The Mathematical Theory of Finite Element Methods

Spring 2021

Brenner & Scott

Chapter 0 Basic Concepts

§0.1 Weak Formulation of BVPs

- (BVP) $\begin{cases} -u''(x) = f(x) \text{ in } I=(0,1) \\ u(0)=0 \text{ and } u'(1)=0 \end{cases}$

$\forall$ test function $v \in V = ?$

$$\int_0^1 f v\,dx = \int_0^1 (-u'' v)\,dx = \int_0^1 u'v'dx - [\underbrace{\boxed{u'(1)}}_{=0} v(1) - u'(0)v(0)]$$

$$= \int_0^1 u'v'dx \quad \text{if } v(0)=0$$

- (WF) or variational formulation (VF) Find $u \in V = \left\{ v \in L^2(0,1) \,\middle|\, \begin{array}{l} \int_0^1 (v')^2 dx < +\infty \\ v(0)=0 \end{array} \right\}$ s.t.

$$a(u,v) \equiv \int_0^1 u'v'dx = \underbrace{\int_0^1 f v\,dx}_{\equiv f(v)} \quad \forall v \in V$$

$u(0)=0$ essential BC

$u'(1)=0$ natural BC

Theorem Assume that $f \in C^0[0,1]$ and that the solution of (WF) $u \in C^2[0,1]$.

$\Rightarrow$ u satisfies (BVP).

Proof $\forall v \in V$

$$\int_0^1 f v\,dx = \int_0^1 u'v'dx = \int_0^1 (-u'')v\,dx - [u'(1)v(1) - u'(0)\underbrace{\boxed{v(0)}}_{=0}]$$

$\Longrightarrow \int_0^1 (f+u'')\,v\,dx = -u'(1)\,v(1) \qquad \forall v \in V$

$= 0, \quad \forall v \in V \cap \{v(1)=0\}$

$\Longrightarrow f(x)+u''(x)=0 \;\; \forall x \in (0,1)$

if not, $\exists\, (x_0,x_1) \subset (0,1)$ s.t. $f+u''>0$ in (x_0,x_1)

since $f+u'' \in C^0[0,1]$

choose $v(x) = (x-x_0)^2 (x-x_1)^2 \in V \cap \{v(1)=0\}$

$\Rightarrow 0 = \int_0^1 (f+u'')\,v\,dx = \int_{x_0}^{x_1} (f+u'')\,v\,dx > 0$

$\forall v \in V,\; u'(1)\,v(1)=0 \Longrightarrow u'(1)=0.$ #

Questions (1) meaning of $u'(x)$ and (2) well-posedness of (WF)

- (MF) Find $u \in V$ s.t. $J(u) = \min_{v \in V} J(v) \iff J(u) \le J(v) \;\forall v \in V$

where $J(v) = \frac{1}{2} a(v,v) - f(v)$ is the energy functional

Theorem (WF) $\iff$ (MF)

assumptions (1) $a(\cdot,\cdot)$ bilinear, $f(\cdot)$ linear
(2) $a(u,v)=a(v,u)$ sym
(3) $a(v,v) \ge 0 \;\; \forall v \in V$

Proof "$\Leftarrow$" $\forall v \in V$, let

$g(t) = J(u+tv) = J(u) + t\,[a(u,v)-f(v)] + \frac{1}{2}t^2 a(v,v)$

$J(u) \le J(v) \;\forall v \in V \Longrightarrow 0 = g'(0) = a(u,v) - f(v)$

"$\Rightarrow$" $\forall v \in V$, $J(v) = J(u+(v-u))$

$= J(u) + t\,[a(u,v-u) - f(v-u)] + \frac{t^2}{2} a(v-u,v-u)$

$= J(u) + \frac{t^2}{2} a(v-u,v-u) \ge J(u)$

#

HW #2, 3, 8

§0.2 Galerkin and Ritz Approximations

Let S be a finite dimensional subspace of V. (GA)

Galerkin approximation   Find $u_s \in S \subset V$ s.t. $\boxed{a(u_s, v) = f(v) \ \forall v \in S.}$

- System of Linear Equations

$$S = \text{span}\{\varphi_i(x)\}_{i=1}^n \Longrightarrow u_s = \sum_{j=1}^n u_j \varphi_j(x)$$

$v = \varphi_i$ $$f(\varphi_i) = a(u_s, \varphi_i) = \sum_{i=1}^n a(\varphi_j, \varphi_i)\, u_j = \left(a(\varphi_1,\varphi_i), \cdots, a(\varphi_n, \varphi_i)\right) \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

$$K_{n\times n} U_{n\times 1} = F_{n\times 1} \quad \text{where } K = (K_{ij}) \text{ with } K_{ij} = a(\varphi_j, \varphi_i)$$

$$U = (u_1, \cdots, u_n)^t, \quad F = (f(\varphi_1), \cdots, f(\varphi_n))^t$$

- property of matrix K

(1) K is symmetric

(2) K is positive definite $\Longleftrightarrow W^t K W > 0 \ \forall \vec{0} \neq W \in R^n$

