1. For the following expressions, determine if they will evaluate as TRUE (logical 1) or FALSE (logical 0) in MATLAB, and explain why. Here `eps` $= 2^{-52}$.

   (a) `(1+2*eps) - 1 == 0`

   Solution. FALSE. `1+2*eps` evaluates as greater than 1, so `(1+2*eps)-1` is different from 0. □

   (b) `(1+eps/2) - 1 == 0`

   Solution. TRUE. `1+eps/2` evaluates as 1, so `(1+eps/2)-1` evaluates as 0. □

   (c) `eps/2 == 0`

   Solution. FALSE. `eps` is the distance from 1 to the next largest floating point number. `eps/2` is $2^{-52}/2 = 2^{-53}$ which is much larger than `realmin`. Thus, `eps/2` does not evaluate to zero. □

2. Assume that $0 < \varepsilon < 2^{-52}$. Consider the following system of equations.

$$\begin{aligned} \varepsilon\, x_1 + 2x_2 &= 4 \\ x_1 - x_2 &= -1 \end{aligned}$$

   (a) Find the exact solution.

   Solution. Either by elimination or using the inverse, we get

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \varepsilon & 2 \\ 1 & -1 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ -1 \end{pmatrix} = \frac{1}{-\varepsilon - 2} \begin{pmatrix} -1 & -2 \\ -1 & \varepsilon \end{pmatrix} \begin{pmatrix} 4 \\ -1 \end{pmatrix} = \frac{1}{2+\varepsilon} \begin{pmatrix} 2 \\ 4+\varepsilon \end{pmatrix}$$

$$\boxed{x_1 = \frac{2}{2+\varepsilon}, \quad x_2 = \frac{4+\varepsilon}{2+\varepsilon}}$$

□

   (b) Determine what the solution will be if you solve this system on a machine using double precision arithmetic *without* partial pivoting.

   Solution. Subtract $1/\varepsilon$ times the first equation from the second to get

$$\begin{aligned} \varepsilon\, x_1 + 2x_2 &= 4 \\ -\left(1 + \frac{2}{\varepsilon}\right) x_2 &= -1 - \frac{4}{\varepsilon} \\ \text{or} \qquad -\frac{2+\varepsilon}{\varepsilon} x_2 &= -\frac{4+\varepsilon}{\varepsilon} \end{aligned}$$

   In the computer the last equation evaluates to

$$-\frac{2}{\varepsilon} x_2 = -\frac{4}{\varepsilon}$$

   which has the solution $x_2 = 2$. Substitution into the first equation gives us

$$\varepsilon x_1 + 2(2) = 4$$

   which has the solution $x_1 = 0$. Thus, without partial pivoting we get $\boxed{(0, 2)}$. □

(c) Determine what the solution will be if you solve the system using double precision arithmetic *with* partial pivoting.

Solution. With partial pivoting we would put the largest element in the pivot position to get

$$\begin{aligned} x_1 - x_2 &= -1 \\ \varepsilon\, x_1 + 2x_2 &= 4 \end{aligned}$$

Now subtract ε times equation one from equation two to get

$$\begin{aligned} x_1 - x_2 &= -1 \\ (2+\varepsilon)x_2 &= 4+\varepsilon \end{aligned}$$

The last equation evaluates to

$$2x_2 = 4$$

which has the solution $x_2 = 2$, as before. Substituting into the first equation, we have

$$x_2 - 2 = -1$$

which has the solution $x_2 = 1$. Thus, with partial pivoting, we get $\boxed{(1,2)}$. □

(d) Which method gave a better result?

Solution. The exact solution is close to $(1,2)$, so partial pivoting gave us a much better approximation. □

