CPSC 303, 2024/25, Term 2, Assignment 1

Released Wednesday, January 15, 2025
Due Wednesday, January 29, 2025, 11:59pm

1. Consider the function $f(x) = \sqrt{1+x}$.

   (a) Show that the two-term Taylor expansion for this function is $T_2(f) = 1 + \frac{x}{2}$.

   Solution:
   First, we can get that

   $$f'(x) = \frac{d}{dx}\sqrt{1+x} = \frac{1}{2\sqrt{1+x}}$$

   Then we can compute, $f(0), f'(0)$.

   $$f(0) = \sqrt{1} = 1, f'(0) = \frac{1}{2\sqrt{1}} = \frac{1}{2}$$

   Thus, since $T_2(f) = f(0)x^0 + f'(0)x^1$,

   $$T_2(f) = 1 \cdot x^0 + \frac{1}{2}x^1 = 1 + \frac{x}{2}$$

   (b) Determine the three-term Taylor expansion of this function about $x_0 = 0$, that is, find

   $$T_3(f) = f(0) + xf'(0) + \frac{x^2}{2}f''(0).$$

   Solution:
   From the $f'(x)$ above, $f''(x)$ is,

   $$f''(x) = \frac{d}{dx}f'(x) = \frac{d}{dx}\frac{1}{2\sqrt{1+x}} = -\frac{1}{4(1+x)^{\frac{3}{2}}}$$

   Therefore, $f''(0) = -\frac{1}{4 \cdot 1^{\frac{3}{2}}} = -\frac{1}{4}$.

   Now we can get $T_3(f)$,

   $$T_3(f) = 1 + \frac{1}{2} \cdot x - \frac{1}{4} \cdot \frac{x^2}{2} = 1 + \frac{1}{2}x - \frac{1}{4}x^2$$

(c) Compute the absolute error and the relative error in using $T_3(f)$ as an algorithm for approximating $f(x)$ for $x = 0.1$, namely $f(0.1) = \sqrt{1.1} = 1.0488....$

Solution:

From the problem we are given that $y = 1.0488$. Now let's compute $\bar{y}$.

$$\bar{y} = 1 + \frac{1}{2} \cdot 0.1 - \frac{1}{4} \cdot (0.1)^2 = 1.0475$$

Now let's compute the absolute error,

$$|\bar{y} - y| = |1.0475 - 1.0488| = 0.0013$$

The relative error is,

$$\left|\frac{\bar{y} - y}{y}\right| = \left|\frac{0.0013}{1.0488}\right| = 0.0012395118$$

(d) Repeat the same computations for the two-term Taylor expansion $T_2(f)$ with $x = 0.1$, and determine which of $T_2(f)$ and $T_3(f)$ gives you a better approximation.

Solution:

Let's calculate the relative and absolute error for $T_2(f)$,

$$|\bar{y} - y| = |1 + \frac{0.1}{2} - 1.0488| = 0.0012$$

$$\left|\frac{\bar{y} - y}{y}\right| = \left|\frac{0.0012}{1.0488}\right| = 0.0011441648$$

Revising what we've computed at (c), we can see that the absolute and relative error is smaller for $T_2(f)$ hence concluding that $T_2(f)$ gives a better approximation.

(e) What are the relative backward errors for $T_2(f)$ and $T_3(f)$?

Solution:

First, for $T_3(f)$, let's search for $\bar{x}$ such that $f(\bar{x}) = \bar{y} = 1.0475$.

$$f(\bar{x}) = \sqrt{1 + \bar{x}} = 1.0475 \rightarrow \bar{x} = 1.0475^2 - 1 = 0.09725625$$

Therefore, the relative backward error for $T_3(f)$ is,

$$\left|\frac{\bar{x} - x}{x}\right| = \left|\frac{0.1 - 0.09725625}{0.1}\right| = 0.0274375$$

Second, for $T_2(f)$,

$$f(\bar{x}) = \sqrt{1 + \bar{x}} = 1.5 \rightarrow \bar{x} = 1.05^2 - 1 = 0.1025$$

$$\left|\frac{\bar{x} - x}{x}\right| = \left|\frac{0.1 - 0.1025}{0.1}\right| = 0.025$$

(f) Use the formula given in the textbook in the framed box on page 5 to explain the difference between the errors for $T_2(f)$ and $T_3(f)$. The last term in that formula (the one involving ξ) represents the error in the approximation and may be useful. We do not know the exact value of ξ but we can still bound the expressions for the error.

