Lupascu Felicia FAF212

Homework 1

Exercise 1.2

Number 12.1875

Value actually stored in float 12.1875

Binary representation 01000001010000110000000000000000

Hexadecimal representation 0x41430000

	Sign	Exponent	Mantissa
Value	+1	2^3	1.5234375
Encoded as	0	130	43090912

Exercise 1.3

Exponent: 7 bits
Sign σ : 1 bit
Mantissa : 16 bits with no heading of the leading bit 1
The arithmetic used chopping

The single ~~form~~ precision IEEE format of x consists of a precision of 24 binary digits.

The exponent e is limited by $-126 < e < 127$

Lupascu Felicia FAF212

Exercise 1.5

Relative error $= \frac{x - fe(x)}{x}$

a) $x_A = 6435.4012$
$x_T = 6435.401163$

Relative error $= 0,000000575\% = 5,749447729446712e^{-9}$
Absolute error $= 0,000037$
Error $= 0,000037$
Significant digits $x_A \approx x_T$: 2 (from error)
The significant digits are 3,7

b) $x_A = 0,007245$ $x_T = 0,00723816$
Relative error $=$ 9,440993788819876e^{-4}
Error $=$ 0,00000684
Significant digits $x_A \approx x_T$ $= 3$ (error)
The significant figures are 6,8,4 for error

c) $x_A = \frac{355}{113}$ $x_T = \pi$
$=$ 3.14159292035398 $=$ 3.14159265358979 3.
Relative error $=$ 8,491367142816 902e^{-8}
Error $=$ 2,667641893008 85e^{-7} $=$ 0,000000266 764189
Significant digits $x_A \approx x_T$: 9

d) $x_A = 2.236$ $x_T = \sqrt{5} = 2,236067977$
Relative Error : 3.04013863102399e^{-5}
Error : 6,797749978969641 e^{-5}
Significant digits $x_A \approx x_T$ =

x_A has m significant digits with respect to x_T if the magnitude of error (x_A) is ≤ 5 units in the $(m+1)^{st}$ digit, beginning with the first nonzero digit in x_T

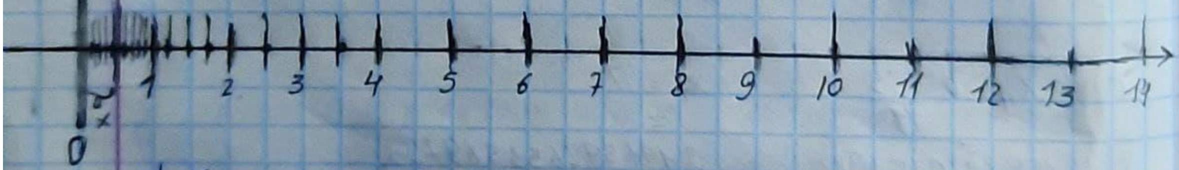
Exercise 1.4.

3 digits in mantissa : 1.00 1.01 1.10 1.11

We should always have 1 as a leading bit

Possibilities for e -3, -2, -1, 0, 1, 2, 3

	e						
$\bar{x}$	-3	-2	-1	0	1	2	3
$(1.00)_2$	$(0.125)_{10}$	$(0.25)_{10}$	$(0.5)_{10}$	$(1)_{10}$	$(2)_{10}$	$(4)_{10}$	$(8)_{10}$
$(1.01)_2$	$(0.15625)_{10}$	$(0.3125)_{10}$	$(0.625)_{10}$	$(1.25)_{10}$	$(2.5)_{10}$	$(5)_{10}$	$(10)_{10}$
$(1.10)_2$	$(0.1875)_{10}$	$(0.375)_{10}$	$(0.75)_{10}$	$(1.5)_{10}$	$(3)_{10}$	$(6)_{10}$	$(12)_{10}$
$(1.11)_2$	$(0.21875)_{10}$	$(0.4375)_{10}$	$(0.875)_{10}$	$(1.75)_{10}$	$(3.5)_{10}$	$7(10)$	$(14)_{10}$



c) $\frac{\pi}{3} = 1,0471975512 \approx 1,047$

$\frac{12}{7} = 1,71428571429 \approx 1,714$

Lupascu Felicia FAF212

Exercise 1.6

Avoid loss of significance errors in formulas

a) $\log(x) - \log(x-1)$ for large values of x

- We can avoid the loss of precision by multyplying the function by conjugate

$$\frac{\log(x) - \log(x-1)}{1} \cdot \frac{\log(x) + \log(x-1)}{\log(x) + \log(x-1)} =$$

$$= \frac{(\log(x) - \log(x-1))(\log(x) + \log(x-1))}{\log(x) + \log(x-1)}.$$

$f(100)_{\text{in first case}} =$ 0,0043648054

$f(100)_{\text{in second case}} =$ 0,00436480540245004

b) $\frac{e^x - 1}{x}$ We have to multply by the conjugate!

Other similar errors are present in calculating other coefficients, and thus they cause a major error in the final answer being calculated

c) $\cos(x+\alpha) - \cos(\alpha)$ for small values

We have to substitute the expression with something simmilar

$$\cos(\alpha) = 1 - 2\sin^2\left(\frac{\alpha}{2}\right).$$

Lupascu Felicia FAF212

Exercise 7.

Assume $|x_T - x_A| < 0.0 \ldots 05$

(10's depend on number of the digits of x_A)

$$f(x_T) - f(x_A) = f'(\xi)(x_T - x_A)$$
$$\approx f'(x_T)(x_T - x_A)$$
$$\approx f'(x_A)(x_T - x_A)$$

a) $\sin(0.521)$

Maxim size of rounding error 0.0005

$x_T - x_A = 0,0005$

$f(x_A) \cdot 0,0005 \approx f(x_T) - f(x_A)$

b) $x_T \in (e^{3.215} \cdot e^{3.225})$

c) $\sqrt{0.0011}$ Maxim size of error is 0,00005

d) $\arcsin(0,5)$

Maxim size of rounding error is 0,05