3X03 Assignment 1

Neelkant Teeluckdharry

September 2024

1 Problem 1

The Taylor series of $\sin x$, centered at $x = 0$, can be written as follows:

$$\sin x \approx x - \frac{x^3}{3!} + \frac{x^5}{5!} ... = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$$

For $\sin x$ to be approximated as x, the subsequent terms of its Taylor series must have sufficiently small error to be negligible. Since the magnitude of the error terms of the sequence are strictly decreasing, this means that bound on the relative error of E_3 must be less than $\frac{10^{-14}}{2}$. Applying Taylor's theorem,

$$|E_3| = \left| \frac{f^{(3)}(z)x^3}{3!} \right| = \left| \frac{x^3 \cos z}{6} \right|$$

for some $0 < z < x$ The relative error is given by:

$$|\delta| = \left| \frac{\sin x - x}{x} \right| = \frac{|E_3|}{|x|} = \frac{|x^2 \cos z|}{6} \leq \frac{x^2}{6} < \frac{10^{-14}}{2}$$

This implies that the range is:

$$\boxed{|x| < \sqrt{3 * 10^{-14}} \approx 1.73 \times 10^{-7}}$$

2 Problem 2

The Taylor series of e^x, centered at $x = 0$ is given by:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

Applying Taylor's theorem, the truncation error is:

$$E_{n+1} = \frac{e^x x^{n+1}}{(n+1)!} < 10^{-10}$$

At $x = 0.5$, this is equivalent to:

$$2^{n+1}(n+1)! > 10^{10} \cdot e^{0.5} \implies \boxed{n \geq 10}$$

This implies at least 10 terms are need to attain such a truncation error.

3 Problem 3a

Consider FP$(10, 3, -2, 2)$. Let $a = 9.98 \times 10^2$, $b = 2.00 \times 10^0$, $c = -1.00 \times 10^0$ be three representable floating point numbers in the system.

$$\text{LHS} = (a + b) + c = (9.98 \times 10^2 + 2.00 \times 10^0) + (-1.00 \times 10^0) \rightarrow \text{overflow}$$

$$\text{RHS} = a + (b + c) = 9.98 \times 10^2 + (2.00 \times 10^0 + -1.00 \times 10^0) = 9.99 \times 10^2$$

4 Problem 3b

Consider an FP$(10, 3, -2, 2)$. Let $a = 1.00 \times 10^2$, $b = 9.99 \times 10^1$, $c = 1.00 \times 10^{-2}$ be three representable floating point numbers in the system.

$$\text{LHS} = (a \times b) \times c = (1.00 \times 10^2 \times 9.99 \times 10^1) \times 1.00 \times 10^{-2} \rightarrow \text{overflow}$$

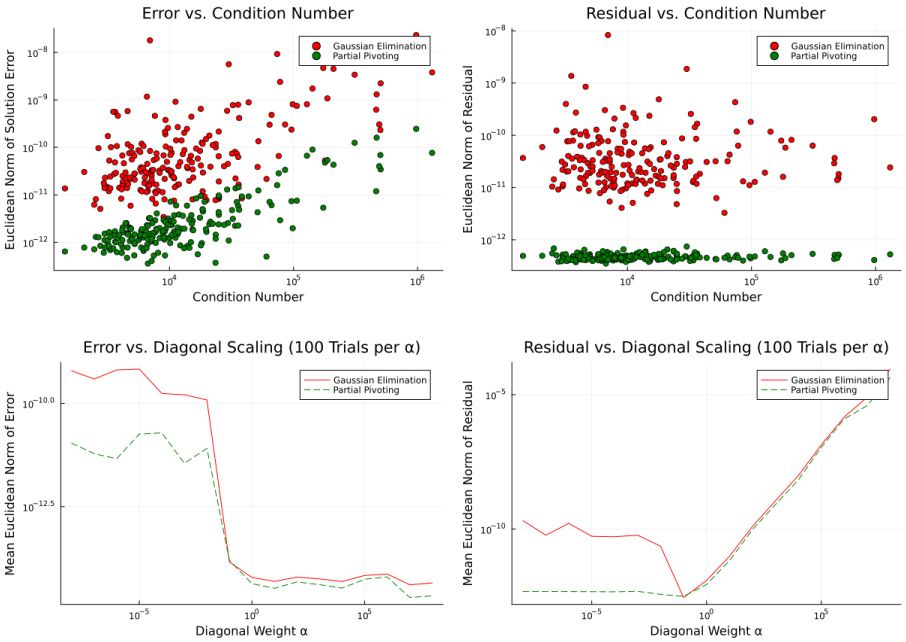
$$\text{RHS} = a \times (b \times c) = 1.00 \times 10^2 \times (9.99 \times 10^1 \times 1.00 \times 10^{-2}) = 9.99 \times 10^1$$

