CSC336 A1

Ramaneek Gill 1000005754

October 3 2014

1.
$Absolute\ error = \hat{y} - y$
$Relative\ error = \frac{Absolute\ error}{y}$

(a)
Absolute error: $2.7 - 2.718281828459046 = -1.828182845904580e-02$
Relative error: $\frac{-1.828182845904580e-02}{2.718281828459046} = -6.725508837105938e-03$

(b) Absolute error: $2.7183 - 2.718281828459046 = 1.817154095418161e-05$
Relative error: $\frac{1.817154095418161e-05}{2.718281828459046} = 6.684936331448306e-06$

(c) Absolute error: $2.718281828 - 2.718281828459046 = -4.590461344378127e-10$
Relative error: $\frac{-4.590461344378127e-10}{2.718281828459046} = -1.688736354088933e-10$

2.
$\beta = 10 \quad p = 3 \quad L = -20 \quad U = +20$

(a)
$1.23 \cdot 10^0 + 5.14 \cdot 10^{-2} = 123 \cdot 10^{-2} + 5.14 \cdot 10^{-2}$
$= 128.14 \cdot 10^{-2} \simeq 1.28 \cdot 10^0$

(b)
$1.58 \cdot 10^1 - 5.41 \cdot 10^{-1} = 1.58 \cdot 10^1 - 0.0541 \cdot 10^1$
$= 1.5259 \cdot 10^1 \simeq 1.53 \cdot 10^1$

(c)
$1.23 \cdot 10^1 + 5.14 \cdot 10^{-4} = 1.23 \cdot 10^1 + 0.0000514 \cdot 10^1$
$= 1.2300514 \cdot 10^1 \simeq 1.23 \cdot 10^1$

(d)
$2.66 \cdot 10^4 - 5.42 \cdot 10^6 = 0.0266 \cdot 10^6 - 5.42 \cdot 10^6$

$= -5.3934 \cdot 10^6 \simeq -5.29 \cdot 10^6$

(e)
$3.76 \cdot 10^{12} - 7.69 \cdot 10^5 = 3.76 \cdot 10^{12} - 0.000000769 \cdot 10^{12}$
$= 3.759999231 \cdot 10^{12} \simeq 3.76 \cdot 10^{12}$

(f)
$1.87 \cdot 10^1 + 4.31 \cdot 10^2 = 0.187 \cdot 10^2 + 4.31 \cdot 10^2$
$= 4.497 \cdot 10^2 \simeq 4.50 \cdot 10^2$

(g)
$1.67 \cdot 10^{10} \times 5.43 \cdot 10^{-15}$
$= 9.068099999999999 \cdot 10^{10-15} \simeq 9.07 \cdot 10^{-5}$

(h)
$-4.67 \cdot 10^{10} \div (1.84 \cdot 10^{-15})$
$= -2.538043478260870 \cdot 10^{10-15} \simeq -2.54 \cdot 10^{-5}$

(i)
$3.86 \cdot 10^{-10} \times 1.23 \cdot 10^{-12} = 4.7478 \cdot 10^{-10-12}$
$= 4.7478 \cdot 10^{-22} = 0.047478 \cdot 10^{-20}$
$\simeq 0.05 \cdot 10^{-20}$

(j)
$2.94 \cdot 10^{-10} \times 6.23 \cdot 10^{-15} = 18.3162 \cdot 10^{-10-15}$
$= 18.3162 \cdot 10^{-25} = 0.000183162 \cdot 10^{-20}$
$\simeq 0$

3.
$f(x) = x^{\frac{1}{3}}$ $\qquad$ $f\prime(x) = \frac{1}{3x^{\frac{2}{3}}}$ $\qquad$ $Condition\ Number \approx \left|\frac{xf\prime(x)}{f(x)}\right|$

Note: The absolute value is omitted from the calculations since all calculations result in positive numbers.

$\frac{x \cdot \frac{1}{3} \cdot x^{\frac{-2}{3}}}{x^{\frac{1}{3}}} = \frac{x \cdot \frac{1}{3}}{x^{\frac{1}{3}} \cdot x^{\frac{2}{3}}} = \frac{x \cdot \frac{1}{3}}{x} = \frac{1}{3}$

The condition number for $f(x) = x^{\frac{1}{3}}$ is $\frac{1}{3}$. The function is well conditioned since a given change in the input will result in a change in the output of about a third of that size. This also shows that the relative forward error is about one third of the relative backwards error. The reciprocal of this is true for the inverse of this function.

