MATH 3351-1 **Numerical Analysis I** **Fall 2025**

Homework 1

Due Date: September 5

Grace Etzel

Complete the assignment in Google Colab and typeset your wite-up in LaTeX. Show all work and code, and provide clear justifications for your answers. Submissions that are messy or poorly organized may be returned with a grade of zero.

link to google colab: link

Problem 1. Rewrite the polynomial

$$P(x) = 4x^6 - 2x^4 - 2x + 4$$

in nested form and evaluate it at $x = -\frac{1}{2}$.

Solution. Recall that we can denote the nested form as, for $P(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x$,

$$P(x) = (((((a)x + b)x + c)x + d)x + \cdots)$$

Then, when $P(x) = 4x^6 - 2x^4 - 2x + 4$,

$$\begin{aligned}
P(x) &= 4x^6 + 0x^5 - 2x^4 + 0x^3 + 0x^2 - 2x + 4 \\
&= (4x^5 + 0x^4 - 2x^3 + 0x^2 + 0x - 2)x + 4 \\
&= ((4x^4 + 0x^3 - 2x^2 + 0x + 0)x - 2)x + 4 \\
&= (((4x^3 + 0x^2 - 2x + 0)x + 0)x - 2)x + 4 \\
&= ((((4x^2 + 0x - 2)x + 0)x + 0)x - 2)x + 4 \\
&= (((((4x + 0)x - 2)x + 0)x + 0)x - 2)x + 4 \\
&= (((((4x)x - 2)x)x)x - 2)x + 4
\end{aligned}$$

Now, let us consider

$$\begin{aligned}
P(\frac{1}{2}) &= (((((4(\frac{1}{2}))(\frac{1}{2})-2)\frac{1}{2})\frac{1}{2})\frac{1}{2}-2)\frac{1}{2}+4 \\
&= (((((2)(\frac{1}{2})-2)\frac{1}{2})\frac{1}{2})\frac{1}{2}-2)\frac{1}{2}+4 \\
&= ((((1-2)\frac{1}{2})\frac{1}{2})\frac{1}{2}-2)\frac{1}{2}+4 \\
&= ((((-1)\frac{1}{2})\frac{1}{2})\frac{1}{2}-2)\frac{1}{2}+4 \\
&= (-\frac{1}{8}-2)\frac{1}{2}+4 \\
&= (-\frac{17}{8})\frac{1}{2}+4 \\
&= -\frac{17}{16}+4 \\
&= \frac{47}{16}.
\end{aligned}$$

Problem 2. Convert each of the following base-10 numbers to base-2 using the method from class (division by 2 for the integer part and multiplication by 2 for the fractional part). Show every step of your work.

(a) $\left(\frac{1}{8}\right)_{10}$

Solution. We can write $\left(\frac{1}{8}\right)_{10} = 0.125$. Then, we multiply by 2, which gives us 0.25. Note that the integer part is 0. We multiply by 2 again, which gives us 0.5, where the integer part is 0. Finally, we multiply by 2 one more time which gives us 1. So, we can write the in base 2 form as 0.001. That is

$$\left(\frac{1}{8}\right)_{2} = 0.001.$$

(b) $(3.2)_{10}$

Solution. We can write $(3.2)_{10} = 3.2$. We want to find this value in base 2. We first consider the integer part 3. We divide it by 2, which gives us 1. Notice that the remainder when we divide by 2 again, will give us 1. So, the base 2 for the integer value is 11. Next, we consider the .2 (the fractional part). When w multiply by 2, we get .4. We multiply it again by 2, we get .8. We multiply by 2 we get, 1.6. Consider the fractional part, we multiply 0.6 by 2 to get 1.2. Notice that we are hitting a pattern.

That is, for the integer part, $(0.2)_2 = 0.\overline{0011}$. Then,

$$\begin{aligned}(3.2)_2 &= (3)_2 + (0.2)_2 \\ &= 11 + 0.\overline{0011} \\ &= 11.\overline{0011}.\end{aligned}$$

Write a Python function that converts a decimal number (including fractional parts) to binary. Use it to check your answers above.

