Numerical Analysis **Spring 2025**

Homework #2 Math and Floating-point Number Representation

Instructor: *Kwangki Kim*

Name : 장준영

Student ID# : 12191529

Problem 1.1 *(FORMAL TAYLOR SERIES FOR f ABOUT c)* What is the Taylor series of the function

$$f(x) = 3x^5 - 2x^4 + 15x^3 + 13x^2 - 12x - 5$$

at the point $c = 2$?

sol) $f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(c)}{n!}(x-c)^n$

① $f(2) = 3\cdot 32 - 2\cdot 16 + 15\cdot 8 + 13\cdot 4 - 12\cdot 2 - 5 = 207$

② $f'(x) = 15x^4 - 8x^3 + 45x^2 + 26x - 12$

$\rightarrow f'(2) = 15\cdot 16 - 8\cdot 8 + 45\cdot 4 + 26\cdot 2 - 12 = 396$

③ $f''(x) = 60x^3 - 24x^2 + 90x + 26$

$\rightarrow f''(2) = 60\cdot 8 - 24\cdot 4 + 90\cdot 2 + 26 = 590$

④ $f^{(3)}(x) = 180x^2 - 48x + 90$

$\rightarrow f^{(3)}(2) = 180\cdot 4 - 48\cdot 2 + 90 = 714$

⑤ $f^{(4)}(x) = 360x - 48$

$\rightarrow f^{(4)}(2) = 720 - 48 = 672$

⑥ $f^{(5)}(x) = 360$

$\rightarrow f^{(5)}(2) = 360$

⑦ $f^{(6)}(x) = 0$

⋮

$$\Rightarrow f_T(x) = f(2) + f'(2)(x-2) + \frac{f''(2)}{2!}(x-2)^2 + \frac{f^{(3)}(2)}{3!}(x-2)^3 + \frac{f^{(4)}(2)}{4!}(x-2)^4 + \frac{f^{(5)}(2)}{5!}(x-2)^5$$

$$\boxed{= 207 + 396(x-2) + \frac{590}{2!}(x-2)^2 + \frac{714}{3!}(x-2)^3 + \frac{672}{4!}(x-2)^4 + \frac{360}{5!}(x-2)^5}$$

Problem 1.2 *(TAYLOR'S THEOREM FOR $f(x)$)*

(a) Derive the Taylor series for e^x at $c = 0$, and prove that it converges to e^x by using Taylor's Theorem.

(b) Derive the formal Taylor series for $f(x) = \ln(1+x)$ at $c = 0$, and determine the range of positive x for which the series represents the function.

(a) $f(x) = e^x$ 라 하면,

$f'(x) = e^x$; $f''(x) = e^x$ ...

$$f_T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \boxed{\sum_{n=0}^{\infty} \frac{1}{n!} x^n}$$

(증명) $R_{n+1}(x) = \dfrac{f^{(n+1)}(\xi)}{(n+1)!} x^{n+1}$ $(\because \xi \in (0, x))$

여기서 $f^{(n+1)}(\xi) = e^{\xi} \le e^{|x|}$

따라서 $|R_{n+1}(x)| \le \dfrac{e^{|x|} x^{n+1}}{(n+1)!}$

$\therefore \displaystyle\lim_{n\to\infty} \frac{e^{|x|} x^{n+1}}{(n+1)!} = 0$ 이므로 $e^x = \displaystyle\sum_{n=0}^{\infty} \frac{x^n}{n!}$

(b) $f(x) = \ln(1+x)$

$f'(x) = \dfrac{1}{1+x} \rightarrow f'(0) = 1$

$f''(x) = -\dfrac{1}{(1+x)^2} \rightarrow f''(0) = -1$

$f^{(3)}(x) = \dfrac{2}{(1+x)^3} \rightarrow f^{(3)}(0) = 2$

$\vdots$

$f^{(n)}(x) = \dfrac{(-1)^{n-1}}{(1+x)^n}(n-1)! \rightarrow f^{(n)}(0) = (-1)^{n-1}(n-1)!$

$$\therefore f_T(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n = \boxed{\sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} x^n}$$

