Problem 1
Show that a real symmetric matrix A is diagonalizable by an orthogonal matrix.

Solution: By the Schur decomposition theorem, for $A \in \mathbb{C}^{m\times m}$ we have

$$A = QUQ^{-1}$$

for upper triangular U, and unitary Q. Since Q is unitary it follows that $QQ^T = I$, as A was real symmetric so it follows that $Q^* = Q^T$. Then this implies that $Q^T = Q^{-1}$. Using this fact, left multiply by Q^T and right multiply by Q to give

$$Q^T AQ = U \tag{1}$$

Thus we only need show that U is a diagonal matrix and we will be done. To that end, let's examine the product UU^T and U^TU. Well by (1)

$$U^T = (Q^T AQ)^T = Q^T A^T Q$$

So

$$\begin{aligned} UU^T &= (Q^T AQ)(Q^T A^T Q) \\ &= Q^T A(QQ^T)A^T Q \\ &= Q^T AA^T Q \end{aligned} \tag{2}$$

Since A is given to be real symmetric, $A = A^T \implies AA^T = A^2 = A^TA$, so really UU^T is

$$UU^T = Q^T A^2 Q$$

With an almost identical argument to the above, notice

$$\begin{aligned} U^TU &= (Q^T A^T Q)(Q^T AQ) \\ &= Q^T A^T(QQ^T)AQ \\ &= Q^T A^T AQ \\ &= Q^T A^2 Q \end{aligned} \tag{3}$$

So $U^TU = UU^T$. However, this can only be true in generality if U is taken to be a diagonal matrix or its own inverse. Since U is upper triangular, it cannot be its own inverse by definition unless it was the identity matrix. U is not the identity matrix since it appeared in the Schur decomposition of A, and if it were we would have as a consequence $A = I$. Thus if $A \neq I$ the only option is that U is a diagonal matrix.

Then we conclude that we have found

$$Q^T AQ = U$$

for orthogonal matrix Q and diagonal matrix U, as desired.

Problem 2
Consider the following system

$$\begin{bmatrix} 1 & 1 \\ \epsilon & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Perform Gaussian elimination with partial pivoting to the modified row-scaled system and discuss what happens. If numerical issues arise, discuss them and assess how to improve the method.

Solution: We modify the the system by multiplying the last row by a constant c, such that $c\epsilon \gg 1$. This is done as if $\epsilon < \epsilon_{\text{mach}}$ our original system would be considered reduced, due to ϵ being interpreted as 0 by a computer in this case. With one potential numerical issue avoided we now turn to work with the augmented matrix,

$$\left[\begin{array}{cc|c} 1 & 1 & 2 \\ c\epsilon & c & c \end{array}\right]$$

Now we apply gaussian elimination with partial pivoting. Since $c\epsilon >> 1$, our max in column 1 is already in the correct position. No row permutations are necessary, we just need to clear $c\epsilon$. To accomplish this, we'll add $-c\epsilon$ row ones to row two, in other words $-c\epsilon R_1 + R_2 \to R_2$, giving

$$\left[\begin{array}{cc|c} 1 & 1 & 2 \\ c\epsilon - c\epsilon & c - c\epsilon & c - 2c\epsilon \end{array}\right] = \left[\begin{array}{cc|c} 1 & 1 & 2 \\ 0 & c(1-\epsilon) & c(1-2\epsilon) \end{array}\right]$$

Here we see hints of a numerical issue. If $\epsilon < \epsilon_{\text{mach}}$, then a computer will interpret $(1-\epsilon)$ as 1 and $(1-2\epsilon)$ as 1 if $\epsilon < \epsilon_{\text{mach}}/2$. If instead ϵ is very small, but not smaller than machine epsilon, we should be ok to proceed. Though back substitution normally happens at this point in our usual algorithm, let's clear the first row second column's entry. To do this, perform $\frac{R_2}{c(1-\epsilon)} + R_1 \to R_1$. Note that this is equivalent to back substituting, just at a different point in time. This gives

$$\left[\begin{array}{cc|c} 1 & 0 & 2 - \frac{c(1-2\epsilon)}{c(1-\epsilon)} \\ 0 & c(1-\epsilon) & c(1-2\epsilon) \end{array}\right]$$

meaning

$$c(1-\epsilon)y = c(1-2\epsilon)$$

or

$$y = \frac{(1-2\epsilon)}{(1-\epsilon)}$$

and

$$x = 2 - \frac{(1-\epsilon)}{(1-2\epsilon)}$$

Thus if $\epsilon < \epsilon_{\text{mach}}/2$ the above will be interpreted as $y = 1, x = 1$. This indeed solves the system for this case, and if $\epsilon > \epsilon_{\text{mach}}$ then our answer will be precisely given by the definitions of x and y containing ϵ.

Problem 3
What can you say about the diagonal entries of a symmetric positive definite matrix?

I claim all entries along the diagonal are positive. I aim to prove this via induction.

Proof. Suppose $A \in \mathbb{F}^{m\times m}$ is symmetric positive definite. Then by definition

$$x^T Ax > 0$$

for all $x \neq 0$, $x \in \mathbb{F}^m$.

Note if $m = 1$, this reduces to scalar multiplication and is obviously true, as we have $cx^2 > 0 \implies c > 0$. I do not take this as my base case as it does not capture the lack of commutativity present in higher dimensions.