4.
a)
$fl\left(\frac{2x}{(1-x)(1+x)}\right)$
$= \frac{2x(1+\delta_1)}{(1-x)(1+\delta_2)(1+x)(1+\delta_3)} \cdot (1+\delta_4)$
$= \frac{2x}{(1-x)(1+x)} \cdot (1+\hat{\delta})$

Note:
$(1+\hat{\delta})$ came from combining all the δ terms in the equation before to simplify the expression.

$\hat{\delta} = \delta_1 + \delta_2 + \delta_3 + \delta_4 + |H.O.T.|$
$|\hat{\delta}| = |\delta_1| + |\delta_2| + |\delta_3| + |\delta_4| + |H.O.T.|$
$\leq \frac{4}{2}\epsilon_{mach} \cdot 1.01 + \frac{1}{2}\epsilon_{mach}$
$\leq \frac{5}{2}\epsilon_{mach} \cdot 1.01$

Now we compare this to...

$fl\left(\frac{1}{(1-x)} - (\frac{1}{(1+x)}\right)$
$= \left(\frac{1}{1-x}\right)(1+\delta_1) - \left(\frac{1}{1+x}\right)(1+\delta_2)$
$\rightarrow (1+\delta_1) \pm (1+\delta_2)$
$= \frac{1}{2}\epsilon_{mach} \pm \frac{1}{2}\epsilon_{mach}$
$\leq \epsilon_{mach}$

This shows that the equation $fl\left(\frac{1}{(1-x)} - (\frac{1}{(1+x)}\right)$ is more accurate in a relative sense since it has less of a rounding error.
b)

5.
a)

```
for i = -25:25

    format long;
    n=0;
    sum=0;
    oldsum=100;
    while (abs(sum-oldsum) > 0.0000000000000001)
      oldsum=sum;
      sum=sum+i^n/factorial(n);
      n=n+1;
    end

    y = (sum - exp(i))/exp(i);
```

```
    disp(['Value of x is ', num2str(i), ', relative error is ', num2str(y),'.']);
end
```

PROGRAM OUTPUT:

```
Value of x is -25, relative error is 58226.187.
Value of x is -24, relative error is 9966.3507.
Value of x is -23, relative error is 66.229.
Value of x is -22, relative error is -115.0737.
Value of x is -21, relative error is 35.3865.
Value of x is -20, relative error is 1.0249.
Value of x is -19, relative error is -0.5442.
Value of x is -18, relative error is 0.04948.
Value of x is -17, relative error is 0.0010196.
Value of x is -16, relative error is 0.00028526.
Value of x is -15, relative error is 1.0354e-05.
Value of x is -14, relative error is -8.6112e-06.
Value of x is -13, relative error is -1.2996e-06.
Value of x is -12, relative error is 6.1217e-08.
Value of x is -11, relative error is 7.6483e-08.
Value of x is -10, relative error is -7.2343e-09.
Value of x is -9, relative error is -5.493e-10.
Value of x is -8, relative error is -1.477e-10.
Value of x is -7, relative error is 1.2601e-11.
Value of x is -6, relative error is -7.2416e-13.
Value of x is -5, relative error is 2.1433e-13.
Value of x is -4, relative error is 1.4775e-14.
Value of x is -3, relative error is 9.756e-16.
Value of x is -2, relative error is 4.1018e-16.
Value of x is -1, relative error is 3.0179e-16.
Value of x is 0, relative error is 0.
Value of x is 1, relative error is 0.
Value of x is 2, relative error is -2.404e-16.
Value of x is 3, relative error is -3.5376e-16.
Value of x is 4, relative error is 5.2056e-16.
Value of x is 5, relative error is -1.915e-16.
Value of x is 6, relative error is 0.
Value of x is 7, relative error is -6.2201e-16.
Value of x is 8, relative error is -1.5255e-16.
Value of x is 9, relative error is 0.
Value of x is 10, relative error is -3.3033e-16.
Value of x is 11, relative error is 0.
Value of x is 12, relative error is -3.5764e-16.
Value of x is 13, relative error is -1.3157e-16.
Value of x is 14, relative error is 1.9361e-16.
Value of x is 15, relative error is 0.
```

```
Value of x is 16, relative error is 0.
Value of x is 17, relative error is 3.0845e-16.
Value of x is 18, relative error is 0.
Value of x is 19, relative error is -1.6698e-16.
Value of x is 20, relative error is -2.4571e-16.
Value of x is 21, relative error is -3.6156e-16.
Value of x is 22, relative error is 2.6602e-16.
Value of x is 23, relative error is 1.9573e-16.
Value of x is 24, relative error is 0.
Value of x is 25, relative error is 0.
```

b)
The more negative the value of x is the less accurate the output of exp1(x) will be. Positive values for x will produce accurate approximations for e^x because there is 0 percent chance of catastrophic cancellation. For negative values whenever the exponent is odd there is a chance for cancellation to occur.

