Emory University

MATH 572 Numerical PDEs

Learning Notes

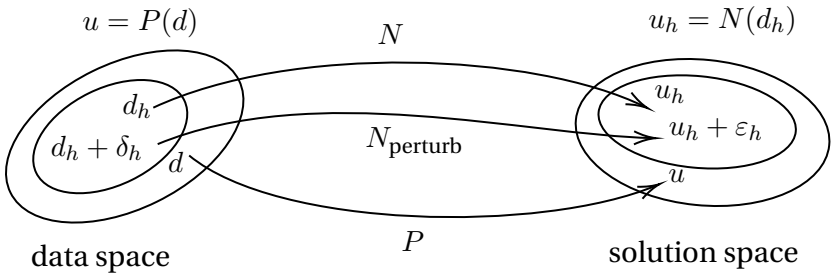
Jiuru Lyu

April 27, 2026

Contents

1 Introduction

1.1 Basic Concept



- Physical Problem: Solution u and data d.
- Numerical Problem: Numerical solution u_h and approximated data d_h, where h measures how fine the approximation is.

Some key concepts are

- *Convergence*: Does $u_h \to u$ when $h \to 0$?
- *Consistency*: Does $N \to P$ when $h \to 0$?
- *Stability*: Does $\varepsilon_h \to 0$ when $\delta_h \to 0$?

Another perspective:

$$P(u,d) = 0 \quad \text{true problem}, \quad P_N(u_N, d_N) = 0 \quad \text{numerical/approximated problem}$$

- Consistency: when $N \to \infty$ ($h \to 0$), $P_N \to P$.

  $$P_N(u, d_N) = \tau \quad \text{(consistency error)}.$$

  When $N \to \infty$, $\tau \to 0$.

- Stability: $P_N(u_N + \delta u, d_N + \delta) = 0$.

  $$|\delta u| \le c|\delta| \implies \text{when } \delta \to 0, \delta u \to 0.$$

- Lax-Richtmyer Theorem: For linear PDE problem, consistency and stability $\implies$ convergence.

1.2 Functional Analysis

Definition 1.2.1 (Functional). A *functional* is $\mathcal{F}(\cdot) : X \to \mathbb{R}$. Let $f(x), g(x) \in X$. A functional is *linear* if $\mathcal{F}(\lambda f(x) + \mu g(x)) = \lambda \mathcal{F}(f(x)) + \mu \mathcal{F}(g(x))$.

Example 1.2.2

For example, $\int_a^b f(x)\,\mathrm{d}x$ is a functional.

Definition 1.2.3 (Form). A *form* is $a(\cdot,\cdot) : X \times X \to \mathbb{R}$, where $f, g \in X$. A form is *bilinear* if:

- $a(\lambda f_1 + \mu f_2, g) = \lambda a(f_1, g) + \mu a(f_2, g)$, and
- $a(f, \lambda g_1 + \mu g_2) = \lambda a(f, g_1) + \mu a(f, g_2)$.

Definition 1.2.4 (Space). A *space* is set with operations defined. If X is a *space*, then

- $f \in X \implies \lambda f \in X$, and
- $f, g \in X \implies \lambda f + \mu g \in X$.

Definition 1.2.5 (Distance). A *distance* is $d(\cdot,\cdot) : X \times X \to \mathbb{R}$ such that $\forall\, f, g, w \in X$:

- $d(f,g) = d(g,f)$, [*symmetry*]
- $d(f,g) \geq 0$ and $d(f,g) = 0 \iff f = g$, and
- $d(f,g) \leq d(f,w) + d(w,g)$ [*triangle inequality*]

A space with a distance is a *metric space*.

Definition 1.2.6 (Completeness). Let (X,d) be a metric space. Let f_n be a sequence in X and f be a candidate limit.

1. The sequence is *convergent* if $d(f_n, f) \to 0$ as $n \to +\infty$.
2. The sequence is *Cauchy* if $d(f_n, f_m) \to 0$ as $n, m \to +\infty$.

In general, $1 \implies 2$, but $2 \not\implies 1$ [$d(f_n, f_m) \leq d(f_n, f) + d(f, f_m)$ *by triangle inequality*].
If in a metric space, $1 \iff 2$, then the metric space is *complete*.

Definition 1.2.7 (Norm). A *norm* is $\|\cdot\|_X : X \to \mathbb{R}$ such that

- $\|f\|_X \geq 0$ and $\|f\|_X = 0 \iff f = 0$,
- $\|\lambda f\|_X = \lambda \|f\|_X$, and
- $\|f + g\|_X \leq \|f\|_X + \|g\|_X$.

A space is *normed* if a norm is defined.
A normed space is also a metric space: $d(f,g) = \|f - g\|_X$.

Definition 1.2.8 (Banach Space). A *Banach space* is a complete normed space.

Definition 1.2.9 (Scalar Product). $(\cdot,\cdot)_X : X \times X \to \mathbb{R}$ is a symmetric bilinear form such that

- $(f,g)_X = (g,f)_X$
- $(f,f)_X \geq 0$ and $(f,f)_X = 0 \iff f = 0$.
- $|(f_1,f_2)_X| \leq (f_1,f_1)_X^{1/2}(f_2,f_2)_X^{1/2}$ [*Cauchy-Swcharz Inequality*]

A scalar product induces a norm:

$$\|f\|_X^2 = (f,f)_X.$$

So, a space with a scalar product is automatically normed.

Example 1.2.10

$$\begin{aligned}(f_1+f_2,f_1+f_2)_X &= (f_1,f_1)_X + 2(f_1,f_2)_X + (f_2,f_2)_X \\ &\leq \|f_1\|_X^2 + 2\|f_1\|_X\|f_2\|_X + \|f_2\|_X^2 \\ &= (\|f_1\|_X + \|f_2\|_X)^2.\end{aligned}$$

Definition 1.2.11 (Hilbert Space). A *Hilbert space* is a Banach space equipped with a scalar product.

Definition 1.2.12 (Orthogonality). The *angle* between two functions is defined as

$$\angle(f_1,f_2) : \cos(\varphi) = \frac{(f_1,f_2)}{\|f_1\|\|f_2\|} = \frac{(f_1,f_2)}{(f_1,f_1)^{1/2}(f_2,f_2)^{1/2}}$$

Two functions are *orthogonal* to each other $\iff (f,g) = 0$. We denote $f \perp g$.

Find the Best Approximation Let $X_L \subset X$, where X_L is "small." X_L and X are both Hilbert spaces. We want to find the *best* approximation of f_L of f in X_L. So, we need to find f_L $s.t.$

$$(f - f_L, v_L) = 0 \quad \forall\, v_L \in X_L.$$

This is an extension of the Pythagorean Theorem. If $g_L \in X_L$:

$$\begin{aligned}\|f-g_L\|^2 = (f-g_L, f-g_L) &= (f-f_L+f_L-g_L, f-f_L+f_L-g_L)\\ &= \|f-f_L\|^2 + 2\underbrace{(f-f_L, f_L-g_L)}_{=0,\quad \text{orthogonality}} + \|f_L-g_L\|^2\\ &= \|f-f_L\|^2 + \|f_L-g_L\|^2\\ &\geq \|f-f_L\|^2.\end{aligned}$$

1.2.1 Lebesgue Integration

Definition 1.2.13 (L^p Spaces).

$$L^p(a,b) = \left\{ f : (L)\int_a^b f^p \,\mathrm{d}x < +\infty \right\}, \quad \text{with } p = 1, 2, \ldots, \in \mathbb{N}.$$

Here, (L) indicates that the integration is done in Lebesgue sense. $L^p(a,b)$ is the collection of functions that can be integrated in interval (a,b) in the Lebesgue sense without blowing up. These are Banach spaces with the norm

$$\|f\|_p = \left(\int_a^b |f|^p \,\mathrm{d}x \right)^{1/p}.$$

Similarly,

$$L^\infty(a,b) = \{f : \text{bounded (up to all-measurable intervals) in } (a,b)\}$$

is also a Banach space with norm

$$\|f\|_\infty = \max_{(a,b)} |f|.$$

Specially, $L^2(a,b)$ or $L^2(\Omega)$ with $\Omega \subseteq \mathbb{R}$ is a Hilbert space:

$$(f,g) = \int_a^b fg \,\mathrm{d}x \quad \text{and} \quad \|f\|_2 = \left(\int_a^b f^2 \,\mathrm{d}x \right)^{1/2} = (f,f)^{1/2}.$$

Definition 1.2.14 (Conjugate Spaces). L^p and L^q are called *conjugated spaces* if we have the function $f \in L^p(a,b)$ and $g \in L^q(a,b)$ with

$$\frac{1}{p} + \frac{1}{q} = 1.$$

We then have

$$(L)\int_a^b fg \,\mathrm{d}x < +\infty \quad \text{or} \quad fg \in L^1(a,b)$$

Remark. L^2 is conjugate with itself. We have

$$|(f,g)_{L^2}| \leq \|f\|_2 \|g\|_2.$$

1.2.2 Distributional Derivatives

Definition 1.2.15 (Dual Spaces). The set of linear bounded functionals obtained on a space X is per se a functional space, that we call *dual space* and denote as X':

$$X' = \{\mathcal{F}(f) \text{ for } f \in X, \text{ linear and } s.t. \|\mathcal{F}(f)\| \leq c\|f\|_X\}.$$

Theorem 1.2.16 Riesz Representation Theorem
In L^2, the dual of L^2 is isometric to L^2. [*This means that any linear and continuous functional in L^2 corresponds to an element within L^2 itself.*]

Definition 1.2.17 (Support). The *support* of a function is where the function $\neq 0$ on the interval or region of the domain of definition. In particular, if the support is closed and bounded region, we say the function has *compact support.*

Definition 1.2.18 (The Space $\mathcal{D}$).

$$\mathcal{D} = \{\text{functions in } \mathcal{C}^\infty(\mathbb{R}^n)\ (n \geq 1) \text{ with compact support}\}.$$

The dual space of $\mathcal{D}$, $\mathcal{D}'$ is defined in the following notion:

$$\mathcal{D}' : \langle T, f\rangle_{\mathcal{D}} \in \mathbb{R}, \text{ where } T \in \mathcal{D}' \text{ is linear and continuous and } f \in \mathcal{D}.$$

T to be continuous is defined in the following sense: Let $\text{supp}(f_n) \subseteq K$ and $f_n \to f$ and $n \to \infty$. Then, T is *continuous* if $\langle T, f_n\rangle \to \langle T, f\rangle$.

Example 1.2.19

Let $H(x)$ be the step function:

$$H(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0. \end{cases}$$

Then,

$$\langle T, f\rangle = \int_{-\infty}^{+\infty} H f \, \mathrm{d}x = \int_0^{+\infty} f \, \mathrm{d}x.$$

The functional $\langle T_0, f\rangle = f(0)$ belongs to $\mathcal{D}'$:

$$|\langle T_0, f\rangle| \leq \max_{x\in\mathbb{R}} |f(x)|.$$

Definition 1.2.20 (Distributions). The objects of $\mathcal{D}'$ are called *distributions* or generalized functions.

Theorem 1.2.21

$$\langle T', f\rangle = -\langle T, f'\rangle .$$

Proof 1. Let $g \in L^2$. Then,

$$\langle g', f\rangle = \int_{\mathbb{R}} g'f = \Big[gf\Big]_{-\infty}^{+\infty} - \int_{\mathbb{R}} gf' = -\langle g, f'\rangle$$

Note that $\Big[gf\Big]_{-\infty}^{+\infty} = 0$ because $\lim_{x\to\infty} f = 0$, and f has compact support. Q.E.D. ■

Example 1.2.22

Consider

$$H(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 & \text{for } x > 0. \end{cases}$$

Then, we can *differentiate* H:

$$\int_{-\infty}^{+\infty} H'f = -\int_{-\infty}^{+\infty} Hf' = -\int_0^{+\infty} f = -\Big[f\Big]_0^{+\infty} = f(0) = \langle T_0, f\rangle .$$

T_0 is called the *Dirac-δ* and it is the mathematical description of an *impulse*. It is *not* a function in the traditional sense.

Definition 1.2.23 (Dirac-δ). *Dirac-δ* is the element in $\mathcal{D}'$ such that

$$\langle \delta, f\rangle = f(0)$$

We know that $H' = \delta$.

Theorem 1.2.24

$$L^2 \subset \mathcal{D}'.$$

1.2.3 Special Spaces

Definition 1.2.25 (Sobolev Spaces). Space $H^k(a,b)$ (or $H^k(\Omega)$) is the space of functions f $s.t.$ it belongs to L^2 together with all the derivatives up to the order k.

$$H^k(a,b) = \left\{ f \text{ } s.t. \text{ } f, f', \dots, f^{(k)} \in L^2 = H^0 \right\}.$$

In multiple dimensions, the derivatives are all the derivatives of order k:

$$\frac{\partial^{\alpha_1+\alpha_2+\cdots+\alpha_m} f}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \cdots \partial^{\alpha_m} x_m} \in L^2 \quad \text{with } \sum_i \alpha_i = k.$$

These spaces are called *Sobolev spaces*. A more generalized definition:

$$W^{k,p} = \left\{ f \text{ } s.t. \text{ } f, f', \dots, f^{(k)} \in L^p \right\}.$$

So, $H^k = W^{k,2}$.

Theorem 1.2.26 Properties of H^k

- $H^k(a,b)$ is a Hilbert space with the scalar product

$$(f,g)_{H^k} = \sum_{i=0}^{k} \left(f^{(i)}, g^{(i)} \right)_{L^2}$$

- $H^1(a,b)$ has the norm

$$\|f\|_{H^1}^2 = \|f\|_{L^2}^2 + \left\|f'\right\|_{L^2}^2$$

- $H^1 \subset L^2$. So,

$$\|f\|_{H^1}^2 \geq \|f\|_{L^2}^2.$$

  In general,

$$H^{k+1} \subset H^k \subset \cdots \subset L^2.$$

Definition 1.2.27 (Trace Operator). *Trace* is an operator that maps a function in H^k to its value of the boundary, and, in general, reduces regularity.

$$\gamma : H^k(\Omega) \to H^{k-\frac{1}{2}}(\partial\Omega)$$

is the (continuous) trace operator.

Definition 1.2.28 (The Space H_0^1). The space $H_0^1(\Omega)$ is the space of H^1 functions with null trace on $\partial\Omega$.

Theorem 1.2.29 Poincare Inequality
$\exists\, c > 0\ s.t.$ for $f \in H_0^1(\Omega)$,

$$\|f\|_{L^2} \leq c\|\nabla f\|_{L^2}.$$

Therefore,

$$\|\nabla f\|_{L^2}^2 \leq \underbrace{\|f\|_{L^2}^2 + \|\nabla f\|_{L^2}^2}_{\|f\|_{H^1}^2} \leq (1+c^2)\|\nabla f\|_{L^2}^2.$$

Hence, in $H_0^1(\Omega)$, the norm of ∇f in L^2 and of f in H^1 are equivalent. This is also true if the function has only a realization of the boundary (with measure > 0) where it vanishes.

Remark.

$$\alpha\|u\|_F \leq \|u\|_S \leq \beta\|u\|_F$$

implies that $\|\cdot\|_F$ and $\|\cdot\|_S$ are equivalent.

Definition 1.2.30 (Dual of H_0^1: H^{-1}).

$$H_0^1 \subset L^2 \subset H^{-1}.$$

Space-Time Functional Spaces For time-dependent problems, we have spaces that describe the two dependencies. For instance:

$$L^p(0,T;H^k(\Omega)) \quad \text{is the space of functions } s.t.\ \|f\|_{H^k}(t) \in L^p(0,T).$$

In short, we denote it as $L^p(H^k)$. We will use $L^2(H_0^1)$ and $L^\infty(L^2)$.

Theorem 1.2.31 Embedding Theorems
If $u \in H^s$ and $\Omega \subset \mathbb{R}^n$,

- $0 < 2s < n$: $H^s \subset L^q$, where $1 \leq q \leq q^* = \dfrac{2n}{n-2s}$ [*continuous embedding*]
- $2s = n$: $H^s \subset L^q \quad \forall\, q : 1 \leq q < +\infty$
- $2s > n$: $H^s \subset C^0(\overline{\Omega})$.

Also, $H^s \subset C^n(\overline{\Omega})$ if $s > n + \dfrac{n}{2}$.

Remark. $H_1^1(\Omega) = \left\{ f \in H^1, f(\partial\Omega) = 1 \right\}$ is not a space. In fact:

$$f_1, f_2 \in H_1^1 \implies f_1 + f_2 \notin H_1^1.$$

1.3 Solving $A\mathbf{x} = \mathbf{b}$

Direction Method

- Gaussian Elimination: $A = LU$.

$$A\mathbf{x} = \mathbf{b} \implies LU\mathbf{x} = \mathbf{b} \implies \begin{cases} L\mathbf{y} = \mathbf{b} \\ U\mathbf{x} = \mathbf{y} \end{cases}$$

- With pivoting: $PAQ = LU$.

$$A\mathbf{x} = \mathbf{b} \implies PA\mathbf{x} = P\mathbf{b} \implies PAQ\underbrace{Q^{-1}\mathbf{x}}_{\mathbf{z}} = P\mathbf{b} = \implies \begin{cases} LU\mathbf{z} = P\mathbf{b} = \mathbf{c} \\ \mathbf{x} = Q\mathbf{z} \end{cases}$$

- Problem:
  1. Complexity $\sim \mathcal{O}(n^3)$
  2. Rounding errors

     Theorem 1.3.1 Perturbation Theorem
     Suppose a perturbation $(A + E)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \boldsymbol{\delta}$. Then, the error

     $$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \chi(A)\left(\frac{\|\boldsymbol{\delta}\|}{\|\mathbf{b}\|} + \frac{\|E\|}{\|A\|}\right),$$

     where $\chi(A) = \|A\|\left\|A^{-1}\right\| \geq 1$ is the condition number of A.
  3. Storage

Iterative Methods

$$\mathbf{x}^{(k)} \longrightarrow \mathbf{x}_{\text{exact}}$$

Example 1.3.2

Given $A\mathbf{x} = \mathbf{b}$ and initial guess $\mathbf{x}^{(0)}$:

$$\mathbf{x}^{(k+1)} = \underbrace{B_k\mathbf{x}^{(k)}}_{\sim\mathcal{O}(n^2)} + \mathbf{c}_k$$

- Total cost: $\sim \mathcal{O}(kn^2)$

- If $k \ll n$, iterative methods will be faster.
- Such methods converge $\iff$ $\rho(B) < 1$, where $\rho(B)$ is the spectral radius of B.

- Richardson method:

$$\begin{aligned} \mathbf{0} = \mathbf{x} - \mathbf{x} &= \mathbf{b} - A\mathbf{x} \\ \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)} &= \mathbf{b} - A\mathbf{x}^{(k)} \\ \implies \mathbf{x}^{(k+1)} &= (I - A)\mathbf{x}^{(k)} + \mathbf{b}. \end{aligned}$$

  Convergent when $\rho(I - A) < 1$.

- What if $\rho(I - A) \geq 1$? If S is SPD, we can introduce a parameter σ:

$$\begin{aligned} \mathbf{0} = \mathbf{x} - \mathbf{x} &= (\mathbf{b} - A\mathbf{x})\sigma \\ \mathbf{x}^{(k+1)} &= (I - \sigma A)\mathbf{x}^{(k)} + \sigma\mathbf{b}. \end{aligned}$$

  In fact, choosing $\sigma = \dfrac{1}{\lambda_{\min} + \lambda_{\max}}$ yilds the best convergence. More generally, choosing $0 < \sigma < \dfrac{2}{\lambda_{\max}}$ gives convergence.

- Preconditioning:

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

  We require $\rho(I - P^{-1}A) < 1$. Ideally, we should choose $P = A$, but it is not practical. In practice, we want some $P \approx A$. For example,

  1. $P = \mathrm{diag}(\mathrm{diag}(A))$ yields the *Jacobi* method.
  2. $P = \mathrm{lower}(A)$ yields the *Gauss-Seidel* method.

- Error: Richardson with Optimal σ for SPD A:

$$\left\|\mathbf{e}^{(k+1)}\right\| \leq \frac{\chi_2(A) - 1}{\chi_2(A) + 1}\left\|\mathbf{e}^{(k)}\right\|,$$

  where $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}_{\text{exact}}$.

- If A is SPD, $A\mathbf{x} = \mathbf{b}$ can be rewritten as a LS problem:

$$\min J \equiv \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top\mathbf{b}.$$

  By FOC: $\boldsymbol{\nabla} J = 0$. So, we can use methods from optimization:

  1. Gradient descent, and

2. Conjugate gradient.

- If A is not SPD, $A^\top A$ is SPD, but $\chi(A^\top A) = (\chi(A))^2$. If we want to use CG, use BiCG instead.
- Krylov subspace methods: GMRES.

1.4 General Statements of PDEs

Definition 1.4.1 (n-th Order PDE).

$$\mathcal{F}\left(u, \frac{\partial u}{\partial x_i}, \frac{\partial^2 u}{\partial x_i \partial x_j}, \dots, \frac{\partial^n u}{\partial^{i_1} x_1 \partial^{i_2} x_2 \cdots \partial^{i_d} x_d}, f\right) = 0,$$

where $i_1 + i_2 + \cdots + i_d = n$, and $d =$ number of independent variables.

Definition 1.4.2 (Linearity of PDEs).

- Linear: all derivatives have linear occurances.
- Quasi-Linear: linear in derivatives of order n. For example,

$$\frac{\partial u}{\partial t} + (1 - 2u)\frac{\partial u}{\partial x} = 0$$

- Semi-Linear: derivatives are linear, and coefficients of derivatives do not depend on lower-dimensional derivatives. For instance,

$$\frac{\partial u}{\partial t} + \sin(x)\frac{\partial u}{\partial x} + u^3 = 0$$

Definition 1.4.3 (Boundary Value Problems/BVP). $B(u)\big|_{\partial\Omega} =$ data is the boundary conditions. Then,

$$\begin{cases} \mathcal{F}(u, \dots, f) \equiv 0 \\ B(u)\big|_{\partial\Omega} = \text{data} \end{cases}$$

defines a *BVP*. An *initial boundary value problem (IBVP)* is a BVP with one dimension being the time dimension.