Solution:

First, let's keep with $x_0 = 0$ and analyze the formula from the textbook.

$$f(h) = \underbrace{f(0) + h \cdot f'(0)}_{T_2(f)(h)} + \frac{h^2}{2!} \cdot f''(\xi_1) = \underbrace{f(0) + h \cdot f'(0) + \frac{h^2}{2!} \cdot f''(0)}_{T_3(f)(h)} + \frac{h^3}{3!} \cdot f'''(\xi_2)$$

where $\xi_1, \xi_2 \in (0, h)$. So we can see the difference among errors,

2. Consider the problem of evaluating the function $f(x) = \tan(x)$.

   (a) Write down the condition number of the problem. You may use the formula

   $$\kappa(x) = \left| \frac{x f'(x)}{f(x)} \right| .$$

   Solution:
   Using the fact that $f(x) = \tan(x)$, we also know that $f'(x) = \sec^2(x)$, then

   $$\kappa = \left| \frac{x' f(x)}{f(x)} \right| = \left| \frac{x \sec^2(x)}{\tan(x)} \right| = \left| \frac{x}{\sin(x)\cos(x)} \right|$$

   (b) Explain why the problem of evaluating $f(x)$ near $\frac{\pi}{2}$ is ill-conditioned, and give $x = 1.57079$ as a specific example to illustrate your point.
   Solution:
   We know that near $\frac{\pi}{2}$,

   $$\lim_{x \to \frac{\pi}{2}} \left| \frac{x}{\sin(x)\cos(x)} \right| = \infty$$

   Using $x = 1.57079$ which is near $\frac{\pi}{2}$,

   $$\kappa = \left| \frac{1.57079}{\sin(1.57079)\cos(1.57079)} \right| = 248,275.7898120366$$

   And this is very large, thus showing that the problem of evaluating $f(x)$ near $\frac{\pi}{2}$ is ill-conditioned.

   (c) Take now $x = 1$, and determine whether the problem of evaluating $f(x) = \tan(x)$ at or near that value is well conditioned.

Solution:

Let's see what κ looks like with $x = 1$,

$$\kappa = \left| \frac{1}{\sin(1)\cos(1)} \right| = 2.1995003406$$

This is rather small, showing that the problem of evaluating $f(x) = \tan(x)$ at or near that value is well conditioned.

3. Consider the floating point system given by $(\beta, t, L, U) = (10, 4, -30, 30)$, using rounding.

(a) What is η, the unit roundoff?

Solution:

From the lecture, we learn that

$$\eta = \frac{1}{2} \cdot \beta^{1-t}$$

therefore,

$$\eta = \frac{1}{2} \cdot (10)^{1-4} = \frac{1}{2} \cdot 10^{-3}$$

(b) What is the smallest positive number in this system?

Solution:

The smallest positive number is,

$$1.000 \times 10^{L} = 1.000 \times 10^{-30}$$

(c) The smallest positive number in the system is added to 1. What is the result of this calculation on this floating point system?

Solution:

The system has a precision of $t = 4$,

$$1.000 + 1.000 \times 10^{-30} = 1.\underbrace{00\ldots00}_{\text{29 zeros}}1$$

The result will be 1.000 due to our precision.

(d) The algebraically smallest number in the system (which is negative) is added to 1. What is the result of this calculation on this floating point system?

Solution:

The algebraically smallest number in the system is -9.999×10^{30},

$$\begin{aligned} 1.000 + (-9.999 \times 10^{30}) &= 0.\underbrace{000\ldots000}_{\text{29 zeros}}1 \times 10^{30} - 9.999 \cdot 10^{30} \\ &= -9.998\underbrace{9\ldots9}_{\text{27nines}} \cdot 10^{30} \\ &= -9.999 \cdot 10^{30}, \quad \text{due to our system.} \end{aligned}$$

due to our system.

(e) What is the result of computing $1 + 1.1 * \eta$?

Solution:

$$\begin{aligned} 1 + 1.1 \cdot \eta &= 1 + 1.1 \cdot 5 \cdot 10^{-4} \\ &= 1 + 5.5 \cdot 10^{-4} \\ &= 1 + 0.00055 \\ &= 1.00055 \\ &= 1.000 \cdot 10^{0}. \end{aligned}$$

(f) What is the largest number in this system?
Solution:
The largest number is,

$$9.999 \times 10^U = 9.999 \times 10^{30}$$

(g) What is the result of computing $(10^{-20})^2$ on this system?
Solution:

$$(10^{-20})^2 = 10^{-20} \cdot 10^{-20} = 10^{-10} \cdot 10^{-30} = 0.\underbrace{00\ldots00}_{9 \text{ zeros}}1 \cdot 10^{-30} = 0 \cdot 10^{-30} = 0$$

