University of Toronto

Faculty of Arts and Science

December 2016 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 7 pages, including this one. **Make sure you have all 7 pages**.

The exam consists of 5 questions. **Answer all 5 questions**.

The mark for each question is listed at the start of the question. Do the questions that you feel are easiest first.

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) A problem that is highly sensitive to small changes in the problem data is poorly conditioned.

(b) In a floating-point number system, the *underflow level* (i.e., UFL in your textbook) is the largest positive floating-point number δ such that $\text{fl}(1+\delta) = 1$, where $\text{fl}(1+\delta)$ is the floating-point value you get when you compute $1 + \delta$ in this floating-point number system.

(c) Let A be an $n \times n$ nonsingular real matrix. If the condition number of A is very large, then the determinant of A must be close to zero.

(d) For a given fixed level of accuracy, a super-linearly convergent iterative method always requires fewer iterations than a linearly convergent method to find a solution to that level of accuracy.

(e) Given three pairs of points (x_1, y_1), (x_2, y_2), (x_3, y_3), where $x_i \in \mathbb{R}$ for $i = 1, 2, 3$, $y_i \in \mathbb{R}$ for $i = 1, 2, 3$ and the x_i are distinct, it is always possible to find a polynomial $p(x)$ of degree 2 or less such that $p(x_i) = y_i$ for $i = 1, 2, 3$.

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

In 250 BC, the Greek mathematician Archimedes estimated the number π as follows. He drew a circle with diameter 1, and hence circumference π. Inside the circle he inscribed a square. The perimeter of the square is smaller than the circumference of the circle, and so it is a lower bound for π. Archimedes then considered an inscribed octagon, 16-gon, 32-gon, etc., each time doubling the number of sides of the inscribed polygon, thereby producing ever better estimates of π, with each estimate less than π. He also performed a similar calculation using polygons that contain the circle (i.e., circumscribed polygons). This also produced a series of estimates converging to π, but with each estimate greater than π. Using 96-sided inscribed and circumscribed polygons, he was able to show that $223/71 < \pi < 22/7$.

In this question, we will focus on the inscribed polygons. There is a recursive formula for these estimates. Let p_n be the perimeter of an inscribed polygon with 2^n sides. It is easy to show that $p_2 = 2\sqrt{2}$ and it is possible (but a little harder) to show that

$$p_{n+1} = 2^n \sqrt{2\left(1 - \sqrt{1 - (p_n/2^n)^2}\right)} \qquad \text{for } n \geq 2 \tag{1}$$

As noted above, $p_n < \pi$ for all $n \geq 2$ and $p_n \to \pi$ as $n \to \infty$.

You don't have to prove the results above. Just take them as facts.

I used formula (1) above to compute p_n and the error $p_n - \pi$ for $n = 2, 3, \ldots, 33$ in MatLab using IEEE double-precision floating-point arithmetic. The results that I obtained are listed below.

n	p_n	$p_n - \pi$
2	2.828427124746190	-3.1317e-01
3	3.061467458920719	-8.0125e-02
4	3.121445152258053	-2.0148e-02
5	3.136548490545941	-5.0442e-03
6	3.140331156954739	-1.2615e-03
7	3.141277250932757	-3.1540e-04
8	3.141513801144145	-7.8852e-05
9	3.141572940367883	-1.9713e-05
10	3.141587725279961	-4.9283e-06
11	3.141591421504635	-1.2321e-06
12	3.141592345611077	-3.0798e-07
13	3.141592576545004	-7.7045e-08
14	3.141592633463248	-2.0127e-08
15	3.141592654807589	1.2178e-09
16	3.141592645321215	-8.2686e-09
17	3.141592607375720	-4.6214e-08

n	p_n	$p_n - \pi$
18	3.141592910939673	2.5735e-07
19	3.141594125195191	1.4716e-06
20	3.141596553704820	3.9001e-06
21	3.141596553704820	3.9001e-06
22	3.141674265021758	8.1611e-05
23	3.141829681889202	2.3703e-04
24	3.142451272494134	8.5862e-04
25	3.142451272494134	8.5862e-04
26	3.162277660168380	2.0685e-02
27	3.162277660168380	2.0685e-02
28	3.464101615137754	3.2251e-01
29	4.000000000000000	8.5841e-01
30	0	-3.1416e+00
31	0	-3.1416e+00
32	0	-3.1416e+00
33	0	-3.1416e+00

