Question 1 (10)

[10 marks]

Recall that a floating-point operation, or *flop*, is an operation of the form $mx + b$. Show how to convert a $(k+1)$-digit base b ($b \neq 10$) positive integer

$$d_k\, d_{k-1} \ldots d_1\, d_0$$

into its base 10 equivalent in k flops or less.

convert into base 10 $(d_k d_{k-1} \ldots d_1 d_0)_b = (d_k b^k + d_{k-1} b^{k-1} \cdots + d_1 b^1 + d_0 b$

(4)

1 flop $= (d_k b^{k-1} + d_{k-1} b^{k-2} \cdots + d_1) b + d_0 b^0$

$+ (d_k b^{k-2} + d_{k-1} b^{k-3} \cdots) + d_1 b$

$+ d_k b^{k-k-2} + d_{k-1} b$

(6)

k flops $+ d_k b$

Question 2 (14.5)

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 5 & 10 \\ 8 & 32 & 8 \\ 1 & 8 & 13 \end{bmatrix}, \quad b = \begin{bmatrix} 7 \\ -16 \\ 6 \end{bmatrix}.$$

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

Swap row 1 and 2

$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ $P_1 A = \begin{bmatrix} 8 & 32 & 8 \\ 2 & 5 & 10 \\ 1 & 8 & 13 \end{bmatrix}$ $L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{2}{8} & 1 & 0 \\ -\frac{1}{8} & 0 & 1 \end{bmatrix}$

$L_1 P_1 A = \begin{bmatrix} 8 & 32 & 8 \\ 0 & -3 & 8 \\ 0 & 4 & 12 \end{bmatrix}$ swap row 2 and 3 $P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ (9)

$P_2 L_1 P_1 A = \begin{bmatrix} 8 & 32 & 8 \\ 0 & 4 & 12 \\ 0 & -3 & 8 \end{bmatrix}$ $L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{3}{4} & 1 \end{bmatrix}$ $L_2 P_2 L_1 P_1 A = \begin{bmatrix} 8 & 32 & 8 \\ 0 & 4 & 12 \\ 0 & 0 & 17 \end{bmatrix} = U$ ✓

$L_2 P_2 L_1 P_1 A = U$

$L_2 P_2 L_1 P_2 P_2 P_1 A = U$

$L_2 \tilde{L}_1 P_2 P_1 A = U$

$P_2 L_1 P_2 = \tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{8} & 1 & 0 \\ -\frac{2}{8} & 0 & 1 \end{bmatrix}$

$\underbrace{P_2 P_1}_{P} A = \underbrace{\tilde{L}_1^{-1} L_2^{-1}}_{L} U$

$\tilde{L}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{8} & 1 & 0 \\ \frac{2}{8} & 0 & 1 \end{bmatrix}$ $L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{3}{4} & 1 \end{bmatrix}$

$P = P_2 P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$ ✓ $L = \tilde{L}_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{8} & 1 & 0 \\ \frac{2}{8} & -\frac{3}{4} & 1 \end{bmatrix}$ ✓

CONTINUED ...

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

$Ax = b$

$PAx = Pb$

$LUx = Pb$

$Ld = Pb$

$$Pb = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 7 \\ -16 \\ 6 \end{bmatrix} = \begin{bmatrix} -16 \\ 6 \\ 7 \end{bmatrix}$$

$Ld = Pb$

$$\begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{8} & 1 & 0 \\ \frac{2}{8} & -\frac{2}{4} & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 6 \\ 7 \end{bmatrix} \Rightarrow \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 8 \\ -3 \end{bmatrix}$$

$Ux = d$

$$\begin{bmatrix} 8 & 32 & 8 \\ 0 & 4 & 12 \\ 0 & 0 & 17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -16 \\ 8 \\ -3 \\ 17 \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \checkmark \\ -\frac{3}{17} \end{bmatrix}$$

✓

$4x_2 - \frac{36}{17} x_3 = 8$

-0.5

$4x_2$

3.5

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 $Ax = b$)?