$\boxed{\therefore (\text{수렴}) : 0 < x \le 1}$

Problem 1.3 *(A technique of calculating the number of terms to be used in a series by just making the $(n+1)$st term less than some tolerance)* If the logarithmic series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots = \sum_{k=1}^{\infty}(-1)^{k-1}\frac{x^k}{k} \quad (-1 < x \leqq 1)$$

is to be used for computing $\underbrace{\ln 2}_{x=1}$ with an error of less than $\frac{1}{2} \times 10^{-6}$, how many terms will be required?

sol) $\ln(1+x) = \sum_{k=1}^{n} (-1)^{k-1} \frac{x^k}{k} + (-1)^n \frac{x^{n+1}}{n+1}$

$$\therefore R_{n+1}(x) = \left|(-1)^n \frac{x^{n+1}}{n+1}\right| = \left|\frac{x^{n+1}}{n+1}\right| \quad (-1 < x \leq 1)$$

$$\leq \frac{1}{n+1} < \frac{1}{2} \times 10^{-6} \quad ; \; n > 1{,}999{,}999$$

$\boxed{\text{답} : 2{,}000{,}000\text{개 이상}}$

Problem 1.4 *(Floating-Point Representation and Errors)* If x, y, and z are machine numbers in a 32-bit word-length computer, what upper bound can be given for the relative roundoff error in computing $z(x+y)$?

sol) 32 bit = 1 (sign) + 8 (exponent) + 23 (mantissa)

1개의 hidden bit 가 존재한다고 가정하자.

총 유효숫자의 개수는 24개 이므로, 1개 연산에서의 오차 $\epsilon \le 2^{-24}$

연산은 총 2번 이루어진다:

① $x+y \rightarrow |\epsilon_1| \le 2^{-24}$

② $z(x+y) \rightarrow |\epsilon_2| \le 2^{-24}$

따라서 전체 상대오차는 대략 $|\epsilon| \le 2 \cdot 2^{-24} = 2^{-23}$ 의 upper bound를 갖는다고 볼 수 있다.

Problem 1.5 *(Floating-Point Representation and Errors)* Critique the following attempt to estimate the relative roundoff error in computing the sum of two real numbers, x and y. In a 32-bit word-length computer, the calculation yields

$$\begin{aligned} z &= \mathrm{fl}[\mathrm{fl}(x) + \mathrm{fl}(y)] \\ &= [x(1+\delta) + y(1+\delta)](1+\delta) \\ &= (x+y)(1+\delta)^2 \\ &\approx (x+y)(1+2\delta) \end{aligned}$$

Therefore, the relative error is bounded as follows:

$$\left| \frac{(x+y)-z}{(x+y)} \right| = \left| \frac{2\delta(x+y)}{(x+y)} \right| = |2\delta| \leqq 2^{-23}$$

Why is this calculation not correct?

sol) 위 계산이 정확하지 않은 이유는 다음과 같다.

- 모든 Roundoff Error를 하나의 δ로 가정함.
- 그러나, 현실에서는 연산의 각 단계마다 다른 오차가 발생한다.
  - $fl(x) = x(1+\delta_1)$
  - $fl(y) = y(1+\delta_2)$
  - $fl(\text{sum}) = \{x(1+\delta_1) + y(1+\delta_2)\}(1+\delta_3) = z$

즉, $z = \{x(1+\delta_1) + y(1+\delta_2)\}(1+\delta_2) = \{(x+y) + (\delta_1 x + \delta_2 y)\}(1+\delta_3)$

따라서 상대오차는:

$$\left| \frac{(x+y)-z}{(x+y)} \right| = \left| \frac{(x+y)\delta_3 + \delta_1 x(1+\delta_3) + \delta_2 y(1+\delta_3)}{(x+y)} \right| = \left| 1 + \frac{\delta_1 x + \delta_2 y}{x+y}(1+\delta_3) \right|$$

- $|\delta_i| < \epsilon$: $|\delta_1 x + \delta_2 y| \le \epsilon(|x|+|y|)$

$$\therefore \left| 1 + \frac{\delta_1 x + \delta_2 y}{x+y}(1+\delta_3) \right| \le \varepsilon + \varepsilon(1+\varepsilon)\cdot\frac{|x|+|y|}{|x+y|}$$

- 일반적인 경우 : $x \simeq y$

$$\therefore \varepsilon + \varepsilon(1+\varepsilon)\cdot\frac{2x}{2x} = \varepsilon+\varepsilon+\varepsilon^2 \simeq 2\varepsilon$$

