

# SOLUTIONS TO SELECTED PROBLEMS ON

## HOMEWORK 1

(PREPARED BY  
KARTIK)

(1)

MA/CSC 427-001: INTRODUCTION  
TO NUMERICAL ANALYSIS I

$$(1) \quad 2 = \frac{2}{2^1} \times 2^1$$

$$= 1 \times 2^1$$

$$= (1.00)_2 \times 2^1$$

MANTISSA = 00000000000000000000000 (23 BITS)

BIASED EXPONENT = 1 + 127

$$= (128) = 10000000$$

SIGN BIT = 0

$$\therefore 2 = \boxed{0 \mid 10000000 \mid 00000000000000000000000}$$

SIGN BIT      BIASED  
                    EXPONENT

MANTISSA

$$(2) \quad 1000 = \frac{1000}{2^9} \times 2^9$$

$$= 1.953125 \times 2^9$$

$$0.953125 \times 2 = 1.90625 \quad 1 \quad \therefore (1.953125)$$

$$0.90625 \times 2 = 1.8125 \quad 1 \quad = (1.111101)_2$$

$$0.8125 \times 2 = 1.625 \quad 1$$

$$0.625 \times 2 = 1.25 \quad 1$$

$$0.25 \times 2 = 0.5 \quad 0$$

$$0.5 \times 2 = 1.0 \quad 1$$

(2)

∴ MANTISSA

$$= 1111010 \dots 0$$

←-----→  
23 BITS

BIASED EXPONENT

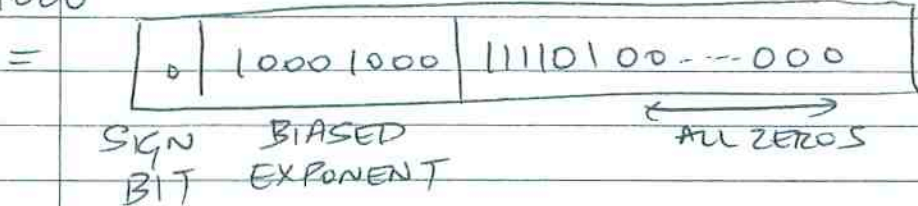
$$= 9 + 127 = 136$$

$$= (10001000)_2$$

SIGN BIT = 0

$$\begin{array}{r|l} 2 & 136 \ 0 \\ 2 & 68 \ 0 \\ 2 & 34 \ 0 \\ 2 & 17 \ 1 \\ 2 & 8 \ 0 \\ 2 & 4 \ 0 \\ 2 & 2 \ 0 \\ & 1 \ 1 \end{array}$$

∴ 1000



$$(3) \quad \frac{1024}{5} = \frac{1024}{5} \times 2^7$$

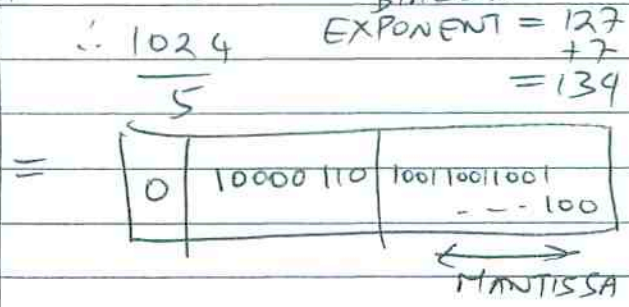
$$= 1.6 \times 2^7$$

$$= (1.10011001100110011001100)_2 \times 2^7$$

←-----→  
MANTISSA

BIASED EXPONENT = 127 + 7 = 134

0.6 x 2 = 1.2	1
0.2 x 2 = 0.4	0
0.4 x 2 = 0.8	0
0.8 x 2 = 1.6	1
0.6 x 2 = 1.2	1



$$\begin{aligned}
 (4) \quad \frac{1}{10} \times 2^{-140} &= 0.1 \times 2^{-140} \\
 &= \frac{0.1}{2^{-4}} \times 2^{-144} \\
 &= 1.6 \times 2^{-144}
 \end{aligned}$$

Underflow error since the smallest positive number in single precision format is  $1.0 \times 2^{-126}$

