MAT 683 Final Homework 6
A. Lutoborski Syracuse University
Fall 2018

Total 100 pts.

Ex 1. Let $z_1, \dots, z_N$ be unimodular complex numbers i.e. $|z_n| =$

$$P(z) = \prod_{n=1}^{N}(z - z_n) = \sum_{n=0}^{N} a_n z^n$$

Let

$$T(\theta) = P(e(\theta)) = \sum_{n=0}^{N} a_n e(n\theta)$$

where $e(x) = \exp(i2\pi x)$.

(a) Show that $a_N = 1$ and that $|a_0| = 1$.

(b) Show that

$$\frac{1}{N}\sum_{k=1}^{N} T(\frac{k}{N} + \alpha) = a_0 + a_N e(N\alpha)$$

(c) Show that there is an α such that $|a_0 + a_N e(N\alpha)| = 2$

(d) Show that if $z_n = e(n/N)$ are the N-th roots of unity then $P(z) = z^N - 1$ and $|T(\theta$ for all θ.

Ex 2. Given an arbitrary function $f : \mathbb{R} \to \mathbb{C}$ and a number $a > 0$, the a-periodizatio is a function f_a defined as

$$f_a(x) = \sum_{n=-\infty}^{\infty} f(x + na),$$

provided that the series converges.

Find the 1-periodization $f_1 : \mathbb{R} \to \mathbb{R}$ of the function

$$f(x) = e^{-|x|}, \quad x \in \mathbb{R}$$

Hint: Find the formula for $g(x) = \sum_{k=-\infty}^{\infty} f(x + k)$ for $0 \le x < 1$. For $0 \le x < 1$ we $f_1(x) = g(x)$ and for values x outside the unit interval f_1 satisfies $f_1(x) = g(x - \lfloor x \rfloor)$.

Ex 3. Let X be an inner product space over $\mathbb{C}$. Show that the following statement equivalent for arbitrary vectors x, y.

(1) $\langle x, y \rangle = 0$

(2) $\|x\| \le \|x + ty\|$ for all $t \in \mathbb{C}$.

(3) $\|x + ty\| = \|x - ty\|$ for all $t \in \mathbb{C}$.

4. Let $\lambda > 0$ be given. We want to solve the equation

$$<\lambda<\frac{1}{e} \qquad x = \lambda e^x$$

(a) Show graphically that if $\lambda < \frac{1}{e}$ then the equation has two positive solutions $0 < \tilde{x}$ If $\lambda > \frac{1}{e}$ there is no solution, if $\lambda = \frac{1}{e}$ there is a single solution.

(b) For $\lambda < \frac{1}{e}$ we consider two iterations:

$$(I1) \qquad x_{n+1} = \lambda e^{x_n}, \qquad n = 0, 1, \dots$$

$$(I2) \qquad x_{n+1} = \ln x_n - \ln \lambda, \qquad n = 0, 1, \dots$$

$(I1)$ converges to $\tilde{x}$ and $(I2)$ converges to $\hat{x}$ when x_0 is near the respective solu-
ze both methods as simple iterations for a fixed point problem for a contractive

ıte (using a calculator) the number $9^{1/3}$ to six decimals, using Newton's method,
$= 2$.
elevant quantities h_0, x_1, M of Kantorovitch's theorem in this case.
orovitch's theorem prove that Newton's method converges.
: $\mathbb{R}^2 \to \mathbb{R}^2$ given by

$$f(\begin{bmatrix} x \\ y \end{bmatrix}) = \begin{bmatrix} x^2 - y - 12 \\ y^2 - x - 11 \end{bmatrix}$$

ne Lipschitz ratio M for the derivative Df that is

$$|Df(p) - Df(q)| \leq M|p - q| \qquad \text{for} \quad p, q \in \mathbb{R}^2$$

Starting at $x_0 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ compute x_1 as one step of Newton's method to solve $f(\begin{bmatrix} x \\ y \end{bmatrix}) = 0$.
) Find a disc which contains a root of the equation.

Ex1: Let $z_1, z_2, \dots, z_N$ be unimodular complex numbers, i.e., $|z_n| = 1$

Let $$p(z) = \prod_{n=1}^{N}(z - z_n) = \sum_{n=0}^{N} a_n z^n$$

Let $T(\theta) = P(e(\theta)) = \sum_{n=0}^{N} a_n e(n\theta)$ where $e(x) = \exp\{i2\pi x) = e^{i2\pi x}$.

a) Show that $a_N = 1$ and that $|a_0| = 1$.

b) Show that $\frac{1}{N}\sum_{\ell=1}^{N} T\left(\frac{\ell}{N} + \alpha\right) = a_0 + a_N e(N\alpha)$

c) Show that there is an α such that $|a_0 + a_N e(N\alpha)| = 2$.

d) Show that if $z_n = e\left(\frac{n}{N}\right)$ are the N^{th} roots of unity then $\begin{cases} P(z) = z^N - 1 \\ T(\theta) \leq 2 \text{ for all } \theta \end{cases}$

a)
$$p(z) = \prod_{n=1}^{N}\left(z - z_n\right) = (z - z_1)(z - z_2)\cdots(z - z_N)$$

⇒ The coefficient that goes with z^N has to be equal to 1 ⇒ $a_N = 1$.

⇒ The term of the polynomial that does not contain z is $(-z_1)(-z_2)\dots(-z_N)$
$$= \underbrace{(-1)^N}_{a_0} z_1 z_2 \dots z_N$$

⇒ $|a_0| = |(-1)^N| = 1$.

b) We first prove that $\frac{1}{q}\sum_{\ell=1}^{q} T\left(\frac{\ell}{q} + \alpha\right) = \sum_{\substack{\ell=-N,N \\ q|n}} a_\ell\, e(\ell\alpha)$

$$\text{LHS} = \frac{1}{q}\sum_{\ell=1}^{q} T\left(\frac{\ell}{q} + \alpha\right) = \frac{1}{q}\sum_{\ell=1}^{q}\sum_{n=0}^{N} a_n\left(e\left(n\left(\frac{\ell}{q} + \alpha\right)\right)\right) =$$

$$= \frac{1}{q}\sum_{n=0}^{N} a_n \underbrace{\sum_{\ell=1}^{q} e\left(\frac{n\ell}{q}\right)}_{=\begin{cases} q & q|n \\ 0 & \text{otherwise} \end{cases}} e(n\alpha) = \frac{1}{q}\, q \sum_{\substack{n=0 \\ q|n}}^{N} a_n e(n\alpha) = \sum_{\substack{n=0 \\ q|n}}^{N} a_n e(n\alpha)$$

* When $q = N$, then

$$\frac{1}{N}\sum_{\ell=1}^{N} T\left(\frac{\ell}{N} + \alpha\right) = \sum_{\substack{n=0 \\ N|n}}^{N} a_n e(n\alpha) = a_0(e(0) + a_N(e(N\alpha))) = a_0 + a_N' e(N\alpha)$$

c> Show that there is an α such that $|a_0 + a_N e(N\alpha)| = 2$.

Since $|a_0| = 1$, $a_N = 1$, this $\Rightarrow |a_0 + e(N\alpha)| = 2$.

- When $a_0 = 1 \Rightarrow |1 + e(N\alpha)| = 2 \Rightarrow e(N\alpha) = 1$.

$$\Rightarrow \cos(2\pi N\alpha) = 1.$$

$$\rightarrow \quad 2\pi N\alpha = k\pi$$

$$\alpha = \frac{k\pi}{2\pi N} = \frac{k}{2N} \quad k \in \mathbb{Z} \text{ } k \text{ even.}$$

- When $a_0 = -1 \Rightarrow |1 + e(N\alpha)| = 2 \Rightarrow e(N\alpha) = -1$.

$$\Rightarrow 2\pi N\alpha = k\pi \quad k \in \mathbb{Z}^+, k \text{ odd}$$

$$\alpha = \frac{k}{2N} \quad k \in \mathbb{Z}^+, k \text{ odd.}$$

d> Show that if $z_n = e(\frac{n}{N})$ are the N^{th} roots of unity then $\begin{cases} P(z) = z^N - 1 \\ |T(\theta)| \le 2. \end{cases}$

Each z_n is the N^{th} root of unity which means they are solutions of $z^N = 1 \Leftrightarrow z^N - 1$ and that each root z_n when $n = \overline{1,N}$ contributes to a linear factor $(z - z_n)$ of $z^N - 1$:

$$z^N - 1 = (z - z_1)(z - z_2)\cdots(z - z_N) = \prod_{n=1}^{N}(z - z_n) = \underline{p(z)}$$

- So we have

$$T(\theta) = P(e(\theta)) \overset{p(z) = z^N - 1}{=\!=\!=} [e(\theta)]^N - 1 = (e^{i2\pi\theta})^N - 1$$

$$|T(\theta)| \le |(e^{i2\pi\theta})^N| + |1| \le 2.$$

E2→ Given an arbitary function $f:\mathbb{R}\longrightarrow\mathbb{C}$

$a>0$

A a_periodication of f is a function f_a defined as $f_a(x) = \sum_{n=-\infty}^{\infty} f(x+na)$ provided that the series converges.

Find a 1 periodization $f_1:\mathbb{R}\longrightarrow\mathbb{R}$ of the function $f(x)=e^{-|x|}$, $x\in\mathbb{R}$.

<u>Hint</u>: Find the formula for $g(x)=\sum_{k=-\infty}^{+\infty} f(x+k)$ for $0\le x<1$.

For $0\le x<1$ we have $f_1(x)=g(x)$ for

for values x outside the unit interval f_1 satisfies $f_1(x)=g(x-\lfloor x\rfloor)$

* Consider $[0,1)\ni x$,

$$\hat{f}(n)=\langle f,e_n\rangle=\int_0^1 f(x)\overline{e_n(x)}\,dx=\int_0^1 f(x)e^{-i2\pi x}\,dx=\int_0^1 e^{-|x|}e^{-i2\pi x}\,dx=$$

$$=\int_0^1 e^{-(1+i2\pi)x}\,dx=\left(\frac{-1}{1+i2\pi}\right)e^{-(1+i2\pi)x}\Big|_0^1=\frac{1-e^{-(1+i2\pi)}}{1+i2\pi}$$

Then we have $g(x)=\sum_{n=-\infty}^{+\infty}\hat{f}(n)e_n(x)=\sum_{n=-\infty}^{+\infty}\frac{1-e^{-(1+i2\pi)}}{1+i2\pi}e^{i2\pi x}$ is a 1- periodic of the function $f(x)=e^{-|x|}$ when $x\in[0,1)$.

* Then for x outside $[0,1)$, take $f_1(x)=g\left(\left(x-\lfloor x\rfloor\right)\right)$

*EX3: Let X be an inner product space over $\mathbb{C}$
Show that the following statements are equivalent for arbitrary vectors x & y
1> $\langle x,y\rangle = 0$
2> $\|x\| \le \|x+ty\|$ for all $t \in \mathbb{C}$
3> $\|x+ty\| = \|x-ty\|$ for all $t \in \mathbb{C}$

We will prove $(1)\Leftrightarrow(2)\Leftrightarrow(3)$ by prooving that $\begin{cases}(1)\Longrightarrow(2)\\(2)\Longrightarrow(3)\\(3)\Longrightarrow(1)\end{cases}$

* Prove $(1)\Longrightarrow(2)$:

We have $\langle x,y\rangle = 0$. Need to prove $\|x\|\le\|x+ty\|$ for all $t\in\mathbb{C}$.

We have
$$\|x+ty\|^2 = \langle x+ty, x+ty\rangle = \langle x+ty, x\rangle + \langle x+ty, ty\rangle$$
$$= \langle x,x\rangle + \langle ty, x\rangle + \langle x, ty\rangle + \langle ty, ty\rangle$$
$$= \|x\|^2 + t\underbrace{\overline{\langle x,y\rangle}}_{=0 \text{ since}(1)} + \bar{t}\underbrace{\langle x,y\rangle}_{=0 \text{ since}(1)} + |t|^2\|y\|^2$$
$$= \|x\|^2 + |t|^2\|y\|^2$$
$$\geqslant \|x\|^2$$

So we have $\left.\begin{array}{l}\|x+ty\|^2 \geqslant \|x\|^2\\ \text{and since } \|\cdot\|\geqslant 0\end{array}\right\} \Rightarrow \|x+ty\| \geqslant \|x\|, \forall t$ which is (2).

* Prove that $(2)\Longrightarrow(3)$

We have $\|x\|\le\|x+ty\|, \forall t\in\mathbb{C}$, we need to prove $\|x+ty\|=\|x-ty\|, \forall t\in\mathbb{C}$

- We have

$$\|x+ty\|^2 \overset{\text{above}}{=\!=} \|x\|^2 + t\overline{\langle x,y\rangle} + \bar{t}\langle x,y\rangle + \|t^2\|\|y\|^2$$
$$\|x-ty\|^2 =\!=\!= \|x\|^2 - \bar{t}\langle x,y\rangle - t\overline{\langle x,y\rangle} + |t|^2\|y\|^2$$

Then if we can ~~we~~ prove that $\langle x,y\rangle = 0$ then it means it is enough to see clearly that $\underset{\forall t\in\mathbb{C}}{\|x+ty\|^2 = \|x-ty\|^2} \Longrightarrow \|x+ty\| = \|x-ty\|, \forall t\in\mathbb{C}$

- So now we want to prove that from (2), $\|x\|\le\|x+ty\|, \forall t\in\mathbb{C}$, we have $\langle x,y\rangle = 0$ by prooving that $\mathrm{Re}\langle x,y\rangle = 0$ and that $\mathrm{Im}\langle x,y\rangle = 0$.

* Let $\|x\| \leq \|x+ty\|$ $\forall t \in \mathbb{C}$. Need to prove $\mathrm{Re}(\langle x,y\rangle) = 0$ $\forall x,y$

- $\|x\| \leq \|x+ty\| \Rightarrow$

$$\Rightarrow \|x+ty\|^2 = \underbrace{\langle x,x\rangle + t^2\langle y,y\rangle + \bar{t}\langle x,y\rangle + t\overline{\langle x,y\rangle} \geq \langle x,x\rangle}_{\text{same}} = \|x\|^2$$

$$\Rightarrow t^2\langle y,y\rangle + \bar{t}\langle x,y\rangle + t\overline{\langle x,y\rangle} \geq 0 \qquad (*)$$

choose t be real $\Rightarrow t^2\langle y,y\rangle + t\langle x,y\rangle + t\overline{\langle x,y\rangle} \geq 0$

$$\Rightarrow t^2\langle y,y\rangle + t\left(\langle x,y\rangle + \overline{\langle x,y\rangle}\right) \geq 0$$

$$\Rightarrow t^2\langle y,y\rangle + 2t\ \mathrm{Re}(\langle x,y\rangle) \geq 0.$$

choose $y \neq 0$ $\Rightarrow t^2 + \underbrace{2\frac{\mathrm{Re}(\langle x,y\rangle)}{\langle y,y\rangle}}_{a} t \geq 0$ $\forall t$

The quadratic equation $t^2 + at \geq 0$ $\forall t$ when $a = 0 \Rightarrow 2\frac{\mathrm{Re}(\langle x,y\rangle)}{\langle y,y\rangle} = 0$

$$\Rightarrow \mathrm{Re}(\langle x,y\rangle) = 0.$$

- Let $\|x\| \leq \|x+ty\|$, $\forall t \in \mathbb{C}$ $\forall x,y$. Need to prove that $\mathrm{Im}(\langle x,y\rangle) = 0$

Similar to above, choose t to be $t = is$, $s > 0, s \in \mathbb{R}$.

Then $(*) \Rightarrow t^2\langle y,y\rangle - t\langle x,y\rangle + t\overline{\langle x,y\rangle} \geq 0$

$$t^2\langle y,y\rangle + t\left(\overline{\langle x,y\rangle} - \langle x,y\rangle\right) \geq 0.$$

$$t^2\langle y,y\rangle + 2t\ \mathrm{Im}\langle x,y\rangle \geq 0.$$

Similarly as above $\Rightarrow \mathrm{Im}\langle x,y\rangle = 0.$

So we have $\mathrm{Re}\langle x,y\rangle = \mathrm{Im}\langle x,y\rangle = 0 \Rightarrow \langle x,y\rangle = 0.$

$\Rightarrow \|x+ty\| = \|x-ty\|$ which is (3).

* Now we want to prove $(3) \Longrightarrow (1)$.

We have $\|x+ty\| = \|x-ty\|$ for all $t \in \mathbb{C}$. We need to prove $\langle x,y\rangle = 0$

- $\|x+ty\| = \|x-ty\| \Leftrightarrow \|x+ty\|^2 = \|x-ty\|^2$ since $\|\cdot\| \geq 0$

$$\Leftrightarrow \|x\|^2 + t\overline{\langle x,y\rangle} + \bar{t}\langle x,y\rangle + |t^2| \, \|y\|^2 = \|x\|^2 - t\overline{\langle x,y\rangle} - \bar{t}\langle x,y\rangle + |t|^2\|y\|^2$$

$$\Rightarrow \quad t\overline{\langle x,y\rangle} + \bar{t}\langle x,y\rangle = 0. \quad (*)$$

- Choose t to be real, then this $(*)$ implies

$$\overline{\langle x,y\rangle} + \langle x,y\rangle = 0$$
$$\Rightarrow 2\operatorname{Re}\langle x,y\rangle = 0$$
$$\Rightarrow \operatorname{Re}\langle x,y\rangle = 0$$

- In $(*)$, choose $t = is$, where $s \in \mathbb{R}$, $s > 0$

then $(*) \Rightarrow is\overline{\langle x,y\rangle} - is\langle x,y\rangle = 0 \quad \forall s > 0$

$$is\left(\overline{\langle x,y\rangle} - \langle x,y\rangle\right) = 0 \quad \forall s > 0$$
$$-is\, 2\operatorname{Im}(\langle x,y\rangle) = 0 \quad \boxed{\forall s > 0}$$
$$\Rightarrow \operatorname{Im}(\langle x,y\rangle) = 0.$$

So we have $\|x+ty\| = \|x-ty\|, \forall t \in \mathbb{C} \Rightarrow \left\{\begin{array}{l}\operatorname{Re}(\langle x,y\rangle) = 0 \\ \operatorname{Im}(\langle x,y\rangle) = 0\end{array}\right\} \Rightarrow \langle x,y\rangle = 0$ which is $\boxed{(1)}$

* In conclusion, we have proved $\left.\begin{array}{l}(1) \Rightarrow (2) \\ (2) \Rightarrow (3) \\ (3) \Rightarrow (1)\end{array}\right\}$ and that the there statements are equivalent.

Problem4: $\lambda > 0$ - given. Want to solve $x = \lambda e^{x}$

a> Show graphically that

$0 < \lambda < \frac{1}{e} \Rightarrow$ 2 positive sols $0 < \tilde{x} < \hat{x}$

$\lambda > \frac{1}{e} \Rightarrow$ no sol

$\lambda = \frac{1}{2} \Rightarrow$ one solution .

```
oneovere=1/exp(1)
%oneovere=0.3678794412

a=-0.1 %a is a constant so that we can add to 1/e so that we can
have
%lambda <,>,= 1/e

lambda=(1/exp(1))
lambda_small=(1/exp(1))-0.1
lambda_big=(1/exp(1))+0.1

hold on
fp = fplot(@(x) lambda*exp(x),[0, 3])
fp = fplot(@(x) lambda_small*exp(x),[0, 3])
fp = fplot(@(x) lambda_big*exp(x),[0, 3])

line=fplot(@(x) x, [0,3])
legend({'lambda=1/e','lambda<1/e',
'lambda>1/e'},'Location','northwest')

title(['the intersection of y=x and y=lambda*(e^x)' ])

hold off
```

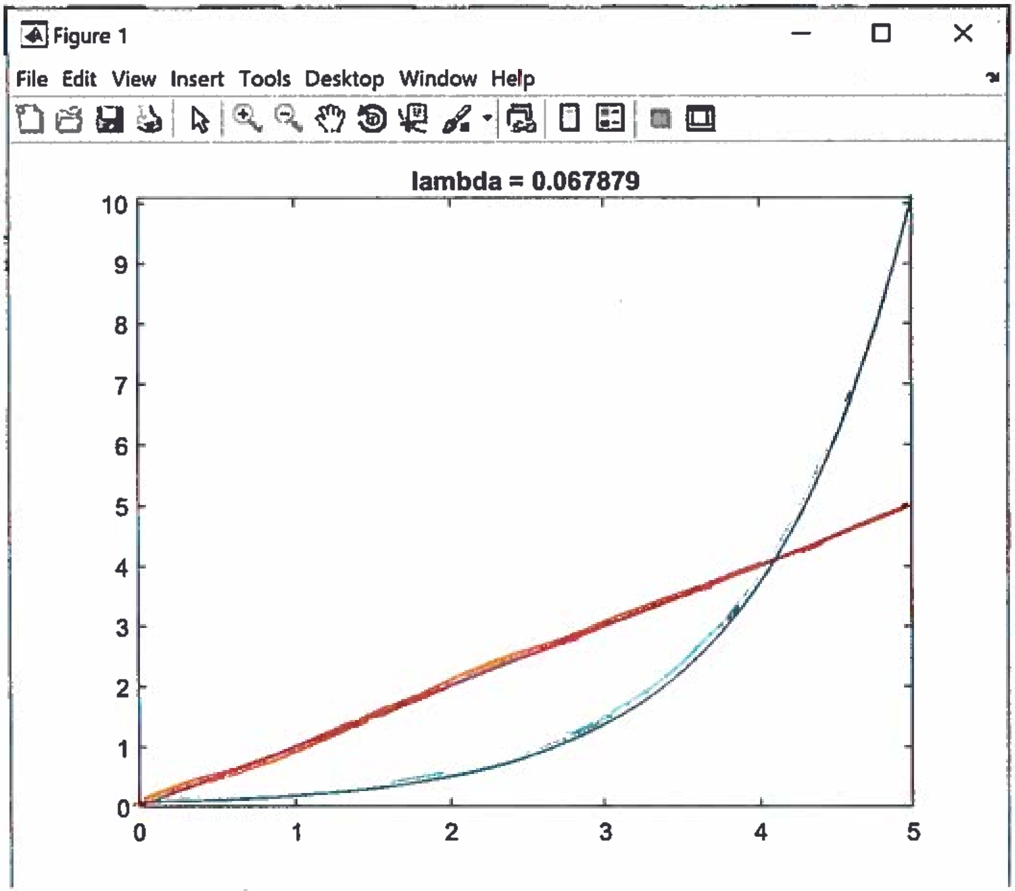


Figure 1
File Edit View Insert Tools Desktop Window Help
lambda = 0.067879
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5

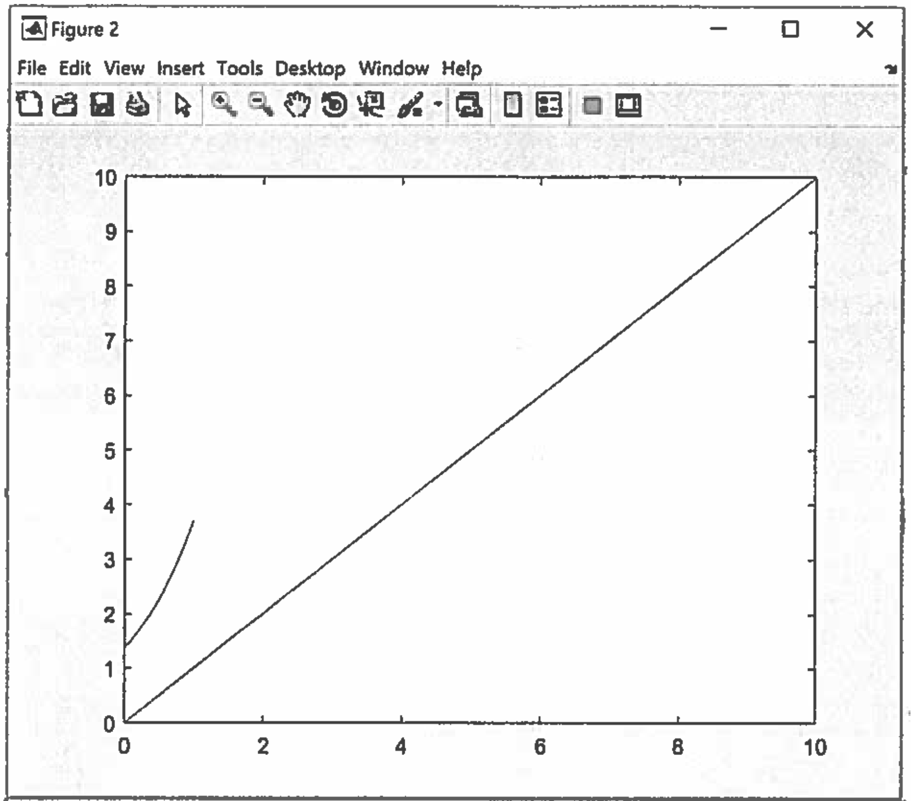


Figure 2
File Edit View Insert Tools Desktop Window Help
0
1
2
3
4
5
6
7
8
9
10
0
2
4
6
8
10

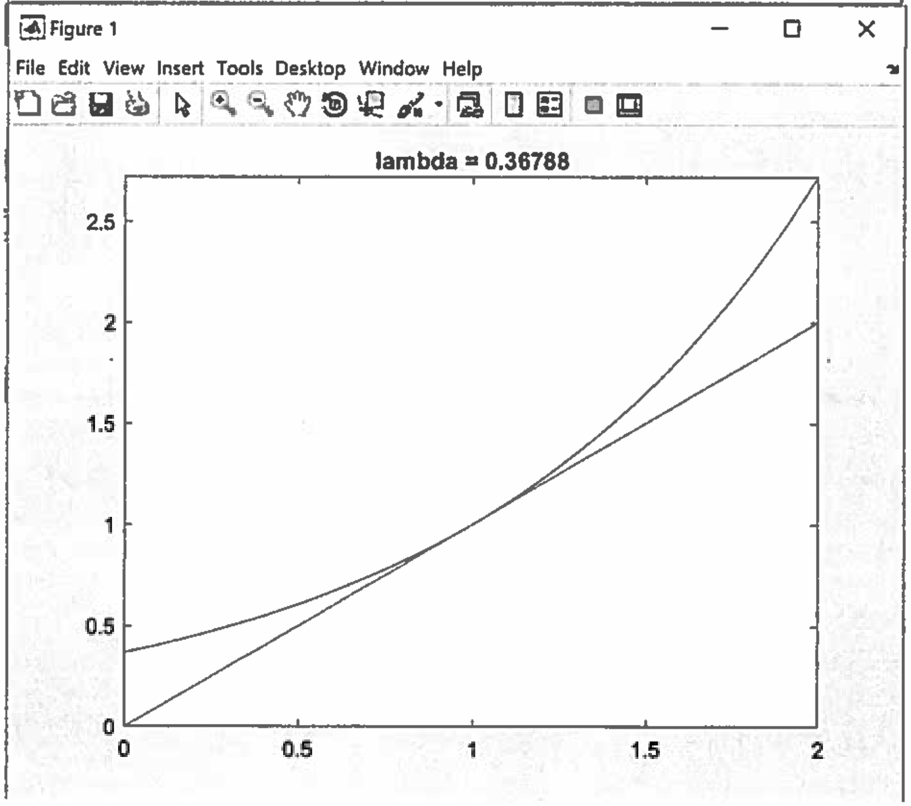


Figure 1
File Edit View Insert Tools Desktop Window Help
lambda = 0.36788
0
0.5
1
1.5
2
2.5
0
0.5
1
1.5
2

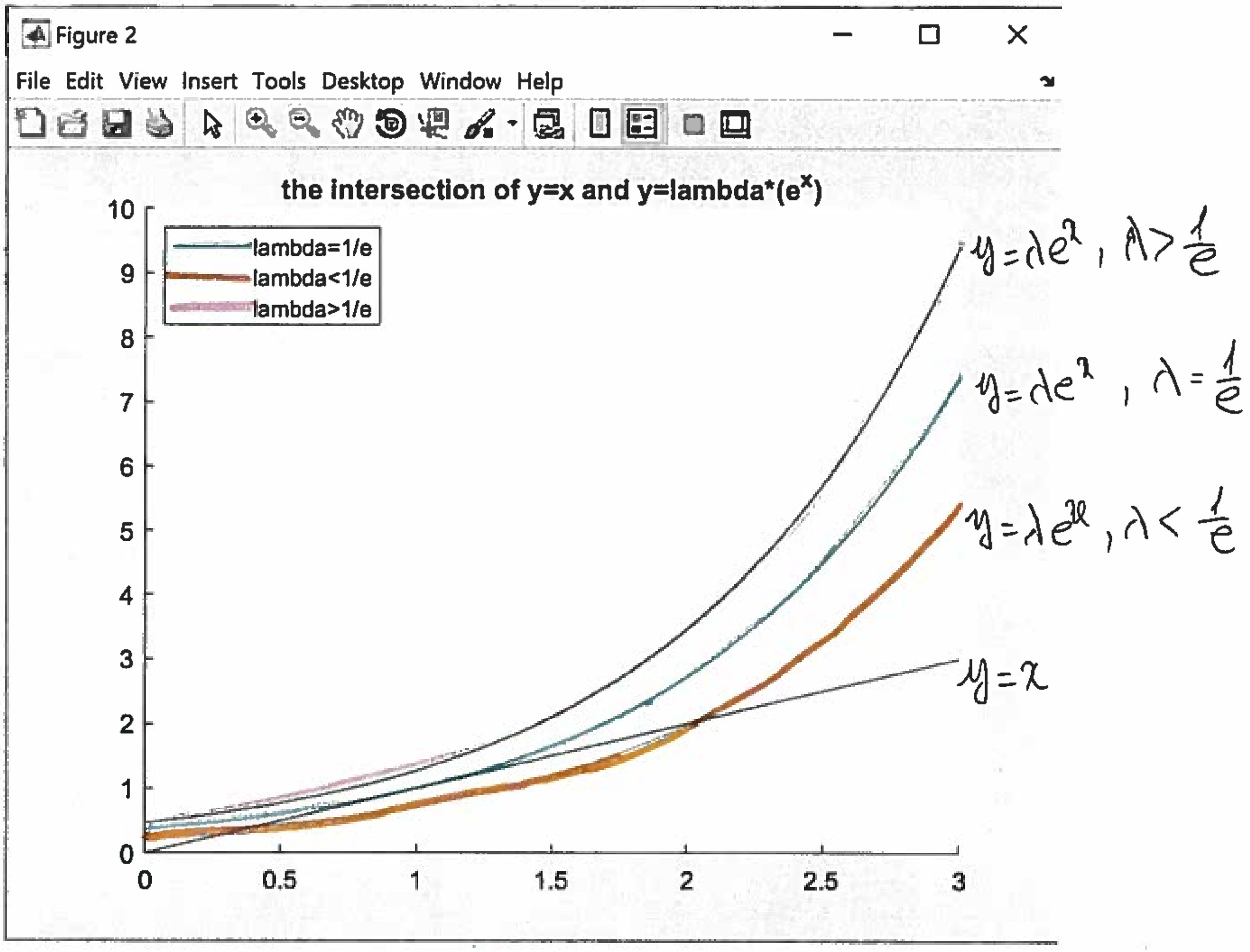


Figure 2
File Edit View Insert Tools Desktop Window Help
the intersection of y=x and y=lambda*(e^x)
lambda=1/e
lambda<1/e
lambda>1/e
10
9
8
7
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
$y=\lambda e^x, \lambda > \frac{1}{e}$
$y=\lambda e^x, \lambda = \frac{1}{e}$
$y=\lambda e^x, \lambda < \frac{1}{e}$
$y=x$

4b) For $\lambda < \frac{1}{e}$, consider two iterations

I1) $x_{n+1} = \lambda e^{x_n}$, $n = 0, 1, \ldots$

I2) $x_{n+1} = \ln x_n - \ln \lambda$, $n = 0, 1 \ldots$

Sho that (I1) $x_n \longrightarrow \tilde{x}$

(I2) $x_n \longrightarrow \hat{x}$

Analyze both methods as ~~sincf~~ simple iteration for a fixed point problem for a contractive mapping.

Consider I1

* Since $f(x) = x - \lambda e^{x} > 0$ when $x = 1$. Take x_0 so that $x_0 - e^{x_0} > 0$.

then $x_{n+1} = \lambda e^{x_n}$

$$\frac{x_{n+1}}{x_n} = \frac{\lambda e^{x_n}}{\lambda e^{x_{n-1}}} = e^{x_n - x_{n-1}}$$

We have $x_1 = \lambda e^{x_0} < x_0$

Hence by induction, $x_{n+1} < x_n$

Thus $\{x_n\}$ is decreasing $\Rightarrow x_n \rightarrow \tilde{x}$

- If we choose x_0 so that $x_0 - \lambda e^{x_0} < 0$

then $x_1 > x_0$

by induction, $x_{n+1} > x_n$

because x_n is bounded above by $-\ln\lambda \Rightarrow x_n \longrightarrow \tilde{x}$.

* Consider I2: $x_{n+1} = \ln x_n - \ln \lambda$

- Where $x_0 \in (-\ln\lambda, +\infty)$

We have $x_{n+1} - x_n = \ln(x_n) - \ln(x_{n-1})$.

If we choose x_0 so that $x_1 = \ln(x_0) - \ln(\lambda) < x_0$.

Then by induction, $\{x_n\}$ is decreasing $\Rightarrow x_n \longrightarrow \hat{x}$ since $\{x_n\}$ is bounded above by x_0

- If we choose x_0 such that $x_1 = \ln(x_0) - \ln(\lambda) > 0$ then $\{x_n\}$ is increasing by induction. $x_{n+1} > x_n$

Since $x_{n+1} = \ln(x_n) - \ln(\lambda) < x_n$ when x_n is large $\Rightarrow$ $\{x_n\}$ is bounded

$$\Rightarrow x_n \longrightarrow \hat{x}.$$

Analyze by contractive theorem.

- If $f(x) = x - \lambda e^x$

$f'(x) = 1 - \lambda e^x$

$f''(x) = -\lambda e^x < 0$

$\Rightarrow f'(x)$ is decreasing

$f'(x) = 0$ if $1 - \lambda e^x = 0$

$\Leftrightarrow x = -\ln \lambda$

x	0		$-\ln \lambda$		$+\infty$
$f'(x)$		$+$	0	$-$	
$f(x)$	$-\lambda\ (<0)$	$\nearrow$	$-\ln\lambda - 1$ if $\lambda < \frac{1}{e}$ >0	$\searrow$	$-\infty$ (<0)

Thus if $\lambda < \frac{1}{e}$ then $f(x) = 0$ has 2 solutions.

* Consider $x_{n+1} = \lambda e^{x_n}$ where $x_0 \in (0, -\ln \lambda)$

Since $1 > \lambda e$ we define $T(x) := \lambda e^x$ where $x \in (0,1)$

$$\Rightarrow T(x) \in (0,1)$$

$\Rightarrow |T(x) - T(y)| = |\lambda e^{z_0}||x - y|$ (by mean value theorem)

$\leq c|x-y|$ since $\sup \lambda e^{z_0} < 1$ $(z_0 \in (0,1))$

$\Rightarrow T^n$ is a contraction map

$\Rightarrow$ There is a fixed point $T(x) = x$ ~~of~~ or $\lambda e^x = x$.

57a) Compute (using a calculator) the number $9^{1/3}$ to six decimal, using Newton's method, starting at $x_0 = 2$.

b) Find the relevant quantities h_0, x_1, M of Kantorovich in this case.

c) Using Kantorowitch's theorem prove that Newton' method converges.

a) We have $x = 9^{1/3} \Leftrightarrow x^3 = 9 \Leftrightarrow f(x) = x^3 - 9 = 0$

So we want to find the solution of $f(x) = x^3 - 9 = 0$ using Newton's method.

* The iterative ~~Newton~~ equation for Newton method is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^3 - 9}{3x_n^2} \qquad f'(x) = 3x^2$$

* Then we have

$$x_0 = 2$$

$$x_1 = 2 - \frac{2^3 - 9}{3 * 2^2} = 2.083333$$

$$x_2 = x_1 - \frac{x_1^3 - 9}{3 * x_1^2} = 2.080089$$

$$\left.\begin{aligned} x_3 &= x_2 - \frac{x_2^3 - 9}{3 * x_2^2} = 2.080084 \\ x_4 &= x_3 - \frac{x_3^3 - 9}{3 * x_3^3} = 2.080084. \end{aligned}\right\} \text{same}$$

So we see that after 3 iteraction we get $x_3 = 2.080084$ and then we always get the same such number.

Then $9^{1/3} \approx 2.080084$.

b) Find the relevant quantities h_0, x_1, M of Kantorovich theorem in this case

- $h_0 = -[Df(x_0)]^{-1} f(x_0) = \frac{-f(x_0)}{f'(x_0)} = \frac{-(2^3-9)}{3 * 2^2} = \frac{8}{3*2^2} = \frac{2}{3} = 0.666667$.
- $x_1 = 2.083333$ (as above)
- $U_1 = B_{|h_0|}(x_1)$ where h_0 and x_1 are as above.

we want to find M so that $|Df(y_1) - Df(y_2)| \le M|y_1 - y_2|, \forall y_1, y_2 \in \overline{U_1}$

$|Df(y_1) - Df(y_2)| = |3y_1^2 - 3y_2^2| = 3|y_1+y_2||y_1-y_2|$

note that for $y_1, y_2 \in \overline{U_1} \Rightarrow |y_1+y_2| \le |y_1| + |y_2| \le 2|2.083333 + 0.666667| = 5.5$

Then $|Df(y_1) - Df(y_2)| \le 3 * 5.5 |y_1 - y_2| = 16.5$ $\qquad M = 16.5$.

c) We have that .

With $x_0 = 2$, $Df(x_0)$ invertible .

① $x_1 = x_0 + h_0$ $\quad U_1 = B_{|h_0|}(x_1)$

② $U_1 \subset U$, and $Df(x)$ satisfy $|Df(y_1) - Df(y_2)| \leq 16.5 |y_1 - y_2|$ $\quad \forall y_1, y_2 \in \overline{U_1}$.

③ $|f(x_0)| \left((Df(x_0))^{-1} \right)^2 M = |2^3 - 9| \left(\frac{1}{3 * 2^2} \right)^2 16.5 \leq \frac{1}{2}$.

Then by Kantorovich's theorem , the Newton's method converges .

* Consider $x_{n+1} = \ln(x_n) - \ln(\lambda)$ where $x_0 \in (-\ln\lambda , +\infty)$

Define $T(x) := \ln x - \ln\lambda$ where $x \in (-\ln\lambda, +\infty)$

→ $|T(x) - T(y)| = \left|\frac{1}{z}\right| |x-y|$ by mean value theorem .

$< c|x-y|$, for $c < 1$ where $z \in (-\ln\lambda, +\infty)$.

Thus T^n is a contraction map ,

therefor there is a fixed point $T(x) = x$ or $\ln(x) - \ln(\lambda) = x$

or $\lambda e^x = x$

* 6) Consider $f:\mathbb{R}^2 \longrightarrow \mathbb{R}^2$ given by

$$f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x^2-y-12 \\ y^2-x-11 \end{bmatrix}$$

a) Find the Lipschitz ratio M for the derivative Df that is

$$|Df(\vec{p}) - Df(\vec{q})| \leq M|\vec{p}-\vec{q}|$$

b) Starting at $\vec{x_0} = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ compute $\vec{x_1}$ as one step of Newton's method to solve $f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)$

c) Find a disc which contains a root of the equation.

a) Since $f\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} x^2-y-12 \\ y^2-x-11 \end{bmatrix}$, we have $Df\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{pmatrix} 2x & -1 \\ -1 & 2y \end{pmatrix}$

- Let $\vec{p} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$ $\vec{q} = \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$, Then

⊕ $|\vec{p}-\vec{q}| = \left|\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right| = \left|\begin{pmatrix} x_1-x_2 \\ y_1-y_2 \end{pmatrix}\right| = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$

⊕ $|Df(\vec{p})-Df(\vec{q})| = \left|\begin{pmatrix} 2(x_1-x_2) & 0 \\ 0 & 2(y_1-y_2) \end{pmatrix}\right| = \sqrt{4(x_1-x_2)^2+4(y_1-y_2)^2}$
$= 2\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$

So the Lipschitz ratio M for the the derivative Df that is $|Df(\vec{p})-Df(\vec{q})| \leq M|\vec{p}$ is $M = 2$.

b)

* First, we want to compute $h_0 = -[Df(x_0)]^{-1} f(x_0)$.

- $\vec{x_0} = \begin{pmatrix} 4 \\ 4 \end{pmatrix} \Rightarrow f(\vec{x_0}) = \begin{pmatrix} 4^2-4-12 \\ 4^2-4-11 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

- $Df(\vec{x_0}) = Df\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \begin{pmatrix} 2*4 & -1 \\ -1 & 2*4 \end{pmatrix} = \begin{pmatrix} 8 & -1 \\ -1 & 8 \end{pmatrix}$

then $\left[Df\begin{pmatrix} 4 \\ 4 \end{pmatrix}\right]^{-1} = \frac{1}{65}\begin{pmatrix} 8 & 1 \\ 1 & 8 \end{pmatrix}$

- $h_0 = -\left(Df\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}\right)\right)^{-1} f\left(\begin{bmatrix} 4 \\ 4 \end{bmatrix}\right) = -\frac{1}{65}\begin{pmatrix} 8 & 1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} = -\frac{1}{65}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

* $x_1 = x_0 + h_0 = \begin{pmatrix} 4 \\ 4 \end{pmatrix} - \frac{1}{65}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{259}{65}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 259/65 \\ 259/65 \end{pmatrix}$

Q7

* Find a disc that contains ~~the~~ a solution of the equation .

- We want to check the Kant equality :

$$|f(x_0)|\,|Df(x_0)^{-1}|^2 M = \left|\begin{pmatrix}0\\1\end{pmatrix}\right|\,\left|\frac{-1}{65}\begin{pmatrix}8 & -1\\-1 & 8\end{pmatrix}\right|^2 2 = \left(\sqrt{0^2+1^2}\right)\left(\frac{1}{65^2}\right)\left(\sqrt{8^2+8^2+1^2+1^2}\right)^2 2$$

$$= \frac{1}{65^2} * 130 * 2 = \frac{4}{65} < \frac{1}{2} \Rightarrow$$ Kant equality holds .

* So now we have the disc that contains a solution is $U_{\left|-\frac{1}{65}\begin{pmatrix}1\\1\end{pmatrix}\right|}(x_0)$.

①

$$\left|\frac{1}{-65}\begin{pmatrix}1\\1\end{pmatrix}\right| = \sqrt{\frac{1}{65^2}+\frac{1}{65^2}} = 0.021757$$

$$\| \quad U_{0.021}\left(\begin{bmatrix}4\\4\end{bmatrix}\right)$$

* Complex exponentials, trigonometric polynomials, discrete Fourier Analysis

① Complex exponentials

* Consider an interval $[0,a]$.

Define $f : [0,a] \longrightarrow \mathbb{C}$ $\quad \langle f, g \rangle = \int_0^a f(t)\,\overline{g(t)}\,dt$

* Define complex exponential

$e_k(t) = e^{i 2\pi k \frac{t}{a}}$ ← complex, orthogonal system

↘ they orthogonal to each other, but no one orthogonal to all of them

* Properties:

- $e_k(t+a) = e_k(t)$ ($e_k(t)$ is a periodic function of real argument t, periodic a.

- $\langle e_k, e_m \rangle = \begin{cases} 0 & k \neq m \\ a & k = m \end{cases}$ $\qquad e^z = 1$ if $\frac{z}{i 2\pi} \in \mathbb{Z}$

* Define $e(x) = e^{i 2\pi x}$, $x \in \mathbb{R}$.

- $e(1) = 1 \quad e(n) = 1$

  $e(x+L) = e(x)$ $\quad e$ is L-periodic

- Let $q > 0$, $q \in \mathbb{Z}$

  Then for any integer n, $\sum_{k=1}^{q} e\left(\frac{nk}{q}\right) = \begin{cases} q & \text{if } q \mid n \\ 0 & \text{otherwise} \end{cases}$.

* Orthogonal polynomials

* Let $(a,b) \subset \mathbb{R}$

Define w is a weight function ; $w(x) > 0$, $\forall x \in (a,b)$; $w \in L^1(a,b)$.

- Define a inner product of 2 functions defined on (a,b).

$$\langle f,g\rangle = \int_a^b f(x)\,g(x)\,w(x)\,dx \qquad \|f\| = \left[\int_a^b |f(x)|^2 w(x)\,dx\right]^{1/2}$$

$$L^2_w(a,b) = \{ f \mid f:(a,b) \longrightarrow \mathbb{R} \,,\, \|f\| < +\infty \}.$$

$$\langle f, gh\rangle = \langle fh, g\rangle$$

* We can't not construct orthogonal polynomial $L^2_w(a,b)$ by Gram Smith.

$p_0(x) = 1$

$p_1(x) = x - \dfrac{\langle 1,x\rangle}{\langle 1,1\rangle}\,1 = x$

$\vdots$

$p_n(x) = x^n - \sum_{i=0}^{n-1} d_{i,n}\, p_i(x) \qquad d_{i,n} = \dfrac{\langle x^n, p_i\rangle}{\langle p_i, p_i\rangle}$

these $p_n(x)$ is monic is NOT orthogonal

* Theorem (Triple recursion formula for constructing orthogonal polynomial)

There exists a unique sequence of polynomial $\{p_n\}_{n=0}^{\infty}$ such that

$$\begin{cases} p_n(x) \text{ is monic of degree } n \\ \langle p_n, q\rangle = 0 \,,\ \forall q \in \mathbb{P}_{n-1} \end{cases} \qquad \begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x. \end{aligned}$$

Such polynomials $\begin{cases} \text{are orthogonal } \langle p_i, p_j\rangle = 0 \,,\ \forall i \neq j \\ p_n(x) = (x-\lambda_n)\,p_{n-1}(x) - \mu_n\, p_{n-2}(x) \qquad n \geq 2 \end{cases}$

$$\lambda_n = \frac{\langle x\,p_{n-1}, p_{n-1}\rangle}{\|p_{n-1}\|^2} \qquad \mu_n = \frac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2}$$

* Properties of monic orthogonal polynomials.

- Let p_n be the n^{th} monic orthogonal polynomial
- Then $\|p_n\| \leq \|s\|$ for any monic polynomial of degree $\leq n$.
- The polynomial p_n has n real distinct roots in (a,b)

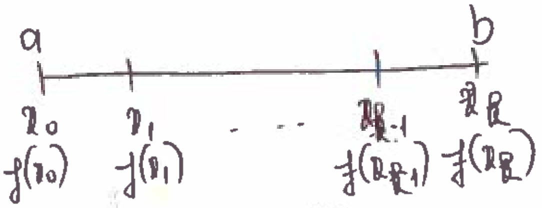
* Chebyshev polynomials are orthogonal polynomials with weight $w(x) = \dfrac{1}{\sqrt{1-x^2}}$

$T_n : (-1,1) \longrightarrow \mathbb{R}$

$$\langle T_m, T_n\rangle = \begin{cases} \|T_n\|^2 & \text{when } n=m \\ 0 & \text{when } n \neq m \end{cases}$$

* Gauss quadratures (About computing integral with weight $w(x)$)

* We want to compute $I(f) = \int_a^b f(x)\, w(x)\, dx$



we want to find $Q(f) = \sum_{i=0}^{l} \lambda_i f(x_i)$

The weight $\lambda_0, \dots, \lambda_l$ are chosen so that $Q(f) = I(f)$ for $f \in \mathbb{P}_l$.

$\Rightarrow$ Gauss quadrature : choose $x_0, \dots, x_l$ so that we can obtain exactness of $\boxed{\text{degree } (2l+1)}$

* Lemma .

Let f be a function in $\mathbb{P}_{2l+1}$

- Then choose the $(l+1)$ nodes :

The $(l+1)$ nodes are $\{x_0, \dots, x_l\}$ which are roots of $P_{l+1}(x)$

$(l+1)$ orthogonal polynomial in $L^2_w(a,b)$.

- The weights $\lambda_0, \dots, \lambda_l$

$$\lambda_i = \int_a^b l_i(x)\, w(x)\, dx \qquad l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{l} \frac{(x - x_j)}{x_i - x_j} \;, i = \overline{0, l} \qquad \boxed{l_0(x) = 1}$$

Then $\begin{cases} Q(f) = \sum_{i=0}^{l} \lambda_i f(x_i) \text{ is exact for } \boxed{f \in \mathbb{P}_{2l+1}} \\ \text{such quadrature is unique.} \end{cases}$

* Gauss Lobatto rule :

* We also want to estimate $I(f) = \int_a^b f(x)\, w(x)\, dx$.

- The notes : $\begin{cases} x_0 = a \quad x_n = b \\ x_1, x_2, \dots, x_{l-2}, x_{l-1} \text{ are roots of orthogonal polynomial with weights .} \end{cases}$

$\overline{W}(x) = (x-a)(x-b)\, w(x)$.

- The weight $\lambda_0, \dots, \lambda_l$.

$$\lambda_i = \int_a^b l_i(x)\, w(x)\, dx \qquad l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{l} \frac{(x - x_j)}{(x_i - x_j)}$$

Then $Q(f)$ is exact for $f \in \mathbb{P}_{(2l-1)}$

* Projection of function onto a span of an orthogonal set

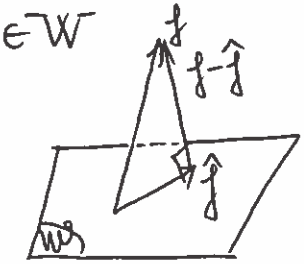
* $f:[a,b] \longrightarrow \mathbb{R}$ $\langle f, g\rangle = \int f(x)g(x)dx$ $\|f\|^2 = \left[\int |f|^2 dx\right]^{1/2}$ $L^2(a,b)$

* Let $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ be an orthogonal set of functions in $\boxed{L^2([a,b])}$

$W = \text{span}\{\varphi_1, \dots, \varphi_N\}$

We define the projection of f on W is a function $\boxed{\hat{f} \in W}$ such that f

$f - \hat{f} \perp \varphi \ \forall \varphi \in W \Leftrightarrow \langle f - \hat{f}, \varphi \rangle = 0, \ \forall \varphi \in W$



* We have $\hat{f} = \sum_{i=1}^{N} \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle} \varphi_i$

orthogonal projection of f on $\text{span}\{\varphi_1, \dots, \varphi_N\}$

* Def (the best orthogonal projection).

We say that $f^* \in W$ is the best approximation of f in W

$\underset{def}{\Leftrightarrow} \|f - f^*\| \leq \|f - \varphi\|, \ \forall \varphi \in W$

* The orthogonal projection f^* is the best approximation of f in W

* Nonlinear equation $f(x)=0$

* Want to solve $f(x)=0$

Instead of solving $f(x)=0$, convert this to problem $g(x) = x$

$$f(x)=0 \iff \alpha f(x) = 0 \iff x = \underbrace{x + \alpha f(x)}_{g(x)} \iff x = g(x)$$

* Consider $g(x) = x$

g must satisfy Brouer theorem assumption $\begin{cases} g:[a,b] \to [a,b] \\ \text{continuous} \end{cases}$.

- An algorithm for solving $g(x) = x$
  - ⊕ Chose a random x_0
  - ⊕ Algorithm $x_1 = g(x_0), \dots, x_k = g(x_{k-1})$ converges by continuity to ξ, $g(\xi)=\xi$

* Sufficient condition for convergence of simple interaction, Contraction mapping theorem

Let $g:[a,b] \to [a,b]$ continuous

sufficient condition: $\exists L$, $L \in (0,1)$, $|g(x) - g(y)| \leq L|x-y|$, $\forall x,y \in [a,b]$

Then $\begin{cases} g \text{ has a unique fixed point } \xi \in [a,b]. \\ \{x_k\} \text{ defined by } x_k = g(x_{k-1}) \text{ converges to } \xi \text{ for any starting point } x_0 \in [a,b] \end{cases}$

* Theorem (Local contraction mapping theorem)

Let $g:[a,b] \to [a,b]$ continuous

g' is continuous in some neighbor of ξ

$|g'(\xi)| < 1$

$\Big\}$ The sequence (x_k), $x_k = g(x_{k-1})$ converges to ξ (provided that x_0 is sufficiently close to ξ)

* Newton Raphson's theorem :

Suppose $f : \mathbb{R} \longrightarrow \mathbb{R}$ differentiable

$f(x^*) = 0$

Suppose that there exists 3 positive constants a_0, a_1, a_2 such that .

1> f is C^1 in $B_a(x^*)$

2> $f'(x)$ is invertible on $B_a(x^*)$, $|f'(x)^{-1}| \leq a_1$.

3> $x \longmapsto f'(x)$ is Lipchitz on $B_a(x^*)$, $|f'(x) - f'(y)| \leq a_2 |x-y|$

Then for any $x_0 \in B_b(x^*)$, $b < \min\{a, \frac{2}{a_1 a_2}\}$

the formula $\left\{ \begin{array}{l} x_{k+1} = x_k - [f'(x_k)]^{-1} f(x_k) \quad \text{(Newton method formula)} \\ \text{is well defined} \\ x_k \longrightarrow x^* \text{ quadratically}, \quad |x_k - x^*| < \frac{2}{a_1 a_2}\left(\frac{1}{2} a_1 a_2 |x - x^*|\right)^{2^k} . \end{array} \right.$

---

* Let $f : U \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^n$, $x_0 \in U$,
$Df(x_0)$ invertible

Define $h_0 = -[Df(x_0)]^{-1} f(x_0)$

$x_1 = x_0 + h_0 \qquad U_1 = B_{|h_0|}(x_1)$

① $|x| = \sqrt{x_i^2} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$

If $\left\{ \begin{array}{l} \overline{U_1} \subset U \\ Df(x) \text{ satisfies Lipchitz condition } |Df(y_1) - Df(y_2)| \leq M |y_1 - y_2| , \forall y_1, y_2 \in \overline{U_1} \text{ ②} \\ \text{Kant equality holds : } \underbrace{|f(x_0)| \, |Df(x_0)^{-1}|^2 M}_{\text{all three don't need to be small}} \leq \frac{1}{2} \quad \text{③} \end{array} \right.$

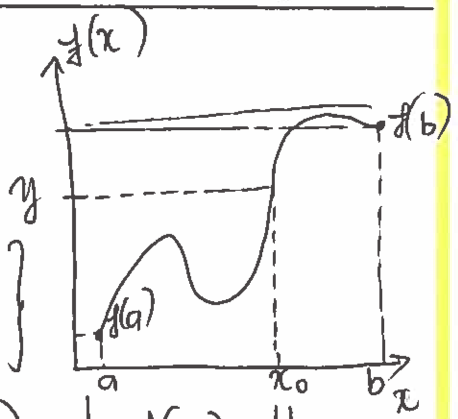
Then $\left\{ \begin{array}{l} \text{the equation } f(x) = 0 \text{ has a unique solution in } \overline{U_1} \\ x_{k+1} = x_k + h_k \text{ converges to this solution .} \end{array} \right.$

*Chapter 3: Solution of nonlinear equations.

*Introduction

- Want to solve a nonlinear equation $\begin{cases} \text{find } x \text{ such that } f(x)=0 \\ \text{find } X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \text{ so that } F(X)=0_{n\times 1}. \end{cases}$
- Example: find x : $x - \tan x = 0$
  $x - a\sin x = b$

* 3.1. Bisection (Interval Halving) method.



*Theorem 1: (Intermidiate value theorem)

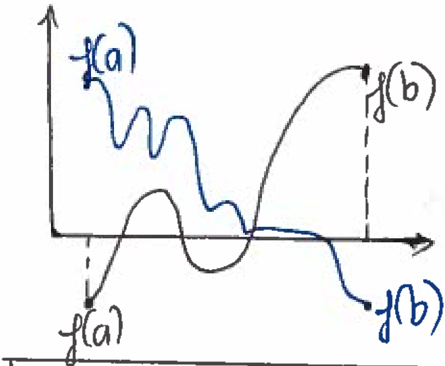
Let $\begin{cases} f:\mathbb{R}\longrightarrow\mathbb{R} \text{ continuous on a domain containing } [a,b] \\ f(a) < f(b) \end{cases}$

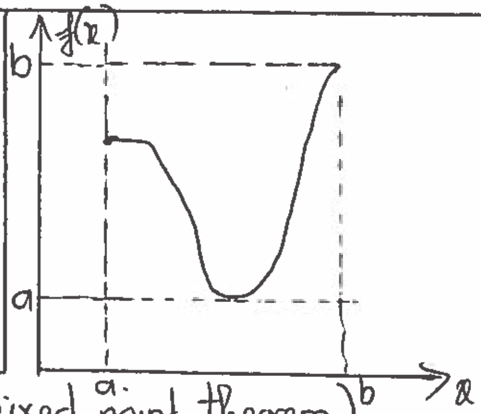
Then for any y s.t $f(a)<y<f(b)$, there exist$_{\wedge}$ $x_0\in(a,b)$ s.t $f(x_0)=y$

at least

*Theorem 2:

Let $\left.\begin{array}{l} f:[a,b]\longrightarrow\mathbb{R} \text{ continuous} \\ f(a)\,f(b) \le 0 \end{array}\right\} \Rightarrow \exists$ (at least) one solution ξ s.t $f(\xi)=0$





*Theorem 3: (Browner's fixed point theorem).

Let $g:\underbrace{[a,b]\longrightarrow[a,b]}_{\text{same}}$ be continuous.

Then: there exists $\xi\in[a,b]$, $f(\overbrace{\xi)=\xi}^{\text{same}}$.

313C Tues/Thursday 3:20- 4:20 pm

MAT 683 Methods of Numerical Analysis I
A. Lutoborski, Syracuse University
Fall 2018.

Classes: Tuesday, Thursday, 11:30-12:20, Carnegie Room 110.

Instructor: Professor Adam Lutoborski, Department of Mathematics, 311 A Carnegie, phone 443-1489, e-mail `alutobor@syr.edu`

Office Hours: Monday 2:00-3:00, Tuesday 10:00-11:00.

Text: "Numerical Analysis. Mathematics of Scientific Computing" by D. Kincaid and W. Cheney, 3rd Edn, AMS 2002.

Prerequisites: MAT 512. MATLAB will be used in some of our homework problems.

Exams, Homeworks, Final Exam: There will be two exams and a cumulative final exam given in this course. Homework will be given every week. Exam 1 will be given after chapters 1,2,6 are covered and Exam 2 after Chapter 7 is covered. Precise date will be announced in class. Dates of the exams will be announced approximately a week before the exam. Final Exam: Wednesday December 11, 12:45-2:45 pm.

Course Grades: Course grades will be determined by: homework= 35%, 2 exams= 40%, final exam= 25%.

Course Description: This is an introductory graduate course in numerical analysis. We cover: computer arithmetic, interpolation and approximation of functions, numerical differentiation and integration, solution of nonlinear equations. The course material will be selected from chapters 1,2,6,7,3 (in that order) of the text.

Course Content:

1. Basic concepts in numerical analysis
   - 1.1 Mathematical preliminaries
   - 1.2 Floating point arithmetic
   - 1.3 Sensitivity analysis
2. Approximation of functions
   - 2.1 Polynomial interpolation
   - 2.2 Hermite interpolation
   - 2.3 Spline interpolation
   - 2.4 Trigonometric interpolation
   - 2.5 Least squares approximation
3. Numerical integration
   - 3.1 Interpolatory quadratures
   - 3.2 Composite quadratures
   - 3.3 Gaussian quadratures
4. Solution of nonlinear equations
   - 4.1 The bisection method

Solve problems

· expand

can't expect can be solved in finit # of steps.

#dimensions is big

→ approximate the sol

4.2 Fixed point iteration

4.3 Newton's method its convergence and modifications

Disability-Related Accomodations: Students who are in need of disability-related academic accommodations must register with the Office of Disability Services (ODS), 804 University Avenue, Room 309, 315-443-4498. Students with authorized disability-related accommodations should provide a current Accommodation Authorization Letter from ODS to the instructor and review those accommodations with the instructor. Accommodations, such as exam administration, are not provided retroactively; therefore, planning for accommodations as early as possible is necessary. For further information, see the ODS website, Office of Disability Services http://disabilityservices.syr.edu/

Academic Integrity: The Syracuse University Academic Integrity Policy holds students accountable for the integrity of the work they submit. Students should be familiar with the Policy and know that it is their responsibility to learn about instructor and general academic expectations with regard to proper citation of sources in written work. The policy also governs the integrity of work submitted in exams and assignments as well as the veracity of signatures on attendance sheets and other verifications of participation in class activities. Serious sanctions can result from academic dishonesty of any sort. For more information and the complete policy, see http://academicintegrity.syr.edu

Religious observances policy: SU religious observances policy recognizes the diversity of faiths represented among the campus community and protects the rights of students, faculty, and staff to observe religious holidays according to their tradition. Under the policy, students are provided an opportunity to make up any examination, study, or work requirements that may be missed due to are religious observance provided they notify their instructors before the end of the second week of classes. For fall and spring semesters, an online notification process is available through MySlice (Student Services → Enrollment → My Religious Observances) from the first day of class until the end of the second week of class.

Mat 683 Method of numerical analysis I

- Most of the problems of "continuous" problems in mathematics can't be solve by finite algorithm
- Numerical analysis constructs algorithms that give solutions converge to approximate answer.
- Science : theory, experiment, computation .

* Discretization

Approximation of functions by simpler functions .
interpolation
series expansion
harmonic analysis (Fouries series, discrete F. transform)
extrapolation
polynomials, orthogonal polynomials, trigonometric polynomial, piecewise polynomials
spline functions
wavelets (we ful these days), sinc, racho, basic function, trigonometric polynomial

Quadrative
Optimalization (minimalization of functions of many variables)
algebraic complexity
parallel algorithm
adaptive algorithm .

L13. Floating point arithmetic

range limited

$\underbrace{2^{31} \approx 2.1 \times 10^{9}}$

Binary numbers

* Fixed-point number (integer)

32 binary digit (alters) → represent 2^{32} integers : from $\underbrace{-2^{31} \text{ to } 2^{31}-1}$

* IEEE floating point numbers : { single precision : 32 ~~bit~~ digits range
{ double precision : 64 digits.

$\underbrace{(-1)^{\text{sign}}}_{\text{1digit}} \underbrace{z}\, 2^{p}$ ← exponential bias.

$z = \underbrace{1.}_{\text{not stored}}\underbrace{xxxx\ldots}_{\text{fraction}}$ $\boxed{1 \le z < 2}$

32

$\underset{\text{|sign|}}{1}\ \underbrace{1,xxx\ x\ldots x}_{23})\ 2^{\overbrace{xxxxxxxx}^{8}}$

	sign	fraction	p	total
• single	1	23	8 bit	32 bit
double	1	52	11	64

Single precision.

• With ~~single precision~~, we use 8bits for p : { 1 bit for sign
{ 7 bit for the magnitude

sign ↓ magnitude.

$2^{\textcircled{1}\overbrace{1111111}} = 2^{1} + \ldots + 2^{7} = 127$

~~⇒ the exponent will range from -127 to 127~~

~~⇒ the only disadvantage for this method : there are 2 representation for 0 exponent +0, -0~~

⇒ We use excess representation / biased format for Exponent bias :

* Excess representation / biased format for Exponent bias : (unique representation for 0)

Exponent bit $0\ \cancel{1} \longrightarrow 127 \longrightarrow \cancel{254}\ 255$

negative 0 positive

Example : reserve -126 127 reserve

→ The range of exponents is
$-12\cancel{7}6 \longrightarrow 12\cancel{8}7$

$1 \longrightarrow 1-127 = -126$

$254 \longrightarrow 254-127 = 127$

$\boxed{\text{expo bias} = 127}$

then
• smallest
2^{-126}
• largest
$\cancel{2}.1\ldots1 \cdot 2^{127}$
$= [\cancel{2} + (1-2^{23})]$

• The largest floating point number which can be represented is

$\underbrace{\cancel{1}.11\ldots1}_{23}\ 2^{\cancel{255}\underbrace{1\ldots1}_{8}}$ $\overset{\text{tailing}}{=} (2-2^{-23})\,2^{127}$ $\approx 3.4 \times 10^{38}$

$= (1-2^{-23}) \times 2^{(127)}$ $) \approx 2 \times 2^{127} = 2^{128} =$ the largest

$\approx 3.4 \times 10^{38}$

• The smallest floating point number

$0.\underbrace{\cancel{1}\ldots0}_{23}\ 2^{0\ldots0} \approx 0.29 \times 2^{-127}$ $(2^{-1}$ bal 0.293×10^{-38}

- Example

⊕ $1.11111 = 1\times 2^{0} + 1\times 2^{-1} + 1\times 2^{-2} + 1\times 2^{-3} + 1\times 2^{-4} + 1\times 2^{-5} =$

⊕ $\underbrace{2}_{10\text{base}} = \underbrace{1.\underbrace{0\ldots0}_{23} \times 2^{1}}_{2\text{ base}}$

---

* Machine epsilon ε_m , relative difference

* The gap between 1 and the next bigger number is called machine epsilon ε_m

- $1 = (-1)^0 \times 1,0\ldots0 \times 2^0$

next bigger $= (-1)^0 \times 1,\underbrace{0..0}_{22}1 \times 2^0$

} This gap is called machine epsilon $\varepsilon_m = 2^{-23}$.

- Consider $x \in \mathbb{R}$

$x = 1 + \frac{\varepsilon_m}{2}$, x can't be represented by any single digit ⇒ need to round.

$1 + 0.4\,\varepsilon_m$: round down to 1

$1 + 0.6\,\varepsilon_m$: round up to $1 + \varepsilon_m$

$1 + 0.5\,\varepsilon_m$: too complicated ⇒ skip.

* relative difference $= \frac{\text{next bigger number} - \text{this number}}{\text{this number}} = \varepsilon_m$

- EX: Find the relative difference between 2^{10} and the next bigger number

$2^{10} = (-1)^0 \times 1.000..0 \times 2^{10}$

next bigger $= (-1)^0 \times 1.\underbrace{00..0}_{22}1 \times 2^{10} = 2^{10} + \underbrace{2^{-23} \times 2^{10}}_{\varepsilon_m}$

relative difference $= \frac{(2^{10} + 2^{-23} \times 2^{10}) - 2^{10}}{2^{10}} = 2^{-23}$.

- $\boxed{\text{Zero:}}$ $\begin{cases} \text{sign} = 0 \text{ or } 1 \ (\text{positive/negative zero}) \\ p = \text{all } 0\text{'s} \\ \text{fraction} = \text{all } 0\text{'s} \end{cases}$

* Double precision numbers :

$2^{11} = 2048$

- $(-1)^{sign} * 1.\underbrace{x \ldots x}_{52\text{ bit}} \cdot 2^{p} \leftarrow$ 11 bits for p

MH : double precision number
$(-1)^{sign} \cdot 1,\underbrace{x \ldots x}_{52} \cdot 2^{\overbrace{p-1023}^{\text{bias}}}$

p: 11 digits : from 0 – 2047

exponent : $\boxed{-1022}$ to $\boxed{102\cancel{3}4}$

fraction has 52 digits

$(2-2^{-52}) \times 10^{1023}$

- largest possible number: $(-1)^0\, 1.1\ldots1 \cdot 2^{1023} \approx 2 \cdot 2^{1023} = \boxed{2^{1024}} \approx 1.8 \cdot 10^{308}$
- smallest possible number: $(-1)^0\, 1.0\ldots0 * \boxed{2^{-1022}} = 2^{-1022} \approx 2.23 * 10^{-308}$?

* Machine precision ← machine epsilon : the gap betwen 1 and the next bigger number

$\boxed{\varepsilon_m = 2^{-52}}$

- $1 = (-1)^0 \times 1.0\ldots0 \times 2^0 =$
  next bigger $= (-1)^0 \times 1.\underbrace{0\ldots01}_{51} \times 2^0 = 2^{-52}$ $\Big\} \Rightarrow \varepsilon_m = 2^{-52}$

(Sometime $\frac{\varepsilon_m}{2}$ is called machine precision .

- $x \in \mathbb{R}$ $fl(x)$: floating point representation / approximation of x

for any x between 1 and $1+\varepsilon_m$
  then $|x - fl(x)| \leq \frac{\varepsilon_m}{2}$

$x + 0.6\,\varepsilon_m \longrightarrow 1 + \varepsilon_m \leftarrow$ floating point representation of x

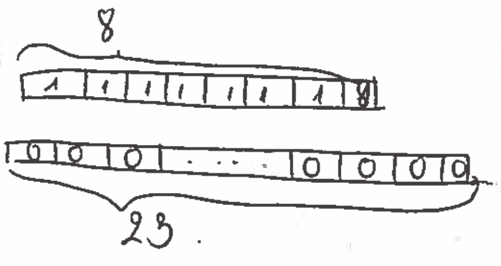
$|x - fl(x)| = |(1+0.6\varepsilon_m) - (1+\varepsilon_n)| = 0.4\,\varepsilon_m$.

---

* Special cases .

- $+\infty$ (positive infinity)
  (overflow problem)

$\begin{cases} \text{sign} = 0 \\ p = 255 \\ \text{fraction} = 0 \end{cases}$



- $-\infty$ (negative infinity)

$\begin{cases} \text{sign} = 1 \\ p = 255 \\ \text{fraction} = 0 \end{cases}$

- divide by zero

$\begin{cases} \text{sign} = 0 \text{ or } 1 \\ p = \text{all } 1 \\ \text{fraction} = \text{anythy except all 0's} \end{cases}$

* Catestophic canclellation : example

- $x_1 = 1 + 0.4\mathcal{E}_m$ $\quad$ $x_2 = 1 + 0.6\mathcal{E}_m$
  $fl(x_1) = 1$ $\quad$ $fl(x_2) = 1 + \mathcal{E}_m$

- $x_2 - x_1 = (1 + 0.6\mathcal{E}_m) - (1 + 0.4\mathcal{E}_m) = 0.2\mathcal{E}_m$
  $fl(x_2) - fl(x_1) = (1 + \mathcal{E}_m) - 1 = \mathcal{E}_m$

- absolute error
  $|(x_2 - x_1) - (fl(x_2) - fl(x_1))| = |0.2\mathcal{E}_m - \mathcal{E}_m| = 0.8\mathcal{E}_m$.

- relative error .

$$\frac{|(x_2 - x_1) - (fl(x_2) - fl(x_1))|}{|x_2 - x_1|} = \frac{0.8\mathcal{E}_m}{0.2\mathcal{E}_m} = 4$$

error 4 times as big as true solution
⇒ catestrophic .

* Summarize :

$$(-1)^{sign} , 1, \overbrace{x\,x \ldots x}^{fraction} \quad 2^{p} \text{ exponent bias .}$$

- exception (non numbers)

p \ fraction	0 0 ... 0	not all zeros .
0 0 ... 0	± 0	under flow
1 ... 1	± ∞ (over flow)	NAN

	# sign bit	# fraction bit	# exponent bit	# total bit	exponent bias
single	1	23	8	32	127
double	1	52	11	64	1023

largest possible	smallest possible	$\mathcal{E}_m$
$2^{128} \approx 3.8 \times 10^{38}$	$2^{-126} \approx 1.18 \times 10^{-38}$	2^{-23}
$2^{1024} \approx 1.8 \times 10^{308}$	$2^{-1022} \approx 2.23\ 10^{-38}$	2^{-52}

* Example : Catastrophic cancellation. good ide not to subtract 2 numbers that are close to each other

$x^2 - 56x + 1 = 0$

$x_1 = 28 + \sqrt{783} = 28 + 27.982 \underbrace{(\pm 0.0005)}_{\text{error}}$

$x_2 = 28 - \sqrt{783} = 28 - 27.982 \underbrace{(\pm 0.0005)}_{\text{error}}$

absolute error : the same

relative error :

x_1 true : $55.9815 \le x_1^{true} \le 55.9825$.

$0.0175 \le x_2^{true} \le 0.0185$.

- Let look at the relative error

$$\frac{|x_1 - x_1^{true}|}{|x_1^{true}|} \le \frac{0.0005}{55.9815} \approx 9.10^{-5}$$

$$\frac{|x_2 - x_2^{true}|}{|x_2^{true}|} \le \frac{0.0005}{0.0175} = 3.10^{-2}$$

⇒ try not to subtract. how to get

- $x_1^{true} * x_2^{true} = 1$. ← gives much better result

$x_2 = \frac{1}{x_1}$ (assume we are happy with our x_1)

$$|x_2 - x_2^{true}| = \left|\frac{1}{x_1} - \frac{1}{x_1^{true}}\right| = \left|\frac{x_1^{true} - x_1}{x_1 \, x_1^{true}}\right| \le \frac{0.0005}{55.9815 * 55.9825} =$$

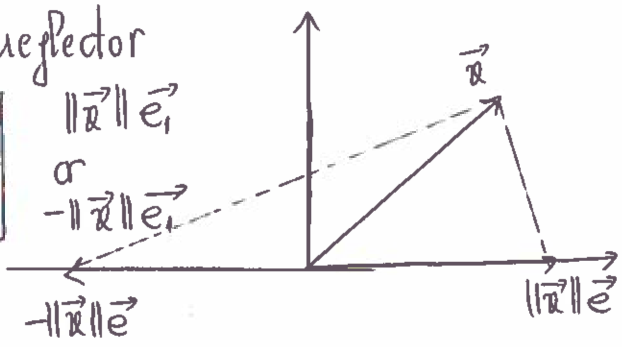
$$= 1.6 * 10^{-7}.$$

$$\frac{|x_2 - x_2^{true}|}{|x_2^{true}|} = \frac{1.6 * 10^{-7}}{0.0175} \approx 9 \cdot 10^{-6}$$

* Example : Householder reflector

$$\begin{bmatrix} * \\ * \\ * \end{bmatrix} \xrightarrow{F} \begin{bmatrix} * \\ 0 \\ 0 \end{bmatrix}$$

$\|\vec{x}\| \vec{e_1}$
or
$-\|\vec{x}\| \vec{e_1}$



we don't want to subtract $\vec{x}$ and $\|\vec{x}\|\vec{e}$ when they are too closed to each other.

* We also don't want to add (too small number) and a (too big number).
(adding 2 numbers whose magnitude are very different) (in finite position)

1 and 2^{-53}

$(-1)^0 \, 1.\underbrace{000\ldots0}_{52} \; 2^0$

2^{-53}
$= (-1)^0 \, 1.\underbrace{0\ldots0}_{52} \; 2^{-53}$

floating point format norm in double position ?

$1 + 2^{-53} = 1.$

* Example : To see that modified Gram Smith is more $>$ accurate Original GramSmit .

- Remind (~~column of A are linearly independent~~)

Original GS

```
for $i = 1 : n$
    $\vec{v_i} = \vec{a_i}$
    for $j = 1 : i-1$
        $r_{ji} = \vec{q_j}^* \vec{a_i}$
        $v_i = \vec{v_i} - r_{ji} \vec{q_j}$
    end
    $r_{ii} = \|v_i\|$
    $\vec{q_i} = \frac{v_i}{r_{ii}}$
end .
```

Modified GS .

```
for $i = 1 : n$
    $v_i = a_i$
end

for $i = 1 : n$
    $r_{ii} = \|v_i\|$
    $\vec{q_i} = \frac{\vec{v_i}}{r_{ii}}$
    for $j = i+1 : n$
        $r_{ij} = \vec{q_i}^* v_j$
        $v_j = \vec{v_j} - r_{ij} q_i$
    end
end .
```

* Let matrix A : $\vec{a_1} = \begin{bmatrix} 1+\varepsilon \\ 1 \\ 1 \end{bmatrix}$ $\vec{a_2} = \begin{bmatrix} 1 \\ 1+\varepsilon \\ 1 \end{bmatrix}$ $\vec{a_3} = \begin{bmatrix} 1 \\ 1 \\ 1+\varepsilon \end{bmatrix}$

ε is 'small' (very closed to be dependent)
$\varepsilon^2 = 0$ (in finite computation).

$\Rightarrow \vec{q_2}^* \vec{q_3} = \frac{1}{2}(1+0+0) = \frac{1}{2} \neq 0$

* Gram Smitch → The GS does not give you orthogonal matrix Q.

- S1 : $\vec{u_1} = \vec{a_1}$

$$\lambda_{11} = \|v_1\| = \sqrt{(1+\varepsilon)^2+1+1} = \sqrt{3+2\varepsilon+\varepsilon^2} \circleddash \sqrt{3+2\varepsilon}$$

$$\varepsilon = \varepsilon_m = 2^{-52} \quad 2\varepsilon = 2 \cdot 2^{-52} = 2^{-51}$$

$3: \quad 1.1\underbrace{00\ldots0}_{51} \times 2^1$

$2\varepsilon: \quad 1.\underbrace{0\ldots\ldots0}_{52} \; 2^{-51}$

$\varepsilon^2: \quad 1.0\ldots\ldots0 \; 2^{-104}$

How to add 2 floating point double precision

$1.1\underbrace{0\ldots\ldots0}_{50}1 \times 2^1 \quad (2^0+2^{-1}+\ldots+2^{-51}) \times 2^1) = 3+2\varepsilon$

$$\boxed{q_1 = \frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix}}$$

- Step 2 $\vec{v_2} = \vec{a_2}$

$$\lambda_{12} = \vec{q_1}^* \vec{a_2} = \frac{1}{\sqrt{3+2\varepsilon}}(1+\varepsilon+1+\varepsilon+1) = \sqrt{3+2\varepsilon}$$

$$v_2 = \begin{bmatrix} 1 \\ 1+\varepsilon \\ 1 \end{bmatrix} - \sqrt{3+2\varepsilon}\,\frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 \\ 1+\varepsilon \\ 1 \end{pmatrix} - \begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\varepsilon \\ \varepsilon \\ 0 \end{pmatrix}$$

$$\lambda_{22} = \|\vec{v_2}\| = \sqrt{\varepsilon^2+\varepsilon^2+0} = \sqrt{2\varepsilon^2} = \sqrt{2}\,\varepsilon$$

$$\boxed{\vec{q_2} = \frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}}$$

- Step 3 $\vec{v_3} = \vec{a_3}$

$$\lambda_{13} = q_1^* \vec{a_3} = \frac{1}{\sqrt{3+2\varepsilon}}(1+\varepsilon+1+1+\varepsilon) = \sqrt{3+2\varepsilon}$$

$$\vec{v_3} = \begin{pmatrix} 1 \\ 1 \\ 1+\varepsilon \end{pmatrix} - \sqrt{3+2\varepsilon}\,\frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\varepsilon \\ 0 \\ \varepsilon \end{pmatrix}$$

$$\lambda_{23} = \vec{q_2}^* \vec{a_3} = \frac{1}{\sqrt{2}}(-1+1+0) = 0$$

$$\lambda_{33} = \sqrt{2}\,\varepsilon \qquad \vec{q_3} = \frac{1}{\sqrt{\ }}\begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

* Modified GS

- Step 1 $\vec{v_1} = \vec{a_1}$, $\vec{v_2} = \vec{a_2}$, $v_3 = \vec{a_3}$

$$\lambda_{11} = \|\vec{v_1}\| = \sqrt{(1+\varepsilon)^2+1+1} = \sqrt{3+2\varepsilon+\varepsilon^2} \circleddash \sqrt{3+2\varepsilon}$$

$$q_1 = \boxed{\frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix}}$$

$$\lambda_{12} = \vec{q_1}^* \vec{v_2} = \frac{1}{\sqrt{3+2\varepsilon}}(1+\varepsilon+1+\varepsilon+1) =$$

$$= \sqrt{3+2\varepsilon}$$

$$\vec{v_2} = \begin{pmatrix} 1 \\ 1+\varepsilon \\ 1 \end{pmatrix} - \sqrt{3+2\varepsilon}\,\frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\varepsilon \\ \varepsilon \\ 0 \end{pmatrix}$$

$$\lambda_{13} = \vec{q_1}^* \vec{v_3} = \frac{1}{\sqrt{3+2\varepsilon}}(1+\varepsilon+1+1+\varepsilon) = \sqrt{3+2\varepsilon}$$

$$\vec{v_3} = \begin{pmatrix} 1 \\ 1 \\ 1+\varepsilon \end{pmatrix} - \sqrt{3+2\varepsilon}\,\frac{1}{\sqrt{3+2\varepsilon}}\begin{pmatrix} 1+\varepsilon \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -\varepsilon \\ 0 \\ \varepsilon \end{pmatrix}$$

- Step 2

$$\lambda_{22} = \|\vec{v_2}\| = \sqrt{\varepsilon^2+\varepsilon^2+0} = \sqrt{2}\,\varepsilon$$

$$\vec{q_2} = \frac{\vec{v_2}}{\lambda_{22}} = \boxed{\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}}$$

now different

$$\lambda_{23} = \vec{q_2}^* \underbrace{\vec{v_3}}_{\vec{a_3}(GS)}$$

in MGS do a little chg

$$= \frac{1}{\sqrt{2}}(\varepsilon+0+0) = \frac{1}{\sqrt{2}}\varepsilon$$

$$\vec{v_3} = \begin{pmatrix} -\varepsilon \\ 0 \\ \varepsilon \end{pmatrix} - \frac{1}{\sqrt{2}}\varepsilon\,\frac{1}{\sqrt{2}}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \left(\ \right)$$

$$= \begin{pmatrix} -\varepsilon \\ 0 \\ \varepsilon \end{pmatrix} - \frac{\varepsilon}{2}\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -\varepsilon/2 \\ -\varepsilon/2 \\ \varepsilon \end{pmatrix}$$

- Step 3 :

$$\lambda_{33} = \sqrt{\frac{\varepsilon^2}{4}+\frac{\varepsilon^2}{4}+\varepsilon^2} = \sqrt{\frac{3}{2}}\,\varepsilon$$

$$q_3 = \frac{\vec{v_3}}{\lambda_{33}} = \ldots = \boxed{\frac{1}{\sqrt{6}}\begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}}$$

$\vec{q_1} + \vec{q_2} = \left| \frac{1}{\sqrt{6+4\varepsilon}} (-L-\varepsilon+L) \right| = \left| \frac{-\varepsilon}{\sqrt{6+4\varepsilon}} \right| \leq \frac{\varepsilon}{\sqrt{6}} = 0$

$\vec{q_2} * \vec{q_3} = \frac{1}{\sqrt{6}} \frac{1}{\sqrt{12}} (1-1+0) = 0$

$\vec{q_1} * \vec{q_1}$ = the same

$\vec{q_1} \vec{q_3} =$

Floating point number system.
(IEEE Standard 754-1985 for Binary Floating Point Arithmetic)

We consider the real numbers of the form

$$x = \pm s 2^E$$

2 is the base or radix. A rational number s, $1 \le s < 2$ is called the significand or mantissa. Integer E is called the exponent, $E_{min} \le E \le E_{max}$. The mantissa, exponent and the sign are represented in a binary format using a $(t+l+1)$-bit word.

$$s = 1 + f, \qquad f = \sum_{i=1}^{t} f_i 2^{-i}, \quad f_i \in \{0,1\}$$

$$E = e - b, \qquad e = \sum_{i=0}^{l-1} e_i 2^i, \quad e_i \in \{0,1\}.$$

In other words $f = (0.f_1 \dots f_t)_2$, $s = (1.f_1 \dots f_t)_2$ and $e = (e_{l-1} \dots e_0)_2$. We store $p \in \{0,1\}$ and $1 - 2p = \pm 1$ is the sign, f, and the integer $e = E + b$, $e \ge 0$ which is the biased exponent, b is a positive integer and is called the bias. Exponent E may be negative or positive adding the bias to it allows us to store a positive integer e. Finally we have the representation

$$x = (1-2p)(1.f_1 \dots f_t)_2 2^E$$

$\mathbb{F}(2, t, E_{min}, E_{max})$ denotes the set of all floating point numbers.

The numbers in the IEEE standard are stored in two formats: single format: 32-bit, $l = 8$, $t = 23$ or double format: 64-bit, $l = 11$, $t = 52$ as in the tables below

p	e_7	$\dots$	e_0	f_1	$\dots$	f_{23}

p	e_{10}	$\dots$	e_0	f_1	$\dots$	f_{52}

The largest e is $e_{max} = 2^l - 1$ and the smallest is $e_{min} = 0$. However we reserve the values $E = e_{max} - b$ and $E = -b$ for special fl numbers and instead we take $E_{max} = e_{max} - 1 - b$ and $E_{min} = -b + 1$, $f = e = 0$ represents ± 0. In single format $E_{max} = 2^8 - 1 - 1 - 127 = 127$, $E_{min} = -127 + 1 = -126$.

	single	double
t: bits in mantissa	23	52
l: bits in exponent	8	11
E_{max}: max exponent	127	1023
E_{min}: min exponent	-126	-1022
b: bias	127	1023

Exponent	Fraction	numerical value	Comments
$E = E_{min} - 1$	$f = 0$	± 0	
$E = E_{min} - 1$	$l = 0,\ f \neq 0$	$\pm(0.f)_2 2^{E_{min}}$	subnormals
$E_{min} \leq E \leq E_{max}$	any f	$\pm(1.f)_2 2^E$	normals
$E = E_{max} + 1$	$f = 0$	$\pm\infty$	like $\frac{1}{0}$ or $\frac{-1}{0}$
$E = E_{max} + 1$	$f \neq 0$	NaN	like $\sqrt{-1}$ or $\frac{0}{0}$

- Only a finite number of rational numbers belong to $\mathbb{F}$.
- The increment between the consecutive fl-numbers in $[2^E, 2^{E+1})$ is $\Delta_E = 2^{E-t}$. The increment doubles from interval $[2^E, 2^{E+1})$ to $[2^{E+1}, 2^{E+2})$.
- There are 2^t fl-numbers in each interval $[2^E, 2^{E+1})$ for all integers E such that $E_{min} \leq E \leq E_{max}$.

Similar statements apply to the negative fl-numbers. There are $2 \cdot 2^t(E_{max} - E_{min} + 1) + 1$ numbers in $\mathbb{F}$. The factors count the number of signs, the number of mantissas, the number of exponents plus 0.

$$x_{min} = 2^{E_{min}} \leq |x| \leq x_{max} = (2 - 2^{-t})2^{E_{max}}$$

Take $f = 0$, $E = E_{min} = -b + 1$ to obtain $x_{min} = 2^{E_{min}}$. To obtain x_{max} take $f = (0.1 \ldots 1)_2$ then

$$1 + f = (1.\underbrace{1 \ldots 1}_{t})_2 = 2^0 + 2^{-1} + \cdots + 2^{-t} = \frac{1 - (2^{-1})^{t+1}}{1 - 2^{-1}} = 2 - 2^{-t}$$

Take such maximal mantissa $s = 2 - 2^{-t}$ and $E_{max} = e_{max} - 1 - b$ to obtain $x_{max} = (2 - 2^{-t})2^{E_{max}}$. In single format $x_{min} = 2^{-126} = 1.2 \cdot 10^{-38}$ $x_{max} = (2 - 2^{23})2^{127}$.

In double format $x_{min} = 2^{-1022} \approx 2.2 \cdot 10^{-308}$, $x_{max} = (2 - 2^{-52})2^{1023} \approx 1.8 \cdot 10^{308}$.

If $x \in \mathbb{F} \cap [2^E, 2^{E+1})$ then $x = (1 + f)2^E$ and $1 + f < 2 - 2^t$. Then the next bigger number in $\mathbb{F} \cap [2^E, 2^{E+1})$ is obtained by making the smallest increment to f by adding $(0.\underbrace{0 \ldots 01}_{t})_2$ which results in the number $(1 + f + 2^{-t})2^E = x + 2^{E-t}$. Hence the increment is

$$\Delta_E = 2^{E-t}$$

In particular $x = 1$ is in the interval $[1, 2)$. The gap between $1 = (1 + 0)2^0$ and the next floating point number is denoted

$$\Delta_0 = 2^{-t} = eps$$

In single precision format $eps = 2^{-23}$. In double precision $eps = 2^{-t} = 2^{-52} \approx 2.2204 \cdot 10^{-16}$.

In $[2^E, 2^{E+1})$ the increment between the consecutive numbers in $\mathbb{F}$ is $\Delta_E = 2^{E-t}$. Since $2^E + 2^t 2^{E-t} = 2^{E+1}$ then there are 2^t fl-numbers in $[2^E, 2^{E+1})$.

The increment $\Delta_E = 2^{E-t}$ doubles from interval $[2^E, 2^{E+1})$ to $[2^{E+1}, 2^{E+2})$, because $\Delta_{E+1} = 2^{E+1-t} = 2 \cdot 2^{E-t}$.

Subnormal numbers.

So far we considered numbers $x = \pm s 2^E$ with $s = (1.f_1 \dots f_t)_2$ which are called normal. This leaves a gap centered around the origin. Subnormal numbers fill the gap and are all evenly spaced. The subnormals are not normalized and are of the format

$$x = \pm f 2^{E_{min}}, \quad f \neq 0$$

The gap between the subnormals is $\Delta = 2^{E_{min}-t}$. The smallest subnormal is $2^{E_{min}-t}$ and the largest $2^{E_{min}-t}(2^t - 1)$. The total amount of positive subnormals is $2^t - 1$. So in single format $(0.1 \dots 0)_2 2^{-126} = 2^{-127}$ is the largest subnormal and $(0.0 \dots 1)_2 2^{-126} = 2^{-23} 2^{-126} = 2^{-149}$ is the smallest subnormal. In double format the smallest subnormal is $2^{-1022-52} = 2^{-1074}$.

Although the range of $\mathbb{F}$ is huge but it is easy to exceed it: $171! \approx 24 \cdot 10^{309}$ is out of range. If $x_0 = 2$, $x_{n+1} = x_n^2$ then for $n \geq 1$ $x_n = 2^{2^n}$ so $x_{10} = 2^{1024}$ which is out of range.

Comparing the fineness of the fl-discretization with the precision to which fundamental constants (Planck constant, gravitational constant, elementary charge) we note that nothing in physics is known to more than 12 digits, thus IEEE numbers are orders of magnitude more precise.

Rounding.

Rounding is an operation which approximates real numbers with suitably chosen nearby floating point numbers. Hence $\mathrm{fl} : \mathbb{R} \to \mathbb{F}$ is a function which we assume satisfies the following requirements

$$\mathrm{fl}(x) = x, \quad x \in \mathbb{F}$$
$$\mathrm{fl}(-x) = -\mathrm{fl}(x), \quad x \in \mathbb{R}$$
$$x_1 \leq x_2 \Rightarrow \mathrm{fl}(x_1) \leq \mathrm{fl}(x_2), \quad x_1, x_2 \in \mathbb{R}$$

Let $x \in [2^E, 2^{E+1})$

$$x = (1.f_1 \dots f_t f_{t+1} \dots)_2 2^E$$

The closest fl number smaller or equal than x is

$$x_- = \max\{y \in \mathbb{F} : y \leq x\} \qquad x_- = (1.f_1 \dots f_t)_2 2^E$$

We define

$$x_+ = \min\{y \in \mathbb{F} : y \geq x\}$$

If $x \notin \mathbb{F}$ then at least one of the bits $f_{t+1}, \ldots$ is nonzero. The closest fl number bigger than x is x_+

$$x_+ = \left((1.f_1 \ldots f_t)_2 + (0.\underbrace{0\ldots01}_{t})_2 \right) 2^E$$

The gap between the two fl-numbers closest to x is $x_+ - x_- = 2^{E-t}$. Denote $\mu = \frac{1}{2}(x_+ - x_-)$.

The standard way of rounding is rounding to the nearest even. Suppose that the significands of x_- and x_+ are given by

$$(1.a_1 \ldots a_t)_2, \qquad (1.b_1 \ldots b_t)_2$$

then exactly one of the digits a_t and b_t is 0 (even).

We define for $x > 0$

$$\mathrm{fl}(x) = \begin{cases} x_- & \text{if} \quad x \in [x_-, \mu) \quad \text{or if} \quad x = \mu \quad \text{and} \quad a_t = 0 \\ x_+ & \text{if} \quad x \in (\mu, x_+] \quad \text{or if} \quad x = \mu \quad \text{and} \quad b_t = 0 \end{cases}$$

Next if $x < 0$ then $\mathrm{fl}(x) = -\mathrm{fl}(-x)$.

Other less precise rounding modes are possible

$$\mathrm{fl}(x) = \begin{cases} x_- & \text{round down} \\ x_+ & \text{round up} \\ \mathrm{sign}(x)|x|_- & \text{round toward } 0 \end{cases}$$

Absolute error in rounding is less than the gap between x_- and x_+ regardless of rounding mode

$$|\mathrm{fl}(x) - x| < 2^{E-t}$$

In round to nearest

$$|\mathrm{fl}(x) - x| < \frac{1}{2} 2^{E-t}$$

If $x = \pm(1.f_1 \ldots f_t f_{t+1} \ldots)_2 2^E$ and $x \in [2^E, 2^{E+1})$ with increment $\Delta_E = 2^{E-t}$ then $|x| \geq 2^E$ and

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq \frac{2^{E-t}}{2^E} = 2^{-t} = eps$$

In round to nearest

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq \frac{\frac{1}{2}2^{E-t}}{2^E} = \frac{1}{2} eps = u$$

In double precision $eps \approx 2.2204 \cdot 10^{-16}$.

Lemma. Let $x \in \mathbb{R}$ be an arbitrary number in the normalized range $x_{min} \leq |x| \leq x_{max}$ of a binary floating point system with precision t. Then rounding to the nearest we obtain

$$\mathrm{fl}(x) = x(1+\delta)$$

for some δ satisfying $|\delta| \leq \frac{1}{2}eps$ where $eps = 2^{-t}$ is the gap between 1 and the next larger floating point number.

Proof. Let $\delta = \frac{\mathrm{fl}(x)-x}{x}$. We then know that $|\delta| \leq \frac{1}{2}eps$. Hence $\mathrm{fl}(x) = \delta x + x$. In other words every number in the normalized range can be represented with a relative error not exceeding the unit roundoff $macheps = \frac{1}{2}eps$.

Floating point arithmetic.

IEEE standard, apart from rounding, provides the correctly rounded arithmetic operations. If $x, y \in \mathbb{F}$ and $\odot \in \{+, -, \cdot, :\}$ then we will denote by $\boxed{\odot}$ the result of $x \odot y$ obtained in the floating point arithmetic. Generally the result of an arithmetic operation on numbers in $\mathbb{F}$ is not a floating point number in $\mathbb{F}$. For example $1, 10 \in \mathbb{F}$ but $1/10 \notin \mathbb{F}$ since $\frac{1}{10} = (0.00011001100\ldots)_2$. Similarly for addition $1, 2^{-53} \in \mathbb{F}$ but $1 + 2^{-53} \notin \mathbb{F}$ however the correctly rounded arithmetic will guarantee that $1 \boxed{:} 10 = \mathrm{fl}(1/10)$. In general we will have that

$$x \boxed{\odot} y = \mathrm{fl}(x \odot y)$$

For $x \odot y$ in the normalized range $x_{min} \leq |x| \leq x_{max}$ we have

$$x \boxed{\odot} y = (x \odot y)(1+\delta)$$

where $|\delta| \leq \frac{1}{2}eps = u$

The reason for producing those slightly perturbed accurate results is that an exact result of the arithmetic operation is normalized, rounded and stored.

Very few of the laws of standard arithmetic are satisfied in floating point arithmetic, most are not. The following operations satisfy standard rules

$$x_1 \boxplus x_2 = x_2 \boxplus x_1 \qquad x_1 \boxdot x_2 = x_2 \boxdot x_1$$

$$x_1 \boxplus (-x_2) = x_1 \boxminus x_2 \qquad x_1 \boxminus (-x_2) = x_1 \boxplus x_2$$

Addition and multiplication are not associative

$$(x_1 \boxplus x_2) \boxplus x_3 \neq x_1 \boxplus (x_2 \boxplus x_3)$$

Distributive laws

$$x \boxdot (y \boxplus z) = (x \boxdot y) \boxplus (x \boxdot z)$$

all fail in general. Multiplication and division are not inverse operations

$$(x_2 \boxdot x_1) \boxed{:} x_1 \neq x_2 \neq (x_2 \boxed{:} x_1) \boxdot x_1$$

Finally addition and subtraction are not distributive with multiplication.

Lemma. (Absorption property) Let $x, y \in \mathbb{F}$, $x > y > 0$. If $y < \frac{1}{4}2^{-t}x$ then

$$\mathrm{fl}(x+y) = x$$

Proof. Let $x = s2^E$, $1 \leq s < 2$. The next floating point number larger than x is $x + \Delta_E = s2^E + 2^{E-t}$. The midpoint μ in $[x, x+\Delta_E]$ is $x + \frac{1}{2}2^{E-t}$. If $x+y$ is smaller than midpoint then (according to rounding to nearest) $x+y$ is rounded down to x. Based on the assumed bound on y and the fact that $s < 2$

$$x + y < x + \frac{1}{4}2^{-t}s2^E < x + \frac{1}{2}2^{E-t} = \mu$$

hence $x+y$ is smaller then μ. Hence rounding down causes the absorption of small number y by large number x.

Suppose that we want to compute approximately $f'(1)$ from the definition

$$f'(1) \approx \frac{f(1+h) - f(1)}{h}$$

Computing the divided quotient

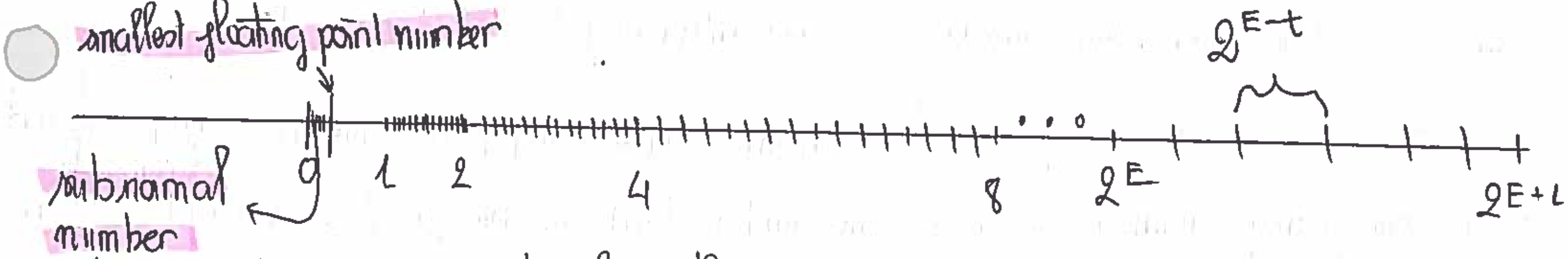
$$\mathrm{fl}\left(\frac{f(1+h) - f(1)}{h}\right) = (f(1 \boxplus h) - f(1)) \boxed{:} h$$

Due to absorption $1 \boxplus h = 1$ for very small h and independent of what f is we obtain $f'(1) \approx 0$.

M. Overton, Numerical Computing with IEEE floating point arithmetic, Tucker, Validated numerics

* Floating point number system .

IEEE754-1985 Standard for Binary floating point



- the ammont of ~~number~~ interval are the same.

* We consider numbers of the form

$$x = \pm s\, 2^E = \pm (1+f)\, 2^{\overset{}{e-\underset{\uparrow \text{bias}}{b}}} = \pm (1.f_1 f_2 \dots f_t)_2\, 2^{e_{l-1} e_{l-2} \dots e_0 - \text{bias}}$$

$$= (1-2p)(1.f_1 f_2 \dots f_t)_2\, 2^{e_{l-1} e_{l-2} \dots e_0 - \text{bias}} .$$

- s : significant / mantisa $1 \leq s < 2$

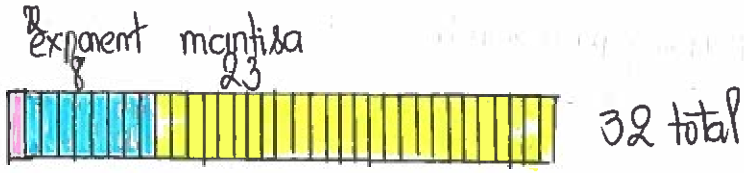
E : exponent $E \in \mathbb{Z}$

$$f = (0.f_1 f_2 \dots f_t)_2 = \sum_{i=1}^{t} f_i\, 2^{-i} = f_1 2^{-1} + f_2 2^{-2} + \dots + f_t 2^{-t}$$

$$e = e_{l-1} e_{l-2} \dots e_0 = e_0 2^0 + e_1 2^1 + \dots + e_{l-1} 2^{l-1} .$$

$1-2p$, $p \in \{0,1\}$: sign of x

* Floating point component .

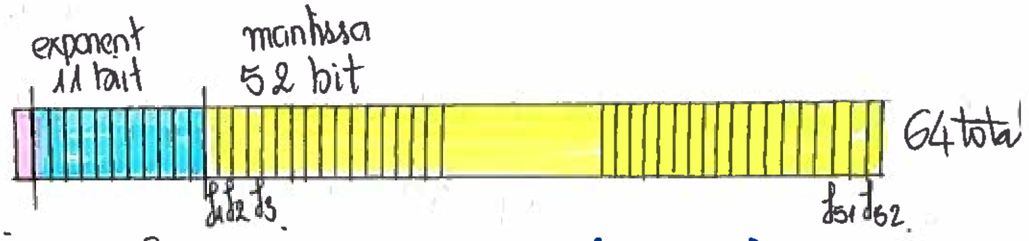


- Single precision

sign 1 exponent 8 fraction 23
$l = 8$ $t = 23$.

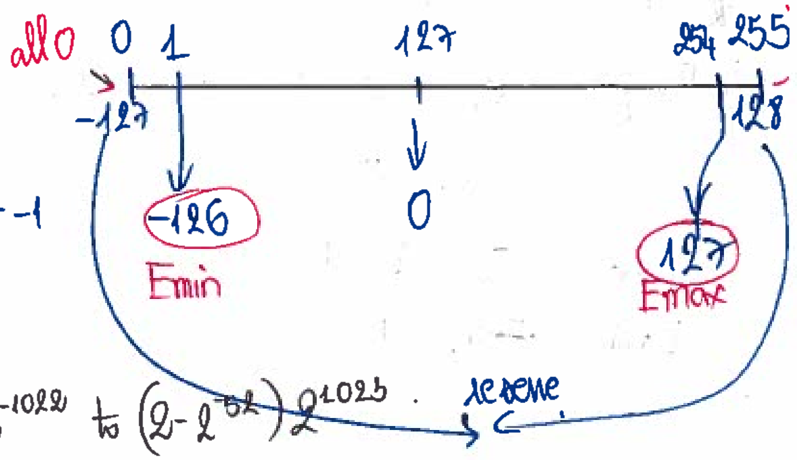
- Double precision .

sign 1 exponent 11 mantissa 52
$l = 11$ $t = 52$



	Single precision	Double precision
t (digit of mantisa)	23	52
l exponent	8	11
Emax	$127 = \frac{2^8}{2} - 1$	$1023 = \frac{2^{11}}{2} - 1$
Emin	$-126 = -b+1$	-1022
bias .	127	1023 .
Range	$\pm 2^{-126}$ to $(2-2^{-23})\, 2^{127}$	$\pm 2^{-1022}$ to $(2-2^{-52})\, 2^{1023}$.

$2^8 = 256 \rightarrow (0 \rightarrow 255)$.



* Only a finite number of rational numbers are in $\mathbb{F}$

↳ The increment between two consecutive number in $[2^E, 2^{E+1})$ is 2^{E-t}

↳ There are 2^t floating point numbers in each interval $[2^E, 2^{E+1})$ at $E_{min} \le E \le E_{max}$

* $x = \pm s\, 2^E = \pm 1.f_1 \dots f_t\, 2^{e-bias}$: normalized floating point numbers ⇒ leave a gap centered around the origin.

x is a subnormal number ⟺ are some number that are bigger than 0 and smaller than $1.0\, 2^{E_{min}}$

fill the gap and are evenly space

is a nonzero floating point number ⟺ $x \pm 0.f_1 \dots f_t\, 2^{E_{min}} = \pm 0.f_1 \dots f_t\, 2^{-1022} = \pm f 2^{E_m}$
$= \pm 0.f\, 2^{-1022}$

- The gap between subnormal numbers is $2^{E_m - t} = 2^{-1022-52} = 2^{-}$

* Some facts about normals

- The smallest subnormal is $2^{E_{min}-t} = 2$
  The largest subnormal is $2^{E_m - t}(2^t - 1)$

- There are $\underset{\substack{\uparrow\\ 2\text{ sides}}}{2}\ \underset{\substack{\uparrow\\ \text{in each interval}}}{2^t}\ [E_{max} - \underset{\substack{\uparrow\\ \#\text{ of mantissa}}}{E_{min}} + 1] + \underset{\substack{\uparrow\\ 0}}{1}$ numbers in $\mathbb{F}$

- The total amount of subnormal is $2^t - 1$.

- Let $x \in \mathbb{F} \cap [2^E, 2^{E+1})$
$= (1+f)\, 2^E$ assume $1 + f < 2 - 2^{-t}$?

$\hat{x} \leftarrow$ the next floating point number that bigger than x.

$$x = (1 + f + 2^{-t})\, 2^E = (1+f)\, 2^E + \underbrace{2^{E-t}}_{\text{gap between numbers in the } [2^E, 2^{E+1}) \text{ interval}} = x + \Delta_E$$

- The gap between number in $[1, 2) = [2^0, 2^{0+1})$ is
$\Delta_0 = 2^{-t} = 2^{-52}$ = epsilon procedure

* $x = \pm s\, 2^E$

The largest $e_{max} = 2^8 - 1 = 2^8 - 1 = 254$

$e_{min} = 0$

$E_{max} = e_{max} - bias = 255 - 127 = 128$

$E_{min} = 0 - bias = -127$ — reserve

⇒ $E_{max} = e_{max} - b - 1$

$E_{min} = \cancel{E} b + 1$

* Exponent encoding (of Double precision)

$e = 00000000001_2$ $2^{1-1023} = 2^{-1022}$: smallest exponent for normal numbers.

$e = 01111111111_2$ $2^{1023-1023} = 2^0$ (zero offset).

$e = 11111111110_2$ $2^{2046-1023} = 2^{1023}$ highest exponent

↑ because the e with all 1 are for reservation

*Double precision example

$0\ 01111111111\ \underbrace{000000\ldots000}_{52} = 2^0 \cdot (1) = 1.$

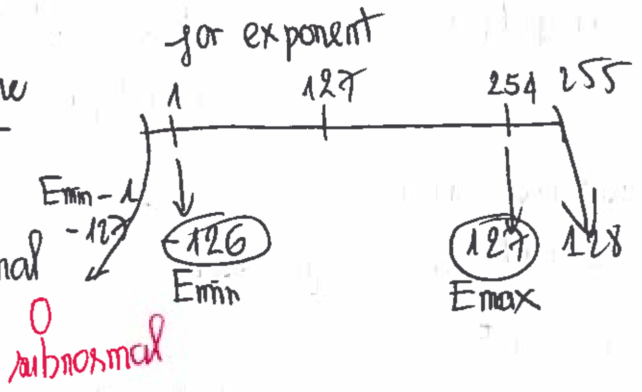
exponent   mantissa.

$0\ 01111111111\ 00\ldots\ldots 0001 = 2^0(1 + 2^{-52}) = 1.\underbrace{0000\ldots000}_{15}2$ smallest number > 1.

$0\ 10000000000\ 000000\ldots000000 = 2^1 \cdot 1 = 2$

*

Exponent	mantissa	numerical value
$E = E_{min} - 1$	$f = 0$	± 0
$E = E_{min} - 1$	$\ell = 0$ $f \neq 0$	$\pm(0.f_2)\,2^{E_{min}}$ subnormal
$E_{min} \leq E \leq E_{max}$	any f	$\pm(1.f)_2$ normal
$E = E_{max} + 1$	$f = 0$	$\pm\infty$ $\frac{1}{0}$, $-\frac{1}{0}$
$E = E_{max} + 1$	$f \neq 0$	NaN $\sqrt{-1}$ $\frac{0}{0}$



* Rounding $\mathbb{R} \xrightarrow{fl} \mathbb{F}$

$$*\begin{cases} fl(x) = x, & x \in \mathbb{F} \\ fl(-x) = -fl(x), & \\ x_1 \leq x_2 & \Rightarrow fl(x_1) \leq fl(x_2) \end{cases}$$

- Let $\underset{\text{can be real}}{x} \in [2^E, 2^{E+1})$ $\underset{\text{can be real}}{x} = (1.f_1 \dots f_t\, f_{t+1} \dots)_2\, 2^E$
- The closed number $\leq x$, denoted by $x_- = \max\{y \in \mathbb{F}, y \leq x\}$.

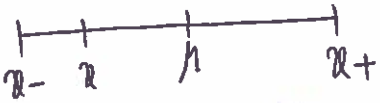
$x_- = (1.f_1 f_2 \dots f_t)_2\, 2^E$

The closed number $\geq x$, denoted by $x_+ = \min\{y \in F, y \geq x\}$.

$x_+ = \{1.f_1 \dots f_t + 2^{-t}\}\, 2^E$.

- Then $x_+ - x_- = 2^{E-t}$ ← when E is bigger, the increment between 2 floating point numbers is bigger

Let $\mu = \frac{1}{2}(x_+ + x_-)$



Then $fl(x)$ round to $x-$ or $x+$ to the nearest one

- Suppose that the significance of $x = 1.(a_1 \dots a_t)_2$ and $x_+ = 1(b_1 \dots b_t)_2$ then we have exactly one of the digit a_t and b_t is 0

$$\boxed{\text{Then } fl(x) = \begin{cases} x_- & \text{if } x \in [x_-, \mu) \text{ or } (x = \mu \text{ and } a_t = 0) \\ x_+ & \text{if } x \in (\mu, x_+] \text{ or } (x = \mu \text{ and } b_t = 0) \end{cases}}$$ (Rounded to the "nearest, even"

* Relative error

$x = \pm(1.f_1 f_2, \dots f_t\, f_{t+1} \dots)\, 2^E$ $x \in (2^E, 2^{E+1})$ $\Delta_E = 2^{E-t}$

$\Rightarrow |x| \geq 2^E$

$$\text{Relative error} = \delta = \frac{|fl(x) - x|}{|x|} \leq \frac{\frac{1}{2}\Delta_E}{|x|} = \frac{\frac{1}{2}2^{E-t}}{2^E} = \frac{1}{2}\underset{\text{machine epsilon}}{e^{-t}} = \frac{1}{2}\varepsilon$$

* Lemma

Let $x \in \mathbb{R}$ in normalized range $|x|_{min} \leq |x| \leq x_{max}$, then rounded to the "nearest even", $fl(x) = x(1 + \underset{\substack{\text{relative rounded error} \\ \text{(can be positive, negative}}}{\delta})$ $\delta \leq \frac{1}{2}\varepsilon = \frac{1}{2}2^{-t}$

$\delta|x| = |fl(x) - x| \Rightarrow fl(x) = x(1+\delta)$

* Lemma:

Let $x \in \mathbb{R}$, in normalized range

$|x|_{min} \leq |x| \leq x_{max}$

then rounded to "nearest even" fl, $fl(x) = x(1+\delta)$, where $\delta \leq \frac{1}{2} eps = \frac{1}{2} 2^{-t}$

relative rounded error

- Proof : from above

$\delta = \frac{|fl(x)-x|}{|x|} < \frac{1}{2} eps$ ~~$fl(x) = x(1+\delta) = x + \delta x$~~

$\delta x = fl(x) - x \Rightarrow fl(x) = x + \delta(x) = x(1+\delta)$

* Floating point arithmetic

$\odot \in \{+, -, \cdot, :\}$ denote $\boxed{\odot}$

very often, the result of an arithmetic operation on 2 floating point number is not a floating point number.

If $x, y \in \mathbb{F}$, $\odot$ is an arithmetic } the following happens : $x \boxed{\odot} y = fl(x \odot y)$

Ex: $x = 1$ $y = 10$ $\odot = :$ then $\frac{x}{y} = \frac{1}{10} \to f\left(\frac{1}{10}\right)$

- $x_1 \boxplus x_2 = x_2 \boxplus x_1$

$x_1 \boxdot x_2 = x_2 \boxdot x_1$

$x_1 \boxplus (-x_2) = x_1 \boxminus x_2$

$x_1 \underset{\boxdot}{\boxplus} (x_2 \underset{\boxdot}{\boxplus} x_3) \neq (x_1 \underset{\boxdot}{\boxplus} x_2) \underset{\boxdot}{\boxplus} x_3$.

the distributive laws do not hold

* Lemma (Absorption property)

$x, y \in \mathbb{F}$, $x > y > 0$

If $y < \frac{1}{4} 2^{-t} x$

} Then $fl(x+y) = x$

* Consider an ex

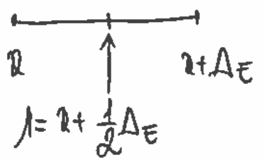
$f'(1) = \frac{fl(1+h) - f(1)}{h}$

- Proof:

Let $x = s 2^E$, $1 \leq s < 2$

The next bigger than x number in $\mathbb{F}$ is $x + 2^{E-t} = s 2^E + 2^{E-t} =$

$= 2^E (s + 2^{-t})$



$\Delta_E = 2^{E-t}$

- If $x + y < \mu \Rightarrow (x+y)$ is rounded to x

$x + y = s 2^E + 2^{E-t} + \frac{1}{4} 2^{-t} x \leq s 2^E + 2^{E-t} + \frac{1}{4} 2^{-t} s 2^E$

$< x + \frac{1}{2} 2^{E-t} = \mu$

$(x + y < x + \frac{1}{4} 2^{-t} s 2^E \overset{x2}{<} x + \frac{1}{4} 2^{-t} 2 . 2^F$

$= x + \frac{1}{2} 2^{E-t} = \mu$.

$= x + \frac{1}{2} \Delta_E = \mu$.

$\Rightarrow fl(x+y) = x$.

* Example

~~fl~~ $f'(1) \approx \dfrac{f(1+h) - f(1)}{h}$

$$fl\left(\frac{f(1+h)-f(1)}{h}\right) = \frac{f(fl(1+h)) - f(1)}{h} = \frac{f(1)-f(1)}{h} = 0$$ ← not a happy computation

* Example : two mathematically equivalent algorithms are inequal not num. equivalent

$A_1(a,b) = a^2 - b^2$

$A_2(a,b) = (a+b)\,(a-b)$ $\qquad$ A_2 is better. $\qquad$ cancelation error

- $fl(a^2-b^2) = \left(a^2(1+\varepsilon_1) - b^2(1+\varepsilon_2)\right)(1+\varepsilon_3)$ $\qquad$ $\varepsilon_1 < \frac{1}{2}eps$

$$= (a^2-b^2)\left[\frac{a^2(1+\varepsilon_1) - b^2(1+\varepsilon_2)}{a^2-b^2}\right](1+\varepsilon_3) = (a^2-b^2)\underbrace{\left[1+\frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2}\right]}_{\substack{\text{if } a \approx b \\ \varepsilon_1 \text{ and } \varepsilon_2 \text{ have opposite sign}}}(1+\varepsilon_3)$$

→ the can be too big.

- $fl((a-b)(a+b)) = \cancel{fl}\left((a-b)(1+\varepsilon_1)(a+b)(1+\varepsilon_2)\right)(1+\varepsilon_3)$

$$= \cancel{fl}(a^2-b^2)(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)$$

$$= (a^2-b^2)\left(1+\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3\right)$$

$$\leq (a^2-b^2)\left(1+3\tfrac{1}{2}eps + O(2^{-2t})\right)$$

next class polynom intepolation

* Polynomial computations $p(x) := a_0 + a_1 x + a_2 x^2 + \dots + a_n x^n$

* What we want to do? : Given p ⇔ given $a_0, a_1, \dots, a_n$, we want to evaluate $p(\overset{\downarrow}{x})$

* Example : simple evaluation of a monomial x^{15}

- way 1 : $\underset{x^n}{x^{15}} = x\underbrace{(x \dots x)}_{14}$ → have to compute $x^2, x^3, x^4, \dots, x^{15}$ ⇒ $\overset{(n-1)}{14}$ muls.
- binary method : $x^2, x^3, x^6, x^7, x^{14}, x^{15}$ ⇒ 6 muls.
- factorization method : 15 = 3×5 ⇒ evaluate $y = x^3 \to 2$ muls.

$x^{15} = y^5$ → use binary method for y
→ compute $y^2, y^4, y^5 \to 3$ muls. } ⇒ 5mu

* Evalute p(x).

- Direct way : compute x^i → (n−1) muls
  compute $a_i x^i$ → n muls.
  add them all → n adds
  } ⇒ (3n − 1) flops
  (we may have to store partial results)

- Horner's algorithm :
  $p(x) = ((\dots(a_n x + a_{n-1})x + \dots + a_1)x) + a_0$

n multiplications
n adds
} ⇒ 2n flops.

⊕ Depends on the polynomial that we want to evaluate
⊕ Horner's algorithm is optimal in normal circumstance
⊕ an algorithm is good if it is stable + solutions converge to the true solution

⊕ $b_n = a_n$
$b_{n-1} = b_n x + a_{n-1}$
$b_{n-2} = b_{n-1}x + a_{n-2}$
⋮
$b_i = b_{i+1}x + a_i$
⋮
$p(x) = b_0 = b_1 x + a_0$
} n adds.

* Explaination :

$$p(y) = a_n y^n + a_{n-1}y^{n-1} + \dots + a_1 y + a_0$$
$$= (b_n y^{n-1} + \dots + b_2 y + b_1)(y - x) + b_0$$

$$\Rightarrow \begin{cases} a_n = b_n \\ a_{n-1} = -b_n x + b_{n-1} \\ \vdots \\ a_1 = -b_2 x + b_1 \\ a_0 = -b_1 x + b_0 \end{cases} \quad \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} = \begin{bmatrix} 1 & -x & & & \\ & 1 & -x & & \\ & & \ddots & \ddots & \\ & & & 1 & -x \\ & & & & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

* Division a polynomial by $(x - \alpha)$: we want $p(x) = Q_\alpha(x)(x - \alpha) + r$.

	a_n	a_{n-1}	- - -	a_1	a_0
α	b_n	b_{n-1}	- - -	b_1	r

⇒ find a number in the second line = sum of
= number above it + α · number on the left.

* Example $p(x) = x^3 - 4x^2 + 3x + 2$ compute p(3)

	1	−4	3	2
3	1	−1	0	2
	b_3	b_2	b_1	b_0

then since $Q_\alpha(x)(x-3) + r$ ⇒ $p(3) = r = 2$.
$(x^2 - x + 0)$

* $(*)\ p(x) = a_n x^n + \cdots + a_1 x + a_0$

$(**)\ p(x) = c_n (x-p)^n + \cdots + c_1(x-p) + c_0 \leftarrow$ an expansion in base p.

Some time we want to convert $(*) \Leftrightarrow (**)$

* Preconditioning (Review : we want to solve $Ax = b \Leftrightarrow \underbrace{MA}_{I}x = Mb$ $\quad A^{-1} \approx M$.

$= Mb$

* Adaption of coefficients :

Let $p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$

- When we use Horner $\Rightarrow$ $\left.\begin{matrix} 4 \text{ muls} \\ 4 \text{ adds} \end{matrix}\right\} \Rightarrow$ 8 flops to compute $p(x)$.

- Can we ~~use~~ "adapt" the coefficients a_k so that the operation count is lower ?

⊕ Find $\alpha_0, \ldots, \alpha_n$ such that if $y = (x+\alpha_0)x + \alpha_1$ then $p(x) = [(y + x + \alpha_2)y + \alpha_3]\alpha_4$.

$= \alpha_4(y^2 + xy + \alpha_2 y + \alpha_3)$

$\Updownarrow$

$$y^2 = (x^2 + 2\alpha_0 x + \alpha_0^2)x^2 + \alpha_1^2 + 2\alpha_1(x+\alpha_0)x.$$

⊕ Compare coefficients :

- at x^4, $a_4 = \alpha_4$

at x^3, $a_3 = \alpha_0(2\alpha_0) + \alpha_4$ $\quad \Rightarrow \alpha_0 = \frac{1}{2}\left(\frac{a_3}{\alpha_4} - 1\right)$

at x^2, $a_2 = \alpha_4(\alpha_0^2 + 2\alpha_1 + \alpha_0 + \alpha_2) \Rightarrow \left.\begin{matrix} \frac{a_2}{a_4} = \alpha_0 + \alpha_0 + 2\alpha_1 + \alpha_2 \\ \beta := \frac{a_2}{a_4} - (\alpha_0)^2 + \alpha_0 \end{matrix}\right\} \Rightarrow \alpha_2 = \beta - 2\alpha_1$

at x, $\alpha_1 = \frac{a_1}{a_4} - \alpha_0 \beta$

$\alpha_3 = \frac{a_0}{a_4} - \alpha_1(\alpha_1 + \alpha_2)$.

preconditioning.

⊕ Then after preconditioning, we have the operators that we have to do is

$y = \underbrace{\underbrace{(x+\alpha_0)}_{1add} x}_{2 mul} + \alpha_1$ $\qquad p(x) = [(y + x + \alpha_2)y + \alpha_3]\alpha_4$. Totally $\Rightarrow \left\{\begin{matrix} 3 \text{ mul} \\ 5 \text{ adds}. \end{matrix}\right.$

1add $\qquad$ 2muls 3ads

- By Horner $\left\{\begin{matrix} 4 \text{ muls} \\ 4 \text{ adds} \end{matrix}\right.$ $\qquad$ - By adaption $\left|\begin{matrix} 3 \text{ mul} \\ 5 \text{ adds}. \end{matrix}\right.$

The total operations have to do are the same

But when mul is more expensive $\Rightarrow$ adaption method is better □

May

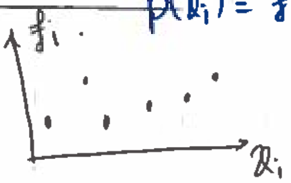
* Chapter 6: Interpolation

6.1 Polynomial interpolation

Given ~~[illegible]~~ distinct numbers

x	x_0	x_1	x_2	x_3	$\cdots$	x_n
value f	f_0	f_1	f_2	f_3	$\cdots$	f_n

⟹ we want to seek a polynomial p of lowest possible degree so that $p(x_i) = f_i, \forall i = \overline{0,n}$



* Theorem on polynomial interpolation

Given $(n+1)$ distinct numbers $x_0, x_1, \dots, x_n$
and numbers $f_0, f_1, \dots, f_n$

There exists a unique polynomial L so that $L(x_i) = f_i, \forall i = \overline{0,n}$

a polynomial of degree at most n

* Proof the existence:

- Consider polynomials $l_0, l_1, \dots, l_n$; $l_i \in \mathbb{P}_n$, $l_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

- We have if a polynomial $p(x) = (x - x_0) q(x) \Leftrightarrow p(x_0) = 0$

If $\left.\begin{matrix} p(x) = (x - x_0) q(x) + r \\ r = 0 \end{matrix}\right\} \Rightarrow (x - x_0) \mid p(x) \Rightarrow p(x)$ has a root at x_0.

- Put
$$l_i(x) = \frac{(x - x_0)(x - x_1) \dots (x - x_{i-1}) \cancel{(x - x_i)} (x - x_{i+1}) \cdots (x - x_n)}{(x_i - x_0)(x_i - x_1) \cdots (x_i - x_{i-1})(x_i - x_{i+1}) \cdots (x_i - x_n)} = \prod_{\substack{l=0 \\ l \neq i}}^{n} \frac{(x - x_l)}{(x_i - x_l)}$$

- Define
$$L(x) := \sum_{i=0}^{n} l_i(x) f_i = l_0(x) f_0 + l_1(x) f_1 + \dots + l_n(x) f_n = \sum_{i=1}^{n} \left[f_i \left(\prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j} \right) \right]$$

Canonical form of Lagrange interpolation

since $L(x_i) = 0 + \dots + 0 + l_i(x_i) f_i + 0 \dots + 0 = 1 f_i = f_i$

Then $L(x)$ is the polynomial that satisfies $L(x_i) = f_i$ ⇒ □ existence.

* Prove the uniqueness

- Assume $\exists L_1(x)$ and $L_2(x)$ that satisfy $\in \mathbb{P}_n(x)$ and $\begin{matrix} L_1(x_i) = f_i \\ L_2(x_i) = f_i \end{matrix}$, $\forall i = \overline{1,n}$

$\Rightarrow L_1(x_i) - L_2(x_i) = 0$

$\Rightarrow (L_1 - L_2) x_i = 0, \forall i = \overline{0,n}$

Since $x_0 \to x_n$ are all distinct ⟹ $(L_1 - L_2)(x)$ has $(n+1)$ zeros. (1)

- Since $(L_1 - L_2) \in \mathbb{P}_n$ ⇒ the degree is at most n ⟹ if $(L_1 - L_2)$ is not a zero polynomial it has to have at most n zeros (2)

$\Rightarrow (L_1 - L_2)(x)$ is a zero polynomial

$\Rightarrow L_1(x) = L_2(x)$ ⇒ □ uniqueness.

*Def

Given $x_0 \quad x_1 \quad \dots \quad x_n$

$f_0 \quad f_1 \dots \quad f_n$

If $f(x_i) = f_i$

and L is a polynomial that $L(x_i) = f_i = f(x_i)$ $\Big\}$ ⇒ L is called an interpolation of f

easy to determine , hard to compute .

* From above

$$L(x) = \sum_{i=1}^{n} f_i \, l_i(x) = \sum_{i=1}^{n} \boxed{f_i \left(\prod_{\substack{j=1 \\ i \neq j}}^{n} \frac{(x-x_j)}{(x_i-x_j)} \right)} = \sum_{i=1}^{n} c_i (x - x_j) \; ;$$

* We want to rewrite $L(x)$ in monomial basic form

hard to determine

easy to compute .

page 313 .

$$L(x) = a_0 + a_1 x + a_2 x^2 + \dots + a_{n-1} x^{n-1} + a_n x^n$$

(a monomial basic $1, x, \dots, x^n$) .

- we have $l_i(x) = \prod_{\substack{j=1 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$

→ question: can a_i be computed from x_i, f_i ? → easy to determine / → hard to compute with .

- We have

$$\begin{cases} L(x_0) = f_0 \Leftrightarrow a_0 + a_1 x_0 + a_2 x_0^2 + \dots + a_n x_0^n = f_0 \\ L(x_1) = f_1 \Leftrightarrow a_0 + a_1 x_1 + a_2 x_1^2 + \dots + a_n x_1^n = f_1 \\ \vdots \\ L(x_n) = f_n \Leftrightarrow a_0 + a_1 x_n + \dots + a_n x_n^n = f_n \end{cases}$$

$$\Rightarrow \underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \dots & x_n^n \end{bmatrix}}_{A:\text{ Vandermonde matrix}} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{bmatrix}$$

this matrix is a nonsingular matrix since the system has a unique solution (by the above theorem,

⇒ $\det(A) \neq 0$, but **A** is often ill conditioned

⇒ $a_i, i = \overline{0,n}$ are inaccurately determined by solving the above system

Newton form of the interpolation polynomial

My

* Given a set of $(n+1)$ data numbers $x_0, x_1, \dots, x_n$ ← distinct

$y_0, y_1, \dots, y_n$

Then the Newton interpolation polynomial is the linear combination of Newton basic polynomials.

$$L_n(x) := \sum_{j=0}^{n} a_j \Pi_j(x) = a_0 1 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \dots + a_n(x-x_0)(x-x_1)\dots(x-x_{n-1})$$

Where Newton basic polynomials are

$$\left.\begin{array}{l} \boxed{\Pi_0(x) := 1} \\ \Pi_1(x) := (x-x_0) \\ \Pi_2(x) := (x-x_0)(x-x_1) \\ \vdots \\ \Pi_n(x) := (x-x_0)(x-x_1)\dots(x-x_{n-1}) \end{array}\right\}$$

they are a basic in $\mathbb{P}^n$ since they are linear independent

$L_n(x) \in \mathbb{P}_n$

$L_{n-1}(x) \in \mathbb{P}_{n-1}$.

Assume that $L_{l-1}(x_i) = y_i$, for $0 \le i \le l-1$

- Now consider

$$L_{0,\dots,l}(x) := \sum_{j=0}^{l} a_j \Pi_j(x) = \sum_{j=0}^{l-1} a_j \Pi_j(x) + a_l \Pi_l(x)$$

$$= L_{0,\dots,l-1}(x) + a_l \Pi_l(x) = L_{l-1}(x) + a_l (x-x_0)(x-x_1)\dots(x-x_{l-1}) \quad (1)$$

$\Rightarrow L_l(x_i) = y_i, \ \forall i = \overline{0, l-1}$

$\Rightarrow \left(L_{(0,\dots,l)}(x_i) - L_{(0,\dots,l-1)}\right)(x_i) = 0$

From (1), we have $L_l(x) - L_{l-1}(x) = a_l \Pi_l(x)$

$$L_{0,\dots,n}(x) = \sum_{l=0}^{n} f[x_0,\dots,x_l]\Pi_l(x)$$

a_l equals to the leading coefficient of L_l

$$a_l = \sum_{i=0}^{l} \frac{f_i}{\prod_{\substack{j=0 \\ j\neq l}}^{l} (x_i - x_j)}$$

$\parallel$

$f[x_0,\dots,x_l]$

↑

l^{th} order devided difference of f

$L_0(x) = f(x_0)$

$L_{0,\dots,l}(x) = L_{0,\dots,l-1}(x) + f[x_0,\dots,x_l]\Pi_l(x)$.

$$L_{0,\dots,n}(x) = f(x_0) + f[x_0,x_1](x-x_0) + f[x_0,x_1,x_2](x-x_0)(x-x_1) + \dots + \dots + f[x_0,x_1,\dots,x_n](x-x_0)(x-x_1)\dots(x-x_{n-1})$$

$$= f(x_0) + f[x_0,x_1]\Pi_1(x) + \dots f[x_0,x_1,x_2]\Pi_2(x) + \dots + \dots + f[x_0,x_1,\dots,x_n]\Pi_n(x).$$

* Theorem (Properties of divided difference)

a> Linearity

If $f(x) = \alpha g(x) + \beta h(x)$

Then $f[x_0, \dots, x_n] = \alpha g[x_0, \dots, x_n] + \beta h[x_0, \dots, x_n]$

b> Commutativity

$f[x_0, \dots, x_l] = f[x_{\sigma(0)}, \dots, x_{\sigma(l)}]$ ← depends on the notes, but not order of the notes commuting then yeild s the same polynomial

c> Recurrence formula

$$f[x_0, \dots, x_l] = \frac{f[x_1, \dots, x_l] - f[x_0, \dots, x_{l-1}]}{x_l - x_0}$$

* Let $L_{0,\dots,l-1} \in \mathbb{P}_{l-1}$, $L_{1,\dots,l} \in \mathbb{P}_l$ be interpolations of f at appropriate nodes

Define $g \in \mathbb{P}_l$.

$$g(x) = \frac{(x - x_0) L_{1,\dots,l}(x) - (x - x_l) L_{0,\dots,l-1}(x)}{x_l - x_0}$$

Check

$g(x) = L_{0,\dots,l}$.

$$g(x_i) = \frac{(x_i - x_0) f_i - (x_i - x_l) f_i}{x_l - x_0} = \frac{f_i(x_i - x_0 - x_i + x_l)}{x_l - x_0} = f_i.$$

$f_i = f(x_i)$.

$$\begin{tikzcd}
x_0 & f_0 \arrow[dr, dash] & & & & \\
x_1 & f_1 \arrow[r, dash] \arrow[dr, dash] & f[x_0, x_1] \arrow[dr, dash] & & & \\
x_2 & f_2 \arrow[r, dash] \arrow[dr, dash] & f[x_1, x_2] \arrow[r, dash] \arrow[dr, dash] & f[x_0, x_1, x_2] \arrow[dr, dash] & & \\
x_3 & f_3 \arrow[r, dash] \arrow[dr, dash] & f[x_2, x_3] \arrow[r, dash] \arrow[dr, dash] & f[x_1, x_2, x_3] \arrow[r, dash] \arrow[dr, dash] & f[x_0, x_1, x_2, x_3] \arrow[dr, dash] & \\
x_4 & f_4 \arrow[r, dash] & f[x_3, x_4] \arrow[r, dash] & f[x_2, x_3, x_4] \arrow[r, dash] & f[x_0, x_1, x_2, x_3] \arrow[r, dash] & f[x_0, x_1, x_2, x_3, x_4]
\end{tikzcd}$$

* Next class

$f(x) - f(x_0) = f'(\xi)(x - x_0)$ (MValue theorem)

$f(x) = f(x_0) + f'(\xi)(x - x_0)$ Lagrange interpolation polynomial

Will prove Lagrange interpolation remainder

* Hemiten interpolation

* with Lagrange intepolation;

Find $L \in \boxed{\mathbb{P}^1}$ such that $L(x_0) = 0 \quad L(x_1) = 1$

Then $L(x) = \dfrac{x - x_0}{x_1 - x_0}$

* For Hermite interpolation:

- Find $h_{0,1} \in \boxed{\mathbb{P}^3}$ such that $\begin{cases} h_{0,1}^{(0)}(x_0) = 0 & h_{0,1}^{(1)}(x_0) = 1 \\ h_{0,1}^{(0)}(x_1) = 0 & h_{0,1}^{(1)}(x_1) = 0 \end{cases}$

Then $h_{0,1}(x) = (x - x_0)\left(\dfrac{x - x_1}{x_0 - x_1}\right)^2$

- Find $h_{0,0}(x)$ such that $\begin{cases} h_{0,0}^{(0)}(x_0) = 1 & h_{0,0}^{(1)}(x_0) = 0 \\ h_{0,0}^{(0)}(x_1) = 0 & h_{0,0}^{(1)}(x_1) = 0 \end{cases}$

Then $h_{0,0}(x) = \underbrace{L_{0,0}(x)}_{\in \mathbb{R}_2} - L_{00}^{(1)}(x_0) \underbrace{h_{0,1}(x)}_{\text{degree } 3}$

$$= \left(\frac{x - x_1}{x_0 - x_1}\right)^2 - \frac{2}{(x_0 - x_1)}(x - x_0)\left(\frac{x - x_1}{x_0 - x_1}\right)^2 = \frac{(x - x_1)^2(2x + x_1 - 3x_0)}{(x_1 - x_0)^3}$$

- $h_{1,1}(x) = (x - x_1)\left(\dfrac{x - x_0}{x_1 - x_0}\right)^2$

$h_{1,0}(x) = \dfrac{(x - x_0)^2(2x + x_0 - 3x_1)}{(x_0 - x_1)^2}$

* 2 points Hermite interpolation form

- Find $H \in \mathbb{P}^3$ s.t $\begin{cases} H^{(0)}(x_0) = f(x_0) & H^{(1)}(x_0) = f'(x_0) \\ H^{(0)}(x_1) = f(x_1) & H'(x_1) = f'(x_1) \end{cases}$ $= \sum_{i=0}^{k} f(x_i) h_{i,0}(x) + \sum_{i=0}^{k} f'(x_i) h_{i,1}(x) + \sum_{i=0}^{k} f''(x_i) h_{i,2}(x)$

Then $H(x) = f(x_0) h_{0,0}(x) + f'(x_0) h_{0,1}(x) + f(x_1) h_{1,0}(x) + f'(x_1) h_{1,1}(x)$

- $H(x) = f(x_0) + f[x_0, 2](x - x_0) + f[x_0, 2; x_1](x - x_0)^2 + f[x_0, 2; x_1, 2](x - x_0)^2(x - x_1)$

$f[x_0, 2] := \dfrac{f^{(1)}(x_0)}{1!}$ | $f[x_0, 2; x_1] = \dfrac{f[x_0, x_1] - f[x_0, 2]}{x_1 - x_0}$

$f[x_0, k] = \dfrac{f^{(k-1)}(x_0)}{(k-1)!}$ | $f[x_0, 2; x_1, 2] = \dfrac{f[x_0, x_1; 2] - f[x_0, 2; x_1]}{x_1 - x_0}$

x_0	$f(x_0)$			
x_0	$f(x_0)$	$f^{(1)}(x_0)$		
x_1	$f(x_1)$	$f[x_0, x_1]$	$f[x_0, 2; x_1]$	
x_1	$f(x_1)$	$f^{(1)}(x_1)$	$f[x_0, x_1; 2]$	$f[x_0, 2; x_1, 2]$

* Theorem 1.

Let $x_0, x_1, \dots, x_k$ be distinct numbers.
Let $m_0, m_1, \dots, m_k$ be intergers, $m_i \geq 1$, $\boxed{\sum_{i=0}^{k} m_i = n+1}$
Let f be a function such that $f^{(m_i-1)}(x_i)$ exist, $i=\overline{1,k}$

$\Bigg\}$ Then there exist a unique polynomial $H \in \mathbb{P}_n$ such that

$$H^{(\ell)}(x_i) = f^{(\ell)}(x_i) \qquad i=\overline{1,k}, \quad \ell=\overline{0,m_i-1}$$

(Note that when $m_i = 1, \forall i = \overline{1,k}$ $\Rightarrow$ we have Lagrange interpolation.

* Theorem 2

Where $\boxed{h_{i,\ell} \in \mathbb{P}_n}$

Then $$H(x) = \sum_{i=0}^{k} \sum_{\ell=0}^{m_i-1} f^{(\ell)}(x_i)\, h_{i,\ell}(x) = \sum_{i=0}^{k} f(x_i)\, h_{i,0}(x) + \sum_{i=0}^{k} f'(x_1)\, h_{i,1}(x) + \sum_{\ell=0}^{k} f''(x_1)\, h_{i,2}(x) + \sum_{i=0}^{k} f'''(x_i)\, h_{i,3}(x) + \dots$$

where $$h_{i,\ell}(x_j) = \begin{cases} 1 & , i=j \text{ and } \ell=m \\ 0 & i\neq j \text{ and } \ell \neq m \end{cases}$$

* How to compute $h_{i,\ell}$:

Define $$L_{i,\ell} = \frac{(x-x_i)^\ell}{\ell!} \prod_{\substack{j=0 \\ j\neq i}}^{k} \left(\frac{x-x_j}{x_i-x_j}\right)^{m_j} , \quad i=\overline{0,k}, \quad \ell = 0, \dots, \boxed{m_i-1}$$

Then $h_{i,\ell}$ polynomials are given by recurrence

$$\begin{cases} h_{i,m_i-1}(x) = L_{i,m_i-1}(x), & i=\overline{0,k} \\ h_{i,m}(x) = L_{i,m}(x) - \displaystyle\sum_{\theta=(m+1)}^{m_i-1} L_{i,m}^{(\theta)}(x_i)\, h_{i,\theta}(x), & \boxed{m = m_i-2, \dots, 0} \end{cases}$$

* Iterpolation error

Remainder in Lagrange interpolation.

real number → My estimate error

complex → harder

* Let ($f \in C^{n+1}([a,b])$) $\{x_0, x_1, \dots, x_n\} \subset (a,b)$ interpolation points.

$\in \mathbb{P}$

Let L_n be the Lagrange interpolation of f at points $x_0, \dots, x_n$; $L_n(x_i) = f_i$, $\forall i = \overline{0,n}$

Then $f(x) - L_n(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!}(x-x_0)\cdots(x-x_n)$, $\xi \in \operatorname{conv}(x_0, \dots, x_n, x)$

$= f[x_0, \dots, x_n, x]$ proved by induction .

* Remark : Mean value theorem
$f(x) - f(x_0) = f'(\xi)(x - x_0)$

* Proof:

- Consider polynomial $q \in \mathbb{P}_{n+1}$ which interpolates f at points $x_0, x_1, \dots, x_n, x$

$$q(t) = L_n(t) + \frac{f(x) - L_n(x)}{\Pi_{n+1}(x)} \Pi_{n+1}(t)$$

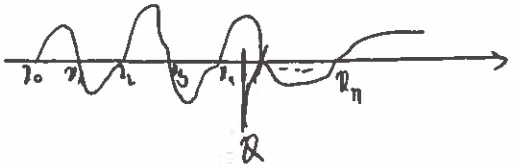
at $x_0, \dots, x_n$ → interpolate

x → also interpolate

when $t = x$, $q(x) = \cancel{L_n(x)} + \dfrac{f(x) - \cancel{L_n(x)}}{\cancel{\Pi_{n+1}(x)}} \cancel{\Pi_{n+1}(x)} = f(x)$

- Let
$E(t) = f(t) - q(t)$

then $E(t) = 0$, $\forall t \in \{x_0, x_1, \dots, x_n, x\}$. ⇒ E has $(n+2)$ zeroes .



⇒ E' has $(n+1)$ zeroes .

⋮

⇒ $E^{(n+1)}$ has one zero , denoted by ξ

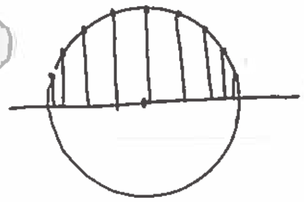
$$0 = E^{(n+1)}(\xi) = f^{(n+1)}(\xi) - \frac{f(x) - L_n(x)}{\Pi_{n+1}(x)} (n+1)!$$

□ ?

* Remark.

when $n \longrightarrow \infty$, then $\dfrac{f^{(n+1)}(\xi)}{(n+1)!} \longrightarrow 0$. ⇒ error ↓ 0

⇒ can improve interpolation by increasing number of interpolation points .



Chebyshev note .

* Hermite interpolation.

*Idea: In this we care only on 2 nodes.

- In Lagrange interpolation, we learned how to solve a simple problem:

Find $L \in \mathbb{P}^1$ such that $L(x_0)=0$ $L(x_1)=1$ x_0, x_1 distinct.

$$L(x_0)=0 \Leftrightarrow L(x) = q(x)(x-x_0) \xrightarrow[L\in\mathbb{P}^1]{} \alpha(x-x_0)$$

$$L(x_1)=1 \Rightarrow L(x_1) = \alpha(x_1-x_0) = 1 \Rightarrow \alpha = \frac{1}{x_1-x_0}$$

$$\Rightarrow L(x) = \frac{1}{(x_1-x_0)}(x-x_0)$$

- How to solve simplest Hermite interpolation?

We want to find $h_{0,1}(x)$, $h_{0,1}\in\mathbb{P}^3$ such that $\begin{cases} h_{0,1}^{(0)}(x_0)=0 \quad \text{and} \quad h_{0,1}^{(1)}(x_0)=1 \\ h_{0,1}^{(0)}(x_1)=0 \quad\quad h_{0,1}^{(1)}(x_1)=0 \end{cases}$

state $p(x)=(x-x_0)(x-x_1)^2$

but $p^{(1)}(x_0) = (x_0-x_1)^2 \neq 1$.

- However $h_{0,1} = \dfrac{p(x)}{(x_0-x_1)^2} = (x-x_0)\left(\dfrac{x-x_1}{x_0-x_1}\right)^2$

the first subscript in $h_{\underset{\text{node}}{0},\underset{\text{derivative}}{1}}$ indicates the node number at which the first derivative is 1 (the second subscript)

- Find $h_{0,0}\in\mathbb{R}^3$ such that $\begin{cases} h_{0,0}^{(0)}(x_0)=1 \quad h_{0,0}^{(1)}(x_1)=0 \\ h_{0,0}^{(0)}(x_0)=0 \quad h_{0,0}^{(1)}(x_1)=0 \end{cases}$

we could easily construct $L_{0,0}(x) = \left(\dfrac{x-x_1}{(x_0-x_1)}\right)^2$ | $L = \left(\dfrac{x-x_0}{x_1-x_0}\right)^2$

$$L_{0,0}^{(1)}(x_0) = \frac{2}{(x_0-x_1)} \neq 0$$

⇒ We want to modify what we have using $h_{0,1}$ (above)

$$h_{0,0}(x) = \underbrace{L_{0,0}(x)}_{\text{degree } 2} - L_{0,0}^{(1)}(x_0)\underbrace{h_{0,1}(x)}_{\text{degree } 3} \quad \text{degree } 3$$

$$= \left(\frac{x-x_1}{x_0-x_1}\right)^2 - \frac{2}{(x_0-x_1)}(x-x_0)\left(\frac{x-x_1}{x_0-x_1}\right)^2 = \frac{(x-x_1)^2(2x+x_1-3x_0)}{(x_1-x_0)^3}$$

- We can easily construct $h_{1,1}, h_{1,0} \in \mathbb{P}^3$

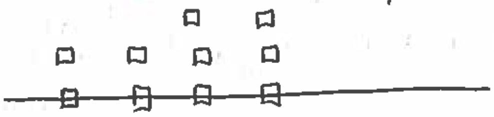
$$h_{1,1}(x) = (x-x_1)\left(\frac{x-x_0}{x_1-x_0}\right)^2$$

$$h_{1,0}(x) = \frac{(x-x_0)^2(2x+x_0-3x_1)}{(x_0-x_1)^3}$$

✱ We can solve Find $\boxed{H \in \mathbb{P}_3}$ ($\boxed{2 \text{ points}}$ Hermite interpolation problem)

$$\begin{cases} H^{(0)}(x_0) = f(x_0) & H^{(1)}(x_0) = f'(x_0) \\ H^{(0)}(x_1) = f(x_1) & H'(x_1) = f'(x_1) \end{cases}$$

Then $H(x) = f(x_0)\,h_{0,0}(x) + f'(x_0)\,h_{0,1}(x) + f(x_1)\,h_{1,0}(x) + f'(x_1)\,h_{1,1}(x)$.



✱ We can write H in Newton's form.

$H(x) = f(x_0) + f[x_0,2](x-x_0) + f[x_0,2;x_1](x-x_0)^2 + f[x_0,2;x_1,2](x-x_0)^2(x-$

$$f[x_0,2] := \frac{f^{(1)}(x_0)}{1!}$$

$$f[x_0,k] := \frac{f^{(k-1)}(x_0)}{\boxed{(k-1)!}}$$

$$f[x_0,2;x_1] := \frac{f[x_0,x_1] - f[x_0,2]}{x_1-x_0}$$

$$f[x_0,2;x_1,2] := \frac{f[x_0,x_1;2] - f[x_0,2;x_1]}{x_1-x_0}$$

	C1	C2	C3	C4
x_0	$f(x_0$			
x_0	$f(x_0)$	$f^{(1)}(x_0)$		
x_1	$f(x_1)$	$f[x_0,x_1]$	$f[x_0,2;x_1]$	
x_1	$f(x_1)$	$f^{(1)}(x_1)$	$f[x_0,x_1;2]$	$f[x_0,2;x_1,2]$

✱ Theorem: $\boxed{n = \sum_{i=1}^{k} m_i - 1}$

Let $x_0, \dots, x_k$ distinct nodes

Let $m_0, \dots, m_k$ integer ≥ 1 $\quad \boxed{\sum_{i=0}^{k} m_i = n+1}$

degree of the polynomial

Let $f^{(m_i-1)}(x_i)$ exists $0 \leq i \leq k$.

Then there exists a unique polynomial $H \in \mathbb{P}_n$ s.t $H^{(l)}(x_i) = f^{(l)}(x_i)$,

$l = 0, \dots, m_i - 1$

* Theorem 1

Let $x_0, \dots, x_k$ be distinct numbers

Let $m_0, \dots, m_k$ be integers, ≥ 1, $\sum_{i=0}^{k} m_i = \overset{\text{degree of the polynomial}}{n+1}$

Let f be a function s.t. $f^{(m_i-1)}(x_i)$ exists $0 \leq i \leq k$.

Then there exist a unique polynomial $H \in \mathbb{P}_n$ such that

$$H^{(\ell)}(x_i) = f^{(\ell)}(x_i) \quad i = \overline{0,\dots,k}, \quad \ell = \overline{0,\dots,m_i-1}$$

* Prove the existence:

- Consider the homogeneous system $H^{(\ell)}(x_i) = 0$ $\quad i = 0, \dots, k, \quad \ell = 0, \dots, m_i - 1$

We want to show $H(x) \equiv 0$

x_i fix

$$H(x) = H(x_i) + H^{(1)}(x_i)(x - x_i) + \dots + \frac{1}{(m_i-1)!} H^{(m_i-1)}(x_i)(x-x_i)^{m_i-1} + \frac{1}{m_i!} H^{(m_i)}(x_i + \theta(x - x_i))\,\cancel{(x - x_i)}\,(x-x_i)^{m_i}$$

Then $H(x) = \prod_{i=0}^{k} (x - x_i)^{m_i} q$

H will be a polynomial of degree n that has n zero $\Rightarrow q = 0 \Rightarrow H(x) \equiv 0$.

* Theorem 2:

$$H(x) = \sum_{i=0}^{k} \sum_{\ell=0}^{m_i-1} f^{(\ell)}(x_i)\, h_{i,\ell}(x) \quad \text{where } h_{i,\ell} \in \mathbb{P}_n$$

$$h^{(m)}_{i,\ell}(x_j) = \begin{cases} 1 & \text{for } i = j \text{ and } m = \ell \\ 0 & \text{for } i \neq j \text{ and } m \neq \ell \end{cases} \qquad \text{for } i = \overline{0,k}, \quad 0 \leq m, \ell \leq m_i - 1 \qquad (*)$$

node notation

which derivative at node i $\neq 0$

* Define polynomial $$L_{i,\ell} = \underbrace{\frac{(x-x_i)^{\ell}}{\ell!}}_{F(x)} \cdot \underbrace{\prod_{\substack{j=0 \\ j \neq i}}^{k} \left(\frac{x - x_j}{x_i - x_j}\right)^{m_j}}_{G(x)} \qquad i = 0, k. \quad \ell = 0, \dots, m_i - 1.$$

$\left[l_i(x)\right]^{m}$

Then $h_{i,\ell}$ polynomials are given by recurrence

$$h_{i,m_i-1}(x) = L_{i,m_i-1}(x) \qquad i = 0, \dots, k$$

and

$$h_{i,m}(x) = L_{i,m}(x) - \sum_{\vartheta = m+1}^{m_i-1} L^{(\vartheta)}_{i,m}(x_i)\, h_{i,\vartheta}(x), \qquad m = m_i - 2, m_i - 3, \dots, 0$$

satisfy (*)

* $L_{i,\ell} \in \mathbb{P}_{n+L+\ell-m_i}$ (can have high or very low degree)

- $L_{i,\ell}$ polynomials satisfy the following interpolation condition.

At note x_i (a)

$$L_{i,\ell}(x_i) = \cdots = L^{(\ell-1)}_{i,\ell}(x_i) = 0 \qquad L^{(\ell)}_{i,\ell}(x_i) = 1.$$

At note x_j

$$L^{(0)}_{i,\ell}(x_j) = \cdots = L^{(m_j-1)}_{i,j}(x_j) = 0 \qquad j \neq i$$

Idea: $L_{i,\ell}$ satisfy ⇒ $h_{i,m}$ satisfies $(*)$.

The tough thing to check is $L^{(\ell)}_{i,\ell}(x_i) = 1$.

- Now verify that $L^{(\ell)}_{i,\ell}(x_i) = 1$.

$$L^{(N)}_{i,\ell}(x_i) = (FG)^{(N)} = \sum_{\ell=0}^{N} \binom{N}{k} F^{N-\ell}_{(x)} G^{(\ell)}_{(x)}$$

Put $N = \ell$

$$= \sum_{\ell=0}^{0} \sim + \sum_{\ell\neq 0}^{\ell} \sim$$

Then

$$L^{(\ell)}_{i,\ell}(x) = F^{(\ell)}(x) G^{(0)}(x) + \underbrace{\sum_{\ell\neq 0} \binom{\ell}{k} F^{(\ell-k)}_{(x)} G^{(\ell)}_{(x)}}_{(x-x_i)R(x)}$$

$$L^{(\ell)}_{i,\ell}(x_i) = F^{(\ell)}(x_i) G^{(0)}(x_i) + 0 = 1\cdot 1 + 0 = 1.$$

* Check that h_{i,m_i-1} and $h_{i,m}$, $m = m_i-2,\dots,0$ satisfy $(*)$ in theorem 2

At each x_i $L_{i,m_i-1}(x)$ satisfies all but cond required of $h_{i,m_i-1}(x)$

To construct the rest of $h_{i,m}$, $m = m_i-2, m_i-3, \dots, 0$

$h_{i,m}$

We modify $L_{i,m}$ by subtracting previously computed $h_{i,m_i-1}, \dots, h_{i,m+1}$

The purpose of modification is to ensure $h^{(\ell)}_{i,m}(x_i) = 0$ $\ell = m+1, \dots, m_i-1$.

$$h^{(\ell)}_{i,m}(x_i) = L^{(\ell)}_{i,m}(x_i) - \sum_{\nu=m+1}^{m_i-1} L^{(\nu)}_{i,m}(x_i)\, h^{(\ell)}_{i,\nu}(x_i) \overset{\text{want to show}}{=} 0, \forall \ell$$

$$= L^{(\ell)}_{i,m}(x) - L^{(\ell)}_{i,m}(x_i) \underbrace{h^{(\ell)}_{i,\ell}(x_i)}_{=1} = 0$$

We can obtain Newton's form of Hermit interpolation.

1) $f[x_0, i] = \dfrac{f^{(i-1)}(x_0)}{(i-1)!}, \quad i \geq 1$

2) $f[x_0, m_0; \dots; x_\ell, m_\ell] = \dfrac{f[x_0, m_0 - 1; \dots; x_\ell, m_\ell] - f[x_0, m_0; \dots; x_\ell, m_\ell - 1]}{x_\ell - x_0}$

* For any i $0 \leq i \leq \ell$, we define

$$s(i) = \begin{cases} 0 & i = 0 \\ m_0 + m_1 + \dots + (m_{i-1}) & \text{for } 0 \lneq i \leq \ell \end{cases}$$

Every integer p, $0 \leq p \leq n$, can be now represented as $p = s(i) + \overset{\text{remainder}}{j}$ $\quad \begin{matrix} 0 \leq i \leq \ell \\ 0 \leq j \leq m_i - 1 \end{matrix}$

Next,

$\Pi_0(x) \equiv 1$

$\Pi_1(x) = \Pi_{s(0)+1}(x) = (x - x_0)$

$\vdots$

$\Pi_{m_0-1}(x) = \Pi_{s(0)+m_0}(x) = (x - x_0)^{m_0 - 1}$

$\Pi_{m_0}(x) = \Pi_{s(1)+0}(x) = (x - x_0)^{m_0}$

$\Pi_{s(i)+j}(x) = (x - x_0)^{m_0} \dots (x - x_{i-1})^{m_\ell - 1} (x - x_i)^j$

$$H(x) = \sum_{p=0}^{m} b_p \Pi_p(x) = \sum_{i=0}^{\ell} \sum_{j=0}^{m_i - 1} b_{s(i)+j} \Pi_{s(i)+j}(x)$$

$$= \sum_{i}^{\ell} \sum_{j}^{m_i - 1} f[x_0, m_0; \dots; x_{i-1}, m_{i-1}; x_i, j+1] (x - x_0)^{m_0} \dots (x - x_{i-1})^{m_{i-1}} (x - x_i)^j$$

* **Example**: $m_0 = 3 \quad m_1 = 2$ $\quad \begin{cases} s(0) = 0 \\ s(1) = m_{\ell-1} + ~~m_1 = 3 + 2 = 5~~ = m_0 = 3. \end{cases}$

$0 = s(0) = 0 + 0 \qquad \Pi_0(x) \equiv 1$

$1 = s(0) + 1 \qquad \Pi_1(x) = (x - x_0)$

$2 = s(0) + 2 = 0 + 2 \qquad \Pi_2(x) = (x - x_0)^2$

$3 = s(1) + 0 = 3 + \qquad \Pi_3(x) = ($

* WL floting point.

```
ℓ = 0
while $(\frac{1}{2})^\ell > 0$
    ℓ = ℓ + 1
end
```

$\ell_{max} = 1075 = E_{min} + t + 1.$

* Example of Hermite interpolation in Newton form

* Example: Let $f(x) = x^4 + 1$ $f'(x) = 4x^3$

We will construct a polynomial $p_5(x)$ such that $p(x_i) = f(x_i)$ $x_i = -1, 0, 1$.
$p'(x_i) = f'(x_i)$ 2.

* We have the Hermite divided difference table is

	$f(z_i)$	$f[\cdot,\cdot]$				
-1	$\boxed{2}$					
-1	2	$\boxed{f'(-1) = -4}$ 1				
0	1	-1	$\boxed{3}$ 2			
0	1	$f'(0) = 0$	1	$\boxed{-2}$ 3=2+1		
1	2	1	1	0	$\boxed{1}$ 4=2+2	
1	2	$f'(1) = 4$	3	2	1	$\boxed{0}$ 5=2+2+1

# of times that we repeat depends on the available derivative

(maximum derivative+1

degree 6

note that we also take the first number of value $f(x_0)$.

the way we write depends on the derivative that we have

$$\Rightarrow H(x) = 2\,4(x+1) + 3(x+1)^2 - 2(x+1)^2 x + 1(x+1)^2 x^2 + 0$$

* Example 2, consider the data with $m=1$, $n_0 = 1$, $n_1 = 2$ given in the following table.

x_i	0	1
$f(x_i)$	1	2
$f'(x_i)$	0	1
$f''(x_i)$	NA	2

z_i	$f(z_i)$	2DD $f[x, y]$	3DD $f[x,y,z]$		
0	1 0				
0	1	$f'(0) = 0$ 1 x_0			
1	2	1	1 2 x_0^2		
1	2	$f'(1) = 1$	0	-1 3 2 1	
1	2	$f'(1) = 1$	$\frac{f''(1)}{2} = 1$	1	2 4 2 2

Correction HW3, P 2,

Suppose that f is a function in $[0,3]$, for which one knows that

$$f(0)=1 \qquad f(1)=2 \qquad f'(1)=-1 \qquad f(3)=f'(3)=0$$

a) Estimate $f(2)$ using Hermite interpolation

b) Estimate the maximum possible error of the answer given in a if one knows, in addition that $f \in C^5[0,3]$ and $|f^{(5)}(x)| \leq M$ on $[0,3]$.

z_i	$f(z_i)$	$f[z_i,z_j]$	$f[\cdot,\cdot,\cdot]$	$f[\cdot,\cdot,\cdot,\cdot]$	
		0			
0	1				
1	2	1^{1}			
1	2	$f'(1)=-1$	-2	$2=1+1$	
3	0	-1	0	$\frac{2}{3}$	$3=1+2$
3	0	$f'(3)=0$	$\frac{1}{2}$	$\frac{1}{4}$	$-\frac{5}{36}$ $4=1+2+1$

Then $H(x) = 1 + 1(x-0) - 2(x-0)(x-1) + \frac{2}{3}(x-0)(x-1)^2 - \frac{5}{36}(x-0)(x-1)^2(x-3)$

$= 1 + \frac{49}{12}x - \frac{155}{36}x^2 + \frac{49}{36}x^3 - \frac{5}{36}x^4$.

We have $H(2) = \frac{11}{18}$.

*

*Chebyshev Polynomials

* Given by recurrence

$T_0(x)=1$

$T_1(x)=x$

$T_{n+1}(x)=2xT_n(x)-T_{n-1}(x)$, $\boxed{n\geq 1}$

The leading for $T_n(x)$ is

$$T_n(x)=\boxed{2^{n-1}}x^n+\cdots$$

$T_2(x)=2x^2-1$

$T_3(x)=4x^3-3x$

$T_4(x)=8x^4-8x^2+1$

$T_5(x)=2^4x^5-20x^3+5x$

$T_6(x)=2^5x^6-48x^4+18x^2-1$

* Four different ways.

* Trigonometric formula

$T_n(x)=\cos(n\arccos x)$, $x\in[-1,1]$.

$\quad=\cos(n\theta)$, where $\theta=\arccos x \iff x=\cos\theta$.

* Complex formula

$T_n(x)=\frac{1}{2}(z^n+z^{-n})=\mathrm{Re}(z^n)$, $x=\mathrm{Re}(z)$, $\boxed{|z|=1}$

$\quad=\cos(n\theta)$ where $z=e^{i\theta}=\cos\theta+i\sin\theta$.

* Transcendental formulas

$T_n(x)=\frac{1}{2}\left[(x+\sqrt{x^2-1})^n+(x-\sqrt{x^2-1})^n\right]$, $\boxed{x\in\mathbb{R}}$

* Hyperbolic formula

$$T_n(x)=\begin{cases}\cosh(n\,\mathrm{arccosh}\,x) & x\geq 1\\ (-1)^n\cosh(n\,\mathrm{arccosh}(-x)) & x\leq -1\end{cases}$$

* The trigonometric formula makes computing zeroes and local extrema of Chebyshev polynomial easy

$T_n(t_{nk})=0 \Rightarrow t_{nk}=\cos\frac{(2k-1)\pi}{2n}$

$k=1,2,\dots,n$

$T_n(s_{nk})=(-1)^k$

$\Rightarrow s_{nk}=\cos\left(\frac{k\pi}{n}\right)$

$k=0,\dots,n$

* Minimal properties of T_n

Let $p\in\mathbb{P}_n$ is a monic polynomial (a monic polynomial is a polynomial with leading coefficient $=1$)

Then $2^{1-n}=\max\limits_{\boxed{-1\leq x\leq 1}}|2^{1-n}T_n(x)|\leq\max\limits_{\boxed{-1\leq x\leq 1}}|p(x)|$

* Minimize the error of Lagrange interpolation by choosing the best notes

- For a given $n\geq 0$, The Chebyshev nodes on the interval $(-1,1)$ are $x_k=\cos\left(\frac{(2k-1)\pi}{2n}\right)$, $k=\overline{1,n}$

These are the roots of the Chebyshev polynomial of the first kind of degree n.

- For nodes for an arbitrary interval $[a,b]$, the nodes are

$$x_k=\frac{1}{2}(a+b)+\frac{1}{2}(b-a)\cos\left(\frac{(2k-1)\pi}{2n}\right),\quad k=\boxed{1},\dots,n$$

* Theorem :

Consider $f:[a,b] \longrightarrow \mathbb{R}$

Let $M_{n+1} = \sup_{a \leq x \leq b} |f^{(n+1)}(x)|$

Then if $x_m = \frac{1}{2}(b+a) + \frac{1}{2}(b-a)\cos\left(\frac{(2\ell-1)\pi}{2n}\right)$, $\ell = \overline{1,n}$

Then we have

$$|f(x) - L(x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!\, 2^{2n+1}}$$

* Chebyshev Polynomials are given by a recurrence

$T_0(x) = 1$

$T_1(x) = x$

$\vdots$

$T_{n+1}(x) = 2x\,T_n(x) - T_{n-1}(x)$ $\quad n \geq 1$

$T_2(x) = 2x^2 - 1$

$T_3(x) = 2x(2x^2-1) - x = 4x^3 - 3x$

$T_4(x) = 8x^4 - 8x^2 + 1$

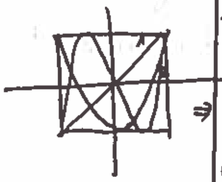
$T_5(x) = 16x^5 - 20x^3 + 5x$

T_n (Leading has 2^{n-1} leading coefficient

* Trigonometric formulas :

$T_n(x) = \cos(n \arccos x)$ $\quad x \in [-1, 1]$

$= \cos(n\theta)$ $\quad x = \cos\theta$



$\theta := \arccos x \Rightarrow x = \cos(\theta)$

$T_n(x) = \cos(n\theta)$

$\Rightarrow T_n(\cos\theta) = \cos(n\theta)$

$\cos(4\theta) = 8\cos^4\theta - 8\cos^2\theta + 1$

* Complex formulas :

$T_n(x) = \frac{1}{2}\left[z^n + z^{-n}\right] = \mathrm{Re}(z^n)$ $\quad x = \mathrm{Re}(z),$ $\quad |z| = 1$

$= \cos(n\theta)$ $\quad$ where $z = e^{i\theta} = \cos\theta + i\sin\theta$

* Transcendental formulas

$$T_n(x) = \frac{1}{2}\left(\left(x + \sqrt{x^2-1}\right)^n + \left(x - \sqrt{x^2-1}\right)^n\right) \quad x \in \mathbb{R}$$

* Hyperbolic formulas

$$T_n(x) = \begin{cases} \cosh(n \operatorname{arccosh} x) & x \geq 1 \\ (-1)^n \cosh(n \operatorname{arccosh}(-x)) & x \leq -1 \end{cases}$$

* Prove the equivalence between reccurence formula and trigonometric formula .

- $T_0(x) = \cos(0 \arccos x) = \cos(0) = 1$

  $T_1(x) = \cos(1 \arccos x) = \cos(\arccos x) = x$

We must verify that

$$\underset{\text{LHS or}}{T_{n+1}} = \cos((n+1)\arccos x) = \overset{\text{RHS}}{2x\,T_{n+1}(x) - T_{n-1}(x)} = \underset{\overset{\parallel}{2\cos\theta\,\cos(n\theta)}}{2x\cos(n\arccos x)} - \cos((n-1)\arccos x$$

$\rightarrow$ NTP $\quad \cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\,\cos(n\theta)$

$\parallel$

~~$\cos(n\theta+\theta)$~~ $= \cos(n\theta)\cos(\theta) - \sin(n\theta)\sin(\theta) + \cos n\theta\cos\theta + \sin n\theta \sin\theta = \text{RHS}$

LHS =

- $T_n(t_{n,\ell}) = 0$ $\quad t_{n,\ell} = \cos\left(\frac{(2\ell-1)\pi}{2n}\right)$ $\quad \ell = 1, \dots, n$

  $T_n(s_{n,\ell}) = (-1)^\ell$ $\quad s_{n\ell} = \cos\left(\frac{\ell\pi}{n}\right)$ $\quad \ell = 0, \dots, n$ $\leftarrow$ The points where T_n attains extrema .

* Now look at the complex formula.

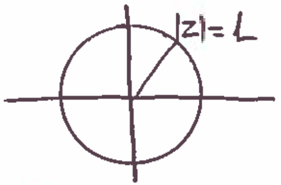
$T_n(x) = \frac{1}{2}(z^n + z^{-n}) = Re(z^n)$ $x = Re(z), |z| = 1$

Recurrence

$T_1(x) = 1$

$T_2(x) = x$

$T_n(x) = 2x\,T_{n+1}(x) + T_{n-2}(x)$. $n \geq 1$



$z = e^{i\theta} = \cos\theta + i\sin\theta$

$x = \underset{\uparrow}{Re(z)} = \cos(\theta)$

$\frac{1}{2}(z^1+z^n) = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \frac{1}{2}(z+z^{-1})$

$z^n = (e^{in\theta}) = \cos(n\theta) + i\sin(n\theta) \Rightarrow T_n(x) = Re(z^n) = \cos(n\theta)$.

$$Re(z^n) = \frac{1}{2}(z^n + z^{-n})$$

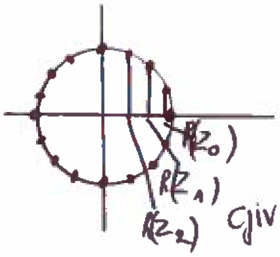
• These Chebyshev points at which T_n attains extreme

$z_\ell = e^{i\frac{\ell 2\pi}{2n}}$, $\ell = 0$

These are $2n$ root of unity

$x_\ell = Re(z_\ell) = \cos\left(\frac{\ell\pi}{n}\right)$ $\ell = 0, \dots, n$.

$x_\ell = s_{n\ell}$ (above)



gives $s_{n\ell}$.

* Minimal properties of T_n

Let $p \in \mathbb{P}_n$ with leading coefficient 1 (monic)

Then $2^{1-n} = \underset{-1\le x\le 1}{\max}\,\underbrace{|2^{1-n}T_n(x)|}_{} \le \underset{-1\le x\le 1}{\max}|p(x)|$

normalize $T_n(x)$ so that it has leading coefficient $= 1$.

when we ~~minimiz~~ normalize the Chebyshev polynomial then it will be the smallest pol among group of pols.

Proof: functional analysis problem (not a normal easy problem)

Put $\|f\| = \underset{-1\le x\le 1}{\max}|f(x)|$

• On $[-1,1]$ $2^{1-n}T_n(x)$ assume extremal values at $y_\ell = \cos\frac{\ell\pi}{n}$

$2^{1-n}T_n(y_\ell) = 2^{1-n}(-1)^\ell$.

By contradiction, suppose there exists $\tilde{p} \in \mathbb{P}_n$ such that $\underset{-1\le p\le 1}{\max}|\tilde{p}(x)| \overset{\text{strictly smaller}}{<} |2^{1-n}T_n(x)| = |2^{1-1}$

Consider $Q(x) = \cancel{(-1)^\ell 2^{1-n}}\; 2^{1-n}T_n(x) - \tilde{p}(x) \in \mathbb{P}_{n-1}$.

At those extremal points

$Q(y_k) = 2^{1-n}T_n(y_k) - \tilde{p}(x) = 2^{1-n}(-1)^k - \tilde{p}(y_k)$ $k = 0, \dots, n$ } slitly pertu but not change the sign

From above $\max|\tilde{p}(x)| < 2^{1-n}$

⇒ The sign of $Q(y_k)$, $sign(Q_{y_k}) = sign\left(2^{1-n}(-1)^k\right) = (-1)^k$.

Sum up ⇒ $\begin{cases} Q(x) \in \mathbb{P}_{n-1} \\ \text{change the sign } (n+1) \text{ time} \Rightarrow \text{has } n \text{ zeros.} \end{cases}$ $\Rightarrow Q \equiv 0 \Rightarrow \tilde{p}(x) \equiv 2^{1-n}T_n(x)$.

contracts with the assumption $\max \tilde{p}(x) < |2^{1-n}|$

* How to minimize the error in Lagrange interpolation by choosing the best nodes?

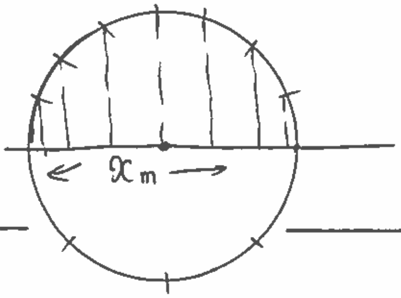
* Theorem :

Consider $f : [a,b] \longrightarrow \mathbb{R}$

Let $M_{n+1} = \sup\limits_{a \le x \le b} |f^{(n+1)}(x)|$

Let $E(x) = f(x) - L_n(x)$

Then if $x_m = \frac{1}{2}\left((b-a)\cos\frac{(2m+1)\pi}{2(n+1)} + b + a\right)$, $m = 0, 1, .., n$ } $|E(x)| \le \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!\, 2^{2n+1}}$.



* Prove :

$\Pi_{n+1}(x) = (x - x_0) \dots (x - x_n) = 2^{-n}T_{n+1}(x)$

$\max\limits_{a \le x \le b} \Pi_{n+1}(x) = 2^{-n}$

- If $x = \frac{1}{2}(b-a)z + b + a$ $-1 \le z \le 1$

$f(x) = f\left[\frac{1}{2}(b-a)z + b + a\right] = \tilde{f}(z)$.

- $\Pi_{n+1}(x) = \frac{(b-a)^{n+1}}{2^{2n+1}}(z - z_0) \dots (z - z_n) = \frac{(b-a)^{n+1}}{2^{n+1}} \frac{T_{n+1}(z)}{2^n}$

$|\Pi_{n+1}(x)| \le \frac{(b-a)^{n+1}}{2^{2n+1}}$

- $f(x) - L_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0) \dots (x - x_n)$

Chebyshev polynomials

September 19, 2018

Chebyshev polynomials are defined by the recurrence formula :

$$\begin{cases} T_0(x) = 1 \\ T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad \text{for } n \geq 1 \end{cases}$$

We have

$$T_2(x) = 2x^2 - 1, \tag{1}$$
$$T_3(x) = 4x^3 - 3x, \tag{2}$$
$$T_4(x) = 8x^4 - 8x^2 + 1, \tag{3}$$
$$T_5(x) = 16x^5 - 20x^3 + 5x, \tag{4}$$
$$T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1 \tag{5}$$

The recurrence formula clearly generates polynomials such that only the even powers of x occur in T_{2k} and only odd powers of x occur in T_{2k-1}. The leading coefficient of T_n is 2^{n-1}, for $n \geq 1$.
We discuss four different ways to represent T_n: trigonometric, complex, transcendental and hyperbolic.

Trigonometric function

$$T_n(x) = \cos(n \arccos x) \qquad \text{for } x \in [-1, 1] \tag{6}$$

Complex function

$$T_n(x) = \frac{1}{2}(z^n + z^{-n}) = \mathrm{Re}(z^n), \qquad \text{for } x = \mathrm{Re}(z),\ |z| = 1 \tag{7}$$

$$T_n(x) = \frac{1}{2}\left((x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n\right) \qquad \text{for } x \in \mathbb{R} \tag{8}$$

$$T_n(x) = \begin{cases} \cosh(n \operatorname{arc\,cosh} x) & \text{for } x \geq 1 \\ (-1)^n \cosh(n \operatorname{arc\,cosh}(-x)) & \text{for } \quad x \leq -1 \end{cases} \tag{9}$$

* Trigonometric function .

The first way to represent T_n on $[-1, 1]$ is as trigonometric functions. Let $0 \leq \theta \leq \pi$ so that if $x = \cos\theta$ then $-1 \leq x \leq 1$ and $\theta = \arccos x$. Then

$\cos(0 \arccos x) = 1 = T_0(x)$, $\cos(\arccos x) = x = T_1(x)$. We need to check that $T_n(x) = \cos(n \arccos x)$ satisfies the 3-term recurrence $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$. Set $\theta = \arccos x$ so that $T_n(x) = \cos n\theta$. We must verify

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos\theta\cos n\theta$$

Rewriting the left we get

$$\cos(n\theta + \theta) + \cos(n\theta - \theta) = 2\cos\theta\cos n\theta$$

because $\cos(\alpha \pm \beta) = \cos\alpha\cos\beta \mp \sin\alpha\sin\beta$. We have that $\cos(n\theta) = T_n(\cos\theta)$ so for example $\cos(5\theta) = 16\cos^5(\theta) - 20\cos^2(\theta) + 5\cos(\theta)$.
The trigonometric representation makes computing the zeroes and local extrema of Chebyshev polynomials easy. Indeed

$$T_n(t_{n,k}) = 0 \qquad t_{n,k} = \cos\frac{(2k-1)\pi}{2n} \qquad k = 1, 2, \ldots, n \qquad (10)$$

$$T_n(s_{n,k}) = (-1)^k \qquad s_{n,k} = \cos\frac{k\pi}{n} \qquad k = 0, 1, \ldots, n \qquad (11)$$

The zeros are computed from the solutions $0 \leq \theta_{n,k} \leq \pi$ of $\cos n\theta = 0$, which are $\theta_{n,k} = \frac{(2k-1)\pi}{2n}$, $k = 1, 2, \ldots, n$. $\frac{\pi}{2n}, \frac{3}{2n}\pi, \ldots, \frac{2n-1}{2n}\pi$. The extrema are computed from the solutions $0 \leq \tilde{\theta}_{nk} \leq \pi$ of $(\cos(n\theta))' = \frac{-n}{\sqrt{1-x^2}}\sin(n\theta)$ which are $\tilde{\theta}_{nk} = 0, \frac{\pi}{n}, \frac{2}{n}\pi, \ldots, \pi$.

Complex function *

Trigonometric definition of Chebyshev polynomials $T_n(x) = \cos(n\theta)$, $\theta = \arccos x$ can be reformulated in terms of complex functions. Let $z = e^{i\theta}$ be complex numbers on the unit circle. Then

$$x = \mathrm{Re}(z) = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \cos\theta$$

Hence

$$\cos(n\theta) = \mathrm{Re}(z^n) = \frac{1}{2}(z^n + z^{-n}), \qquad |z| = 1$$

Consider the $n+1$ points $\{z_j\}$ on the upper half of the unit circle in the complex plane

$$z_j = e^{i\frac{j\pi}{n}}, \qquad j = 0, \ldots, n$$

Points $\{z_j\}$ are equispaced and divide the upper part of the unit circle into n equal parts. They may be interpreted as first $n+1$ of the $2n$-th roots of unity

$$z_j = e^{i\frac{j2\pi}{2n}}, \qquad j = 0, \ldots, n$$

The real parts of points $\{z_j\}$

$$x_j = \mathrm{Re}(z_j) = \frac{1}{2}(z_j + z_j^{-1}) = \cos(\frac{j\pi}{n}), \qquad j = 0, \ldots, n$$

are called Chebyshev points. They are contained in $(-1, 1)$ and cluster near 1 and -1. The Chebyshev points are the local extrema of the n-th Chebyshev polynomial T_n in $[-1, 1]$. Above they were denoted $s_{n,j}$.

Orthogonality of Chebyshev polynomials.

We consider the unitary space $L^2_w([-1,1])$ with the inner product $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)w(x)\,dx$ where $w(x) = (1-x^2)^{-1/2}$.

$$\langle T_i, T_j\rangle = \begin{cases} 0 & \text{if} \quad i \neq j \\ \frac{1}{2}\pi & \text{if} \quad i = j \neq 0 \\ \pi & \text{if} \quad i = j = 0 \end{cases}$$

We set $\theta = \arccos x$, $\mathrm{d}\theta = -(1-x^2)^{-1/2}\mathrm{d}x$. So when $x \in [-1,1]$ then $\pi \geq \theta \geq 0$. Using again the trigonometric identity $2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta)$ we get that

$$\begin{aligned}\langle T_i, T_j\rangle &= \int_0^\pi \cos(i\theta)\cos(j\theta)\,\mathrm{d}\theta \\ &= \frac{1}{2}\left(\int_0^\pi \cos(i-j)\theta\,d\theta + \int_0^\pi \cos(i+j)\theta\,\mathrm{d}\theta\right)\end{aligned}$$

If $i \neq j$ denoting $(i \pm j)\theta = \upsilon$ we have $\upsilon \in [0, (i\pm j)\pi]$ and hence

$$\int_0^\pi \cos(i\pm j)\theta\,\mathrm{d}\theta = \frac{1}{i\pm j}\int_0^{(i\pm j)\pi} \cos\upsilon\,\mathrm{d}\upsilon = \frac{1}{i\pm j}[\sin\upsilon]_{\upsilon=0}^{\upsilon=(i\pm j)\pi} = 0.$$

If $i = j \neq 0$

$$\frac{1}{2}\left(\int_0^\pi \cos 0\,d\theta + \int_0^\pi \cos(2i\theta)\,d\theta\right) = \frac{1}{2}(\pi + 0) = \frac{\pi}{2}$$

Second interesting way to represent T_n is to use transcendental functions.

$$T_n(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n\right) \quad \text{for} \quad x \in \mathbb{R}$$

This also provides a closed form solution for the algebraic recurrence which defined the Chebyshev polynomials.
If $|x| \leq 1$ then the transcendental formula gives the same as before because setting $x = \cos\theta$ in the above formula using Euler's formula, we obtain a complex valued expression

$$\begin{aligned}T_n(x) &= \frac{1}{2}\left((\cos\theta + i\sin\theta)^n + (\cos\theta - i\sin\theta)^n\right) \\ &= \frac{1}{2}(\cos n\theta + i\sin n\theta + \cos n\theta - i\sin n\theta) \\ &= \cos n\theta\end{aligned}$$

The transcendental definition corresponds to the algebraic one also if $|x| \geq 1$.

$$T_n(x) = \frac{1}{2}\left((x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n\right) = \frac{1}{2}(y^n + w^n)$$

We now show that the recurrence $T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x)$ holds. We have that $y + w = 2x$ and $yw = 1$. The recurrence will hold if we show that

$$y^n + w^n = (y+w)(y^{n-1} + w^{n-1}) - (y^{n-2} + w^{n-2})$$

expanding the right side of the above and simplifying

$$\begin{aligned} y^n + w^n &= y^n + yw^{n-1} + wy^{n-1} + w^n - y^{n-2} - w^{n-2} \\ 0 &= (yw-1)(w^{n-2} + y^{n-2}) \end{aligned}$$

* Hyperbolic functions

The last equality holds because $yw = 1$.
The third useful way to represent T_n is to use hyperbolic functions:

$$T_n(x) = \cosh\left(n \operatorname{arccosh} x\right) \quad \text{for} \quad x > 1, \quad n \geq 0.$$

where $\cosh x = 1/2(e^x + e^{-x})$.
To check that $T_n(x) = \cosh(n \operatorname{arc}\cosh x)$ for $x \geq 1$ satisfies the 3-term recurrence set $\theta = \operatorname{arc}\cosh x$ so that $T_n(x) = \cosh n\theta$. We must verify that

$$\cosh(n\theta) + \cosh((n-2)\theta) = 2\cosh\theta\cosh(n-1)\theta$$

We write the left side as $\cosh((n-1)\theta+\theta)+\cosh((n-1)\theta-\theta)$ and this simplifies the expression on the right because $\cosh(\alpha \pm \beta) = \cosh\alpha\cosh\beta \pm \sinh\alpha\sinh\beta$ (unlike in the addition formula for cos).
To check that the recurrence holds for $T_n(x) = (-1)^n\cosh(n\operatorname{arc}\cosh(-x))$ for $x \leq -1$ we set $\operatorname{arc}\cosh(-x) = \theta$ and we must verify that $(-1)^n\cosh n\theta + (-1)^{n-2}\cosh((n-2)\theta) = 2(-\cosh\theta)(-1)^{n-1}\cosh((n-1)\theta)$ which is the same as in the case $x \geq 1$.

Rodrigues formula for Chebyshev polynomials. We want to establish an explicit formula for Chebyshev polynomials of the form

$$T_n(x) = \frac{(-1)^n(1-x^2)^{\frac{1}{2}}}{(2n-1)!!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left((1-x^2)^{n-\frac{1}{2}}\right), \qquad n = 0, 1, \ldots$$

where we set the double factorial $(-1)!! = 1$.

Proof. (a) We start with the transcendental formula for Chebyshev polynomials

$$T_n(x) = \frac{1}{2}((x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n)$$

Use $x \pm \sqrt{x^2-1} = \frac{1}{2}(\sqrt{x+1} \pm \sqrt{x-1})^2$ to obtain the formula

$$T_n(x) = 2^{-n-1}((\sqrt{x+1}+\sqrt{x-1})^{2n} + (\sqrt{x+1}-\sqrt{x-1})^{2n})$$

Next using the above and binomial formula we show that the "binomial formula" for T_n holds:

$$T_n(x) = 2^{-n}\sum_{k=0}^{n}\binom{2n}{2k}(x+1)^{n-k}(x-1)^k, \qquad n = 0, 1, \ldots$$

Finally we will use the Leibniz differentiation rule to compute the n-derivative in the Rodrigues formula to show it follows from the binomial formula.

$$\begin{aligned} T_n(x) &= \frac{1}{2}((x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n) \\ &= 2^{-n-1}((\sqrt{x+1}+\sqrt{x-1})^{2n} + (\sqrt{x+1}-\sqrt{x-1})^{2n}) \end{aligned}$$

(b) Using the binomial formula

$$\begin{aligned} T_n(x) &= 2^{-n-1}((\sqrt{x+1}+\sqrt{x-1})^{2n} + (\sqrt{x+1}-\sqrt{x-1})^{2n}) \\ &= 2^{-n-1}\left(\sum_{j=0}^{2n}\binom{2n}{j}(x+1)^{\frac{2n-j}{2}}(x-1)^{\frac{j}{2}} + \sum_{j=0}^{2n}\binom{2n}{j}(x+1)^{\frac{2n-j}{2}}(-1)^j(x-1)^{\frac{j}{2}}\right) \\ &= 2^{-n-1}\left(2\sum_{j=0,2,\ldots}^{2n}\binom{2n}{j}(x+1)^{\frac{2n-j}{2}}(x-1)^{\frac{j}{2}}\right) \\ &= 2^{-n}\sum_{k=0}^{n}\binom{2n}{2k}(x+1)^{n-k}(x-1)^k \end{aligned}$$

The last part of the proof demonstrates that the formula obtained in (b) is the Rodrigues formula for Chebyshev polynomials. We evaluate the the n-th derivative in the Rodrigues formula using the Leibniz differentiation rule.

$$\frac{d^n}{dx^n}\left((1-x^2)^{n-\frac{1}{2}}\right) = \sum_{k=0}^{n}\binom{n}{k}\frac{d^k}{dx^k}(1+x)^{n-\frac{1}{2}}\frac{d^{n-k}}{dx^{n-k}}(1-x)^{n-\frac{1}{2}}$$

$$= \sum_{k=0}^{n}\binom{n}{k}(n-\tfrac{1}{2})\ldots(n-k+\tfrac{1}{2})(1+x)^{n-k-\frac{1}{2}}(-1)^{n-k}(n-\tfrac{1}{2})\ldots(k+\tfrac{1}{2})(1-x)^{k-\frac{1}{2}}$$

Next we multiply the previous formula by $(-1)^n(1-x^2)^{\frac{1}{2}}$ which changes signs and makes the exponents integer

$$= (-1)^n(1-x^2)^{\frac{1}{2}}\frac{d^n}{dx^n}\left((1-x^2)^{n-\frac{1}{2}}\right) =$$

$$= \sum_{k=0}^{n}(-1)^k\binom{n}{k}(n-\tfrac{1}{2})\ldots(n-k+\tfrac{1}{2})(n-\frac{1}{2})\ldots(k+\tfrac{1}{2})(1+x)^{n-k}(1-x)^k$$

$$2^{-n}\sum_{k=0}^{n}\binom{n}{k}(2n-1)\ldots(2n-2k+1)\cdot(2n-1)\ldots(2k+1)(x+1)^{n-k}(x-1)^k$$

We examine the coefficient in the sum we obtained above

$$2^{-n}\binom{n}{k}(2n-1)(2n-3)\dots(2n-2k+1)\cdot(2n-1)(2n-3)\dots(2k+1)=$$

$$2^{-n}\frac{2\cdot 4\cdot\ldots\cdot 2n}{(2\cdot 4\cdot\ldots\cdot 2k)(2\cdot 4\cdot\ldots\cdot(2n-2k)}\frac{(1\cdot 3\cdot\ldots\cdot(2n-1))^2}{(1\cdot 3\cdot\ldots\cdot(2n-2k-1)(1\cdot 3\cdot\ldots\cdot(2k-1)}$$

$$=2^{-n}(2n-1)!!\frac{(2n)!}{(2k)!(2n-2k)!}$$

$$=2^{-n}(2n-1)!!\binom{2n}{2k}$$

Finally

$$\frac{(-1)^n(1-x^2)^{\frac{1}{2}}}{(2n-1)!!}\frac{d^n}{dx^n}\left((1-x^2)^{n-\frac{1}{2}}\right)=2^{-n}\sum_{k=0}^{n}\binom{2n}{2k}(x+1)^{n-k}(x-1)^k=T_n(x)$$

Sometimes the coefficient in the Rodrigues formula is written differently

$$\frac{(-1)^n}{(2n-1)!!}=\frac{(-2)^n n!}{(2n)!}$$

Indeed

$$\frac{(-2)^n n!}{(2n)!}=\frac{(-1)^n 2^n n!}{(2n\dots(n+1))n!}=\frac{(-1)^n 2^n n!}{2n(2n-2)\dots 2\cdot(2n-1)!!}=\frac{(-1)^n 2^n n!}{2^n\cdot n!(2n-1)!!}=\frac{(-1)^n}{(2n-1)!!}$$

Minimal property of T_n.

Minimal property of T_n.

Theorem. Let p be an n-th degree polynomial with leading coefficient 1. Then

$$2^{1-n}=\max_{-1\le x\le 1}|2^{1-n}T_n(x)|\le\max_{-1\le x\le 1}|p(x)|$$

Proof. On $[-1,1]$ $2^{1-n}T_n(x)$ assumes its extremal values at points $y_k=\cos\frac{k\pi}{n}$

$$2^{1-n}T_n(y_k)=2^{1-n}(-1)^k \qquad k=0,1,\dots,n$$

Suppose by contradiction that there exists a monic polynomial $\widetilde{p}\in\mathbb{P}_n$ such that

$$\max_{-1\le x\le 1}|\widetilde{p}(x)|<2^{1-n}$$

Consider $Q(x)=2^{1-n}T_n(x)-\widetilde{p}(x)$ which is a polynomial of degree $n-1$. We have

$$Q(y_k)=(-1)^k 2^{1-n}-\widetilde{p}(y_k), \qquad k=0,1,\dots,n$$

Due to the fact that we assumed that the norm of $\widetilde{p}$ is small

$$\operatorname{sign}(Q(y_k))=(-1)^k$$

Due to the intermediate value theorem Q has n zeros and hence $Q\equiv 0$. Thus $2^{1-n}T_n=\widetilde{p}$. But this would imply that

$$\max_{-1\le x\le 1}|\widetilde{p}(x)|=2^{1-n}$$

contradicting the assumption that the norm of $\widetilde{p}$ is small.

Paszkowski p 18, Montgomery p 239-388, Kostrykin II p163

* Bernstein polynomials

- $n \geq 1, n \geq 0$

$$B_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}, \quad k = 0,\dots,n$$

$$B_{0,0}(x) \equiv 1$$

- $n = 1$

$B_{1,0}(x) = 1-x \qquad B_{1,1}(x) = x$

- $n = 2$

$B_{2,0}(x) = (1-x)^2 \qquad B_{2,1}(x) = 2(1-x)x \qquad B_{2,2}(x) = x^2$

* Properties:

$D_n : [0,1]$

- $B_{n,k}(x) \geq 0 \qquad B_{n,k}(0) = 0 \quad B_{n,k}(1) = 0$
- $\sum_{k=0}^{n} B_{n,k}(x) = 1, \quad x \in \mathbb{R}$ ← partition of unity
- $B_{n,k}(x) = B_{n,n-k}(1-x), \ k = 0,\dots,n$ Symmetry
- $\begin{cases} B_{n,k}(x) = x B_{n-1,k-1}(x) + (1-x) B_{n-1,k}, \quad k = \overline{0,n} \\ B_{n-1,-1}(x) \equiv 0 \qquad B_{n-1,n}(x) \equiv 0 \end{cases}$ Recurrence
- $B_{n,k}$ has exactly one max in $[0,1]$, attain at $x = \frac{k}{n}$
- $\{B_{n,k}\}_{0 \leq k \leq n}$ are basic $\mathbb{P}_n$

* Proof:

- Partition of unity

$$1 = (x + (1-x))^n = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k}$$

- Symmetry

$$\binom{n}{k} = \binom{n}{n-k}$$

- Recurrence

$$\binom{n}{k} \underset{\uparrow}{=} \binom{n-1}{k-1} + \binom{n-1}{k}$$

multiplied by $x^k (1-x)^{n-k}$

- Basic: Why are they independent?

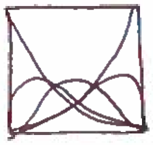
Consider the linear combination $\sum_{k=0}^{n} b_k B_{n,k}(x) = 0$

We want to show that $b_0 = b_1 = \dots = b_n = 0$

$\sum_{l=0}^{n} b_l B_{n,l}(x) = 0$, want to show that $b_0 = b_1 = \dots = b_n = 0$

- $x = 1 \Rightarrow \underset{\underset{0}{\shortparallel}}{\text{LHS}} = \sum_{l=0}^{n} b_l B_{n,l}(1) = b_n B_{n,k}(1) = b_n$
- $B_{n,l}$, $l = \overline{0,\dots,n-1}$ is divisable by $(1-x)$

  $0 = \sum_{l=0}^{n-1} b_l B_{n,l}(x)$ divide by $(1-x)$, then set $x=1$ $b_{n-1} = 0$



* About mcx

Put $x = \frac{t-a}{b-a} \Rightarrow$ can be extended into any interval $[a,b]$

* Bernstein polynomials associated with $f \in C[0,1]$

- $\underset{\underset{\text{parameter here}}{\uparrow}}{B_n}(f)(x) = \sum_{l=0}^{n} f\left(\frac{l}{n}\right)\binom{n}{l} x^l (1-x)^{n-l}$ → may be closed to f
- $\boxed{\underset{\underset{\text{Bernstein poperator}}{\uparrow}}{B_n} : C[0,1] \longrightarrow \mathbb{P}_n}$

* Properties:

- B is linear $B_n(f+g)(x) = B_n(f)(x) + B_n(g)(x)$
- $B_n(f)(x) \xrightarrow{n\to\infty} f$
- B is monotone

  $f \le h \Rightarrow B_n(f)(x) \le B_n(h)(x)$
- We want to show $B_n(f) = f$ if $f(t) = t^j$, $j = 0,1$

  We will also need to evaluate $B_n(f)$ for $f(t) = t^2$
- $B_n(1)(x) = \sum_{l=0}^{n} \binom{n}{l} x^l (1-x)^{n-k} \underset{\text{above property}}{=\!=\!=} 1$ (id 1)

Replace n by (n-1), next multiply by nx

$$nx = \sum_{l=0}^{n-1} n\binom{n-1}{l} x^{l+1} (1-x)^{n-(k+1)}$$

$$= \sum_{l=0}^{n-1} (l+1)\binom{n}{l+1} x^{l+1} (1-x)^{n-(k+1)}$$

(id2)

$$= \sum_{s=1}^{n} s\binom{n}{s} x^s (1-x)^{n-s}$$

$$= \sum_{l=0}^{n} k\binom{n}{l} x^l (1-x)^{n-k}$$

$$\Rightarrow x = \sum_{l=0}^{n} \frac{l}{n}\binom{n}{l} x^l (1-x)^{n-l} = B_n(t)(x).$$ for $f(t)=t$

$B_n(f)(x) = f(x)$.

- Replace n by $(n-1)$ in id2

obtain

$$\sum_{l=0}^{n} l^2 \binom{n}{l} x^l (1-x)^{n-l} = (n-1)nx^2 + nx \qquad \text{(id3)}$$

dividing by n^2

$$B_n(t^2)(x) = x^2 + \frac{x-x^2}{n} \xrightarrow{n\to\infty} x^2$$

$$\xrightarrow{n\to\infty} x^2.$$

$f \geq 0$ then $H(f) \geq 0$

*** Theorem (Korowkin)** is also true for $C[0,1] \to C[0,1]$

Let $H_n(f) : C[a,b] \longrightarrow C[a,b]$ be a sequence of monotone operators

$$\|H_n(t^\nu)(x) - x^\nu\| \xrightarrow{n\to\infty} 0, \text{ for } \nu = 0,1,2$$

$\|H_n(f_\nu) - f_\nu\| \xrightarrow{n\to\infty} 0$, when $f = t^\nu, \nu = 0,1,2$

* Remind $\|f\| = \sup_{a\leq x\leq b} |f(x)|$

Then for any continuous function $f \in C[a,b]$

$$\|H_n(f) - f\| \xrightarrow{n\to\infty} 0$$

* Proof Let $t, x \in [a,b]$ ε

x fix t variable

NTP $\varepsilon > 0, \exists \delta$ s.t $|f(t) - f(x)| \leq \frac{\varepsilon}{2} + \frac{2M}{\delta^2}(t-x)^2$, $\underline{M = \|f\|}$

NTP $f(x) - \frac{\varepsilon}{2} - \frac{2M}{\delta^2}(t-x)^2 \leq f(t) < f(x) + \frac{\varepsilon}{2} + \frac{2M}{\delta^2}(t-x)^2$ $(*)$

$|A| < b$

$-b < A < b$

$H(-b) \leq H(A) < H(b)$

$|H(A)| < H(b)$

- f is uniformly continuous on $[a,b]$.

$\Rightarrow \forall \varepsilon > 0, \exists \delta > 0, |t-x| < \delta$ then $|f(t) - f(x)| \leq \frac{\varepsilon}{2}$ $\Rightarrow$ done $(*)$

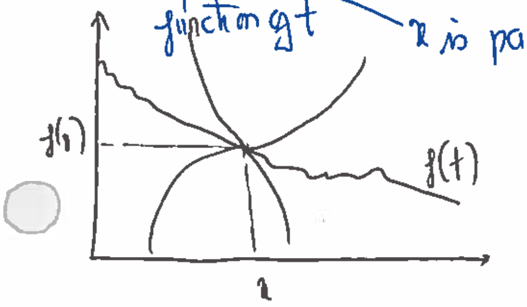
- If $t - x > \delta$, $|f(t) - f(x)| \leq |f(t)| + |f(x)| \leq 2M$

$\Rightarrow |f(t) - f(x)| \leq 2M = 2M\frac{\delta^2}{\delta^2} \leq 2M\frac{(t-x)^2}{\delta^2}$ $\Rightarrow (*)$

function of t x is parameter.

- what is good for f

$\Rightarrow$ good for monotone operator.



* Next step : Apply H to f

- $\left|H_n(f)(x) - \boxed{f(x)} H_n(1)(x)\right| \leq \frac{\varepsilon}{2} H_n(1)(x) + \frac{2M}{\delta^2} H_n((t-x)^2)$ $(**)$

- $\left|H_n(f)(x) - f(x)\right| \leq \underbrace{\left|H_n(f)(x) - f(x) H_n(1)(x)\right|} + \left|f(x) H_n(1)(x) - 1\right|$

$\uparrow$ want to prove that $H_n f \to f$

$\overset{(**)}{\leq} \frac{\varepsilon}{2} H_n(1)(x) + \frac{2M}{\delta^2} H_n((t-x)^2) + |f(x) H_n(1)(x) - 1|$

$= \frac{\varepsilon}{2}(H_n(1)(x) - 1 + 1) + |f(x) H_n(1)(x) - 1| + \frac{2M}{\delta^2}\Big[(x^2 H_n(1)(x) - x^2) + 2(x^2 - x H_n(t)(x))\Big] + (H_n(t^2)(x) - x^2)$

- We then will prove that

$\|H_n(f)(x) - f(x)\| \leq \frac{\varepsilon}{2} + \left(\frac{\varepsilon}{2} + \|f\| + \frac{2M}{\delta^2}\|x\|^2\right)\|H_n(1)(x) - 1\| +$

$+ \frac{4M}{\delta^2}\|x\| \, \|H_n(t)(x) - x\| + \frac{2M}{\delta^2}\|H_1(t^2)(x - x^2)\|$

So when $n > N$

$\|H_n(f)(x) - f(x)\| \leq \varepsilon$

$\leq \underbrace{\frac{\varepsilon}{2} + \frac{\varepsilon}{2}|H_n(1)(x) - 1| + |f(x)|\,|H_n(1)(x) - 1| + \frac{2M}{\delta^2}|x^2 H_n(1)(x) - x^2|} +$

$\frac{4M}{\delta^2}|x^2 - x H_n(t)(x)| + \frac{2M}{\delta}(H_n(t^2)(x) - x^2)$

~~$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2}\underbrace{|H_n(1)(x) - 1|}_{\text{eventually} \to 0} + |f(x)|\,|H_n(1)(x) - 1| + \frac{2M}{\delta^2}$~~

$\leq \frac{\varepsilon}{2} + \left(\frac{\varepsilon}{2} + \|f\| + \frac{2M}{\delta^2}\|x\|^2\right)\|H_n(1)(x) - 1\| + \frac{4M}{\delta^2}\|x\|\,\|H_n(t)(x) - x\| + \frac{2M}{\delta^2}\|H_1(t^2)(x - x^2)\|$

for n large enough, δ small enough

* Based on the theorem $\|D_n(f) - f\| \xrightarrow[n\to\infty]{} 0$

+ Using Bernstein polynomials to construct p. Bernstein curves.

Let $\lambda_0,\dots,\lambda_n \in \mathbb{R}^d$ be given (n+1) points ← called control points

The Bernstein curve of degree n, $n \geq 1$ is a parametric curve

$C^n_\lambda : [0,1] \longrightarrow \mathbb{R}^d$ given by

$C^n_\lambda(t) = \sum_{k=0}^{n} \lambda_k B_{n,k}(t)$, where $\lambda = (\lambda_0, \dots, \lambda_n)$ ← keep the order

$\in (0,1)$ polynomial of degree n

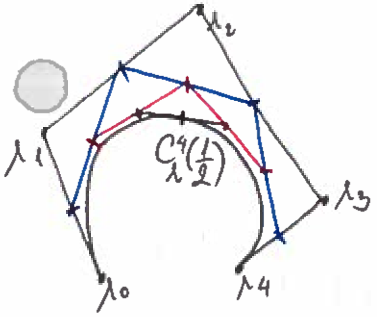
* We have

- $C^n_\lambda(0) = \lambda_0$ $\quad C^n_\lambda(1) = \lambda_n$

because $C^n_\lambda(0) = \sum_{k=0}^{n} \lambda_0 B_{0,k}(t) = \lambda_0 \underbrace{\sum_{k=0}^{n} B_{0,k}(t)}_{=1} = \lambda_0$

$C^n_\lambda(1) = \sum$

* De Casteljau algorithm for computing $C^n_\lambda(\overset{\in(0,1)}{t})$ (ex: $C^4_\lambda(\frac{1}{2})$)



n = 4. construct a polynomial of degree 4 in t

we compute the value at specific t

we don't compute the parameter of the polynomial

- Algorithm: we construct a sequence of vectors (better than using $C^n_\lambda(t)$)

$\lambda_0^{(1)} = (1-t)\lambda_0 + t\lambda_1$, ... , $\lambda_{n-1}^{(1)} = (1-t)\lambda_{n-1} + t\lambda_n$

$\lambda_0^{(2)} = (1-t)\lambda_0^{(1)} + t\lambda_1^{(1)}$, ... , $\lambda_{n-2}^{(2)} = (1-t)\lambda_{n-2}^{(1)} + t\lambda_{n-1}^{(1)}$

⋮

$\lambda_0^{(n)} = (1-t)\lambda_0^{(n-1)} + t\lambda_1^{(n-1)}$

We define $\boxed{d^{(n)}_\lambda(t) = \lambda_0^{(n)}}$

We want to show that $d^n_\lambda(t) = C^{\{n\}}_\lambda(t)$ to show that the Bernstein curve

* Now show that

$\boxed{\text{Lemma: } C^n_\lambda(t) = d^n_\lambda(t).}$

Remind $\binom{n}{k} = \binom{n}{n-k}$

$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$

* Induction prove

- $\underline{n = 1}$:

$\lambda = (\lambda_0, \lambda_1)$

$$C^1_\lambda(t) = \lambda_0 \underbrace{B_{1,0}(t)}_{(1-t)} + \lambda_1 \underbrace{B_{1,1}(t)}_{t} = (1-t)\lambda_0 + t\lambda_1$$

- Assume lemma holds for $(n-1)$: $C^{n-1}_\lambda(t) = d^{(n-1)}_\lambda(t)$
- ~~Consider when $i = (n)$~~

Let $\lambda_- = (\lambda_0, \dots, \lambda_{n-1})$ $\quad \lambda_+ = (\lambda_1, \dots, \lambda_n)$

$\lambda_0^{(n-1)} = C^{n-1}_{\lambda_-}$ $\quad \lambda_1^{(n-1)} = C^{n-1}_{\lambda_+}$

- We need to show $C^{(n)}_\lambda(t) = d^n_\lambda(t) \Leftrightarrow$ NTS $\sum_{k=0}^{n} \lambda_k B_{n,k}(t) = \lambda_0^{(n)}$

$$\text{LHS} = \sum_{k=0}^{n} \lambda_k \binom{n}{k} t^k (1-t)^{n-k} \quad \text{• RHS} = \lambda_0^{(n)} = (1-t)\lambda_0^{(n-1)} + t\lambda_1^{(n-1)}$$

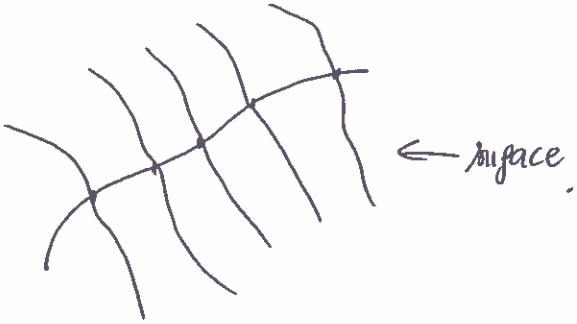
$$= (1-t)\sum_{k=0}^{\boxed{n-1}} \lambda_k \binom{n-1}{k} t^k (1-t)^{n-1-k} + t\sum_{k=0}^{\boxed{n-1}} \lambda_{k+1} \binom{n-1}{k} t^k (1-t)^{n-1-k}.$$

$$= \underbrace{\sum_{k=0}^{n-1} \lambda_k \binom{n-1}{k} t^k (1-t)^{n-k}} + \sum_{k=0}^{n-1} \lambda_{k+1} \binom{n-1}{k} t^{k+1} (1-t)^{n-1-k}.$$

$$= \sum_{k=0}^{n-1} \lambda_k \binom{n-1}{k} t^k (1-t)^{n-k} + \sum_{k=1}^{n} \lambda_k \binom{n-1}{k-1} t^k (1-t)^{n-k}.$$

$$= \sum_{k=1}^{n-1} \lambda_k \underbrace{\left[\binom{n-1}{k} + \binom{n-1}{k}\right]}_{\binom{n}{k}} t^k (1-t)^{n-k} + \lambda_0 (1-t)^n + \lambda_n t^n$$

$$= \sum_{k=0}^{n} \lambda_k \binom{n}{k} t^k (1-t)^{n-k}$$



* Numerical diggerentiation

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h}$$

$$f'(x) \simeq L'(x)$$

$$L(x) = \frac{x - x_1}{-h} f(x_0) + \frac{x - x_0}{h} f(x_1) \qquad x_0,\ x_1 = x_0 + h$$

$$f(x) = L(x) + (x-x_0)(x-x_1)\frac{f''(\xi)}{2!} = L(x) + R(x).$$

$$f'(x) = L'(x) + R'(x)$$

$$f'(x) = \frac{f(x_0+h) - f(x_0)}{h} + \frac{1}{2}(x-x_0)(x-x_1)\frac{d}{dx}\left(f''(\xi_x)\right) + \frac{1}{2}\left(2x - (x_0+x_1)\right) f''(\xi_x)$$

$$\boxed{f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h} - \frac{h}{2} f''(\xi_x)}$$

$$\underset{\text{computable}}{\frac{1}{2h}\left(f(x_0+h) - f(x_0-h)\right)} = \underset{\text{want to}}{f'(x_0)} + \underset{\text{error}}{\frac{h^2}{6} f^{(3)}(x_0) + O(h^4)}$$

$$* \quad f'(x_0) = \frac{1}{h}\left[f(x_0+h) - f(x_0)\right] - \frac{h}{2} f''(\xi_x)$$

$$f(x_0 \pm h) = f(x_0) \pm h f''(x_0) + \frac{h^2}{2} f''(x_0) \pm \frac{h^3}{6} f''(x_0) + \frac{h^4}{4!} f^{(4)}(x_0) + O(h^5)$$

- Subtact $f(x_0+h)$ and $f(x_0-h)$

$$\frac{1}{2h}\left[f(x_0+h) - f(x_0-h)\right] = f'(x_0) + \frac{h^2}{2} f'''(x_0) + O(h^4)$$

- $F(h) = f'(x_0) + T_1 h^2 + O(h^4)$

Richardson's extrapolation

$$F(h) = T_0 + T_1 h^p + O(h^\lambda), \quad \lambda > p$$

We want to know T_0, $F(h)$ easily computable.

- Take $0 < b < 1$

Compute

$$F(h) = T_0 + T_1 h^p + O(h^\lambda)$$

$$\underline{F(bh) = T_0 + T_1 (bh)^p + O(h^\lambda)}$$

$$F(bh) - F(h) = T_1 (b^p - 1) + O(h^\lambda)$$

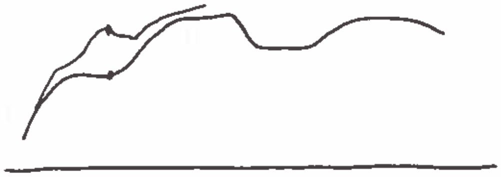
$$T_1 h^p = \frac{F(h) - F(bh)}{1 - b^p} + \mathcal{O}(h^r)$$

$$T_0 = F(h) + \frac{F(bh) - F(h)}{1 - b^p} + \mathcal{O}(h^2)$$

* Interpolation by pieuise polynomials

6.5 6.6 6.7

* Disadvantage of polynomial interpolation :



- If we change, even only one note, we have to redo the whole thing.

-

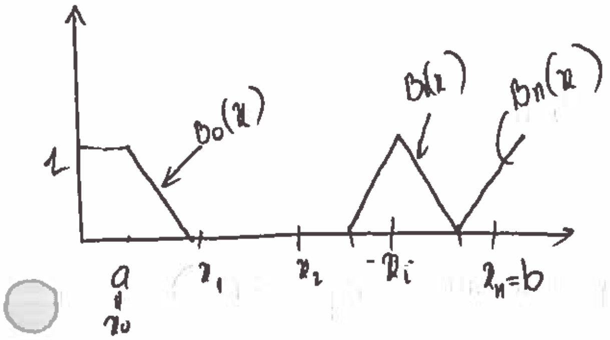
* Try to interpolate the notes locally.

* Example :

Consider a class of functions $\overset{\text{continuous}}{S_1^0(\Delta)}$ $\quad$ $\Delta \overset{\text{note}}{\doteq} \{x_0, \dots, x_n\}$

$a = x_0, x_1, x_2, \dots, x_n = b$

On each interval $[x_i, x_{i+1}]$, a function from $S_1^0(\Delta)$ { is a polynomial of degree ≤ 1. (degree 1.)
{ is continuous



- $$B_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & [x_{i-1}, x_i] \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & [x_i, x_{i+1}] \end{cases} \quad = 1 - \frac{|x - x_{i+1}|}{h_i}$$

$1 \leq i \leq n-1$

- $B_i(x_j) = \delta(i - j)$

$$\begin{cases} \delta(0) = 1 \\ \delta(\ell) = 0, \ \ell \neq 0 \end{cases}$$

$B_0(x) = \dfrac{x - x_1}{x_0 - x_1}$, on $[x_0, x_1]$

$B_n(x) = \dfrac{x - x_{n-1}}{x_n - x_{n-1}}$, on $[x_{n-1}, x_n]$

$\{B_i\}_{i=0}^{n}$ are basic in $S_0^1(\Delta)$!

Proof
- How many interval do I have

$$\underset{\text{coeff}}{2n} - \underset{\substack{\text{constraint} \\ \text{on each note}}}{(n-1)} = 2n - n + 1 = n + 1.$$

- We want to show that $\{B_i\}$ are linearly independent

$$\sum_{i=0}^{n} c_i B_i(x) = 0 \xrightarrow{\text{need}} c_i = 0$$

① $x = x_\ell \quad c_\ell B_\ell(x_\ell) = c_\ell = 0$

* We define the $S_1^0(\Delta)$ interpolant of f as $L(f)$

$L(f) = \sum_{i=0}^{n} f(x_i) B_i(x)$ a nonlocal interpolant

(on $[x_i, x_{i+1}]$, $L(f)$ depends only on $f(x_i)$ $f(x_{i+1})$)

- $L(f)\Big|_{[x_i, x_{i+1}]}(x) = f(x_i) + (x - x_i) f[x_i, x_{i+1}]$

Interpolation error :

$$|f(x) - L(f)(x)| = |(x - x_i)(x - x_{i+1}) f[x_i, x_{i+1}, x]|$$

$$\leq \left(\frac{h}{2}\right)^2 \sup_{a \leq x \leq b} \left|\frac{f''(\xi)}{2}\right|$$

$$\|f\| = \sup_{a \leq x \leq b} |f(x)|$$

$$\|Lf\| = \max_{0 \leq i \leq n} \sup_{x_i \leq x \leq x_{i+1}} |L(f)(x)| = \max_{0 \leq i \leq n} |f(x_i)| \leq \|f\|$$

$\inf_{g \in S_1^0(\Delta)} \|f - g\| \leftarrow$ we call the approximation error for approximating f by $S_1^0(\Delta)$

* Let $g \in S_1^0(\Delta)$ $L(g) = g$

$$\|f - L(f)\| = \|f - g - L(f) + L(g)\| \leq \|f - g\| + \|L(f - g)\| \leq 2\|f - g\|$$

$$\inf_{g \in S_1^0} \|f - g\| \leq \|f - L(f)\| \leq 2\|f - g\|$$

*

* Piecewise cubic function

- $S_3^1(\Delta) = \{ g \in C^1[a,b] , g|_{[x_{i-1}, x_{i+1}]} \in \mathbb{P}_3 \}$ ← first derivative continuous).

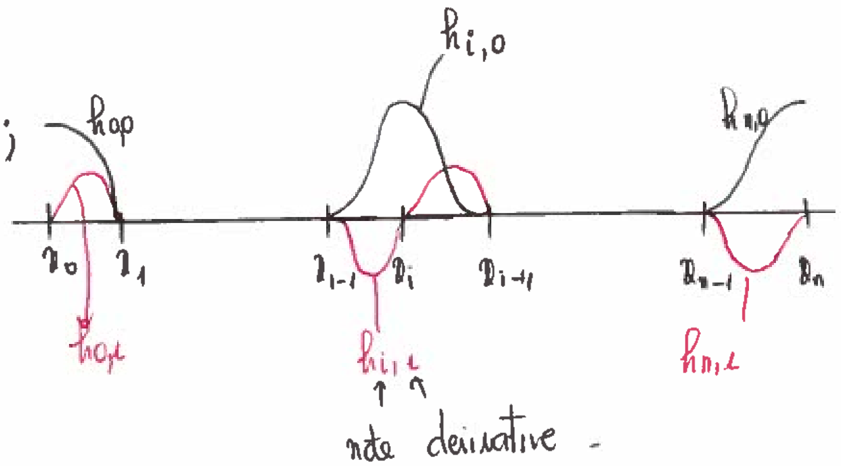
On each interval, g is given by 4 parameters

- ⇒ Dimension of the space is $\dim S_3^1(\Delta) = 2(n+1)$.

$4n - 2(n-1) = 4n - 2n + 2 = 2(n+1)$.

- If we prescribe the conditions:

$$\begin{cases} g(x_i) = f(x_i) & ; \ i = 0, \dots, n \ ; \\ g^{(1)}(x_i) = f^{(1)}(x_i) & ; \ i = 0, 1, \dots, n \end{cases}$$



these functions are the basic in $S_3^{(1)}(\Delta)$

* Idea of natural splines. $S_m^{(m-1)}(\Delta)$

We consider $S_m^{(m-1)}(\Delta)$

- The dimension of $S_m^{(m-1)}(\Delta)$ is $\dim \overset{\text{constraint}}{S_m^{(m-1)}} = n + m$

n interval, each interval ⇒ degree m,

$$\underbrace{n(m+1)}_{\text{parameter}} - \underbrace{(n-1)}_{\text{interior node}} \underbrace{m}_{\text{derivative}} = nm + n - nm + m = n + m$$

- What are the interpolation ~~and~~ condition ~~and~~ to determine such spline S

$S(x_i) = f(x_i) \quad i = \overline{0, \dots, n}$ $\qquad n + m - n - 1 = m - 1$.

Periodic spline $\quad S^{(l)}(a) = S^{(l)}(b) \quad , l = 1, \dots, m-1$

For $m = 2l - 1$

$\Rightarrow S^{(l+j)}(a) = S^{(l+j)}(b) = 0 \qquad j = 0, 1, \dots, l-2$

called a natural spline.

* The B_spline with degree 2 are Hermitian, that we studied before.

* Interpolation by piecwise polynomicals

* Disadvantage of polynomial interpolation:

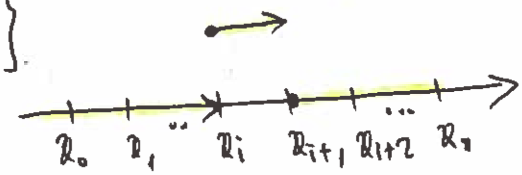
$\begin{cases} \text{if we change, even only on note, we have to redo the whole things.} \\ \text{the degree is high} \end{cases}$

* So we want to interpolate the notes locally. Also, we consider the notes $\Delta = \{x_0, \dots, x_n\}$

Define $S_k^r = \{ \text{all splines that} \begin{cases} \text{have degree } k \\ \text{have continuous } r\text{-th derivative} \end{cases} \}$

* B-splines of degree 0:

$S_1^0(\Delta) = \{ B_i^{(0)}(x),\ B_i^0(x) = \begin{cases} 1 & x_i \leq x \leq x_{i+1} \\ 0 & \text{otherwise} \end{cases} \}$.



- So we have $B_i(x)$ has some properties

1) $\chi_{B_i(x)} = [x_i, x_{i+1})$

2) $B_i(x) \geq 0, \forall i, \forall x$

3) $\sum_{i=-\infty}^{\infty} B_i^{(0)}(x) = 1$, for all x.

4) $\{B_i(x)\}$ forms a basic for $S_1^0(\Delta)$, all splines of degree 0,

* The function $B_i^{(0)}(x)$ are the starting point for a recursive definition of all of the higher degree B-splines.

$$B_i^{(k)}(x) = \frac{x - x_{i-1}}{\underbrace{x_{i-1+k} - x_{i-1}}} B_{i-1}^{(k-1)}(x) + \frac{x_{i+k} - x}{x_{i+k} - x_i} B_i^{(k-1)}(x), \quad k \geq 1.$$

Put $V_{i-1}^k(x) = \dfrac{x - x_{i-1}}{x_{i-1+k} - x_{i-1}}$ then $V_i^{(k)} = \dfrac{x - x_i}{x_{i+k} - x_i}$

$$1 - V_i^{(k)} = \frac{x_{i+k} - x_i - x + x_i}{x_{i+k} - x} = \frac{x_{i+k} - x}{x_{i+k} - x}.$$

Then we have

$$B_i^k(x) = V_{i-1}^k(x) B_{i-1}^{(k-1)}(x) + (1 - V_i^k)(x) B_i^{(k-1)}(x).$$

* B splines of degree 1.

continuous clenal

$S^0_1(\Delta) = \{ B_i^1(x)$ so that $B_i^1(x) \begin{cases} \text{is a polynomial of degree} \leq 1 \\ \text{is a continuous function} \end{cases}$.

degree

* We define $B_i^1(x)$ by the recursive formula.

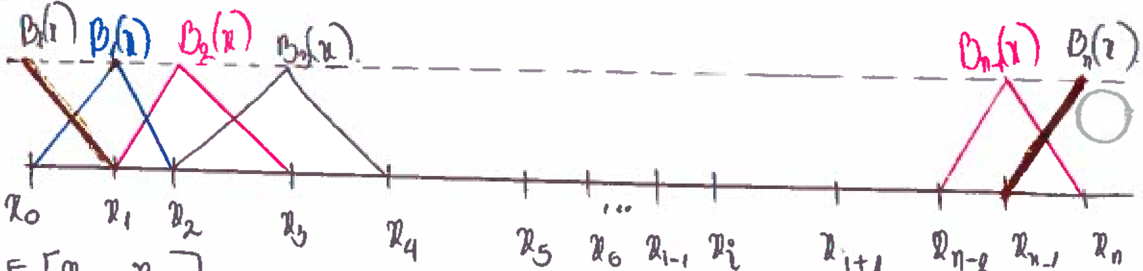
$$B_i^{(1)}(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} B^{(0)}_{i-1}(x) + \frac{x_{i+1} - x}{x_{i+1} - x_i} B_i(x).$$

$i = \overline{1, n-1}$

$$= \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i] \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}]. \\ 0, & \text{otherwise} \end{cases} \quad \text{for} \quad i = \overline{1, n-1}$$

- $B_0^{(1)}(x) = \dfrac{x_1 - x}{x_1 - x_0}$ $x \in [x_0, x_1]$

$B_n^{(1)}(x) = \dfrac{x - x_{n-1}}{x_n - x_{n-1}}$ for $x \in [x_{n-1}, x_n]$



* Then we have $B_i(x_j) = \delta(i-j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

① $\chi_{B_i(x)} = [t_{i-1}, t_{i+1}]$.

② $B_i^{(1)}(x) \geq 0, \forall x$

3) $B_i^{(1)}(x)$ is continuous and is differentiable at every point except t_{i-1}, t_i, x_{i+1}.

④ $\sum_{i=-\infty}^{+\infty} B_i^1(x) = 1$.

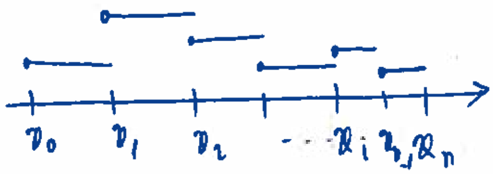
* Spline interpolation .

Given $(n+1)$ points $\{x_0, x_1, .., x_n\}$. $x_0 < x_1 < \cdots < x_n$

* A Spline S^k_m - a spline function of degree k, having notes $x_0, x_1, .., x_n$ satisfies .

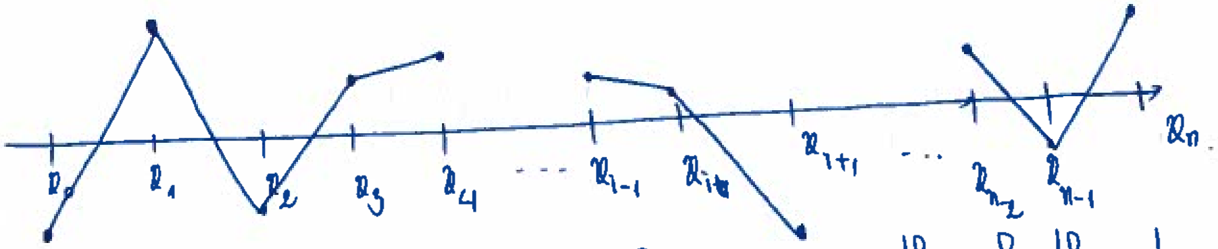
$$\begin{cases} S \text{ is a polynomial of degree } \le m \text{ on each interval } [x_i, x_{i+1}) \\ S \text{ has a continuous } k\text{st derivative on } [x_0, x_n] , \ (k \le (m-1)) \end{cases}$$

* Example of Splines of degree 0 .



$$S(x) = \begin{cases} S_0(x) = c_0 & x \in [x_0, x_1) \\ S_1(x) = c_1 & x \in [x_1, x_2) \\ \vdots & \vdots \\ S_{n-1}(x) = c_{n-1} & x \in [x_{n-1}, x_n) \end{cases}$$

* Note that when we want to find $S_1(x)$



- Consider $(a, f(a))$, $(b, f(b))$, the line that goes through these two points has equation:

$$y = f(a) + (f(b) - f(a)) \frac{x - x_a}{x_b - x_a} = f(a)\left[1 - \frac{x - x_a}{x_b - x_a}\right] + f(b)\left[\frac{x - x_a}{x_b - x_a}\right] =$$

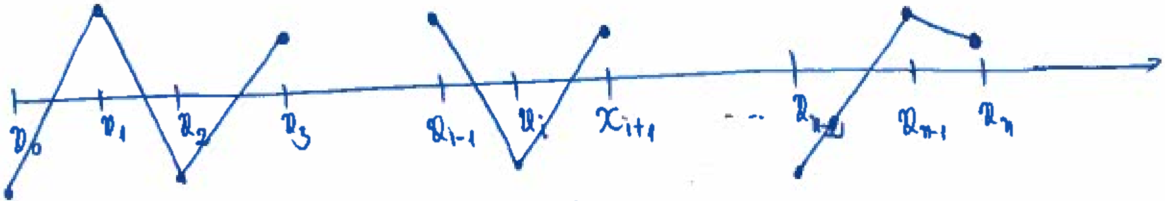
$$= f(a)\frac{x_b - x}{x_b - x_a} + f(b)\frac{x - x_a}{x_b - x_a} = f(a)\frac{x - x_b}{x_a - x_b} + f(b)\frac{x - x_a}{x_b - x_a}$$

To sum up,

The line goes through $(x_a, f(a))$, $(x_b, f(b))$ has equation

$$f(x) = f(a)\frac{x - x_b}{x_a - x_b} + f(b)\frac{x - x_a}{x_b - x_a}.$$

=



Then the Spline interpolation has equation .

$$f(x) = \begin{cases} f(x_0)\,\dfrac{x-x_1}{x_0-x_1} + f(x_1)\,\dfrac{x-x_0}{x_1-x_0} & , x\in[x_0, x_1] \\ f(x_1)\,\dfrac{x-x_2}{x_1-x_2} + f(x_2)\,\dfrac{x-x_1}{x_2-x_1} & , x\in[x_1, x_2] \\ \vdots & \\ f(x_{n-1})\,\dfrac{x-x_n}{x_{n-1}-x_n} + f(x_n)\,\dfrac{x-x_{n-1}}{x_n-x_{n-1}}. & , x\in[x_{n-1}, x_n] \end{cases}$$

$$f(x_2)\,\frac{x-x_3}{x_2-x_3} + f(x_3)\,\frac{x-x_2}{x_3-x_2} \qquad x\in[x_2, x_3].$$

Natural cubic splines

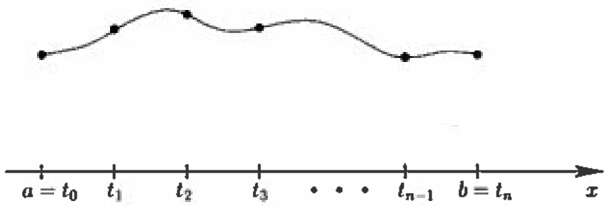
Arne Morten Kvarving

Department of Mathematical Sciences
Norwegian University of Science and Technology

October 21 2008

Motivation

- We are given a "large" dataset, i.e. a function sampled in many points.
- We want to find an approximation in-between these points.
- Until now we have seen one way to do this, namely high order interpolation - we express the solution over the whole domain as one polynomial of degree N for $N+1$ data points.



Motivation

- Let us consider the function

$$f(x) = \frac{1}{1+x^2}.$$

Known as Runge's example.

- While what we illustrate with this function is valid in general, this particular function is constructed to really amplify the problem.

Motivation

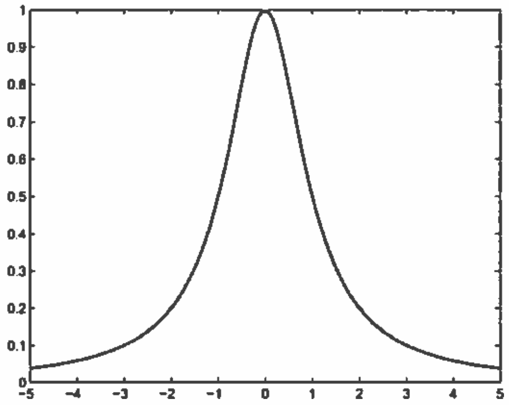


Figure: Runge's example plotted on a grid with 100 equidistantly spaced grid points.

Motivation

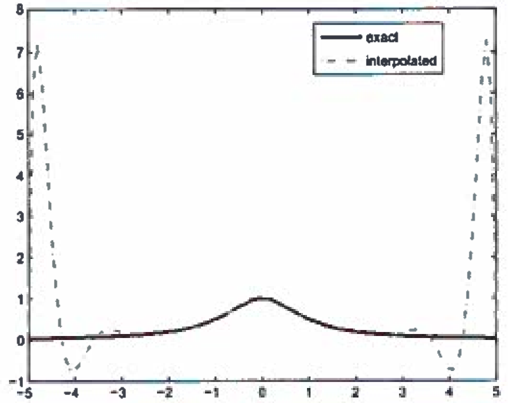
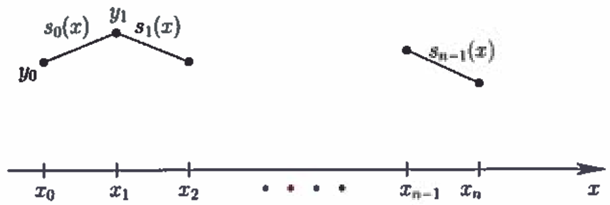


Figure: Runge's example interpolated using a 15th order polynomial based on equidistant sample points.

Motivation

- It turns out that high order interpolation using a global polynomial often exhibit these oscillations hence it is "dangerous" to use (in particular on equidistant grids).
- Another strategy is to use piecewise linear interpolation. For instance, piecewise linear interpolation.



Motivation

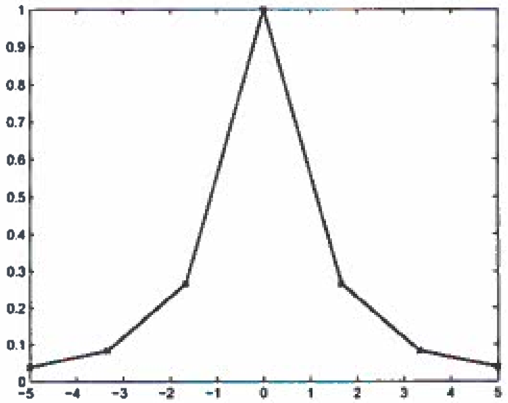
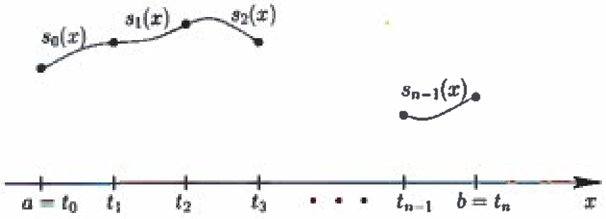


Figure: Runge's example interpolated using piecewise linear interpolation. We have used 7 points to interpolate the function in order to ensure that we can actually see the discontinuities on the plot.

A better strategy - spline interpolation

- We would like to avoid the Runge phenomenon for large datasets ⇒ we cannot do higher order interpolation.
- The solution to this is using piecewise polynomial interpolation.
- However piecewise linear is not a good choice as the regularity of the solution is only C^0.
- These desires lead to splines and spline interpolation.



Splines - definition

A function $S(x)$ is a spline of degree k on $[a, b]$ if

- $S \in C^{k-1}[a, b]$.
- $a = t_0 < t_1 < \cdots < t_n = b$ and

$$S(x) = \begin{cases} S_0(x), & t_0 \le x \le t_1 \\ S_1(x), & t_1 \le x \le t_2 \\ \vdots & \\ S_{n-1}(x), & t_{n-1} \le x \le t_n \end{cases}$$

where $S_i(x) \in \mathbb{P}^k$.

Cubic spline

$$S(x) = \begin{cases} S_0(x) = & a_0x^3 + b_0x^2 + c_0x + d_0, \qquad t_0 \le x \le t_1 \\ \vdots & \\ S_{n-1}(x) = & a_{n-1}x^3 + b_{n-1}x^2 + c_{n-1}x + d_{n-1}, \;\; t_{n-1} \le x \le t_n. \end{cases}$$

which satisfies

$$S(x) \in C^2[t_0, t_n] : \left. \begin{matrix} S_{i-1}(x_i) = S_i(x_i) \\ S'_{i-1}(x_i) = S'_i(x_i) \\ S''_{i-1}(x_i) = S''_i(x_i) \end{matrix} \right\} , i = 1, 2, \cdots , n-1.$$

we need to find

Cubic spline - interpolation

Given $(x_i, y_i)_{i=0}^n$. Task: Find $S(x)$ such that it is a cubic spline interpolant.

we need to compute 3(n-1) coefficient.

- The requirement that it is to be a cubic spline gives us $3(n-1)$ equations. ~~degree (knotes-1)~~
- In addition we require that

$$S(x_i) = y_i, \qquad i = 0, \cdots , n$$

which gives $n+1$ equations.

- This means we have $4n - 2$ equations in total.
- We have $4n$ degrees of freedom $(a_i, b_i, c_i, d_i)_{i=0}^{n-1}$.
- Thus we have 2 degrees of freedom left.

Cubic spline - interpolation

We can use these to define different subtypes of cubic splines:

- $S''(t_0) = S''(t_n) = 0$ - natural cubic spline.
- $S'(t_0), S'(t_n)$ given - clamped cubic spline.
- $$\left. \begin{matrix} S_0'''(t_1) = S_1'''(t_1) \\ S''_{n-2}(t_{n-1}) = S'''_{n-1}(t_{n-1}) \end{matrix} \right\} \text{ - Not a knot condition (MATLAB)}$$

Natural cubic splines

Task: Find $S(x)$ such that it is a natural cubic spline.

- Let $t_i = x_i, i = 0, \cdots, n$.
- Let $z_i = S''(x_i), i = 0, \cdots, n$. This means the condition that it is a natural cubic spline is simply expressed as $z_0 = z_n = 0$.
- Now, since $S(x)$ is a third order polynomial we know that $S''(x)$ is a linear spline which interpolates (t_i, z_i).
- Hence one strategy is to first construct the linear spline interpolant $S''(x)$, and then integrate that twice to obtain $S(x)$.

Natural cubic splines

- The linear spline is simply expressed as

$$S_i''(x) = z_i \frac{x - t_{i+1}}{t_i - t_{i+1}} + z_{i+1} \frac{x - t_i}{t_{i+1} - t_i}.$$

- We introduce $h_i = t_{i+1} - t_i, i = 0, \cdots, n$ which leads to

$$S''(x) = z_{i+1} \frac{x - t_i}{h_i} + z_i \frac{t_{i+1} - x}{h_i}.$$

- We now integrate twice

$$\begin{aligned} S_i(x) = {} & \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \frac{z_i}{6h_i}(t_{i+1} - x)^3 \\ & + C_i (x - t_i) + D_i (t_{i+1} - x) . \end{aligned}$$

Natural cubic splines

- Interpolation gives:

$$S_i(t_i) = y_i \Rightarrow \frac{z_i}{6} h_i^2 + D_i h_i = y_i, i = 0, \cdots, n.$$

- Continuity yields:

$$S_i(t_{i+1}) = y_{i+1} \Rightarrow \frac{z_{i+1}}{6} h_i^2 + C_i h_i = y_{i+1}.$$

Natural cubic splines

- We insert these expressions to find the following form of the system

$$\begin{aligned} S_i(x) = {} & \frac{z_{i+1}}{6h_i}(x - t_i)^3 + \frac{z_i}{6h_i}(t_{i+1} - x)^3 \\ & + \left(\frac{y_{i+1}}{h_i} - \frac{z_{i+1}}{6} h_i \right)(x - t_i) \\ & + \left(\frac{y_i}{h_i} - \frac{h_i}{6} z_i \right)(t_{i+1} - x) . \end{aligned}$$

- We then take the derivative.

Natural cubic splines

- The derivative reads

$$S_i'(x) = \frac{z_{i+1}}{2h_i}(x - t_i)^2 - \frac{z_i}{2h_i}(t_{i+1} - x)^2$$
$$+ \underbrace{\frac{1}{h_i}(y_{i+1} - y_i)}_{b_i} - \frac{h_i}{6}(z_{i+1} - z_i)\,.$$

- In our abscissas this gives

$$S_i'(t_i) = -\frac{1}{2}z_i h_i + b_i - \frac{h_i}{6}z_{i+1} + \frac{1}{6}h_i z_i$$
$$S_i'(t_{i+1}) = \frac{z_{i+1}}{2}h_i + b_i - \frac{h_i}{6}z_{i+1} + \frac{1}{6}h_i z_i$$
$$S_{i-1}(t_i) = \frac{1}{3}z_i h_{i+1} + \frac{1}{6}h_{i-1}z_{i-1} + b_{i-1}$$
$$S_i'(t_i) = S_{i-1}(t_i) \Rightarrow$$
$$6(b_i - b_{i-1}) = h_{i-1}z_{-1} + 2(h_{i-1} + h_i)z_i + h_i z_{i+1}\,.$$

Natural cubic splines - algorithm

This means that we can find our solution using the following procedure:

- First do some precalculations

$$\begin{aligned}
h_i &= t_{i+1} - t_i, & i &= 0, \cdots, n-1\\
b_i &= \frac{1}{h_i}(y_{i+1} - y_i)\,, & i &= 0, \cdots, n-1\\
v_i &= 2(h_{i-1} + h_i)\,, & i &= 1, \ldots, n-1\\
u_i &= 6(b_i - b_{i-1})\,, & i &= 1, \cdots, n-1\\
z_0 &= z_n = 0
\end{aligned}$$

Natural cubic splines - algorithm

- Then solve the tridiagonal system

$$\begin{bmatrix} v_1 & h_1 & & & & \\ h_1 & v_2 & h_2 & & & \\ & h_2 & v_3 & h_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & h_{n-2} \\ & & & & h_{n-2} & v_{n-1} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_{n-2} \\ z_{n-1} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_{n-2} \\ u_{n-1} \end{bmatrix}.$$

Natural cubic splines - example

- Given the dataset

i	0	1	2	3
x_i	0.9	1.3	1.9	2.1
y_i	1.3	1.5	1.85	2.1
$h_i = x_{i+1} - x_i$	0.4	0.6	0.2	
$b_i = \frac{1}{h_i}(y_{i+1} - y_i)$	0.5	0.5833	1.25	
$v_i = 2(h_{i-1} + h_i)$		2.0	1.6	
$u_i = 6(b_i - b_{i-1})$		0.5	4	

- The linear system reads

$$\begin{bmatrix} 2.0 & 0.4 \\ 0.4 & 1.6 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.4 \end{bmatrix}$$

Natural cubic splines - example

- We find $z_0 = 0.5, z_1 = 0.125$. This gives us our spline functions

$$S_0(x) = 0.208\,(x-0.9)^3 + 3.78\,(x-0.9) + 3.25\,(1.3-x)$$
$$S_1(x) = 0.035\,(x-1.3)^3 + 0.139\,(1.9-x)^3 + 0.664 - 0.62x$$
$$S_2(x) = 0.104\,(x-1.9)^3 + 10.5\,(x-1.9) + 9.25\,(2.1-x)$$

Gaussian elimination of tridiagonal systems

- Assume we are given a general tridiagonal system

$$\begin{bmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

- First elimination (second row) yields

$$\begin{bmatrix} d_1 & c_1 & & & \\ 0 & \tilde{d}_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & & & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix}, \begin{bmatrix} b_1 \\ \tilde{b}_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}, \qquad \begin{aligned} \tilde{d}_2 &= d_2 - \frac{a_1}{d_1}c_1 \\ \tilde{b}_2 &= b_2 - \frac{a_1}{d_1}b_1 \end{aligned}$$

Gaussian elimination of tridiagonal systems

- This means that the elimination stage is

```
for $i = 2, \cdots, n$
    $m = a_{i-1}/d_{i-1}$
    $\tilde{d}_i = d_i - mc_{i-1}$
    $\tilde{b}_i = b_i - mb_{i-1}$
end
```

- And the backward substitution reads

```
$x_n = \tilde{b}_n/d_n$
for $i = n-1, \cdots, 1$
    $x_i = \left(\tilde{b}_i - c_i x_{i+1}\right)/\tilde{d}_i$
end
```

where $\tilde{b}_1 = b_1$.

Gaussian elimination of tridiagonal systems

- This will work out fine as long as $\tilde{d}_i \neq 0$.
- Assume that $|d_i| > |a_{i-1}| + |c_i|$ - i.e. diagonal dominance.
- For the eliminated system diagonal domiance means that

$$|\tilde{d}_i| < |c_i|.$$

- We now want to show that diagonal domiance of the original system implies that the eliminated system is also diagonal dominant.

Gaussian elimination of tridiagonal systems

- We now assume that $|\tilde{d}_{i-1}| > |c_{i-1}|$. This is obviously satisfied for $\tilde{d}_1 = d_1$.

$$\begin{aligned}|\tilde{d}_i| = |d_i - \frac{a_{i-1}}{\tilde{d}_{i-1}} c_{i-1}| &\geq |d_i| - \frac{|a_{i-1}|}{|\tilde{d}_{i-1}|}|c_{i-1}| \\ &> |a_{i-1} - |c_i| - |a_{i-1}| = |c_i|.\end{aligned}$$

- Hence the diagonal domiance is preserved which means that $\tilde{d}_i \neq 0$. The algorithm produces a unique solution.

Why cubic splines?

- Now to motivate why we use cubic splines.
- First, let us introduce a measure for the smoothness of a function:

$$\mu(f) = \int_a^b (f''(x))^2 \, dx. \qquad (1)$$

- We then have the following theorem

Theorem

Given interpolation data $(t_i, y_i)_{i=0}^n$. Among all functions $f \in C^2[a,b]$ which interpolates (t_i, y_i), the natural cubic spline is the smoothest, where smoothness is measured through (1).

Why cubic splines?

- We need to prove that

$$\mu(f) \geq \mu(S) \,\forall\, f \in C^2[a,b].$$

- Introduce

$$g(x) = S(x) - f(x), \qquad \begin{array}{l} g(x) \in C^2[a,b] \\ g(t_i) = 0, i = 0, \cdots, n. \end{array}$$

- Inserting this yields

$$\begin{aligned}\mu(f) &= \int_a^b (S''(x) - g''(x))^2 \, dx \\ &= \mu(S) + \mu(g) - 2\int_a^b S''(x) g''(x) \, dx\end{aligned}$$

Now since $\mu(g) > 0$, we have proved our result if we can show that

$$\int_a^b S''(x) g''(x) \, dx = 0.$$

Why cubic splines?

- We have that

$$\int_a^b S''(x) g''(x) \, dx = g'(x) S''(x)|_a^b - \int_a^b g'(x) S'''(x) \, dx$$

First part on the right hand side is zero since $z_0 = z_n = 0$. Second part we split in an integral over each subdomain

$$\begin{aligned}-\int_a^b g'(x) S'''(x) \, dx &= -\sum_{i=0}^{n-1} \int_{t^i}^{t^{i+1}} g'(x) S'''(x) \, dx \\ &= -\sum_{i=0}^{n-1} 6a_i \int_{t_i}^{t_{i+1}} g'(x) \, dx \\ &= -\sum_{i=0}^{n-1} 6a_i \, g(x)|_{t_i}^{t_{i+1}} = 0.\end{aligned}$$

Cubic spline result

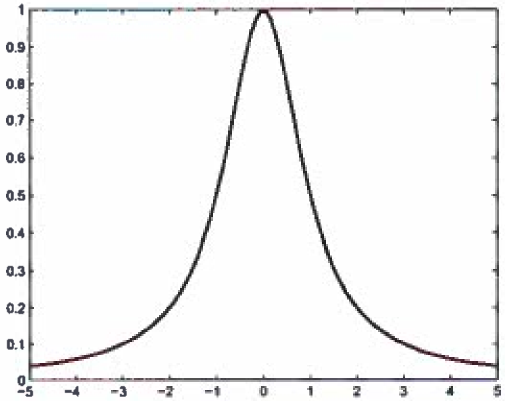


Figure: Runge's example interpolated using cubic spline interpolation based on 15 equidistant samples.

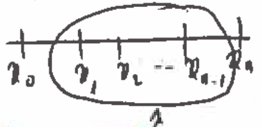
* Splines The motivation comes from Runge's example. $\Delta = \{ \overset{a}{\overset{\shortparallel}{x_0}}, \dots, \overset{b}{\overset{\shortparallel}{x_n}} \}$

$S_m^{\ell}(\Delta) = \{ S : S\big|_{[x_i, x_{i+1})} \in \mathbb{P}_m, i = 0, \dots, n-1 \quad S \in C^{\ell}(x_0, x_n) \}$

$\{ S_m^{\ell}(\Delta), S \in \mathbb{P}_m, S \in C^{\ell}(\Delta)$
$\ell \leq m-1$

- $\dim(S_m^{m-1}(\Delta)) = n+m$
- if $m = 3$, there are cubic splines.

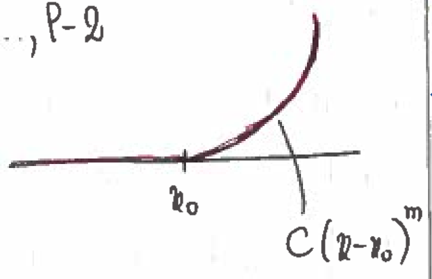
- 1) S is a periodic spline
  $S^{(\ell)}(a) = S^{(\ell)}(b)$ $\quad \boxed{\ell = \overline{1, \dots, m-1}}$ (continuity of ℓ^{th} derivative happens in all midde nodes



2) S natural spline
$m = 2\ell - 1$
$S^{(\ell+j)}(a) = S^{(\ell+j)}(b) = \boxed{0}$ $j = \overline{0, \dots, \ell-2}$

* Example: for the set of cubic spline $S_3^2 \in \mathbb{P}_3 \in C^1$
Then $S^{(\ell)}(a) = S^{(\ell)}(b), \forall \ell = \overline{1, 2}$
$S^{(2)}(a) \underset{\uparrow}{=} S^{(2)}(b) = 0$.
natural cubic spline

* Example
- The simplest spline of $S_{\boxed{m}}^{\boxed{m-1}}$



$S(x) \equiv 0$ if $x < x_0$

$$\left.\begin{array}{l} S^{(0)}(x_0) = 0 \\ S^{(1)}(x_0) = 0 \\ \vdots \\ S^{(m-1)}(x_0) = 0 \end{array}\right) C(x-x_0)^m$$

* Example: Fitting the spline for 3 points.

* Consider $S_2^1(\Delta)$ $\quad \boxed{\Delta = \{0, 1, 2\}}$

$$S_2^1(x) = \begin{cases} a_{0,2}x^2 + a_{0,1}x + a_{00} & x \in [0,1) \\ a_{1,2}x^2 + a_{1,1}x + a_{10} & x \in [1,2) \end{cases}$$

$(\text{degree}+1)(n-1)$.
$3 \times 2 = 6$.

We must compute 6 coefficients a_{ij}

- Continuity of S at $x = 1$ $\quad a_{0,2} + a_{01} + a_{00} = a_{12} + a_{11} + a_{10}$
- Continuity of S' at $x = 1$ $\quad 2a_{02} + a_{01} = 2a_{12} + a_{11}$
- Let $\begin{cases} S(0) = f(0) = f_0 \\ S(1) = f_1 \\ S(2) = f_2 \\ S'(0) = f_0' \end{cases}$ $\Rightarrow$ We have 6 equations
  However not all equation can be solve

$$\begin{cases} a_{00} = f_0 \\ a_{02} + a_{01} + a_{00} = f_1 \\ 4a_{12} + 2a_{01} + a_{00} = f_2 \\ a_{01} = f_0^{(1)} \end{cases}$$

So we have to solve the system of equations:

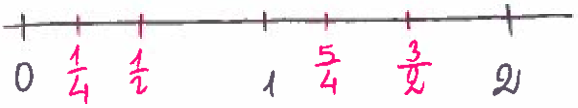
$$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 2 & 1 & 0 & -2 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 4 & 2 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{02} \\ a_{01} \\ a_{00} \\ a_{12} \\ a_{11} \\ a_{10} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ f_0 \\ f_1 \\ f_2 \\ f_0' \end{bmatrix}$$

invertible.

Case 2

- Match the spline values at four points $\frac{1}{4}, \frac{1}{2}, \frac{5}{4}, \frac{3}{2}$

We still need to find S $\begin{cases} \text{cont at } 1 \\ S' \text{cont at } 1 \end{cases}$



$$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 2 & 1 & 0 & -2 & -1 & 0 \\ \frac{1}{16} & \frac{1}{4} & 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{25}{16} & \frac{5}{4} & 1 \\ 0 & 0 & 0 & \frac{9}{4} & \frac{3}{4} & 1 \end{bmatrix} \begin{bmatrix} a_{02} \\ a_{01} \\ a_{00} \\ a_{12} \\ a_{11} \\ a_{10} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ f_{1/4} \\ f_{1/2} \\ f_{5/4} \\ f_{3/2} \end{bmatrix}$$

invertible

- Case 3 Take point $0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}$

$$\begin{bmatrix} 1 & 1 & 1 & -1 & -1 & -1 \\ 2 & 1 & 0 & -2 & -1 & 0 \\ \frac{1}{16} & \frac{1}{4} & 1 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{1}{64} & \frac{1}{4} & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \\ \\ \\ \\ \\ \\ \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ f_{1/4} \\ f_{1/2} \\ f_0 . \\ f_{1/8} \end{bmatrix}$$

$Ax = b$

$\|x\| \leq c\|b\|$

$b \to 0$ then $\begin{cases} \text{uniqueness} \\ \underset{\sim}{x} = \underset{\sim}{0} \end{cases}$

not invertible

not solvable

* Computation of the $\boxed{\text{cubic}}$ natural spline.

x_0
x_1
x_2
$\dots x_{n-1}$
x_n

we have to find n splines, each splines has 4 coeffici (degree+1)

$\Rightarrow$ We need $(\underbrace{\text{degree}+1}_{4})n = 4n$ coeffi

Let $x_0 < x_1 < x_2 < \dots < x_n$

Let k be a positive integers

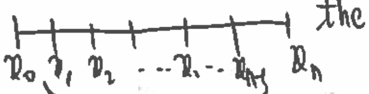
Find S of degree $k \geq 1$ with knots $x_0, \dots, x_n$ such that:

$\begin{cases} (1) \text{ on each } [x_i, x_{i+1}),\ S \text{ is a polynomial of degree} \leq k. \\ (2)\ S^{(k-1)}(x) \text{ is continuous on } [x_0, x_n] \end{cases}$

* For cubic natural spline $\begin{cases} S \text{ is a polynomial of degree} \leq 3. \quad S|_{[x_i, x_{i+1})} = S_i \in \mathbb{P}_3 \\ S^{(0)}, S^{(1)}, S^{(2)} \text{ are continuous on } [x_0, x_n] \end{cases}$

such S has 4n coefficients

- So now we need to find 4n coefficients.

+ In notes  the cond $S(x_0) = S(x_1) \Rightarrow$ We have 2n conds)

+ S' in nodes $x_1, \dots, x_{n-1}$ $\Rightarrow$ gives us $(n-1)$ conds.

+ S'' in nodos $x_1, \dots, x_{n-1}$ $\Rightarrow$ gives us $(n-1)$

we have $4(n-1)$ conditions

$\Rightarrow$ we need 2_degree of freedom

If we impose $S(x_i) = f_i$, $i = \overline{0, n}$ $\Rightarrow$ $(n+1)$ conds.

$\Rightarrow$ we need to impose 2 more conditions $\begin{cases} S^{(2)}(x_0) = 0 \\ S^{(2)}(x_n) = 0 \end{cases}$ or $\begin{cases} S^{(1)}(x_0) = f_0^1 \\ S^{(1)}(x_n) = f_n^1 \end{cases}$

* Call $S''(x_j) = M_j$: moment of S.

We will show how to determine the coefficients of S in terms of moments.

$h_{i+1} := x_{i+1} - x_i$

on $[x_i, x_{i+1})$ $S''(\cdot)$ is an affine function

- $$S''(x) = M_i \frac{x_{i+1} - x}{x_{i+1} - x_i} + M_{i+1} \frac{x - x_i}{x_{i+1} - x_i} = M_i \frac{x_{i+1} - x}{h_{i+1}} + M_{i+1} \frac{x - x_i}{h_{i+1}} \qquad (1)$$

- To determine S integrate

$$S'(x) = -M_i \frac{(x_{i+1} - x)^2}{2h_{i+1}} + M_{i+1} \frac{(x - x_i)^2}{2h_{i+1}} + A_i \qquad (2)$$

$$S(x) = M_i \frac{(x_{i+1} - x_i)^3}{6h_{i+1}} + M_{i+1} \frac{(x - x_i)^3}{6h_{i+1}} + A_i(x - x_i) + B_i \qquad (3)$$

* From $S(x_i) = f_i$, we find A_i, B_i

Set $x = x_i$ in (3)

$$f_i = S(x_i) = M_i \frac{(x_{i+1}-x_i)^3}{6h_{i+1}} = M_i \frac{h_i^2}{6} + B_i \qquad (4)$$

$$f_{i+1} = S(x_{i+1}) = M_{i+1} \frac{h_{i+1}^2}{6} + A_i h_{i+1} + B_i \qquad (5)$$

$$\Rightarrow \begin{cases} B_i = f_i - M_i \dfrac{h_{i+1}^2}{6} \\ A_i = \dfrac{f_{i+1}-f_i}{h_{i+1}} - \dfrac{h_{i+1}}{6}(M_{i+1}-M_i) \end{cases}$$

If we have M_i, M_{i+1} $\Rightarrow$ we have A_i, B_i (6)

If we

* On $[x_i, x_{i+1})$, Let $S(x) = a_i + b_i(x-x_i) + c_i(x-x_i)^2 + d_i(x-x_i)^3$ (7)

We want to compute a_i, b_i, c_i, d_i in terms of M_i, f_i

~~From (2)~~ $f_i = S(x_i) = a_i$

$$\frac{M_i}{2} = c_i$$

- From (2)

$$b_i = S'(x_i) = -M_i \frac{h_{i+1}}{2} + A_i = \frac{f_{i+1}-f_i}{h_{i+1}} - 2\,\frac{M_i + M_{i+1}}{6}\,h_{i+1}$$

- From (7) and (1)

$$d_i = \frac{M_{i+1}-M_i}{6h_{i+1}}$$

* Now compute $\left|\begin{array}{l} M_i \text{ in terms of } f_0, \dots, f_n \\ M_0 = M_n = 0 \end{array}\right.$

$$S'(x_i^-) = S'(x_i^+)$$

- From (2) and (6), (use (2) where A_i is computed by (6))

$$S'(x) = -M_i \frac{(x_{i+1}-x)^2}{2h_{i+1}} + M_{i+1} \frac{(x-x_i)^2}{2h_{i+1}} + \frac{f_{i+1}-f_i}{h_{i+1}} - \frac{h_{i+1}}{6}(M_{i+1}-M_i)$$

on $[x_i, x_{i+1}$ (8)

Impose $S'(x_i^-) = S'(x_i^+)$, on (8), we have

$$\frac{h_i}{6} M_{i-1} + \frac{h_i + h_{i+1}}{3} M_i + \frac{h_{i+1}}{6} M_{i+1} = \frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{i-1}}{h_i}$$

interior nodes. $i = \overline{1, n-1}$

•

$$\mu_i M_{i-1} + 2M_i + \lambda_i M_{i+1} = \delta_i$$

where $\mu_i = \frac{h_i}{h_i + h_{i+1}}$ $\lambda_i = \frac{h_{i+1}}{h_i + h_{i+1}}$ $\mu_i + \lambda_i = 1$.

$$\delta_i = \frac{6}{h_i + h_{i+1}} \left(\frac{f_{i+1} - f_i}{h_{i+1}} - \frac{f_i - f_{-1}}{h_i} \right)$$

So we have the equation.

$$\underbrace{\begin{bmatrix} 2 & \lambda_0 & & & & \\ \mu_1 & 2 & \lambda_1 & & & \\ & \mu_2 & 2 & \lambda_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & \mu_n & 2 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} \delta_0 \\ \delta_1 \\ \delta_1 \\ \vdots \\ \delta_{n-1} \\ \delta_n \end{bmatrix}}_{b}$$

Find M_i, $i = \overline{1, n-1}$ ← interior

Find A_i, B_i

Find a_i, b_i, c_i, d_i

* Lemma

If $T\underset{\sim}{x} = \underset{\sim}{b}$ then $\max\limits_i |x_i| \le \max\limits_i |b_i|$

this says that the system $T\underset{\sim}{x} = \underset{\sim}{0}$ has a unique solution
↑ existence

* Proof:

Let r be such that $|x_r| = \max\limits_i |x_i|$

• Take the r^{th} equation of $Tx = b$

$$\mu_r x_{r-1} + 2x_r + \lambda_r x_{r+1} = b_r$$

Then
$$\max b_i \ge |b_r| \ge 2|x_r| - \mu_r |x_{r-1}| - \lambda_r |x_{r+1}| \ge 2|x_r| - \mu_r |x_r| - \lambda_r |x_r| = (2 - \mu_r - \lambda_r)|x_r| = |x_r|$$

* Computing the natural cubic interpolating Spline

The system that we need to solve to find

$$\underbrace{\begin{bmatrix} 2 & \lambda_0 & & & & \\ \mu_1 & 2 & \lambda_1 & & & \\ & \mu_2 & 2 & \lambda_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & & \mu_n & 2 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} M_0 \\ \boxed{\begin{matrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \end{matrix}} \\ M_n \end{bmatrix}}_{\underset{\sim}{M}} = \underbrace{\begin{bmatrix} \delta_0 \\ \delta_1 \\ \\ \vdots \\ \delta_{n-1} \\ \delta_n \end{bmatrix}}_{\underset{\sim}{\delta}}$$

$M_0 = M_n = 0$

$\lambda_0 = 0$

$\mu_n = 0$

$\delta_0 = \delta_1 = 0$

$\mu_i = \dfrac{h_i}{h_i + h_{i+1}} \qquad \lambda_i = \dfrac{h_{i+1}}{h_i + h_{i+1}}$

$\mu_i + \lambda_i = 1, \quad i = \overline{1, n-1}$

* Extremal property of cubic spline interpolation $\quad x_0 = a \quad x_n = b$

Let $S \in N_3$ be such that $\begin{cases} S_i = f(x_i) & i = 0, 1, 2, .., n \\ S \text{ is affine for } & x \geq x_n,\ x \leq x_0 \end{cases}$

$\uparrow$ degree

*** Theorem:**

Let $g \in C^2(\mathbb{R})$ which interpolates f $\begin{cases} g(x_i) = f(x_i), \ i = \overline{0, n} \\ g''(a) = g''(b) = 0 \end{cases}$

We have

$$\int_a^b (g''(x))^2 dx \geq \int_a^b (S''(x))^2 dx$$

flatter → $f'' \to 0$

convex → f'' bigger

* **Proof:**

$$\int_a^b (g''(x) - S''(x))^2 dx = \int_a^b [g''(x)]^2 dx - 2\int_a^b S''(g'' - S'') dx - \int_a^b (S''(x))^2 dx$$

- $$\int_a^b S''(g'' - S'') dx \xlongequal[\text{by part}]{\text{integration}} \sum_{i=0}^{n-1} \underbrace{\int_{x_i}^{x_{i+1}} S''(g'' - S'') dx}$$

$$\sim = \int_{x_i}^{x_{i+1}} S''(g'' - S'') dx = S''(g' - S')\Big|_{x_i}^{x_{i+1}} - \underbrace{\int_{x_i}^{x_{i+1}} \underbrace{S'''(x)}_{\text{constant}} [g'(x) - S'(x)] dx}_{k_i \int_{x_i}^{x_{i+1}} [g'(x) - S'(x)] dx \ = \ k [g(x) - S_i} = 0$$

$$\Rightarrow \int_a^b (g''-S'')^2 dx = \int_a^b [g''(x)]^2 dx - \int_a^b [S''(x)]^2 dx$$

$$\Rightarrow \int_a^b [g''(x)]^2 dx = \int_a^b [S''(x)]^2 dx + \underbrace{\int_a^b (g''-S'')^2 dx}_{\geqslant 0}$$

$$\Rightarrow \int_a^b [g''(x)]^2 \geqslant \int_a^b [S''(x)]^2 dx \qquad \square \text{proof}$$

1 B-splines.

We begin with a remainder about divided differences. We had

$$f[t_0,\ldots,t_n]=\sum_{j=0}^{n}f(t_j)\prod_{\substack{s=0\\ s\neq j}}^{n}(t_j-t_s)^{-1}$$

It may also be useful to treat a divided difference as a linear functional which transforms f into a number $f[t_0,\ldots,t_n]$ and denote such functional as

$$\delta^n(t_0,\ldots,t_n)f=f[t_0,\ldots,t_n]$$

If $f(x,y)$ is a function of two variables, then we can apply $\delta^n(t_0,\ldots,t_n)$ to $f(\cdot,y)$ which is a function of x and obtain $\delta^n_x(t_0,\ldots,t_n)f(\cdot,y)$. In general

$$\frac{\partial^l}{\partial y^l}\delta^n(t_0,\ldots,t_n)f(\cdot,y)=\delta^n(t_0,\ldots,t_n)\frac{\partial^l f(\cdot,y)}{\partial y^l}$$

*To define a B-spline we will need some prototypical splines. Let

$$(t-x)_+=\max\{t-x,0\}=\begin{cases}t-x & t>x\\ 0 & t\leq x\end{cases}$$

and the powers of the above function

$$(t-x)^r_+=\begin{cases}(t-x)^r & t>x\\ 0 & t\leq x\end{cases}$$

and in particular

$$(t-x)^0_+=\begin{cases}1 & t>x\\ 0 & t\leq x\end{cases}$$

*Hence the characteristic function of $\mathbb{R}_+$ is

$$t^0_+(t)=\begin{cases}0 & t\leq 0\\ 1 & 0<t\end{cases}$$

A simplest spline of degree r with node t_0 is a function $t\mapsto(t-t_0)^r_+$ which is in C^{r-1}.

A strictly increasing sequence of knots is prescribed

$$\cdots<t_{-1}<t_0<t_1<\cdots$$

where $\lim_{i\to\pm\infty}=\pm\infty$.

A B-spline B_i^r of degree r is given as a function of x by

$$B_i^r(x) = (t_{i+r+1} - t_i)\delta_t^{r+1}(t_i, \dots, t_{i+r+1})(t - x)_+^r$$

for $x \in \mathbb{R}$. More explicitly

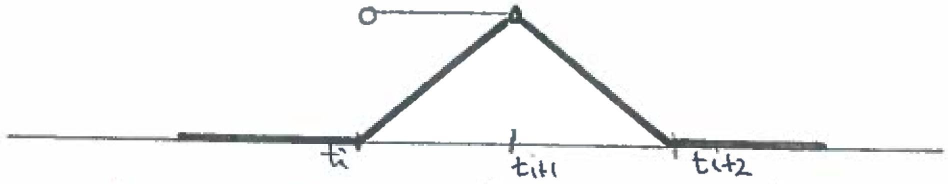
$$B_i^r(x) = (t_{i+r+1} - t_i) \sum_{j=i}^{i+r+1} \left((t_j - x)_+^r \prod_{\substack{s=i \\ s \neq j}}^{i+r+1} (t_j - t_s)^{-1} \right)$$

The simplest 0-degree spline $B_i^0(x)$ is a piecewise constant, left-continuous function

$$B_i^0(x) = (t_{i+1} - x)_+^0 - (t_i - x)_+^0 = \begin{cases} 1 - 1 = 0, & x \leq t_i \\ 1 - 0 = 1, & t_i < x \leq t_{i+1} \\ 0 - 0 = 0, & t_{i+1} < x \end{cases}$$

The piecewise continuous tent function is given by

$$B_i^1(x) = (t_{i+2} - t_i) \left(\frac{(t_i - x)_+}{(t_i - t_{i+1})(t_i - t_{i+2})} + \frac{(t_{i+1} - x)_+}{(t_{i+1} - t_i)(t_{i+1} - t_{i+2})} + \frac{(t_{i+2} - x)_+}{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} \right)$$



A necessary tool to obtain a recurrence formula for the B-splines is the formula for the divided difference of the product of two functions

Lemma. (Leibniz formula for divided differences) Let $f(t) = g(t)h(t)$, then

$$f[t_i, \dots, t_{i+k}] = \sum_{r=i}^{i+k} g[t_i, \dots, t_r]h[t_r, \dots, t_{i+k}]$$

Proof. This formula is a difference analog of $f^{(k)} = \sum_{r=0}^{k} \binom{k}{r} g^{(r)} h^{(k-r)}$. Consider $G(t)$ which is the polynomial of degree k interpolating the function g at $k+1$ points $t_i, t_{i+1} \dots t_{i+k}$ and written in Newton's form

$$G(t) = g(t_i) + \sum_{r=i+1}^{i+k} g[t_i, \dots, t_r](t - t_i) \dots (t - t_{r-1})$$

Let $H(t)$ be the polynomial of degree k interpolating the function h at the same $k+1$ points $t_i, t_{i+1}, \ldots, t_{i+k}$. This time we include the interpolation points in the Newton's formula beginning with t_{i+k} ending with t_i

$$H(t) = h(t_{i+k}) + \sum_{s=i}^{i+k-1} h[t_s, \ldots, t_{i+k}](t - t_{s+1}) \ldots (t - t_{i+k})$$

Suppose that $G(t) = \sum_{r=i}^{i+k} a_r$ and $H(t) = \sum_{s=i}^{i+k} b_s$. The polynomial $F(t) = G(t)H(t)$ of degree $2k$ interpolates $f(t) = g(t)h(t)$ at points $t_i, t_{i+1}, \ldots, t_{i+k}$ because G interpolates g and H interpolates h.

$$F(t) = (\sum_{r=i}^{i+k} a_r)(\sum_{s=i}^{i+k} b_s) = \underbrace{\sum_{r \leq s} a_r b_s}_{P_1(t)} + \underbrace{\sum_{r > s} a_r b_s}_{P_2(t)}$$

We will examine now the polynomials $P_2(t)$ and $P_1(t)$. A term a_r, $r = i, \ldots, i+k$ contains the product $(t - t_i) \ldots (t - t_{r-1})$ of degree $r - i$. A term b_s, $s \geq r$ contains the product $(t - t_{s+1}) \ldots (t - t_{i+k})$ of degree $i + k - s$. Hence $a_r b_s$ is of degree $k - s + r$. When $r > s$ then $r - 1 \geq s$ and each term a_r contains at least the factors $(t - t_i) \ldots (t - t_s)$. As a result in $P_2(t) = \sum_{r>s} a_r b_s$ each product $a_r b_s$ contains a product $(t - t_i) \ldots (t - t_s)(t - t_{s+1}) \ldots (t - t_{i+k})$. Hence

$$P_2(t_j) = 0. \qquad j = i, \ldots, i+k$$

and consequently

$$\delta^k(t_i, \ldots, t_{i+k})P_2 = 0$$

We now apply δ^k to the equation $F = GH = P_1 + P_2$. By linearity of δ^k

$$\delta^k F = \delta^k P_1 + \delta^k P_2$$

F interpolates f and hence $\delta^k F = \delta^k f$ so

$$\delta^k f = \delta^k P_1$$

P_1 is of degree k because $P_1 = \sum_{r \leq s} a_r b_s$ where a_r is of degree $r - i$ and b_s of degree $i + k - s$ and hence $a_r b_s$ is of degree $i + k - s + r - i \leq k - s + s = k$. The leading coefficient of P_1 is a sum of leading coefficients in polynomials $a_r b_s$ of degree k

$$\sum_{r=i}^{i+k} a_r b_r = \sum_{r=i}^{i+k} g[t_i, \ldots, t_r]h[t_r, \ldots, t_{i+k}](t - t_i) \ldots \widehat{(t - t_r)} \ldots (t - t_{i+k})$$

Hence

$$\delta^k P_1 = \sum_{r=i}^{i+k} g[t_i, \ldots, t_r]h[t_r, \ldots, t_{i+k}]$$

and

$$\delta^k(t_i,\dots,t_{i+k})f = \sum_{r=i}^{i+k} g[t_i,\dots,t_r]h[t_r,\dots,t_{i+k}]$$

Recurrence relation (de Boor, Cox).

We will derive now the recurrence relation for the B-splines which is equivalent to the definition and is an extremely useful tool in establishing various properties of spline functions.

$$B_i^r(x) = \frac{x-t_i}{t_{i+r}-t_i}B_i^{r-1}(x) + \frac{t_{i+r+1}-x}{t_{i+r+1}-t_{i+1}}B_{i+1}^{r-1}(x)$$

The proof uses the formula for the divided difference of the product. We have that $(t-x)_+^r = (t-x)(t-x)_+^{r-1}$. Denote $g(t) = t-x$ so that $g(t_i) = t_i - x$, $g[t_i,t_{i+1}] = 1$ and $g[t_i,\dots,t_j] = 0$ for $j > i+1$.

$$\delta_i^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^r = \delta_i^{r+1}(t_i,\dots,t_{i+r+1})[(t-x)(t-x)^{r-1}]_+$$

$$= g[t_i]\delta_i^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^{r-1} + g[t_i,t_{i+1}]\delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1}$$

$$= (t_i - x)\frac{\delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1} - \delta_i^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1}}{t_{i+r+1}-t_i} +$$

$$+ \delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1}$$

$$= \frac{x-t_i}{t_{i+r+1}-t_i}\delta_i^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1} + \left(\frac{t_i-x}{t_{i+r+1}-t_i}+1\right)\delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1}$$

$$= \frac{x-t_i}{t_{i+r+1}-t_i}\delta_i^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1} + \frac{t_{i+r+1}-x}{t_{i+r+1}-t_i}\delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1}$$

Since

$$B_i^r(x) = (t_{i+r+1}-t_i)\delta_i^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^r$$

by applying this definition twice with r replaced by $r-1$ and values i and $i+1$ we get

$$\delta_i^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1} = \frac{B_i^{r-1}(x)}{t_{i+r}-t_i}, \qquad \delta_i^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1} = \frac{B_{i+1}^{r-1}(x)}{t_{i+r+1}-t_{i+1}}$$

multipying both sides of the chain equality above by $t_{i+r+1}-t_i$ we obtain the recurrence.

Compact support.

$$B_i^r(x) = 0 \quad \text{for} \quad x \notin (t_i, t_{i+r+1}), \quad r \geq 0$$

We are proving that r-degree B-spline based at t_i has support in $r+1$ consecutive intervals. For $x < t_i \leq t \leq t_{i+r+1}$ we have $(t-x)_+^r = (t-x)^r$ is a polynomial of degree r in t. Therefore its $r+1$ order divided difference based at points $t_i, \dots, t_{i+r+1}$ is 0

$$\delta_t^{r+1}(t_i, \dots, t_{i+r+1})(t-x)^r = 0$$

and hence $B_i^r(x) = 0$.

For $t_i \leq t \leq t_{i+r+1} < x$ we have $(t-x)_+^r \equiv 0$ so its $r+1$ order divided difference based at points $t_i, \dots, t_{i+r+1}$ is 0 and again $B_i^r(x) = 0$.

Positivity in (t_i, t_{i+r+1}) **.**

$$B_i^r(x) > 0 \quad \text{for} \quad x \in (t_i, t_{i+r+1}), \quad r \geq 0$$

Induction. 0-order spline B_i^0 is positive on (t_i, t_{i+1}). Assume that is true for $r-1$ order spline. We will use the recurrence relation

$$B_i^r(x) = \frac{x - t_i}{t_{i+r} - t_i} B_i^{r-1}(x) + \frac{t_{i+r+1} - x}{t_{i+r+1} - t_{i+1}} B_{i+1}^{r-1}(x)$$

$(r-1)$-order B splines are positive in r consecutive intervals

$$B_i^{r-1}(x) = 0 \quad \text{if} \quad x \notin (t_i, t_{i+r})$$
$$B_{i+1}^{r-1}(x) = 0 \quad \text{if} \quad x \notin (t_{i+1}, t_{i+r+1})$$

We want to show that one of the terms in the recurrence is positive and another nonnegative. First consider $t_i < x < t_{i+r}$. Then first term is positive based on induction hypothesis. The second term B_{i+1}^{r-1} is positive on (t_{i+1}, t_{i+r}) but not on (t_i, t_{i+1}), so second term is nonnegative. Next consider $t_{i+r} \leq x < t_{i+r+1}$. We have $B_i^{r-1}(x) = 0$ so the first term is 0 and second is positive.

Partition of unity.

$$\sum_{j=-\infty}^{\infty} B_j^r(x) = 1 \quad \text{for} \quad x \in \mathbb{R}$$

For each x the infinite sum contains only finitely many nonzero terms. If $t_i \leq x < t_{i+1}$ then only $B_{i-r}^r, \dots, B_i^r$ have supports intersecting (t_i, t_{i+1})

$$\sum_{j=-\infty}^{\infty} B_j^r(x) = \sum_{j=i-r}^{i} B_j^r(x)$$

Based on the recursive definition of divided differences

$$B_j^r(x) = \delta_t^r(t_{j+1}, \dots, t_{j+r+1})(t-x)_+^r - \delta_t^r(t_j, \dots, t_{j+r})(t-x)_+^r$$

We have a telescoping sum where the above fragments of the first and last terms in the sum do not cancel

$$\sum_{j=i-r}^{i} B_j^r(x) = \delta_t^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^r - \delta_t^r(t_{i-r},\dots,t_i)(t-x)_+^r = 1-0$$

We need to explain the last equality. We assumed $t_i \le x < t_{i+1}$ so $(t-x)_+^r = (t-x)^r$ when $t_{i+1} \le t \le t_{i+r+1}$. Hence $\delta_t^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^r = 1$ which is the leading coefficient of t in $(t-x)^r$. As to the second term for $t_i \le x < t_{i+1}$ and for $t_{i-r} \le t \le t_i$ the function $(t-x)_+^r = 0$ so its divided difference vanishes too. When $x = t_i$ the function $B_j^0(x)$ may have jumps but is right continuous and hence the telescoping sum is either 1 or has right limit 1 as $t \to t_i^+$.

Derivative of B-spline.

For $r \ge 2$ we have

$$(B_i^r)'(x) = r\left(\frac{B_i^{r-1}(x)}{t_{i+r}-t_i} - \frac{B_{i+1}^{r-1}(x)}{t_{i+r+1}-t_{i+1}}\right)$$

When $r = 1$ the formula is true except for $x = t_i, t_{i+1}, t_{i+2}$.

From the formula $B_i^r(x) = (t_{i+r+1}-t_i)\delta_t^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^r$

$$\begin{aligned}(B_i^r)'(x) &= (t_{i+r+1}-t_i)\delta_t^{r+1}(t_i,\dots,t_{i+r+1})\frac{\partial}{\partial x}(t-x)_+^r \\ &= -r(t_{i+r+1}-t_i)\delta_t^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^{r-1}\end{aligned}$$

From the recursive definition of divided difference δ_t^{r+1} we have

$$\delta_t^{r+1}(t_i,\dots,t_{i+r+1})(t-x)_+^{r-1} = \frac{\delta_t^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1} - \delta_t^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1}}{t_{i+r+1}-t_i}$$

Hence

$$\begin{aligned}(B_i^r)'(x) &= -r\left(\delta_t^r(t_{i+1},\dots,t_{i+r+1})(t-x)_+^{r-1} - \delta_t^r(t_i,\dots,t_{i+r})(t-x)_+^{r-1}\right) \\ &= -r\left(\frac{B_{i+1}^{r-1}(x)}{t_{i+r+1}-t_{i+1}} - \frac{B_i^{r-1}(x)}{t_{i+r}-t_i}\right) \\ &= r\left(\frac{B_i^{r-1}(x)}{t_{i+r}-t_i} - \frac{B_{i+1}^{r-1}(x)}{t_{i+r+1}-t_{i+1}}\right)\end{aligned}$$

Linear independence of B-splines.

Lemma. The set of $r+1$ B-splines $\{B_j^r, B_{j+1}^r, \dots, B_{j+r}^r\}$ of degree r is linearly independent on a single interval (t_{j+r}, t_{j+r+1}).

