MTH 351 Homework 2

Philip Warton

January 27, 2020

1.

We want to use numerical solutions to compute the integral:
$I = \int_1^2 \frac{e^x - 1}{x} dx$

Let $f(x) = \frac{e^x - 1}{x}$.

(a)

To compute a Taylor polynomial for $f(x)$ let us first find the Taylor polynomial for e^x around $x_0 = 0$. Since e^x and all of its derivatives are equal to 1 when $x = 0$, we can write this polynomial as $e^x \approx 1 + \frac{x}{1!} + \frac{x^2}{2!} + ... + \frac{x^n}{n!}$. Now let us replace e^x with it's Taylor polynomial approximation. Let $p_n(x)$ denote the Taylor polynomial for $f(x)$, and we have

$$\begin{aligned} p_n(x) &= \frac{1 + \frac{x}{1!} + \frac{x^2}{2!} + ... + \frac{x^n}{n!} - 1}{x} \\ &= \frac{1}{1!} + \frac{x}{2!} + \frac{x^2}{3!} + ... + \frac{x^{n-1}}{n!} \\ &= \sum_{k=1}^{n} \frac{x^{n-1}}{n!} \end{aligned}$$

(b)

To write an approximation of I, let us replace $f(x)$ with $p_n(x)$ as follows:

$$\begin{aligned} I &= \int_1^2 \frac{e^x - 1}{x} dx \\ &\approx \int_1^2 p_n(x) dx \\ &= \int_1^2 \sum_{k=1}^{n} \frac{x^{n-1}}{n!} dx \\ &= \sum_{k=1}^{n} \int_1^2 \frac{x^{n-1}}{n!} dx \\ &= \sum_{k=1}^{n} \left(\frac{x^n}{n(n!)} \Big|_1^2 \right) \\ &= \sum_{k=1}^{n} \frac{2^n - 1}{n(n!)} \end{aligned}$$

(c)

Now we wish to find some n sufficiently large so that our error is less than $\epsilon = 10^{-5}$. Let the error term of our polynomial $p_n(x)$ be $r_n(x) = \dfrac{e^c x^n}{(n+1)!}$ where c lies between $x_0 = 0$ and x. Since our x is evaluated at 2 and 1, we can bound our error above by chooseing $c = 2$. We then have $\dfrac{e^c x^n}{(n+1)!} \leqslant \dfrac{e^2 x^n}{(n+1)!}$. By taking the integral of $r_n(x)$ we can produce an error term. We can write

$$
\begin{aligned}
I_{\text{error}} &= \int_1^2 \frac{e^c x^n}{(n+1!)} dx \\
&\leqslant \int_1^2 \frac{e^2 x^n}{(n+1!)} dx \\
&= \frac{e^2}{(n+1)!} \int_1^2 x^n dx \\
&= \frac{e^2}{(n+1)!} \left(\frac{x^{n+1}}{n+1} \Big|_1^2 \right) \\
&= \frac{e^2}{(n+1)!} \left(\frac{2^{n+1} - 1}{n+1} \right) \\
&= \frac{e^2 (2^{n+1} - 1)}{(n+1)!(n+1)}
\end{aligned}
$$

Let this be less than $\epsilon = 10^{-5}$ and find the smallest possible n. After using a calculator we find that $n \geqslant 11$ is sufficiently large.

(d)

After verifying that $n \geqslant 11$ worked, note that $n = 10$ was a smaller number that was also within our epsilon error. It may be the case that choosing a $c = 0$ for the Langrange error term is sufficient. See Matlab code attached on canvas.

2.

Let our model for numbers be as follows: $c_0 \;\; b_1 \; b_2 \; b_3 \; b_4 \;\; a_1 \; a_2 \; a_3$ where c_0 is the sign part, $b_1 \; b_2 \; b_3 \; b_4$ is the exponent part, and $a_1 \; a_2 \; a_3$ is the mantissa part.

(a)

Find the dynamic range and the machine epsilon.

To find the dynamic range let us find both the largest and smallest strictly positive numbers that can be represented by our model. First we wish to find the smallest number. Let $c_0 = 0$ so that our number is positive. Let $b_1 b_2 b_3 b_4 = 0000$ so that our exponent will be its lowest possible value 2^{-6}. Finally let $a_1 a_2 a_3 = 001$ such that the number will

be as small as possible without being zero. From this we have

$$\begin{aligned} x &= (1)(0.001)_2(2^{-6}) \\ &= 1 * 2^{-3} * 2^{-6} \\ &= 2^{-9} \end{aligned}$$

Therefore our smallest number is 2^{-9}.

