Problem 1 (10%)

Give the single binary representation of the number 5.3. We follow the IEEE standard and use rounding even.

Solution

Because

\[ 5 = (101)_2, \quad 0.3 = (0.01001)_2, \]

we can know that

\[ 5.3 = 101.01001 = 1.0101001 \times 2^2. \]

- The sign bit: 0.
- The exponent part: \( 2 + 127 = 129 \Rightarrow 10000001. \)
- The mantissa part: 010, 1001, 1001, 1001, 1010 by using rounding even.
Problem 2 (55%)

Assume we want to use one byte for storing a floating-point number. We consider

- 1 sign bit
- 4 bits for significand
- 3 bits for exponents
- normalized representation
- biased exponents
- largest/smallest exponents reserved for special quantities and/or denormalized numbers

Other settings are the same as IEEE standard. Answer the following questions.

(a) (5%) Range of valid exponents.

(b) (5%) Largest and smallest positive floating-point number.

(c) (10%) Largest and smallest positive denormalized number.

(d) (5%) For the number considered in problem 1, what is the binary presentation?

(e) (5%) The machine epsilon.

(f) (15%) Consider rounding even. What is the largest relative error in subtracting two positive values? Assume no guard digit and no denormalized numbers. You must rigorously explain why your obtained answer is the largest.

(g) (10%) Assume exactly rounded operation with rounding even. Can you give an example where without denormalized number

\[ x \neq y \text{ but } x - y = 0 \]

and with denormalized number

\[ x - y \neq 0. \]

Note that with/without denormalized numbers we assume that the range of exponents is the same as that in (a).
Solution

(a) $-2$ to $3$.

(b) Largest: $1.1111 \times 2^3$.
     Smallest: $1.0000 \times 2^{-2}$.

(c) Largest: $0.1111 \times 2^{-2}$.
     Smallest: $0.0001 \times 2^{-2}$.

(d) From problem 1, we know that

\[ 5.3 = 1.0101001 \times 2^2. \]

- The sign bit: 0.
- The exponent part: $2 + 3 = 5 \Rightarrow 101$.
- The mantissa part: 0101 by using rounding even.

(e) In binary system, we have $\beta = 2$.
Because we consider normalized representation $\Rightarrow p = 5$.
Therefore, the machine epsilon

\[ \epsilon = \beta^{-p+1}/2 \]
\[ = 2^{-5}. \]

Common mistake:

Many wrongly think $p = 4$.

(f) The relative error is

\[ \frac{|(x \oplus y) - (x - y)|}{|(x - y)|}. \]

Without loss of generality, we assume

\[ x > y \text{ and } x = 1.xxxx. \]

Assume $y$ is rounded to $\hat{y}$. The error is

\[ |(x - \hat{y}) - (x - y)| = |y - \hat{y}| \leq 0.00001. \]
For $|x - y|$, the smallest possible value is by making $x$ and $y$ as close as possible. Because $x = 1.xxxx$, the smallest $x - y$ is

$$0.00001.$$  

We show that

$$x = 1.0000 \text{ and } y = 0.11111$$

satisfy the above bound. We have

$$\hat{y} = 1.0000$$

by rounding even. Thus

$$x \oplus y = 0.$$  

The relative error is

$$\frac{|0 - 2^{-5}|}{|2^{-5}|} = 1.$$  

**Common mistake:**

1. Some use the theorem in slides to show that the error can be “as large as” 1. However, this does not mean that you already have the largest error.

2. If your description is not clearly written, you do not get all the points.

(g) Consider

$$x = 1.0000 \text{ and } y = 0.11111.$$  

Clearly,

$$x \neq y.$$  

The exact result is

$$x - y = 0.00001 < 1.0000 \times 2^{-2}.$$  

Without denormalized numbers, we have

$$x \oplus y = 0.$$  

With denormalized numbers, we have

$$x \oplus y = 0.00001 = 0.0010 \times 2^{-2}.$$  

4
Problem 3 (15%)  

For the floating-point system considered in problem 2, if rounding up is considered, is there an example so that \( x_n \) continually increases to overflow when we do  
\[
  x_0 = x, \quad x_1 = (x_0 \ominus y) \oplus y, \ldots, \quad x_n = (x_{n-1} \ominus y) \oplus y.
\]

Assume \( \ominus \) and \( \oplus \) are exactly rounded using rounding up.

