NUMERICAL ANALYSIS

Big-O: $f(n) = O(\phi(n))$ means $\left|\frac{f(n)}{\phi(n)}\right|$ bounded for sufficiently large n, i.e. $\exists\, C, N_0$ s.t. $\forall n \geq N_0$, $\left|\frac{f(n)}{\phi(n)}\right| \leq C$.

↳ if $f(x) = O(\phi(x))$ as $x \to x_0$, then $\left|\frac{f(x)}{\phi(x)}\right|$ bounded in a neighbourhood of x_0 $(|x - x_0| \leq r)$.

FLOATING POINT: $F_{\sigma, Q, S}$, σ = exp shift, Q = # of exp bits, S = # of significand bits

$$F^{\text{normal}}_{\sigma, Q, S} = \{\pm 2^{q-\sigma} \times (1.b_1 b_2 \cdots b_S)_2 : 1 \leq q < 2^Q - 1\}$$

$$F^{\text{sub}}_{\sigma, Q, S} = \{\pm 2^{1-\sigma} \times (0.b_1 b_2 \cdots b_S)_2\}, \quad F^{\text{special}} = \{\infty, -\infty, \text{NaN}\}.$$

↳ first bit is signed bit → $0 = +ve$, $1 = -ve$.

e.g. $\frac{1}{3} = \sum_{k=0}^{\infty} \frac{1}{2^{2k}} = (0.010101\ldots)_2 = 2^{-2}(1.0101\ldots)_2$ ← rounding

$= 2^{125-127}(1.0101\ldots)_2$ etc.

• For $F_{15,5,10}$: subnormal range → $(2^{-14}, 2^{-24}]$ (always discrete)

$\uparrow 2^{1-\sigma-S}$

MACHINE EPSILON: $\varepsilon_M = 2^{-S}$ → "max error that can occur when rounding to the unit value".

MAX possible NORMAL number

↳ $= 2^{2^Q - 2 - \sigma} \times (1.111\ldots)_2 = 2^{2^Q - 2 - \sigma} \cdot (2 - \varepsilon_M)$

e.g. $m(y) = \min\{x \in F_{127,8,23} : x > y\}$ → smallest single precision number greater than y.

↳ $m(2) - 2$, next float after 2 is $2 \times (1 + 2^{-23}) \Rightarrow m(2) - 2 = 2^{-22}$

↳ $m(1024) - 1024 = 2^{10}(1 + 2^{-23}) - 2^{10} = 2^{-13}$.

OPERATIONS: $x \oplus y = fl(x+y)$, $x \ominus y = fl(x-y)$, $x \otimes y = fl(x \times y)$, $x \oslash y = fl(x/y)$. $\Big\}$ $x, y \in F$.

ABSOLUTE / RELATIVE ERROR: approximating x by $\tilde{x}$,

$\tilde{x} = x + \delta_a = x(1+\delta_r)$ then $|\delta_a|, |\delta_r|$ are absolute + relative errors.

$$fl(x) = x(1+\delta_x) \text{ where } |\delta_x^{nearest}| \leq \frac{\varepsilon_m}{2}, \ |\delta_x^{up/down}| < \varepsilon_M.$$

e.g. $(fl(1.1) \oplus fl(1.2)) \oplus fl(1.3) = 3.6 + C\varepsilon_M$ for some C.

DIVIDED DIFFERENCES: $f'(x) \approx \dfrac{f(x+h) - f(x)}{h}$ sufficiently small h.

