Floating-point numbers

Ágnes Baran, Csaba Noszály

How can we trust in machine computations?

Exercise 1

Examine the value of the (logical) expression: $0.4 - 0.5 + 0.1 == 0$.

What is the value of $0.1 - 0.5 + 0.4 == 0$?

Exercise 2

What is the theoretical (expected) value of x after performing the following algorithm:

```
x=1/3;
for i=1:40
    x=4*x-1;
end
```

Exercise 3

Examine values of the following expressions:

$$2^{66} + 1 == 2^{66},\ 2^{66} + 100 == 2^{66},\ 2^{66} + 10000 == 2^{66}$$

Exercise 4

What are the results of algorithms below?

```
a=0;
for i=1:5
    a=a+0.2;
end
a==1
```

```
a=1;
for i=1:5
    a=a-0.2;
end
a==0
```

Try to explain!

Exercise 5

(a) Write a code that computes the machine epsilon!

(b) Read the help of the function `eps`! What is the value of eps(1)?

Exercise 6

(a) Write a code that computes ε_0!

(b) What is the value of `eps(0)`?

Exercise 7

Examine the values of `realmin` and `realmax`! What is `realmin('single')` and `realmax('single')`?

For a given a, t, k_+, k_- the floating-point numbers is finite subset of the real interval $[-M_\infty, M_\infty]$

Exercise 8

Let $a = 2,\ t = 4,\ k_- = -3,\ k_+ = 2.$

(a) Draw all positive (normalized) numbers from the system!

(b) What is the value of M_∞, ε_0 és ε_1?

(c) What is the distance of two neighbouring numbers?

Exercise 9

Examine again the values of the following expressions:

$$2^{66} + 1 == 2^{66},\ 2^{66} + 10 == 2^{66},\ 2^{66} + 100 == 2^{66},$$
$$2^{66} + 1000 == 2^{66},\ 2^{66} + 10000 == 2^{66}$$

Try to find the smallest $n > 0$ for which $2^{66} + n == 2^{66}$ is `false`! What is the value of `eps(2^66)`?

Exercise 10

Let $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Compute the corresponding floating point numbers for:

$$0.4, \quad 0.3, \quad \frac{1}{3}, \quad 0.7, \quad \frac{1}{32}$$

Exercise 11

Examine the value of expression $0.4 - 0.5 + 0.1 == 0$! Explain! Examine the value of expression $0.1 - 0.5 + 0.4 == 0$! Explain!

Exercise 12

Let $a = 2$, $t = 4$, $k_- = -3$, $k_+ = 2$. Try to find positive $x \neq y$ floating point numbers, for which:

(f) $x + y < M_\infty$, but $x + y$ is not a floating point number.

(g) $fl(x + y) = x$.

Exercise 13

What will be the value of x after executing the code below?

```
x=1/3;
for i=1:40
    x=4*x-1;
end
```

Why is so different what we see?

Exercise 14

The code below modifies and restores the value of x by successive squarerooting and squareing. In theory x remains the same. What we see in practice? Why?

```
for i=1:60
    x=sqrt(x);
end
for i=1:60
    x=x^2;
end
```