Numerical Methods 2026 — Midterm 1

Solutions

Problem 1 (10 pts). Consider the following matrix

$$A = \begin{bmatrix} 4 & 2 & 8 & 12 \\ 2 & 10 & 10 & 9 \\ 8 & 10 & 36 & 34 \\ 12 & 9 & 34 & 45 \end{bmatrix}.$$

Show every step of performing Cholesky factorization on A with the outer product form. Hint: the resulting L contains only integer values.

Solution.

Step 1: Calculate $\sqrt{\alpha} = \sqrt{4} = 2$, so that

$$L^{(1)} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 6 & 0 & 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{v} = \begin{bmatrix} 2 \\ 8 \\ 12 \end{bmatrix}$$

in this step. Also, we calculate

$$B - \frac{\boldsymbol{v} \cdot \boldsymbol{v}^T}{\alpha} = \begin{bmatrix} 10 & 10 & 9 \\ 10 & 36 & 34 \\ 9 & 34 & 45 \end{bmatrix} - \begin{bmatrix} 1 & 4 & 6 \\ 4 & 16 & 24 \\ 6 & 24 & 36 \end{bmatrix} = \begin{bmatrix} 9 & 6 & 3 \\ 6 & 20 & 10 \\ 3 & 10 & 9 \end{bmatrix}.$$

Thus, we have

$$A^{(1)} = \begin{bmatrix} 2 & 1 & 4 & 6 \\ 1 & 9 & 6 & 3 \\ 4 & 6 & 20 & 10 \\ 6 & 3 & 10 & 9 \end{bmatrix}.$$

Step 2: Now we re-do the calculation on $A^{(1)}$, we have $\sqrt{\alpha^{(1)}} = \sqrt{9} = 3$,

$$L^{(2)} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 4 & 2 & 1 & 0 \\ 6 & 1 & 0 & 1 \end{bmatrix}, \text{ and } \boldsymbol{v}^{(1)} = \begin{bmatrix} 6 \\ 3 \end{bmatrix}$$

in this step. Then, calculate

$$B^{(1)} - \frac{\boldsymbol{v}^{(1)} \cdot \left(\boldsymbol{v}^{(1)}\right)^T}{\alpha^{(1)}} = \begin{bmatrix} 20 & 10 \\ 10 & 9 \end{bmatrix} - \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 16 & 8 \\ 8 & 8 \end{bmatrix}.$$

Therefore,

$$A^{(2)} = \begin{bmatrix} 2 & 1 & 4 & 6 \\ 1 & 3 & 2 & 1 \\ 4 & 2 & 16 & 8 \\ 6 & 1 & 8 & 8 \end{bmatrix}.$$

Step 3: Similarly, we have $\sqrt{\alpha^{(2)}} = \sqrt{16} = 4$,

$$L^{(3)} = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 6 & 1 & 4 & 0 \\ 9 & 3 & 2 & 1 \end{bmatrix} \text{ and, } \boldsymbol{v}^{(2)} = \begin{bmatrix} 8 \end{bmatrix}$$

in this step. Furthermore,

$$B^{(2)} - \frac{\boldsymbol{v}^{(2)} \cdot \left(\boldsymbol{v}^{(2)}\right)^T}{\alpha^{(2)}} = \begin{bmatrix} 8 \end{bmatrix} - \begin{bmatrix} 4 \end{bmatrix} = \begin{bmatrix} 4 \end{bmatrix}.$$

Therefore,

$$A^{(3)} = \begin{bmatrix} 2 & 1 & 4 & 6 \\ 1 & 3 & 2 & 1 \\ 4 & 2 & 4 & 2 \\ 6 & 1 & 2 & 4 \end{bmatrix}.$$

Step 4: In the final, $\sqrt{\alpha^{(3)}} = \sqrt{4} = 2$, so that

$$L^{(4)} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 4 & 2 & 4 & 0 \\ 6 & 1 & 2 & 2 \end{bmatrix}.$$

Problem 2 (30 pts). Consider the following matrix:

$$A = \begin{bmatrix} 2 & 3 & 12 \\ 4 & 5 & 20 \\ 1 & 2 & 5 \end{bmatrix} \tag{1}$$

(a) (5 pts) In pivoted LU factorization, please give

$$P_1, M_1, P_2, M_2, U$$

such that

$$M_2 P_2 M_1 P_1 A = U.$$

Note that you should choose the pivot which has the largest absolute value. Hint: The resulting U has only one fraction value.