Let $w(x) = \sum_{i=1}^n w_i \varphi_i(x)$ with $W = (w_1, \cdots, w_n)^t$

$$W^t K W = \sum_{i,j=1}^n w_i w_j K_{ij} = \sum_{i,j=1}^n a(w_i \varphi_i, w_j \varphi_j) = a(w, w) \geq 0$$

$$a(w,w) = \int_0^1 (w')^2 dx = 0 \Rightarrow w' = 0 \Rightarrow w \equiv \text{const.} \xRightarrow{w(0)=0} w \equiv 0.$$

Theorem Given $f \in L^2(0,1)$, (GA) has a unique solution.

Ritz Approximation Find $u_S \in S$ s.t. $\boxed{J(u_S) = \min_{v \in S} J(v)}$ (RA)

$\forall\, v \in S$, $\quad v(x) = \sum_{i=1}^{n} v_i \varphi_i(x)$

$$J(v) = \frac{1}{2} a(v,v) - f(v) = \frac{1}{2} \sum_{i,j=1}^{n} v_i v_j \, a(\varphi_i, \varphi_j) - \sum_{i=1}^{n} v_i f(\varphi_i)$$

$$= \frac{1}{2} V^t K V - V^t F$$

$$\nabla_V J(v) = KV - F \Longrightarrow KU = F$$

- Gradient Descent starting at an initial u^0

$$u^k = u^{k-1} - \tau \nabla J(u^{k-1}) \quad \text{for } k = 1, 2, \cdots$$

HW #1

§0.3 Error Estimates

Let $u \in V$ be the solution of (WF): $a(u,v) = f(v) \quad \forall\, v \in V$

Let $u_S \in S \subset V$ be the solution of (GA): $a(u_S, v) = f(v) \quad \forall\, v \in S \subset V$

$$\|u - u_S\|_? \le ?$$

- where to start?

error equation $\quad a(u - u_S, v) = 0 \quad \forall\, v \in S$ (the orthogonal property)

- the energy norm

$$\|v\|_E = \sqrt{a(v,v)} = \int_0^1 (v')^2 dx$$

- a tool (the Schwarz inequality)

$$|a(u,v)| \leq \|u\|_E \|v\|_E \quad \forall u, v \in V.$$

Theorem $\|u-u_S\|_E = \min_{v\in S} \|u-v\|_E.$

Proof Galerkin Approximation $\quad a(u_S, v) = f(v) \quad \forall v \in S$

$$\|u-u_S\|_E^2 = a(u-u_S, u-u_S) = a(u-u_S, u-v) \quad \forall v \in S$$
$$\leq \|u-u_S\|_E \|u-v\|_E$$

Ritz Approximation $\quad J(u_S) \leq J(v) \quad \forall v \in S$

$$\|u-u_S\|_E^2 = 2\big(J(u_S) - J(u)\big)$$
$$\leq 2\big(J(v) - J(u)\big) = \|u-v\|_E^2 \quad \forall v \in S. \qquad \#$$

Theorem Assume that $\inf_{v\in S} \|w-v\|_E \leq \varepsilon \|w''\| \quad \forall w$ s.t. $\|w''\| < +\infty$

$$\|w\| = \sqrt{\int_0^1 w^2 dx}$$

$$\Longrightarrow \|u-u_S\| \leq \varepsilon^2 \|u''\| = \varepsilon^2 \|f\|.$$

Proof How to connect $\|u-u_S\|$ with (WF)?

a tool (the duality argument)

the adjoint problem $\begin{cases} -w'' = u-u_S \text{ in } (0,1) \\ w(0) = 0 \text{ and } w'(1) = 0 \end{cases} \Longrightarrow$ Find $w \in V$ s.t. $a(w,v) = (u-u_S, v) \ \forall v \in V$

$$\|u-u_S\|^2 = a(w, u-u_S) = a(w-v, u-u_S) \quad \forall v \in S$$
$$\leq \|w-v\|_E \|u-u_S\|_E$$

$$\leq \varepsilon \|w''\| \, \varepsilon \|u''\| = \varepsilon^2 \|u-u_S\| \, \|f\| . \qquad \#$$

Possible choice of S

linear spaces: polynomials, piecewise polynomials (spline, finite element, etc), trigonometric functions, rational functions, wavelet, ...

set: free knot spline, neural network, ...

HW Let $J(v) = \frac{1}{2} a(v,v) - f(v)$ for $v \in P$, where P is a subset of V. Assume that $a(\cdot,\cdot)$ is symmetric bilinear and that $f(\cdot)$ is linear.

