University of Toronto Scarborough
Department of Computer and Mathematical Sciences

December Examinations 2020

CSC C37H3 F

Duration—3 hours

This is a take-home exam. Complete your solutions, and submit no later than the Registrar's scheduled finish time for the exam, **following the instructions given on the final exam page of the course website.**

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

This exam consists of 7 questions. **Make sure the copy you have downloaded has 14 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 exam 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 exam was written by the person whose name and student # appear above.

Signature: _____

	Your grade
1.	_____ / 15
2.	_____ / 10
3.	_____ / 15
4.	_____ / 15
5.	_____ / 15
6.	_____ / 15
7.	_____ / 15
Total	_____ / 100

Question 1

[15 marks]

Consider the normalized floating-point system $\mathbb{R}_3(3, 1)$ with *limited* exponent range $-1 \leq e \leq 1$.

a. What is the smallest positive (nonzero) number representable? Give your answer in both base-3 and base-10.

b. What is the largest positive number representable? Give your answer in both base-3 and base-10.

c. 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 $fl(x) \in \mathbb{R}_3(3, 1)$ in this floating-point system? Give your answer in base-10.

d. What is the floating-point representation of $(407)_{10}$ in this system? (*Hint*: Does the representation exist?)

e. What is the floating-point representation of $(0.567)_{10}$ in this system? Give your answer in base-3. Recall that there are only *three* base-3 digits in the mantissa.

f. List all possible normalized, non-zero mantissas in this system. In total, how many floating-point numbers are representable? Recall that the exponent range is *limited*.

CONTINUED . . .

Question 2

[10 marks]

A floating-point operation, or *flop*, is an operation of the form $mx + b$. Show how to convert a $(k+1)$-digit base b ($b \neq 10$) positive integer

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

into its base 10 equivalent in k flops or less.

CONTINUED . . .

Question 3

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 5 & 10 \\ 8 & 32 & 8 \\ 1 & 8 & 13 \end{bmatrix}, \quad b = \begin{bmatrix} 7 \\ -16 \\ 6 \end{bmatrix}.$$

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

CONTINUED . . .

b. Use the factorization computed in (**a**) to solve the system.

c. Instead of first computing the $PA = LU$ factorization, we could have solved the system above by processing the left and right-hand sides simultaneously with Gauss transforms and permutations. Would this alternate approach incur any extra cost? **Explain.** We are solving one system only.

Question 4

[15 marks]

For $f : \mathbb{R} \to \mathbb{R}$, recall Newton's method for solving $f(x) = 0$ starting from an initial guess x_0:

$$x_{k+1} = x_k - f(x_k)/f'(x_k),\ k = 0, 1, \ldots,\ \text{until convergence.}$$

a. Describe two possible tests for "convergence".

b. What is the purpose of the following fixed-point iteration? (*Hint*: Identify it as the Newton iteration for a certain function.)

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

c. Design a Newton iteration for computing $\sqrt[3]{\alpha}$, $\alpha > 0$. Your iteration will *not* converge for certain starting values. Identify one such value, and prove that your iteration will not converge from that value.

Question 5

[15 marks]

Determine p, q and r so that the order of the fixed-point iteration

$$x_{k+1} = px_k + \frac{qa}{x_k^2} + \frac{ra^2}{x_k^5}$$

for computing $a^{1/3}$ becomes as high as possible. For this choice of p, q and r, indicate how the error in x_{k+1} depends on the error in x_k.

(*Hint*: When deciding the conditions your iteration must satisfy, remember that it is necessary for $a^{1/3}$ to be a fixed point.)

CONTINUED . . .

Question 6

[15 marks]

For the simple interpolation problem

$$p(x_i) = y(x_i),\ i = 0, \ldots, n,$$

we stated (but did not derive) the polynomial interpolation error formula:

$$E(x) = y(x) - p(x) = \frac{y^{(n+1)}(\xi)}{(n+1)\,!}\, W(x) \tag{1}$$

where

$$W(x) = \prod_{i=0}^{n}(x - x_i) \ \text{ and } \ \xi \in \text{span}\{x, x_0, x_1, \ldots, x_n\}.$$

We argued that for certain underlying functions $y(x)$, the upper bound for $E(x)$ can increase dramatically as the number of interpolation points increases.

a. Identify the two main parts of (1) by boxing and circling, as was done in lecture. Which part of (1) contributes most to the increase in the upper bound for $E(x)$? Explain.

b. Can the growth in either part of (1) be controlled? If so, how can it be controlled?

c. A classic example of an underlying function which results in the upper bound for $E(x)$ to increase rapidly is the *Runge function* $y(x) = 1/(1 + 25x^2)$. For $p(x) \in \mathcal{P}_n$ interpolating equally spaced points in $[-1, 1]$, sketch a graph illustrating the "Runge Phenomenon".

d. Consider Weierstrass's theorem:

> "Let $y \in \mathcal{C}[a, b]$. For each $\epsilon > 0$ there exists a polynomial $p(x)$ of degree N_ϵ (N_ϵ depends on ϵ) such that $\|y - p\|_\infty < \epsilon$."

This theorem says that any continuous function on the interval $[a, b]$ may be approximated as closely as we like by some polynomial. Yet, as we saw in lecture and as (hopefully) you have illustrated in (c), Runge's function *cannot* be accurately interpolated with a single polynomial on the interval $[-1, 1]$.

Does the Runge Phenomenon contradict Weierstrass's theorem? Explain.

CONTINUED . . .

Question 7

[15 marks]

Consider the data points $\{(0, 3), (1, 7), (2, 37), (3, 141)\}$.

a. Set up the Vandermonde system for determining the monomial-basis form of the polynomial which interpolates these data points. Do not solve the system.

b. Derive the Newton form of the interpolating polynomial. Show all of your work, including the divided-difference table.

c. Derive the Lagrange form of the interpolating polynomial. Verify it is the same polynomial as in (b).

d. Briefly discuss the relative efficiency of the methods in (a), (b), and (c). Which method is best if we need to include additional data points? Explain.

CONTINUED . . .

e. Construct the linear spline (i.e., the piecewise linear interpolant) which interpolates all four data points $\{(0, 3), (1, 7), (2, 37), (3, 141)\}$.

END OF EXAM