2.3 Sources of Error

Two major sources of error with numerical Gaussian elimination:

1 Ill-conditioning & 2 "Swamping".

In order to develop the notion of "error" in relation to Gaussian elimination we must first discuss "Norms".

Norms

Recall from linear algebra the notion of a vector norm.

For $\vec{v} \in \mathbb{R}^2$, $\vec{v} = \langle a, b \rangle$ we define

$\|\vec{v}\| = \sqrt{a^2+b^2}$; This called the Euclidean or 2-Norm of a vector. ($\|\vec{v}\|_2$ denoted)

Depending on the application/use, there are a variety of types of vector norms (we will see, by analogy, matrix norms shortly).

In general a vector norm satisfies 3 properties:

1. $\|\vec{x}\| \geq 0$ with equality iff $\vec{x} = \vec{0}$.
2. $\forall\, \alpha \in \mathbb{R}$ & $\vec{v} \in \mathbb{R}^n$, $\|\alpha \vec{v}\| = |\alpha| \cdot \|\vec{v}\|$
3. $\forall\, \vec{u}, \vec{v} \in \mathbb{R}^n$, $\|\vec{u} + \vec{v}\| \leq \|\vec{u}\| + \|\vec{v}\|$ (Triangle inequality)

We now define the $\boxed{\infty\text{-Norm}}$:

def. $\boxed{\|\vec{x}\|_\infty = \operatorname{Max} |x_i|}$ where $1 \leq i \leq n$

i.e. $\|\vec{x}\|_\infty$ is the maximum of the abs. value of the components of $\vec{x}$.

Def. The definition of Backward Error (BE) & Forward Error (FE) are defined for systems of linear equations analogously with Ch.1.

Let $\vec{x}_a$ be an approximate solution of the linear system: $A\vec{x} = \vec{b}$. The residual is the vector: $\boxed{\vec{r} = \vec{b} - A\vec{x}_a}$.

The $\boxed{\text{Backward Error} = \|\vec{b} - A\vec{x}_a\|_\infty}$ &

the $\boxed{\text{Forward Error} = \|\vec{x} - \vec{x}_a\|_\infty}$

Ex. Find BE & FE for the approximate solution: $\vec{x}_a = \langle 1, 1 \rangle$ of the following system:

$$\begin{bmatrix} 1 & 1 \\ 3 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

The correct solution is: $\vec{x} = \langle 2, 1 \rangle$.

$$\underbrace{\|\vec{b} - A\vec{x}_a\|_\infty}_{BE} = \left\| \begin{bmatrix} 3 \\ 2 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 3 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|_\infty = \left\| \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right\|_\infty = 3.$$

$$FE = \|\vec{x} - \vec{x}_a\|_\infty = \left\| \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\|_\infty = \left\| \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|_\infty = 1.$$

Note that FE & BE can be <u>different orders of magnitude</u>.

Ex. Let $\vec{x}_a = \langle -1, 3.0001 \rangle$ for the system:

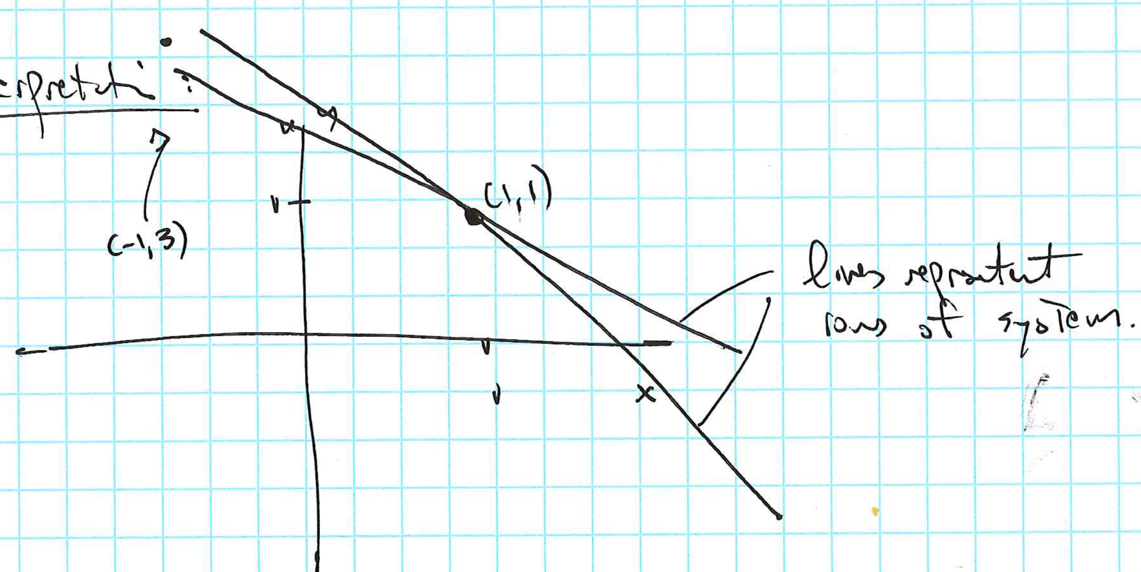
$$x_1 + x_2 = 2$$
$$1.001x_1 + x_2 = 2.001$$

