Numerical Analysis formulae

Page No.: Date: youvn

LAGRANGE'S INTERPOLATION TECHNIQUE

- **Theorem (by Lagrange):**

  Take $n+1$ distinct points $x_0, x_1, \ldots, x_n \in \mathbb{R}$ and take real nos. $f_0, f_1, \ldots, f_n$.
  Then $\exists$ a UNIQUE Polynomial $p \in \mathbb{P}_n$ such that $p(x_i) = f_i$ for $i = 0, 1, \ldots, n$.

- **Fundamental Theorem of Algebra:**

  A polynomial of degree n can have at most n real zeros (or exactly n complex zeros)

Polynomial Interpolation — Recipe of Lagrange:

- Start with the $(n+1)$ given pairs
  $(x_0, f_0), (x_1, f_1), (x_2, f_2), \ldots, (x_n, f_n)$

- Build the $n+1$ Lagrange polynomials, each of degree n:
  $$\mathcal{L}_k^n(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{(x - x_j)}{(x_k - x_j)}$$

- Construct the polynomial interpolant $p \in \mathbb{P}_n$ as a weighted sum:
  $$p(x) = \sum_{k=0}^{n} f_k \mathcal{L}_k^n(x)$$

ERROR COMMITTED BY POLYNOMIAL APPROX.

- To measure closeness between two graphs f and g, we use $\max_{x \in [a,b]} |f(x) - g(x)|$

  Eg) $\max_{x \in [a,b]} |f|$ measures how far the fn is from zero.

- We define "Maximum Norm" $\|\cdot\|$ as:
$$\|f\| = \max_{x \in [a,b]} |f(x)|$$

- Weierstrass Approximation Theorem:

  Take a fn. $f \in C[a,b]$.
  Given a real no. $\epsilon > 0$, $\exists$ a polynomial p such that $\|f - p\| < \epsilon$

(not a constructive proof)
In order to find p, we pick points on x-axis (using Chebychev's idea; mostly preferred), then construct p using Lagrange's recipe.

- Theorem:

  Take $f \in C^{n+1}[a,b]$.
  Let $x_0, x_1, \ldots, x_n$ be $n+1$ distinct points in $[a,b]$. And let $p \in \mathbb{P}_n$ be such that
  $$p(x_i) = f(x_i) \quad \text{for } i = 0, 1, \ldots, n.$$
  Then $\forall x \in [a,b]$, $\exists$ a point $\xi_x \in [a,b]$ such that
  $$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{k=0}^{n} (x - x_k)$$

(Thus Error seems to depend on interpolation pts. x_k)

Using the theorem:

$$\max_{x\in[a,b]} |f(x)-p(x)| \le \frac{1}{(n+1)!} \|f^{(n+1)}\| \max_{x\in[a,b]} \prod_{k=0}^{n} |x-x_k|$$

∴ the error depends only on the no. of interpolation pts. and

$$\max_{x\in[a,b]} \prod_{k=0}^{n} |x-x_k|$$

NEWTON'S DIVIDED DIFFERENCES

Suppose $f \in C[a,b]$. Let $p \in \mathbb{P}^n$ be the polynomial that agrees with f at $x_0, x_1, \ldots, x_n$.
Objective is to determine n+1 coefficients $a_0, a_1, \ldots, a_n$ of

$$p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \ldots + a_n(x-x_0)\cdots(x-x_{n-1})$$

$$p(x_i) = f(x_i) \quad \text{for } i = 0, 1, \ldots, n$$

Clearly, $a_0 = p(x_0) = f(x_0)$

$$a_1 = \frac{f(x_1)-f(x_0)}{x_1-x_0}$$

Definition of divided difference:

Give distinct points $x_0, x_1, \ldots, x_n$ in $[a,b]$. Let $p \in \mathbb{P}^n$ be the polynomial that interpolates f at those points. The coefficient of x^n in p is called the "divided difference" i.e. $f[x_0, x_1, \ldots, x_n]$

$$f[x_0, x_1, \ldots, x_n] = \sum_{k=0}^{n} f(x_k) \prod_{\substack{j=0 \\ j\neq k}}^{n} \frac{1}{(x_k - x_j)}$$

Page No.:
Date:

Recurrence Relation for divided difference :

$$f[x_0, x_1, \cdots, x_{m+1}] = \frac{f[x_1, x_2, \cdots, x_{m+1}] - f[x_0, x_1, \cdots, x_m]}{x_{m+1} - x_0}$$

Thus,

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, x_1, \cdots, x_n] \prod_{k=0}^{n-1} (x - x_k)$$

satisfying $P_n(x_i) = f(x_i)$ for $i = 0, 1, \cdots, n$.

Divided difference table :-

$f[x_0]$	$f[x_0, x_1]$	$f[x_0, x_1, x_2]$	$f[x_0, x_1, \cdots, x_n]$
$f[x_1]$	$f[x_1, x_2]$		
$\vdots$	$\vdots$		
$f[x_n]$	$f[x_{n-1}, x_n]$		

OPERATION COUNT

Lagrange : $p(x) = f(x_0) \prod_{j \neq 0} \frac{(x - x_j)}{(x_0 - x_j)} + \cdots + f(x_n) \prod_{j \neq n} \frac{(x - x_j)}{(x_n - x_j)}$

Addition	Subtraction	Multiplication	Division
n	$2n(n+1)$	$2n^2 + n - 1$	$n+1$

Newton : $p(x) = f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \cdots, x_n]$.

Addition	Subtraction	Multiplication	Division	$\prod_{0}^{n-1} (x - x_j)$
n	$\frac{3n(n+1)}{2}$	$\frac{n(n+1)}{2}$	$\frac{n(n+1)}{2}$	

PIECEWISE / SPLINE INTERPOLATION

To interpolate a $f \in C[a,b]$.
Pick some points from $[a,b]$

$$a = x_0 < x_1 < x_2 \cdots < x_n = b$$

$\therefore$ Data points are $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$

Fix an integer $m \leq n$
On each subinterval $[x_{i-1}, x_i]$, build $\psi \in C[a,b]$ such that $\psi(x_{i-1}) = f_{i-1}$, $\psi(x_i) = f_i$

$$\psi(x) = \begin{cases} a_0^1 + a_1^1 x + a_2^1 x^2 + \cdots + a_m^1 x^m & \text{on } [x_0, x_1] \\ a_0^2 + a_1^2 x + a_2^2 x^2 + \cdots + a_m^2 x^m & \text{on } [x_1, x_2] \\ \vdots \\ a_0^n + a_1^n x + a_2^n x^2 + \cdots + a_m^n x^m & \text{on } [x_{n-1}, x_n] \end{cases}$$

considering the sub-intervals $[x_{i-1}, x_i]$ for $i = 1, 2, \cdots, n$

Hence, there are $n \cdot (m+1)$ coeffs to be determined:

- $n+1$ interpolation conditions:

$$\psi(x_i) = f(x_i) = f_i \quad \text{for } i = 0, 1, \cdots, n$$

- continuity of derivatives at interior pts:

$$\lim_{h \to 0^+} \psi(x_i - h) = \lim_{h \to 0^+} \psi(x_i + h) \quad i = 1, \cdots, n-1$$

$$\lim_{h \to 0^+} \frac{d\psi}{dx}(x_i - h) = \lim_{h \to 0^+} \frac{d\psi}{dx}(x_i + h) \quad "$$

$$\vdots$$

$$\lim_{h \to 0^+} \frac{d^{m-1}\psi}{dx^{m-1}}(x_i - h) = \lim_{h \to 0^+} \frac{d^{m-1}\psi}{dx^{m-1}}(x_i + h) \quad "$$

$\therefore$ ~~$n(m+1) = (m+1)$~~ $(n-1)\,m$ conditions.

Hence, total of $(n+1)n + (1-n)$ conditions
In need of $(n-1)$ more conditions.

