Roll No: EE18B067 Name:Abhishek Sekar

Rollnumber: EE18B067

Date: August 16, 2021

1. [Representation of Real Numbers on the computer] Real numbers are represented on the computer as follows:



where

- S is the sign bit with it being 0 for positive numbers and 1 for negative numbers.
- A,B,C are the first three significant digits in the decimal expansion with the decimal point occurring between A and B.
- All digits after the third significant digit are chopped off.
- E is the exponent in base 10 with a bias of 5.
- Number Representations

Number	S	Α	В	С	E
+ 0	0	0	Anything	Anything	Anything
- 0	1	0	Anything	Anything	Anything
$+\infty$	0	9	9	9	9
- ∞	1	9	9	9	9
NaN	Not 0 or 1	Anything	Anything	Anything	Anything

(a) How many non-zero Floating Point numbers (FPN) can be represented by our machine?

Solution

We can try solving this by considering the number of choices we have for the values S,A,B,C and E which will result in a non-zero floating point number.

For a non-zero floating point number,

- S can take one of a total of 2 permissible values, namely $\{0,1\}$. S taking any other value is used to represent NaN which is not a floating point number.
- A can take one of a total of 9 permissible values, namely $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. A is not permitted to take the value 0, as that is a requirement to represent zero on the machine.
- **B**,**C** and **E** can take one of a total of 10 permissible values, namely {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} as there is no restriction on them when it comes to representing zero on the machine thereby permitting all possible choices.

Employing the fundamental principle of counting, the total number of ways in which we can represent a non-zero floating point number is the product of the number of ways in which each individual value S,A,B,C,E can be represented for the given purpose.

Therefore, the number of non-zero Floating Point Numbers that can be represented on the machine is $2x9x10x10x10 = 18x10^3$ or 18000 numbers of which 17998 are normal floating point numbers and the other two are $+\infty$ and $-\infty$.

- (b) How many FPN are in the following intervals?
 - i. (9, 10)

Solution:

We can try employing the same approach as in the previous subdivision to find the total number of FPN.

(9, 10):

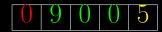
• Representation of the lower bound 9:

To find the machine representation of 9, let us express it in scientific notation with 3 significant digits (as that is the extent to which the machine can represent the number).

Rollnumber: EE18B067

 $9 = 9.00 \times 10^{0}$

From this we can see that 9 is represented in the below form.



- S is 0 as 9 is positive.
- A is 9 as it is the first digit(digit before the decimal point) in scientific notation.
- B and C are 0 as they're the first two significant digits that trail the decimal point.
- E is 5 as the exponent in scientific notation is 0. Adding the bias 5 to the exponent value, we get the value of E to be 5.

• Representation of the upper bound 10:

Expressing 10 in scientific notation to three significant digits, we get,

 $10 = 1.00 \times 10^{1}$

From this we can see that 10 is represented in the below form.



- S is 0 as 10 is positive.
- A is 1 as it is the first digit(digit before the decimal point) in scientific notation.
- B and C are 0 as they're the first two significant digits that trail the decimal point.
- E is 6 as the exponent in scientific notation is 1. Adding the bias 5 to the exponent value, we get the value of E to be 6.

• Identifying restrictions on S,A,B,C,E through the above two representations

- As both 9 and 10 are positive, any number between them is also positive therefore, S remains 0. Therefore, the possible values of S includes only 0.
- As the interval is an open one, meaning 9 and 10 are not included, the term A necessarily has to take the value 9 as anything else would not represent floating point numbers lying in the interval as the largest FPN below 10 starts with the digit 9 and the smallest FPN above 9 also starts with the digit 9. Therefore, the possible values of A includes only 9.
- B and C can take any value with the sole exception of both of them being 0, ie, $\{B,C\} = \{0,0\}$ as that would represent 9 which doesn't lie inside the interval. Therefore, the possible values of B,C include those from $\{0,1,2,3,4,5,6,7,8,9\}$ obeying the constraint $\{B,C\} \neq \{0,0\}$.
- E necessarily has to take the value 5 as the E term remains constant throughout the interval and only changes at the endpoint 10 which is not part of the interval. Therefore, the possible values of E includes only 5.

• Computing the answer

Therefore by the fundamental principle of counting, the total number of floating point numbers lying between 9 and 10 are, 1x1x(10x10-1)x1 = 99 FPN. (10x10 - 1)x1 = 99 F

ii. (10, 11)

Solution:

(10, 11):

• Representation of the lower bound 10:

Expressing 10 in scientific notation to three significant digits, we get,

 $10 = 1.00 \times 10^{1}$

From this we can see that 10 is represented in the below form.



Rollnumber: EE18B067

- S is 0 as 10 is positive.
- A is 1 as it is the first digit(digit before the decimal point) in scientific notation.
- B and C are 0 as they're the first two significant digits that trail the decimal point.
- E is 6 as the exponent in scientific notation is 1. Adding the bias 5 to the exponent value, we get the value of E to be 6.

• Representation of the upper bound 11:

Expressing 10 in scientific notation to three significant digits, we get,

 $11 = 1.10 \times 10^{1}$

From this we can see that 11 is represented in the below form.



- S is 0 as 11 is positive.
- A is 1 as it is the first digit (digit before the decimal point) in scientific notation.
- B is 1 and C is 0 as they're the first two significant digits that trail the decimal point.
- E is 6 as the exponent in scientific notation is 1. Adding the bias 5 to the exponent value, we get the value of E to be 6.

• Identifying restrictions on S,A,B,C,E through the above two representations

- As both 10 and 11 are positive, any number between them is also positive therefore, S remains 0.
 Therefore, the possible values of S includes only 0.
- As the interval is an open one, the term A necessarily has to take the value 1 as anything else would not represent floating point numbers lying in the interval as the largest FPN below 11 starts with the digit 1 and the smallest FPN above 10 also starts with the digit 1. Therefore, the possible values of A includes only 1.
- B necessarily has to take the value 0 due to the same reason above for A applied to the second significant digit. C however can take any value besides 0 as that would represent 10 which isn't part of the interval. Therefore, the possible values of B includes only 0 and that of C is {1, 2, 3, 4, 5, 6, 7, 8, 9}.
- E necessarily has to take the value 6 as the E term remains constant throughout the interval. Therefore, the possible values of E includes only 5.

• Computing the answer

Therefore by the fundamental principle of counting, the total number of floating point numbers lying between 10 and 11 are, 1x1x1x9x1 = 9 FPN.

iii. (0, 1)

Solution:

(0, 1):

• Representation of the lower bound 0:

From the table, we can see that +0 (as the interval is between 0 and 1) is represented in the below form.

Where X indicates that any digit can be put in that place.

• Representation of the upper bound 1:

Expressing 1 in scientific notation to three significant digits, we get,

 $1 = 1.00 \times 10^0$

From this we can see that 1 is represented in the below form.



- S is 0 as 1 is positive.
- A is 1 as it is the first digit(digit before the decimal point) in scientific notation.
- B and C are 0 as they're the first two significant digits that trail the decimal point.
- E is 5 as the exponent in scientific notation is 0. Adding the bias 5 to the exponent value, we get the value of E to be 5.

• Identifying restrictions on S,A,B,C,E through the above two representations

- As we're looking at the interval between 0 and 1 which comprises only positive numbers, S remains 0. Therefore, the possible values of S includes only 0.
- As the interval is an open one, the term A cannot take the value 0 as 0 is an endpoint of the interval. A can take any other value since with an appropriate exponent term, the number will remain in the interval. Therefore, the possible values of A are $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$.
- B and C can take any value whatsoever as there is no inherent constraint on them here as in the previous parts. For any value of B and C, with an appropriate A and E value, the number will remain in the interval. Therefore, the possible values of B,C are {0,1,2,3,4,5,6,7,8,9}.
- E has to take values less than 5. If E takes the value 5, it automatically represents a number lying outside the interval. This is because A is the digit appearing before the decimal point in the scientific notation, and for any permitted value of A, the FPN will be greater than or equal to 1 for an E value of 5 or higher. Additionally, the values of E less than 5 belong to numbers with negative exponents which lie in the given interval. Therefore, the possible values of E are {0,1,2,3,4}.

