Problem 1:

Q1.

$$e^x = \sum_{k=0}^{n} \frac{f^{(k)}(x)}{k!} h^k + E_{n+1} \quad at\ x=0,\ h=0.5$$

$$= 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{f^{(n+1)}(\xi)}{(n+1)!} 0.5^{n+1}$$

If let $E_{n+1} < 10^{-10}$, notice E_{n+1} is bounded by

$$\frac{e^{0.5} \cdot 0.5^{n+1}}{(n+1)!} \qquad \xi \in [x, x+h]$$

$\therefore \xi_{max} = x+h = 0.5$

$$\therefore \text{let } \frac{e^{0.5} \cdot 0.5^{n+1}}{(n+1)!} < 10^{-10}$$

$\therefore$ just try:

let $n=9$: $\frac{e^{0.5} \cdot 0.5^{9+1}}{10!} = 4.44 \times 10^{-10} > 1 \times 10^{-10}$

try $n=10$: $\frac{e^{0.5} \cdot 0.5^{11}}{11!} = 2.02 \times 10^{-11}$

$= 0.202 \times 10^{-10} < 1 \times 10^{-10}$

$\therefore n = 10.$

Problem 2:

Q2.

(a) Assume $a = 1.23$ $(t=3)$,

$b = 1.000 \times 10^3$ $(t=4)$

$c = -1.000 \times 10^3$ $(t=4)$. Take $t=3$, $\beta = 10$.

$(a+b)+c$: $fl[fl(fl(a)+fl(b)) + fl(c)]$

$= fl[fl[1.23 + 1.00\times10^3] - 1.00\times10^3]$

$= fl[fl[1.00123 \times 10^3] - 1.00\times10^3]$

$= fl[1.00\times10^3 - 1.00\times10^3]$

$= 0.00$

$a+(b+c) =$

$fl[fl(a) + fl[fl(b) + fl(c)]]$

$= fl[1.23 + fl[1.00\times10^3 - 1.00\times10^3]]$

$= 1.23$

$\therefore$ proved : $(a+b)+c \neq a+(b+c)$ in FP.

(b) Assume

$a = 371$

$b = 0.276$ $\quad t=3$

$c = 8.07$

$(a \cdot b) \cdot c =$

$fl[fl(fl(a) \cdot fl(b)) \cdot fl(c)]$

$= fl[fl[371 \times 0.28] \times 8.07]$

$= 839$

$a \cdot (b \cdot c) =$

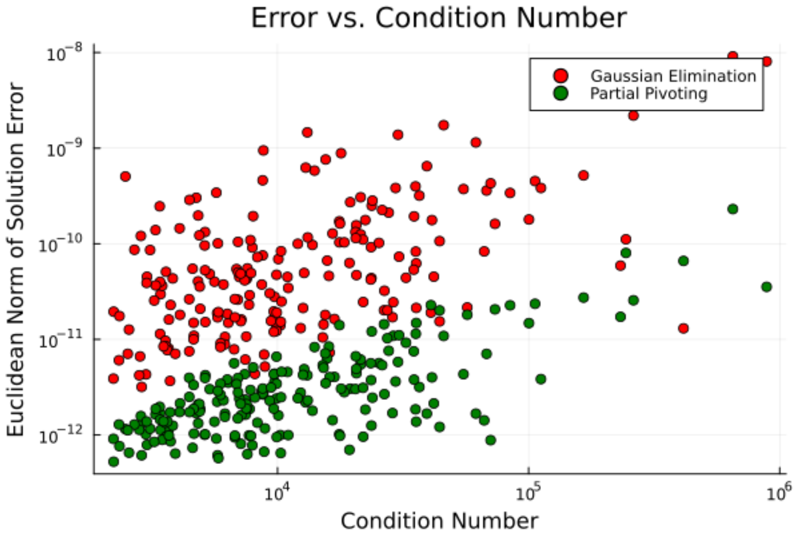
$fl[fl(a) \cdot fl[fl(b) \cdot fl(c)]]$

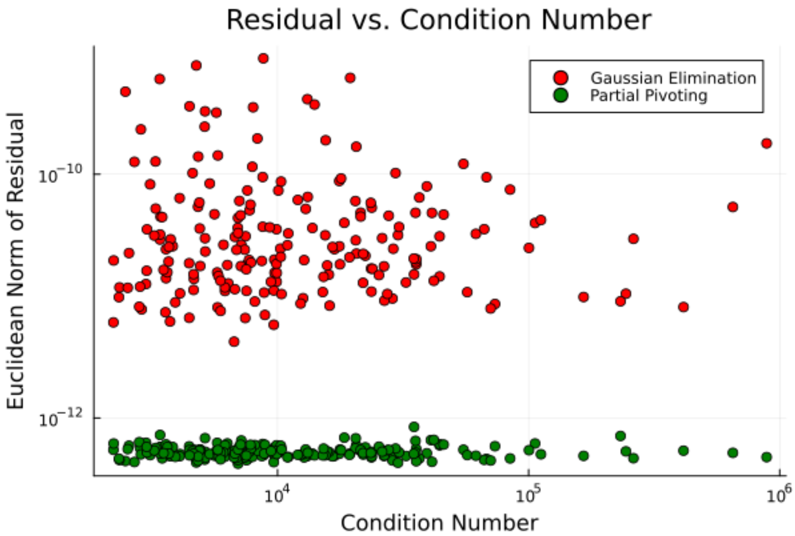
$= fl[371 \times fl[0.28 \times 8.07]]$

$= 838$

$\therefore$ proved : $(a \cdot b) \cdot c \neq a \cdot (b \cdot c)$ in FP arithmetic