Problem 3. Find the first 7 bits of the binary representation of e. Do you expect this number will eventually repeat? Why or why not? (Recall e is the transcendental number $e = 2.7182818284\ldots$).

Solution. We find the first 7 bits using manually. Consider the integer part, 2. When we divide 2 by 2, we get 1 with remainder 0. Next, we consider the fractional part, 0.718281828. When we multiple this by 2, we get 1.436563656. We multiply by 2 a second time so that we get $0.436563656 \cdot 2 = 0.873127312$. We multiple by 2 a third time to get $0.873127312 \cdot 2 = 1.746254624$. We multiply by 2 a fourth time to get $0.746254624 \cdot 2 = 1.492509248$. We multiple a fifth time to get $0.492509248 \cdot 2 = 0.985018496$. We multiply a sixth time to get $0.985018496 \cdot 2 = 1.970036992$. We multiply by 2 one last time to get $0.9700369922 = 1.940073984$. That is, for the fractional part, we have0.1011011. Then, for the first seven bits,

$$\begin{aligned}(2.718281828)_2 &= (2)_2 + (0.718281828)_2 \\ &= 10.1011011\end{aligned}$$

We also use the Python code to check our answer. We should not expect this number to repeat because it is a transcendental irrational number. So, its binary formulation will have an infinite non-repeating form.

Problem 4. Convert the decimal number $x = 100.2$ to binary and express the number as the floating point number $fl(x)$ using the IEEE Rounding to Nearest Rule. Compute the absolute and relative errors between $fl(x)$ and the original value x.

Solution. When we convert $(100.2)_{10}$ to $(100.2)_2$, we split up the fractional and integer parts. We have already computed the fractional part in a previous problem, so we will only consider the integer part. We divide $100/2 = 50$. Next, we divide $50/2 = 25$. Then, we divide $25/2 = 12$ with remainder 1. Next, we divide $12/6 = 3$. Then, we divide $3/2 = 1$ with remainder 1. Lastly, we divide $1/2 = 0$ with remainder 1. That is, we get $100_{10} = 1100100_2$.

So,

$$\begin{aligned}(100.2)_2 &= (100)_2 + (0.2)_2 \\ &= 1100100 + 0.\overline{0011} \\ &= 1100100.\overline{0011}.\end{aligned}$$

We can express this in binary scientific notation as $1100100.\overline{0011} = 1.100100\overline{0011} \times 2^6$. Notice that the tail is 0.4×2^{-48}. Using the IEEE Rounding to the Nearest Rule, we get $2^{-52} \times 2^6 = 2^{26}$. We can compute the error, which is represented as $error = 2^{-46} - 0.8 \times 2^{-46} = 0.2 \times 2^{-46}$. Therefore,

$$fl(100.2) = 100.2 + (0.2 \times 2^{-46})$$

The absolute error gives us

$$\begin{aligned}\text{absolute error} &= |fl(x) - x| \\ &= |(100.2 + 0.2 \times 2^{-46}) - 100.2| \\ &= 0.2 \times 2^{-46}.\end{aligned}$$

Then, the relative error gives us

$$\begin{aligned}\text{relative error} &= \frac{|fl(x) - x|}{|x|} \\ &= \frac{|(100.2 + 0.2 \times 2^{-46}) - 100.2|}{100.2} \\ &= \frac{1}{100} \cdot 2^{-46}\end{aligned}$$

Problem 5. Convert each of the following binary numbers to base 10 using the method discussed in class. Be sure to show every step of your work.

(a) $(110111.001)_2$

Solution. Similar to how we executed the base 10 to base 2, we will split it up by integer and fractional parts. First, the integer part, which gives us $2^5 + 2^4 + 2^2 + 2^1 + 2^0 = 32 + 16 + 4 + 2 + 1 = 55$. For the fractional part, $2^{-3} = \frac{1}{8} = 0.125$. That is, $(110111.001)_{10} = (110111)_{10} + (0.001)_{10} = 55 + 0.125 = 55.125$.

