Homework1 for SI211: Numerical Analysis

Mao Song
SIST
ShanghaiTech University
Student ID: 2019233134
maosong@shanghaitech.edu.cn

Abstract

This is the solution for Homework 1 of SI211: Numerical Analysis, which is taught by Boris Houska.

1 Floating point numbers

What is the bit representations of the floating point number 5.25 using the *IEEE standard* for double precision numbers? Solve this problem with pen and paper first. Use a computer program to verify your result.

Solution 1. Since $5.25 > 0.0$, the sign indicator is 0. Now that $5 = (101)_{(2)}$, and $0.25 = (0.010)_2$, so

$$5.25 = (101.010)_{(2)} = (1.01010 \times 2^2)_{(2)} \tag{1}$$

thus the characteristic is $1023 + 2 = 1025 = (10000000001)_{(2)}$, the mantissa is

$$.0101000000000000000000000000000000000000000000000000 \tag{2}$$

Combine the above results, the number 5.25 is represented as

$$0\ 10000000001\ 0101000000000000000000000000000000000000000000000000 \tag{3}$$

We can use Julia function *bitstring(5.25)* to verify the above result.

2 Numerical evaluation error

Evaluate the function

$$f(x) = \frac{1 - \cos(x)}{x^2} \tag{4}$$

with a compute program of your choice using the standard *IEEE double precision floating point format*. Plot the numerical result on a logarithmic scale for $x \in \left[10^{-15}, 10^{-1}\right]$.

Solution 2. We use MATLAB to solve this problem, first we sample 10000 points uniformly distributed in $\left[10^{-15}, 10^{-1}\right]$, then we calculate $f(x)$ directly from the definition, we then obtain the following results, see Fig 2.

3 Taylor expansion

Consider again the function

$$f(x) = \frac{1 - \cos(x)}{x^2} \tag{5}$$

from the first homework problem. Can you approximate f by using a Taylor approximation? Does this help you to evaluate f with higher accuracy?

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

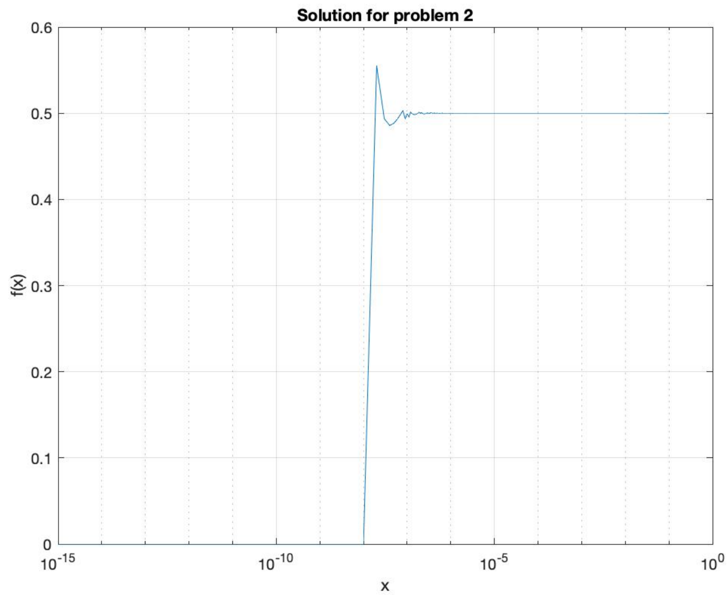

Figure 1: The solution to problem 2

Solution 3. We consider the Maclaurin expansion of $\cos(x)$, which is given by

$$\cos(x) = 1 + \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} \tag{6}$$

thus

$$f(x) = \frac{1}{2} - \frac{1}{24}x^2 + \frac{1}{720}x^4 - \cdots \tag{7}$$

We use MATLAB to create a plot with first n terms of $f(x)$ to see the impact of n ,see Fig 3. Since

