University of Toronto

Faculty of Arts and Science

December 2014 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.

Write your answers in the exam booklets provided.

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 pages**.

The exam consists of 6 questions. **Answer all 6 questions**. Each question is worth 10 marks.

To pass this course, you need a total mark for the course of at least 50% and you must receive at least 35% on this the Final Exam.

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) Using higher-precision arithmetic will make a poorly conditioned problem better conditioned.

(b) If A is a symmetric matrix, then $\|A\|_1 = \|A\|_\infty$.

(c) Let A be an $n \times n$ nonsingular real matrix. If the determinant of A is close zero, then A must be badly conditioned (i.e., cond(A) is very large).

(d) If an iterative method for finding the root of an equation gains more than 1 bit of accuracy per iteration, then it has a superlinear rate of convergence.

(e) Suppose we approximate a function $f(x)$ on an interval $[a, b]$ of finite length (i.e., $b - a$ is finite) by a polynomial $p_n(x)$ of degree n or less that interpolates $f(x)$ at the $n + 1$ evenly spaced points

$$x_i = a + \frac{i}{n}(b - a), \quad \text{for } i = 0, 1, 2, \ldots, n,$$

in $[a, b]$. (Note that $x_i \in [a, b]$, for $i = 0, 1, 2, \ldots, n$, $x_i - x_{i-1} = 1/n$, for $i = 1, 2, \ldots, n$ and $p_n(x_i) = f(x_i)$, for $i = 0, 1, 2, \ldots, n$.) Is it always the case that

$$\lim_{n\to\infty} \max_{x\in[a,b]} |f(x) - p_n(x)| = 0$$

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

Kepler's equation

$$M = x - e\sin(x) \qquad (1)$$

arises in celestial mechanics. For this question, let M and e be constants, with $0 < e < 1$, and let x be a variable.

(a) Show that Kepler's equation (1) has exactly one solution, x^*, and that $x^* \in [M - 1, M + 1]$.

(b) In this part, you can assume that Kepler's equation (1) has exactly one solution, x^*, and that $x^* \in [M - 1, M + 1]$ even if you did not prove it in part (a).
Let x_0 be any point in the interval $[M - 1, M + 1]$ and let

$$x_{n+1} = g(x_n), \quad \text{for } n = 0, 1, 2, \ldots$$

where

$$g(x) = M + e\sin(x)$$

Show that $x_n \to x^*$ as $n \to \infty$.

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

You often need to compute the midpoint, m, of an interval $[a, b]$, where $a \in \mathbb{R}$, $b \in \mathbb{R}$ and $a \leq b$. For example, you need to compute m in the bisection method.

There are several mathematically equivalent ways to compute m. One possibility is

$$m = \frac{a+b}{2} \tag{2}$$

Another possibility is

$$m = a + \frac{b-a}{2} \tag{3}$$

Your book notes in §5.5.1 that, from a computational point-of-view, (3) can be much better than (2) in some cases.

(a) Assume that you are using a 3-decimal-digit floating-point number system with an exponent range 10^{-10} to 10^{+10}. That is, the nonzero normalized floating-point numbers in this system are of the form $\pm d_1.d_2d_3 \times 10^n$, where $0 \leq d_i \leq 9$ for $i = 1, 2, 3$, $d_1 \neq 0$ and $-10 \leq n \leq 10$. There is a special representation for 0.

Also assume that this 3-decimal-digit floating-point number system correctly implements the round-to-nearest rounding mode (see below for a definition of round-to-nearest rounding mode).

Give an example that shows that there are normalized floating-point numbers, a and b, in this 3-decimal-digit floating-point number system such that $a < b$ and

$$\mathrm{fl}\left(\frac{a+b}{2}\right) \notin [a, b]$$

In this example, a, b and $\mathrm{fl}(\frac{a+b}{2})$ should all be normalized floating-point numbers in this 3-decimal-digit floating-point number system. In particular, there should be no overflow or underflow in the computation of $\mathrm{fl}(\frac{a+b}{2})$.

(b) Consider any floating-point number system that correctly implements the round-to-nearest rounding mode (see below for a definition of round-to-nearest rounding mode). Let a and b be any two normalized floating-point numbers in this floating-point number system with $a \leq b$. Show that

$$\mathrm{fl}\left(a + \frac{b-a}{2}\right) \in [a, b]$$

provided that no overflows or underflows occur in the computation above.

State any other assumptions that you need to make in proving this result.

By round-to-nearest rounding mode, I mean that, if op is any one of the four basic arithmetic operations, $+$, $-$, $*$, $/$, and x and y are any two floating-point numbers in this floating-point number system, then $\mathrm{fl}(x \text{ op } y)$ is the floating-point number closest to the true value, x op y, provided that an overflow does not occur in the computation $\mathrm{fl}(x \text{ op } y)$.