ERROR BOUND IN LINEAR SPLINES

Theorem:

Take $f \in C^2[a,b]$.
Let $S_L(x)$ be the linear spline interpolating f at the $(n+1)$ knots.

$$a = x_0 < x_1 \cdots < x_n = b$$

Let $h > 0$ be defined as

$$h = \max_{1 \le i \le n} h_i = \max_{1 \le i \le n} (x_i - x_{i-1})$$

Then the following error bound holds true:

$$\|f - S_L\| \le \frac{h^2}{8} \|f''\|$$

Note: "Natural" cubic splines are those in which the second derivatives at endpoints of domain, i.e. $S_0''(x_0) = S_{n-1}''(x_n) = 0$

NATURAL CUBIC SPLINES

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

From condition (i), $d_i = f_i$, $i = 0, 1, \cdots, n-1$

Define new variables

$$\sigma_i = S''(x_i) \quad \text{for } i = 0, 1, \cdots, n$$

For the case of equally spaced interpolation points,

$$h = x_{i+1} - x_i$$

Since the spline is natural,

$$\sigma_0 = \sigma_n = 0$$

Note that $\sigma_i = s''(x_i) = 2b_i$

i.e. $b_i = \frac{\sigma_i}{2}$

Using Condition (v),

$$s_i''(x_{i+1}) = s_{i+1}''(x_{i+1})$$

$$\therefore \sigma_{i+1} = 6a_i h + 2b_i$$

$$\therefore a_i = \frac{\sigma_{i+1} - \sigma_i}{6h}, \quad i = 0, 1, \ldots, n-1$$

$\therefore$ As of now, $f_{i+1} = \left(\frac{\sigma_{i+1} - \sigma_i}{6h}\right) h^3 + \frac{\sigma_i}{2} h^2 + c_i h + f_i$

This yields $\frac{f_{i+1} - f_i}{h} = \left(\frac{\sigma_{i+1} + 2\sigma_i}{6}\right) h + c_i$

$$\Rightarrow c_i = \frac{f_{i+1} - f_i}{h} - \frac{h}{6}(2\sigma_i + \sigma_{i+1})$$

Using Condition (iv),

$$s_i'(x_{i+1}) = s_{i+1}'(x_{i+1})$$

$$3a_i h^2 + 2b_i h + c_i = c_{i+1}$$

Finally we get: $\sigma_{i-1} + 4\sigma_i + \sigma_{i+1} = \frac{6}{h^2}(f_{i-1} - 2f_i + f_{i+1})$

for $i = 1, \ldots, n-1$

$$\begin{pmatrix} 4 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 4 & 1 & & \vdots & \vdots \\ 0 & 1 & 4 & 1 & \vdots & \vdots \\ \vdots & \vdots & & \ddots & 0 & 0 \\ 0 & \vdots & & & 1 & 0 \\ 0 & 0 & \cdots & & 1 & 4 & 1 \\ 0 & 0 & \cdots & & 1 & 4 \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \vdots \\ \sigma_{n-1} \end{pmatrix} = \frac{6}{h^2} \begin{pmatrix} f_0 - 2f_1 + f_2 \\ f_1 - 2f_2 + f_3 \\ \vdots \\ f_{n-2} - 2f_{n-1} + f_n \end{pmatrix}$$

↓

Diagonally Dominant

Invertible

$$a_i = f_i$$

$$b_i = \frac{\sigma_i}{2}$$

$$a_i = \frac{\sigma_{i+1} - \sigma_i}{6h}$$

$$c_i = \frac{f_{i+1} - f_i}{h} - \frac{h}{6}(2\sigma_i + \sigma_{i+1})$$

ERROR ESTIMATE IN NATURAL CUBIC SPLINE

Let $f \in C^4[a, b]$

Let $s \in C^2[a, b]$ be the natural cubic spline that interpolates f at $(n+1)$ equally - spaced knots.

$$a = x_0 < x_1 \cdots < x_n = b$$

$$h = |x_{i+1} - x_i| \quad , i = 0, 1, \cdots, n-1$$

Then the following error bound holds true:

$$\|f - s\| \leq \|f^{(iv)}\| h^4$$

NUMERICAL INTEGRATION

Given a real-valued fn. f, compute

$$\int_a^b f(x)\,dx$$

Best method is to find antiderivative F then do $F(b) - F(a)$. But that may not always be possible.

NUMERICAL QUADRATURE

Partition $[a, b]$ into n parts

$$a = x_0 < x_1 < x_2 \cdots < x_n = b$$

Interpolate f at these $n+1$ points by some $p \in \mathbb{P}^{\wedge}$.

$$\int_a^b f(x)\,dx \approx \int_a^b p(x)\,dx$$

NEWTON-COTES INTEGRATION FORMULAE

$n+1$ equally spaced partitioning points.

$$p(x) = \sum_{i=0}^{n} f(x_i)\, l_i(x), \quad l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{(x - x_k)}{(x_i - x_k)}$$

$x_i = a + ih$

Introduce variable t such that $x = a + th$

$$t = \frac{x-a}{h} = n\,\frac{x-a}{b-a} \in [0, n]$$

$$\therefore \frac{x - x_k}{x_i - x_k} = \frac{t-k}{i-k}$$

$$\therefore l_i(x=t) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{t-k}{i-k} = \varphi_i(t)$$

$$\int_a^b L_i(x)\,dx = h \int_0^n \varphi_i(t)\,dt$$

Quadrature Weights:

$$w_i = \int_0^n \varphi_i(t)\,dt$$

$$\therefore \int_a^b f(x)\,dx \approx h \sum_{i=0}^{n} w_i f(x_i)$$

Notice that $w_i \in \mathbb{Q}$ and depend only on n.

$w_i \not\leftrightarrow f$

$w_i \not\leftrightarrow a, b$ i.e. boundaries of integral.

$$\sum_{i=0}^{n} w_i = n$$

Trapezium Rule ($n=1$)

$$w_0 = \int_0^1 \frac{t-1}{0-1}\,dt = \left(\frac{-t^2}{2} + t\right)\Big|_0^1 = \frac{1}{2}$$

$$w_1 = \int_0^1 \frac{t-0}{1-0}\,dt = \frac{1}{2}$$

$$\therefore \int_a^b f(x)\,dx \approx \frac{b-a}{2}\left(f(a)+f(b)\right)$$

Simpson's Rule ($n=2$)

$$w_0 = \int_a^b \frac{(t-1)(t-2)}{(0-1)(0-2)}\,dt = \frac{1}{3}$$

$$w_1 = 4/3$$

$$w_2 = 1/3$$

$$\int_a^b f(x)dx \approx \frac{h}{3}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$$

Quadrature Weights

n	w_0	w_1	w_2	w_3	w_4
1	1/2	1/2			
2	1/3	4/3	1/3		
3	3/8	9/8	9/8	3/8	
4	14/45	64/45	24/45	64/45	14/45

$$I_{p_1} = \frac{h}{2}(f(a) + f(b)) \quad \{\text{Trapezoid Rule}\}$$

$$I_{p_2} = \frac{h}{3}(f(a) + 4f(a+h) + f(b)) \quad \{\text{Simpson's Rule}\}$$

$$I_{p_3} = \frac{3h}{8}(f(a) + 3f(a+h) + 3f(a+2h) + f(b))$$

$$I_{p_4} = \frac{h}{45}(14f(a) + 64f(a+h) + 24f(a+2h) + 64f(a+3h) + 14f(b))$$

{Dividing $[a,b]$ in 4 parts}

Error Estimate in Newton-Cotes

Take $f \in C^{n+1}[a,b]$

$$|I_f - I_{p_n}| \leq \frac{1}{(n+1)!}\|f^{(n+1)}\| \int_a^b \prod_{i=0}^{n} |x - x_i|\, dx$$

- Hence, error in trapezoidal rule $\leq \frac{1}{2}\|f''\|\frac{(b-a)^3}{6}$

- Newton Cotes formula need not give convergence as $n \to \infty$, $\because$ weights may be negative.
Hence, we use Gaussian quadrature $G_n(f) =$

$$W_i := \int_a^b [l_i(x)]^2 dx = \int_a^b \prod_{\substack{k=0 \\ k \neq i}}^{n} \left(\frac{x - x_k}{x_i - x_k}\right)^2 dx$$

Here quadrature points are not equally spaced.
Gaussian quadrature weights are non-negative.

Gauss quadrature - Convergence Theorem:
Let $f \in C[a,b]$
Let $G_n(f)$ be the Gaussian quadrature formula of order n. Then $\lim_{n\to\infty} |G_n(f) - I_f| = 0$

- Composite Newton-Cotes

alternative to building Newton Cotes for large n.
- Apply Newton-Cotes to small set of quadr. points.
- Only requirement is Continuity.