Problem 1.6 *(Floating-point numbers)* Identify the floating-point numbers corresponding to the following bit strings:

1 8 23

(a) 00000000000000000000000000000000

(b) 10000000000000000000000000000000

(c) 01111111100000000000000000000000

(d) 11111111100000000000000000000000

(e) 00000000100000000000000000000000

(f) 01000000101100000000000000000000

(g) 00111111100000000000000000000000

(h) 00111101110011001100110011001100

■ 구조 : $(-1)^s \times (1.f) \times 2^{c-127}$

(a) 0
- 부호: 0 (양수)
- 지수 : 0
- 가수 : $(1.0\cdots)_2 = 1$

(b) 0
- 부호: 1 (음수)
- 지수:
- 가수: $(1.0\cdots)_2 = 1$

(c) $+\infty$
- 부호: 0 (양수)
- 지수: $(11111111)_2 = 255 \to c-127 = 128$
- 가수: $(1.0\cdots)_2 = 1$

(d) $-\infty$
- 부호 : 1 (음수)
- 지수: $(1111\ 1111)_2 = 255 \to c-127 = 128$
- 가수: $(1.0\cdots)_2 = 1$

(e) 2^{-126} (가장 작은 정규화된 양의 실수)
- 부호: 0 (양수)
- 지수: $(0000\ 0001)_2 = 1 \to c-127 = -126$
- 가수: $(1.0\cdots)_2 = 1$

(f) 5.5
- 부호 : 0 (양수)
- 지수 : $(1000\ 0001)_2 = 129 \to c-127 = 2$
- 가수: $(1.011)_2 = 1.375$

$\to 1.375 \times 2^2 = 5.5$

(g) 1
- 부호: 0 (양수)
- 지수: $(0111\ 1111)_2 = 127 \to c-127 = 0$
- 가수: $(1.0\cdots)_2 = 1$

$\to 1 \times 2^0 = 1$

(h) 0.1
- 부호: 0 (양수)
- 지수: $(0111\ 1011)_2 = 123 \to c-127 = -4$
- 가수: $(1.1001100110011001\cdots)_2$

$\to (1.1001100110011001\cdots)_2 \times 2^{-4} = (0.0001100110011001\cdots)_2 = 0.1$

Problem 1.7 *(Exactly how many significant binary digits are lost in the subtraction x`y when x is close to y?)* In the subtraction 37.593621−37.584216, how many bits of significance will be lost?

(Hint) Let x and y be normalized floating-point machine numbers, where $x > y > 0$. If $2^{-p} \leqq 1 - (y/x) \leqq 2^{-q}$ for some positive integers p and q, then at most p and at least q significant binary bits are lost in the subtraction $x - y$. **Proof:** We prove the second part of the theorem and leave the first as an exercise. To this end, let $x = r \times 2^n$ and $y = s \times 2^m$, where $\frac{1}{2} \leqq r, s < 1$. (This is the normalized binary floating-point form.) Since $y < x$, the computer may have to shift y before carrying out the subtraction. In any case, y must first be expressed with the same exponent as x. Hence, $y = (s2^{m-n}) \times 2^n$ and

$$x - y = \left(r - s2^{m-n}\right) \times 2^n$$

The mantissa of this number satisfies the equations and inequality

$$r - s2^{m-n} = r\left(1 - \frac{s2^m}{r2^n}\right) = r\left(1 - \frac{y}{x}\right) < 2^{-q}$$

Hence, to normalize the representation of $x-y$, a shift of at least q bits to the left is necessary. Then at least q (spurious) zeros are supplied on the right-hand end of the mantissa. This means that at least q bits of precision have been lost. ■

가짜

sol) $x = 37.593621$, $y = 37.584216$ 이라 하자.

$$1 - \frac{y}{x} = 1 - \frac{37.584216}{37.593621} = 0.0002501754$$

$$\therefore 2^{-p} \leq 0.0002501754 \leq 2^{-q}$$

$-\log_2(0.0002501754) \simeq 11.96$ 이므로 $p = 12,\ q = 11$

$\boxed{\text{답 : 최소 11 비트, 최대 12비트 손실}}$