Computing the answer

Therefore by the fundamental principle of counting, the total number of floating point numbers lying between 0 and 1 are, 1x9x10x10x5 = 4500 FPN.

(c) Identify the smallest positive and largest positive FPN on the machine.

Solution:

Smallest positive FPN:

For the smallest positive FPN, we would like the following:

- The number should be one with the largest negative exponent in order to be as small as possible. Therefore, we have E=0, thereby the actual exponent being -5 (subtracting the bias).
- A,B,C have to take the values 1,0,0 respectively as for a given exponent E, 1.00×10^{E} is smaller than any other number with the same exponent.
- S has to take the value 0 as the number is positive.

With these, the representation of the smallest positive FPN on the machine is being 1.00×10^{-5} .

s 0 1 0 0 with its value

Rollnumber: EE18B067

Largest positive FPN:

If the question is referring to the largest positive FPN, then it is $+\infty$.

However in the case it is referring to the largest positive normal FPN, meaning finite floating point number we should

tread as we did for the smallest positive FPN.

For the largest positive normal FPN, we would like the following:

- The number should be one with the largest positive exponent in order to be as large as possible. Therefore, we have E = 9, thereby the actual exponent being +4 (subtracting the bias).
- A,B,C have to take the values 9,9,8 respectively as for a given exponent E, 9.98×10^E is larger than any other number with the same exponent (considering only 3 significant digits) with the sole exception being 9.99×10^E . However, the exception requires A,B,C,E = 9 which happens to be the representation of ∞ in our machine and hence this combination is not permitted to represent a normal FPN.
- S has to take the value 0 as the number is positive.

With these, the representation of the largest positive FPN on the machine is 0 99819 with its value being 9.98×10^4 .

(d) Identify the machine precision.

Solution:

Machine precision ϵ_m is defined as the difference between the number 1 and the smallest number exceeding 1 represented on the machine.

Referring to a previous subdivision, we can see that 1 is represented as $\begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}$ in the machine.

We are now required to find the next number greater than 1 which is representable on the machine. For this, we should keep the values of S,A,B and E intact and increase C by the smallest possible step. This is because, changing S,A or E will make the number drastically larger or smaller than 1 which defeats our purpose. Since C represents the number appearing after B in the decimal point, it is more sensitive to change when compared to B.

 1.01×10^{0} .

0 1 0 1 5 with a value of

Rollnumber: EE18B067

Therefore, employing the definition of machine precision, $\epsilon_m = 1.01 - 1 = 10^{-2}$

Therefore, the number just greater than 1 representable on the machine is

(e) What is the smallest positive integer not representable exactly on this machine?

Solution

Using the fact that any number with more than three significant digits gets stripped down to its first three significant digits in our machine, logically, the smallest positive integer not representable exactly would be the smallest one with more than 3 significant digits.

This number would be 1001 as in scientific notation, it is expressed as 1.001×10^3 , but in our machine this gets chopped off to 1.00×10^3 and hence is not representable exactly.

Any positive integer below 1001 can be expressed accurately using just three significant digits and is therefore representable exactly on the machine.

(f) Consider solving the following recurrence on our machine:

$$a_{n+1} = 5a_n - 4a_{n-1}$$

with $a_1 = a_2 = 2.932$. Compute a_n for $n \in \{3, 4, 5, 6, 7\}$ on our machine.