4. [10 marks: 5 marks for each part]

Consider the linear system $Ax = b$, where

$$A = \begin{pmatrix} 1 & 2 & -3 \\ -1 & 1 & 2 \\ 3 & 0 & -3 \end{pmatrix} \qquad \text{and} \qquad b = \begin{pmatrix} -3 \\ -1 \\ 3 \end{pmatrix}$$

(a) Compute the LU factorization with partial pivoting of the matrix A. That is, compute a permutation matrix P, a unit-lower-triangular matrix L for which all elements $L_{i,j}$ satisfy $|L_{i,j}| \leq 1$ and an upper triangular matrix U such that $PA = LU$.

Show all your calculations.

(b) Use the LU factorization of the matrix A from part (a) to solve the linear system $Ax = b$.

Show all your calculations.

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

Let A be an $n \times n$ nonsingular matrix and let U and V be $n \times m$ matrices with $m \leq n$. Let I_m be the $m \times m$ identity matrix and assume that the $m \times m$ matrix $I_m - V^T A^{-1} U$ is nonsingular. The Woodbury formula

$$(A - UV^T)^{-1} = A^{-1} + A^{-1}U(I_m - V^T A^{-1} U)^{-1} V^T A^{-1} \tag{4}$$

is a generalization of the Sherman-Morrison formula that you used in Assignment 3.

(a) Show that the right side of (4) really is the inverse of $A - UV^T$.

Hint: Multiply the right side of (4) by $A - UV^T$ and show that the result simplifies to I_n, the $n \times n$ identity matrix.

(b) Assume that you have already computed the LU factorization of A. That is, you have a permutation matrix P, a unit lower triangular matrix L for which all elements $L_{i,j}$ satisfy $|L_{i,j}| \leq 1$ and an upper triangular matrix U such that $PA = LU$.

Suppose you now need to solve $(A - UV^T)x = b$. Assume that $m \ll n$. Show how you can use (4) to compute x in much less time than computing an LU factorization of $A - UV^T$ and using it to solve $(A - UV^T)x = b$.

Explain why the numerical method you propose is much more efficient that computing an LU factorization of $A - UV^T$ and using it to solve $(A - UV^T)x = b$.

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

Inverse Interpolation is another way to find roots of an equation. It is described in §5.5.5 of your textbook, but we didn't discuss it in class.

Suppose you want to find a root of $f(x)$. That is, you want to find a point x^* such that $f(x^*) = 0$. In inverse interpolation, we assume that the inverse of the function f exists near the root x^* and we interpolate the inverse function.

The inverse of f is a function f^{-1} such that, if $y = f(x)$, then $f^{-1}(y) = x$. In particular, note that, since $0 = f(x^*)$, $f^{-1}(0) = x^*$.

In inverse interpolation, we approximate $f^{-1}(y)$ by an interpolating polynomial $p_n(y)$ and then approximation $x^* = f^{-1}(0)$ by $p_n(0)$.

To be more specific, assume that $x_n, x_{n-1}, \ldots, x_{n-k}$ are all approximations to the root x^*. Let $y_{n-i} = f(x_{n-i})$, for $i = 0, 1, \ldots, k$. If the inverse function $f^{-1}(y)$ exits near the root x^*, then the y_{n-i}, for $i = 0, 1, \ldots, k$, must be distinct. (This is a key assumption.) So, $f^{-1}(y_{n-i}) = x_{n-i}$, for $i = 0, 1, \ldots, k$. Now find the polynomial $p_n(y)$ of degree k of less that satisfies

$$p_n(y_{n-i}) = x_{n-i}, \quad \text{for } i = 0, 1, \ldots, k \tag{5}$$

Then let

$$x_{n+1} = p_n(0) \tag{6}$$

To start the method, we need $k + 1$ initial guesses $x_0, x_1, \ldots, x_k$ for the root x^*. Then we find hopefully better and better approximations x_{n+1} to the root x^* by the following algorithm.

```
for $n = k, k+1, k+2, \ldots$
    Find the polynomial $p_n(y)$ of degree $k$ of less that satisfies (5)
    $x_{n+1} = p_n(0)$
end
```

(a) Show that inverse interpolation with $k = 1$ is equivalent to the secant method

$$x_{n+1} = x_n - f(x_n)\,\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}$$

(b) For $k = 2$, write out the formula (6) in terms of the values x_{n-i} and $y_{n-i} = f(x_{n-i})$, for $i = 0, 1, 2$.

Have a Happy Holiday

Total Marks = 60

Total Pages = 5