Assignment 1

Department of Computational and Data Sciences
DS226: Introduction to Computing for Artificial Intelligence and Machine Learning
Name: Vikramaditya Mishra[CDS-21757]

1 Question 1.

Solution: (a) To define a function which maps

$$U\text{toST}_w : \mathbb{Y} \rightarrow \mathbb{Z},$$

We need to find inverse function $U_w^{-1}(\cdot)$ first,

$$U_w^{-1}(\mathbf{y}) : \mathbb{Y} \rightarrow \mathbb{X},$$

Let y_{w-1} is given unsigned integer for $U_w^{-1}(\mathbf{y})$, where $\mathbf{y} \in \mathbb{Y}$ and $\mathbf{x}\,(x_{w-1}, x_{w-2}, \ldots, x_0) \in \mathbb{X}$ is output of the function.

$$=> x_i = \lfloor \frac{y_i}{2^i} \rfloor \quad -(1)$$

$$=> y_{i-1} = y_i mod 2^i \quad -(2)$$

$$=> x_0 = y_0 \quad -(3)$$

where $1 \leq i \leq w-1$ and $\lfloor . \rfloor$ is greatest integer function.

As given in the question, that $\mathbf{x} \in \mathbb{X}$ can also be used to represent a negative integer (signed) data type using Two'scomplement encoding, defined by the function

$$ST_w : \mathbb{X} \rightarrow \mathbb{Z}, \quad ST_w(\overrightarrow{\mathbf{x}}) \doteq -x_{w-1}2^{w-1} + \sum_{i=0}^{w-2} x_i 2^i,$$

Subsitute x_i in above equation, using equations $(1), (2) and (3)$,

$$U\text{toST}_w : \mathbb{Y} \rightarrow \mathbb{Z}, \quad UtoST_w(\overrightarrow{\mathbf{y}}) \doteq -\lfloor \frac{y_{w-1}}{2^{w-1}} \rfloor 2^{w-1} + \sum_{i=0}^{w-2} \lfloor \frac{y_i}{2^i} \rfloor 2^i$$

(b) To define a function which maps

$$ST_w toU : \mathbb{Z} \rightarrow \mathbb{Y},$$

We need to find inverse function $ST_w^{-1}(\cdot)$ first,

$$ST_w^{-1}(\mathbf{z}) : \mathbb{Z} \rightarrow \mathbb{X},$$

Let z_{w-1} is given signed two's complement integer for $ST_w^{-1}(\mathbf{z})$, where $\mathbf{z} \in \mathbb{Z}$ and $\mathbf{x}\,(x_{w-1}, x_{w-2}, \ldots, x_0) \in \mathbb{X}$ is output of the function.

Case-(i): $z_{w-1} < 0$

$$=> x_{w-1} = 1 \quad -(4)$$

$$=> z_{w-2} = z_{w-1} + 2^{w-1} \quad -(5)$$

$$=> x_i = \lfloor \frac{z_i}{2^i} \rfloor \quad -(6)$$

$$=> z_{i-1} = z_i mod 2^i \quad -(7)$$

$$=> x_0 = z_0 \quad -(8)$$

where $1 \leq i \leq w-2$ and $\lfloor . \rfloor$ is greatest integer function.
Case-(ii): $z_{w-1} \geq 0$

$$=> x_{w-1} = 0 \quad -(9)$$

$$=> z_{w-2} = z_{w-1} \quad -(10)$$

$$=> x_i = \lfloor \frac{z_i}{2^i} \rfloor \quad -(11)$$

$$=> z_{i-1} = z_i mod 2^i \quad -(12)$$

$$=> x_0 = z_0 \quad -(13)$$

where $1 \leq i \leq w-2$ and $\lfloor . \rfloor$ is greatest integer function.
As given in the question, the unsigned integer data representation can be defined as a function

$$U_w : \mathbb{X} \to \mathbb{Y}, \quad U_w(\mathbf{x}) \doteq \sum_{i=0}^{w-1} x_i 2^i.$$

