Homework 1

Due February 16, 19.00

Problem 1.1

All readings are posted on the course web page.

a) Read Lectures 0, 1.1, 1.2, 1.3.

b) Review how to do binary to decimal and vice-versa conversions.

c) Read definitions of Numerical Analysis by K. Atkinson and L. Trethenen.

d) Read the history of IEEE standard 754.

e) Read the paper on some common bugs related to computer representation of numbers.

Problem 1.2

Write the binary single precision IEEE floating-point expression for the number 12.1875. Specify sign σ, exponent E and mantissa.

Problem 1.3

Some microcomputers in the past used a binary floating-point format with 7 bits for the exponent and 1 bit for the sign σ. The mantissa contained 16 bits, with no hiding of the leading bit 1. The arithmetic used chopping. Determine the accuracy of the representation by finding the following:

a) machine epsilon;

b) integer M;

c) accuracy of the chopping operation.

Problem 1.4

Consider a binary floating-point representation with mantissa containing 3 digits without hiding the leading 1 and $-3_{10} \leqslant e \leqslant 3_{10}$.

a) List all numbers that can be stored exactly together with their decimal value.

b) Plot these numbers on real axis.

c) For this arithmetic, specify what are the corresponding floating-point representation of $\pi/3$ and $12/7$ if rounding is used.

d) Repeat c) if chopping is being used.

Problem 1.5

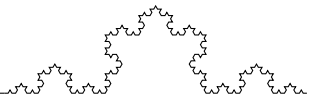
Calculate the error, relative error and the number of significant digits in the following approximations $x_A \approx x_T$.

a) $x_A = 6435.4012, \quad x_T = 6435.401163;$

b) $x_A = 0.007245, \quad x_T = 0.00723816;$

c) $x_A = 355/113, \quad x_T = \pi;$

a) $x_A = 2.236, \quad x_T = \sqrt{5}.$



Problem 1.6
Avoid loss-of-significance errors in the following formulas

$a)$ $\log(x) - \log(x-1)$ for large values of x;
$b)$ $\dfrac{e^x - 1}{x}$ for small values of x;
$c)$ $\cos(x+a) - \cos(a)$ for small values of x;

Problem 1.7
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 $Rel(x_A)$:

$a)$ $\sin(0.521)$;
$b)$ $e^{3.22}$;
$c)$ $\sqrt{0.0011}$;
$d)$ $\arcsin(0.5)$.