2.3 Sources of Error

Two major sources of error in numerical Gaussian elimination:

1. Ill-conditioning
2. "Swapping"

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 \( \mathbf{v} \in \mathbb{R}^n \), \( \mathbf{v} = (v_1, v_2, \ldots, v_n) \) we define

\[
\| \mathbf{v} \| = \sqrt{a_1^2 + \ldots + a_n^2} ; \text{ This called the Euclidean or 2-Norm of a vector. (\(\| \mathbf{v} \|_2\) is 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 the following properties:

1. \( \| \mathbf{v} \| \geq 0 \) with equality if and only if \( \mathbf{v} = \mathbf{0} \).
2. \( \forall \mathbf{v}, \mathbf{w} \in \mathbb{R}^n, \| \mathbf{v} + \mathbf{w} \| \leq \| \mathbf{v} \| + \| \mathbf{w} \| \) (Triangle inequality)
3. \( \forall \alpha \in \mathbb{R} \), \( \| \alpha \mathbf{v} \| = |\alpha| \| \mathbf{v} \| \)

We now define the \( \infty \)-Norm:

\[
\| \mathbf{v} \|_\infty = \max |v_i| \quad \text{where} \quad i \in \mathbb{N}
\]

i.e. \( \| \mathbf{v} \|_\infty \) is the maximum of the absolute value of the components of \( \mathbf{v} \).
Def. The definition of Backward Error (BE) and Forward Error (FE) are defined for systems of linear equations analogously with Ch. 1.

Let $\vec{x}_0$ be an approximate solution of the linear system $A\vec{x} = \vec{b}$. The residual $\vec{r}$ is the vector: $\vec{r} = \vec{b} - A\vec{x}_0$.

The Backward Error is $\|\vec{r}\|_\infty = \|\vec{b} - A\vec{x}_0\|_\infty$.

The Forward Error is $\|\vec{x} - \vec{x}_0\|_\infty$.

Ex. Find $BE$ & $FE$ for the approximate solution $\vec{x}_0 = \langle 1, 1 \rangle$ 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 $\vec{x} = \langle 2, 1 \rangle$.

$BE = \|\vec{b} - A\vec{x}_0\|_\infty = \|\begin{bmatrix}
3 \\
2
\end{bmatrix} - \begin{bmatrix}
1 & 1 \\
3 & -4
\end{bmatrix}\begin{bmatrix}
1 \\
1
\end{bmatrix}\|_\infty = \|\begin{bmatrix}
2 \\
1
\end{bmatrix}\|_\infty = 3$.

$FE = \|\vec{x} - \vec{x}_0\|_\infty = \|\begin{bmatrix}
2 \\
1
\end{bmatrix} - \begin{bmatrix}
1 \\
1
\end{bmatrix}\|_\infty = \|\begin{bmatrix}
1 \\
0
\end{bmatrix}\|_\infty = 1$. 
Note that FE & BE can be different orders of magnitude.

Ex.

Let \( \tilde{x}_0 = \langle -1, 3, 0, 001 \rangle \) for the system:

\[
\begin{align*}
    x_1 + x_2 &= 2, \\
    1.001x_1 + x_2 &= 2.001
\end{align*}
\]

Using Gaussian elimination, we find: \( \bar{x} = \langle 1, 1 \rangle \).

\[
\| A \tilde{x}_0 - \bar{x} \|_\infty = \| \begin{bmatrix} 2 & 0.001 \\ 1.001 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \|_\infty = 0.001
\]

\[
\| \tilde{x} - \bar{x} \|_\infty = \| \begin{bmatrix} 1 \\ 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \|_\infty = \| \begin{bmatrix} 2 \\ 2.0001 \end{bmatrix} \|_\infty = 2.0001
\]

**Geometric Interpretation:**

The two lines represent the two equations of the system.

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

\[\begin{align*}
\text{Relative backward error of } A \tilde{x} = \bar{x} \text{ is:} & \quad \frac{\| A \tilde{x}_0 - \bar{x} \|_\infty}{\| \tilde{x}_0 \|_\infty} \\
\text{Relative forward error is:} & \quad \frac{\| \tilde{x} - \bar{x} \|_\infty}{\| \tilde{x} \|_\infty}
\end{align*}\]
The error magnification factor for \( A \tilde{x} = \tilde{b} \)
is the ratio of the two:

\[
\text{EMF} = \frac{\text{Rel. FE}}{\text{Rel. BE}} = \frac{\| \tilde{x} - \tilde{x}_0 \|_{\infty}}{\| \tilde{x} - \tilde{x}_0 \|_{1}} = \frac{\| \tilde{b} - \tilde{b}_0 \|_{\infty}}{\| \tilde{b} - \tilde{b}_0 \|_{2}}
\]

For the previous example, then, the rel. BE = \( \frac{1}{2,000} \) = .0005%.

A rel. normal error is: \( \frac{2,000}{1} = 200\% \)

so \( \text{EMF} = \frac{2,000}{1} = 2,000 \). (Type EMF!)

\[
\text{Def.} \quad \text{The condition number of an } \text{matrix } \text{is the maximum possible error magnification factor for solving: } A\tilde{x} = \tilde{b} \text{ over all choices of } \tilde{b}.
\]

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:

\[
\text{Def.} \quad \| A \|_\infty = \text{Maximum absolute row sum}
\]

Then:

\[
\text{cond} (A) = \| A \|_\infty \| A^{-1} \|_1
\]
Using the theorem, we compute (c) and (d) for the previous system.

\[ A = \begin{bmatrix} 1 & 0.001 \\ 0 & 1 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} 1 & -0.001 \\ -0.001 & 1 \end{bmatrix} \]

\[ \|A\|_1 = 2.001 \quad \|A^{-1}\|_1 = 2.001 \quad \text{cond}(A) = 1.00 \cdot \|A\|_1 \cdot \|A^{-1}\|_1 = 4.000 \]

Note that we get EMF as before. Since the condition number reveals the max. possible error in the error magnification for any \( T \) for this system, \( \text{cond}(A) \leq 40.004 \).

In general, \( \|\text{cond}(A)\|_1 \leq 10^k \), we should expect to lose about \( k \) digits of accuracy in our solution.

Just as before, we can extend the idea of a norm to matrices where the following (3) properties are satisfied:

1. \( \|A\|_1 \geq 0 \) with equality iff \( A = 0 \)
2. \( \forall a \in \mathbb{R}, \|aA\|_1 = |a| \cdot \|A\|_1 \)
3. \( \text{for matrices } A, B, \|A + B\|_1 \leq \|A\|_1 + \|B\|_1 \)

Note: we define the \( \|x\|_1 \) norm of a vector \( \|x\|_1 \), \( \|x\|_1 = |x_1| + |x_2| + \ldots + |x_n| \)

\[ \|A\|_1 = \max \text{ absolute column sum} \]
In addition, a matrix norm is said to be an operator norm if \( \|A\| \) can be written in terms of a particular vector norm as:

\[
\|A\| = \max_{\|x\| = 1} \|Ax\| \quad (x \neq 0)
\]

This yields the inequality:

\[
\|A\| \cdot \|x\| = \|Ax\|
\]

We now give the main theorem of the section:

\[
\text{r}e(c) = \|A\| \cdot \|A^*\|
\]

\[\text{Proof:}\]

Let \( A(x - x_n) = \bar{x} \) and \( \Delta x = x - x_n \).

By (2) it follows that:

\[
\|A(x - x_n)\| = \|\bar{x}\| \geq \frac{1}{\|A\|} \|\bar{x}\|
\]

Also, by (2):

\[
\|A\| \cdot \|x\| = \|Ax\| = \|x - x_n\|
\]

\[
\frac{1}{\|A\|} \cdot \|x\| \geq \|x - x_n\|
\]

Multiply (1) by \( \frac{1}{\|A\|} \):

\[
\frac{1}{\|A\|} \cdot \|x - x_n\| \leq \|\bar{x}\| \leq \|A\| \cdot \|x\|
\]

(Recall: EMF = \[\frac{\|x - x_n\|}{\|x\|}\])

\[
\text{EMF} \leq \|A\| \cdot \|A^*\|
\]

So EMF has \( \|A\| \cdot \|A^*\| \) as upper bound!

(1) One can also show, using the operator norm definition, that the "worst case" is always attainable.
A second common source of error in classical Gaussian elimination is much easier to fix. (swapping).

Ex.

\[
\begin{align*}
10^{-20} x_1 + x_2 &= 1 \\
 x_1 + 2x_2 &= 4
\end{align*}
\]

We solve this system two ways.

I Exact solution:

\[
\begin{bmatrix}
10^{-20} & 1 & 1 \\
2 & 4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
10^{-20} \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
2.2 \\
-2 \times 10^{-20}
\end{bmatrix}
\]

II Computer version of Gaussian elimination

\[
\begin{bmatrix}
10^{-20} & 1 & 1 \\
2 & 4
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
10^{-20} \\
0
\end{bmatrix}
\]

\[
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
= 
\begin{bmatrix}
2.2 \\
-2 \times 10^{-20}
\end{bmatrix}
\]

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: \(x_1 = 0, x_2 = 1\)

This solution has a very large relative error!
Q: How do we fix this problem?
A: Apply row exchanges + Gaussian elimination.

$$\begin{bmatrix} 1 & 2 & 1 & 4 \\ 10^{-20} & 1 & 1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 2 & & 2 \\ 0 & 1 - 2 \times 10^{-20} & 1 - 1 \times 10^{-20} & 1 - 1 \times 10^{-20} \end{bmatrix}$$

Note that $1 - 2 \times 10^{-20}$ is stored as $1$ & $1 - 1 \times 10^{-20}$ is stored as $2$. Thus, the solution is $x_1 = 2, x_2 = 1$, 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 2 (thereby leading to an incorrect answer).

"Version III" of the problem completes elimination without swamping because the multiplier here is very small ($10^{-20}$). Hence, 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 problems of large multipliers & swamping. (See 2.4: Partial Pivoting.)