Problem 1

(4+4+2 points)

a) Identify the problematic subtraction and give an algebraically equivalent formulation of $1 - \sqrt{1 - \epsilon}$ that does not suffer from this problematic subtraction for $|\epsilon| \ll 1$. Briefly explain your choice.

Answer: In $1 \stackrel{\downarrow}{-} \sqrt{1 - \epsilon}$ the initial subtraction (arrow on top) is problematic. $\sqrt{1 - \epsilon}$ results in a value close to 1 for $|\epsilon| \ll 1$. Subtracting numbers of similar size can lead to a loss of precision. An algebraically equivalent function is:

$$1 - \sqrt{1 - \epsilon} \times \frac{1 + \sqrt{1 - \epsilon}}{1 + \sqrt{1 - \epsilon}}$$
$$= 1 - (1 - \epsilon) \times \frac{1}{1 + \sqrt{1 - \epsilon}}$$
$$= \frac{\epsilon}{1 + \sqrt{1 - \epsilon}}$$

It works as we are doing an addition on numbers of similar magnitude.

b) Use $\epsilon = 0.0002$ and write it in floating point representation with a mantissa of k = 4 digits precision and base b = 10. Show that your equivalent formulation is much more accurate for $\epsilon = 0.0002$ compared to the original formulation in a).

Answer: First, using $1 - \sqrt{1 - \epsilon}$:

$$= 0.1 \times 10 - \sqrt{0.1 \times 10 - 0.2 \times 10^{-3}}$$

$$= 0.1 \times 10 - \sqrt{0.9998}$$

$$= 1 - 0.9998$$

$$= 0.0002$$

$$= 0.2000 \times 10^{-3}$$

Now, using $\frac{\epsilon}{1+\sqrt{1-\epsilon}}$:

$$= \frac{0.2 \times 10^{-3}}{0.1 \times 10 + \sqrt{0.1 \times 10 - 0.2 \times 10^{-3}}}$$

$$= \frac{0.2 \times 10^{-3}}{1 + 0.9998}$$

$$= \frac{0.2 \times 10^{-3}}{1 + 0.0999 \times 10^{1}}$$

$$= 0.10000 \times 10^{-3}$$

The actual value is $1.000005005 \times 10^{-4} = 0.1000005005 \times 10^{-3}$. So, the error for original formula is:

$$\approx 0.999949995 \times 10^{-4}$$

The error for the algebraic alternative is:

$$\approx 0.50005\times 10^{-8}$$

So, the original formula is worse than the algebraic alternative.

c) Use the theorem that was shown in class to predict the number of lost significant **bits** (i.e. to base 2) when executing the problematic subtraction for $\epsilon = 0.0002$.

Answer:

$$2^{-p} \le 1 - \frac{y}{x} \le 2^{-q}$$

$$2^{-p} \le 1 - \frac{\sqrt{1 - \epsilon}}{1} \le 2^{-q}$$

$$2^{-p} \le 1.000005005 \times 10^{-4} \le 2^{-q}$$

$$2^{-13} \le 1.000005005 \times 10^{-4} \le 2^{-12}$$

You lose at least 12 bits and at maximum 13 bits.

Problem 2

(5 points) Find the solution to the system of equations using Gaussian elimination:

$$3x + y - z = 1$$

$$x - y + z = -3$$

$$2x + y + z = 0$$

Answer: The augmented matrix is:

$$\begin{bmatrix} 3 & 1 & -1 & 1 \\ 1 & -1 & 1 & -3 \\ 2 & 1 & 1 & 0 \end{bmatrix}$$

$$R_2 \to R_2 + -\frac{1}{3}R_1$$

 $R_3 \to R_3 + -\frac{2}{3}R_1$

$$\begin{bmatrix} 3 & 1 & -1 & 1 \\ 0 & -\frac{4}{3} & \frac{4}{3} & -\frac{10}{3} \\ 0 & \frac{1}{3} & \frac{5}{3} & -\frac{2}{3} \end{bmatrix}$$

$$R_2 \rightarrow R_2 - R_1$$
 and $R_3 \rightarrow R_3 - 2R_1$

$$\begin{bmatrix} 1 & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ 0 & -\frac{4}{3} & \frac{4}{3} & -\frac{10}{3} \\ 0 & \frac{1}{3} & \frac{5}{3} & -\frac{2}{3} \end{bmatrix}$$

$$R_3 \rightarrow R_3 + \frac{1}{4}R_2$$

$$\begin{bmatrix} 1 & \frac{1}{3} & -\frac{1}{3} & \frac{1}{3} \\ 0 & -\frac{4}{3} & \frac{4}{3} & -\frac{10}{3} \\ 0 & 0 & 2 & -\frac{3}{2} \end{bmatrix}$$

Now,

$$z = -\frac{\frac{3}{2}}{2} = -\frac{3}{4}$$

$$y = \frac{\frac{-10}{3} - \frac{4}{3} \cdot -\frac{3}{4}}{-\frac{4}{3}}$$
$$= -\frac{7}{3} \times \frac{3}{4}$$

$$x = \frac{\frac{1}{3} + \frac{1}{3} \cdot -\frac{3}{4} - \frac{1}{3} \cdot \frac{7}{4}}{1}$$

$$= \frac{4}{12} - \frac{3}{12} - \frac{7}{12}$$

$$= -\frac{6}{12}$$

$$= -\frac{1}{2}$$

Problem 3

(2+8 points) Let
$$A = \begin{bmatrix} 2 & 0 & -1 & 2 \\ 2 & 7 & -11 & -2 \\ 4 & 1 & -6 & 2 \\ 2 & 1 & -3 & 0 \end{bmatrix}$$
 and $b = \begin{bmatrix} 4 \\ 3 \\ 5 \\ 1 \end{bmatrix}$

a) Check if Gaussian elimination with scaled partial pivoting can be applied to solve Ax = b. Answer:

This method can be applied if a unique solution exists. A linear system of equations Ax = b, where A is a square matrix, has a unique solution if A has an inverse, which means $det A \neq 0$.

$$det A = (2) \cdot det \begin{bmatrix} 7 & -11 & -2 \\ 1 & -6 & 2 \\ 1 & -3 & 0 \end{bmatrix} - (0) \cdot det \begin{bmatrix} 2 & -11 & -2 \\ 4 & -6 & 2 \\ 2 & -3 & 0 \end{bmatrix} + (-1) \cdot det \begin{bmatrix} 2 & 7 & -2 \\ 4 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} - (2) \cdot det \begin{bmatrix} 2 & 7 & -11 \\ 4 & 1 & -6 \\ 2 & 1 & -3 \end{bmatrix}$$

Now, we need to find determinants for each 3×3 matrix:

$$det \begin{bmatrix} 7 & -11 & -2 \\ 1 & -6 & 2 \\ 1 & -3 & 0 \end{bmatrix} = (7) \cdot det \begin{bmatrix} -6 & 2 \\ -3 & 0 \end{bmatrix} - (-11) \cdot \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} + (-2) \cdot \begin{bmatrix} 1 & -6 \\ 1 & -3 \end{bmatrix}$$
$$= (7) \cdot (6) + (11) \cdot (-2) - (2) \cdot (3)$$
$$= 42 - 22 - 6$$
$$= 14$$

$$det \begin{bmatrix} 2 & -11 & -2 \\ 4 & -6 & 2 \\ 2 & -3 & 0 \end{bmatrix} = (2) \cdot det \begin{bmatrix} -6 & 2 \\ -3 & 0 \end{bmatrix} - (-11) \cdot det \begin{bmatrix} 4 & 2 \\ 2 & 0 \end{bmatrix} + (-2) \cdot det \begin{bmatrix} 4 & -6 \\ 2 & -3 \end{bmatrix}$$
$$= (2) \cdot (6) + (11) \cdot (-4) - (2) \cdot (0)$$
$$= 12 - 44$$

$$\det \begin{bmatrix} 2 & 7 & -2 \\ 4 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} = (2) \cdot \det \begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix} - (7) \cdot \det \begin{bmatrix} 4 & 2 \\ 2 & 0 \end{bmatrix} + (-2) \cdot \det \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$$
$$= (2) \cdot (-2) - (7) \cdot (-4) - (2) \cdot (2)$$
$$= 20$$

$$\det \begin{bmatrix} 2 & 7 & -11 \\ 4 & 1 & -6 \\ 2 & 1 & -3 \end{bmatrix} = (2) \cdot \det \begin{bmatrix} 1 & -6 \\ 1 & -3 \end{bmatrix} - (7) \cdot \det \begin{bmatrix} 4 & -6 \\ 2 & -3 \end{bmatrix} + (-11) \cdot \det \begin{bmatrix} 4 & 1 \\ 2 & 1 \end{bmatrix}$$
$$= (2) \cdot (3) - (7) \cdot (0) - 11 \cdot (2)$$
$$= -16$$

Therefore,

$$det A = (2) \cdot (14) - (0) \cdot (-32) + (-1) \cdot (20) - (2) \cdot (-16)$$
$$= 28 - 20 + 32$$
$$= 40$$

Therefore, $det A \neq 0$, which means A is invertible/not singular.

b) Solve Ax = b for x by Gaussian elimination with scaled partial pivoting. **Answer:** First, we write the matrix in the following notation:

$$[A|b] = \begin{bmatrix} 2 & 0 & -1 & 2 & | & 4 \\ 2 & 7 & -11 & -2 & | & 3 \\ 4 & 1 & -6 & 2 & | & 5 \\ 2 & 1 & -3 & 0 & | & 1 \end{bmatrix}$$

Now, we need to find the maximum absolute values from each row:

$$S = (2, 11, 6, 3)$$

Using S, we can now find the Relative Pivots:

$$(|a_i/S_i|) = (2/2 2/11 4/6 2/3) = (1 2/11 2/3 2/3)$$

The largest pivot is in the first row. Therefore [A|b] and S can remain the same, no swapping is required.

We will now perform the following row operations:

$$R2 - R1 => R2$$

$$R3 - 2 \cdot R1 => R3$$

$$R4 - R1 => R4$$

The new matrix looks as follows:

$$\begin{bmatrix}
2 & 0 & -1 & 2 & | & 4 \\
0 & 7 & 0 & -4 & | & -1 \\
0 & 1 & -4 & -2 & | & -3 \\
0 & 1 & -2 & -2 & | & -3
\end{bmatrix}$$

S looks as follows:

$$S = (2, 11, 6, 3)$$

And we have the following Relative Pivots:

$$(|a_i/S_i|) = (0/2, 7/11, 1/6, 1/3) = (0, 0.636, 0.167, 0.333)$$

7/11 is the largest value and it's in the second position, so we don't need to change the order of the rows. Again, no swapping is required.

Now that we have verified the required position of the rows, we are ready to apply the following row operations:

$$R3 - 1/7 \cdot R2 => R3$$

 $R4 - 1/7 \cdot R2 => R4$

After applying these row operations, the matrix looks as follows:

$$\begin{bmatrix}
2 & 0 & -1 & 2 & 4 \\
0 & 7 & 0 & -4 & -1 \\
0 & 0 & -4 & -10/7 & -20/7 \\
0 & 0 & -2 & -10/7 & -20/7
\end{bmatrix}$$

S and the Relative Pivots look as follows:

$$S = (2, 11, 6, 3)$$

$$(a_i/S_i) = (1/2, 0, 4/6, 2/3) = (0.5, 0, 0.667, 0.667)$$

Both third and fourth entries have equal and highest pivots, so change in order is not required and there's no need to swap.

We now apply the following row operation:

$$R4 - 1/2 \cdot R3 => R4$$

On applying this operation, we have the following results:

$$\begin{bmatrix}
2 & 0 & -1 & 2 & 4 \\
0 & 7 & 0 & -4 & -1 \\
0 & 0 & -4 & -10/7 & -20/7 \\
0 & 0 & 0 & -5/7 & -10/7
\end{bmatrix}$$

$$S = (2, 11, 6, 3)$$

From this point, we can apply Backward Substitution to obtain the final values we're seeking.

