ASE-4046 Exercise 1 (Computer Arithmetic)

Problem 1

Find the result when the following arithmetic operations are computed on a processor that uses the floating point number system $(\beta, t, L, U) = (10, 3, -9, 9)$.

(a) $23.1454545 + 0.232976$

(b) $23.1454545/0.232976$

Problem 2

Reformulate or approximate the following MATLAB expressions to avoid subtraction of nearly equal quantities when $x \simeq 0$. Compute their values for the specified x value.

(a) `exp(x)-exp(-x)` $\quad x = 10^{-5}$

(b) `1/(sqrt(1+x^2)-sqrt(1-x^2))` $\quad x = 10^{-2}$

Problem 3

An *extended double* binary floating point number has $t = 63$ bits in the mantissa and exponent range $(L, U) = (-2^{14}, 2^{14})$. What are

(a) the unit roundoff μ

(b) `realmin`, the smallest positive number that can be represented as a normalised floating point number

(c) `realmax`, the largest number that can be represented as a normalised floating point number

Problem 4

The code

```
d = sqrt(b^2-4*a*c);
r1 = -(b-d)/(2*a);
r2 = -(b+d)/(2*a);
```

is supposed to compute the roots of a quadratic polynomial $ax^2 + bx + c$. It *fails* to accurately compute the roots of the polynomial $x^2 - 10^9 x + 1$, which has 2 positive real roots. Fix the code.

Answers 1. (a) 2.338×10^1, (b) 9.936×10^1
2. (a) `2.000000000033334e-05`, (b) `9.999999987499998e+03`
3. (a) 5.421×10^{-20}, (b) 8.405×10^{-4933}, (c) 2.379×10^{4932}