

Solution For Assignment 1

Mohan Zhang, 1002748716, morgan.zhang@mail.utoronto.ca

October 10, 2016

1 Question 1.

$T = 3.14159265358979$

(a)

$A = 3.1$

Approximate absolute errors $|A - T| = 0.04159265358979 \approx 0.0416 = 4.16 \times 10^{-2}$

Approximate relative errors $\frac{|A-T|}{T} = \frac{0.04159265358979}{3.14159265358979} = 0.013239352830247875 \approx 0.0132 = 1.32 \times 10^{-2}$

(b)

$A = 3.142$

Approximate absolute errors $|A - T| = 0.0004073464102098967 \approx 0.000407 = 4.07 \times 10^{-4}$

Approximate relative errors $\frac{|A-T|}{T} = 0.00012966238947128807 \approx 0.000130 = 1.30 \times 10^{-4}$

(c)

$A = 3.14159265$

Approximate absolute errors $|A - T| = 358979 \times 10^{-9} \approx 3.59 \times 10^{-9}$

Approximate relative errors $\frac{|A-T|}{T} = 1.1426655822646306 \times 10^{-9} \approx 1.14 \times 10^{-9}$

2 Question 2.

(a)

result $= 5.8427 \times 10^0 \approx 5.84 \times 10^0$

(b)

result $= 3.3524 \times 10^1 \approx 3.35 \times 10^1$

(c)

result $\approx 4.15 \times 10^1$

(d)

result $\approx -3.78 \times 10^6$

(e)

result $\approx 4.53 \times 10^{12}$

(f)

result = $5.703 \times 10^2 \approx 5.70 \times 10^2$

(g)

result = $9.158100000000001 \times 10^{-5} \approx 9.16 \times 10^{-5}$

(h)

result = $-1.1885593220338984 \times 10^{25} = -\text{Inf}$ (i)

result = $9.5821 \times 10^{-22} \approx 0.10 \times 10^{-20}$

(j)

result = $3.30792 \times 10^{-24} \approx 0$

3 Question 3.

for a small change h (h can be positive or negative, when $x+h$ is greater than or equal to 0) $f(x) = x^{1/4}$, $\text{Cond} = \frac{(f(x+h)-f(x))/f(x)}{(x+h-x)/x} = \frac{(f(x+h)-f(x))/f(x)}{h/x} = \frac{f(x+h)-f(x)}{h} \cdot \frac{x}{f(x)}$ So,

$\lim_{h \rightarrow 0} \text{Cond} = f'(x) \cdot \frac{x}{f(x)} = 1/4$, which is less than 1, in other words, not much bigger than 1. So, $f(x)$ is well-conditioned.

4 Question 4.

Claim: the statement is true. There is no rounding error.

(a)

For example, when convert 5 to the nearest IEEE floating-point number, we'll get $(1) \times 1.0100...0 \times 2^2$, there is no rounding error. the IEEE double precision floating point number has, 1 sign digit, 11 exponent digits, 52 fraction digits. So, between $(-1) \times 1.11...1 \times 2^{52}$ and $(1) \times 1.11...1 \times 2^{52}$, every integer can represent precisely, which are, as we can see, are $-(2^{53}-1)$ and $2^{53}-1$ respectively. So, $\text{fl}(m) = m$, provided that $|m| \leq 2^{53}-1$.

(b)

So, when ever we multiply two integer, we got a integer back. Whenever $|m \times n| \leq 2^{53} - 1$, the result can be represent by a IEEE double float number precisely by (a). So, we know that the conclusion is correct.

5 Question 5.

(a)

Please See the attached file.

(b)

$\exp1(x)$ approximates well when $x = -17:25$, but when $x = -25:-16$, the error is not insignificant. The reason, or, the rounding error, is because: