$eps \cdot b^{1-t} = 6^{1-7} = 6^{-6}$

Q1: a. $(0.1000000)_6 \times 6^{(-55)_6}$ Normalized

$t=7$

$(0.0000001)_6 \times 6^{(-55)_6}$

b. $(0.5555555)_6 \times 6^{(55)_6}$

$t=7$

c. chopping — (RRO) $\delta < b^{1-t}$ (Delta)

rounding — $|\delta| \leq \frac{1}{2} b^{(1-t)}$

$$\left| \frac{fl(x) - x}{|x|} \right| < \frac{1}{2} b^{1-t} = \frac{1}{2} \times 6^{1-7} = \frac{1}{93312}$$

d.

Numerator	denominator	quotient	remainder
407	6	67	5
67	6	11	1
11	6	1	5
1	6	0	1

$(407)_{10} = (1515)_6$

$fl(407) = (0.151500)_6 \times 6^{(04)_6}$

For fraction part 407.1

$(1515) . (0.0333)$

Additional Question: A terminating decimal fraction with n digit must have a terminating binary fraction representation base 6

counter example : $(0.1)_{10} \Rightarrow (0.0\overline{3})_6$

Q2: a) Given a matrix A and a permutation matrix P, the goal is to find matrixes L and U such that $PA = LU$

Starting with matrix P and A:

$$P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$P_1 A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 6 & 6 \\ 1 & 7 & 6 \\ 4 & 12 & 12 \end{bmatrix} = \begin{bmatrix} 4 & 12 & 12 \\ 1 & 7 & 6 \\ 2 & 6 & 6 \end{bmatrix}$$

Now, to find L and U, we perform Gaussian Elimination:

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix} \qquad L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}$$

$$L_1 P_1 A_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{4} & 1 & 0 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 12 & 12 \\ 1 & 7 & 6 \\ 2 & 6 & 6 \end{bmatrix} = \begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix} = U$$

$$L_1 P_1 A = U$$

$$\underbrace{P_1 A}_{P} = \underbrace{L_1^{-1} U}_{L} \Rightarrow PA = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 6 & 6 \\ 1 & 7 & 6 \\ 4 & 12 & 12 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix} = LU$$

b) To solve $Ax = b \Rightarrow P_1 A x = Pb \Rightarrow \tilde{A}x = \tilde{b}$

$$\underbrace{L_1^{-1} U}_{d} x = \tilde{b}$$

$$L_1^{-1} \cdot d = \tilde{b}$$

Now solve $L_1^{-1} d = \tilde{b}$ for d

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{4} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} 40 \\ 25 \\ 20 \end{bmatrix}$$

$$\begin{cases} d_1 = 40 \\ d_2 = 15 \\ d_3 = 0 \end{cases} \quad d = \begin{bmatrix} 40 \\ 15 \\ 0 \end{bmatrix}$$

Scanned with CamScanner

Then, solve $U \cdot X = d$ for x

$$\begin{bmatrix} 4 & 12 & 12 \\ 0 & 4 & 3 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 40 \\ 15 \\ 0 \end{bmatrix}$$

$$\begin{cases} 4x_1 + 12x_2 + 12x_3 = 40 \\ 4x_2 + 3x_3 = 15 \end{cases}$$

Let $x_3 = b$

$$x_2 = \frac{15 - 3b}{4}$$

$$x_1 = 10 - 3x_2 - 3x_3$$

$$= 10 - \frac{45 - 9b}{4} - 3b$$

$$= 10 - \frac{45}{4} - \frac{3}{4}b$$

Note: $\begin{bmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{bmatrix}_{4\times4}$

$$L_2 P_2 L_1 P_1 A = U$$

$$L_2 (P_2 L_1 P_2) P_2 P_1 A = U$$

$\tilde{L}_1$

$$L_2 \tilde{L}_1 P_2 P_1 A = U$$

$$P_2 P_1 A = \tilde{L}_1^{-1} L_2^{-1} U$$

$$\tilde{L}_1 = P_2 L_1 P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ x & 1 & 0 & 0 \\ y & 0 & 1 & 0 \\ z & 0 & 0 & 1 \end{bmatrix}$$

swap row 2 and row 4

$$\tilde{L}_1^{-1} = \begin{bmatrix} 1 & & & \\ -x & 1 & & \emptyset \\ -y & & 1 & \\ -z & \emptyset & & 1 \end{bmatrix}$$

flip the sign

Q3: This is just a simple application of the algebraic rules

$$PA = LU$$
$$\det(PA) = \det(LU)$$
$$\det(P)\cdot\det(A) = \det(L)\cdot\det(U)$$
$$\det(A) = \frac{1}{\det(P)}\det(L)\cdot\det(U)$$

