Digital version

CS112 Numerical Analysis - Homework 1

Mher Saribekyan A09210183

February 6, 2026

Problem 1

Assume $x = \frac{4040}{9}$ (in the base-10 representation).

a) Write x in the normalized form;

$$x = \frac{4040}{9} = 0.44\overline{8}... \times 10^3$$

b) Find the floating-point approximations of x, $fl(x)$, using both the $k = 6$ digit chopping and rounding methods. Estimate the absolute and relative errors of these approximations.

$$fl(x) = 0.448889 \times 10^3$$

$$abserr(x, fl(x)) = |x - fl(x)| = 0.1 \times 10^{-2}$$

$$abserr(x, fl(x)) = \frac{|x - fl(x)|}{|x|} = \frac{0.1 \times 10^{-2}}{\frac{4040}{9}} \approx 0.223 \times 10^{-5}$$

Problem 2

Assume numbers are represented in a floating point one-digit normalized decimal form using chopping, where the exponent can be either -1, 0, or 1.

a) List all machine numbers (including 0).

$$0.d_1, \times 10^n, \text{ where } d_1 \in 0, 1, ..., 9 \text{ and } n \in -1, 0, 1$$

$$0, 0.01, 0.02, ..., 0.09, 0.1, 0.2, 0.3, ...0.9, 1, 2, ..., 9$$

b) Describe all numbers for which overflow occurs. Explain!

The largest (absolute value) number we can store is $0.9 \times 10^1 = 9$, therefore for numbers $x > 9$ and $x < -9$ it will overflow, as it cannot be stored with our given floating point format.

c) Describe all numbers for which underflow occurs. Explain!

The smallest (non-zero absolute value) number we can store is $0.1 \times 10^{-1} = 0.01$, therefore for numbers $x \in (0, 0.01)$ and $x \in (-0.01, 0)$ it will overflow, as it cannot be stored with our given floating point format.

Problem 3

Let $a = 0.121$, $b = 0.358$, $c = 0.3$. Assuming a 2-digit decimal floating point arithmetic with rounding is used,

$$(a + b) \cdot c \approx (a \oplus b) \otimes c$$

$$(a + b) \cdot c \approx (a \otimes c) \oplus (b \otimes c)$$

Are these approximations different? If your answer is Yes, choose the better one.

$$(a + b) \cdot c \approx (0.12 \oplus 0.36) \otimes 0.30 \approx 0.48 \otimes 0.30 \approx 0.14$$

$$(a + b) \cdot c \approx (0.12 \otimes 0.30) \oplus (0.36 \otimes 0.30) \approx 0.36 \times 10^{-1} \oplus 0.11 \approx 0.15$$

The first approximation has less steps, and therefore possibly better: $(a + b) \cdot c \approx 0.14$

Problem 4

Assume we use 4 digit rounding to present numbers in floating-point form and to perform machine operations (in the base-10 representation). Give an example of numbers a and b such that

a) $(a \oplus b) \oslash 2 \notin [a, b]$;

$$a := 0.10011, b := 0.10012 \implies (a \oplus b) \oslash 2 = (0.1001 \oplus 0.1001) \oslash 2 = 0.1001 \notin [0.10011, 0.10012]$$

b) $a > 2,\ b > 2$ and $a \ominus b = a$.

$$a := 0.9999 \times 10^5,\ b := 0.3 \times 10^1 \implies a \ominus b = 99990 - 3 = 99987 \approx 0.9999 \times 10^5$$

Problem 5

Identify for which values of x there is a subtraction of nearly equal numbers, and find an alternative form that avoids the mentioned problem.

a) $\sqrt{x^2 + 1} + x$

As $x \to -\infty$, we get $\sqrt{x^2 + 1} + x \approx \sqrt{x \cdot x} + x \approx |x| \oplus x \overset{x \leq 0}{=} x \ominus x \approx 0$

$$\sqrt{x^2 + 1} + x = \frac{(\sqrt{x^2 + 1} + x)(\sqrt{x^2 + 1} - x)}{\sqrt{x^2 + 1} - x} = \frac{1}{\sqrt{x^2 + 1} - x}$$

b) $e^x - e$

For $x \to 1$, we get $e^x - e \to e - e$.

