Caelan Osman
MEEN 273 Sec 4
January 31, 2020

Homework 4

Exercise 1 Convert 3200 to a float in IEEE representation

A float has one bit for sign(positive or negative), 8 bits for it's exponent with a bias of 127, and 23 bits for the mantissa (not including the shadow bit).

$$\text{number} = (\text{sign}) \times 2^{(\text{exponent}-\text{bias})} \times (\text{fraction})$$

$$3200 = (+) \times 2^{(138-137)} \times \text{fraction}$$

Because the number is positive, the first bit is 0. Converting the exponent into binary gives us

$$138 = 2^7 + 2^3 + 2 = 10001010$$

Now for determining the mantissa:

$$3200 = 2^{11} \times x \text{ which implies that } x = \frac{3200}{2^{11}} = 1.5625$$

The one is represented by the shadow bit, so we have to determine the rest.

$$1.5626 = 1 + 1 \times \frac{1}{2} + 0 \times \frac{1}{4} + 0 \times \frac{1}{8} + 1 \times \frac{1}{16} + 0 + 0 + 0....... + 0$$

This gives us or final representation in floating point

$$3200 = 01000101010010000000000000000000$$

Exercise 2 Things change a little bit for a double, still one bit for the sign, but now we have 11 bits for the exponent with a bias of 1023 and a 52 bits for the mantissa (again, not including the shadow bit).

$$3200 = (+) \times 2^{(1034-1023)} \times (\text{fraction})$$

$$1034 = 2^{10} + 0 + ... + 0 + 2^3 + 0 + 2 + 0 = 10000001010$$

$$\frac{3200}{2^{11}} = 1.5625 = 1 + 1 \times \frac{1}{2} + 0 \times \frac{1}{4} + 0 \times \frac{1}{8} + 1 \times \frac{1}{16} + 0 + 0 + 0....... + 0$$

combining this all we get our double representation for 3200.

$$3200 = 0100000010101001000000000000000000000000000000000000000000000000$$

Exercise 3 Convert the number -0.0264892578125 into a float using the IEEE floating point representation

In this case the first bit will be one as our number is negative. We need the nest smallest number that can be written as 2^e so this gives us.

$$-0.0264892578125 > 2^{-6}$$

Which give us

$$101111001$$

for the first nine bits. Now we just have to find the fraction.

$$\frac{0.0264892578125}{2^{-6}} = 1.69531250 = 1 + 0.69531250$$

$$0.69531250 \times 2 = \mathbf{1}.139062500$$

$$0.139062500 \times 2 = \mathbf{0}.7812500$$

$$0.7812500 \times 2 = \mathbf{1}.5625$$

Note that .5625 is the same as what we found for **Exercise 1** so we can just sub in this representation which is then just followed by trailing 0s.

$$-0.0264892578125 = 10111100110110010000000000000000$$

Exercise 4 What is this number (in IEEE representation) 01010010101011111111000011110000 converted to decimal?

The first bit is 0, so we know this number is gonna be positive, we also note that it is a 32 bit binary number, so the exponent will be represented by 8 bits the bias is 127 and the mantissa is represented by the remaining 23 bits (not including the shadow bit).

$$\text{number} = (\text{sign}) \times 2^{(\text{exponent}-\text{bias})} \times (\text{fraction})$$

$$\text{exponent} = 10100101 - 127 = 2^7 + 2^5 + 2^2 + 1 = 165 - 127$$

$$2^{165-127} = 2^{38}$$

$$\text{fraction} = 01011111111000011110000 = 1 + 0 \times \frac{1}{2} + 1 \times \frac{1}{4} + 0 \times \frac{1}{8} + 1 \times \frac{1}{16} + 1 \times \frac{1}{32} + 1 \times \frac{1}{64}$$

$$+1\times\frac{1}{128}+1\times\frac{1}{256}+1\times\frac{1}{512}+1\times\frac{1}{1024}+1\times\frac{1}{2048}+0\times\frac{1}{4096}+0\times\frac{1}{8192}+0\times\frac{1}{16384}+0\times\frac{1}{32768}$$

$$+1 \times \frac{1}{2^{16}} + 1 \times \frac{1}{2^{17}} + 1 \times \frac{1}{2^{18}} + 1 \times \frac{1}{2^{19}} + 0 \times \frac{1}{2^{20}} + 0 \times \frac{1}{2^{21}} + 0 \times \frac{1}{2^{22}} + 0 \times \frac{1}{2^{23}}$$

$$= 3.7783076864 \times 10^{11}$$

Exercise 5 What is the machine epsilon for a computer that dedicates 17 bits for the mantissa (the 17 is without the shadow bit).