(b) (5 pts) Following (a), what are the P and L such that $PA = LU$?

(c) (10 pts) Recall that in pivoted LU factorization, we choose the largest absolute value in the column and swap two rows. Here, we consider an extension of the pivoted LU factorization method that selects the largest absolute value in both row and column. If the largest absolute value occurs in the column, swap the two rows. Otherwise, swap the two columns. Thus, at each step we have PAQ, where

- one of P or Q is I.
- If P is not I, it is a permutation matrix to swap the two rows.
- If Q is not I, it is a permutation matrix to swap the two columns.

For example, consider the matrix

$$\begin{bmatrix} 1 & 3 & 9 \\ 5 & 20 & 10 \\ 8 & 2 & 5 \end{bmatrix}.$$

To choosing the first pivot, we compare the entries in the first row and first column, and select the element with the largest absolute value. The entry at position $(1, 3)$ has the largest absolute value among all entries in the first row and the first column. Therefore,

$$Q_1' = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \text{ and } P_1' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Follow the similar idea of (a). For the matrix A in (1), please find

$$P_1, Q_1, M_1, P_2, Q_2, M_2, \text{ and } U$$

so that

$$M_2 P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = U. \tag{2}$$

Note that you should choose the pivot which has the largest absolute value. If there are more than two candidates of pivot, we choose the one that is the closest to the diagonal part.

Hint: Note that the resulting

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & 1/6 \\ 0 & 0 & u_{33} \end{bmatrix}.$$

(d) (5 pts) Earlier, we were able to find P and L such that $PA = LU$ in (b). Now, consider the same matrix A in (1) and set $Q = Q_1 Q_2$. For the example in (c), can you find P such that (2) can be written as $PAQ = LU$?

(e) (5 pts) Following (d), solving the linear system

$$A\boldsymbol{x} = \boldsymbol{b}$$

is equivalent to solve

$$P^{-1} L U Q^{-1} \boldsymbol{x} = \boldsymbol{b}.$$

Therefore, we can get the solution $\boldsymbol{x}^*$ by the following steps.

(i) Solve $L\boldsymbol{y} = P\boldsymbol{b}$ to get $\boldsymbol{y}^*$, where $\boldsymbol{y}^* = UQ^{-1}\boldsymbol{x}$.

(ii) Solve $U\boldsymbol{z} = \boldsymbol{y}^*$ to get $\boldsymbol{z}^*$, where $\boldsymbol{z}^* = Q^{-1}\boldsymbol{x}$.

(iii) Solve $Q^{-1}\boldsymbol{x} = \boldsymbol{z}^*$ to get $\boldsymbol{x}^*$.

Now, we have

$$\boldsymbol{b} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Please solve

$$A\boldsymbol{x} = \boldsymbol{b}$$

by the aforementioned steps. Note that you can verify your solution $\boldsymbol{x}^*$ by checking whether $A\boldsymbol{x}^*$ is equal to $\boldsymbol{b}$.

Solution.

