University of Toronto at Scarborough
**CSCC37—Numerical Algorithms for Computational Mathematics, Fall 2020**

# Term Test

**Duration:** 120 minutes (estimated)

**Date and Time:** Saturday 14 November, 4:00–8:00 p.m.

**Aids allowed:** Open-book. All aids are allowed.

This is a take-home test. Complete your solutions, and submit no later than the scheduled finish time for the test, **following the instructions given on the term test page of the course website.** (Note that you are given more than the estimated time to complete the test.)

This test consists of 7 questions. **Make sure your copy has 13 pages (including this one).** Write your answers in the spaces provided. You will be rewarded for concise, well-thought-out answers, rather than long rambling ones. Please write legibly.

Take a few minutes before you begin the test to read through each question, and then start with the question(s) you find easiest.

Name: Jingyun Long
(Circle your family name.)

UTORid: longjin2

Student #: 1005543114

Tutorial section: TUT04

**YOU MUST SIGN THE FOLLOWING:**

I declare that this test was written by the person whose name and student # appear above.

Signature: Jingyun

**Your grade**

| | | | |
|---|---|---|---|
| *1.* | / 10 | *5.* | / 10 |
| *2.* | / 5 | *6.* | / 10 |
| *3.* | / 15 | *7.* | / 15 |
| *4.* | / 15 | | |
| | | **Total** | / 80 |

Graded by

## Question 1

[10 marks]

Design a computer that is able to store $(0.1)_{10}$ exactly. A floating point number on your computer must be represented internally in a base less than 10, and must have a mantissa with a finite number of digits. (**Hint:** Is this possible?)

If we allow fractional bases, then we can simply do $R_{0.1}(1,1)$, so $(0.1)_{10} = (1)_{0.1}$.

If we can only have integer bases, so $b \in \{1, \dots, 9\}$, then it is not possible since digits are also integers.

$$0.1 = \frac{1}{10} = \underbrace{\frac{1}{2} \cdot \frac{1}{5}}_{\substack{\text{base 5 with}\\ \text{digit 0.5:}\\ \text{can't do}}} = \underbrace{2 \cdot \frac{1}{20}}_{\substack{\text{base 20}\\ \text{with digit 2:}\\ \text{possible but}\\ 1 \leq b \leq 9}} = 3 \cdot \frac{1}{30} = \dots$$

(or base 2 with digit 0.2)

only way to represent $(0.1)_{10}$ exactly and finitely in a base $< 10$

It is possible however with bases $> 10$, e.g. $b = 20$, $b = 30$

## Question 2

[5 marks]

What numbers are representable with a finite expression in the binary system but are not finitely representable in the decimal system? **Justify your answer.**

None, since any finite binary number can be written as a finite decimal number.

If we have some finite binary # $x = (d_n \ldots d_0 . d_{-1} \ldots d_{-m})_2$

we can write in decimal as

$$\sum_{i=0}^{n} \underbrace{d_i}_{\in\{0,1\}} \cdot \underbrace{2^i}_{\text{"+"-ve power of 2, is finite}} + \sum_{i=1}^{m} \underbrace{d_{-i}}_{\in\{0,1\}} \cdot \underbrace{2^{-i}}_{\text{"-"-ve power of 2, is finite since}}$$

$$\frac{1}{2^i} = \underbrace{5^i}_{\text{finite}} \cdot \frac{1}{10^i}$$

since any finite binary number can be written as a sum of finite terms, then they are all finitely representable in decimal.

## Question 3

[15 marks]

We wish to evaluate $f(x) = \beta - \sqrt{\beta^2 - x^2}$ where $\beta \gg 0$ and $|x| < \beta$.

**a.** Identify the range of $x$ where $f(x)$ suffers from potential subtractive cancellation. An approximate range will suffice.