This is especially true for large negative numbers since their values start to become very small and are catastrophically cancelling each other out every other time (due to the odd exponents). Since they are becoming smaller and smaller we are also starting to lose accuracy because they start to approach machine epsiolon, this implies that the smaller the number the greater it is affected by rounding errors since the percentage of change is large for very small number but small for very large numbers.

Large negative numbers greatly affect the computation of exp1(x) because of catastrophic cancellation. The larger the number the greater the chance (and magnitude) for catastrophic cancellation to occur, this creates a very large relative error because some of the summations are essentially 'missing' from the computation. When the absolute error (which seems very small) is divided by the actual result for exp(x) it creates an incredibly large relative error for large negative numbers of x since exp(x) is incredibly small.

c)

```
for i = -25:25

    format long;
    n=0;
    sum=0;
    oldsum=100;

    if i < 0
      while (abs(sum-oldsum) > 0.0000000000000001)
        oldsum=sum;
```

```
        sum = sum + abs(i)^n/factorial(n);
        n=n+1;
      end
      sum = 1/sum;

    else
      while (abs(sum-oldsum) > 0.0000000000000001)
        oldsum=sum;
        sum = sum + i^n/factorial(n);
        n=n+1;
      end
    end

    y = (sum - exp(i))/exp(i);
    disp(['Value of x is ', num2str(i), ', relative error is ', num2str(y, 20),'.']);
end

PROGRAM OUTPUT:
Value of x is -25, relative error is -1.1633018894671471824e-16.
Value of x is -24, relative error is 0.
Value of x is -23, relative error is -2.5189726512115696584e-16.
Value of x is -22, relative error is -1.8533565025077174694e-16.
Value of x is -21, relative error is 4.0908705266039873827e-16.
Value of x is -20, relative error is 2.0065962176423984623e-16.
Value of x is -19, relative error is 0.
Value of x is -18, relative error is 0.
Value of x is -17, relative error is -3.1968813783826973311e-16.
Value of x is -16, relative error is 0.
Value of x is -15, relative error is 0.
Value of x is -14, relative error is -2.546613627991847494e-16.
Value of x is -13, relative error is 1.8736935966904402478e-16.
Value of x is -12, relative error is 4.13576011966193391e-16.
Value of x is -11, relative error is 0.
Value of x is -10, relative error is 2.9851427583633870839e-16.
Value of x is -9, relative error is -2.1963452995274011146e-16.
Value of x is -8, relative error is 1.6159805628193289999e-16.
Value of x is -7, relative error is 7.1338323167266504429e-16.
Value of x is -6, relative error is 0.
Value of x is -5, relative error is 2.5745579123912058761e-16.
Value of x is -4, relative error is -3.788507704295968289e-16.
Value of x is -3, relative error is 2.7874281942602091471e-16.
Value of x is -2, relative error is 2.0508750528199364057e-16.
Value of x is -1, relative error is -1.5089495366877009823e-16.
Value of x is 0, relative error is 0.
Value of x is 1, relative error is 0.
Value of x is 2, relative error is -2.4040375598952703896e-16.
```

```
Value of x is 3, relative error is -3.5375839763577193127e-16.
Value of x is 4, relative error is 5.2056176652781067952e-16.
Value of x is 5, relative error is -1.9150397176546979952e-16.
Value of x is 6, relative error is 0.
Value of x is 7, relative error is -6.220138622355873523e-16.
Value of x is 8, relative error is -1.5255074136007890648e-16.
Value of x is 9, relative error is 0.
Value of x is 10, relative error is -3.303279646387443601e-16.
Value of x is 11, relative error is 0.
Value of x is 12, relative error is -3.5764022924286603267e-16.
Value of x is 13, relative error is -1.3156848767429207449e-16.
Value of x is 14, relative error is 1.936053668855614674e-16.
Value of x is 15, relative error is 0.
Value of x is 16, relative error is 0.
Value of x is 17, relative error is 3.084493964005385958 6e-16.
Value of x is 18, relative error is 0.
Value of x is 19, relative error is -1.6697634570411643382e-16.
Value of x is 20, relative error is -2.4570865898591969847e-16.
Value of x is 21, relative error is -3.6156465663489854958e-16.
Value of x is 22, relative error is 2.660244076603817595e-16.
Value of x is 23, relative error is 1.9572982085613041273e-16.
Value of x is 24, relative error is 0.
Value of x is 25, relative error is 0.
```