoint quadratic function and node function.

  Generalization: $X_h^r := \left\{V_h \in \mathcal{C}^0\left(\overline{\Omega}\right) : V_h|_{k_j} \in \mathbb{P}_r \quad \forall\, k_j \in T_h\right\}$, where h is the level of discretization, $\mathbb{P}_r$ is the set of polynomials with degree r, and T_h is the triangulation/mesh.

Definition 4.3.3 (Interpolant). The interpolant of v in the space X_h^r is the function $\Pi_h^r(V)$ $s.t.$

$$\Pi_h^r(v(x_i)) = v(x_i) \quad \forall\, x_i \text{ node of partition } T_h.$$

Theorem 4.3.4

Let $v \in \mathcal{H}^{r+1}(I)$ with $r \geq 1$, and let $\Pi_h^r(v) \in X_h^r$. Then, the following estimates hold

$$\|v - \Pi_h^r(v)\|_{\mathcal{H}^k(I)} \leq C_{k,r} h^{r+1-k} \|v\|_{\mathcal{H}^{r+1}(I)} \qquad \text{for } k = 0, 1.$$

Theorem 4.3.5

Let $u \in V$ be the exact solution of the variational problem via the finite element approximation of order r, where $V_h = X_h^r \cap V$. Moreover, let $u \in \mathcal{H}^{p+1}(I)$ for $r \leq p$. Then, we have a priori estimate

$$\|u - u_h\|_V \leq \frac{M}{\alpha} C h^r \|u\|_{\mathcal{H}^{r+1}(I)},$$

where the constant $\dfrac{M}{\alpha}$ comes from Cea Lemma.

Remark 2. (Implication of Theorem 4.3.5). Increasing r too much will not help us gain faster speed on convergence.

r	$u \in \mathcal{H}^1$	$u \in \mathcal{H}^2$	$u \in \mathcal{H}^3$	$u \in \mathcal{H}^4$
1	convergence	$\boxed{h}$	h	h
2	convergence	h	$\boxed{h^2}$	h^2
3	convergence	h	h^2	$\boxed{h^3}$
4	convergence	h	h^2	h^3

So, $\|u - u_h\|_{\mathcal{H}^1} \leq C h^s \|u\|_{\mathcal{H}^{s+1}}$, where $s = \min\{r, p\}$.

Example 4.3.6

Consider the problem

$$-u'' = f \quad x \in (0, 1).$$

The exact solution is given by

$$u_{\text{ex}} = \begin{cases} \sin\left(\pi\left(x - \dfrac{1}{3}\right)\right), & x \leq \dfrac{1}{3} \\ 1 - \cos\left(\pi\left(x - \dfrac{1}{3}\right)\right) + \pi\left(x - \dfrac{1}{3}\right). & \end{cases} \tag{S}$$

- Recall: $u_{\text{ex}} \in \mathcal{H}^{s+1}(0,1)$. Let u_h be the solution of FE in $\mathbb{P}^q$. The accuracy is summarized as

	$s=1$	$s=2$	$s=3$
$q=1$	$\boxed{1}$	1	1
$q=2$	1	$\boxed{2}$	2
$q=3$	1	2	$\boxed{3}$

We know that the boxed denotes the optimal selection, and

$$\|u_{\text{ex}} - u_h\| \le Ch^{\min\{s,q\}}.$$

- Question: what is the space of (S)?
  1. (S) is continuous
  2. First derivative is also continuous.
     Second derivative is not continuous but $\in \mathcal{L}^2(0,1)$.
     Third derivative is not in $\mathcal{L}^2(0,1)$.
  3. So, $u_{\text{ex}} \in \mathcal{H}^2(0,1)$.

  Hence, $s=1$. Regardless of the degree of FE we use, the order of convergence should be only *linear*.