This is a pretty simple problem we just apply $E = 2^{-n}$ where n is the number of bits in the mantissa without the shadow bit.

$$E = 2^{-17} \approx 7.6293945 \times 10^{-6}$$

Exercise 6 Perform Taylor Series expansion about $x = 2$ for the following equation (use three terms, i.e. must be second order and x is in radians):

$$f(x) = \ln x \cos x + x^{\frac{7}{2}}$$

We recall from calculus the Taylor Series expansion formula is

$$f(x) = f(a) + \frac{df(a)}{dx}\Delta x + \frac{1}{2!}\frac{d^2 f(a)}{dx^2}\Delta x^2 + \frac{1}{3!}\frac{d^3 f(a)}{dx^3}\Delta x^3 + ... \tag{1}$$

As the question asks we only want a second order polynomial about $x = 2$. We first find our derivatives.

$$\frac{df(x)}{dx} = \frac{d}{dx}(\ln x \cos x + x^{\frac{7}{2}}) = \frac{\cos x}{x} - \sin x \ln x + \frac{7x^{5/2}}{2} = \frac{2\cos x + 7x^{\frac{7}{2}} - 2x \sin x \ln x}{2x}$$

$$\frac{d^2 f(x)}{dx^2} = \frac{d}{dx}\left(\frac{2\cos x + 7x^{\frac{7}{2}} - 2x \sin x \ln x}{2x}\right)$$

$$= \frac{-4x^2 \cos(x) \ln(x) - 8x \sin(x) + 35x^{\frac{7}{2}} - 4\cos(x)}{4x^2}$$

Looking at $f(2)$

$$f(a) = \ln 2 \cos 2 + 2^{\frac{7}{2}} \cong 11.025257 \tag{2}$$

looking at $\frac{df(2)}{dx}$

$$\frac{df(2)}{dx} = \frac{2\cos 2 + 7 * 2^{\frac{7}{2}} - 2 * 2 \sin 2 \ln 2}{2 * 2} \cong 18.9606 \tag{3}$$

finally looking at $\frac{d^2 f(2)}{dx^2}$

$$\frac{d^2 f(2)}{dx^2} = \frac{-4 * 2^2 \cos(2) \ln(2) - 8 * 2 \sin(2) + 35 * 2^{\frac{7}{2}} - 4\cos(2)}{4 * 2^2} \cong 24.2319 \tag{4}$$

putting this all together we get

$$f(x) \cong 11.02527 + 18.9606(x-2) + \frac{24.2319(x-2)^2}{2!} \tag{5}$$

Exercise 7 Using forward finite difference what is the slope at x=2 for the same equation above? Us $\Delta x = 0.01$. What is the error in the value compared to the true slope?

We recall the formula for forward finite difference is

$$\frac{df(x_i)}{dx} \cong \frac{f(x_i + \Delta x) - f(x_i)}{\Delta x} \tag{6}$$

using $\Delta x = 0.01$ and $x_i = 2$

$$\frac{df(x_i)}{dx} \cong \frac{f(2 + 0.01) - f(2)}{0.01} \cong 19.0821 \tag{7}$$

from **Exercise 6** we have the actual value is $\cong 18.9606$ so the error is

$$18.9606 - 19.0821 \cong -0.1215 \text{ or about } 6\%.$$

Exercise 8 Using Centered Finite Difference what is the slope at $x = 2$ for the same equation above? Use $\Delta x = 0.01$. What is the error in this value compared to the true slope?

The formula for centered finite difference is

$$\frac{df(x_i)}{dx} \cong \frac{f(x_i + \Delta x) - f(x_i - \Delta x)}{2\Delta x} \tag{8}$$

plugging in $x = 2$ and $\Delta x = 0.01$

$$\frac{df(x_i)}{dx} \cong \frac{f(2 + 0.01) - f(2 - 0.01)}{2 * 0.01} \cong 18.9609 \tag{9}$$

we find the error to be

$$\cong 18.9606 - 18.9609 = -0.0003794 \text{ or about } 0.002\%$$