Assume that $J(u) = \min\limits_{v \in V} J(v)$

Prove that $$a(u-v, u-v) = 2\big(J(v) - J(u)\big).$$

§0.4 Piecewise Polynomials — Finite Element Method

- a partion of the domain $\Delta: \ 0 = x_0 < x_1 < \cdots < x_n = 1$ $\quad$ $0=x_0 \quad x_1 \quad x_2 \quad \cdots \quad x_n = 1$

  set subinterval $I_i = (x_{i-1}, x_i)$ and the length $h_i = x_i - x_{i-1}$

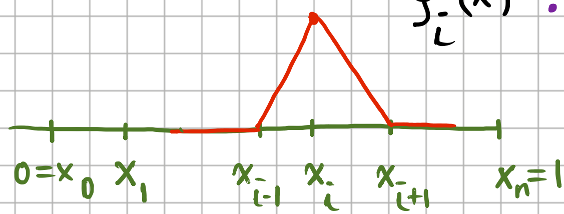
- linear finite element space (continuous)

$$S = S_1^0(\Delta) = \left\{ v \in C^0[0,1] \,\middle|\, v|_{I_i} \text{ is a linear poly. and } v(0) = 0 \right\}$$

$$\dim S = \underbrace{n}_{\#\text{ of intervals}} \cdot \overbrace{2}^{\text{DoF at each interval}} - \underbrace{(n-1)}_{\#\text{ of interior nodes}} - \overbrace{1}^{\text{B.C.}} = n$$

nodal basis functions

$\varphi_i(x)$: (1) $\varphi_i(x) \in S$ and (2) $\varphi_i(x_j) = \delta_{ij} = \begin{cases} 1, & j=i \\ 0, & j \neq i \end{cases}$



$$\Longrightarrow \varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x_{i-1} \leq x \leq x_i \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x_i \leq x \leq x_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

properties

(1) $\{\varphi_i(x)\}_{i=1}^{n}$ is linearly indep.

$$\sum_{i=1}^{n} c_i \varphi_i(x) \equiv 0 \text{ in } [0,1] \Longrightarrow 0 = \sum_{i=1} c_i \varphi_i(x_j) = c_j$$

$$\Longrightarrow S = \text{span}\{\varphi_i(x)\}_{i=1}^{n}$$

(2) $$\begin{cases} -\varphi_i''(x) = 0 \text{ in } (x_{i-1}, x_i) \cup (x_i, x_{i+1}) \\ \varphi_i(x_{i-1}) = 0, \ \varphi_i(x_i) = 1, \ \varphi_i(x_{i+1}) = 0 \end{cases}$$

- interpolant Given $v \in C^0[0,1]$ and $v(0)=0$, the interpolant v_I of v is defined by (1) $v_I \in S$ and (2) $v_I(x_i) = v(x_i)$ for $i = 0, 1, \cdots, n$

Theorem For $u \in C^2[0,1] \cap V$, we have

(E1) $$\|u - u_I\|_E \leq C h \|u''\|$$

where $h = \max_{1 \leq i \leq n} h_i$ and C is a positive constant indep. of h.

(E2) $\qquad \|u-u_I\|_\infty \le c h^2 \|u''\|_{max}$ where $\|v\|_{max} = \max_{x\in[0,1]} |v(x)|$

Proof I of (E2) $\qquad$ Let $e(x) = u(x) - u_I(x)$.

$\forall x \in [x_{i-1}, x_i] = I_i$

$$e(x_{i-1}) = e(x_i) = 0 \xrightarrow{\exists A(x) = ?} e(x) = A(x)(x-x_{i-1})(x-x_i)$$

$A(x) = ?$ auxiliary function $\quad g(t) = e(t) - A(x)(t-x_{i-1})(t-x_i)$

$$0 = g(x_{i-1}) = g(x_i) = g(x) \Longrightarrow \exists \xi \in (0,1) \text{ s.t.}$$

$$0 = g''(\xi) = u''(\xi) - A(x)\cdot 2 \Longrightarrow A(x) = \frac{1}{2}u''(\xi)$$

$$e(x) = \frac{1}{2}u''(\xi)(x-x_{i-1})(x-x_i)$$

$$\max_{x\in I_i} |e(x)| = \frac{1}{2}\max_{x\in I_i} \left|u''(\xi)(x-x_{i-1})(x-x_i)\right| \le \frac{1}{2}\|u''\|_{\infty, I_i} \max_{x\in I_i} |(x-x_{i-1})(x-x_i)|$$

$$= \frac{1}{8} h_i^2 \|u''\|_{\infty, I_i}$$

$\Longrightarrow$ (E2)

