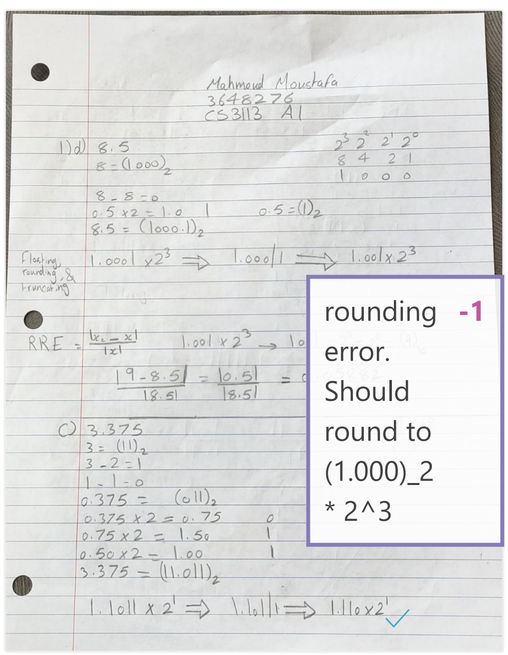
Q1

9 / 10

Give a binary floating point representation with a 4 bit mantissa $(b_1 . b_2 b_3 b_4 \times 2^{\epsilon})$ of the following numbers (use rounding to truncate). Calculate the relative rounding error in each case.

```
(a) 3.9
(b) 0.19
(c) 3.375
(d) 8.5
```



$1.110 \times 2^1 \rightarrow 11.10 = 2^1 + 2^0 . 2^{-1} = 3.5$

$RRE = \frac{|3.5 - 3.375|}{|3.375|} = \frac{|0.125|}{|3.375|} = 0.03704$ ✓

$2^1 \; 2^0$
$2 \; 1$

a) 3.9

$3 = (11)_2$

$3 - 2 = 1$

$1 - 1 = 0$

$0.9 = (0.1\overline{1100})$

$0.9 \times 2 = 1.8$	1
$0.8 \times 2 = 1.6$	1
$0.6 \times 2 = 1.2$	1
$0.2 \times 2 = 0.4$	0
$0.4 \times 2 = 0.8$	0
$0.8 \times 2 = 1.6$	1

$3.9 = (11.1\overline{1100})_2$

$1.11\overline{1100} \times 2^1 \Rightarrow 1.111|1 \Rightarrow 10.00 \Rightarrow 1.000 \times 2^2$

$1.000 \times 2^2 \rightarrow 100.0 = 2^2 = 4$ ✓

$RRE = \frac{|4 - 3.9|}{|3.9|} = \frac{|0.1|}{|3.9|} = 0.02564$ ✓

b) ~~0.9~~

~~$0.9 = (0.1\overline{1100})$~~	
~~$0.9 \times 2 = 1.8$~~	1
~~$0.8 \times 2 = 1.6$~~	1
~~$0.6 \times 2 = 1.2$~~	1
~~$0.2 \times 2 = 0.4$~~	0
~~$0.4 \times 2 = 0.8$~~	0
~~$0.8 \times 2 = 1.6$~~	1

$\cancel{1.\overline{1100} \times 2^{-1}}$

$\cancel{1.\overline{1100} \times 2^{-1} \rightarrow 0.1\overline{1100} = 2^{-1} + 2^{-2} + 2^{-3} = 0.875}$

$\cancel{RRE = \frac{|0.875 - 0.91|}{|0.91|} = 0.02778}$

1) b) 0.19

$0.19 = (0.0011000010 1\ldots\ldots)_2$

$0.19 \times 2 = 0.38$	0
$0.38 \times 2 = 0.76$	0
$0.76 \times 2 = 1.52$	1
$0.52 \times 2 = 1.04$	1
$0.04 \times 2 = 0.08$	0
$0.08 \times 2 = 0.16$	0
$0.16 \times 2 = 0.32$	0
$0.32 \times 2 = 0.64$	0
$0.64 \times 2 = 1.28$	1
$0.28 \times 2 = 0.56$	0
$0.56 \times 2 = 1.12$	1

$1.1000010 1 \times 2^{-3} \Rightarrow 1.100 \times 2^{-3}$

$1.100 \times 2^{-3} \rightarrow 0.0011 = 2^{-3} + 2^{-4} = 0.1875$

$$RRE = \frac{|0.1875 - 0.19|}{|0.19|} = \frac{0.0025}{0.19} \; 0.01316$$

Q2

5 / 5

Perform the multiplication $42.00 \times 1.2344 \times 1600$ in two different orders, using 4 digits and rounding, **base 10**. In both cases, compute the relative error of the result.