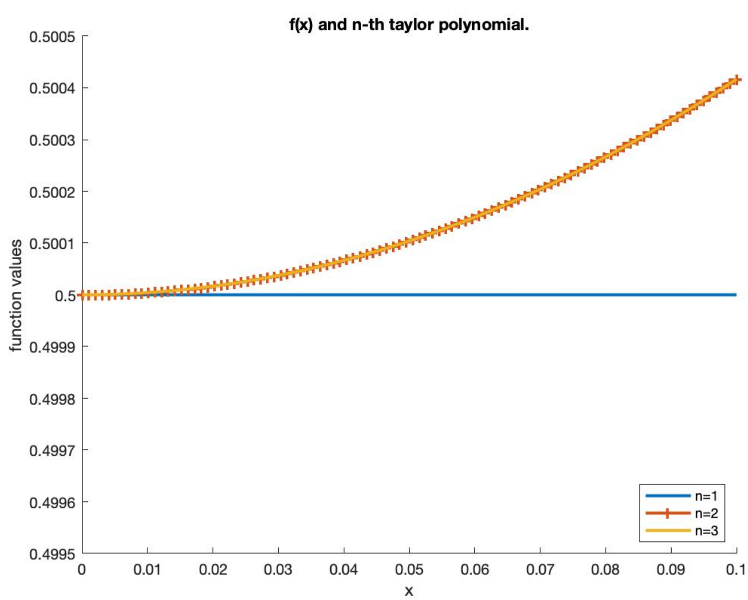

Figure 2: The solution to problem 3

we have

$$\lim_{x\to 0} f(x) = \min_{x\to 0} \frac{1-\cos(x)}{x} = \frac{1}{2} \tag{8}$$

As the Figure shows, the result is numerically more stable than directly computing the function value, but it tends to lose this property when x becomes large.

4 Numerical differentiation

Consider the five-point differentiation formula

$$f'(x) \approx \frac{1}{12h}\left[-25f(x_0) + 48f(x_0+h) - 36f(x_0+2h) + 16f(x_0+3h) - 3f(x_0+4h)\right] \tag{9}$$

1. What is the mathematical approximation error of this formula?
2. For which values of h would you expect that this formula leads to a minimum approximation error taking both the mathematical as well as the numerical approximation error into account?
3. Implement the above differentiation formula in a compute program of your choice and use it to find the derivative of the test function $f(x) = e^x$ at $x = 1$. Plot the total derivative evaluation error as a function of h and intepret your results.

Solution 4. The solution is given as follows:

1. First, we expand 5-th Taylor polynomial for $f(x_0 + ih), i = 1, 2, 3, 4$:

$$f(x_0+ih) = f(x_0) + if'(x_0)h + \frac{i^2 f''(x_0)}{2}h^2 + \frac{i^3 f^{(3)}(x_0)}{6}h^3 \tag{10}$$

$$+ \frac{i^4 f^{(4)}(x_0)}{24}h^4 + \frac{i^5 f^{(5)}(x_0)}{120}h^5 + o(h^6), i = 1,2,3,4 \tag{11}$$

   substitute the above equations into the approximation of $f'(x)$, we have

$$f'(x) \approx \frac{1}{12h}\left[12f'(x_0)h + \frac{12}{5}f^{(5)}(x_0)h^5 + o(h^6)\right] = f'(x_0) + \frac{1}{5}f^{(5)}(x_0)h^4 + o(h^5) \tag{12}$$

   reiwrite the above equation in the form of Taylor theorem, we have

$$f'(x) = f'(x_0) + \frac{1}{5}f^{(5)}(\xi(x))h^4 \tag{13}$$

   where $\xi(x)$ is between x and $x + 4h$. Thus the truncation error is $o(h^4)$.

2. If f is well conditioned, we choose

$$h^* \approx \arg\min_h \left(h^4 + \frac{eps}{h}\right) = (eps)^{1/5} \tag{14}$$

3. We use the rounding manner , which leaves 9 digits to calculate the values of $f(x)$, thus the round-off error is bounded by $eps = 5 \times 10^{-10}$. Now our best h^* is given by

$$h^* \approx (eps)^{1/5} \approx 0.01 \tag{15}$$

   The result is now shown as in Fig 3, from which we can see that the result is consistent with the theoritical analysis.

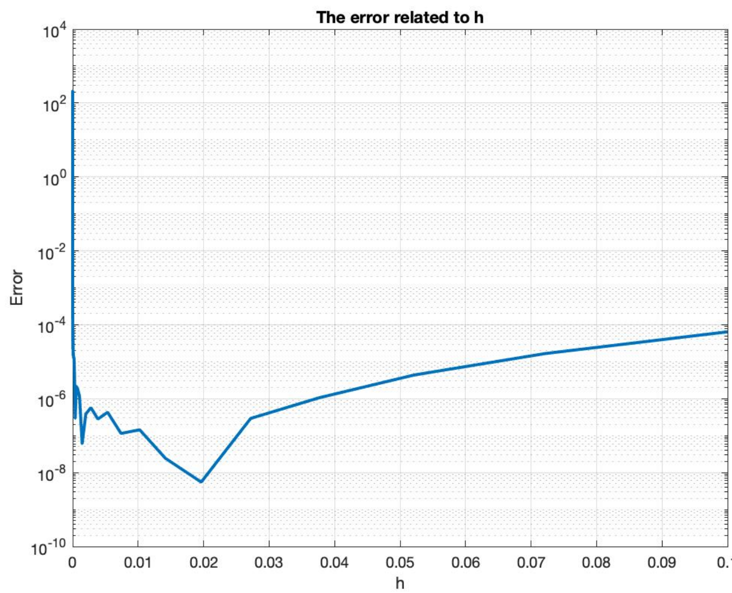

Figure 3: The solution to problem 4