Proof. When $r = 0$ then $\{B_j^0\}$ is 1 on (t_j, t_{j+1}) and is linearly independent. Suppose that Lemma holds for $r-1$. We want to show that if

$$S(x) = \sum_{i=0}^{r} c_{j+i} B_{j+i}^r(x)$$

and if $S|_{(t_{j+r}, t_{j+r+1})} = 0$ then $c_j = \ldots = c_{j+r} = 0$.
We begin by finding S' from $(B_{j+i}^r)'$

$$S'(x) = r \sum_{i=1}^{r} \frac{c_{j+i} - c_{j+i-1}}{t_{j+i+r} - t_{j+i}} B_{j+i}^{r-1}$$

To derive this formula we differentiate the recurrence relation for $B_{j+i}^r(x)$.

$$(B_{j+i}^r)'(x) = \left(\frac{B_{j+i}^{r-1}}{t_{j+i+r} - t_{j+i}} - \frac{B_{j+i+1}^{r-1}}{t_{j+i+r+1} - t_{j+i+1}} \right)$$

$$\begin{aligned} S'(x) &= r \left(\sum_{i=0}^{r} c_{j+i} \frac{B_{j+i}^{r-1}}{t_{j+i+r} - t_{j+i}} - \sum_{i=0}^{r} c_{j+i} \frac{B_{j+i+1}^{r-1}}{t_{j+i+r+1} - t_{j+i+1}} \right) \\ &= r \left(\sum_{i=0}^{r} c_{j+i} \frac{B_{j+i}^{r-1}}{t_{j+i+r} - t_{j+i}} - \sum_{s=1}^{r+1} c_{j+s-1} \frac{B_{j+s}^{r-1}}{t_{j+s+r} - t_{j+s}} \right) \end{aligned}$$

The first term in the first sum vanishes because when $i = 0$ spline B_j^{r-1} has support in (t_j, t_{j+r}) and doesn't contribute to $S(x)$ on (t_{j+r}, t_{j+r+1}). The last term in the second sum vanishes on (t_{j+r}, t_{j+r+1}) because spline B_{j+r+1}^{r-1} has support in (t_{j+r+1}, t_{j+2r+1}). We replace subscript s by i, both sums become $\sum_{i=1}^{r}$ and can be combined into a single sum which is the final formula for S'.

From the formula for $S'(x)$ and from the inductive hypothesis of linear independence of $\{B_{j+1}^{r-1}, \ldots, B_{j+r}^{r-1}\}$ we obtain that

$$c_j = \ldots = c_{j+r} = \lambda$$

Due to partition of unity property of B-splines

$$S(x) = \lambda \sum_{i=0}^{r} B_{j+i}^r(x) = \lambda, \qquad x \in (t_{j+r}, t_{j+r+1})$$

Since we assumed that $S|_{(t_{j+r}, t_{j+r+1})} = 0$ then $\lambda = 0$ and hence $c_j = \ldots = c_{j+r} = 0$.

Lemma. The set of $r+n$ B-splines $\{B_{-r}^r, B_{-r+1}^r, \ldots, B_{n-1}^r\}$ of degree r is linearly independent on an interval (t_0, t_n).

Proof. Let $S = \sum_{i=-r}^{n-1} c_i B_i^r$ and suppose $S|_{(t_0,t_n)} = 0$. On (t_0, t_1) only $B_{-r}^r, \dots, B_0^r$ are nonzero and hence

$$0 = S|_{(t_0,t_1)} = \sum_{-r}^{0} c_i B_i^r|_{(t_0,t_1)}$$

By previous lemma on (t_0, t_1) $r+1$ splines $B_{-r}^r, \dots, B_0^r$ are linearly independent so $c_{-r} = \dots = c_0 = 0$. We do not know if $c_1, \dots, c_{n-1}$ are 0. Let $1 \leq j \leq n-1$ be the first index so that $c_j \neq 0$. Hence $(t_j, t_{j+1}) \subseteq (t_0, t_n)$ and for all $x \in (t_j, t_{j+1})$ due to assumption

$$0 = S(x) = \sum_{i=j}^{n-1} c_i B_i^r(x) = c_j B_j^r(x) \neq 0$$

because on (t_j, t_{j+1}) all $B_{j+1}^r, \dots, B_{n-1}^r$ are zero and because $c_j \neq 0$ and $B_j^r(x) \neq 0$. Hence all $c_i = 0$.

*Divided different

* Expanded form of divided difference.

$$f[x_0] = f(x_0)$$

$$f[x_0, x_1] = \frac{f(x_0)}{(x_0 - x_1)} + \frac{f(x_1)}{(x_1 - x_0)}$$

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

$$f[x_0, x_1, \dots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ i \neq j}}^{n} (x_i - x_j)} = \sum_{i=0}^{n} \left(f(x_i) \prod_{\substack{s=0 \\ s \neq j}}^{n} (x_i - x_s)^{-1} \right)$$

* Given $(k+1)$ data points.

$(x_0, f_0) \dots (x_n, f_n)$

$$[f_\nu] := f_\nu$$

$$[f_\nu, \dots, f_{\nu+k}] = \frac{[f_{\nu+1}, \dots, f_{\nu+k}] - f[\nu, \dots, \nu+k-1]}{x_{\nu+k} - x_\nu}$$

$$f[x_\nu, \dots, x_{\nu+k}] = \frac{f[x_{\nu+1}, \dots, x_{\nu+k}] - f[x_\nu, \dots, x_{\nu+k-1}]}{x_{\nu+k} - x_\nu}$$

* Properties of divided difference.

a) Linearity

If $f(x) = \alpha g(x) + \beta h(x)$

then $f[x_0, \dots, x_n] = \alpha g[x_0, \dots, x_n] + \beta h[x_0, \dots, x_n]$.

b) Commutation

$$f[x_0, \dots, x_n] = f[\sigma(x_0), \dots, \sigma(x_n)]$$

c) Recurrence formula.

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

* BSplines

*We have the formula of divided difference

$$f[x_0,\dots,x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{\prod_{\substack{s=0\\ s\neq j}}^{n}(x_j-x_s)}$$

- It may be useful to treat divided difference as a linear function with transform f into a number $f[t_0,\dots,t_n]$.

$\delta^n(t_0,\dots,t_n)f = f[t_0,\dots,t_n]$ ← when f is a function of 1 variable

- If $f(x,y)$ is a function of two variables $f(\cdot,y)$ is a function of x

$\rightarrow$ obtain $\delta^n_x(t_0,\dots,t_n)f(\cdot,y)$

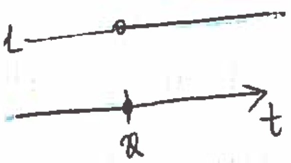
Note: When f is a polynomial of degree $(n-1)$ then $\Rightarrow \delta^n(t_0,\dots,t_n)f = 0$

In general $\frac{\partial^l}{\partial y^l}\delta^n(t_0,\dots,t_n)f(\cdot,y) = \delta^n(t_0,\dots,t_n)\frac{\partial^l f(\cdot,y)}{\partial y^l}$

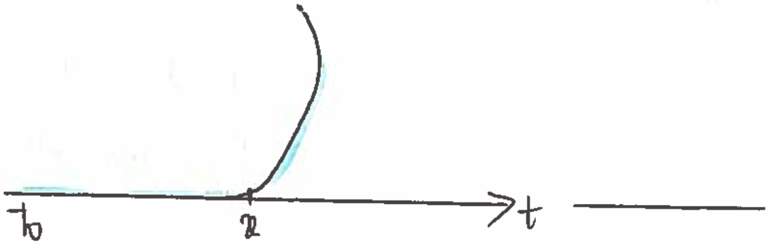
* To define a B_spline, we will need some prototypical splines.

Let $(t-x)_+ = \begin{cases} t-x & t>x \\ 0 & t\leq x\end{cases}$

$(t-x)^0_+ = \begin{cases} 1 & t>x \\ 0 & t\leq x\end{cases}$



$(t-x)^r_+ = \begin{cases} (t-x)^r & t>x \\ 0 & t\leq x\end{cases}$



- A simplest spline of degree r with node t_0 is a function $t \longmapsto (t-t_0)^r_+ \in C^{r-1}$

* A strictly increasing of knots is prescribed $\dots t_{-1}<t_0<t_1<\dots$ $\lim_{i\to\pm\infty}|t_i| \infty$

a Bspline B_i^r of degree r is given as a function of x

$$B_i^r(x) = (t_{i+r+1}-t_i)\,\delta_t^{r+1}(t_i,\dots,t_{i+r+1})(t-x)^r_+ \qquad r\geq 0$$

$$= (t_{i+r+1}-t_i)\sum_{j=i}^{i+r+1}\frac{(t_j-x)^r_+}{\prod_{\substack{s=i\\ s\neq j}}^{i+r+1}(t_j-t_s)}$$

note that later on, we can have t receives values from t_i to t_{i+r+1} which means $t_i\leq t\leq t_{i+r+1}$

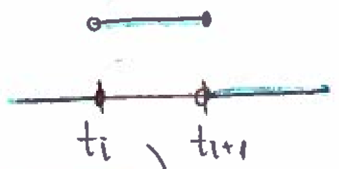
$$B_i^{(r)}(x) = \delta_t^r(t_{i+1},\dots,t_{i+r+1})(t-x)^r_+ - \delta_t^r(t_i,\dots,t_{i+r})(t-x)^r_+$$

* Example :

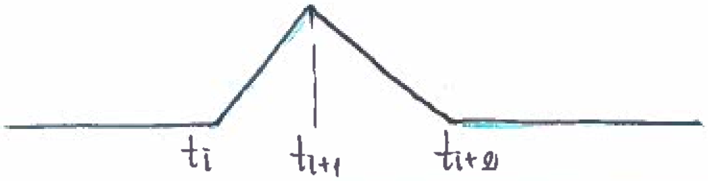
- The simplest 0-degree $B_i^0(x)$ is a piece wise constant, left continuous function

$$B_i^0(x) = (t_{i+1}-t_i)\left[\sum_{j=i}^{i+1} \frac{(t_j-x)_+^0}{\prod_{\substack{s=i\\ s\neq j}}^{i+1}(t_j-t_s)}\right] = [t_{i+1}-t_i]\left[\frac{(t_i-x)_+^0}{(t_i-t_{i+1})} + \frac{(t_{i+1}-x)_+^0}{(t_{i+1}-t_i)}\right]$$

$$= (t_{i+1}-x)_+^0 - (t_i-x)_+^0 = \begin{cases} 1-1=0 & x\leq t_i \\ 1-0=1 & t_i<x\leq t_{i+1} \\ 0-0=0 & t_{i+1}<x \end{cases}$$



- $$B_i^1(x) = (t_{i+2}-t_i)\left(\frac{(t_i-x)_+}{(t_i-t_{i+1})(t_i-t_{i+2})} + \frac{(t_{i+1}-x)_+}{(t_{i+1}-t_i)(t_{i+1}-t_{i+2})} + \frac{(t_{i+2}-x)_+}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}\right)$$



* Lemma (Leibniz formula for divided differences).
(This is a necessary tool to obtain a recurrence formula for B-splines).

Let $f(t) = g(t)\,h(t)$

Then $f[t_i,\dots,t_{i+k}] = \sum_{r=1}^{i+k} g[t_i,\dots,t_r]\,h[t_r,\dots,t_{i+k}]$.

$$= g[t_i]\,h[t_i,\dots,t_{i+k}] + g[t_i,t_{i+1}]\,h[t_{i+1},\dots,t_{i+k}]$$
$$g[t_i,t_{i+1},t_{i+2}]\,h[t_{i+2},\dots,t_{i+k}] + \dots + g[t_i,\dots t_{i+k}]\,h[t_{i+k}]$$

* Theorem (Recurrence formula for B-spline) (Cox De Boor)

$$B_i^r(x) = \frac{x-t_i}{t_{i+r}-t_i}B_i^{r-1}(x) + \frac{t_{i+r+1}-x}{t_{i+r+1}-t_{i+1}}B_{i+1}^{r-1}(x).$$

* Properties :

① Compact support : For any $r\geq 0$, $B_i^r(x)>0$ when $x\in(t_i,t_{i+r+1})$
$B_i^r(x)=0$ when $x\notin(t_i,t_{i+r+1})$

② Partition of unity $\sum_{j=-\infty}^{+\infty} B_j^r(x) = 1$ for $x\in\mathbb{R}$.

③ Linear independent :
$\underbrace{\{B_j^r, B_{j+1}^r,\dots,B_{j+r}^r\}}_{(r+1)\text{ Bsplines of degree } r}$ is linearly independent on a single interval (t_{j+r}, t_{j+r+1})

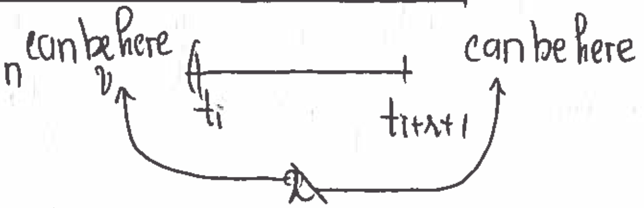
It is important to say which spline on which interval.

$\{B_{-r}^r, B_{-r+1},\dots, B_{n-1}^r\}$ (set $r+n$ Bsplines of degree r) is linearly independent on the interval $(t_0,t$

* Properties of Bsplines

* Compact support
For $\lambda \geq 0$, then $B_i^\lambda(x) = 0$ for $x \notin (t_i, t_{i+\lambda+1})$

* Prove : when $x \notin (t_i, t_{i+\lambda+1})$, then



- Case 1 : When $x < t_i \leq t \leq t_{i+\lambda+1}$

$\Rightarrow (t-x)_+^\lambda = (t-x)^\lambda$ is a polynomial of degree $\boxed{\lambda}$

$\Rightarrow \delta_t^{\lambda+1}(t_0, \dots, t_n)(t-x)^\lambda = 0$ since $\delta_t^{\lambda+1}$ is a $\boxed{(\lambda+1)}$ order divided difference based at points $t_i, \dots, t_{i+\lambda+1}$. (similar to derivative)

- Case 2 : When $t_i \leq t \leq t_{i+\lambda+1} < x$

then $(t-x)_+^\lambda = 0$

$\Rightarrow \delta_t^{\lambda+1}(t_0, \dots t_n)(t-x)_+^\lambda = 0$

* Positivity of $B_i^\lambda(x)$: $B_i^\lambda(x) > 0$ when $t_i < x < t_{i+\lambda+1}$

* We we prove by induction :

- $\lambda = 0$ true
- Assume possibility holds for $(\lambda-1)$, we have $B_i^{\lambda-1}(x) > 0$, when $t_i < x < t_{i+\lambda}$

which means we have $B_i^{\lambda-1}(x) = 0$ for $x \notin (t_i, t_{i+\lambda})$

$B_{i+1}^{\lambda-1}(x) = 0$ for $x \notin (t_{i+1}, t_{i+\lambda+1})$

- So now we want to prove that it is true for λ, which means :

we need to prove that $B_i^\lambda(x) > 0$ for $t_i < x < t_{i+\lambda+1}$

⊕ We we consider the recurrence formula

$$B_i^\lambda(x) = \frac{x-t_i}{t_{i+\lambda+1}-t_i} B_i^{(\lambda-1)}(x) + \frac{t_{i+\lambda+1}-x}{t_{i+\lambda+1}-t_{i+1}} B_{i+1}^{(\lambda-1)}(x).$$

we want to show that $\left\{ \begin{array}{l} \text{one of the term of the recurrence is positive} \\ \text{another term is nonnegative} \end{array} \right.$.

$\Rightarrow B_\lambda^\lambda(x) > 0$ when $t_i < x < t_{i+\lambda}$ □.

② Partition of unity $\sum_{j=-\infty}^{+\infty} B_j^\lambda(x) = 1$ for $x \in \mathbb{R}$.

Note that for each x, the infinite sum contain only finitely many nonzer terms When $t_i \leq x < t_{i+1}$, then (t_i, t_{i+1}) is the intersecting the support of $B^\lambda_{i-\lambda}, \dots, B^\lambda_i$ only which means we have $\sum_{j=-\infty}^{+\infty} B_j^\lambda(x) = \sum_{j=i-\lambda}^{i} B_j^\lambda(x)$

* So now we want to prove that $\sum_{j=i-\lambda}^{i} B_j^\lambda(x) = 1$ for $x \in [t_i, t_{i+1})$. (1)

- Based on the recurrence definition of divided differences, we have

$$B_j^\lambda(x) = \delta_t^\lambda(t_{j+1}, \dots, t_{j+\lambda+1})(t-x)_+^\lambda - \delta_t^\lambda(t_j, \dots, t_{j+\lambda})(t-x)_+^\lambda \quad (2).$$

Substitute (2) to (1), we have a telescoping sum where the above fragments of the first and the last terms in the sum do not cancel

$$\Rightarrow \sum_{j=i-\lambda}^{i} B_j^\lambda(x) = \delta_t^\lambda(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^\lambda - \delta_t^\lambda(t_{i-\lambda}, \dots, t_i)(t-x)_+^\lambda = 1 - 0 = 1.$$

* Now we want to explain the last equality

- Since $x \in [t_i, t_{i+1}) \to (t-x)_+^\lambda \overset{\uparrow}{=} (t-x)^\lambda \Rightarrow \delta_t^\lambda(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^\lambda = 1$.

  $\Rightarrow$ when $t_{i+1} \leq t \leq t_{i+\lambda+1}$).

- Since $x \in [t_i, t_{i+1}) \Rightarrow (t-x)_+^\lambda \overset{\uparrow}{=} 0 \Rightarrow \delta_t^\lambda(t_{i-\lambda}, \dots, t_i)(t-x)_+^\lambda = 0$.

  when $t_{i-\lambda} < t < t_i$

* So, in conclusion, we have proved that $\sum_{j=-\infty}^{+\infty} B_j^\lambda(x) = \sum_{j=i-\lambda}^{i} B_j^\lambda(x)$ for $x \in [t_i, t_{i+1})$.

but since $[t_i, t_{i+1})$, $i \in -\infty, +\infty)$ create partition of $\mathbb{R}$, which mean it is true for $x \in \mathbb{R}$.

*Linear independence of B_splines.

Lemma 1:

$\underbrace{\{B_j^r, B_{j+1}^r, \dots, B_{j+r}^r\}}_{\text{The set of } (r+1) \text{ B\_splines of degree } r}$ is linearly independent on a single interval (t_{j+r}, t_{j+r+1})

Lemma 2:

$\underbrace{\{B_{-r}^r, B_{-r+1}^r, \dots, B_{n-1}^r\}}_{\text{the set of } (r+n) \text{ B\_splines of degree } r}$ is linearly independent on an interval (t_0, t_n).

* Prove lemma 1: we will prove this by induction.

- When $r = 0$, then we have $\{B_j^0\}$, $\mathcal{B}$ contains only one B_spline and is linearly independent in (t_j, t_{j+1})
- Assume that the assumption is hold for $r-1$, which mean we have $(B_j^{r-1}, \dots, B_{j+r-1}^{r-1})$ is linearly independent in (t_{j+r-1}, t_{j+r})

which means, Put $A(x) = \sum_{i=0}^{r-1} c_{j+i} B_{j+i}^{r-1}(x)$ then if $A\big|_{(t_{j+r-1}, t_{j+r})} = 0$ then $c_{j+i} = 0$ $\forall i = \overline{1,r}$

- Now we want to prove that

- When $r > s$ $r-1 \geq s$

So each term a_r contains at least the factors $(t-t_i)\dots(t-t_s)$

$$P_2(t) = \underbrace{\sum_{r>s} a_r b_s}$$

each term $a_r b_s$ contains $(t-t_i)\dots(t-t_s)(t-t_{s+1})\dots(t-t_{i+k})$

$\Rightarrow P_2(t_j) = 0 \ , \ j = \overline{i, i+k}$

Hence $\delta^\ell(t_j, \dots t_{i+\ell}) P_2 = 0$

- $F = GH = P_1 + P_2$

$$\delta^\ell F = \delta^\ell P_1 + \underbrace{\delta^\ell P_2}_{=0} = \delta^\ell P_1$$

interested in leading coefficient of P_1

- $P_1(t) = \sum_{r \leq s} a_r b_s$

of degree $r-i$ of degree $i+k-s$

of degree $k-s+r \leq k-s+s = k$.

The leading coefficient in P_1 is a sum of leading coefficients in the term of $a_r b_s$ of degree k

- $$\sum_{r=i}^{i+\ell} a_r b_r = \sum_{r=i}^{i+k} g[t_i, \dots t_r]\ h[t_r, \dots, t_{i+\ell}]\ (t-t_i)\dots(t-t_r)\dots(t-t_{i+\ell})$$

Hence $$\delta^\ell P_1 = \sum_{r=i}^{i+\ell} g[t_i, \dots, t_r]\ h[t_r, \dots, t_{i+\ell}] = \delta^\ell f$$

* Recurrence formula for B spline (cox, Deboor)

$$B_i^{\lambda}(x) = \frac{x-t_i}{t_{i+\lambda}-t_i} B_i^{\lambda-1}(x) + \frac{t_{i+\lambda+1}-x}{t_{i+\lambda+1}-t_{i+1}} B_{i+1}^{\lambda-1}(x)$$

$$(t-x)_+^{\lambda} = (t-x)(t-x)_+^{\lambda-1}$$

$g(t) = t - x$

$g(t_i) = t_i - x$

$$\begin{cases} g[t_i, t_{i+1}] = 1 \\ g[t_i, \dots, t_j] = 0 \qquad j > i+1 \end{cases}$$

$$\bullet\ \delta_t^{\lambda+1}(t_i, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda} = \overset{(t_i-x)}{\overset{\shortparallel}{g(t_i)}}\, \delta_t^{\lambda+1}(t_i, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1} + \overbrace{g[t_i, t_{i+1}]}^{1}\, \delta_t^{\lambda}(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1}$$

$$= (t_i - x)\, \frac{\delta^{\lambda}(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1} - \delta^{\lambda}(t_i, \dots, t_{i+\lambda})(t-x)_+^{\lambda-1}}{t_{i+\lambda+1} - t_i} + \delta_t^{\lambda}(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1}$$

$$= \frac{(x-t_i)}{t_{i+\lambda+1}-t_i}\, \delta_t^{\lambda}(t_i, \dots, t_{i+\lambda})(t-x)_+^{\lambda-1} + \underbrace{\left(\frac{(t_i-x)}{t_{i+\lambda+1}-t_i} + 1\right)}_{\frac{t_{i+\lambda+1}-x}{t_{i+\lambda+1}-t_i}} \delta_t^{\lambda}(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1}$$

$$= \frac{(x-t_i)}{t_{i+\lambda+1}-t_i}\, \delta_t^{\lambda}(t_i, \dots, t_{i+\lambda})(t-x)_+^{\lambda-1} + \frac{t_{i+\lambda+1}-x}{t_{i+\lambda+1}-t_i}\, \delta_t^{\lambda}(t_{i+1}, \dots, t_{i+\lambda+1})(t-x)_+^{\lambda-1}$$

7.2 Numerical integeation based on interpolation

* Idea: we want to compute $\int_a^b f(x)\,dx$, where a, b finite, where $f(x)$ is hard to compute

$\Rightarrow$ We simplifier the problem by finding $g \approx f$, and so that $\int_a^b f(x)dx \approx \int_a^b g(x)dx$

* Integration via polynomial interpolation :

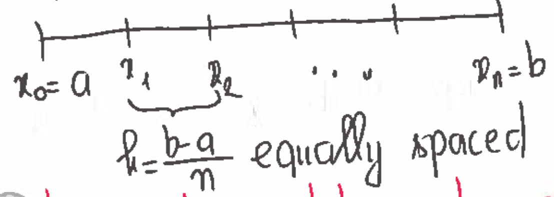
* We have $f(x) \approx L_n(x) = \sum_{i=0}^{n} f(x_i)\, l_i(x)$, $l_i(x) = \prod_{\substack{j=0\\ j\neq i}}^{n} \frac{(x-x_j)}{x_i - x_j}$

then we have

$$\int_a^b f(x)dx \approx Q(f) = \int_a^b L_n(x)dx = \int_a^b \sum_{i=0}^{n} f(x_i)\, l_i(x)\, dx = \sum_{i=0}^{n} f(x_i) \underbrace{\int_a^b l_i(x)dx}_{\lambda_i \text{ (weight)}} = \sum_{i=0}^{n} \lambda_i f(x_i)$$

$I(f) = \int_a^b f(x)dx$

• If the nodes are equally spaced



$h = \frac{b-a}{n}$ equally spaced

$$\boxed{\Rightarrow (*) \int_a^b f(x)dx = \sum_{i=0}^{n} \lambda_i f_i \text{ is called Newton-Cotes formula}}$$

* Method of undetermined coefficients (to determine the weight λ_i)

From above (*) $\int_a^b f(x)dx = \sum_{i=0}^{n} \lambda_i f_i$, $\begin{cases} \text{we already have } f_i \\ \text{we want to } \underline{\text{compute}}\ \lambda_i \end{cases}$

• We apply (*) for $f(x) = x^j$

$\Rightarrow Q(x^j) = I(x^j)$

$\sum \lambda_i x_i^j = \underbrace{\int_a^b x^j dx}_{\text{can compute}}$ $\Rightarrow$ we can construct a Vandermode matrix to find λ_i, $i = \overline{0,n}$

* Definition

Given $p \in \mathbb{P}_\alpha$,
We say that Q has a degree of exactness of degree α iff $Q(p) = I(p)$, $p \in \mathbb{P}_\alpha$

* $\Rightarrow$ Hence, interpolating quadratures of exactness of degree n (in $(n+1)$ nodes and L_n)

* Define error for numerical integration based on Lagrange interpolation

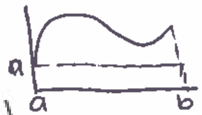
$$E(f) = I(f) - Q(f) = \int_a^b [f(x) - L_n(x)]\,dx = \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x-x_0)\cdots(x-x_n)\,dx .$$

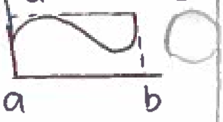
* Constant interpolation

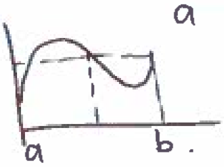
Let $x_0 = a$ $x_1 = b$ $h = b-a$ $\frac{a+b}{2} = x_{1/2}$.

Then we have $Q(f)$ can be

$$Q(f) = \begin{cases} (b-a)f(a) = h f(a) & \leftarrow \text{left rectangle rule} \\ (b-a)f(b) = h f(b) & \leftarrow \text{right rectangle rule} \\ (b-a)f(x_{\frac{1}{2}}) & \leftarrow \text{midpoint rule} \end{cases}$$







* We will prove that for the mid-point rule, then

$$\int_a^b f(x)\,dx = h f(x_{1/2}) + \underbrace{\frac{1}{24} h^3 f''(\eta)}_{\text{error}}$$

* Proof: Suppose $L_0(f) = f(x_{1/2})$ remind: the note here

then $f(x) - L_0(f)(x) = f'(\xi)(x - x_{1/2})$

then $E(f) = \int_a^b f(x) - L_0(f)(x)\,dx = \int_a^b f'(\xi)(x - x_{1/2})\,dx$

$$\Rightarrow |E(f)| \le \int_a^b |f'(\xi)|\,|x - x_{1/2}|\,dx = |f'(\xi)| \frac{(x - x_{1/2})^2}{2}\Big|_a^b = \frac{|f'(\xi)|}{2}\left(|b - x_{1/2}|^2 - |a - x_{1/2}|^2\right)$$

$$\le \frac{|f'(\xi)|}{2}\, 2\left(|b - x_{1/2}|^2\right) = |f'(\xi)| \frac{(b-a)^2}{4}$$

* Exactness of midpoint rule for affine function follows from the Hermite interpolation formula

$$H(x) = f(x_{1/2}) + (x - x_{1/2}) f'(x_{1/2}) = L_0(f) + (x - x_{1/2}) f'(x_{1/2})$$

which is a Hermite interpolation of f at $x_{1/2}$ $H(x_{1/2}) = f(x_{1/2})$

$H'(x_{1/2}) = f'(x_{1/2})$.

- $Q(f) = I(L_0) = I(H)$

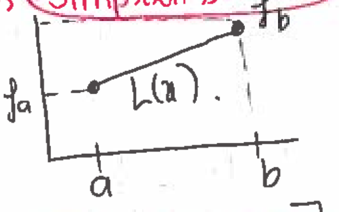
$$f(x) - H(x) = \frac{1}{2!} f''(\xi_x)(x - x_{1/2})^2$$

$$\Rightarrow E(f) = I(f) - Q(f) = I(f - H) = \int_a^b \frac{1}{2!} f''(\xi_x)(x - x_{1/2})^2\,dx = \frac{h^3}{24} f''(\eta)$$

- Remind:

φ, ψ are continuous on $[a,b]$, ψ doesn't change sign $\Big\} \Rightarrow \exists \eta,\ \int_a^b \varphi(x)\psi(x)\,dx = \varphi(\eta)\int_a^b \psi(x)\,dx$.

* Find the error of estimating intagral by linear intepolation & Simpson's rule



Give trapezoidal rule ; Take $x_0 = a$ $x_1 = b$

$$L(x) = \frac{x-b}{a-b} f(a) + \frac{x-a}{b-a} f(b).$$

Then $Q(f) = I(L) = \frac{f(a)}{a-b} \underbrace{\int_a^b (x-b)\,dx}_{=\frac{1}{2}(b-a)^2} + \frac{f(b)}{b-a} \underbrace{\int_a^b (x-a)\,dx}_{=\frac{1}{2}(b-a)^2} = \frac{b-a}{2}\left[f(a) + f(b) \right]$

Then

$$I(f) - Q(f) = \int_a^b \frac{f''(\xi_x)}{2!}(x-a)(x-b)\,dx = -\frac{(b-a)^3}{12} f''(\xi).$$

* Error(Eliminate error of Simpsone's rule.

- Define a Hemite intepolant of f that satisfies $\begin{cases} H_3(x_i) = f(x_i) & , i = 0,1,2 \\ H_3'(x_1) = f'(x_1) \end{cases}$

$$H_3(x) = L_2(x) + K(x-x_0)(x-x_1)(x-x_2)$$

- Find K

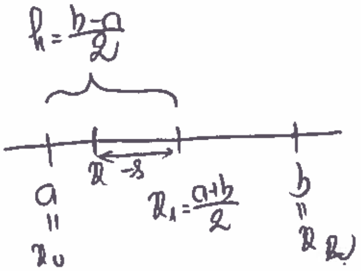
$$L_2'(x_1) = \frac{f_2 - f_0}{2h}$$

$$H_3'(x_1) = \frac{f_2 - f_0}{2h} - Kh^2 = f'(x_1)$$

$$\rightarrow K = \frac{1}{h^2}\left(\frac{f_2 - f_0}{2h} - f'(x_1) \right)$$

$$\rightarrow I(H_3) = I(L_2) + K \underbrace{\int_{-h}^{h} (s+h)s(s-h)\,ds}_{K\int_{-h}^{h} s(s^2-h^2)\,ds = 0} = I(L_2)$$

* Quadratic interpolant Simpson's rule



- $x_0 = a$ $\quad x_1 = \frac{a+b}{2}$ $\quad x_2 = b$ $\quad h = \frac{b-a}{2}$

$$L_2(x) = \frac{(x-x_1)(x-x_2)}{(-h)\ (-2h)} f_0 + \frac{(x-x_0)(x-x_2)}{(h)\ (-h)} f_1 + \frac{(x-x_0)(x-x_1)}{2h\ \ h} f_2$$

- $x - x_0 = x - x_1 + h = s+h$ $\qquad s := x - x_1$.

$x - x_2 = s - h$.

- $\int_{x_0}^{x_2} (x-x_1)(x-x_2)\,dx = \int_{-h}^{h} s(s-h)\,ds = \int_{-h}^{h} s^2\,dx = \frac{1}{3} s^3 \Big|_{-h}^{h} = \frac{2}{3} h^3$

$\int_0^{x_2} (x-x_0)(x-x_2)\,dx = \frac{4}{3} h^3 \qquad \int_0^{x_2} (x-x_0)(x-x_1)\,dx = \frac{2}{3} h^3$

So we have $Q(f) = I(L_2) = h\left(\frac{1}{3} f_0 + \frac{4}{3} f_1 + \frac{1}{3} f_2\right) = \frac{b-a}{6}\left(f_0 + 4f_1 + f_2\right)$

* Errorelminate the error of Simpson

- Define a Hermite the ~~error~~ Hermite interpolation of f $\quad \begin{cases} H_3(x_i) = f(x_i) & i=0,1,2 \\ H_3'(x_1) = f'(x_1) \end{cases}$

$H_3(x) = L_2(x) + K(x-x_0)(x-x_1)(x-x_2)$

- Find K

$$\left.\begin{aligned} L_2'(x_1) &= \frac{f_2-f_0}{2h} \\ H_3'(x_1) &= \frac{f_2-f_0}{2h} - Kh^2 = f'(x_1) \end{aligned}\right\} \to K = \frac{1}{h^2}\left[\frac{f_2-f_0}{2h} - f'(x_1)\right]$$

- $I(H_3) = I(L_2) + K\int_{-h}^{h} (s+h)s(s-h)\,ds = I(L_2) + K\underbrace{\int_{-h}^{h} s(s^2-h^2)\,ds}_{=0} = I(L_2)$.

- Error

$E(f) = I(f-L_2) = I(f-H_3)$

$f(x) = H_3(x) + (x-x_0)(x-x_1)^2(x-x_2)\,\frac{f^{(4)}(\xi)}{4!}$

$\to E(f) = -\frac{(b-a)^5}{2880} f^{(4)}(\eta) \underset{f\in\Pi_3}{\overset{\text{iff}}{=\!=\!=}} 0$

* Orthogonal polynomials.

* Let $(a,b) \subset \mathbb{R}$.

Let $w(x) > 0$ on $x \in (a,b)$, $w(x) \in L^1(a,b)$.

Call w a weight function

We defined inner product of two functions defined on (a,b):

$$\langle f,g\rangle = \int_a^b f(x)\, g(x)\, w(x)\, dx \qquad L^2_w(a,b) = \{ f:(a,b) \to \mathbb{R},\ \|f\| < \infty \}$$

$$\|f\| = \left[\int_a^b f^2(x)\, w(x)\, dx\right]^{1/2}$$

property: $\langle fh, g\rangle = \langle f, hg\rangle$

* We construct orthogonal polynomial in $L^2_w(a,b)$

$p_0(x) = 1$

$p_1(x) = x - \frac{\langle 1, x\rangle}{\langle 1, 1\rangle} 1$ (projection of x onto 1

$\vdots$

$p_n(x) = x^n - \sum_{i=0}^{n-1} \lambda_{i,n} p_i(x)$ $\qquad \lambda_{i,n} = \frac{\langle x^n, p_i\rangle}{\langle p_i, p_i\rangle}$

↑ unstable

(Modified Gramm Smith is stable)

← not orthogonal

* Properties:

- $p_n(x)$ is monic
- such system is not orthonormal, can be normalized by [dividing by norm
  change leading coefficient

* Theorem (triple recursion formula)

There exists a unique sequence of polynomials $\{p_n\}_{n=0}^{\infty}$ such that

$$\begin{cases} p_n(x) \text{ is monic of degree } n \\ \langle p_n(x)\, q\rangle = 0, \ \forall q \in \mathbb{P}_{n-1} \end{cases}$$

Such polynomials are orthogonal $\langle p_i, p_j\rangle = 0$ $\quad i \neq j$

and satisfy $p_n(x) = (x - \lambda_n) p_{n-1}(x) - \mu_n p_{n-2}(x)$ $\quad n \geq 2$

where $\lambda_n = \frac{\langle x p_{n-1}, p_{n-1}\rangle}{\|p_{n-1}\|^2}$ $\quad \mu_n = \frac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2}$

* Theorem (Triple recursion formula

There exists a unique sequence of polynomial $\{p_n\}_{n=0}^{\infty}$ such that

$$\begin{cases} p_n(x) \text{ is monic of degree } n \\ \langle p_n, q\rangle = 0, \quad \forall q \in \mathbb{P}_{n-1} \end{cases}$$

Such polynomials are orthogonal $\langle p_i, p_j \rangle = 0$, $i \neq j$ (1)

satisfies $p_n(x) = (x - \lambda_n) p_{n-1}(x) - \mu_n p_{n-2}(x)$ $n \geq 2$ (2)

$$\lambda_n = \frac{\langle x p_{n-1}, p_{n-1}\rangle}{\|p_{n-1}\|^2} \qquad \mu_n = \frac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2}$$

* Proof (1) We first prove the recurrence formula

We have $x^n \in \operatorname{span}\{p_0, \dots, p_n\} = \mathbb{P}_n$

If $q \in \mathbb{P}_{n-1}$, then $\langle p_n, q \rangle = 0$

(2) Consider a polynomial

$$p_n(x) - x p_{n-1}(x) = \sum_{i=0}^{n-1} a_i p_i(x)$$ ← we want to show that this sum is actually shorter than how it looks (There are only 2 non zero coefficients)

We want to show that only a_{n-1} and a_{n-2} nonzero

* Multiply by p_i, $i = 0, n-1$

$$p_i \text{ RHS} = p_i \sum_{i=1}^{n-1} a_i p_i(x) = a_i \langle p_i, p_i \rangle = -\langle x p_{n-1}, p_i \rangle = -\langle p_{n-1}, x p_i \rangle \qquad (*)$$

- For $i \leq (n-3)$, then

$\langle p_{n-1}, x p_i(x) \rangle = 0 \Rightarrow a_i = 0$ if $i \leq (n-3)$.

- If $i = (n-1)$

$$(*) \Rightarrow a_{n-1} \langle p_{n-1}, p_{n-1} \rangle = -\langle p_{n-1}, x p_{n-1} \rangle$$

$$a_{n-1} = -\frac{\langle x p_{n-1}, p_{n-1} \rangle}{\langle p_{n-1}, p_{n-1} \rangle} = -\frac{\langle x p_{n-1}, p_{n-1} \rangle}{\|p_{n-1}\|} = -\lambda_n$$

- If $i = (n-2)$

$$(*) \Rightarrow a_{n-2} \langle p_{n-2}, p_{n-2} \rangle = -\langle p_{n-1}, x p_{n-2} \rangle = -\langle p_{n-1}, p_{n-1} - p_{n-1} + x p_{n-2} \rangle$$

$$= -\langle p_{n-1}, p_{n-1} \rangle - \langle p_{n-1}, \underbrace{-p_{n-1} + x p_{n-2}}_{\in \mathbb{P}_{n-2}} \rangle = -\langle p_{n-1}, p_{n-1} \rangle$$

$$\Rightarrow a_{n-2} = -\frac{\langle p_{n-1}, p_{n-1} \rangle}{\langle p_{n-2}, p_{n-2} \rangle} = -\frac{\|p_{n-1}\|^2}{\|p_{n-2}\|^2} = -\mu_n$$

* Extremal property of monic orthogonal polynomials

Let p_n be n^{th} monic orthogonal polynomial

Then $\|p_n\| \le \|s\|$ for any monic polynomial of degree $\le n$

* Proof:

Suppose that

$$S(x) = p_n(x) - \sum_{i=0}^{n-1} c_i p_i(x) = p_n + rest$$

$$\|s(x)\| = \|p_n\| + \|rest\| \Rightarrow \|p_n\| \le \|s\|$$

* Root of orthogonal polynomials

The polynomial p_n has n real distinct roots in (a,b)

* Proof by contradiction:

- Let $x_1, \dots, x_k$ be real root of $p_n(x)$ which are of odd multiplicity
- If $k = n$ our statement is true
- If $1 \le k < n$ then

$$q(x) = (x - x_1)\dots(x - x_k)$$

then $\int_a^b p_n(x) q(x) w(x)\, dx = 0$ $\Rightarrow$ $p_n q$ has only roots of even multiplicity 2

hence $p_n q$ doesn't change sign

hence $p_n(x) \equiv 0$

* For Chebyshev polynomial are orthogonal polynomials with weight $w(x) = \frac{1}{\sqrt{1-x^2}}$

$T_n (-1, 1) \longrightarrow R$:

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

$$\langle T_n, T_m \rangle = \begin{cases} \|T_n\|^2 & \text{when } m = n \\ 0 & \text{when } m \ne n \end{cases}$$

* Gauss quadratures. (About computing integral with weight $w(x)$)

* $I(f) = \int_a^b f(x)\, w(x)\, dx$

We look for $Q(f) = \sum_{i=0}^{\ell} \lambda_i f_i(x_i)$ — notes in (a,b) $\quad \lambda_i > 0$

The weights $\lambda_0, \dots, \lambda_k$ are chosen in such a way that $Q(f)$ is exact for $f \in \mathbb{P}_k$

Gauss quadratures: chose $x_0, \dots, x_\ell$ so as to obtain exactness of degree $(2\ell+1)$

* Lemma:

A quadrature rule Q with nodes $\{x_0, \dots, x_\ell\}$ which are equal to the root of $P_{\ell+1}(x)$ which the $(\ell+1)$ orthogonal polynomial in $L^2_w(a,b)$ and with weight $\lambda_0, \dots, \lambda_\ell$

$$\lambda_i = \int_a^b \ell_i(x)\, w(x)\, dx \qquad \ell_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{\ell} \frac{(x - x_j)}{(x_i - x_j)}, \; i = 0, \dots, \ell \qquad \ell_0(x) = 1$$

is exact for polynomial of degree $\cancel{\ell}(2\ell+1)$ $\qquad Q(f) = I(f) \quad f \in \mathbb{P}_{2\ell+1}$

Such quadrature is unique $\qquad Q$ is not exact for $\mathbb{P}_{2\ell+2}$

* Proof:

Define the nodes $x_0, x_1, \dots, x_\ell$ as roots of $P_{\ell+1}$

$$\lambda_i = \int_a^b \ell_i(x) w(x)\, dx$$

- First, show that Q is exact for $f \in \mathbb{P}_k$

$$Q(f) = \sum_{i=0}^{k} \lambda_i f(x_i) = \sum_{i=0}^{k} \left[\left(\int_a^b \ell_i(x)\, w(x) dx \right) f(x_i) \right] = \int_a^b \underbrace{\sum_{i=0}^{k} \ell_i(x) f(x_i)}\, w(x)\, dx$$

$$= \int_a^b L(f)(x)\, w(x) = \int_a^b f(x)\, w(x)\, dx = I(f)$$

- $0 = Q(P_{\ell+1}) = \langle 1, P_{\ell+1} \rangle = \int_a^b P_{\ell+1}(x)\, w(x)\, dx = I(P_{\ell+1})$

- If $f \in \mathbb{P}_{2\ell+1}$, then upon dividing it by $P_{\ell+1}$

$$f(x) = q(x)\, P_{\ell+1}(x) + r(x) \qquad q, r \in \mathbb{P}_\ell$$

Suppose that $P_{k+1}(x) = (x-x_0)\dots(x-x_k)$

$f(x_i) = r(x_i) \qquad i=0,\dots,k$

$$\int_a^b f(x)\, w(x)\, dx = \underbrace{\int_a^b q(x)\, P_{k+1}(x)\, w(x)\, dx}_{\substack{=0 \text{ because } q\in P_k \\ P_{k+1}(x) \text{ orthogonal}}} + \underbrace{\int_a^b r(x)\, w(x)\, dx}_{Q(r) = Q(f).}$$

$\Rightarrow I(f) = Q(f)$

* Prove uniqueness.

Take $\Pi(x) = (x-x_0)\dots(x-x_k)$ deg $k+1$

$\Pi^2(x)$ = is of degree $(2k+2)$

If $I(\Pi^2) = Q(\Pi^2)$

$0 < I(\Pi^2) = Q(\Pi^2) = 0$

$\lambda_i > 0, \quad \lambda_i = \sum_{j=0}^{k} \lambda_j\, l_i^2(x_j) = Q(\overset{\deg 2k}{l_i^2}) = \int_a^b l_i^2(x)\, w(x)\, dx > 0$

- Error $I(f) - Q(f)$ (

Let $\boxed{H \in P_{2k+1}}$ be the Hermite interpolation for f

$H^{(l)}(x) = f^{(l)}(x_i) \qquad l=0,1 \quad i=0,\dots,k.$

$f(x) = H(x) + f[x_0,2;\dots;x_k,2,x]\,\Pi^2(x)$

- Since Q is exact for H in P_{2k+1}

$I(H) = Q(H) = Q(f)$

$E(f) = I(f) - Q(f) = I(f) - I(H) = I(f-H) =$

$$= \int_a^b f[x_0,2;\dots;x_k,2,x]\,\Pi^2(x)\, w(x)\, dx$$

$$= f[x_0,2;\dots,x_k,2,\xi] \int_a^b \Pi^2(x)\, w(x)\, dx$$

$$= \frac{1}{(2k+2)!} f^{(2k+2)}(\eta) \int_a^b \Pi^2(x)\, w(x)\, dx$$

*Gauss Lobatto rule

* We want to estimate $I(f) = \int_a^b f(x)\, w(x)\, dx$

- Choose nodes:

$$\begin{cases} x_0 = a \qquad x_\ell = b \qquad (\ell-1) \\ x_1, \dots, x_{\ell-1} \text{ are roots of orthogonal polynomials on } (a,b) \text{ with } \overline{W}(x) = (x-a)(b-x)\, w(x) \end{cases}$$

- Choose weights

$$\lambda_i = \int_a^b l_i(x)\, w(x)\, dx \qquad l_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \frac{(x-x_j)}{(x_i - x_j)}$$

Then $Q(f)$ is exact for polynomial $f \in \mathbb{P}_{2\ell-1}$

* Proof:

- Because Q is interpolating, it is exact for $p \in \mathbb{P}_\ell$. (1).
- Consider $\langle f, g\rangle = \int_a^b f(x)\, g(x)\, \overline{W}(x)\, dx$
  - $p_0, p_1, \dots, p_{\ell-1}$ orthogonal polynomials $\Rightarrow$ it has as many roots as the degree

so $p_{\ell-1} = (x - x_1) \dots (x - x_{\ell-1})$ $\leftarrow$ since we chose $x_1, \dots, x_{\ell-1}$ are root of $p_{\ell-1}$

* Let $f \in \mathbb{P}_{2\ell-1}$, we want to show that Q is of Gauss Lobatto exact

take $f \in \mathbb{P}_{2\ell-1}$ and devide f by $(x-a)(b-x)\, p_{\ell-1}(x)$

- $f(x) = \underbrace{q(x)}_{\in \mathbb{P}_{\ell-2}} (x-a)(b-x)\, p_{\ell-1}(x) + \underbrace{r(x)}_{\in \mathbb{P}_\ell}$

- Then we have

$$\int_a^b f(x)\, w(x)\, dx = \int_a^b q(x)(x-a)(b-x)\, p_{\ell-1} w(x)\, dx + \int r(x)\, w(x)\, dx.$$

$$= \underbrace{\int_a^b q(x)\, p_{\ell-1}(x)\, \overline{W}(x)\, dx}_{=0 \text{ since orthogonal with base } \overline{W}(x).} + \int_a^b r(x)\, w(x)\, dx \quad = Q(r) = \sum_{i=0}^{\ell} \lambda_i r(x_i) = \sum_{i=0}^{\ell} \lambda_i f(x_i) = Q(f). \; (1)$$

and note that since we chose $x_1, \dots, x_{\ell-1}$ are roots of $p_{\ell-1}(x)$.

then $f(x) = q(x) \underbrace{(x-a)(b-x)\, p_{\ell-1}(x)}_{=0 \text{ when } x = x_0, \dots, x_\ell} + r(x) \Rightarrow f(x_j) = r(x_j),\ \forall j = \overline{0, \ell}$

* Example : Let $a>0$. Consider $Q(f) = \lambda_1 f(-a) + \lambda_2 f(0) + \lambda_3 f(a)$ for computing $I(f) = \int_{-1}^{1} f(x)\,dx$

a) Determine the weight $\lambda_1, \lambda_2, \lambda_3$ (in terms of a) so that Q is exact for polynomial of degree ≤ 2

b) For what $a>0$, the weight are positive ?

c) Show that for $a=\sqrt{\frac{3}{5}}$, Q is exact for polynomial of degree ≤ 5 .

a) Q is exact for polynomials of degree ≤ 2 ⇒ is exact. for $f = 1, x$ and x^2

- $f=1$ then • $I(f) = \int_{-1}^{1} 1\,dx = 2$ • $Q(f) = \lambda_1+\lambda_2+\lambda_3$ ⇒ $\lambda_1+\lambda_2+\lambda_3 = 2$.
- $f = x$ • $I(f) = \int_{-1}^{1} x\,dx = \frac{x^2}{2}\Big|_{-1}^{1} = 0$ • $Q(f) = -a\lambda_1 + a\lambda_3$ ⇒ $-a\lambda_1 + a\lambda_3 = 0$
- $f = x^2$ • $I(f) = \int_{-1}^{1} x^2\,dx = \frac{x^3}{3}\Big|_{-1}^{1} = \frac{2}{3}$ • $Q(f) = a^2\lambda_1 + a^2\lambda_3$ ⇒ $a^2\lambda_1 + a^2\lambda_3 = \frac{2}{3}$.

So we have
$$\begin{pmatrix} 1 & 1 & 1 \\ -a & 0 & a \\ a^2 & 0 & a^2 \end{pmatrix}\begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 2/3 \end{pmatrix} \Rightarrow \begin{cases} \lambda_1 = \lambda_3 \\ \lambda_1 + \lambda_3 = \frac{2}{3a^2} \\ \lambda_2 = 2 - \lambda_1 - \lambda_3 = 2 - 2\lambda_1 = 2 - 2\frac{1}{3a^2} = 2\frac{(3a^2-1)}{3a^2} \end{cases} \Rightarrow \lambda_1 = \lambda_3 = \frac{1}{3a^2}$$

b) The weights are positive when $3a^2 - 1 > 0 \Rightarrow 3a^2 > 1 \Rightarrow \begin{cases} a < -\frac{1}{\sqrt{3}} \\ a > \frac{1}{\sqrt{3}} \end{cases} \xrightarrow[a>0]{\text{since}} a > \frac{1}{\sqrt{3}}$

c) • By the way we construct Q, we have Q is exact for polynomial of degree ≤ 2 .

We now want to prove that when $a = \sqrt{\frac{3}{3}}$, Q is also exact for polynomial of degree $= 3, 4, 5$.

when $a = \sqrt{\frac{3}{5}}$ $\lambda_1 = \lambda_3 = \frac{5}{9}$ $\lambda_2 = \frac{8}{9}$

Then $Q(f) = \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{26}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right)$.

*We now want to check that Q is exact for polynomial of degree 3, 4, 5 .

- We have when $f = x^3$ (degree 3) and $f = x^4$ (degree $=4$) then

  $I(f) = \int_{-1}^{1}$ odd function $= 0$ $Q(f) = 0$ ⇒ exact .

- when $f = x^4$ (degree 4) .

  $I(f) = \int_{-1}^{1} x^4\,dx = \frac{2}{5}$ $Q(f) = \frac{5}{9} \times 2 \left(\sqrt{\frac{3}{5}}\right)^4 = \frac{2}{5}$ ⇒ exact ⇒ done .

* Legendre polynomials : (are generated using triple recursion formula) in the open interval $(-1,1)$ and weight function $w(x) = 1$

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,dx$$

Gauss Chebyshev
$(-1,1)$ and $w(x) = \frac{1}{\sqrt{1-x^2}}$

- $P_0(x) = 1$

$$P_1(x) = x - \frac{\langle x, 1\rangle}{\langle 1, 1\rangle} 1 = x - \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} 1\,dx} 1 = x$$

$$P_2(x) = \left(x - \frac{\langle x x, x\rangle}{\langle x, x\rangle}\right) x - \frac{\langle x, x\rangle}{\langle 1, 1\rangle} 1 = \left(x - \frac{\int_{-1}^{1} x^3\,dx}{\int x^2\,dx}\right) x - \left(\frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} dx}\right) 1 = x^2 - \frac{1}{3}$$

$$P_3(x) = \dots = x^3 - \frac{3}{5} x \ \dots$$

* Gauss Hermite quadrature | orthogonal with $[-\infty, \infty]$ $w(x) = e^{-x^2}$

* We want to compute $I(f) = \int_{-\infty}^{+\infty} f(x)\, e^{-x^2}\,dx$

* We want to find the point x_0, x_1 and weight A_0, A_1 such that $Q(f) = A_0 f(x_0) + A_1 f(x_1)$ is exact for polynomial of degree ≤ 3

$$= \frac{\sqrt{\pi}}{2} f\left(-\frac{1}{\sqrt{2}}\right) + \frac{\sqrt{\pi}}{2} f\left(\frac{1}{\sqrt{2}}\right)$$ ← only 2 points 2 weights that can help $Q(f)$ exacts for polynomial of degree ≤ 3

* We want $Q(f)$ so that it is exact for polynomial of degree ≤ 3

⇒ We want Q such that $Q(x^i) = I(x^i)$, when $i = 0, 1, 2, 3$

- $I(x^0) = \int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}$

$$I(x^1) = \int_{-\infty}^{+\infty} x\, e^{-x^2}\,dx = 0$$

$$I(x^2) = \int_{-\infty}^{+\infty} x^2\, e^{-x^2}\,dx = \sqrt{\frac{\pi}{2}}$$

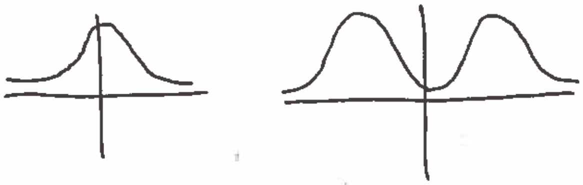
$$I(x^3) = \int_{-\infty}^{+\infty} x^3 e^{-x^2}\,dx = 0$$

- Prove $I(x^2) = \sqrt{\frac{\pi}{2}}$

$$I(x^2) = \int_{-\infty}^{+\infty} x^2\, e^{-x^2}\,dx = \frac{-1}{2}\int_{-\infty}^{+\infty} x\left(e^{-x^2}\right)'\,dx =$$

$$= -\frac{1}{2}\left[x\, e^{-x^2}\Big|_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} e^{-x^2}\,dx \right] =$$

$$= -\frac{1}{2}(-\pi) = \sqrt{\frac{\pi}{2}}.$$



* Since $Q(x^i) = I(x^i)$ we have

$i=0 \quad Q(x^0) = A_0 + A_1 = I(x^0) = \sqrt{\pi}$

$i=1 \quad Q(x^1) = A_0x_0 + A_1x_1 = I(x^1) = 0$

$i=2 \quad Q(x^2) = A_0x_0^2 + A_1x_1^2 = I(x^2) = \sqrt{\frac{\pi}{2}}$

$i=3 \quad Q(x^3) = A_0x_0^3 + A_1x_1^3 = I(x^3) = 0$

$$\Rightarrow \begin{cases} A_0 + A_1 = \sqrt{\pi} & (1) \\ A_0x_0 + A_1x_1 = 0 & (2) \\ A_0x_0^2 + A_1x_1^2 = \sqrt{\frac{\pi}{2}} & (3) \\ A_0x_0^3 + A_1x_1^3 = 0 & (4) \end{cases}$$

constants

* By Gauss quadrature, x_0, x_1 are roots of $P_{k+1}(x) = \pi(x) = x^2 + p(x) + q$
and we want to find the constants p and q.

and note that x_0, x_1 are roots $\Rightarrow \boxed{\begin{array}{l} x_0^2 + px_0 + q = 0 \\ x_1^2 + px_1 + q = 0 \end{array}}$

- mul (1) by q, mul (2) by p, mul (3) by 1 and add.

$$\begin{cases} qA_0 + qA_1 = q\sqrt{\pi} \\ pA_0x_0 + pA_1x_1 = 0 \\ A_0x_0^2 + A_1x_1^2 = \sqrt{\frac{\pi}{2}} \end{cases} \Rightarrow A_0\underbrace{(x_0^2 + px_0 + q)}_{=0} + A_1\underbrace{(x_1^2 + px_1 + q)}_{=0} = \sqrt{\pi}\left(q + \frac{1}{2}\right)$$

$$\Rightarrow q = -\frac{1}{2}.$$

- Now we want to find p

mul (2) by q, mul (3) by p, mul (4) by (1) and add

$$\begin{cases} A_0qx_0 + A_1qx_1 = 0 \\ A_0px_0^2 + A_1px_1^2 = p\sqrt{\frac{\pi}{2}} \\ A_0x_0^3 + A_1x_1^3 = 0 \end{cases} \rightarrow A_0x_0\underbrace{(x_0^2 + px_0 + q)}_{=0} + A_1x_1\underbrace{(x_1^2 + p(x_1) + q)}_{=0} = \frac{\pi}{2}p$$

$$\Rightarrow p = 0$$

- So we have $\pi(x) = x^2 - \frac{1}{2}$ and x_0 and x_1 are roots of this polynomial $\Rightarrow \boxed{x_0 = -\frac{1}{\sqrt{2}}, \; x_1 = \frac{1}{\sqrt{2}}}$

* So now we want to find A_0 and A_1.

$$\begin{cases} \sqrt{\pi} = A_0 + A_1 \\ 0 = A_0x_0 + A_1x_1 \end{cases} \Rightarrow \begin{cases} A_0 = \frac{\sqrt{\pi}}{2} \\ A_1 = \frac{\sqrt{\pi}}{2} \end{cases}$$

So we have $Q(f) = \frac{\sqrt{\pi}}{2} f\left(-\frac{1}{\sqrt{2}}\right) + \frac{\sqrt{\pi}}{2} f\left(\frac{1}{\sqrt{2}}\right)$ □

Computing Gauss quadrature using Haar condition for the system of orthogona polynomials.

* Consider

point values of function.

$$P = \begin{bmatrix} P_0(x_0) & & P_0(x_\ell) \\ & & \\ P_\ell(x_0) & & P_\ell(x_\ell) \end{bmatrix} \qquad \{p_0, \dots, p_\ell\} \quad \{x_0, \dots, x_\ell\}$$

If $p_0, p_1, \dots, p_k$ are orthogonal polynomials then P is invertible.

(rows are linearly independent /columns.)

* Proof

Suppose that $p_0, \dots, p_\ell$ are orthogonal but P is not invertible.

P is not invertible $\Rightarrow$ rows are linearly dependent.

$\underset{\sim}{c}^T P = \underset{\sim}{0}$ and $\underset{\sim}{c} \neq \underset{\sim}{0}$.

$$-[\;] = -$$

- we have $c^T P = \left[c_1 (p_0(x_0) + \dots + p_\ell(x_0)) \,\middle|\, \dots \,\middle|\, c_\ell (p_0(x_\ell) \dots + p_\ell(x_\ell)) \right]$

$$= \left[\cancel{q} \; q_1(x) \,\middle|\, q_2(x) \,\middle|\, \dots \,\middle|\, q_\ell(1) \right] = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}$$

* If the weights of the Gauss quadrature $Q(f)$ are computed from

$$P \begin{bmatrix} \lambda_0 \\ \vdots \\ \lambda_k \end{bmatrix} = \begin{bmatrix} \langle p_0, p_0 \rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- If $Q(f) = \sum_{i=0}^{k} \lambda_i f(x_i)$

then $Q(f) = I(f) = \int_a^b f(x) w(x)\,dx \qquad f \in \mathbb{P}_{2\ell+1}$

- when $p \in \mathbb{P}_{2\ell+1}$ $\quad p(x) = p_{\ell+1}(x) q(x) + r(x) \qquad q(x) = \sum_{j=0}^{k} \alpha_j p_j \qquad r(x) = \sum_{j=1}^{k} \beta_j p_j$

Then $I(p) = \langle p, 1 \rangle = \langle p_{\ell+1} q + r, 1 \rangle = \langle p_{\ell+1} q, 1 \rangle + \langle r, 1 \rangle =$ all of β_j (expect β_0) are orthogonal to 1 ?

$$= \underbrace{\langle p_{\ell+1}, q \rangle}_{=0} + \langle r, 1 \rangle = \langle r, 1 \rangle$$

$$= \cancel{\sum_{j=1}^{k}} \beta_0 \langle p_0, p_0 \rangle$$

$$Q(p) = \sum_{i=0}^{k} \lambda_i \, p(x_i) = \sum_{i=0}^{\ell} \lambda_i \, r(x_i) = \sum_{j=0}^{k} \beta_j \left(\sum_{i=0}^{k} \lambda_i \, p_j(x_i) \right) = \beta_0 \langle p_0, p_0 \rangle$$

$$\sum_{i=0}^{\ell} p_n(x_i) \, \lambda_i = \begin{cases} \langle p_0, p_0 \rangle & i=0 \\ 0 & n = 1, \dots, k \end{cases}$$

+ Quardrature weight function .

$w(x) = 1$ on ~~$[a,b]$~~ Gauss Legendre $(-1, 1)$

$w(x) = \sqrt{1-x^2}$ on $[-1,1]$

$w(x) = \dfrac{1}{\sqrt{1-x^2}}$ on $[-1,1]$ Gauss Chebyshev

$w(x) = \exp(-x^2)$ on $(-1,1)$ Gauss Hemite .

$\langle x, y \rangle \geq 0, \forall x, y$

* Approximation in an inner product space

- Let V be a vector space in $\boxed{\mathbb{R}}$.

$\langle\cdot,\cdot\rangle : V\times V \longrightarrow \boxed{\mathbb{R}}$

$$\begin{cases} \langle u,u\rangle \geq 0 \quad , \langle u,u\rangle = 0 \Leftrightarrow u=0 \\ \langle u,v\rangle = \langle v,u\rangle \\ \langle u+w, v\rangle = \langle w,v\rangle + \langle w, v\rangle \\ \langle \lambda u, v\rangle = \lambda\langle u,v\rangle \end{cases}$$

also in the second argument.

- If V is over $\boxed{\mathbb{C}}$, $\langle\cdot,\cdot\rangle : V\times V \longrightarrow \boxed{\mathbb{C}}$

$\langle \lambda u,v\rangle = \lambda\langle u,v\rangle$

$\langle u,v\rangle = \overline{\langle v,u\rangle}$

$\langle u,\lambda v\rangle = \bar{\lambda}\langle u,v\rangle$

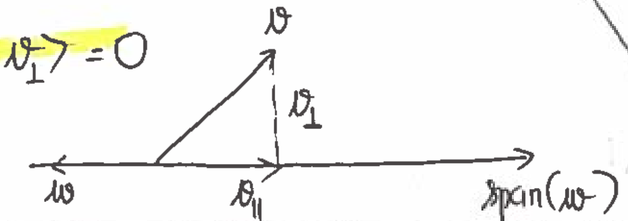
$= \overline{\langle \lambda v,u\rangle} = \bar{\lambda}\,\overline{\langle v,u\rangle} =$

- Orthogonal projection and Cauchy Schwartz

Let $v, w \in V$ $\quad \boxed{w\neq 0}$

There exist unique vector $v_{\parallel}, v_{\perp} \in V$ such that.

$$\begin{cases} v = v_{\parallel} + v_{\perp} \\ v_{\parallel} = cw \\ \langle v_{\perp}, w\rangle = 0 \end{cases} \quad \rightarrow \langle v_{\parallel}, v_{\perp}\rangle = 0$$



* Show the uniqueness of the decomposition:

Let $v = v_{\parallel}' + v_{\perp}' = v_{\parallel} + v_{\perp}$

because $\left.\begin{array}{l} \langle v_{\perp}', w\rangle = 0 \\ \langle v_{\perp}, w\rangle = 0 \end{array}\right\} \Rightarrow \langle v_{\perp}' - v_{\perp}, \underbrace{w}_{\vec{w}\neq\vec{0}}\rangle = 0 \Big\} \Rightarrow v_{\perp}' = v_{\perp}$

$\Rightarrow v_{\parallel}' - v_{\parallel} = v_{\perp} - v_{\perp}' = \vec{0} \Rightarrow v_{\parallel}' = v_{\parallel}$ -

$\Big\} \Rightarrow$ the uniqueness.

* Show the existence

Put $c = \dfrac{\langle v,w\rangle}{\|w\|^2}$ $\quad v_{\parallel} = cw \quad v_{\perp} = v - cw$

- So now we only have to check $\langle v_{\perp}, w\rangle = 0$

$\langle v_{\perp}, w\rangle = \langle v - cw, w\rangle = \langle v,w\rangle - c\langle w,w\rangle = \langle v,w\rangle - \dfrac{\langle v,w\rangle}{\|w\|^2}\|w\|^2 = 0$

* Geometric prove of Cauchy Schwarz : $|\langle v, w\rangle| \leqslant \|w\| \, \|v\|$ $\forall \vec{v}, \vec{w} \in V$
where $\|v\| = \langle v, v\rangle^{1/2}$

- If $\underset{\sim}{w} = \underset{\sim}{0}$ ⇒ true.
- Consider when $\underset{\sim}{w} \neq \underset{\sim}{0}$

Use $v = v_\parallel + v_\perp$ → $\|v\|^2 = \|v_\parallel\|^2 + \|v_\perp\|^2 + 2\underbrace{\langle v_\parallel, v_\perp\rangle}_{=0} = \|v_\parallel\|^2 + \|v_\perp\|^2 \geqslant \|v_\parallel\|^2$

$v_\parallel = cw = \frac{\langle v, w\rangle}{\|w\|^2} w$ | $> \frac{|\langle v, w\rangle|^2}{\|w\|^4}\|w\|^2 = \frac{|\langle v, w\rangle|^2}{\|w\|^2}$

⇒ $\|v\|^2 \|w\|^2 \geqslant |\langle v, w\rangle|^2$

⇒ what we need to prove □

The Cauchy Schwarz expresses how u is linearly dependent of v
If $u = \alpha v$ then we have equality in C-S ($\vec{u} = \alpha\vec{v} \Leftrightarrow \vec{u}$ and $\vec{v}$ have the same direction and $\frac{\vec{u}}{\|u\|} = \pm \frac{\vec{v}}{\|\vec{v}\|}$)

$\langle u + \lambda v, u + \lambda v\rangle \geqslant 0$

$\langle u, u\rangle + \bar{\lambda}\langle u, v\rangle + \lambda(\langle v, u\rangle + \bar{\lambda}\langle v, v\rangle) \geqslant 0$

$\forall \lambda \in \mathbb{C}$, take $\lambda = -\frac{\langle u, v\rangle}{\langle v, v\rangle}$

If equality holds in C-S and if $\lambda = -\frac{\langle u, v\rangle}{\langle v, v\rangle}$ $\langle u + \lambda v, u + \lambda v\rangle = 0$
⇒ $u + \lambda v = 0$

$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ How do we express the proprotionality of vectors.

$$\frac{u_i}{v_i} = \frac{u_j}{v_j} \Leftrightarrow u_i v_j = u_j v_i \quad \forall i, j$$

$$\sum_{i,j}(u_i v_j - u_j v_i)^2 = 0$$

$$= \sum_i \sum_j (u_i^2 v_j^2 + u_j^2 v_i^2 - 2 u_i v_j u_j v_i) =$$

$$= \left(\sum_i u_i^2\right)\left(\sum_j v_j^2\right) + \left(\sum_j u_j^2\right)\left(\sum_i v_i^2\right) - 2\left(\sum_i u_i v_i\right)\sum_j (u_j v_j)$$

$$= 2\left(\sum_i u_i^2\right)\left(\sum v_i^2\right) - 2\left(\sum_i u_i v_i\right)^2 = 0$$

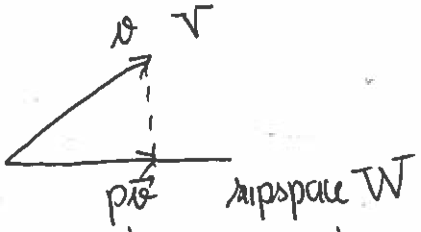
* When $\vec{u} = \alpha \vec{v}$, we have $\vec{u}\|v\| - \vec{v}\|\vec{u}\| = \vec{0}$

$\Leftrightarrow \left\| \vec{u}\|\vec{v}\| - \vec{v}\|\vec{u}\| \right\| = 0$.

$\Leftrightarrow \langle \vec{u}\|\vec{v}\| - \vec{v}\|\vec{u}\| , \vec{u}\|\vec{v}\| - \vec{v}\|\vec{u}\| \rangle = 0$

$\Leftrightarrow 2\|u\|^2\|v\|^2 - \|u\|\|v\|\left(\langle u,v\rangle + \langle v,u\rangle\right) > 0$

This inequality remains true if v is replaced by αv where $\langle u, \alpha v\rangle = |\langle u,v\rangle|$



(projection of v onto subspace W

$\|v - pv\|$ is minimal

$\|v - pv\| \leq \|v - w\|$, $\forall w \in W$

* Example of projecting a function onto a function spanby

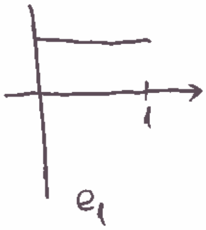
Consider $f \in L^2([0,1])$, $f(x) = \begin{cases} x, & 0 \leq x \leq \frac{1}{2} \\ -x+1, & \frac{1}{2} \leq x < 1 \end{cases}$

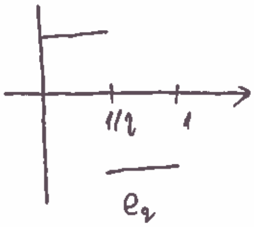
Find the projection of f onto space $W = \text{span}\{e_1, e_2, e_3, e_4\}$.