$\because$ U, L are Δ matrix
(upper triangle, lower triangle)

$\therefore \det(\Delta)$ is the product of it's diagonal entries

$$\det(L) = \prod_{i=1}^{n} l_{ii} = 1 \;;\; \det(U) = \prod_{i=1}^{n} u_{ii}$$

$\det(P) = \pm 1$ depending on whether there are even/odd # of row interchanges during pivoting

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 0 - 0 + 1\cdot(0-1) = -1$$ ← odd

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cdot & \cdot \\ 0 & \cdot & 1 \end{bmatrix} = 1\cdot 1 - 0 + 0 = 1$$

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = 0 - 1\cdot(0-1) + 0 = 1$$ ← even

Hence: $\det(A) = (\pm 1)\prod_{i=1}^{n} u_{ii}$

$d_k \cdot b^k + d_{k-1} \cdot b^{k-1} + \cdots + d_0 \cdot b^0$

$d_1 \cdot b^1$

(take 3 mins)

Q4: Given a $(k+1)$-digit base-b number

$$d_k d_{k-1} \cdots d_1 d_0,$$

We want to convert it to its base-10 equivalent, using k flops or less.

We use Horner's method to evaluate the polynomial represented by the number. The process is:

1. Initialize the result $R := d_k$, the most significant digit
2. For each digit d_i from $i = k-1$ down to 0:
   - Perform a floating-point operation (flop) by computing $R = b \cdot R + d_i$
3. After the last iteration, R is base-10 equivalent of the base-b number

The algorithm is:

$R = d_k$
for i from $k-1$ down to 0 d_0 ← k iterations
    $R = b * R + d_i$
end for

The variable R now contains the base-10 equivalent of the base-b number after k iterations, which corresponds to k flop

$$\Big(\big((d_k b + d_{k-1})b + d_{k-2}\big)b + d_{k-3} \cdots\Big)$$

$d_k \cdot b^k + d_k \cdot b^{k-1} + \cdots + d_1 b^1 + d_0 \cdot 1$

↑ goes k iter ↑ goes $k-1$ iter ↑ goes 1 iter ↑ at last

Q5. Treat y as a constant and the determine a FP of the iteration in terms of y.

$$x_{k+1} = 2x_k - x_k^2 y \iff y = \frac{2x_k - x_{k+1}}{x_k^2} = \frac{2}{x_k} - \frac{x_{k+1}}{x_k^2}$$

As $k \to \infty$, $x_{k+1} = x_k = \hat{x}$ when converge happen

$$y = \frac{2}{\hat{x}} - \frac{\hat{x}}{\hat{x}^2}$$

$$y = \frac{1}{\hat{x}}$$

$$\hat{x} = \frac{1}{y}$$

Therefore, $\frac{1}{y}$ is the fixed point of this iteration

Since this is a Newton's iteration, all the FP are roots of $F(x)$?? we need to know $F(x)$. ← "certain function"

To find $F(x)$: ★Question: when we using Newton's method to find convergence, what are we solving in terms of $F(x)$?

Given $g(x) = 2x - x^2 y \Rightarrow x - \frac{F(x)}{F'(x)}$ in Newton's method

$$= x - (x^2 y - x)$$

$$\therefore x^2 y - x = \frac{F(x)}{F'(x)}$$

It takes some tries, and you should find, A trick common from MATA37 !

$$F(x) = \frac{1}{x} - y \quad F'(x) = -x^{(-2)}$$

$$\hookrightarrow \frac{F(x)}{F'(x)} = \frac{\frac{1-xy}{x}}{-x^{-2}} = x - x^2 y \checkmark$$

Find "certain function"

$\therefore$ The root of $F(x)$: $\frac{1}{x} - y = 0 \quad x = \frac{1}{y}$

$\therefore$ The purpose of this fix point iteration is to find the root of $\frac{1}{x} - y$

