CSC336 Assignment 1

Zikun Chen

September 7, 2018

Question 1

A: floating point approximation
T: true value
A.E.: absolute error
R.E.: relative error

(a)

$$T = 2.71828182845905$$

$$A = 2.72$$

$$A.E. = A - T = (2.72 - 2.71828182845905) = 0.00171817154 \approx 1.72 \times 10^{-3}$$

$$R.E. = \frac{A - T}{T} = \frac{0.00171817154}{2.71828182845905} \approx 0.00063207998 \approx 6.32 \times 10^{-4}$$

(b)

$$T = 2.71828182845905$$

$$A = 2.718$$

$$A.E. = 2.718 - 2.71828182845905 = -0.00028182845 \approx -2.82 \times 10^{-4}$$

$$R.E. = \frac{-0.00028182845}{2.71828182845905} \approx -0.00010367889 \approx -1.04 \times 10^{-4}$$

(c)

$$T = 2.71828182845905$$

$$A = 2.718283$$

$$A.E. = 2.71828183 - 2.71828182845905 = 1.54095003 \times 10^{-9} \approx 1.54 \times 10^{-9}$$

$$R.E. = \frac{1.54095003 \times 10^{-9}}{2.71828182845905} \approx 5.6688384 \times 10^{-10} \approx 5.67 \times 10^{-10}$$

Question 2

(a) $4.26 \cdot 10^0$

(b) $6.45 \cdot 10^1$

(c) $5.61 \cdot 10^1$

(d) $-3.77 \cdot 10^6$

(e) $7.51 \cdot 10^{12}$

(f) $8.80 \cdot 10^2$

(g) $2.60 \cdot 10^{-4}$

(h) Underflow to subnormal floating point number $0.12 \cdot 10^{-20}$

(i) Underflow to 0

(j) Overflow to -Inf

Question 3

(a) From class, we know that if $\hat{x}$ is very close to x (small change in x):

$$\text{Relative Error} = \frac{f(\hat{x}) - f(x)}{f(x)} \approx \frac{xf'(x)}{f(x)} \cdot \frac{\hat{x} - x}{x}$$

where $k = \frac{xf'(x)}{f(x)}$ is the condition number, the size of which indicates how much the function f is ill-conditioned.

For $f(x) = log_e(x)$, $\frac{xf'(x)}{f(x)} = \frac{x\frac{1}{x}}{log_e(x)} = \frac{1}{log_e(x)}$

When x is close to 1, k is either -Inf or +Inf. Hence, it is ill-conditioned for x close to 1.

When x is close to 10, k is close to $\frac{1}{log_e(10)} \approx 0.434$. Hence, it is well-conditioned when x close to 10.

(b) **Matlab Program:**

```
function [rel_change] = log_rel_change(x1, change)
x2 = x1 + change;
rel_change = (log(x2)-log(x1))/log(x1);
end

fprintf('\nQuestion 3\n');
x1 = 1 + 1e-10;
x10 = 10 + 1e-10;
change = 1e-10;
```

```
re1 = log_rel_change(x1, change);
re10 = log_rel_change(x10, change);
fprintf('For x = %.10f, the relative change in log(x) is %1.0f.\n', x1, re1);
fprintf('For x = %.10f, the relative change in log(x) is %.4e.\n', x10, re10);
```

Output:

```
Question 3
For x = 1.0000000001, the relative change in log(x) is 1.
For x = 10.0000000001, the relative change in log(x) is 4.3429e-12.
```

Explanation: The computation results support the theoretical predictions form a) since the outputs show that the relative change in the output is 1 when x is close to 1 (ill-conditioned), and $4.34 \cdot 10^{-12}$ when x is close to 10 (well-conditioned).

This is based on the fact that following part a), we chose x to be very close to 1 and 10 (10^{-10} more than 1 and 10), and the relative changes in x are very small as well (relative change for input $= \frac{10^{-10}}{x}$).

