

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 expressions? 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) $6.01 \cdot 10^3 + 5.20 \cdot 10^3$
- (b) $3.48 \cdot 10^3 - 1.21 \cdot 10^1$
- (c) $1.01 \cdot 10^3 \times 5.02 \cdot 10^{-4}$
- (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. [10 marks: 5 marks for each part]

Jim 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

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

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 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. [5 marks]

Consider the function

$$f(x) = 1 + (\sin(x))^2$$

Is this function well-conditioned for all $x \in [-100, 100]$ or are there some values of $x \in [-100, 100]$ for which $f(x)$ is ill-conditioned? By well-conditioned here, we mean that a small relative change in x produces a small relative change in $f(x)$.

Justify your answer.

4. [5 marks]

Are there vectors x and $y \in \mathbb{R}^2$ for which $\|x\|_1 < \|y\|_1$ and $\|x\|_\infty > \|y\|_\infty$?

Either give an example of an x and $y \in \mathbb{R}^2$ that satisfies these two inequalities or explain why no such x and y can exist.

5. [5 marks]

I mentioned in class that $\|x\|_\infty \leq \|x\|_1$ for all vectors $x \in \mathbb{R}^n$.

Does a similar result hold for matrices?

That is, does $\|A\|_\infty \leq \|A\|_1$ hold for all matrices $A \in \mathbb{R}^{n \times n}$?

Justify your answer.

6. [5 marks]

We showed in class that, if $Ax = b$ and $A\hat{x} = \hat{b}$, where x, b, \hat{x} and $\hat{b} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ is nonsingular matrix, then

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \text{cond}(A) \frac{\|b - \hat{b}\|}{\|b\|}$$

I mentioned in class that you could also show that

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

Show that (1) holds.