5 Problem 8

5.1 Answers

The trends observed in the top plots, depicting Solution Error Norm $\stackrel{\text{def}}{=} \|x - x^*\|$ vs Condition Number as well as Residual Norm $\stackrel{\text{def}}{=} \|b - Ax^*\|$ vs Condition Number, are the following:

- Euclidean Norms of both solutions and residuals are usually larger when using Gaussian Elimination as opposed to Partial Pivoting. This suggests that pivoting is more numerically stable than normal Gaussian Elimination.
- Euclidean Norm of solution error increases with condition number (within the sampled range of values).
- The Residual Norm remained constant at an order of magnitude of 10^{-12} when using Partial Pivoting, while samples using Gaussian Elimination varied from $10^{-10} - 10^{-11}$.
- There is a larger spread in the data collected using Gaussian Elimination.



These observations are supported by the following arguments:

$$||x^* - \tilde{x}||_2 = ||A^{-1}r||_2 \leq ||A^{-1}||_2 \cdot ||r||_2 = \frac{cond(A) \cdot ||r||_2}{||A||_2} \tag{1}$$

Obs. 2: Inequality (1) shows a direct proportionality between condition number and the bound on the solution error, which provides some justification to the observation. Furthermore,

$$\|r\|_2 = \|A(x^* - \tilde{x})\|_2 \geq \|A(x^* - x_{optimal})\| \tag{2}$$

Obs. 3: A is a randomly generated matrix of size 200 by 200. When using partial pivoting, each row operation will introduce the least possible error, as the best possible pivot is chosen. In the best case (with no cancellation errors), since the total number of operations is fixed for a matrix of this size, there will exist a lower bound on the norm of the residual r, as shown in (2). The lower bound will be achieved when the calculation of x is as numerically stable as possible, which it does when using partial pivoting. In contrast, using unpivoted Gaussian Elimination gives more unpredictable results, as a random pivot is chosen for each row operation. Therefore, the results will have more variability, which explains the results obtained in the plot.

The trends observed in the bottom plots, depicting Solution Error and Residual vs Diagonal Scaling respectively, are the following:

- As $\alpha \geq 0.1$ (roughly), both the error and residual are close using Gaussian Elimination and Partial Pivoting, indicating that both methods are able to provide good solutions.
- As $\alpha < 0.1$ (roughly), the difference between both the Error and Residual when using Gaussian Elimination.

These observations are supported by the following arguments:

Since $(A+\alpha I)x = \lambda x + \alpha x = (\lambda+\alpha)x$ and $A^{-1}Ax = A^{-1}\lambda x \implies A^{-1}x = \frac{1}{\lambda}x$.

It can be shown that

$$\|A+\alpha I\|^{-1} = \frac{1}{\min_{i=1,..n}|\lambda_i+\alpha|} \tag{3}$$

$$\|A+\alpha I\| = \max_{i=1,..n}|\lambda_i+\alpha| \tag{4}$$

And,

$$cond(A+\alpha I) = \frac{\max_{i=1,..,n}|\lambda_i+\alpha|}{\min_{i=1,...,n}|\lambda_i+\alpha|} \tag{5}$$

Obs 1: Thus, as α grows, the system becomes diagonally dominant and and the condition number will approach 1. In this case, the matrix will become well-conditioned and less sensitive to errors, making it more suitable to solve using unpivoted Gaussian Elimination. This will reduce both the condition number and absolute solution error.

Obs 2: As shown in (5), if $\alpha < 1$, the condition number could become quite large, especially if A has small eigenvalues. This leads to an ill-conditioned system, where even small errors are amplified, resulting in a large solution error, as demonstrated by the inequality in (1). In this case, using unpivoted Gaussian Elimination will likely cause the both the solution error and residual norms to increase compared to using partial pivoting. This is because partial pivoting improves numerical stability, by selecting large pivot elements and prevents small divisions which can lead to large numerical errors (as the the size of the Float64 type limits the precision of the stored matrix entry). This demonstrates how pivoting improves the performance of Gaussian Elimination, as it facilitates solving ill-conditioned systems.