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 AM to noon).

Do **NOT** turn this are **TOLD** to start.

Answer **ALL** Quest

Write your answers 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 [illegible] of the form $\pm d_1.d_2d_3 \cdot 10^n$ where $d_i \in \{0, 1, 2, \ldots, 9\}$ for $i = 1, 2, 3$ [illegible] $-8, \ldots, 10\}$. The normalized floating-point numbers in this system [illegible] nonzero numbers of the form $\pm d_1.d_2d_3 \cdot 10^n$ with $d_1 \neq 0$. The s[illegible] have $n = -10$, $d_1 = 0$ and $d_i \neq 0$ for at least one of $i = 2$ or $i =$ [illegible] floating-point number system, this number system also has the tw[illegible]-Infty and $-$Infty, which stand for numbers that are too large in m[illegible] sitive 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 the 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) $(5.26 \cdot 10^2) + (2.57 \cdot 10^1)$

(b) $(5.03 \cdot 10^5) \times (2.02 \cdot 10^{-2})$

(c) $(4.04 \cdot 10^{-7}) \times (-3.03 \cdot 10^{-5})$

(d) $(-3.03 \cdot 10^0) \times (5.85 \cdot 10^4)$

(e) $\big((7.06 \cdot 10^6) \times (2.03 \cdot 10^4)\big) - \big((4.02 \cdot 10^5) \times (3.01 \cdot 10^5)\big)$

2. [5 marks]

Assume both x and y are positive real numbers and $x \neq y$. In this case, we would expect some cancellation in computing $\log_e(x) - \log_e(y)$. On the other hand,

$$\log_e(x) - \log_e(y) = \log_e(x/y)$$

and $\log_e(x/y)$ [illegible]tions (hence no cancellation). Does this mean that computing log [illegible]ve a more accurate result than computing $\log_e(x) - \log_e(y)$?

Hint: for what [illegible]on $\log_e$ poorly conditioned?

3. [5 marks]

Assume

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

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]

Recall that, for any real number $p \geq 1$, the p-norm of a vector $x \in \mathbb{R}^n$ is

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

For any real $n \times n$ matrix A, the matrix p-norm subordinate to this vector p-norm is

$$\max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

Assuming A is nonsingular,

$$\text{cond}_p(A) = \|A\|_p \, \|A^{-1}\|_p$$

Let D be a real $n \times n$ diagonal matrix. That is,

$$D = \begin{pmatrix} d_1 & 0 & 0 & \cdots & 0 \\ 0 & d_2 & 0 & \cdots & 0 \\ 0 & 0 & d_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & d_n \end{pmatrix}$$

where $d_i \in \mathbb{R}$ for $i = 1, 2, \ldots, n$. In addition, assume that $d_i \neq 0$ for $i = 1, 2, \ldots, n$. I mentioned in class that

$$\text{cond}_p(D) = \frac{\max\{|d_i| : i = 1, 2, \ldots, n\}}{\min\{|d_i| : i = 1, 2, \ldots, n\}} \tag{1}$$

Explain why (1) is true.