4.4 Advection Diffusion and Reaction in 1D

4.4.1 Advection Diffusion

$$-\mu u'' + \beta u' = f \qquad \mu > 0,\ \mu \in \mathbb{R}^+,\ \beta \in \mathbb{R}.$$

- With FD:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \beta \frac{u_{i+1} - u_{i-1}}{2\Delta x} = f_i \tag{FD}$$

If $f=0$, $u(0)=0$, and $u(1)=1$, we get that

$$u_{\text{ex}} = \frac{e^{(\beta/\mu)x} - 1}{e^{(\beta/\mu)} - 1}.$$

We also know (FD) is table when $\mathbb{P}_e = \dfrac{|\beta|\Delta x}{2\mu} > 1$.

We can also consider the upwind scheme to make (FD) stable regardless of $\mathbb{P}_e$:

$$\beta u' \approx \begin{cases} \beta \dfrac{u_i - u_{i-1}}{\Delta x}, & \beta > 0 \\ \beta \dfrac{u_{i+1} - u_i}{\Delta x}, & \beta < 0. \end{cases}$$

- With Linear FEM: the formulation is

$$-\mu \cancel{\left[u'v\right]_0^1} + \mu \int_0^1 u'v' + \int_0^1 \beta u'v = \int fv.$$

With $u_h = \sum_j u_j \varphi_j(x)$, where φ_j is linear, we get

$$\int_0^1 u'v' = \mu \underbrace{\int_0^1 \varphi_j' \cdot \varphi_i'}_{\text{constant}} + \beta \underbrace{\int_0^1 \varphi_j' \varphi_i}_{\text{linear}}$$

The FEM equation is

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x} + \beta \frac{u_{i+1} - u_{i-1}}{2} = 0 \tag{FEM}$$

Note that

$$\frac{1}{\Delta x}(\text{FEM}) = (\text{FD}).$$

So, FEM is also suffering from oscillations, and we require $\mathbb{P}_e < 1$.

- FEM with upwind scheme:

Change μ to $\mu(1 + \mathbb{P}_e)$. Or, in general, the Scharfetter-Gummel (SG) Method:

$$\mu^* = \mu(1 + \Phi(\mathbb{P}_e)).$$

Then,

$$\mathbb{P}_{\text{upw}} = \frac{|\beta|\Delta x}{2\mu_{\text{upw}}} = \frac{|\beta|\Delta x}{2\mu(1 + \mathbb{P}_e)} = \frac{\mathbb{P}_e}{1 + \mathbb{P}_e} < 1 \quad \forall\, \Delta x.$$

4.4.2 Advection Reation

$$-\mu'' + \sigma u = f, \qquad f \in \mathcal{L}^2(0,1),\ \sigma > 0.$$

- With FD:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x^2} + \sigma u_i = f(x_i).$$

Form a system:

$$A_d + \sigma I = f.$$

1. If $\sigma = 0$: only diffusion
2. $\lambda(A_d)$, $\rho(A_d)$ $\perp\!\!\!\perp$ of Δx
3. $\lambda(A_d + \sigma I) = \lambda(A_d) + \sigma$, $\perp\!\!\!\perp$ of $\Delta x \implies$ no oscilations.

- Linear FEM:

$$-\mu \frac{u_{i+1} - 2u_i + u_{i-1}}{\Delta x} + \frac{\sigma \Delta x}{6}(u_{i+1} + 4u_i + u_{i-1}).$$

1. We can have instability: The condition is

$$\mathbb{P}_e = \frac{\sigma \Delta x^2}{6\mu} < 1.$$

   we need to enforce the roots of the characteristic polynomials to be > 0.

2. Compare with AD:

AD	AR
] $\mathbb{P}_e = \frac{\vert\beta\vert \Delta x}{2\mu} < 1$	$\mathbb{P}_e = \frac{\sigma \Delta x^2}{6\mu} < 1$
$\Delta x < \frac{2\mu}{\vert\beta\vert}$	$\Delta x < \sqrt{\frac{6\mu}{\sigma}}$

   Suppose $\frac{\mu}{|\beta|}, \frac{\mu}{\sigma} \sim \mathcal{O}(10^{-6})$. Then, $\Delta x_{\text{AD}} < \mathcal{O}(10^{-6})$ is hard to achieve. However, $\Delta x_{\text{AR}} < \mathcal{O}(10^{-3})$ is easier.