- Composite Trapezoidal Rule:

For $m=2$: $h\left(\frac{1}{2}f(a) + \frac{1}{2}f(a+h)\right) + h\left(\frac{1}{2}f(a+h) + \frac{1}{2}f(b)\right)$

$= h\left(\frac{1}{2}f(a) + f(a+h) + \frac{1}{2}f(b)\right)$

For $m \geq 2$: $h = \frac{b-a}{m}$

$\therefore C_{P_1}(f) = h\left(\frac{1}{2}f(a) + f(a+h) + f(a+2h) + \cdots + f(a+(m-1)h) + \frac{1}{2}f(b)\right)$

- Composite Simpson's $\frac{1}{3}$ Rule:

Divide $[a,b]$ into $2m$ subintervals with $m \geq 2$

For $m=2$: $\frac{h}{3}\left(f(a) + 4f(a+h) + f(a+2h)\right) + \frac{h}{3}\left(f(a+2h) + 4f(a+3h) + f(b)\right)$

$h = \frac{b-a}{2m}$

For $m \geq 2$: $C_{P_2(f)} = \frac{h}{3}\left(f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + 2f(a+4h) + \cdots + 2f(a+(2m-2)h) + 4f(a+(2m-1)h) + f(b)\right)$

- Error estimate for Composite Trapezo: $|C_{P_1}(f) - I_f| \leq \frac{m}{12}\|f''\|h^3 = \frac{b-a}{12}\|f''\|h^2$. $\therefore h \to 0$ is preferred.

NUMERICAL SOLUTIONS TO ODEs

- Peano's Theorem for existence of a solution for $y' = f(t,y)$, $y(0) = y_0$:

  Suppose $f: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.
  f is continuous in $(-\delta, \delta) \times (y_0 - \eta, y_0 + \eta)$
  for some $\delta > 0$, $\eta > 0$
  $\Rightarrow \exists\ \varepsilon > 0$ and a function $y \in C^1(-\varepsilon, \varepsilon)$!
  such that $y' = f(t,y)$, $y(0) = y_0$

- Lipschitz Continuity

  A function $g: [a,b] \to \mathbb{R}$ is said to be Lipschitz continuous if

  $$|g(x) - g(y)| \leq L|x-y| \quad \forall\ x, y \in [a,b]$$

  with some positive constant L.

- Cauchy - Lipschitz Uniqueness Theorem

  Suppose $f(t,y)$ be continuous in both variables and let it be Lipschitz Continuous in y variable. Then $\exists$ an $\varepsilon > 0$ and a unique function $y \in C^1(-\varepsilon, \varepsilon)$ such that

  $$y' = f(t,y), \quad y(0) = y_0$$

$$y' = f(t, y)$$

$$y(t) = y_0 + \int_0^t f(s, y(s))\, ds \qquad \forall\, t \geq 0$$

Numerical Scheme

Define mesh points $t_n = nh$

$y_n \approx y(t_n)$

Euler Method (Explicit method)

$$y_{n+1} = y_n + h f(t_n, y_n)$$

$$f(s, y(s)) \approx f(t_n, y_n)$$

Trapezoidal Method (implicit method)

$$y_{n+1} = y_n + \frac{h}{2}\left(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\right)$$

$$f(s, y(s)) \approx \frac{f(t_n, y_n) + f(t_{n+1}, y_{n+1})}{2}$$

{i.e. $y_{n+1} - \frac{h}{2} f(t_{n+1}, y_{n+1}) = y_n - \frac{h}{2} f(t_n, y_n)$}

ORDER OF A NUMERICAL METHOD

Numerical Method given by the recurrence relation $y_{n+1} = F(t, f, y_0, y_1, \cdots, y_n, y_{n+1})$.
The method is of order p if

$$y(t_{n+1}) - F(t, f, y(t_1), y(t_2), \cdots, y(t_{n+1})) = Ch^{p+1}$$

where $y(t)$ is the exact solution to the ODE $y' = f(t, y)$

- TAYLOR SERIES OF g around pt. a is given by

$$g(t) = g(a) + \frac{g'(a)(t-a)}{1!} + \frac{g''(a)(t-a)^2}{2!} + \cdots + \frac{g^{(n)}(a)(t-a)^n}{n!} + \underbrace{\frac{g^{(n+1)}(\xi_t)(t-a)^{n+1}}{(n+1)!}}_{R_n(t)}$$

for some $\xi_t \in [a, t]$

$R_n(t) \equiv$ remainder term

Eg) Euler's Method is of order 1.
Eg) Trapezoidal Method is of order 2.

Note: Method of order p recovers every polynomial solution of degree p or less.

ERRORS IN SOLUTIONS

- CONVERGENT Numerical Method $\Rightarrow \lim_{h \to 0^+} \max_{n=0,1,\cdots,\lfloor T/h \rfloor} |e_{n,h}| = 0$

- In Euler's Method, local truncation error $\leq \frac{h^2}{2} y''(\xi_t)$

if $M = \max_{\xi_t \in [0,T]} y''(\xi_t)$ $\Rightarrow$ local trunc. err. $\leq M\frac{h^2}{2}$

Also, $|e_{n,h}| \leq \frac{Ch}{L}\left((1+Lh)^n - 1\right)$, $n = 0, 1, \cdots, \lfloor T/h \rfloor$

ONE-STEP METHOD: approximation at n^{th} time step depends on approximation at $(n-1)^{th}$ time step.

eg: Euler, Trapezoidal

TWO-STEP METHOD: approximation at $(n+1)^{th}$ time step depends on approximation at n^{th} & $(n-1)^{th}$ time steps.

$$y(t_{n+1}) - y(t_{n-1}) = \int_{t_{n-1}}^{t_{n+1}} f(s, y(s))\,ds$$

eg: Simpson's $\frac{1}{3}$ rule

eg) Adams-Bashforth Method:

$$y_{n+4} = y_{n+3} + \frac{h}{24}\left(55 f_{n+3} - 59 f_{n+2} + 37 f_{n+1} - 9 f_n\right)$$

→ 4-step
→ Explicit

eg) Adams-Moulton Method:

$$y_{n+3} = y_{n+2} + \frac{h}{24}\left(9 f_{n+3} + 19 f_{n+2} - 5 f_{n+1} - 9 f_n\right)$$

→ 3-step
→ implicit.

- General s-STEP method: $\sum_{k=0}^{s} \alpha_k y_{n+k} = h \sum_{k=0}^{s} \beta_k f(t_{n+k}, y_{n+k})$

where each α_i and β_i is given.

$\beta_s = 0 \Leftrightarrow$ Explicit method
$\beta_s \neq 0 \Leftrightarrow$ Implicit method

- Adam's recipe for s-Step Method

Suppose we have the following approximations:

$$y_n, y_{n+1}, \ldots, y_{n+s-1}$$

$$y(t_{n+s}) = y(t_{n+s-1}) + \int_{t_{n+s-1}}^{t_{n+s}} f(\tau, y(\tau))\, d\tau$$

Interpolate $f(t_i, y_i)$ with $\psi(t_i)$

$$\psi(t) = \sum_{k=0}^{s-1} \psi_k(t) f(t_{n+k}, y_{n+k})$$

$$\psi_k(t) = \prod_{\substack{l=0 \\ l \neq k}}^{s-1} \left(\frac{t - t_{n+l}}{t_{n+k} - t_{n+l}} \right)$$

So $y(t_{n+s}) = y(t_{n+s-1}) + \sum_{k=0}^{s-1} \int_{t_{n+s-1}}^{t_{n+s}} \psi_k(\tau) f(t_{n+k}, y_{n+k})\, d\tau$

$$y_{n+s} = y_{n+s-1} + h \sum_{k=0}^{s-1} \beta_k f(t_{n+k}, y_{n+k})$$

$$\beta_k = \frac{1}{h} \int_{t_{n+s-1}}^{t_{n+s}} \psi_k(\tau)\, d\tau = \frac{1}{h} \int_0^h \psi_k(t_{n+s-1} + \tau)\, d\tau$$

- Adam's recipe for 2-step method:

$$y_{n+2} = y_{n+1} + h\beta_0 f(t_n, y_n) + h\beta_1 f(t_{n+1}, y_{n+1}) \quad \text{for } n = 0, 1, \ldots$$

$$\beta_0 = \frac{1}{h} \int_{t_n}^{t_{n+1}} \left(\frac{\tau - t_{n+1}}{t_n - t_{n+1}} \right) d\tau$$

$$\beta_1 = \frac{1}{h} \int_{t_n}^{t_{n+1}} \left(\frac{\tau - t_n}{t_{n+1} - t_n} \right) d\tau$$

