

CSC336: Assignment 1

Junjie Cheng, 1002770539

October 12 2017

1. (a) Absolute error = $A - T = 2.72 - 2.71828182845905 = 1.72 * 10^{-3}$
Relative error = $\frac{A-T}{T} = 6.32 * 10^{-4}$
(b) Absolute error = $A - T = 2.718 - 2.71828182845905 = -2.82 * 10^{-4}$
Relative error = $\frac{A-T}{T} = -1.04 * 10^{-4}$
(c) Absolute error = $A - T = 2.71828183 - 2.71828182845905 = 1.54 * 10^{-9}$
Relative error = $\frac{A-T}{T} = 5.67 * 10^{-10}$
2. (a) $4.21 * 10^0 + 5.47 * 10^{-2} = 4.2647 * 10^0$. It will be rounded to $4.26 * 10^0$
(b) $6.52 * 10^1 - 7.27 * 10^1 0^{-1} = 6.4473 * 10^1$. It will be rounded to $6.45 * 10^1$
(c) $5.61 * 10^1 + 6.67 * 10^{-4} = 5.6100667 * 10^1$. It will be rounded to $5.61 * 10^1$
(d) $4.52 * 10^4 - 3.82 * 10^6 = -3.7748 * 10^6$. It will be rounded to $-3.77 * 10^6$
(e) $7.51 * 10^{12} - 5.25 * 10^5$ will be rounded to $7.51 * 10^{12}$
(f) $3.82 * 10^1 + 8.42 * 10^2 = 8.803 * 10^2$. It will be rounded to $8.80 * 10^2$
(g) $4.47 * 10^{10} * 5.81 * 10^{15} = 2.59707 * 10^{-4}$. It will be rounded to $2.60 * 10^{-4}$
(h) $2.41 * 10^{10} * 4.81 * 10^{12} = 1.15921 * 10^{-21} = 0.115921 * 10^{-20}$. It will be rounded to $0.12 * 10^{-20}$, a subnormal floating point.
(i) $6.37 * 10^{10} * 5.28 * 10 * 15 = -3.36336 * 10^{-24}$. It will be rounded to 0 because of underflow.
(j) $6.27 * 10^{10} / (2.72 * 10^{15}) = -2.305 * 10^{25}$. It will be rounded to -Inf because of overflow.

3. (a) Let Δx denote the change of input x (i.e. $\Delta x = \hat{x} - x$).

For very small Δx , we can write $f(\hat{x}) - f(x) = \Delta x f'(x)$

RelativeForwardError

$$\begin{aligned} &= \frac{f(\hat{x}) - f(x)}{f(x)} \\ &= \frac{f(x + \Delta x) - f(x)}{f(x)} = \frac{f'(x)(\hat{x} - x)}{f(x)} = \frac{xf'(\hat{x})}{f(x)} \times \frac{\hat{x} - x}{x} = \frac{1}{\log_e(x)} \times \frac{\hat{x} - x}{x} \end{aligned}$$

When x is close to 1, the condition number $|\frac{1}{\log_e(x)}|$ approaches Inf. The function is ill-conditioned in a relative sense with respect to small relative changes in the value of the input argument x for x close to 1.

When x is close to 10, the condition number $|\frac{1}{\log_e(x)}| \approx \frac{1}{\log_e 10} \approx 0.4343$. The function is well-conditioned in a relative sense with respect to small relative changes in the value of the input argument x for x close to 10.

- (b) See attachments for code and result.

In part (a), we calculated that the condition number can (almost) be expressed as $|\frac{1}{\log_e(x)}|$. In the computational result of part (b), the condition numbers are very close for the same

x . Also, the condition numbers are very large when $x = 1$ and small when $x = 10$, this also agrees with the calculation we have done in part (a).

