*Numerical Methods in Engineering and Applied Science*

Homework 1

**By Buchnev Arseniy**

1. **Check if the following equality holds using MATLAB:**

**0.1 + 0.2 == 0.3**

**Try to represent distinct numbers using fprintf(‘%.20e’, x) and explain the result.**

**Solution:**

**Graphical user interface, text, application

Description automatically generated**

**This happens due to floating point numbers having finite accuracy in n-bit representation. The number represented in computer is the closest floating-point number that can be written in binary:**

**Text, letter

Description automatically generated**

**Obviously, this combination will not hold the truth check.**

1. **Check the associativity of summation:**

**(0.1 + 0.2) + 0.3 == 0.1 + (0.2 + 0.3)**

**Find the exponents of these numbers in decimal format for 32-bit float according to IEEE Standard with the help of online converter. Explain the loss of accuracy during floating point addition on binary level. Is the loss of accuracy always guaranteed?**

**Solution:**

**As it may seem obvious,**

**A picture containing text

Description automatically generated.**

**Let us check the binary repr. online:**

**0.1 Graphical user interface, text, application, email

Description automatically generated**

**0.2 Graphical user interface, text, application, email

Description automatically generated**

**0.3 Graphical user interface, text, application, email

Description automatically generated**

**0.5 (= 0.2 + 0.3)**

**Graphical user interface, application

Description automatically generated with medium confidence**

**Graphical user interface

Description automatically generated with medium confidence**

**As we can see, the loss of accuracy won’t occur, if the real number can be represented exactly as a combination of powers of 2, which 0.5 obviously can (2­-1). In the example above we have (0.2 + 0.3) which wraps into 0.5, which, in turn, is represented exactly. Not so for (0.1 + 0.2) = 0.3.**

1. **Check if the following equalities hold using MATLAB:**

**(253 + 1) - 253 == 1**

**(253 + 2) - 253 == 2**

**Solution:**

**We can find the answer in the IEEE 754 standard before running the code:**

**A screenshot of a computer

Description automatically generated**

**Thus, our 253 + 1 will be rounded to a multiple of 2, causing the equality to not hold:**

**Graphical user interface, text, application

Description automatically generated**