- Order of general s-step method:

method: $\sum_{k=0}^{s} \alpha_k y_{n+k} = h \sum_{k=0}^{s} \beta_k f(t_{n+k}, y_{n+k})$ for $n = 0, 1, \ldots$

Method is of order $p \geq 1$ iff:

$$\sum_{k=0}^{s} \alpha_k y(t_{n+k}) - h \sum_{k=0}^{s} \beta_k f(t_{n+k}, y(t_{n+k})) = O(h^{p+1})$$

- Theorem: Order of method is $p \geq 1$ ~~only if~~ then we should have:

$$\sum_{k=0}^{s} \alpha_k = 0$$

$$\sum_{k=0}^{s} k^m \alpha_k = \sum_{k=0}^{s} k^{m-1} \beta_k \quad \text{for } m = 1, \ldots, p$$

$$\sum_{k=0}^{s} k^{p+1} \alpha_k \neq (p+1) \sum_{k=0}^{s} k^p \beta_k$$

- Convergence of s-STEP method:

Define Ist characteristic polynomial $\rho(z) = \sum_{k=0}^{s} \alpha_k z^k$

" IInd " " $\sigma(z) = \sum_{k=0}^{s} \beta_k z^k$

- Dahlquist Equivalence Theorem:

An s^- step method is convergent iff

* order of method p is ≥ 1
* roots of $\rho(z)$ lie in the closed unit disk $\{z \in \mathbb{C} \mid |z| \leq 1\}$ with those lying on $|z| = 1$ being SIMPLE.

Page No.:
Date:

Runge-Kutta Methods

$$y' = f(t, y(t)) \quad , \quad y(0) = y_0$$

Integrating over $(0, h)$ yields

$$y(h) - y_0 = \int_0^h f(s, y(s))\, ds$$

Change of variable $s = h\tau$ yields

$$y(h) - y_0 = h \int_0^1 f(h\tau, y(h\tau))\, d\tau$$

We approximate: $\int_0^1 f(h\tau, y(h\tau))\, d\tau \approx \sum_{i=1}^{v} b_i f(c_i h, y(c_i h))$
via quadrature formula

We don't know value of $y(c_i h)$.
So approximate: $\xi_i \approx y(c_i h)$ for $i = 1, \ldots, v$

For $i = 1$, we take: $c_1 = 0$ i.e. $\xi_1 = y_0$
For $i = 2$, we take: $\xi_2 = \xi_1 + a_{2,1} f(0, \xi_1)$
For $i = 3$, we take: $\xi_3 = y_0 + h a_{3,2} f(0, \xi_1) + h a_{3,2} f(c_2 h, \xi_2)$

$$\vdots$$

$$\xi_v = y_0 + h \sum_{i=1}^{v-1} a_{v,i} f(c_i h, \xi_i)$$

$\therefore$ Problem reduces to "v-stage Runge-Kutta Method":

$$y_1 = y_0 + h \sum b_i f(c_i h, \xi_i)$$

where $\xi_i = y_0 + h \sum_{j=1}^{i-1} a_{i,j} f(c_j h, \xi_j)$

Here $A = (a_{ij})$ is called RK matrix;
$b_i \rightarrow$ RK weights
$c_i \rightarrow$ RK nodes

Generalizing: $y_{n+1} = y_n + h \sum_{i=1}^{v} b_i f(t_n + c_i h, \xi_i)$

$$\xi_i = y_n + h \sum_{j=1}^{i-1} a_{i,j} f(t_n + c_j h, \xi_j)$$

$a_{i,j}$, b_i & c_i are same for all subintervals $[t_n, t_{n+1}]$

* Comparing with Taylor Expansion of $y(t_{n+1})$

$$y(t_{n+1}) = y(t_n) + hf(t_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_n, y(t_n)) + f(t_n, y(t_n))\frac{\partial f}{\partial y}(t_n, y(t_n))\right) + \mathcal{O}(h^3)$$

Example : 2-stage Runge-Kutta Method

$v = 2$. We need to find $c_2, b_1, b_2, a_{2,1}$ $(c_1 = 0)$

We will find these constants imposing the order is ≥ 2.

$$y_{n+1} = y_n + hb_1 f(t_n, y_n) + hb_2 f(t_n + c_2 h, \xi_2)$$

where $\xi_2 = \xi_1 + ha_{2,1} f(t_n, y_n) = y_n + ha_{2,1} f(t_n, y_n)$

Employing Taylor series in 2-variables,

$$f(t_n + c_2 h, \xi_2) = f(t_n + c_2 h, y_n + ha_{2,1} f(t_n, y_n))$$
$$= f(t_n, y_n) + h\left(c_2 \frac{\partial f}{\partial t}(t_n, y_n) + a_{2,1} f(t_n, y_n)\frac{\partial f}{\partial y}(t_n, y_n)\right) + \mathcal{O}(h^2)$$

Substituting this we get:

$$y_{n+1} = y_n + hb_1 f(t_n, y_n) + hb_2 f(t_n + c_2 h, \xi_2)$$
$$= y_n + hb_1 f(t_n, y_n) + hb^2\left(f(t_n, y_n) + h\left(c_2 \frac{\partial f}{\partial t} + a_{2,1} f \frac{\partial f}{\partial y}\right) + \mathcal{O}(h^2)\right)$$
$$= y_n + h(b_1 + b_2) f(t_n, y_n) + h^2 b_2\left(c_2 \frac{\partial f}{\partial t}(t_n, y_n) + a_{2,1} f(t_n, y_n)\frac{\partial f}{\partial y}(t_n, y_n)\right) + \mathcal{O}(h^3)$$

* order 2 $\Rightarrow$ $b_1 + b_2 = 1$, $a_{2,1} = c_2$, $b_2 c_2 = 1/2$

We can take $b_1 = b_2 = 1/2$, $a_{2,1} = c_2 = 1$

∴ Solution: $\boxed{y_{n+1} = y_n + \frac{h}{2}\left(f(t_n, y_n) + f(t_n + h, y_n + hf(t_n, y_n))\right)}$

- RK Tableau :

c	A
	b^T

2-stage Runge Kutta :

0	
1	1
	$\frac{1}{2}$ $\frac{1}{2}$

4-stage RK :

0				
1/2	1/2			
1/2	0	1/2		
1	0	0	1	
	1/6	1/3	1/3	1/6

NUMERICAL SOLUTIONS TO SYSTEM OF LINEAR EQUATIONS

$A\mathbf{x} = \mathbf{b}$ where $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$ and $A \in M_{m,n}(\mathbb{R})$

- Underdetermined System ($m < n$)
  - NO solution
  - Infinite solutions
- Overdetermined System ($m > n$)
  - No solution
  - Solutions Exist (if no. of lin. indep. equations $\leq$ no. of unknowns)
- Square System ($m = n$)
  - A is invertible : Unique soln.
  - A is not invertible
    - no soln.
    - infinite solns.

{Note: A is invertible iff null space $\text{null space}(A) = \{0\}$}

- In square system, if A is invertible, unique solution given by $x = A^{-1}b$

- Cramer's Formula : $x_i = \dfrac{\det(A_i)}{\det(A)}$ for $i = 1, \ldots, n$

  where A_i is formed by replacing i^{th} column of A by b.

Special Case: A is upper Δ^{lar} then solve for x by back-substitution.

$$x_i = \frac{b_i - \sum_{j=i+1}^{N} a_{ij} x_j}{a_{ii}}$$

i.e. $x_n = \dfrac{b_n}{a_{nn}}$, $x_{n-1} = \dfrac{b_{n-1} - a_{n-1,n} x_n}{a_{n-1,n-1}}$

(Similarly for lower Δ^{lar} matrices)

- Gaussian Elimination is a direct method to solve the system.

→ Transform given system to an upper Δ^{lar} system by Elimination.

i.e. $Ax = b$. We want to find a matrix B such that BA is upper Δ^{lar} then solve the system $BAx = Bb$ by back-substitution.

- An Existence Theorem :

A → $n \times n$ matrix

$\exists$ at least one non-singular matrix B such that BA is upper Δ^{lar}.

