

UNIVERSITY OF TORONTO

Faculty of Arts and Science

DECEMBER 2006 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 4 pages, including this one. Make sure you have all 4.

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) If two real numbers are each exactly representable as floating-point numbers, then their sum will also be exactly representable as a floating-point number.
- (b) For any  $x$ , if you compute the sum

$$S = \sum_{i=0}^n \frac{x^i}{i!}$$

in MatLab for a large enough  $n$ , then  $S$  will be a good approximation to  $e^x$ , provided that no overflows or underflows occur in computing either  $S$  or  $e^x$ .

- (c) If  $A$  is an  $n \times n$  nonsingular matrix, then  $\text{cond}(A) = \text{cond}(A^{-1})$ .
- (d) Since multipliers in Gaussian elimination with partial pivoting are bounded by 1 in magnitude, the entries in the successive reduced matrices in Gaussian elimination cannot grow in magnitude.
- (e) If  $\hat{x}$  is an approximation to a root  $x^*$  of a function  $f(x)$  and if  $|f(\hat{x})|$  is small, then  $\hat{x}$  must be close to  $x^*$ .

2. [5 marks]

Sam computed the expression

$$(2.0 + 1.0\text{e-}14) - 2.0$$

in MatLab and was very surprised to get back the result

$$1.021405182655144\text{e-}14$$

The answer in exact arithmetic is clearly  $1.0\text{e-}14$ . So the computed answer has two correct digits only.

Explain why MatLab computes such an inaccurate answer. In particular, why does the computed answer have about two correct digits only?

In answering this question, you may find it useful to recall that MatLab uses IEEE double-precision floating-point arithmetic in this computation. This is a base 2 system in which each normalized nonzero floating-point number has 53 binary digits in its significand. Thus *machine epsilon* for this system is  $2^{-52} \approx 2.22 \times 10^{-16}$ .

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

Bill wrote the MatLab function

```
function [r1,r2] = roots(a,b,c)
    r1 = ( -b + sqrt(b^2 - 4*a*c) ) / (2*a) ;
    r2 = ( -b - sqrt(b^2 - 4*a*c) ) / (2*a) ;
```

to compute the two roots  $r_1$  and  $r_2$  of the quadratic  $ax^2 + bx + c$ .

For  $a = c = 1$  and  $b = 10^7$ , his function returned the values

$$r_1 = -9.9652 \times 10^{-8} \quad r_2 = -1.0000 \times 10^{+7}$$

However, he knew that something was wrong, because he remembered from a high-school math course that the true roots  $r_1$  and  $r_2$  of the quadratic  $ax^2 + bx + c$  satisfy  $ar_1r_2 = c$ , but his computed roots satisfied  $ar_1r_2 = 0.99652$ , while  $c = 1$ . So he knew that at least one of the two computed roots must be inaccurate.

Bill checked his function carefully, but he couldn't find anything wrong with it.

- (a) Are both computed roots inaccurate or just one? If one, which one is inaccurate and which is accurate? Justify your answer.
- (b) Advise Bill on how to modify his function so that both computed roots are accurate. Explain why you believe your modification will produce accurate values for both roots.
4. [15 marks: 5 marks for each part]

- (a) Use the maximum norm (also called the max norm, infinity norm or  $\infty$ -norm) to calculate the condition number of the matrix

$$A = \begin{pmatrix} 15 & 5 \\ 1 & 2 \end{pmatrix}$$

You may use without verification that

$$A^{-1} = \frac{1}{25} \begin{pmatrix} 2 & -5 \\ -1 & 15 \end{pmatrix}$$

Show all your calculations.

- (b) Consider solving the system of linear equations  $Ax = b$ , where  $A$  is given above in part (a) and

$$b = \begin{pmatrix} 5 \\ 1 \end{pmatrix}$$

Suppose that each element in both  $A$  and  $b$  is known only to within  $\pm 1.0 \times 10^{-5}$ . Moreover, assume that you have access to a numerically stable linear equation solver and that all the arithmetic calculations are done exactly (i.e., without any roundoff error).

How accurately can you expect to be able to compute  $x$ ? Justify your answer.

Note that you are not asked to compute  $x$ .

- (c) How does your answer to part (b) change if you continue to assume that each element in both  $A$  and  $b$  is known only to within  $\pm 1.0 \times 10^{-5}$ , but you now assume also that all arithmetic calculations are done using double-precision IEEE binary floating-point arithmetic, rather than in exact arithmetic? Justify your answer.

Note again that you are not asked to compute  $x$ .

Hint: recall that *machine epsilon* is  $2^{-52} \approx 2.22 \times 10^{-16}$  in double-precision IEEE binary floating-point arithmetic.

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

Consider the linear system  $Ax = b$  where

$$A = \begin{pmatrix} 3 & 10 & 5 \\ 6 & 12 & 6 \\ 2 & 2 & 2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 6 \\ 4 \end{pmatrix}$$

(a) Compute the LU factorization with partial pivoting of the matrix  $A$ .

That is, compute a permutation matrix  $P$ , a unit-lower-triangular matrix  $L$  with all elements less than or equal to 1 in magnitude, 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.

6. [12 marks: 5 marks each for parts (a) and (b); 2 marks for part (c)]

The Cray 1 supercomputer did not have a divide unit. Instead, to compute  $a/b$  for  $b \neq 0$ , it first computed the reciprocal  $r = 1/b$  and then computed the product  $a \cdot r$ .

To compute the reciprocal  $r = 1/b$ , for  $b \neq 0$ , it used the fact that  $r$  is the solution of the equation

$$f(x) = 1/x - b = 0 \tag{1}$$

It first found an initial approximation  $r_0$  to  $r$  that was accurate to about half the digits in a floating-point number. Then, using  $r_0$  as an initial guess, it did one iteration of Newton's method applied to (1) to compute a final approximation  $r_1$  to  $r$ .

(a) Show that, if you apply Newton's method to (1), it is possible to re-arrange the terms in the resulting formula so that no divisions are required.

Write the formula in as computationally effective form as possible.

(b) Show that the error satisfies

$$\frac{r - r_1}{r} = \left( \frac{r - r_0}{r} \right)^2 \tag{2}$$

where  $r = 1/b$ .

(c) Why does (2) imply that  $r_1$  has roughly twice as many correct digits as  $r_0$  has?

Total Marks = 62

Total Pages = 4

Have a Happy Holiday