(2) EXERCISE Set 12, Page 27, Problem 13

Only part 13(a) is done here

$$\frac{1}{3} x^2 - \frac{123}{4} x + \frac{1}{6} = 0$$

$$x_1 = \frac{123}{4} + \sqrt{\left(\frac{123}{4}\right)^2 - 4\left(\frac{1}{3}\right)\left(\frac{1}{6}\right)}$$

$$2\left(\frac{1}{3}\right)$$

$$x_1 = 92.24457963$$

$$x_2 = \frac{-2c}{b - \sqrt{b^2 - 4ac}}$$

→ We use this formula for  $x_2$  since we are subtracting two nearly equal nos in the regular formula for  $x_2$

$$= 5.420372688 \times 10^{-3}$$

$$x_2 = -2\left(\frac{1}{6}\right)$$

$$-\frac{123}{4} - \sqrt{\left(\frac{123}{4}\right)^2 - 4\left(\frac{1}{3}\right)\left(\frac{1}{6}\right)}$$

(4)

ROUNDING

Using 4 digit arithmetic

$$\begin{aligned} f_l(x_1) &= \frac{30.75 + \sqrt{(30.75)^2 - 4(0.3333)(0.1667)}}{2(0.3333)} \\ &= \frac{(30.75 + 30.75)}{10.6666} \\ &= 92.26 \end{aligned}$$

SIMILARLY

$$f_l(x_2) = \frac{-2c}{(b - \sqrt{b^2 - 4ac})} = \frac{-0.3333}{(-30.75 - 30.75)}$$

$$= 0.005419$$

Abs error  
in  $x_1 = |x_1 - f_l(x_1)|$   
 $= 0.01542$

Relative

error in  $x_1 = \frac{|x_1 - f_l(x_1)|}{|x_1|}$

$$= \frac{|92.24457963 - 92.26|}{|92.24457963|}$$

$$= 1.672 \times 10^{-4}$$

(5)

Absolute

$$\text{error in } x_2 = |x_2 - f_1(x_2)|$$

$$= |0.005420372688 - 0.005419|$$

$$= 1.372688 \times 10^{-6}$$

Relative

error in  $x_2$

$$= \frac{|x_2 - f_1(x_2)|}{|x_2|}$$

$$= \frac{1.372688 \times 10^{-6}}{0.005420372688}$$

$$= 2.53246055 \times 10^{-4}$$

(3) EXERCISE SET 1.2, PAGE 27, PROBLEM 14

Only part 14(a) is done here

Using 4 digit chopping

$$f_1(x_1) = \frac{(30.75 + 30.74)}{(0.6666)} = 92.24$$

$$\text{Abs error in } x_1 = |x_1 - f_1(x_1)| = \left| \frac{92.24457963}{-92.24} \right|$$

∴ Relative

error in  $x_1$

$$= 4.57963 \times 10^{-3}$$

$$= \frac{|x_1 - f_1(x_1)|}{|x_1|} = \frac{4.57963 \times 10^{-3}}{92.24457963}$$

$$= 4.96466 \times 10^{-5}$$

(6)

$$f_l(x_2) = \frac{-2c}{b - \sqrt{b^2 - 4ac}}$$
$$= \frac{-0.3333}{(-30.75 - 30.74)}$$
$$= 0.00542$$

Absolute error

$$\ln x_2 = |x_2 - f_l(x_2)|$$
$$= \left| 5.420372688 \times 10^{-3} - 0.00542 \right|$$
$$= 3.9355992 \times 10^{-7}$$

Relative error

$$\ln x_2 = \frac{|x_2 - f_l(x_2)|}{|x_2|}$$
$$= \frac{3.9355992 \times 10^{-7}}{5.420372688 \times 10^{-3}}$$
$$= 7.2607 \times 10^{-5}$$

7

(15)