Definition 1.4.4 (Well-Posedness). Our problem is *well-posed (WP)* when the solution

- exists,
- is unique, and
- depends continuously on the data [*Suppose* $\mathcal{F}(u, \dots, f) = 0$. *Perturb the problem:* $\mathcal{F}(u^*, \dots, f + \delta f = 0$. *Does* $u^* \to u$ *as* $\delta f \to 0$*?*]

2 Poisson Equation

$$\begin{cases} -\Delta u = f \qquad \text{and some BCs} \\ \text{OR:} \quad -\dfrac{\partial^2 u}{\partial x_1^2} - \dfrac{\partial^2 u}{\partial x_2^2} - \dfrac{\partial^2 u}{\partial x_3^2} = f. \end{cases}$$

Boundary Conditions:

- Dirichlet $u(\partial\Omega) = g$.

  With lifting function, we can turn Dirichlet into a homogeneous condition: Let $\mathscr{L}_g \in \Omega \, s.t. \, \mathscr{L}_g(\partial\Omega) = g$. Then,

  $$u = \overset{\circ}{u} + \mathscr{L}_g \implies \begin{cases} -\Delta \overset{\circ}{u} = f + \Delta \mathscr{L}_g \\ \overset{\circ}{u}(\partial\Omega) = 0 \end{cases}$$

  So, only homogeneous Dirichlet BC will be discussed.

- Neumann: $\boldsymbol{\nabla} u \cdot \mathbf{n} = g$.

  Poisson equations with Neumann BCs are not WP. The solutions are not unique by adding constants.

- Robin: $\boldsymbol{\nabla} u \cdot \mathbf{n} + \alpha u(\partial\Omega) = g$.

2.1 Poisson Equation as a Minimization Problem

Consider the energy:

$$J(u) = \frac{1}{2}\int_\Omega |\boldsymbol{\nabla} u|^2 - \int_\Omega fu$$

with boundary condition $u(\partial\Omega) = 0$.

Goal: Find the configuration (u) that minimizes the energy. i.e.,

$$\min J(u).$$

Solve $\boldsymbol{\nabla} J = 0$. Let v be a function such that $v(\partial\Omega) = 0$. Then,

$$\begin{aligned} \boldsymbol{\nabla} J &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}(J(u+\varepsilon v) - J(u)) \\ &= \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \cdot \frac{1}{2}\int_\Omega \cancel{|\boldsymbol{\nabla} u|^2} + \varepsilon^2 |\boldsymbol{\nabla} v|^2 + 2\varepsilon \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v - \cancel{\int_\Omega fu} - \varepsilon \int_\Omega fv + \frac{1}{2}\cancel{|\boldsymbol{\nabla} u|^2} + \cancel{\int_\Omega fu} \\ &= \lim_{\varepsilon\to 0} \frac{1}{2}\int_\Omega \cancel{\varepsilon |\boldsymbol{\nabla} v|^2} + 2\boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v - \int_\Omega fv \\ &= \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v - \int_\Omega fv = 0 \quad \forall \, v \, s.t. \, v(\partial\Omega) = 0. \end{aligned}$$

By Greens' Theorem and Gauss Theorem,

$$\int_{\partial\Omega} \mathbf{v} \cdot \mathbf{n} = \int_\Omega \boldsymbol{\nabla} \cdot (\mathbf{v}) \quad \text{and} \quad \int_\Omega \boldsymbol{\nabla} \cdot (\boldsymbol{\beta}\rho) = \int_\Omega \boldsymbol{\nabla}\rho \cdot \boldsymbol{\beta} + \int_\Omega \boldsymbol{\nabla} \cdot \boldsymbol{\beta} \cdot \rho.$$

So,

$$\int_\Omega \boldsymbol{\nabla}\rho \cdot \boldsymbol{\beta} = \int_{\partial\Omega} \rho\boldsymbol{\beta}\cdot\mathbf{n} - \int (\boldsymbol{\nabla}\cdot\boldsymbol{\beta})\cdot\rho.$$

Substitute $\rho = v$ and $\boldsymbol{\beta} = \boldsymbol{\nabla}u$:

$$\int_\Omega \boldsymbol{\nabla}u\cdot\boldsymbol{\nabla}v = \underbrace{\int_{\partial\Omega} v\boldsymbol{\nabla}u\cdot\mathbf{n}}_{=0 \text{ b/c } v(\partial v)=0} - \int_\Omega (\Delta u)v.$$

Hence, we have

$$\begin{aligned} -\int_\Omega (\Delta u)v - \int_\Omega fv &= 0 \\ \int_\Omega (-\Delta u - f)v &= 0 \\ \boxed{-\Delta u = f}&. \end{aligned}$$

Remark.

- Strong formulation:
$$-\Delta u = f$$
requires second derivatives (Laplace operator)
- Week formulation:
$$\int_\Omega \boldsymbol{\nabla}u\boldsymbol{\nabla}v - \int_\Omega fv = 0 \quad \forall\, v\ s.t.\ v(\partial\Omega) = 0$$
only needs first-order derivatives (Gradient).

If u solves $\int_\Omega \boldsymbol{\nabla}u\boldsymbol{\nabla}v - \int_\Omega fv = 0 \quad \forall\, v\ s.t.\ v(\partial\Omega) = 0$, then $J(u) \le J(u+w)$.

2.2 Finite Differences

Two mathematically equivalent strong formulations may be different numerically:

$$-\Delta u = f.$$

On the other hand, we can write

$$-\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}u) = f \implies \begin{cases} \boldsymbol{\nabla}u = \boldsymbol{\sigma} \\ \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = f. \end{cases}$$

2.2.1 Poisson Equation in 1D

$$\begin{cases} -u'' = f, \quad x \in [0,1] \\ u(0) = u(1) = 0 \end{cases}$$

- Max-Min Principle
- Discretization:

$$-u''(x_j) = f(x_j).$$

Collocation/Mesh. $|h|$ is the step size.

- **Claim**

$$(u')' \approx \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2}$$

Proof 1. Denote $u(x_{j+1}) = u_{j+1}$. By Taylor expansion:

$$\begin{aligned} u_{j+1} &= u_j + u'(x_j)h + \frac{1}{2}u''(x_j)h^2 + \frac{1}{3!}u'''(x_j)h^3 + \frac{1}{4!}u^{(4)}(x_j)h^4 + \mathcal{O}(h^5) \\ u_{j-1} &= u_j - u'(x_j)h + \frac{1}{2}u''(x_j)h^2 - \frac{1}{3!}u'''(x_j)h^3 + \frac{1}{4!}u^{(4)}(x_j)h^4 + \mathcal{O}(h^5) \\ u_{j+1} + u_{j-1} &= 2u_j + u''(x_j)h^2 + \frac{2}{4!}u^{(4)}h^4 + \mathcal{O}(h^6) \\ u''(x_j) &= \frac{u_{j+1} + u_{j-1} - 2u_j}{h^2} + \underbrace{\frac{1}{12}u^{(4)}(x_j)h^2}_{\tau(h),\text{ consistency error}} \end{aligned}$$

Hence, when we make the mesh finer ($h \to \frac{1}{2}h$), the consistency error decrease by 4 times (order of 2). Q.E.D. ■

- Form the numerical problem: Except for the endpoints, we can write

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f_j.$$

Denote

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{bmatrix}.$$

Define

$$A = \frac{1}{h^2}\begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}.$$

Then, we get $\boxed{A\mathbf{u} = \mathbf{f}}$.

- **Proposition** *A is SPD (symmetric positive definite) $\implies$ all eigenvalues of A are positive.*

  Proof 2. We will show A is symmetric, and A is positive definite.

  1. Symmetry is trivial.
  2. Positivity: [*WTS:* $\forall\, \mathbf{x} \neq 0, \mathbf{x}^\top A\mathbf{x} > 0$.]

$$\begin{aligned}\mathbf{x}^\top A\mathbf{x} &= \frac{1}{h^2}\left(2x_1^2 - 2x_1x_2 + x_2^2 + \cdots\right)\\ &= \frac{1}{h^2}\left(x_1^2 + (x_1 - x_2)^2 + \cdots + (x_j - x_{j+1})^2 + \cdots + x_{N-1}^2\right)\\ &\geq 0.\end{aligned}$$

  If we have a constant vector: $\mathbf{x} = \begin{bmatrix}\mathbf{x}_* \cdots \mathbf{x}_*\end{bmatrix}^\top$. Then,

$$\mathbf{x}^\top A\mathbf{x} = \frac{1}{2}(\mathbf{x}_*^2 + \mathbf{x}_*^2) = 0 \iff \mathbf{x}_* = 0.$$

  Hence, $\forall\, \mathbf{x} \neq 0,\ \mathbf{x}^\top A\mathbf{x} \geq 0$.

  So, A is SPD. Q.E.D. ■

- **Corollary** *A is non-singular, and the numerical problem is well-posed.*

- Error Analysis:

$$\begin{aligned}\mathbf{u}_{\text{num}} = A^{-1}\mathbf{f} &\implies \|\mathbf{u}_{\text{num}}\| \leq \left\|A^{-1}\right\|\|\mathbf{f}\|\\ A\mathbf{u}_{\text{ex}} = \mathbf{f} + \boldsymbol{\tau}, &\qquad \begin{bmatrix}\tau_j\end{bmatrix} = \begin{bmatrix}cf''(x_j)h^2\end{bmatrix}\\ \mathbf{e} = \mathbf{u}_{\text{ex}} - \mathbf{u}_{\text{num}} &= A^{-1}\boldsymbol{\tau}\\ \|\mathbf{e}\| \leq \left\|A^{-1}\right\| &\underbrace{\|\boldsymbol{\tau}\|}_{\mathcal{O}(h^2)}\end{aligned}$$

  But this is not enough: A^{-1} also depends on h.

  Proposition $\left\|A^{-1}\right\|_2 \perp\!\!\!\perp h$.

  Proof 3. Recall: $\|A\|_2 = \rho(A) = \max_i |\lambda_i|$, the spectral radius. Then,

$$\left\|A^{-1}\right\|_2 = \frac{1}{\min\limits_i |\lambda_i|} = \rho\left(A^{-1}\right).$$

  [*WTS: eigenvalues of A (or A^{-1}) are independent of h*].

$$A = \frac{1}{h^2}\,\text{trid}(-1, 2, -1), \quad T = \text{trid}(-1, 2, -1).$$

Solve $T\mathbf{x} = \lambda\mathbf{x}$:

$$(T - \lambda I)\mathbf{x} = 0$$
$$x_{j-1} - (2-\lambda)x_j + x_{j+1} = 0.$$

Characteristic equation:

$$1 - (2-\lambda)\rho^2 + \rho^2 = 0$$
$$x_j = a\rho_1^j + b\rho_2^j \quad \text{or} \quad x_j = a\rho^j + jb\rho^j.$$

Imposing $\Delta \geq 0$, we get $\lambda \geq 4$. Imposing BC: $x_0 = x_N = 0$, we get $a = b = 0$, and thus $\mathbf{x} = 0$. Trivial.

Hence, we don't have eigenvalues ≥ 4, and we should only consider complex solutions.

$$\rho_{1,2} = \frac{2 - \lambda \pm \mathrm{i}\sqrt{2}\sqrt{4-\lambda}}{2} = \cos(\theta) \pm \mathrm{i}\sin(\theta) \quad [\textit{Enforcing } |rho| = 1]$$

So,

$$x_j = a\rho_1^j + b\rho_2^j.$$

Since $x_0 = x_N = 0$:

$$\begin{cases} a + b = 0 \\ a\rho_1^N + B\rho_2^N = 0 \end{cases} \implies \rho_1^N - \overline{\rho}_1^N = 0.$$

$$cos(N\theta) + \mathrm{i}\sin(N\theta) - \cos(N\theta) + \mathrm{i}\sin(N\theta) = 0 \implies \theta = \frac{j\pi}{N}, \quad j = 1, \ldots, N-1.$$

Going back, we have

$$\lambda_j = -\frac{1}{\rho} + \rho + 2 = 2\left(1 - \cos\left(\frac{j\pi}{N}\right)\right)$$
$$\cos(2\alpha) = 1 - 2\sin^2(\alpha)$$
$$2\sin^2(\alpha) = 1 - \cos(2\alpha) \implies \lambda_j = 4\sin^2\left(\frac{j\pi}{2N}\right)$$

Hence, eigenvalues of A:

$$\lambda = \frac{4}{h^2}\sin^2\left(\frac{\pi}{2N}\right) = \frac{4}{h^2}\sin^2\left(\frac{\pi}{2}h\right) \quad [h = \frac{1}{N}]$$

When $h \to 0$, $\sin^2\left(\frac{\pi}{2}h\right) \to \left(\frac{\pi}{2}h\right)^2$. So,

$$\lambda = \frac{4}{h^2}\left(\frac{\pi}{2}\right)^2 h^2 = \pi^2 \perp\!\!\!\perp h.$$

Q.E.D. ■

Corollary *Our numerical method is convergent.*

$$\mathbf{e} = A^{-1}\boldsymbol{\tau}.$$

When $h \to 0,\ \boldsymbol{\tau} \to 0,\ \left\|A^{-1}\right\|_2 \to \pi^2$. Then,

$$\|\mathbf{e}\|_2 \leq \left\|A^{-1}\right\|_2 \|\boldsymbol{\tau}\|_2 = \pi^2 \|\boldsymbol{\tau}\|_2.$$
$$\|\boldsymbol{\tau}\|_2^2 = \sum_{i=1}^{N-1} \tau_i^2 = \sum_{i=1}^{N-1} c^2 \big(f''(x_j)\big)^2 h^4.$$

If f'' is bounded, $\|f''\| \leq \widehat{c}$. Then, combing the constants and the sum, we have

$$\|\boldsymbol{\tau}\|_2^2 \leq \widetilde{c}(N-1)h^4.$$

Recall that $N = \dfrac{1}{h}$. Then, $\|\boldsymbol{\tau}\|_2^2 \leq \widetilde{c}h^3 \implies \|\boldsymbol{\tau}\|_2 \sim \mathcal{O}\big(h^{3/2}\big)$. So,

$$\|\mathbf{e}\|_2 \sim \mathcal{O}\Big(h^{3/2}\Big).$$

- Practical Notes:

  When constructing A, we first assemble the full matrix, and then replace the boundaries with actual BCs.

- Another perspective of FD Convergence.

Definition 2.2.1 (h-Norm). Let v_h be a grid function.

$$\|v_h\|_h^2 = \left(\sum_{j=1}^{N-1} v_j^2 + \frac{1}{2}v_0^2 + \frac{1}{2}v_N^2\right) h,$$

and

$$L_h v_h = \frac{v_{j+1} - 2v_j + v_{j-1}}{h^2}.$$

Some properties include:

1. $(L_h v_h, w_h) = (v_h, L_h w_h)$
2. $(L_h v_h, w_h) = \displaystyle\sum_{j=1}^{N-1} \frac{v_{j+1} - v_j}{h} \cdot \frac{w_{j+1} - w_j}{h}.$
3. $h(L_h v_h, v_h)$ is a norm. Denote this norm as $|||\cdot|||_h$.

   Proof 4.

   $$h(L_h v_h, v_h) = h \sum \left(\frac{v_{j+1} - v_j}{h}\right)^2.$$

   Non-negativity and other properties can be easily shown. Q.E.D. ■

4. $c\|v_h\|_h^2 \leq |||v_h|||_h^2$.
5. Problem: $L_h u_h = f_h$.

$$
\begin{aligned}
c\|u_h\|_h^2 \leq h(L_h u_h, u_h) \leq \|f_h\|_h \|u_h\|_h &\implies \|u_h\| \leq c\|f_h\|_h \\
c\|\mathbf{e}_h\|^2 \leq h(L_h \mathbf{e}_h, \mathbf{e}_h) \leq \|\boldsymbol{\tau}\|_h \|\mathbf{e}_h\|_h &\implies \|\mathbf{e}_h\|_h \leq \frac{1}{c}\|\boldsymbol{\tau}\|_h \\
\|\boldsymbol{\tau}\|_h^2 = h\left|\sum_{j=1}^{N} c_j h^4\right| &\leq h \cdot N \cdot \max_{j=1,\dots,N} c_j \cdot h^4 \\
&= h \cdot \frac{1}{h} \max_{j=1,\dots,N} c_j \cdot h^4 \\
&= h^4 \max_{j=1,\dots,N} c_j \\
\|\boldsymbol{\tau}\|_h \leq \sqrt{h^4 \max_{j=1,\dots,N} c_j} &= h^2 \sqrt{\max_{j=1,\dots,N} c_j} \sim \mathcal{O}(h^2).
\end{aligned}
$$

Remark 5. (What's Missing Here?).

$$\text{Differential Problem} \xrightarrow{①} \underbrace{A\mathbf{u} = \mathbf{b}}_{\substack{\text{Discretized} \\ \text{linear system}}} \xrightarrow{②} \mathbf{u}_{\text{num}} \approx A^{-1}\mathbf{b}.$$

1. In ①: Convergence analysis
2. In ②: numerical linear algebra errors and rounding errors. (We were not discussing these previously).

2.3 Galerkin Methods

Set-Up

$$\begin{cases} -\Delta u = f \\ u(\partial\Omega) = 0 \end{cases}$$

Weak formulation: Find u *s.t.*

$$\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x \quad \forall\, v \; s.t. \; v(\partial\Omega) = 0.$$

Requirements $\boldsymbol{\nabla} u \in L^2$, $\boldsymbol{\nabla} v \in L^2 \implies u, v \in H_0^1$, and $f \in H^{-1}$ (dual of H_0^1, functional applied to H_0^1.)

Lemma 2.1 Max-Milgram Lemma (Sufficient Condition) Consider the problem $a(u,v) = \mathcal{F}(v)$, where $u, v \in X$. If

- $a(\cdot,\cdot)$ is a bilinear form,
- bounded: $\exists\, M > 0 \; s.t. \; |a(u,v)| \leq M\|u\|_X \|v\|_X$,

- coercivity: $a(u,u) \geq \alpha \|u\|_X^2 \quad \forall\, u \in X$, and
- $\mathcal{F}(\cdot)$ is a continuous linear functional: $\|\mathcal{F}(v)\|_X \leq \gamma \|v\|_X \quad \forall\, v \in X$,

then, the problem is well-posed. i.e., ①②③④ $\Longrightarrow$ well-posedness (sufficient but not necessary).

Theorem 2.3.2
Poisson equation is well-posed.

Proof 1.

$$\begin{cases} -\Delta u = f \\ u(\partial\Omega) = 0 \end{cases}$$

$\Longrightarrow \forall\, v \in H_0^1$, find $u \in H_0^1$ *s.t.*

$$\underbrace{\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v}_{a(u,v)} - \underbrace{\int_\Omega f v}_{\mathcal{F}(v)} = 0.$$

- $\|u\|_{H^1}^2 = \|u\|_{L^2}^2 + \|\boldsymbol{\nabla} u\|_{L^2}^2$
- $\left| \int_\Omega f v \right| \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{H^1} \|v\|_{H^1}$
- $\forall\, u \in X = H_0^1$,

$$\int_\Omega |\boldsymbol{\nabla} u|^2 = a(u,u) \geq \alpha \|u\|_X^2.$$

By Poincaré Inequality,

$$C_p \int_\Omega u^2 \leq \int_\Omega |\boldsymbol{\nabla} u|^2$$
$$C_p \|u\|_{L^2}^2 \leq \|\boldsymbol{\nabla} u\|_{L^2}^2 = \underbrace{|u|_{H_1}^2}_{\textit{semi-norm}}$$
$$\|u\|_{H^1}^2 = \int_\Omega u^2 + \int_\Omega \boldsymbol{\nabla} u^2 \leq \left(\frac{1}{C_p} + 1 \right) \int_\Omega |\boldsymbol{\nabla} u|^2$$

So, we have coercivity.

By Lax-Milgram Lemma, Poisson equation is WP. Q.E.D. ■

Remark. $a(u,v) = \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v$ is a scalar product in H_0^1:

- symmetric, and
- non-negativity (by coercivity).

The norm induced by $a(u,u) = \|u\|_a^2$ is equivalent to $\|u\|_{H^1}^2$:

$$C\|u\|_{H^1}^2 \leq a(u,u) \leq \|u\|_{H^1}^2.$$

If we change the problem and the space:

$$\begin{cases} -\Delta u = f \\ \boldsymbol{\nabla} u \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \quad \text{(Neumann)} \end{cases}$$

$\implies$ Find $u \in H^1$ *s.t.* $\quad \forall\, v \in H^1$:

$$\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega f v.$$

Now, we don't have Poincaré Inequality anymore, and $a(u,v)$ is not coercive. We cannot apply LM Lemma. In fact, the problem is not WP. Any $(u + \text{constant})$ will be a solution.

Hw to make the problem WP?

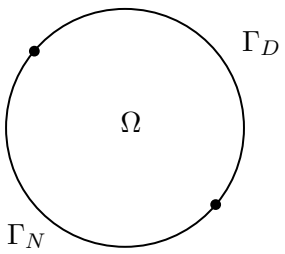
$$\begin{cases} -\Delta u + \sigma u = f, \quad \sigma > 0 \\ \boldsymbol{\nabla} u \cdot \mathbf{n} = 0 \quad \text{on } \partial\Omega \end{cases}$$

$$\implies \underbrace{\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v + \sigma \int_\Omega uv}_{a(u,v)} = \int_\Omega fv$$

$$|a(u,v)| \le \max(\sigma, 1)\|u\|_{H^1}\|v\|_{H^1}$$
$$\min(1,\sigma)\|u\|^2_{H^1} \le a(u,v).$$

$\implies$ $a(u,v)$ is coercive, and the problem is WP.

Now, let's consider mixed boundary conditions (i.e., Robin conditions).



$$\begin{cases} -\Delta u = f \\ u(\Gamma_D) = 0 \\ \boldsymbol{\nabla} u \cdot \mathbf{n}\big|_{\Gamma_N} = 0, \end{cases}$$

where $\overline{\Gamma_D \cup \Gamma_N} = \partial\Omega$, and $\text{measure}(\Gamma_D \cap \Gamma_N) = 0$.

$$-\int_\Omega \Delta u \cdot v = -\int_{\Gamma_N} \underbrace{\boldsymbol{\nabla} u \cdot \mathbf{n}}_{=0} v + \int_{\Gamma_D} \boldsymbol{\nabla} u \cdot \mathbf{n} \underbrace{v}_{=0} + \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v.$$

Problem: Find $u \in H^1_{\Gamma_D}$ *s.t.* $\forall\, v \in H^1_{\Gamma_D}$ *s.t.*

$$\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega fv.$$

This problem is WP because we have Poicaré on $H^1_{\Gamma_D}$.

Now, consider the true problem: Find $u \in X = H_0^1$ *s.t.*

$$\int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega f v \quad \forall\, v \in X \implies a(u,v) = \mathcal{F}(v).$$

Assumption $a(u,v) = a(v,u)$ (symmetry) and all LM assumptions.