Proof of (E1) $\qquad$ the scaling argument + Poincaré's inequality

$\forall x \in I_i = [x_{i-1}, x_i]$

$$\int_{x_{i-1}}^{x_i} e'(x)^2 dx \overset{?}{\le} c h_i^2 \int_{x_{i-1}}^{x_i} e''(x)^2 dx \qquad \boxed{e''(x) = u''(x)}$$

$$\| \qquad\qquad\qquad \|$$

$$\int_0^1 (h_i^{-1}\hat{e}')^2 h_i d\hat{x} \qquad c h_i^2 \int_0^1 (h_i^{-2}\hat{e}'')^2 h_i d\hat{x}$$

change of variable

$x = x_{i-1} + h_i \hat{x} : [0,1] \to I_i$

$dx = h_i d\hat{x}$

$e(x) = e(x_{i-1} + h_i\hat{x}) = \hat{e}(\hat{x})$

$e'(x) = \frac{d\hat{e}}{d\hat{x}} \cdot \frac{d\hat{x}}{dx} = h_i^{-1}\hat{e}'$

$$\Longrightarrow \boxed{\int_0^1 \hat{e}'(\hat{x})^2 d\hat{x} \overset{?}{\le} c\int_0^1 \hat{e}''(\hat{x})^2 d\hat{x} \qquad \text{(P-I)}}$$

Poincaré's inequality $\quad \hat{e}(0)=\hat{e}(1)=0 \Rightarrow \exists\, \xi \in (0,1)$ s.t. $e'(\xi)=0$

$$\Longrightarrow \quad \hat{e}'(\hat{x}) = \int_{\xi}^{\hat{x}} \hat{e}''(\ )d$$

Cauchy-Schwarz inequality

$$|\hat{e}'(\hat{x})| = \left|\int_{\xi}^{\hat{x}} \hat{e}''(t)dt\right| \leq \left|\int_{\xi}^{\hat{x}} 1^2 dt\right|^{\frac{1}{2}} \left|\int_{\xi}^{\hat{x}} \hat{e}''(t)^2 dt\right|^{\frac{1}{2}}$$

$$\leq |\hat{x}-\xi|^{\frac{1}{2}} \left(\int_0^1 \hat{e}''(t)^2 dt\right)^{\frac{1}{2}}$$

$$\int_0^1 \hat{e}'(\hat{x})^2 d\hat{x} \leq \int_0^1 |\hat{x}-\xi|^{\frac{1}{2}\cdot 2} d\hat{x} \int_0^1 \hat{e}''(t)^2 dt \leq \frac{1}{2}\int_0^1 \hat{e}''(\hat{x})^2 d\hat{x}$$

#

HW #6, 9, 10

Corollary $\quad \|u-u_S\|_E \leq c\,h\,\|u''\|$

$$\|u-u_S\| \leq c\,h^2\,\|u''\|$$

Remark Interpolant operator $\mathcal{I}: C^0[0,1] \to S$ is defined by

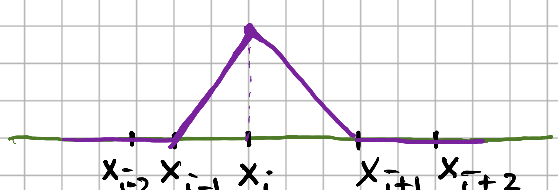
$$\mathcal{I}v = v_I.$$

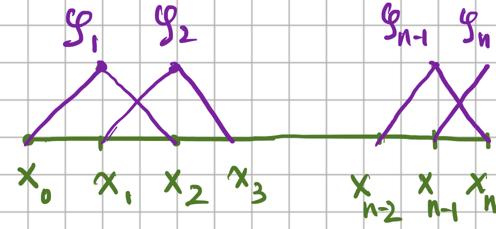
$\mathcal{I}$ is a projection $\Longleftrightarrow \mathcal{I}^2 = \mathcal{I}$

- **Discretization** for $S = \text{span}\{\varphi_i(x)\}_{i=1}^{n}$

For each i

$$a(\varphi_i,\varphi_j) = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx = \begin{cases} \int_{x_{i-1}}^{x_i} \varphi_i'\varphi_{i-1}'dx = \frac{1}{h_i}\cdot\left(-\frac{1}{h_i}\right)h_i = -h_i^{-1}, & j=i-1 \\ \int_{x_i}^{x_{i+1}} \varphi_i'\varphi_{i+1}'dx = \left(-\frac{1}{h_{i+1}}\right)\frac{1}{h_{i+1}}\cdot h_{i+1} = -h_{i+1}^{-1} & j=i+1 \\ \int_{x_{i-1}}^{x_{i+1}} \varphi_i'\varphi_i'dx = \frac{1}{h_i}+\frac{1}{h_{i+1}} & j=i \\ 0 & \text{otherwise} \end{cases}$$





$$f(\varphi_i) = \int_0^1 f \varphi_i \, dx = \int_{x_{i-1}}^{x_{i+1}} f(x)\varphi_i(x)\, dx = F_i$$

$$KU = \begin{bmatrix} h_1^{-1}+h_2^{-1} & -h_2^{-1} & & & & 0 \\ -h_2^{-1} & h_2^{-1}+h_3^{-1} & -h_3^{-1} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ 0 & & & -h_{n-1}^{-1} & h_{n-1}^{-1}+h_n^{-1} & h_n^{-1} \\ & & & & -h_{n-1}^{-1} & h_n^{-1} \end{bmatrix} = F$$