- a_1, a_2 would be chopped to three significant digits to begin with.
- At each step in the recurrence, $5a_n$ and $4a_{n-1}$ would be chopped down to the first three significant digits before the subtraction is performed.

Solution:

• Initialization:

 $a_1, a_2 = 2.932$. As this has 4 significant digits, in our machine, a_1, a_2 will be chopped down to 3 significant digits

and represented as 0 2935 with a value of 2.93×10^{0} .

• n = 2:

- Multiplication

As n = 2, a_n is a_2 represented as 2.93×10^0 on the machine. Therefore the representation of $5 \times a_n$ will be $5 \times 2.93 \times 10^0 = 1.465 \times 10^1$ chopped to three significant digits which is 1.46×10^1 .

Similarly doing the same for the a_{n-1} term a_1 , we have, $4xa_1 4x2.93x10^0 = 1.172x10^1$ represented as $1.17x10^1$ on the machine.

- Subtraction

Subtracting the machine representation of $4a_1$ from that of $5a_2$ we get 2.90×10^0 which can be accurately represented in the machine as it has three significant digits.

- Update

Therefore, a_3 will be represented as $0 \mid 2 \mid 9 \mid 0 \mid 5$



with a value of 2.90×10^{0} .

Rollnumber: EE18B067

• n = 3:

- Multiplication

As n = 3, a_n is a_3 . Therefore the representation of $5xa_n$ will be $5x2.90x10^0 = 1.45x10^1$ which already has 3 significant digits and therefore no chopping takes place.

Similarly doing the same for the a_{n-1} term a_2 , we have, $4xa_2 4x2.93x10^0 = 1.172x10^1$ represented as $1.17x10^1$ on the machine.

- Subtraction

Subtracting the machine representation of $4a_2$ from that of $5a_3$ we get 2.80×10^0 which can be accurately represented in the machine as it has three significant digits.

- Update

Therefore, a_4 will be represented as 0 2 8 0 5 with a value of 2.80×10^0 .

• n = 4:

- Multiplication

As n = 4, a_n is a_4 . Therefore the representation of $5xa_n$ will be $5x2.80x10^0 = 1.40x10^1$ which already has 3 significant digits and therefore no chopping takes place.

Similarly doing the same for the a_{n-1} term a_3 , we have, $4xa_3 4x2.90x10^0 = 1.16x10^1$ represented as the same on the machine.

- Subtraction

Subtracting the machine representation of $4a_3$ from that of $5a_4$ we get 2.40×10^0 which can be accurately represented in the machine as it has three significant digits.

- Update

Therefore, a_5 will be represented as 0 2 4 0 5 with a value of 2.40×10^0 .

• n = 5:

- Multiplication

As n = 5, a_n is a_5 . Therefore the representation of $5xa_n$ will be $5x2.40x10^0 = 1.20x10^1$ which already has 3 significant digits and therefore no chopping takes place.

Similarly doing the same for the a_{n-1} term a_4 , we have, $4xa_4 4x2.80x10^0 = 1.12x10^1$ represented as the same on the machine.

- Subtraction

Subtracting the machine representation of $4a_4$ from that of $5a_5$ we get 8.00×10^{-1} which can be accurately represented in the machine as it has three significant digits.

- Update

Therefore, a_6 will be represented as



with a value of 8.00×10^{-1} .

• n = 6:

- Multiplication

As n = 6, a_n is a_6 . Therefore the representation of $5xa_n$ will be $5x8.00x10^{-1} = 4.00x10^0$ which already has 3 significant digits and therefore no chopping takes place.

Similarly doing the same for the a_{n-1} term a_5 , we have, $4xa_5 4x2.40x10^0 = 9.60x10^0$ represented as the same on the machine.

- Subtraction

Subtracting the machine representation of $4a_5$ from that of $5a_6$ we get -5.60×10^0 which can be accurately represented in the machine as it has three significant digits.