• error in approximating derivative:

$$\left| f'(x) - \frac{f(x+h) - f(x)}{h} \right| \leq \frac{Mh}{2} \quad \text{for } M = \sup_{x \leq t \leq x+h} |f''(t)|$$

$$\boxed{\text{Taylor's thm: } f(x+h) = f(x) + f'(x)h + \frac{f''(t)}{2}h^2}$$

for some $t \in [x, x+h]$. □

OTHER DIFFERENCES: (leftside, central) $\dfrac{f(x) - f(x-h)}{h}$, $\dfrac{f(x+h) - f(x-h)}{2h}$

DIVIDED DIFFERENCE ERROR BOUND: assume $f^{FP}(x) = f(x) + \delta_x^f$

and $|\delta_x^f| \leq c\varepsilon_m$, THEN $(f^{FP}(x+h) \ominus f^{FP}(x)) \oslash h = f'(x) + \delta_{x,h}^{FD}$

since: $= \dfrac{f(x+h) + \delta_{x+h}^f - f(x) - \delta_x^f}{h}(1+\delta_1)$ assuming $h = 2^{-n}$, $n \leq S$ $\Rightarrow x \oplus h = x + h$.

$$= \frac{f(x+h) - f(x)}{h}(1+\delta_1) + \frac{\delta_{x+h}^f - \delta_x^f}{h}(1+\delta_1) \rightarrow |\delta_1| \leq \frac{\varepsilon_M}{2}.$$

Taylor's Thm

$$= f'(x) + \underbrace{f'(x)\delta_1 + \frac{f''(t)}{2}h(1+\delta_1) + \frac{\delta_{x+h}^f - \delta_x^f}{h}(1+\delta_1)}$$

$$\leq \frac{|f'(x)|}{2}\varepsilon_m + Mh + \frac{4c\varepsilon_m}{h}, \quad M = \sup_{x \leq t \leq x+h} |f''(t)|$$

using bound $|1+\delta_1| \leq 2$.

N/B: choose $h \propto \sqrt{\varepsilon_m}$ in divided differences, i.e. $h = k \cdot \sqrt{\varepsilon_m}$.

DUAL NUMBERS: $a + b\varepsilon$, $a \in \mathbb{R}$, $b \in \mathbb{R}$, $\varepsilon^2 = 0$.

$\boxed{\text{POLYS on DUAL NUMBERS: } p(a+b\varepsilon) = p(a) + b \cdot p'(a) \cdot \varepsilon}$ → for poly p.

e.g. for $p(x) = (x-1)\cdot(x-2) + x^2$, computing $p'(2)$:

$$p(2+\varepsilon) = (1+\varepsilon)\cdot\varepsilon + (2+\varepsilon)^2 = \varepsilon + 4 + 4\varepsilon = 4 + \overbrace{5}^{p'(2)}\varepsilon$$

$\boxed{\text{DUAL EXTENSION: } f(a+b\varepsilon) = f(a) + b f'(a)\cdot\varepsilon}$

↳ dual extension preserved under chain (composition) and product.

MATRICES: mul_rows and mul_columns for $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$,

$$Ax = \begin{bmatrix} \sum_{j=1}^{n} a_{1j}x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj}x_j \end{bmatrix} = x_1 \cdot \underline{a}_1 + x_2 \cdot \underline{a}_2 + \dots + x_n \cdot \underline{a}_n$$

↖ how we normally do it by hand.

BACK-SUBSTITUTION: $\underline{U}\,\underline{x} = \underline{b}$ → $\begin{pmatrix} u_{11} & \cdots & u_{1n} \\ & \ddots & \vdots \\ & & u_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$

$$\Rightarrow x_k \cdot u_{kk} = b_k - \sum_{j=k+1}^{n} u_{kj} x_j$$

- DIAGONAL matrices have bandwidths $(l, u) = (0,0)$.
- BIDIAGONAL has $(l,u) = (1,0)$ for lower bidiagonal and $(l,u) = (0,1)$ for upper bidiagonal.
- TRIDIAGONAL has $(l,u) = (1,1)$.

UNITARY: $U(n) = \{\underline{Q} \in \mathbb{C}^{n\times n} : \underline{Q}^*\underline{Q} = \underline{I}\}$ (for $\underline{Q}^* = \overline{\underline{Q}^T}$)

↳ norm-preserving → $\|\underline{Q}\,\underline{x}\| = \|\underline{x}\|$, and all evalues of $\underline{Q} = \pm 1$.