- Diagonal Submatrix (definition) :

matrix $A = (A_{ij})$

Δ^k = diagonal submatrix of order k.

$$= \begin{pmatrix} A_{11} & \cdots & A_{1k} \\ \vdots & & \\ A_{k1} & \cdots & A_{kk} \end{pmatrix}$$

Theorem :

Let A be a matrix s.t. all submatrices of order Δ^k $\forall$ k from $1, \ldots, n$ are invertible, then Gaussian elimination does not need pivoting strategy.

Elimination Algorithm : $A_{ij}^{(k+1)} = A_{ij}^{(k)} - m_{ik} A_{j}^{(k)}$

$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)}$

- LU factorization.

$$\mathbb{A} = \mathbb{L}\mathbb{U}$$

where $L_{ij} = \begin{cases} 1, & i=j \\ m_{ij}, & i>j \\ 0, & \text{otherwise} \end{cases}$

$\mathbb{U} \equiv \mathbb{A}^{(n)}$ (i.e. obtained at the end of Gaussian Elimination)

- When can we do LU factori. ?
Ans: When we can do Gaussian Elim. without pivoting.
$\therefore$ LU $\Leftrightarrow$ Gauss Elim. without pivoting.

- LU factori. is unique

- Algorithm :

$$U_{1j} = A_{1j}$$

$$U_{jj} = A_{jj} - \sum_{k=1}^{j-1} L_{jk} U_{kj}$$

$$L_{j+1,j} = \frac{A_{j+1,j} - \sum_{k=1}^{j-1} L_{j+1,k} U_{kj}}{U_{jj}}$$

$$L_{nj} = \frac{A_{nj} - \sum_{k=1}^{j-1} L_{nk} U_{kj}}{U_{jj}}$$

- Cholesky Factorization

$A = BB^T$ where B is lower Δ^{lar}

Applicable for :
- Real entries
- Symmetric A
- Positive definite A. i.e. $\langle Ay, y \rangle > 0 \ \forall \ y \neq 0$ (i.e. Eigen values > 0)

- Uniqueness Theorem (& Existence):
$A \rightarrow$ real, symmetric, +ve definite matrix.
$\exists$ a unique real lower Δ^{lar} matrix B with positive diagonal entries s.t. $A = BB^T$

- Positive-Definiteness test of a symmetric real matrix A:
Cholesky algorithm gives the required B iff A is also +ve definite.

- Algorithm:

$$B_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} B_{jk}^2}$$

$$B_{j+1,j} = \frac{A_{j+1,j} - \sum_{k=1}^{j-1} B_{j+1,k} B_{jk}}{B_{jj}}$$

$$B_{n,j} = \frac{A_{nj} - \sum_{k=1}^{j-1} B_{nk} B_{jk}}{B_{jj}}$$

- QR Factorization

$A = QR$

where Q is orthogonal i.e. $Q^{-1} = Q^T$ & R is upper Δ^{lar}

$Ax = b$ reduces to $Rx = Q^T b$ which can be solved by back-substitution.

- Existence Theorem:

Any square matrix A can be QR factorized.
if A is invertible then R is invertible.

- Uniqueness: If R is assumed to have positive diagonal entries then the pair (Q, R) is unique.

- Algorithm: (for invertible A)

Denote columns of A by $A_1, A_2, \ldots, A_n$.
Column vectors are linearly indep.

Build an orthonormal basis using Gram-Schmidt.

$$q_1 = \frac{A_1}{\|A_1\|}$$

$$q_2 = \frac{A_2 - \langle q_1, A_2 \rangle q_1}{\| \quad " \quad \|}$$

$$\vdots$$

$$q_i = \frac{A_i - \sum_{k=1}^{i-1} \langle q_k, A_i \rangle q_k}{\| \quad " \quad \|} \quad \text{for } 1 \le i \le n$$

Now make Q such that its i^{th} column is q_i.

To make R:

$$R_{ki} = \begin{cases} \langle q_k, A_i \rangle & , \ 1 \le k \le i-1 \\ \| A_i - \sum_{k=1}^{i-1} \langle q_k, A_i \rangle q_k \| & , \ k = i \\ 0 & , \ k > i \end{cases}$$

NORMS

We use norms to estimate the error between true solution (x) and approximate solution ($\tilde{x}$) to $Ax = b$

$e = x - \tilde{x}$

- **Norms on $\mathbb{R}^n$:**

Norm is a fn. $\|\cdot\| : \mathbb{R}^n \to [0, \infty)$ such that

$\|x\| > 0 \quad \forall\, x \neq 0$

$\|\alpha x\| = |\alpha| \|x\| \quad \forall\, \alpha \in \mathbb{R},\ x \in \mathbb{R}^n$

$\|x + y\| \leq \|x\| + \|y\|$

- ℓ^p-norm for $p \in [1, \infty)$: (norm is on $\mathbb{R}^n$)

$$\|x\|_{\ell^p} = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

- ℓ^∞-norm:

$$\|x\|_{\ell^\infty} = \max_{1 \leq i \leq n} |x_i|$$

- Equivalent norms:

Two norms $\|\cdot\|$ and $\|\cdot\|'$ are equivalent if $\exists\, m > 0$ and $\exists\, M > 0$ such that

$$m\|x\| \leq \|x\|' \leq M\|x\| \quad \forall\, x \in \mathbb{R}^n$$

- $\|\cdot\|_{\ell^\infty}$ and $\|\cdot\|_{\ell^p}$ norms are equivalent $\therefore$ $\|\cdot\|_{\ell^p}$ & $\|\cdot\|_{\ell^{p_2}}$ are equivalent

$$\|x\|_{\ell^\infty} \leq \|x\|_{\ell^p} \leq \sqrt[p]{n}\, \|x\|_{\ell^\infty}$$

- $\|\cdot\|_{\ell^1}$ and $\|\cdot\|_{\ell^2}$ are equivalent:

$$\|x\|_{\ell^2} \leq \|x\|_{\ell^1} \leq \sqrt{n}\, \|x\|_{\ell^2}$$

- All norms on $\mathbb{R}^n$ are equivalent.

- **Norms on $M_{m,n}(\mathbb{R})$**

- ℓ^p-norms:

  Let $A \in M_{m,n}(\mathbb{R})$.

  $$\|A\|_{\ell^p} = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |A_{ij}|^p \right)^{1/p} \quad \text{for } p \in [1, \infty)$$

  $$\|A\|_{\ell^\infty} = \max_{\substack{1 \le i \le m \\ 1 \le j \le n}} |A_{ij}|$$

- $L_{p,q}$-norms:

  $$\|A\|_{L_{p,q}} = \left(\sum_{j=1}^{n} \left(\sum_{i=1}^{m} |A_{ij}|^p \right)^{q/p} \right)^{1/q} \quad \text{for } p, q \in [1, \infty)$$

  → Collect ℓ^p-norm of each column vector of size m.
  → Take ℓ^q-norm of this collection.

- $\|A\|_{L_{p,p}} = \|A\|_{\ell^p}$

- $\|A\|_{\ell^2}$ is called Frobenius norm.
  $\|AB\|_{\ell^2} \le \|A\|_{\ell^2} \|B\|_{\ell^2}$

- **Matrix Norm**

A norm $\|\cdot\|$ defined on $M_n(\mathbb{R})$ is Matrix norm if $\forall A, B \in M_n(\mathbb{R})$,

$$\|AB\| \le \|A\| \|B\|$$

- $\|I_n\| \ge 1 \quad (\because \|A\| = \|AI_n\| \le \|A\| \|I_n\|)$
- Frobenius norm is matrix norm.
- $\|\cdot\|_{\ell^\infty}$ is not matrix norm. (Take A to have all ones. $\therefore A^2$ has all n)

- **Subordinate Matrix Norm:**
  Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$. It induces a norm on M_n defined by

$$\|A\| = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|}$$

This is referred to as norm on M_n subordinate to the vector norm $\|\cdot\|$.