Question 4

(a) for $x = 5 * 10^{-17} = \frac{1}{2}\varepsilon_{machine}$,

$$fl(1-x) = 1$$

$$fl(1+x) = 1$$

$$fl(\frac{1}{1-x}) = 1$$

$$fl(\frac{1}{1+x}) = 1$$

$$A = fl(\frac{1}{1-x} - \frac{1}{1+x}) = 0$$

$$T = \frac{1}{1-x} - \frac{1}{1+x} = 0 \text{ for } x \neq 0$$

$$R.E. = \frac{A-T}{T} = \frac{-T}{T} = -1$$

(b) The alternative expression is:

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

$$
\begin{aligned}
fl(\frac{2*x}{(1-x)*(1+x)}) &= \frac{fl(2*x)}{fl((1-x)(1+\delta_1)*(1+x)(1+\delta_2))} \\
&= fl(\frac{2x(1+\delta_4)}{(1-x)(1+\delta_1)*(1+x)(1+\delta_2)(1+\delta_3)}) \\
&= \frac{2x(1+\delta_4)}{(1-x)(1+x)(1+\delta_1)(1+\delta_2)(1+\delta_3)}(1+\delta_5) \\
&= \frac{2x}{(1-x)(1+x)}\frac{(1+\delta_4)(1+\delta_5)}{(1+\delta_1)(1+\delta_2)(1+\delta 3)}
\end{aligned}
$$

where $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5 \leq \frac{1}{2}\varepsilon_{machine}$
Let

$$1+\tilde{\delta}_1 = (1+\delta_4)(1+\delta_5) = 1+\delta_4+\delta_5+\delta_4\delta_5$$

$$1+\tilde{\delta}_2 = (1+\delta_1)(1+\delta_2)(1+\delta_3) = 1+\delta_1+\delta_2+\delta_3+\delta_1\delta_2+\delta_2\delta_3+\delta_1\delta_3+\delta_1\delta_2\delta_3$$

so

$$
\begin{aligned}
|\tilde{\delta}_1| &= |\delta_4+\delta_5+\delta_4\delta_5| \\
&\leq |\delta_4|+|\delta_5|+|\delta_4\delta_5| \\
&\leq \varepsilon_{machine}+\frac{1}{4}\varepsilon_{machine}^2 \\
&= (1+\frac{1}{4}\varepsilon_{machine})\varepsilon_{machine} \\
&\leq 1.01\varepsilon_{machine}
\end{aligned}
$$

and

$$
\begin{aligned}
|\tilde{\delta}_2| &= |\delta_1+\delta_2+\delta_3+\delta_1\delta_2+\delta_2\delta_3+\delta_1\delta_3+\delta_1\delta_2\delta_3| \\
&\leq |\delta_1|+|\delta_2|+|\delta_1\delta_2|+|\delta_2\delta_3|+|\delta_2\delta_3|+|\delta_1\delta_2\delta_3| \\
&\leq \frac{3}{2}\varepsilon_{machine}+\frac{3}{4}\varepsilon_{machine}^2+\frac{1}{8}\varepsilon_{machine}^3 \\
&= (1.5+\frac{3}{4}\varepsilon_{machine}+\frac{1}{8}\varepsilon_{machine}^2)\varepsilon_{machine} \\
&\leq 1.51\varepsilon_{machine}
\end{aligned}
$$

Therefore, we define δ^* such that:

$$
\begin{aligned}
1+\delta^* &= \frac{(1+\delta_4)(1+\delta_5)}{(1+\delta_1)(1+\delta_2)(1+\delta 3)} \\
&= \frac{1+\tilde{\delta}_1}{1+\tilde{\delta}_2} \\
&< (1+\tilde{\delta}_1)(1+\tilde{\delta}_2) && \text{since } (1+\tilde{\delta}_1)>0 \text{ and } \tilde{\delta}_2>0 \\
&= 1+\tilde{\delta}_1+\tilde{\delta}_2+\tilde{\delta}_1\tilde{\delta}_2
\end{aligned}
$$

$$\begin{aligned}|\delta^*| &= |\tilde{\delta}_1 + \tilde{\delta}_2 + \tilde{\delta}_1\tilde{\delta}_2| \\ &\leq |\tilde{\delta}_1| + |\tilde{\delta}_2| + |\tilde{\delta}_1\tilde{\delta}_2| \\ &\leq 1.01\varepsilon_{machine} + 1.51\varepsilon_{machine} + 1.5251\varepsilon_{machine}^2 \\ &= (2.52 + 1.5251\varepsilon_{machine})\varepsilon_{machine} \\ &\leq 2.6\varepsilon_{machine}\end{aligned}$$

