1 Practice

Solution 1.1. Suppose that $f \in C^{n+1}[a,b]$, meaning that f is $(n+1)$-times differentiable on the interval $[a,b]$, and let $x_0 \in [a,b]$. Then for any $x \in [a,b]$, there is $\xi(x) \in [a,b]$ such that

$$f(x) = f(x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x-x_0)^{n+1}.$$

The main usage of Taylor's theorem in numerical analysis is to approximate a complicated function by a (relatively much simpler) polynomial function.

Solution 1.2. The first four derivatives of cosine are:

$$f'(x) = -\sin(x) \qquad f''(x) = -\cos(x) \qquad f'''(x) = \sin(x) \qquad f''''(x) = \cos(x).$$

Evaluating at 0 gives

$$f'(0) = 0 \qquad f''(0) = -1 \qquad f'''(0) = 0 \qquad f''''(x) = 1.$$

Thus the third order Taylor approximation (left term below) with error term (right term below) is

$$f(x) = \left(1 - \frac{1}{2}x^2\right) + \left(\frac{1}{24}x^4\cos(\xi(x))\right)$$

where $\xi(x)$ is an unknown function of x. Using the approximation, we obtain

$$f(0.01) \approx 1 - \frac{1}{2}(0.01)^2 = 0.99995.$$

Based on the error term, we can obtain the bound

$$\left|\frac{1}{24}(0.01)^4\cos(\xi(0.01))\right| \leq \frac{1}{24}(0.01)^4 \approx 4.2 \times 10^{-10}$$

Solution 1.3. The main limitation of floating point arithmetic is that it cannot represent numbers exactly. As a result, even though we often take for granted that arithmetic operations (e.g. addition, subtraction, multiplication, division, exponentiation) can be evaluated easily and exactly which is very false.

In the IEEE 754 floating point standard, ε represents the increments in the mantissa of the floating point representation. This is a fundamental limit on the accuracy of floating point representations. For example, for real numbers between 1 and 2, the most accurate floating point approximation could have error up to ε depending on the choice of rounding.

Solution 1.4. Since $4 \leq 8 - 5\varepsilon < 8$, we know that $8 - 5\varepsilon \in [4, 8)$, so is represented by the characteristic $c = 1025$, so the floating point numbers in $[4, 8)$ are

$$4, 4+4\varepsilon, 4+8\varepsilon, \ldots, 8-8\varepsilon, 8-4\varepsilon.$$

Thus, the closest floating point number to $8 - 5\varepsilon$ is $8 - 4\varepsilon$, so we find $fl(8-5\varepsilon) = 8 - 4\varepsilon$. Alternatively, under the "round towards 0" convention, we would round $8 - 5\varepsilon$ down to $8 - 8\varepsilon$. I forgot to include this in the original solutions, and I have gone back to add credit to solutions which gave this answer.

Solution 1.5. First, to find which characteristic corresponds to x, we compute $c = \lfloor \log_2(|x|) \rfloor$ which is the largest power of 2 smaller than x. Then $x \in [2^c, 2^{c+1})$ and the minimum increment between floating point numbers in this interval is $2^c\varepsilon$, so this is also the maximum roundoff error. Thus,

$$|fl(x) - x| \leq 2^{\log_2(|x|)}\varepsilon = |x|\varepsilon.$$

2 Numerical Differentiation

The error of our approximation is

$$\left| f'(x) - \frac{fl(f(x+h_0)) - fl(f(x))}{h_0} \right|.$$

To simplify notation, define $a = f(x+h_0)$ and $b = f(x)$. Then

$$\begin{aligned} \left| f'(x) - \frac{fl(a) - fl(b)}{h_0} \right| &= \left| f'(x) - \frac{f(x+h_0) - f(x)}{h_0} + \frac{a-b}{h_0} - \frac{fl(a) - fl(b)}{h_0} \right| \\ &\leq \left| f'(x) - \frac{f(x+h_0) - f(x)}{h_0} \right| + \left| \frac{a-b}{h_0} - \frac{fl(a) - fl(b)}{h_0} \right|. \end{aligned}$$

Solution 2.1. Problem 2.1 deals with the left error term above using Taylor's theorem to first order with second order error. Here we are expanding $f(x+h_0)$ around x:

$$f(x+h) = f(x) + f'(x)(x+h_0) + f''(\xi(x+h))\frac{(x+h_0-x)^2}{2}.$$

Plugging this in above gives us

$$\begin{aligned} \left| f'(x) - \frac{f(x+h_0) - f(x)}{h_0} \right| &= \left| f'(x) - \frac{f(x) + f'(x)(x+h_0-x) + f''(\xi(x+h_0))\frac{(x+h_0-x)^2}{2} - f(x)}{h_0} \right| \\ &= \left| f''(\xi(x+h_0))\frac{h_0}{2} \right| \\ &\leq \max_x |f''(x)|\frac{h_0}{2}. \end{aligned}$$

Solution 2.2. Problem 2.2 deals with the right error term using floating point approximations. Recall that floating point numbers come in intervals according to the characteristic. Carefully analyzing this structure gives the inequality

$$|fl(x) - x| \leq |x|\varepsilon$$

for any real number x, where $\varepsilon = 2^{-52}$, which we will use freely.