(a) Step 1: The pivot is 4, so we have

$$P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

such that

$$P_1 A = \begin{bmatrix} 4 & 5 & 20 \\ 2 & 3 & 12 \\ 1 & 2 & 5 \end{bmatrix}.$$

Then, we can calculate

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

so that

$$M_1 P_1 A = \begin{bmatrix} -4 & -5 & 20 \\ 0 & 1/2 & 2 \\ 0 & 3/4 & 0 \end{bmatrix}.$$

Step 2: The pivot is 3/4, so

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

and $P_2 M_1 P_1 A$ is

$$\begin{bmatrix} -4 & -5 & 20 \\ 0 & 3/4 & 0 \\ 0 & 1/2 & 2 \end{bmatrix}.$$

Thus, M_2 is calculated by

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

and

$$M_2 P_2 M_1 P_1 A = U = \begin{bmatrix} 4 & 5 & 20 \\ 0 & 3/4 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

(b) We have

$$P = P_2 P_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

and

$$L = P_2 M_1^{-1} P_2^{-1} M_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1/4 & 1 & 0 \\ 1/2 & 2/3 & 1 \end{bmatrix}$$

(c) Step 1: The pivot is 12, so

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } Q_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

such that

$$P_1 \cdot A \cdot Q_1 = \begin{bmatrix} 12 & 3 & 2 \\ 20 & 5 & 4 \\ 5 & 2 & 1 \end{bmatrix}.$$

Thereby, we can calculate

$$M_1 = \begin{bmatrix} 1 & 0 & 0 \\ -5/3 & 1 & 0 \\ -5/12 & 0 & 1 \end{bmatrix}$$

and

$$M_1 P_1 \cdot A \cdot Q_1 = \begin{bmatrix} 12 & 3 & 2 \\ 0 & 0 & 2/3 \\ 0 & 3/4 & 1/6 \end{bmatrix}.$$

Step 2: Now the pivot is 3/4, so

$$P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \text{ and } Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

such that

$$P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = \begin{bmatrix} 12 & 3 & 2 \\ 0 & 3/4 & 1/6 \\ 0 & 0 & 2/3 \end{bmatrix}.$$

Hence, we have

$$M_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and then

$$M_2 P_2 M_1 P_1 \cdot A \cdot Q_1 Q_2 = \begin{bmatrix} 12 & 3 & 2 \\ 0 & 3/4 & 1/6 \\ 0 & 0 & 2/3 \end{bmatrix}.$$

(d) We have

$$Q = Q_1 Q_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Since

$$M_2 P_2 M_1 P_1 A Q = U,$$

we can derive

$$AQ = (M_2 P_2 M_1 P_1)^{-1} U = \begin{bmatrix} 1 & 0 & 0 \\ 5/3 & 0 & 1 \\ 5/12 & 1 & 0 \end{bmatrix} U.$$

Now consider

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

Then we have

$$PAQ = P(M_2 P_2 M_1 P_1)^{-1} U = P \begin{bmatrix} 1 & 0 & 0 \\ 5/3 & 0 & 1 \\ 5/12 & 1 & 0 \end{bmatrix} U = \begin{bmatrix} 1 & 0 & 0 \\ 5/12 & 1 & 0 \\ 5/3 & 0 & 1 \end{bmatrix} U = LU.$$

That is,

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 5/12 & 1 & 0 \\ 5/3 & 0 & 1 \end{bmatrix}.$$

(e) In step (i), we have to solve

$$\begin{aligned} & L\boldsymbol{y} = P\boldsymbol{b} \\ \Rightarrow & \begin{bmatrix} 1 & 0 & 0 \\ 5/12 & 1 & 0 \\ 5/3 & 0 & 1 \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \boldsymbol{b} \\ \Rightarrow & \begin{bmatrix} 1 & 0 & 0 \\ 5/12 & 1 & 0 \\ 5/3 & 0 & 1 \end{bmatrix} \boldsymbol{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \end{aligned}$$

and the solution $\boldsymbol{y}^*$ can be calculated as

$$\boldsymbol{y}^* = \begin{bmatrix} 1 \\ 7/12 \\ -5/3 \end{bmatrix}.$$

In step (ii), we have to solve

$$\begin{aligned} & U\boldsymbol{z} = \boldsymbol{y}^* \\ \Rightarrow & \begin{bmatrix} 12 & 3 & 2 \\ 0 & 3/4 & 1/6 \\ 0 & 0 & 2/3 \end{bmatrix} \boldsymbol{z} = \begin{bmatrix} 1 \\ 7/12 \\ -5/3 \end{bmatrix}. \end{aligned}$$