$$e^x - e = e(e^{x-1} - 1)$$

Problem 6

We all know that

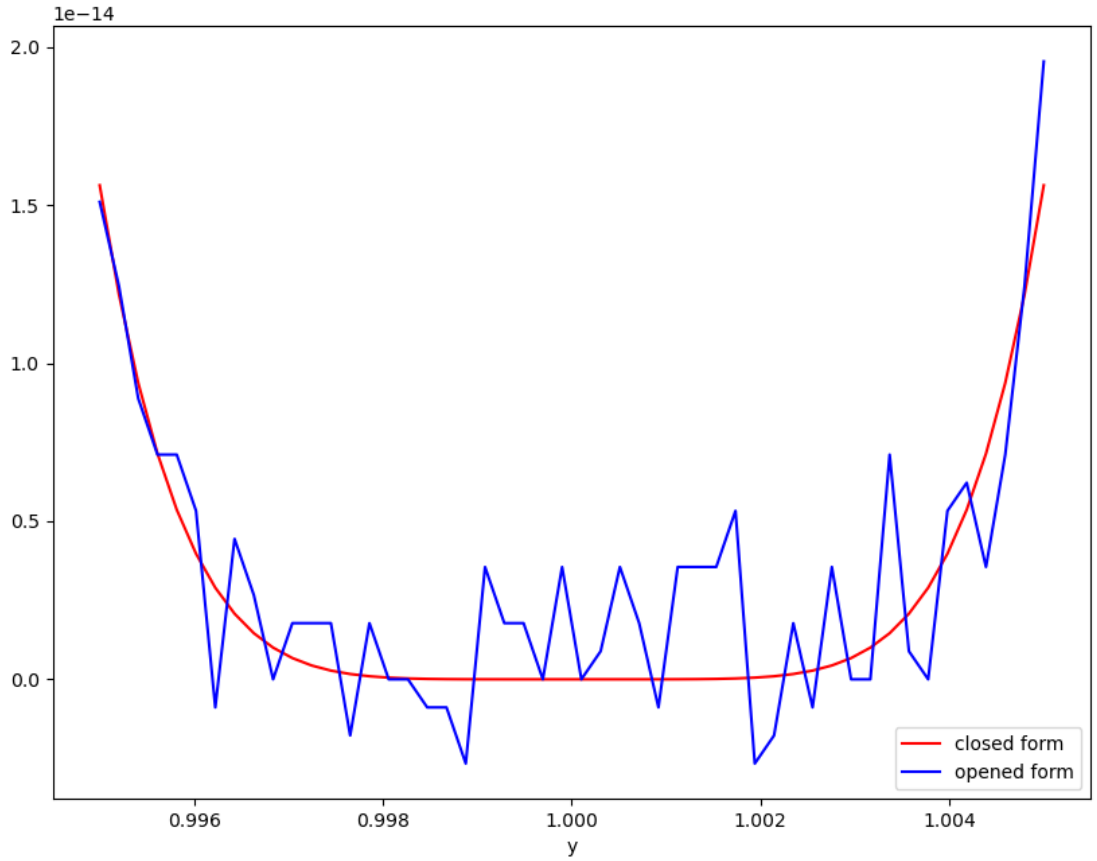
$$(x-1)^6 = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1.$$

a) Plot the graph of $P(x) = (x-1)^6$ on the interval $[0.995, 1.005]$.

b) Plot the graph of $P(x) = x^6 - 6x^5 + 15x^4 - 20x^3 + 15x^2 - 6x + 1$ on the interval $[0.995, 1.005]$.

c) Try to explain the difference between the results. Which one is better and why?

```
import numpy as np
import matplotlib.pyplot as plt

f1 = lambda x : (x - 1)**6
f2 = lambda x : (x**6 - 6*(x**5) + 15*(x**4) - 20*(x**3) + 15*(x**2) - 6*x + 1)

xs = np.linspace(0.995, 1.005)
f1s = f1(xs)
f2s = f2(xs)

plt.xlabel("x")
plt.xlabel("y")

plt.plot(xs, f1s, color="red")
plt.plot(xs, f2s, color="blue")

plt.legend(["closed form", "opened form"], loc="lower right")

plt.show()
```



The function in a) consists of one power function and one subtraction function, which results in a lower error than the function at b), which consists of multiple power functions, multiplications, additions and subtractions. The y value is in the order of 10^{-14}, and hence the floating point error is noticable when doing multiple calculations.