You can see from the table above that the computed $p_n > \pi$ for $n = 15$ and several other values of $n > 15$. This violates the theoretical result that $p_n < \pi$ for all $n \geq 2$. Moreover, the error in the computed approximation p_n to π grows in magnitude for $n > 15$, rather than converging to zero, as it would if the calculation were done in exact arithmetic. Furthermore, the computed $p_n = 0$ for $n = 30, 31, 32, 33$ and is easy to see from formula (1) that p_n will be zero for all $n > 33$ as well, since, from formula (1), it follows easily that, if $p_n = 0$, then $p_{n+1} = 0$.

(a) Explain why formula (1) produces such poor approximations to π when computed using IEEE double-precision floating-point arithmetic.

As part of your answer to this question, explain why the computed $p_n = 0$ for $n \geq 30$.

In answering this question, you can use the numerical results from the table above. For example, you can claim without proof that all computed $p_n \in [0, 5]$.

Note: although rounding error should play a role in your answer to this question, there should be more to your explanation than just saying that there is rounding error in the computation, since there is rounding error in almost all floating-point computations, but most of them are accurate.

(b) Find a formula that is mathematically equivalent to formula (1), but does not suffer from the extreme loss of accuracy that we see in the numerical results above for formula (1).

3. [20 marks: 5 marks for each part]

Consider the matrix

$$A = \begin{pmatrix} -1 & -2 & 1 \\ 4 & 4 & -4 \\ 2 & -1 & -5 \end{pmatrix}$$

(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) above to solve the linear system $Ax = b$, where

$$b = \begin{pmatrix} -2 \\ 4 \\ -4 \end{pmatrix}$$

Show all your calculations.

(c) Suppose we change the (3,3) element of A from -5 to 1 to yield a new matrix

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

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

Note that the u and v that satisfy $\hat{A} = A - uv^T$ are not unique. To see this, let $\tilde{u} = \alpha u$ and $\tilde{v} = v/\alpha$ for any nonzero real number α. Then $\tilde{u}$ and $\tilde{v}$ also satisfy $\hat{A} = A - \tilde{u}\tilde{v}^T$.

Suggestion: given this freedom in choosing u and v, choose u and v above so that the calculations in part (d) below work out easily.

(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.

4. [15 marks: 5 marks for each part]

Consider the function $f(x) = 2 + \cos(x) - \mathrm{e}^x$ for $x \in \mathbb{R}$. Throughout this question, assume that x is measured in radians when computing $\cos(x)$ or $\sin(x)$.

You might may find the following table of values helpful in answering the questions below.

x	$2 + \cos(x)$	e^x
0.50000	2.87758	1.64872
0.60000	2.82534	1.82212
0.70000	2.76484	2.01375
0.80000	2.69671	2.22554
0.90000	2.62161	2.45960
1.00000	2.54030	2.71828
1.10000	2.45360	3.00417
1.20000	2.36236	3.32012
1.30000	2.26750	3.66930
1.40000	2.16997	4.05520
1.50000	2.07074	4.48169

(a) Find an interval of length at most 0.1 that contains a root of $f(x)$.
Justify your answer.

(b) How many roots does $f(x)$ have?
Justify your answer.

(c) Explain how you can use Newton's method to find a root of $f(x)$. In particular,

- give Newton's iteration to find a root of $f(x)$, and
- give a value for a good initial guess x_0 to start Newton's iteration for $f(x)$ and explain why you think your initial guess is a good choice for x_0.

5. [5 marks]

Find a polynomial $p(x)$ of degree 3 or less that satisfies

$$\begin{aligned} p(0) &= 1 \\ p(1) &= 0 \\ p(-1) &= 2 \\ p(2) &= 5 \end{aligned}$$

Have a Happy Holiday

Total Marks = 60

Total Pages = 7