Subsitute x_i in above equation, using equations $(4) to (13)$ to get ,

$$STtoU_w : \mathbb{Z} \to \mathbb{Y}, \quad STtoU_w(\mathbf{z}) \doteq 2^{w-1} + \sum_{i=0}^{w-2} \lfloor \frac{z_i}{2^i} \rfloor 2^i \qquad (Case-(i))$$

$$STtoU_w : \mathbb{Z} \to \mathbb{Y}, \quad STtoU_w(\mathbf{z}) \doteq \sum_{i=0}^{w-2} \lfloor \frac{z_i}{2^i} \rfloor 2^i \qquad (Case-(ii))$$

2 Question 2.

Solution: (a)

Numbers	IEEE single precision float format	Hexadecimal
86.125	0 10000101 01011000100000000000000	56.2
0.523	0 01111110 00001011110001101010100	0.85E353F7
-0	1 00000000 00000000000000000000000	0

.

(b) $2^{23} - 1$, The interval $[-2^{-12}, -2^{-11}]$ will be represented as

$\{-2^{-12}, -2^{-12}(1+2^{-23}), -2^{-12}(1+2\times 2^{-23}), -2^{-12}(1+3\times 2^{-23}), ..., -2^{-12}(1+(2^{23}-1)\times 2^{-23}), -2^{-11}\}$

Answer does not changes, The interval $[-2^{-13}, -2^{-12}]$ will be represented as

$\{-2^{-13}, -2^{-13}(1+2^{-23}), -2^{-13}(1+2\times 2^{-23}), -2^{-13}(1+3\times 2^{-23}), ..., -2^{-13}(1+(2^{23}-1)\times 2^{-23}), -2^{-12}\}$

(c)

n=2^{24}+1 = 16777217, The interval $[2^{24}, 2^{25}]$ will be represented as $\{2^{24}, 2^{24}(1+2^{-23}), 2^{24}(1+2\times 2^{-23}), ..., 2^{25}\}$.

$2^{24}(1+2^{-23}) = 2^{24}+2; 2^{24}+1$ is missing.

For 32-bit signed integer representation, n=2147483648.

For 32-bit unsigned integer representation, n= 4294967296.

3 Question 3.

Solution:(a)

i. 4 bits

ii.$1_{10} = 0001_2$, As the nation achieve democracy, we have to subtract 2 points from aggression value. For the given nation we have to subtract 2 from 1, result will be 1111_2 which is 15_10.

iii. Wrap around error. To solve this error we can use signed 2's complement form(4+1 bits). Now after subtracting 2 we will get -1, then for using nuclear weapons we will add 10 and final score will be 9.
(b)

i. 32-bit if unsigned integer and 33-bit if signed integer.

ii.497 days

iii.1,069,446,856,703

(c).

Stage	16-bit IEEE floating point	Decimal	16-bit signed integer
I	0110001111111011	1021.510	0000001111111101
II	0110011111101100	202810	0000011111101100
III	0111001101101101	15208	0011101101101000
IV	0111100000011111	33760	1000000000000000
V	0111101000111111	5116810	1000000000000000

(d)

(i)Binary representation of 0.1 = 0.0001 1001 1001 1001 1001 1001 100.....
=>0.1-x=0.0000 0000 0000 0000 0000 0000 [1100].....
=>0.1-x=0.1100[1100].......................... $\times 10^{-23}$

(ii)error= 0.1100[1100].......................... $\times 10^{-23}$

$= 9.53764 \times 10^{-8}$

4 Question 4.

Solution:(a)

$$y = 0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n$$

There will be 2 case for rounding off:

Case(i):If $0 \leq d_{k+1} < 5$, then $d'_k = d_k$,

So, $\psi_k(y) = 0.d_1 d_2 \cdots d_k \times 10^n$,

$$Relativeerror = \left|\frac{y - \psi_k(y)}{y}\right|$$

Put the values of y and $\psi_k(y)$ in above equation

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n - 0.d_1 d_2 \cdots d_k \times 10^n}{0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n}\right|$$