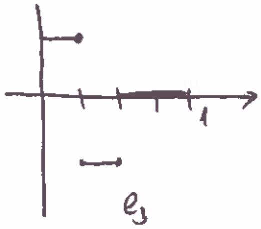
$e_1(x) = \phi(x)$, $e_2(x) = \psi(x)$, $e_3(x) = \psi(2x)$, $e_4(x) = \psi(2x-1)$ ~~wher~~

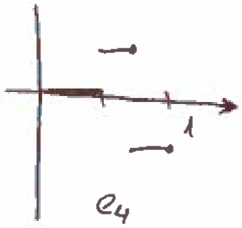
$\phi(x) = 1$, $x \in [0,1]$

$\psi(x) = \begin{cases} 1, & 0 \leq x < \frac{1}{2} \\ -1, & \frac{1}{2} \leq x < 1 \\ 0, & \text{otherwise} \end{cases}$









* We have the projection of f onto space spanned by e_1, e_2, e_3 and e_4 is

$$f^*(x) = \sum_{i=1}^{4} \frac{\langle f, e_i \rangle}{\langle e_i, e_i \rangle} e_i \longleftarrow$$

we can only do this when $\{e_1, e_2, e_3, e_4$ are orthogonal (don't need to be orthonormal)

- $\langle e_1, e_1 \rangle = \int_0^1 dx = 1$ $\quad$ $\langle f, e_1 \rangle = \int_0^{1/2} x\, dx + \int_{1/2}^{1} (-x+1)\, dx = \frac{1}{4}$

$\langle e_2, e_2 \rangle = 1$ $\quad$ $\langle f, e_2 \rangle = \int_0^{1/2} f(x) + \int_{1/2}^{1} -f(x)\, dx = 0$

$\langle e_3, e_3 \rangle = 1/2$ $\quad$ $\langle f, e_3 \rangle = -\frac{1}{2} \frac{1}{2} \frac{1}{4} = -\frac{1}{16}$

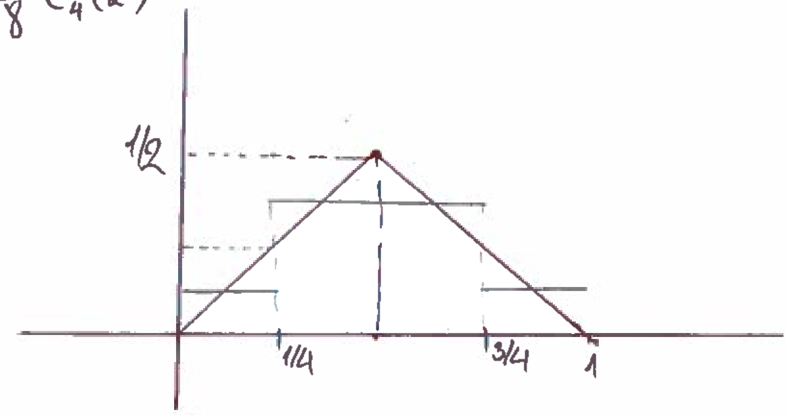
$\langle e_4, e_4 \rangle = \frac{1}{2}$ $\quad$ $\langle f, e_4 \rangle = \frac{1}{16}$

- So we have

$$f^* = \frac{1/4}{1} e_1(x) + \frac{0}{1} e_2(x) + \frac{(-1/16)}{(1/2)} e_3(x) + \frac{1/16}{1/2} e_4(x)$$

$$= \frac{1}{4} e_1(x) - \frac{1}{8} e_3(x) + \frac{1}{8} e_4(x)$$

$$= \begin{cases} \frac{1}{8} & 0 < x < \frac{1}{4} \\ \frac{3}{8} & \frac{1}{4} < x < \frac{1}{2} \\ \frac{3}{8} & \frac{1}{2} < x < \frac{3}{4} \\ \frac{1}{8} & \frac{3}{4} < x < 1 \end{cases}$$



* Projection of function onto a span of an orthogonal set

* $f : [a,b] \longrightarrow R$

$$\langle f, g \rangle = \int_a^b f(x)\, g(x)\, dx \qquad \|f\|^2 = \int_a^b [f(x)]^2 dx \quad < +\infty$$

$$L^2([a,b])$$

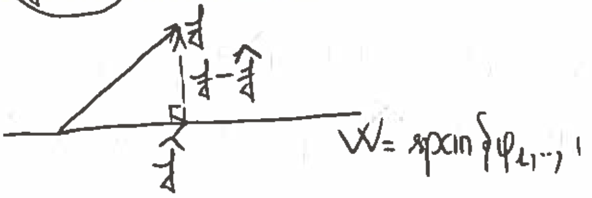
* Let $\{\varphi_1, \dots, \varphi_N\}$ be an orthogonal set of functions in $L^2([a,b])$

$W = \text{span}\{\varphi_1, \dots, \varphi_N\}$

We define the projection of f onto W as an element $\hat{f} \in W$ such that

$\langle f - \hat{f}, \varphi \rangle = 0, \quad \forall \varphi \in W$

- Find $\hat{f} = \sum_{i=1}^{N} \hat{c}_i \varphi_i \Rightarrow$ want to find $\widehat{c_i}$, $i = \overline{1,N}$



Take $\varphi = \varphi_i$, $\langle f - \hat{f}, \varphi_i \rangle = 0$, $\forall i = \overline{1,N}$

$$\Rightarrow \langle f - \sum_{i=1}^{N} \hat{c}_i \varphi_i, \varphi_i \rangle = 0 = 0, \quad \forall i = \overline{1,N}$$

$$\Rightarrow \langle f, \varphi_i \rangle - \sum_{i=1}^{N} \hat{c}_i \langle \varphi_i, \varphi_i \rangle = 0$$

$$\Rightarrow \sum_{j=1}^{N} \langle \varphi_j, \varphi_i \rangle \hat{c}_j = \langle f, \varphi_i \rangle, \quad i = \overline{1,N}$$

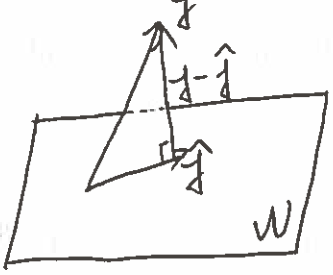
$$\begin{bmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_2, \varphi_1 \rangle & \cdots & \langle \varphi_N, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_2 \rangle & \langle \varphi_2, \varphi_2 \rangle & \cdots & \langle \varphi_N, \varphi_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_2, \varphi_N \rangle & \langle \varphi_2, \varphi_N \rangle & \cdots & \langle \varphi_N, \varphi_N \rangle \end{bmatrix} \begin{bmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \vdots \\ \hat{\varphi}_N \end{bmatrix} = \begin{bmatrix} \langle f, \varphi_1 \rangle \\ \langle f, \varphi_2 \rangle \\ \vdots \\ \langle f, \varphi_N \rangle \end{bmatrix}$$

Because $\{\varphi_1, \dots, \varphi_N\}$ is orthogonal

$$\Rightarrow \langle \varphi_i, \varphi_i \rangle \hat{c}_i = \langle f, \varphi_i \rangle, \quad i = \overline{1,N}$$

$$\Rightarrow \hat{c}_i = \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle},$$

and so, $\hat{f} = \sum_{i=1}^{N} \frac{\langle f, \varphi_i \rangle}{\langle \varphi_i, \varphi_i \rangle} \varphi_i$



Def:

ve say that $f^* \in W$ is the best approximation of f in W if

$\|f - f^*\| \leq \|f - \varphi\|$, $\forall \varphi \in W$

— f^* don't need to be unique in common vect spa

— in L^2 it is unique.

Theorem: (orthogonal projection theorem)

f $f \in L^2([a,b])$ then $f^* = \hat{f}$, where $\hat{f}$ is the orthogonal projection of f onto W

(orthogonal projection of f is the best approximation of f in W)

Proof: Suppose that $f^* = \hat{f}$ is the ~~best~~ , prove that $f^* = \hat{f}$ is the best projection .

$|f - \varphi| = |f - \hat{f}^* + \hat{f}^* - \varphi$

ien $\langle f - \hat{f}^*, \varphi \rangle = 0$, $\forall \varphi \in W$

cause $(f^* - \varphi) \in W \Rightarrow \langle f - \hat{f}, \hat{f} - \varphi \rangle = 0$, $\forall \varphi \in W$

$\to \|f - \varphi\|^2 = \|f - \hat{f}\|^2 + \|\hat{f} - \varphi\|^2 + 2\underbrace{\langle f - \hat{f}, \hat{f}, \varphi \rangle}_{=0}$

$\Rightarrow \|f - \varphi\|^2 > |f - \hat{f}|$

Let f^* is the best approximation , prove that $f^* = \hat{f}$

ince we need to prove that $f^* = \hat{f}$, we need to prove that $\langle f - f^*, \varphi \rangle = 0$, $\forall \varphi \in W$

is enough to prove that $\langle f - f^*, \varphi_i \rangle$, $i = \overline{1, n}$ $(f - f^*)$ is orthogonal to each 1 dim subspace of W

$d = d(f, W) = \|f - f^*\| = \|f - f^* - 0\| \geqslant d(f - f^*, W_1) \geqslant d(f - f^*, W)$

$\|f - f^* - 0\| = d(f - f^*, W_1)$

says that the zero vector is the best approximation of $f - f^*$ in W_1

ce there exists a projection of $f - f^*$ onto W_1, then from part the projection of $f - f^*$ to W_1 is 0

Let

$$\mathcal{G} = \begin{bmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_2, \varphi_1 \rangle & \cdots & \langle \varphi_n, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_n \rangle & \langle \varphi_2, \varphi_n \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle \end{bmatrix}$$

Gramm matrix .

Let

$$\underset{\text{Gramm matrix}}{G} = \begin{bmatrix} \langle \varphi_1, \varphi_1 \rangle & \langle \varphi_2, \varphi_1 \rangle & \cdots & \langle \varphi_N, \varphi_1 \rangle \\ & & & \\ \langle \varphi_1, \varphi_N \rangle & \langle \varphi_2, \varphi_N \rangle & \cdots & \langle \varphi_N, \varphi_N \rangle \end{bmatrix}$$

* Gramm matrix properties

1> Gramm matrix is a Hermitian nonegative definite matrix.

2> If $\varphi_1, \varphi_2, \dots, \varphi_N$ is linearly independent ⟹ G is positive definite.

3> Every possitive definite matrix is a Gramm matrix of a particular basic wrt a parti

proc

- Prove that G is positive definite when $\varphi_1, \dots, \varphi_N$ are linearly independent.

Need to prove that $\langle x, Gx \rangle > 0$ for $x \neq 0$

$$(Gx)_i = \sum_{j=1}^{n} G_{ij}\, x_j = \sum_{j=1}^{n} \langle \varphi_j, \varphi_i \rangle x_j$$

$$\langle x, Gx \rangle = \sum_{i=1}^{N} x_i (Gx)_i = \sum_i \sum_j x_i \overline{x_j} \langle \varphi_i, \varphi_j \rangle = \Big\langle \sum_i x_i \varphi_i , \sum_j x_j \varphi_j \Big\rangle =$$

$$= \Big\| \sum_{i=1}^{N} x_i \varphi_i \Big\|^2 = \|x\|^2 > 0$$

Generalities

Theorem: (Fundamental theorem on approximation)

et $W \subset V$ (W is a finite dimensional subspace of V which is norm space)

$\|\cdot\| : V \longrightarrow R$
$$\begin{cases} \|v\| \geq 0, \quad \|v\| = 0 \Leftrightarrow v = 0 \\ \|\alpha v\| = |\alpha| \, \|v\| \\ \|u+v\| \leq \|u\| + \|v\| \end{cases}$$

ien for any $f \in V$ there exists a best approximating (optimal) element $h^* \in W$

$\|f-h^*\| \leq \|f-h\| \qquad \forall h \in W$

$\underset{\sim}{0} \in W$

$\|f-h^*\| \leq \|f-0\| = \|f\|$

t $K = \{ h \in W, \ \|f-h\| \leq \|f\| \}$ is a closed ball in a finite dimensional subspace W

K is compact

e may minimize $\|f-h\|$

$\|f-h\|$ is a continuous function of h $\qquad \Big| \|f-(h+g)\| - \|f-h\| \Big| \leq \|g\|$

We say that V is stricly convex space iff (def)

$\|f\| = \|g\| = 1$ and $f \neq g \quad \rightarrow \quad \|f+g\| < 2$

Every inner product space V is stricly convex by parallelogram rule

$\|f+g\|^2 = 2(\|f\|^2 + \|g\|^2) - \|f-g\|^2$

$\left.\begin{array}{r} \text{if } \|f\| = \|g\| = 1 \\ f \neq g \end{array}\right\} \Rightarrow \|f+g\|^2 = 4 - \|f-g\|^2 \quad \Rightarrow \|f+g\| = \sqrt{4 - \|f-g\|^2} < 2$

$$\begin{cases} C[0,1] \text{ is not stricly convex} \\ \|f\| = \max\limits_{0 \leq x \leq 1} |f(x)| \end{cases}$$

let $f = 1$ $\qquad \|f\| = 1 \qquad \rightarrow \quad f+g = 1+t$

$g(t) = t \qquad \|g\| = 1 \qquad\qquad \|f+g\| = 2.$

$f \neq g$

- Prove that the unit ball in the space $C[0,1]$, $\|f\| = \max\limits_{0\le x\le 1} |f(x)|$ has a segment in

$\lambda f + (1-\lambda) g \quad 0 \leq \lambda \leq 1$ a segment in the space

- $\|\lambda f + (1-\lambda) g\| = 1$

* The approximation problem in a strictly convex named space

* The approximation problem in a strictly convex named space, has a unique solution

$$\left.\begin{array}{l}\|x\| \leq r \\ \|y\| \leq r\end{array}\right\} \Rightarrow \|x+y\| \leqslant 2r$$ → then we say the norm and the space is strictly convex

- Prove (by contradiction)
Suppose $g_1 \neq g_2$ are both best approximation elements for f

⊕ $\|f - g_2\| = \|f - g_1\| = E_W(f) = \inf\limits_{h\in W} \|f - h\|$

⊕ $\|2f - (g_1 + g_2)\| < 2E_W(f)$

$$2\|f - \underbrace{\frac{g_1+g_2}{2}}\| < E_W(f)$$

Linear LS, data fitting

* $(t_1, b_1), \dots, (t_m, b_m) \qquad t_i \in \mathbb{R}^k \quad b_i \in \mathbb{R}$

- Suppose that these data points represent some underlying function $f(t)$ such that

$f(t_i) = b_i \qquad i = \overline{1,m}$

We want to approximate by a linear combination of $\varphi_1, \varphi_2, \dots, \varphi_n$

- We will identify f with a vector in $\mathbb{R}^m$, $\sum\limits_{i=0}^{n} c_i \varphi_i(t)$ is called a model, c_i are param of a model

a) The model can be used to interpolate f and approximately predict values of f at points other than t_i

b) data compression: a few of model coefficients offer less storage than all of the data points

c) Smoothing: If φ are smooth the data b_i is noisy

Consider the column vectors b, f^*, φ_i

$$f^* = \sum_{i=1}^{n} c_i \varphi_i$$

$$\varphi_i = [\varphi_i(t_1) \qquad \varphi_i(t_m)]^T$$

$$b = [b_1, \dots, b_m]^T$$

$$\langle b, \varphi \rangle = \sum_{k=1}^{m} b(t_k)\,\varphi(t_k)$$

by the projection theorem $\langle b - f^*, \varphi_i \rangle = 0 \quad i = 1, .., n$

$$\begin{bmatrix} \sum_{j=1}^{n} c_j \varphi_j(t_1) \\ \vdots \\ \sum_{j=1}^{n} c_j \varphi_j(t_n) \end{bmatrix}, \begin{bmatrix} \varphi_i(t_1) \\ \\ \varphi_i(t_m) \end{bmatrix} \Big\rangle = \Big\langle \begin{bmatrix} b_1 \\ \\ b_n \end{bmatrix}, \begin{bmatrix} \varphi_i(t_1) \\ \\ \varphi_i(t_m) \end{bmatrix} \Big\rangle, \quad i = \overline{1, n}$$

$$\sum_{k=1}^{m} \Big(\sum_{j=1}^{n} c_j \varphi_j(t_k)\, \varphi_i(t_k) \Big) = \sum_{k=1}^{m} b_k \varphi_i(t_k) \iff A^T A c = A^T b$$

is

$$A^T A c = A^T b$$

$$[A^T A]_{ij} = \sum_{k=1}^{m} \varphi_i(t_k)\varphi_j(t_k) \quad \text{where } A = \begin{bmatrix} \varphi_1(t_1) & \varphi_n(t_1) \\ & \\ \varphi_1(t_m) & \varphi_n(t_m) \end{bmatrix}$$

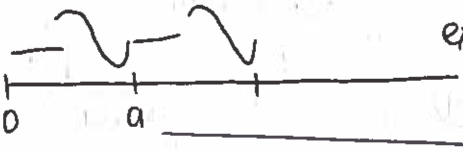
The residual $r = b - Ac$

* Complex exponentials, trigonometric polynomials, discrete Fourier Analysis

① * Complex exponentials

* Consider $f:[0,a]\to\mathbb{C}$: we can do periodic expansion to extend from $[0,a]\xrightarrow{to}\mathbb{R}$

$\langle f,g\rangle = \int_0^a f(x)\overline{g(x)}\,dx$



ex: $e^{i\varphi} = \cos\varphi + i\sin\varphi$ is a 2π periodic

* Define $e_\ell(t) = e^{i2\pi\ell\frac{t}{a}}$, $\ell \equiv$ frequency

* Properties

- $e_\ell(t+a) = e_\ell(t)$ ($e_\ell(t)$ is a periodic function ~~of~~ t with periodic a).

$$e_\ell(t+a) = e^{i2\pi\ell\frac{(t+a)}{a}} = e^{i2\pi\ell\frac{t}{a} + i2\pi\ell} = e^{i2\pi\ell\frac{t}{a}}\underbrace{e^{i2\pi\ell}}_{=1} = e^{i2\pi\ell\frac{t}{a}} = e_\ell(t)$$

- $\langle e_\ell, e_m\rangle = \begin{cases} 0 & , \ell\neq m \\ a & , \ell = m\end{cases}$ (complex exponentials is orthogonal)

$$\langle e_\ell, e_m\rangle = \int_0^a e_\ell(t)\overline{e_m(t)}\,dt = \int_0^a e^{i2\pi(\ell-m)\frac{t}{a}}\,dt = \frac{a}{i\,2\pi(\ell-m)}\left[e^{i2\pi(\ell-m)\frac{t}{a}}\right]\Bigg|_{t=}^{t=}$$

$$= \frac{a}{i(2\pi)(\ell-m)}\left[e^{i2\pi(\ell-m)} - 1\right] = \begin{cases} 0 & \ell\neq m \\ a & \ell = m\end{cases}$$

Remind $e^z = 1$ when $\frac{z}{\ell 2\pi}\in\mathbb{Z}$

* Define $e(x) = e^{i2\pi x}$, $x\in\mathbb{R}$

- $e(1) = 1$
- $e(x+1) = e(x)$ ($e(x)$ is a periodic-1 function).
- $e(n) = 1$, when $n\in\mathbb{Z}$
- Let $q>0$, $q\in\mathbb{Z}^+$

Then for any integer n, $\sum_{k=1}^{q} e\left(\frac{nk}{q}\right) = \begin{cases} q & q\mid n \Leftrightarrow \frac{n}{q} \text{ is an integer} \\ 0 & \text{otherwise}\end{cases}$

- Proof

⊕ When $q\mid n$, we have $\frac{n}{q}\in\mathbb{Z} \Rightarrow \frac{nk}{q}\in\mathbb{Z} \Rightarrow \underbrace{e^{i2\pi\underbrace{\frac{nk}{q}}_{\in\mathbb{Z}}}}_{e(\frac{nk}{q})} = 1 \Rightarrow \sum_{k=1}^{q} e\left(\frac{nk}{q}\right) =$

⊕ When $q\nmid n$

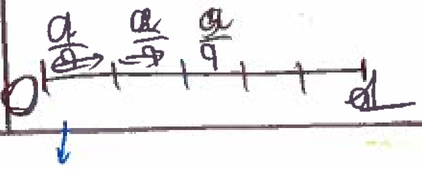
the term $e\left(\frac{nk}{q}\right)$ forms a geometric sequence with $e\left(\frac{n}{q}\right)\neq 1$.

Then $\sum_{\ell=1}^{n} e^{\frac{n\ell}{q}} = e^{\#}\sum_{\ell=1}^{n} e^{\ell} = e^{\frac{n}{q}}\,\frac{e^{\frac{nq}{q}}-1}{e(\frac{n}{q})-1} = 0$

Remind $1+\lambda+\cdots+\lambda^{q-1} = \frac{1-\lambda^q}{1-\lambda}$

Trigonometric polynomials is when we consider function with periodic = 1

Define $T(x) = \sum_{n=-N}^{N} t_n e(nx)$



sample with equal distance points

Theorem:

Suppose ($q \in \mathbb{Z}^+$, $q < N$), then $\frac{1}{q} \sum_{a=1}^{q} T\left(\frac{a}{q}\right) = \begin{cases} \sum_{\substack{-N \le n \le N \\ q \mid n}} t_n & \text{(when } n \text{ is a number that } q \mid n\text{).} \\ 0 & \text{otherwise .} \end{cases}$

If ($q > N$), then $\frac{1}{q} \sum_{a=1}^{q} T\left(\frac{a}{q}\right) = t_0$

Proof:

$$\frac{1}{q}\sum_{a=1}^{q} T\left(\frac{a}{q}\right) = \frac{1}{q}\sum_{a=1}^{q}\sum_{n=-N}^{N} t_n e\left(n\frac{a}{q}\right) = \frac{1}{q}\sum_{n=-N}^{N} t_n \underbrace{\sum_{a=1}^{q} e\left(\frac{na}{q}\right)}_{= \begin{cases} q & \text{when } q \mid n \\ 0 & \text{other wise} \end{cases}} = \begin{cases} \sum_{\substack{n=-N \\ q \mid n}}^{N} t_n \\ 0 & \text{others.} \end{cases}$$

Complex exponentials when considering 1-periodic functi

When consider 1-periodic function .

Then $e(nx) = e^{i2\pi nx} = e^{i2\pi n \frac{x}{1}} = e_n(x) \Rightarrow e(nx) = e_n(x)$

$\{e_n\}$ is an orthoramal system $\langle e_n, e_m\rangle = \begin{cases} 0 & \text{when } n \neq m \\ 1 & \text{when } n = m \end{cases}$

When f is a 1-periodic function,

define $\hat{f}_n = \langle f, e_n\rangle = \int_0^1 f(x)\overline{e_n(x)}\,dx = \int_0^1 f(x)\overline{e(nx)}\,dx = \int_0^1 f(x)e^{-i2\pi nx}\,dx$

Fourier coefficient .

then $F(f(x))$

Fourier approximation of f

* Then the next 2 or 3 pages :

Theorem :

- When ~~$N < q < 2N$~~ $q < 2N$, $-N \le \ell \le N$: $\frac{1}{q}\sum_{a=1}^{q} e\left(\frac{n\ell}{q}\right)T\left(\frac{a}{q}\right) = \sum_{-N \le n \le N} t_n$

  $q > 2N$, $-N \le \ell \le N$: $\frac{1}{q}\sum_{a=1}^{q} e\left(\frac{n\ell}{q}\right)T\left(\frac{a}{q}\right) = t_\ell$

② * Trigonometric polynomials

$$T(x) = \sum_{n=-N}^{N} t_n\, e(nx) \quad - (2N+1 \text{ coefficients})$$

* Theorem

Suppose $q \in \mathbb{Z}$, $q > 0$ then

$$\frac{1}{q}\sum_{a=1}^{q} T\left(\frac{a}{q}\right) = \sum_{\substack{-N \le n \le N \\ q|n}} t_n$$

sample with equal distance points

obtain some coefficient

Remind $\sum e\left(\frac{n\ell}{q}\right) = \begin{cases} q & q|n \\ 0 \end{cases}$

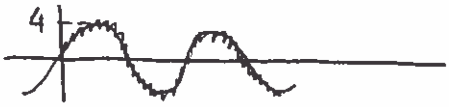
* Proof:

$$\frac{1}{q}\sum_{n=-N}^{N} t_n \underbrace{\sum_{a=1}^{q} e\left(\frac{na}{q}\right)}_{=q \text{ if } q|n} = \sum_{\substack{-N \le n \le N \\ q|n}} t_n$$

* Corollary: If $q > N$, then $\frac{1}{q}\sum_{a=1}^{q} T\left(\frac{a}{q}\right) = t_0$

sampling points

* Example:



$f(t) = 4\sin t + 0.3\sin(683t)$

Store 4 0.3

frequency 1 683

- Complex exponentials

$e(x) = e^{i2\pi x}$ — 1 periodic function < no shift, no localization

$e_n(x) = e(nx) = e^{i2\pi n x}$ $\quad n \in \mathbb{Z}$ $\quad \langle e_n, e_m \rangle = \begin{cases} 0 & n \neq m \\ 1 & n = m \end{cases}$

frequency

$$\langle f, g \rangle = \int_0^1 f(x)\, \overline{g(x)}\, dx$$

Any linear combination of e_n is a 1-periodic function

$$\underset{z}{F(f(x))} = \sum_{n=-\infty}^{\infty} \underset{\in \mathbb{C}}{\hat{f}(n)}\, e_n(x) = \lim_{N \to +\infty} \sum_{n=-N}^{N} \hat{f}(n)\, e_n(x)$$

$$\hat{f}(n) = \langle f, e_n \rangle = \int_0^1 f(x)\, e_n(-nx)\, dx = \int_0^1 f(x)\, e^{-i2\pi n x}\, dx$$

Fourier coefficient of f

Remark
If we take

$e'_n(x) = e^{i2\pi \frac{nx}{T}}$ is T- periodic, such $e'_n(x)$ is orthogonal in the sense of $\langle f, g\rangle = \int_0^T f(x)\overline{g(x)}\,dx$

that $\langle e'_n, e'_m\rangle = \begin{cases} 0 & n \neq m \\ T & n = m \end{cases}$

The finite version of the Fourier series of f is its partial sum

$$T(x) = \sum_{n=-N}^{N} t_n e_n(x)$$

ch function (1- periodic) is called trigonometric polynomial

If $t_n = \hat{f}(n)$ then $T(x)$ is a partial sum of F_f

and $\|f - T\| \leq \|f - g\| \quad \forall g \in \text{Span}\{e_{-N}, \dots, e_0, \dots, e_N\}$ ⇒ bad

chang f locally
then $\langle f, e_n\rangle$ changed
⇒ all coefficients have to be ~~de~~ compute

Suppose $f : \mathbb{R} \longrightarrow \mathbb{C}$ non periodic

nalysis is done

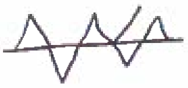
$$\hat{f}(\xi) = \int_{-\infty}^{+\infty} f(x)\, e^{-i2\pi\xi x}\, dx$$: The Fourier transform of f

frequency domain.

Synthesis is $f(x) = \int_{-\infty}^{+\infty} \hat{f}(\xi)\, e^{i2\pi\xi x}\, d\xi$ inverse of the transform

in stead of Σ

Since $e_n(x)$ are never 0, their support is all $\mathbb{R}$
hence, integrated against f all values of f are taken into account for each $\hat{f}(n)$
A local change in f affects all Fourier coefficients.
Complex exponential are well localized in frequency
badly localized in time



DFT → discrete Fourier transform (will study)

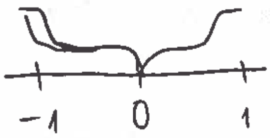
* We want a basic (in stead of $e_n(x)$) which is better localized in tim e
better localized in frequency.

$\psi \in L^2(\mathbb{R})$ wavelet

$$\psi_{j,k}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$$ $j,j,k \in \mathbb{Z}$

frequency shift localize

smooth function



even cosin xe

odd sin xe

$$f(x) = \sum_{j,k\in\mathbb{Z}} \left[\langle f, \psi_{j,k} \rangle \psi_{j,k}(x) \right]$$

$$T(x) = \sum_{n=-N}^{N} t_n \underbrace{e(nx)}_{\text{trigonometric function}}$$

* t polynomial of degree $\leq N$

why polynomial? only a finite number of roles. 3

$2n+1$ coefficients.

This function has at most $2N$ roots in $[0,1]$ ✓ (only count roots inside of the period)

- $T(x) = t_{-N}\, e(-Nx) + \cdots + t_N (e(Nx)$
  $= e(-Nx)\left[t_{-N} + t_{-N+1} e((-N+1)x) + \cdots + t_N e(2Nx) \right]$

Let $p(z) = \sum_{n=0}^{2N} t_{n-N} z^n$

then

$$T(x) = e(-Nx)\, P(e(x))$$

P has at most $2N$ roots on the unit circle $|z| = 1$

~~even number since z root then $\bar{z}$ root~~

- Sampling and interpolation of T

Sampling and interpolation of T

Theorem

ippose that $q \in \mathbb{Z}$, $q>0$

ippose $\boxed{q>2N}$, q has to be large enough

f $-N \le k \le N$

en $$\frac{1}{q}\sum_{a=1}^{q} e\left(-\frac{ka}{q}\right) T\left(\frac{a}{q}\right) = t_k$$

Lemma

$1>0$, $$\sum_{k=1}^{q} e\left(\frac{nk}{q}\right) = \begin{cases} q & q \mid n \\ 0 & \text{otherwise} \end{cases}$$

had theorem 1 :

$$\frac{1}{q}\sum_{a=1}^{q} T\left(\frac{a}{q}\right) = \sum_{-N\le n\le N} t_n$$

$>N$ then $$\frac{1}{q}\sum_{a=1}^{q} T\left(\frac{a}{q}\right) = t_0$$

Theorem 2 $\boxed{q<2N}$

$$\frac{1}{q}\sum_{a=1}^{q} e\left(-\frac{ka}{q}\right) T\left(\frac{a}{q}\right) \sum_{-N\le n\le N} t_n$$

$$T = \sum_{-N}^{N} t_n \left(e(nx)\right)$$

Proof of theorem 2

$$\frac{1}{q}\sum_{n=-N}^{N} t_n \underbrace{\sum_{a=1}^{q} e\left(\frac{(n-k)a}{q}\right)}_{= q \text{ when } q \mid (n-k)}$$

if $q \mid (n-k)$ sum is q

$$= \frac{1}{q} q \sum_{\substack{n=-N \\ n\equiv k \pmod q}}^{N} t_n$$

* Complex exponentials, trigonometric polynomial, Fourier series, DFT (Discrete F transf

*Consider 1 periodic functions:

$$f : \mathbb{R} \to \mathbb{C}$$

$$x \mapsto f(x) = f(x+n) \;, \forall n \in \mathbb{Z} \qquad = f(x+1)$$

* Complex exponentials

$e_n(x) = e(nx)$ $\qquad e(x) =$

$e_n(x) = e^{i2\pi nx}$

$\{e_n\}_{n\in\mathbb{Z}}$ this is an orthogonal system

* $\hat{f}(n) = \langle f, e_n \rangle = \int_0^1 f(x)\overline{e_n(x)}\,dx = \int_0^1 f(x)\, e^{-i2\pi nx}\,dx$

$$f(x) = \sum_{n=-\infty}^{+\infty} \hat{f}(n)\, e_n(x) = \lim_{N\to\infty} \sum_{n=-N}^{N} \hat{f}(n)\, e_n(x)$$

$T(x) = \sum_{n=-N}^{N} t_n\, e_n(x)$: trigonometric polynomial ← this is an 1 periodic function.

* Example of trigonometric polynomial

- Dirichlet kernel $D_N(x) = \sum_{n=-N}^{N} e_n(x) = 1 + 2\sum_{n=1}^{N} \cos(2\pi nx)$ $\quad \{ D_N(0) = 1 + 2N$

$\Big(e_n(x) + e_{-n}(x) = e^{i2\pi x} + e^{-i2\pi x} = \cos(2\pi nx) + i\sin(2\pi nx) + \cos(2\pi nx) - i\sin(2$

$= 2\cos(2\pi nx)$

$D_N(0) = 1 + 2N$

- We can express D_N in closed form $\qquad e^{u} - e^{-u} = 2i\sin(2\pi u)$

$$D_N(x) = \begin{cases} \dfrac{\sin((2N+1)\pi x)}{\sin(\pi x)} & x \notin \mathbb{Z} \\ 2N+1 & x \in \mathbb{Z} \end{cases}$$

$$D_N(x) = e(-Nx) * \sum_{n=0}^{2N} e(nx) = e(-Nx)\, \frac{e((2N+1)x)}{e(x) - 1} =$$

$$= \frac{\cancel{e(-Nx)}\; \cancel{e((N+\frac{1}{2})x)}\, \left[e((N+\frac{1}{2})x) - e(-(N+\frac{1}{2})x) \right]}{\cancel{e(\frac{x}{2})}\, \left[e(\frac{x}{2}) - e(-\frac{x}{2}) \right]} = \text{what we need}$$

- The zeros of D_N : $\frac{1}{2N+1}, \frac{2}{2N+1}, \ldots, \frac{2N}{2N+1}$

Lemma:

$$\sum_{k=0}^{q-1} e\left(\frac{nk}{q}\right) = \begin{cases} q & q \mid n \\ 0 & \text{otherwise} \end{cases}$$

- If $q > 2N$, $-N \leq k \leq N$ $\quad \frac{1}{q}\sum_{a=0}^{q-1}\left(-\frac{ka}{q}\right)T\left(\frac{a}{q}\right) = t_k$

e can retrieve T from its samples (equal distance samples)

$$T(x) = \sum_{n=-N}^{N} \underbrace{\left(\frac{1}{q}\sum_{a=0}^{q-1} e\left(-\frac{an}{q}\right)T\left(\frac{a}{q}\right)\right)}_{t_n} e_n(x)$$

$$= \frac{1}{q}\sum_{a=0}^{q-1} T\left(\frac{a}{q}\right)\sum_{n=-N}^{N} e\left(n\left(x-\frac{a}{q}\right)\right) = \frac{1}{q}\sum_{a=0}^{q-1} T\left(\frac{a}{q}\right) D_N\left(x-\frac{a}{q}\right)$$

? what are kernels?

Let $q = 2N+1$

Let $c(1), \dots, c(2N+1)$ be given

We can define $U(x)$ a trigonometric polynomial of degree N such that

$$U\left(\frac{a}{2N+1}\right) = c(a) \qquad a = 1, 2, \dots, 2N+1$$

$$u(x) = \frac{1}{2N+1}\sum_{a=0}^{2N} c(a)\, D_N\left(x - \frac{a}{2N+1}\right)$$

$$U\left(\frac{b}{2N+1}\right) = \frac{1}{2N+1}\sum_{a=0}^{2N} c(a)\, D_N\left(\frac{b-a}{2N+1}\right)$$

$D_N(0) = 2N+1.$

* DFT Discrete Fourier Transformation

* Discrete ~~Fourier~~ complex exponential are going to be arithmetic functions.

- $f : \mathbb{Z} \longrightarrow \mathbb{C}$

, which are $\boxed{N\text{- periodic}}$ $N \in \mathbb{Z}$, $N > 0$

$f(\ell) = f(\ell + mN)$ $m \in \mathbb{Z}$ finite discrete with N values

* when a function is periodic N

$e_n(\ell) = e\left(\frac{n\ell}{N}\right)$ arithmetic function $\boxed{N\text{- periodic}}$

These discrete complex exponential are orthogonal

$$\langle e_n, e_m \rangle = \sum_{\ell=0}^{N-1} e_n(\ell)\, \overline{e_m(\ell)}$$

- We define $\hat{f}(n) = \langle f, e_n \rangle = \sum_{\ell=0}^{N-1} f(\ell)\, e\left(-\frac{n\ell}{N}\right)$

such arithmetic function $\hat{f}$ is also N periodic

and is called DFT of f

- $f(\ell) = \frac{1}{N} \sum_{n=0}^{N-1} \hat{f}(n)\, e_n(\ell)$

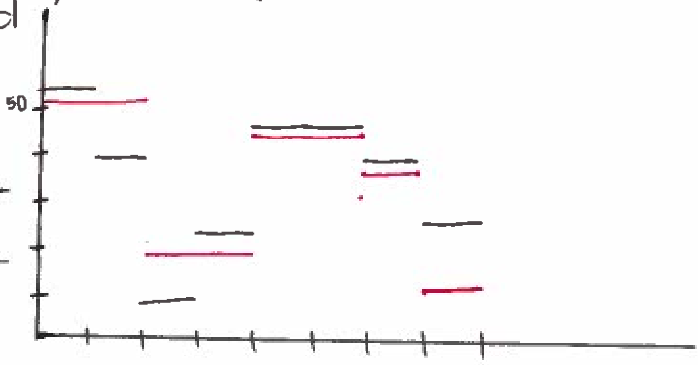
* Example: Having a signal (Multi resolution)

56 40 8 24 48 48 40 16 ← even number of them

$s = \frac{a+b}{2}$ $d = \frac{a-b}{2} = a - s$ $\left.\begin{matrix} a = s + d \\ b = s - d \end{matrix}\right)$ backward formula

forward formulas

56	40	8	24	48	48	40	16
48	16	$\overset{s}{48}$	28 |	8	-8	$\overset{d}{0}$	12
32	$\overset{s}{38}$ |	16	$\overset{d}{10}$ |	8	-8	$\overset{d}{0}$	12
$\overset{s}{35}$ |	$\overset{d}{-3}$ |	16	$\overset{d}{10}$ |	8	-8	$\overset{d}{0}$	12



Reconstruction with threadhold 9

51	51	19	19	45	45	37	13
51	19	45	25 |	0	0	0	12
35	35 |	16	10 |	0	0	0	12
35 |	0 |	16	10 |	0	0	0	12

* Some basic theorems about solving equation $f(x) = 0$ involving $f : \mathbb{R} \longrightarrow \mathbb{R}$

Break: read C3

Sol of nonlinear equations $f(x) = 0$

Ex: $x^2 - a = 0$

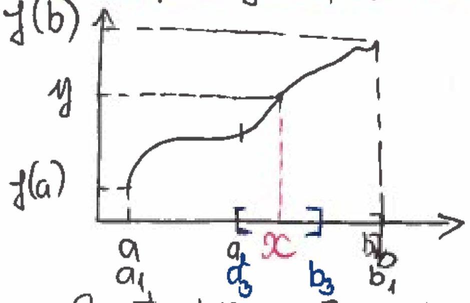
$a_n x^n + \cdots + a_1 x + a_0 = 0$

(real domains only)

* Theorem 1 (Intermidiate value theorem)

Let $f: \mathbb{R} \longrightarrow \mathbb{R}$ continuous on a domain containing $[a, b]$ with $f(a) < f(b)$

Then for any y, $f(a) < y < f(b)$, there exists $x_0 \in (a, b)$ such that $f(x_0) = y$



Proof

* Divide and conquer

Construct a sequence of interval

$[a_1, b_1] = [a, b]$, $[a_2, b_2], \ldots$

such that $f(a_k) < y < f(b_k)$

- Construct $[a_2, b_2]$: Let $[a_2, b_2]$ is one of the half of $[a_1, b_1]$.

  If $f\left(\frac{a_1+b_1}{2}\right) < y$ then $[a_2, b_2] = \left[\frac{a_1+b_1}{2}, b_1\right]$

  $f\left(\frac{a_1+b_2}{2} > y\right)$ then $[a_2, b_2] = \left[a_1, \frac{a_1+b_1}{2}\right]$

  $f\left(\frac{a_1+b_1}{2}\right) = y$ then $x = \frac{a_1+b_1}{2}$

- Then by induction $b_k - a_k = 2^{1-k}(b_1 - a_1)$

Nested intervals with diameters $\longrightarrow 0$, $\lim_{k \to \infty} a_k = x_0$

- Remark: no more than 52 steps in matlab

thigo

quadrature

projection onto space

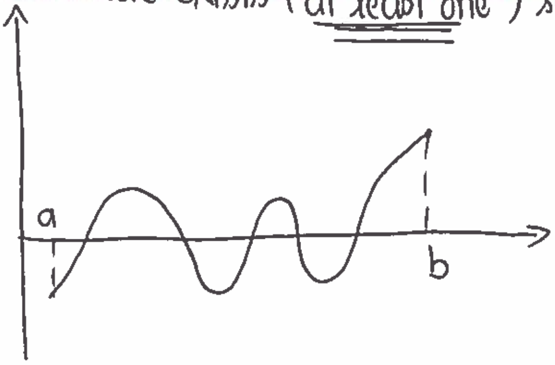
Thursday Nov 2

* Theorem 2: (Bisection – Interval Halving) Method

Let $f: [a, b] \longrightarrow \mathbb{R}$ continuous

such that $f(a) f(b) \leq 0$

then there exists (at least one) solution ξ s.t. $f(\xi) = 0$



- Proof: • $f(a) = 0$ or $f(b) = 0 \Rightarrow$ a or b are solutions

If $f(a) f(b) \neq 0$

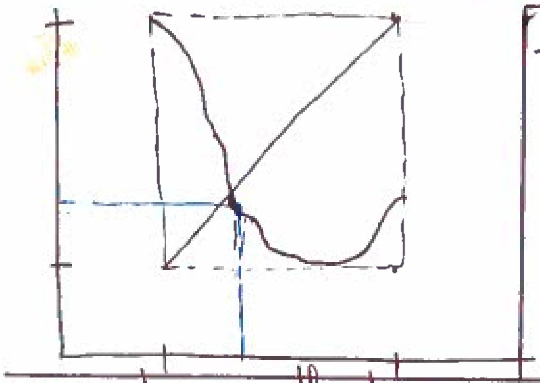
then 0 belongs to the interval having endpoints are $f(a)$ or $f(b)$

Then by intermidiate theorem $\Rightarrow \exists!\ \xi$ is a sol of $f(x)$

Theorem 3 (Browner's fixed point theorem).

Let $g:\underbrace{[a,b]\longrightarrow[a,b]}_{\text{same}}$ be continuous

ien there exists $\xi\in[a,b]$ such that $\overset{\text{same}}{g(\xi)=\xi}$.



*Proof.

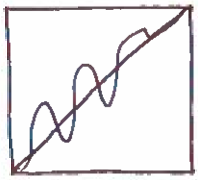
Let $f(x)=x-g(x)$

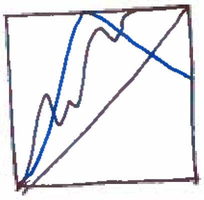
Then $f(a)a-g(a)\leq 0$

$f(b)=b-g(b)\geq 0$

Then $f(a)\,f(b)\leq 0$

By theorem 2 ⇒ sol exists.





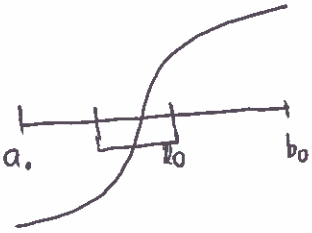
Bisection method:

- $f(x)$ 0 $f(a)\,f(b)<0$

Bisection $[a,b]=[a_0,b_0]$ $x_0=a_0+\frac{1}{2}[b_0-a_0]$

$$(a_1,b_1)=\begin{cases}(x_0,b_0) & f(x_0)f(b_0)<0\\ (a_0,x_0) & f(x_0)f(a_0)<0\end{cases}$$

If $|f(x_0)f(a_0)>0$ then use induction $f(a_0)\,f(b_0)<0$



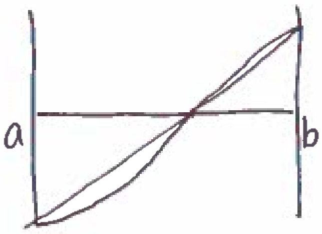
- Let's view the bisection method as follow

Take an affine function $L(x)$ such that

$L(a_k)=-1$ $L(b_k)=1$

$L(x)=-1+(x-a_k)\frac{2}{b_k-a_k}$

$L(x)=0$ given $x_1=\frac{a_k+b_k}{2}$



Improved version of Bisection method (False position)

$L(a_k)=f(a_k)$ $L(b_k)=f(b_k)$

$$L(x)=f(a_k)+(x-a_k)\frac{f(b_k)-f(a_k)}{b_k-a_k}$$

$L(x_k)=0$

$$x_k=a_k-\frac{f(a_k)}{f(b_k)-f(a_k)}(b_k-a_k)$$

* $f(x)=0$ $\underbrace{L(f)(x)}_{H(f)}$: is an interpolation of f

$H(x) = f(x_n) + f'(x_n)(x - x_n)$ $f(x_n) = H(x_n)$ $f'(x_n) = H^{(1)}(x_n)$

$H(x) = 0$ $0 = f(x_n) + f'(x_n)(x - x_n)$

$$-f(x_n) = f'(x_n)(x - x_n)$$

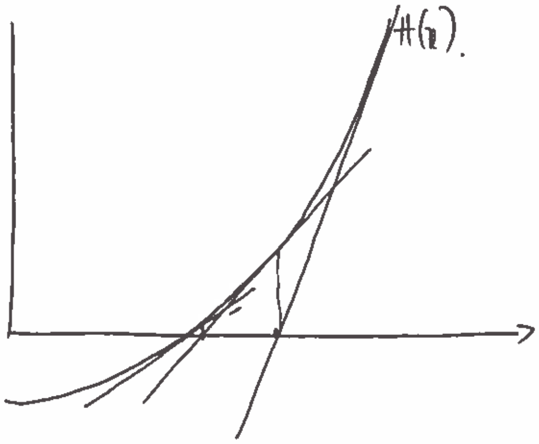
$$\frac{-f(x_n)}{f'(x_n)} = x - x_n$$

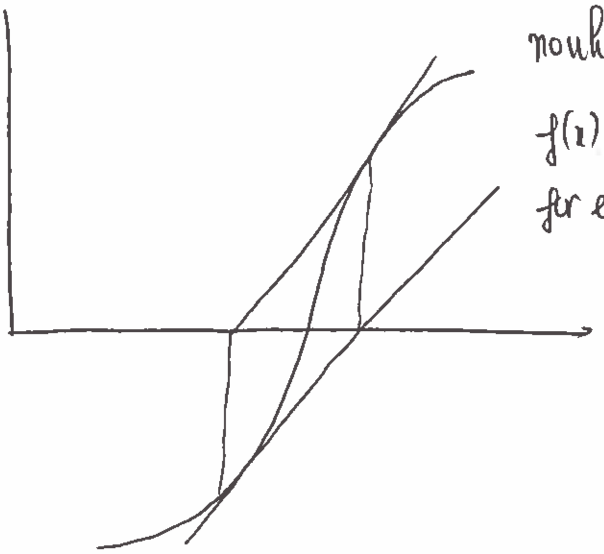
Newton method (interpolation)

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$$x_1 = a_1 - \frac{f(a_1)}{\frac{f(b_1)-f(a_1)}{b_1 - a_1}}$$

* Newton method ? (cond for applying Newton method ?)





nowhere

$f(x) = \sqrt{x}$

for example

rool out the functions that $f'(x) \approx f(x)$

* L Kantorovich Global convergence theorem .

Exam 2 : (5 problem)

Gaussian quadrature exactness
error term
notes & weights.

est approximation in L^2 is the projection. (HW5)

igonometric polynomials and their interpolatory

- Λ

* Nonlinear equation $f(x)=0$

- Let $g(x) = x$, finding fixed points of a map $\qquad g(x) =$
- We can start at a general equation $f(x)= 0$ and convert it to a fixed point probl

$f(x)=0$

$\alpha f(x)=0$

$x = \underbrace{x+\alpha f(x)}$,

$x = g(x)$

$g:[a,b]\to[a,b]$

g continuous

* g must satisfies Brower theorem assumption.

We have an algorithm for solving $x = g(x)$, $x_{k+1} = g(x_k)$

such sequence $(x_{k+1} = g(x_k))$ converges by continuity to ξ such that $\xi = g(\xi)$

- Sufficient condition for convergence of simple iteration:

There exists L, $0<L<1$

$$|g(x)-g(y)| < L|x-y| \text{ for all } x,y \in [a,b]$$

* Theorem (contraction mapping theorem)

$g:[a,b] \longrightarrow [a,b]$ is continuous

Then g has a unique fixed point $\xi \in [a,b]$

Moreover $\{x_k\}$ defined by simple iteration converges to ξ for any starting point x $\in$

- Proof: • Proof the uniqueness.

Suppose $\xi, \eta \in [a,b]$ are both fixed points

$$|\xi-\eta| = |g(\xi)-g(\eta)| \le L|\xi-\eta|$$

$\Rightarrow \underbrace{(1-L)}_{>0}|\xi-\eta| \le 0$ then $\xi=\eta \Rightarrow$ uniqueness.

Proof the convergence

- Let $x_0 \in [a,b]$, $x_{k+1} = g(x_k)$, show that $x_k \longrightarrow \xi$.

$$|x_k-\xi| = |g(x_{k-1})-g(\xi)| \le L|x_{k-1}-\xi|$$

then $|x_k-\xi| \le \underset{\substack{\uparrow \\ 0<L<1}}{L^k}|x_0-\xi|$ so we have $x_k \to \xi$.

Theorem (Local contraction mapping theorem

~~If g is differentiable on (a,b)~~,

~~then there exists an L such that~~

Let $g:[a,b]\longrightarrow[a,b]$ continuous

g' is continuous in some neighbor of ξ

$|g'(\xi)|<1$

Then the sequence $\{x_k\}$ $x_{k+1}=g(x_k)$ converges to ξ, provided that x_0 is sufficiently close to ξ.

Proof:

Let g' is continuous in $[\xi-h, \xi+h]$

Since $|g'(\xi)|<1$, then by continuity, we can find a smaller interval $I_{\xi,\delta}$

$I_\delta=[\xi-\delta, \xi+\delta]$, $0\leq\delta\leq h$

such that

$|g'(x)|\leq L$ with $L<1$.

Take $L=\frac{1}{2}(1+|g'(\xi)|)$ and chose $\delta\leq h$ such that

$|g'(x)-g'(\xi)|\leq\frac{1}{2}(1-|g'(\xi)|)$ in I_δ

For $x\in I_\delta$, $|g'(x)|\leq|g'(x)-g'(\xi)|+|g'(\xi)|$

$$\leq\frac{1}{2}(1-g'(\xi))+|g'(\xi)|$$

$$=\frac{1}{2}(1+|g'(\xi)|)=L<1$$

$x_k\in I_\delta$, $x_{k+1}-\xi=g(x_k)-g(\xi)=g'(\eta_k)(x_k-\xi)$

$$|x_{k+1}-\xi|\leq L|x_k-\xi|\leq L\delta<\delta$$

; $x_0\in I_\delta \Rightarrow x_k\in I_\delta$

$$|x_k-\xi|\leq L^k|x_0-\xi|$$

- $f: \mathbb{R} \longrightarrow \mathbb{R}$, f is C^1 function.

$f(x) = 0$

f has to be differentiable ↓

$x_0 \in \mathbb{R}$

$$\begin{cases} H(x) = f(x_0) + f'(x_0)(x - x_0) \\ H(x_0) = f(x_0) \\ H'(x_0) = f'(x_0) \end{cases}$$

Hermiton interpolation

$f: \mathbb{R}^n \longrightarrow \mathbb{R}^n$

$x \in \mathbb{R}^n$

$x_0 \in \mathbb{R}^n$

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

$$\begin{cases} H(\underset{\sim}{x}) = f(\underset{\sim}{x_0}) + Df(x_0)(\underset{\sim}{x} - \underset{\sim}{x_0}) \\ | \quad = \quad | \quad + \ \square \ | \\ H(x_0) = f(x_0) \\ \frac{\partial i}{\partial x_i} H(x_0) = \frac{\partial i}{\partial x_j} f(x_0) \end{cases}$$

$$Df = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} \\ \\ \frac{\partial f_n}{\partial x_1} \end{bmatrix}$$

- Let $H(x) = 0$

$\Leftrightarrow f(x_0) + f'(x_0)(x - x_0) = 0$

$f'(x_0)(x - x_0) = -f(x_0)$

$(x - x_0) = -\dfrac{f(x_0)}{f'(x_0)} = -[f'(x_0)]^{-1} f(x_0)$

write this way in case we have a matrix ↓ then this is an inverse of a mat

$$\boxed{x_1 = x_0 - [f'(x_0)]^{-1} f(x_0)}$$

* Newton Raphson's theorem.

Suppose $f: \mathbb{R} \longrightarrow \mathbb{R}$ differentiable

$f(x^*) = 0$

Suppose that there exists 3 constants positive a, a_1, a_2 such that.

1 f is C^1 in $B_a(x^*)$

2) $f'(x)$ is invertible ($f'(x) \neq 0$) on $B_a(x^*)$

$|f'(x)^{-1}| \leq a_1$

3) $x \longmapsto f'(x)$ is Lipchitz on $B_a(x^*)$

$|f'(x) - f'(y)| \leq a_2 |x - y|$

Then for any $x_0 \in B_b(x^*)$, $b < \min\{a, \frac{2}{a_1 a_2}\}$

then the $x_{k+1} = x_k - f'(x_k)^{-1} f(x_k)$ (Newton method formula)

is well defined and converges quadratically to x^*

$$|x_k - x^*| \leq \frac{2}{a_1 a_2} \left(\frac{1}{2} a_1 a_2 |x - x^*| \right)^{2^k}$$

$f(x^*) = 0 \qquad x_\ell \in B_b(x^*)$

$$f'(x_\ell)(x_{\ell+1} - x^*) = \underbrace{f(x^*)}_{=0} - f(x_\ell) + f'(x_\ell)(x_\ell - x^*)$$

have to replace $x_{\ell+1}$ by the formula of Newton method then we have the formula

$$|x_{\ell+1} - x^*| \leq [|f'(x_\ell)|]^{-1} \underbrace{|f(x^*) - f(x_\ell) + f(x_\ell)(x_\ell - x^*)|}_{A}$$

$$= \left| \int_0^1 \left(f'(x_\ell) - f'(tx_\ell + (1-t)x^*)\right)(x_\ell - x^*)\,dt \right|$$

$$= \left| f'(x_\ell)(x_\ell - x^*) - \underbrace{\int_0^1 f'(\underbrace{tx_\ell + (1-t)x^*}_{g(t)}) \underbrace{(x_\ell - x^*)}_{g'(t)}\,dt} \right|$$

$$= \int_0^1 f'(g(t))\, g'(t)\,dt = [f(g(t))]\Big|_0^1 = f(g(1)) - f(g(0))$$

$$= f(x_\ell) - f(x^*)$$

$$= \left| f'(x_\ell)(x_\ell - x^*) - f(x_\ell) + f(x^*) \right|$$

so

$$|x_{\ell+1} - x^*| \leq \underbrace{[|f'(x_\ell)|]^{-1}}_{\leq a_1 \text{ by assumption}} \cdot \underbrace{\left| \int_0^1 [f'(x_\ell) - f(tx_\ell + (1-t)x^*)](x_\ell - x^*)\,dt \right|}$$

$$\leq |x_\ell - x^*| \underbrace{\int_0^1 |f'(x_\ell) - f'(tx_\ell + (1-t)x^*)|\,dt}_{\leq a_2 |x_\ell - tx_\ell + (1-t)x^*|}$$

* Exam $x^2 = 0$ run the Newton method.

$x^2 - b = 0$

*Newton's method for solving $f(x) = 0$

$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$

Global convergence theorem of L. Kantorovich

*Newton' method

• $x_{k+1} = x_k - [Df(x_k)]^{-1} f(x_k) = x_k + h_k$ where $h_k = -[Df(x_k)]^{-1} f(x_k$

• $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$

$$f = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_n(x_1, \dots, x_n) \end{bmatrix}$$

*Theorem

• Let $x_0 \in U$, $U \subset \mathbb{R}^n$, $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$

$Df(x_0)$ is invertible $\quad$ f invertible.

Define $h_0 = -[Df(x_0)]^{-1} f(x_0)$ $\quad$ (1)

① $x_1 = x_0 + h_0$ $\quad$ $U_1 = B_{|h_0|}(x_1)$ $\quad$ $|x| = \sqrt{x_i^2}$

If $\overline{U}_1 \subset U$ and if the derivative $Df(x)$ satisfies Lipchitz condition

② $|Df(y_1) - Df(y_2)| \leq M |y_1 - y_2|$ $\quad$ $y_1, y_2 \in \overline{U}_1$ $\quad$ (2)

• Kant equality holds

all three don't need to small to land

③ $\underbrace{|f(x_0)| |[Df(x_0)]^{-1}|^2 M}_{} \leq \frac{1}{2}$

Then the equation $f(x) = 0$ has a unique solution in $\overline{U}_1$ and the Newton's method $x_{k+1} = x_k + h_k$ converges to this solution.

* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ $\quad$ $|f(x_0)|$ has units r

units are u $\quad$ units are r

~~$[Df(x_0)]^{-1}$~~

$|f(x_0)|$ has unit r

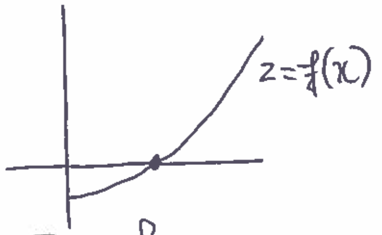
$[Df(x_0)]$ has units $\frac{r}{u}$

$[Df(x_0)]^{-1}$ " " $\frac{u}{r}$

$\frac{f(x+h) - f(x)}{h}$ $\quad$ $f'(x)$ has units $\frac{r}{u}$

$|[Df(x_0)]^{-1}|^2$ $\quad$ $\frac{u^2}{r^2}$

M has unit $\frac{\frac{r}{u}}{u} = \frac{r}{u^2}$



Example: want to solve to find $\begin{pmatrix}x\\y\end{pmatrix}$ so that .

? namoga matix .

$$f\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = \begin{bmatrix}x^2-y-2\\y^2-x-6\end{bmatrix} \quad \surd \quad f\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = \begin{pmatrix}0\\0\end{pmatrix}$$

art at $x_0 = \begin{bmatrix}2\\3\end{bmatrix}$ $\quad Df\left(\begin{bmatrix}x\\y\end{bmatrix}\right) = \begin{bmatrix}2x & -1\\-1 & 2y\end{bmatrix}$

$$Df\left(\begin{bmatrix}x_1\\y_1\end{bmatrix}\right) - Df\left(\begin{bmatrix}x_2\\y_2\end{bmatrix}\right) = \begin{bmatrix}2(x_1-x_2) & 0\\0 & 2(y_1-y_2)\end{bmatrix}$$

$$\left|Df\left(\begin{bmatrix}x_1\\y_1\end{bmatrix}\right) - Df\left(\begin{bmatrix}x_2\\y_2\end{bmatrix}\right)\right| = \sqrt{4(x_1-x_2)^2+4(y_1-y_2)^2} \ominus 2\left|\begin{bmatrix}x_1\\y_1\end{bmatrix}-\begin{bmatrix}x_2\\y_2\end{bmatrix}\right| \qquad M=2$$

$$_0 = -[Df(x_0)]^{-1} f(x_0) = -\frac{1}{23}\begin{pmatrix}6 & 1\\1 & 4\end{pmatrix}\begin{bmatrix}-1\\1\end{bmatrix} =$$

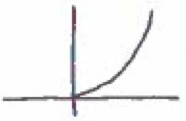
$$Df(x_0) = \begin{bmatrix}4 & -1\\-1 & 6\end{bmatrix} \quad [Df(x_0)]^{-1} = \frac{1}{23}\begin{pmatrix}6 & 1\\1 & 4\end{pmatrix} \Big)$$

$$|f(x_0)| \quad |[Df(x_0)]^{-1}|^2 \quad M$$

$$\sqrt{2} \quad \frac{54}{23^2} \quad 2 = \sqrt{2}\,\frac{108}{529} = 0.2888 < \frac{1}{2}$$

$f(x) = x^2$

$x_0 = 1$



$$x_1 = x_0 - [f'(x_0)]^{-1} f(x_0) = x_0 - \frac{1}{2x_0}\,x_0 = 1 - \frac{1}{2} = \frac{1}{2}$$

* Proof. Hypotheses (1), (2) and (3) are satisfied.

We must prove 4 statements

① $Df(x_{i+1})$ is invertible $\quad h_{i+1} = -[Df(x_i)]^{-1} f(x_i)$

② $|h_{i+1}| \leq \frac{1}{2}|h_i|$

③ $|f(x_{i+1})|\,|[Df(x_{i+1})^{-1}|^2 \leq |f(x_i)|\,|Df(x_i)^{-1}|^2$

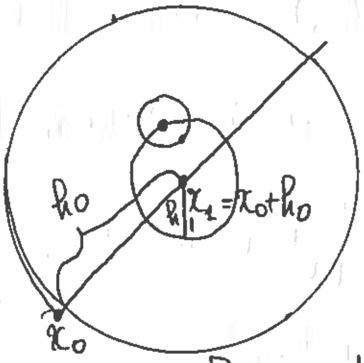
④ $|f(x_{i+1})| \leq \frac{M}{2}|h_i|^2$

$$x_i = x_0 + \underbrace{\sum_{k=0}^{i} h_k}_{\text{if this series converges absolutely} \Rightarrow \text{it converges.}}$$

From absolute convergence of a series follows convergence ②

if $\sum_{k=0}^{\infty} |h_k| \quad \Rightarrow x_0 + \sum_{k=0}^{\infty} h_k = x^*$

$|h_i| \leq \frac{|h_0|}{2^i}$



$x_2 = x_1 + h_1 \qquad |h_1| \leq \frac{1}{2}|h_0|$

- From ④,

$|f(x_{i+1})| \leq \frac{M}{2}|h_i|^2 \leq \frac{M}{2^{i+1}}|h_0|^2$

hence $f(x_i) \longrightarrow 0$ so $f(x^*) = 0$

We need 3 lemas:

* Lemma 1 (Taylor's) ~~formula~~)

$\left\{\begin{array}{l} \text{If } f \text{ is differentiable} \\ \text{If } |Df(x) - Df(y)| \leq M(x-y) \qquad x, y \in \overline{U} \end{array}\right\}$

- Then for $x, y \in \mathcal{U}$

$$|f(x) - f(y) - Df(x)(y-x)| \leq \frac{M}{2}|y-x|^2$$

- Let $h = y - x \qquad g(t) = f(x + th)$

$f(x+h) - f(x) = g(1) - g(0) = \int_0^1 g'(t)\,dt$

$g'(t) = Df(x+th)h = Df(x)h = Df(x)h + \left(Df(x+th)h - Df(x)h\right) dt$

$$(x+h)-f(x) = Df(x)h + \int_0^1 \left(Df(x+th)h - Df(x)h\right) dt$$

$$f(x+h)-f(x)-Df(x)h| = \left|\int_0^1 [Df(x+th)h - Df(x)h]\, dt\right| \leq \int_0^1 M|h|^2 t\, dt$$

$$= M|h|^2 \int_0^1 t\, dt = \frac{M}{2}|h|^2$$

Lemma 2:

$$Df(x_{i+1})]^{-1}| \leqslant 2|[Df(x_i)]^{-1}|$$

$$f(x_i)]^{-1}[Df(x_{i+1})] \approx I$$

$Df(y)$ doesn't vary too much.

$$ut\ |A| = |I - [Df(x_i)]^{-1}[Df(x_{i+1})]| =$$

$$= |[Df(x_i)]^{-1} Df(x_i) - [Df(x_i)]^{-1}[Df(x_{i+1})]|$$

$|AB| \leq |A||B|$.

$$= |[Df(x_i)]^{-1} [Df(x_i) - Df(x_{i+1})]|$$

$$\leqslant |[Df(x_i)]^{-1}|\ |[Df(x_i) - Df(x_{i+1})]|$$

$$= |[Df(x_i)]^{-1}|\ M\ |h_i|$$

$h_i = [Df(x_i)]^{-1} f(x_i)$

$$= |Df(x_i)^{-1}|^2\, M\, f(x_i)$$

$$\underset{(3)}{\leqslant} \frac{1}{2}.$$

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots$$

also true for matrices.

if $|X| < 1$

$$(I-X)^{-1} = I + X + X^2 + \cdots$$

$I-A$ is invertible (since A has small norm)

$ut\ B = (I-A)^{-1}$, $B(I-A) = I$.

$$)| = |I-A| = \left|\sum_{i=0}^{\infty} A^i\right| \leq \sum_{i=0}^{\infty} |A^i| \leqslant \frac{1}{1-\frac{1}{2}} = 2.$$

$$(I-A) = [Df(x_i)]^{-1}[Df(x_{i+1})]$$

$$[Df(x_i)](I-A) Df(x_{i+1}) \overset{m}{=}$$

$\Rightarrow$ Lemma 2.

* Lemma 3. (statement ④) (gonna use lemma 1)

$$\boxed{|f(x_{i+1})| \le \frac{M}{2}|h_i|^2}$$

From lemma 1,

$$|f(x_{i+1}) - f(x_i) - Df(x_i)\,h_i| \le \frac{M}{2}|h_i|^2$$

$$h_i = -[Df(x_i)]^{-1} f(x_i)$$

$$-Df^{-1}(x_i)\,h_i =$$

$$\Rightarrow |f(x_{i+1})| \le |f(x_{i+1}) - f(x_i) - Df(x_i)\,h_i| \le \frac{n}{2}|h_i|^2 \quad \square .$$

* Prove statement 2 : $|h_{i+1}| \le \frac{1}{2}|h_i|$

$$|h_{i+1}| = \left|[Df(x_{i+1})]^{-1} f(x_{i+1})\right| \le \left|[Df(x_{i+1})]^{-1}\right| |f(x_{i+1})| \overset{④}{\le} \frac{n}{2}|h_i|^2 \left|[Df(x_{i+1})]^{-1}\right|$$

$$= |h_i| \frac{n}{2}|h_i| \left|[Df(x_{i+1})^{-1}]\right|$$

$$= |h_i| \frac{M}{2} \left(-[Df(x_i)]^{-1}\right) f(x_i) \left|[Df(x_{i+1})^{-1}]\right|$$

$$\underset{\text{lemma2}}{=} |h_i| \frac{n}{\cancel{2}} \left([Df(x_i)]^{-1}\right) f(x_i)\, \cancel{2}\, |Df(x_i)^{-1}|$$

$$= \underbrace{f(x_i)|Df(x_i)^{-1}|^2 M \quad |h_i|}$$

$$\underset{(3)}{\le} \frac{1}{2}|h_i|$$

* Prove statement 3 : $|f(x_{i+1})|\, |[Df(x_{i+1})^{-1}]|^2 \le |f(x_i)|\, |Df(x_i)^{-1}|^2$

From lemma 3,

$$|f(x_{i+1})|\,|Df(x_{i+1})^{-1}|^2 \underset{\substack{lm3\\lm2}}{\le} \frac{n}{2}|h_i|^2\, 2\,|Df(x_i)^{-1}|^2 \le$$

$$\le 2\,|Df(x_i)^{-1}|^2\, n\, |-Df(x_i)^{-1} f(x_i)|^2 = |Df(x_i)^{-1}|^2 |f(x_i)|$$

* In conclusion, from the 4 proved statements, $x_k \longrightarrow$ solution.

We still need to prove the uniqueness.

:- <u>Prove the uniqueness</u>: in $\overline{U_1}$

;how that if $y \in \overline{U_1}$
and $f(y) = 0$

hen $|y - x_{i+1}| \leqslant \frac{1}{2}|y - x_i|$

Then Taylor's formula. $\underset{\overset{\parallel}{0}}{f(y)} = f(x_i) + Df(x_i)(y - x_i) + r_i$

$$|y - x_i| = -[Df(x_i)]^{-1} f(x_i) + [Df(x_i)]^{-1} r_i$$

$$= h_i - [Df(x_i)]^{-1} r_i$$

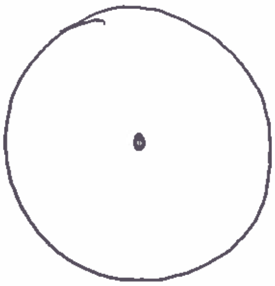
$$|r_i| = |f(y) - f(x_i) - Df(y - x_i)| \underset{\text{lemma1}}{\leq} \frac{M}{2}|y - x_i|^2$$

$$y - x_{i+1} = -[Df(x_i)]^{-1} r_i \qquad (a)$$

$$|y - x_{i+1}| \leq |[Df(x_i)]^{-1}| \frac{M}{2} |y - x_i|^2 \underset{(b)}{\leq} \frac{1}{2}|y - x_i|$$

$$|y - x_1| \leq |[Df(x_0)^{-1}]| \frac{M}{2} |y - x_0|^2 \leq |[Df(x_0)^{-1}| M |h_0| |y - x_0| \overset{\text{asump(3)}}{\leq} \frac{1}{2}|y - x_0|$$

~~$|y - x_0| \leq 2|h_0|$~~



Then by induction ,

$\ast$ $|y - x_{i+1}| \leq \frac{1}{2}|y - x_i|$

$$\frac{|y - x_{i+1}|}{|y - x_i|} \leq |Df(x_i)^{-1}| \frac{M}{2} |y - x_i| \quad \cdots \quad \leq \frac{1}{2} .$$

Mat683
Homework 1
Tran Le

MAT 683 Homework 1
A. Lutoborski. Syracuse University. Fall 2018

1. (10 pts) The floating point system $\mathbb{F}(2, 52, -1022, 1023)$ (IEEE double precision) includes many integers but not all of the integers in its range.
(a) What is the largest integer N such that all integers in the interval $[-N, N]$ are represented exactly in $\mathbb{F}$?
(b) What is the smallest positive integer n that does not belong to $\mathbb{F}$.
Hint: Begin by representing first few nonzero integers in $\mathbb{F}$.

