Numerical Methods 2022 — Midterm 1

Solutions

Problem 1 (5 pts). Consider the following value:

$$-13.375$$

Give the bit-string representation if we store it as an IEEE double precision number. You must explain the details instead of just giving the bit string.

Solution.

The binary representation of 13.375 can be calculated as 1101.011, which is equal to 1.101011×2^3. Since IEEE double uses biased representation for the exponent, the stored value of that should be $3 + 1023 = 10000000010$. Therefore, the IEEE double representation should be

$$\underbrace{1}_{\text{sign}}\ \underbrace{10000000010}_{\text{exponent}}\ \underbrace{101011\overbrace{0\ldots0}^{46}}_{\text{mantissa}}.$$

Problem 2 (20 pts). Let us assume

$$\beta = 2$$

and rounding even. Give an example of two binary values x and y, so that two guard digits are not enough to implement an exactly rounded operation for $x - y$. Note that in any place where rounding is needed, you must consider rounding even.

Solution.

Suppose that $p = 5$. If we have

$$x = 1.1111 \times 2^4$$

and

$$y = 1.1001 \times 2^0,$$

the exactly operation for $x - y$ is

$$x - y = 1.1111 \times 2^4 - 0.00011001 \times 2^4 = 1.11010111 \times 2^4.$$

This value is rounded to

$$1.1101 \times 2^4,$$

which is the output of the exactly round operation. However, when the two guard digits are applied, the exact result of floating point subtract $x - y$ becomes

$$1.111100 \times 2^4 - 0.000110 \times 2^4 = 1.110110 \times 2^4,$$

where is rounded to

$$1.1110 \times 2^4$$

in the computer.

Problem 3 (15 pts). We discussed an example in the course slides to show that rounding even is better than rounding up in calculating

$$(x \ominus y) \oplus y.$$

Can you give an example where rounding up is better than rounding even under one calculation of

$$(x \ominus y) \oplus y?$$

If not, please prove that rounding up cannot be better than rounding even. Let us assume $\beta = 10$, and you can choose the value p of your floating-point system. Note that x and y can be negative numbers.

Solution.

Let us take $p = 2$. When $x = -9.8$ and $y = 7 \times 10^{-1}$, we have the exactly result

$$x - y = -1.05 \times 10^{1}.$$

Thus, rounding up takes

$$x \ominus y = -1.1 \times 10^{1},$$

and rounding even takes

$$x \ominus y = -1.0 \times 10^{1}.$$

After that, for rounding up, we take the calculation

$$(x \ominus y) + y = -1.1 \times 10^{1} + 7 \times 10^{-1} = -1.03 \times 10^{1}$$

exactly, so rounding up takes

$$(x \ominus y) \oplus y = -1.0 \times 10^{1}.$$

For rounding even, since we take the calculation

$$(x \ominus y) + y = -1.0 \times 10^{1} + 7 \times 10^{-1} = -9.3$$

exactly, rounding even takes

$$(x \ominus y) \oplus y = -9.3.$$

To compare these two rounding results, we have

$$|x - (-10)| = 0.2 < 0.5 = |x - (-9.3)|.$$

Thus, we have an example that rounding up is better than rounding even.

Problem 4 (25 pts). Consider the following matrix:

$$A = \begin{bmatrix} 1 & 4 & -2/3 & -2 \\ 1 & 4 & -1 & -1 \\ 4 & 12 & -8 & 4 \\ 2 & 8 & 0 & 4 \end{bmatrix}$$

Please conduct pivoted LU factorization on A to answer the following questions. Choose the pivot that has the largest absolute value.

(a) (10 pts) Give

$$P_1, M_1, P_2, M_2, P_3, M_3, U$$

such that

$$M_3P_3M_2P_2M_1P_1A = U.$$

Hint: The calculated U should only contain integers.

(b) (10 pts) Following (a), what are the P and L such that $PA = LU$?

(c) (5 pts) Using the P, L, U obtained from (b), find the solution x for the linear system

$$Ax = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

by solving two triangular systems.

Solution.