The Galerkin Methods solve the numerical problem in $X_N \subset X$: Find $u_N \in X_N$ *s.t.*

$$a(u_N, v_N) = \mathcal{F}(v_N), \quad \forall\, v_N \in X_N.$$

This problem is WP, which can be trivially proven by LM assumptions.

Proposition 2.3 *This problem is strongly consistent (consistency error is* 0*).*

Proof 2. Note that $a(u,v) = \mathcal{F}(v)$ and $a(u_N, v_N) = \mathcal{F}(v_N)$. Then,

$$a(u, v_N) = \mathcal{F}(v_N) \quad \text{because } X_N \subset X.$$

The consistence error is 0. So, the problem is strongly consistent. We could also show

$$a(u - u_N, v_N) = 0 \quad \forall\, v_N \in X_N.$$

Q.E.D. ■

Proposition 2.4 *This problem is stable.*

Proof 3.

$$\alpha \|u_N\|_{H^1}^{\cancel{2}} \le a(u_N, u_N) \le \|f\| \cdot \cancel{\|u_N\|_{H^1}}.$$

So, it is clearly stable. Q.E.D. ■

Lemma 2.5 Cèa Lemma

$$\|u - u_N\|_{H^1} \le K \inf_{w_N \in X_N} \|u - w_N\|_{H^1}.$$

Proof 4.

$$\alpha \|u - u_N\|_{H^1}^2 \le a(u - u_N, u - u_N) = a(u - u_N, u - w_N) + a(u - u_N, w_N - u_N)$$

[*Bilinearity,* $w_N \in X_N$]

$$\alpha \|u - u_N\|_{H^1}^{\cancel{2}} \le a(u - u_N, u - w_N) \le M \cancel{\|u - u_N\|_{H^1}} \cdot \|u - w_N\|_{H^1}$$

[*Continuity*]

$$\|u - u_N\|_{H^1} \le \frac{M}{\alpha} \|u - w_N\|_{H^1}$$

$$\|u - u_N\|_{H^1} \le K \|u - w_N\|_{H^1} \quad \forall\, w_N \in X_N$$

$$\|u - u_N\|_{H^1} \le K \inf_{w_N \in X_N} \|u - w_N\|_{H^1}.$$

Q.E.D. ■

This Lemma implies that Galerkin solution is not the best one, but its H_1 norm error goes to 0 as quick as the best one. i.e., it has the same rate of convergence as the best approximation in X_N. When

$N \to \infty$, $X_N \to X$, and $\inf_{w_N \in X_n} \|u - w_N\|_{H^1} \to 0$.

Claim 2.6 P_N : Find $u_N \in X_N$, $a(u_N, v_n) = \mathcal{F}(v_N) \quad \forall\, v_N \in X_N$ is WP.

Proof 5. We can write $u_N = \sum_{j=0}^{N-1} c_j \varphi_j(\mathbf{x})$, where $\varphi_j(\cdot)$ are basis functions. Then, $X_n = \operatorname{span}\{\varphi_j\}_{j=0}^{N-1}$.

So,

$$a(u_N, v_N) = a\left(\sum_{j=0}^{N-1} c_j \varphi_j(\mathbf{x}), v_N\right) = \mathcal{F}(v_N)$$

$$\sum_{j=0}^{N-1} c_j a(\varphi_j, v_N) = \mathcal{F}(v_N) \qquad [\textit{Bilinearity}]$$

Instead of testing on $v_N \in X_N$, as $X_N = \operatorname{span}\{\varphi_j\}_{j=0}^{N-1}$, we just need to prove it works for all φ_i's.

$$\sum_{j=0}^{N-1} c_j a(\varphi_j, \varphi_i) = \mathcal{F}(\varphi_i) \quad [\textit{Take } v_N = \varphi_i]$$

Define

$$A \coloneqq \Big[a(\varphi_j, \varphi_i)\Big]_{i,j}, \quad \mathbf{c} = \Big[c_j\Big]_j, \quad \mathbf{b} = \Big[\mathcal{F}(\varphi_i)\Big]_i.$$

So, P_N becomes

$$A\mathbf{c} = \mathbf{b}.$$

Claim 2.7 A is non-singular

- A is symmetric: $a(\varphi_j, \varphi_i) = a(\varphi_i, \varphi_j)$ [*symmetry*]
- A is positive-definite: Assume $\mathbf{c} \neq 0$, then

$$\begin{aligned}
\mathbf{c}^\top A\mathbf{c} = \sum_{i,j} c_j A_{ij} c_i &= \sum_{i,j} c_j a(\varphi_j, \varphi_i) c_i \\
&= \sum_{i,j} a(c_j \varphi_j, c_i \varphi_i) && [\textit{bilinearity}] \\
&= a\left(\sum_{j=0}^{N-1} c_j \varphi_j, \sum_{i=0}^{N-1} c_i \varphi_i\right) && [\textit{bilinearity}] \\
&= a(u_N, u_N) && [u_N = \sum_{i=0}^{N-1} c_i \varphi_i] \\
&\geq \alpha \|u_N\|_{H^1}^2 \geq 0 && [\textit{coercivity}]
\end{aligned}$$

So, A is SPD, and thus A is non-singular. $\implies$ P_N is WP. Q.E.D. ■

Remark. What if the problem is non-linear?

$$-\Delta u + u^3 = f \implies \int_\Omega \nabla u \nabla v + \int_\Omega u^3 v = f.$$

- This problem is WP (although we don't have symmetry).
- Solving: linearize, then discretize:

$$\int_\Omega \nabla u^{(k+1)} \nabla v + \int_\Omega \left(u^{(k)}\right)^2 u^{(k+1)} v = f$$

  We can use root finding. For example, Newton's method.

- Solving: discretize, then linearize

$$u_N = \sum_{j=0}^{N-1} c_j \varphi_j, \quad X_N = \operatorname{span}\{\varphi\}_{j=0}^{N-1}.$$

  Then, we form a non-linear equation $F(\mathbf{c}) = 0$. We can also use Newton's method.

2.4 Finite Element: One Possible Choice of X_N

Recall: we need to compute

$$a(\varphi_j, \varphi_i) = \int_\Omega \nabla \varphi_j \nabla \varphi_i.$$

So, we want functions that are easy to differentiate and integrate. $\implies$ Polynomials!

Example 2.4.1 Runge Counterexample

Consider the function $f = \dfrac{1}{x^2+1}, \quad x \in [-5, 5]$.

If we interpolate $f(x)$ with euispaced points and polynomials,

$$\inf \|u - w_N\| \xrightarrow{N\to\infty} 0$$

will not be held.

Solutions:

- Optimize the position of nodes: Gaussian interpolation, Chebysev nodes.
- Piecewise interpolation: locally fit piecewise polynomial.

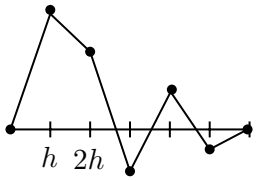
2.4.1 Finite Element in 1D

$$-u'' = f \implies \int_0^1 u'v' = \int_0^1 fv$$

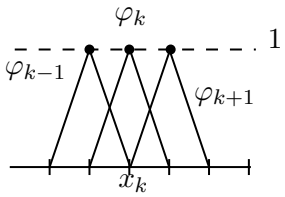
with $u(0) = u(1) = v(0) = v(1) = 0$, where

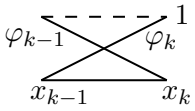
$$X_N = \{u_h \mid u_h \in \mathcal{C}([0,1]) \text{ and } u_h \in \mathbb{P}^1(I_h)\},$$

where $\mathbb{P}^1$ denotes polynomial of order 1. So, $X_N^1 \subset H_0^1(0,1)$, and $\dim(X_N) = N - 1 = \frac{1}{h} - 1$. So, when $N \to +\infty$, $h \to 0$, and $\dim(X_N) \to +\infty$.



When doing interpolation, let $y = f(x)$ be the true function, and $y_i = f(x_i)$ be the interpolant. There are many ways to construct the interpolant. Finite Element will use the *Lagrange Polynomial* (Hat functions):





$$\varphi_k(x) = \begin{cases} \dfrac{x - x_{k-1}}{x_k - x_{k-1}} = \dfrac{x - x_{k-1}}{h}, & x \in [x_{k-1}, x_k] \\ \dfrac{x_{k+1} - x}{x_{k+1} - x_k} = \dfrac{x_{k+1} - x}{h}, & x \in [x_k, x_{k+1}] \\ 0, & \text{o/w.} \end{cases}$$

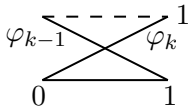
Note that

$$\varphi_i(x_k) = \delta_{ik} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases}$$

is the Kronecker-δ function. So, if $v_N = \sum_{i=0}^{N-1} y_i \varphi_i$, we have

$$v_N(x_k) = \sum_{i=0}^{N-1} y_i \varphi_i(x_k) = y_k.$$

FE In Practice We will set everything up on a reference interval $[0,1]$ with $\widehat{\varphi}_0 = 1 - \widehat{x}$ and $\widehat{\varphi}_1 = \widehat{x}$. We will use the bijiection $x = \text{map}(\widehat{x}) = a\widehat{x} + b$ such that $\text{map}(0) = x_{j-1}$ and $\text{map}(1) = x_j$. Then,



$$\varphi_j = \begin{cases} \text{map}(\widehat{\varphi}_0) & \text{in} \quad (x_{j-1}, x_j) \\ \text{map}(\widehat{\varphi}_1) & \text{in} \quad (x_j, x_{j+1}). \end{cases}$$

Also,

$$\widehat{x} = \frac{x - b}{a} = \frac{x - x_{j-1}}{h}.$$

Convergence Theorem and Its Application Let $u(x) \in \mathcal{C}^0([0,1])$. Denote $\Pi_u^1(x) \in X_h^1$ as

$$\Pi_u^1(x) = \big\{\text{continuous function, linear } s.t. \text{ on } I_k = [x_k, x_{k+1}],\ \Pi_u^1(I_k) \in \mathbb{P}^1\big\},$$

where $\mathbb{P}^1$ denotes polynomial of order 1. Then,

$$\Pi_u^1(x) = \sum_{j=0}^{N} u(x_j)\varphi_j(x_j).$$

To measure error, consider the following norms:

$$\big\|u - \Pi_u^1\big\|_{L^2} \quad \text{and} \quad \big\|\nabla(u - \Pi_u^1)\big\|_{L^2} = \underbrace{\big|u - \Pi_u^1\big|_{H^1}}_{\text{semi-norm}}$$

Theorem 2.4.2
If $u \in H^2([0,1])$, then

$$\begin{aligned} \big\|u - \Pi_u^1\big\|_{L^2} &\le C_0 |u|_{H^2} h^2 = C\big\|u''\big\|_{L^2} h^2 \\ \big|u - \Pi_u^1\big|_{H^1} &\le C_1 \big\|u''\big\|_{L^2} h \\ \implies \big\|u - \Pi_u^1\big\|_{H^1} &\le C\big\|u''\big\|_{L^2} h. \end{aligned}$$

Proof 1. $u \in H^2([0,1]) \implies u \in \mathcal{C}^1([0,1])$.

- Let's work locally on the interval $[x_j, x_{j+1}]$. Denote that

$$e = u - \Pi_u^1, \quad e(x_j) = 0, \quad \text{and } e(x_{j+1}) = 0.$$

By Rolle's Theorem, $\exists\, \xi_j\ s.t.\ e'(\xi_j) = 0$. By Fundamental Theorem of Calculus,

$$\begin{aligned} e'(x) &= \int_{\xi_j}^{x} e''(x)\,\mathrm{d}x = \int_{\xi_j}^{x} u''(x)\,\mathrm{d}x && [(\Pi_u^1)'' = 0] \\ \big|e'(x)\big| &\le \left|\int_{x_j}^{x_{j+1}} u''(x)\,\mathrm{d}x\right| \\ &\le \left(\int_{x_j}^{x_{j+1}} 1^2\right)^{1/2} \left(\int_{x_j}^{x_{j+1}} (u'')^2\right)^{1/2} = h^{1/2}\big\|u''\big\|_{L^2(I_j)}. && [\textit{Cauchy-Schwarz}] \\ \implies \big\|e'(x)\big\|_{L^2(I_j)}^2 &= \int_{x_j}^{x_{j+1}} \big|e'(x)\big|^2 \le h \cdot h\big\|u''\big\|_{L^2(I_j)}^2 = h^2\big\|u''\big\|_{L^2(I_j)}^2. \end{aligned}$$

Now, let's move to global:

$$\|e'\|_{L^2(I)}^2 = \sum_j \|e'\|_{L^2(I_j)}^2 \le h^2 \|u''\|_{L^2(I)}^2.$$
$$\boxed{|e|_{H^1} = \|e'\|_{L^2} \le Ch\|u''\|_{L^2} = Ch|u|_{H^2} \sim \mathcal{O}(h)}.$$

- Let's repeat the same argument:

$$|e(x)| = \left|\int_{x_j}^{x_{j+1}} e' \,\mathrm{d}x\right| \le \left(\int_{x_j}^{x_{j+1}} 1^2\right)^{1/2} \left(\int_{x_j}^{x_{j+1}} (e')^2\right)^{1/2}$$
$$e^2(x) = (e(x))^2 \le h\|e'\|_{L^2(I_j)}$$
$$\int_{x_j}^{x_{j+1}} e^2(x) \le h^2\|e'\|_{L^2(I_j)}^2 \le h^4\|u''\|_{L^2(I_j)} \qquad [\textit{Use conclusion from ①}]$$

Globally,

$$\|e\|_{L^2}^2 = \sum_j \|e\|_{L^2(I_j)}^2 \le h^4 |u''|_{L^2(I)}^2$$
$$\implies \boxed{\|e\|_{L^2} \le Ch^2\|u''\|_{L^2} \sim \mathcal{O}(h^2)}.$$

Q.E.D. ■

Applications of Theorem 2.4.2:

- Consider FE problem $a(u_h, v_h) = \mathcal{F}(v_h)$. For $u_h \in X_h^1$, we have

$$\|u - u_h\|_{H^1} \le \frac{M}{\alpha} \inf_{w_h \in X_h^1} \|u - w_h\|_{H^1} \qquad [\textit{Cea Lemma}]$$
$$\le \frac{M}{\alpha}\|u - \Pi_{u,h}^1\|_{H^1} \qquad [\textit{Theorem 2.4.2}]$$
$$\le \frac{M}{\alpha} h\|u''\|_{H^2}$$
$$\|u - u_h\|_{L^2} \le \mathcal{O}(h^2).$$

- If we know the error is bad locally, we can locally refine h to get better error.

Advection-Reaction Problem, Local Assembly, and Global Assembly

$$\begin{cases} -u'' + \sigma u = f, \quad \sigma > 0 \\ u(0) = u(1) = 0. \end{cases}$$

Weak formulation:

$$\int_0^1 u'v' + \int_0^1 \sigma uv = \int_0^1 fv \quad \forall\, v \in H_0^1.$$

- Matris A:

$$A_{ij} = a(\varphi_i, \varphi_j) = \int_0^1 \varphi_j' \varphi_i' + \underbrace{\int_0^1 \sigma \varphi_j \varphi_i}_{\text{mass matrix}}$$

$$b_i = \int_0^1 f \varphi_i$$

How to integrate numerically? We do it interval by interval. Suppose we have first order Lagrange polynomials. Then,

$$\int_0^1 \sum_j \int_{I_j},$$

where

$$\int_{I_k} = \underbrace{\int_{I_k} \varphi_i' \varphi_j'}_{\text{constant}} + \int_{I_k} \sigma \varphi_i \varphi_j.$$

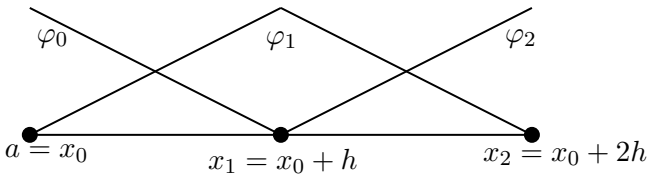
Note that

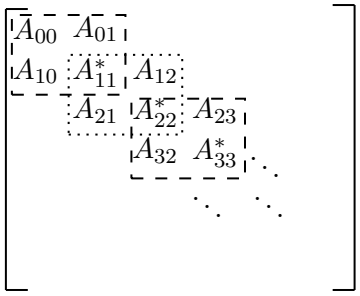
$$\int_{I_k} \varphi_\ell' \varphi_r' = 0 \quad \text{if} \quad \ell, r \neq k, k+1.$$

So, the matrix will be tridiagonal, and

$$A_{ij} = \int_0^1 \varphi_j' \varphi_i' + \int_0^1 \sigma \varphi_j \varphi_i = \int_{x_{j-1}}^{x_{j+1}} \varphi_j' \varphi_i' + \int_{x_{j-1}}^{x_{j+1}} \varphi_j \varphi_i.$$

Let's only consider the first interval:





$$A_{00} = \int_{x_0}^{x_1} \varphi_0' \varphi_0' + \sigma \int_{x_0}^{x_1} \varphi_0 \varphi_0$$

$$A_{01} = \int_{x_0}^{x_1} \varphi_0' \varphi_1' + \sigma \int_{x_0}^{x_1} \varphi_0 \varphi_1$$

$$A_{11} = \int_{x_0}^{x_2} \varphi_1' \varphi_1' + \sigma \int_{x_0}^{x_2} \varphi_1 \varphi_1 = \underbrace{\int_{x_0}^{x_1} \varphi_1' \varphi_1' + \sigma \int_{x_0}^{x_1} \varphi_1 \varphi_1}_{=: A_{11}^*} + \underbrace{\int_{x_1}^{x_2} \varphi_1' \varphi_1' + \sigma \int_{x_1}^{x_2} \varphi_1 \varphi_1}_{\text{will be calculated in the next interval}}$$

$$A_{10} = \int_{x_0}^{x_1} \varphi_1' \varphi_0' + \sigma \int_{x_0}^{x_1} \varphi_1 \varphi_0 = A_{01}.$$

We will get the rest of A_{11} in the second interval.

- Local assembly: we will form the 2×2 matrix locally:

$$\begin{aligned} a_{00} &= a(\widehat{\varphi}_0, \widehat{\varphi}_0)_{\text{local}} \\ a_{10} &= (\widehat{\varphi}_0, \widehat{\varphi}_1)_{\text{local}} = a_{01} \\ a_{11} &= a(\widehat{\varphi}_1, \widehat{\varphi}_1)_{\text{local}} \end{aligned}$$

We can compute the integrals on a reference interval $[0, 1]$:

$$a_{ii} = \int_{x_i}^{x_{i+1}} \varphi_i' \varphi_i' + \sigma \int_{x_i}^{x_{i+1}} \varphi_i \varphi_i$$

Consider the map: $x = h\widehat{x} + x_i \in [0, 1]$.

$$\frac{\partial \varphi_i}{\partial x} = \frac{\partial \widehat{\varphi}_i}{\partial \widehat{x}} \cdot \frac{\partial \widehat{x}}{\partial x}, \quad \frac{\mathrm{d}x}{\mathrm{d}\widehat{x}} = h.$$

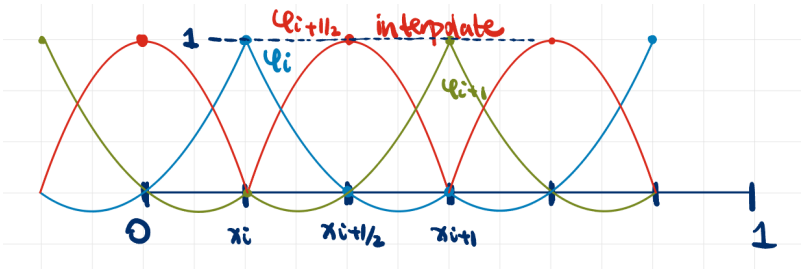
Then,

$$a_{ii} = \int_0^1 \frac{\partial \widehat{\varphi}_i}{\partial \widehat{x}} \cdot \frac{\partial \widehat{x}}{\partial x} \cdot \frac{\partial \widehat{\varphi}_i}{\partial \widehat{x}} \cdot \frac{\partial \widehat{x}}{\partial x} h \, \mathrm{d}\widehat{x} + \sigma \int_0^1 \widehat{\varphi}_i \widehat{\varphi}_i h \, \mathrm{d}\widehat{x}.$$

Now, everything is pre-computable. To compute these integrals efficiently, we will use *Gaussian Quadrature*.

- Global assembly: additivity of integrals.
- Treatment of the boundary conditions: after global assembly, we will impose the BCs.

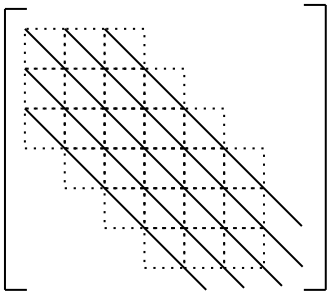
Use Quadratic Functions Instead



$$\begin{aligned} \varphi_i &= \frac{(x - x_{i+1})(x - x_{i+1/2})}{(x_i - x_{i+1})(x_i - x_{i+1/2})} \\ \varphi_{i+1/2} &= \frac{(x - x_i)(x - x_{i+1})}{(x_{i+1/2} - x_i)(x_{i+1/2} - x_{i+1})} \\ \varphi_{i+1} &= \frac{(x - x_i)(x - x_{i+1/2})}{(x_{i+1} - x_i)(x_{i+1} - x_{i+1/2})} \end{aligned}$$

- Local assembly: 3×3 matrices.

- Global assembly: pena-diagonal (five diagonal entries):



Convergence and Error

$$\begin{aligned}
\mathbb{P}^1: &\quad e_{H^1} \sim \mathcal{O}(h) \qquad e_{L^2} \sim \mathcal{O}(h^2) \\
\mathbb{P}^2: &\quad e_{H^1} \sim \mathcal{O}(h^2) \qquad e_{L^2} \sim \mathcal{O}(h^3) \\
\mathbb{P}^3: &\quad e_{H^1} \sim \mathcal{O}(h^3) \qquad e_{L^2} \sim \mathcal{O}(h^4)
\end{aligned}$$

Theorem 2.4.3 Convergence of FE in 1D
Suppose $u \in H^{s+1}(0,1)$ with $s \geq 1$. For FE, we work in X_h^q. Then,

$$\begin{aligned}
\|e\|_{H^1} &\leq C\|u\|_{H^{\min(q,s)}} h^{\min(q,s)} \\
\|e\|_{L^2} &\leq C\|u\|_{H^{\min(q,s)}} h^{\min(q,s)+1}.
\end{aligned}$$

Remark. This implies that increasing the degree of polynomials in FE makes sense only when the solution is regular enough.