18 ZEROS      26 ZEROS  
↔                      ↔

(a) 0 10000001010 100100110...0 0...0

↓  
1024 + 2 + 8  
1034

2<sup>1</sup> + 2<sup>-4</sup> + 2<sup>-7</sup> + 2<sup>-8</sup>  
= 1.57421875

1034 - 1023  
= (11)<sub>10</sub>

NUMBER = 1.57421875 × 2<sup>11</sup>  
= 3224

18 ZEROS      26 ZEROS  
↔                      ↔

(b) 1 10000001010 100100110...0 0...0

1034 - 1023

2<sup>1</sup> + 2<sup>-4</sup> + 2<sup>-7</sup> + 2<sup>-8</sup>  
= 1.57421875

- = (11)<sub>10</sub>

NUMBER = -1.57421875 × 2<sup>11</sup>  
= -3224

10 ONES  
↔

18 ZEROS  
↔

26 ZEROS  
↔

(c) 0 011...1 010100110...0 0...0

↓  
1023

2<sup>2</sup> + 2<sup>-4</sup> + 2<sup>-7</sup> + 2<sup>-8</sup>

1023 - 1023  
= (0)<sub>10</sub>

= 1.32421875

NUMBER = 1.32421875

10 ONES  
↔

18 ZEROS  
↔

25 ZEROS  
↔

(d) 0 01...1 01010110...0 0...01

1023

2<sup>2</sup> + 2<sup>-4</sup> + 2<sup>-7</sup> + 2<sup>-8</sup> + 2<sup>-52</sup>

1023 - 1023  
= (0)<sub>10</sub>

= 1.3242187500000002

(5) EXERCISE SET 1.2, PAGE 27, PROBLEM 16

(a) Next largest and smallest numbers from 3224

are  $3224 + 2^{-52}$  and  $3224 - 2^{-52}$   
 which are  $3224.000000000000000002$  and  
 $3223.999999999999999998$ , respectively

(b) Next largest and smallest numbers from -3224

are  $-3224 + 2^{-52}$  and  $-3224 - 2^{-52}$   
 which are  
 $-3223.999999999999999998$  and  
 $-3224.000000000000000002$ ,  
 respectively

(c) Next largest and smallest numbers from 1.32421875

are  
 $1.32421875 + 2^{-52}$  and  $1.32421875 - 2^{-52}$   
 which are  
 $1.3242187500000002$  and  
 $1.3242187499999998$   
 respectively

(d) Next largest and smallest

numbers from  $1.3242187500000002$  are  
 $1.3242187500000002 + 2^{-52}$  and  $1.3242187500000002 - 2^{-52}$   
 which are  $1.3242187500000004$  and  $1.3242187500000000$ , respectively

Best results all obtained when  $h = \sqrt{\epsilon_{mach}}$ .

Now let us consider the central difference formula for the difference quotient which is

$$\frac{f(x+h) - f(x-h)}{2h}$$

We have

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x)$$

$$+ \frac{h^3}{6} f'''(\xi_1(x))$$

for some  $\xi_1(x)$  between  $x$  and  $x+h$

Similarly

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2} f''(x)$$

$$- \frac{h^3}{6} f'''(\xi_2(x))$$

for some  $\xi_2(x)$  between  $x$  and  $x-h$

$$\therefore \left[ \frac{f(x+h) - f(x-h)}{2h} - f'(x) = \frac{h^2}{12} (f'''(\xi_1) + f'''(\xi_2)) \right] \rightarrow (2)$$

So the discretization error is  $O(h^2)$  instead of  $O(h)$  giving more accurate results

Q (6) EXPLANATION OF RESULTS :-  
WHY CENTRAL DIFFERENCE  
GIVES A BETTER APPROXIMATION ?

We have

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi(x))$$

for some  $\xi(x)$  between  $x$  and  $x+h$

$$\therefore \left[ \frac{f(x+h) - f(x)}{h} - f'(x) = \frac{h}{2} f''(\xi(x)) \right] \rightarrow (1)$$

$\therefore$  If  $h$  is reduced by a factor of 10 then the discretization error (quantity on lhs of (1)) also reduced by about a factor of 10. (not exactly because  $\xi(x)$  between  $x$  and  $x+h$  changes when  $h$  changes)

$\therefore$  Discretization error is  $O(h)$ .

However when  $h = \epsilon_{\text{MACHINE}}$  (MACHINE EPSILON)

then  $x+h = x$

$\therefore \frac{f(x+h) - f(x)}{h} = 0$  and so the results become meaningless when  $h$  gets smaller and is closed to  $\epsilon_{\text{MACH}}$ .