Cameron Beneteau
400274092
beneteac

MECHTRON 3X03: A1

Due: Oct/6/23

1. $$\left| \frac{\sin x - x}{\sin x} \right| \leq 0.5 \times 10^{-14} = 5 \times 10^{-15}$$

Taylor Series Approximation (centered at 0)

$$\sin(x) = \sin(0) + \cos(0)\,x - \frac{1}{2}\sin(0)\,x^2 - \frac{1}{6}\cos(0)\,x^3$$

(0) (1) (0) (1)

$$+ \frac{1}{24}\sin(0)\,x^4 + \frac{1}{120}\cos(0)\,x^5 + \cdots$$ (negligible)

$$\sin(x) \approx x - \frac{\xi^3}{6} \quad \text{where} \quad \xi \in [0, x]$$

$$\left| \frac{\frac{-\xi^3}{6}}{x - \frac{\xi^3}{6}} \right| \leq 5 \times 10^{-15}$$

$$\left| \frac{-\xi^3}{6x - \xi^3} \right| \leq 5 \times 10^{-15}$$

$$\frac{\xi^3}{6x - \xi^3} \leq 5 \times 10^{-15}$$

$$\xi^3 \leq (5 \times 10^{-15})(6x - \xi^3)$$

$$\xi^3 \leq 3x \times 10^{-14} - \xi^3(5 \times 10^{-15})$$

$$\xi^3(1 + 5 \times 10^{-15}) \leq 3x \times 10^{-14}$$

Since $\xi \in [0, x]$, $\xi = x$ will maximize error

$$x^3(1 + 5 \times 10^{-15}) \leq 3x \times 10^{-14}$$

Let $1 + 5 \times 10^{-15} \approx 1$ and $x \neq 0$ (to divide)

$$x^2 \leq 3 \times 10^{-14}$$

$$\boxed{|x| \leq \sqrt{3} \cdot 10^{-7}}$$

2a) $f(x) = e^{x+2h}$ Taylor Series at x

$$f(x) = \sum_{K=0}^{\infty} \frac{f^{(K)}(c)}{K!}(x-c)^K$$

$x = x+2h$ and $c = x$

$$f(x+2h) = \sum_{K=0}^{\infty} \frac{f^{(K)}(x)}{K!}(2h)^K$$

$$\frac{d^K}{dx^K}\left(e^{x+2h}\right) = e^{x+2h} \text{ for all } K$$

$$f(x+2h) = \sum_{K=0}^{\infty} \frac{e^{x+2h}}{K!}(2h)^K$$

$$= e^{x+2h} + 2h \cdot e^{x+2h} + \frac{4h^2 \cdot e^{x+2h}}{2} + \cdots$$

2b) $f(x) = \sin(x - 3h)$ Taylor Series at x

$$f(x) = \sum_{K=0}^{\infty} \frac{f^{(K)}(c)}{K!}(x-c)^K \qquad \begin{array}{l} x = x - 3h \\ c = x \end{array}$$

$$f(x-3h) = \sum_{K=0}^{\infty} \frac{f^{(K)}(x)}{K!}(-3h)^K$$

$$\frac{d}{dx}(\sin x) = \cos x \qquad \frac{d}{dx}(\cos x) = -\sin x$$

$$\frac{d}{dx}(-\sin x) = -\cos x \qquad \frac{d}{dx}(-\cos x) = \sin x$$

$$\frac{d}{dx}(x - 3h) = 1$$

$$f(x-3h) = \sin(x-3h) - 3h \cdot \cos(x-3h) - \frac{9h^2 \cdot \sin(x-3h)}{2} + \frac{27h^3 \cos(x-3h)}{6} + \cdots$$

3) $f(x) = e^x$ approximate by truncated Maclaurin
For $x = 0.5$, how many terms for abs acc $\leq 10^{-10}$

$$e^x = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k = 1 + x + x^2/2 + x^3/6 + x^4/24 + x^5/120 + \cdots$$

$$E_{n+1} = \frac{f^{(n+1)}(\xi)}{(n+1)!} \xi^{n+1} \qquad \left(\text{Note: } f^{(n+1)}(\xi) = e^{\xi} \text{ for any } n\right)$$

$$E_{n+1} = \frac{e^{\xi}(\xi)^{n+1}}{(n+1)!} \leq 10^{-10} = \frac{e^{0.5} \cdot 0.5^{n+1}}{(n+1)!}$$