3. Consider the diagonal matrix

$$A = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix}$$

where the diagonal entries are positive and arranged in decreasing order: $\sigma_1 \geq \sigma_2 > 0$. Show that the norm $\|A\|_1 = \sigma_1$, the norm of the inverse is $\left\|A^{-1}\right\|_1 = \dfrac{1}{\sigma_2}$, and hence the condition number is the ratio of the largest to smallest diagonal element:

$$\kappa_1(A) = \frac{\sigma_1}{\sigma_n}$$

Solution. Consider the norm of the vector $A\mathbf{x}$:

$$\begin{aligned} \|A\mathbf{x}\|_1 &= \|(\sigma_1 x_1, \sigma_2 x_2)\|_1 = |\sigma_1 x_1| + |\sigma_2 x_2| = \sigma_1|x_1| + \sigma_2|x_2| \\ &= \sigma_1\left(|x_1| + \frac{\sigma_2}{\sigma_1}|x_2|\right) \\ &\leq \sigma_1\left(|x_1| + |x_2|\right) \qquad \left(\text{since } \frac{\sigma_2}{\sigma_1} \leq 1\right) \\ &= \sigma_1\|\mathbf{x}\|_1 \end{aligned}$$

Thus

$$\frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq \sigma_1$$

Moreover, when $\mathbf{x} = (1,0)$, we have

$$\frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \sigma_1$$

Thus, since $\|A\mathbf{x}\|/\|\mathbf{x}\|$ is bounded by σ_1, and this bound is achieved, we have

$$\|A\|_1 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \sigma_1$$

For A^{-1}, since A is diagonal,

$$A^{-1} = \begin{pmatrix} 1/\sigma_1 & 0 \\ 0 & 1/\sigma_2 \end{pmatrix}$$

This is again a diagonal matrix whose largest diagonal element is $1/\sigma_2$. Thus, by the same reasoning as above,

$$\left\|A^{-1}\right\|_1 = \frac{1}{\sigma_2}$$

The condition number is

$$\kappa_1(A) = \|A\|_1 \left\|A^{-1}\right\|_1 = \frac{\sigma_1}{\sigma_2}$$

□

4. Consider the iteration scheme

$$\begin{aligned} x_{n+1} &= x_n - hy_n \\ y_{n+1} &= y_n + hx_n \end{aligned}$$

Find the eigenvalues and explain why (x_n, y_n) will go to infinity (in norm) as $n \to \infty$, starting with any initial condition (x_0, y_0) as long as one of x_0 or y_0 is nonzero.

Solution. Write this as

$$\mathbf{x}_{n+1} = A\mathbf{x}_n$$

where

$$A = \begin{pmatrix} 1 & -h \\ h & 1 \end{pmatrix}$$

The eigenvalues satisfy

$$0 = \det(A - \lambda I) = \begin{vmatrix} 1-\lambda & -h \\ h & 1-\lambda \end{vmatrix} = (1-\lambda)^2 + h^2$$

Thus the eigenvalues are

$$\lambda_{1,2} = 1 \pm hi$$

These have magnitude

$$|\lambda| = \sqrt{1+h^2} > 1$$

The solution can be written as

$$\mathbf{x}_n = A^n \mathbf{x}_0$$

Since both eigenvalues satisfy $|\lambda| > 1$,

$$\|A^n \mathbf{x}_0\| \to \infty$$

as $n \to \infty$, as long as $\mathbf{x}_0 \neq \mathbf{0}$.

□

5. Show that the equation

$$e^x + \sin x = 4$$

has a solution in the interval $[1, 2]$. Write out the first two steps of Netwon's method with starting guess $x_0 = 1$.

Solution. Write the equation as

$$0 = f(x) = e^x + \sin x - 4$$

Note that

$$\begin{aligned} f(1) &= e + \sin 1 - 4 < 3 + \sin 1 - 4 < 0 \\ f(2) &= e^2 + \sin 2 - 4 > 8 + \sin 2 - 4 > 0 \end{aligned}$$

Thus f changes sign on $[1, 2]$. Since e^x and $\sin x$ are continuous, f is continuous. Thus, by the Intermediate Value Theorem, f must cross 0 somewhere in $[1, 2]$.

The Newton iteration is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

which in this case is

$$x_{n+1} = x_n - \frac{e^{x_n} - \sin x_n - 4}{e^{x_n} - \cos x_n}$$

With $x_0 = 1$, we get

$$\begin{aligned} x_1 &= 1 - \frac{e - \sin 1 - 4}{e - \cos 1} \\ x_2 &= 1 - \frac{e - \sin 1 - 4}{e - \cos 1} - \frac{e^{x_1} - \sin x_1 - 4}{e^{x_1} - \cos x_1} \end{aligned}$$

$\square$