Numerical Analysis

Homework 1. Answers to some problems.

Problem 1

$(2.8)_{10} = (10.110011001100110011001100110011\ldots)_2$
$= (1.0110011001100110011001100110011\ldots)_2 \cdot 2^1$
$\sigma = +1;\ e = 1; \overline{x} = (1.0110011001100110011001100110011\ldots)_2$

In single precision IEEE format $E = e + 127 = 128_{10} = (10000000)_2$;

1	10000000	01100110011001100110011

Double precission is done similarly.

$\underbrace{(111\ldots1)_2}_{21-8=13} = 2^{12} + 2^{11} + \ldots 2 + 2^0 = \frac{2^{13}-1}{2-1} = 2^{13} - 1 = 8191$

Problem 2 *Machine epsilon is the difference between* 1 *and the next larger number that can be stored in the given format. Therefore, in this format,*

Machine epsilon $= (1.00000\ldots.0001)_2 - 1_2$ $(29\ zeros) = (0.00000\ldots0001)_2 = 2^{-30}$

M, by definition given in class, is the largest integer having the property that any integer x smaller than M can be stored exactly in this floating-point format.

If 31 is the number of binary digits in the significand, then all integers less than or equal to $(11\ldots1)_2$ can be stored exactly, where this significand contains 31 digits, all 1. It equals to $2^{31} - 1$.
In addition, the number $2^{31} = (1.0\ldots0)_2 \cdot 2^{31}$ also stores exactly.
Note that there not enough digits to store $2^{31} + 1$, as this would require 32 binary digits in the significand.
Therefore $M = 2^{31} \approx 2.147 \cdot 10^9$.

The largest number that can be represented in this format is $(1.111\ldots111)_2 \cdot 2^e$, where significand contains 31 ones, and the exponent $e = (11111111)_2 - 127_{10} = 255 - 127 = 128_{10}$. It is approximately 10^{37}.

Since rounding is used, the bounds for ε are

$$-2^{-31} \leq \varepsilon \leq 2^{-31}$$

Problem 3 *List of numbers in decimal format:*

$0.5, 0.625, 0.75, 0.875,$
$1, 1.25, 1.5, 1.75,$
$2, 2.5,$
$3, 3.5,$
$4, 5, 6, 7, 8,$
$10, 12, 14$

$\pi/4 \approx 0.785398163$. In this arithmetic it will be stored as 0.75 if both rounding and chopping are used.

$14/5 = 2.8$ In this arithmetic it will be stored as 3 if rounding is used.and 2.5 in case of chopping.

Problem 4 *a) Error* $= 28.254 - 28.271 = -0.017$; $\quad Rel(x_A) = -\frac{0.017}{28.254} = -0.00060168$.

Three significant digits;
Similarly
b) Three significant digits; c) Three significant digits; d) Four significant digits.

Problem 5

$$\begin{aligned}
&a) \quad Use \quad \log(x+1) - \log(x) = \log\frac{x+1}{x} \\
&b) \quad Use \quad Taylor \quad decomp. \quad for \quad e^x \\
&c) \quad Use \ \sin(x+a) - \sin(a) = 2\cos\frac{x+2a}{2}\sin\frac{x}{2} \\
&d) \quad Use \quad a-b = \frac{a^3-b^3}{a^2+ab+b^2} \\
&e) \quad Use \quad = \frac{\left(\sqrt{1+\frac{1}{x}}-1\right)\left(\sqrt{1+\frac{1}{x}}+1\right)}{\sqrt{1+\frac{1}{x}}+1} = \frac{\sqrt{x}}{\sqrt{x+1}+\sqrt{x}}
\end{aligned}$$

Problem 6 *Use formula*

$$|f(x_T) - f(x_A)| \approx |f\text{"}(x_A)| \cdot |x_T - x_A|$$

a) Since x_A was correctly rounded, $|x_T - x_A| \le 0.0005$. Therefore, $|f(x_T) - f(x_A)| \le |\sin 1.4713| \cdot 0.0005 \approx 4.9761 \cdot 10^{-4}$
$|rel(f(x_A))| \approx \frac{|f\text{"}(x_A)| \cdot |x_T - x_A|}{|f(x_A)|} \approx 0.0051$
b,c,d) are done similarly.

Problem 7 *Define*

$$f(x) = \int_0^x \frac{t^2}{1+t^4}dt. \quad Then \quad f'(x) = \frac{x^2}{1+x^4}$$

Thus $f(\pi) - f(22/7) \approx f'(x_A)(x_T - x_A) = f'\left(\frac{22}{7}\right)\left(\pi - \frac{22}{7}\right) = \frac{\left(\frac{22}{7}\right)^2}{1+\left(\frac{22}{7}\right)^4}\left(\pi - \frac{22}{7}\right) \approx -1.2672 \cdot 10^{-4}$.