3. Can we avoid this condition? We can do so by using trapezoidal rule.

$$\sigma \int_0^1 \varphi_i \varphi_j \, \mathrm{d}x = \begin{cases} 0, & j \neq 0, i \pm 1 \\ \frac{\sigma}{6}\Delta x, & j = i \pm 1 \\ \\ \frac{2\sigma}{3}\Delta x, & j = i \end{cases}$$

   If we compute this integral with trapezoidal rule:

$$(T) \int_a^b f \approx \frac{f(a) + f(b)}{2}(b - a) \qquad \text{(Trapezoidal)}$$

Then,

$$(T)\int_0^1 \varphi_i\varphi_j = \begin{cases} 0, & j \neq i, i \pm 1 \\ 0, & j = i \pm 1 \\ \Delta x, & j = i. \end{cases}$$

So,

$$\sigma(T)\int_0^1 \varphi_i\varphi_j = \begin{cases} 0, & i \neq j \\ \sigma\Delta x, & i = j \end{cases} \implies \sigma I \text{ matrix representation}$$

Then, the FE formula becomes

$$\begin{aligned} & -\mu\frac{u_{i+1} - 2u_i + u_{i+1}}{\Delta x} + \sigma u_i \Delta x = f_i \\ \implies \ & \Delta x \underbrace{\left(-\mu\frac{u_{i+1} - 2u_i + u_{i+1}}{\Delta x^2} + \sigma u_i\right)}_{\text{FD formula, stable}} = f_i. \end{aligned}$$

This procedure is called *Mass Lumping*.

– Mass matrix:

$$(T)\int_0^1 \varphi_i\varphi_j$$

– Lumping:
  Original approximation is given by

$$\frac{\sigma}{6}(u_{i+1} + 4u_i + u_{i-1})\Delta x$$

  When moving u_{i+1} and u_{i-1} to u_i, we get

$$\frac{\sigma}{6}(6u_i)\Delta x = \sigma u_i \Delta x.$$

Mass lumping stabilizes the FE solution for AR problem.

4.4.3 Generalization

- Recall:

  Exact problem: Find $u \in V \ s.t. \ a(u, v) = \mathcal{F}(v) \quad \forall\, v \in V$.

  Numerical problem: Find $u_h \in V_h \ s.t. \ a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall\, v_h \in V_h$.

- What happens if we do upwind or mass lumping?

A modification to the numerical problem:

$$\text{Find } u_h \in V_h \; s.t. \; a_h(u_h, v_h) = \mathcal{F}_h(v_h) \quad \forall\, v_h \in V_h,$$

where

1. upwind:

$$a_h(u_h, v_h) = a(u_h, v_h) + \frac{|\beta| h}{2\mu} \int_0^1 u_h' v_h'$$

2. mass lumping:

$$\begin{aligned} a_h(u_h, v_j) &= (T) \int_0^1 \mu u_h' v_h' + (T) \int_0^1 \beta u_h' v_h + (T) \int_0^1 u_h v_h \\ &= a(u_h, v_h) + \underbrace{(T) \int_0^1 - \int_0^1}_{\text{integration error}} \end{aligned}$$

This is called the *generalized Galerkin shceme*.

- Under generalized Galerkin, we don't have strong consistency anymore:

$$a_h(u - u_h, v_h) \neq 0.$$

$$\begin{cases} a(u, v_h) = \mathcal{F}(v_h) \\ a_h(u_h, v_h) = \mathcal{F}_h(v_h). \end{cases}$$

$$\implies a_h(u_h, v_h) = a(u_h, v_h) + \delta(u_h, v_h),$$

where $\delta(u_h, v_h) = \delta_{\mathcal{F}}(v_h)$.

