Assignment:2
Due:15/08/18

Definitions

• **Resolution:** For a $Q_{n.m}$ format, $1/(2^m)$ is the resolution

• **Precision:** The length of the register/ size of the datatype is called pre-
cision

• **Range:** The range of a $Q_{n.m}$ for unsigned numbers is 0 to 2^n -2^{-m}

Learn Notation by an example

• **Example :** Denote Π in fixed point notation using 8 bit and 16 bit registers

• Multiply the number by 2^m, after deciding on the Q format that you want
to use.

• Quantize the number. Again to revert back, divide the rounded number
by 2^m

• **Example:** Denote - π/8 in Fixed Point.

- π/8 =-0.392699081698724

since the integer part is zero, and only one bit is required for signed rep-
resentation, choose $Q_{2.14}$ format , so fp=round(− π/8 2^{14}) =**-6434**

Again to Floating Point domain, $-6434/2^{14}$ = -0.392700195312500

Error is 1.113613776027034 10^{-6}

Basic arithmetic operations:Addition/Subtraction

• To add two numbers in fixed point notation, both the numbers should be
in the same Q format.

• **Exercise :** Add π/4 and π/8 using 16 bit prec. **Solution :** x= 3π/8

X = round(x 2^{14}) = 12868

y= π/8

$Y = \text{round}(y \quad 2^{14}) = 6434$ Though we can proceed with $Q_{1.15}$ format, the
addition introduces a carry/extra bit, because of which we choose $Q_{2.14}$
$Z = X + Y = 19302$
$z = Z/2^{14} = 1.178100585937500$

• Adding $Q_{n.m}$ and $Q_{n.m}$ will result in a number that has $Q_{n+1.m}$

Assignment:
• **Exercise:** Repeat the above example with 8 bit prec
• **Exercise:** Add $\pi/6$ and $-\pi/8$ using 8 and 16 bit prec.
• **Exercise:** Write a fixed point c code for addition and subtraction