PERMUTATION MATRIX is ORTHOGONAL → $P_\sigma^T = P_{\sigma^{-1}} = P_\sigma^{-1}$.

ROTATIONS: $SO(2) = \{Q \in O(2) : \det Q = 2\}$.

$$Q_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{OR} \quad Q = \frac{1}{\sqrt{a^2+b^2}}\begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$ → $(\tan\theta = \frac{b}{a})$

REFLECTION: $Q_v = I - 2\underline{v}\cdot\underline{v}^*$ (where $\|\underline{v}\| = 1$).

note: $\underline{v}\cdot\underline{v}^T$ rank 1 matrix $\Rightarrow$ Q_v rank 1 perturbation of I.

$\det Q_v = -1$ (since -1 is evalue with multiplicity 1).

HOUSEHOLDER REFLECTION maps to axis: $\underline{Q_x\,\underline{x} = \pm\|\underline{x}\|\,\underline{e}_1}$

↳ choose opposite sign of x_1 for reflection.

e.g. Q s.t. $Q\begin{pmatrix}2\\1\\2\end{pmatrix} = \begin{pmatrix}-3\\0\\0\end{pmatrix}$, $\underline{x} = \begin{pmatrix}2\\1\\2\end{pmatrix}$ $\|x\| = 3$

$\Rightarrow \underline{\underline{y} = \mp\|\underline{x}\|\,\underline{e}_1 + \underline{x}} = \begin{pmatrix}3\\0\\0\end{pmatrix} + \begin{pmatrix}2\\1\\3\end{pmatrix} = \begin{pmatrix}5\\1\\2\end{pmatrix}$

$\Rightarrow \underline{w} = \dfrac{\underline{y}}{\|\underline{y}\|} = \dfrac{1}{\sqrt{30}}\begin{pmatrix}5\\1\\2\end{pmatrix} \Rightarrow Q = I - 2\underline{w}\cdot\underline{w}^T = \dfrac{1}{15}\begin{pmatrix}-10 & -5 & -10\\ -5 & 14 & -2\\ -10 & -2 & 11\end{pmatrix}.$

QR FACTORISATION: $A = QR = (q_1|\cdots|q_m)\begin{pmatrix}\times & \cdots & \times\\ & \ddots & \vdots\\ & & \times\\ & & 0\\ & & \vdots\\ & & 0\end{pmatrix} \leftarrow m\times n$

reduced QR: $A = \hat{Q}\hat{R} = \underset{(m\times n)}{(q_1|\cdots|q_n)}\underset{(n\times n)}{\begin{pmatrix}\times & \cdots & \times\\ & \ddots & \vdots\\ & & \times\end{pmatrix}}$

$A = QR \Rightarrow \underline{A^{-1}b = R^{-1}Q^Tb}$ $\quad(Ax = b \Rightarrow x = A^{-1}b)$

$\boxed{\text{given } QR = A\text{, THEN } \underline{x = \hat{R}^{-1}\hat{Q}^Tb} \text{ minimises } \|Ax-b\|.}$

(norm minimised if it is zero provided full column rank of A

CHOLESKY: $A = LU = LL^T$ e.g. $A = \begin{pmatrix}2 & 2 & 1\\ 2 & 3 & 2\\ 1 & 2 & 2\end{pmatrix}$.

here, $\alpha_1 = 2$ and $\underline{v}_1 = \begin{pmatrix}2\\1\end{pmatrix}$ $\Rightarrow A_2 = \begin{pmatrix}3 & 2\\ 2 & 2\end{pmatrix} - \underset{\uparrow\,\alpha_1}{\dfrac{1}{2}}\underset{\uparrow\,\underline{v}_1}{\begin{pmatrix}2\\1\end{pmatrix}}(2\ 1) = \begin{pmatrix}1 & 1\\ 1 & 3/2\end{pmatrix}$

then $\alpha_2 = 1$, $\underline{v}_2 = 1$ $\Rightarrow A_3 = \dfrac{3}{2} - 1 = \dfrac{1}{2}. = \alpha_3$