(a) Step 1. Exchange the first and third row since 4 is has the largest absolute value:

$$P_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 1 & 4 & -1 & -1 \\ 1 & 4 & -2/3 & -2 \\ 2 & 8 & 0 & 4 \end{bmatrix}$$

Step 2. Do Gaussian elimination on the first column:

$$M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/4 & 1 & 0 & 0 \\ -1/4 & 0 & 1 & 0 \\ -1/2 & 0 & 0 & 1 \end{bmatrix}, M_1P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 1 & 1 & -2 \\ 0 & 1 & 4/3 & -3 \\ 0 & 2 & 4 & 2 \end{bmatrix}$$

Step 3. Exchange row 2 and 4 since 2 is larger:

$$P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, P_2M_1P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 2 & 4 & 2 \\ 0 & 1 & 4/3 & -3 \\ 0 & 1 & 1 & -2 \end{bmatrix}$$

Step 4. Do Gaussian elimination on column 2:

$$M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 \\ 0 & -1/2 & 0 & 1 \end{bmatrix}, M_2P_2M_1P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & -2/3 & -4 \\ 0 & 0 & -1 & -3 \end{bmatrix}$$

Step 5. Exchange row 3 and 4 since -1 is larger in absolute value:

$$P_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, P_3M_2P_2M_1P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & -2/3 & -4 \end{bmatrix}$$

Step 6. Do Gaussian elimination on column 3:

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2/3 & 1 \end{bmatrix}, M_3P_3M_2P_2M_1P_1A = \begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & 0 & -2 \end{bmatrix} = U$$

(b) We have

$$P = P_3P_2P_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

According to slide "linear_LU3", L can be calculated as

$$\begin{bmatrix} (P_3P_2M_1^{-1})_{:,1} & (P_3M_2^{-1})_{:,2} & (M_3^{-1})_{:,3} & I_{:,4} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1/4 & 1/2 & 1 & 0 \\ 1/4 & 1/2 & 2/3 & 1 \end{bmatrix}$$

where M_1^{-1}, M_2^{-1} and M_3^{-1} can be obtained by flipping the sign of the off-diagonal element of M_1, M_2 and M_3, respectively.

(c) Because $Ax = b \iff PAx = Pb \iff L(Ux) = Pb$. First, we solve

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 1/4 & 1/2 & 1 & 0 \\ 1/4 & 1/2 & 2/3 & 1 \end{bmatrix} y = Pb = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Trivially, we get $y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ by forward substitution. Then, we solve

$$\begin{bmatrix} 4 & 12 & -8 & 4 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & -1 & -3 \\ 0 & 0 & 0 & -2 \end{bmatrix} x = y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

by back substitution and get

$$\begin{aligned} x_4 &= -0.5 \\ x_3 &= -(0 + -\frac{3}{2}) = 1.5 \\ x_2 &= \frac{0 - 2 \times \frac{-1}{2} - 4 \times 1.5}{2} = -2.5 \\ x_1 &= \frac{0 - 4 \times -0.5 + 8 \times 1.5 - 12 \times -2.5}{4} = 11. \end{aligned}$$

Problem 5 (35 pts). Consider the following matrix:

$$A = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 2 \end{bmatrix}$$

(a) (5 pts) Prove that A is positive semi-definite but not positive definite. If you calculate A's eigenvalues, you should calculate it by hand and show the details.

(b) (5 pts) Perform Cholesky factorization (the outer product form) and show that the procedure fails at a certain point.

(c) (5 pts) On page 5 of our slide "linear_chol1", we assumed that

$$B - \frac{vv^T}{\alpha} = \bar{L}\bar{L}^T. \tag{1}$$

When the procedure in (b) fails, if you can observe an $\bar{L}$ satisfying (1), can you still generate an L such that $A = LL^T$?

(d) (5 pts) Is the L satisfying $A = LL^T$ unique? If your answer is yes, provide a proof. Otherwise, give matrices $L_1 \neq L_2$ such that $A = L_1L_1^T = L_2L_2^T$.

(e) (10 pts) To fix the problem in (b), we can consider a "pivoted" version of Cholesky factorization. In this new verion, at each stage we move the **largest diagonal element** to the $(1,1)$ position of the remaining sub-matrix. Note that this can be done by P^TAP where P is a permutation matrix. Please redo Cholesky factorization with pivoting and show the details in each step.

Note that the procedure stops when all remaining diagonal elements are zeros. By this procedure, the obtained L has the largest number of non-zero diagonal elements.