The system of equations look as follows:

We can compute the results as follows:

$$-5/7 \cdot x_4 = -10/7 => x_4 = 2$$

$$-4 \cdot x_3 - 10/7 \cdot x_4 = -1 => -4 \cdot x_3 - 10/7 \cdot 2 => x_3 = 0$$

$$7 \cdot x_2 - 4 \cdot x_4 = -1 => 7 \cdot x_2 - 4 \cdot 2 = -1 => 7 \cdot x_2 = 7 => x_2 = 1$$

$$2 \cdot x_1 - x_3 + 2 \cdot x_4 = 4 => 2 \cdot x_1 - 0 + 2 \cdot 2 = 4 => x_1 = 0$$

Problem 4

(3+4+3 points) Let
$$A = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix}$$

a) Show that matrix A is positive definite.

Answer:

Method 1:

A square matrix is called positive definite if it's symmetric and all its eigenvalues are positive, that is $\lambda > 0$.

To check symmetry, we find the transpose.

$$A^T = \left[\begin{array}{rrr} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{array} \right] = A$$

Now that we have confirmed that the matrix is symmetric, we need to confirm that its eigenvalues are positive. We do this as follows:

$$|A - \lambda \cdot I| = \begin{vmatrix} (1 - \lambda) & 0 & 2 \\ 0 & (4 - \lambda) & 2 \\ 2 & 2 & (15 - \lambda) \end{vmatrix} = 0$$

We need to simplify $|A - \lambda \cdot I|$ in order to solve for λ .

$$|A - \lambda \cdot I| = (1 - \lambda) \cdot \begin{vmatrix} (4 - \lambda) & 2 \\ 2 & 15 - \lambda \end{vmatrix} - (0) \cdot \begin{vmatrix} 0 & 2 \\ 2 & (15 - \lambda) \end{vmatrix} + (2) \cdot \begin{vmatrix} 0 & (4 - \lambda) \\ 2 & 2 \end{vmatrix}$$

$$= (1 - \lambda) \cdot [(4 - \lambda)(15 - \lambda) - 4] + (2) \cdot [0 - 2(4 - \lambda)]$$

$$= (1 - \lambda) \cdot (60 - 15\lambda - 4\lambda + \lambda^2 - 4) - 4 \cdot (4 - \lambda)$$

$$= (1 - \lambda) \cdot (\lambda^2 - 19\lambda + 56) - 16 + 4\lambda$$

$$= \lambda^2 - 19\lambda + 56 - \lambda^3 + 19\lambda^2 - 56\lambda - 16 + 4\lambda$$

$$= -\lambda^3 + 20\lambda^2 - 71\lambda + 40$$

From the above calculations, we have found the following relation:

$$|A - \lambda \cdot I| = -\lambda^3 + 20\lambda^2 - 71\lambda + 40 = 0$$

On solving this, we find the following values for λ :

$$\lambda_1 = 15.618$$
 $\lambda_2 = 3.688$
 $\lambda_3 = 0.695$

Hence, we have verified that the matrix is symmetric and that all its eigenvalues are positive. Therefore, the matrix is positive definite.

Method 2: For A to be positive definite, z^TAz is strictly positive for every nonzero

vector z. We execute this as follows:

$$z^{T}Az = \begin{bmatrix} z_{1} & z_{2} & z_{3} \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix} \cdot \begin{bmatrix} z_{1} \\ z_{2} \\ z_{3} \end{bmatrix}$$

$$= \begin{bmatrix} z_{1} & z_{2} & z_{3} \end{bmatrix} \cdot \begin{bmatrix} z_{1} + 2z_{3} \\ 4z_{2} + 2z_{3} \\ 2z_{1} + 2z_{2} + 15z_{3} \end{bmatrix}$$

$$= z_{1}^{2} + 2z_{1}z_{3} + 4z_{2}^{2} + 2z_{2}z_{3} + 2z_{1}z_{3} + 15z_{3}^{2}$$

$$= (z_{1}^{2} + 4z_{1}z_{3} + 4z_{3}^{2}) + (4z_{2}^{2} + 4z_{2}z_{3} + z_{3}^{2}) + 10z_{3}^{2}$$

$$= (z_{1} + 2z_{3})^{2} + (2z_{2} + z_{3})^{2} + 10z_{3}^{2}$$

Since $z \neq 0$, it confirms that the following is always true:

$$(z_1 + 2z_3)^2 + (2z_2 + z_3)^2 + 10z_3^2 > 0$$

Hence, $z^T A z$ is strictly positive, which means A is positive definite.

b) Compute the LU decomposition of A.