Update

Therefore, a_7 will be represented as

|--|--|

with a value of -5.60×10^0 .

Rollnumber: EE18B067

2. [Recurrence relation of an Integral]

Consider the following integral: $I_n = \int_0^1 x^{2n} \sin(\pi x) dx$

(a) Obtain a recurrence for I_n in terms of I_{n-1} .

Solution:

Observe that the integral $\int_0^1 x^{2n} \sin(\pi x) dx$ can be also viewed as $\int_0^1 x^{2n} d\left(\frac{-\cos(\pi x)}{\pi}\right)$ as $d\left(\frac{-\cos(\pi x)}{\pi}\right) = \sin(\pi x) dx$. We can use this relationship to integrate the integral by parts.

$$\int_{a}^{b} u dv = uv|_{a}^{b} - \int_{a}^{b} v du \tag{1}$$

Using 1 on our integral, we have,

$$\int_0^1 x^{2n} \sin(\pi x) dx = \int_0^1 x^{2n} d\left(\frac{-\cos(\pi x)}{\pi}\right)$$

$$\Rightarrow x^{2n} \left(\frac{-\cos(\pi x)}{\pi}\right) \Big|_0^1 - \int_0^1 \left(\frac{-\cos(\pi x)}{\pi}\right) d\left(x^{2n}\right)$$

$$\Rightarrow x^{2n} \left(\frac{-\cos(\pi x)}{\pi}\right) \Big|_0^1 - \int_0^1 \left(\frac{-\cos(\pi x)}{\pi}\right) 2n\left(x^{2n-1}\right) dx$$

Now employing 1 on the reduced integral obtained above, we have,

$$\int_{0}^{1} \left(\frac{\cos\left(\pi x\right)}{\pi}\right) 2n\left(x^{2n-1}\right) dx = \int_{0}^{1} 2n\left(x^{2n-1}\right) d\left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) \left(\operatorname{As} d\left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) = \left(\frac{\cos\left(\pi x\right)}{\pi}\right) dx\right)$$

$$\Rightarrow 2n\left(x^{2n-1}\right) \left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) \Big|_{0}^{1} - \int_{0}^{1} \left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) d2n\left(x^{2n-1}\right)$$

$$\Rightarrow 2n\left(x^{2n-1}\right) \left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) \Big|_{0}^{1} - \int_{0}^{1} \left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) (2n) \cdot (2n-1) \left(x^{2n-2}\right) dx$$

$$\Rightarrow 2n\left(x^{2n-1}\right) \left(\frac{\sin\left(\pi x\right)}{\pi^{2}}\right) \Big|_{0}^{1} - \frac{2n \cdot 2n - 1}{\pi^{2}} I_{n-1}$$

Plugging this in the larger expression we had earlier obtained, we get,

$$\begin{split} I_n &= x^{2n} \left(\frac{-\cos\left(\pi x\right)}{\pi} \right) \bigg|_0^1 + 2n \left(x^{2n-1} \right) \left(\frac{\sin\left(\pi x\right)}{\pi^2} \right) \bigg|_0^1 - \frac{2n \cdot 2n - 1}{\pi^2} I_{n-1} \\ &\Rightarrow \left(\frac{1}{\pi} - 0 \right) + (0 - 0) - \frac{2n \cdot 2n - 1}{\pi^2} I_{n-1} \text{ (Computing the expressions within the limits)} \end{split}$$

Which gives us the recurrence,

$$I_n = \frac{1}{\pi} - \frac{(2n) \cdot (2n-1)}{\pi^2} I_{n-1} \tag{2}$$

Rollnumber: EE18B067

(b) Evaluate I_0 by hand.