HW Let $S_0(\Delta) = \left\{ v \in L^2(I) \,\middle|\, v|_{I_i} \in P_0(I_i) \right\}$

Prove that for any $u(x) \in C^1[0,1]$,

$$\min_{v \in S_0(\Delta)} \|u - v\| \le C h \|u'\| \quad \text{for both } L^2(I) \text{ and } L^\infty(I) \text{ norm}$$

§0.5 Fixed, Adaptive, and Free Meshes

For any $u(x)$ defined on $I = [0,1]$ and a given tolerance $\varepsilon > 0$,

find a partition $\Delta: 0 = x_0 < x_1 < \cdots < x_n = 1$

with the least mesh points (the smallest n) s.t.

$$\min_{v \in S(\Delta)} \|u - v\| \le \varepsilon \|u\|.$$

Example $f(x) = x^r$ on $[0,1]$ for $r \in (0,1)$

(1) uniform partition $\Rightarrow \inf\limits_{v \in S_0^{-1}(\Delta_n)} \|f - v\|_\infty \le c\left(\frac{1}{n}\right)^r$; (2) find Δ_n s.t. $\inf\limits_{v \in S_0^{-1}(\Delta_n)} \|f - v\|_\infty \le c\left(\frac{1}{n}\right)$.

- fixed mesh vs adaptive mesh (a simple case)

Assume that $0 < \int_0^1 |u'|\,dx = \|u'\|_{L^1(I)} < +\infty$

Let $S_0^{-1}(\Delta_n) = \left\{ v \in L^2(I) \,\middle|\, v|_{I_i} \in P_0(I_i) \right\}$, where Δ_n is a partition of the interval $I = [0,1]$: $0 = x_0 < x_1 < \cdots < x_n = 1$.

Theorem

(1) For any fixed Δ_n, there exists a $u(x)$ such that

$$Q(u, \Delta_n) = \inf_{v \in S_0^{-1}(\Delta_n)} \|u - v\|_\infty \Big/ \|u'\|_{L^1(I)} \ge \frac{1}{2}$$

(2) For any function $u(x)$, there exists a partition Δ_n s.t.

$$Q(u, \Delta_n) \le \frac{1}{n}.$$

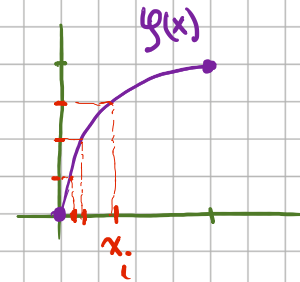
Proof (1) For any fixed Δ_n, let $u(x)$ be a ramp function

$$\Rightarrow \|u'\|_{L^1(I)} = \int_{x_0}^{x_1} |u'|\,dx = \int_{x_0}^{x_1} \frac{1}{x_1 - x_0}\,dx = 1$$

The best $L^\infty(I)$ approximation is the step function $V(x)$

$$\Rightarrow \inf_{v \in S(\Delta_n)} \|u - v\|_\infty = \max_{x \in [x_0, x_1]} |u(x) - v(x)| = \frac{1}{2}$$

(2) Assume that $\|u'\|_{L^1(I)}=1$. Let $\varphi(x)=\int_0^x |u'|\,dx$.



$\Rightarrow \varphi(0)=0,\ \varphi(1)=1$

$\varphi(x)$ is continuous and non-decreasing function

- partition the range of $\varphi(x)$, $[0,1]=[\varphi(0),\varphi(1)]$, by a uniform partition $\varphi(0)=0<\frac{1}{n}<\frac{2}{n}<\cdots<\frac{n}{n}=1=\varphi(1)$

$\xrightarrow{\varphi(x)\nearrow}$ $\exists\ 0=x_0<x_1<\cdots<x_n=1$ such that

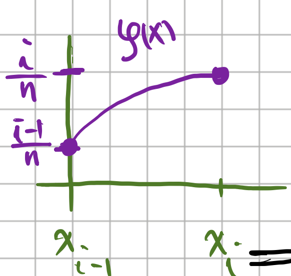
$\boxed{x_i=?}$

$$\varphi(x_i)=\frac{i}{n}$$

- approximation $\hat{u}(x)\in S_0^{-1}(\Delta_n)$ and $\hat{u}\big|_{I_i}=u(x_{i-1})$

On $I_i=[x_{i-1},x_i]$

$$|u(x)-\hat{u}(x)|=|u(x)-u(x_{i-1})|=\left|\int_{x_{i-1}}^{x_i}u'(x)\,dx\right|\leq\int_{x_{i-1}}^{x_i}|u'(x)|\,dx$$



$$=\varphi(x_i)-\varphi(x_{i-1})=\frac{1}{n}$$

$$\Rightarrow \|u-\hat{v}\|_{\infty,I}=\max_{1\leq i\leq n}\|u-\hat{v}\|_{\infty,I_i}\leq\frac{1}{n}$$

$$\Rightarrow \inf_{v\in S(\Delta)}\|u-v\|_{\infty,I}\leq\|u-\hat{v}\|_{\infty,I}\leq\frac{1}{n}$$

$$\Rightarrow \frac{\inf\|u-v\|_{\infty,I}}{\|u'\|_{L^1(I)}}\leq\frac{1}{n}.$$

#

How to construct mesh Δ_n s.t. $\inf_{v \in S(\Delta)} \|u-v\|_{L^2(I)} \le \frac{1}{n} \|u'\|_{L^1(I)}$?

(1) adaptive mesh refinement   Solve $\to$ Estimate $\to$ Mark $\to$ Refine

starting with Δ_{n_0}

(i) compute $u_k \in S(\Delta_{n_k})$ s.t. $\|u-u_k\|_{L^2(I)} = \min_{v \in S(\Delta_k)} \|u-v\|_{L^2(I)}$

(ii) compute indicators

$$\eta_K = \|u-u_k\|_{L^2(K)}$$ for each subinterval $K \in \mathcal{T}_k$,

$\mathcal{T}_k$ is the collection of subintervals of Δ_{n_k}

(iii) if $\eta = \sqrt{\sum_{K \in \mathcal{T}_k} \eta_K^2} \le \varepsilon$, then stop; otherwise, go to (iv)

(iv) determine the set of subintervals for refinement

$$\hat{\mathcal{T}}_k = \{ K \in \mathcal{T}_k : \ \eta_K \text{ is } \overset{?}{\underline{\text{large}}} \}$$

(v) refine $K \in \hat{\mathcal{T}}_k$ to generate a new mesh $\mathcal{T}_{k+1}$ or Δ_{n_k},

then go to step (i).

Remark when u is unknown like solution of PDE,

how to estimate η_K ?

(2) free-knot spline functions

$$S_n = \left\{ v(x) \in C^0[0,1] \;\middle|\; \begin{array}{l} v(x) \text{ is linear on each subinterval } I_i = [x_{i-1}, x_i] \\ \text{and } x_i \in [0,1] \text{ is parameter to be determined} \end{array} \right\}$$

$\forall\, v(x) \in S_n$

$$v(x) = \sum_{i=0}^{n} v_i \varphi_i(x) \quad \text{with } \varphi_i(x) = \begin{cases} h_i^{-1}(x - x_{i-1}), & \text{in } I_i \\ h_{i+1}^{-1}(x_{i+1} - x), & \text{in } I_{i+1} \\ 0, & \text{otherwise} \end{cases}$$

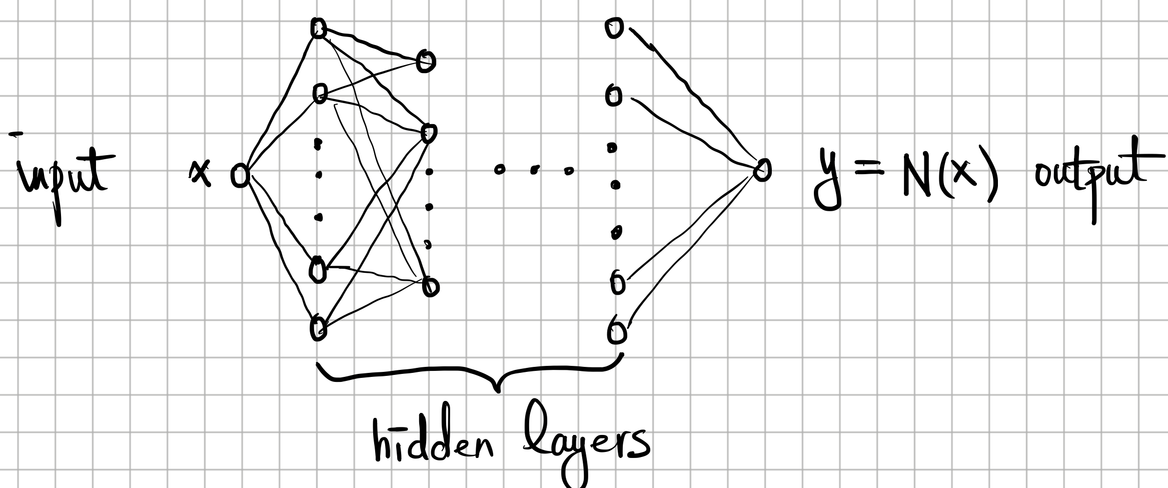
where $\{v_i\}_{i=0}^{n}$ and $\{x_i\}_{i=0}^{n}$ are parameters to be determined.