~~$x \approx 0 \Rightarrow x \in [-\epsilon, \epsilon]$ since $\sqrt{\beta^2 - x^2} \approx \beta$~~ $\beta - \beta = 0$

$(|x| \approx \beta \Rightarrow x \in (-\beta, -\beta+\epsilon] \cup [\beta-\epsilon, \beta)$ since $x^2 \approx \beta^2)$

**b.** Evaluate $\lim_{x\to\alpha} \text{cond}(f(x))$, where $\alpha$ is the smallest number, in absolute value, within the range you identified in (**a**). Does the condition number reflect the potential for subtractive cancellation? **Explain.**

$$f'(x) = \frac{-1}{2\sqrt{\beta^2 - x^2}}(-2x) = \frac{x}{\sqrt{\beta^2 - x^2}}$$

$$\lim_{x\to\beta} \text{cond}(f(x)) = \lim_{x\to\beta}\left(\left|\frac{x}{\sqrt{\beta^2-x^2}}\cdot x\right| \cdot \left|\frac{1}{\beta - \sqrt{\beta^2-x^2}}\right|\right)$$

$$= \lim_{x\to\beta}\left(\left|\frac{x^2}{\sqrt{\beta^2-x^2}(\beta - \sqrt{\beta^2-x^2})}\right|\right)$$

$$= \frac{\beta^2}{0(\beta - 0)} = \infty$$

Yes, cond(f(x)) blows up when x is near $\beta$ which reflects the potential for subtractive cancellation.

**c.** Derive an alternate form of $f(x)$, stable for evaluation in the range you identified in (**a**).

$$f(x) = \beta - \sqrt{\beta^2 - x^2}, \quad \beta >> 0, \quad |x| < \beta$$

$$= \beta - \sqrt{x^2\left(\frac{\beta^2}{x^2} - 1\right)}$$

$$= \beta - |x|\sqrt{\frac{\beta^2}{x^2} - 1} \qquad |x| \approx \beta$$

or

$$= \beta - \sqrt{(\beta - x)(\beta + x)}$$

## Question 4

[15 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 6 & 6 \\ 3 & 5 & 12 \\ 6 & 6 & 12 \end{bmatrix}, \quad b = \begin{bmatrix} 20 \\ 25 \\ 30 \end{bmatrix}.$$

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

$$P_{13}A = \begin{bmatrix} 6 & 6 & 12 \\ 3 & 5 & 12 \\ 2 & 6 & 6 \end{bmatrix}, \quad L_1 = \begin{bmatrix} 1 & 0 & 0 \\ \frac{-1}{2} & 1 & 0 \\ \frac{-1}{3} & 0 & 1 \end{bmatrix}, \quad L_1 P_{13} A = \begin{bmatrix} 6 & 6 & 12 \\ 0 & 2 & 6 \\ 0 & 4 & 2 \end{bmatrix}$$

$$P_{23} L_1 P_{13} A = \begin{bmatrix} 6 & 6 & 12 \\ 0 & 4 & 2 \\ 0 & 2 & 6 \end{bmatrix}, \quad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{-1}{2} & 1 \end{bmatrix}, \quad L_2 P_{23} L_1 P_{13} A = \begin{bmatrix} 6 & 6 & 12 \\ 0 & 4 & 2 \\ 0 & 0 & 5 \end{bmatrix} = U$$

$$L_2 P_{23} L_1 P_{13} A = L_2 \underbrace{P_{23} L_1 P_{23}} P_{23} P_{13} A = L_2 \tilde{L}_1 \underbrace{P_{23} P_{13}}_{P} A = U \Rightarrow PA = \underbrace{\tilde{L}_1^{-1} L_2^{-1}}_{L} U$$

$$\tilde{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ \frac{-1}{3} & 1 & 0 \\ \frac{-1}{2} & 0 & 1 \end{bmatrix} \qquad P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$L = \tilde{L}_1^{-1} L_2^{-1}$ ← sum of negated multipliers

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

$$\Rightarrow L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{3} & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 6 & 6 & 12 \\ 0 & 4 & 2 \\ 0 & 0 & 5 \end{bmatrix}, \quad P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

CONTINUED ...

