CSCC37 A1

Leo (Si) Wang

Oct 17, 2022

Question 1

Case 1: * + * doesn't have carry over

Meaning * + * does not have more than 1 digit in it's representation in base b.
Given $* + * \neq *$, we can say $* \neq 0$, and also that # is even.
If there is no carry over, we know $\# + \# = \#\diamond$
For any digit d_0 we know in any base b, $d_0 + d_0 \leq (1(b-2))_b$ for any base $b > 1$ The only way for $\# + \# = \#\diamond$ to work is if $\# = 1$ and $b = 2$.
is not even, plus the given the number of symbols, the base must be at least 3.
$\therefore$ we go to case 2:

Case 2: * + * does have carry over

As stated in case 1, the number of the digit in carry over when adding 2 single digits is always at most 1 in any base (if you count 0 as a carry over anyways). Thus we have:

$$* + * = (1\#)_b$$
$$\Rightarrow \# + \# + 1 = \#\diamond$$
If we know $d_0 + d_0 \leq (1(b-2))_b$,
Then we know $d_0 + d_0 + 1 \leq (1(b-1))_b$
$$\Rightarrow \# = 1$$
$$\Rightarrow 1 + 1 + 1 = (1\diamond)_b$$
$\Rightarrow \diamond = 0$, and $b = 3$ (Since there are 3 symbols, $b = 2$ is impossible)
For the sake of completion here is the solution to *
$$* + * = (11)_3$$
$$\Rightarrow * = 2$$

$\therefore$ the base of the system is 3 $\blacksquare$

Question 2

Try to store $(0.1)_{10}$ in b = 3 to 9:

$b = 3$

Multiplier	Decimal	Integer
3	0.1	0
3	0.3	0
3	0.9	2
3	0.7	2
3	0.1	0

$\therefore (0.1)_3 = 0.\overline{0022}$ and cannot be exactly stored in a base 3 machine.

$b = 4$

Multiplier	Decimal	Integer
4	0.1	0
4	0.4	1
4	0.6	2
4	0.4	1

$\therefore (0.1)_4 = 0.0\overline{12}$ and cannot be exactly stored in a base 4 machine.

$b = 5$

Multiplier	Decimal	Integer
5	0.1	0
5	0.5	2
5	0.5	2

$\therefore (0.1)_5 = 0.0\overline{2}$ and cannot be exactly stored in a base 5 machine.

$b = 6$

Multiplier	Decimal	Integer
6	0.1	0
6	0.6	3
6	0.6	3

$\therefore (0.1)_6 = 0.0\overline{3}$ and cannot be exactly stored in a base 6 machine.

$b = 7$

Multiplier	Decimal	Integer
7	0.1	0
7	0.7	4
7	0.9	6
7	0.3	2
7	0.1	0

$\therefore (0.1)_7 = 0.\overline{0462}$ and cannot be exactly stored in a base 7 machine.

$b = 8$

Multiplier	Decimal	Integer
8	0.1	0
8	0.8	6
8	0.4	3
8	0.2	1
8	0.6	4
8	0.8	6

$\therefore (0.1)_8 = 0.0\overline{6314}$ and cannot be exactly stored in a base 8 machine.

$b = 9$

Multiplier	Decimal	Integer
9	0.1	0
9	0.9	8
9	0.1	0
9	0.9	8

$\therefore (0.1)_9 = 0.\overline{08}$ and cannot be exactly stored in a base 9 machine.

In general, for $3 \leq b < 10$ there is no $n \in \mathbb{N}$ st b^n is divisible by 10.

$\therefore$ (0.1) cannot be represented exactly on a machine of any base 2-9 ■

Question 3

Definition of ϵ: The smallest non-normalized floating point number s.t. $1 + \epsilon > 1$
Remember that 1 in any base b, 1 can be represented as $(1)_b = (.1)_b \times b^{(1)_b}$

Case 1: The FP system chops digits

Given the representation of 1, it's easy to see that the smallest number that can be added which the definition of ϵ still holds is

$$(.\underbrace{00\cdots01}_{t})_b \times b^{(1)_b}$$
$$= (.\underbrace{10\cdots00}_{t})_b \times b^{(1-(t-1))_b}$$
$$= b^{(1-t)_b}$$
$$\therefore \epsilon = b^{(1-t)_b}, \text{ for chopping FP systems}$$

Case 2: The FP system rounds digits

The FP system will round if $d_{t+1} \geq \frac{1}{2}b$
Given the representation of 1, we need to make it such that the digit $d_{t+1} = \frac{1}{2}$.
Take the previous answer for chopping systems and half the amount the required amount:

$$\frac{1}{2}(.\underbrace{00\cdots01}_{t})_b \times b^{(1)_b}$$
$$= \frac{1}{2}(.\underbrace{10\cdots00}_{t})_b \times b^{(1-(t-1))_b}$$
$$= \frac{1}{2}b^{(1-t)_b}$$
$$\therefore \epsilon = \frac{1}{2}b^{(1-t)_b}, \text{ for rounding FP systems}$$

