0 Instructions

This problem set is **due on Tuesday, September 10 by midnight**. Submit your solutions on Canvas. Include the names of everyone you worked with on this problem set. Include any code you used to solve the problems as part of your submission.

1 Practice

Problem 1.1. State Taylor's theorem as well as one reason for its usefulness in your own words.

Problem 1.2. Give the third order Taylor approximation with fourth order error term for $f(x) = \cos(x)$ centered at $x_0 = 0$. Use your answer to approximate $\cos(0.01)$ and give an upper bound for the error of this approximation.

Problem 1.3. Explain in your own words the main limitation of floating point arithmetic as well as the importance of the quantity $\varepsilon = 2^{-52}$.

Problem 1.4. Compute $fl(8 - 5\varepsilon)$.

Problem 1.5. A useful error bound for floating point approximation is $|fl(x) - x| \leq |x|\varepsilon$ if there is no overflow. You don't need to give a full proof, but give a rough justification for this fact. Sometimes, we may write this in relative form as

$$\frac{|fl(x) - x|}{|x|} \leq \varepsilon$$

or equivalently in the form

$$fl(x) = x(1 + \delta) \text{ for some } |\delta| \leq \varepsilon.$$

2 Numerical Differentiation

Recall that the derivative $f'(x)$ of a differentiable function $f(x)$ is defined by

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}.$$

In calculus classes, we never consider the numerical aspects of "letting h go to 0." One way to approximate the limit above is to choose a small value h_0 for h:

$$f'(x) \approx \frac{f(x + h_0) - f(x)}{h_0}. \tag{1}$$

Problem 2.1. Use Taylor's theorem to bound the error of the approximation (1) in terms of h_0. In other words, use Taylor's theorem to find an upper bound for the quantity

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

in terms of h_0 (and possibly f and its derivatives).

In addition to the approximation error, we also incur error from floating point approximations used to represent these numbers in a computer. While we have freedom to choose h_0, the function $f(x)$ will generally be beyond our control. Therefore, the best we can do numerically is (assuming we choose h_0 to be a floating point number)

$$f'(x) \approx \frac{fl(f(x+h_0)) - fl(f(x))}{h_0}. \tag{2}$$

Problem 2.2. Bound the error in approximation (2) in terms of $\varepsilon = 2^{-52}$ and h_0. In other words, find an upper bound for the quantity

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

in terms of ε and h_0 (and possibly f and its derivatives).

Problem 2.3. Based on your error bounds for the previous two problems, identify an optimal choice of h_0 in terms of ε to minimize the overall error. (Feel free to drop constants and only think about h_0 and ε.) What is the approximate error for this choice of h_0?

Problem 2.4. Let $f(x) = e^x$. The exact value of $f'(1)$ is e. Using the approximation (1), plot the error

$$e - \frac{f(x+h_0) - f(x)}{h_0}$$

as a function of h_0. What do you notice? (It may be helpful to plot the log of the error as a function of $\log(h_0)$.) Explain the features of your graph and whether or not it agrees with your work above. You should notice in your log-scaled graph that the log error eventually becomes constant at $y = 1$. Explain why this happens.

This problem just showed that using the "forward difference" approximation of a derivative has a fundamental limit in terms of the accuracy that can be achieved. This is rather unsatisfying if we need to work with derivatives at higher precisions.

Problem 2.5. We can get an improvement in accuracy by using the "centered difference" approximation to the derivative

$$f'(x) \approx \frac{f(x+h_0) - f(x-h_0)}{2h_0}.$$

Using similar ideas as before, find error bounds coming from the approximation of the derivative and the floating point approximations. What is the optimal choice of h_0 here? (As before, feel free to drop constants.)

3 An Infinite Sum

Infinite sums can be difficult to evaluate, and one strategy is to evaluate partial sums to see whether they converge. Consider the infinite sum

$$S = \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

and let S_n denote the partial sums

$$S_n = \sum_{k=1}^{n} \frac{1}{k^2}.$$

As in the first problem, there is a tension between two components of error when approximating this infinite sum:

1. If we don't sum enough terms, we will not get an accurate estimation. This error term is approximately $\frac{1}{n}$.
2. If we sum too many terms, we will accumulate floating point errors. This error term is approximately εn.

Problem 3.1. Use the rough estimates of the errors above to determine the optimal value of n in terms of ε. For this n, give an estimate for the error $|S_n - S|$.

This infinite sum gives one way to approximate π numerically as $\pi \approx \sqrt{6S_n}$, although we will see a much better approach on the next homework. There is another subtlety to consider in this problem: the sum S_n can be computed in many ways. Consider the "left-to-right" sum

$$((\dots((1 + \frac{1}{4}) + \frac{1}{9}) + \dots) + \frac{1}{n^2})$$

and the "right-to-left" sum

$$(((\dots(\frac{1}{n^2} + \frac{1}{(n-1)^2}) + \frac{1}{(n-2)^2}) + \dots) + 1).$$

Mathematically, these expressions are equivalent, but as floating point expressions they are very different.

Problem 3.2. Which of the two methods above (left-to-right vs right-to-left) do you think will produce the most accurate floating point representation of the partial sum S_n? You do not need to prove this exactly, but you should explain your reasoning.

Problem 3.3 (extra credit)**.** Show that the error of the left-to-right partial sum is bounded by $2n\varepsilon$ while the error of the right-to-left partial sum is bounded by $(3 + \ln(n))\varepsilon$.