- For Galerkin method: we have *Cea Lemma*

$$\|u - u_h\|_{\mathcal{H}^1} \leq C \inf_{w_h \in V_h} \|u - w_h\|.$$

- For generalized Galerkin method: we have *Strang Lemma*:

$$\begin{aligned} \|u - u_h\|_{\mathcal{H}^1} \leq C_1 &\inf_{w_h \in V_h} \|u - w_h\| && \text{[form Cea]} \\ &+ C_2 \inf_{w_h \in V_h} \sup_{v_h \in V_h} |a_h(w_h, v_h) - a(w_h, v_h)| \\ &+ C_3 \sup_{v_h \in V_h} |\mathcal{F}_h(v_h) - \mathcal{F}(v_h)| \end{aligned}$$

- For upwind:
$$\mathcal{O}(h^q) + \mathcal{O}(h) + 0,$$
where $q = \min\{s, p\}$. This implies that regardless what s and p we have, the upwind will only produce a convergence rate of linear.

- For SG: $\mathcal{O}(h^2)$

- For mass lumping:
$$\mathcal{O}(h^q) + \mathcal{O}(h^2) + \mathcal{O}(h^2).$$

4.5 2D Problems

4.5.1 Poisson Problem in 2D

$$\begin{cases} -\mu\Delta u = f \\ u(\partial\Omega) = u_D \end{cases}$$

- Weak formulation:

  1. Green's Formula:
$$\int_\Omega \boldsymbol{\nabla} u \cdot w = \int_{\partial\Omega} w\mu u - \int_\Omega \boldsymbol{\nabla} w \cdot u$$
$$\int_\Omega \boldsymbol{\nabla} w \cdot u = \int_{\partial\Omega} w \cdot \mu u - \int_\Omega \boldsymbol{\nabla} u \cdot w.$$

  μ is normal to $\partial\Omega$, a standard unit vector. We further have
$$\boldsymbol{\nabla} \cdot w = \frac{\partial w_0}{\partial x} + \frac{\partial w_1}{\partial y} + \frac{\partial w_2}{\partial z}$$
$$= \sum_{i=0}^{2} \frac{\partial w_i}{\partial x_i}.$$

  So,
$$-\mu \int_\Omega \overbrace{\Delta u}^{\boldsymbol{\nabla} w} \cdot v \,\mathrm{d}w = \int_\Omega fv \qquad \Delta u = \boldsymbol{\nabla} \cdot (\underbrace{\boldsymbol{\nabla} u}_{w})$$
$$\underbrace{-\mu \int_{\partial\Omega} \boldsymbol{\nabla} u \cdot uv}_{v(\partial\Omega)=0} + \mu \int_\Omega \overbrace{\boldsymbol{\nabla} u}^{w} \cdot \boldsymbol{\nabla} v = \int_\Omega fv \qquad \forall\, v \in \mathcal{H}_0^1(\Omega).$$
$$\mu \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v = \int_\Omega fv.$$

- FE: Suppose $V_h \subset V$. Find $u_h \in V_h$ $s.t.$

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall\, v_h \in V_h,$$

where

$$a(u_h, v_h) = \mu \int_\Omega \nabla u \cdot \nabla v \quad \text{and} \quad \mathcal{F}(v_h) = \int_\Omega f v.$$

1. FEM in $\mathbb{P}^1$: u_h is a piecewise linear function in Ω.

   Lemma *If a function is $\mathcal{C}^0(\Omega)$, then it is $\mathcal{H}^1(\Omega) \equiv V$.*

   Assumption, we have no handing nodes (a node that is both an interior of some lines and the vertex of the others) or overlapping triangles.
   On each T_k, u_h is linear:

   $$u_h = a_k x_0 + b_k x_1 + c_k.$$

   Each u_j is determined by the three vertices, and the continuity is for free.

