

This is a **closed-book test**: no books, no notes, no calculators allowed.

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

1. [5 marks]

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 with subnormal numbers. 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 $d_i \in \{0, 1, 2, \dots, 9\}$ for $i = 1, 2, 3$ and $n \in \{-10, -9, -8, \dots, 10\}$. For normalized nonzero numbers, $d_1 \neq 0$. For subnormal nonzero numbers, $n = -10$, $d_1 = 0$ 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 number system.

In the floating-point number system described above, what is the result of each of the following floating-point arithmetic operations? Write your answer as a normalized number in this floating-point system, if possible, or as a subnormal number in this floating-point system in the case of gradual underflow, or as +infty or -infty in the case of overflow.

- (a) $5.26 \cdot 10^3 + 6.58 \cdot 10^3$
- (b) $5.28 \cdot 10^1 - 2.16 \cdot 10^{-1}$
- (c) $2.25 \cdot 10^2 \times 3.45 \cdot 10^{-3}$
- (d) $-3.52 \cdot 10^5 \times 4.25 \cdot 10^6$
- (e) $3.45 \cdot 10^{-6} \times 5.27 \cdot 10^{-6}$

2. [5 marks]

In Assignment 1, you wrote a MatLab function to compute the sum of the series

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

from left to right until the accumulated sum stops changing. If you did this correctly, you would find that your function computes a very accurate approximation to the true value of the sum for positive x of small to medium magnitude, but a very inaccurate approximation to the true value of the sum for negative x , unless x is small in magnitude. For example, the computed value of the sum is very accurate for $x = 25$, but very inaccurate for $x = -25$.

Explain why rounding errors make the computation very inaccurate for $x = -25$, but do not have a serious affect on the accuracy for $x = 25$.

3. [5 marks]

Use the matrix infinity norm to compute the condition number of the matrix

$$A = \begin{pmatrix} 3 & 1 \\ 2 & 4 \end{pmatrix}$$

Show all your computations.

4. [5 marks]

Consider the system of linear equations $Ax = b$, where

$$A = \begin{pmatrix} 3.01 & 1 & 1 \\ 1 & 2.99 & 1 \\ 1 & 1 & 3.01 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

It's easy to see that an approximate solution to this system is

$$\tilde{x} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}$$

in the sense that the associated residual

$$r = b - A\tilde{x} = \begin{pmatrix} -0.002 \\ 0.002 \\ -0.002 \end{pmatrix}$$

is small. The condition number of A in the infinity-norm is approximately 3.014.

Can you conclude from this that \tilde{x} is close to the true solution x of $Ax = b$? (By close, I mean that the relative difference between x and \tilde{x} in the infinity norm is about 10^{-2} or less.)

Justify your answer.