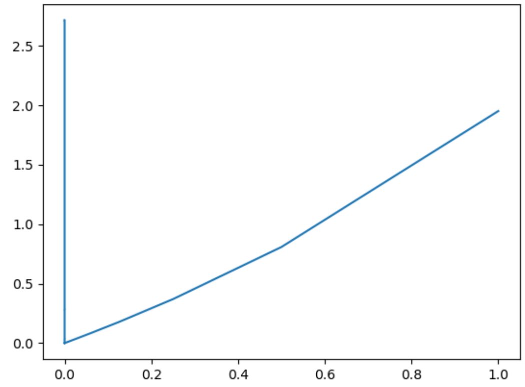
$$\begin{aligned} \left| \frac{a-b}{h_0} - \frac{fl(a) - fl(b)}{h_0} \right| &= \left| \frac{a - fl(a)}{h_0} - \frac{b - fl(b)}{h_0} \right| \\ &\leq \left| \frac{a - fl(a)}{h_0} \right| + \left| \frac{b - fl(b)}{h_0} \right| \\ &\leq \frac{|a|\varepsilon}{h_0} + \frac{|b|\varepsilon}{h_0} = (|a| + |b|)\frac{\varepsilon}{h_0}. \end{aligned}$$

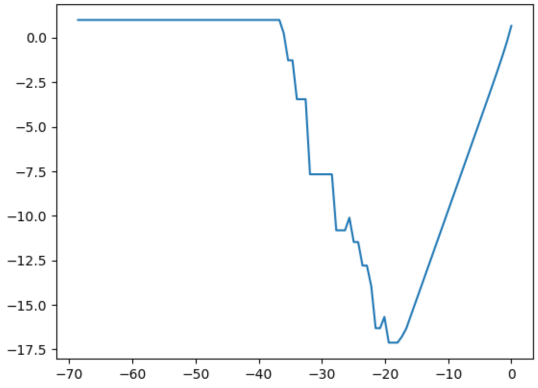
Solution 2.3. The sum of the two error terms we computed above is

$$\max_x |f''(x)|\frac{h_0}{2} + (|a| + |b|)\frac{\varepsilon}{h_0}.$$

Dropping all constants from this, we have $h_0 + \frac{\varepsilon}{h_0}$ which is minimized at $h_0 = \sqrt{\varepsilon}$. (Of course, this will depend on the actual values of the constants, but for a rough guess it's fine). The approximate error for this choice is $\sqrt{\varepsilon} + \frac{\varepsilon}{\sqrt{\varepsilon}} = 2\sqrt{\varepsilon}$.

Solution 2.4. The plots below give the error vs h_0 (above) and the log error vs $\log(h_0)$ (below). In the following, we will only refer to the second graph because it is more interpretable.





Roughly speaking, the graph contains three regions. In the leftmost region, corresponding to small values of h_0 (approximately $h_0 < e^{-36} \approx 2^{-52}$), we see that

$$fl(1 + h_0) = 1$$

due to floating point rounding errors. This means that we compute $e^{1+h_0} - e^1 = e^1 - e^1 = 0$, meaning that the error will be equal to the constant e. Once we take a (natural) log, this becomes 1.

In the middle region, we see that the graph looks very jumpy. This is due to the fact that the error in this region is dominated by floating point issues arising from problem 2.2

and further that the error increases with decreasing h_0. Since the roundoff error $x - fl(x)$ is not continuous, this results in the erratic behavior of the graph.

Finally, in the rightmost region, the graph looks fairly smooth and linear. This is due to the fact that the error in this region is dominated by the Taylor approximation error which grows linearly with h_0.

Solution 2.5. Problem 2.5 is essentially the same as problems 2.1-2.3. For notational convenience, we'll set $a = f(x + h_0)$ and $b = f(x - h_0)$. Then

$$\begin{aligned} \left| f'(x) - \frac{fl(a) - fl(b)}{2h_0} \right| &= \left| f'(x) - \frac{f(x+h_0) - f(x-h_0)}{2h_0} + \frac{a-b}{2h_0} - \frac{fl(a) - fl(b)}{2h_0} \right| \\ &\leq \left| f'(x) - \frac{f(x+h_0) - f(x-h_0)}{2h_0} \right| + \left| \frac{a-b}{2h_0} - \frac{fl(a) - fl(b)}{2h_0} \right| \end{aligned}$$

We handle the left error term similarly to before. However, we can get better performance by using a second order Taylor approximation with third order error isntead.

$$\begin{aligned} f(x \pm h_0) &= f(x) + f'(x)(x \pm h_0 - x) + f''(x)\frac{(x \pm h_0 - x)^2}{2} + f'''(\xi(x \pm h))\frac{(x \pm h_0 - x)^3}{6} \\ &= f(x) \pm f'(x)h_0 + f''(x)\frac{h_0^2}{2} \pm f'''(\xi(x \pm h))\frac{h_0^3}{6}. \end{aligned}$$

