Homework 1

Problem 1.1

Let d_1d_2/m_1m_2 be your birthday. Find the binary single precision IEEE floating-point expression for the number $(d_1m_2)_{10}$. Also, find the binary double precision IEEE floating-point expression for the number $(d_2m_1)_{10}$. In both cases, specify significand, exponent and σ. Convert $(111\ldots1)_2$ to decimal form with the parentheses enclosing $(21 - m_1m_2)$ ones.

Problem 1.2

Some microcomputers in the past used a binary floating-point format with 8 bits for the exponent and 1 bit for the sign σ. The significand contained 31 bits, with no hiding of the leading bit 1. The arithmetic used rounding. To determine the accuracy of the representation, find the machine epsilon, integer M, and the largest number that can be represented exactly in this floating-point format. Also, find the accuracy of the rounding operation.

Problem 1.3

Consider a binary floating-point representation with significand containing 3 digits without hiding the leading 1 and $-1_{10} \le e \le 3_{10}$. List all numbers that can be stored exactly together with their decimal value. Plot these numbers on real axis. For this arithmetic, specify what are the corresponding floating-point representation of $\pi/4$ and $14/5$ if a) rounding is used; b) chopping is used?

Problem 1.4

Calculate the error, relative error and the number of significant digits in the following approximations $x_A \approx x_T$.

$$\begin{aligned}
&a) \quad x_A = 28.271, \quad x_T = 28.254;\\
&b) \quad x_A = 0.028271, \quad x_T = 0.028254;\\
&c) \quad x_A = 19/7, \quad x_T = e;\\
&d) \quad x_A = 1.414, \quad x_T = \sqrt{2}.
\end{aligned}$$

Problem 1.5

Avoid loss-of-significance errors in the following formulas

$$\begin{aligned}
&a) \quad \log(x+1) - \log(x) \quad \text{for large values of } x;\\
&b) \quad \frac{e^x - 1}{x} \quad \text{for small values of } x;\\
&c) \quad \sin(x+a) - \sin(a) \quad \text{for small values of } x;\\
&d) \quad \sqrt[3]{x+1} - \sqrt[3]{x} \quad \text{for large values of } x;\\
&e) \quad \sqrt{1 + \frac{1}{x} - 1} \quad \text{for large values of } x.
\end{aligned}$$

Problem 1.6

In the following function evaluations $f(x_A)$, assume the numbers x_A are correctly rounded to the number of digits shown. Bound the error $f(x_T) - f(x_A)$ and the relative error $Rel(x_A)$:

$$\begin{aligned}
&a) \quad cos(1.473);\\
&b) \quad e^{2.231};\\
&c) \quad \sqrt{0.0275};\\
&d) \quad \arctan(4.7869).
\end{aligned}$$

Problem 1.7

Extra Bonus point Bound

$$\int_0^{\pi} \frac{t^2}{1+t^4}\, dt - \int_0^{22/7} \frac{t^2}{1+t^4}\, dt.$$

Hint: Define function

$$f(x) = \int_0^{x} \frac{t^2}{1+t^4}\, dt.$$

and then bound $f(\pi) - f(22/7)$.