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{0.d_{k+1} d_{k+2} \cdots}{0.1}\right| \times 10^{-k}$$

As $d_{k+1} < 5$,

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{0.d_{k+1} d_{k+2} \cdots}{0.1}\right| \times 10^{-k} < 5 \times 10^{-k}$$

$$=> \left|\frac{y - \psi_k(y)}{y}\right| < 0.5 \times 10^1 \times 10^{-k}$$

$$=> \left|\frac{y - \psi_k(y)}{y}\right| < 0.5 \times 10^{1-k} \quad -(1)$$

Case(ii):If $5 \leq d_{k+1} \leq 9$, then $d'_k = d_k + 1$,

So, $\psi_k(y) = 0.d_1 d_2 \cdots (d_k + 1) \times 10^n$,

$$Relativeerror = \left|\frac{y - \psi_k(y)}{y}\right|$$

Put the values of y and $\psi_k(y)$ in above equation

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n - 0.d_1 d_2 \cdots (d_k + 1) \times 10^n}{0.d_1 d_2 \cdots d_k d_{k+1} d_{k+2} \cdots \times 10^n}\right|$$

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{1 - 0.d_{k+1} d_{k+2} \cdots}{0.1}\right| \times 10^{-k}$$

As $9 \geq d_{k+1} \geq 5$,

$$=> \left|\frac{y - \psi_k(y)}{y}\right| = \left|\frac{1 - 0.d_{k+1} d_{k+2} \cdots}{0.1}\right| \times 10^{-k} \leq 5 \times 10^{-k}$$

$$=> \left|\frac{y - \psi_k(y)}{y}\right| \leq 0.5 \times 10^1 \times 10^{-k}$$

$$=> \left|\frac{y-\psi_k(y)}{y}\right| \leq 0.5 \times 10^{1-k} \quad -(2)$$

Combining (1) and (2) we get,

$$=> \left|\frac{y-\psi_k(y)}{y}\right| \leq 0.5 \times 10^{1-k}$$

(b) Largest number that can be represented using given representation is 0.9999×10^{15}.

To avoid overflow,

$$=> \frac{p!}{3!(p-3)!} \leq 0.9999 \times 10^{15}$$

Simplify above equation,

$$=> \frac{p(p-1)(p-2)(p-3)!}{3!(p-3)!} \leq 0.9999 \times 10^{15}$$

$$=> \frac{p(p-1)(p-2)}{3!} \leq 0.9999 \times 10^{15}$$

$$=> p(p-1)(p-2) \leq 3! \times 0.9999 \times 10^{15}$$

$$=> p^3 - 3p^2 + 2p \leq 5.9994 \times 10^{15}$$

to find upper bound,

$$=> p^3 - 3p^2 + 2p = 5.9994 \times 10^{15}$$

Solve above equation to get $p = 181707.00201$

So largest value of p for overflow can be avoided is 181707 or 0.1817×10^6

(c) Exponential average $f_e(a,b) = \frac{e^a - e^b}{a-b}$

$=> a = 1.0166$ and $b = 1.0116$

Put the values of a and b in exponetial average equation

$f_e(1.0166, 1.0116) = \frac{e^{1.0166} - e^{1.0116}}{1.0166 - 1.0116}$

$f_e(1.0166, 1.0116) = 2.75688 \quad$ (actual value)

We can also approximate the exponential by $e^z = \frac{6+2z}{6-4z+z^2}$ (which is [2/1] Padè approximation of e^z)

Now calculate the approximate values of $e^{1.0166}$ and $e^{1.0116}$ usin above equation,