To find the largest possible number let us maximize each component without x being infinity. Of course let $c_0 = 0$ as our number must be positive. Then let $b_1b_2b_3b_4 = 1110$ as this as large as our exponent can be without making x represent infinity. From there, let $a_1a_2a_3 = 111$ so that it is a large as possible. This gives us

$$\begin{aligned} x &= (1)(1.111)_2(2^7) \\ &= (2 - 2^{-3})(2^7) \\ &= (\frac{15}{8})(2^7) \end{aligned}$$

Given we have $x_{\text{min}} = 2^{-9}$ and $x_{\text{max}} = (\frac{15}{8})(2^7)$, we can say that the dynamic range is about $\frac{((\frac{15}{8})(2^7)}{2^{-9}} = (\frac{15}{8})2^{16} = 122880$

Now to find the machine epsilon we must take the smallest number larger than 1 and find the difference between 1 and that number. As our exponent increases, the precision of our number decreases. If our exponent is less than 0, then our number is guarenteed to be less then one, therefore let our exponent equal zero. To make our exponent zero, E must equal 7. Let our mantissa part be the smallest possible non-zero number. Altogether we have $c_0 = 0$, $b_1b_2b_3b_4 = 0111$, and $a_1a_2a_3 = 001$. We then get

$$\begin{aligned} x &= (1)(1.001)_2(2^0) \\ &= (1.001)_2 \\ &= 1 + \frac{1}{2^3} \\ &= \frac{9}{8} \end{aligned}$$

Thus $1 - \frac{9}{8} = \frac{1}{8}$ is our machine epsilon.

(b)

Write the numbers represented by the following bit sequences:

$\boxed{11001001}$

From this we have $c_0 = 1$ therefore the sign will be negative. Our exponent part will be $2^{(1001)_2-7}$. For the mantissa part we have $(1.001)_2$. This gives us $(-1)(\frac{9}{8})(2^{9-7}) = \frac{-9}{8} * 2^2 = \frac{-9}{2} = -4.5$.

$\boxed{00000000}$

Therefore we have a sign part as (1), and with our exponent part $E = 0$, we have our exponent part as 2^{-6} and no leading 1 on our mantissa part. Thus we have $(0.000)_2$ as our mantissa part. With $x = (1)(0.000)_2(2^{-6}) = 0$, we have this number representing 0.

$\boxed{11111000}$

With the sign part being (-1) and $(1111)_2 = E = 15$, we have that $x = -\infty$, regardless of mantissa part.

3.

(a)

Represent the following numbers in the style presented in problem 2.

$\boxed{x = 1}$

$$\begin{aligned} x &= 1 \\ &= (1)(1)(1) \\ &= (1)(1.000)_2(2^0) \end{aligned}$$

Therefore let $c_0 = 0, b_1b_2b_3b_4 = 0111, a_1a_2a_3 = 000$ and our number is 00111000.

$\boxed{x = 5.5}$

$$\begin{aligned} x &= 5.5 \\ &= (101.1)_2 \\ &= (1.011)_2(2^2) \\ &= (1)(1.011)_2(2^2) \end{aligned}$$

Thus we have $c_0 = 0, b_1b_2b_3b_4 = 1001, a_1a_2a_3 = 011$ and our number is 01001011.

$\boxed{x = 12.9}$

$$\begin{aligned} x &= 12.9 \\ &= (1100.1110011...)_2 \\ &= (1.10011100...)_2(2^3) \\ &\approx (1)(1.101)_2(2^3) \end{aligned}$$

Thus we have $c_0 = 0, b_1b_2b_3b_4 = 1010, a_1a_2a_3 = 101$ and our number is 01010101.

$\boxed{x = 1000}$
From problem 2 we know that the largest possible number we can represent is $15 * 2^4$ = 240. Therefore, we must use the exponent part representing infinity, and with the number being positive it must be a positive infinity. Given this we can write our sign part as (-1), our exponent part as 1111, and our mantissa part as any collection of numbers. The number can be written as $11111a_1a_2a_3$ where a_n can be either 1 or 0.

$\boxed{x = 0.0001}$
We can write this number as $\frac{1}{10000}$. The smalles possible number we can represent we know to be $2^{-9} = \frac{1}{512}$. Therefore this number is too small and will be rounded to zero. We can write this number as 00000000.

(b)

Find the smallest number that is larger than the following numbers, as in the format from problem 2.

$\boxed{x > 5.5}$ To find the smallest number greater than 5.5, let us add one to our mantissa part, and leave the sign and exponent parts alone. If we were to change the sign, our number would be less than 5.5, and if we were to change the exponent, our number would be off by at least one binary place. Therefore we replace our former mantissa part 011 with 100. This gives us $(1.100)_2 * 2^2 = (6)_{10}$. This number is written as 01001111.

$\boxed{x > 12.9}$ Due to rounding, our previous representation of 12.9 was equal to 13 and therefore larger than 12.9. For our purposes, however, let us find the next number larger than that. Once again we must take our previous number and add one to our mantissa part. This gives us $(1.110)_2 * 2^3 = 14$. This number is represented by 01010110.

$\boxed{x > 100.25}$ Let us find the binary representation of 100.25. We can write $(100.25)_{10} = (1100100.01)_2$. Our representation of this would be $(1.101)_2 * 2^6$. This would make our exponent part (1101). The smallest number greater than this would be $(1.110)_2 * 2^6 = 104$. We write this as 01101110.