a) $(42.00 \times 1.2344) \times 1600$

b) $42.00 \times (1.2344 \times 1600)$

2) a) $(42.00 \times 1.2344) \times 1600$

$= (42.00 \times 1.234) \times 1600$

$= 51.828 \times 1600$

$= 51.83 \times 1600$

$= 82928$

$= 8.293 \times 10^4 \Rightarrow$ Rounded x_c $\qquad$ 82951.68 x

$RRE = \frac{|x_c - x|}{|x|}$

$= \frac{|82930 - 82951.68|}{|82951.68|}$

$= 0.0002614$

b) $42.00 \times (1.2344 \times 1600)$

$= 42.00 \times (1.234 \times 1600)$

$= 42.00 \times 1974.4$

$= 42.00 \times 1974$

$= 82908$

$= 8.291 \times 10^4 \Rightarrow x_c$ rounded $\qquad$ $82951.68 \rightarrow x$

$RRE = \frac{|82910 - 82951.68|}{|82951.68|}$

$= 0.0005025$

good! **5**

Q3

3 / 4

Subtract $\sqrt{8280}$ from 91 using 4 digits base 10 and rounding. Compute the relative error and comment on its magnitude.

3) 4 digit base 10 & rounding

$91 - \sqrt{8280}$

$\sqrt{8280} = 90.99$ rounded

$91 - 90.99$

$= 0.01$

$91 - \sqrt{8280} = 0.00549467138$

$$RRE = \frac{|0.01 - 0.00549467138|}{|0.00549467138|}$$

$= 0.8199$

RRE is too high: 81.99 %

Ans: large error due to catastrophic cancellation -1

Q4

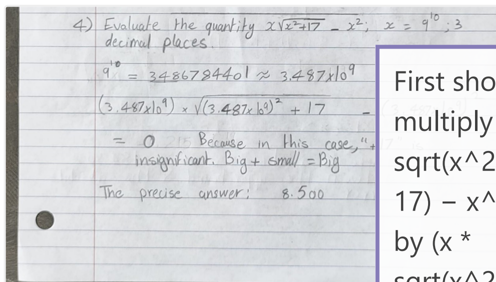
0 / 5

(textbook, section 0.4, exercise 4)

Evaluate the quantity

$x\sqrt{x^2+17} - x^2$ where $x = 9^{10}$,

correct to at least 3 decimal places.



4) Evaluate the quantity $x\sqrt{x^2+17} - x^2$; $x = 9^{10}$; 3 decimal places.

$9^{10} = 3486784401 \approx 3.487 \times 10^9$

$(3.487 \times 10^9) \times \sqrt{(3.487 \times 10^9)^2 + 17} -$

$= 0$ Because in this case, "+

insignificant. Big + small = Big

The precise answer: 8.500

First should multiply x * sqrt(x^2 + 17) − x^2 by (x * sqrt(x^2 +17) + x^2)/(x * sqrt(x^2 +17) + x^2). After simplifying, then substituting x = 9^10. **-5**

Q5

9 / 10

Consider a hypothetical 8 bit floating point machine representation with a sign bit, a 3 bit exponent, and a 4 bit mantissa ($se_1e_2e_3b_1b_2b_3b_4$), where the exponent bias is 3 (add 3 to exponent of number to form machine representation). Recall that actual mantissa has 5 bits, since the leading 1 is not stored on the machine.

(a) What is the number $e \approx 2.718$ in this 8-bit format?

(b) What is the number that $(10100111)_2$ represents in this 8-bit format?

(c) What is the upper bound of the relative error when representing a real number in this 8-bit format?

5)

a) 2.718

2^1	2^0
2	1
1	0

$2 = (10)_2$

0.718 –

$0.718 \times 2 = 1.436$	1
$0.436 \times 2 = 0.872$	0
$0.872 \times 2 = 1.744$	1
$0.744 \times 2 = 1.488$	1
$0.488 \times 2 = 0.976$	0
$0.976 \times 2 = 1.952$	1
$0.952 \times 2 = 1.904$	1
$0.904 \times 2 = 1.808$	1
$0.808 \times 2 = 1.616$	1
$0.616 \times 2 = 1.232$	1
$0.232 \times 2 = 0.464$	0
$0.464 \times 2 = 0.928$	0
$0.928 \times 2 = 1.856$	1
$0.856 \times 2 = 1.712$	1
	⋮