4. *For this question, you may find it helpful to read about cancellation errors in Section 2.3 of the textbook and review our discussion in the lecture of approximating a derivative (see Section 1.2 in the textbook).* Consider the function

$$f(x) = \frac{1 - \cos(x)}{x^2}.$$

(a) Show that $0 \le f(x) < \frac{1}{2}$ for all $x \ne 0$.
Solution:
Given that $f(x) = \frac{1-\cos(x)}{x^2}$, we already know that $f(x) \ge 0$ since $1 - \cos(x) \ge 0$ and $x^2 \ge 0$. Now let's look at the behavior near $x = 0$,

$$\lim_{x \to 0} \frac{1 - \cos(x)}{x^2} = \lim_{x \to 0} \frac{\sin^2(x)}{x^2} \cdot \frac{1}{1 + \cos(x)} = \frac{1}{2}$$

Let's use the trigonometric property that $\cos(x) = 1 - 2\sin^2(\frac{x}{2})$, also we know that for any $t \in \mathbb{R}, \sin^2(t) \le t^2$. Since $x \ne 0$,

$$\sin^2(x) \le x^2 \to \frac{\sin^2(x)}{x^2} \le 1$$

If we replace x with $\frac{x}{2}$ we get,

$$\sin^2(\frac{x}{2}) \le (\frac{x}{2})^2 \to \frac{\sin^2(\frac{x}{2})}{x^2} \le \frac{1}{4} \to \frac{\frac{1-\cos(x)}{2}}{x^2} \le \frac{1}{4}$$

$$\therefore \frac{1 - \cos(x)}{x^2} \le \frac{1}{2}$$

However, since we know that the value $\frac{1}{2}$ is achieved at the limit near $x = 0$, excluding $x = 0$ we get,

$$\therefore \quad 0 \le \frac{1 - \cos(x)}{x^2} < \frac{1}{2}$$

as required.

(b) The MATLAB command

```
x=single(3e-4);
```

generates a variable x in single precision, with value $x = 3 \cdot 10^{-4}$. Write a short MATLAB script that computes $f(x)$ for the above particular value of x in single precision. Now, repeat the same calculation in double precision (MATLAB's default). (Use `format long` to see more digits in your output.) Explain the difference in the results and provide a well-justified reason for this difference, based on analyzing the error and the properties of the single precision and the double precision floating point systems. What goes wrong with the single-precision computation?

(c) Use the formula

$$\cos(x) = 1 - 2\sin^2\left(\frac{x}{2}\right)$$

to rewrite the formula for $f(x)$. Repeat your calculations for the same value of x in single and double precision. Explain the results.

5. Suppose we are given the four data points $(-1,1), (0,1), (1,2), (2,0)$. In determining the interpolating polynomials below use mainly pen and paper, but you may use MATLAB to solve linear systems or perform any calculations. You may also use the MATLAB command `polyfit` or any other MATLAB command to verify the correctness of your results.

(a) Determine the interpolating cubic polynomial using the monomial basis.

Solution:

We have $n = 3$, if we construct a linear system we get,

$$A = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 0 \end{bmatrix}$$

with the interpolant $p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$.

If we solve for the linear system $Ac = \mathbf{y}$ where $c = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}$

$$c = \begin{bmatrix} 1 \\ \frac{7}{6} \\ \frac{1}{2} \\ -\frac{2}{3} \end{bmatrix}$$

Hence, we get the cubic polynomial,

$$\therefore p(x) = 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3$$

(b) Determine the interpolating cubic polynomial using the Lagrange basis.

Solution:

Let's construct the four basis functions, $\phi_0(x), \phi_1(x), \phi_2(x), \phi_3(x)$.

$$\phi_0(x) = \frac{x(x-1)(x-2)}{-(-1-1)(-1-2)} = -\frac{x(x-1)(x-2)}{6}$$

$$\phi_1(x) = \frac{(x+1)(x-1)(x-2)}{(0-1)(0-2)} = \frac{(x+1)(x-1)(x-2)}{2}$$

$$\phi_3(x) = \frac{(x+1)x(x-2)}{(1+1)(1-2)} = -\frac{x(x+1)(x-2)}{2}$$

$$\phi_4(x) = \frac{(x+1)x(x-1)}{(2+1)2(2-1)} = \frac{x(x+1)(x-1)}{6}$$

By letting $c_j = y_j$,

$$\therefore p(x) = -\frac{x(x-1)(x-2)}{6} + \frac{(x+1)(x-1)(x-2)}{2} - x(x+1)(x-2)$$

