University of Toronto at Scarborough
CSCC37—Numerical Algorithms for Computational Mathematics, Fall 2021

Term Test

Duration: 90 minutes (estimated)

Date and Time: Wednesday 27 October, 17:00-19:00ET.

Aids allowed: Open-book. All aids are allowed.

This is a take-home test. Complete your solutions, and submit no later than the scheduled finish time for the test, **following the instructions given on the term test page of the course website.** (Note that you are given more than the estimated time to complete the test.)

This test consists of 5 questions. **Make sure your copy has 10 pages (including this one).** Write your answers in the spaces provided. You will be rewarded for concise, well-thought-out answers, rather than long rambling ones. Please write legibly.

Take a few minutes before you begin the test to read through each question, and then start with the question(s) you find easiest.

Name: _____ UTORid: _____
(Circle your family name.)

Student #: _____ Tutorial section: _____

YOU MUST SIGN THE FOLLOWING:

I declare that this test was written by the person whose name and student # appear above.

Signature: _____

Your grade

1.	_____ / 10	*4.*	_____ / 10
2.	_____ / 10	*5.*	_____ / 10
3.	_____ / 10		
		Total	_____ / 50

Graded by

Question 1

[10 marks]

Consider the normalized floating-point system $\mathbb{R}_6(7, 2)$ with exponent range $-(55)_6 \leq e \leq (55)_6$.

a. What is the smallest positive (nonzero) number representable? Give your answer in base-6.

b. What is the largest positive number representable? Give your answer in base-6.

c. What is the floating-point representation of $(407)_{10}$ in this system? Give your answer in base-6.

CONTINUED . . .

d. What is the floating-point representation of $(0.9)_{10}$ in this system? Give your answer in base-6.

Note: Recall that there are only 7 base-6 digits in the mantissa. If necessary, use rounding to truncate the mantissa.

e. Combing the results of (**c**) and (**d**), what is the floating-point representation of $(407.9)_{10}$ in this system? Give your answer in base-6. Again, if necessary, use rounding to truncate the mantissa.

f. Assuming round-to-nearest, what is the tightest upper bound on the relative error $|fl(x) - x|/|x|$ when $x \in \mathbb{R}$ is stored as $fl(x) \in \mathbb{R}_6(7, 2)$ in this floating-point system? Give your answer in base-10.

CONTINUED . . .

Question 2

[10 marks]

When each of the following expressions is evaluated using floating-point arithmetic, poor results are obtained for a certain range of values of x. In each instance, identify this range and provide an alternate expression that can be used for such values of x.

Note: As you are not given a specific floating-point system $\mathbb{R}_b(t, s)$, you can only give an *approximate* range where poor results are obtained. You may use words such as "close to" when identifying the range.

a. $(2 - (2 - x))$, initially evaluated according to the parenthetical precedence

b. $\sqrt{1 + x} - \sqrt{1 - x}$

CONTINUED . . .

c. $1 - \sin(x)$

d. $e^x - 1$

CONTINUED . . .

Question 3

[10 marks]

Consider the linear system $Ax = b,\ A \in \mathbb{R}^{3\times 3}, x, b \in \mathbb{R}^3$ with

$$A = \begin{bmatrix} 2 & 6 & 6 \\ 1 & 7 & 6 \\ 4 & 12 & 12 \end{bmatrix}, \quad b = \begin{bmatrix} 20 \\ 25 \\ 40 \end{bmatrix}.$$

a. Compute the $PA = LU$ factorization of A. Use exact arithmetic. Show all intermediate calculations, including Gauss transforms and permutation matrices.

b. Use the factorization computed in (**a**) to solve the system, if a solution exists. If there is no solution, **you must justify why there is no solution**.

CONTINUED . . .

Question 4

[10 marks]

Let $A \in \mathbb{R}^{n \times n}$ be a nonsingular matrix, and let $x_i, t_i \in \mathbb{R}^n, i = 1, \ldots, k$.

a. Explain how Gaussian elimination with partial pivoting can be used to solve the k linear systems

$$Ax_1 = t_1,\ Ax_2 = t_2,\ \ldots,\ Ax_k = t_k$$

as *efficiently as possible*. Derive the complexity of your approach by counting *flops* (multiplication/addition pairs).

Notes: (1) You do **not** need to give the details of Gaussian elimination—you may assume the $PA = LU$ factorization exists at the cost (complexity) quoted in lecture. (2) Your final operation count will include both n and k.

b. For $k = n$ and a suitable choice for $t_1, t_2, \ldots, t_n$, your algorithm in (**a**) can be used to compute A^{-1}. Explain how this can be done.

c. A possible scheme for solving $Ax = b$ is to first compute A^{-1} using (**b**), and then compute $x = A^{-1}b$. Is this scheme preferable to the Gaussian elimination algorithm as discussed in lecture? Give operation counts to justify your answer.

CONTINUED . . .

Question 5

[10 marks]

Let $\hat{x}$ be a computed solution to $Ax = b,\ A \in \mathbb{R}^{n\times n}, x, b \in \mathbb{R}^n$. The following bound for the relative error in $\hat{x}$ was derived in class:

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \text{cond}(A)\, \frac{\|r\|}{\|b\|},$$

where $r \in \mathbb{R}^n, r = b - A\hat{x}$. Starting with the equations $Ax = b$ and $A\hat{x} = b - r$, derive a *lower* bound for $\|x - \hat{x}\|/\|x\|$. What do these bounds tell us about the reliability of $\hat{x}$?

END OF TEST