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

$A\vec{x} = \vec{b}$

$\Rightarrow PA\vec{x} = P\vec{b}$

$\Rightarrow LU\vec{x} = P\vec{b}$

Let $\vec{d} = U\vec{x}$

$\Rightarrow L\vec{d} = P\vec{b}$

$P\vec{b} = P_{23}P_{13}\begin{bmatrix}20\\25\\30\end{bmatrix} = P_{23}\begin{bmatrix}30\\25\\20\end{bmatrix} = \begin{bmatrix}30\\20\\25\end{bmatrix}$

$\Rightarrow \begin{bmatrix}1 & 0 & 0\\ \frac{1}{3} & 1 & 0\\ \frac{1}{2} & \frac{1}{2} & 1\end{bmatrix}\begin{bmatrix}d_1\\d_2\\d_3\end{bmatrix} = \begin{bmatrix}30\\20\\25\end{bmatrix}$

$d_1 = 30$

$10 + d_2 = 20 \Rightarrow d_2 = 10$

$15 + 5 + d_3 = 25 \Rightarrow d_3 = 5$

$\vec{d} = \begin{bmatrix}30\\10\\5\end{bmatrix}$

$U\vec{x} = \vec{d}$

$\Rightarrow \begin{bmatrix}6 & 6 & 12\\ 0 & 4 & 2\\ 0 & 0 & 5\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix} = \begin{bmatrix}30\\10\\5\end{bmatrix}$

$x_3 = 1$

$4x_2 + 2 = 10 \Rightarrow x_2 = 2$

$6x_1 + 12 + 12 = 30 \Rightarrow x_1 = 1$

$\vec{x} = \begin{bmatrix}1\\2\\1\end{bmatrix}$

**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 $Ax = b$)?

Computing $PA = LU$ first is more stable as able to do pivoting to avoid large multipliers.

After doing costly computation of $LU$, can easily solve $A\vec{x} = \vec{b}$ for multiple $\vec{b}$.