Trial and Error until $E_{n+1} \leq 10^{-10}$

n	7	8	9	10
E_{n+1}	$1.6e^{-7}$	$8.9e^{-9}$	$4.4e^{-10}$	$2.0e^{-11}$

Need at least 10 terms for error $\leq 10^{-10}$

4) $1-a^2$ IEEE 754 double precision (64-bits)
what vals of 'a' evaluate $1-a^2 = 1$

- no unit roundoff in a
- roundoff error in a^2: δ_1
- roundoff error in $1-a^2$: δ_2

$$fl(a^2) = fl(a^2)(1+\delta_1)$$

$$fl(1-a^2) = Fl(1-fl(a^2))(1+\delta_1)(1+\delta_2)$$

$$= (1-a^2)(1+\delta_1)(1+\delta_2)$$

$$= (1-a^2)(1+\delta_1+\delta_2+\delta_1\delta_2)$$

$$= (1-a^2)(1+\delta)$$

assume negligible $\rightarrow 0$

$$\delta = \delta_1 + \delta_2 + \delta_1\delta_2 \qquad |\delta_1|, |\delta_2| \leq u \qquad |\delta| \leq 2u$$

$$u = \frac{\epsilon_{mach}}{2} \quad \text{so} \quad (1-a^2)(1+\epsilon_{mach})$$

$$1 = (1 - a^2)(1 + \epsilon_{mach})$$

$$1 = 1 - a^2\epsilon_{mach} - a^2 + \epsilon_{mach}$$

$$a^2\epsilon_{mach} + a^2 = \epsilon_{mach}$$

$$a^2(\epsilon_{mach} + 1) = \epsilon_{mach}$$

$$a = \sqrt{\frac{\epsilon_{mach}}{\epsilon_{mach} + 1}}$$

$$\boxed{|a| \leq \sqrt{\frac{\epsilon_{mach}}{1 + \epsilon_{mach}}}}$$

Note: If we can't assume $\delta_1 \cdot \delta_2 \approx 0$, then

$$1 = (1 - a^2)\left(1 + \epsilon_{mach} + \left(\frac{\epsilon_{mach}}{2}\right)^2\right)$$

which gives us

$$|a| \leq \sqrt{\frac{\epsilon_{mach} + \frac{1}{4}\epsilon_{mach}^2}{1 + \epsilon_{mach} + \frac{1}{4}\epsilon_{mach}^2}}$$

5a) $(a+b)+c \neq a+(b+c)$

$a = 0.0002 \quad b = 4000 \quad c = -4000 \quad t = 4$

$(a+b)+c$

$= (0.0002 + 4000) - 4000$

$= 4000.0002 - 4000$

$= 4000 - 4000$

$= 0$

roundoff looses precision

$fl(4000.0002) = 4000$

$a+(b+c)$

$= 0.0002 + (4000 - 4000)$

$= 0.0002 + 0$

$= 0.0002$

no roundoff error

$a = 0.0002 \quad b = 4000 \quad c = -4000 \quad t = 4$

$(a+b)+c \neq a+(b+c) \quad 0 \neq 0.0002$

b) $(a \cdot b) \cdot c \neq a \cdot (b \cdot c)$

$a = 1.234 \quad b = 2.003 \quad c = 3.456 \quad t = 4$

$(a \cdot b) \cdot c$

$= (1.234 \cdot 2.003)(3.456)$

$= 2.471702 \cdot 3.456$

$\approx 2.472 \cdot 3.456$

$= 8.543232$

$= 8.543$

roundoff looses precision

$fl(2.471702) = 2.472$

$a \cdot (b \cdot c)$

$= 1.234 \cdot (2.003 \cdot 3.456)$

$\approx 1.234 \cdot 6.922368$

$= 1.234 \cdot 6.922$

$= 8.541748$

$= 8.542$

less roundoff error

$a = 1.234 \quad b = 2.003 \quad c = 3456 \quad t = 4$

$(a \cdot b) \cdot c \neq a \cdot (b \cdot c) \quad 8.543 \neq 8.542$

6a) $|x_i - \tilde{x}_i|$ vs. $|x_i - \hat{x}_i|$

$\tilde{x}_i$: Rounding error from one addition calculation AND rounding error from all previous terms. Error grows with i since each each new term accumulates the rounding errors of previous.