   $$u_h(x_0, x_1) = \sum c_j \varphi_j(x_0, x_1), \quad \text{where } \varphi_j(x_0, x_1) = \begin{cases} 1, & (x_0, x_1) \in p_j \\ 0, & \text{o/w.} \end{cases}$$

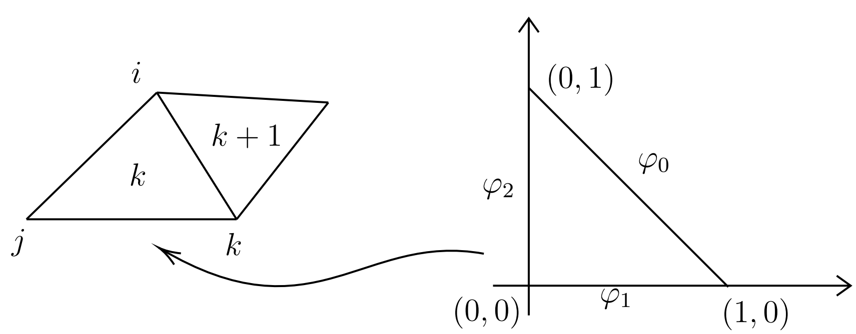
   So,

   $$u_h(x_0, x_1) = \sum u_j \varphi_j(x_0, x_1).$$

   Then, the FEM discretized problem is

   $$\sum u_j a(\varphi_i, \varphi_j) = \mathcal{F}(\varphi_j)$$
   $$\implies Au = b$$

   $\star$ Loop over elements: Reference element



$$\begin{cases} \varphi_0 = 1 - \widehat{x} - \widehat{y} \\ \varphi_1 = \widehat{x} \\ \varphi_2 = \widehat{y} \end{cases} \xrightarrow{\text{transform to physical elements}} \begin{cases} x_0(0,0) \to x_0(i \\ x_0(1,0) \to x_0(j) \\ x_0(0,1) \to x_0(k). \end{cases}$$

The mapping:

$$x_0(\widehat{x}, \widehat{y}) = x_0(i)\widehat{\varphi}_0(\widehat{x}, \widehat{y}) + x_0(j)\widehat{\varphi}_1(\widehat{x}, \widehat{y}) + x_0(j)\widehat{\varphi}_2(\widehat{x}, \widehat{y}).$$

Change of variable:

$$\boldsymbol{\nabla}_{x_0,x_1} = J^{-1}\boldsymbol{\nabla}\widehat{x}, \widehat{y}$$

Then,

$$\int_{T_h} \boldsymbol{\nabla}\varphi_j \boldsymbol{\nabla}\varphi_i \,\mathrm{d}(x_0, x_1) = \int_{\widehat{T}} J^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\varphi_\alpha J^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\varphi_\beta |J|\, d(\widehat{x}, \widehat{y}),$$

where $\alpha, \beta = 0, 1, 2$. So, the submatrix to add is 3×3.

4.5.2 Advection Diffusion in Multidimension

We want to model polutant concentration:

$$-\mu\Delta u + \beta \cdot \boldsymbol{\nabla} u + \sigma u = f,$$

where if μ depends on u, $\mu = -\boldsymbol{\nabla} \cdot (\mu \cdot \boldsymbol{\nabla} u)$, β models for wind, σ models biological consumption. The initial condition is given by $u(\Gamma_D) = \text{data}_D$. The Péclet is

$$\mathbb{P}_e = \frac{\|\beta\|h}{2\mu} < 1.$$

- With upwind method: $\mu \to \mu^* = \mu(1 + \mathbb{P}_e)$. We can compute

$$\mathbb{P}_e^* = \frac{\|\beta\|h}{2\mu^*} = \frac{\|\beta\|h}{2\mu(1 + \mathbb{P}_e)} = \frac{\mathbb{P}_e}{1 + \mathbb{P}_e} < 1 \quad \forall\, h.$$

$$\mu^* = \mu\left(1 + \frac{\|\beta\|h}{2\mu}\right).$$

- If the wind is only along x:

$$-\mu^* \frac{\partial^2 u}{\partial x^2} - \mu^* \frac{\partial^2 u}{\partial y^2} \quad \text{is a bad implementation}$$

Here, the second μ^* related to y is not helping at all. It affects accuracy. So, we consider the following method

$$-\mu^* \frac{\partial^2 u}{\partial x^2} - \mu \frac{\partial^2 u}{\partial y^2},$$

which is a better practical implementation.

- Generally: Streamline Diffusion.

$$-\mu\Delta u + \beta\nabla u + \sigma u = \frac{h}{2}\nabla\cdot\left((\beta\cdot\nabla u)\frac{\beta}{\|\beta\|}\right) = f.$$

Weak formulation:

$$\mu\int_\Omega \nabla u\cdot\nabla v + \int_\Omega \beta\nabla u\cdot v + \int_\Omega \sigma uv + \underbrace{\frac{h}{2}\int_\Omega (\beta\cdot\nabla u)(\beta\cdot\nabla v)\frac{1}{\|\beta\|}}_{\text{normalizing along } \beta\text{, direction of wind}} = \int_\Omega fv.$$

Theorem 4.5.1 Strang Lemma

For generalized Galerkin method, we have consistency in the following way:

$$\begin{aligned}\|u-u_h\|_{\mathcal{H}^1} \leq C_1 &\inf_{w_h\in V_h}\|u-w_h\| && \text{[form Cea]}\\ &+ C_2\inf_{w_h\in V_h}\sup_{v_h\in V_h}|a_h(w_h,v_h)-a(w_h,v_h)|\\ &+ C_3\sup_{v_h\in V_h}|\mathcal{F}_h(v_h)-\mathcal{F}(v_h)|\end{aligned}$$

Theorem 4.5.2 Strong Consistent Methods (Thomas Jr. Hughes)

$$\underbrace{a(u,v)+\ell_h(u,v)}_{a_h(u,v)} = \underbrace{\mathcal{F}(\cdot,v)+g_h(\cdot,v)}_{\mathcal{F}_h(v)},$$

where $\ell_h(u,v) = g_h(v)$.

$$-\mu\Delta u + \beta\cdot\nabla u + \sigma u - f = 0$$

$$\sum_{T_k} K(-\mu\Delta u + \beta\cdot\nabla u + \sigma u - f, -\mu\Delta v + \beta\cdot\nabla v + \sigma u) = 0,$$

where K depends on h and j.

4.6 Time Dependent Problems

- 1D heat equation:

$$\frac{\partial u}{\partial t} - \mu\frac{\partial^2 u}{\partial x^2} + \beta\frac{\partial u}{\partial x} + \sigma u = 0.$$

- Multiple dimension:

$$\frac{\partial u}{\partial t} - \nabla\cdot(\mu\nabla u) + \beta\nabla u + \sigma u = 0.$$

with boundary condition $u(\partial\Omega) = 0$ and initial condition $u(x, y, 0) = u_0(x, y)$.

- General approach: FD in time and FE in space.
- Variational formulation: $V = \mathcal{H}_0^1(\Omega)$ and $v \in V$:

$$\int_\Omega \frac{\partial u}{\partial t} v + \int_\Omega \mu \boldsymbol{\nabla} u \boldsymbol{\nabla} v + \int_\Omega \beta \cdot \boldsymbol{\nabla} uv + \int_\Omega \sigma uv = \int_\Omega fv \quad \forall\, v \in V,$$

where

$$-\int_\Omega \boldsymbol{\nabla} \cdot (\mu \boldsymbol{\nabla} u) v = -\int_\Omega \mu \boldsymbol{\nabla} u \cdot uv + \int_\Omega \mu \boldsymbol{\nabla} u \boldsymbol{\nabla} v,$$

if μ is not space dependent.