Table 1: Summary of Rate of Convergence in H^1, Given Different q and s

	$s=1$	2	3	4	5	$\cdots$
$q=1$	$\boxed{1}$	1	1	1	1	
2	1	$\boxed{2}$	2	2	2	
3	1	2	$\boxed{3}$	3	3	
4	1	2	3	$\boxed{4}$	4	
$\vdots$						

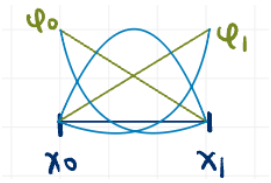
Optimal choices: with the regularity we have (s), take the minimal optimal polynomial degree (q). There are two ways to reduce error: (1) reduce h, and (2) increase q.

Reusing Matrices

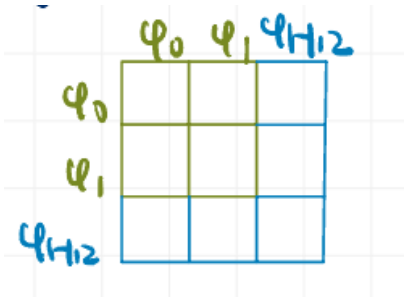
Set-up We have built linear FE matrices

$$A_L \mathbf{u}_L = \mathbf{b}_L.$$

Goal Set up quadratic FE ($A_q \mathbf{u}_Q = \mathbf{b}_Q$) from reusing A_L



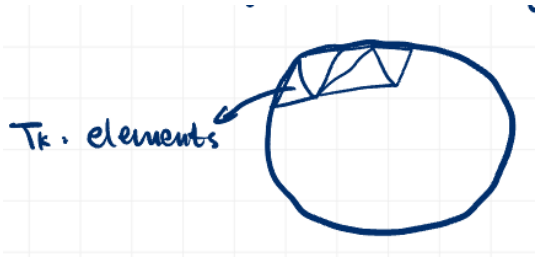
$$\begin{aligned} \varphi_0 &= 1 - \widehat{x} \\ \varphi_1 &= \widehat{x} \end{aligned} \implies \varphi_{H,2} = \varphi_0 \varphi_1 = (1-\widehat{x})\widehat{x}$$



Conclusion: What is a Finite Element? We have an interval I. A finite element is built on $K \in \mathcal{P}(I)$, some partition of I. On each K, we build basis polynomials of degree q. i.e., $u(K) \in \mathbb{P}^q$.

2.4.2 Finite Elements in Multiple Dimension

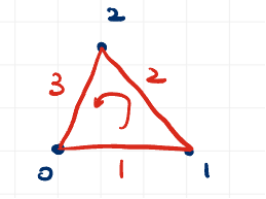
Let's divide the region into triangles:



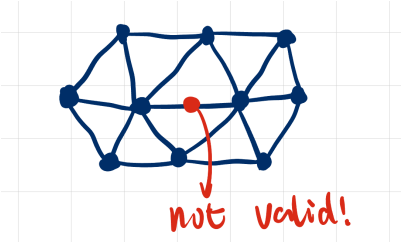
$$\Omega = \bigcup_{k=1}^{N} T_k,$$

though this relationship could be false in reality.

Triangle = Vertex + Edge



Requirement Vertices are not on edges of another triangle

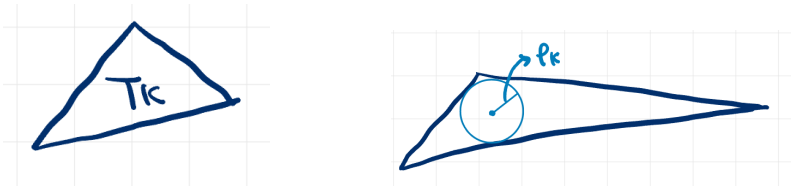


Property If $v \in \mathcal{C}^0(\overline{\Omega})$ and $\in H^k(T_k)$, then $v \in H^k(\Omega)$

How large is the mesh? We have two measurements:

- $h = \max |\mathbf{x}_1 - \mathbf{x}_2|$ for $\mathbf{x}_1, \mathbf{x}_2 \in T_k$

- $\dfrac{h_k}{\rho_k} < \delta$, independent of k.



(a) (b)

Recall the problem:

$$\begin{cases} -\Delta u = f \\ u(\partial\Omega) = 0 \end{cases} \implies \int \nabla u \nabla \varphi = \int f\varphi,$$

we aim to build

$$\Omega = \bigcup_{k=1}^{N} T_k \qquad \text{and} \qquad X_k^h = \left\{ v_n \in H^1 : v_h(T_k) \in \mathbb{P}^k \right\}.$$

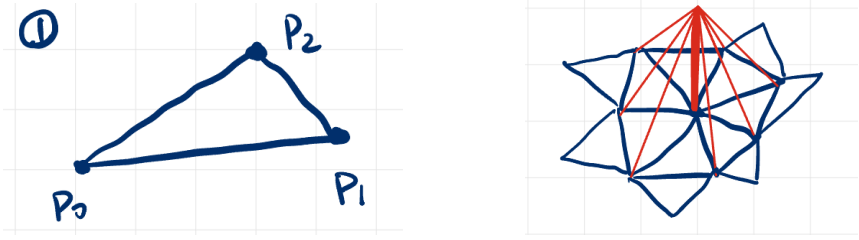
Linear Polynomial

$$ax + by + c \qquad \left(\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \right).$$

We could rewrite

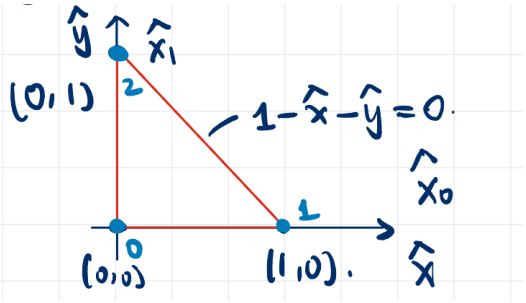
$$P_k(\mathbf{x}) - P_k(P_j) = c_j,$$

which will give us a "tent" function.



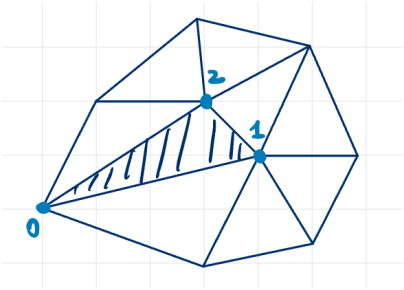
(c) (d)

We will work on a reference interval:



$$\widehat{\varphi}_0 = 1 - \widehat{x} - \widehat{y}$$
$$\widehat{\varphi}_1 = \widehat{x}$$
$$\widehat{\varphi}_2 = \widehat{y}.$$

$$\mathbf{x} = B_k \widehat{\mathbf{x}} + \mathbf{c}_k$$



$$x(\widehat{x},\widehat{y}) = \begin{cases} x(0,0) = x_0 \\ x(1,0) = x_1 \\ x(0,1) = x_2. \end{cases}$$

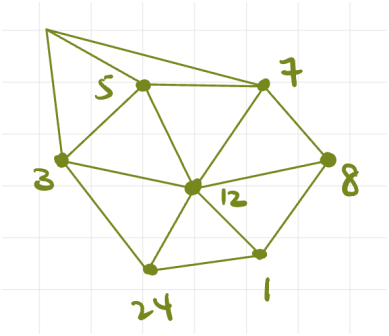
$$x = x_0\widehat{\varphi}_0(\widehat{x},\widehat{y}) + x_1\widehat{\varphi}_1(\widehat{x},\widehat{y}) + x_2\widehat{\varphi}_2(\widehat{x},\widehat{y})$$
$$y = y_0\widehat{\varphi}_0(\widehat{x},\widehat{y}) + y_1\widehat{\varphi}_1(\widehat{x},\widehat{y}) + y_2\widehat{\varphi}_2(\widehat{x},\widehat{y})$$

Then, we can build the matrix by

$$A_{ij} = \int_\Omega \nabla\varphi_j \nabla\varphi_i = \sum_k \int_{T_k} \nabla\varphi_j \nabla\varphi_i.$$

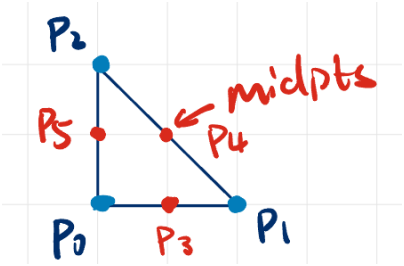
This matrix is symmetric, but we don't have clear pattern.

Example 2.4.4



Node 12 sees nodes $1, 8, 7, 5, 3$, and 24. With a different mesh generation and numbering, we have a different matrix structure.

$\mathbb{P}^2$ Functions



$$ax^2 + bxy + cy^2 + dx + ey + f$$

$$\widehat{\varphi}_0 = 2\underbrace{(1-\widehat{x}-\widehat{y})}_{\text{line b/w} P_1P_2}\underbrace{\left(\frac{1}{2}-\widehat{x}-\widehat{y}\right)}_{\text{line b/w} P_3P_5}$$
$$\widehat{\varphi}_1 = 2\widehat{x}\left(\widehat{x}-\frac{1}{2}\right)$$
$$\widehat{\varphi}_2 = 2\widehat{y}\left(\widehat{y}-\frac{1}{2}\right)$$
$$\widehat{\varphi}_3 = 4(1-\widehat{x}-\widehat{y})\widehat{x}$$
$$\widehat{\varphi}_4 = 4\widehat{x}\widehat{y}$$
$$\widehat{\varphi}_5 = 4(1-\widehat{x}-\widehat{y})\widehat{y}$$

On a general element, $\mathbf{x} = B_k\widehat{\mathbf{x}} + \mathbf{c}_k$. Transforming the Integral using Chain rule:

$$\left.\begin{aligned}\frac{\partial\varphi}{\partial x} &= \frac{\partial\varphi}{\partial\widehat{x}}\cdot\frac{\partial\widehat{x}}{\partial x} + \frac{\partial\varphi}{\partial\widehat{y}}\cdot\frac{\partial\widehat{y}}{\partial x}\\ \frac{\partial\varphi}{\partial y} &= \frac{\partial\varphi}{\partial\widehat{x}}\cdot\frac{\partial\widehat{x}}{\partial y} + \frac{\partial\varphi}{\partial\widehat{y}}\cdot\frac{\partial\widehat{y}}{\partial y}\end{aligned}\right\}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\varphi = B_k^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\widehat{\varphi}.$$

Since $\mathbf{x} = B_k\widehat{\mathbf{x}} + \mathbf{c}$, we have $\widehat{\mathbf{x}} = B_k^{-1}\widehat{\mathbf{x}} + \mathbf{d}$. So,

$$\frac{\partial\widehat{x}}{\partial x} = \left(B_k^{-1}\right)_{1,1},\quad \frac{\partial\widehat{x}}{\partial y} = \left(B_k^{-1}\right)_{1,2}$$

$$\begin{aligned}\int_K \boldsymbol{\nabla}\varphi_j\cdot\boldsymbol{\nabla}\varphi_i\,\mathrm{d}\mathbf{x} &= \int_{\widehat{K}_\triangle}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\widehat{\varphi}_j J_k^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\widehat{\varphi}_i J_k^{-1}|J_k|\,\mathrm{d}\widehat{\mathbf{x}}\\ &= \int_{\widehat{K}_\triangle}\left(B_k^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\widehat{\varphi}_i\right)\left(B_k^{-1}\boldsymbol{\nabla}_{\widehat{x},\widehat{y}}\widehat{\varphi}_j\right)|\det(B_k)|\,\mathrm{d}\widehat{\mathbf{x}}\end{aligned}$$

We can also transform the semi-norm:

$$\begin{aligned}\int_K(\boldsymbol{\nabla}u)^2\,\mathrm{d}x &= \int_{\widehat{K}}\left(B_k^{-1}\widehat{\boldsymbol{\nabla}}u\right)^2|\det B_k|\,\mathrm{d}\widehat{x}\\ |u|^2_{H^1(K)} &\le \left\|B_k^{-1}\right\|^2|\det B_k||u|^2_{H^1(\widehat{K})}\\ |u|_{H^1(K)} &\le \left\|B_k^{-1}\right\||\det B_k|^{\frac{1}{2}}|u|_{H^1(\widehat{K})}\\ |u|_{H^1(\widehat{K})} &\le \|B_k\||\det B_k|^{-\frac{1}{2}}|u|_{H^1(K)}\\ &\qquad\text{[semi-norms in } H^1(K) \text{ and in } H^1(\widehat{K}) \text{ are equivalent]}\\ |u|_{H^m(K)} &\le \left\|B_k^{-1}\right\|^m|\det B_k|^{\frac{1}{2}}|u|_{H^m(\widehat{K})}.\end{aligned}$$

Bounding $\|B_k\|$: $\widehat{x}\to x$:

$$\begin{aligned}\|B_k\| &= \sup_{\|\xi\|=\widehat{\rho}}\frac{\|B\xi\|}{\widehat{\rho}} \le \frac{h_k}{\widehat{\rho}}\\ \left\|B_k^{-1}\right\| &\le \cdots \le \frac{\widehat{h}}{\rho_k} = \frac{\sqrt{2}}{\delta_k}.\end{aligned}$$

Theorem 2.4.5 Convergence of FE in Multiple Dimension
Suppose $u\in H^{s+1}$ and $u_h\in\mathbb{P}^p = X_h^p$. The error is given by

$$\|u-u_h\|_{H^1}\le Ch^q|u|_{H^{q+1}},$$

where $q=\min(p,s)$.

Proof 2. Strategies:

1. Global $\to$ Local: K

2. $K \to \widehat{K}$
3. Error in $\widehat{K}$
4. Go back to $\widehat{K}$ use previous inequalities
5. $K \to \Omega = \bigcup_{k=1}^{\infty} T_k$

Recall some results: Ceà Lemma:

$$\|u - u_h\| \le C \inf_{w_h \in V_h} \|u - w_h\| \le C\big\|u - \Pi_h^p u\big\|,$$

where Π_h^p is the polynomial interpolation of order p. To move from ② $\to$ ③ in our strategy:

$$\int_K (\boldsymbol{\nabla} u - \boldsymbol{\nabla}\Pi_u^p u)^2 \,\mathrm{d}\mathbf{x} \to \int_{\widehat{K}} \left(\boldsymbol{\nabla}\widehat{u} - \boldsymbol{\nabla}\Pi_h^p \widehat{u}\right)^2 \mathrm{d}\widehat{\mathbf{x}}$$

Lemma 2.6 Bramble-Hilbert $u \in H^{r+1}$, $\mathcal{L}(u) : H^{r+1} \to H^m$, $\mathcal{L}(p) = 0 \quad \forall\, p \in \mathbb{P}^r$. Then,

$$\|\mathcal{L}(v)\|_{H^m} \le C \inf_{p\in\mathbb{P}} \|u + p\|_{H^{r+1}}$$

Proof.

$$\|\mathcal{L}(v)\| \le \|\mathcal{L}\| \cdot \|v\|$$

Since $\mathcal{L}(p) = 0$, we know that

$$\|\mathcal{L}(v+p)\| \le \|\mathcal{L}\| \cdot \|v+p\| \le \|\mathcal{L}\| \inf_{p\in\mathbb{P}} \|v+p\|.$$

□

Lemma 2.7 Deny-Lions

$$\inf \|u + p\|_{H^{r+1}} \le C|u|_{H^{r+1}}.$$

So, on $\widehat{K}$, we have

$$\big|u - u_{\widehat{h}}\big|_{H^m(\widehat{K})} \le C|u|_{H^{r+1}(\widehat{K})}.$$

Let's go to K: Recall that

$$|u|_{H^{r+1}(\widehat{K})} \le \|B_k\| \cdot |\det B_k|^{-\frac{1}{2}} |u|_{H^{r+1}(K)}$$
$$|u|_{H^{r+1}(K)} \le \big\|B_k^{-1}\big\| \cdot |\det B_k|^{\frac{1}{2}} |u|_{H^{r+1}(\widehat{K})}.$$

So,

$$|e|_{H^m(K)} \le C\big\|B_k^{-1}\big\|^m \cancel{|\det B_k|^{\frac{1}{2}}} \|B_k\|^{r+1} \cancel{|\det B_k|^{-\frac{1}{2}}} |u|_{H^{r+1}(K)}.$$

Note that

$$\big\|B_k^{-1}\big\| \le \frac{h_k}{\rho_k} \quad \text{and} \quad \|B_k\| \le \frac{h_k}{\widehat{\rho}},$$

we have that

$$\begin{aligned}|e|_{H^m(K)} &\le C\left(\frac{h_k}{\rho_k}\right)^m h_k^{r+1-m}|u|_{H^{r+1}(K)} \\ &\quad [\textit{Assumption: mesh is regular, so } \frac{h_k}{\rho_k} < \delta \quad (\perp\!\!\!\perp\ k)] \\ &\le C\delta^m h_k^{r+1-m}|u|_{H^{r+1}(K)} \\ &= \widetilde{C} h_k^{r+1-m}|u|_{H^{r+1}(K)}.\end{aligned}$$

- If $m = 1$:

$$|e|_{H^1(K)} \le Ch^r|u|_{H^{r+1}(K)}.$$

- If $m = 0$:

$$|e|_{L^2(K)} \le Ch^{r+1}|u|_{H^{r+1}(K)}.$$

Q.E.D. ■

Proposition 2.8

$$\|e_{FEM}\|_{L^2} \le Ch^{r+1}|u|_{H^{r+1}}.$$

Proof 3. Introduce the auxiliary problem (Aubin-Nitsche Trick):

$$\begin{cases} -\Delta u = f \\ u(\partial\Omega) = 0 \end{cases} \implies \begin{cases} -\Delta\varphi = e \\ \varphi(\partial\Omega) = 0. \end{cases}$$

Elliptic regularity:

$$\begin{cases} -\Delta u = f \in L^2 \\ u(\partial\Omega) = 0 \end{cases} \implies \|u\|_{H^2\cap H_0^1} \le C\|f\|_{L^2} \quad (\|u\|_{H^2} \le C\|\Delta u\|_{L^2}).$$

In weak formulation, $a(\varphi, v) = (e_h, v)$. So,

$$\begin{aligned}\|e_h\|^2 &= a(\varphi, e_h) = a(e_h, \varphi) = a(e_h, \varphi - \varphi_h) \\ \|e_h\|_{L^2}^{\not 2} &\le Ch|\varphi|_{H^2}\|e_h\|_{H^1} && [\textit{Interpolation error, } \|\varphi - \varphi_h\|_{H^1} \le Ch|\varphi|_{H^2}] \\ &\le \cancel{\|e_h\|_{L^2}} h\|e_h\|_{H^1} && [-\Delta\varphi = e_h \implies |\varphi|_{H^2} \le C\|e_h\|_{L^2}] \\ \|e_h\|_{L^2} &\le h\|e_h\|_{H^1}.\end{aligned}$$

Q.E.D. ■

Proposition 2.9 *The matrix A is SPD, sparse, and* $\mathrm{cond}_2(A) \sim \mathcal{O}(h^{-2})$.

Remark. Since A is SPD, solving $A\mathbf{x} = \mathbf{b}$ is equivalent to

$$\min \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top\mathbf{x}.$$

We could solve using CG:

$$\left\|\mathbf{e}_{\text{CG}}^{(k+1)}\right\| \leq \frac{\sqrt{\text{cond}_2(A)}-1}{\sqrt{\text{cond}_2(A)}+1}\left\|\mathbf{e}_{\text{CG}}^{(k)}\right\|.$$

To accelerate convergence, use pre-conditioner P^{-1} $s.t.$

$$\text{cond}_2\left(P^{-1}A\right) \ll \text{cond}_2(A).$$

We solve $P^{-1}A\mathbf{x} = P^{-1}\mathbf{b}$ instand.

Proof 4. (of Proposition 2.9)

$$\text{cond}_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

In Rayleigh coefficient form, eigenvalues are

$$\frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} = \frac{\mathbf{x}^\top(\lambda\mathbf{x})}{\mathbf{x}^\top\mathbf{x}} = \lambda,$$

where $\mathbf{x}^\top A\mathbf{x} = a(v_h, v_h)$. Note that $v_h = \sum_i x_i\varphi_i$. Let d denote the dimension of the problem. Then,

$$\alpha_1 h^d \mathbf{x}^\top\mathbf{x} \leq \|v_h\|_{L^2}^2 \leq \alpha_2 h^2 \mathbf{x}^\top\mathbf{x} \tag{1}$$

Also, inverse inequality gives

$$c_1\|v_h\| \leq \|\boldsymbol{\nabla} v_h\|_{L^2} \leq ch^{-1}\|v_h\|_{L^2} \tag{2}$$

Combine the two inequalities,

$$c_1\frac{h^d\cdot\cancel{\|v_h\|_{L^2}^2}}{\cancel{\|v_h\|_{L^2}^2}} \leq \frac{a(v_h,v_h)}{\mathbf{x}^\top\cdot\mathbf{x}} \leq c_2\frac{h^d\|v_h\|_{H^1}^2}{\|v_h\|_{L^2}^2} \qquad \left[\begin{array}{c}\text{① } \textit{Divid (1) by } \mathbf{x}^\top\mathbf{x} \\ \text{② } \textit{By continuity}\end{array}\right]$$

$$\leq c_2\frac{h^d h^{-2}\cancel{\|v_h\|_{L^2}^2}}{\cancel{\|v_h\|_{L^2}^2}} \qquad \left[\begin{array}{c}\text{① } \textit{Poicaré: } \|v_h\|_{H^1}^2 \leq C\|\boldsymbol{\nabla} v_h\|_{L^2}^2 \\ \text{② } \textit{By (2)}\end{array}\right]$$

So,

$$c_1 h^d \leq \underbrace{\frac{a(v_h,v_h)}{\mathbf{x}^\top\mathbf{x}}}_{\text{eigenvalue}} \leq c_2 h^{d-2}.$$

So, $\lambda_{\min} = c_1 h^d$ and $\lambda_{\max} = c_2 h^{d-2}$. Then,

$$\text{cond}_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{c_2 h^{d-2}}{c_1 h^d} = ch^{-2}.$$

Q.E.D. ■

2.5 Mixed Problems

$$\begin{cases} -\Delta u = f \\ u(\partial u) = 0. \end{cases}$$

Post-processing: we are also interested in $\boldsymbol{\nabla} u$.

$$u_h = \sum_i v_i \varphi_i(x,y,z) \quad \text{from FE}$$
$$\boldsymbol{\nabla} u_h = \sum_i v_i \boldsymbol{\nabla} \varphi_i.$$

Problem with this method:

- No control over accuracy:
$$u_h \sim \mathcal{O}(h^2) \implies \boldsymbol{\nabla} u_h \sim \mathcal{O}(h).$$
- $\boldsymbol{\nabla} u_h$ is usually not continuous on the domain.