Base case: m = 2. Well

$$x^T Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 & a_{21}x_1 + a_{22}x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{4}$$

$$= a_{11}x^2 + a_{22}x^2 + a_{12}x_1x_2 + a_{21}x_1x_2$$

Since A is supposed to be symmetric positive definite, we know the above to be greater than zero. That is,

$$a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2 + a_{21}x_1x_2 > 0$$
$$\implies a_{11}x_1^2 + a_{22}x_2^2 > -x_1x_2(a_{12} + a_{21})$$

recall the only requirements placed on x are that the entries $x_1 \neq 0 \neq x_2$, or more precisely they cannot both be zero at the same time. Since there exist values of x_1, x_2 such that the right hand side evaluates to be negative (such as $x_1 = 1, x_2 = 1$), it must be that

$$a_{11} > 0 \quad \text{and} \quad a_{22} > 0$$

or we've violated the definition of symmetric positive definite.

From here we can observe an important pattern that will hold as we scale up our dimensions - this product can always be written in an analogous form. That is, we can have squares of entries of x with diagonals of A as coefficients on the left and a difference of non diagonal entries multiplied with non squared entries of x on the right hand side. As we scale we will have more terms on the right hand side, that is it won't be a single product, but the argument that there exist values of x_i to make the right hand side negative hold since we have infinite choices of x.

Inductive hypothesis: Suppose for $m = n$, symmetric positive definite $A^{m\times m}$ will have positive entries on the diagonal.

Inductive step: Show this holds for $m = n + 1$. We actually did most the heavy lifting in the base case. By the concluding observations of said case, we'll be able to write the left hand side of $x^T Ax > 0$ as

$$a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + \ldots a_{nn}x_n^2 + a_{n+1\,n+1}x_{n+1}^2$$

Where the first $1 \ldots n$ terms are handled by our inductive hypothesis. The right hand side will be the difference of many terms of the form $x_1 \cdot \ldots \cdot x_{n+1}(a_{12} + a_{21} + \ldots a_{n-1,n+1})$, often missing a term in the middle depending on what dimension we are in. Though this could be written down rigorously, it is not important. The key thing to note is that the right hand side will have those products of x_i in front, which can be chosen such that the right hand side is negative, which means that $a_{n+1,n+1} > 0$ since by our inductive hypothesis the previous n a_n terms were greater than zero.

Thus we have shown for arbitrary $m \in \mathbb{N}$, it must be that symmetric positive definite $A \in \mathbb{F}^{m\times m}$ has positive entries along the diagonal.

□

Problem 4
Consider the matrix $A \in \mathbb{C}^{m\times m}$ written in block form

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

With the added criteria A has LU decomposition if and only if the upper left $k \times k$ block matrix $A_{1:k,1:k}$ is nonsingular for each k, $1 \leq k \leq m$.

Solution: **(a):** First we verify the given formula. Observe

$$\begin{bmatrix} I & 0 \\ -A_{21}A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} IA_{11} + 0 \cdot A_{21} & IA_{12} + 0 \cdot A_{22} \\ -A_{21}A_{11}^{-1}A_{11} + IA_{21} & -A_{21}A_{11}^{-1}A_{12} + A_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} \\ -A_{21} + A_{21} & -A_{21}A_{11}^{-1}A_{12} + A_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

as desired.

(b): Now to verify the second given formula, which follows as a result of Gaussian elimination. To that end, we start with

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

In order to row reduce, we'll need to get rid of A_{21} in our first column. To do this, we'll need to add $-A_{21}A_{11}^{-1}$ times row one to row two. In less words, $-A_{21}A_{11}^{-1}R_1 + R_2 \to R_2$. This yields

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} - A_{21}A_{11}^{-1}A_{11} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

Which shows when we have a matrix in this form, it must be that the block matrix D in the lower right corner is

$$D = A_{22} - A_{21}A_{11}^{-1}A_{12}$$

as desired.

Problem 5
Consider solving $Ax = b$, with A and b complex valued of order m. That is, $A \in \mathbb{C}^{m\times m}$ and $b \in \mathbb{C}^m$. Modify this problem to be a problem where you only solve a *real* square system of order $2m$. Then compare the storage and number of floating-point operations for the real-valued method in (a) vs solving the original complex valued system $Ax = b$

Solution: Consider that if $Ax = b$, we may rewrite the system as

$$A = A_1 + iA_2$$

where $A_1 = Re(A)$ and $A_2 = Im(A)$

We may do the same thing for b, x which gives us

$$(A_1 + iA_2)(x_1 + ix_2) = b_1 + ib_2$$

where x_1, b_1 denote the real portion of their respective variables and x_2, b_2 denote the imaginary part. Then

$$A_1x_1 + A_1ix_2 + iA_2x_1 + iA_2ix_2 = b_1 + ib_2$$

or

$$A_1x_1 - A_2x_2 + i(A_1x_2 + A_2x_1) = b_1 + ib_2$$

Clearly A_1x_1 and A_2x_2 will introduce no i terms. Recall that A_2 was taken to be the imaginary part of A, which confusingly is real by definition (hence the i we factor out front). Then the only way the above can hold is if

$$A_1x_1 + A_2x_2 = b_1 \text{ and } A_1x_2 + A_2x_1 = b_2$$

Since all terms in the above equations are real valued (we were able to divide through by i in the rhs equation), we have two real square systems to solve each of order m. That is, the total order is $m+m = 2m$.

Now to compare the operation counts. We are given that Gaussian elimination requires $O(m^3/3)$ operations to solve for a real valued $m \times m$ system. Observe that for the real decomposition arrived at in part (a) we still still need to do Gaussian elimination, meaning it works out to be $O(2m^3/3)$.

Next consider that multiplication for real valued elements is a single operation. For complex valued multiplication, we have four operations - one for each product as a result of the distributive property. Thus the Gaussian elimination with complex entries could be as bad as $O(4m^3/3)$.

Then in conclusion, the real decomposition method is a whole $2/3m^3$ faster in order that going for immediate Gaussian elimination with complex entries.