Solution:

$$I_{0} = \int_{0}^{1} \sin(\pi x) dx$$
 (Substituting n = 0 in I_{n})
$$\Rightarrow \frac{\pi}{\pi} \int_{0}^{1} \sin(\pi x) dx$$
 (Multiplying and dividing by π)
$$\Rightarrow \frac{1}{\pi} \int_{0}^{1} \sin(\pi x) d\pi x$$

$$\Rightarrow \frac{-1}{\pi} \cos(\pi x) \Big|_{0}^{1}$$
 (As $\int \sin(\pi x) d\pi x = -\cos(\pi x)$)
$$\Rightarrow \frac{1}{\pi} (\cos(0) - \cos(\pi))$$
 (Evaluating within the limits)
$$\Rightarrow \frac{2}{\pi}$$

Therefore $I_0 = \frac{2}{\pi}$

(c) Use the recurrence to obtain I_n for $n \in \{1, 2, ..., 15\}$ in MATLAB/Python.

Solution:

The recurrence was coded in python and the resulting output was rounded off to 7 decimal places to make it easier for us to compare it with the wolfram alpha output. The table below shows the values obtained.

Values of the integral obtained using the recurrence relation

I_0	$6.36619800*10^{-1}$
I_1	$1.89303700 * 10^{-1}$
I_2	$8.81441000*10^{-2}$
I_3	$5.03839000*10^{-2}$
I_4	$3.24325000*10^{-2}$
I_5	$2.25607000*10^{-2}$
I_6	$1.65740000*10^{-2}$
I_7	$1.26785000*10^{-2}$
I_8	$1.00056000*10^{-2}$
I_9	$8.09380000*10^{-3}$
I_{10}	$6.68030000*10^{-3}$
I_{11}	$5.60460000*10^{-3}$
I_{12}	$4.85120000*10^{-3}$
I_{13}	$-1.18600000 * 10^{-3}$
I_{14}	$4.09155000*10^{-1}$
I_{15}	$-3.57484695*10^{+1}$

(d) Use Wolfram Alpha to obtain I_n by directly performing the integral for $n \in \{1, 2, ..., 15\}$.

Solutions

The approximate values of the integral obtained through Wolfram Alpha is given here as the original expressions involving

 π are rather very large to be expressed here.

Values of the integral obtained using Wolfram Alpha

Rollnumber: EE18B067

I_0	$6.3662 * 10^{-1}$
I_1	$1.8930*10^{-1}$
I_2	$8.8144 * 10^{-2}$
I_3	$5.0384 * 10^{-2}$
I_4	$3.2433 * 10^{-2}$
I_5	$2.2561 * 10^{-2}$
I_6	$1.6574 * 10^{-2}$
I_7	$1.2679 * 10^{-2}$
I_8	$1.0006 * 10^{-2}$
I_9	$8.0938 * 10^{-3}$
I_{10}	$6.6802 * 10^{-3}$
I_{11}	$5.6060*10^{-3}$
I_{12}	$4.7708 * 10^{-3}$
I_{13}	$4.1089 * 10^{-3}$
I_{14}	$3.5754 * 10^{-3}$
I_{15}	$3.1393 * 10^{-3}$

(e) Explain your observations.

Solution:

Comparing the values obtained directly from Wolfram Alpha and those obtained from python using the recurrence relation and tabulating the absolute difference across terms(rounded to 5 decimal places) where it is non-zero, we have,

The absolute differences between the values

I_{12}	$8.000000 * 10^{-5}$
I_{13}	$5.290000 * 10^{-3}$
I_{14}	$4.055800*10^{-1}$
I_{15}	$3.575161 * 10^{+1}$

- We can see that the integral values computed on python using the recurrence relation differ a lot from the wolfram alpha values as n grows larger and larger.
- This disparity is observed since every integral involves a term incorporating π which is an irrational number.
- π being irrational is not accurately representable on the machine. This causes the recurrence relation to produce erroneous values where the error gets larger as n grows larger.
- This inaccurate representation even leads to results that even contradict logic. For instance, we can see from the recurrence relation, that the sequence comprising the integral values is a decreasing sequence. However, this condition is not observed in the sequence comprising the values obtained through python.