$\Rightarrow L = \begin{pmatrix}\sqrt{2} & & \\ \sqrt{2} & 1 & \\ 1/\sqrt{2} & 1 & 1/\sqrt{2}\end{pmatrix} = \begin{pmatrix}\sqrt{\alpha_1} & & \\ \dfrac{v_1}{\sqrt{\alpha_1}} & \sqrt{\alpha_2} & \\ & \dfrac{v_2}{\sqrt{\alpha_2}} & \sqrt{\alpha_3}\end{pmatrix}.$

P-NORM: $\|x\|_p = \left(\sum_k |x_k|^p\right)^{1/p}$, $\|x\|_\infty = \max_k |x_k|$

$(\|x+y\| \leq \|x\| + \|y\|$; $\|c\underline{x}\| = |c|\cdot\|\underline{x}\|$; $\|x\| = 0 \Rightarrow \underline{x} = \underline{0})$

FROBENIUS-NORM: $\underline{\|A\|_F = \sqrt{\sum_{k=1}^{m}\sum_{j=1}^{n}|a_{kj}|^2}}$

SINGULAR VALUE DECOMPOSITION : $A = U \cdot \Sigma \cdot V^*$

if $\lambda_1, \dots, \lambda_n$ evalues of $A^* A$ (decreasing order) then,

$$\Sigma = \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_n} \end{pmatrix} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$$

note : $\|A\|_2 = \sigma_1$ and $\|A^{-1}\|_2 = \sigma_n^{-1} = \frac{1}{\sigma_n}$.

CONDITION NUMBER : for an error in the input, you can use condition number to bound error in the output.

- FORWARD ERROR ANALYSIS → bound error in output.
- BACKWARD ERROR ANALYSIS → bound error in input.

$$\boxed{\kappa(A) = \|A\|_2 \cdot \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n} \quad (= \kappa(A^{-1}))}$$

and $\kappa_p(A) = \|A\|_p \cdot \|A^{-1}\|_p$.

Assume backward error bound $\|\delta A\| \leq \|A\| \cdot \varepsilon$. Then,

for $(A + \delta A)x = Ax + \delta_f$:

forward error bound is → $\|\delta_f\| \leqslant \|Ax\| \cdot \kappa(A) \cdot \varepsilon$

↳ given error in input, we can find error in output,

FOURIER EXPANSIONS : $f(\theta) = \sum_{k=-\infty}^{\infty} \hat{f}_k \cdot e^{ik\theta}$ → Fourier-Taylor if it goes from 0 to ∞.

for $\hat{f}_k = \langle e^{ik\theta}, f \rangle = \int_0^{2\pi} \frac{1}{2\pi} e^{-ik\theta} \cdot f(\theta)\, d\theta$

(A funct^n has bounded 2-norm : $\|f\|_2 = \sqrt{\int_0^{2\pi} |f(\theta)|^2 d\theta} < \infty$

iff sum of $|\hat{f}_k|^2$ are bounded : $\|\hat{f}\|_2 = \sqrt{\sum_{k=-\infty}^{\infty} |\hat{f}_k|^2} < \infty$.)

N/B : use INTEGRATION BY PARTS for bounding $\hat{f}_k$ with $f(0) = f(2\pi)$
$f'(0) = f'(2\pi)$ etc.

TRAPEZIUM RULE : $\int_0^{2\pi} f(\theta)\, d\theta \approx \frac{2\pi}{n} \sum_{j=0}^{n-1} f(\theta_j)$.

DISCRETE FOURIER COEFFS: $\hat{f}_k^n = \frac{1}{n} \sum_{j=0}^{n-1} e^{-ik\theta_j} f(\theta_j)$

ALIASING: $\forall p \in \mathbb{Z}$, $\hat{f}_k^n = \hat{f}_{k+pn}^n$ $\Rightarrow$ knowing $\hat{f}_0^n, \dots, \hat{f}_{n-1}^n$ enough to know all.