The solution $\boldsymbol{z}^*$ can be calculated as

$$\boldsymbol{z}^* = \begin{bmatrix} 1/6 \\ 4/3 \\ -5/2 \end{bmatrix}.$$

In step (iii), we have to solve

$$Q^{-1}\boldsymbol{x} = \boldsymbol{z}^*,$$

which implies

$$\boldsymbol{x} = Q\boldsymbol{z}^* = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \boldsymbol{z}^*.$$

Thus,

$$\boldsymbol{x}^* = \begin{bmatrix} -5/2 \\ 4/3 \\ 1/6 \end{bmatrix}.$$

Verify:

$$A\boldsymbol{x}^* = \begin{bmatrix} 2 & 3 & 12 \\ 4 & 5 & 20 \\ 1 & 2 & 5 \end{bmatrix} \begin{bmatrix} -5/2 \\ 4/3 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \boldsymbol{b}.$$

Problem 3 (35 pts). TensorFloat-32 (TF32) is a reduced-precision floating-point format used for matrix multiplication on NVIDIA Tensor Cores. In TF32 computation, an input stored in IEEE 754 single precision (FP32) is **rounded before multiplication** so that it uses a normalized representation with:

- 1 sign bit,
- 8 exponent bits, and
- 10 significand bits.

The exponent range is the same as that of IEEE 754 single precision, so the exponent bias is 127. For normalized numbers, the value is interpreted as

$$(-1)^s \times 2^{E-127} \times (1.f)_2$$

where s is the sign bit, E is the unsigned exponent range, and f is the 10-bit significand range.

Assume throughout this problem that:

1. the input number is first represented exactly in FP32,
2. it is then rounded to TF32 using rounding even.

Answer the following questions.

(a) (5 pts) Round the FP32 number 347.6875 to TF32 and write the resulting:

- sign bit,
- exponent bits, and
- significand bits.

Show your work. (Given: $347 = 2^8 + 2^6 + 2^4 + 2^3 + 2^1 + 2^0$)

(b) (5 pts) What is the absolute rounding error introduced when converting 347.6875 from FP32 to TF32?

(c) (5 pts) Let x be a normalized FP32 number with exponent e, and let y be the result of rounding x to TF32 using rounding even. Derive an upper bound on the relative error

$$\frac{|y-x|}{|x|}$$

by considering the following steps.

Step 1: Find the maximum absolute error $|y-x|$.

Step 2: Find the minimum possible value of $|x|$.

Step 3: Combine the above results to have the upper bound.

(d) (10 pts) Consider the following values given in binary

$$a = 1.00000\ 00001\ 1_2, \qquad b = 1.00000\ 00000\ 1_2, \qquad c = 1.00000\ 00010\ 1101_2.$$

Let us follow the Tensor Core workflow to compute

$$d = c - a \times b$$

by the following steps.

Step 1: Round a and b to TF32.
Note: In Steps 2 and 3, a and b use their TF32-rounded values, while c remains unchanged.

Step 2: Compute the exact value of $a \times b$ in the binary number system.

Step 3: Compute the exact value of $c - a \times b$ in the binary number system with Step 2's result. Hint: the resulting $c - a \times b$ should be positive.

Step 4: Write the result of Step 3 into d in FP32 format.

Please show all intermediate steps in details and give the final result in FP32 format (i.e., the bit string of d).

(e) (10 pts) IEEE half precision (FP16) uses

- 1 sign bit,
- 5 exponent bits with bias 15,
- 10 significand bits.