Using Gaussian elimination, we find: $\vec{x} = \langle 1, 1 \rangle$.

$$\underbrace{\|\vec{b} - A\vec{x}_a\|_\infty}_{BE} = \left\| \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1.0001 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3.0001 \end{bmatrix} \right\|_\infty = \underline{.0001}$$

$$FE = \|\vec{x} - \vec{x}_a\|_\infty = \left\| \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} -1 \\ 3.0001 \end{bmatrix} \right\|_\infty = \left\| \begin{bmatrix} 2 \\ -2.0001 \end{bmatrix} \right\|_\infty = \underline{2.0001}$$

<u>Geometric Interpretation</u>:



<u>Note</u>: even though (-1,3) <u>almost</u> represents a point of intersection with the two lines, it is, nevertheless, far from the true solution: (1,1).

def. The <u>relative backward error</u> of $A\vec{x} = \vec{b}$ is:

$$\frac{\|\vec{r}\|_\infty}{\|\vec{b}\|_\infty}$$

The <u>relative forward error</u> is: $\dfrac{\|\vec{x} - \vec{x}_a\|_\infty}{\|\vec{x}\|_\infty}$.

def. The error magnification factor for $A\vec{x}=\vec{b}$

is the ratio of the two:

$$\text{EMF} = \frac{\text{Rel. FE}}{\text{Rel. BE}} = \frac{\dfrac{\|\vec{x}-\vec{x}_a\|_\infty}{\|\vec{x}\|_\infty}}{\dfrac{\|\vec{r}\|_\infty}{\|\vec{b}\|_\infty}}$$

For the previous example, then, the rel. BE $= \dfrac{.0001}{2.0001} = .005\%$,

& the rel. forward error is: $\dfrac{2.0001}{1} \doteq 200\%$

so $\text{EMF} = \dfrac{2.0001}{\dfrac{.0001}{2.0001}} \doteq 40{,}004.$ (large EMF!)

Def. The condition number of an $n\times n$ matrix: $\text{cond}(A)$,

is the maximum possible error magnification factor for

solving: $A\vec{x}=\vec{b}$ over all choices of $\vec{b}$.

To compactly define the condition number of a matrix A,

we first need the definition of a matrix norm,

defined analogously with a vector norm:

def. $\|A\|_\infty$ = maximum absolute row sum.

Thm:

$$\boxed{\text{cond}(A) = \|A\|\cdot\|A^{-1}\|}$$

Using the Theorem, we compute cond(A) for the previous system.

$$A = \begin{bmatrix} 1 & 1 \\ 1.0001 & 1 \end{bmatrix}; \quad A^{-1} = \begin{bmatrix} -10000 & 10000 \\ -10001 & -10000 \end{bmatrix}$$

$$\|A\|_\infty = \underline{2.0001} \qquad \|A^{-1}\|_\infty = \underline{20{,}001}$$

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\| = (2.0001)(20{,}001) \approx \boxed{40{,}004}.$$

Note that we get EMF, as before. Since the condition number reveals the max. possible emf, so the error magnification for any $\vec{b}$ for this system $\leq$ 40,004.

In general, if cond(A) $\approx 10^k$, we should expect to lose about k digits of accuracy in our solution.

Note that the $\boxed{\text{Hilbert Matrix}}$: $H_{ij} = \frac{1}{i+j-1}$ is a well-known example of a matrix w/ large condition number.

Just as before, we can extend the idea of a $\boxed{\text{norm to matrices}}$ where the following (3) properties are satisfied:

(1) $\|A\| \geq 0$ with equality iff $A = 0$

(2) $\forall\, \alpha \in \mathbb{R}$, $\|\alpha A\| = |\alpha| \cdot \|A\|$

(3) for matrices A, B. $\|A+B\| \leq \|A\| + \|B\|$.

Note: We define the $\boxed{\text{1-Norm}}$ of a vector $\|\vec{x}\|_1$ as

$\boxed{\|\vec{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|}$ & $\boxed{\|A\|_1 = \text{max absolute column sum}}$.