$\hat{x}_i$: Possible rounding error from one calculation of addition/multiplication. Each term will only ever have the rounding error of its own calculation.

$\tilde{x}_0 = a$

$\tilde{x}_1 = fl(a+h) = (a+h)(1+\delta_1)$

$\tilde{x}_2 = fl(\tilde{x}_1+h) = (a+h)(1+\delta_1)(1+\delta_2)$

$\vdots \qquad \vdots \qquad \vdots$

$\tilde{x}_n = fl(\tilde{x}_{n-1}+h) = (\tilde{x}_{n-1}+h)(1+\delta)^{n-1}$

$\hat{x}_0 = a$

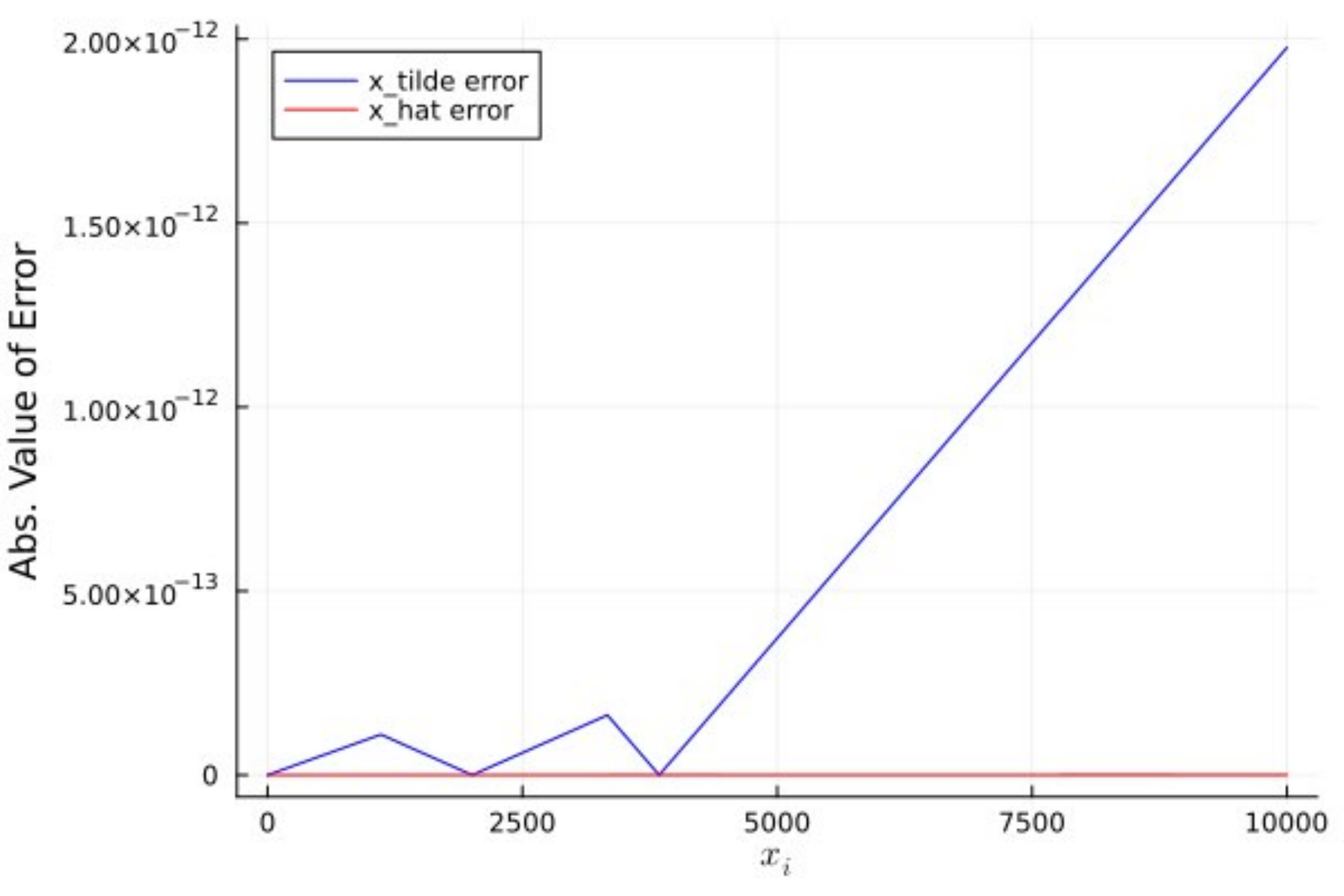
$\hat{x}_1 = fl(a+1h)(1+\delta_1)(1+\delta_2)$