Like FP32, both TF32 and FP16 reserve certain exponent values for special cases:

Exponent bits	Interpretation
all zeros	denormalized numbers and zero
all ones	infinity and NaN

Now, suppose that a **positive** value x cannot be represented as a normalized FP16 number, while it can be represented as a normalized TF32 number. Please give the possible range of x.

Solution.

(a) First, let us convert 347.6875 to binary. For the integer part, we have

$$347 = 256 + 64 + 16 + 8 + 2 + 1 = 2^8 + 2^6 + 2^4 + 2^3 + 2^1 + 2^0,$$

which implies that

$$347_{10} = 101011011_2.$$

For the fractional part, we have

$$0.6875 = 0.5 + 0.125 + 0.0625 = 2^{-1} + 2^{-3} + 2^{-4},$$

which implies that

$$0.6875_{10} = 0.1011_2.$$

Therefore,

$$347.6875_{10} = 101011011.1011_2.$$

Second, we do the normalization

$$101011011.1011_2 = 1.010110111011_2 \times 2^8.$$

Thus, we have

- sign bit: $s = 0$
- true exponent: $e = 8$
- stored exponent bits:
$$E = e + 127 = 8 + 127 = 135 = 10000111_2$$
- significand bits after the leading 1:
$$01011\ 01110\ 11_2$$

Finally, because TF32 keeps 10 significand bits, we have to round

$$01011\ 01110\ 11_2$$

by the required rules. The discarded bits begin with 1 (the 11th bit), and the remaining discarded bits are not all zero, so under the rounding even rule, we round up. Thus, the final 10-bit significand range becomes

$$01011\ 01111_2.$$

Overall, the TF32 representation is

$$\boxed{0\ 10000111\ 0101101111},$$

so the answer is

- sign bit: $\boxed{0}$
- exponent bits: $\boxed{10000111}$
- significand bits: $\boxed{0101101111}$

(b) The original value is

$$x = 1.010110111011_2 \times 2^8.$$

The rounded TF32 value is

$$y = 1.0101101111_2 \times 2^8.$$

so the difference between y and x is:

$$\begin{aligned} y - x &= 1.010110111100_2 \times 2^8 - 1.010110111011_2 \times 2^8 \\ &= (1.010110111100_2 - 1.010110111011_2) \times 2^8 \\ &= 0.000000000001_2 \times 2^8 \\ &= 2^{-12} \times 2^8 \\ &= 2^{-4} \\ &= 0.0625. \end{aligned}$$

(c) Let x be a normalized FP32 number, and y be the TF32-rounded number of x under rounding even. The relative error is defined as:

$$\left| \frac{y - x}{x} \right|$$

Since TF32 stores 10 significand bits, for a number with exponent e, the ulps is:

$$\text{ulps} = 2^{-10} \times 2^e$$

Under rounding even, the absolute rounding error $|y - x|$ is bounded by half of the ulps:

$$|y - x| \leq \frac{1}{2}\text{ulps} = 2^{-11} \times 2^e$$

The minimum possible magnitude for x in this exponent range is when all significand bits are zero:

$$|x|_{\min} = 1.0_2 \times 2^e = 2^e$$

Therefore, an upper bound on the relative rounding error is:

$$\left| \frac{y - x}{x} \right| \leq \frac{2^{-11} \times 2^e}{2^e} = 2^{-11}$$

(d) First, we round the inputs to TF32. For $a = 1.00000\ 00001\ 1_2$, the bits after the 10th significand bit are exactly halfway. The 10th bit is 1, so we round up (rounding even) as

$$a_{\text{TF32}} = 1.00000\ 00010_2.$$

For $b = 1.00000\ 00000\ 1_2$, this is also a tie case. However, the 10th bit is 0, so we round down as

$$b_{\text{TF32}} = 1.00000\ 00000_2$$

Second, we compute the product

$$a_{\text{TF32}} \times b_{\text{TF32}} = 1.00000\ 00010_2 \times 1.00000\ 00000_2 = 1.000000001_2.$$

