ADITYA CHETAN
2016217

Dt.
Pg.
B+

Scientific Computing

Assignment 1

PROBLEM 1

(a) The floating point system that I will be using is,

$$(10, 2, -3, 3)$$

Preliminaries required:

Overflow, $OFL = 10^{(3+1)}(1 - 10^{-2})$

$= 10^4(1 - 0.01)$

$= 0.99 \times 10^4$

$= 9.9 \times 10^3$ in our f.p. system

ε_{MT}

Considering rounding ~~by truncation to nearest~~ by chopping

Machine epsilon, $\varepsilon_{MT} = 10^{(1-2)}$

$= 0.1$

Considering rounding to nearest, Machine epsilon,

$\varepsilon_{MN} = \frac{1}{2}\varepsilon_{MT} = 0.5 \times 10^{-2}$

(b) now, for the numbers,

Let

now, for the numbers,

let $a = 2.5 \times 10^0$

$b = OFL = 9.9 \times 10^3$

now, using (i)

$$m = \frac{(a+b)}{2.0}$$

$$= \frac{(2.5 \times 10^0 + 9.9 \times 10^3)}{2.0}$$ → this is where the problem occurs.

Here, the numerator is not representable in our floating point format and thus, the above formula would fail to compute the mid-point. The reason is that numerator computes to a no. greater than OFL of the fp system.

from (ii)

$$m = a + \left(\frac{b-a}{2}\right)$$

$$= 2.5 \times 10^0 + \left(\frac{9.9 \times 10^3 - 2.5}{2}\right)$$

$$= 2.5 + \frac{9.900 - 2.5}{2} \times 10^3$$

$$= 2.5 \times 10^0 + \frac{9.8 \times 10^3}{2}$$ ← [from rounding by chopping]

$$= 2.5 \times 10^0 + 4.9 \times 10^3$$

$$= 4.9 \times 10^3$$

For reference, the actual mid-pt. of the 2 nos. is 4951.25

Hence, even though with an error, it is calculable.

Dt. _____ B+
Pg. _____

~~(c)~~

ALITER: Consider $a = 2.5 \times 10^0 + 6 \times 10^{-2}$

$b = 2.6 \times 10^0 + 6 \times 10^{-2}$

Clearly, $a = x + \varepsilon$ where $\varepsilon < \varepsilon_{MT}$

$b = y + \varepsilon$ " $\varepsilon < \varepsilon_{MT}$

now, using (i)

$$m = \frac{a+b}{2.0} = \frac{2.5 + 2.6 + 1.2 \times 10^{-1}}{2.0}$$

$$= \frac{5.1 + 1.2 \times 10^{-1}}{2.0}$$

$$= \frac{5.1 + 1 \times 10^{-3}}{2.0}$$

[using def^n of ε_{MT} in a system of rounding by truncation]

this is where the problem occurs.

$= 2.6$ (by rounding by truncation)

which is not the right answer.

using (ii),

$$m = a + \frac{(b-a)}{2}$$

$$= 2.5 + 0.2 \times 10^{-1} + \left(\frac{2.6 - 2.5}{2}\right)$$

$$= 2.5 + (0.06 \times 10^{-1}) + 5 \times 10^{-3}$$

$$= 2.5 + 6 \times 10^{-3} + 5 \times 10^{-3}$$

$$= 2.5 + 1.1 \times 10^{-2} > 2.5$$

$\therefore$ Hence (ii) is more suitable.

Premium

Dt. _____ B+
Pg. _____

PROBLEM 2

(c) A possible reason for deterioration of error when using the "bottom-up" recurrence relation in (b) is that ~~the~~ in this recurrence, ∵ we are adding the terms and hence, the precomputed errors in the addends get summed up and increases overall, leading to a larger loss of precision as ~~it~~ we progress. This is because ~~the~~ ~~the~~ error keeps getting accumulated at each step leading to a ~~large overall~~ fairly noticeable loss in precision near ~~x~~ $n = 30$.

(d) No, I don't ~~think~~ think that the modified recurrence will result in a similar loss in precision. This is because we are substracting the terms on the RHS, ~~to~~ and hence, the ~~error te~~ ~~to~~ precomputed error terms would also get substracted leading to a smaller overall error.

NOTE: here by pre-computed errors, I mean the error in the terms from the previous computations.

Dt.
Pg. B+

PROBLEM 3

(a) ~~Nos. with~~ Consider the floating pt. system:

$(2, 4, -3, 3)$

(a) nos. with unique representⁿ

0.111×10^{-3}

(b) no. with 2 representations.

0.111×10^{-2}
1.110×10^{-3}

(c) no. with 3 representations.