2. (10 pts) Let $x \in \mathbb{R}$. If $\square : \mathbb{R} \to \mathbb{F}$ satisfies two axioms

$$x \in \mathbb{F} \Rightarrow \square(x) = x$$

$$x, y \in \mathbb{R} \quad \text{and} \quad x \le y \Rightarrow \square(x) \le \square(y)$$

then the interval spanned by x and $\square(x)$ contains no points of $\mathbb{F}$ in its interior.

3. (10 pts) What is the last value k displayed by the following scripts? Explain based on floating point number system.

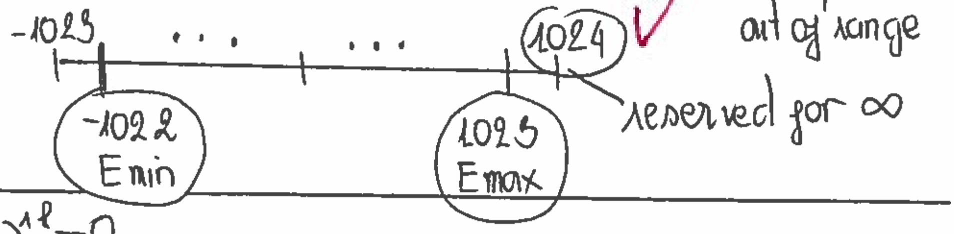
a>
```
>> k=0;
>> while (1+1/2^k)>1
          k=k+1
   end
```

The while loop stops at $k = 53$, since in $[2^0, 2^{0+1})$ k $\Delta_0 = 2^{-t} = 2^{-52}$ ← this is the gap between 1 and the next bigger floating point number.
↑ machine epsilon

$fl(1 + 2^{-53}) = 1$

b>
```
>> k=0;
>> while 2^k<inf
          k=k+1
   end
```

The while loop stops at $k = 1024$ since we have $2^{1024} \geq i$



c>
>> k=0;
>> while ~~2^k<inf~~ $(\frac{1}{2})^{k} > 0$.
          k=k+1
   end

The c> while loop stops at $k = 1075$ since we have in double format, the smallest subnormal is $2^{E_{min}-t} = 2^{-1022-52} = 2^{-1074}$

10

1, The floating point system $F(2, 52, -1022, 1023)$ includes many integer but not all of the intergers in its range.

a, What is the largest integer N s.t all integers in the interval $[-N, N]$ are represented exactly in $\mathbb{F}$

b, What is the smallest positive integer n that does not belong to $\mathbb{F}$

---

a, We have the increment between two consecutive floating point numbers in $[2^E, 2^{E+1}$ is $2^{E-t} = 2^{E-52}$

- So if we want to have that all integers in $[2^E, 2^{E+1}]$ can be represented exactly in $\mathbb{F}$ when $2^{E-t} = \underset{\text{increment between 2 consecutive number}}{1} \Leftrightarrow E = t = 52$.

$\Rightarrow$ The largest integer N that all integers in the interval can be represented exactly i

$2^{E+1} = 2^{t+1} = 2^{52+1} = 2^{53}$

More explanation:

(With all the interval that have the form $[2^L, 2^{L+1})$ where $L < E = t = 52$

(then the increment between 2 consecutive floating point numbers is $2^{L-t} < 2^{E-t}$.)

b, From the analysing from part a, we have the smallest integer that does not belong to $\mathbb{F}$ is $2^{53} + 1$. ✓ □ 1đ

2, Let $x \in \mathbb{R}$.

If $\square : \mathbb{R} \longrightarrow \mathbb{F}$ satisfies two axioms $\begin{cases} x \in \mathbb{F} \rightarrow \square x = x \\ x, y \in \mathbb{R} \text{ and } x \leq y \Rightarrow \square x \leq \square y . \end{cases}$

Then the interval spanned by x and $\square x$ contains no points of $\mathbb{F}$ in its interi

---

- Case 1: If $x \in \mathbb{F}$ then we have $\square x = x \Rightarrow$ there is just one point ✓
- Case 2: If $\begin{cases} x \in \mathbb{R} \\ x \notin \mathbb{F} \end{cases}$ then there is a gap between x and $\square x$. ✓

Now let $y \in \mathbb{F}$, then

⊕ If $x \leq y \leq \square x$ then $\begin{cases} x \leq y \\ y \leq \square x \end{cases} \Rightarrow \begin{cases} \square x \leq \square y \\ \square y \leq \square x \end{cases} \Rightarrow \square y = \square x \xrightarrow{y \in \mathbb{F}} y = \square$
$\Rightarrow \square y = y$

1đ

⊕ If $\square x \leq y < x \Leftrightarrow \begin{cases} \square x \leq y \\ y < x \end{cases} \Rightarrow \begin{cases} \square x \leq \square y \\ \square y \leq \square x \end{cases} \Rightarrow \square y = \square x \Longrightarrow y = \square$

this mean y has to be equal to $\Box x$

$\Rightarrow$ There is no $y \in \mathbb{F}$ in the interior of the interval spanned by x and $\Box x$.

1	10
2	10
3	10
	30

MAT 683 Homework 1
A. Lutoborski. Syracuse University. Fall 2018

1.(10 pts) The floating point system $\mathbb{F}(2, 52, -1022, 1023)$ (IEEE double precision) includes many integers but not all of the integers in its range.
(a) What is the largest integer N such that all integers in the interval $[-N, N]$ are represented exactly in $\mathbb{F}$?
(b) What is the smallest positive integer n that does not belong to $\mathbb{F}$.
Hint: Begin by representing first few nonzero integers in $\mathbb{F}$.

Solution. Consider an integer number with the largest mantissa. The largest mantissa is

$$2^0 + 2^{-1} + \ldots + 2^{-52} = \frac{1 - 2^{-52}}{1 - 2^{-1}} = 2 - 2^{-52}$$

Multiplying largest mantissa by 2^{52} we get $2^{53} - 1$. The next integer is $N = 2^{53}$ also representable exactly. All integers $\leq N$ are in $\mathbb{F}$. There are integers in $[N, 2^{54}]$ which cannot be represented exactly, and the smallest such number is $N + 1 = 2^{53} + 1$.

2. (10 pts) Let $x \in \mathbb{R}$. If $\Box : \mathbb{R} \to \mathbb{F}$ satisfies two axioms

$$x \in \mathbb{F} \Rightarrow \Box(x) = x$$

$$x, y \in \mathbb{R} \quad \text{and} \quad x \leq y \Rightarrow \Box(x) \leq \Box(y)$$

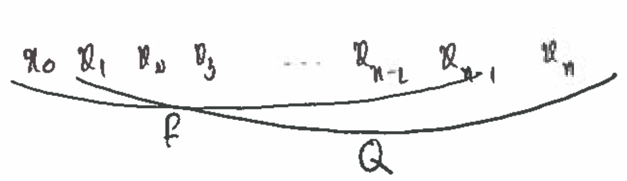
then the interval spanned by x and $\Box(x)$ contains no points of $\mathbb{F}$ in its interior.
Solution. When $x \in \mathbb{F}$ then the claim is true. Assume $x \notin \mathbb{F}$ and (without loss of generality) $x < \Box(x)$. Suppose the claim is false and there is $y \in \mathbb{F}$ such that $x < y < \Box(x)$. By first axiom $\Box(y) = y$. By second from $x < y$ follows that $\Box(x) \leq y$ which is a contradiction to $y < \Box(x)$.
3. (10 pts) What is the last value k displayed by the following scripts? Explain based on floating point number system.

```
>> k=0;
>> while (1+1/2^k)>1
          k=k+1
 end
```

```
>> k=0;
>> while 2^k<inf
          k=k+1
   end
```

```
>> k=0;
>> while  (1/2)^k>0
          k=k+1
   end
```

Solution. (a) $k_{max} = 53$ that is $k_{max} = t + 1$. When $k = 52$ then $1 + 2^{-52} = (1.\dots 1)_2$. The number $1 + 2^{-53}$ is in the midpoint between 1 and $1 + 2^{-52}$ which are both floating point numbers. Rounding to nearest even we round down and $fl(1 + 2^{-53}) = 1$. The inequality is not satisfied and the script stops at $k_{max} = 53$.
(b) $k_{max} = 1024$ that is $k_{max} = E_{max} + 1$.
(c) $k_{max} = 1075$ that is $k_{max} = -E_{min} + t + 1$. The smallest positive subnormal is $2^{-1022-52} = 2^{-1074}$.



Due 09/20

MAT 683 Homework 2
A. Lutoborski. Syracuse University. Fall 2018

1.(20 pts) Assume that $P, Q \in \mathbb{P}_{n-1}$ interpolate f at $x_0, \dots, x_{n-1}$ and $x_1, \dots, x_n$ respectively and all points are distinct. The nodes $x_1, \dots, x_{n-1}$ are "common" nodes of both polynomials. Then $L \in \mathbb{P}_n$ and

$$L(x) = \frac{(x-x_0)Q(x) - (x-x_n)P(x)}{x_n - x_0} = \frac{(x-x_0)L_{1,\dots,n}(x) - (x-x_n)L_{0,\dots}}{x_n - x_0}.$$

interpolates f at $x_0, \dots, x_n$. The above successive linear interpolation formula constructs an interpolation polynomial $L = L_{0,\dots,n}$ of degree n as a convex combination of two interpolants P and Q which both interpolate at nodes $x_1, \dots, x_{n-1}$. The coefficients in the convex combination are not constants but polynomials of degree 1 and the combination becomes a polynomial of degree n. The simplest example of the formula is when $\boxed{n = 1}$, $P(x_0) = f(x_0)$ and $Q(x_0) = f(x_1)$. Then

$$L(x) = \frac{\overbrace{(x-x_0)}^{~~L_1(x)~~}f(x_1) - (x-x_1)f(x_0)}{x_1 - x_0} = \overbrace{\frac{x-x_1}{x_0-x_1}}^{l_0(x)}f(x_0) + \overbrace{\frac{x-x_0}{x_1-x_0}}^{l_1(x)}f(x_1) = l_0(x)f(x_0) + l_1(x$$

which is the canonical form of the Lagrange interpolation polynomial $L_{0,1}(x) = l_0(x)f(x_0) + l_1(x)f(x_1)$.

Successive linear interpolation formula is used in the Neville's method:

		$J=1$ $I=1,n$	$J=2$ x_3-x_1 $x_{(i)}-x_{(j)}$	$J=3$ $P_{i,J+1}=$	$P(I)$
x_0	$f(x_0) = L_0(x)$				
x_1 $I-J-1$	$f(x_1) = L_1(x)$	$L_{0,1}(x)$			
x_2	$f(x_2) = L_2(x)$	$L_{1,2}(x)$ $P_{i-1,j}$	$L_{0,1,2}(x)$		
x_3 $x(i)$	$f(x_3) = L_3(x)$	$L_{2,3}(x)$ $P_{i,j}$	$L_{1,2,3}(x)$ P_{43} $P_{i,J+1}$	$L_{0,1,2,3}(x)$	$L_{0,1,2,3,n-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_n	$f(x_n) = L_n(x)$	$L_{n-1,n}(x)$	$L_{n-2,n-1,n}(x)$	$L_{n-3,\dots,n}(x)$	$\dots$

Q P

$L_{0,\dots,n}(x$

and in the Aitken method:

	$J=1$					
x_0	$f(x_0) \overset{P_{11}}{=} L_0(x)$					
x_1	$f(x_1) \overset{P_{21}}{=} L_1(x)$	$L_{0,1}(x)$ P_{22}				
x_2	$f(x_2) \overset{P_{31}}{=} L_2(x)$	$L_{0,2}(x)$ P_{32}	$L_{0,1,2}(x)$			
x_3	$f(x_3) \overset{P_{41}}{=} L_3(x)$	$L_{0,3}(x)$ P_{42}	$L_{0,1,3}(x)$	$L_{0,1,2,3}(x)$		
$\vdots$	P_{n1}	$\vdots$	$\vdots$	$\vdots$	L_{01234}	$L_{0,1,\dots,}$
x_n	$f(x_n) \overset{f}{=} L_n(x)$	$L_{0,n}(x)$ P_{n2}	$L_{0,1,n}(x)$	$L_{0,1,2,n}(x)$	$\dots$ L_{0123n}	$L_{0,\dots,n}(x)$

$L_{012\ n-1}$
$L_{0123\ n}$

L_0

$P(i,1) = f$

$i = j+1:n$

$$P_{(i,j+1)} = \frac{P(i,j)(x-x_j) - P(j,j)(x-x_i)}{x_i - x_j}$$

row ↑ col.

? (a) Show that $L_{0,\ldots,n}(x)$ values generated in the table via the successive linear interpolation formula are indeed the values of the Lagrange interpolating polynomial.

✓ (b)Use the following scripts for Aitken method to determine the reciprocal of $x = 1.03$ from a table of $k = 10$ equispaced points in $[1, 2]$ for $f(x) = 1/x$.

(c) Explain the advantages of the above Aitken algorithm over the cardinal Lagrange interpolation at the same nodes for the same function values. operation cnt

(d) Modify the Aitken script to obtain a script for Neville's method. Compare the results for (b). What function would be more accurately interpolated by Aitken than by Neville?

(a)

```
function[Q,R]=aitken(x,f,xval)
n=length(x); P=zeros(n);
P(:,1)=f;
for j=1:n-1  ← for colum
    for i=j+1:n  ← row        x - x(i-j-1)    j i-1 for i.        x(i) - x(i-j-1)
        P(i,j+1)=(P(i,j)*(xval-x(j))-P(j,j)*(xval-x(i)))/(x(i)-x(j));
    end   ↑ in th.
end
Q=P(n,n); R=[x.' P]
```

$$\frac{\left(\overset{Q}{P_{ij}}(x - x_{(j)}) - \overset{P}{(P_{jj})}(x - x_i)\right)}{x_i - x_j}$$

(b)

```
%runaitken
x=1:.2:2; f=1./x;
[interpval  table]=aitken(x,f,1.03);
fprintf('Interpolated value=%10.8f\n\n', interpval)
disp('Table=')
disp(table)
```

P(:,1) j=1..n

f(x0) = P(0,1)

f(x1) P(1,1)

f(x2) = P(2,1)

f(x3) P(3,1)

⋮

f(xn) P(n,1)

```
for j = 1 : n
    for i = 1 : n
        P(i, i+j) = P(i, i+j-1) * (x(i+j) - xval)
                    - (x(i) - xval) * ( P(i+1, i+j) ) ) / (x_{i+j} - x_i
    end
end
```

Mat68.
#W8
Tran Le

20

1) Given $\underbrace{x_0, \overbrace{x_1, x_2, .., x_{n-1}}^{Q_{1,..,n}(x)}}_{P_{0,..,n-1}(x)}, x_n$ distinct .

$P_{0,..,n-1}(x) \in \mathbb{P}_{n-1}$ interpolates $x_0, .., x_{n-1}$

$Q_{1,..,n}(x) \in \mathbb{P}_{n-1}$ interpolates $x_1, .., x_n$.

Then $L(x) = \dfrac{(x-x_0)\, Q_{1,..,n}(x) - (x-x_n)\, P_{0,..,n-1}(x)}{(x_n - x_0)}$

a) Show that $L_{0,..,n}(x)$ values generated in the table are indeed the values of Lagrang interpolating polynomial.

*① We have $L_{0,..,n}(x) \in \mathbb{P}_n$

*② We want to prove that $L_{0,..,n}(x)$ interpolates f at $\overbrace{x_0, x_1, .., x_{n-1}, x_n}^{(n+1) \text{ points}}$:

- We have $L_{(0,..,n)}(x)$ interpolates f at $x_1, .., x_{n-1}$ since

$$L_{0,..,n}(x_i) = \frac{(x_i - x_0)\, Q_{1,..,n}(x_i) - (x_i - x_n)\, P_{0,..,n-1}(x_i)}{x_n - x_0} =$$

$$= \frac{(x_i - x_0) f(x_i) - (x_i - x_n) f(x_i)}{x_n - x_0} = \frac{(x_n - x_0) f(x_i)}{x_n - x_0} = f(x_i),\ i = \overline{1,}$$

- At node x_0

$$L_{(0,..,n)}(x_0) = \frac{\overset{0}{\overset{\shortparallel}{(x_0 - x_0)}}\, Q_{1,..,n}(x) - (x_0 - x_n)\, \overset{f(x_0)}{\overset{\shortparallel}{P_{0,..,n-1}(x_0)}}}{x_n - x_0} = \frac{-(x_0 - x_n) f(x_0)}{x_n - x_0} = f($$

- At node x_n

$$L_{(0,..,n)}(x_n) = \frac{(x_n - x_0)\, \overset{f(x_n)}{\overset{\shortparallel}{Q_{1,..,n}(x_n)}} - \overset{0}{\overset{\shortparallel}{(x_n - x_n)}}\, P_{0,..,n-1}(x_n)}{x_n - x_0} = \frac{(x_n - x_0) f(x_n)}{x_n - x_0} = f(x_n$$

* Then from ① and ② and the theorem that there exist a unique polynomial ~~of~~ degree n that interpolate f at $x_0, .., x_n$ $\Rightarrow$ the above L is the same with Lagrange interpolating polynomial.

b)

```
function[Q,R]=aitken(x,f,xval)
%input:
%x: row vector containing points x0,x1, ..., xn
%f: the actual function that we want to approcimate it values at
those xn points
%xval: the value x that we want to know the approcimated value
at
%Output:
%Q the approcimate value of f at xval
%R the table value of the polynomial at xval
n=length(x);
P=zeros(n);
P(:,1)=f;
for j=1:n-1
    for i=j+1:n
        P(i,j+1)=(P(i,j)*(xval-x(j))-P(j,j)*(xval-x(i)))/(x(i)-
x(j));
    end
end
Q=P(n,n);
R=[x.' P];
```

From the results (on next page) we have that in this case, when $f = \frac{1}{x}$, the results are the same when we use Aitkens and Nevill

d)

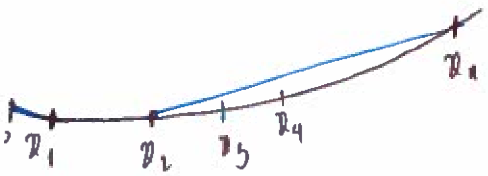
```
function[Q,R]=neville(x,f,xval)
%input:
%x: row vector containing points x0,x1, ..., xn
%f: the actual function that we want to approcimate it values at
those xn points
%xval: the value x that we want to know the approcimated value
at
%Output:
n=length(x);
P=zeros(n);
P(:,1)=f;
for j=1:n-1
    for i=j+1:n
        P(i,j+1)=(P(i,j)*(xval-x(i-j))-P(i-1,j)*(xval-
x(i)))/(x(i)-x(i-j));
    end
end
Q=P(n,n);
R=[x.' P];
```

* What function would be more accurately interpolated by Aiken than by Neville? ✓

I don't really get a precise answer for this question, but I get the idea that the difference in accuracy of Aiken method and Neville method based on the order of nodes that we choose to interpolate. → explain more next pag

With the function that f_i are not much different :

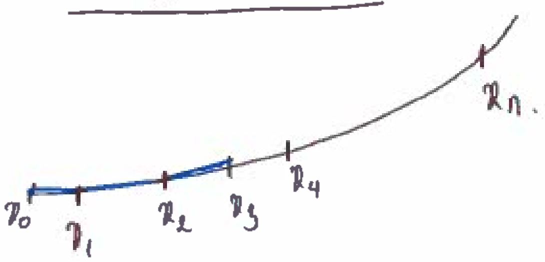
Aitken method



L_{012n}

Interpolates sume first notes and the last node x_n

Neville method .

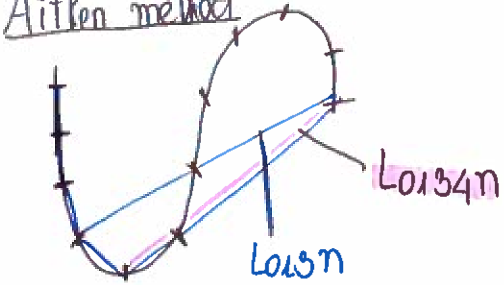


interpolate ~~the~~ the near by first notes
interpolate the near by last notes .

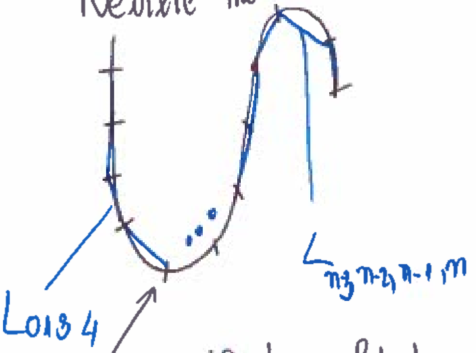
Then in this case Aitken method would be more accurate .

With the function that the values of f_i when i's are small and when i's are big are different

Aitken method



Neville method

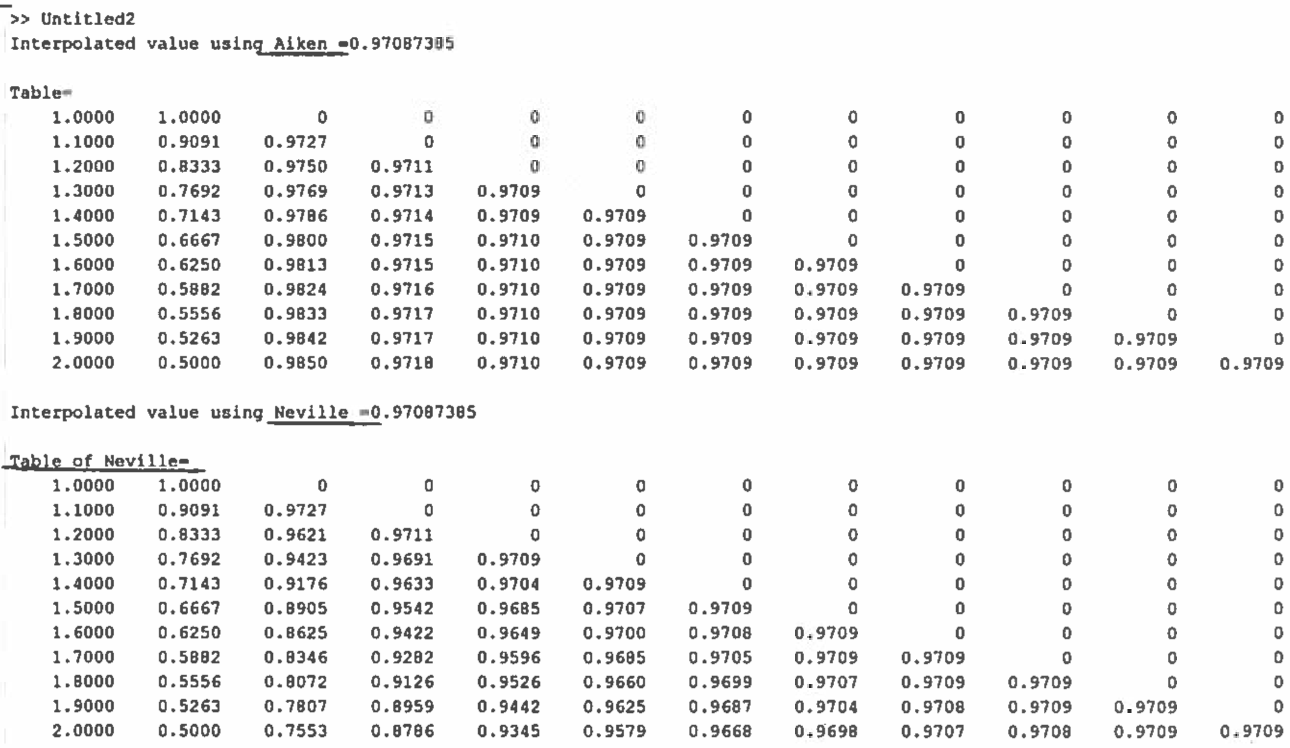


Then in this case the Neville method would be more accurate .

% Run Aitken and Neville methods .

```
%runaiken
x=1:.1:2;
f=1./x;
[interpval table]=aitken(x,f,1.03);
fprintf('Interpolated value using Aiken =%10.8f\n\n', interpval)
disp('Table=')
disp(table)
%run neville
~~x=1:.1:2;~~
~~f=1./x;~~
[interpval_n table_n]=neville(x,f,1.03);
fprintf('Interpolated value using Neville =%10.8f\n\n',
interpval_n)
disp('Table of Neville=')
disp(table_n)
%
%Compute the error of Aiken and Neiville
xval=1.03
errorA=abs(interpval-1/xval)
errorN=abs(interpval-1/xval)
```

Results.



```
>> Untitled2
Interpolated value using Aiken =0.97087385

Table=
    1.0000    1.0000         0         0         0         0         0         0         0         0         0         0
    1.1000    0.9091    0.9727         0         0         0         0         0         0         0         0         0
    1.2000    0.8333    0.9750    0.9711         0         0         0         0         0         0         0         0
    1.3000    0.7692    0.9769    0.9713    0.9709         0         0         0         0         0         0         0
    1.4000    0.7143    0.9786    0.9714    0.9709    0.9709         0         0         0         0         0         0
    1.5000    0.6667    0.9800    0.9715    0.9710    0.9709    0.9709         0         0         0         0         0
    1.6000    0.6250    0.9813    0.9715    0.9710    0.9709    0.9709    0.9709         0         0         0         0
    1.7000    0.5882    0.9824    0.9716    0.9710    0.9709    0.9709    0.9709    0.9709         0         0         0
    1.8000    0.5556    0.9833    0.9717    0.9710    0.9709    0.9709    0.9709    0.9709    0.9709         0         0
    1.9000    0.5263    0.9842    0.9717    0.9710    0.9709    0.9709    0.9709    0.9709    0.9709    0.9709         0
    2.0000    0.5000    0.9850    0.9718    0.9710    0.9709    0.9709    0.9709    0.9709    0.9709    0.9709    0.9709

Interpolated value using Neville =0.97087385

Table of Neville=
    1.0000    1.0000         0         0         0         0         0         0         0         0         0         0
    1.1000    0.9091    0.9727         0         0         0         0         0         0         0         0         0
    1.2000    0.8333    0.9621    0.9711         0         0         0         0         0         0         0         0
    1.3000    0.7692    0.9423    0.9691    0.9709         0         0         0         0         0         0         0
    1.4000    0.7143    0.9176    0.9633    0.9704    0.9709         0         0         0         0         0         0
    1.5000    0.6667    0.8905    0.9542    0.9685    0.9707    0.9709         0         0         0         0         0
    1.6000    0.6250    0.8625    0.9422    0.9649    0.9700    0.9708    0.9709         0         0         0         0
    1.7000    0.5882    0.8346    0.9282    0.9596    0.9685    0.9705    0.9709    0.9709         0         0         0
    1.8000    0.5556    0.8072    0.9126    0.9526    0.9660    0.9699    0.9707    0.9709    0.9709         0         0
    1.9000    0.5263    0.7807    0.8959    0.9442    0.9625    0.9687    0.9704    0.9708    0.9709    0.9709         0
    2.0000    0.5000    0.7553    0.8786    0.9345    0.9579    0.9668    0.9698    0.9707    0.9708    0.9709    0.9709
```

c) Explain the advantage of the above Aithen algorithm over the cardinal Lagrange interpolation at the same nodes for the same function values.

* First, we compute the operation counts of Aithen algorithm.

At each $L = \dfrac{(\overset{\textcircled{1}}{x - x_0})\overset{\textcircled{2}}{}Q \overset{\textcircled{3}}{-} (\overset{\textcircled{4}}{x - x_n})\overset{\textcircled{5}}{}P}{\underset{\textcircled{6}}{x_n - x_0}} \textcircled{7}$ $\Rightarrow$ 7 operators.

To fulfill the table, we need

$$7(n + (n-1) + \cdots + 2 + 1) = \frac{7(n)(n+1)}{2} \text{ operators} \doteq 3.5n^2 + 3.5n$$

* Second, we want to comptue the operation counts of Lagrange interpolating polyno

- $l_i(x) = \dfrac{(x-x_0)(x-x_1)\ldots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\ldots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)} =$ n subtractions and (n-1) mu / n subs and (n-1) mul

$\uparrow$

$(2(2n-2)+1 = 4n-3$

- $L(x) = \sum_{i=0}^{n} l_i(x) f_i$

for each i from 0 to n, we have to compute $l_i(x) \leftarrow$ costs $4n-3$ } $\Rightarrow 4n-2$
multiplys l_i with $f_i \leftarrow$ 1 muls.

- we do it for i from $0 \to n$ $\Rightarrow$ $(n+1)(4n-2)$ operations.

- then we add all $l_i f_i$, for $i = \overline{0, n}$ $\Rightarrow$ need $(n+1)$ adds.

$\Rightarrow$ Totally, we need $(n+1)+(n+1)(4n-2) = 4n^2 + 3n - 4$.

* So we have

Aiken algorithm's operation counts $= O(3.5n^2)$ $\leftarrow$ more effective

Lagrange algorithm's operation count $= O(4n^2)$

Mat682
TranLe
#W19

Homework 19 and 20 (Thursday, April 19th)

Problem 1. The standard Chebyshev polynomials for $k = 0, 1, 2, \ldots$ are given by

$$\tau_k(t) = \begin{cases} \cos(k\cos^{-1} t) & \text{for } t \in [-1, 1], \\ \cosh(k\cosh^{-1} t) & \text{for } t > 1, \\ (-1)^k \tau_k(-t) & \text{for } t < -1. \end{cases} \tag{1}$$

By considering the trigonometric and hyperbolic identities

$$\cos(k \pm 1)\theta = \cos k\theta \cos\theta \mp \sin k\theta \sin\theta,$$
$$\cosh(k \pm 1)\theta = \cosh k\theta \cosh\theta \pm \sinh k\theta \sinh\theta,$$

prove that the Chebyshev polynomials $\tau_k(t)$ satisfy the three-term recurrence

$$\tau_{k+1}(t) = 2t\tau_k(t) - \tau_{k-1}(t) \tag{2}$$

(for the cases $|t| \leq 1$ and $|t| > 1$, separately). Then by induction or otherwise, prove that

$$\tau_k(t) = \frac{1}{2}\left[\left(t + \sqrt{t^2 - 1}\right)^k + \left(t - \sqrt{t^2 - 1}\right)^k\right]. \tag{3}$$

Problem 2. The polynomial that achieves the minimization

$$\min_{p_k \in \Pi_k, p_k(0)=1} \max_{z \in [a,b]} |p_k(z)| \tag{4}$$

is known to be the shifted and scaled Chebyshev polynomial

$$\chi_k(t) = \frac{\tau_k\left(\frac{b+a}{b-a} - \frac{2t}{b-a}\right)}{\tau_k\left(\frac{b+a}{b-a}\right)}. \tag{5}$$

Prove that

$$\|\mathrm{e}^{(k)}\|_A \leq 2\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^k \|\mathrm{e}^{(0)}\|_A. \tag{6}$$

Notations are the same as used in class. That is, Π_k is the set of all polynomials whose degree is no more than k, $\mathrm{e}^{(k)}$ is the error associated with the kth iterate of CG, and κ is the condition number of A. (The majority of the proof of (6) was given in class. The point of this exercise is to fill in some gaps.)

*HW19 / 20 > Problem 17

* Prove (2): $T_k(t)$ satisfy the three-term recurrence.

$$T_{k+1}(t) = 2t\, T_k(t) - T_{k-1}(t) \qquad (2).$$

* We have the LHS (2):

$$\text{LHS} = T_{k+1}(t) = \begin{cases} \cos((k+1)\cos^{-1}(t)) & t \in [-1,1] \\ \cosh((k+1)\cosh^{-1}(t)) & t > 1 \\ (-1)^{k+1}\, T_{k+1}(-t) & t < -1 \end{cases}$$

$$= \begin{cases} \cos(k\cos^{-1}(t))\, \cos(\overset{t}{\cancel{\cos^{-1}(t)}}) - \sin(k\cos^{-1}(t))\, \sin(\cos^{-1}(t)), & t\in \\ \cosh(k\cosh^{-1}(t))\, \cosh(\overset{t}{\cancel{\cos^{-1}(t)}}) + \sinh(k\cosh^{-1}(t))\, \sinh(\cosh^{-1}(t)), & \\ (-1)^{k+1}\, T_{k+1}(-t) & t < 1 \end{cases}$$

* Consider the RHS of (2)

- $$2t\, T_k(t) = \begin{cases} 2t\, \cos(k\cos^{-1}(t)) & t\in[-1,1] \\ 2t\, \cosh(k\cosh^{-1}(t)) & t > 1 \\ 2t\, (-1)^k\, T_k(-t) & t < -1 \end{cases}$$

- $$T_{k-1}(t) = \begin{cases} \cos((k-1)\cos^{-1}(t)) & t\in[-1,1] \\ \cosh((k-1)\cosh^{-1}(t)) & t > 1 \\ (-1)^{k-1}\, T_{k-1}(-t) & t < -1 \end{cases}$$

$$= \begin{cases} \cos(k\cos^{-1}(t))\, \cos(\overset{t}{\cancel{\cos^{-1}(t)}}) + \sin(k\cos^{-1}(t))\, \sin(\cos^{-1}((t))) & t\in[-1 \\ \cosh(k\cos^{-1}(t))\, \cosh(\overset{t}{\cancel{\cosh^{-1}(t)}}) - \sinh(k\cosh^{-1}(t))\, \sinh(\cosh^{-1}(t)), & t> \\ (-1)^{k-1}\, T_{k-1}(-t) & t < -1 \end{cases}$$

- $$\text{RHS} = 2t\, T_k(t) - T_{k-1}(t) = \begin{cases} t\cos(k\cos^{-1}(t)) - \sin(k\cos^{-1}(t))\, \sin(\cos^{-1}(t)) & t\in[-1, \\ t\cosh(k\cosh^{-1}(t)) + \sinh(k\cosh^{-1}(t))\, \sinh(\cosh^{-1}(t)) & t> \\ 2t\,(-1)^k\, T_k(-t) - (-1)^{k-1}\, T_{k-1}(-t), & t < -1. \end{cases}$$

* So we have LHS = RHS in case $t \in [-1,1]$ and $t > 1$.

In care $t < -1$,

hen ℓ is even, $t < -1$, • $T_{\ell+1}(t) = -T_{\ell+1}(\overbrace{-t}^{>1})$

$$\bullet\ 2t(-1)^{\ell} T_{\ell}(-t) - (-1)^{\ell-1} T_{\ell-1}(-t) = 2t\, T_{\ell}\underbrace{(-t)}_{>1} + T_{\ell-1}\underbrace{(-t)}_{>1} =$$

$$= 2t\left(\cosh(\ell \cosh^{-1}(-t))\right) + (-t)(\cosh(\ell \cos^{-1}(\overset{>1}{-t})) - \sinh(\ell \overset{>1}{\cosh^{-1}}(-t))\sinh(\cosh^{-1}(-t$$

$$= t\left(\cosh(\ell \cosh^{-1}(-t))\right) - \sinh(\ell \cosh^{-1}(-t))\sinh(\cosh^{-1}(-t))$$

$$= -T_{\ell+1}(-t)$$

Similarly for the case when ℓ is odd

→ The Chebyshev polynomial $T_{\ell}(t)$ satisfy the three term recurrence □.

* Problem 2 : We want to prove that $\|\vec{e}^{(\ell)}\|_A \le 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{\ell} \|\vec{e}^{(0)}\|_A$

From class we know that

$$\|\vec{e}^{(\ell)}\|_A = \left\{ \min_{\substack{P_\ell \in \Pi_\ell \\ P_\ell(0)=1}} \max_j |P_\ell(\lambda_j)\|_A \right\} \|\vec{e}^{(0)}\|_A \le \max_j |\chi_\ell(\lambda_j)| \quad \|\vec{e}^{(0)}\|_A$$

$$\le \max_{t\in[a,b]} |\chi_\ell(t)| \quad \|\vec{e}^{(0)}\|_A \quad , \text{ where } \begin{cases} a = \lambda_{min} \\ b = \lambda_{max} \end{cases}$$

* So now we need to prove that $\max_{t\in[a,b]} |\chi_\ell(t)| \le 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{\ell}$.

- Consider $\chi_\ell(t)$, we have

$$|\chi_\ell(t)| = \frac{\left|T_\ell\left(\frac{b+a}{b-a} - \frac{2t}{b-a}\right)\right|}{\left|T\left(\frac{b+a}{b-a}\right)\right|} \qquad (*)$$

- We have $\underbrace{\frac{b+a}{b-a} - \frac{2b}{b-a}}_{-1} \le \frac{b+a}{b-a} - \frac{2t}{b-a} \le \underbrace{\frac{b+a}{b-a} - \frac{2a}{b-a}}_{1}$ (since note that $a<t<b$

(*

$$\Rightarrow \left.\begin{matrix} -1 \le \frac{b+a}{b-a} - \frac{2t}{b-a} \le 1 \end{matrix}\right\} \Rightarrow \left|T_\ell\left(\frac{b+a}{b-a} - \frac{2t}{b-a}\right)\right| = |\cos(\sim)| \le 1$$

Then by (1) of Problem 1, we have

- We also have

$$T_\ell\left(\frac{b+a}{b-a}\right) \overset{\text{by (3)}}{=\!=} \frac{1}{2}\left[\left(\frac{b+a}{b-a} + \sqrt{\frac{4ab}{(b-a)^2}}\right)^{\ell} + \left(\frac{b+a}{b-a} - \sqrt{\frac{4ab}{(b-a)^2}}\right)^{\ell}\right]$$

$$= \frac{1}{2}\left[\left(\frac{(\sqrt{a}+\sqrt{b})^2}{b-a}\right)^{\ell} + \left(\frac{(\sqrt{b}-\sqrt{a})^2}{b-a}\right)^{\ell}\right] = \frac{1}{2}\left[\left(\frac{\sqrt{a}+\sqrt{b}}{\sqrt{b}-\sqrt{a}}\right)^{\ell} + \underbrace{\left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{a}+\sqrt{b}}\right)^{\ell}}_{\geq 0}\right]$$

$$\geq \frac{1}{2}\left[\left(\frac{\sqrt{a}+\sqrt{b}}{\sqrt{b}-\sqrt{a}}\right)^{\ell}\right] \qquad (***)$$

Then from (*)+(**)+(***) we have

$$|\chi_\ell(t)| \le 2\left(\frac{\sqrt{b}-\sqrt{a}}{\sqrt{a}+\sqrt{b}}\right)^{\ell} = 2\left(\frac{\sqrt{\frac{b}{a}}-1}{1+\sqrt{\frac{b}{a}}}\right)^{\ell} = 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{\ell} \quad \text{since } K = \frac{\lambda_{max}}{\lambda_{min}}$$

$$\Rightarrow \max_{t\in[a,b]} \chi_\ell(t) \le 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{\ell}$$

$$\Rightarrow \|\vec{e}^{(\ell)}\|_A \le \max_{t\in[a,b]} \chi_\ell(t) \|\vec{e}^{(0)}\|_A \le 2\left(\frac{\sqrt{K}-1}{\sqrt{K}+1}\right)^{\ell} \|\vec{e}^{(0)}\|$$

This is what we need to prove.

* Prove (3): $T_\ell(t) = \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell} + \left(t-\sqrt{t^2-1}\right)^{\ell}\right]$ (3)

- $T_{(0)}(t) = \begin{cases} \cos(0) = 1 & t \in [-1,1] \\ \cosh(0) = 1 & t > 1 \\ (-1)^0\, T_{(0)}(-t) & \end{cases}$

- Induction hypothesis (3) is true for all $\ell = \overline{0,n}$, which means we have ⇒

→ $T_{\ell-1} = \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell-1} + \left(t-\sqrt{t^2-1}\right)^{\ell-1}\right]$

$T_{\ell} = \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell} + \left(t-\sqrt{t^2-1}\right)^{\ell}\right]$

- We want to prove (3) is also true when i = ℓ+1, which means we want to prove $T_{\ell+1}(t) = \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell+1} + \left(t-\sqrt{t^2-1}\right)^{\ell+1}\right]$.

- Now we will prove the above claim.

We have by (2) that

$$T_{\ell+1}(t) = 2t\,T_\ell(t) - T_{\ell-1}(t)$$

$$= t\left[\left(t+\sqrt{t^2-1}\right)^{\ell-1} + \left(t-\sqrt{t^2-1}\right)^{\ell-1}\right] - \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell-1} + \left(t-\sqrt{t^2-1}\right)^{\ell-1}\right]$$

$$= \left(t+\sqrt{t^2-1}\right)^{\ell-1}\left[t\left(t+\sqrt{t^2-1}\right) - \frac{1}{2}\right] + \left(t-\sqrt{t^2-1}\right)^{\ell-1}\left[t\left(t-\sqrt{t^2-1}\right) - \frac{1}{2}\right]$$

⊕ We have $\frac{1}{2}\left(t+\sqrt{t^2-1}\right)^2 = \frac{1}{2}\left[t^2+t^2-1+2t\sqrt{t^2-1}\right] = t\left(t+\sqrt{t^2-1}\right) - \frac{1}{2}$

$\frac{1}{2}\left(t-\sqrt{t^2-1}\right)^2 = \frac{1}{2}\left[t^2+t^2-1-2t\sqrt{t^2-1}\right] = t\left(t-\sqrt{t^2-1}\right) - \frac{1}{2}$

Then

$$= \frac{1}{2}\left[\left(t+\sqrt{t^2-1}\right)^{\ell+1} + \left(t-\sqrt{t^2-1}\right)^{\ell+1}\right]$$

Thus (3) is also true when i = ℓ+1.

⟹ By induction (3) is ~~c~~true □.

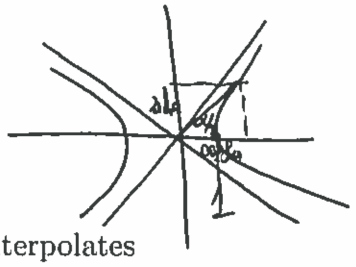
is an increasing funct.

$\sinh(x) = \frac{e^{x}-e^{-x}}{2} = \frac{e^{2x}-1}{2e^{x}}$ $\cosh(x) = \frac{e^{x}+e^{-x}}{2} = \frac{e^{2x}+1}{2e^{x}}$ $\cosh^2 x - \sinh^2 x = 1$

$[\sinh(x)]' = \cosh(x)$ $[\cosh(x)]' = \sinh x$

$\cosh(0) = 1$

Due Tues

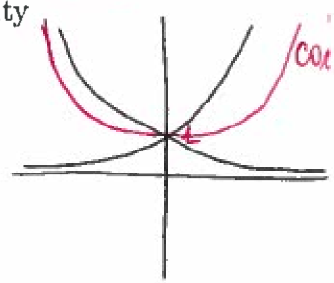


MAT 683 Homework 3

A. Lutoborski. Syracuse University. Fall 2018

1. Kincaid #14, p 325 Let $p \in \mathbb{P}_{n-1}$ be a polynomial that interpolates $f(x) = \sinh x$ at any set of n nodes in the interval $[-1, 1]$ assuming that one of the nodes is 0. Prove that the error satisfies on $[-1, 1]$ the inequality

$$|f(x) - p(x)| \leq \frac{2^n}{n!}|f(x)|$$

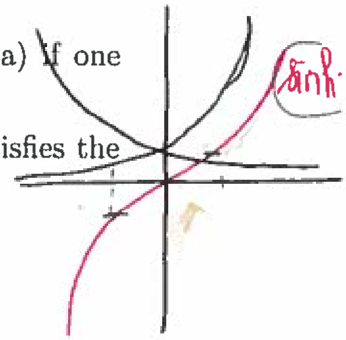


2. Suppose f is a function on $[0, 3]$ for which one knows that

$$f(0) = 1, \quad f(1) = 2, \quad f'(1) = -1, \quad f(3) = f'(3) = 0.$$

(a) Estimate $f(2)$ using Hermite interpolation

(b) Estimate the maximum possible error of the answer given in (a) if one knows, in addition that $f \in C^5[0, 3]$ and $|f^{(5)}(x)| \leq M$ on $[0, 3]$.



3. Maxflat filter. Find a third degree polynomial $H(x)$ which satisfies the conditions

$$H(0) = 1, \quad H'(0) = 0, \quad H(1) = 0, \quad H'(1) = 0.$$

(*) Find a polynomial H of degree $2p - 1$ which satisfies

$$H^{(k)}(0) = \delta(k), \quad 0 \leq k < p$$
$$H^{(k)}(1) = \delta(\ell) \quad 0 \leq k < p$$

$H^{(\ell)}(0) =$

where $\delta(0) = 1$, and $\delta(k) = 0$ for $k \neq 0$.

4. Node polynomial for Chebyshev points. Show that

$$p(x) = 2^{-n}(T_{n+1}(x) - T_{n-1}(x)), \qquad n \geq 1$$

is the unique monic polynomial in $\mathbb{P}_{n+1}$ with zeros at the $n + 1$ Chebyshev points $x_j = \cos(\frac{j\pi}{n})$, $0 \leq j \leq n$ which are the points at which $T_n(x_j) = 0$.

Lecture note 5 :

	$x_0 = 0$	$x_1 = 1$
$f^{(0)}(x_i)$	1	1
$f^{(1)}(x_1)$	0	0
	0	0
$f^{(p-1)}(x_i)$	0	0

1) Kindcaid #14 p325 .

Le Tran

Let $p\in \mathbb{P}_{n-1}$ be a polynomial that interpolates $f(x)=\sinh x$ at any set of n nodes in the interval $[-1,1]$, assuming that one of the node is 0.

Prove that the error : $|f(x)-p(x)| \le \frac{2^n}{n!} f(x)'$

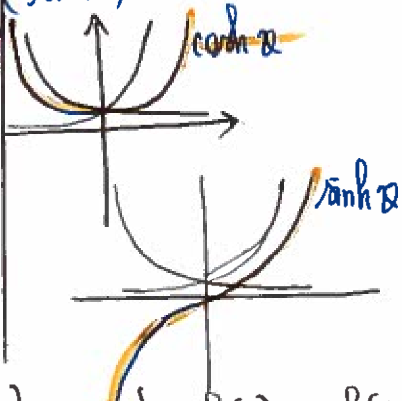
* Review about functions: $\sinh(x)$ and $\cosh(x)$

$\sinh(x) = \frac{e^x-e^{-x}}{2} = \frac{e^{2x}-}{2}$

$\cosh(x) = \frac{e^x+e^{-x}}{2} = \frac{e^{2x}+}{2}$

$(\sinh x)' = \cosh(x)$

$(\cosh(x))' = \sinh(x)$



* We have since $f(x)=\sinh(x)\in C^\infty[-1,1]$.

$$|f(x)-p(x)| \le \frac{|f^{(n)}(\xi)|}{n!}\underbrace{|(x-x_0)|\cdots|(x-0)|\cdots|(x-x_{n-1})|}_{n \text{ nodes , one of the node is } 0}$$

• note that for all x_i , $|x-x_i| \le |+1-(-1)| = 2$

Then $|f(x)-p(x)| \le \frac{\|f^{(n)}\|}{n!} 2^{n-1} |x|$ (1)

* Now consider $\|f^{(n)}\|$

$$\|f^{(n)}\| = \max_{\xi\in[-1,1]} |f^{(n)}(\xi)| = \max_{\xi\in[-1,1]}\{|\sinh(\xi)|, |\cosh(\xi)|\} = \max\{\sinh(1), \cosh(1$$

$$= \max\left\{\underset{1.175}{\frac{e^1-e^{-1}}{2}}, \underset{1.54}{\frac{e^1+e^{-1}}{2}}\right\} \le 2 \quad (2)$$

✓

* Now we want to prove that $|x| \le |\sinh(x)|$ for $x\in[-1,1]$. (3)

It suffices to prove that $\sinh(x) \geq x$ when $x\in[0,1]$

$g(x) = \sinh(x) - x = \frac{e^x-e^{-x}}{2} - x$

$g'(x) = \frac{e^x+e^{-x}}{2} - 1 > 0, \forall x\in[0,1]$.

$\Rightarrow g(x) \geq g(0) = \frac{e^0-e^0}{2} - 0 = 0 \Rightarrow \sinh(x) \geq x$

✓

10

* Then from (1) and (2) and (3) :

$$|f(x)-g(x)| \le \frac{\|f^{(n)}(x)\|}{n!} 2^{+n-1} \underbrace{|x|}_{\le|\sinh(x)|=|f(x)|} \le \frac{2}{n!} 2^{n-1} |f(x)| = \frac{2^n}{n!}|f(x)| \quad \square$$

2> Suppose f is a function on $[0,3]$ for which one knows that

$$f(0)=1\ ;\ f(1)=2 \quad f'(1)=-1\ ;\ f(3)=f'(3)=0$$

a> Estimate $f(2)$ using Hermite interpolation.

b> Estimate the maximum possible error of the answer given in a> if one knows, in addition, that $\left|\begin{array}{l} f\in C^5[0,3] \\ |f^{(5)}(x)|\le M \text{ on } [0,3]. \end{array}\right.$

a> We first want to find a polynomial that can interpolate:

$x_0=0$	$x_1=1$	$x_2=3$	$k=2$
$m_0=1$	$m_1=2$	$m_2=2$	$\Rightarrow$ The polynomial $p(x)\in\mathbb{P}_{5-1}=\mathbb{P}_4$.

Then $H(x)=\underbrace{f(x_0)}_{=1}h_{0,0}(x)+\underbrace{f(x_1)}_{=2}h_{1,0}(x)+\underbrace{f(x_2)}_{=0}h_{2,0}(x)+\underbrace{f'(x_1)}_{-1}h_{1,1}(x)+\underbrace{f'(x_2)}_{=0}h_{2,1}(x)$

$=$

$=h_{0,0}(x)+2h_{1,0}(x)-h_{1,1}(x)$.

where

- $h_{1,1}(x)=L_{1,1}(x)\underset{\substack{L_{i,l}\text{ where}\\ i=1\\ l=1}}{=}\dfrac{(x-x_1)^1}{1!}\prod_{j=0,2}\left(\dfrac{x-x_j}{x_1-x_j}\right)^{m_j}=(x-x_1)\left(\dfrac{x-x_0}{x_1-x_0}\right)^1\left(\dfrac{x-x_2}{x_1-x_2}\right)^2$

  $=(x-1)\left(\dfrac{x-0}{1-0}\right)^1\left(\dfrac{x-3}{1-3}\right)^2=(x-1)x^1\dfrac{(x-3)^2}{4}$

- $h_{0,1}(x)=L_{0,1}(x)\underset{\substack{L_{i,l}\ i=0\\ l=1}}{=}\dfrac{(x-x_0)^1}{1!}\prod_{j=1,2}\left(\dfrac{x-x_j}{x_0-x_j}\right)^{m_j}=(x-x_0)\left(\dfrac{x-x_1}{x_0-x_1}\right)^2\left(\dfrac{x-x_2}{x_0-x_2}\right)^2=$

  $=(x-0)\left(\dfrac{x-1}{0-1}\right)^2\left(\dfrac{x-3}{-3}\right)^2=\dfrac{x(x-1)^2(x-3)^2}{9}$

- $h_{0,0}=L_{0,0}(x)\underset{\substack{L_{i,l}\ i=0\\ l=0}}{=}\prod_{j=1,2}\left(\dfrac{x-x_j}{x_0-x_j}\right)^{m_j}=\left(\dfrac{x-x_1}{x_0-x_1}\right)^{m_1}\left(\dfrac{x-x_2}{x_0-x_2}\right)^{m_2}=\left(\dfrac{x-1}{0-1}\right)^2\left(\dfrac{x-3}{-3}\right)^2=$

  $=\dfrac{(x-1)^2(x-3)^2}{9}$

- $h_{1,0}(x)=L_{1,0}(x)-L'_{1,0}(x_1)\,h_{11}(x)$

$L_{1,0}(x)=\prod_{j=0,2}\left(\dfrac{x-x_j}{x_1-x_j}\right)^{m_j}=\left(\dfrac{x-x_0}{x_1-x_0}\right)^{m_0}\left(\dfrac{x-x_2}{x_1-x_2}\right)^{m_1}=\left(\dfrac{x}{1-0}\right)^1\left(\dfrac{x-3}{1-3}\right)^2=\dfrac{x(x-3)}{4}$

$L'_{1,0}(x)=\dfrac{1}{4}\left[(x-3)^2+x\,2x\right]=\dfrac{1}{4}\left[(x-3)^2+2x^2\right]$

$$L'_{1,0}(x_1) = L'_{1,0}(1) = \frac{1}{4}\left[(1-3)^2 + 2\cdot 1^2\right] = \frac{1}{4}\left(2^2+2\right) = \frac{3}{2}$$

$$h_{1,0}(x) = L_{1,0}(x) - L'_{1,0}(x_1)\, h_{11}(x)$$

$$= \frac{x(x-3)^2}{4} - \frac{3}{2}(x)(x-1)\frac{(x-3)^2}{4} = \frac{x(x-3)^2}{4}\left[1-\frac{3}{2}(x-1)\right] =$$

$$= \frac{x(x-3)^2(5-3x)}{8}$$

then:

$$H(x) = h_{0,0}(x) + 2h_{1,0}(x) - h_{11}(x)$$

$$= \frac{(x-1)^2(x-3)^2}{9} + 2\,\frac{x(x-3)^2(5-3x)}{8} - \frac{x(x-1)(x-3)^2}{4}$$

8

en

$$f(2) \simeq H(2) = \frac{1}{9} + \frac{1}{4}2(-1) - \frac{2}{4} = \frac{1}{9} - \frac{1}{2} - \frac{1}{2} = \frac{1}{9} - 1 = -\frac{8}{9}$$

b) Estimate the maximum possible error of the answer given in a)
if $\begin{cases} f \in C^5[0,3] \\ |f^{(5)}(x)| \le M \text{ on } [0,3]. \end{cases}$

Theorem:

Let $x_0,\dots,x_k$ distinct notes.
$m_0,\dots,m_k$ integer ≥ 1, $\sum_{i=0}^{k} m_i = n+1$

Put $\alpha = k + \sum_{i=0}^{k}(m_i-1)$

Then if $f \in C^{\alpha+1}(a,b)$, we need $H \in P_\alpha$ and.

$$|f(x) - H_{(\alpha)}(x)| \le \frac{\|f^{(\alpha+1)}\|}{(\alpha+1)!}(x-x_0)^{m_0}(x-x_1)^{m_1}\dots(x-x_k)^{m_k}.$$

✓

With our problem $k=2$ $\quad x_0=0 \quad x_1=1 \quad x_2=3. \quad \alpha = 2+(1-1)+(2-1)+(2-1)$
$\quad m_0=1 \quad m_1=2 \quad m_2=2. \quad = 4$

Then if $f \in C^{\alpha+1}[0,3] = C^5[0,3]$, we have

$$|f(x) - H_4(x)| \le \frac{\|f^{(5)}\|}{5!}3^1 3^2 3^2 = \frac{M\,3^5}{5!}$$ ✓

3> Max flat filter

a> Find a third degree polynomial $H(x)$ which satisfies the conditions

$H(0)=1$, $H'(0)=0$, $H(1)=0$, $H'(1)=0$

b> Find a polynomial H of degree $(2p-1)$ which satisfies $\quad n = 2p-1$

$$\begin{cases} H^{(l)}(0) = \delta(l) & 0 \le l < p \\ H^{(l)}(1) = \delta(l) & 0 \le l < p \end{cases} \quad \text{where} \begin{cases} \delta(0) = 1 \\ \delta(l) = 0 \text{ for } k \neq 0. \end{cases}$$

a> $$H(x) = \underbrace{f(x_0)}_{=1} h_{0,0}(x) + \underbrace{f'(x_0)}_{=0} h_{0,1}(x) + \underbrace{f(x_1)}_{=0} h_{1,0}(x) + \underbrace{f'(x_1)}_{=0} h_{11}(x)$$

$$= h_{0,0}(x) = \frac{(x-x_1)^2(2x+x_1-3x_0)}{(x_1-x_0)^2} = (1-x)^2(1+2x)$$

b> With this problem, we want to find a polynomial H that interpolates two note $x_0 = 0$, $x_1 = 1$ that satisfis .

x_i \ H	$x_0 = 0$	$x_1 = 1$
$f(x_i)$	1	1
$f'(x_i)$	0	0
$\vdots$		
$f^{(p-1)}(x_i)$	0	0
	$m_0 = (p-1)$	$m_1 = (p-1)$

Then

$$H(x) = \sum_{i=0}^{k} \sum_{l=0}^{m_i-1} f^{(l)}(x_1)\, h_{il}(x)$$

$$= \sum_{i=0}^{1} f(x_i) h_{i,0}(x) + \underbrace{\sum_{i=0}^{1} f'(x_1) h_{i,1}(x) + \cdots + \sum_{i=0}^{1} f^{(p-1)}(x_1) h_{i,p1}(x)}_{=0 \text{ because } f^{(l)}(x_1) = \; \forall l = \overline{1,p-1},\; i = \overline{0,1}}$$

$$= h_{0,0}(x) + h_{1,0}(x).$$

* Now we want to compute $h_{0,m_0-1}(x)$

$$h_{0,m_0-1}(x) = L_{0,m_0-1} = L_{0,m_0-1}$$

$$L_{0,m_0-1}(x) \underset{\substack{i=0 \\ l=m_0-1.}}{=} \cancel{(x} \left(\frac{x-x_i}{l!}\right)^{l} \prod_{\substack{j=0 \\ j\neq i}}^{k} \left(\frac{x-x_j}{x_i-x_j}\right)^{m_j} \underset{\substack{i=0 \\ l=m_0-1 \\ =p-2 \\ m_1=p-1}}{=} \frac{(x-x_0)^{p-2}}{(p-2)!} \left(\frac{x-x_1}{x_0-x_1}\right)^{(p-1)} \underset{\substack{x_0=0 \\ x_1=1 \\ m_1=p-1}}{=}$$

$$= \frac{(x)^{p-2}}{(p-2)!} \frac{(x-1)^{p-1}}{(-1)^{p-1}}.$$

9

$* \; h_{0,m_0-2}$

$$\underbrace{h_{0,m_0-2}}_{m=m_0-2} = L_{0,m_0-2} - \sum_{\theta=m+1}^{m_0-1} L_{i,m}^{(\theta)} \, h_{i,\theta}(x) = L_{0,m_0-2} - L_{0,m_0-2}^{(m_0-1)}(x_0) \, h_{0,m_0-1}(x$$

- Note that

$$L_{0,m} = \frac{(x-x_0)^m}{(m)!} \left(\frac{x-x_1}{x_i-x_1}\right)^{m_1} = \frac{(x-x_0)^m (x-x_1)^{p-1}}{m! \; (x_0-x_1)^{p-1}}.$$

4) (Node polynomial for Chebyshev points.)

Show that $p(x) = 2^{-n}\left(T_{n+1}(x) - T_{n-1}(x)\right)$, $n \geq 1$.

is the unique monic polynomial in $\mathbb{P}_{n+1}$

with zeros at the $n+1$ Chebyshev points $x_j = \cos\left(\frac{j\pi}{n}\right)$, $0 \leq j \leq n$

which are the points at which are the points at which $T_n(x_j) = 0$

* $T_n(x) = \cos(n \arccos x) = \cos(n\theta)$ where $\theta = \arccos x \Leftrightarrow x = \cos\theta$.

Then we have

$$2^{-n}\left(T_{n+1}(x) - T_{n-1}(x)\right) = 2^{-n}\left[\cos[(n+1)x] - \cos[(n-1)x]\right]$$

$$= 2^{-n}(-2)\sin\left(\frac{(n+1)\theta + (n-1)\theta}{2}\right)\sin\left(\frac{(n+1)\theta - (n-1)\theta}{2}\right) \quad \checkmark$$

$$= -(2)^{-n+1}\sin(n\theta)\sin(\theta)$$

* Then the solution are

$$\left[\begin{array}{l}\sin(n\theta) = 0 \\ \sin\theta = 0\end{array}\right. \Leftrightarrow \left[\begin{array}{l}n\theta = j\pi \\ \theta = k\pi\end{array}\right. \Rightarrow \left[\begin{array}{l}\theta = \frac{j\pi}{n} \\ \theta = k\pi\end{array}\right. \Rightarrow x = \cos(\theta) = \cos\left(\frac{j\pi}{n}\right), 0 \leq j \leq r$$

10

1	10
2	8
3	9
4	10
	37

40 pts.

MAT 683 Homework 3
A. Lutoborski. Syracuse University. Fall 2018

SOLUTIONS

1. Kincaid #14, p 325 Let $p \in \mathbb{P}_{n-1}$ be a polynomial that interpolates $f(x) = \sinh x$ at any set of n nodes in the interval $[-1, 1]$ assuming that one of the nodes is 0. Prove that the error satisfies on $[-1, 1]$ the inequality

$$|f(x) - p(x)| \leq \frac{2^n}{n!}|f(x)|$$

Solution. $\sinh x = \frac{1}{2}(e^x - e^{-x})$. Assume that $x_0 = 0$ then

$$|f(x) - p(x)| \leq \frac{1}{n!}|f^{(n)}(\xi_x)||x| \prod_{i=1}^{n-1} |x - x_i|$$

Since $f^{(n)}(x) = \sinh x$ for n even and $f^{(n)}(x) = \cosh x$ for n odd then

$$|f^{(n)}(x)| \leq \max\{\sinh 1, \cosh 1\} = C = \cosh 1 = 1.5431 \leq 2$$

on $[-1, 1]$. Hence

$$|f(x) - p(x)| \leq \frac{C}{n!}|x|2^{n-1}$$

and

$$\frac{|f(x) - p(x)|}{|f(x)|} \leq \frac{C}{n!}\frac{|x|}{|\sinh x|}2^{n-1}$$

But $\frac{|x|}{|\sinh x|} \leq 1$ on $[-1, 1]$ and

$$|f(x) - p(x)| \leq \frac{2^n}{n!}|f(x)|$$

2. Suppose f is a function on $[0, 3]$ for which one knows that

$$f(0) = 1, \quad f(1) = 2, \quad f'(1) = -1, \quad f(3) = f'(3) = 0.$$

(a) Estimate $f(2)$ using Hermite interpolation
(b) Estimate the maximum possible error of the answer given in (a) if one knows, in addition that $f \in C^3[0, 3]$ and $|f^{(5)}(x)| \leq M$ on $[0, 3]$.

Solution. Let $H \in \mathbb{P}_4$ be the Hermite interpolation polynomial for the above data in Newton's form:

$$\begin{array}{cccccc}
0 & 1 & & & & \\
1 & 2 & 1 & & & \\
1 & 2 & -1 & -2 & & \\
3 & 0 & -1 & 0 & \frac{2}{3} & \\
3 & 0 & 0 & -\frac{1}{2} & \frac{1}{4} & -\frac{5}{36}
\end{array}$$

$$\begin{aligned} H(x) &= 1 + x - 2x(x-1) + \frac{2}{3}x(x-1)^2 - \frac{5}{56}x(x-1)^2(x-3) \\ &= 1 + \frac{49}{12}x - \frac{155}{36}x^2 + \frac{49}{36}x^3 - \frac{5}{36}x^4. \end{aligned}$$

We have $H(2) = \frac{11}{18}$ and to estimate the error we use

$$f(x) - H(x) = x(x-1)^2(x-3)^2 \frac{f^{(5)}(\xi(x))}{5\,!}$$

3. Maxflat filter. Find a third degree polynomial $H(x)$ which satisfies the conditions

$$H(0) = 1, \quad H'(0) = 0, \quad H(1) = 0, \quad H'(1) = 0.$$

(*) Find a polynomial H of degree $2p - 1$ which satisfies

$$\begin{aligned} H^{(k)}(0) &= \delta(k), \qquad 0 \le k < p \\ H^{(k)}(1) &= 0, \qquad 0 \le k < p \end{aligned}$$

where $\delta(0) = 1$, and $\delta(k) = 0$ for $k \neq 0$.

Solution. An elementary way is to find H in Newton's form

$$\begin{array}{ccccc} 0 & 1 & & & \\ 0 & 1 & 0 & & \\ 1 & 0 & -1 & -1 & \\ 1 & 0 & 0 & 1 & 2 \end{array}$$

From it

$$H(x) = 1 + 0(x-0) - 1(x-0)^2 + 2x^2(x-1) = 1 - 3x^2 + 2x^3$$

We may also write

$$H(x) = (1-x)^2(1+2x).$$

What is truly amazing is that

$$(1-x)^{-2} = 1 + 2x + \mathcal{O}(x^2),$$

so that

$$H(x) = (1-x^2)((1-x)^{-2} + \mathcal{O}(x^3)) = 1 + \mathcal{O}(x^2),$$

looking at H in this form allows us to check easily that H satisfies the necessary conditions $H(0) = 1$ and $H'(0) = 0$ at $x = 0$. This the key step in the design of the max-flat filter in signal processing.

(*) We have the binomial series

$$(1-x)^{-p} = \sum_{k=0}^{\infty} \binom{p+k-1}{k} x^k$$

Hence

$$(1-x)^{-p} = \sum_{k=0}^{p-1} \binom{p+k-1}{k} x^k + \mathcal{O}(x^p) = Q(x) + \mathcal{O}(x^p)$$

Define

$$H(x) = (1-x)^{-p}Q(x) = (1-x)^{-p} \sum_{k=0}^{p-1} \binom{p+k-1}{k} x^k$$

Also

$$H(x) = (1-x)^p Q(x) = (1-x)^p((1-x)^{-p} + \mathcal{O}(x^p)) = 1 + \mathcal{O}(x^p)$$

The last formula shows that the interpolation conditions for H at $x = 0$ are satisfied. The penultimate formula shows that the interpolation conditions for H at $x = 1$ are satisfied.

4. Node polynomial for Chebyshev points. Show that

$$p(x) = 2^{-n}(T_{n+1}(x) - T_{n-1}(x)), \qquad n \geq 1$$

is the unique monic polynomial in $\mathbb{P}_{n+1}$ with zeros at the $n+1$ Chebyshev points $x_j = \cos(\frac{j\pi}{n})$, $0 \leq j \leq n$ which are the points at which $T_n(x_j) = 0$.
Solution. Clearly p is monic. We need a trigonometric formula

$$\cos\alpha - \cos\beta = -2\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$$

Let $\alpha = (n+1)\theta$ and $\beta = (n-1)\theta$ where $\theta = \arccos x$. Hence

$$p(x) = 2^{-n}(-2\sin(n\theta)\sin\theta))$$

Clearly $p(0) = p(\pi) = 0$. Next we find the roots of $\sin(n\theta)$, $0 \leq \theta \leq \pi$. $\sin(n\theta) = 0$ when $\theta = \frac{k\pi}{n}$ and $k = 0, \ldots, n$. For $k > n$ one gets $\theta > \pi$.

MAT 683 Exam 1
A. Lutoborski. Syracuse University. Fall 2018

1. (10 pts) Calculate x^{55} in less than 10 multiplications.
2. (15 pts) Evaluate efficiently an odd power polynomial

$$p(x) = a_1x + a_3x^3 + \ldots + a_{2n+1}x^{2n+1}$$

3. (15 pts) The formula $f(N) = N^2$ for $1 \leq N \leq 7$ generates the numbers $1, 4, 9, 16, 25, 36, 49$.
(a) Find a rule $g(N)$ that will generate the same first seven numbers but produce 1 as the eight term.
(b) Find a rule $h(N)$ that inserts 44 in the place of 16 and 36 in the first sequence.

4. (15 pts) Suppose that we want to estimate 1.5! from the values $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$. Find the Newton's formula for cubic Lagrange interpolant of $x!$ and compute $L(1.5)$.
5. (15 pts) Draw on 3 separate figures of planar Bezier curves

$$c(t) = \sum_{k=0}^{3} p_k B_{3,k}(t), \qquad t \in [0,1]$$

where $B_{3,k}(t) = \binom{3}{k} t^k (1-t)^{3-k}$ with control points p_k at the vertices of a unit square:
(a) $p_0 = (0,0)$, $p_1 = (0,1)$, $p_2 = (1,1)$, $p_3 = (1,0)$
(b) $p_0 = (0,0)$, $p_1 = (1,1)$, $p_2 = (0,1)$, $p_3 = (1,0)$
(c) $p_0 = (0,0)$, $p_1 = (0,1)$, $p_2 = (1,0)$, $p_3 = (1,1)$.
Which Bezier curve has a highest midpoint (highest above the x-axis) midpoint?
6. (15 pts) (a) Show that the Chebyshev polynomials have the semigroup property

$$T_m(T_n(x)) = T_{mn}(x), \qquad m, n > 0$$

(b) Show that the equation $T_n(x) = x$ has n roots and find them if $n = 4$.
7. (15 pts) Determine $p \in \mathbb{P}_3$ in

$$s(x) = \begin{cases} p(x) & \text{if } \ 0 \leq x \leq 1 \\ (2-x)^3 & \text{if } \ 1 \leq x \leq 2. \end{cases}$$

such that $s(0) = 0$ and s is a cubic spline in $\mathbb{S}_3^2(\Delta)$ on the subdivision $\Delta = [0,1] \cup [1,2]$ of the interval $[0,2]$. Do you get a natural spline?