What to do? Auxiliary problem

$$\begin{cases} \boldsymbol{\nabla} u - \boldsymbol{\sigma} = 0 \qquad (\boldsymbol{\nabla} u = \boldsymbol{\sigma}) \\ -\boldsymbol{\nabla} \cdot \boldsymbol{\sigma} = f. \end{cases}$$

- This problem is also a minimization problem:

$$J^* = \int_\Omega |\mathbf{p}|^2, \qquad W_f = \underbrace{\{\boldsymbol{\nabla} \cdot \mathbf{w} + f = 0\}}_{\text{constraint}}$$

$$\min_{\mathbf{p} \in W_f} J^*.$$

Lagrangian multiplier: $\mathcal{L} = \int_\Omega |\mathbf{p}|^2 + \int v(\boldsymbol{\nabla} \cdot \mathbf{p} + f)$. Then, the problem becomes

$$\min_{\mathbf{p}} \max_{v} \mathcal{L} \qquad \leftarrow \text{saddle point problem}$$

$\max_v$: pushing v to be large, we force $\boldsymbol{\nabla} \cdot \mathbf{p} + f = 0$.

- FOC: Find $\boldsymbol{\sigma} \in H(\text{div}; \Omega) \equiv \left\{ \sigma_i \in L^2, \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} = \sum \frac{\partial \sigma_i}{\partial x_i} \in L^2 \right\} = Y$, $u \in L^2(\Omega) = X$ $s.t.$

$$\begin{cases} \int \boldsymbol{\sigma} \cdot \mathbf{p} + \int \boldsymbol{\nabla} \cdot \mathbf{p} u = 0 \\ \int_\Omega v(\boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + f) = 0 \end{cases} \qquad \forall\, v \in X, \mathbf{p} \in Y.$$

One can verify that

$$\begin{cases} -\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = f \\ \boldsymbol{\sigma} - \boldsymbol{\nabla}\mathbf{u} = 0 \end{cases} \implies \begin{array}{l} \int v(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f) = 0 \\ \int (\boldsymbol{\sigma} - \boldsymbol{\nabla}\mathbf{u})\cdot\mathbf{p} = 0. \end{array} \quad \textit{[Multiply by test functions and integrate]}$$

For the second integral, consider integration by parts: $\int \boldsymbol{\sigma}\cdot p + \int \boldsymbol{\nabla}\cdot\mathbf{p}u = 0.$

- The Lagrangian multiplier problem is well-posed.

  Proposition 2.1 $\inf-\sup$ **Condition/LBB** $\forall\, v \in X, \quad \exists\, \mathbf{p} \in Y \ s.t. \ \exists \beta > 0 \ s.t.$

$$\left| \int_\Omega v\boldsymbol{\nabla}\cdot\mathbf{p} \right| \geq \beta \|v\|_X \|\mathbf{p}\|_Y,$$

  where $\|\mathbf{p}\|_{H(\mathrm{div})} = \|\mathbf{p}\|_{L^2} + \|\boldsymbol{\nabla}\cdot\mathbf{p}\|_{L^2}$

$$\iff \inf_{\mathbf{p}\in Y} \sup_{v\in X} \frac{\left| \int_\Omega v\boldsymbol{\nabla}\cdot\mathbf{p} \right|}{\|v\|_X \|\mathbf{p}\|_Y} \geq \beta.$$

  Proof 1. (Sketch) Consider the auxiliary problem:

$$\begin{cases} -\Delta\psi = v \\ \psi(\partial\Omega) = 0, \end{cases} \qquad \mathbf{p} = \boldsymbol{\nabla}\psi.$$

  Q.E.D. ■

  With the inf-sup condition, the Lagrangian multiplier problem is well-posed.

- Is the numerical problem well-posed?

  Consider $X_h \subset X$ and $Y_h \subset Y$. Then,

$$\begin{cases} \int_\Omega v_h(\boldsymbol{\nabla}\cdot\mathbf{p}_h + f) = 0 \\ \int_\Omega \mathbf{p}_h\cdot\boldsymbol{\sigma}_h + (\boldsymbol{\nabla}\cdot\mathbf{p}_h)v_h = 0. \end{cases}$$

  We cannot select X_h and Y_h as we wish. We have to choose X_h and Y_h $s.t.$ the inf-sup condition holds. i.e., $\forall\, v_h \in X_h, \quad \exists\, \mathbf{p}_h \in Y_h \ s.t.$

$$\left| \int_\Omega \boldsymbol{\nabla}\cdot\mathbf{p}_h v_h \right| \geq \beta \|v_h\|_X \|\mathbf{p}_h\|_Y.$$

- Reviart-Thomas:

$$\begin{cases} v_h \in \mathbb{P}^{k-1} \\ \mathbf{p}_h \in \mathbb{D}^k \equiv \left[\mathbb{P}^{k-1}\right]^d + \mathbf{x}\mathbb{P}^{k-1}. \end{cases}$$

This couple is inf-sup compatible.

$$\|v - v_h\|_{L^2} - \|\mathbf{p} - \mathbf{p}_h\|_{H^d} \sim \mathcal{O}(h^k).$$

There are other compatible couples, such as DBM.

Remark 2. (Messages from this Section).

- We can reformulate a second order problem as a system of first order problems, but the requirement of regularity is different.
- In first order formulation, we lose coercivity, and we have an inf-sup (min-max/saddle point) problem. So, we need to find the right pair for discretization.

2.6 Remarks on FE

- Why triangles?

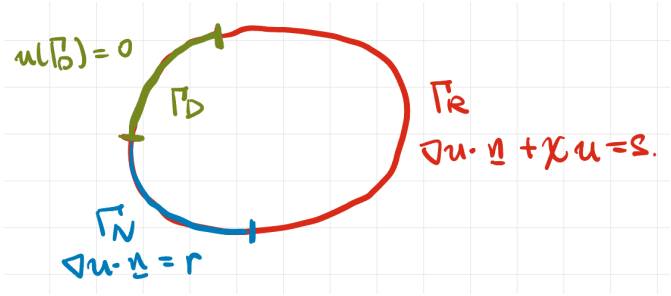
  We can also use squares, and we have to modify the polynomials accordingly:

  $$\mathbb{Q}^1 = axy + bx + cy + d \qquad \leftarrow \text{Bilinear polynomials}$$

  If we use pentagon, we use *Mimetic finite difference/Viral finite elements.*

- Weak formulation with General BCs:

$$\begin{cases} u(\Gamma_D) = 0 & \\ \boldsymbol{\nabla} u \cdot \mathbf{n} = r & \text{on } \Gamma_N \\ \boldsymbol{\nabla} u \cdot \mathbf{n} + \chi u = s & \text{on } \Gamma_R \end{cases}$$



$$-\Delta u = f \qquad \text{where } u \in H^1_0 + \mathscr{L},$$

where $\mathscr{L}$ is a lifting function. $\mathscr{L}_g$ is a $H^1(\Omega)$ $s.t.$ $\mathscr{L}_g(\Gamma_D) = g$. So,

$$\begin{aligned} \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v &= \int_\Omega fv + \underbrace{\int_{\partial\Omega} \boldsymbol{\nabla} u \cdot \mathbf{n} v}_{=0 \quad \text{on } \Gamma_D} \quad \forall\, v \in H^1_0 \\ &= \int_\Omega fv + \int_{\Gamma_N \cup \Gamma_R} \boldsymbol{\nabla} u \cdot \mathbf{n} v \end{aligned}$$

$$\int_\Omega \nabla u \nabla v = \int_\Omega fv + \int_{\Gamma_N} rv + \int_{\Gamma_R} sv - \int_{\Gamma_R} \chi uv.$$

So, the problem becomes: Find $u \in \mathscr{L} + H_0^1$ $s.t.$

$$\int_\Omega \nabla u \nabla v + \int_{\Gamma_R} \chi uv = \int_\Omega fv + \int_{\Gamma_N} rv + \int_{\Gamma_R} sv \quad \forall\, v \in H_0^1.$$

2.7 Spectral Method

We are still in Galerkin framework.

$$a(u_h, v_h) = \mathcal{F}(v_h) \quad \text{on some } X_h.$$

$$\inf_{w_h \in X_h} \|u - w_h\| \xrightarrow{h\to 0} 0.$$

Idea Use global polynomials (Gaussian interpolation) on special nodes.

- Gaussian Interpolation:
$$\int_a^b f \approx \int_a^b \pi_f.$$

1. $\int_a^b f(x) = f(\overline{x})(b-a)$. Do we have an optimal choice for $\overline{x}$?
$$\overline{x} = \frac{a+b}{2}.$$

2. We want two nodes:
$$\int_a^b f(x) = f(x)\underbrace{\frac{x-x_1}{x_2-x_1}}_{w_1} + f(x_2)\underbrace{\frac{x-x_2}{x_2-x_1}}_{w_2}$$
We can simply take $x_1 = a$ and $x_2 = b$ to recover the Trapezoidal rule, but can we find the optimal x_1 and x_2?

3. In general,
$$\int_a^b f(x) \approx \sum_i w_i f(x_i).$$
This problem is not easy to solve... But we can recover it from orthogonal polynomials.
[*Why Orthogonal polynomials? Let's see some Fourier series*

$$u(x) = \sum_{j=1}^{\infty} c_j \sin(j\pi x)$$
$$c_j = \frac{\int u(x)\sin(j\pi x)\,\mathrm{d}x}{\int \sin^2(j\pi x)\,\mathrm{d}x}$$

We can truncate:

$$u = \underbrace{\sum_{j=1}^{N} c_j \sin(j\pi x)}_{u_N} + \underbrace{\sum_{j=N+1}^{\infty} c_j \sin(j\pi x)}_{\text{Remainder}}$$

Because of orthogonality, Remainder is orthogonal to u_N. Up to N terms, u_N is the best approximation of u.]

Bottleneck: good for periodic problems, but technical difficulties when imposing BCs.

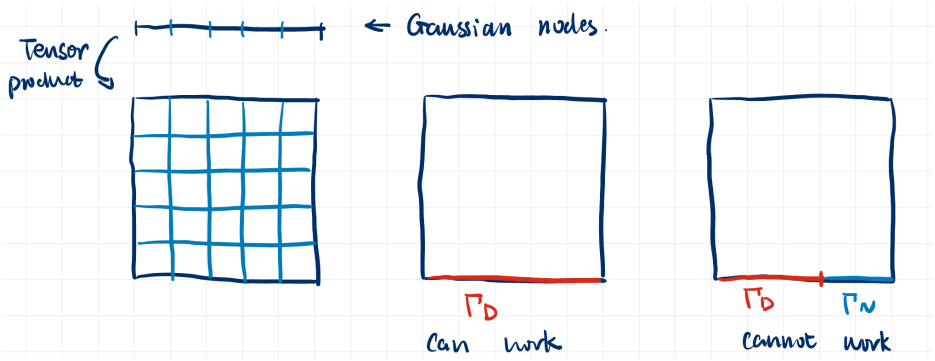
- Legendre polynomial on $L^2(-1,1)$:

$$\int_{-1}^{+1} \pi_L^K \pi_L^R = 0 \quad \text{if } K \neq R.$$

- Gauss Theory: interpolating nodes are always inside the domain. Nothing on the boundary is provided.

  Gauss-Lobatto: Correction to Gauss Theory that allows nodes to be on the boundary.

  However, we still require the BC to be the same on one edge.



  Ceà Lemma:

$$\inf_{p_N \in \mathbb{Q}_N} \|u - p_N\|_{H^1} \leq C \underbrace{N^{-(s-1)}}_{=N^{s-1}} |u|_{H^{s+1}} \qquad u \in H^{s+1}$$

$$u \in \mathcal{C}^\infty \implies Ce^{-N}.$$

  This is *exponential convergence*: If the solution is regular enough, spectral method would work well.

- $\int_{-1}^{+1} u_N' v_N' \,\mathrm{d}x = \boxed{\int_{-1}^{+1} f v_N \,\mathrm{d}x}$, where $u_N, v_N \in \mathbb{P}^{N-1}$, and $\int_{-1}^{+1} f v_N \,\mathrm{d}x$ gives the source of error: we don't know degree of f. We usually never compute this exactly.

$$\int_{-1}^{+1} u_N' v_N' \,\mathrm{d}x = \sum w_N u_N' v_N' \qquad \leftarrow \text{no quadrature error here}$$

To approximate $\int_{-1}^{+1} f v_N \,\mathrm{d}x$, we use GNI (Galerkin Numerical Integration).

Lemma 2.1 Strang Lemma

$$\|u - u_N\|_H \leq \underbrace{\text{Discretization}}_{C \inf_{w_N} \|u - w_N\|} + \underbrace{\text{Quadrature Error.}}_{\substack{\text{Could be controlled} \\ \text{by using Gaussian quadrature}}}$$

3 Advection-Diffusion Problem

$$\underbrace{-\mu\Delta u}_{\text{diffusion}} + \underbrace{\beta\nabla u}_{\text{advection}} + \underbrace{\sigma u}_{\text{reaction}} = f.$$

- Diffusion problem (Laplace) was derived from

$$\min J(u) \equiv \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega fv.$$

- β breaks the symmetry.

3.1 Finite Difference of 1D Problem

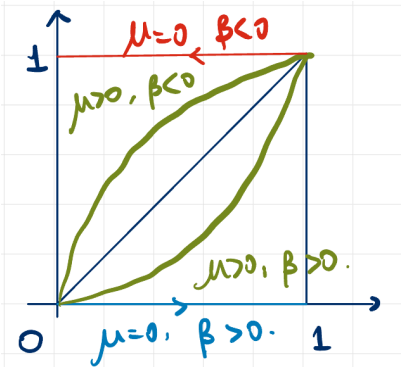
Consider

$$\begin{cases} -\mu u'' + \beta u' = 0, \quad \mu \geq 0 \\ u(0) = 0 \quad \text{and} \quad u(1) = 1. \end{cases}$$

Then,

$$u = \frac{e^{\beta/\mu x} - 1}{e^{\beta/\mu} - 1}$$

is the analytical solution.



- FD Discretization:

$$-\mu \underbrace{\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}}_{\sim\mathcal{O}(h^2)} + \beta \underbrace{\frac{u_{i+1} - u_{i-1}}{2h}}_{\substack{\text{centered difference} \\ \sim\mathcal{O}(h^2)}} = 0.$$

  To show order of convergence, use Taylor's expansion.

- If β is large, then we have oscillation, and we need finer mesh.

  How small h needs to be to prevent oscillatory solutions?

  Use second order difference equation to solve:

$$\begin{cases} \underbrace{\left(\frac{\mu}{h^2} + \frac{\beta}{2h}\right)}_{a} u_{i+1} \underbrace{-\frac{2\mu}{h^2}}_{b} u_i + \underbrace{\left(\frac{\mu}{h^2} - \frac{\beta}{2h}\right)}_{c} u_{i-1} = 0 \\ u_0 = 0, \quad u_N = 1 \end{cases}$$

So, the second order difference equation is

$$a\rho^2 + b\rho + c = 0$$
$$u_i = \alpha\rho_1^i + \beta\rho_2^i \implies u_0 = \alpha + \beta = 0 \implies \alpha = -\beta.$$

Assume $\rho_1 = 0$:

$$\text{If } \frac{c}{a} = \rho_1\rho_2 < 0 \implies \rho_2 \text{ is negative} \implies \text{oscillations}$$

1. $\mathbb{P}_e = \dfrac{\beta h}{2\mu} > 1$ (Pèclet): Oscillations
2. $\mathbb{P}_e < 1 \implies c > 0 \implies$ no oscillations.

So, we require step size:

$$\frac{\beta h}{2\mu} < 1 \implies h < \frac{2\mu}{\beta}.$$

Problem: If $\dfrac{\mu}{\beta} = 10^{-6}$, we need over 1 million points...

- Can we have a better scheme? A upwind scheme.

  This scheme breaks the centered symmetry.

$$\beta u' \approx \beta\frac{u_i - u_{i-1}}{h}$$

1. Accuracy: first-order $\sim \mathcal{O}(h)$

   We sacrifice the accuracy to allow larger step size.
2. This method is *physically consistent*, although we lose something mathematically.
3. Prove this scheme avoids oscillations:

   Method I Repeat the second order difference equation analysis.

   Method II

   Theorem 3.1.1
   Upwind never oscillates.

   Proof 1.

$$\begin{aligned}\beta\frac{u_i - u_{i-1}}{h} &= \underbrace{\frac{\beta}{2}\frac{u_{i+1} - u_{i-1}}{h}}_{\text{centered approx.}} - \frac{\beta}{2h}u_{i+1} + \frac{\beta u_i}{h} - \frac{\beta}{2h}u_{i-1} \\ &= \beta\frac{u_{i+1} - u_{i-1}}{2h} - \frac{\beta h}{2}\underbrace{\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2}}_{\text{approx. of } u''} \\ &= \text{centered approx.} + \frac{\beta h}{2}(-u'')\end{aligned}$$

So, the original problem: $-\mu u'' + \beta u' = 0$ becomes

$$\begin{aligned}
\text{UPW} \quad -\mu\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + \frac{\beta}{h}(u_i - u_{i-1}) &= 0 \\
-\mu\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + \frac{\beta}{2h}(u_{i+1}-u_{i-1}) - \frac{\beta h}{2}\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} &= 0 \\
-\left(\mu + \frac{\beta h}{2}\right)\frac{u_{i+1}-2u_i+u_{i-1}}{h^2} + \frac{\beta}{2h}(u_{i+1}-u_{i-1}) &= 0 \qquad [\mathbb{P}_e = \frac{\beta h}{2\mu}] \\
-\underbrace{\mu(1+\mathbb{P}_e)}_{\mu^*} u'' + \beta u' &= 0
\end{aligned}$$

Pèclet number of this new problem:

$$\mathbb{P}_e^* = \mathbb{P}_e^{\text{UPW}} = \frac{\beta h}{2\mu^*} = \frac{\beta h}{2\mu(1+\mathbb{P}_e)} = \frac{\mathbb{P}_e}{1+\mathbb{P}_e} < 1.$$

Q.E.D. ■

4. Is this method stable?

   We have a perturbed problem:

   $$-\left(\mu + \frac{\beta h}{2}\right)u'' + \beta u' = 0$$

   Note that $\frac{\beta h}{2} \to 0$ as $h \to 0$. So, we are stable.

   $$\text{upwind} = \text{centered}(\mu(1+\mathbb{P}_e)).$$

5. Solution is always consistent, regardless of choice of h.
6. Can we have second order upwind?

   Use Taylor's expansion to design a second order approximation of u_i', using only u_{i-2} and u_{i-1}.
7. If $\beta < 0$, the wind is backward. Use

   $$\frac{u_{i+1}-u_i}{h}.$$

3.2 Finite Difference in 2D+

Wind is a vector.

- $\boldsymbol{\beta} = \begin{bmatrix}\beta & 0\end{bmatrix}$, horizontal wind.

  $$-\mu(1+\mathbb{P}_e)\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u = 0.$$

  In this case, only apply upwind to $\frac{\partial u}{\partial x}$, as the wind is horizontal. No change on y direction.

- $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \beta_2\end{bmatrix}$.

  $$\beta_1\frac{\partial u}{\partial x} + \beta_2\frac{\partial u}{\partial y}.$$

Streamline Diffusion:

$$-\mu\boldsymbol{\nabla}\cdot(\boldsymbol{\nabla}u)+\beta\cdot\boldsymbol{\nabla}u-\frac{\frac{1}{2}h}{\|\boldsymbol{\beta}\|}\boldsymbol{\nabla}\cdot((\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)\boldsymbol{\beta})=f.$$

Weak formulation:

$$-\frac{\frac{1}{2}h}{\|\boldsymbol{\beta}\|}\int_\Omega\boldsymbol{\nabla}((\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)\boldsymbol{\beta})v\,\mathrm{d}w=\frac{\frac{1}{2}h}{\|\boldsymbol{\beta}\|}\int_\Omega(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}v)\,\mathrm{d}w.$$

3.3 Finite Elements in 1D

Consider

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

- What is a weak formulation?

$$v(-\mu u''+\beta u'+\sigma u)=0, \quad \text{for } v(0)=v(1)=0.$$

Define $\ell(x)=x$. Then, $u=\mathring{u}+\ell(x)$, then $\mathring{u}(0)=\mathring{u}(1)=0$.

$$\begin{aligned}\big(-\mu\mathring{u}''+\beta\mathring{u}'+\sigma\mathring{u}\big)v&=v(-\beta\cdot1-\sigma\ell)\\ \int_0^1-\mu\mathring{u}''v+\beta\mathring{u}'v+\sigma\mathring{u}v&=\int_0^1v(-\beta\cdot1-\sigma\ell)\\ \underbrace{\int_0^1\mu\mathring{u}'v'+\beta\mathring{u}'v+\sigma\mathring{u}v}_{a(\mathring{u},v)}&=\underbrace{-\int_0^1(\beta+\sigma\ell)v}_{\mathcal{F}(v)}\end{aligned}$$

So, the variational form: $a(\mathring{u},v)=\mathcal{F}(v)$, where

$$a(\mathring{u},v)=\int_0^1\mu\mathring{u}'v'+\beta\mathring{u}'v+\sigma uv.$$

Still, $\beta\mathring{u}'v$ is breaking the symmetry here.

1. $a(\mathring{u},v)$ is not symmetric: $a(\mathring{u},v)\neq a(v,\mathring{u})$.
2. $a(\mathring{u},v)$ is coercive: $a(u,u)\geq\alpha\|u\|_{H^1}^2$.

   Proof 1.

   $$a(u,u)=\int_0^1\mu(u')^2+\sigma u^2+\beta u'u.$$

   Note that $u'u=\frac{1}{2}(u^2)'$, so $\int_0^1\beta u'u=\frac{1}{2}\int_0^1\beta(u^2)'$. Using integration by parts,

   $$\frac{1}{2}\int_0^1\beta(u^2)'=\frac{1}{2}\underbrace{\Big[\beta u^2\Big]_0^1}_{=0}-\frac{1}{2}\int_0^1\beta'u^2$$

So,

$$\begin{aligned} a(u,u) &= \int_0^1 \mu(u')^2 + \sigma u^2 - \frac{1}{2}\beta' u^2 \\ &= \int_0^1 \mu(u')^2 + \left(\sigma - \frac{1}{2}\beta'\right)u^2 \end{aligned}$$

Assumption: $\sigma - \frac{1}{2}\beta' \geq \rho$.