Q6: (1) Let P_a be the polynomial that interpolates $\{(-1,4),(0,6),(1,12)\}$

⇒ Using divided-difference table, we have

x_i	$y[x_i]$	$y[x_{i+1},x_i]$	$y[x_{i+2},x_{i+1},x_i]$	$y[x_3,x_2,x_1,x_0]$
$x_1 = -1$	4	$\frac{6-4}{0-(-1)} = 2$	$\frac{y[x_2,x_1]-y[x_1,x_0]}{x_{i+2}-x_i} = \frac{6-2}{1-(-1)} = 2$	
$x_2 = 0$	6	$\frac{12-6}{1-0} = 6$		$\frac{y[x_3,x_2,x_1]-y[x_2,x_1,x_0]}{x_3-x_0} = 0$
$x_3 = 1$	12	$y[x_4,x_3]$	$y[x_4,x_3,x_2]$	
2	16			

By Newton Polynomial thm:

$$P_a(x) = y[-1] + (x-(-1))\cdot y[0,-1] + (x-0)(x-(-1))\,y[1,0,-1]$$

$$= 4 + 2(x+1) + 2x(x+1)$$

(2) Let P_b be the polynomial that interpolates $\{(-1,4),(0,6),(1,12)\}$

⇒ Using Lagrange basis, we have

$$l_0(x) = \prod_{\substack{j=0 \\ (j\neq 0)}}^{2} \frac{x-x_j}{x_0-x_j} = \left(\frac{x-0}{-1-0}\right)\cdot\left(\frac{x-1}{-1-1}\right) = \frac{x(x-1)}{2}$$

$$l_1(x) = \prod_{\substack{j=0 \\ (j\neq 1)}}^{2} \frac{x-x_j}{x_1-x_j} = \left(\frac{x-(-1)}{0-(-1)}\right)\cdot\left(\frac{x-1}{0-1}\right) = \frac{(x+1)(x-1)}{-1}$$

$$l_2(x) = \prod_{\substack{j=0 \\ (j\neq 2)}}^{2} \frac{x-x_j}{x_2-x_j} = \frac{x-(-1)}{1-(-1)}\cdot\frac{x-0}{1-0} = \frac{(x+1)x}{2}$$

Now, by Lagrange method of interpolation $P(x) = \sum_{i=0}^{n} l_i(x)\,y_i$

$$P_b = 4\cdot\frac{x(x-1)}{2} + 6\cdot\frac{(x+1)(x-1)}{-1} + 12\,\frac{(x+1)x}{2}$$

$$= 2x(x-1) - 6(x+1)(x-1) + 6x(x+1)$$

(3). By Uniqueness of Polynomial Interpolation Theorem
For any sets $\{x_i : i=0,1,\cdots n\}$ and $\{y_i : i=0,1,\cdots n\}$ where x_i is distinct, y_i not necessarily distinct.
$\Rightarrow \exists$ a unique polynomial $p \in P_n$ s.t. $p(x_i) = y_i$ for $i=0,1\cdots n$

we have $\{x_i\}_0^2 = \{-1, 0, 1\}$ which 3 distinct x_i's
and $\{y_i\}_0^2 = \{4, 6, 12\}$

It's clear that $p_a(x) \in P_2$ where p_a interpolates the given data points by Newton Method.

Similarly, $p_b(x) \in P_2$ where p_b interpolates the same data points by lagrange basis...

By our theorem above, there is only one polynomial $p \in P_2$ that interpolates given 3 data points. $\Rightarrow$ p_a and p_b must be same

(4). • Consider the langrange method. If we add another data point, we would need to re-compute all $l_i(x)$, so the previous polynomial does not reduce our work.

• Consider the divided-difference method.
Note that from the previously computing polynomial p_3, we only need to compute one extra term.

$$(x-x_1)(x-x_2)(x-x_3)\, y[x_4, x_3, x_2, x_1]$$

• Consider the Vandermonde method.
It is obvious that append a new data point will result a new system to solve (Vandermonde matrix changed! with a new.)

$$\underbrace{\begin{bmatrix} & & & \\ & & & \\ & & V & \\ & & & \end{bmatrix}}_{(\text{row}+1)} \underbrace{\begin{bmatrix} a_0 \\ \vdots \\ a_n \\ a_{n+1} \end{bmatrix}}_{(\text{column}+1)} = \begin{bmatrix} y_0 \\ \vdots \\ \text{previous} \\ y_{n+1} \end{bmatrix}$$

previous interpolant cannot be used to reduce the work.

