University of Toronto

Faculty of Arts and Science

December 2010 Examinations

CSC 336 H1F — Numerical Methods

Duration — 3 hours

No Aids Allowed

Answer ALL Questions

Do **NOT** turn this page over until you are **TOLD** to start.

Please fill-in **ALL** the information requested on the front cover of **EACH** exam booklet that you use.

The exam consists of 5 pages, including this one. Make sure you have all 5.

The exam consists of 6 questions. **Answer all 6 questions**. The mark for each question is listed at the start of the question. Do the questions that you feel are easiest first.

The exam was written with the intention that you would have ample time to complete it. You will be rewarded for concise well-thought-out answers, rather than long rambling ones. **We seek quality rather than quantity**.

Moreover, an answer that contains relevant and correct information as well as irrelevant or incorrect information will be awarded fewer marks than one that contains the same relevant and correct information only.

Write legibly. Unreadable answers are worthless.

1. [10 marks; 2 marks for each part]

   For each of the five statements below, say whether the statement is **true** or **false** and briefly justify your answer.

   (a) A problem is ill-conditioned if its solution is highly sensitive to small changes in the problem data.

   (b) In the IEEE double-precision floating-point number system, addition is not always commutative. That is, you can find double-precision floating-point numbers a and b for which $\mathrm{fl}(a+b) \neq \mathrm{fl}(b+a)$.

   (c) The product of two lower-triangular $n \times n$ matrices is a lower-triangular $n \times n$ matrix.

   (d) If a linear system is well-conditioned, then pivoting is not necessary in Gaussian elimination.

   (e) Newton's method is an example of a fixed-point iteration scheme.

2. [10 marks: 5 marks for each part]

   Consider the expression

$$\sqrt{1+x} - \sqrt{1-x} \tag{1}$$

   for $x \in [-1, 1]$.

   (a) For what range of values of x does expression (1) produce inaccurate results (in a relative error sense) in IEEE double-precision floating-point arithmetic?

   Justify your answer.

   (b) Give another expression that is mathematically equal to (1) for which the computation is more accurate (in a relative error sense) in floating-point arithmetic for the problematic range of x identified in part (a).

   Explain why you believe your new expression is more accurate.

3. [10 marks: 5 marks for each part]

In Assignment 1, you computed an approximation to e^x by summing the terms in the series

$$1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

from left to right until the sum did not change. You can approximate $\sin(x)$ in a similar way by summing the terms in the series

$$x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

from left to right until the sum does not change.

(a) Assume that you are using IEEE double-precision floating-point arithmetic to compute the approximation to $\sin(x)$ described above. For what range of values of x do you expect this algorithm to produce an accurate approximation to $\sin(x)$ and for what range of values of x do you expect this algorithm to produce an inaccurate approximation to $\sin(x)$?

Justify your answer.

(b) Observe that you can decompose x as $x = y + n\pi$ for $y \in [-\pi/2, \pi/2]$ and n an integer. You can compute n by rounding x/π to the nearest integer (the `round` statement in MatLab will do this) and setting $y = x - n\pi$. Then $\sin(x) = (-1)^n \sin(y)$.

You don't have to prove the results in the paragraph above. I'm giving them to you as background information.

Explain how you can use the observation in the first paragraph of part (b) to improve the algorithm for approximating $\sin(x)$ described at the beginning of this question.

Explain why you believe this new algorithm will be accurate for all normalized IEEE double-precision floating-point numbers of small to medium magnitude (e.g., $|x| \leq 10^6$).

4. [20 marks: 5 marks for each part]

Consider the matrix

$$A = \begin{pmatrix} 2 & 4 & 2 \\ 4 & 8 & 2 \\ -2 & -1 & 1 \end{pmatrix}$$

(a) Using partial pivoting, compute the LU factorization of A. That is, compute the 3×3 permutation matrix P, the 3×3 unit-lower-triangular matrix L and the 3×3 upper-triangular matrix U such that $PA = LU$.

Show all your calculations.

(b) Use the LU factorization of A computed in part (a) to solve the linear system $Ax = b$, where

$$b = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}$$

Show all your calculations.

(c) Suppose we change the (3,1) element of A from -2 to 1 to yield a new matrix

$$\hat{A} = \begin{pmatrix} 2 & 4 & 2 \\ 4 & 8 & 2 \\ 1 & -1 & 1 \end{pmatrix}$$

Note that all the elements of A and $\hat{A}$ are equal except for the (3,1) element.

Find two vectors u and v such the $\hat{A} = A - uv^T$. Thus, A and $\hat{A}$ differ by a rank 1 update.

(d) Use the Sherman-Morrison formula

$$(A - uv^T)^{-1} = A^{-1} + \frac{A^{-1}uv^TA^{-1}}{1 - v^TA^{-1}u} \tag{2}$$

to solve $\hat{A}\hat{x} = b$, where $\hat{A}$ is the matrix in part (c) and b is the vector in part (b). Do not compute any inverses explicitly in the Sherman-Morrison formula (2). Instead, use the LU factorization from part (a) whenever you need to solve a linear system with the matrix A.

Show all your calculations.

5. [5 marks]

Assume that you have already computed the LU factorization of an $n \times n$ nonsingular matrix A. That is, you have computer an $n \times n$ permutation matrix P, an $n \times n$ unit-lower-triangular matrix L and an $n \times n$ upper-triangular matrix U such that $PA = LU$. How can you use this LU factorization to solve $A^T x = b$, where A^T is the transpose of A?

6. [10 marks: 5 marks for each part]

Most computer languages have a built-in function to compute $\sqrt{a}$ for any real number $a \geq 0$. An effective way of derive such a built-in function is to apply Newton's method to find the non-negative root of the quadratic

$$f(x) = x^2 - a.$$

This root is, of course, $\sqrt{a}$.

(a) Show that Newton's method for finding a root of $f(x) = x^2 - a$ is mathematically equivalent to the computationally-efficient fixed-point iteration

$$x_{k+1} = \frac{1}{2}\left(x_k + \frac{a}{x_k}\right) \tag{3}$$

Show all of the calculations necessary to justify your answer.

(The fixed-point iteration (3) for computing $\sqrt{a}$ is often called *Heron's Formula* and has been know since antiquity.)

(b) Show that the fixed-point iteration (3) is quadratically convergent if $a > 0$, but only linearly convergent if $a = 0$.

Show all of the calculations necessary to support your answer. Don't just quote the result that Newton's method is quadratically convergent for any simple root of $f(x)$.

(Note that the linear convergence of the fixed-point iteration (3) for $a = 0$ is not a problem in practice, since we know $\sqrt{0} = 0$, so we don't need to use the iteration (3) to compute $\sqrt{0}$.)

Have a Happy Holiday

Total Marks = 65

Total Pages = 5