- Can also be written as $\|A\| = \sup_{y \in \mathbb{R}^n, \|y\|=1} \|Ay\|$
- $\exists\, z \in \mathbb{R}^n \setminus \{0\}$ such that $\|A\| = \frac{\|Az\|}{\|z\|}$
- Subordinate norm is a matrix norm, i.e. $\|AB\| \leq \|A\|\|B\|$
- Subordinate matrix norm of I_n is 1 ($\because \|I_n\| = \sup \frac{\|I_n x\|}{\|x\|} = 1$)
- Frobenius norm of identity $= \sqrt{n}$ $\therefore$ it is not subordinate to any vector norm.
- Let $\|\cdot\|_p$ denote the matrix norm subordinate to any vector norm $\|\cdot\|_p$. $\|A\|_p = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_p}{\|x\|_p}$
- For a unitary matrix A (i.e. $A^{-1} = A^T$), $\|A\|_2 = 1$
- For any unitary matrix A and $B \in M_n$, we have
  $\|AB\|_2 = \|BA\|_2 = \|B\|_2$

for $\|\cdot\|_2$:

$\therefore$ $\|\cdot\|_2$-norm is invariant under multiplication by unitary matrices.

{ $\|\cdot\|_2$ norm is useful for theoretical computations, because of inner product structure }

- If A is a diagonal matrix: $A = \text{diag}(a_1, a_2, \ldots, a_n)$

$$\|Ax\|_2 = \max_{1 \leq i \leq n} |a_i|$$

NORMAL MATRICES

- Adjoint A^* of a matrix $A \in M(\mathbb{C})$ is
$$A^*_{ij} = \overline{A}_{ji}$$

- A is normal if $AA^* = A^*A$

- Theorem (Characterization of Normal Matrices):
Let $S \subset M_n(\mathbb{C})$ denote set of all normal matrices. Define set $T = \{A \in M_n(\mathbb{C}) \text{ s.t. } A = U \operatorname{diag}(\lambda_1, \cdots, \lambda_n) U^*\}$ with U unitary (i.e. $U^{-1} = U^*$) and λ_i eigenvalues of A. Then $S = T$.

- For a normal matrix A,
$$\|A\|_2 = \|U \operatorname{diag}(\lambda_1, \cdots, \lambda_n) U^*\|_2 = \|\operatorname{diag}(\lambda_1, \cdots, \lambda_n)\|_2$$
$$= \max_{1 \le i \le n} |\lambda_i|$$

SPECTRAL RADIUS AND MATRIX NORMS

- Spectral Radius:
Maximum of moduli of Eigen values of a matrix A is spectral radius of A: $\rho(A) = \max_{1 \le i \le n} |\lambda_i|$

- $\|A\|_2 = \rho(A)$ $\{A \text{ is normal}\}$ (A)
- ρ cannot be a norm $\because \rho_A = 0 \not\Rightarrow A = 0$
- For any "matrix norm" $\|\cdot\|$ on $M_n(\mathbb{C})$, $\rho(A) \le \|A\|$
- For a given $A \in M_n$ and any $\epsilon > 0$, $\exists$ a subordinate matrix norm $\|\cdot\|$ such that
$$\rho(A) \le \|A\| \le \rho(A) + \epsilon.$$
(The matrix norm depends on A, ϵ)

- Practical recipe for computing $\|\cdot\|_1$ of a given matrix:

$$\|A\|_1 = \max_{1 \le j \le n} \left(\sum_{i=1}^{n} |A_{ij}| \right)$$

(Hence $\|\cdot\|_1$ is also called column-sum norm)

- $\|A\|_\infty = \max_{1 \le i \le n} \left(\sum_{j=1}^{n} |A_{ij}| \right)$

(Row-sum norm).

- Practical upper bound for spectral radius:

$$\rho(A) \le \min(\text{column-sum norm}, \text{row-sum norm})$$

- HERMITIAN MATRIX: $A = A^*$
  - $\therefore$ Eigenvalue is always real.
  - $\therefore$ Normal matrix
  - Proposition: $A \in M_n(\mathbb{C})$ is Hermitian iff $\exists$ a unitary matrix U and real eigenvalues of A: $\lambda_1, \lambda_2, \cdots, \lambda_n$ such that
$$A = U \, \text{diag}(\lambda_1, \cdots, \lambda_n) \, U^*$$

  - For any $A \in M_n(\mathbb{C})$, A^*A is Hermitian & All eigenvalues of A^*A are real and non-negative.

- SINGULAR VALUES: of a matrix $A \in M_n(\mathbb{C})$ are the non-negative square roots of the n eigenvalues of A^*A {by convention, we do not mention '0' as a sing. val even if it is}.

  - Singular values of a normal matrix are the moduli of it's eigenvalues ($\because$ a normal matrix A can be written as $UDU^* = A$, where $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$
  - $\|A\|_2 = \max_{1 \le i \le n} \sqrt{\lambda_i} = \max(\sqrt{\text{eigenval. of } A^*A}) =$ largest singular val. of A

• MATRIX SEQUENCES

Convergence:

A sequence $\{A^{(k)}\}$ of matrices converges to a limit A if $\lim_{k\to\infty} \|A^{(k)} - A\| = 0$ for some norm on $M_n(\mathbb{C})$.

It is also denoted by $\lim_{k\to\infty} A^{(k)} = A$

- Choice of norm is irrelevant here because all norms are equivalent in finite dimensional space.

• Sequence of iterated powers: $A^{(k)} = A^k$

The following are equivalent :-

i) $\lim_{k\to\infty} A^k = 0$ → zero matrix

ii) $\lim_{k\to\infty} A^k x = 0 \quad \forall x \in \mathbb{C}^n$

iii) $\rho(A) < 1$

iv) $\exists$ at least one subordinate matrix norm $\|\cdot\|$ such that $\|A\| < 1$.

• Geometric matrix series:

$$S^{(0)} = I_n \; ; \; S^{(k)} = I_n + A + \cdots + A^k$$

If $\{S^{(k)}\}$ converges, we get geometric series $\sum_{k=0}^{\infty} A^k = \lim_{k\to\infty} S^{(k)}$

- Geometric series converges iff $\rho(A) < 1$.

If $\rho(A) < 1$, then $I_n - A$ is invertible and

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

- Any invertible matrix has a neighbourhood in which there are invertible matrices.

Take an invertible matrix $A \in M_n(\mathbb{C})$. Take a subordinate matrix norm $\|\cdot\|$. Then the matrix B, s.t. $\|A - B\| < \frac{1}{\|A^{-1}\|}$, is invertible.

* - $\text{cond}(A) = \|A\|\,\|A^{-1}\| \geq \rho(A)\,\rho(A^{-1})$
  The lower bound is attained if A is normal ($\|A\|_2 = \rho(A)$)
  $\text{cond}_2(A) = \rho(A)\,\rho(A^{-1}) = |\lambda_{max}|/|\lambda_{min}|$

Relative Error: Let $y \in \mathbb{C}^n$ and $\tilde{y} \in \mathbb{C}^n$ be its perturbed value. Relative error wrt norm $\|\cdot\|$ on $\mathbb{C}^n$ is defined as $\dfrac{\|\tilde{y} - y\|}{\|y\|}$

Condition Number: of an invertible matrix $A \in M_n(\mathbb{C})$ relative to a subordinate matrix norm $\|\cdot\|$ is
$\text{cond}(A) = \|A\|\,\|A^{-1}\|$

- $\text{cond}_p(A) = \|A\|_p\,\|A^{-1}\|_p$
- $\text{cond}(I_n) = 1$
- $\text{cond}(A) \geq 1$ $\quad (\because \|AA^{-1}\| \leq \|A\|\,\|A^{-1}\|)$
- $\text{cond}(A) = \text{cond}(A^{-1})$
- $\text{cond}(\alpha A) = \text{cond}(A)$ $\quad \forall\ \alpha \neq 0$

- If A is invertible and $A + B$ is invertible, what is $(A+B)^{-1}$: $(A+B)^{-1} = (I_n + A^{-1}B)^{-1} A^{-1}$

- Condition Number as Amplification factor:
$$\frac{\|x_\epsilon - x\|}{\|x\|} \leq \|A\|\,\|A^{-1}\| \left\{ \frac{\|b_\epsilon - b\|}{\|b\|} + \frac{\|A_\epsilon - A\|}{\|A\|} \right\} + O(\epsilon^2)$$
→ Estimating relative errors in solution by relative errors in data.
→ $\text{cond}(A) = \|A\|\,\|A^{-1}\|$ acts as amplifying factor.