Solution

Yes. Consider \( x = 10002 = 1.0000 \times 2^3 \), \( y = -0.012 = -1.0 \times 2^{-2} \), \( x_1 \) is calculated as follows:

\[
  x - y = 1000.01, \quad x \ominus y = 1000.1, \\
  (x \ominus y) + y = 1000.01, \quad x_1 = (x \ominus y) \oplus y = 1000.1.
\]

And \( x_2 \):

\[
  x_1 - y = 1000.11, \quad x_1 \ominus y = 1001.0, \\
  (x_1 \ominus y) + y = 1000.11, \quad x_2 = (x_1 \ominus y) \oplus y = 1001.0.
\]

Clearly the value is increased by 0.1 until overflow.

Problem 4 (20%)  

Consider the following matrix

\[
\begin{bmatrix}
-4 & 7 & -3 & 8 \\
6 & -3 & 3 & -5 \\
12 & 6 & -6 & -6 \\
-4 & 16 & -4 & 14
\end{bmatrix}
\]

Conduct LU factorization with pivoting.

(a) (10%) Give \( P_1, M_1, P_2, M_2, P_3, M_3 \) such that

\[
M_3P_3M_2P_2M_1P_1A = U.
\]

(b) (10%) What are \( P, L, U \) such that

\[
PA = LU?
\]
Solution

Because with pivoting, we first need to calculate

\[
\begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
-4 & 7 & -3 & 8 \\
6 & -3 & 3 & -5 \\
12 & 6 & -6 & -6 \\
-4 & 16 & -4 & 14
\end{bmatrix}
= \begin{bmatrix}
12 & 6 & -6 & -6 \\
6 & -3 & 3 & -5 \\
-4 & 7 & -3 & 8 \\
-4 & 16 & -4 & 14
\end{bmatrix}.
\]

Then,

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
-1/2 & 1 & 0 & 0 \\
1/3 & 0 & 1 & 0 \\
1/3 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
12 & 6 & -6 & -6 \\
6 & -3 & 3 & -5 \\
-4 & 7 & -3 & 8 \\
-4 & 16 & -4 & 14
\end{bmatrix}
= \begin{bmatrix}
12 & 6 & -6 & -6 \\
0 & -6 & 6 & -2 \\
0 & 9 & -5 & 6 \\
0 & 9 & -5 & 6
\end{bmatrix}.
\]

Similarly, we can derive

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
12 & 6 & -6 & -6 \\
0 & -6 & 6 & -2 \\
0 & 9 & -5 & 6 \\
0 & 18 & -6 & 12
\end{bmatrix}
= \begin{bmatrix}
12 & 6 & -6 & -6 \\
0 & 18 & -6 & 12 \\
0 & 9 & -5 & 6 \\
0 & -6 & 6 & 2
\end{bmatrix}.
\]

In the end, we have

\[
\begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
-4 & 7 & -3 & 8 \\
6 & -3 & 3 & -5 \\
12 & 6 & -6 & -6 \\
-4 & 16 & -4 & 14
\end{bmatrix}
= \begin{bmatrix}
1 & 0 & 0 & 0 \\
1/3 & 1 & 0 & 0 \\
1/2 & -1/3 & 1 & 0 \\
-1/3 & 1/2 & -1/2 & 1
\end{bmatrix}
\begin{bmatrix}
12 & 6 & -6 & -6 \\
0 & 18 & -6 & 12 \\
0 & 0 & 4 & 2 \\
0 & 0 & 0 & 1
\end{bmatrix}.
\]
\[ \begin{align*} 
P_1 &= \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\
M_1 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 \\ 1/3 & 0 & 1 & 0 \\ 1/3 & 0 & 0 & 1 \end{bmatrix} \\
P_2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \\
M_2 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 \\ 0 & 1/3 & 0 & 1 \end{bmatrix} \\
P_3 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \\
M_3 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1 \end{bmatrix} \\
\end{align*} \]

(b)

\[ P = P_3P_2P_1 \]
\[ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1/2 & 1 \end{bmatrix} \]

\[ = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \]
From (1), we know

\[ L = \begin{bmatrix}
1 & 0 & 0 & 0 \\
-1/3 & 1 & 0 & 0 \\
1/2 & -1/3 & 1 & 0 \\
-1/3 & 1/2 & -1/2 & 1
\end{bmatrix} \]

and

\[ U = \begin{bmatrix}
12 & 6 & -6 & -6 \\
0 & 18 & -6 & 12 \\
0 & 0 & 4 & 2 \\
0 & 0 & 0 & 1
\end{bmatrix} \]