Third, we subtract the above result from $c = 1.00000000101101_2$. The binary points for subtraction can be calculated by

$$\begin{array}{r} +1.00000000101101_2 \\ -1.00000000100000_2 \\ \hline 0.00000000001101_2 \end{array}$$

and be normalized to

$$0.00000000001101_2 = 1.101_2 \times 2^{-11}.$$

Finally, we convert this value to FP32 format:

- **Sign bit:** 0 (since the number is positive).
- **Exponent:** The true exponent is $e = -11$, so the stored exponent is

$$E = -11 + 127 = 116 = 01110\ 100_2.$$

- **significand bits:** The significand after the leading 1 is 101. Padding with zeros to fill 23 bits gives

$$10100\ 00000\ 00000\ 00000\ 000_2.$$

Therefore, the result in FP32 format is

0	01110 100	10100 00000 00000 00000 000

(e) For normalized FP16 numbers, by excluding the reserved exponent values, the exponent range satisfies

$$1 \le E \le 30$$

with bias 15. Therefore the true exponent range is

$$-14 \le e \le 15.$$

Hence, the smallest positive normalized FP16 number is

$$1.0_2 \times 2^{-14} = 2^{-14},$$

and the largest positive normalized FP16 number is

$$1.1111111111_2 \times 2^{15} = (2 - 2^{-10}) \times 2^{15}.$$

Thus, a positive number x cannot be represented as a normalized FP16 number if

$$0 < x < 2^{-14} \quad \text{or} \quad x > (2 - 2^{-10}) \times 2^{15}.$$

For normalized TF32 numbers, the exponent range satisfies

$$1 \le E \le 254,$$

with bias 127. Therefore the true exponent range is

$$-126 \le e \le 127.$$

Hence, the smallest positive normalized TF32 number is

$$1.0_2 \times 2^{-126} = 2^{-126},$$

and the largest positive normalized TF32 number is

$$1.1111111111_2 \times 2^{127} = (2 - 2^{-10}) \times 2^{127}.$$

Thus, a positive number that is not representable as a normalized FP16 value but is a valid normalized TF32 value falls in the following range:

$$\boxed{2^{-126} \leq x < 2^{-14} \quad \text{or} \quad (2 - 2^{-10}) \times 2^{15} < x \leq (2 - 2^{-10}) \times 2^{127}}.$$

Problem 4 (25 pts). Assume that we have a real, invertible, $n \times n$ matrix A. We say that the matrix A has a singular value decomposition (SVD) if there exist matrices U, Σ, V such that

$$A = U\Sigma V^T,$$

where

- U, V are $n \times n$ orthogonal matrices satisfying

$$U^T U = UU^T = I, V^T V = VV^T = I, \text{ and}$$

- Σ is an $n \times n$ diagonal matrix with $\Sigma_{ii} = \sigma_i > 0$.

(a) (5 pts) The Frobenius norm of a matrix is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2}.$$

Prove that

$$\|A\|_F = \sqrt{\text{trace}(A^T A)}. \tag{3}$$

Note that the definition of trace$(\cdot)$ is the sum of a square matrix's diagonal values, i.e.,

$$\text{trace}(B) = \sum_{i=1}^{n} B_{ii}, \text{where } B \text{ is a } n \times n \text{ matrix.}$$

(b) (10 pts) Prove that

$$\|A\|_F = \sqrt{\sum_{i=1}^{n} \sigma_i^2}.$$

Hint: You may directly use the derived result in (3) and consider the SVD formulation.

(c) (10 pts) If $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, show that

$$\|\boldsymbol{u}\boldsymbol{v}^T\|_2 = \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2 = \|\boldsymbol{u}\boldsymbol{v}^T\|_F,$$

where $\|\cdot\|_2$ is the 2-norm.

Solution.