$\hat{x}_2 = fl(a+2h)(1+\delta_1)(1+\delta_2)$

$\vdots \qquad \vdots \qquad \vdots$

$\hat{x}_n = fl(a+nh)(1+\delta_1)(1+\delta_2)$

As the accumulated rounding error in $|x_i - \tilde{x}_i|$ grows with i, $|x_i - \hat{x}_i|$ is more accurate.


x_tilde error
x_hat error
Abs. Value of Error
2.00×10^−12
1.50×10^−12
1.00×10^−12
5.00×10^−13
0
0
2500
5000
7500
10000
x_i

7) $f'(x) \approx \dfrac{f(x+h)-f(x-h)}{2h}$, $h>0$, $f'''(x)$ cont. on $[x-h, x+h]$

a)

① $f(x+h) \approx f(x) + h\cdot f'(x) + \dfrac{h^2\cdot f''(x)}{2} + \dfrac{h^3\cdot f'''(\xi)}{6}$

② $f(x-h) \approx f(x) - h\cdot f'(x) + \dfrac{h^2\cdot f''(x)}{2} - \dfrac{h^3\cdot f'''(\xi)}{6}$

① − ②

$$f(x+h)-f(x-h) \approx 2h\cdot f'(x) + \frac{2h^3\cdot f'''(\xi)}{6}$$

$$\frac{f(x+h)-f(x-h)}{2h} \approx f'(x) + \frac{h^2\cdot f'''(\xi)}{6}$$

$$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h} - \frac{h^2\cdot f'''(\xi)}{6}$$

Truncation error is $-\dfrac{h^2\cdot f'''(\xi)}{6}$, $\xi \in [x-h, x+h]$

b) $$f'(x) \approx \frac{f(x+h)-f(x-h)}{2h} - \frac{h^2 \cdot f'''(\xi)}{6} \quad (1)$$

Let $f_1 = f(x+h)$ for some δ_1

Let $f_2 = f(x-h)$ for some δ_2

$$\frac{f_1 - f_2}{2h} = \frac{f(x+h)-f(x-h)}{2h} + \frac{\delta_1 - \delta_2}{2h} \quad (2)$$

(1) - (2)

$$f'(x) - \frac{f_1 - f_2}{2h} = -\frac{h^2 \cdot f'''(\xi)}{6} - \frac{\delta_1 - \delta_2}{2h}$$

Denote M the max of $|f''(\xi)|$ for $\xi \in [x-h, x+h]$ and assume $|\delta_1|, |\delta_2| \leq \epsilon_{mach}$

$$\left| f'(x) - \frac{f_1 - f_2}{2h} \right| = \left| -\frac{h^2 \cdot f'''(\xi)}{6} - \frac{\delta_1 - \delta_2}{2h} \right|$$

$$\leq \left| \frac{h^2 \cdot f'''(\xi)}{6} \right| + \left| \frac{\delta_1 - \delta_2}{2h} \right|$$

$$\leq \frac{Mh^2}{6} + \frac{\epsilon_{mach}}{h}$$

Where M is the max of $f'''(\xi)$, $\xi \in [x-h, x+h]$

$$g(h) = \frac{Mh^2}{6} + \frac{\epsilon_{mach}}{h}$$

Using first order optimolity condition:

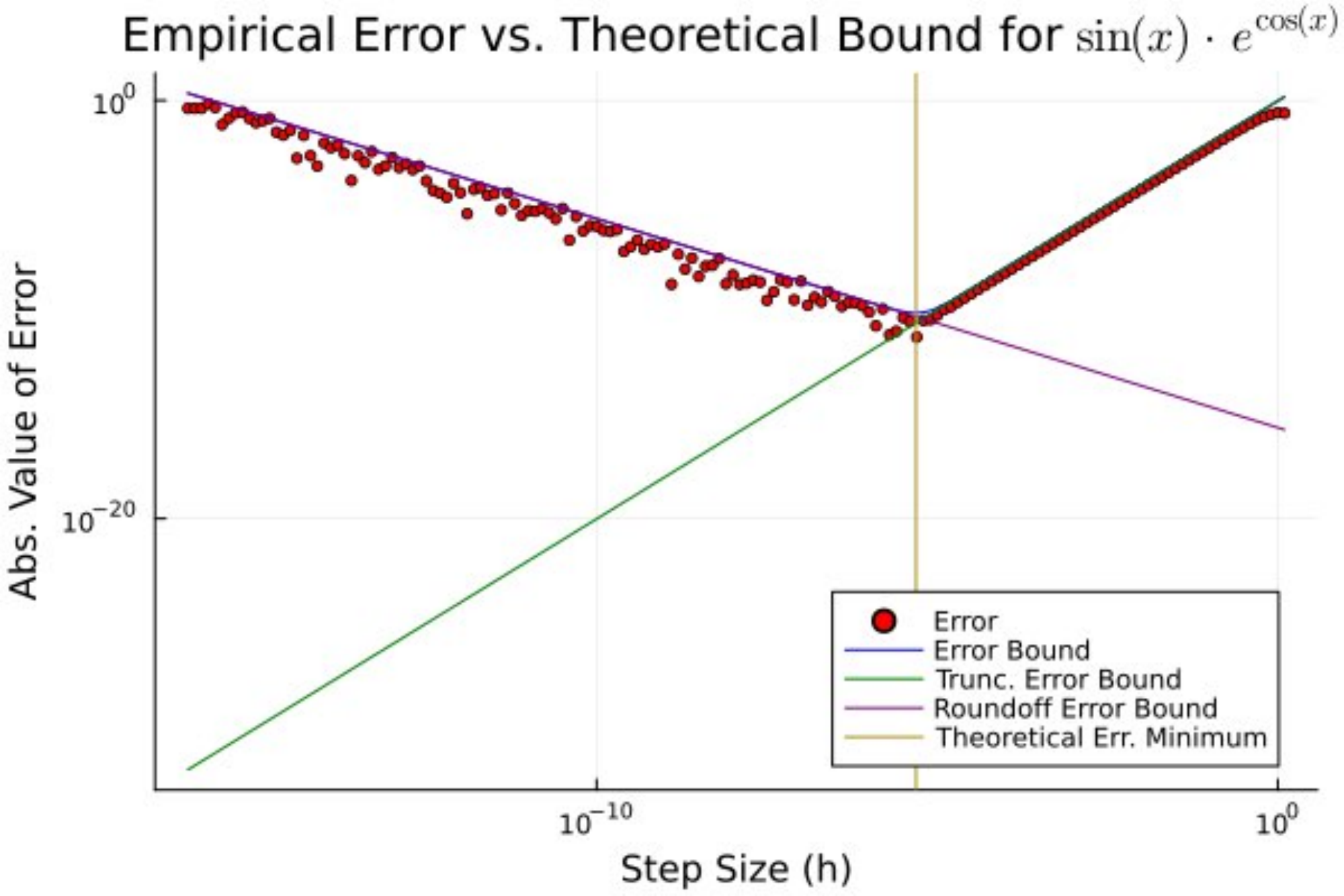
$$g'(h) = \frac{d}{dh}(g(h)) = 0$$

$$g'(h) = 0 = \frac{Mh}{3} - \frac{\epsilon_{mach}}{h^2}$$

$$\frac{Mh}{3} = \frac{\epsilon_{mach}}{h^2}$$

$$Mh^3 = 3\epsilon_{mach}$$

$$h = \sqrt[3]{\frac{3\epsilon_{mach}}{M}}$$ is the value of h that gives the smallest error


Empirical Error vs. Theoretical Bound for $\sin(x) \cdot e^{\cos(x)}$
Abs. Value of Error
10^{0}
10^{-20}
10^{-10}
10^{0}
Step Size (h)
Error
Error Bound
Trunc. Error Bound
Roundoff Error Bound
Theoretical Err. Minimum

8a) The error in $f(x)$ can be large because:

1. Catastrophic cancellation
   Occurs when subtracting nearby numbers that contain roundoff. As we know, numbers in a computer are in floating-point representation, which has roundoff error. When calculating $f(x)$ as x approaches 0, $e^x \to 1$ and $x \to 0$, giving us an approximation for the numerator of $1-0-1$. Since the magnitudes of the two 1 terms are close, and contain roundoff, this may lead to a case of catastrophic cancellation, resulting in a relatively large error for the function of $f(x)$.

