MATH2059 Numerical Methods

HOMEWORK FOR MIDTERM EXAM

Spring 2019-2020

Start 14:00 – End 17:00

Name:
Student ID:

Q1	**/ 25**
Q2	**/ 25**
Q3	**/ 25**
Q4	**/ 25**
Total	**/ 100**

- Please use A4 size white blank sheets to write your answers with handwriting. You can print this booklet and write your answers on it, if you have a printer.
- Please write your name and student ID at the top of all your answer sheets. Please copy the above grading table to the first page as well.
- Please copy one of the sentences below to the first page of your answer sheet in handwriting and sign under the sentence.

  "Sorular tarafımca çözülmüş olup, hiçbir yolla başka bir şahıstan yardım alınmamış ve başka bir şahısla paylaşılmamıştır."

  "The questions have been answered by myself, and I have neither given nor received any unauthorized help in any way."
- There are 4 questions and 5 pages in this handout.
- You have 3 hours to complete your answers and upload them to UES in a single pdf document. The document should be named as "**name_surname.pdf**". You should pay attention to the following:
  - Use a scanner or a **scanner application on your mobile phone** (Adobe Scan etc.) to scan your answers into a single pdf file. **Do not take photos** of the pages.
  - Scan each page one by one in upright **portrait orientation** (NOT landscape).
  - Make sure that all pages are clear and **readable**.
  - Do not delay your submission to the last minute as there may be unexpected problems. Try to begin uploading your file at least 10 minutes before the deadline (17:00).
- **Chrome browser** is recommended to minimize any problems.
- SHOW ALL YOUR WORK CLEARLY AND STEP BY STEP TO RECEIVE CREDIT. Good luck!

QUESTION 1

Floating-Point Representation. Fractional quantities are typically represented in computers using *floating-point format*. In this approach, which is very much like scientific notation, the number is expressed as

$$\pm s \times b^e$$

where s = the *significand* (or *mantissa*), b = the base of the number system being used, and e = the exponent.

Prior to being expressed in this form, the number is *normalized* by moving the decimal place over so that only one significant digit is to the left of the decimal point. This is done so computer memory is not wasted on storing useless nonsignificant zeros. For example, a value like 0.005678 could be represented in a wasteful manner as 0.005678×10^0. However, normalization would yield 5.678×10^{-3} which eliminates the useless zeroes.

Suppose that we have a hypothetical base-10 computer with a 7-digit word size. Assume that one digit is used for the sign, three for the exponent, and three for the mantissa. For simplicity, assume that one of the exponent digits is used for its sign, leaving two digits for its magnitude:

$$s_1 d_2 . d_3 d_4 \times 10^{s_0 d_0 d_1}$$

where s_0 and s_1 represent the signs, $d_0 d_1$ represent the magnitude of the exponent, and digits d_2, d_3, d_4 are used for the magnitude of the significand.

(a) What is the largest positive number that can be represented on this computer?
(b) What is the smallest possible positive number?
(c) What is the smallest possible negative number?
(d) Suppose we want to represent the number $2^{-7} = 0.0078125$ on this computer. What would be the round-off error expressed as the absolute relative error?
(e) What is the minimum possible modification to this computer so that 2^{-7} can be represented exactly (with no roud-off error)?
(f) Explain the concepts of "underflow" and "overflow" for this computer. When do underflow and overflow occur?
(g) What would be the "machine epsilon" for this computer? (Note that machine epsilon is the smallest possible number such that $1 + \varepsilon \neq 1$) on the machine).

QUESTION 2 - Roots: Bisection Method

(a) The following procedure finds the simple root of $f(x) = 0$ in the interval $[a, b]$ using the bisection method so that the residual (the absolute value of the function at the current root estimate) is less than the given positive quantity ϵ. Some parts are missing; fill in the blanks.

1. Set $a^{(1)} = a, b^{(1)} = b$, and $i = 0$.
2. $i = i + 1$.
3. Calculate $x_m =$ _____.
4. If _____ $\leq \epsilon$, take the desired root as $\hat{x} = x_m$ and stop. Otherwise, continue to next step.
5. If _____, set $a^{(i+1)} = x_m$ and $b^{(i+1)} = b^{(i)}$, and go to step 2.
6. If _____, set $b^{(i+1)} = x_m$ and $a^{(i+1)} = a^{(i)}$, and go to step 2.

(b) Solve $f(x) = x - x^{1/3} - 2 = 0$ using bisection method, with $a = 3, b = 4$. Perform 5 iterations and fill in the following table. (Note: You must show your calculations at each iteration to receive credit.)

iteration	a	b	x_m	$f(a)$	$f(x_m)$
0	3	4	-	-	-
1	3	4			
2					
3					
4					
5					

QUESTION 3 – Roots: Fixed point iteration

Solve $f(x) = x - x^{1/3} - 2 = 0$ using fixed point iteration using two different update formulas. Let us represent the first update formula as:

$$x_{new} = g_1(x_{old})$$

And the second update formula as:

$$x_{new} = g_2(x_{old})$$

(a) Determine the functions $g_1(x)$ and $g_2(x)$ and write the two update formulas.

(b) Perform 5 iterations using the two update methods and fill in the below table. Start with the initial guess of 3. Let k denote the iteration number.

k	$g_1(k-1)$	$g_2(k-1)$
0	3	3
1		
2		
3		
4		
5		

(c) Compare your results in (b) and comment on the convergence behavior of the two functions and give the reasons.

(d) Compare the fixed-point iteration results with the bisection results in Question 2, and comment on the convergence behavior of the two methods.

QUESTION 4

(a) In IEEE double precision arithmetic the associative rule for summation may be violated. We will use the two methods to compute $1 + 10^{17} - 10^{17}$

i. Find the result of $1 + (10^{17} - 10^{17})$ = ……

ii. Find the result of $(1 + 10^{17}) - 10^{17}$ = ……

iii. Clearly explain why the results of (i) and (ii) are different.

(b) In IEEE double precision arithmetic the associative rule for multiplication may be violated. We will use two methods to compute $10^{155}10^{155}10^{-250}$.

i. Find the result of $10^{155}(10^{155}10^{-250})$ = ……..

ii. Find the result of $(10^{155}10^{155})10^{-250}$ = ……

iii. Clearly explain why the results of (i) and (ii) are different.