Answer:

Method 1:

Compute U:

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix} \underbrace{(R3 - 2R1 = R3)}_{} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 2 & 11 \end{bmatrix} \underbrace{(R3 - 1/2 \cdot R2 = R3)}_{} \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 10 \end{bmatrix} = U$$

$$A = LU = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ L_1 & 1 & 0 \\ L_2 & L_3 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ L_1 & 4 & 2L_1 + 2 \\ L_2 & 4L_3 & 2L_2 + 2L_3 + 10 \end{bmatrix}$$

Comparing these entries with A, we can find the values of L, which are the following:

$$L_1 = 0$$

$$L_2 = 2$$

$$L_3 = 1/2$$

Therefore, we have established the following relation:

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1/2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 10 \end{bmatrix}$$

Method 2:

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ L_{11} & 1 & 0 \\ L_{21} & L_{22} & 1 \end{bmatrix} \cdot \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{bmatrix}$$

$$= \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{11} \cdot L_{11} & U_{12} \cdot L_{11} + U_{22} & U_{13} \cdot L_{11} + U_{23} \\ L_{21} \cdot U_{11} & U_{12} \cdot L_{21} + L_{22} \cdot U_{22} & U_{13} \cdot L_{21} + U_{23} \cdot L_{22} + U_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix}$$

From this, we can obtain all the requird values as follows:

$$U_{11} = 1$$

$$U_{12} = 0$$

$$U_{13} = 2$$

$$U_{11}L_{11} = 0 \qquad => L11 = 0$$

$$U_{12}L_{12} + U_{22} = 4 \qquad => U_{22} = 4$$

$$U_{13}L_{11} + U_{23} = 2 \qquad => U_{23} = 2$$

$$L_{21}U_{11} = 2 \qquad => L_{21} = 2$$

$$U_{12}L_{21} + L_{22}U_{22} = 2 \qquad => L_{22} = 1/2$$

$$U_{13}L_{21} + U_{23}L_{22} + U_{33} = 15 \qquad => U_{33} = 10$$

Therefore, we have established the following relation:

$$A = LU = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1/2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 10 \end{bmatrix}$$

c) Compute the Cholesky decomposition of A.

Answer:

Method 1:

Cholesky Decomposition:

$$A = \widetilde{L}\widetilde{L}^T = (LD^{1/2})(D^{1/2}L^T) = LDL^T = LU$$

We have already found U and L in the previous part. From the above relation, we can

deduce the following:

$$DL^{T} = U \to D \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 0 & 0 & 10 \end{bmatrix}$$
$$\to D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 10 \end{bmatrix}$$
$$\to D^{1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \sqrt{10} \end{bmatrix}$$

$$\widetilde{L} = LD^{1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 1/2 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & \sqrt{10} \end{bmatrix}$$

$$\widetilde{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 2 & 1 & \sqrt{10} \end{bmatrix}$$

Method 2:

A is symmetric and positive definite, hence we can directly apply Cholesky Decomposition on it.

$$A = \widetilde{L}\widetilde{L}^{T}$$

$$= \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \cdot \begin{bmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{bmatrix}$$

$$= \begin{bmatrix} L_{11}^{2} & L_{11}L_{21} & L_{11}L21 \\ L_{11}L_{21} & L_{21}^{2} + L_{22}^{2} & L_{21}L_{31} + L_{22}L32 \\ L_{31}L_{11} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^{2} + L_{32}^{2} + L_{33}^{2} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 15 \end{bmatrix}$$

From comparing these matrices, we can compute the elements of \widetilde{L} , as follows:

$$L_{11}^2 = 1 \to L_{11} = 1$$

$$L_{11}L_{21} = 0 \to L_{21} = 0$$

$$L_{11}L_{31} = 2 \to L_{31} = 2$$

$$L_{21}^2 + L_{22}^2 = 4 \to L_{22} = 2$$

$$L_{21}L_{31} + L_{22}L_{32} = 2 \to L_{32} = 1$$

$$L_{31}^2 + L_{32}^2 + L_{33}^2 = 15 \to L_{33} = \sqrt{10}$$

Finally, we have found the matrix \widetilde{L} .

$$\widetilde{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 2 & 1 & \sqrt{10} \end{bmatrix}$$