(a) Let us consider the sth diagonal entry of A^TA:

$$(A^TA)_{ss} = \sum_{t=1}^{n} A_{ts}A_{ts} = \sum_{t=1}^{n} A_{ts}^2.$$

Hence, we can derive the trace of A^TA as

$$\text{trace}(A^TA) = \sum_{s=1}^{n}(A^TA)_{ss} = \sum_{s=1}^{n}\sum_{t=1}^{n} A_{ts}^2. \tag{4}$$

By taking

$$t = i \text{ and } s = j,$$

we can have

$$(4) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2 = \|A\|_F^2 \geq 0.$$

Therefore, we have

$$\sqrt{\text{trace}(A^TA)} = \|A\|_F.$$

(b) By the derived result in (a), we proved that

$$\|A\|_F = \sqrt{\text{trace}(A^TA)}$$

With the SVD of A, we can derive that

$$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^TU^TU\Sigma V^T = V\Sigma\Sigma V^T$$

Hence

$$\text{trace}(A^TA) = \sum_{i=1}^{n}(A^TA)_{ii} = \sum_{i=1}^{n}\sum_{k=1}^{n} V_{ik}(\Sigma\Sigma V^T)_{ki} = \sum_{i=1}^{n}\sum_{k=1}^{n} V_{ik}\Sigma_{kk}^2V_{ki}^T = \sum_{k=1}^{n}\Sigma_{kk}^2\sum_{i=1}^{n} V_{ik}^2$$

Since V is an orthogonal matrix, its columns are orthonormal, which implies:

$$\sum_{i=1}^{n} V_{ik}^2 = 1,$$

Therefore

$$\|A\|_F = \sqrt{\sum_{k=1}^{n}\sum_{i=1}^{n}\Sigma_{kk}^2V_{ik}^2} = \sqrt{\sum_{k=1}^{n}\Sigma_{kk}^2} = \sqrt{\sum_{k=1}^{n}\sigma_k^2}.$$

(c) Let us define

$$A = \boldsymbol{u}\boldsymbol{v}^T$$

and separate the proof into two parts.

- **The First Equality.** By the definition of 2-norm, we have

$$\|A\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \max_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2.$$

  Substituting $A = \boldsymbol{u}\boldsymbol{v}^T$ into the above expression

$$\max_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{x}\|_2 = \max_{\|\boldsymbol{x}\|_2=1} |\boldsymbol{v}^T\boldsymbol{x}|\|\boldsymbol{u}\|_2 \tag{5}$$

  By using the Cauchy-Schwarz inequality, we have

$$|\boldsymbol{v}^T\boldsymbol{x}| \leq \|\boldsymbol{v}\|_2\|\boldsymbol{x}\|_2.$$

  Equality is attained by choosing $\boldsymbol{x} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}$ when $\boldsymbol{v} \neq 0$:

$$|\boldsymbol{v}^T\boldsymbol{x}| = \|\boldsymbol{v}\|_2\|\boldsymbol{x}\|_2$$

  So we can derive

$$(5) = \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$$

  If $\boldsymbol{v} = 0$, then clearly

$$\|\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{x}\|_2 = 0 = \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$$

  Therefore, we have done the first equality.

- **The Second Equality.** By the definition of Frobenius norm, we have

$$\|\boldsymbol{u}\boldsymbol{v}^T\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(\boldsymbol{u}\boldsymbol{v}^T)_{ij}^2}. \tag{6}$$

  Since $A_{ij} = (\boldsymbol{u}\boldsymbol{v}^T)_{ij} = u_i v_j$, we can obtain

$$(6) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} u_i^2 v_j^2} = \sqrt{\sum_{i=1}^{n} u_i^2 \sum_{j=1}^{n} v_j^2} = \sqrt{\|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2} = \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2$$

  Therefore, we have done the second equality.

Overall, we complete the proof of

$$\|\boldsymbol{u}\boldsymbol{v}^T\|_2 = \|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 = \|\boldsymbol{u}\boldsymbol{v}^T\|_F.$$