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

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]

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 $d_i \in \{0, 1, 2, \ldots , 9\}$ for $i = 1, 2, 3$ and $n \in \{-10, -9, -8, \ldots , 10\}$. 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 $+\text{Infty}$ and $-\text{Infty}$ which stand for numbers that are too large in magnitude (either positive or negative, respectively) to represent in this floating-point 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 $+\text{Infty}$ or $-\text{Infty}$ in the case of overflow.

(a) $(4.53 \cdot 10^3) - (3.21 \cdot 10^3)$
(b) $(3.65 \cdot 10^3) + (4.86 \cdot 10^2)$
(c) $(1.01 \cdot 10^2) \times (4.15 \cdot 10^{-3})$
(d) $(5.21 \cdot 10^{-6}) \times (-2.02 \cdot 10^{-6})$
(e) $(-4.13 \cdot 10^4) \times (-3.03 \cdot 10^6)$
2. [10 marks: 5 marks for each part]

Jim wrote the MatLab function

```matlab
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

\[
\begin{align*}
    r_1 & = 1.0000 \times 10^{+7} \\
    r_2 & = 9.9652 \times 10^{-8}
\end{align*}
\]

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
\( a \cdot r_1 \cdot r_2 = c \), but his computed roots satisfied \( a \cdot r_1 \cdot r_2 = 0.99652 \), while \( c = 1 \). So he knew that at least one of the two roots he calculated must be inaccurate.

Jim checked his function carefully, but he couldn’t find anything wrong with it.

(a) Why did Jim’s function compute such an inaccurate result?

[Note: although rounding error should play a role in your answer, 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) Advise Jim 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.
3. [6 marks]

Let

\[ A = \begin{pmatrix} 1 & -2 \\ -3 & 4 \end{pmatrix} \]

Calculate \( \|A\|_1 \), \( \|A\|_2 \) and \( \|A\|_\infty \).

Show all your calculations.

For the 2-norm, there will be some square roots in your answer. Just leave them as square roots; you don’t need to evaluate them.

4. [5 marks]

Let \( Ax = b \) and \( A\hat{x} = \hat{b} \), where \( A \) is a real \( n \times n \) nonsingular matrix and \( x, \hat{x}, b \) and \( \hat{b} \) are all real vectors of length \( n \) (i.e., \( x, \hat{x}, b \) and \( \hat{b} \) are all vectors in \( \mathbb{R}^n \)). I showed in class that

\[ \frac{\|x - \hat{x}\|}{\|x\|} \leq \text{cond}(A) \frac{\|b - \hat{b}\|}{\|b\|} \] (1)

where

\[ \text{cond}(A) = \|A\| \cdot \|A^{-1}\| \] (2)

and the matrix norm in (2) is \textit{subordinate to} the vector norm in (1).

I also mentioned in class that

\[ \frac{1}{\text{cond}(A)} \frac{\|b - \hat{b}\|}{\|b\|} \leq \frac{\|x - \hat{x}\|}{\|x\|} \] (3)

Show that (3) is true.