PROBLEM SET #1

APC 523/MAE 507/AST 523 : Numerical Algorithms for Scientific Computing

Vivek Kumar

March 13, 2019

1 Error in (symmetric) rounding vs chopping

Assertion: When mapping a real number x to a nearby machine number in $\mathbb{R}(p,q)$, the upper bound in the relative error for symmetric rounding is:

$$\left|\frac{x - \mathrm{rd}(x)}{x}\right| \leq 2^{-p}$$

Proof:
Consider the number x to be represented as:

$$x = \pm\left(\sum_{l=1}^{\infty} b_{-l}2^{-l}\right)2^{e}$$

If the number is to be rounded to p terms, two cases arise:

CASE I. The $(p+1)^{\mathrm{th}}$ is 0.
In this scenario the difference between the true value and the rounded value is given by:

$$x - \mathrm{rd}(x) = \pm\left(\sum_{l=p+2}^{\infty} b_{-l}2^{-l}\right)2^{e}$$

The maximum relative error can then be computed as:

$$\begin{aligned}\max\left|\frac{x - \mathrm{rd}(x)}{x}\right| &= \frac{\max|x - \mathrm{rd}(x)|}{\min|x|}\\ &= \frac{2^{-p-1}2^{e}}{2^{-1}2^{e}}\\ &= 2^{-p}\end{aligned}$$

which is what we set to prove.

CASE II. The $(p+1)^{\mathrm{th}}$ is 1.
In this scenario the maximum difference between the true and the rounded value is obtained as:

$$\max|x - \mathrm{rd}(x)| = \left(2^{-p} - 2^{-p-1}\right)2^{e}$$

This is the case we all the leading terms from $(p+2)$ are 1. Hence the maximum relative error can be computed as before:

$$\begin{aligned}\max\left|\frac{x - \mathrm{rd}(x)}{x}\right| &= \frac{\max|x - \mathrm{rd}(x)|}{\min|x|}\\ &= \frac{\left(2^{-p} - 2^{-p-1}\right)2^{e}}{2^{-1}2^{e}}\\ &= 2^{-p}\end{aligned}$$

which is what we set to prove

Both the cases show that the maximum symmetric rounding off error is 2^{-p}