Therefore,

$$fl(\frac{2 * x}{(1-x) * (1+x)}) = \frac{2x}{(1-x)(1+x)}(1+\delta^*)$$

where $|\delta^*| \leq 2.6\varepsilon_{machine} \forall x \neq \pm 1$

Question 5

(a) **Matlab Program:**

```
function [sum] = exp1(x)
    n = 0;
    sum = 1;
    oldsum = 0;
    while sum~=oldsum
        oldsum=sum;
        sum = sum + (x^n)/factorial (n);
        n = n + 1;
    end
end

fprintf(’\nexp1\n’);
for x = -25:25
    re1 = (exp1(x)-exp(x))/exp(x);
    fprintf(’x = %1.0f, R.E. = %.4e. ’, x, re1);
    if (mod(x,2)==0)
        fprintf(’\n’);
    end
end
```

Output:

```
exp1
x = -25, R.E. = 7.2005e+10. x = -24, R.E. = 2.6489e+10.
x = -23, R.E. = 9.7448e+09. x = -22, R.E. = 3.5849e+09.
x = -21, R.E. = 1.3188e+09. x = -20, R.E. = 4.8517e+08.
```

```
x = -19, R.E. = 1.7848e+08. x = -18, R.E. = 6.5660e+07.
x = -17, R.E. = 2.4155e+07. x = -16, R.E. = 8.8861e+06.
x = -15, R.E. = 3.2690e+06. x = -14, R.E. = 1.2026e+06.
x = -13, R.E. = 4.4241e+05. x = -12, R.E. = 1.6275e+05.
x = -11, R.E. = 5.9874e+04. x = -10, R.E. = 2.2026e+04.
x = -9, R.E. = 8.1031e+03. x = -8, R.E. = 2.9810e+03.
x = -7, R.E. = 1.0966e+03. x = -6, R.E. = 4.0343e+02.
x = -5, R.E. = 1.4841e+02. x = -4, R.E. = 5.4598e+01.
x = -3, R.E. = 2.0086e+01. x = -2, R.E. = 7.3891e+00.
x = -1, R.E. = 2.7183e+00. x = 0, R.E. = 1.0000e+00.
x = 1, R.E. = 3.6788e-01. x = 2, R.E. = 1.3534e-01.
x = 3, R.E. = 4.9787e-02. x = 4, R.E. = 1.8316e-02.
x = 5, R.E. = 6.7379e-03. x = 6, R.E. = 2.4788e-03.
x = 7, R.E. = 9.1188e-04. x = 8, R.E. = 3.3546e-04.
x = 9, R.E. = 1.2341e-04. x = 10, R.E. = 4.5400e-05.
x = 11, R.E. = 1.6702e-05. x = 12, R.E. = 6.1442e-06.
x = 13, R.E. = 2.2603e-06. x = 14, R.E. = 8.3153e-07.
x = 15, R.E. = 3.0590e-07. x = 16, R.E. = 1.1254e-07.
x = 17, R.E. = 4.1399e-08. x = 18, R.E. = 1.5230e-08.
x = 19, R.E. = 5.6028e-09. x = 20, R.E. = 2.0612e-09.
x = 21, R.E. = 7.5826e-10. x = 22, R.E. = 2.7895e-10.
x = 23, R.E. = 1.0262e-10. x = 24, R.E. = 3.7751e-11.
x = 25, R.E. = 1.3888e-11.
```

(b) For x positive, the function produce accurate approximation to e^x.

For $x \leq 0$, the function produce poor approximation to e^x.