Then,

$$\begin{aligned} a(u,u) &\geq \int_0^1 \mu(u')^2 + \rho u^2 \\ &\geq \mu\|u'\|_{L^2}^2 + \rho\|u\|_{L^2}^2 \\ &\geq \alpha\|u\|_{H^1}^2, \quad \alpha = \min\{\rho,\mu\}. \end{aligned}$$

Q.E.D. ■

3. $a(\mathring{u},v)$ is continuous: $|a(u,v)| \leq \gamma\|u\|_{H^1}\|v\|_{H^1}$.

Proof 2.

$$\mu\int u'v' + \beta\int u'v + \sigma\int uv \leq \max\{\mu,|\beta|,\sigma\}\|u\|_{H^1}\|v\|_{H^1}$$

Q.E.D. ■

Remark.

- Usually, $\mu < \rho$
- For a convection-dominated problem, $|\beta|$ will be the largest
- For a reaction dominated problem, σ will be the largest.

4. Strong consistent:

$$a(u-u_h, w_h) = 0.$$

Lemma 3.1 Cèa

$$\|u-u_h\| \leq \left(\frac{C_{\text{continuous}}}{C_{\text{coercivity}}}\right)\frac{\gamma}{\alpha}\inf_{w_h\in X_h}\|u-w_h\|.$$

Usually, $\frac{\gamma}{\alpha}$ can be huge, and so Cèa lemma is not very helpful.

- Linear FE:

$$A = \Big[a(\varphi_i,\varphi_j)\Big].$$

$$\mu\int_0^1 \varphi_i'\varphi_{i\pm1}', \quad \beta\int_0^1 \varphi_{i\pm1}'\varphi_i, \quad \text{and} \quad \sigma\int_0^1 \varphi_i\varphi_{i\pm1}.$$

The FE system is

$$\frac{\mu}{h}(u_{i+1} - 2u_i + u_{i-1}) + \frac{\beta}{2h}(u_{i+1} - u_{i-1}) + \frac{\sigma h}{6}(u_{i+1} + 4u_i + u_{i-1}) = 0.$$

1. Suppose $\sigma = 0$:
$$\frac{1}{h}(\text{FE}) = \text{FD}$$
So, FE suffer in the same way as FD. To prevent oscillation, we also require
$$\mathbb{P}_e = \frac{|\beta|h}{2\mu} < 1.$$

2. As lesson from FD: add a diffusive term $-\frac{\beta h}{2}u''$. In weak formulation:
$$a(u_h, v_h) + \frac{\beta h}{2}\int_0^1 u_h' v_h' = \mathcal{F}(v_h).$$
This is exactly the upwind FE method.

3. What we lose with UPW-FE: No strong consistency, so we don't have Cèa Lemma anymore.

4. What we still have: Strang Lemma.
Lemma 3.2 Strang
$$\|u - u_h\| \leq C \inf_{w_h} \|u - w_h\| + \sup_{w_h}(\underbrace{|a(u, w_h) - a_h(u, w_h)|}_{\frac{\beta h}{2}\int_0^1 u_h' v_h'})$$
Implication: Regardless of the order of polynomial, due to the additional term, the convergence rate is always $\sim \mathcal{O}(h)$. So, it makes no sense to use higher order than linear.

- Scharfter-Gammal: Design a better μ such that
$$\mu^{\text{smart}} = \mu(1 + \Phi(\mathbb{P}_e)).$$
Problem: Only work for 1D problems.

3.4 Convection-Dominated problem

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

Analytical approach: change variable $u \mapsto w$ $s.t.$
$$\frac{\partial w}{\partial t} - \chi\frac{\partial^2 w}{\partial x^2} = 0.$$

Can we don the same in our numerical method? *Slotboom Variable*

Proof 1. In 1D,

$$-\mu u'' + \beta u' = 0, \quad \mu \geq 0.$$

Suppose $u = \rho e^{\sigma x}$, where σ is to be determined (Slotboom variable). Then,

$$\begin{aligned} u' &= \rho' e^{\sigma x} + \sigma \rho e^{\sigma x} \\ u'' &= \rho'' e^{\sigma x} + 2\rho' \sigma e^{\sigma x} + \sigma^2 \rho e^{\sigma x}. \end{aligned}$$

Then,

$$\begin{aligned} e^{\sigma x}\left[-\mu\rho'' - 2\mu\rho'\sigma - \mu\sigma^2\rho + \beta\rho' + \beta\sigma\rho\right] &= 0 \\ -2\mu\sigma + \beta = 0 \implies \sigma &= \frac{\beta}{2\mu} \\ \beta\sigma - \mu\sigma^2 = \beta\frac{\beta}{2\mu} - \mu\frac{\beta^2}{4\mu^2} &= \frac{\beta^2}{4\mu}. \end{aligned}$$

So,

$$-\mu\rho'' + \frac{\beta^2}{4\mu}\rho = 0.$$

Weak formulation:

$$\int \mu\rho' v + \frac{\beta^2}{4\mu}\int \rho v = 0.$$

Q.E.D. ■

This method is better tan traditional UPW-FE, but still may suffer from instabilities. Moreover, the weak formulation solves for ρ. To get u back, we need exponential functions: more errors.

3.5 Strongly Consistent Stabilization

$$a(u,v) + b_h(u,v) = \mathcal{F}(v)$$

Generalized Galerkin:

$$a_h(u,v) = \mathcal{F}_h(v).$$

- Streamline Diffusion:

$$b_h = \frac{ch}{\|\boldsymbol{\beta}\|}\int (\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}v)\,\mathrm{d}\Omega.$$

- By Strang's Lemma, as long as $b_h \to 0$ as $h \to 0$, we still have convergence:

$$\|u_h - u\| \leq \underbrace{1}_{\text{Cèa}\ \inf_{w_h\in V_h}\|u-w_h\|} + \underbrace{2}_{|a-a_h|\ =\frac{ch^2}{\|\boldsymbol{\beta}\|}(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}v)} + \underbrace{3}_{|\mathcal{F}-\mathcal{F}_h|=0}$$

- We don't have strong consistency:

$$a_h(u,v_h) = a(u,v_h) + \underbrace{b_h(u,v_h)}_{\neq 0} \neq \mathcal{F}(v_h)$$

- Strongly Consistent Stabilization (Hughes and Brooks):

$$a(u,v) = \mathcal{F}(v).$$

1. Let's add something that will vanish if the solution is regular enough: residual in the strong form

$$a(u,v) + (a(u,v) - \mathcal{F}(v)) = \mathcal{F}(v).$$

2. In strong form: $-\mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u - f$.

   Go back to weak:

$$\int_\Omega (-\mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u - f)\mathscr{L}v,$$

   where $\mathscr{L}$ is an operator on v.

   We cannot do this at continuous level, but it is perfectly fine at the discrete level:

$$\sum_K \delta_K \int_K (-\mu\Delta u_h + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u_h - f)\mathscr{L}v_h.$$

   [*If we take $\mathscr{L}v_h = \boldsymbol{\beta}\cdot\boldsymbol{\nabla}v_h$, we have exactly the streamline diffusion.*]

3. What is $\mathscr{L}$?

$$A = \frac{1}{2}\underbrace{\left(A + A^\top\right)}_{\substack{\text{symmetric:}\\ B=B^\top}} + \frac{1}{2}\underbrace{\left(A - A^\top\right)}_{\substack{\text{skew-symmetric}\\ B=-B^\top}}$$

   Goal: Write $\mathscr{L} = -\mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u = \mathscr{L}_s + \mathscr{L}_{ss}$.

$$\mathscr{L}_s = \int_\Omega \mu\boldsymbol{\nabla}u\boldsymbol{\nabla}v$$

$$\int_\Omega (\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)v = -\int_\Omega u\boldsymbol{\nabla}\cdot(\boldsymbol{\beta}v) = -\int_\Omega u\boldsymbol{\beta}\cdot\boldsymbol{\nabla}v - \underbrace{\int_\Omega uv\boldsymbol{\nabla}\cdot(\boldsymbol{\beta})}_{\text{symmetric}}$$

$$\underbrace{\int_\Omega \beta\cdot\boldsymbol{\nabla}uv + \frac{1}{2}\int_\Omega uv\boldsymbol{\nabla}\cdot(\boldsymbol{\beta})}_{\mathcal{G}(u,v)=-\mathcal{G}(u,v),\text{ skew symmetric}} = -\int_\Omega u\boldsymbol{\beta}\cdot\boldsymbol{\nabla}v - \frac{1}{2}\int_\Omega uv\boldsymbol{\nabla}\cdot(\boldsymbol{\beta}).$$

   So,

$$\mathscr{L} = \underbrace{\mu\int_\Omega \boldsymbol{\nabla}u\boldsymbol{\nabla}v - \frac{1}{2}\int_\Omega uv\boldsymbol{\nabla}\cdot(\boldsymbol{\beta})}_{\mathscr{L}_s} + \underbrace{\int(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}u)v + \frac{1}{2}\int_\Omega uv\boldsymbol{\nabla}\cdot(\boldsymbol{\beta})}_{\mathscr{L}_{ss}}.$$

4. SUPG: Streamline Upwind Petrov Galerkin:

$$\sum \delta_k \int_k (\mathscr{L}u - f)(\mathscr{L}_{ss}v)\,\mathrm{d}w,$$

   * $\mathscr{L}_u = -\mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u$
   * $\mathscr{L}_s = -\mu\Delta u - \frac{1}{2}(\boldsymbol{\nabla}\cdot\boldsymbol{\beta})u$
   * $\mathscr{L}_{ss} = \beta\cdot\boldsymbol{\nabla}u + \frac{1}{2}(\boldsymbol{\nabla}\cdot\boldsymbol{\beta})u$

5. GaLS: Galerkin Least squares

$$\mathscr{L}_\rho = \rho\mathscr{L}_s + \mathscr{L}_{ss}$$

   * $\rho = 0$: SUPG
   * $\rho = 1$: GaLS
   * $\rho = -1$: Douglas-Wang (DW)

6. For all the methods:

$$\|u_h - u_{\text{ex}}\|_{*\rho} \le Ch^{r+\frac{1}{2}}|u|_{H^{s+1}}, \quad r = \min\{p, s\},$$

   where $\|\cdot\|_{*\rho}$ is not a standard norm, and it depends on ρ.

   Implication: If the solution is regular enough, we can still benefit from using high order polynomials.

$$\|u\|_{\substack{\rho=1\\ \text{GaLS}}} \equiv \mu\|\boldsymbol{\nabla}u_h\|_{L^2}^2 + \min\{\boldsymbol{\nabla}\cdot\boldsymbol{\beta}, \sigma\}\|u_h\|_{L^2}^2 + \sum_K \delta_k\|\mathscr{L}u_h\|_{L^2(T_u)}$$

3.6 Reaction-Dominated Problem

$$-\mu\Delta u + \sigma u = f.$$

- FD:

$$-\mu\frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} + \sigma u_i = f_i,$$

  and the matrix will be a good matrix:

$$A = \mu\begin{bmatrix} 2 & 1 & \\ -1 & \ddots & \ddots \\ & \ddots & \ddots \end{bmatrix} + \sigma I$$

- FE:

$$\sigma\int_0^1 \varphi_i\varphi_j \quad \to \quad \frac{h}{6}\begin{bmatrix} 4 & 1 & \\ 1 & \ddots & \ddots \\ & \ddots & \ddots \end{bmatrix} \leftarrow \text{Mass Matrix.}$$

  We need $\mathbb{P}_e = \dfrac{|\sigma|h^2}{6\mu} < 1$ to avoid oscillation.

We are in trouble, but not as much as in convection-domination,

$$h < \sqrt{6}\sqrt{\frac{\mu}{|\sigma|}},$$

where the square root is helping a lot.

Notice: Instead of solving $\int_0^1 \varphi_i\varphi_j$ exactly, we use a trapezoidal rule:

$$(T)\int_i^{i+1} \varphi_i\varphi_j = \frac{\varphi_i\varphi_j(x_{i+1}) + \varphi_i\varphi_j(x_i)}{2}h.$$

This scheme produces the identity matrix I.

- Mass Lumping: Make the mass matrix M a diagonal matrix by summing the off-diagonal terms to the diagonal.

1.

$$\begin{aligned}\sigma\int_0^1 \varphi_i\varphi_j \approx \sigma(T)\int_0^1 \varphi_i\varphi_j &= \sigma(T)\int_{x_{i-1}}^{x_{i+1}} \varphi_i\varphi_j \\ &= \sigma\left((T)\int_{x_{i-1}}^{x_i} \varphi_i\varphi_{i-1} + \underbrace{\int_{x_i}^{x_{i+1}} \varphi_i\varphi_{i-1}}_{=0} + \int_{x_{i-1}}^{x_i} \varphi_i\varphi_{i+1} + \int_{x_i}^{x_{i+1}} \varphi_i\varphi_{i+1}\right) \\ &= \sigma\frac{h}{2}(\varphi_i\varphi_{i-1}(x_i) + \varphi_i\varphi_{i-1}(x_{i-1})) + 0 \\ &= 0.\end{aligned}$$

So, the off-diagonal terms are 0. For the diagonal term,

$$\begin{aligned}\sigma(T)\int_{xi-1}^{x_i} \varphi_i^2 + \sigma(T)\int_{x_i}^{x_{i+1}} \varphi_i^2 &= \frac{\sigma h}{2}\left(\varphi_i^2(x_{i-1}) + \varphi_i^2(x_i)\right) + \frac{\sigma h}{2}\left(\varphi_i^2(x_i) + \varphi_i^2(x_{i+1})\right) \\ &= \sigma h.\end{aligned}$$

Therefore, $FD = FE + ML$ (mass lumping)

2. **Lemma 3.1 Strang's Lemma**

$$-\mu\underbrace{\int_0^1 \varphi_i'\varphi_j'}_{\mathbb{P}^0} + \beta\underbrace{\int_0^1 \varphi_0'\varphi_j}_{\mathbb{P}^1} + \sigma\underbrace{\int_0^1 \varphi_i\varphi_j}_{\mathbb{P}^2}.$$

Mass lumping (ML) can be viewed as a **Generalized Galerkin Method**:
$a(u,v) = \mathcal{F}(v) \implies a_h(u,v) = \mathcal{F}_h(v)$, and $\sigma\int uv - \sigma(T)\int uv \sim \mathcal{O}(h^2)$. So, we can go up to quadratic FE and achieve $\mathcal{O}(h^2)$ with mass lumping.

4 Parabolic Equations

$$\frac{\partial u}{\partial t} - \mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla} u + \sigma u = f, \quad \mathbf{x}\in\Omega, \quad \mu \geq 0$$
$$+ \text{B.C.} : u(\partial\Omega) = g \quad \text{(Can be Dirichlet, Neumann, Robin)}$$
$$+ \text{I.C.}$$

4.1 Weak Formulation and Well-Posedness

$$\left(\frac{\partial u}{\partial t} - \mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla} u + \sigma u\right)v = fv$$
$$\int_0^T\int_\Omega \left(\frac{\partial u}{\partial t} - \mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla} u + \sigma u\right)v = \int_0^T\int_\Omega fv$$
$$\iint \frac{\partial u}{\partial t}v - \mu\iint \Delta uv + \iint \boldsymbol{\beta}\cdot\boldsymbol{\nabla} uv + \iint \sigma uv = \iint fv$$

- By Green's formula:
$$-\iint \Delta uv = \iint \boldsymbol{\nabla} u\boldsymbol{\nabla} v$$
- If $\Omega \perp\!\!\!\perp t$, then
$$\int_0^T\int_\Omega \frac{\partial u}{\partial t}v = \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega uv$$

Theorem 4.1.1

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} uv = \int_\Omega \frac{\partial(uv)}{\partial t} + \int_{\partial\Omega}(\mathbf{w}\cdot\mathbf{n})uv,$$

where $\mathbf{w} = \dfrac{\partial\Omega(t)}{\partial t}$.

- The bilinear form is
$$\int_0^T \left(\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega uv + \underbrace{\mu\int_\Omega \boldsymbol{\nabla} u\boldsymbol{\nabla} v + \int_\Omega \boldsymbol{\beta}\cdot\boldsymbol{\nabla} uv + \int_\Omega \sigma uv}_{a(u,v)}\right) = \int_0^T \underbrace{\int_\Omega fv}_{\mathcal{F}(v)},$$
where $u, v \in H_0^1$.
- **Lemma 4.2 Lax-Milgram Lemma Extension** If $a(u,v)$ is bilinear, continuous in $H_0^1(\Omega)$, and $\exists\,\alpha > 0,\ \lambda \geq 0$ such that
$$a(u,u) + \lambda\|u\|_{L^2}^2 \geq \alpha\|u\|_{H^1}^2, \qquad \text{(Weak Coercivity)}$$
and $\mathcal{F}(v)$ is continuous, then the weak formulation is WP. [*When $\lambda = 0$, we have strong coercivity.*]

Example 4.1.3

For elliptic problem,

$$-\mu u'' + \sigma u = f \quad \text{is WP requires} \quad \sigma \geq 0.$$

For parabolic problem,

$$\frac{\partial u}{\partial t} - \mu \frac{\partial^2 u}{\partial x^2} + \sigma u = f \quad \text{is WP} \quad \forall\, \sigma.$$

So, time-dependent problem is more likely to be WP.

- What space we want to use for time?

1. $\int_0^T \int_\Omega uv$. If $u = v$,

$$\int_0^T \frac{\mathrm{d}}{\mathrm{d}t} \|u\|_{L^2}^2 = \|u\|_{L^2}^2(T) - \|u\|_{L^2}^2(0)$$

Space: $L^\infty(0,T;L^2(\Omega)) \equiv L^\infty(L^2)$.

2. $\int_0^T \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v$. If $u = v$,

$$\int_0^T \|u\|_{H^1}^2(t)$$

Space: $L^2(0,T;H_0^1(\Omega)) \equiv L^2(H_0^1)$.

Theorem 4.1.4

Suppose $f \in L^2(L^2)$ and $u_0 \in L^2(\Omega)$. If $a(u,v)$ is weakly coercive, then there exists a unique $u \in L^2(H_0^1) \cap L^\infty(L^2)$, solution to the problem.

Theorem 4.1.5

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = \mathcal{F}(v) \tag{P}$$

Suppose $u(\mathbf{x},0) = u_0(\mathbf{x}) \in L^2(\Omega)$, $f \in L^2(L^2)$. If

- $a(\cdot,\cdot)$ bilinear, continuous, weakly coercive,

$$\exists\, \alpha, \lambda > 0, \quad a(u,u) + \lambda \|u\|_{L^2}^2 \geq \alpha \|u\|^2.$$

- $\mathcal{F}(v)$ is continuous linear functional.

Then, (P) is WP.

Example 4.1.6

$$
\begin{aligned}
w &= e^{-\lambda t}u \\
\frac{\partial w}{\partial t} &= e^{-\lambda t}\frac{\partial u}{\partial t} - \lambda e^{-\lambda t}u \\
e^{-\lambda t}\frac{\partial u}{\partial t} &= \frac{\partial w}{\partial t} - \lambda e^{-\lambda t}u \\
e^{-\lambda t}\left(\frac{\partial u}{\partial t}, v\right) + \left(e^{-\lambda t}u, v\right) &= e^{-\lambda t}\mathcal{F}(v) \\
\left(\frac{\partial w}{\partial t} + \lambda w, v\right) + a(w,v) &= \mathcal{G}(v) \\
\left(\frac{\partial w}{\partial t}, v\right) + \underbrace{\lambda(w,v) + a(w,v)}_{\widetilde{a}(w,v)\geq\alpha\|w\|_{H^1}^2} &= \mathcal{G}(v)
\end{aligned}
$$

A Priori Bounds

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u,v) = \mathcal{F}(v).$$

Take $v = u$.

- If Ω is time independent,
$$\int_\Omega \frac{\partial u}{\partial t}u = \frac{1}{2}\int_\Omega \frac{\partial}{\partial t} = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|_{L^2}^2.$$
- By weak coercivity,
$$a(u,u) \geq \alpha\|u\|_{H^1}^2 + \lambda\|u\|_{L^2}^2.$$
- $\mathcal{F}(u) = \int fu \leq \|f\|_{L^2}\|u\|_{L^2}$

So,

$$
\begin{aligned}
\frac{1}{2}\int_0^T \frac{\mathrm{d}}{\mathrm{d}t}\|u\|_{L^2}^2 + \alpha\int_0^T \|u\|_{H^1}^2 + \lambda\int_0^T \|u\|_{L^2}^2 &\leq \int_0^T \|f\|_{L^2}\|u\|_{L^2} \\
ah \leq \varepsilon a^2 + \frac{1}{4\varepsilon}b^2 \quad &\left[\textit{Young Inequality}, 0 \leq \left(\sqrt{\varepsilon}a - \frac{1}{2\sqrt{\varepsilon}}b^2\right)^2\right] \\
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 + \alpha\|u\|_{H^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2 &\leq \frac{1}{2\lambda}\|f\|_{L^2}^2 \\
\frac{1}{2}\int_0^T \|u\|^2 + \alpha\int_0^T \|u\|_{H^2}^2 + \frac{\lambda}{2}\int_0^T \|u\|_{L^2}^2 &\leq \frac{1}{2\lambda}\int_0^T \|f\|_{L^2}^2 \\
\|u\|_{L^2}^2(T) + 2\alpha\int_0^T \|u\|_{H^2}^2 + \lambda\int_0^T \|u\|_{L^2}^2 &\leq \underbrace{\frac{1}{\lambda}\int_0^T \|f\|_{L^2}^2 + \|u_0\|_{L^2}^2}_{\text{data}}
\end{aligned}
$$

Lemma 4.7 Granwall Lemma

$$g(t) \le \alpha + \int_0^t \beta g(\tau)\,\mathrm{d}\tau, \quad \beta \ge 0$$
$$g(t) \implies \alpha + \int_0^t \alpha\beta e^{\int_\tau^t \beta(\xi)\,\mathrm{d}\xi}\,\mathrm{d}\tau.$$

4.2 Finite Element and Semi-Discretization

Solving Systems of ODEs

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = M^{-1}A\mathbf{u} + M^{-1}\mathbf{b}, \quad \mathbf{u} = \Big[u(x_i)\Big]$$
$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = A\mathbf{u} + \mathbf{b} \quad \text{(more generally)}$$

- Absolute stability in ODE:

  A general ODE can be written as
  $$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t,y).$$
  Consider only the ODE
  $$\frac{\mathrm{d}y}{\mathrm{d}t} = \lambda y, \quad \lambda < 0.$$
  Then, $y(t) \to 0$ as $t \to \infty$. So, the numerical solution $y_{\Delta t}$ does the same.

  1. Backward Euler (BE):
  $$\frac{y^{n+1}-y^n}{\Delta t} \approx \lambda y^{n+1} \implies y^{n+1}(1+|\lambda|\Delta t) = y^n$$
  $$y^{n+1} = \frac{1}{1+|\lambda|\Delta t}y^n$$
  $$y^{n+1} = \left(\frac{1}{1+|\lambda|\Delta t}\right)^{n+1} y^0 \to 0 \quad \text{as } n \to \infty.$$

  2. Forward Euler (FE):
  $$\frac{y^{n+1}-y^n}{\Delta t} \approx \lambda y^n \implies y^{n+1} = (1+\lambda\Delta t)y^n.$$

     For absolute stability, we require
  $$|1+\lambda\Delta t| < 1$$
  $$|1-|\lambda|\Delta t| < 1$$
  $$-1 < 1-|\lambda|\Delta t < 1$$
  $$\Delta t < \frac{2}{|\lambda|}$$

- Moving to systems:

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = A\mathbf{y}(+\mathbf{b}), \quad \lambda = \mathrm{eig}(A), \quad \mathrm{Re}(\lambda) < 0.$$

Diagonalization: $TAT^{-1} = D$. Then, $D = TAT^{-1}$. Denote $\mathbf{w} = T\mathbf{y}$. We have

$$\frac{\mathrm{d}\mathbf{w}}{\mathrm{d}t} = D\mathbf{w}.$$

So, we require

$$\Delta t < \frac{2}{\max|\lambda_i|}.$$

- Absolute stability has nothing to do with accuracy.

Semi-Discretization

$$\left(\frac{\partial u}{\partial t}, v\right) + a(u, v) = \mathcal{F}(v).$$

Let $V_h \subset H_0^1$, $u_h = \sum u_j(t)\varphi_j(\mathbf{x})$ (separation of variable), and $v_h = \varphi_j(\mathbf{x})$.

The numerical problem is

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

$$\left(\sum \frac{\mathrm{d}u_j}{\mathrm{d}t}\varphi_j, \varphi_i\right) + \underbrace{\sum u_j(t)}_{\mathbf{u}(t)} \underbrace{a(\varphi_j, \varphi_i)}_{A} = \underbrace{\mathcal{F}(\varphi_i)}_{\mathbf{f}}$$

$$\underbrace{\frac{\mathrm{d}}{\mathrm{d}t}u_j(t)}_{\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t}} \underbrace{(\varphi_j, \varphi_i)}_{\substack{M \\ \text{mass matrix}}} + A\mathbf{u}(t) = \mathbf{f}$$

$$M\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + A\mathbf{u} = \mathbf{f}$$

$$\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} = \underbrace{-M^{-1}A}_{\substack{\text{negative} \\ \text{eigenvalue}}}\mathbf{u} + M^{-1}\mathbf{f}$$

Convergence Define $e_h = u - u_h$.

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) = 0.$$

Assume strong coercivity,

$$\begin{aligned}
\alpha\|e_h\|_{H^1}^2 \le a(e_h, e_h) &= a(e_h, u - u_h) \\
&= a(e_h, u - w_h) + a(e_h, \underbrace{w_h - u_h}_{=0, \in V_h}) \\
&= \underbrace{a(e_h, u - w_h)}_{①} - \underbrace{\left(\frac{\partial e_h}{\partial t}, w_h - u_h\right)}_{②}
\end{aligned}$$

- Term ①:

$$\begin{aligned} a(e_h, u - w_h) &\le \gamma \|e_h\| \|u - w_h\| && \text{[Continuity]} \\ &\le \frac{\alpha}{2}\|e_h\|_{H^1}^2 + C\|u - w_h\|_{H^1}^2 && \text{[Young Inequality]} \end{aligned}$$

- Term ②:

$$\begin{aligned} -\left(\frac{\partial e_h}{\partial t}, w_h - u_h\right) &= -\left(\frac{\partial e_h}{\partial t}, w_h - u + u - u_h\right) \\ &= -\left(\frac{\partial e_h}{\partial t}, w_h - u\right) - \left(\frac{\partial e_h}{\partial t}, e_h\right) \\ &= -\left(\frac{\partial e_h}{\partial t}, w_h - u\right) - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_{L^2}^2 \end{aligned}$$

Combined, we have

$$\begin{aligned} \alpha\|e_h\|_{H^1}^2 &\le \frac{\alpha}{2}\|e_h\|_{H^1}^2 + C\|u - w_h\|_{H^1}^2 - \left(\frac{\partial e_h}{\partial t}, w_h - u\right) - \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_{L^2}^2 \\ \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_{L^2}^2 + \frac{\alpha}{2}\|e_h\|_{H^1}^2 &\le C\|u - w_h\|_{H^1}^2 - \left(\frac{\partial e_h}{\partial t}, w_h - u\right) \\ \int_0^T \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|e_h\|_{L^2}^2 + \int_0^T \frac{\alpha}{2}\|e_h\|_{H^1}^2 &\le C\int_0^T \|u - w_h\|_{H^1}^2 - \int_0^T \left(\frac{\partial e_h}{\partial t}, w_h - u\right). \end{aligned}$$

So, the rate of convergence is of order $\sim \mathcal{O}(h^2)$.

Θ-Method

$$\int_{t^n}^{t^{n+1}} \text{RHS} = \Delta t\theta \text{RHS}(t^{n+1}) + \Delta t(1-\theta)\text{RHS}(t^n), \quad \theta \in (0, 1].$$

- Common choices of θ:
  1. $\theta = 1$: Backward Euler
  2. $\theta = \frac{1}{2}$: Crank-Nicolson/Trapezoidal Rule

$$\int_{t^n}^{t^{n+1}} f(t) = \frac{f(t^{n+1}) + f(t^n)}{2}\Delta t$$

  3. $\theta = 0$: Forward Euler

- Stability Analysis:

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

Use diagonalization and substitution, we get

$$(1-\theta)\lambda_i w_i^{n+1} = (1 + (1-\theta)\Delta t\lambda_i)w_i^n$$

1. $\theta \ge \frac{1}{2}$: Unconditionally stable.

2. $\theta < \frac{1}{2}$: Conditionally stable. Condition is given by

$$\Delta t < \frac{2}{|\lambda_i|(1-2\theta)}$$

Theorem 4.2.1
u_n of FEM and Θ method is such that

$$\|u-u_h\|_{L^\infty(L^2)} + \|u-u_h\|_{L^2(H^1)} \leq C\Big(\Delta t^{p(\theta)} + h^r\Big),$$

where

$$p(\theta) = \begin{cases} 2, & \theta = \frac{1}{2} \\ 1, & \text{o/w} \end{cases}$$

and $r = \min\{\deg, s\}$.

Summary

$$\frac{\partial u}{\partial t} - \mu\Delta u = f \quad \text{on } \Omega.$$

$$\int_\Omega \frac{\partial u}{\partial t} v + \mu \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega fv$$
$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega uv + \mu \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega fv$$

Define $\mathbf{u}(\mathbf{x},t) = \sum u_j(t)\rho_j(\mathbf{x})$, $M = \int_\Omega \varphi_i\varphi_j$, and $K = \mu\int_\Omega \boldsymbol{\nabla}\varphi_i\boldsymbol{\nabla}\varphi_j$. We have

$$M\frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} + K\mathbf{u} = \mathbf{b}.$$

With Θ-method:

$$M\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + \theta K\mathbf{u}^{n+1} + (1-\theta)K\mathbf{u}^n = \theta\mathbf{b}^{n+1} + (1-\theta)\mathbf{u}^n$$
$$\underbrace{(M+\Delta\theta K)}_{\text{left matrix}}\mathbf{u}^{n+1} = \underbrace{(M-\Delta t(1-\theta)K)}_{\text{right matrix}}u^n + \Delta t\theta\mathbf{b} + (1-\theta)\Delta t\mathbf{u}^n$$

- When $\theta < \frac{1}{2}$, Θ-method is conditionally stable. The condition is given by

$$\Delta t < \frac{2}{\rho(1-2\theta)},$$

where $\rho = \max\left(\left|\mathrm{eig}(M^{-1}K)\right|\right) \sim \mathcal{O}(h^{-2})$. So, $\Delta t \sim \mathcal{O}(h^2)$. [*So, if we want to refine h, we need*

to refine Δ at the same time. For example,

$$(\Delta, h) \to \left(\frac{\Delta t}{4}, \frac{h}{2}\right).$$

] This is annoying! So, we don't use explicit method a lot for parabolic problems.

- On the other hand, running full implicit (FI) solver requires three loops. This is inefficient. What we can do is to run a semi-implicit (SI) solver that only requires two loops.
- However, semi-implicit is not unconditionally stable. The condition is given by $\Delta t < \mathcal{O}(h)$. [*Generally, this is easy to fulfill.*]

4.3 Space-Time Finite Element (Space-Time FEM)

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

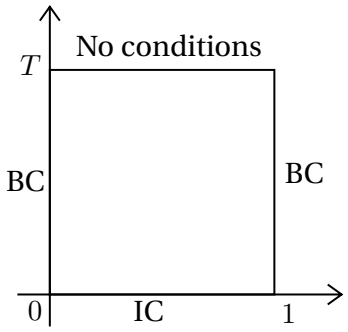
Attempt: think of this as a 2D problem:

$$-\mu \frac{\partial^2 u}{\partial x^2} - 0 \cdot \frac{\partial^2 u}{\partial y^2} + 0 \cdot \frac{\partial u}{\partial x} + 1 \cdot \frac{\partial u}{\partial y} = f.$$

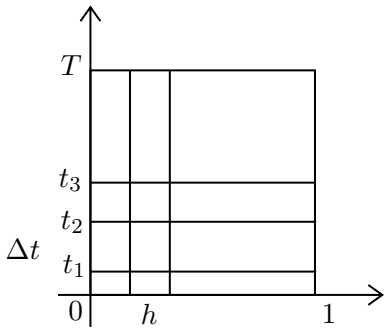
We are back in advection-diffusion problem.

Problem with this attempt:

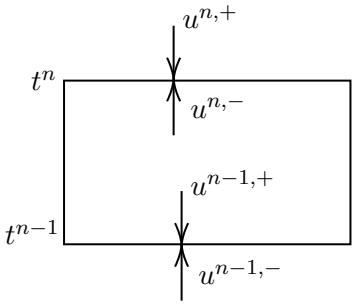
- Time might be huge, so it will be a large (and imbalanced) mesh.
- The physics is different. For time, we have an initial condition, stead of a boundary condition.
- The problem is not WP (time is not reversible).



Decouple the time slab: For each time, we can have different space discretization level



- Problem: we will have "jumping" solutions.



  See exercise book for reference. Eventually, we will get backward Euler.

- Nowadays, people do not use space-time FEM anymore becase it is too complicaed.

5 Least-Square FEM and PINNs

5.1 Least-Square FEM (LS-FEM)

$$\begin{cases} -\Delta u = f \quad \text{in } \Omega \\ u(\partial\Omega) = g. \end{cases}$$

Our Approach:

- Pointwise FD:
$$-\underbrace{\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{\Delta x^2}}_{\partial^2 u/\partial x^2} - \underbrace{\frac{u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta y^2}}_{\partial^2 u/\partial y^2} = f_{i,j}$$

- Weak formulation (Variational): FEM
$$\frac{1}{2}\int_\Omega |\boldsymbol{\nabla} u|^2 - \int_\Omega fu \implies \text{Find } u \in V \ s.t. \ \int_\Omega \boldsymbol{\nabla} u \boldsymbol{\nabla} v = \int_\Omega fv \quad \forall\, v \in V.$$

- Minimization problem:
$$\min_u \frac{1}{2}\int_\Omega |\boldsymbol{\nabla} u|^2 - \int_\Omega fu,$$
where $u \in$ Deep Neural Network $\to$ Deep Ritz

- Least Square:
$$J(u; f, g) = w_1 \| -\Delta u - f\|^2_{L^2(\Omega)} + w_2 \|u - g\|_{\partial\Omega},$$
where $f \in L^2(\Omega)$ and $u \in H^2$. The problem is find u such that $J(u; f, g) \le J(u + \delta v; f, g)$.

1. Let $u \in \mathscr{L}_g \oplus H^1_0 \cap H^2$ (when $g = 0$, $u \in H^1_0 \cap H^2$). Consider $J = \frac{1}{2}\|\Delta u + f\|^2$.
Goal: $J(u; f) \le J(u + \delta v; f)$.
Note that
$$\begin{aligned} J(u + \delta v; f) - J(u; f) &= \frac{1}{2}\int_\Omega (\Delta u + \delta\Delta v + f)(\Delta u + \delta\Delta v + f) - \frac{1}{2}\int (\Delta u + f)(\Delta u + f) \\ &= \int_\Omega \Delta v(\Delta u + f) + \delta(\Delta v)^2 \end{aligned}$$
So,
$$\begin{aligned} \lim_{\delta\to 0}\frac{1}{\delta}(J(u + \delta v; f) - J(u; f)) &= \lim_{\delta\to 0}\int_\Omega \Delta v(\Delta u + f) + \delta(\Delta v)^2 \\ &= \int_\Omega \Delta v(\Delta u + f) \\ &= 0. \end{aligned}$$
So, Euler-Lagrangian condition is
$$\int_\Omega (\Delta u + f)\Delta v = 0 \implies \int_\Omega \Delta u \Delta v = \int_\Omega f\Delta v.$$

Numerically, with FEM, $u_h \in V_h$, and

$$\int_\Omega \Delta u_h \Delta v_h = \int_\Omega f \Delta v_h.$$

But, the problem is $V_h \not\subset V$. So, this method is not practical.

2. Let $X = H^2 \cap H_0^1$ and $Y = L^2$. Norm equivalence gives

$$\alpha_1 \|u\|_X^2 \leq \underbrace{J(u;0)}_{\|\Delta u + f\|_Y} \leq \alpha_2 \|u\|_X^2$$

Then,

$$\alpha_1 \|u - u_h\|_X^2 \leq \|\Delta u + \Delta u_h\|_Y^2 = \|\Delta u + f\|_Y^2 = J(u;f) \leq \alpha_2 \|u - u_h\|_X^2.$$

So, $\exists, \alpha_1, \alpha_2 > 0$ such that

$$\alpha_1 \|u\|_{H^2 \cap H_0^1}^2 \leq \|\Delta u\|_{L^2}^2 \leq \alpha_2 \|u\|_{H^2 \cap H_0^1}^2.$$

Proof 1. For the second inequality,

$$\|u\|_{H^2 \cap H_0^1}^2 = \|u\|_{L^2}^2 + \|\boldsymbol{\nabla} u\|_{L^2}^2 + \left\|D^2 u\right\|^2,$$

where $\left\|D^2 u\right\|^2 = \|\Delta u\|^2 + \left\|\dfrac{\partial^2 u}{\partial x \partial y}\right\|^2$.
For the first inequality, it depicts "coercivity."

$$\int_\Omega u \Delta u = \underbrace{\int_{\partial\Omega} (\boldsymbol{\nabla} u \cdot \mathbf{n}) u}_{=0} - \int \boldsymbol{\nabla} u \boldsymbol{\nabla} v$$

$$\alpha \|u\|_{L^2}^2 \leq \|\boldsymbol{\nabla} u\|_{L^2}^2 = \left| \int \Omega u \Delta u \right| \leq \|u\|_{L^2} \|\Delta u\|_{L^2} \leq \frac{\alpha}{2} \|u\|_{L^2}^2 + C \|\Delta u\|_{L^2}^2$$

$$\beta \|u\|_{L^2}^2 \leq \|\Delta u\|_{L^2}^2 \implies \|\boldsymbol{\nabla} u\|_{L^2}^2 \leq \gamma \|\Delta u\|_{L^2}^2$$

Q.E.D. ■

Pavel Bacher and Max Gurnzburger Consider the first order version of the same problem:

$$\begin{cases} \boldsymbol{\sigma} - \boldsymbol{\nabla} u = 0 \\ \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + f = 0, \end{cases}$$

where $u \in H_0^1$, $\boldsymbol{\sigma} \in L^2$, and $\boldsymbol{\nabla} \cdot \boldsymbol{\sigma} \in L^2$, and

$$\frac{\partial \sigma_1}{\partial x_1} + \frac{\partial \sigma_2}{\partial x_2} + \frac{\partial \sigma_3}{\partial x_3} \in L^2 \implies \boldsymbol{\sigma} \in H(\Omega; \mathrm{div}).$$

- Variational formulation:

$$J = \frac{1}{2}\|\boldsymbol{\sigma} - \boldsymbol{\nabla} u\|_{L^2}^2 + \|\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f\|_{L^2}^2.$$

Find $\boldsymbol{\sigma}, u$ such that $J(\boldsymbol{\sigma}, u; f) \leq J(\boldsymbol{\rho}, v; f) \quad \forall\, \boldsymbol{\rho} \in H(\mathrm{div}), v \in H^1$. The first order condition is given by

$$\begin{cases} \int (\boldsymbol{\sigma} - \boldsymbol{\nabla} u)\boldsymbol{\rho} + \int (\boldsymbol{\nabla}\cdot\boldsymbol{\rho})(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f) = 0 \\ \int (\boldsymbol{\sigma} - \boldsymbol{\nabla} u)\cdot\boldsymbol{\nabla} v = 0. \end{cases}$$

We can use linear FEM: $V = H^1(\Omega)$, $Q = H(\mathrm{div})$. Choose $Q_h \subset Q$ and $V_h \subset V$.

- Do we have norm equivalence? Yes.

$$\|\boldsymbol{\sigma}\|_{H(\mathrm{div}}^2 = \|\boldsymbol{\sigma}\|_{L^2}^2 + \|\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}\|_{L^2}^2$$
$$\alpha_1\Big(\|u\|_{H^1}^2 + \|\boldsymbol{\sigma}\|_{H(\mathrm{div})}^2\Big) \leq J(\boldsymbol{\sigma}, u; 0) \leq \alpha_2\Big(\|u\|_{H^1}^2 + \|\boldsymbol{\sigma}\|_{H(\mathrm{div})}^2\Big).$$

- Benefits of LS formulation, and compare it with the mixed method.
Recall that we had mixed method

$$\begin{cases} \int_\Omega \boldsymbol{\sigma}\cdot\boldsymbol{\rho} + u\boldsymbol{\nabla}\cdot\boldsymbol{\rho} = 0 \\ \int_\Omega (\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f)v = 0. \end{cases}$$

1. Why we don't go with this formulation?
We require an inf-sup condition for mixed method, but we don't need this condition anymore with the LS formulation.
2. Why?
The solution to the LS problem is the normal equations

$$A^\top A\mathbf{u} = A^\top \mathbf{b},$$

where $A^\top A$ is SPD. So, we don't need inf-sup condition.
3. Some practical questions:
   - We include some additional quadrature error:

$$\int \approx \sum.$$

   - With boundary conditions,

$$\|\text{Residual}\|_{L^2}^2 + \|\text{Residual of BCs}\|_{L^2}^2$$

We have no Poincaré Inequality, and norm equivalence is much more harder to obtain. So, we don't have control in J, which means no control in error.

– What if we show advection-diffusion-reaction problem with LS?

$$\mathscr{L}u = -\mu\Delta u + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}u(+\gamma u)$$
$$\int_\Omega (\mathscr{L}u - f)\mathscr{L}v = 0$$

(a) We've seen this in GaLS, in which we consider it elementwise and it served as a stabilization term.

(b) With LS, assume $u_h \in H^2 \cap H_0^1$. The normal equations are SPD. So, we have no oscillations.

The first order formulation is

$$\begin{cases} \boldsymbol{\sigma} - (-\mu\boldsymbol{\nabla}u + \boldsymbol{\beta}u) = 0 \\ \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f = 0 \end{cases}$$

We can also prove norm equivalence.

- A curl free formulation for Poisson equation:

$$\begin{cases} \boldsymbol{\sigma} - \boldsymbol{\nabla}u = 0 \\ \boldsymbol{\nabla}\cdot\boldsymbol{\sigma} + f = 0 \\ \boldsymbol{\nabla}\times\boldsymbol{\sigma} = 0 \end{cases}$$

1. $(u, \boldsymbol{\sigma}) \in H_0^1 \times [H^1]^d$
2. $\boldsymbol{\nabla}\times\boldsymbol{\sigma} = 0$ is always true in theory, but it may not be true numerically.

5.2 Physics-Informed Neural Networks (PINNs)

Set-up:

$$-\Delta u = f \implies -\Delta u - f = 0$$
$$\sum w_i |\Delta u(x_i) + f(x_i)|^2 = 0$$

Consider u a neural network $\mathrm{NN}(\theta)$.

- Adding BC: $u = g$ on $\partial\Omega$.

$$\min w_1 \sum w_i |\Delta u(x_i) + f(x_i)|^2 + w_2 \sum \widetilde{w}_j |u(x_j) - g(x_j)|^2$$

- Strength: We don't need geometry information. We don't need a mesh! (But it is always better if we could have some geometry information...)
- It is natural to use PINNs for Data Assimilation problems. We just add an extra term to measure data mesh mismatch.

Numerical Evidence:

- The architecture of the NN has a limited impact on accuracy and efficiency
- Strong enforcement of BC works better than the weak one (including the BC term in the minimization problem)
- First-order formulations work better and with ReLU networks.
- No need for inf-sup for the mixed formulation.
- No need of special formulation for the stabilization.
- Great potential for the Data Assimilation.