(c) Determine the interpolating cubic polynomial using the Newton basis. Generate both the lower triangular system and the divided difference table.
Solution:
The newton basis will have the basis functions,

$$\phi_0(x) = 1, \phi_1(x) = (x+1), \phi_2(x) = (x+1)x, \phi_3(x) = (x+1)x(x-1)$$

Let's construct the linear system,

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 1 & 3 & 6 & 6 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 0 \end{bmatrix}$$

Then,

$$c = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \\ -\frac{2}{3} \end{bmatrix}$$

Giving us the interpolating cubic polynomial,

$$\therefore p(x) = 1 + \frac{1}{2}x(x+1) - \frac{2}{3}x(x-1)(x+1)$$

(d) Show that the three representations give the same polynomial.
Solution:
Let's first look at the polynomial from (a),

$$p_a(x) = 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3$$

from (b),

$$p_b(x) = -\frac{x(x-1)(x-2)}{6} + \frac{(x+1)(x-1)(x-2)}{2} - x(x+1)(x-2)$$

$$= (-\frac{1}{6}x^3 + \frac{1}{2}x^2 - \frac{1}{3}x) + (\frac{1}{2}x^3 - x^2 - \frac{1}{2}x + 1) - (x^3 - x^2 - 2x)$$

$$= 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3$$

from (c),

$$p_c(x) = 1 + \frac{1}{2}x(x+1) - \frac{2}{3}x(x-1)(x+1)$$
$$= 1 + \frac{1}{2}x + \frac{1}{2}x^2 - \frac{2}{3}x^3 + \frac{2}{3}x = 1 + \frac{7}{6}x + \frac{1}{2}x^2 - \frac{2}{3}x^3$$

Thus, we can see that $p_a(x) = p_b(x) = p_c(x)$, showing the three representations give the same polynomial.

6. Suppose we want to approximate e^x on $[0, 1]$, by using polynomial interpolation with $x_0 = 0, x_1 = \frac{1}{2}$ and $x_2 = 1$. Let $p_2(x)$ denote the interpolating polynomial.

   (a) Find the interpolating polynomial using your favourite technique.
   Solution:
   Let's do the Netwon interpolation, the base functions are,

   $$\phi_0(x) = 1, \phi_1(x) = x, \phi_2(x) = x(x - \frac{1}{2})$$

   Then we solve for the system,

   $$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & \frac{1}{2} & 0 \\ 1 & 1 & \frac{1}{2} \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 \\ e^{\frac{1}{2}} \\ e \end{bmatrix}$$

   Which gives us,

   $$\mathbf{c} = \begin{bmatrix} 1 \\ 2e^{\frac{1}{2}} - 2 \\ 2e - 4e^{\frac{1}{2}} + 2 \end{bmatrix}$$

   Therefore, our interpolating polynomial is $p_2(x) = 1 + (2e^{\frac{1}{2}} - 2)x + (2e - 4e^{\frac{1}{2}} + 2)x(x - \frac{1}{2})$.

   (b) Find an upper bound for the error

   $$\max_{0 \le x \le 1} |e^x - p_2(x)|.$$

   Solution:
   We want to find $\max_{0 \le x \le 1} |e^x - p_2(x)| = \max_{t \in [0,1]} \frac{|f^{(3)}(t)|}{3!} \max_{s \in [0,1]} \left| \prod_{j=0}^{2} (s - x_j) \right|$ since $f(x) = e^x$ and $n = 2$ in our case.

   $$f^{(3)}(t) = e^t \to \max_{t \in [0,1]} \frac{|f^{(3)}(t)|}{3!} = \frac{e}{6}$$

   For,

   $$\max_{s \in [0,1]} \left| \prod_{j=0}^{2} (s - x_j) \right| = \max_{s \in [0,1]} \left| s(s - \frac{1}{2})(s - 1) \right|$$

   We need to find the maximum of $\left| s(s - \frac{1}{2})(s - 1) \right|$ within the range [0,1].

   $$\frac{d}{ds} s(s - \frac{1}{2})(s - 1) = 3s^2 - 3s + \frac{1}{2} = 0 \to s = \frac{3 - \sqrt{6}}{6} \in [0, 1] \text{ local maximum}$$

$|s(0)| = 0, |s(1)| = 0.$

Thus,

$$\max_{s\in[0,1]} s(s-\frac{1}{2})(s-1) = \frac{\sqrt{3}}{36}$$

Finally, the upperbound for the error is,

$$\therefore \max_{0\leq x\leq 1} |e^x - p_2(x)| = \frac{\sqrt{3}}{36}\cdot\frac{e}{6} = \frac{\sqrt{3}e}{216}$$