The function produce very poor approximation for negative value of x's, especially when x has large magnitude. This is because when $x < 0$, the true of e^{-x} will be very small, especially when x is further from zero, the significant digits lie in the last few digits on the right.

However, when $x < 0$ the Taylor series is alternating in signs and increasing in magnitude. This will cause catastrophic cancellation when we add new terms to the sum (opposite signs). The significant small digits of the new term, which should eventually contribute to the final result, will be lost after the addition or subtraction of the new term to the sum.

On the other hand, for $x > 0$, the Taylor series approximation avoids the numerical instability of summing over large terms with opposite signs. And adding n positive number together will not produce a significant rounding error.

(c) **Matlab Program:**

```
function [sum] = exp2(x)
    n = 0;
    sum = 1;
    oldsum = 0;
    if x < 0
        x = abs(x);
        while sum˜=oldsum
            oldsum=sum;
            sum = sum + (xˆn)/factorial (n);
            n = n + 1;
        end
        sum = 1/sum;
    elseif x == 0
        sum = 1;
    else
        while sum˜=oldsum
           oldsum=sum;
            sum = sum + (xˆn)/factorial (n);
           n = n + 1;
        end
    end
end

fprintf(’\nexp2\n’);
for x = -25:25
    re2 = (exp2(x)-exp(x))/exp(x);
    fprintf(’x = %1.0f, R.E. = %.4e. ’, x, re2);
    if (mod(x,2)==0)
        fprintf(’\n’);
    end
end
```

Output:

```
exp2
x = -25, R.E. = -1.3888e-11. x = -24, R.E. = -3.7751e-11.
x = -23, R.E. = -1.0262e-10. x = -22, R.E. = -2.7895e-10.
x = -21, R.E. = -7.5826e-10. x = -20, R.E. = -2.0612e-09.
x = -19, R.E. = -5.6028e-09. x = -18, R.E. = -1.5230e-08.
x = -17, R.E. = -4.1399e-08. x = -16, R.E. = -1.1254e-07.
x = -15, R.E. = -3.0590e-07. x = -14, R.E. = -8.3153e-07.
x = -13, R.E. = -2.2603e-06. x = -12, R.E. = -6.1442e-06.
x = -11, R.E. = -1.6701e-05. x = -10, R.E. = -4.5398e-05.
x = -9, R.E. = -1.2339e-04. x = -8, R.E. = -3.3535e-04.
x = -7, R.E. = -9.1105e-04. x = -6, R.E. = -2.4726e-03.
x = -5, R.E. = -6.6929e-03. x = -4, R.E. = -1.7986e-02.
x = -3, R.E. = -4.7426e-02. x = -2, R.E. = -1.1920e-01.
x = -1, R.E. = -2.6894e-01. x = 0, R.E. = 0.0000e+00.
x = 1, R.E. = 3.6788e-01. x = 2, R.E. = 1.3534e-01.
```

```
x = 3, R.E. = 4.9787e-02. x = 4, R.E. = 1.8316e-02.
x = 5, R.E. = 6.7379e-03. x = 6, R.E. = 2.4788e-03.
x = 7, R.E. = 9.1188e-04. x = 8, R.E. = 3.3546e-04.
x = 9, R.E. = 1.2341e-04. x = 10, R.E. = 4.5400e-05.
x = 11, R.E. = 1.6702e-05. x = 12, R.E. = 6.1442e-06.
x = 13, R.E. = 2.2603e-06. x = 14, R.E. = 8.3153e-07.
x = 15, R.E. = 3.0590e-07. x = 16, R.E. = 1.1254e-07.
x = 17, R.E. = 4.1399e-08. x = 18, R.E. = 1.5230e-08.
x = 19, R.E. = 5.6028e-09. x = 20, R.E. = 2.0612e-09.
x = 21, R.E. = 7.5826e-10. x = 22, R.E. = 2.7895e-10.
x = 23, R.E. = 1.0262e-10. x = 24, R.E. = 3.7751e-11.
x = 25, R.E. = 1.3888e-11.
```