~~0.100×10^{1}~~ 0.110×10^{-1}
1.100×10^{-2}
0.011×10^{0}

(d) no. with 4 representations.

0.100×10^{1}
0.010×10^{2}
0.001×10^{3}
~~1.000×10^{0}~~
1.000×10^{0}

Dt. _____ B+
Pg. _____

PROBLEM 4

Let, $B = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$

→ (1) Swap rows 1 & 4.

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} 0+0+0+a_{41} & 0+0+0+a_{42} & 0+0+0+a_{43} & 0+0+0+a_{44} \\ 0+a_{21}+0+0 & 0+a_{22}+0+0 & 0+a_{23}+0+0 & 0+a_{24}+0+0 \\ 0+0+a_{31}+0 & 0+0+a_{32}+0 & 0+0+a_{33}+0 & 0+0+a_{34}+0 \\ a_{11}+0+0+0 & a_{12}+0+0+0 & a_{13}+0+0+0 & a_{14}+0+0+0 \end{bmatrix}$$

$$= \begin{bmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \end{bmatrix}$$

∴ Let $A_1 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

Changing the 1st & 4th row of the unit matrix I_4 causes 1st & 4th row of the multiplied matrix to swap. Also, ∵ we want to operate on rows of B, left multiplication is necessary, as only by left multiplication can I manipulate the rows to change order.

* $I_n \rightarrow n \times n$ unit matrix.

Dt. Pg. B+

(2) Add 2 times column 3 to column 1.

Consider A_1BC_1 where $C_1 = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

~~∵ we are manipulating columns, we need to right multiply. Also, this~~

(3) Swap columns 2 & 3

Consider $(A_1BC_1)C_2$, where $C_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

~~Also~~ causes 3rd col. elements to appear in 2nd ←

dot product with transpose of row vector causes ← 2nd col. elements to appear in 2nd.

(4) Add 5 times row 2 to row 4.

Consider $A_2(A_1BC_1C_2)$, where $A_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 5 & 0 & 1 \end{bmatrix}$

(similar reason as (1))

↑ dot prod. of $\begin{bmatrix} 0 \\ 5 \\ 0 \\ 1 \end{bmatrix}$ with $\begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \rightarrow 5b+d$ ←

(5) Delete column 2

Consider $(A_2A_1BC_1C_2)C_3$, where $C_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

∵ resultant is a 4×3 matrix, C_3 must be 4×3.

$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ Causes column 2 elements to appear so removed it.

~~This guide~~

(6) Delete ~~Columns~~ row 3.

Consider $A_3(A_2A_1BC_1C_2C_3)$, where $A_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

resultant is a 3×3 matrix, so A_3 must be 3×4.

$\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}^T$ ~~so~~ causes 3rd row-elements to appear in multiplication ~~last~~ of 4×4 matrix with I_4, so removed it ($I_4 \rightarrow$ 4×4 unit matrix)

Dt.:
Pg.

B+

Hence, the final expression is,

$$A_3\left(\left(A_2\left(\left((A_1 B)C_1\right)C_2\right)\right)C_3\right) = A_3A_2A_1BC_1C_2C_3$$

$$\downarrow$$

$$\left(\because \text{matrix multiplicat}^{n} \text{ is associative}\right)$$

with values of $A_1, A_2, A_3, C_1, C_2, C_3$ as given.

(b) $\because$ Matrix mul. is associative,

$$A_3A_2A_1BC_1C_2C_3 = \left(A_3(A_2A_1)\right)B\left(C_1(C_2C_3)\right)$$

$$= ABC$$

where,

$$A = \left(A_3(A_2A_1)\right) = A_3A_2A_1$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 5 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 5 & 0 & 0 \end{bmatrix}$$

$$C = \left(C_1(C_2C_3)\right) = C_1C_2C_3$$

$$= \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Dt. _____ B+
Pg. _____

$$\Rightarrow C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\therefore A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 5 & 0 & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(★) Reasons for part (a)

(2) to manipulate the columns, right multiplication is necessary. Also, for $AB = C$, $c_{ij} = a_i^T b_j$, where a_i is the i^{th} row of A & b_j is the j^{th} column of B.

for any row of A, R_i of A, B & 1^{st} col^m of C_1, C_j, $R_i^T C_j$ results in $2k_{i3} + k_{i1}$. Hence, C_1 must be $\begin{bmatrix} 1 \\ 0 \\ 2 \\ 0 \end{bmatrix}$. rest all columns of C_1 have to be such, that if C_j is the j^{th} column, then C_j has a 1 at j^{th} place & rest all 0's to preserve the positions of element.