Approx in terms of exact: $\boxed{\hat{f}_k^n = \dots + \hat{f}_{k-2n} + \hat{f}_{k-n} + \hat{f}_k + \hat{f}_{k+n} + \dots}$

ORTHOGONAL POLYNOMIALS: non-periodic $f(x) = \sum_{k=0}^{\infty} c_k \cdot p_k(x)$

GRADED: if p_n is degree n in set $\{p_0(x), p_1(x), \dots, p_n(x), \dots\}$

$\langle f, g \rangle := \int_a^b f(x) g(x) w(x)\, dx$ for weight $w(x)$ and $x \in (a,b)$

↳ graded polynomial basis $\{p_0(x), p_1(x), \dots\}$ are OPs if $\langle p_n, p_m \rangle = 0$ $(n \neq m)$.

PROP: for poly $r(x)$ degree n, and orthogonal $\{p_n(x)\}$

$$\Rightarrow r(x) = \sum_{k=0}^{n} \frac{\langle p_k, r \rangle}{\| p_k \|^2} \cdot p_k(x)$$

↳ if r satisfies: $0 = \langle p_0, r \rangle = \dots = \langle p_n, r \rangle \Rightarrow r = 0$.

ORTHOGONAL TO LOWER DEGREE: for $p(x) = k_n x^n + O(x^{n-1})$ with $\langle p, f_m \rangle = 0$ $\forall f_m$ degree $m < n$, iff $p = k_n \pi_n(x)$ where $\pi_n(x)$ monic OP.

$$\boxed{\begin{aligned} &\text{3-TERM-RECURRENCE: } x \cdot p_0(x) = a_0 \cdot p_0(x) + b_0 \cdot p_1(x) \\ &x \cdot p_n(x) = c_{n-1} \cdot p_{n-1}(x) + a_n \cdot p_n(x) + b_n \cdot p_{n+1}(x) \\ &\text{where } \{p_n\} \text{ are OPs.} \end{aligned}}$$

↳ if $\{\pi_n\}$ are monic OPs, then $b_n = 1$.

e.g. monic $\pi_0(x), \dots, \pi_3(x)$. $x\pi_0 = a_0 \pi_0 + \pi_1$

$\Rightarrow a_0 = \frac{\langle \pi_0, x\pi_0 \rangle}{\| \pi_0 \|^2}$, then find π_1.

$x\pi_1 = c_0 \pi_0 + a_1 \pi_1 + \pi_2 \Rightarrow c_0 = \frac{\langle \pi_0, x\pi_1 \rangle}{\| \pi_0 \|^2}$, $a_1 = \frac{\langle \pi_1, x\pi_1 \rangle}{\| \pi_1 \|^2}$ etc.

JACOBI MATRIX: $x \cdot f(x) = x \cdot \sum_{k=0}^{\infty} c_k \cdot P_k(x) = x \cdot [P_0 | P_1 | \dots] \cdot \begin{bmatrix} c_0 \\ c_1 \\ \vdots \end{bmatrix}$

$= x \cdot \underline{P}(x) \cdot \vec{c} = \underline{P}(x) \cdot \underbrace{X \vec{c}}$

$X = J^T.$

(3-term-recurrence)

$$x\, P(x) = P(x) \begin{pmatrix} a_0 & c_0 & & \\ b_0 & a_1 & c_1 & \\ & b_1 & a_2 & \ddots \\ & & b_2 & \ddots \end{pmatrix}$$

← Jacobi matrix is this TRANSPOSE.

J for OPs

(but for OPs, J symmetric) ✓

TO SHOW OP: check: ① graded, ② orthogonal w.r.t. $w(x)$ to all lower degree.