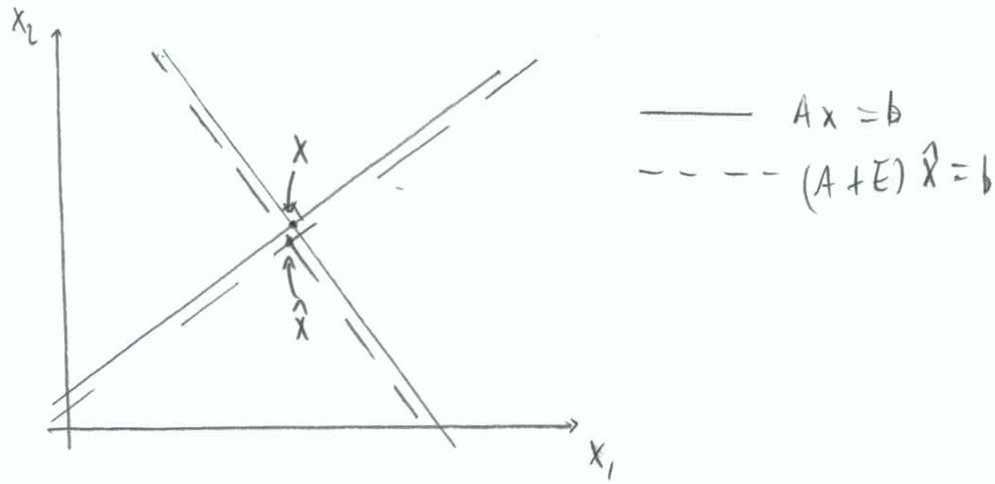
**CONTINUED ...**

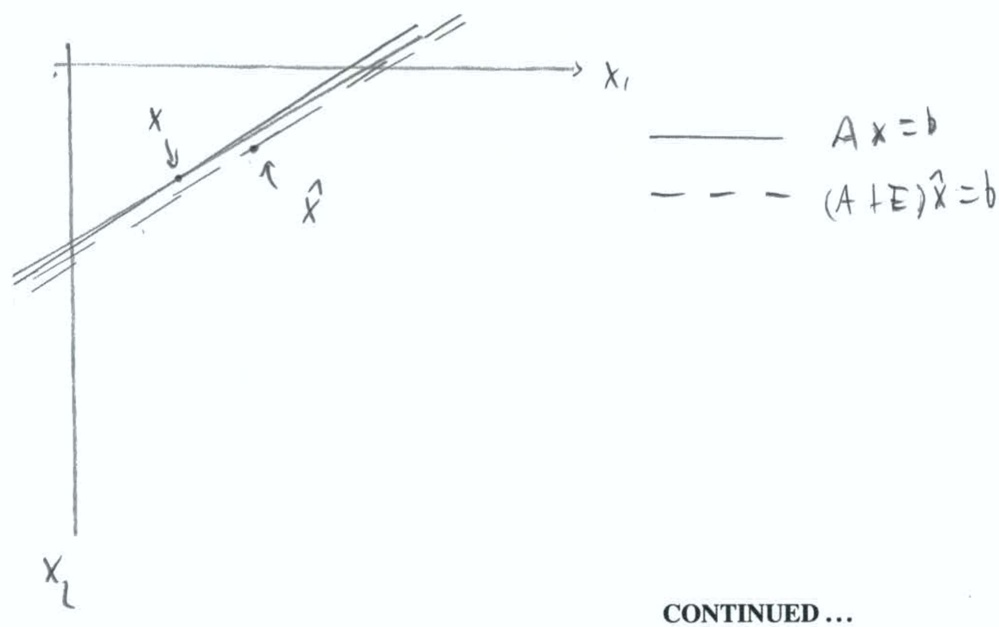
## Question 5

[10 marks]

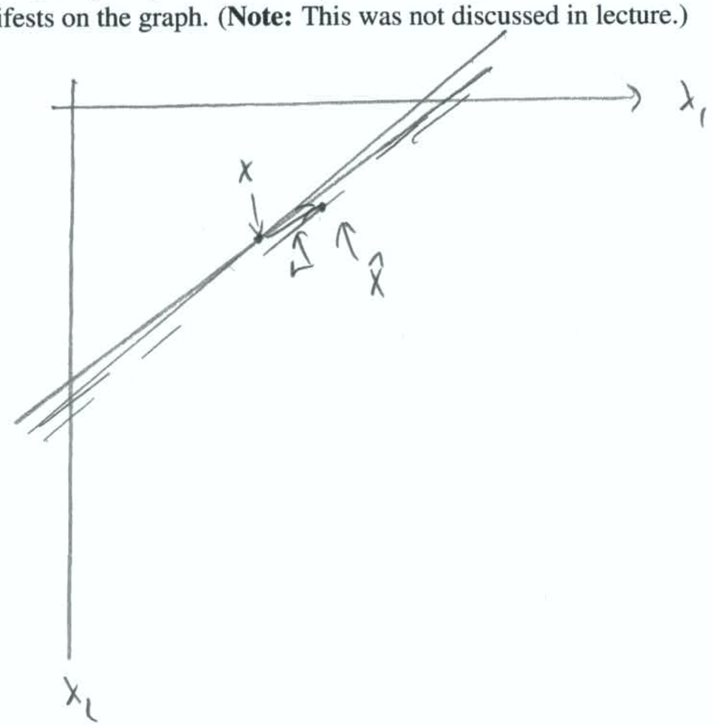
Recall in lecture we discussed the geometric interpretation of the manifestation of round-off error during the Gaussian Elimination/LU factorization process. We drew two graphs depicting the intersection of lines which represented, respectively, the solution of a poorly conditioned and a perfectly conditioned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

