

1. The standard double precision floating point numbers can be written down as

$$x = \pm(1.b_1b_2\dots b_{52})_2 \times 2^q = \pm \left(1 + \frac{b_1}{2} + \frac{b_2}{4} \dots + \frac{b_{52}}{2^{52}}\right) \times 2^q$$

for some integer  $q$  (which is also restricted, but this does not concern us here), where  $b_j \in \{0, 1\}$  for  $j = 1, 2, \dots, 52$ . Find  $\epsilon_1 > 0$  and  $\epsilon_2 > 0$ , such that

$$fl(y) = 2, \quad 2 - \epsilon_1 < y < 2 + \epsilon_2.$$

2. Derive the following formula:

$$\pi = 16 \arctan\left(\frac{1}{5}\right) - 4 \arctan\left(\frac{1}{239}\right).$$

Write a MATLAB program for computing  $\arctan(x)$  by Taylor expansion, and use it to calculate  $\pi$  (also in MATLAB). Submit your programs and the MATLAB results.

3. When you subtract two nearly equal floating point numbers, the result will not have full precision, i.e., there will be less correct digits. This is called “loss of significance” caused by subtraction. For each of the following functions, find the values of  $x$  at which there will be loss of significance, and suggest how to avoid loss of significance.

(a)  $e^x - \sin(x) - \cos(x),$

(b)  $(\cos(x) - e^{-x})/\sin(x),$

(c)  $\frac{\sqrt{1+x^2}-1}{x^2} - \frac{x^2 \sin(x)}{x - \tan(x)},$

(d)  $\frac{e^{2x}-1}{2x}.$

4. Let  $p_n = \int_0^1 x^n e^x dx$  for  $n = 1, 2, \dots$ , show that  $1 = p_1 > p_2 > p_3 > \dots > 0$  and

$$p_{n+1} = e - (n+1)p_n.$$

Write a MATLAB program for computing  $p_n$  for  $1 \leq n \leq 20$ . Submit the MATLAB program and results. Explain your results.