We can add some regularity: $\mathcal{L}^2(0, T; \mathcal{H}_0^1(\Omega)) = \mathcal{L}^2(\mathcal{H}^1)$ and $\mathcal{L}^\infty(0, T; \mathcal{L}^2(\Omega)) = \mathcal{L}^\infty(\mathcal{L}^2)$. Then, the problem becomes: Find $u \in \mathcal{L}^2(\mathcal{H}_0^1) \cap \mathcal{L}^\infty(\mathcal{L}^2)\ s.t.$

$$\left(\frac{\partial u}{\partial t}, v\right) = a(u, v) = (f, v) \quad \forall\, v \in V = \mathcal{H}_0^1(\Omega).$$

By Lax-Milgram, this problem is:

1. Continuous for $a(\cdot, \cdot)$ and $\mathcal{F}(\cdot)$,
2. Weak coercive.

So, the problem is well-posed.

- Numerical problem: $V_h \subset V = \mathcal{H}_0^1(\Omega)$.

Find $u_h \in \mathcal{L}^2(V_h) \cap \mathcal{L}^\infty(\mathcal{L}^2)\ s.t.$

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) = \mathcal{F}(v_h) \quad \forall\, v_h \in V_h,$$

where $u_h(x, y, t) = \sum u_j^{(t)} \varphi_j(x, y)$.

- Solution from separation of variables:

$$u = T(t)X(x),$$

where T represents time and X represents space.

$$\begin{aligned} &\frac{\mathrm{d}T}{\mathrm{d}t} X - \frac{\partial^2 X}{\partial x^2} T = 0 \\ &\frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t} - \frac{1}{X}\frac{\mathrm{d}^2 X}{\mathrm{d}x^2} = K \quad \leftarrow \text{separable} \end{aligned}$$

So, we have

$$u = \sum_{j=0}^{\infty} T_j X_j(x).$$

A numerical solution will be

$$u = \sum_{j=0}^{N} T_j X_j(x).$$

The error is

$$e = \sum_{j=N+1}^{\infty} T_j X_j(x),$$

decays with a factor of e^{-N}. Not bad, but the problem is that this approach only works on a specific type of problem: separable.

- A more generic method:

$$\sum_j \frac{\mathrm{d}u_i}{\mathrm{d}t} \underbrace{(\varphi_j, \varphi_i)}_{\text{mass matrix}} + \sum u_j(t) \underbrace{a(\varphi_j, \varphi_i)}_{A} = b_j(t)$$

$$M \cdot \frac{\mathrm{d}u}{\mathrm{d}t} + Au = b$$

$$M \frac{1}{\Delta t}\big(u^{n+1} - u^n\big) + Au^{n+1} = b^{n+1}$$

$$\left(\frac{1}{\Delta t} M + A\right) u^1 = b^1 + \frac{1}{\Delta t} M u^0$$

$$\left(\frac{1}{\Delta t} M + A\right) u^{n+1} = b^{n+1} + \frac{1}{\Delta t} M u^0.$$

We can solve this system by θ method.

$$\frac{1}{\Delta t} M\big(u^{n+1} - u^n\big) + \theta A u^{n+1} + (1-\theta) A u^n = \theta b^{n+1} + (1-\theta) b^n$$

$$\left(\frac{1}{\Delta t} M + \theta A\right) u^{n+1} = \theta b^{n+1} + (1-\theta) b^n + \left(\frac{1}{\Delta t} M - (1-\theta) A\right) u^n.$$

- CFL condition for stability:

$$\frac{\Delta t}{\Delta x} |a| \le c < 1,$$

  1. For LX: $c = \frac{1}{\sqrt{3}}$
  2. For UPW: $c = \frac{1}{3}$.

- Wave equation: Leap frog can be incorporated with FEM. Also need to satisfy CFL conditions.