- Well-conditioned Matrix: $\text{cond}(A) \approx 1$.
  Ill-conditioned " : $\text{cond}(A) \gg 1$

Well conditioned Matrices: • If A is unitary ($A^{-1} = A^*$),
$\|A\|_2 = \|A^*\|_2 = 1 \;\therefore \text{cond}_2(A) = 1$
• $\text{cond}_2(\text{rotation matrix}) = 1$ (rotation matrix is unitary)

* - $\|A\|_2 = \mu_{max}$ (largest sing. val. of A) (We know this). $\text{cond}_2(A) = \|A\|_2\,\|A^{-1}\|_2 = \mu_{max}/\mu_{min}$

PAGE No.
DATE / /

ITERATIVE METHODS FOR LINEAR SYSTEMS $Ax = b$

$x^{(0)} \to x^{(1)} \ldots \to \infty$ such that sequence converges to desired soln. x. i.e. $\lim_{k\to\infty} \|x^{(k)} - x\| = 0$ for some norm on $\mathbb{R}^n$

- error @ k^{th} iteration : $e^{(k)} = x^{(k)} - x$ for $k = 0, 1, 2, \ldots$
- residual @ k^{th} iteration : $r^{(k)} = Ax^{(k)} - b$ for $k = 0, 1, \ldots$

- Convergent iterative method : One that converges to exact solution for any choice of initial vector.

$$\lim_{k\to\infty} x^{(k)} = x \iff \lim_{k\to\infty} e^{(k)} = 0 \iff \lim_{k\to\infty} r^{(k)} = 0$$

- Splitting of A : $A = M - N$, M is invertible
  (M, N) is called splitting

- Method : Initiate with $x^{(0)} \in \mathbb{R}^n$.
  Compute $x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$, $k = 0, 1, \ldots$
  $k \to \infty$ yields $Mx = Nx + b$ i.e. $Ax = b$
  { $M^{-1}N$ is called iteration matrix }

- Convergence Theorem : The above method converges iff
  $\rho(M^{-1}N) < 1$

- Error : $e^{(k)} = (M^{-1}N)^k e^{(0)}$
  Rate of Convergence Theorem : $\|e^{(k)}\| \sim \|e^{(0)}\| [\rho(M^{-1}N)]^k$ for $k >> 1$
  (any arbitrary norm on $\mathbb{R}^n$)

$\therefore$ smaller ρ => better convergence.

PAGE No.
DATE / /

- **Richardson's Iterative Method:**

Take $M_1 = \frac{1}{\alpha} I_n$, $N_1 = \frac{1}{\alpha} I_n - A_1$, $\alpha \neq 0$

Then $M_1^{-1} N_1 = I_n - \alpha A_1$

$x^{(k+1)} = x^{(k)} + \alpha (b - A_1 x^{(k)})$

- When does the method converge?

$M_1^{-1} N_1$ has eigen values $(1 - \alpha\lambda_i)$ where λ_i are eigen values of A_1.

∴ method converges iff $|1 - \lambda_i \alpha| < 1$ $\forall$ eigen vals. of A_1 λ_i

- Real symmetric A_1 with real eigenvalues:

$\lambda_1 \leq \lambda_2 \leq \cdots\cdots \leq \lambda_n$

- if $\lambda_i \lambda_j < 0$ → method doesn't converge for any α.
- if $\lambda_i > 0$ $\forall i$ then " converges iff $0 < \alpha < \frac{2}{\lambda_n}$
- if $\lambda_i < 0$ $\forall i$ " " " " $\frac{2}{\lambda_1} < \alpha < 0$

- Optimal α for A_1 with $0 < \lambda_1 \leq \cdots\cdots \leq \lambda_n$ (+ve. Eigen vals.)?

$\tilde{\alpha} = \frac{2}{\lambda_1 + \lambda_n}$. This gives minimum ρ

$\rho(I_n - \tilde{\alpha} A_1) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$

- if further A_1 is normal, $\rho(I_n - \tilde{\alpha} A_1)$

$= \frac{\lambda_n/\lambda_1 - 1}{\lambda_n/\lambda_1 + 1} = \frac{\text{cond}_2(A_1) - 1}{\text{cond}_2(A_1) + 1}$ → increasing fn. of cond_2.

∴ More well-conditioned the matrix, faster the convergence of Richardson method.

PAGE No.
DATE / /

- **Jacobi Method :**

$A = M - N$

$M = D = \text{diag}(A_{11}, \dots, A_{nn})$, $A_{ii} \neq 0$ $\because M$ is invertible

$N = D - A$

$\therefore M^{-1}N = I_n - D^{-1}A$

$x^{(k+1)} = x^{(k)} + D^{-1}(b - Ax^{(k)})$

- For what matrices A does the method converge ?

if A is "strictly row-diagonally dominant" or "strictly column-diagonally dominant".

(In either of these cases, $\rho(M^{-1}N) < 1$

- Defn : Strictly row diagonally-dominant matrix A if $|A_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^{n} |A_{ik}|$

Strictly column diagonally-dominant if $|A_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^{n} |A_{ki}|$

PAGE No.
DATE / /

Gauss-Seidel Method:

$A = M_1 - N_1$

$M_1 = D - E \quad ; \quad N_1 = F$

$D = \text{diag}(A_{11}, \dots, A_{nn}) \quad , \quad A_{ii} \neq 0$

$E_{ij} = \begin{cases} -A_{ij} & , i > j \\ 0 & \text{otherwise} \end{cases}$; $F_{ij} = \begin{cases} -A_{ij} & , i < j \\ 0 & \text{otherwise} \end{cases}$

E: strictly lower Δ^{lar} ; F: strictly upper Δ^{lar}

$M_1^{-1} N_1 = G_T = (I_n - L)^{-1} U$

where $L = D^{-1} E$ (L: strictly lower Δ^{lar}) $\quad U = D^{-1} F$ (U: strictly upper Δ^{lar})

For Jacobi method,
J = iteration matrix
$= L + U$

- When does the method converge?

if A is "strictly row diagonally-dominant".

- Which method performs better and how?

Stein-Rosenburg Theorem: if J is non-negative, then one and only one of these hold:

- $\rho(G_T) = \rho(J) = 0$
- $0 < \rho(G_T) < \rho(J) < 1$
- $\rho(G_T) = \rho(J) = 1$
- $\rho(G_T) > \rho(J) > 1$

i.e. Either Jacobi & Gauss-Seidel converge or both diverge.
When both converge, Gauss-Seidel outperforms Jacobi.

PAGE No.
DATE / /

- **Relaxed Gauss-Seidel Method:**

$$A = M - N$$

$$M = \frac{1}{\omega} D - E \quad ; \quad N = \left(\frac{1}{\omega} - 1\right) D + F$$

{E and F same as Gauss-Seidel}

$$M^{-1}N = \left(\frac{1}{\omega} D - E\right)^{-1} \left(\left(\frac{1-\omega}{\omega}\right) D + F\right)$$

i.e. $$G_\omega = \left(\frac{1}{\omega} I_n - L\right)^{-1} \left(\left(\frac{1-\omega}{\omega}\right) I_n + U\right)$$

- **Thm:** For any A, we have $\rho(G_\omega) \geq |\omega - 1|$ $\forall \omega \neq 0$

So what conditions for convergence?

convergence $\Rightarrow 0 < \omega < 2$

$\omega \in \mathbb{R} \setminus (0, 2) \Rightarrow$ no convergence

- If A is symmetric positive definite, $A \in M_n(\mathbb{R})$ then for any $\omega \in (0, 2)$: $\rho(G_\omega) < 1$.

- Some other lemmas for positive definiteness & iteration matrices.

- A be Hermitian, positive definite
  (M, N) be splitting of A.
  Suppose $M + M^* - A$ is positive definite.

Eigenvalue → i) Let λ be eigenvalue of $H \equiv A^{-1}(2M - A)$.
Then $\text{Real}(\lambda) > 0$

Spectral radius → ii) Take $H = A^{-1}(2M - A)$. Then

$$(H - I_n)(H + I_n)^{-1} = M^{-1}N$$

And, $$\rho(M^{-1}N) < 1$$

PAGE No.
DATE / /

EIGEN-VALUES OF A MATRIX

Eigen pair (λ, x) where $x \neq 0$ associated with A if $Ax = \lambda x$

- How to find Eigen vals. ?
  $|A - \lambda I_n| = 0$
- characteristic polynomial is of degree n.
  Can have at most n roots, along with repeated roots if any. i.e. $\lambda_i = \lambda_j$ for some $i \neq j$.