Since PA=LU factorization makes the qestion easier to solve as we need to get solutions using upper and lower triangular matrices which is easy. Also, Gaussian Elimination with partial pivoting makes the error $\|E\| \leq R \cdot eps \cdot \|A\|$ which is very small. The LU factorization gives us the ability to further solve $Ay = c$, $Az = d$ **CONTINUED...** with most of the work already done.

2

Question 3 (10)

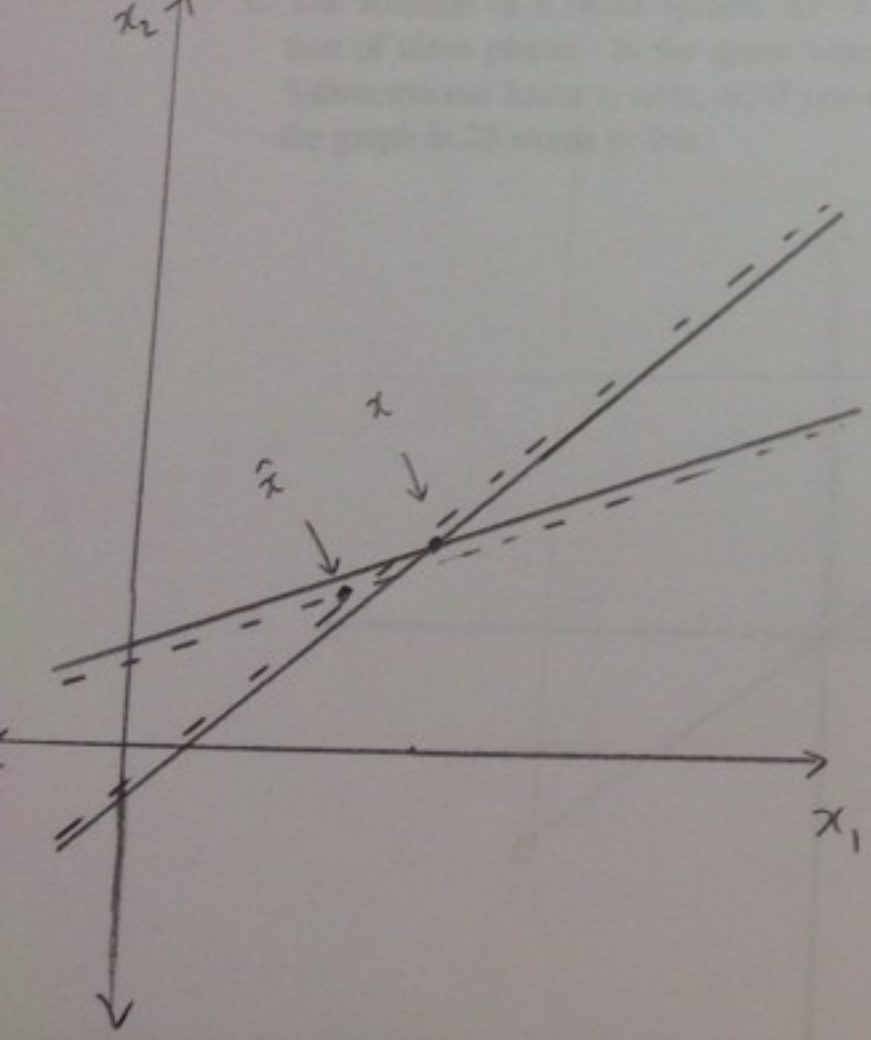
[10 marks]