2. Division by a small number
   As x approaches 0, x^2 will be even smaller. Therefore, small errors in the numerator of $f(x)$ $(e^x - x - 1)$ can be amplified when dividing by a very small x^2 denominator.

b) Taylor Series Approximation of e^x:

$$f(x) = \sum_{K=0}^{\infty} \frac{f^{(K)}(c)}{K!}(x-c)^K \qquad \begin{array}{l} x = x \\ c = 0 \end{array}$$

$$f(x) = \sum_{K=0}^{\infty} \frac{x^K}{K!}$$

$$e^x \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

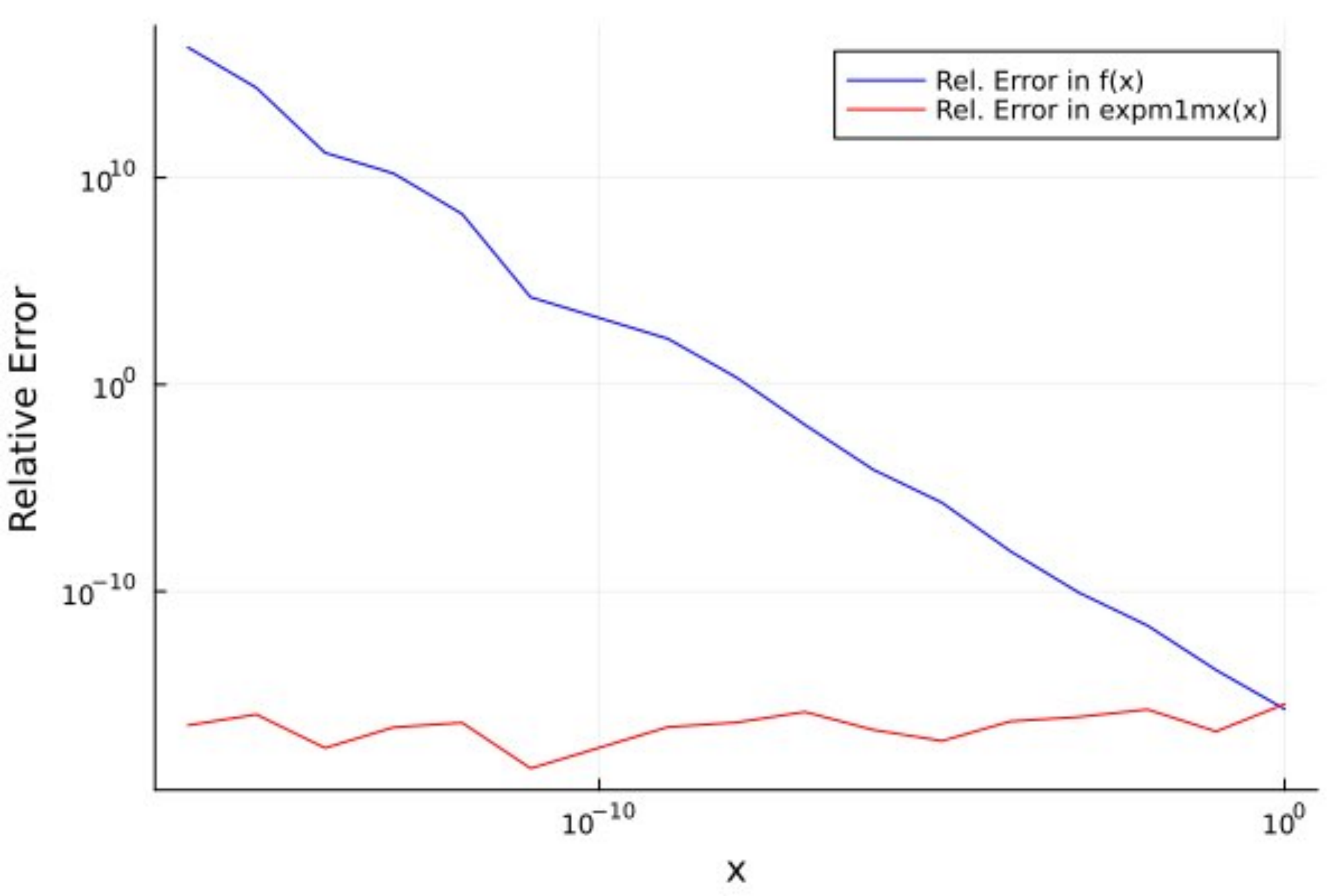
Subtract $1+x$

$$e^x - x - 1 \approx \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

Divide by x^2

$$\frac{e^x - x - 1}{x^2} \approx \frac{1}{2!} + \frac{x}{3!} + \frac{x^2}{4!} + \cdots$$