③ have right normalisation constant k_n.

e.g. for Chebyshev → $T_0(x) = 1$, $T_n(x) = 2^{n-1} x^n + O(x^{n-1})$

on $[-1,1]$ with $w(x) = \frac{1}{\sqrt{1-x^2}}$.

check $\langle t_n, t_m \rangle = 0$; show $p_{n+1} = 2x\, p_n - p_{n-1}$ ⇒ graded since multiply by $2x$ each time starting with $p_0 = 1$; and $k_n = 2^{n-1}$. □

LAGRANGE BASIS POLYNOMIAL: $\ell_k(x) = \prod_{j \neq k} \frac{x - x_j}{x_k - x_j} = \frac{(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}$

↳ $\ell_k(x_k) = 1$ and $\ell_k(x_j) = 0$ for $j \neq k$.

$\boxed{\text{LAGRANGE INTERPOLATION: } p(x) = f(x_1)\cdot \ell_1(x) + \dots + f(x_n)\cdot \ell_n(x)}$

e.g. $\exp(x)$ at $0,1,2$: $p(x) = \ell_1(x) + e\, \ell_2(x) + e^2 \ell_3(x)$

$$= \frac{(x-1)(x-2)}{(-1)(-2)} + e \frac{x(x-2)}{(-1)} + e^2 \frac{x(x-1)}{2}.$$

INTERPOLATORY QUADRATURE RULE: $\int_a^b f(x) \cdot w(x)\, dx = \sum_{j=1}^{n} w_j \cdot f(x_j)$

where, $w_j = \int_a^b \ell_j(x) \cdot w(x)\, dx$

GAUSSIAN QUADRATURE is interpolatory quadrature rule at roots of OP, $q_n(x)$ where $f(x) = \sum_{0}^{n-1} c_k q_k(x)$.

GAUSS QUADRATURE: $\displaystyle \underline{\int_a^b f(x)\,w(x)\,dx \approx \sum_{j=1}^{n} w_j \cdot f(x_j)}$

where $\underline{\displaystyle w_j = \frac{1}{q_0(x_j)^2 + \dots + q_{n-1}(x_j)^2}}$, ← $\underline{q_i(x)\text{ ORTHONORMAL}}$

TRUNCATED JACOBI MATRIX: $\underline{J_n = \begin{pmatrix} a_0 & b_0 & & \\ b_0 & \ddots & \ddots & \\ & \ddots & a_{n-2} & b_{n-2} \\ & & b_{n-2} & a_{n-1} \end{pmatrix}}$ ← symmetric replacing c_n with b_n. From a_0 to a_{n-1}.

↳ roots of OP are eigenvalues of truncated Jacobi matrix.

e.g. **INTERPOLATION via QUADRATURE** (for chebyshev), $n=3$

↳ $\displaystyle w_j = \frac{1}{q_0(x_j)^2 + q_1(x_j)^2 + q_2(x_j)^2}$, $x_0 = -\frac{\sqrt{3}}{2}$, $x_1 = 0$, $x_2 = \frac{\sqrt{3}}{2}$.

normalise

$$= \frac{1}{\frac{1}{\pi} + \frac{2}{\pi}x_j^2 + \frac{2}{\pi}(2x_j^2-1)} = \frac{\pi}{3} \quad \text{for all roots}$$

setup $\underline{Q = \begin{array}{c} \begin{array}{ccc} x_0 & x_2 & x_3 \end{array} \\ \begin{pmatrix} w_0 q_0(x_1) & w_2 q_0(x_2) & w_3 q_0(x_3) \\ w_1 q_1(x_1) & w_2 q_1(x_2) & w_3 q_1(x_3) \\ w_1 q_2(x_1) & w_2 q_2(x_2) & w_3 q_2(x_3) \end{pmatrix} \end{array}}$

then if we want to expand x^2 in $q(x)$, $\underline{Q\begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{pmatrix} = \begin{pmatrix} \text{coeff of } q_0(x) \\ \text{coeff of } q_1(x) \\ \text{coeff of } q_2(x) \end{pmatrix}}$,