$\therefore \epsilon$ is a bound for relative round off error δ ■

Question 4

We know that $||\overrightarrow{x}||_\infty = max(|x_1|, \cdots, |x_n|)$ for vector $\overrightarrow{x} = (x_1, \cdots, x_n)$
Suppose for the sake of the proof that $x_n \neq 0$ is the largest absolute element in $\overrightarrow{x}$, meaning $|x_n| = max(|x_1|, \cdots, |x_n|)$

$$\begin{aligned}
||\overrightarrow{x}||_p &= \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \\
&= \left(\sum_{i=1}^{n} \left|\frac{x_n x_i}{x_n}\right|^p\right)^{\frac{1}{p}} \\
&= \left(|x_n|^p \sum_{i=1}^{n} \left|\frac{x_i}{x_n}\right|^p\right)^{\frac{1}{p}} \\
&= |x_n| \left(\sum_{i=1}^{n} \left|\frac{x_i}{x_n}\right|^p\right)^{\frac{1}{p}} \\
&= |x_n| \left(1 + \sum_{i=1}^{n-1} \left|\frac{x_i}{x_n}\right|^p\right)^{\frac{1}{p}}
\end{aligned}$$

Because x_n is the largest element, we know $0 \leq \left|\frac{x_i}{x_n}\right| \leq 1$ for $1 \leq i \leq n-1$
We can use squeeze theorem to get the result:

$$\begin{gathered}
||\overrightarrow{x}||_p \geq |x_n|(1)^{\frac{1}{p}} \\
||\overrightarrow{x}||_p \leq |x_n|(n)^{\frac{1}{p}} \\
\lim_{p\to\infty} |x_n|(1)^{\frac{1}{p}} \leq \lim_{p\to\infty} ||\overrightarrow{x}||_p \leq \lim_{p\to\infty} |x_n|(n)^{\frac{1}{p}} \\
|x_n| \leq \lim_{p\to\infty} ||\overrightarrow{x}||_p \leq |x_n| \\
\therefore \text{By squeeze theorem } \lim_{p\to\infty} ||\overrightarrow{x}||_p = |x_n| \\
\\
\text{Thus } ||\overrightarrow{x}||_\infty = \lim_{p\to\infty} ||\overrightarrow{x}||_p
\end{gathered}$$

■

Question 5

For $n = 2$ we have: $AB = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} + a_{12}b_{22} \\ 0 & a_{22}b_{22} \end{bmatrix}$
Thus we can calculate:

$$\begin{aligned}
&fl(AB) = fl(A) \times fl(B) \\
&= \begin{bmatrix} fl(a_{11}) & fl(a_{12}) \\ 0 & fl(a_{22}) \end{bmatrix} \times \begin{bmatrix} fl(b_{11}) & fl(b_{12}) \\ 0 & fl(b_{22}) \end{bmatrix} \\
&= \begin{bmatrix} fl(a_{11})fl(b_{11}) & fl(a_{11})fl(b_{12}) + fl(a_{12})fl(b_{22}) \\ 0 & fl(a_{22})fl(b_{22}) \end{bmatrix} \\
&= \begin{bmatrix} a_{11}b_{11}\,(1-\delta_3) & ((a_{11}b_{12})(1-\delta_1) + (a_{12}b_{22})(1-\delta_2))\,(1-\delta_4) \\ 0 & (a_{22}b_{22})\,(1-\delta_5) \end{bmatrix} \text{(Given by the derivations done in lecture)} \\
&= \underbrace{\begin{bmatrix} a_{11} & a_{12}(1-\delta_2)(1-\delta_4) \\ 0 & a_{22}(1-\delta_5) \end{bmatrix}}_{\hat{A}} \times \underbrace{\begin{bmatrix} b_{11}(1-\delta_3) & b_{12}(1-\delta_1)(1-\delta_4) \\ 0 & b_{22} \end{bmatrix}}_{\hat{B}}
\end{aligned}$$

We can see $\hat{A}$ and $\hat{B}$ resemble A and B closely, but with slight offsets in some components.
Let $\hat{A} = A + E_A$ and $\hat{B} = B + E_B$

Define $E_A = \begin{bmatrix} 0 & a_{12}(-\delta_2 - \delta_4 + \delta_2\delta_4) \\ 0 & a_{22}(-\delta_5) \end{bmatrix}$, $E_B = \begin{bmatrix} b_{11}(-\delta_3) & b_{12}(-\delta_1 - \delta_4 + \delta_1\delta_4) \\ 0 & 0 \end{bmatrix}$
Let δ_A be the bound on the offset for A such that $|\delta_A| = max(|-\delta_2 - \delta_4 + \delta_2\delta_4|, |\delta_5|)$
Let δ_B be the bound on the offset for B such that $|\delta_B| = max(|-\delta_1 - \delta_4 + \delta_1\delta_4|, |\delta_3|)$

Then we know:

$$||E_A|| \leq ||\delta_A A|| = |\delta_A| \; ||A||$$
$$||E_B|| \leq ||\delta_B B|| = |\delta_B| \; ||B||$$

Thus $||E_A||$ and $||E_B||$ are bounded by a small multiple of A and B respectively and $fl(AB) = \hat{A}\hat{B}$ $\blacksquare$