Now we have a simple series to implement in code, with very few sources of error.


Rel. Error in f(x)
Rel. Error in expm1mx(x)
Relative Error
10^{10}
10^{0}
10^{-10}
10^{-10}
10^{0}
x

9) Compare $g(x)$ and $h(x)$ Outputs

Looking at step-by-step calculations

For $x = 1 \times 10^{-10}$

$g(x)$	$h(x)$
e^x	e^x
1.0000000001	1.0000000001
$e^x - 1$	$e^x - x$
$1.0000000827403701 \times 10^{-10}$	1.0
$e^x - 1 - x$	$e^x - x - 1$
$8.274037096265818 \times 10^{-18}$	0.0
$e^x - 1 - x / x^2$	$e^x - x - 1 / x^2$
827.4037096265816	0.0

Analyzing the step by step calculations above, it is clear the discrepancy between the outputs of $g(x)$ and $h(x)$ lie in the order of operations in their numerators, specifically:

$g(x)$: $e^x - 1 - x$ vs. $h(x)$: $e^x - x - 1$

In both, $e^x = 1.0000000001$, which can be written as $(1 + 1 \times 10^{-10})$ clearly showing that this value is equal to $1 + x$, as $x = 1 \times 10^{-10}$. By inspection, subtracting $(1+x)$ from $1+x$ should give us 0, but this is not the case in $g(x)$.

Here, $e^x - 1$ in $g(x)$ suffers from Catastrophic cancellation: subtracting nearby numbers containing roundoff, resulting in a relatively large error. Conversely, $e^x - x$ in $h(x)$ subtracts 1×10^{-10} from $e^x = 1.0000000001$, numbers that are not near, which is why $h(x)$ doesn't suffer this same issue

This is why the computed value of $e^x - 1 - x$ in $g(x)$ is $8.274037096265818 \times 10^{-18}$ instead of 0, what we see for $e^x - x - 1$ in $h(x)$.

Furthermore, this error in the numerator of $g(x)$ is amplified by division of a small number. As x approaches 0, x^2 will be even smaller. Therefore, small errors in the numerator of $g(x)$ $(e^x - 1 - x)$ can be magnified when dividing by a very small x^2 denominator. This is exactly what we see in $g(x)$ as the division by x^2 brings the output to 827.4037096265816, Since the numerator of $h(x)$ calculates to 0, division by a small x^2 value has no effect on the final output.

For $x = z^{-33} = 1.1641532182693481 \times 10^{-10}$

$g(x)$	$h(x)$
e^x	e^x
1.0000000001164153	1.0000000001164153
$e^x - 1$	$e^x - x$
$1.1641532182693481 \times 10^{-10}$	1.0
$e^x - 1 - x$	$e^x - x - 1$
0.0	0.0
$e^x - 1 - x / x^2$	$e^x - x - 1 / x^2$
0.0	0.0

Analyzing the step by step calculations above, we can use the same explanation as when x was 1×10^{-16} to explain why the output of $h(x)$ remains 0, but $g(x)$ has something interesting.

In the case of $g(x)$, its output is 0 because we "got lucky" with our x value. We can see the calculation of e^x is 1.0000000001164153 and $e^x - 1$ is $1.1641532182693481 \times 10^{-10}$. Coincidentally, $e^x - 1$ is the exact same value of our x input of 2^{-33}. Therefore, when subtracting x from $e^x - 1$, we could just say this is $x - x$, which is 0. Because $e^x - 1$ is exactly equal to x in the computers floating-point representation, $e^x - 1 - x = 0$

To compare, if x was 2^{-32} or 2^{-34}, the calculation of $e^x - 1 - x$ would not line up as nicely, which would give $g(x)$ a non-zero output.