Condition number of the sitffness matrix:

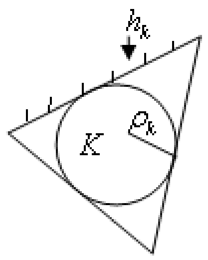
Assume the triangulation $J_h$ satisfies

$h_k \ge \beta_1 h \quad h = \max_{k \in J_h} h_k,\ h_k =$ diameter of element $k$

$\frac{\rho_k}{h_k} \ge \beta_2 \quad \rho_k =$ radius of the circle incribed in $k = \int \nabla u \nabla v dx$

Consider bilinear form $a(u,v)$ satisfying the coercivity and continuity on $H^1_{2,0}(\Omega)$ $\left(V_h \subset H^1_{2,0}\right)$

$$\begin{cases} (i)\, a(u,v) \ge \partial \|u\|^2 \\ (ii)\, a(u,v) \le \beta \|u\| \|v\| \quad \left(\Omega \subset R^2\right) \end{cases}$$




The following Lemma holds

Lemma 1. $\exists$ constants $c$ and $C$(depends on $\alpha, \beta$) such that for all

$$v_h = \sum_{i=1}^m v_i \varphi_i \in V_h,\ \text{the following inequalities hold.}$$

$$(10)\quad ch^2 |\vec{v}|^2 \le \|v_h\|^2 \le Ch^2 |\vec{v}|^2 \quad \left(\vec{v} = (v_1, v_2, \cdots, v_m)\right)$$

$$(11)\, a(v_h, v_h) \equiv \int_\Omega |\nabla v_h|^2 dx \le ch^{-2} \|v_h\|^2 \quad \left(\begin{array}{l}\text{inverse estimate compare with poincare} \\ \text{inequality } \|v\|_{L^2} < \|\nabla v\|_{L^2}\end{array}\right)$$

(prove: Skip. See claes Johnson's section 7.7)

(exercise)

With the help of the Lemma 1, we can show the condition number of the stiff matric $K$ is $O\left(h^{-2}\right)$.

$cond\left(K\right)=\|K\|\|K^{-1}\|_{(\nwarrow \text{matrix norm})}$ $\quad \|K\|=\sup\limits_{x\in R^n}\left(\dfrac{\|Ax\|}{\|x\|}\right)$

Since

$$\frac{\vec{v}^T K\vec{v}}{\left|\vec{v}\right|^2}=\frac{a\left(v_h,v_h\right)}{\left|\vec{v}\right|^2}\overset{(11)}{\le} ch^{-2}\frac{\|v_h\|^2}{\left|\vec{v}\right|^2}\overset{(10)}{\le} c^* \underset{\vec{v}}{\overset{\sup}{\Rightarrow}} \lambda_{\max}\le c^{**}\begin{pmatrix}\vec{v}:\text{evector corresponding}\\ \text{to max evalue}\end{pmatrix}$$

and $\dfrac{\vec{v}^T K\vec{v}}{\left|\vec{v}\right|^2}=\dfrac{a\left(v_h,v_h\right)}{\left|\vec{v}\right|^2}\ge\partial\dfrac{\|v_h\|^2}{\left|\vec{v}\right|^2}\ge c^{**}h^2\underset{\inf}{\Rightarrow}\lambda_{\max}\ge c^{**}h^2\begin{pmatrix}\text{v: evector corresponding}\\ \text{to min evalue}\end{pmatrix}$

$\Rightarrow \lambda_{\max}\left(K\right)\le C^*$ and $\lambda_{\min}\left(K\right)\ge c^{**}h^2$

$\Rightarrow cond\left(K\right)\le \widetilde{Ch^{-2}}\left(\widetilde{C}=\dfrac{c^*}{c^{**}}\right)$

$\Big($exercise: In 3D, $ch^3\left|\vec{v}\right|\le\|u\|^2<ch^3\left|v\right|\ \Rightarrow cond\left(K\right)<\widetilde{Ch^2}\Big)$

Remark:

(1) Recall that when solving linear system $Ax=b$ by iterative method, (let $x$ be the iterative solution i.e. $Ax=b+\Delta b$), we have

$\dfrac{\left|\Delta x\right|}{\left|x\right|}\le cond\left(A\right)\dfrac{\left|r\right|}{\left|b\right|}$ $\quad$ here $r=b-Ax$ (the residual)

If the relative error is required to be less than $\varepsilon$, the relative residual $\left(\dfrac{\left|r\right|}{\left|b\right|}\right)$

should be required to be less that $\dfrac{\varepsilon}{cond\left(A\right)}\underset{\left(\uparrow \text{FEM stiff matirx}\right)}{\approx}\varepsilon\cdot h^2$

$\Rightarrow \underset{\uparrow\text{vector}}{\dfrac{\left|\Delta x\right|}{\left|x\right|}}<\varepsilon\Rightarrow\left|\Delta x\right|<\varepsilon\left|x\right|\underset{(10)}{<}\varepsilon\|x_h\|h^{-1}$

(2) How large $\varepsilon$ should be?

Recall that $\|x_{\text{trur}} - x_{\text{FEM}}\|_{L^2} \underset{\uparrow \text{Duality argument}}{<} ch^2 \|x_{\text{true}}\|_{H^2(\Omega)} \underset{\substack{\updownarrow \text{Assuming} \\ \text{regularity} \\ \text{estimate}}}{<} ch^2 \|f\|$

($x_{\text{FEM}}$ is the FEM solution, $x_{true}$ is the PDE solution)

Since we don't want $|\Delta x| > \|x_{\text{true}} - x_{\text{FEM}}\|$, $\left(\begin{array}{r} \text{i.e. we want } |\Delta x| \ll h^2 \|f\| \\ |\Delta x| < \varepsilon |x_{\text{FEM}}| \end{array}\right)$

$(\Delta x = x_{\text{iter}} - x_{\text{FEM}})$

we set $\varepsilon \ll \underbrace{\frac{\|f\|}{|x_{\text{FEM}}|} \cdot h^2}_{v}$

$$\Rightarrow \varepsilon \ll \underset{\substack{\downarrow \\ \text{r dep on } f \\ \text{and } x_{\text{true}}}}{r} \cdot h^3, \quad \frac{\|f\|}{\|x_{\text{FEM}}\|} h^3 > \frac{\|f\|}{\|x_{\text{true}}\|} \cdot \frac{\|x_{\text{true}}\|}{\|x_{\text{true}}\| + ch^2 \|x_{\text{true}}\|_{H^2}} h^3$$

The a posteriori error estimation:

$$-\Delta u + cu = f$$

consider $a(u,v) = \int \nabla u \nabla v + c \int uv$ and $u$ be the solution of

$a(u,v) = \int fv dx = L(f)$ $\|u\|_E$ :the energe norm\

let $\begin{cases} e = u - u^h \\ e^h = u^h - u^h, \text{ here } u^h = Iu \text{ (interpolant of } u\text{)} \\ \eta = u^h - u \end{cases}$

clearly we have $e = e^h + \eta$

Since

$$\|e\|_E^2 = \left|a(e,e)\right| = \left|a\left(e^h+\eta,e\right)\right| \overset{\substack{\text{orthogonality} \\ (a(x_h,e)=0\ \varphi_h \in V_h)}}{=} \left|a(\eta,e)\right|$$

$$= \left|a\left(\eta,u-u^h\right)\right|$$

$$= \left|a(\eta,u)-a\left(\eta,u^h\right)\right|$$

$$= \left|L(\eta)-\left(\int_\Omega \nabla\eta\nabla u^h + c\int_\Omega \eta u^h\right)\right|$$

