Homework 1
CS6490

Darpan Gaur
CO21BTECH11004

Pair 1

Part a

- **INT32** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.
- **FP32** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.

Part b

Range

- **INT32** : The range of INT32 is $-2^{31} = -2147483648$ to $2^{31} - 1 = 2147483647$.
- **FP32** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Positive minimum value 0|00000000|00000000000000000000001 $\approx 1.4 \cdot 10^{-45}$ and maximum value 0|11111110|11111111111111111111111 $\approx 3.4 \cdot 10^{38}$. Negative minimum value 1|1111110|11111111111111111111111 $\approx -3.4 \cdot 10^{38}$, and maximum value 1|00000000|00000000000000000000001 $\approx -1.4 \cdot 10^{-45}$. Also, zero is represented as 0|00000000|00000000000000000000000.

Precision

- **INT32** : Here, minimum distance between two numbers is 1, so precision is 1.
- **FP32** : For this precision is not constant, depends on exponent and mantissa. High precision for small numbers.
  Eg:- 0|00000000|00000000000000000000000 and 0|00000000|00000000000000000000001 have difference of $\approx 1.4 \cdot 10^{-45}$.
  Similarly, low precision $\approx 2.02 \cdot 10^{31}$. Hence, precision varies in range $1.4 \cdot 10^{-45}$ to $2.03 \cdot 10^{31}$.

Part c

99999999 can be represented in INT32 but not in FP32, equivalent FP32 value is $1.0000000 \cdot 10^8$, with error of 1.

Part d

1.25 can be represented in FP32 but not in INT32, equivalent INT32 value is 1, with error of 0.25.

Pair 2

Part a

- **INT32 (18-14)** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.

- **FP32** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.

Part b

Range

- **INT32 (18-14)** : Zero is represented with all bits as 0. Positive minimum value $000000000000000000.00000000000001 \approx 6.1035 \cdot 10^{-5}$ and maximum value $011111111111111111.11111111111111 \approx 131071.99993896484$. Negative minimum value $100000000000000000.00000000000001 \approx -131071.99993896484$, and maximum value $111111111111111111.11111111111111 \approx -6.1035 \cdot 10^{-5}$.

- **FP32** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Positive minimum value $\approx 1.4 \cdot 10^{-45}$ and maximum value $\approx 3.4 \cdot 10^{38}$. Negative minimum value $\approx -3.4 \cdot 10^{38}$, and maximum value $\approx -1.4 \cdot 10^{-45}$. Also, zero is represented.

Precision

- **INT32 (18-14)** : Precision is constant and equal to $2^{-14} \approx 6.1035 \cdot 10^{-5}$.

- **FP32** : For this precision is not constant, depends on exponent and mantissa. Precision varies in range $1.4 \cdot 10^{-45}$ to $2.03 \cdot 10^{31}$.

Part c

14295.515563964844 can be represented in INT32 (18-14) but not in FP32, equivalent FP32 value is 14295.515625, with error of 0.000061035.

Part d

1.0099999904632568359375 can be represented in FP32 but not in INT32 (18-14), equivalent INT32 (18-14) value is 1, with error of 0.25.

Pair 3

Part a

- **INT32 (14-18)** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.
- **FP16** : As there are 16 bits, it can represent $2^{16} = 65536$ unique numbers.

Part b

Range

- **INT32 (14-18)** : Zero is represented with all bits as 0. Smallest positive value $= 2^{-18} = 3.815 \cdot 10^{-6}$, largest positive value
  01111111111111|111111111111111111 ≈ 8191.999996.
  Smallest negative value $= -8191.999996$, largest negative value $= -3.815 \cdot 10^{-6}$.
- **FP16** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Positive minimum value $\approx 5.96 \cdot 10^{-8}$ and maximum value ≈ 65504. Negative minimum value ≈ -65504, and maximum value $-5.96 \cdot 10^{-8}$. Also, zero is represented.

Precision

- **INT32 (14-18)** : Precision is constant and equal to $2^{-18} \approx 3.815 \cdot 10^{-6}$.
- **FP16** : For this precision is not constant, depends on exponent and mantissa.
  High precision for small numbers.
  Eg:- 0|00000|0000000000 and 0|00000|0000000001 have difference of $\approx 5.96 \cdot 10^{-8}$. Similarly, low precision = 32.

Part c

9999 can be represented in INT32 (14-18) but not in FP16, equivalent FP16 value is 10000, with error of 1.

Part d

65504 can be represented in FP16 but not in INT32 (14-18).

Pair 4

Part a

- **INT32 (14-18)** : As there are 32 bits, it can represent $2^{32} = 4294967296$ unique numbers.
- **bfloat16** : As there are 16 bits, it can represent $2^{16} = 65536$ unique numbers.

