Tasks

1. For each MATLAB command below, write and explain what will be printed on the screen after it is executed:

(a) 1/0-1/0

(b) (1 + eps + eps^2 == 1 + eps^2 + eps)

(c) realmin/2^55

(d) 1-eps/2+eps/2

(e) realmax('single ')+realmax('double')

(f) double(realmax('single'))+realmax('double')

(g) Especially here, write down what the variables a, b, c, d will contain after the commands are executed:

a=(10+eps+eps^2==10+eps^2+eps);

b=(1/0)*(2/Inf);

c=(eps*eps^2==eps^3); d = (NaN==NaN)

2. Explain why double-precision floating-point numbers are more than 2 times more accurate than single-precision arithmetic?

3. Construct a MATLAB function myfloats(ekth,man,fp_type) that takes as input an integer ekth, a vector mantis with elements {0,1} and a third argument (string) fp_type with only allowed values fp64, fp32, fp16, bfloat16. The function checks the input data against fp_type and if the length of the vector and the size of the exponent are valid, returns in decimal form the floating-point numbers corresponding to the selected (from the string) representation. Otherwise, an informational message is printed indicating the reason for the failure (e.g. exponent out of range, mantissa length incompatible, fp type invalid).

4. Given the floating-point numbers system with base 2, minimum exponent -3, maximum exponent 4, and 4 bits for the tail (mantissa) of which the first (most significant) corresponds to 20 and the rest to 21 to 23. What is the maximum number that can be represented? What is the minimum normalized number that can be represented?

5. Calculate the Jacobian register of the function f ([x, y, z]) = x2 + xy + 2z.

6. Calculate the Jacobian register of the function f ([x, y, z]) = [2x + 3y - 1, 3x - 2y - 2]⊤. Generalize and compute the Jacobian register of the function f (x) = Ax - b for given A ∈ Rnn, x, b ∈ Rn.

7. Show that the calculation (x+y)+z is back stable. Then show that the 0 returned by the calculation 10^20+10-10^20 in MATLAB is the exact result of the same operation performed in infinite precision arithmetic on somewhat different data. In particular, find an example of such data.

8. Find the condition index of the functions ax, sin(x), ex, x/(x + a) (see the answers at https:/ en.wikipedia.org/wiki/Condition_number).

9. Show that computing the value of a polynomial αnxn + - - - - α1x + α0 for α0, ..., αn and x by the Horner algorithm is back stable. For convenience, show it for n = 3 and then you can generalize.

10. Show that the calculation of the function

$$f([a,b,c,d,e,f]) = \begin{pmatrix} ad & ae & af \\ bd & be & bf \\ cd & ce & cf \end{pmatrix}$$

for real a, b, c, d cannot be back stable.

11. Consider suitable ways to calculate e^x - 1 for small values of |x|. Warning: Investigate the values computed by exp(x)-1, expm1(x) and the first terms of the Taylor series for ex - 1 for various values of x ∈ (0, 1]. In particular, look at the values for x = 10p for several values of p = 1, 2, ..., 40. What do you observe?