ESE2014: Fixed-point Arithmetic

Name: Shreya Mamadapur Student ID: C0774035
Instructor: Takis Zourntos

In the following exercises, if you need a scaling factor, choose it so that it is the nearest power of two, that makes |x_norm| less than one. This makes your calculations easier (why?).

1) Assume a 16-bit word, with an 8-bit fraction, i.e., M=8. Provide the fixed-point representations for the following numbers. In each case, what is the error associated with the representation?

a) 3.14159

b) 0.2378

c) 5.125

d) 125.32

ANSWER

a) x_norm = 3.14159 and M=8
Fixed-point representation is in the form: **x = x_norm * (Scaling factor) * 2^M**
To make |x_norm| less than 1, we can choose the scaling factor of 2^-2

x = 3.14159 * (1/ 2^2) * 2^8
= 3.14159/4 * (256)
x = 201.06176

x_actual= 201*(2^2)/2^8
= 3.140625
Error = |x_norm - x_actual|
= 0.000965

b) x_norm = 0.2378 and M=8
Let's choose a scaling factor of 2^2
x = 0.2378 * (2^2) * 2^8
x = 243.5072

x_actual= 243/(2^2*2^8)
= 0.2373046875
Error = |x_norm - x_actual|
= 0.000495312

c) x_norm = 5.125 and M=8

ESE2014: Fixed-point Arithmetic

Let’s choose a scaling factor of 2^-3
x = 5.125 * (1/ 2^3) * 2^8
x = 164
x_actual= 164*(2^3/2^8)
= 5.125
Error = |x_norm - x_actual|
= 0

d) x_norm = 125.32 and M=8
Let’s choose a scaling factor of 2^-7
x = 125.32 * (1/ 2^7) * 2^8
x = 250.64

x_actual= 250*(2^7/2^8)
= 125
Error = |x_norm - x_actual|
= 0.32

2) repeat the above, but use a 10-bit fraction, i.e., M=10.

ANSWER

e) x_norm = 3.14159 and M=10
Fixed-point representation is in the form: **x = x_norm * (Scaling factor) * 2^M**
To make |x_norm| less than 1, we can choose the scaling factor of 2^-2

x = 3.14159 * (1/ 2^2) * 2^10
= 3.14159/4 * (1024)
x = 804.24704

x_actual= 804*(2^2)/2^10
= 3.140625
Error = |x_norm - x_actual|
= 0.000965

f) x_norm = 0.2378 and M=10
Let’s choose a scaling factor of 2^2
x = 0.2378 * (2^2) * 2^10
x = 974.0288

x_actual= 974/(2^2*2^10)
= 0.237792968
Error = |x_norm - x_actual|
= 0.000007031

ESE2014: Fixed-point Arithmetic

g) x_norm = 5.125 and M=10
Let's choose a scaling factor of 2^-3
x = 5.125 * (1/ 2^3) * 2^10
x = 328
x_actual= 328*(2^3/2^10)
= 5.125
Error = |x_norm - x_actual|
= 0

h) x_norm = 125.32 and M=10
Let's choose a scaling factor of 2^-7
x = 125.32 * (1/ 2^7) * 2^10
x = 1002.56

x_actual= 1002*(2^7/2^10)
= 125.25
Error = |x_norm - x_actual|
= 0.07

3) repeat the above, but assume a 32-bit word and 16-bit fraction. How do the errors compare with the 16-bit, M=8, case?

ANSWER

i) x_norm = 3.14159 and M=16
Fixed-point representation is in the form: **x = x_norm * (Scaling factor) * 2^M**
To make |x_norm| less than 1, we can choose the scaling factor of 2^-2

x = 3.14159 * (1/ 2^2) * 2^16
= 3.14159/4 * (65536)
x = 51471.81056

x_actual= 51471*(2^2)/2^16
= 3.141540527
Error = |x_norm - x_actual|
= 0.000049472

j) x_norm = 0.2378 and M=16
Let's choose a scaling factor of 2^2
x = 0.2378 * (2^2) * 2^16
x = 62337.8432

x_actual= 62337/(2^2*2^16)
= 0.237796783

ESE2014: Fixed-point Arithmetic

Error = |x_norm - x_actual|
= 0.000003216

k) x_norm = 5.125 and M=16
Let's choose a scaling factor of 2^-3
x = 5.125 * (1/ 2^3) * 2^16
x = 41984
x_actual= 41984*(2^3/2^16)
= 5.125
Error = |x_norm - x_actual|
= 0

l) x_norm = 125.32 and M=16
Let's choose a scaling factor of 2^-7
x = 125.32 * (1/ 2^7) * 2^16
x = 64163.84

x_actual= 64163*(2^7/2^16)
= 125.3183594
Error = |x_norm - x_actual|
= 0.001640625

As we can see from the above examples, for an n-bit word, higher the M, more robust/error-free the system.