- $\det(A) = \lambda_1 \lambda_2 \cdots \lambda_n$
  $\text{Tr}(A) = \lambda_1 + \cdots + \lambda_n$

- Eigen values of Δ^{lar} matrix are its diagonal values.
- " " " " real symm. matrix are real.
- " " " orthogonal matrix satisfy $|\lambda| = 1$.

- "A and B are similar" means $B = P^{-1} A P$ for some invertible P.

- $\max_i |\lambda_i| = \rho(A) \leq \|A\|$, where $\|\cdot\|$ is any matrix norm
  i.e. all Eigen vals. lie in the ball centred around origin in $\mathbb{C}$ with radius $\|A\|$.

- Gershgorin Circle Theorem:
  Any Eigenvalue λ of matrix A belongs to the set $\bigcup_{i=1}^{n} D_i$.

  where $D_i = \{ z \in \mathbb{C} \mid |z - A_{ii}| \leq R_i \}$
  and $R_i = \sum_{j \neq i} |A_{ij}|$

ROOTS OF NONLINEAR EQUATIONS

- **Intermediate Value Theorem**

  f be continuous $[a,b] \to \mathbb{R}$.
  If $f(a)\,f(b) \leq 0$ then $\exists\ \xi \in [a,b]$ s.t. $f(\xi) = 0$

- Fixed pt: $y \in \mathbb{R}$ is a fixed pt. of $\psi : \mathbb{R} \to \mathbb{R}$ if $\psi(y) = y$

- **Fixed point theorem**

  g be continuous fn $g : [a,b] \to [a,b]$.
  Then $\exists\ \xi \in [a,b]$ s.t. $\xi = g(\xi)$

- Finding roots of $f(x) = 0$ is equivalent to solving $x - g(x) = 0$ for some cont. fn g.
  Then $f(\xi) = 0 \iff \xi = g(\xi)$

- Iterative process to get fixed pt of continuous fn. g: $(g : [a,b] \to [a,b])$
  Idea is to approximate fixed pt. ξ by sequence $x^{(k)}$.

  $x^{(0)} \in [a,b]$ = initial guess.
  Further iterates computed using (to ξ step), $x^{(k+1)} = g(x^{(k)})$
  Suppose the sequence converges, then to show that it converges to the fixed pt. of g:

$$\xi = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} g(x^{(k)}) = g\left(\lim_{k\to\infty} x^{(k)}\right) = g(\xi$$

  $\therefore$ ξ must be fixed pt. of g.

For what functions g does the sequence of iterates converge?

- **Contraction mapping theorem:**

- If $g:[a,b] \to [a,b]$ is a contraction on $[a,b]$, then g has a "unique" fixed pt. $\xi \in [a,b]$.
Also, sequence of iterates that is given by
$$x^{(k)} = g(x^{(k-1)}), \quad k = 1, 2, \ldots$$
converges to fixed pt. ξ for any initial guess $x^{(0)} \in [a,b]$

- Contraction: Let g be cont. fn. $g:[a,b] \to \mathbb{R}$.
g is called a contraction if $\exists$ a constant $L \in (0,1)$ such that
$$|g(x) - g(y)| \leq L|x-y| \quad \forall x, y \in [a,b]$$

Can we lift the condition of contraction on entire $[a,b]$?

- **Theorem:**

continuous fn. $g:[a,b] \to [a,b]$.
Let $\xi \in [a,b]$ be fixed pt. of g.
Assume g' is continuous in $[\xi - h, \xi + h]$ for some $h > 0$. Assume $|g'(\xi)| < 1$
Then the sequence of iterates $x^{(k+1)} = g(x^{(k)})$ converges to ξ provided initial guess $x^{(0)}$ is sufficiently close to ξ.

- On the other hand, in above thm. if we assume $|g'(\xi)| > 1$, then sequence of iterates given by $x^{(k+1)} = g(x^{(k)})$ does not converge to ξ for whatever initial guess $x^{(0)} \neq \xi$

* Conclusion: continuous fn $f:[a,b]\to\mathbb{R}$.
Simple iteration $x^{(k+1)} = f(x^{(k)})$
Relaxation iteration $x^{(k+1)} = x^{(k)} - \lambda f(x^{(k)})$

How to approximate zeros of function f?

- Relaxation iteration:

Continuous fn. $f:[a,b]\to\mathbb{R}$.
initial guess $= x^{(0)} \in [a,b]$
Iteration $x^{(k+1)} = x^{(k)} - \lambda f(x^{(k)})$, $k=0,1,2,\cdots$
$\lambda \neq 0$ is some real parameter.

- How do we know the above sequence converges to root of f?

Suppose the sequence converges to ξ.

$$\xi = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} x^{(k)} - \lambda \cdot \lim_{k\to\infty} f(x^{(k)}) = \xi - \lambda \lim_{k\to\infty} f(x^{(k)})$$

$$\Rightarrow f(\xi) = 0$$

Relaxation iteration to find zeros of f $\iff$
Simple iteration to find fixed pts. of $g(x) = x - \lambda f(x)$

- What should value of λ be to guarantee convergence?

Suppose f is differentiable about ξ.
$\therefore g$ is also diff. around ξ. $g' = 1 - \lambda f'$
$|g'(\xi)| = |1 - \lambda f'(\xi)| < 1$ if:
- λ has same sign as $f'(\xi)$
- λ is not too large.

Theorem: cont. fn. $f:[a,b]\to[a,b]$
Let $\xi \in [a,b]$ be zero of f, i.e. $f(\xi)=0$.
Assume f' is continuous around ξ and let $f'(\xi)\neq 0$.
Then $\exists\ \lambda>0,\ \delta>0$ s.t. sequence of iterates given by $x^{(k+1)} = x^{(k)} - \lambda f(x^{(k)})$ converges to ξ for whatever initial guess $x^{(0)} \in [\xi-\delta, \xi+\delta]$

*

* • **Bisection Method**: Suppose a_k & b_k s.t. $f(a_k)f(b_k)<0$. Define $c_k = \frac{a_k+b_k}{2}$. Define $(a_{k+1}, b_{k+1}) = \begin{cases} (a_k, c_k) & \text{if } f(c_k)f(b_k)>0 \\ (c_k, b_k) & \text{if } f(c_k)f(b_k)<0 \end{cases}$ if $f(c_k)\neq 0$

• **Secant Method**: Replace $f'(x_k)$ in Newton's meth. by $f'(x_k) \approx \frac{f(x_k)-f(x_{k-1})}{x_k - x_{k-1}}$

Gives $e^{(k+1)} \leq \frac{2}{3} e^{(k)}$

• **Error sequence** (for sequence of iterates):

For any sequence of iterates $x^{(k)}$ trying to approximate a point ξ, the error sequence is

$$e^{(k)} = x^{(k)} - \xi \quad \text{for } k = 0, 1, 2, \ldots$$

• **Order of convergence** of sequence of iterates is p (≥ 1) if $\exists$ constant $C \geq 0$ (with $C<1$ if $p=1$) and integer N s.t. $\forall k \geq N$:

$$|e^{(k+1)}| \leq C |e^{(k)}|^p$$

$p=1 \Rightarrow$ linear convergence

$p=2 \Rightarrow$ quadratic convergence

- What is the order of simple iteration $x^{(k+1)} = g(x^{(k)})$?

$$|e^{(k+1)}| = |x^{(k+1)} - \xi| = |g(x^{(k)}) - g(\xi)| \leq L |e^{(k)}|$$

$\therefore$ linear convergence.

• **Newton-Raphson Method**

It is a relaxed iteration with best convergence.

$$g(x) = x - \frac{f(x)}{f'(x)}$$

- When does this method converge?

Thm: Let $f: [a,b] \to \mathbb{R}$ be C^2 on I_δ on $[\xi-\delta, \xi+\delta]$ for some $\delta > 0$. Suppose $f(\xi)=0$, $f'(\xi)\neq 0$, $f''(\xi)\neq 0$. Suppose further $\exists$ constant M s.t. $\frac{|f''(x)|}{|f'(y)|} \leq M \; \forall x, y \in I_\delta$

If initial guess $x^{(0)}$ is s.t. $|x^{(0)} - \xi| \leq h$ with $h = \min\{\delta, \frac{1}{M}\}$, then iteration sequence converges quadratically to ξ. $\lim_{k\to\infty} \frac{e^{(k+1)}}{(e^{(k)})^2} \leq \frac{M}{2}$

*