(b) $(111.\overline{001})_2$.

Solution. Similar to how we executed the base 10 to base 2, we will split it up by integer and fractional parts. First, the integer part gives us $2^2 + 2^1 + 2^0 = 7$. Next, we consider the fractional part. The fractional part has an infinite tail. To do this, we find $\frac{\text{value of repeating part in base 10}}{2^{\text{length of repeating part}}-1}$. Notice that 001 in base 10 is 1, so that $\frac{\text{value of repeating part in base 10}}{2^{\text{length of repeating part}}-1} = \frac{1}{7}$. Then, $(111.\overline{001})_{10} = (111)_{10} + (0.\overline{001})_{10} = 7 + \frac{1}{7} = \frac{50}{7}$.

Next, write a Python function that converts a binary number to its decimal (base 10) equivalent, and use it to verify your results above.

Problem 6. Determine for which values of x the following expressions contain a subtraction of nearly equal numbers and find an alternate form that avoids the problem.

(a) $\dfrac{1 - \sec(x)}{\tan^2(x)}$

Solution. The values that contain nearly equal numbers are x values that approach $0 \pm 2k\pi$ for any $k \in \mathbb{Z}$. First, note that $\tan^2(x) = \sec^2(x) - 1 = -(1 - \sec(x))(1 + \sec(x))$. Then,

$$\begin{aligned} \frac{1 - \sec(x)}{\tan^2(x)} &= \frac{1 - \sec(x)}{-(1 - sec(x))(1 + \sec(x))} \\ &= -\frac{1}{1 + \sec(x)}. \end{aligned}$$

(b) $\dfrac{\sqrt{1+x} - 1}{x}$

Solution. The values that contain nearly equal numbers are x values that approach 0.

$$\begin{aligned} \frac{\sqrt{1+x} - 1}{x} &= \frac{\sqrt{1+x} - 1}{x} \cdot \frac{\sqrt{1+x} + 1}{\sqrt{1+x} + 1} \\ &= \frac{x}{x(\sqrt{1+x} + 1)} \\ &= \frac{1}{\sqrt{1+x} + 1}. \end{aligned}$$

Next, implement the numerically stable reformulation and compare its outputs with those of the original expression (e.g., near the problematic points). Use this comparison to verify your results above.

Problem 7. Find the Taylor polynomial of degree 5 about the point $x = 0$ for $f(x) = \cos(3x)$. Determine the order of the remainder term using Big-O notation.

Solution. We are interested in finding the maclaurin series of degree 5 polynomial. In general the Maclaurin series takes the form

$$f(h) = \sum_{k=0}^{n} \frac{f^{(k)}(0)}{k!} h^k + R_n.$$

We are interested in setting $n = 5$ so that

$$f(h) = \sum_{k=0}^{5} \frac{f^{(k)}(0)}{k!} h^k + R_n.$$

Set $f(x) = \cos(3x)$. Then, for $x = 0$, we can find the derivatives.

$$\begin{aligned} f(0) &= 1 \\ f'(0) &= 0 \\ f''(0) &= -9 \\ f'''(0) &= 0 \\ f^{(4)}(0) &= 81 \\ f^{(5)}(0) &= 0. \end{aligned}$$

Then,

$$\begin{aligned} f(h) &= 1 - \frac{9}{2}h^2 + \frac{81}{4!}h^4 + R_n \\ &= 1 - \frac{9}{2}h^2 + \frac{27}{8}h^4 + R_n. \end{aligned}$$

Next, we want to consider the remainder, which is defined as

$$\begin{aligned} R_n &= \frac{f^{n+1}(\xi)}{(n+1)!} h^{n+1} \\ &= \frac{f^6(\xi)}{6!} h^6 \\ &= O(h^6), \end{aligned}$$

where $\xi \in (0, h)$. Finally, using the big notation, we have that our final answer is

$$f(h) = 1 - \frac{9}{2}h^2 + \frac{27}{8}h^4 + O(h^6)$$