$2.718 = (10.1011011111\ldots)$

$= 1.01011011111 \times 2^1$

$= 1.0110 \times 2^1$

sign = 0

bias = 3

exponent = 3+1 = 4

mantissa = 0110

0100 0110 $\Rightarrow$ 8-bit representation (format)

b) (10100111) to decimal

$\underbrace{1}_{\text{sign}}\ \underbrace{010}_{\text{exponent}}\ \underbrace{0111}_{\text{mantissa}}$

exponent = 2 - 3 = -1 $(010)_2$ - bias

mantissa = 1.0111

$1.0111 \times 2^{-1} = 0.10111 = 2^{-1} + 2^{-3} + 2^{-4} + 2^{-5}$

$= 0.71875$

sign not correct

c)

$\epsilon_{mach} = 2^{-b} \to 2^{-4} = \frac{1}{16}$

upper bound of the relative error $= \frac{1}{2}$ (

$= \frac{1}{2}$

$= \frac{1}{32}$

d)

Q6

6 / 6

(textbook, section 0.5, exercise 2) Find c satisfying the Mean Value Theorem for $f(x)$ on the interval $[0, 1]$. (a) $f(x) = e^x$ (b) $f(x) = x^2$ (c) $f(x) = 1/(x + 1)$

6) $[0,1]$

a) $f(x) = e^x \Rightarrow f'(x) = e^x$

$$f'(c) = \frac{(f(b) - f(a))}{b - a}$$

$$= \frac{e^1 - e^0}{1 - 0} = \frac{e^1 - 1}{1} = e^1 - 1 = e - 1$$

$f'(x) = e^x$

Solve $e^c = e - 1 = \ln(e^c) = \ln(e - 1)$

$c = \ln(e - 1) = 0.5413$

b) $f(x) = x^2 \Rightarrow f'(x) = 2x$

$$f'(c) = \frac{(f(b) - f(a))}{b - a} = \frac{1 - 0}{1} = 1$$

$f'(x) = 2x$

Solve $2c = 1$

$c = \frac{1}{2} = 0.5$

c) $f(x) = \frac{1}{x+1} \Longrightarrow f'(x) = \frac{-1}{(x+1)^2}$

$f'(c) = \frac{f(b) - f(a)}{b - a}$

$f(b) = 0.5$
$f(a) = 1$

$= \frac{0.5 - 1}{1 - 0} = \frac{-0.5}{1} = -0.5$

Solve $\frac{-1}{(c+1)^2} = -0.5 =$

$\frac{1}{(c+1)^2} = 0.5 =$

$1 = 0.5 \times (c+1)^2$

$2 = (c+1)^2$

$\sqrt{2} = c+1$

$\sqrt{2} - 1 = c$

$0.4142 = c$

✓✓✓ in interval

$-\sqrt{2} = c+1$

$-\sqrt{2} - 1 = c$

$-2.4142 = c$ ✗

less than 0

Q7

7 / 8

Determine the second-degree Taylor polynomial and associated remainder term for the function $f(x) = e^{-x^2}$, expanding about 0, and use it to estimate $e^{-0.1^2}$. Compare the upper bound of the remainder to the exact absolute error.

7) $f(x) = e^{-x^2}$

$f'(x) = (e^{-x^2})(-2x)$

$f''(x) = (e^{-x^2})(-2x)(-2x) + (-2)(e^{-x^2}) = e^{-x^2}(4x^2-2)$

$f'''(x) = (e^{-x^2})(-2x)(4x^2-2) + (e^{-x^2})(8x) = e^{-x^2}((-8x^3+4x) + (8x)) = e^{-x^2}(-8x^3+12x)$

$f(0) = 1 \quad f'(0) = 0 \quad f''(0) = -2 \quad f'''(0) = 0$

$P(0.1)_2 = 1 + 0(0.1-0) + \frac{-2}{2!}(0.1-0)^2 = 1 - \frac{\cancel{2}(0.1)^2}{\cancel{2}}$

$= 0.99$

absolute error $= |0.99 - 0.9900498| = 0.0000498$

$f'''(c) < f'''(x) \qquad \frac{f'''(c)(x-x_0)}{3!}$

$\boxed{\begin{matrix} x_0 < c < x \\ 0 < c < 0.1 \end{matrix}}$

$\frac{e^{-0.1^2}(-8(0.1)^3 + 12(0.1))}{3!} = 0.01848$

upper bound

$0.01848 > 0.0000498$

There is a difference of 0.0184302 between the upper bound & the absolute error

remainder not correct