In addition, a <u>matrix norm</u> is said to be an $\boxed{\text{operator norm}}$ if $\|A\|$ can be written in terms of a particular vector norm as:

$$\|A\| = \max \frac{\|A\vec{x}\|}{\|\vec{x}\|} \qquad (\vec{x} \neq \vec{0})$$

This yields the inequality: $\boxed{\|A\| \cdot \|\vec{x}\| \geq \|A\vec{x}\|}$ $(*)$

We now prove the main theorem of the section: <u>$\mathrm{cond}(A) = \|A\| \cdot \|A^{-1}\|$</u>

$\boxed{\text{Proof}}$: Let $A(\vec{x} - \vec{x}_a) = \vec{r}$ & $A\vec{x} = \vec{b}$.

By $(*)$ it follows that: $\boxed{\|A^{-1}\| \|\vec{r}\| \geq \|\vec{x} - \vec{x}_a\|}$ (1)

Also, by $(*)$: $\|A\| \|\vec{x}\| \geq \|A\vec{x}\| = \|\vec{b}\|$

$$\rightarrow \boxed{\frac{1}{\|\vec{b}\|} \geq \frac{1}{\|A\| \|\vec{x}\|}} \qquad (2)$$

$$\frac{\|A\|}{\|\vec{b}\|} \geq \frac{1}{\|\vec{x}\|}$$

Multiply (1) by $\frac{1}{\|\vec{x}\|}$:

$$\frac{\|\vec{x} - \vec{x}_a\|}{\|\vec{x}\|} \leq \frac{\|A^{-1}\| \|\vec{r}\|}{\|\vec{x}\|} \leq \frac{\|A^{-1}\| \cdot \|\vec{r}\|}{\|\vec{b}\|} \cdot \|A\|$$

$$\left(\text{Recall: } \mathrm{EMF} = \frac{\frac{\|\vec{x} - \vec{x}_a\|}{\|\vec{x}\|}}{\frac{\|r\|}{\|b\|}} \right) \Rightarrow \boxed{\mathrm{EMF} \leq \|A\| \cdot \|A^{-1}\|}$$

So EMF has <u>$\|A\| \cdot \|A^{-1}\|$</u> <u>as upper bound</u>!

(*) One can also show, using the operator norm definition that this "worst case" is <u>always attainable</u>.

Swamping

A second common source of error in classical Gaussian elimination is much easier to fix. (swamping).

Ex.

$$\left.\begin{aligned} 10^{-20} x_1 + x_2 &= 1 \\ x_1 + 2x_2 &= 4 \end{aligned}\right\}$$ We solve this system two ways.

(I) Exact solution:

$$\left[\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 2 & 4 \end{array}\right] \sim \left[\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 0 & 2-10^{20} & 4-10^{20} \end{array}\right]$$

$$\rightarrow \boxed{x_1 \simeq 2, \; x_2 \simeq 1.}$$

(II) Computer version of Gaussian elimination

$$\left[\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 1 & 2 & 4 \end{array}\right] \sim \left[\begin{array}{cc|c} 10^{-20} & 1 & 1 \\ 0 & 2-10^{20} & 4-10^{20} \end{array}\right]$$

(*) Note that $2-10^{20} = -10^{20}$ due to rounding, and similarly $4-10^{20}$ is stored as -10^{20}.

The "computer" solution is consequently: $\boxed{x_1 = 0, \; x_2 = 1}$

This solution has a very large <u>relative error</u>!

Q: How do we fix this problem?

A: Apply row exchanges + Gaussian elimination.

III $\left[\begin{array}{cc|c} 1 & 2 & 4 \\ 10^{-20} & 1 & 1 \end{array}\right] \sim \left[\begin{array}{cc|c} 1 & 2 & 4 \\ 0 & 1-2*10^{-20} & 1-4*10^{-20} \end{array}\right]$

Note that $1-2*10^{-20}$ is stored as 1 &

$1-4*10^{-20}$ is stored as 1, This gives the

solution: $\boxed{x_1 = 2, \ x_2 = 1}$, as needed.

In summary, "version II" failed because we used a large multiplier (10^{20}) during Gaussian elimination. This multiplier effectively "swamped" the bottom equation, so that after "row replacement" our system essentially consists of two copies of row 1 (thereby leading to an incorrect answer).

"Version III" of the problem completes elimination without swamping because the multiplier here is very small (10^{-20}). Here, Gaussian elimination basically preserves the linear independence of the rows of the original system.

Big Idea: Multipliers in Gaussian elimination should be kept small to preserve independence / avoid swamping. If we carefully implement row swaps ("partial pivoting") we can avoid the perils of large multipliers & swamping. See 2.4: Partial Pivoting.