Question 1

[10 marks]

Let $x,\ y \in \mathcal{R}$. Recall that $fl(x), fl(y) \in \mathcal{R}_b(t,s)$ denote the floating-point representations of x and y, respectively, where $fl(x) = x(1-\delta_x)$, $fl(y) = y(1-\delta_y)$, and $\delta_x,\ \delta_y$ quantify the relative roundoff errors in the respective representations.

In lecture, we showed that a typical computer estimates the product of x and y as

$$fl(fl(x)\cdot fl(y)) = (x\cdot y)(1-\delta_.)$$

where $|\delta_.| \le 3$ eps. Using similar techniques, derive a tight error bound for computer division.

$$fl\left(\frac{fl(x)}{fl(y)}\right) = \frac{x\cdot(1-\delta_x)}{y\cdot(1-\delta_y)}\cdot(1-\delta_{x/y}) = \frac{x}{y}\cdot\frac{(1-\delta_x)(1-\delta_{x/y})}{1-\delta_y} =$$

$$= \frac{x}{y}\cdot\frac{1-\delta_x-\delta_{x/y}+\delta_x\cdot\delta_{x/y}}{1-\delta_y} =$$

$$= \left[\text{since } \delta < \text{EPS} \Rightarrow \delta_x\cdot\delta_{x/y} - \text{insignificant, also } \cancel{1-\delta_y = 1}\right] =$$

No!

$$= \frac{x}{y}\cdot(1-\delta_x-\delta_{x/y}) = \frac{x}{y}\cdot(1-\delta_{division}),\ \text{where } \delta_{division} = \delta_x+\delta_{x/y}$$

bound?

3

Question 2

[10 marks]

Let $\hat{x}$ be a computed solution to $Ax = b,\ A \in \mathcal{R}^{n\times n}$. The following bound for the relative error in $\hat{x}$ was derived in class:

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

where $r = b - A\hat{x}$. Starting with the equations $Ax = b$ and $A\hat{x} = b - r$, derive a *lower* bound for $\|x - \hat{x}\|/\|x\|$. What do these bounds tell us about the reliability of $\hat{x}$?

$$\begin{cases} A\underline{x} = b \\ A\underline{\hat{x}} = \underline{b} - \underline{r} \end{cases} \Rightarrow A(\underline{x} - \underline{\hat{x}}) = \underline{r} \Leftrightarrow A^{-1} \cdot A(\underline{x} - \underline{\hat{x}}) = A^{-1} \cdot \underline{r} \Leftrightarrow$$

$$\Leftrightarrow \text{since } A\underline{x} = b \Leftrightarrow \frac{\|A^{-1}A \cdot (\underline{x} - \underline{\hat{x}})\|}{\|A \cdot \underline{x}\|} = \frac{\|A^{-1} \cdot \underline{r}\|}{\|\underline{b}\|} \Leftrightarrow$$

$$\Leftrightarrow \text{since } \|A^{-1}\| = \frac{1}{\|A\|} \Rightarrow \frac{\|A^{-1}A(\underline{x} - \underline{\hat{x}})\|}{\|A \cdot \underline{x}\|} = \frac{\|\underline{r}\|}{\|A\,\underline{b}\|} \Leftrightarrow$$

$$\Leftrightarrow \frac{\|A^{-1}\|\|A\| \cdot \|\underline{x} - \underline{\hat{x}}\|}{\|\underline{x}\|} \geq \frac{\|\underline{r}\|}{\|\underline{b}\|} \Leftrightarrow \frac{\|\underline{x} - \underline{\hat{x}}\|}{\|\underline{x}\|} \geq \frac{\|\underline{r}\|}{\|\underline{b}\|} \cdot \frac{1}{\|A^{-1}\| \cdot \|A\|}$$

$$\Rightarrow \frac{1}{\text{cond}(A)} \cdot \frac{\|\underline{r}\|}{\|\underline{b}\|} \leq \frac{\|\underline{x} - \underline{\hat{x}}\|}{\|\underline{x}\|};$$

the reliability of $\hat{x}$ depends ~~one~~ ~~one~~ on the size of cond (A), if cond (A) is large a ~~small~~ $\hat{x}$ is not reliable; (even if the norm of the residual $\|r\|$ is small);
however if cond (A) is small (close to 1) $\hat{x}$ is a reliable solution;

8

Question 3

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 3 & 5 & 9 \\ 4 & 4 & 4 \\ 1 & 5 & 5 \end{bmatrix}, \quad b = \begin{bmatrix} 40 \\ 24 \\ 26 \end{bmatrix}.$$

a. Compute the $PA = LU$ factorization of A. Use exact arithmetic. Show all intermediate calculations, including Gauss transforms and permutation matrices.

① $P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (interchange rows 1 and 2) $\quad P_1 \cdot A = \begin{bmatrix} 4 & 4 & 4 \\ 3 & 5 & 9 \\ 1 & 5 & 5 \end{bmatrix}$

$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{3}{4} & 1 & 0 \\ -\frac{1}{4} & 0 & 1 \end{bmatrix} \qquad L_1 \cdot P_1 \cdot A = \begin{bmatrix} 4 & 4 & 4 \\ 0 & 2 & 6 \\ 0 & \cancel{1}\,4 & \cancel{1}\,4 \end{bmatrix}$

② no need to ~~compute~~ interchange rows; 2 - maximum pivot; $\to P_2 = I_3$;

$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{2} & 1 \end{bmatrix}$; $L_2 \cdot L_1 \cdot P_1 \cdot A = \begin{bmatrix} 4 & 4 & 4 \\ 0 & 2 & 6 \\ 0 & 0 & -2 \end{bmatrix}$,

$$\Rightarrow \quad L_2 \cdot L_1 \cdot P_1 \cdot A = U \iff \underbrace{P_1}_{P} \cdot A = \underbrace{L_1^{-1} \cdot L_2^{-1}}_{L} \cdot U$$

$$\Rightarrow \quad P \cdot A = LU \iff \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 3 & 5 & 9 \\ 4 & 4 & 4 \\ 1 & 5 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{3}{4} & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix} \cdot \begin{bmatrix} 4 & 4 & 4 \\ 0 & 2 & 6 \\ 0 & 0 & -2 \end{bmatrix}$$

Wrong LU, right method.