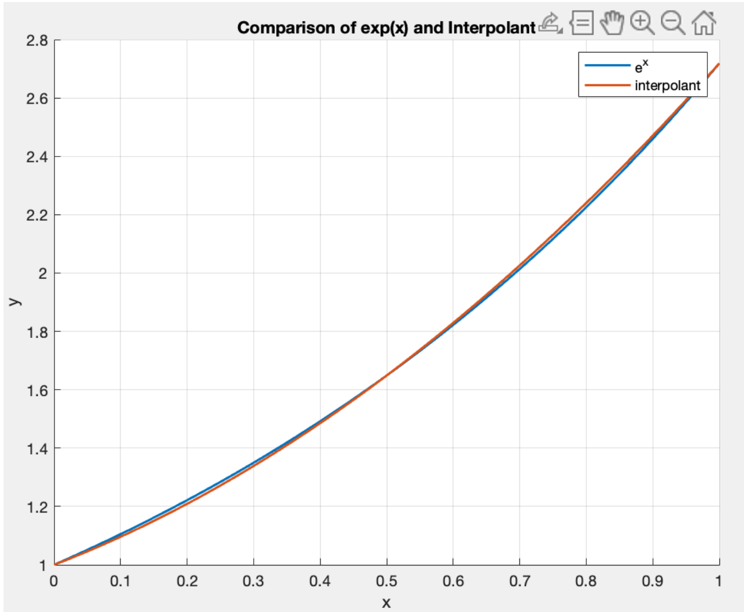
(c) Plot the function e^x and the interpolant you found, both on the same plot, using the commands `plot` and `hold`.

Solution:

The following code was used to plot the function e^x and the interpolant,

```
a = 1;
b = 2 * exp(1/2) - 2;
c = 2 * exp(1) - 4 * exp(1/2) + 2;
exponential = @(x) exp(x);
interpolant = @(x) a + b .* x + c .* x .* (x - 1/2);

hold on
fplot(exponential, [0 1], 'DisplayName', 'e^x', 'LineWidth', 1.5)
fplot(interpolant, [0 1], 'DisplayName', 'interpolant', 'LineWidth', 1.5)

legend show

xlabel('x');
ylabel('y');
title('Comparison of exp(x) and Interpolant');
grid on;

hold off;
```

Which gives the plot,



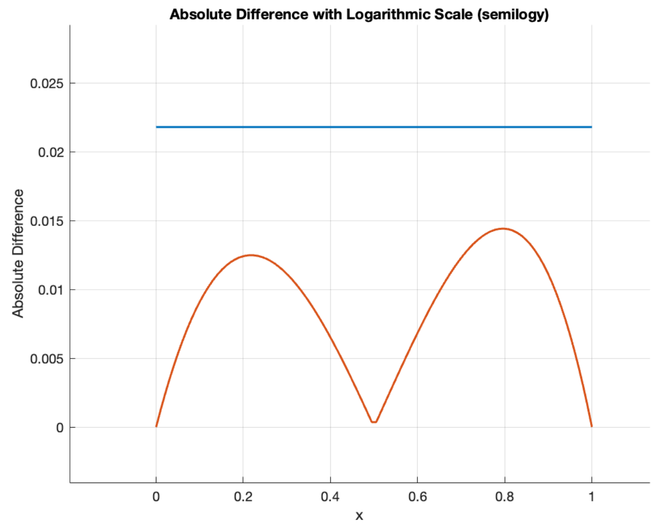
(d) Plot the absolute error $|e^x - p_2(x)|$ on the interval using logarithmic scale (the command `semilogy`) and briefly compare the error to the error bound you found in part (b).

Solution:

The following code was used to plot the absolute difference between the function e^x and the interpolant $p_2(x)$ and the error bound we found in part (b).

```
a = 1;
b = 2 * exp(1/2) - 2;
c = 2 * exp(1) - 4 * exp(1/2) + 2;
exponential = @(x) exp(x);
interpolant = @(x) a + b .* x + c .* x .* (x - 1/2);

abs_diff = @(x) abs(exp(x) - (a + b .* x + c .* x .* (x - 1/2)));

x_values = linspace(0, 1, 100);

y_values = abs_diff(x_values);

current_error_bound = sqrt(3) * exp(1) / 216;
error_bound_line = current_error_bound * ones(size(x_values));
display(current_error_bound)


hold on;
semilogy(x_values, error_bound_line, 'LineWidth', 1.5);
semilogy(x_values, y_values, 'LineWidth', 1.5);
xlabel('x');
ylabel('Absolute Difference');
title('Absolute Difference with Logarithmic Scale');
grid on;
```

When plotted, we get the following plot,



We can see that the absolute error of the interpolant and the function is lower than the maximum bound that we have computed.