- the best least-squares approximation: given $u(x)$ on $[0,1]$, find u_n s.t.

$$\|u - u_n\|_{L^2(I)} = \inf_{v \in S_n} \|u - u_n\|_{L^2(I)} = \inf_{\vec{\theta} \in R^N} \|u - u_n(\vec{\theta})\|_{L^2(I)}$$

where $\vec{\theta} = \{\vec{v}, \vec{x}\} \in R^{2(n+1)}$.

§0.6 Neural Network



Homework (Computer Project : fixed, adaptive, and free meshes)

Let $u(x) = x^r$ be a function defined on $[0,1]$ with $r \in (0,2)$.

Use the continuous piecewise linear polynomials defined on the uniform, adaptive, and free meshes to solve

(1) the best least-squares approximation : find $u_n \in S_1^0(\Delta_n)$ s.t.

$$\|u-u_n\| = \min_{v \in S_1^0(\Delta_n)} \|u-v\|, \quad \text{where } \|v\| = \sqrt{\int_0^1 v^2\,dx}$$

(2) the Poisson equation : find $u_n \in S_1^0(\Delta_n)$ s.t.

$$\begin{cases} -u''(x) = f(x) & \text{in } I = (0,1) \\ u(0) = 0 \text{ and } u'(1) = r \end{cases}$$

where $f(x) = r(1-r)\,x^{r-2}$.

- For $r = 0.1, 0.4, 0.8, 1.2, 1.6$, examine approximation accuracies using these three approaches for a fixed number of DoF.
- Given a prescribed tolerance $\varepsilon > 0$, examine the number of DoF $n(\varepsilon)$ s.t.

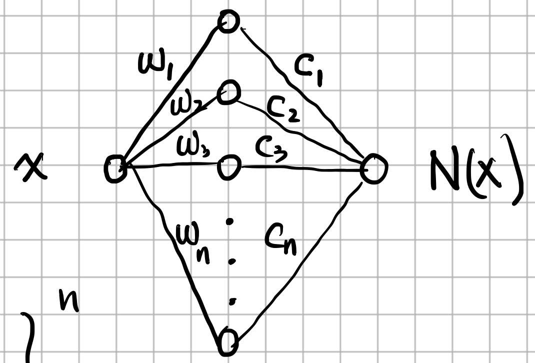
(1) $\|u-u_n\| \le \varepsilon \|u\|$

(2) $\|u-u_n\|_E \le \varepsilon \|u\|_E$.

$$\|v\|_E = \sqrt{\int_0^1 (v')^2\,dx}$$

- <u>two-layer neural network</u>

$$N(x) = \sum_{i=1}^{n} c_i \, \sigma(w_i x - b_i) + c_0$$



- input weights $\{w_i\}_{i=1}^{n}$ and bias $\{b_i\}_{i=1}^{n}$

  output weights $\{c_i\}_{i=1}^{n}$ and bias c_0

$\sigma(t)$ — activation function

<u>examples</u>

- ReLU (Rectified Linear Unit)

$$\sigma(t) = \max\{0, t\} = \begin{cases} 0, & t<0 \\ t, & t>0 \end{cases}$$

- spline $\quad \sigma_k(t) = \max\{0, t^k\}$

- sigmoid (logistic, Gaussian, or arctan)

$$\lim_{t\to-\infty} \sigma(t) = 0 \quad \text{and} \quad \lim_{t\to+\infty} \sigma(t) = 1$$

logistic $\quad \sigma(t) = \dfrac{1}{1+e^{-t}}$

Gaussian $\quad \sigma(t) = \dfrac{1}{(2\pi)^{1/2}} \displaystyle\int_{-\infty}^{t} e^{-\frac{y^2}{2}} \, dy$

arctan $\quad \sigma(t) = \dfrac{1}{\pi} \arctan(t) + \dfrac{1}{2}$

- ReLU two-layer NN

one dimension
$$\sigma(w_i x - b_i) = |w_i|\, \sigma\left(\frac{w_i}{|w_i|} x - \frac{b_i}{|w_i|}\right)$$

$$\mathcal{M}_n(\sigma) = \left\{ c_0 + \sum_{i=1}^{n} c_i\, \sigma(x - b_i) : c_i, b_i \in R \right\} = \text{free knot spline}$$

d-dimension
$$\mathcal{M}_n(\sigma) = \left\{ \vec{c}_0 + \sum_{i=1}^{n} \vec{c}_i\, \sigma(\vec{w}_i \cdot \vec{x} - b_i) : \vec{c}_i \in R^d,\ b_i \in R,\ \vec{w}_i \in \mathcal{S}^d \right\}$$

where $\mathcal{S}^d = \left\{ \vec{x} \in R^d : |\vec{x}| = 1 \right\}$.