**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.

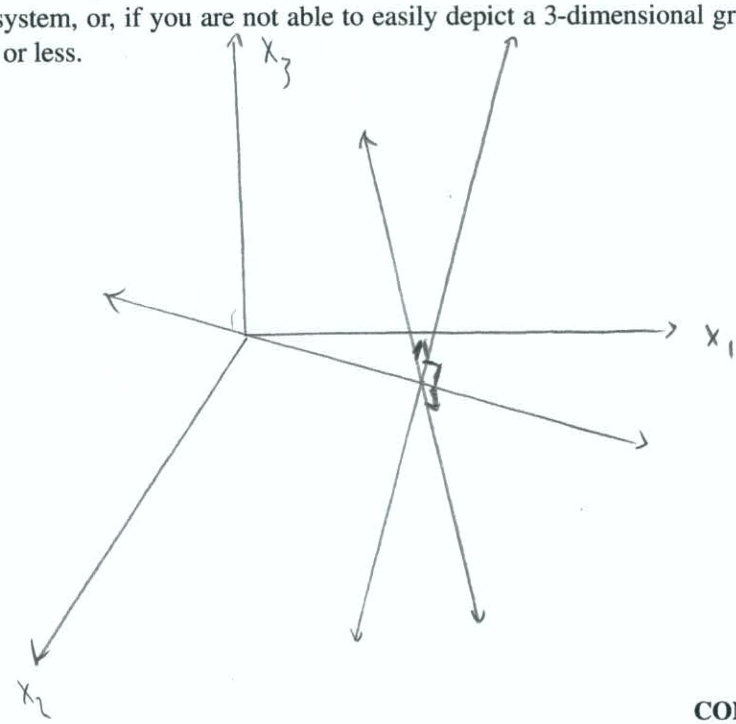


**b.** Copy the graph representing the poorly conditioned system to the space below. Show how the residual vector $r = b - A\hat{x}$ manifests on the graph. (**Note:** This was not discussed in lecture.)




**c.** The solution of a linear system $Ax = b, A \in \mathcal{R}^{3\times 3}, x, b \in \mathcal{R}^3$ 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.



# Question 6

[10 marks]

Consider the linear system $Ax = b$ where

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2+\epsilon \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \tag{1}$$

and $0 \le \epsilon < 1$.

**a.** Derive a formula for $\text{cond}_1(A)$, the 1-norm condition number of $A$. What is $\lim_{\epsilon \to 0} \text{cond}_1(A)$?

$\|A\|_1$ = max absolute column sum

$= \max(3, 6+\epsilon) = 6$, b/c $\epsilon \geq 0$

$$A^{-1} = \frac{1}{\underbrace{2(2+\epsilon) - 4(1)}_{= 4+\epsilon-4 = \epsilon}} \begin{bmatrix} 2+\epsilon & -1 \\ -4 & 2 \end{bmatrix} = \begin{bmatrix} \frac{2}{\epsilon}+1 & -\frac{1}{\epsilon} \\ \frac{-4}{\epsilon} & \frac{2}{\epsilon} \end{bmatrix}$$