Plugging this into the above and simplifying gives

$$\begin{aligned} \left| f'(x) - \frac{f(x+h_0) - f(x-h_0)}{2h_0} \right| &= \left| f'(x) - \frac{f(x) + f'(x)h_0 + f''(x)\frac{h_0^2}{2} + f'''(\xi(x+h))\frac{h_0^3}{6}}{2h_0} \right. \\ &\qquad \left. + \frac{f(x) - f'(x)h_0 + f''(x)\frac{h_0^2}{2} - f'''(\xi(x-h))\frac{h_0^3}{6}}{2h_0} \right| \\ &= \left| -f'''(\xi(x+h))\frac{h_0^2}{12} - f'''(\xi(x-h))\frac{h_0^2}{12} \right| \\ &\leq 2 \max_x |f'''(x)| \frac{h_0^2}{12}. \end{aligned}$$

For the floating point portion of the error, we proceed similarly as before

$$\left| \frac{a-b}{2h_0} - \frac{fl(a) - fl(b)}{2h_0} \right| \leq \left| \frac{a - fl(a)}{2h_0} \right| + \left| \frac{b - fl(b)}{2h_0} \right| \leq \frac{|a| + |b|}{2} \frac{\varepsilon}{h_0}$$

The sum of the two error terms we computed is

$$\max_x |f'''(x)| \frac{h_0^2}{6} + \frac{|a| + |b|}{2} \frac{\varepsilon}{h_0}.$$

Again throwing away constants, we obtain $h_0^2 + \frac{\varepsilon}{h_0}$ which is minimized at $h_0 = \left(\frac{\varepsilon}{2}\right)^{\frac{1}{3}}$ yielding an error

$$\left(\frac{\varepsilon}{2}\right)^{\frac{2}{3}} +^3 \sqrt{2}\varepsilon^{\frac{2}{3}} \approx \varepsilon^{\frac{2}{3}}.$$

This error is a bit better than before at little extra cost.

3 An Infinite Sum

Solution 3.1. The rough estimates we obtain give an error that behaves like $\frac{1}{n} + n\varepsilon$. Optimizing this expression for n yields $n = \frac{1}{\sqrt{\varepsilon}}$, meaning that our error is approximately $2\sqrt{\varepsilon}$.

Solution 3.2. The right-to-left sum will achieve the more accurate floating point representation for this problem. The reason is that the right-to-left sum adds smaller numbers first, where we have more accuracy in the floating point representation before going on to add the larger numbers. This gives a chance for the smaller numbers to "accumulate" before they are insignificant due to roundoff error.

While it's not the optimal way to sum numbers in general (actually, identifying the optimal summation pattern to reduce floating point roundoff errors is tremendously complicated), summing numbers from smaller to larger generally works well in practice.

Solution 3.3. We'll demonstrate the left-to-right sum. The other summation is similar and left as an exercise for the curious reader.

Let $A_n = fl(A_{n-1} + fl(a_n))$ be our approximation to the partial sum S_n, where $a_n = \frac{1}{n^2}$ denotes the n^{th} term of the partial sum. Then we want to understand the errors $E_n = A_n - S_n$. Starting with $A_{n+1} - S_{n+1}$, we obtain

$$\begin{aligned}
|E_{n+1}| &= |A_{n+1} - S_{n+1}| = |fl(A_n + fl(a_{n+1})) - S_{n+1}| \\
&= |fl(A_n + fl(a_{n+1})) - S_n - a_{n+1} + A_n - A_n + fl(a_{n+1}) - fl(a_{n+1})| \\
&= |fl(A_n + fl(a_{n+1})) - (A_n + fl(a_{n+1})) + (A_n - S_n) + (fl(a_{n+1}) - a_{n+1})| \\
&\leq |fl(A_n + fl(a_{n+1}))|\varepsilon + |E_n| + |a_{n+1}|\varepsilon
\end{aligned}$$

Note that while we typically use the inequality $|x - fl(x)| \leq |x|\varepsilon$, we can use the same reasoning to obtain $|x - fl(x)| \leq |fl(x)|\varepsilon$ as well.

$$= |E_n| + (A_{n+1} + a_{n+1})\varepsilon.$$

Observe first that when $n = 1$, we have $A_1 = a_1 = 1 = S_1$. This means $E_1 = 0$, so we can stop at $i = 2$. Adding together these inequalities from $i = 2$ to $i = n$ gives

$$|E_n| \leq \varepsilon \sum_{i=2}^{n} A_i + a_i \leq \varepsilon(A_n + \sum_{i=2}^{n} A_i) \leq \varepsilon(2 + 2(n-1)) = 2n\varepsilon.$$

(Technically, we should be careful here to argue that $A_i \leq 2$, but I'll leave this detail for the curious reader.)