Problem 7:





7.(a)

From the scatter plots, as the condition number of matrix A increases, the norm of the solution error also increases. The data points for Gaussian elimination without pivoting show larger errors than those with partial pivoting. When the condition number is small, the errors are tightly clustered near zero, but they become more scattered as the condition number grows. That is because the formula shows the relative error is bounded by:

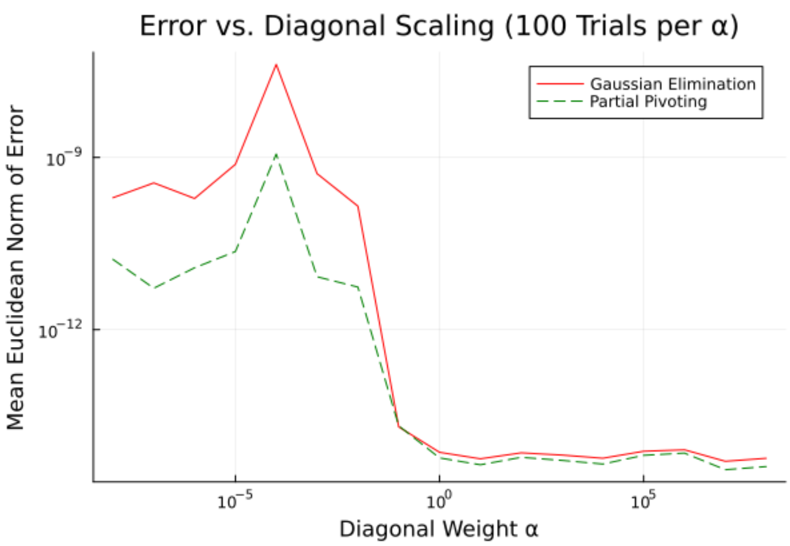
$$\frac{|x^* - \tilde{x}|}{|\tilde{x}|} \lesssim cond(A) * \frac{||\mathrm{r}||}{||\mathrm{b}||}$$

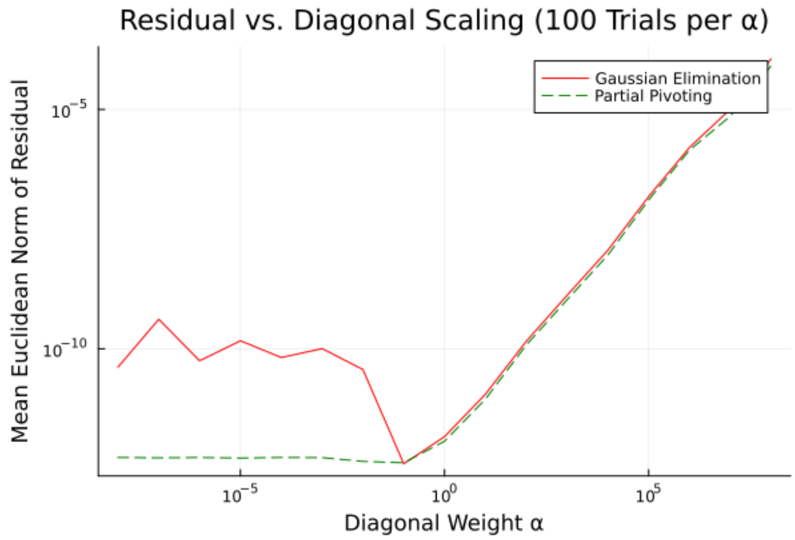
The residual norms, however, do not increase as clearly with the condition number. The residual is influenced mainly by floating-point round-off and not directly by cond(A). For Gaussian elimination without pivoting, residuals show a wider spread, while partial pivoting produces smaller and more consistent residuals.

$$r = b - A\tilde{x}$$

Mathematically, the relative error is approximately proportional to the condition number times the relative residual. Therefore, as cond(A) increases, the solution error can become large even if the residual remains small. Partial pivoting reduces the risk of dividing by small pivots, which prevents severe round-off and cancellation errors. This leads to more accurate solutions and smaller residuals compared with naive Gaussian elimination.

Q7 (b)





When examining the effect of the diagonal scaling factor α, both the residual and the error are larger for naive Gaussian elimination when α is small. At low α, the matrix behaves similarly to the unscaled system, and small pivots can appear, causing significant round-off error. Partial pivoting improves accuracy in this regime.

As α increases, the diagonal dominance of the matrix improves. The error for both methods decreases because the condition number of $A + \alpha I$ becomes smaller. However, the residual norm gradually increases. This occurs because in finite-precision arithmetic, large α values amplify rounding effects in the computation of $r = b - (A + \alpha I)x$, so the residual appears larger even though the actual error becomes smaller.

When α is sufficiently large, $A + \alpha I$ is dominated by its diagonal terms, so both Gaussian elimination and partial pivoting perform similarly—the matrix is well-conditioned, and pivoting provides little additional benefit.