(f) (5 pts) Is it possible that overall we have $P^TAP = LL^T$ for some permutation matrix P where L is the matrix obtained in (e)?

Solution.

(a) We can find the eigenvalues by solving for λs satisfying $\det(A - \lambda I) = 0$:

$$\begin{aligned} & \det\left(\begin{bmatrix} 1-\lambda & -1 & 1 \\ -1 & 1-\lambda & -1 \\ 1 & -1 & 2-\lambda \end{bmatrix}\right) = 0 \\ \iff & (1-\lambda)^2(2-\lambda) + 2 - 2(1-\lambda) - (2-\lambda) = 0 \\ \iff & (2-\lambda)(\lambda^2 - 2\lambda + 1 - 1) + 2\lambda = 0 \\ \iff & (2-\lambda)\lambda(\lambda - 2) + 2\lambda = 0 \\ \iff & \lambda(\lambda^2 - 4\lambda + 2) = 0 \end{aligned}$$

Solving for λ, we get the eigenvalues of A are 0 and $2 \pm \sqrt{2}$. Since they are all non-negative, A is positive semi-definite. However, there is a zero eigenvalue, so A is not positive definite.

Alternative solution:
It can be observed that

$$A = \begin{bmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Therefore, for any vector $v = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix}$, we have $v^T A v = (v_1 - v_2 + v_3)^2 + v_3^2 \geq 0$. By definition, A is positive semi-definite. Further, consider the case where $v_1 = v_2 \neq 0$ and $v_3 = 0$. We have $v^T A v = 0$ for some $v \neq 0$. Thus A is not positive definite.

(b) First, we calculate the first column of L:

$$L_{11} = \sqrt{\alpha} = \sqrt{A_{11}} = 1$$
$$L_{2:3,1} = v/\sqrt{\alpha} = A_{2:3,1} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$
$$A^{(2)} = B - \frac{vv^T}{\alpha} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

Then, we run into a divide-by-zero problem when calculating the factorization for the sub-matrix $A^{(2)}$:

$$\alpha_2 = \sqrt{0} = 0$$
$$v_2/\alpha_2 = [0]/0$$

Therefore, the procedure fails.

(c) We can oberve that $B - \frac{vv^T}{\alpha}$ can actually be factorized:

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}^T.$$

Therefore, (1) can be satisfied with $\bar{L} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$. According to page 5 of slide "linear_chol1", we can combine $\bar{L}$ with the first column we calculated in (b) to obtain a full Cholesky factorization for A:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

(d) Continuing from (c), we can observe that the matrix $\bar{L}$ is not unique since $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ also satisfies $B - \frac{vv^T}{\alpha} = \bar{L}\ \bar{L}^T$. Combining the result for the first column in (c), we get another factorization for A:

$$L_2 = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 1 & 0 \end{bmatrix} \neq L$$

(e) Step 1: Since 2 is the largest diagonal element we use $P_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ to move it to the top left corner:

$$A^{(1)} = P_1^T A P_1 = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix}$$

Let v and α be $A^{(1)}_{1,1}$ and $A^{(1)}_{2:3,1}$, respectively. Then

$$\begin{aligned} L_{11} =& \sqrt{\alpha} = \sqrt{2} \\ L_{2:3,1} =& v/\sqrt{\alpha} = \begin{bmatrix} \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \\ A^{(2)} =& A^{(1)}_{2:3,2:3} - \frac{vv^T}{\alpha} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \end{aligned}$$

Step 2: Applying the same rule to the sub-matrix $A^{(2)}$, we use $P_2 = I$ to permute $A^{(2)}$ (no change since the diagonal are all $\frac{1}{2}$). Let v and α be $A^{(2)}_{1,1}$ and $A^{(2)}_{2,1}$, we get

$$\begin{aligned} L_{22} =& \sqrt{\alpha} = 1/\sqrt{2} \\ L_{32} =& v/\sqrt{\alpha} = -1/\sqrt{2} \\ A^{(3)} =& \big[1/2\big] - \big[1/2\big] = \big[0\big] = L_{33}. \end{aligned}$$

We get the factorization $L = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \end{bmatrix}$.

(f) Using the L from (e), we get

$$LL^T = \begin{bmatrix} 2 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix} = P^T A P$$

where $P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. Therefore, it is possible (and generally right, although we do not prove it here).