$=> e^{1.0166} = \frac{6+2\times1.0166}{6-4\times1.0166+1.0166^2}$

$=> e^{1.0166} = 2.70744 \quad -(1)$

$=> e^{1.0116} = \frac{6+2\times1.0116}{6-4\times1.0116+1.0116^2}$

$=> e^{1.0116} = 2.69512 \quad -(2)$

Approximate value of $f_e(1.0166, 1.0116)$ using equation (1) and (2),

$f_e(1.0166, 1.0116) = \frac{e^{1.0166}-e^{1.0116}}{1.0166-1.0116}$

$f_e(1.0166, 1.0116) = 2.46513 \quad$ (approximate value)

Aboslute rounding error= $|acutalvalue - approximatevalue|$

Aboslute rounding error= $|2.75688 - 2.46513|$

Aboslute rounding error= 0.29175 -(3)

Relative rounding error $=\left|\frac{actualvalue-approximatevalue}{actualvalue}\right|$

Relative rounding error $=\left|\frac{2.75688-2.46513}{2.75688}\right|$

Relative rounding error =0.10583 -(4)

5 Question 5.

Solution:(a)

floating point operation per iteration = 2 (1- addittion and 1- multiplication)

number of iteration = n

Total floating point point operation = 2n

(b)

floating point operations for inner loop = 2n (from previous part)

number of iteration for outer loop = m

Total number of floating point operation = 2nm

(c)

Floating point operations for inner most loop = 2n (from part a)

Floating point operations for midder loop = $r \times 2n$ = 2nr (no. of iteration ×floating point operations per iteration)

Total floating point operations = $m \times 2nr$ = 2mnr

(d)Consider the matrices $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{B} \in \mathbb{R}^{n\times r}, \mathbf{C} \in \mathbb{R}^{r\times p}$ and the vector $\mathbf{x} \in \mathbb{R}^{r}$.

(i)To minimize the calculation we will calculate $A(Bx)$.

Let $\mathbf{d} \in \mathbb{R}^n$ is a vector which is result of the product of the matrix **B** and the vector **x**. Further $\mathbf{e} \in \mathbb{R}^m$ is the result of the product of the matrix **A** and the vector **d**. We can use following pseudocode to calculate the multiplication:

Require: $B \in \mathbb{R}^{n\times r}$ and $x \in \mathbb{R}^r$

$d \leftarrow 0, d \in \mathbb{R}^n$

for i = 1 to n do

for j = 1 to r do

$d_i \leftarrow d_i + b_{ij}x_j$

end for

end for

Require: $A \in \mathbb{R}^{m\times n}$ and $d \in \mathbb{R}^n$

$e \leftarrow 0, e \in \mathbb{R}^m$

for i = 1 to m do

for j = 1 to n do

$e_i \leftarrow e_i + a_{ij}d_j$

end for

end for

According to the method used in part (b), total number of floating point operations = $2mn + 2nr$

For the given value of m,n and r, minimum number of operations = 105000000.

(ii)To minimize the calculation we will calculate $A(BC)$.

Let $\mathbf{D} \in \mathbb{R}^{n\times p}$ is a matrix which is result of the product of the matrices **B** and **C**. Further $\mathbf{E} \in \mathbb{R}^{m\times p}$ is the result of the product of the matrices **A** and **D**. We can use following pseudocode to calculate the multiplication:

Require: $B \in \mathbb{R}^{n\times r}$ and $C \in \mathbb{R}^{r\times p}$

$D \leftarrow 0, D \in \mathbb{R}^{n\times p}$

for i = 1 to n do

for j = 1 to p do

for k = 1 to r do

$d_{ij} \leftarrow d_{ij} + b_{ik}c_{kj}$

end for

end for

end for

Require: $A \in \mathbb{R}^{m\times n}$ and $D \in \mathbb{R}^{n\times p}$

$E \leftarrow 0, E \in \mathbb{R}^{m\times p}$

for i = 1 to m do

for j = 1 to p do

for k = 1 to n do

$e_{ij} \leftarrow e_{ij} + a_{ik}d_{kj}$

end for

end for

end for

According to the method used in part (c), total number of floating point operations $=2mnp + 2nrp$

For the given value of m,n,r and p, minimum number of operations = 15750000000.