Universal Approximation Theory (Cybenko (89) and Hornik-Stinchcombe-White (89))

$\mathcal{M}(\sigma) = \left\{ v(\vec{x}) \in \mathcal{M}_n(\sigma) : n \in \mathbb{Z}_+ \right\}$ is dense in $C(K)$ where K is a compact set in R^d provided that σ is not a polynomial.

- the best least-squares approximation

Given $f(x) \in H^1(\Omega)$, find $f_n \in \mathcal{M}_n(\sigma)$ such that

$$\|f(\cdot) - f_n(\cdot\,; \vec{\theta}^*)\| = \min_{v \in \mathcal{M}_n(\sigma)} \|f - v\| = \min_{\vec{\theta} \in R^N} \underbrace{\|f(\cdot) - v(\cdot\,; \vec{\theta})\|}_{J(\vec{\theta})}$$

where $\vec{\theta} = (\vec{c}, \vec{b})$.

$$= \min_{\vec{\theta} \in R^N} J(\vec{\theta})$$

- Order of approximation

Petrushev (1998) For any $f \in H^s(\Omega)$ with $s = 2 + \frac{d-1}{2}$,

$$\inf_{\varphi \in M_n(\sigma)} \|f-\varphi\|_{L^2(\Omega)} \leq c \left(\frac{1}{n}\right)^{\frac{1}{2}+\frac{3}{2d}} \|f\|_{H^s(\Omega)}$$

Siegel-Xu (2021) For any $v \in \mathcal{B}^s(\Omega) = \{\varphi: \Omega \to \mathbb{C} \mid \|\varphi\|_{\mathcal{B}^s(\Omega)} < +\infty\}$, where

$$\|\varphi\|_{\mathcal{B}^s(\Omega)} = \inf_{\varphi_e|_\Omega = \varphi} \int_{R^d} (1+|\vec{\xi}|^s) \, |\hat{\varphi}_e(\vec{\xi})| \, d\vec{\xi}$$

$$\Rightarrow \quad \inf_{\varphi \in M_n(\sigma)} \|f-\varphi\|_{H^m(\Omega)} \leq c(\log n) \left(\frac{1}{n}\right)^{2-m} \|f\|_{\mathcal{B}^s(\Omega)}, \quad m=0,1$$

- Effect of Numerical Integration

$$\int_\Omega f(\vec{x})\,d\vec{x} \approx Q_g(f). \qquad \text{set } \|f\|_g = \sqrt{Q_g(f^2)}$$

Find $f_g \in M_n(\sigma)$ s.t. $\|f-f_g\|_g = \min_{\varphi \in M_n(\sigma)} \|f-\varphi\|_g$

$$\|f-f_g\| \leq c \left\{ \inf_{\varphi \in M_n(\sigma)} \left[\|f-\varphi\| + \sup_{w \in M_n(\sigma)} \frac{|(I-Q_g)(\varphi w)|}{\|w\|} \right] + \sup_{w \in M_n(\sigma)} \frac{|(I-Q_g)(fw)|}{\|w\|} \right\}$$

- training $\qquad \min_{\vec{\theta} \in R^N} g(\vec{\theta})$

Gradient Descent Given initial $\vec{\theta}_1$, $\qquad \vec{\theta}_{k+1} = \vec{\theta}_k - \tau \nabla f(\vec{\theta}_k)$

Adam

$$\vec{\theta}_{k+1} = \vec{\theta}_k - \tau \frac{\nabla \hat{f}(\vec{\theta}_k)}{\left|\nabla \tilde{f}(\vec{\theta}_k)\right|},$$

where
$$\nabla \hat{f}(\vec{\theta}_k) = \frac{1-\beta_1}{1-\beta_1^k}\left[\beta_1^{k-1}\nabla f(\vec{\theta}_1) + \beta_1^{k-2}\nabla f(\vec{\theta}_2) + \cdots + \beta_1 \nabla f(\vec{\theta}_{k-1}) + \nabla f(\vec{\theta}_k)\right]$$
$$\nabla \tilde{f}(\vec{\theta}_k) = \frac{1-\beta_2}{1-\beta_2^k}\left[\beta_2^{k-1}\nabla f(\vec{\theta}_1) + \cdots + \nabla f(\vec{\theta}_k)\right].$$

Conclusions