M 441: Homework 1

Nathan Stouffer

Problem 1

Problem. The real number $x = 11/8$ has decimal representation $x_d = 1.375$ and binary represenatation $x_b = 1.011$. Compute each of the following relative errors in decimal.

$$E_d = \frac{|x_d - chop(x_d, 3)|}{x_d} \qquad E_b = \frac{|x_b - chop(x_b, 3)|}{x_b}$$

Answer. Let's begin with $E_d = \frac{|1.375 - chop(1.375, 3)|}{1.375} = \frac{|1.375 - 1.37|}{1.375} = \frac{0.005}{1.375} = \frac{1}{275} = 0.0036.$

Seperately, we can compue E_b. In binary, the numerator is $|1.011 - chop(1.011, 3)| = |1.011 - 1.01| = 0.001$. Translating to decimal, we get $1/8$ for the numerator. So E_b is

$$E_b = \frac{1/8}{11/8} = \frac{1}{8} * \frac{8}{11} = \frac{1}{11}$$

Problem 2

Problem. Suppose one can compute $\sqrt{x}$ exactly but an error of $\delta > 0$ is incurred by some finite representation $\hat{x}$ of x.

a) For $\delta > 0$ find a uniform upper bound on the absolute error $E_a = |\sqrt{x} - \sqrt{\hat{x}}|$ valid for all $x \in [0, 1]$.

b) If $\delta = 10^{-6}$ what does a) imply the upper bound on E_a is on $[0, 1]$?

Answer.

a) We begin by finding a uniform upper bound for the absolute error $E_a = |\sqrt{x} - \sqrt{\hat{x}}|$:

$$E_a = |\sqrt{x} - \sqrt{\hat{x}}| = |\sqrt{x} - \sqrt{\hat{x}}| * \frac{\sqrt{x} + \sqrt{\hat{x}}}{\sqrt{x} + \sqrt{\hat{x}}} = \frac{|(\sqrt{x} - \sqrt{\hat{x}}) * (\sqrt{x} + \sqrt{\hat{x}})|}{\sqrt{x} + \sqrt{\hat{x}}} = \frac{|x - \hat{x}|}{\sqrt{x} + \sqrt{\hat{x}}}$$

Now we substitute $\hat{x} = x + \delta$:

$$E_a = \frac{|x - x - \delta|}{\sqrt{x} + \sqrt{x + \delta}} = \frac{\delta}{\sqrt{x} + \sqrt{x + \delta}}$$

Since $x \in [0, 1]$,

$$E_a \leq max\{E_a\} = E(0, \delta) = E(\delta) = \frac{\delta}{\sqrt{0} + \sqrt{0 + \delta}} = \frac{\delta}{\sqrt{\delta}} = \sqrt{\delta}$$

Therefore, $E_a \leq E(\delta) = \sqrt{\delta}\ \ \forall x \in [0, 1]$.

b) If $\delta = 10^{-6}$, then part a) implies that the upper bound for E_a on $[0, 1]$ is

$$\sqrt{\delta} = \sqrt{10^{-6}} = (10^{-6})^{1/2} = 10^{-6/2} = 10^{-3}$$

Problem 3

Problem. The Taylor series for $f(x) = ln(1+x)$ is

$$ln(1+x) = \sum_{k=1}^{n}(-1)^{k-1}\frac{x^k}{k} + E_n(\zeta, x) = P_n(x) + E_n(\zeta, x)$$

and converges for $x \in (-1, 1]$.

a) Use the Alternating Series Test to bound the error $|E_n|$ by $\hat{E}_n$. Use $\hat{E}_n$ to find an n sufficiently large so that

$$|ln(2) - P_n(1)| \leq \hat{E}_n \leq 10^{-6}$$

Here $x = 1$.

b) One can accelerate the series convergence rate using the following identity

$$ln(2) = ln(e * 2/e) = 1 + ln(2/e) = 1 + ln(1 + (2/e - 1)) = 1 + ln(1+x)$$

Answer.

a) We begin by finding a bound $\hat{E}_n$ for $|E_n|$. By the Alternating Series Theorem,

$$|E_n| = |f(x) - P_n(x)| \leq |a_{n+1}|$$

So we can compute the upper bound as

$$|E_n| = |(-1)^{n+1-1}\frac{x^{n+1}}{n+1}| = \frac{|x^{n+1}|}{n+1} \leq \hat{E}_n = \frac{1}{n+1}$$

For $|ln(2) - P_n(1)| \leq \hat{E}_n \leq 10^{-6}$, we must have $n \geq 999999$ so that

$$\hat{E}_n \leq 1/(999999+1) = 1/1000000 = 10^{-6}$$

b) Using the identity $ln(2) = 1 + ln(1+x)$ where $x = 2/e - 1$, we can accelerate the convergence. Consider the following table generated using MATLAB and this new formula.

n	$1 + P_n(x)$	$E_a = \|1 + P_n(x) - ln(2)\|$
1	0.735758882342885	0.042611701782939
2	0.700847198212544	0.007700017652599
3	0.694697129923282	0.001549949363336
4	0.693478304234465	0.000331123674520
5	0.693220653144671	0.000073472584726
6	0.693163918134727	0.000016737574782
7	0.693151068086924	0.000003887526978
8	0.693148097014804	0.000000916454859
9	0.693147399166433	0.000000218606488
10	0.693147233206223	0.000000052646278