University of Toronto at Scarborough

CSCC37—Numerical Algorithms for Computational Mathematics, Fall 2020

Term Test

Duration: 120 minutes (estimated)

Date and Time: Saturday 14 November, 4:00–8:00 p.m.

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 7 questions. **Make sure your copy has 13 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	*5.*	______ / 10
2.	______ / 5	*6.*	______ / 10
3.	______ / 15	*7.*	______ / 15
4.	______ / 15		
		Total	______ / 80

Graded by [ ]

Question 1

[10 marks]

Design a computer that is able to store $(0.1)_{10}$ exactly. A floating point number on your computer must be represented internally in a base less than 10, and must have a mantissa with a finite number of digits. (**Hint:** Is this possible?)

Question 2

[5 marks]

What numbers are representable with a finite expression in the binary system but are not finitely representable in the decimal system? **Justify your answer.**

Question 3

[15 marks]

We wish to evaluate $f(x) = \beta - \sqrt{\beta^2 - x^2}$ where $\beta \gg 0$ and $|x| < \beta$.

a. Identify the range of x where $f(x)$ suffers from potential subtractive cancellation. An approximate range will suffice.

b. Evaluate $\lim_{x \to \alpha} \text{cond}(f(x))$, where α is the smallest number, in absolute value, within the range you identified in (**a**). Does the condition number reflect the potential for subtractive cancellation? **Explain.**

c. Derive an alternate form of $f(x)$, stable for evaluation in the range you identified in (**a**).

Question 4

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 6 & 6 \\ 3 & 5 & 12 \\ 6 & 6 & 12 \end{bmatrix}, \quad b = \begin{bmatrix} 20 \\ 25 \\ 30 \end{bmatrix}.$$

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.

c. Why is Gaussian Elimination usually implemented as in this question (i.e., $PA = LU$ is computed separately, and then the factorization is used to solve $Ax = b$)?

Question 5

[10 marks]

Recall in lecture we discussed the geometric interpretation of the manifestation of round-off error during the Gaussian Elimination/LU factorization process. We drew two graphs depicting the intersection of lines which represented, respectively, the solution of a poorly conditioned and a perfectly conditioned linear system $Ax = b, A \in \mathcal{R}^{2\times 2}, x, b \in \mathcal{R}^2$.

a. Reproduce the graphs below. As in lecture, draw the true systems with solid lines and the systems resulting from roundoff error with dashed lines. Clearly label the true solution and the approximate solutions on each graph.

b. Copy the graph representing the poorly conditioned system to the space below. Show how the residual vector $r = b - A\hat{x}$ manifests on the graph. (**Note:** This was not discussed in lecture.)

c. The solution of a linear system $Ax = b, A \in \mathcal{R}^{3\times 3}, x, b \in \mathcal{R}^3$ is the line or point of intersection of three planes. In the space below, either draw a graph representing a *perfectly* conditioned 3-dimensional linear system, or, if you are not able to easily depict a 3-dimensional graph, describe the graph in 25 words or less.

Question 6

[10 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2+\epsilon \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \tag{1}$$

and $0 \leq \epsilon < 1$.

a. Derive a formula for $\text{cond}_1(A)$, the 1-norm condition number of A. What is $\lim_{\epsilon\to 0} \text{cond}_1(A)$?

b. Sketch a graph illustrating the general trend of (1) as $\epsilon \to 0$. (Since you are not given specific values for the right-hand side b, you cannot pin down exact x and y intercepts.) Also show on the graph the potential effect(s) of small perturbations in the coefficients of A, such as those introduced when (1) is solved on a computer using Gaussian Elimination with partial pivoting.

Question 7

[15 marks]

Consider the iterative improvement algorithm discussed in lecture:

Solve $Ax = b$ for initial approximation $\hat{x}_0$.
for $i = 0, 1, \ldots$ until convergence
  compute $r_i = b - A\hat{x}_i$
  solve $Az_i = r_i$
  update $\hat{x}_{i+1} = \hat{x}_i + z_i$
end for

We gave an intuitive explanation of why this algorithm could improve the initial approximate solution $\hat{x}_0$, but we were vague on the conditions required for convergence. In this question, you will attempt to derive the precise conditions required for convergence.

Starting with $Az_i = r_i$ and $(A+E)\hat{z}_i = r_i$, derive a formula showing how the absolute error in the $(i+1)$-st iterate $\|\hat{x}_{i+1} - x\|$ is bound by a multiple of the absolute error in the i-th iterate $\|\hat{x}_i - x\|$. Argue that the magnitude of this multiple is dependent on the condition of A, and is less than one (hence convergence) if A is well-conditioned.

[Continue your answer on the next page if necessary . . .]

[…continue your answer to #7 here.]

END OF TEST