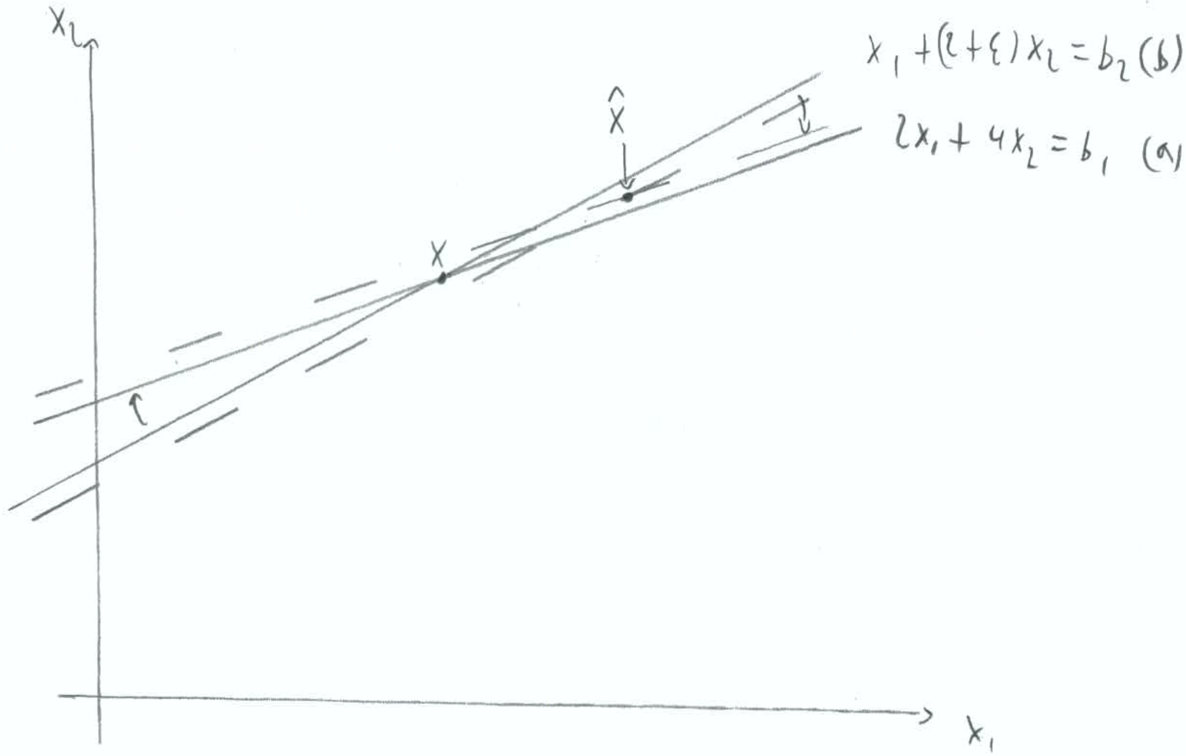
$$\|A^{-1}\|_1 = \max\left(\frac{2}{\epsilon} + 1 - \frac{4}{\epsilon}, \frac{-1}{\epsilon} + \frac{2}{\epsilon}\right)$$

$$= \max\Big(\underbrace{1 - \frac{2}{\epsilon}}_{\in(-\infty,-1)}, \underbrace{\frac{1}{\epsilon}}_{\in(1,\infty)}\Big) = \frac{1}{\epsilon}$$

$$\text{cond}_1(A) = \|A\|_1 \|A^{-1}\|_1 = 6 \cdot \frac{1}{\epsilon} = \frac{6}{\epsilon}$$

$$\lim_{\epsilon \to 0} \text{cond}_1(A) = \lim_{\epsilon \to 0} \frac{6}{\epsilon} = \infty$$

**b.** Sketch a graph illustrating the general trend of (1) as $\epsilon \to 0$. (Since you are not given specific values for the right-hand side $b$, you cannot pin down exact $x$ and $y$ intercepts.) Also show on the graph the potential effect(s) of small perturbations in the coefficients of $A$, such as those introduced when (1) is solved on a computer using Gaussian Elimination with partial pivoting.




Slope of (b) approaches (a) as $\epsilon \to 0$ until overlap ($\infty$ soln) or parallel, non-overlapping (no soln)

Since lines are nearly parallel, small perturbations greatly affect result.

## Question 7

[15 marks]

Consider the iterative improvement algorithm discussed in lecture:

> 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

We gave an intuitive explanation of why this algorithm could improve the initial approximate solution $\hat{x}_0$, but we were vague on the conditions required for convergence. In this question, you will attempt to derive the precise conditions required for convergence.

Starting with $Az_i = r_i$ and $(A+E)\hat{z}_i = r_i$, derive a formula showing how the absolute error in the $(i+1)$-st iterate $\|\hat{x}_{i+1} - x\|$ is bound by a multiple of the absolute error in the $i$-th iterate $\|\hat{x}_i - x\|$. Argue that the magnitude of this multiple is dependent on the condition of $A$, and is less than one (hence convergence) if $A$ is well-conditioned.

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

$$\|\hat{x}_{i+1} - x\| = \|\hat{x}_i + z_i - x\| \leq \|\hat{x}_i - x\| + \|z_i\|$$
$$= \|\hat{x}_i - x\| + \|A^{-1} r_i\|$$

[Continue your answer on the next page if necessary ...]