6

b. Use the factorization computed in **(a)** to solve the system.

$\underline{A} \cdot \underline{x} = \underline{b} \iff PA\underline{x} = P\underline{b} \iff L \cdot U \cdot \underline{x} = P\underline{b} \iff \begin{cases} L\cdot\underline{y} = P\underline{b} \\ U\underline{x} = \underline{y} \end{cases}$

× wrong P.

$$L \cdot \underline{y} = P \cdot \underline{b} \iff \begin{bmatrix} 1 & 0 & 0 \\ 3/4 & 1 & 0 \\ 1/4 & 1/2 & 1 \end{bmatrix} \cdot \underline{y} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 40 \\ 24 \\ 26 \end{bmatrix} \iff$$

$$\iff \begin{bmatrix} 1 & 0 & 0 \\ 3/4 & 1 & 0 \\ 1/4 & 1/2 & 1 \end{bmatrix} \cdot \underline{y} = \begin{bmatrix} 24 \\ 40 \\ 26 \end{bmatrix} \Rightarrow \underline{y} = \begin{bmatrix} 24 \\ 22 \\ 9 \end{bmatrix};$$

$$U \cdot \underline{x} = \underline{y} \iff \begin{bmatrix} 4 & 4 & 4 \\ 0 & 2 & 6 \\ 0 & 0 & -2 \end{bmatrix} \cdot \underline{x} = \begin{bmatrix} 24 \\ 22 \\ 9 \end{bmatrix} \Rightarrow \underline{x} = \begin{bmatrix} -14 \\ 49/2 \\ -9/2 \end{bmatrix};$$

✓

Right method.

4.

c. Why is Gaussian Elimination usually implemented as in this question (i.e., $PA = LU$ is computed separately, and then the factorization is used to solve $Ax = b$)?

Because it can be reused for any $\underline{b}$ in $A\underline{x} = \underline{b}$; making the calculation of solutions ~~in~~ for systems using the same matrix A, more efficient;
(because the factorization is the most expensive part of ~~the~~ finding the solution to $A\underline{x} = \underline{b}$);

✓ 2

Question 4

[5 marks]

Consider the iterative improvement algorithm discussed in tutorial:

Solve $Ax = b$ for initial approximation $\hat{x}_0$.
for $i = 0, 1, \ldots$ until convergence
  compute $r_i = b - A\hat{x}_i$
  solve $Az_i = r_i$
  update $\hat{x}_{i+1} = \hat{x}_i + z_i$
end for

After the first iteration of this algorithm,

$$\begin{aligned} \hat{x}_1 &= \hat{x}_0 + z_0 \\ &= \hat{x}_0 + A^{-1}r_0 \\ &= \hat{x}_0 + A^{-1}(b - \underbrace{A\hat{x}_0}_{=b}) \\ &= \hat{x}_0 + A^{-1}b - A^{-1}A\hat{x}_0 \\ &= \hat{x}_0 + x - \hat{x}_0 \\ &= x \end{aligned}$$

$\Rightarrow A^{-1}(b - A\hat{x}_0) = 0$;

This alone is not a fallacy.

Apparently the algorithm converges to the true solution x in just one iteration! What is the fallacy in this argument?

2

Question 5

[15 marks]

Consider the functions $f(x) = 1 - 1/(2x)$ and $g(x) = 2x(1-x)$.

a. How many roots does f have? Are the roots of f fixed-points of g? Are there more fixed points of g than roots of f? **Justify your answers.**

~~f doesn't have 2 any roots, ... f(x) ≠ 0, ∀x ∈ ℝ \ {0};~~

- $f(x) = \frac{2x-1}{2x}$ ~~=>~~, $x \in \mathbb{R} \setminus \{0\}$, $f(x) = 0$ if $x = \frac{1}{2}$;

~~g ... x ... f(x) ... 2x-1 / 2x ... 2x(x-1)~~

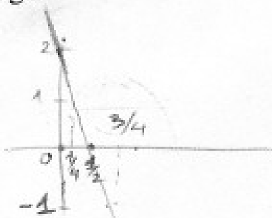
- $g(\frac{1}{2}) = 2 \cdot \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{2}$ => for $x = \frac{1}{2}$ $g(x) = x$ => the root of $f(x)$ ~~are~~ is a fixed point of g;

- yes, ex: $g(0) = 0$;

5

b. Using an appropriate theorem proven in lecture, determine the region of local convergence of the fixed-point iteration $x_{k+1} = g(x_k)$, $k = 0, 1, \ldots$, with $g(x)$ as defined above. In other words, find the interval on the x-axis for which the iteration is *guaranteed* to converge.

using the Fixed point theorem:



$g'(x) = -4x + 2$; $|g'(x)| < 1$, $\forall x \in (\frac{1}{4}, \frac{3}{4})$;

$g(x) \in (\frac{1}{4}, \frac{3}{4})$, $\forall x \in (\frac{1}{4}, \frac{3}{4})$;

the fixed-point iteration will converge on $(\frac{1}{4}, \frac{3}{4})$;

10