Part b

Range

- **INT32 (14-18)** : Zero is represented with all bits as 0. Smallest positive value $= 2^{-18} = 3.815 \cdot 10^{-6}$, largest positive value
  $01111111111111|111111111111111111 \approx 8191.999996$.
  Smallest negative value $= -8191.999996$, largest negative value $= -3.815 \cdot 10^{-6}$.
- **bfloat16** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Smallest positive value $0|00000000|0000001 \approx 9.1835 \cdot 10^{-41}$ and
  largest positive value $0|11111110|111111111 \approx 3.3895314 \cdot 10^{38}$. Smallest negative value $1|1111110|111111111 \approx -3.3895314 \cdot 10^{38}$, and
  largest negative value $1|00000000|0000001 \approx -9.1835 \cdot 10^{-41}$. Also, zero is represented.

Precision

- **INT32 (14-18)** : Precision is constant and equal to $2^{-18} \approx 3.815 \cdot 10^{-6}$.
- **bfloat16** : For this precision is not constant, depends on exponent and mantissa.
  High precision for small numbers.
  Eg:- $0|00000|0000000000$ and $0|00000|0000000001$ have difference of $\approx 9.1835 \cdot 10^{-41}$.
  Similarly, low precision $\approx 1.329 \cdot 10^{36}$.

Part c

4096.00390625 can be represented in INT32 (14-18) but not in bfloat16, equivalent bfloat16 value is 4128, with error of 32.

Part d

Max value of bfloat16 is $3.3895314 \cdot 10^{38}$, which can't be represented in INT32 (14-18).

Pair 5

Part a

- **FP16** : As there are 16 bits, it can represent $2^{16} = 65536$ unique numbers.
- **bfloat16** : As there are 16 bits, it can represent $2^{16} = 65536$ unique numbers.

Part b

Range

- **FP16** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Positive minimum value $\approx 5.96 \cdot 10^{-8}$ and maximum value ≈ 65504. Negative minimum value ≈ -65504, and maximum value $-5.96 \cdot 10^{-8}$. Also, zero is represented.

- **bfloat16** : Here, special values like $+\infty$, $-\infty$ and NaN are also included. Smallest positive value $\approx 9.1835 \cdot 10^{-41}$ and largest positive value $\approx 3.3895314 \cdot 10^{38}$. Smallest negative value $\approx -3.3895314 \cdot 10^{38}$, and largest negative value $\approx -9.1835 \cdot 10^{-41}$. Also, zero is represented.

Precision

- **FP16** : For this precision is not constant, depends on exponent and mantissa.
  High precision for small numbers.
  Eg:- 0|00000|0000000000 and 0|00000|0000000001 have difference of $\approx 5.96 \cdot 10^{-8}$. Similarly, low precision $= 32$.

- **bfloat16** : For this precision is not constant, depends on exponent and mantissa.
  High precision for small numbers.
  Eg:- 0|00000|0000000000 and 0|00000|0000000001 have difference of $\approx 9.1835 \cdot 10^{-41}$. Similarly, low precision $\approx 1.329 \cdot 10^{36}$.

Part c

$5.96 \cdot 10^{-8}$ can be represented in FP16 but not in bfloat16.

Part d

$3.3895314 \cdot 10^{38}$ can be represented in bfloat16 but not in FP16.

Pair 6

Part a

- **INT8** : As there are 8 bits, it can represent $2^8 = 256$ unique numbers.

- **FP16** : As there are 16 bits, it can represent $2^{16} = 65536$ unique numbers.

Part b

Range

- **INT8** : The range of INT8 is $-2^7 = -128$ to $2^7 - 1 = 127$.

- **FP16** : Here, special values like $+\infty$, $-\infty$ and NaN are also included.
  Positive minimum value $0|00000|0000000001 \approx 5.96 \cdot 10^{-8}$ and
  maximum value $0|11110|1111111111 \approx 65504$.
  Negative minimum value $1|11110|1111111111 \approx -65504$, and
  maximum value $1|00000|0000000001 \approx -5.96 \cdot 10^{-8}$. Also, zero is represented as $0|00000|0000000000$.

Precision

- **INT8** : Here, minimum distance between two numbers is 1, so precision is 1.
- **FP16** : For this precision is not constant, depends on exponent and mantissa.
  High precision for small numbers.
  Eg:- $0|00000|0000000000$ and $0|00000|0000000001$ have difference of $\approx 5.96 \cdot 10^{-8}$. Similarly, low precision $= 32$.

Part c

9999 can be represented in INT8 but not in FP16, equivalent FP16 value is 10000, with error of 1.

Part d

1.25 can be represented in FP16 but not in INT8, equivalent INT8 value is 1, with error of 0.25.