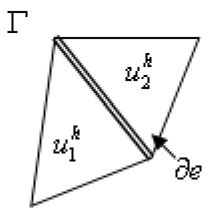
$$= \left|L(\eta)-\sum_{K\in J_h}\int_e \nabla\eta\nabla u^h + c\int_\Omega \eta h^h\right|$$

$$\overset{\left(\substack{\text{integration}\\ \text{by part}}\right)}{=} \left|L(\eta)+\int_\Omega \eta\left(\Delta u^h - cu^h\right)+\sum_{\substack{s\in\partial\\ K\in J_h}}\int_K \eta \underbrace{\left[\!\left[\nabla u^h\cdot\overrightarrow{n_1}\right]\!\right]}_{(*)}\right|$$

$$\le \underbrace{\left|\left\langle \eta, r^n\right\rangle\right|}_{(I)} + \sum_{s=1}^{\#\text{of edges}}\left|\int_s \underbrace{\eta(x)\underbrace{\left[\!\left[\nabla u_k^h\cdot\vec{n}_s\right]\!\right]}_{J_{k,s}^h}}_{(II)}\right|$$

$$r^h = f+\Delta u_h - cu_h$$

$$\left((*)\,\text{flux jump}: \int_{\partial e_1}\left(\nabla u_2^h-\nabla u_1^h\right)\eta = \left[\!\left[\underbrace{\left(\nabla u_2^h-\nabla u_1^h\right)}_{J}\cdot n\right]\!\right]\right)$$



Using more advance interpolation estimation, it can be shown

$$\begin{cases} \|u - Iu\|_{L^2(K)} \le c_1 h_k \|e\|_{\underset{\text{energy norm}}{\underset{\uparrow}{E}}, \underset{\text{element}}{\underset{\uparrow}{K}}} \\ \|u - Iu\|_{L^2(\partial K)} \le c_2 h_k^{\frac{1}{2}} \|e\|_{E,K} \end{cases}$$

$$\Rightarrow (I) \le \|h^{-1}\eta\| \|hr^h\|_{L^2} \le c \|hr^h\|_{L^2} \|e\|_E$$

$$(II) \le \sum_{K \in J_h} \sum_{s \to \partial K} h_k^{-\frac{1}{2}} \eta(x_k) \cdot h^{\frac{1}{2}} J_{k,s}^h$$

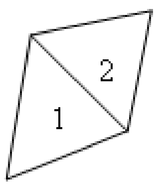
$$\le \sum_{K \in J_h} \sum_{s \to \partial K} c_2 \|e\|_{E,K} \left[ \int_{\partial s} \left( h^{\frac{1}{2}} J_{K,s}^h \right)^2 ds \right]^{\frac{1}{2}} \le \sum_{k \in J_h} \widetilde{c_2 \|e\|_{E,K}} \left( h J_{k,s}^h \right)$$

$$\le \widetilde{c_2} \left( \sum_{k \in J_h} \|e\|_{E,K}^2 \right)^{\frac{1}{2}} \left( \sum_{k \in J_h} \left( h J_{k,s}^h \right)^2 \right)^{\frac{1}{2}}$$

let $R_h = \sum_{k \in J_h} \left( J_{k,s}^h \right)$, we have

$$\|e\|_E^2 \le \tilde{\tilde{c}} \left( \|hr^h\|_{L^2(\Omega)} + \sum_{\substack{K \in J_h \\ s \in \text{edges of } J_h}} \underset{\substack{\Uparrow \\ \text{only depends on } u_h}}{h_s \cdot J_{K,s}^h} \right)$$

here $J_{K,s}^h = \begin{cases} \frac{1}{2} \left| \nabla u_1^h \cdot \overrightarrow{n_s} - \nabla u_2^h \cdot \overrightarrow{n_s} \right|, & s \text{ in the interior of } \Omega \\ 0 & \text{if } s \text{ on the } \partial\Omega \end{cases}$

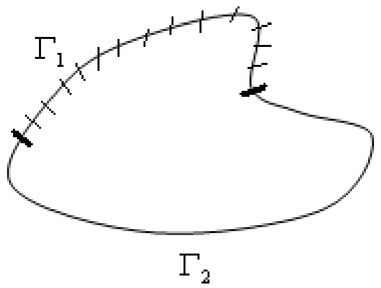



FEM for parabolic problem

consider the heat equation

$$(+) \begin{cases} \dot{u} - div(\mu \nabla u) = f \quad \text{in } \Omega \times I \\ u = 0 \text{ on } \Gamma_1 \times I \\ \mu \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma_2 \times I \\ u(x,0) = \dot{u}(x) \quad x \in \Omega \end{cases}$$