4. (a) *RelativeError*

$$\begin{aligned}
&= \frac{\left(\frac{1}{1-x} - \frac{1}{1+x}\right) - \left(\frac{1}{(1-x)(1+\sigma_1)}(1+\sigma_2) - \frac{1}{(1+x)(1+\sigma_3)}(1+\sigma_4)\right)(1+\sigma_5)}{\frac{1}{1-x} - \frac{1}{1+x}} \\
&= \frac{2x(1+\sigma_1)(1+\sigma_3) + (1-x)(1+\sigma_1)(1+\sigma_4)(1+\sigma_5) - (1+x)(1+\sigma_3)(1+\sigma_2)(1+\sigma_5)}{(1-x^2)(1+\sigma_1)(1+\sigma_3)} \\
&= \frac{2x}{1-x^2} \\
&= 1 + \frac{(1+\sigma_2)(1+\sigma_3)(1+\sigma_5) - (1+\sigma_1)(1+\sigma_4)(1+\sigma_5)}{2} - \frac{(1+\sigma_2)(1+\sigma_3)(1+\sigma_5) + (1+\sigma_1)(1+\sigma_4)(1+\sigma_5)}{2x(1+\sigma_1)(1+\sigma_3)}
\end{aligned}$$

Note that the last part of the relative error is inverse proportional to x . That is, when x is extremely small, the absolute value of relative error can be very large.

(b) We choose $\frac{2x}{(1-x)(1+x)} = \frac{1}{1-x} - \frac{1}{1+x}$.

$$\begin{aligned}
&\text{RelativeError} \\
&= \frac{\frac{2x}{(1-x)(1+x)} - \frac{(2x)(1+\sigma_1)}{((1-x)(1+\sigma_2)(1+x)(1+\sigma_3))(1+\sigma_5)}(1+\sigma_4)}{\frac{2x}{(1-x)(1+x)}} \\
&= 1 - \frac{\frac{(2x)(1+\sigma_1)}{((1-x)(1+\sigma_2)(1+x)(1+\sigma_3))(1+\sigma_5)}(1+\sigma_4)}{\frac{2x}{(1-x)(1+x)}} \\
&= 1 - \frac{(1+\sigma_1)(1+\sigma_4)}{(1+\sigma_2)(1+\sigma_3)(1+\sigma_5)} \\
&= \frac{\sigma_2 + \sigma_3 + \sigma_5 + \sigma_2\sigma_3 + \sigma_2\sigma_5 + \sigma_3\sigma_5 + \sigma_2\sigma_3\sigma_5 - \sigma_1 - \sigma_4 - \sigma_1\sigma_4}{1 + \sigma_2 + \sigma_3 + \sigma_5 + \sigma_2\sigma_3 + \sigma_2\sigma_5 + \sigma_3\sigma_5 + \sigma_2\sigma_3\sigma_5}
\end{aligned}$$

Note that $|\sigma_i| < \epsilon_{\text{machine}}$, so in the worst case, we have

$$|RelativeError| = \frac{5\epsilon_{\text{machine}} + 4\epsilon_{\text{machine}}^2 + \epsilon_{\text{machine}}^3}{1 - 3\epsilon_{\text{machine}} - 3\epsilon_{\text{machine}}^2 - 3\epsilon_{\text{machine}}^3}$$

Since $\epsilon_{\text{machine}} \ll 1$, $|RelativeError| < 6\epsilon_{\text{machine}}$ in the worst case. So, we can conclude that $\frac{2x}{(1-x)(1+x)}$ has a very small relative error for all values of x , provided that there is no overflow or underflow.

5. (a) See attachments for code and output.

(b) For $x > -18$, the function produce very accurate approximations, (when $x = -18$ the approximation is also kind of accurate as the relative error is around only 0.05).

For $x < -19$, the relative error gets much larger. Of course rounding error contributes at lot on the poor approximations, the main reason that the function performs well on $x > 19$ but badly on $x < -19$ is that, $\exp(x)$ gets too small for $x < -19$. When denominator gets too small, the whole fraction denoting the relative number gets large.

On the other hand, when $|x|$ gets larger, it will require a higher i for $\frac{x^i}{i!}$ to be insignificant so that the accumulated sum stops changing. Much more operations are needed to compute each $\frac{x^i}{i!}$ with a very high i , resulting a higher error for $\frac{x^i}{i!}$. This can make the relative error grows even faster for more negative x . However, the denominator $\exp(x)$ gets too small is the main reason for the poor approximation for $x < -19$.

(c) See attachments for code and output.