LS Formulation of PINNs:

- $\mathcal{L} : X \to Y$, $\mathcal{L}u = f$ in Ω, $u = 0$ on $\partial\Omega$.
  Assume BC is strongly encoded in X (we manually set the boundary entries to 0).
- **Hypothesis: (Norm Equivalence)** $\exists$ two positive constants α_1, α_2 such that $\forall\, v \in X$,

$$\alpha_1 \|v\|_X^2 \leq \|\mathcal{L}v\|_Y^2 \leq \alpha_2 \|v\|_X^2.$$

- Define $\mathcal{J}(u, f) = \|\mathcal{L}u - f\|_Y^2$. THe problem is reformulated as finding $u \in X$ such that

$$u = \arg\min_{v \in X} \mathcal{J}(v, f).$$

- In PINNs, replace $\mathcal{J}(v, f)$ with

$$\mathcal{J}_N(v, f) = \frac{|\Omega|}{N} \sum_{i=1}^{N} |\mathcal{L}v(x_i) - f(x_i)|^2.$$

  For $Y = L^2(\Omega)$, $\mathcal{J}_N(v, f)$ is a quadrature approximation of $\mathcal{J}(v, f)$.
- Ceà Lemma in PINNs:
  If the adopted NN generates functions in $X_{\text{NN}} \subset X$, then,

$$\|u_{\text{PINN}} - u\|^2 \leq C_1 + C_2 N^{-s/2} + \frac{\alpha_2}{\alpha_1} \inf_{u_{\text{NN}} \in X_{\text{NN}}} \|u - u_{\text{NN}}\|_X,$$

  where s depends on the regularity of the residual $\mathcal{L}u - f$ in Y and the distribution of the quadrature nodes, and C_1 depends on the accuracy of the optimization algorithm.

Consequences for the Poisson Problem:

- The second order (traditional) formulation works ($\alpha_1 \|v\|_{H^2}^2 \leq \|\Delta v\|_{L^2}^2 \leq \|v\|_{H^2}^2$) only if the activation function are regular enough. [*ReLU does not work, but* $\tanh$ *works.*]
- The Poincaré inequality is critical: A weak enforcement of the BC suffers because the coercivity control works in weaker spaces ($H^{1/2}(\Omega)$).

- The first order traditional div-grad formulation works for NN generating functions in H^1 and $H(\mathrm{div})$. So, ReLU works here!

$$u \in H_0^1(\Omega), \quad \boldsymbol{\sigma} \in H(\Omega;\mathrm{div}) \; s.t. \begin{cases} \boldsymbol{\nabla} u - \boldsymbol{\sigma} = 0 \\ -\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = f \end{cases}$$

$$\|v\|_{H^1} + \|\boldsymbol{\tau}\|_{H(\mathrm{div})} \leq \|\boldsymbol{\nabla} v - \boldsymbol{\tau}\|_{L^2} + \|\boldsymbol{\nabla}\cdot\boldsymbol{\tau}\|_{L^2}$$

- The (redundant) curl-free formulation works with a stronger control ($X = H_0^1 \cap [H^1]^d$):

$$u \in H_0^1(\Omega), \quad \boldsymbol{\sigma} \in H(\Omega;\mathrm{div}) \; s.t. \begin{cases} \boldsymbol{\nabla} u - \boldsymbol{\sigma} = 0 \\ \boldsymbol{\nabla}\times\boldsymbol{\sigma} = 0 \\ -\boldsymbol{\nabla}\cdot\boldsymbol{\sigma} = f \end{cases}$$

$$\|v\|_{H^1} + \|\boldsymbol{\tau}\|_{H^1} \leq \|\boldsymbol{\nabla} v - \boldsymbol{\tau}\|_{L^2} + \|\boldsymbol{\nabla}\cdot\boldsymbol{\tau}\|_{L^2} + \|\boldsymbol{\nabla}\times\boldsymbol{\tau}\|_{L^2}.$$

Remark 1. (Comments on LS-FEM and PINNs).

- For LS-FEM, we have the norm equivalence

$$\alpha_1\|u\|_X^2 \leq J(u;0,0) \leq \alpha_2\|u\|_X^2 \quad \forall\, u \in X$$

  1. With this norm equivalence, we have control of error.
  2. If we don't have Poincaré Inequalities, it is hard to prove norm equivalence. We should impose strong enforcement of BC to have this norm equivalence result.

- Source of error:

  1. (LS)-FEM: $\underbrace{\text{Discretization Error}}_{\text{dominating}}$, Consistency, Quadrature Error, Linear Algebra Error.
  2. PINNs: Discretization Error, Quadrature Error, $\underbrace{\text{Optimization Error}}_{\text{dominating}}$.

- PINNs are excellent in solving inverse problems but questionable in solving forward problems.

6 Hyperbolic Problems

6.1 Finite Differences

- We discretize the semi-plane $\{(x,t) : 0 < t, -\infty < x < \infty\}$ with time step Δt and mesh size h.
- The grid nodes (x_j, t^n) denotes where $x_j = jh, \quad j \in \mathbb{Z}$ and $t^n = n\Delta t, \quad n \in \mathbb{N}$.
- Denote the numerical solution u_j^n that approximates $u(x_j, t^n)$ for any j and n.
- For simplicity, set $\lambda = \dfrac{\Delta t}{h}$.

6.1.1 Conservation Laws

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0.$$

- Backward Euler/Centered (BE/C): Implicit method.

$$\left.\frac{\partial u}{\partial t}\right|_{t^{n+1}} \approx \frac{u_j^{n+1} - u_j^n}{\Delta t} \qquad \text{and} \qquad a\frac{\partial u^{n+1}}{\partial x} \approx \frac{a}{2}\frac{u_{j+1}^{n+1} - u_{j-1}^{n+1}}{h} \quad \leftarrow \text{centered difference}$$

$$\boxed{u_j^{n+1} + \frac{\lambda}{2}a\Big(u_{j+1}^{n+1} - u_{j-1}^{n+1}\Big) = u_j^n}.$$

  At each time step, the solution is obtained by solving a linear system.

- Upwind (UPW): Explicit method.

$$a\frac{\partial u}{\partial x} = \frac{a}{2}\frac{u_{j+1}^n - u_{j-1}^n}{h} - \frac{|a|}{2}\cdot h\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}$$
$$u_j^{n+1} = u_j^n - \frac{\lambda}{2}a\big(u_{j+1}^n - u_{j-1}^n\big) + \frac{\lambda}{2}|a|\big(u_{j+1}^n - 2u_j^n + u_{j-1}^n\big)$$

- Lax-Wendroff (LW): Explicit method.

  This method is based on Taylor expansion.

$$u_j^{n+1} = u_j^n + \Delta t\left.\frac{\partial u}{\partial t}\right|_{t^n} + \frac{\Delta t^2}{2}\left.\frac{\partial^2 u}{\partial t^2}\right|_{t^n} + \cdots$$
$$\frac{\partial u}{\partial t} = -a\frac{\partial u}{\partial x} \implies \frac{\partial^2 u}{\partial x \partial t} = -a\frac{\partial^2 u}{\partial x^2} \implies \frac{\partial^2 u}{\partial t^2} = -a\frac{\partial^2 u}{\partial t \partial x} = -a\left(-a\frac{\partial^2 u}{\partial x^2}\right) = a^2\frac{\partial^2 u}{\partial x^2}$$

  So,

$$\begin{aligned} u_j^{n+1} &= u_j + \Delta t\left(-a\left.\frac{\partial u}{\partial x}\right|_{t^n}\right) + \frac{\Delta t^2}{2}a^2\left(\left.\frac{\partial^2 u}{\partial x^2}\right|_{t^n}\right) \\ &= u_j^n - \frac{\Delta t}{2}a\frac{u_{j+1}^n - u_{j-1}^n}{h} + \frac{\Delta t^2}{2}a^2\frac{u_{j+1}^n - 2u_j^n + u_{j+1}^n}{h^2} \\ &= u_j^n - \frac{\lambda}{2}a\big(u_{j+1}^n - u_{j-1}^n\big) + \frac{\lambda^2}{2}a^2\big(u_{j+1}^n - 2u_j^n + u_{j+1}^n\big). \end{aligned}$$

- For UPW and LW, they are explicit methods because solution in t^{n+1} depends on the solutions at the previous time steps, so no linear systems need to be solved.
- The additional term in UPW and LW can be considered as some diffusive term (which are purely numerical). They can also be viewed as stabilizing terms.
- Generalization of UPW/LW to more complicated PDEs: Consider the generalized conservation law:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} q(u) = 0.$$

UPW and LW can be written in the form

$$\boxed{u_j^{n+1} = u_j^n - \lambda\Big(H_{j+1/2}^n - H_{j-1/2}^n\Big)},$$

with a suitable choice of $H_{j+1/2} = H(u_j, u_{j+1})$:

1. UPW:
$$H_{j+1/2} = \frac{1}{2}[a(u_{j+1} + u_j) - |a|(u_{j+1} - u_j)]$$
2. LW:
$$H_{j+1/2} = \frac{1}{2}\big[a(u_{j+1} + u_j) - \lambda a^2(u_{j+1} - u_j)\big].$$

$H(\cdot,\cdot)$ is called *numerical flux*.

6.1.2 Methods for Linear System

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

- Generalization of the scalar case:

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \lambda\Big(\mathbf{H}_{j+1/2}^n - \mathbf{H}_{j-1/2}^n\Big),$$

where $\mathbf{u}_j^n$ is the approximation vector of $\mathbf{u}(x_j, t^n)$ and $\mathbf{H}_{j+1/2}$ is the *vector numerical flux*. The numerical flux is obtained by generalization of the scalar case.

- Upwind for Systems (UPW-S):

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{\lambda}{2}A(\mathbf{u}_{j+1}^n - \mathbf{u}_{j-1}^n) + \frac{\lambda}{2}|A|\big(\mathbf{u}_{j+1}^n - 2\mathbf{u}_j^n + \mathbf{u}_{j-1}^n\big).$$

The numerical flux is

$$\mathbf{H}_{j+1/2} = \frac{1}{2}[A(\mathbf{u}_{j+1} + \mathbf{u}_j) - |A|(\mathbf{u}_{j+1} - \mathbf{u}_j)].$$

- Lax-Wendroff for Systems (LW-S):

$$\mathbf{u}_j^{n+1} = \mathbf{u}_j^n - \frac{\lambda}{2}A\big(\mathbf{u}_{j+1}^n - \mathbf{u}_{j-1}^n\big) + \frac{\lambda^2}{2}A^2\big(\mathbf{u}_{j+1}^n - 2\mathbf{u}_j^n + \mathbf{u}_{j-1}^n\big),$$

and the associated numerical flux is

$$\mathbf{H}_{j+1/2} = \frac{1}{2}\big[A(\mathbf{u}_{j+1} - \mathbf{u}_j) - \lambda A^2(\mathbf{u}_{j+1} - \mathbf{u}_j)\big].$$

6.1.3 Wave Equation

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

- Wave equation can be written as system of conservation laws. So, we can use UPW or LW for systems.
- Leap-Frog:

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \gamma^2 \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2} = 0$$
$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = (\gamma\lambda)^2\big(u_{j+1}^n - 2u_j^n + u_{j-1}^n\big).$$

- (Optimal Newmark):

$$u_j^{n+1} - 2u_j^n + u_j^{n-1} = \frac{(\gamma\lambda)^2}{4}\Big(w_j^{n-1} + 2w_j^n + w_j^{n+1}\Big),$$

with $w_j^n = u_{j+1}^n - 2u_j^n + u_{j-1}^n$.

6.2 Analysis of FD Methods

6.2.1 Consistency

Definition 6.2.1 (Local Truncation Error/LTE). *Local Truncation Error (LTE)* of a numerical scheme is the residual obtained when the exact solution is plugged into the scheme.

From the classical Taylor expansions, we have

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

The scalar problem $\dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0$ gives

$$0 = \frac{\partial u}{\partial t} + a\left.\frac{\partial u}{\partial x}\right|_{(x_j, t^{n+1})} = \underbrace{\frac{u(x_j, {}^{n+1}) - u(x_j, t^n)}{\Delta t} + a\frac{u(x_{j+1}, t^{n+1}) - u(x_{j-1}, t^{n+1})}{2h}}_{\text{BE/C scheme}} + \underbrace{\boxed{\tau_j^{n+1}(\Delta t, h)}}_{\text{Truncation error}}$$

Definition 6.2.2 (Global Truncation Error). The global truncation error is given by

$$\tau(\Delta t, h) = \max_{j,n} \left|\tau_j^n\right|.$$

Definition 6.2.3 (Consistency). If $\tau(\Delta t, h)$ tends to 0 for Δt and h independently going to 0, we say that the numerical scheme is *consistent*. Suppose the solution of the exact problem is smooth enough. The method is said to be of *order* p *in time and of order* q *in space*, where p, q are integers, if we have

$$\tau(\Delta t, h) = \mathcal{O}(\Delta t^p + h^q).$$

Theorem 6.2.4 Order of Convergence of FD Schemes

- BE/C: $\mathcal{O}(\Delta t + h^2)$ [h^2 *because we used centered difference in space*]
- UPW: $\mathcal{O}(\Delta t + h)$
- LW: $\mathcal{O}(\Delta t^2 + h^2 + \Delta t h)$ [$\Delta t^2 + h^2$ *is the dominating terms. Usually,* Δt *and* h *scales in the same way. So,* $\Delta t h^2$ *is a third order term.*]

6.2.2 Convergence

Definition 6.2.5 (Convergence). The scheme is said ot be *convergent* if

$$\lim_{\Delta t, h \to 0} \max_{j,n} \left|u(x_j, t^n) - u_j^n\right| = 0.$$

Theorem 6.2.6 Lax-Richtmyer's Equivalence Theorem
For linear problems, a consistent method is convergent $\iff$ it is stable.

Hence, convergence is guaranteed provided that numerical errors are kept under control by acting on h and Δt.

6.2.3 Stability

Definition 6.2.7 (Stability). A numerical method is *stable* if for any time T, there are two constants $C_T > 0$ (possibly depending on T, but not on Δt or h) and $\delta_0 > 0$ such that

$$\|\mathbf{u}^n\|_\Delta \leq C_T \left\|\mathbf{u}^0\right\|_\Delta,$$

for all n such that $n\Delta t \leq T$, for all $\Delta t, h$ such that $0 < \Delta t, h \leq \delta_0$, and for all initial data $\mathbf{u}^0$.

Note that in the above definition, $\|\cdot\|_\Delta$ is a suitable *discrete norm*. For example,

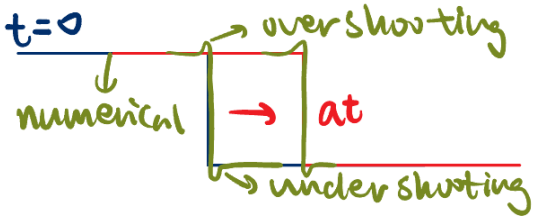
$$\|\mathbf{v}\|_{\Delta,p} = \left(h \sum_{j=-\infty}^{\infty} |v_j|^p \right)^{1/p} \quad \text{for } p = 1, 2 \quad \text{and} \quad \|\mathbf{v}\|_{\Delta,\infty} = \sup_j |v_j|.$$

Definition 6.2.8 (Strongly Stable). If the scheme is stable and

$$\|\mathbf{u}^n\|_\Delta \leq \left\|\mathbf{u}^{n-1}\right\|_\Delta \quad \forall\, n \geq 1,$$

(i.e., $C_T = 1$), then the scheme is *strongly stable* with respect to $\|\cdot\|_\Delta$.

Strong stability is useful for asymptotic in time computations ($T \gg 1$). Actually, it ensures that the numerical solution is bounded $\forall\, T$. If we don't have strong stability, we will suffer from overshooting and undershooting problems:



Definition 6.2.9 ((Un)Conditional Stability).

- The numerical method is *conditionally stable* if its stability result can hold under restrictions on h and Δt.
- It is *unconditionally stable* otherwise.

BE/C The method BE/C is *unconditionally* strongly stable in norm $\|\cdot\|_{\Delta,2}$.

CFL Condition An explicit scheme is never unconditionally stable.

A necessary condition is that Δt and h fulfill the *CFL condition*.

Definition 6.2.10 (CFL (Courant-Friedrichs-Levy) Condition). For the problem

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0,$$

the CFL condition is given by

$$|a\lambda| \leq 1 \implies \Delta t \leq \frac{h}{|a|},$$

where a is the velocity, and $\lambda = \dfrac{\Delta t}{h}$. The adimensional number $a\lambda$ is the *CFL number*.

Remark.

- Recall that for parabolic problems, we require $\Delta t \leq ch^2$, so the condition is harder to fulfill. But here, we only require $\Delta \leq ch$, os it is easier to fulfill, and the results of explicit methods will not be very bad.
- If a is not constant, CFL condition reads

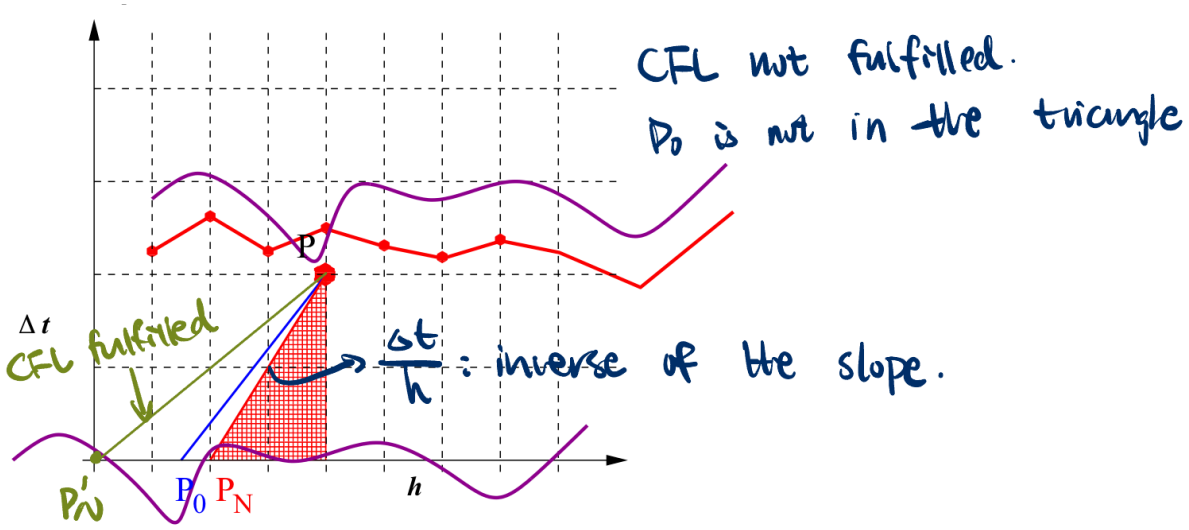
$$\Delta t \leq \frac{\Delta x}{\sup\limits_{x\in\mathbb{R},t>0} |a(x,t)|}.$$

- For a system, CFL is given by

$$\left|\lambda_k \frac{\Delta t}{\Delta x}\right| \leq 1, \quad k = 1, \ldots, p,$$

where $\{\lambda_k \mid k = 1, \ldots, p\}$ are the eigenvalues of A.

Graphically, CFL requires that propagation velocity of the numerical solution $(h/\Delta t)$ is greater than the velocity of the analytical solution $(|a|)$. If CFL condition is satisfied, all the data affecting the exact analytical solution in a point also affect the numerical one.



We prevent initial data to be perturbed so to affect the exact solution and not the numerical one, mandatory for the convergence.

[*If CFL is satisfied: numerical domain of convergence contains the exact domain of convergence*].

UPW and LW

Theorem 6.2.11
If CFL is satisfied, UPW and LW are *strongly stable* is the norm $\|\cdot\|_{\Delta,1}$.

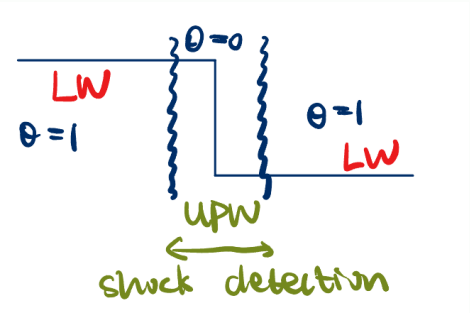
Theorem 6.2.12 Discrete Maximum Principle
Under CFL condition, UPW satisfies also the inequality

$$\|\mathbf{u}^n\|_{\Delta,\infty} \leq \left\|\mathbf{u}^0\right\|_{\Delta,\infty} \quad \forall\, n \geq 0.$$

So, UPW is strongly stable in the norm $\|\cdot\|_{\Delta,\infty}$.

Remark.

- So, we have no undershooting or overshooting for UPW with CFL.
- For UPW, we trade accuracy with stability: it is less accurate but more stable.
- Improvements: shock capturing methods: combine LW and UPW
  1. Numerical flux will be
  $$\theta H_{\text{LW}} + (1-\theta) H_{\text{UPW}},$$
  where $\theta = 1$ for regular solution, and $\theta = 0$ for poor regularity.
  2. Depending on how we detect shock/irregularity, we form different methods.



For Wave Equation

- Leap-Frog is stable under CFL condition
$$\Delta t \leq \frac{\Delta x}{|\gamma|}.$$
- Optimal Newmark method is unconditionally stable.

6.3 Von Neumann Analysis of FD Methods

Analyze the stability in $\|\cdot\|_{\Delta,2}$

$$\begin{aligned} u_0(x) &= \sum_{k=-\infty}^{\infty} a_k e^{\mathrm{i}kx} \\ a_k &= \frac{1}{2\pi}\int_0^{2\pi} u_0(x) e^{-\mathrm{i}kx}\,\mathrm{d}x \\ u_j^n &= \sum_{k=-\infty}^{\infty} a_k e^{\mathrm{i}kjh}\gamma_k^n, \quad j = 0, \pm 1, \pm 2, \ldots, n \geq 2. \end{aligned}$$

Exact solution to

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0$$

is given by

$$u(x, t^n) = u_0(x - an\Delta t), \quad \forall\, n \geq 0,\ \forall\, x \in \mathbb{R}.$$

With Taylor expansion, we have

$$u(x_j, t^n) = \sum a_k e^{\mathrm{i}kjh} g_k^n, \quad g_k = e^{\mathrm{i}ak\Delta t},$$

where g_k^n is called the *time travel coefficient*.

γ_j corresponds to g_k, depends on the numerical schemes.

- g_k and γ_k are complex numbers, so we can analyze its magnitude and phase angle.
- Analyzing its magnitude, phase angle, and frequency tells us how solution evolve over time.
- Under CFL, $|\gamma_k| < 1$ for all methods.
    1. **Dissipation**: how many damping/amplitude.
    2. **Dispersion**: how much delay/velocity.

6.4 Finite Elements

FEMs for hyperbolic problems are less stable. We also need to do "semi-discretization." Depending on how we do time discretization, we form different methods.

- UPW-FEM:
$$\text{CFL} < \frac{1}{3}$$
- LW-FEM:
$$\text{CFL} < \frac{1}{\sqrt{3}}.$$

Under CFL, the two methods are conditionally strongly stable.