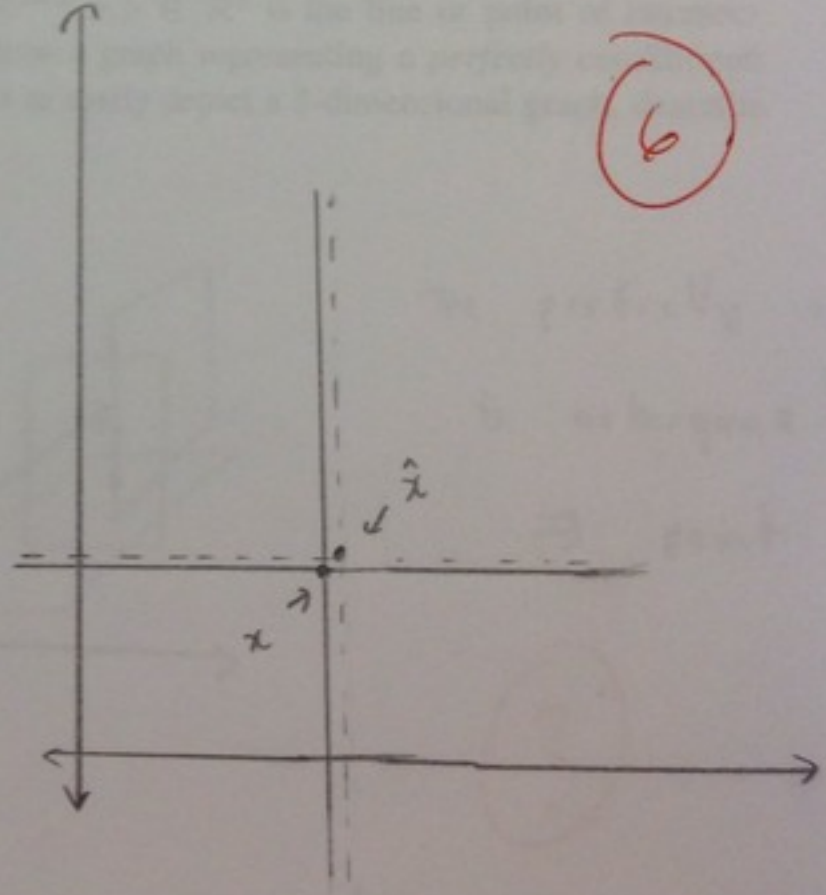
Recall in lecture we discussed the geometric interpretation of the manifestation of round-off error during the Gaussian Elimination/LU factorization process. We drew two graphs depicting the intersection of lines which represented, respectively, the solution of a poorly conditioned and a perfectly conditioned linear system $Ax = b, A \in \mathcal{R}^{2\times2}, x, b \in \mathcal{R}^2$.

a. Reproduce the graphs below. As in lecture, draw the true systems with solid lines and the systems resulting from roundoff error with dashed lines. Clearly label the true solution and the approximate solutions on each graph.

$Ax = b$ is true solution

$A(\hat{x} + E) = b$ is solution with round off error

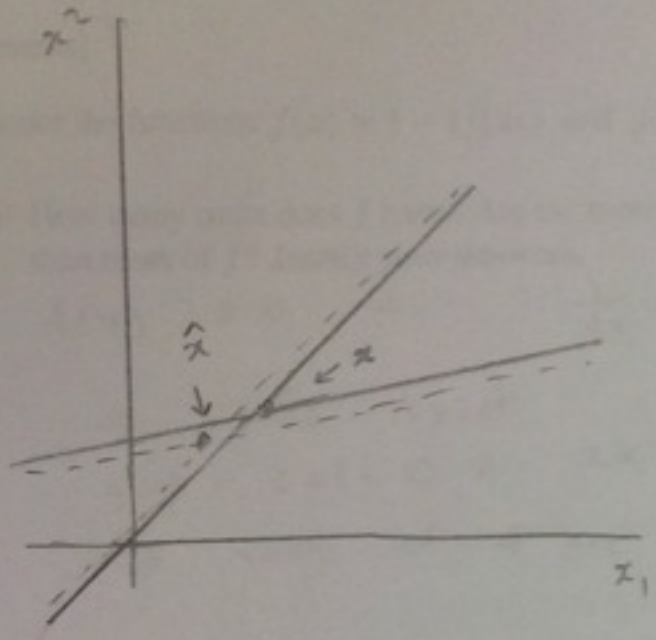




(6)

CONTINUED ...

b. Copy the graph representing the poorly conditioned system to the space below. Show how the residual vector $r = b - A\hat{x}$ manifests on the graph. (**Note:** This was not discussed in lecture.)