MAT 683 Homework 4
A. Lutoborski. Syracuse University. Fall 2018

1. Interpolation by translates of absolute value function. (20 pts)
Let $x_0 < x_1 < \ldots < x_n$ be real numbers. Let f_i, $i = 0, \ldots, n$ be given. Define

$$m_j = \frac{f_j - f_{j-1}}{x_j - x_{j-1}}, \quad 1 \leq j \leq n \qquad m_{n+1} = -m_0 = \frac{f_n + f_0}{x_n - x_0}$$

$$a_j = \frac{m_{j+1} - m_j}{2}, \qquad 0 \leq j \leq n$$

Show that the function $S(x) = \sum_{j=0}^{n} a_j |x - x_j|$ interpolates the data $f_0, \ldots, f_n$ at points $x_0, \ldots, x_n$.

Solution. We want to show that

$$S(x_i) = f_i, \qquad i = 0, \ldots, n.$$

From the third formula

$$\begin{aligned} 2S(x) &= \sum_{j=0}^{n} (m_{j+1} - m_j)|x - x_j| \\ &= \sum_{j=1}^{n+1} m_j |x - x_{j-1}| - \sum_{j=0}^{n} m_j |x - x_j| \\ &= \sum_{j=1}^{n} m_j (|x - x_{j-1}| - |x - x_j|) + m_{n+1}|x - x_n| - m_0 |x - x_0| \end{aligned}$$

For $x = x_0$ we get

$$\begin{aligned} 2S(x_0) &= \sum_{j=1}^{n} m_j (x_{j-1} - x_j) + m_{n+1}(x_n - x_0) \\ &= \sum_{j=1}^{n} (f_{j-1} - f_j) + (f_n + f_0) \\ &= 2f_0 \end{aligned}$$

Similarly we show that $2S(x_n) = 2f_n$.
For $0 < i < n$ we return to the formula

$$2S(x) = \sum_{j=1}^{n} m_j (|x - x_{j-1}| - |x - x_j|) + m_{n+1}|x - x_n| - m_0 |x - x_0|$$

and substitute $x = x_i$.

$$2S(x_i) = \sum_{j=1}^{i} m_j((x_i - x_{j-1}) - (x_i - x_j)) + \sum_{j=i+1}^{n} m_j((x_{j-1} - x_i) - (x_j - x_i))$$
$$+ m_{n+1}(x_n - x_i) - m_0(x_i - x_0) =$$
$$= \sum_{j=1}^{i} m_j(x_j - x_i) - \sum_{j=i+1}^{n} m_j(x_j - x_{j-1}) + m_{n+1}(x_n - x_0)$$
$$= \sum_{j=1}^{i} (f_j - f_{j-1}) - \sum_{j=i+1}^{n} (f_j - f_{j-1}) + m_{n+1}(x_n - x_0)$$
$$= (f_i - f_0) - (f_n - f_i) + (f_n + f_0)$$
$$= 2f_i$$

2.Kincaid $\sharp$ 2 p. 374 (10 pts) Prove that if $t_m \leq x < t_{m+1}$ then

$$\sum_{i=-\infty}^{\infty} c_i B_i^k(x) = \sum_{i=m-k}^{m} c_i B_i^k(x)$$

Solution. supp$B_i^k = [t_i, t_{i+k}]$. Hence if $i \geq m+1$ or if $i + k + 1 \leq m$ then $(t_m, t_{m+1}) \cap \mathrm{supp} B_i^k = \emptyset$.

3.Kincaid $\sharp$ 8 p. 375 (10 pts)

4. Partition of unity (10 pts) Show that

$$\sum_{i=-r}^{n} B_i^r(x) = 1, \qquad x \in (t_0, t_n)$$

Solution. Induction. When $r = 0$ the formula is true. Assume the formula holds for $r - 1$. In order to prove our formula we use the recurrence relation for B splines.

$$\sum_{i=-r}^{n} B_i^r(x) = \frac{x - t_{-r}}{t_0 - t_{-r}} B_{-r}^{r-1}(x) + \sum_{i=-r+1}^{n} B_i^{r-1}(x) + \frac{t_{n+r+1} - x}{t_{n+r+1} - t_{n+1}} B_{n+1}^{r-1}(x)$$

The sum on the left side, upon using the recurrence, telescopes giving the middle sum on the right. The very first and last terms from B_{-r}^r and B_n^r are left as first and third term on the right side. $B_{-r}^{r-1}(x)$ has support in $[t_{-r}, t_0]$ and $B_{n+1}^{r-1}(x)$ has support in $[t_{n+1}, t_{n+r+1}]$ and so both of these functions vanish on (t_0, t_n). By the induction hypothesis $\sum_{i=-r+1}^{n} B_i^{r-1}(x) = 1$.

MAT 683 Exam 1 SOLUTIONS
A. Lutoborski. Syracuse University. Fall 2018

1. (10 pts) Calculate x^{55} in less than 10 multiplications.
Solution. We can compute x^{55} in 8 multiplications. To compute $y = x^5$ we need 3 multiplications: x^2, x^4, x^5. Next we compute $y^{11} = (y^2)^5 y$. For this we need 5 multiplications: 1 mult to compute y^2, 3 mults to compute $(y^2)^5$ and 1 to compute $(y^2)^5 y$.
It is easier to get the result in 9 multiplications: $x^2, x^3, x^6, x^{12}, x^{13} x^{26}, x^{27}, x^{54}, x^{55}$.
2. (15 pts) Evaluate efficiently an odd power polynomial

$$p(x) = a_1 x + a_3 x^3 + \ldots + a_{2n+1} x^{2n+1}$$

Solution. Set $y = x^2$, then

$$p(x) = ((\ldots(a_{2n+1} y + a_{2n-1})y + \ldots + a_3)y + a_1)x$$

3. (15 pts) The formula $f(N) = N^2$ for $1 \leq N \leq 7$ generates the numbers $1, 4, 9, 16, 25, 36, 49$.
(a) Find a rule $g(N)$ that will generate the same first seven numbers but produce 1 as the eight term.
(b) Find a rule $h(N)$ that inserts 44 in the place of 16 and 36 in the first sequence.

Solution. (a)

$$g(N) = N^2 + \frac{(N-1)(N-2)\ldots(N-7)}{5040}(-63)$$

Hence $g(8) = 64 + \frac{5040}{5040}(-63) = 1$.
(b)

$$h(N) = N^2 + \frac{(N-1)(N-2)(N-3)(N-5)(N-6)(N-7)}{-36} \cdot 28$$
$$+ \frac{(N-1)(N-2)(N-3)(N-4)(N-5)(N-7)}{-120} \cdot 8$$

4. (15 pts) Suppose that we want to estimate 1.5! from the values $0! = 1$, $1! = 1$, $2! = 2$, $3! = 6$. Find the Newton's formula for cubic Lagrange interpolant of $x!$ and compute $L(1.5)$.
Solution. We will use the Newton's formula for cubic interpolant $L(x)$ of $x!$.

$$\begin{array}{ccccc}
0 & 1 & & & \\
1 & 1 & & & \\
2 & 2 & 1 & \frac{1}{2} & \\
3 & 6 & 4 & \frac{3}{2} & \frac{1}{3}
\end{array}$$

$$L(x) = 1 + \frac{1}{2}x(x-1) + \frac{1}{3}x(x-1)(x-2), \qquad L(1.5) = 1.25$$

This is simpler than evaluating an integral $\Gamma(2.5) = 1.3293$.

5. (15 pts) Draw on 3 separate figures of planar Bezier curves

$$c(t) = \sum_{k=0}^{3} p_k B_{3,k}(t), \qquad t \in [0,1]$$

where $B_{3,k}(t) = \binom{3}{k} t^k (1-t)^{3-k}$ with control points p_k at the vertices of a unit square:

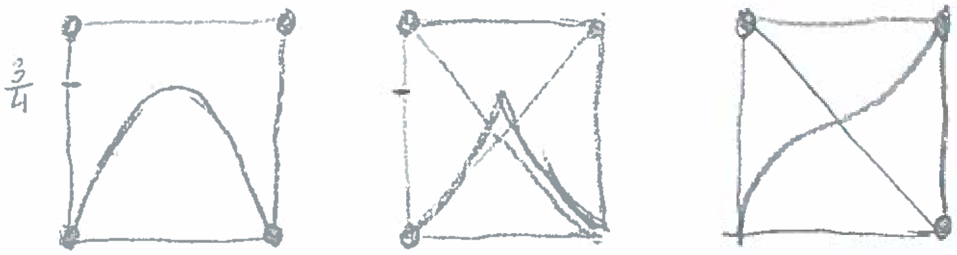
(a) $p_0 = (0,0)$, $p_1 = (0,1)$, $p_2 = (1,1)$, $p_3 = (1,0)$

(b) $p_0 = (0,0)$, $p_1 = (1,1)$, $p_2 = (0,1)$, $p_3 = (1,0)$

(c) $p_0 = (0,0)$, $p_1 = (0,1)$, $p_2 = (1,0)$, $p_3 = (1,1)$.

Which Bezier curve has a highest midpoint (highest above the x-axis) midpoint?

Solution.



$p_0^{(3)} = c(\frac{1}{2}) = (\frac{1}{2}, \frac{3}{4})$ in case (a) and (b).

6. (15 pts) (a) Show that the Chebyshev polynomials have the semigroup property

$$T_m(T_n(x)) = T_{mn}(x), \qquad m, n > 0$$

(b) Show that the equation $T_n(x) = x$ has n roots and find them if $n = 4$.

Solution. (a)

$$T_m(T_n(x)) = \cos(m \arccos(\cos n \arccos x) = \cos(mn \arccos x) = T_{mn}(x)$$

(b) Show that the equation $T_n(x) = x$ has n roots in $[-1, 1]$.

Solution. Converting the equation to trigonometric form

$$\cos(n\theta) - \cos\theta = 0, \qquad \cos\theta = x, \qquad 0 \le \theta \le \pi.$$

But $\cos\alpha - \cos\beta = -\sin\frac{\alpha+\beta}{2}\sin\frac{\alpha-\beta}{2}$ hence the equation is

$$-2\sin\frac{(n+1)\theta}{2}\sin\frac{(n-1)\theta}{2} = 0$$

Solving the equation for all $0 \leq \theta \leq 2\pi$ we get

$$\theta_k = \frac{2\pi}{n+1}k, \qquad k = 0, 1, \ldots, n$$

$$\tilde{\theta}_k = \frac{2\pi}{n-1}k, \qquad k = 0, 1, \ldots, n-2$$

The solutions we obtained are in $[0, 2\pi]$ but θ is restricted to $[0, \pi]$. However $\cos\theta$ is an even function on $[0, 2\pi]$ with respect to the midpoint π. This means that if θ and α are symmetric with respect to the midpoint π and $\theta - \pi = -\alpha + \pi$ then $\cos\theta = \cos\alpha$. Therefore we can restrict θ to $[0, \pi]$ to obtain all possible roots in x.

If n is even we have

$$\theta_k = \frac{2\pi}{n+1}k, \qquad k = 0, 1, \ldots, n/2$$

$$\tilde{\theta}_k = \frac{2\pi}{n-1}k, \qquad k = 0, 1, \ldots, n/2 - 1$$

and the n roots of $T_n(x) = x$ are $x_k = \cos\theta_k$, $\tilde{x}_k = \cos\tilde{\theta}_k$

For n odd

$$\theta_k = \frac{2\pi}{n-1}k, \qquad k = 0, 1, \ldots, (n+1)/2$$

$$\tilde{\theta}_k = \frac{2\pi}{n-1}k, \qquad k = 0, 1, \ldots, (n-3)/2$$

7. (15 pts) Determine $p \in \mathbb{P}_3$ in

$$s(x) = \begin{cases} p(x) & \text{if } \ 0 \leq x \leq 1 \\ (2-x)^3 & \text{if } \ 1 \leq x \leq 2. \end{cases}$$

such that $s(0) = 0$ and s is a cubic spline in $\mathbb{S}_3^2(\Delta)$ on the subdivision $\Delta = [0, 1] \cup [1, 2]$ of the interval $[0, 2]$. Do you get a natural spline?

Solution. We have to find a $p \in \mathbb{P}_3$ such that

$$p(0) = 0, \quad p(1) = 1, \quad p'(1) = s'(1) = -3, \quad p''(1) = s''(1) = 6.$$

The above four Hermite interpolation conditions determine p uniquely. Let $p(x) = x(ax^2 + bx + c)$ so that $p(0) = 0$. The remaining three conditions are:

$$\begin{aligned} a + b + c &= 1, \\ 3a + 2b + c &= -3 \\ 3a + b &= 3 \end{aligned}$$

which gives $a = 7$, $b = -18$, $c = 12$ so

$$p(x) = 7x^3 - 18x^2 + 12x.$$

Since $p''(0) = -36$ s is not a natural spline.

Tianle

MAT 683 Homework 4
A. Lutoborski. Syracuse University. Fall 2018

1. Interpolation by translates of absolute value function. (20 pts) Let $x_0 < x_1 < \ldots < x_n$ be real numbers. Let f_i, $i = 0, \ldots, n$ be given. Define

$$m_j = \frac{f_j - f_{j-1}}{x_j - x_{j-1}}, \quad 1 \le j \le n \qquad m_{n+1} = -m_0 = \frac{f_n + f_0}{x_n - x_0}$$

$$a_j = \frac{m_{j+1} - m_j}{2}, \qquad 0 \le j \le n$$

Show that the function $S(x) = \sum_{j=0}^{n} a_j |x - x_j|$ interpolates the data $f_0, \ldots, f_n$ at points $x_0, \ldots, x_n$.

2.Kincaid ♯ 2 p. 374 (10 pts)

3.Kincaid ♯ 8 p. 375 (10 pts)

4.Kincaid ♯ 28 p. 376 (10 pts)

1	8
2	10
3	10
4	0
	28

1, Interpolation by translates of absolute value function.

Let $x_0 < x_1 < \dots < x_n$ be real numbers
$f_1 \quad f_2 \quad f_n$ be given.

Define
$$\begin{cases} m_j = \dfrac{f_j - f_{j-1}}{x_j - x_{j-1}}, \; 1 \le j \le n \\ m_{n+1} = -m_0 = \dfrac{f_n + f_0}{x_n - x_0} \end{cases}$$

$$a_j = \frac{m_{j+1} - m_j}{2}, \; 0 \le j \le n$$

Show that the function $S(x) = \sum_{j=0}^{n} a_j |x - x_j|$ interpolates the data $f_0, \dots, f_n$ at points $x_0, \dots, x_n$

* I have tried many ways to solve this problem, like:

  1, induction

  2, Use the property that the equation of a line that goes through (x_a, f_a) and (x_b, f_b) has equation:

$$f(x) = f_a \frac{x - x_b}{x_a - x_b} + f_b \frac{x - x_a}{x_b - x_a}$$

I finally came up with an answer by writeingdown the system of equations to find $\vec{a} = [a_1, a_2, \dots, a_n]^T$ and use Gaussian elimination to solve it. But I think this is quite a boring and tedious way. It would be nice if I knew another way to do it. I think substituding may work.

* We have $S(x) = \sum_{j=0}^{n} a_j |x - x_j|$. We want to find $\vec{a} = [a_0, \dots, a_n]^T$ that satisfies

$$\begin{cases} S(x_0) = \underbrace{a_0(x_0 - x_0)}_{=0} + a_1(x_1 - x_0) + a_2(x_2 - x_0) + \dots + a_{n-1}(x_{n-1} - x_0) + a_n(x_n - x_0) = f_0 \\ S(x_1) = a_0(x_1 - x_0) + 0 + a_2(x_2 - x_1) + \dots + a_{n-1}(x_{n-1} - x_1) + a_n(x_n - x_1) = f_1 \\ S(x_2) = a_0(x_2 - x_0) + a_1(x_2 - x_1) + 0 + \dots + a_{n-1}(x_{n-1} - x_2) + a_n(x_n - x_2) = f_2 \\ \vdots \\ S(x_{n-1}) = a_0(x_{n-1} - x_0 + a_1(x_{n-1} - x_1)\; a_2(x_{n-1} - x_2) + \dots + 0 \quad a_n(x_n - x_{n-1}) = f_{n-1} \\ S(x_n) = a_0(x_n - x_0) + a_1(x_n - x_1) + a_2(x_n - x_2) + \dots + a_{n-1}(x_n - x_{n-1}) + 0 = f_n \end{cases}$$

we have

$$\left[\begin{array}{cccccccc|c|l}
) & x_1-x_0 & x_2-x_0 & x_3-x_0 & x_4-x_0 & x_{n-2}-x_0 & x_{n-1}-x_0 & x_n-x_0 & f_0 & \text{Row 1} \\
{}_1-x_0 & 0 & x_2-x_1 & x_3-x_1 & x_4-x_1 & x_{n-2}-x_1 & x_{n-1}-x_1 & x_n-x_1 & f_1 & \text{Row 2} \\
{}_2-x_0 & x_2-x_1 & 0 & x_3-x_2 & x_4-x_2 & x_{n-2}-x_2 & x_{n-1}-x_2 & x_n-x_2 & f_2 & \text{Row 3} \\
-x_0 & x_3-x_1 & x_3-x_2 & 0 & x_4-x_3 & x_{n-2}-x_3 & x_{n-1}-x_3 & x_n-x_3 & f_3 & \text{Row 4} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \\
{}_{n-2}-x_0 & x_{n-2}-x_1 & x_{n-2}-x_2 & x_{n-2}-x_3 & x_{n-2}-x_4 & 0 & x_{n-1}-x_{n-2} & x_n-x_{n-2} & & \\
{}_{-1}-x_0 & x_{n-1}-x_1 & x_{n-1}-x_2 & x_{n-1}-x_3 & x_{n-1}-x_4 & x_{n-1}-x_{n-2} & 0 & x_n-x_{n-1} & f_{n-1} & \text{Rn} \\
-x_0 & x_n-x_1 & x_n-x_2 & x_n-x_3 & x_n-x_4 & x_n-x_{n-2} & x_n-x_{n-1} & 0 & f_n. & \text{Row}_{n+}
\end{array}\right]$$

$(*)$

$$\left[\begin{array}{ccccccccc|c|l}
) & x_1-x_0 & x_2-x_0 & x_3-x_0 & x_4-x_0 & --- & x_{n-2}-x_0 & x_{n-1}-x_0 & x_n-x_0 & f_0 & \\
-x_0 & -(x_1-x_0) & -(x_1-x_0) & -(x_1-x_0) & -x_1-x_0 & --- & -(x_1-x_0) & -(x_1-x_0) & -(x_1-x_0) & f_1-f_0 & \text{Row 2}-\text{Ro} \\
-x_1 & x_1-x_2 & -(x_2-x_1) & -(x_2-x_1) & -(x_2-x_1) & --- & -(x_2-x_1) & -(x_2-x_1) & -(x_2-x_1) & f_2-f_1 & \text{Row 3}-\text{Ro} \\
-x_2 & x_3-x_2 & x_3-x_2 & -(x_3-x_2) & -(x_3-x_2) & -- & -(x_3-x_2) & -(x_3-x_2) & -(x_3-x_2) & f_3-f_2 & \text{Row 4}-\text{Ro} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\
{}_{-1}-x_{n-2} & x_{n-1}-x_{n-2} & x_{n-1}-x_{n-2} & x_{n-1}-x_{n-2} & x_{n-1}-x_{n-2} & --- & x_{n-1}-x_{n-2} & -(x_{n-1}-x_{n-2}) & -(x_{n-1}-x_{n-2}) & f_n-f_{n-2} & \text{Row}_n-\text{R}_{n-} \\
-x_{n-1} & x_n-x_{n-1} & x_n-x_{n-1} & x_n-x_{n-1} & x_n-x_{n-1} & -- & x_n-x_{n-1} & x_n-x_{n-1} & -(x_n-x_{n-1}) & f_n-f_{n-1} & \text{Row}_{n+1}-\text{Ro}
\end{array}\right]$$

$$\begin{array}{c|cccccccc|c}
R_1 & 0 & x_1-x_0 & x_2-x_0 & x_3-x_0 & x_4-x_0 & x_{n-2}-x_0 & x_{n-1}-x_0 & x_n-x_0 & f_0 \\
R_2 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & \frac{f_1-f_0}{(x_1-x_0)} \\
R_3 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & \frac{f_2-f_1}{(x_2-x_1)} \\
R_4 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & \frac{f_3-f_2}{x_3-x_2} \\
 & & & & & (**) & & & & \\
R_n & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & \\
R_{n+1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & \frac{f_n-f_{n-}}{x_n-x_{n-}}
\end{array}$$

$$\begin{array}{c|ccccccccc|c}
R_1 & 0 & x_1-x_0 & x_2-x_0 & x_3-x_0 & x_4-x_0 & & x_{n-2}-x_0 & x_{n-1}-x_0 & x_n-x_0 & f_0 \\
R_2 & 1 & -1 & -1 & -1 & -1 & & -1 & -1 & -1 & \frac{f_1-f_0}{x_1-x_0} \\
R_3-R_2 & 0 & 2 & 0 & 0 & 0 & & 0 & 0 & 0 & \overset{m_2}{\frac{f_2-f_1}{x_2-x_1}} - \frac{f_1}{x_1} \\
R_4-R_3 & 0 & 0 & 2 & 0 & 0 & & 0 & 0 & 0 & \overset{m_3}{\frac{f_3-f_2}{x_3-x_2}} - \frac{f_2}{x} \\
 & & & & & (***) & & & & & \\
 & 0 & 0 & 0 & 0 & 0 & ---- & 0 & 2 & 0 & \overset{m_n}{\frac{f_n-f_{n-1}}{x_n-x_{n-1}}} - \overset{m}{\frac{f_{n-1}}{x_{n-1}}}
\end{array}$$

$a_1 = \frac{m_2-m_1}{2}$

$a_2 = \frac{m_3-m_2}{2}$

$\vdots$

$a_{n-1} = \frac{m_n-m_{n-2}}{2}$

So, up to now, we have proved that

$$a_i = \frac{m_i - m_{i-1}}{2}, \ i = \overline{1, n-1}$$

We now need to find a_0 and a_n

this was given to us.

We need $S(x_i)=f_i$

yes this is very tough

nsider the system (*) from the last 2 page, take Row_{n+1} + Row 1, we have

$$\begin{bmatrix} x_n - x_0 & x_n - x_0 & x_n - x_0 & x_n - x_0 & x_n - x_0 & & x_n - x_0 & \vdots & f_n + f_0 \end{bmatrix}$$

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & & 1 & | & \dfrac{f_n + f_0}{x_n - x_0} \end{matrix} \quad (****$$

en subtract this row by, the last row in (**), we have

$$\left[\begin{matrix} 0 & 0 & 0 & \dots & - - - - & 2 & \overset{m_{n+1}}{\dfrac{f_n + f_0}{x_n - x_0}} - \overset{m_n}{\dfrac{f_n - f_{n-1}}{x_n - x_{n-1}}} \end{matrix} \right.$$

$$\Rightarrow a_n = \frac{m_{n+1} - m_n}{2} \quad \square \text{ for } a_n$$

Now find a_0

dd the row (****) with the second ~~of~~ row of the system (***)

$$\Rightarrow \left[\begin{matrix} 2 & 0 & 0 & 0 & \dots - - & 0 & \overset{-m_0}{\dfrac{f_n + f_0}{x_n - x_0}} + \overset{m_1}{\dfrac{f_1 - f_0}{x_1 - x_0}} \end{matrix} \right.$$

$$\Rightarrow a_0 = \frac{m_1 - m_0}{2} \quad \square \text{ for } a_0$$

Q7 Kincaid #2 p 374 (10 points).

Prove that if $t_m \leq x < t_{m+1}$, then

$$\sum_{i=-\infty}^{+\infty} c_i B_i^k(x) = \sum_{i=m-k}^{m} c_i B_i^k(x) .$$

* We have a property that the support of $B_i^k(x)$ is (t_i, t_{i+k+1})

$\Rightarrow$ the sum $\sum_{i=-\infty}^{+\infty} c_i B_i^\ell(x)$ only have some finitely elements that are nonzero

* Consider when $t_m \leq x < t_{m+1}$, this is the intersection of the supports of $B_{m-\ell}^\ell, B_{m-k+1}^k, \dots, B_m^k$

When $j \notin \{m-k, m-\ell+1, .., m\}$ then $B_j^\ell(x) = 0$ when $t_m \leq x < t_{m+1}$.

So we have $\sum_{i=-\infty}^{+\infty} c_i B_i^\ell(x) = \sum_{i=m-k}^{m} c_i B_i^\ell(x)$ □.

✓

Kincard #8 P375 (*)

Prove that if $\sum_{i=-\infty}^{\infty} c_i B_i^\ell(x) = 0, \forall x$ then $c_i = 0$ for all i

This answer uses the Lemma 1 from your note :

$\underbrace{\{B_j^k, \dots, B_{j+k}^k\}}$ is linearly independent on the interval (t_{j+k}, t_{j+k+1}).

he set of $(k+1)$ Bsplines of degree k

$$\left.\begin{array}{l}\text{We have that } (*) \text{ is true for all } x \\ \{[t_m, t_{m+1})\}_m \text{ creates a partition for } \mathbb{R}\end{array}\right\} \Rightarrow (*) \text{ is true for any } [t_m, t_{m+1})$$

From problem 2,

$$\left.\begin{array}{l}\displaystyle\sum_{i=m-\ell}^{m} c_i B_i^\ell(x) = \sum_{i=-\infty}^{+\infty} c_i B_i^\ell(x) = 0 \quad \text{for } x \in [t_m, t_{m+1}) \\ \text{rom Lemma 1, } \{B_{m-\ell}^\ell, \dots, B_m^k\} \text{ is linearly independent in } [t_m, t_{m+1})\end{array}\right\} \Rightarrow$$

⇒ So we have $c_{m-\ell} = \dots = c_m = 0$

Since $(*)$ is true for all $[t_m, t_{m+1})$ ⇒ $c_{m-\ell} = \dots = c_m = 0, \forall m$

which means $c_i = 0, \forall i$ □3.

47 Kincaid #28 p376

Prove that $\sum_{i=0}^{n} B_i^k(x) = 1 \qquad t_k \le x \le t_{k+n}$

* Consider when $x \in [t_k, t_{k+n}]$, this is the intersection of the supports of $B_0^\ell(x), B_1^\ell(x), \dots, B_n^\ell(x)$.

Since $B_0^\ell(x)$ has support is (t_0, t_{0+k+1})
$B_1^\ell(x)$ has support is (t_1, t_{1+k+1})
$\vdots$
$B_n^\ell(x)$ has support is (t_n, t_{k+n+1})

0

So we have $\sum_{i=0}^{n} B_i^\ell(x) = \sum_{i=-\infty}^{+\infty} B_i^\ell(x) \underset{\uparrow}{=} 1$

a property that we have learned in class.

Pbm from Kincaid is wrong:

take $t_K = K \quad r=2 \quad n=2$

to obtain contradiction.

1	8
2	10
3	10
4	0
	28

Tian Le

MAT 683 Homework 5
A. Lutoborski. Syracuse University. Fall 2018

Least squares approximation.

3. Gaussian Quadrature Error Estimate. Let $Q(f) = \sum_{j=0}^{k} \lambda_j f(x_j)$ be the Gaussian quadrature on $(-1,1)$ with weight $w(x) \equiv 1$ such that

$$Q(p) = \int_{-1}^{1} p(x)\,\mathrm{d}x \qquad \forall p \in \mathbb{P}_{2k+1}.$$

Show that

$$\left| Q(f) - \int_{-1}^{1} f(x)\,\mathrm{d}x \right| \leq 4 \inf_{p \in \mathbb{P}_{2k+1}} \Big(\sup_{-1 \leq x \leq 1} |f(x) - p(x)| \Big)$$

2. Approximate x^2 in $L^2(0,1)$ by a combination of $1, x$ and by by a combination of x^{100}, x^{101}. Which approximation gives a smaller approximation error? Explain the reasons. Plot the approximations on the same graph.

3. Least squares regression line from bivariate data. In statistics the least squares regression line for predicting y from x is given by $y = bx + a$ where

$$b = r\frac{s_y}{s_x}, \qquad a = \bar{y} - b\bar{x}.$$

and

$$\bar{x} = \frac{x_1 + \cdots + x_m}{m}, \quad s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_m - \bar{x})^2}{m},$$

$$r = \frac{\frac{1}{m}\left((x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_m - \bar{x})(y_m - \bar{y})\right)}{s_x s_y}.$$

are respectively the mean, variance and correlation. Show how a and b are obtained via least squares approximation.

4. (Kincaid p 404,#3)

1) Gaussian Quadrature Error Estimate

*Homework Tran Le

Let $Q(f) = \sum_{j=0}^{l} \lambda_j f(x_j)$ be the Gaussian quadrature on $(-1,1)$ with weight $w(x) \equiv 1$ such that $Q(p) = \int_{-1}^{1} p(x)\,dx \quad \forall p \in \mathbb{P}_{2l+1}$

Show that: $$\left| Q(f) - \int_{-1}^{1} f(x)\,dx \right| \leq 4 \inf_{p \in \mathbb{P}_{2l+1}} \left(\sup_{-1 \leq x \leq 1} |f(x) - p(x)| \right)$$

Weierstrass th says that a continuous f can be arbitrarily well approximated b a polynomial of per ver hi d

* We have by Weirtioos theorem $\exists\, p \in \mathbb{P}_{2l+1}$ such that $\|f - g\|_\infty \leq \varepsilon$ and

$$\left| Q(f) - \int_{-1}^{1} f(x)\,dx \right| \leq \underbrace{\left| Q(f) - Q(p) \right|}_{(1)} + \underbrace{\left| Q(p) - \int_{-1}^{1} p(x)\,dx \right|}_{(2)} + \underbrace{\left| \int_{-1}^{1} p(x)\,dx - \int_{-1}^{1} f(x)\,dx \right|}_{(3)}$$

$$(1) = |Q(f) - Q(p)| = \left| \sum_{i=0}^{l} \lambda_i f(x_i) + \sum_{i=0}^{l} \lambda_i p(x_i) \right| \leq \sum_{i=0}^{l} \lambda_i |f(x_i) - p(x_i)|$$

$$\leq \varepsilon \underbrace{\sum_{i=0}^{l} \lambda_i (1)}_{= \int_{-1}^{1} 1\,dx \text{ since it is exact for degree } 0} = \varepsilon \int_{-1}^{1} dx = 2\varepsilon$$

✓

$$(2) = \left| Q(p) - \int_{-1}^{1} p(x)\,dx \right| = 0 \quad \text{since the assumption}$$

$$(3) = \left| \int_{-1}^{1} p(x)\,dx - \int_{-1}^{1} f(x)\,dx \right| \leq \int_{-1}^{1} |p(x) - f(x)|\,dx \leq \varepsilon \int_{-1}^{1} dx = 2\varepsilon$$

$$\Rightarrow \left| Q(f) - \int_{-1}^{1} f(x)\,dx \right| \leq 2\varepsilon + 2\varepsilon = 4\varepsilon = 4 \inf_{p \in \mathbb{P}_{2l+1}} \left(\sup |f(x) - p(x)| \right) \quad \square$$

2> Correction:

Approximate x^2 in $L^2(0,1)$ by a combination of $1, x$

and by a combination of x^{100}, x^{101}.

which approximation gives a smaller approximation error. Explain.

* Find the approximation of x^2 in $L^2(0,1)$ by a combination of 1 and x

Put $\phi_1 = 1$ in $L^2(0,1)$

$\phi_2 = x$ in $L^2(0,1)$

According to the orthogonal projection theorem, we have

$f^* = c_1\phi_1 + c_2\phi_2$ where

$$\begin{cases} \langle f^*, \phi_1\rangle = \langle f, \phi_1\rangle \\ \langle f^*, \phi_2\rangle = \langle f, \phi_2\rangle \end{cases} \Leftrightarrow \begin{cases} \langle c_1\phi_1 + c_2\phi_2, \phi_1\rangle = \langle x^2, \phi_1\rangle \\ \langle c_1\phi_1 + c_2\phi_2, \phi_2\rangle = \langle x^2, \phi_2\rangle \end{cases}$$

$$\Leftrightarrow \left\{ \begin{bmatrix} \langle\phi_1,\phi_1\rangle & \langle\phi_2,\phi_1\rangle \\ \langle\phi_1,\phi_2\rangle & \langle\phi_2,\phi_2\rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \langle x^2,\phi_1\rangle \\ \langle x^2,\phi_2\rangle \end{bmatrix} \right.$$

- $\langle\phi_1,\phi_1\rangle = \int_0^1 1\,dx = 1 \qquad \langle\phi_2,\phi_1\rangle = \langle\phi_1,\phi_2\rangle = \int_0^1 1x\,dx = \frac{x^2}{2}\Big|_0^1 = \frac{1}{2}$

$\langle x^2,\phi_1\rangle = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3} \qquad \langle x^2,\phi_2\rangle = \int_0^1 x^3\,dx = \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$.

$$\Rightarrow \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/4 \end{pmatrix} \qquad \langle\phi_2,\phi_2\rangle = \int_0^1 x^2 = \frac{1}{3}.$$

$\Rightarrow c_1 = -1/6$ and $c_2 = 1$.

Then $f^*(x) = -\frac{1}{6} + x$

* The error:

$$\|f - f^*\|^2_{L^2(0,1)} = \left[\int_0^1 \left(x^2 - x + \frac{1}{6}\right)^2 dx\right] = \left(\frac{x^3}{3} - \frac{x^2}{2} + \frac{x}{6}\right)^2\Big|_0^1 =$$

$$= \frac{1}{3} \cancel{- \frac{1}{2} + \frac{1}{6} =}$$

2) Approximate x^2 in $L^2(0,1)$ by a combination of $1, x$ and by a combination of x^{100}, x^{101}

Which approximation gives a smaller approximation error ? Explain

Plot the approximation on the same graph

Put $f(x) = x^2$

* We first want to find the approximation of f by a combination of $e_1(x)=1$ $e_2(x)=$

$$f^*_e(x) = \frac{\langle f,e_1\rangle}{\langle e_1,e_1\rangle} e_1(x) + \frac{\langle f,e_2\rangle}{\langle e_2,e_2\rangle} e_2(x)$$

not orthog

$$\langle e_1,e_1\rangle = \int_0^1 1\,dx = 1 \qquad \langle f,e_1\rangle = \int_0^1 x^2\,dx = \frac{x^3}{3}\Big|_0^1 = \frac{1}{3}$$

$$\langle e_2,e_2\rangle = \int_0^1 x^2\,dx = \frac{1}{3} \qquad \langle f,e_2\rangle = \int_0^1 x^3\,dx = \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$$

NOT correct

Then $f^*_e(x) = \frac{1}{3} + \frac{\frac{1}{4}}{\frac{1}{3}}x = \begin{cases} \left(\frac{1}{3} + \frac{3}{4}x\right) ? & x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$

* Now we want to find the approximation of f by a combination of $k_1(x) = x^{100}$ and $k_2(x) = x$

$$f^*_k(x) = \frac{\langle f,k_1\rangle}{\langle k_1,k_1\rangle} k_1(x) + \frac{\langle f,k_2\rangle}{\langle k_2,k_2\rangle} k_2(x)$$

not orthogo

$$\langle k_1,k_1\rangle = \int_0^1 x^{200}\,dx = \frac{x^{201}}{201}\Big|_0^1 = \frac{1}{201} \qquad \langle k_2,k_2\rangle = \int_0^1 x^{202} = \frac{1}{203}$$

$$\langle f,k_1\rangle = \int_0^1 x^{102}\,dx = \frac{1}{103} \qquad \langle f,k_2\rangle = \int_0^1 x^{103} = \frac{1}{104}$$

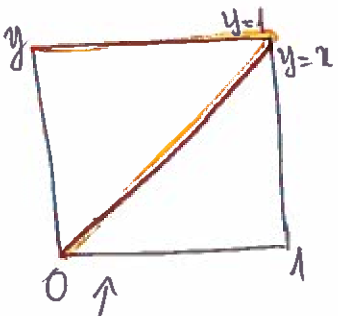
not correc

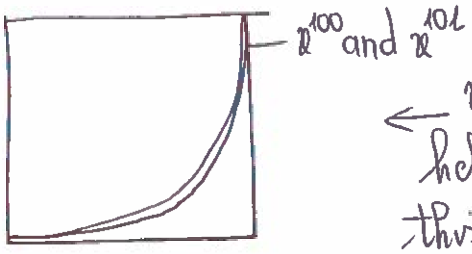
then $f^*_k(x) = \frac{\frac{1}{103}}{\frac{1}{201}} x^{100} + \frac{\frac{1}{104}}{\frac{1}{203}} x^{101} = \begin{cases} \frac{201}{103}x^{100} + \frac{203}{104}x^{101}, & x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$

Which approximation gives smaller error ?

The approximation of $f = x^2$ onto $e = \text{span}\{1, x\}$ gives smaller error than the approximation of f onto $\ell = \text{span}\{x^{100} \text{ and } x^{101}\}$.

ince we have





← Your explaination helps me find out this really interesting observe.

the space e spanned by $\{1, x\}$ can capture more of x^2

By computing the error

$$\|f - f^*_e\|^2 = \int_0^1 \left[x^2 - \left(\frac{1}{3} + \frac{3}{4}x\right)\right]^2 dx$$

nd

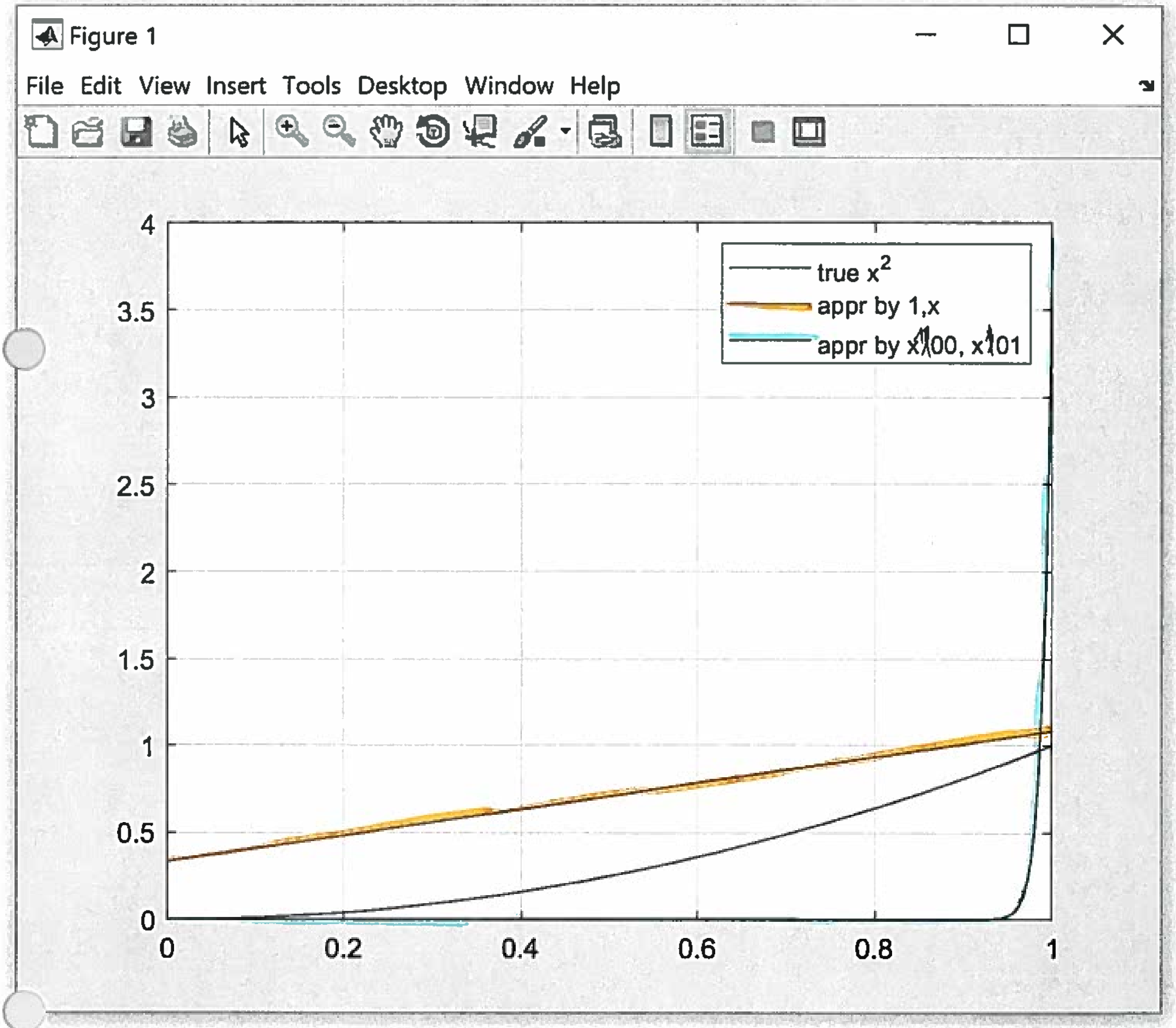
$$\|f - f^*_\ell\|^2 = \int_0^1 \left[x^2 - \left(\frac{201}{103}x^{100} + \frac{203}{104}x^{101}\right)\right]^2 dx$$

hen since $x \in (0,1)$ $\quad \frac{201}{103}x^{100} \ll \frac{1}{3}$

$$\frac{203}{104}x^{100} \ll \frac{3}{4}x.$$

$$\Rightarrow \underset{\substack{\uparrow \\ \text{smaller} \\ \text{error}}}{\|f - f^*_e\|^2} < \underset{\substack{\uparrow \\ \text{bigger} \\ \text{error.}}}{\|f - f^*_\ell\|^2}$$

```
fplot(@(x) x^2,[0 1],'k')
hold on
fplot(@(x) 1/3+3/4*x,[0 1],'r')
hold on
fplot(@(x) (201/103)*(x^100)+(203/104)*(x^101),[0 1],'b')
legend('true x^2', 'appr by 1,x','appr by x^100, x^101')
hold off
grid on
```



3> Least square regression line from bivariate data

In statistic, the least square regression line for predicting y from x is given by $y = bx + a$ where $\bar{x} = \dfrac{\sum_{i=1}^{m} x_i}{m}$

where $b = r\dfrac{s_y}{s_x}$ $a = \bar{y} - b\bar{x}$ $\quad s_x^2 = \dfrac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m}$

Show that a and b are obtained via least square approximation $\quad r = \dfrac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{m s_x s_y}$

* Assume that we have a real observed data includes (x_i, y_i) $i = \overline{1,m}$

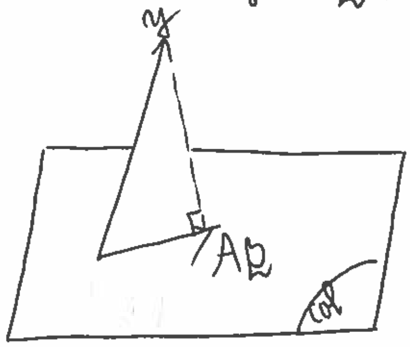
We want to solve the equation $\underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}}_{\text{put } = A} \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{=\beta} = \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}}_{\text{put} = y}$ to find $\beta = \begin{pmatrix} a \\ b \end{pmatrix}$ ✓

since $m \gg 2$, the equation $A\beta = y$ has no solution.

$\Rightarrow$ We want to consider the least square problem, which is the problem that find β so that it can minimize $\|A\beta - y\|_2$

$\Rightarrow$ We want to find $\hat{\beta}$ so that $A\hat{\beta} = \mathrm{Proj}^{\perp}_{\mathrm{col}\,A}(\vec{y})$ (orthogonal projection of $\vec{y}$ over space spaned by column of A)

$\uparrow$ since the orthogonal projection is the best ~~pro~~ approximation.



$\Rightarrow y - A\hat{\beta} \perp \mathrm{col}(A)$

$\Rightarrow A^T(y - A\hat{\beta}) = 0$

$A^T y - A^T A\hat{\beta} = 0$

$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \hat{\beta} = (A^T A)^{-1} A^T \vec{y}$

$$\Rightarrow \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (A^T A)^{-1}(A^T\vec{y}) = \left[\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix}\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix}\right]^{-1}\left[\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_m \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}\right]$$

$$= \underbrace{\begin{bmatrix} \sum_{i=1}^{m} 1 & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{bmatrix}}_{M}^{-1} \begin{bmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{bmatrix} = M^{-1}\begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}$$

Now we want to find M^{-1}

$$\text{et}(M) = \sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} 1 - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} x_i = m \sum_{i=1}^{m} x_i^2 - (m\bar{x})^2 = m^2\left(\frac{1}{m}\sum_{i=1}^{m} x_i^2 - \bar{x}^2\right)$$

$$= m^2\left(\frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2\right) = m \sum_{i=1}^{m}(x_i - \bar{x})^2$$

en

$$M^{-1} = \frac{1}{m\sum_{i=1}^{m}(x_i - \bar{x})^2}\begin{pmatrix} \sum_{i=1}^{m} x_i^2 & -\sum_{i=1}^{m} x_i \\ -\sum_{i=1}^{m} x_i & \sum_{i=1}^{m} 1 \end{pmatrix}$$

So we have

$$\begin{pmatrix}\hat{a} \\ \hat{b}\end{pmatrix} = \frac{1}{m\sum_{i=1}^{m}(x_i - \bar{x})^2}\begin{pmatrix} \sum_{i=1}^{m} x_i^2 & -\sum_{i=1}^{m} x_i \\ -\sum_{i=1}^{m} x_i & \sum 1 \end{pmatrix}_{2\times 2}\begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i y_i \end{pmatrix}_{2\times 1}$$

So we have

$$\hat{b} = \frac{1}{m\sum_{i=1}^{m}(x_i - \bar{x})^2}\left(-\sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i + \sum_{i=1}^{m} 1 \sum_{i=1}^{m} x_i y_i\right) =$$

$$= \frac{\frac{1}{m}\left(-\sum x_i \sum y_i + m \sum x_i y_i\right)}{m\underbrace{\left(\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{m}\right)}_{s_x^2}} = \frac{\frac{1}{m}\left(-m\bar{x}\, m\bar{y} + m\sum x_i y_i\right)}{m\, s_x^2} =$$

$$= \frac{-m\bar{x}\bar{y} + \sum x_i y_i}{m\, s_x^2}$$

we have

$$\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i y_i) - \sum \bar{x} y_i - \sum \bar{y} x_i + \sum \bar{x}\bar{y} = \qquad (*)$$

$$= \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + \sum_{i=1}^{m}\bar{x}\bar{y}$$

$$= \sum x_i y_i - \bar{x}\, m\bar{y} - \bar{y}\, m\bar{x} + m\bar{x}\bar{y} = \sum x_i y_i - m\bar{x}\bar{y} =$$

$$\Rightarrow \hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{m\, s_x^2} = \underbrace{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{m\, s_x\, s_y}}_{r}\frac{s_y}{s_x} = r\frac{s_y}{s_x} \quad \square \text{ for } \hat{b}$$

- So now we want to compute $\hat{a}$

$$\hat{a} = \frac{1}{m\sum_{i=1}^{m}(x_i - \bar{x})^2}\left[\sum_{i=1}^{m} y_i \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} x_i y_i\right]$$

$$\hat{a} = \frac{m\bar{y}\sum_{i=1}^{m} x_i^2 - m\bar{x}\sum x_i y_i}{m\sum_{i=1}^{m}(x_i-\bar{x})^2} \xrightarrow[\text{subtract } m\bar{x}\, m\bar{x}\bar{y}]{\text{add and}} \frac{m\bar{y}\sum_{i=1}^{m} x_i^2 - m\bar{x}\sum x_i y_i + m\bar{x}\, m\bar{x}\bar{y} - m\bar{x}}{m\sum_{i=1}^{m}(x_i-\bar{x})^2}$$

$$= \frac{m\bar{y}\sum x_i^2 - m^2\bar{x}^2\bar{y}}{m\sum_{i=1}^{m}(x_i-\bar{x})^2} - \frac{m\bar{x}\left[\sum x_i y_i - m\bar{x}\bar{y}\right]}{m\sum_{i=1}^{m}(x_i-\bar{x})^2} = \sum(x_i-\bar{x})(y_i-\bar{y}) \text{ by } (a$$

$$= \frac{\cancel{m}\bar{y}\,\cancel{\left[\sum x_i^2 - m\bar{x}^2\right]}}{\cancel{m}\cancel{\sum_{i=1}^{m}(x_i-\bar{x})^2}} - \frac{m\sum(x_i-\bar{x})(y_i-\bar{y})}{m\sum_{i=1}^{m}(x_i-\bar{x})^2}\,\bar{x}$$

$= \bar{y} - r\bar{x}$ □ done for $\hat{a}$.

✓

/0

47> Kincaid #3/404.

Suppose that we wish to approximate an even function by a polynomial of degree $\leq n$ using the norm $\| f\| = \left(\int_{-1}^{1} |f(x)|^2 dx\right)^{1/2}$

Prove that the best approximation is also even. Generalize.

* We want to prove that the best approximation of f has to be an even function

⇒ it suffices to prove that the distance of f and an even function is smaller than the distance of f and any odd function.

* Let e be an even function

o be an odd function

We need to prove that $\| f-e\|^2 \leq \| f-(e+o)\|^2$

- $\| f(-e\|^2 = \int_{-1}^{1} | f(x)-e(x)|^2 dx$

- $\| f-(e+o)\|^2 = \int_{-1}^{1} [f(x) - e(x) - o(x)]^2 dx$

$$= \int_{-1}^{1} [f(x)-e(x)]^2 dx - \underbrace{2\int_{-1}^{1} \underbrace{[f(x)-e(x)]}_{\text{even}} o(x)\, dx}_{\substack{\text{odd function} \\ =0 \text{ since } \int_{-1}^{1} \text{odd function } dx.}} + \underbrace{\int_{-1}^{1} (o(x))^2 dx}_{\geq 0}.$$

$$\geq \int_{-1}^{1} [f(x)-e(x)]^2 dx$$

⇒ $\| f-e\|^2 \leq \| f-(e+o)\|^2$

⇒ The best approximation has to be an even function □

10

1	6
2	4
3	10
4	10
	30

MAT 683 Homework 5
A. Lutoborski. Syracuse University. Fall 2018

Least squares approximation. SOLUTIONS.

1. Gaussian Quadrature Error Estimate. Let $Q(f) = \sum_{j=0}^{k} \lambda_j f(x_j)$ be the Gaussian quadrature on $(-1, 1)$ with weight $w(x) \equiv 1$ such that

$$Q(p) = \int_{-1}^{1} p(x)\,\mathrm{d}x \qquad \forall p \in \mathbb{P}_{2k+1}.$$

Show that

$$\left|Q(f) - \int_{-1}^{1} f(x)\,\mathrm{d}x\right| \leq 4 \inf_{p\in\mathbb{P}_{2k+1}} \left(\sup_{-1\leq x\leq 1} |f(x) - p(x)|\right)$$

Solution. The relevant information about the quadrature Q is that, as for every Gaussian quadrature, it is exact for all $p \in \mathbb{P}_{2k+1}$ and that its weights $\lambda_0, \ldots, \lambda_k$ are positive. Since Q is exact for $f \equiv 1$ then $\sum_{j=0}^{k} \lambda_j = 2 = \int_{-1}^{1} 1\,\mathrm{d}x$.

$$\begin{aligned}\left|Q(f) - \int_{-1}^{1} f(x)\,\mathrm{d}x\right| &= \left|Q(f) - Q(p) - \left(\int_{-1}^{1} f(x)\,\mathrm{d}x - \int_{-1}^{1} p(x)\,\mathrm{d}x\right)\right| \\ &\leq |Q(f) - Q(p)| + \left|\int_{-1}^{1} (f(x) - p(x))\,\mathrm{d}x\right| \\ &\leq \sum_{j=0}^{k} \lambda_j |f(x_j) - p(x_j)| + \int_{-1}^{1} |f(x) - p(x)|\,\mathrm{d}x \\ &\leq \sup_{-1\leq x\leq 1} |f(x) - p(x)| \sum_{j=0}^{k} \lambda_j + \sup_{-1\leq x\leq 1} |f(x) - p(x)| \int_{-1}^{1} 1\,\mathrm{d}x \\ &= \sup_{-1\leq x\leq 1} |f(x) - p(x)|(2 + 2).\end{aligned}$$

Finally we take the infimum over $p \in \mathbb{P}_{2k+1}$ of the right side of the inequality. As a consequence we have $\lim_{k\to\infty} Q_k(f) = \int_{-1}^{1} f(x)\,\mathrm{d}x$ because by Weierstrass theorem $\lim_{k\to\infty} \inf_{p\in\mathbb{P}_{2k+1}} \sup_{-1\leq x\leq 1} |f(x) - p(x)| = 0$.

2. Approximate x^2 in $L^2(0, 1)$ by a combination of $1, x$ and by by a combination of x^{100}, x^{101}. Which approximation gives a smaller approximation error? Explain the reasons. Plot the approximations on the same graph.

Solution. According to the orthogonal projection theorem to find the best approximation $f^* = c_1\phi_1 + c_2\phi_2$ to the function $f(x) = x^2$ in both cases: $\phi_1(x) = 1, \phi_2(x) = x$ and $\phi_1(x) = x^{100}, \phi_2(x) = x^{101}$ we need to solve the normal system

$$\begin{bmatrix} \langle \phi_1, \phi_1 \rangle & \langle \phi_2, \phi_1 \rangle \\ \langle \phi_1, \phi_2 \rangle & \langle \phi_2, \phi_2 \rangle \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \langle x^2, \phi_1 \rangle \\ \langle x^2, \phi_2 \rangle \end{bmatrix}.$$

The system in the first case is

$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{4} \end{bmatrix}$$

This gives $c_1 = -\frac{1}{6}$ $c_2 = 1$ and

$$f^*(x) = -\frac{1}{6} + x$$

with error

$$\|f - f^*\|^2_{L^2[0,1]} = \frac{1}{180} \approx 0.0055556.$$

The system in the second case is

$$\begin{bmatrix} \frac{1}{201} & \frac{1}{202} \\ \frac{1}{202} & \frac{1}{203} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{103} \\ \frac{1}{104} \end{bmatrix}$$

This gives $c_1 = \frac{2009799}{5356}$, $c_2 = -\frac{1004647}{2678}$ and

$$f^*(x) = \frac{2009799}{5356} x^{100} - \frac{1004647}{2678} x^{101},$$

with error

$$\|f - f^*\|^2_{L^2[0,1]} = \frac{23532201}{143433680} \approx 0.164063.$$

Therefore the approximation with high powers is less effective since they are 'less linearly independent'.

3. Least squares regression line from bivariate data. In statistics the least squares regression line for predicting y from x is given by $y = bx + a$ where

$$b = r\frac{s_y}{s_x}, \qquad a = \bar{y} - b\bar{x}.$$

and

$$\bar{x} = \frac{x_1 + \cdots + x_m}{m}, \quad s_x^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_m - \bar{x})^2}{m},$$
$$r = \frac{\frac{1}{m}((x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_m - \bar{x})(y_m - \bar{y}))}{s_x s_y}.$$

are respectively the mean, variance and correlation. Show how a and b are obtained via least squares approximation.

Solution. Our task is to find an affine function $f : \mathbb{R} \to \mathbb{R}$, $y = f(x) = bx + a$ whose graph is close to the data points $(x_1, y_1), \ldots, (x_m, y_m)$. Denote the residuals $r_i = y_i - f(x_i)$. In the context of least squares data fitting we want to minimize $\|r\|^2 = \sum_{i=1}^m r_i^2$.

$$r = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = y - Ac.$$

The minimizer $c = [a, b]^T$ of $\|y - Ac\|^2$ is given as the solution of the system of normal equations:

$$A^T Ac = A^T y.$$

Explicitly the system is

$$\begin{bmatrix} m & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Since $\det(A^T A) = m \sum x_i^2 - (\sum x_i)^2$ then the Cramer formulas give us the solution in terms of the right hand side of the system:

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{m \sum x_i^2 - (\sum x_i)^2}, \qquad b = \frac{m \sum x_i y_i - \sum x_i \sum y_i}{m \sum x_i^2 - (\sum x_i)^2}.$$

We have to assume here that not all x_i are equal or in other words that A is of full column rank.

In order to solve the system in terms of the means, variances and correlation we solve the first equation for a:

$$a = \bar{y} - b\bar{x}.$$

This gives us the formula for a that we need. If we substitute the above value of a in the second equation we obtain

$$b = \frac{\langle x, y \rangle - m\bar{x}\bar{y}}{\|x\|^2 - m\bar{x}^2}.$$

The question is why $b = r\frac{s_y}{s_x}$?

From the Pythagorean theorem we have:

$$\|x\|^2 - m\bar{x}^2 = (x_1 - \bar{x})^2 + \ldots + (x_m - \bar{x})^2 = ms_x^2.$$

Denote $e = [1, \ldots, 1]^T \in \mathbb{R}^m$.

$$\begin{aligned} ms_x s_y \cdot r &= (x_1 - \bar{x})(y_1 - \bar{y}) + \ldots + (x_m - \bar{x})(y_m - \bar{y}) \\ &= \langle x - \bar{x}e, y - \bar{y}e \rangle \\ &= \langle x, y \rangle + m\bar{x}\bar{y} - m\bar{y}\bar{x} - m\bar{x}\bar{y} \\ &= \langle x, y \rangle - m\bar{x}\bar{y}. \end{aligned}$$

Substituting the above computed quantities in the numerator and denominator of b we obtain $b = r\frac{s_y}{s_x}$.

4. (Kincaid p 404,#3) Approximate an even function by a polynomial of degree n using $\|f\| = (\int_{-1}^{1} |f(x)|^2)^{1/2}$. Prove that the best approximation is also even. Generalize.

Solution. We generalize by taking a positive weight $w(x)$ which is an even function on $(-1,1)$. Let $\phi_0, \phi_1, \phi_2, \ldots, \phi_n$ be obtained by Gram-Schmidt from $1, x, x^2, \ldots$. We have that if j is even (odd) than ϕ_j is even (odd). This can be shown by induction from

$$\phi_k(x) = x^k - a_0\phi_0(x) - \ldots - a_{k-1}\phi_{k-1}(x)$$

where

$$a_j = \frac{\int_{-1}^{1} x^k \phi_j(x) w(x)\, \mathrm{d}x}{\int_{-1}^{1} (\phi_j(x))^2 w(x)\, \mathrm{d}x}$$

Suppose the result is true for $j = 0, \ldots, k-1$ then with w even if k+j is odd then the numerator is 0. Let $p_n(x) = \gamma_0\phi_0(x) + \ldots + \gamma_n\phi_n(x)$ be the BAP for f. Then

$$\gamma_j = \frac{\int_{-1}^{1} f(x)\phi_j(x) w(x)\, \mathrm{d}x}{\int_{-1}^{1} (\phi_j(x))^2 w(x)\, \mathrm{d}x}$$

If f is even then for $2j-1$ ϕ_{2j-1} is odd and $\gamma_{2j-1} = 0$ and hence p_n is even.

MAT 683 Exam 2

A. Lutoborski. Syracuse University. Fall 2018

1. (20 pts) Suppose that $f : [0,1] \to \mathbb{R}$ has a continuous second derivative on $[0,1]$. Show that there is $\xi \in (0,1)$ such that

$$\int_0^1 x f(x)\,\mathrm{d}x = \frac{1}{2} f(\frac{2}{3}) + \frac{1}{72} f''(\xi)$$

Hint: Gauss quadrature on $(0,1)$ with weight $w(x)$ and nodes $x_0, \ldots, x_k$ has error $I(f) - Q(f) = \frac{1}{(2k+2)!} f^{(2k+2)}(\xi) \int_0^1 \Pi^2(x) w(x) \mathrm{d}x$, $\Pi(x) = \prod_{i=0}^{k} (x - x_i)^2$.

2. (20 pts) Consider $f \in L^2(\mathbb{R})$ and let $U \subset L^2(\mathbb{R})$ be a subspace of even functions i.e. $U = \{g \in L^2(\mathbb{R}) : g(x) = g(-x)\}$. Let

$$f_e(x) = \frac{1}{2}(f(x) + f(-x))$$

(a) Show that $f_e \in U$.
(b) Show that

$$\langle f - f_e, g \rangle = 0 \quad \text{for all} \quad g \in U$$

(c) Explain why

$$\|f - f_e\| \le \min_{g \in U} \|f - g\|$$

3. (20 pts) Let V be a normed vector space and $W \subset V$ be a finite dimensional subspace. Element $h^* \in W$ is a best approximant to $f \in V$ if $\|f - h^*\| \le E_W(f) = \inf_{h \in W} \|f - h\|$. Show that the set S of best approximants to $f \in V$ is convex, that is if $h_1, h_2 \in S$ then $\alpha h_1 + \beta h_2 \in S$ if $\alpha, \beta \ge 0$ and $\alpha + \beta = 1$.

4. (20 pts) Let $x_1, x_2, \ldots, x_{2N+1}$ be distinct points in $[0,1)$. For $1 \le r \le 2N+1$ put

$$T_r(x) = \prod_{\substack{j=1 \\ j \ne r}}^{2N+1} \frac{\sin \pi(x - x_j)}{\sin \pi(x_r - x_j)}$$

(a) Show that $T_r(x)$, $x \in \mathbb{R}$ is a 1-periodic function on $\mathbb{R}$.
(b) Show that T_r is an $2N$-degree trigonometric polynomial.
(c) Show that if $T(x) = \sum_{r=1}^{2N+1} c_r T_r(x)$ then

$$T(x_r) = c_r \qquad 1 \le r \le 2N+1$$

5. (20 pts) Let $T(x) = \sum_{n=-N}^{N} t_n e(nx)$ be a trigonometric polynomial, where $e(x) = \exp(i2\pi x)$. Let $q \in \mathbb{Z}$, $q > 0$, $\alpha \in \mathbb{R}$. Show that

$$\frac{1}{q} \sum_{a=0}^{q-1} T\left(\frac{a}{q} + \alpha\right) = \sum_{\substack{-N \le n \le N \\ q \mid n}} t_n e(n\alpha)$$

MAT 683 Exam 2 SOLUTIONS
A. Lutoborski. Syracuse University. Fall 2018

1. (20 pts) Suppose that $f : [0,1] \to \mathbb{R}$ has a continuous second derivative on $[0,1]$. Show that there is $\xi \in (0,1)$ such that

$$\int_0^1 x f(x)\,\mathrm{d}x = \frac{1}{2} f(\frac{2}{3}) + \frac{1}{72} f''(\xi)$$

Hint: Gauss quadrature on $(0,1)$ with weight $w(x)$ and nodes $x_0, \ldots, x_k$ has error $I(f) - Q(f) = \frac{1}{(2k+2)!} f^{(2k+2)}(\xi) \int_0^1 \Pi^2(x) w(x)\mathrm{d}x$, $\Pi(x) = \prod_{i=0}^{k}(x - x_i)^2$.
Solution. We consider the right side of the formula as a sum of Gaussian quadrature with one node and remainder term in the quadrature. We have $w(x) = x$. To obtain a Gaussian quadrature with one node we need an orthogonal polynomial of degree 1 on $(0,1)$ with weight w. By orthogonalizing monomials 1 and x we obtain $p_1(x) = x - 2/3$ whose root is $x_0 = 2/3$. To obtain the weight λ_0 we need to integrate the cardinal Lagrange interpolating polynomial $l_0(x) = 1$

$$\lambda_0 = \int_0^1 1 \cdot x\,\mathrm{d}x = \frac{1}{2}$$

Finally the remainder: $\Pi(x) = x - \frac{2}{3}$

$$\int_0^1 \Pi^2(x) w(x)\mathrm{d}x = \int_0^1 (x - \frac{2}{3})^2 x\,\mathrm{d}x = \frac{1}{36}$$

So that $I(f) - Q(f) = 1/72 f^{(2)}(\xi)$.

2. (20 pts) Consider $f \in L^2(\mathbb{R})$ and let $U \subset L^2(\mathbb{R})$ be a subspace of even functions i.e. $U = \{g \in L^2(\mathbb{R}) : g(x) = g(-x)\}$. Let

$$f_e(x) = \frac{1}{2}(f(x) + f(-x))$$

(a) Show that $f_e \in U$.
(b) Show that

$$\langle f - f_e, g \rangle = 0 \quad \text{for all} \quad g \in U$$

(c) Explain why

$$\|f - f_e\| \le \min_{g \in U} \|f - g\|$$

Solution.

$$\begin{aligned} \langle f - f_e, g \rangle &= \int_{-\infty}^{\infty} \frac{1}{2} f(x) g(x)\,\mathrm{d}x - \int_{-\infty}^{\infty} \frac{1}{2} f(-x) g(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{2} f(x) g(x)\,\mathrm{d}x - \int_{-\infty}^{\infty} \frac{1}{2} f(-x) g(-x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} \frac{1}{2} f(x) g(x)\,\mathrm{d}x - \int_{-\infty}^{\infty} \frac{1}{2} f(y) g(y)\,\mathrm{d}y = 0 \end{aligned}$$

Since U is a closed subspace in $L^2(\mathbb{R})$ then the projection f_e of f onto U is the best approximation of f.

3. (20 pts) Let V be a normed vector space and $W \subset V$ be a finite dimensional subspace. Element $h^* \in W$ is a best approximant to $f \in V$ if $\|f - h^*\| \leq E_W(f) = \inf_{h \in W} \|f - h\|$. Show that the set S of best approximants to $f \in V$ is convex, that is if $h_1, h_2 \in S$ then $\alpha h_1 + \beta h_2 \in S$ if $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$.

Solution. A function h is the best approximant of f in the subspace W if

$$\|f - h\| \leq \inf_{g \in W} \|f - g\| = E_W(f)$$

Suppose that $h_1, h_2 \in W$ are both best approximants of f and suppose that $\alpha h_1 + \beta h_2$ and $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. Then

$$\begin{aligned}\|f - (\alpha h_1 + \beta h_2)\| &= \|(\alpha + \beta) f - (\alpha h_1 + \beta h_2)\| = \|\alpha(f - h_1) + \beta(f - h_2)\| \\ &\leq \alpha\|f - h_1\| + \beta\|f - h_2\| = (\alpha + \beta) E_W(f) = E_W(f)\end{aligned}$$

4. (20 pts) Let $x_1, x_2, \ldots, x_{2N+1}$ be distinct points in $[0, 1)$. For $1 \leq r \leq 2N+1$ put

$$T_r(x) = \prod_{\substack{j=1 \\ j \neq r}}^{2N+1} \frac{\sin \pi(x - x_j)}{\sin \pi(x_r - x_j)}$$

(a) Show that $T_r(x)$, $x \in \mathbb{R}$ is a 1-periodic function on $\mathbb{R}$.
(b) Show that T_r is an $2N$-degree trigonometric polynomial.
(c) Show that if $T(x) = \sum_{r=1}^{2N+1} c_r T_r(x)$ then

$$T(x_r) = c_r \qquad 1 \leq r \leq 2N + 1$$

Solution. (a)

$$\sin(\pi(x + 1 - x_j)) = -\sin(\pi(x - x_j))$$

Hence each factor in $T_r(x)$ is not 1-periodic, but a product of $2N$ such factors is 1-periodic and hence $T_r(x + 1) = T_r(x)$.

(b) The $\sin(\pi x)$ function is a trigonometric polynomial of degree 1 because

$$\sin(\pi x) = \frac{1}{2i}\left(e(x/2) - e(-x/2)\right)$$

Similarly $\cos(\pi x)$ and hence $\sin(\pi(x - x_j))$ are trigonometric polynomials of degree 1. Multiplication of two trigonometric polynomials gives a trigonometric polynomial whose degree is the sum of the degrees of the factors. Therefore $T_r(x)$ is a product of $2N$ trigonometric polynomials of degree 1 and hence is a polynomial of degree $2N$.

5. (20 pts) Let $T(x) = \sum_{n=-N}^{N} t_n e(nx)$ be a trigonometric polynomial, where $e(x) = \exp(i2\pi x)$. Let $q \in \mathbb{Z}$, $q > 0$, $\alpha \in \mathbb{R}$. Show that

$$\frac{1}{q}\sum_{a=0}^{q-1} T\left(\frac{a}{q} + \alpha\right) = \sum_{\substack{-N \leq n \leq N \\ q|n}} t_n e(n\alpha)$$

Solution. We will use the basic identity about sampling of polynomial $e(nx)$

$$\sum_{a=0}^{q-1} e(\frac{na}{q}) = \begin{cases} q & q|n \\ 0 & q\nmid n \end{cases}$$

$$\begin{aligned}\sum_{a=0}^{q-1} T\left(\frac{a}{q}+\alpha\right) &= \sum_{n=-N}^{N} t_n \sum_{a=0}^{q-1} e(\frac{na}{q})e(n\alpha) \\ &= \sum_{n=-N}^{N} t_n e(n\alpha) \sum_{a=0}^{q-1} e(\frac{na}{q}) \\ &= q \sum_{\substack{-N\leq n\leq N \\ q|n}} t_n e(n\alpha)\end{aligned}$$