CSC 336 MIDTERM TEST 20 October 2017.

This is a **closed-book test**: no books, no notes, no calculators, no phones, no tablets, no computers (of any kind) allowed.

Duration of the test: 50 minutes (11:10 AM to noon).

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

Answer **ALL** Questions.

Write your answers in the test booklets provided.

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

The test consists of 4 pages, including this one. Make sure you have all 4 pages.

The test consists of 4 questions. **Answer all 4 questions**. The mark for each question is listed at the start of the question.

The test 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. [5 marks: 1 mark for each answer]

Consider a floating-point number system with parameters $\beta = 10$, $p = 3$, $L = -10$ and $U = +10$ that uses the *round-to-nearest* rounding rule and allows gradual underflow to subnormal numbers as well as underflow to zero. That is, the numbers in the system include zero and nonzero numbers of the form $\pm d_1.d_2d_3 \cdot 10^n$ where $d_i \in \{0, 1, 2, \ldots, 9\}$ for $i = 1$, 2, 3 and $n \in \{-10, -9, -8, \ldots, 10\}$. The normalized floating-point numbers in this system include 0 and the nonzero numbers of the form $\pm d_1.d_2d_3 \cdot 10^n$ with $d_1 \neq 0$. The subnormal numbers have $n = -10$, $d_1 = 0$ and $d_i \neq 0$ for $i = 2$ or 3. Like the IEEE floating-point number system, this number system also has the two special numbers +Infty and −Infty which stand for numbers that are too large in magnitude (either positive or negative, respectively) to represent in this floating-point system. The system also has a NaN, which stands for "not-a-number".

In the floating-point number system described above, what is the result of each of each of floating-point arithmetic operations (a)–(e) below? Write your answer as

- a normalized number in this floating-point system, if possible,
- a subnormal number in this floating-point system in the case of gradual underflow,
- zero in the case that the true answer is zero or there is an underflow to zero,
- +Infty or −Infty in the case of overflow,
- NaN if the result of the computation is not any of the above.

(a) $(7.54 \cdot 10^2) + (4.26 \cdot 10^1)$

(b) $(1.01 \cdot 10^5) \times (4.04 \cdot 10^{-2})$

(c) $(5.25 \cdot 10^{-7}) \times (-2.02 \cdot 10^{-5})$

(d) $(-6.06 \cdot 10^6) \times (2.02 \cdot 10^4)$

(e) $\big((6.06 \cdot 10^6) \times (2.02 \cdot 10^4)\big) - \big((5.05 \cdot 10^5) \times (3.03 \cdot 10^5)\big)$

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

Consider the expression

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

and assume $x \geq 0$.

(a) Suppose the expression (1) is computed using IEEE double-precision floating-point arithmetic. Give an example of a value of $x \geq 0$ for which the computed value of (1) has a very large relative error.

There should be no overflow of underflow in your example, just standard rounding errors.

Show that your example has a very large relative error.

(b) Find another expression that is mathematically equal to (1), but your new expression has a very small relative error (when computed using IEEE double-precision floating-point arithmetic) for all values of $x \geq 0$, provided there is no overflow or underflow in evaluating your new expression.

Explain why you believe your new expression has a very small relative error (when computed using IEEE double-precision floating-point arithmetic) for all values of $x \geq 0$, provided there is no overflow or underflow in evaluating your new expression.

3. [5 marks]

Assume

$$x = \begin{pmatrix} 3 \\ -5 \\ 1 \end{pmatrix} \qquad \text{and} \qquad A = \begin{pmatrix} 3 & -2 & 1 \\ -2 & 5 & -2 \\ 1 & -3 & -4 \end{pmatrix}$$

Give the value of each of the following norms.

(a) $\|x\|_1$

(b) $\|x\|_2$

(c) $\|x\|_\infty$

(d) $\|A\|_1$

(e) $\|A\|_\infty$

4. [5 marks]

Suppose that $\|\cdot\|_v$ is a norm for vectors in $\mathbb{R}^n$ (i.e., real vectors of dimension n). As we discussed in class, a matrix norm for real $n \times n$ matrices subordinate to (or induced by) this vector norm is defined by

$$\|A\|_m = \max_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v} \tag{2}$$

(In class, we didn't include the subscripts v and m on the norms, but I have added them here to make it clear which are the vector norms and which is the subordinate matrix norm.)

Show that, for this vector norm and associated matrix norm,

$$\|Ax\|_v \leq \|A\|_m \|x\|_v \tag{3}$$

for all $x \in \mathbb{R}^n$ and all real $n \times n$ matrices A.

Note that I told you in class that (3) is true and we used it several times, but I don't think we ever proved that it is true.