(initial boundary value problem)




$1-D$ Model problem:

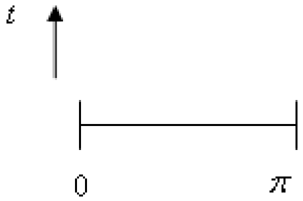
$$(+)\begin{cases} \dfrac{d}{dt}u-\dfrac{d^2u}{dx^2}=f \\ u(0,t)=u(\pi,t)=0 \\ u(x,0)=u^0 \end{cases}$$

In case $f=0$, by seperation of variables

consider $u(x,t)=e^{-u^2t}h(x)\ \Rightarrow \dfrac{\partial u}{\partial t}=-w^2e^{-w^2t}h(x),\ \dfrac{\partial^2 u}{\partial x^2}=e^{-w^2t}h''(x)$

$$\Rightarrow -w^2e^{-w^2t}h(x)-e^{-w^2t}h''(x)=0$$

$$\Rightarrow e^{-w^2x}\left(h''(x)+w^2h(x)\right)=0$$



Consider $h(x) = de^{iwx}$ for any $d \in \mathbb{C}$

From $u(0,t) = u(\pi,t) = 0,$ we have $h(x) = (-i)e^{iwx}$

$\underset{\text{real part}}{\Rightarrow}$ $h(x) = \sin wx$

Therefore a general solution of $(+)$ has the following form

$$u(x,t) = \sum_{w} we^{-w^2 t} \sin wx \quad -(**)$$

Since $u(x,0) = \sum_{w} c_w \sin wx = u^0 \Rightarrow \underset{\substack{\uparrow \\ \text{fourier coefficient}}}{c_w} = \frac{2}{\pi}\int(\sin wx)u^0 dx$

$\Rightarrow u(x,t) = \sum_{w=1}^{\infty} c_w e^{-w^2 t} \sin wx$

$\Rightarrow$ each component $\sin wx$ lives on a time scale $o(w^{-2})$

$\Rightarrow$ High freq modes quickly get damped! $-(1)$

$\Rightarrow$ The solution $u$ becomes smoother as $t \to \infty$

But $u(x,t)$ will not be smooth for $t \ll 1$

$$\left\|\dot{u}(t)\right\| = \left(\int_{\Omega} u^2(x,t)\,dx\right)^{\frac{1}{2}} = \left\|\frac{\partial^2 u}{\partial x^2}\right\| = \left\|\sum c_w w^2 e^{-iwt} \sin wx\right\| \to \infty \text{ as } t \to 0 \quad -(2)$$

$\left\|\dot{u}(t)\right\| = \left\|\frac{\partial^2 u}{\partial x^2}\right\| \to 0$ the rate is depends on how small $\underset{\substack{\uparrow \\ \text{multitude of} \\ \text{high freq modes}}}{c_w}$ is for large $w$

$\Rightarrow$ In general smooth initial $u^0$ gives small $c_w$ for large $w_2$

An initial phase for $t$ small where certain derivatives of $u$ are large is called an "initial transient"

Observation: based on $(1)\&(2)$

$(i)$ initial transient $\Rightarrow$ small time step for discretization of $\frac{d}{df}u$

if rough initial $u^0$ (oscillating discontinuity etc.)

$\Rightarrow$ small mech size for discretization of $u_{xx}$

$(ii)$ smooth $u(x,t)$ as $t \to \infty \Rightarrow$ largertime step and larger mesh size.