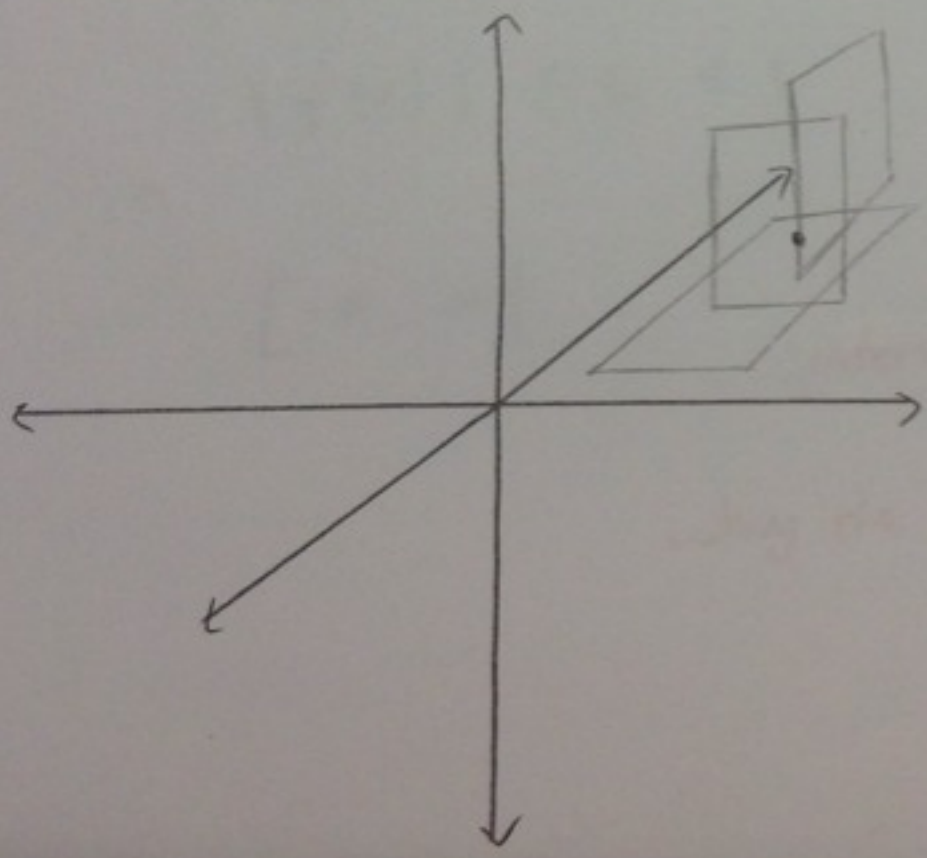
$r_1 = b - A\hat{x}_1$

$r_2 = b - A\hat{x}_2$

⇒ Error contained in both r_1 and r_2

1

c. The solution of a linear system $Ax = b, A \in \mathcal{R}^{3\times3}, x, b \in \mathcal{R}^3$ is the line or point of intersection of three planes. In the space below, either draw a graph representing a *perfectly* conditioned 3-dimensional linear system, or, if you are not able to easily depict a 3-dimensional graph, describe the graph in 25 words or less.



the perfectly conditioned is orthogonal planes ⇒ point solution

3

CONTINUED ...

Question 4

(5)

[15 marks]

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

$f(x) = \frac{2x}{2x} - \frac{1}{2x} = \frac{2x-1}{2x}$.

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.**

(3)

$f(x) = 0 \Rightarrow 1 - \frac{1}{2x} = 0 \Rightarrow 1 = \frac{1}{2x} \Rightarrow 2x = 1 \Rightarrow x = \frac{1}{2}$

$\Rightarrow$ f has 1 root ✓

$g(x) = 2x(1-x) \Rightarrow 2x - 2x^2 \Rightarrow x + x - 2x^2 \Rightarrow x + x(1-2x)$

$\Rightarrow$ roots of f are fixed points of g because $g(x) = x$

are ther more fixed points? (-2)

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.

$[a,b]$ s.t $g(x) \in [a,b]$ $x \in [a,b]$

(2)

$|g'(x)| \leq L < 1$

$[-\infty, \infty]$

interval? (-4)

why the interval is a sink? (-4)

END OF EXAM