Scanned with CamScanner

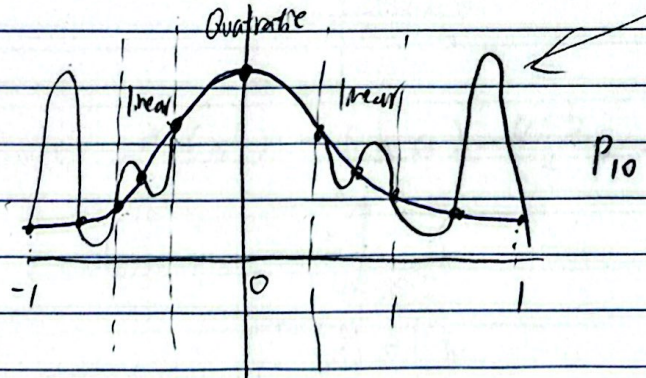
Q7: Error in Polynomial Interpolation.

$E(x) = y(x) - p(x)$

Interpolating polynomial

What is a runge phenomenon?

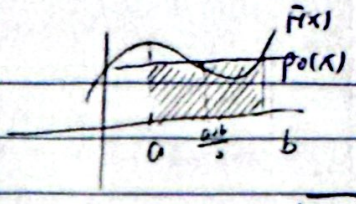
Briefly speaking: More interpolant data points cause higher degree polynomical and result worse fitting. Oscillation in some area



Solution to interpolate the Runge Function:

- Use piecewise polynomial interpolation
- fitting lower degree polynomal $P_0 \rightarrow P_6$
- choose the interpolation points more judiciously
  ⇒ Chebyshev interpolant
  more data points on left side and righ side and less points in the middle.

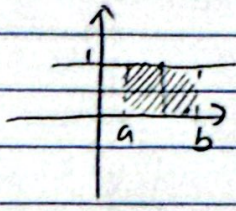
Q8: Consider $I(f) = \int_a^b f(x)\,dx$



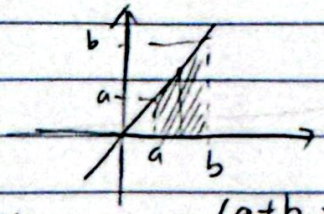
- Midpoint Rule: $I(f) \approx M(f) = I(p_0) = \underbrace{(b-a)}_{\text{weight } A_0} f(\underbrace{\frac{a+b}{2}}_{\text{node } x_0})$

precision: $m = 1$ because:

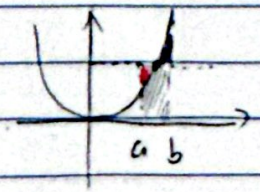
when $f = x^0$ $\quad M(x^0) = M(1) = (b-a)\cdot 1 = I(x^0) = \int_a^b 1\,dx$



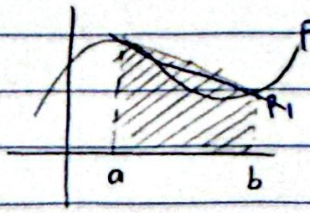
when $f = x^1$ $\quad M(x^1) = (b-a)\cdot\frac{a+b}{2} = \frac{b^2-a^2}{2} = I(x) = \int_a^b x\,dx$



when $f = x^2$ $\quad M(x^2) = (b-a)\left(\frac{a+b}{2}\right)^2 \neq \frac{b^3-a^3}{3} = I(x^2) = \int_a^b x^2\,dx$



- Trapezoidal rule: $I(f) \approx T(f) = I(p_1) = (b-a)\dfrac{[f(a)+f(b)]}{2}$



precision: $m = 1$ $\quad A_0 = \frac{b-a}{2}$ $\quad x_0 = a$

$A_1 = \frac{b-a}{2}$ $\quad x_1 = b$

when $f = x^0$ $\quad T(x^0) = (b-a)\cdot\frac{1+1}{2} = b-a = I(x^0)$

when $f = x^1$ $\quad T(x^1) = (b-a)\cdot\frac{(a+b)}{2} = \frac{b^2-a^2}{2} = I(x^1)$

when $f = x^2$ $\quad T(x^2) = (b-a)\cdot\frac{a^2+b^2}{2} \neq \frac{b^3-a^3}{3} = I(x^2)$

- Simpson's Rule $\quad I(f) \approx S(f) = I(p_2) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{(a+b)}{2}\right) + f(b)\right)$

precision: $m = 3$

$x_0 = a$ $\quad x_1 = \frac{a+b}{2}$ $\quad x_2 = b$

$A_0 = \frac{b-a}{6}$ $\quad A_1 = \frac{2}{3}(a+b)$ $\quad A_2 = \frac{5}{6}$

Scanned with CamScanner