In general, we have the following estimation

$$\begin{cases} \|u(t)\| \le \|u^0\| & t \in I \\ \|\dot{u}(t)\| \le \frac{c}{t}\|u^0\| & t \in I \end{cases} \quad -(3)$$

Exercise: (1) prove (3) by using (**)

(2) prove (3) using Energy method

$$u\left(\frac{du}{dt} - \frac{d^2u}{dx^2}\right) = 0 \overset{(9)}{\Rightarrow} \int \frac{1}{2}\frac{du^2}{dt} + \int |\nabla u|^2 = 0$$

$$\Rightarrow \frac{1}{2}\frac{d}{dt}\|u\|^2 = -\|\nabla u\|^2 \underset{\text{poincare ineq.}}{<} -\|u\|^2$$

$$\Rightarrow \|u\| < e^{-t}\|u^0\| \Rightarrow \|u\| < \|u^0\|$$

Exercise: using Energy method to prove $\|\dot{u}\| \le \frac{1}{t}\|u^0\|$

Semi-discretization in space for $(+)$ with dinichlet data $(\Gamma_2 = \phi)$

consider $u_h(x,t) = \sum_{i=1}^{m} \psi_i(t)\varphi_i(x)$ $\begin{pmatrix} \text{in previous example} \\ \psi_i(t) = c_w e^{-w^2 t} \\ \varphi_i(x) = \sin wx \quad m = \infty \end{pmatrix}$

$$\Rightarrow \sum_{i=1}^{m} \dot{\psi}_i(x)\underbrace{(\varphi_i, \varphi_j)}_{M} + \sum_{i=1}^{m} \psi_i(t)\underbrace{\langle \nabla\varphi_i, \nabla\varphi_j \rangle}_{K}$$

$$\begin{cases} = \underbrace{\langle f(t), \varphi_j \rangle}_{F}, \quad j = 1 \sim m \\ \sum_{i=1}^{m} \psi_i(0)(\varphi_i, \varphi_j) = \underbrace{(u^0, \varphi_j)}_{U^0} \quad j = 1 \sim m \end{cases} \qquad \vec{\psi} = (\psi_1, \psi_2, \cdots \psi_m)$$

$$\underset{(4)}{\Rightarrow} \begin{cases} M\dot{\vec{\psi}} + K\vec{\psi} = F(t) \\ M\vec{\psi}(0) = U^0 \end{cases} \Rightarrow \text{system of ordianry equation}$$

Recall that the condition of $M$ $(\chi(M) = \Delta(1))$

and the condition of $K$ $(\chi(k) = o(h^{-2}))$ as $h \to 0$

cosider $\vec{\psi} = e^{-i\vec{w}t}\vec{v}$ for the homogeneous case $M\dot{\vec{\psi}} + K\vec{\psi} = 0$

$$\Rightarrow \left[-(Mi)\vec{w} + K\right]\vec{v} = 0 \underset{\substack{\text{Solve the} \\ \text{genearlized} \\ \text{eigenvalue} \\ \text{problem gives}}}{\Rightarrow} \text{ the eigenvalue } w \text{ and } V_w$$

the general solution can now be expressed as

$\vec{\psi} = \sum_{j=1}^{m} c_j e^{iw_j t} v_j$ here $\vec{c} = \{c_j\}_{j=1\sim m}$ satisfies $M \cdot V \cdot \vec{c} = U^0,\ V = [v_1, v_2, \cdots v_m]$

Similar to the model problem,

one has smooth mode corresponding to the $\min_{j=1\sim m}\{w_j\} \approx o(1)$

and highly oscillatory mode corresponding to the $\max_{j=1\sim m}\{w_j\} \approx o(h^{-2})$

$\vec{\psi}$ has components live on time scales from $\underset{(\text{smooth mode})}{o(h^{-2})}$ to $\underset{(\text{oscillatory mode})}{o(1)}$

(or say from $o(1)$ to $o(h^2)$)

As mentioned previously, one should use methods which adapt the size of each time step according to the smoothness of $\vec{\psi}$. Moreover, if explicit time discretization is employed, one should choose the step as small as $o(h^2)$ in order to prevent potential instability. $\left( \begin{array}{l} \Rightarrow \text{trancation error} \\ \Rightarrow o(\Delta t) = o(h^2) \end{array} \right)$

To avoid extreme small time step, implicit time discretization such as implicit Euler or Crnak-Nicolson method can be employed.