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 $(p+1)^{\text{th}}$ 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 $(p+1)^{\text{th}}$ 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}

2 An accurate implementation of e^x

(a) Compute the value of $e^{5.5}$ by working out the terms of the infinite series upto $n = 30$ by rounding upto 5-significant figures
The true value of `exp(5.5)` is 244.691932. Here the data generated is presented here

n	numerator	denominator	n^{th} term	value
0	1.0000	1.0000	1.0000	1.0000
1	5.50000E+00	1.00000E+00	5.50000E+00	6.50000E+00
2	3.02500E+01	2.00000E+00	1.51250E+01	2.16250E+01
3	1.66380E+02	6.00000E+00	2.77300E+01	4.93550E+01
4	9.15090E+02	2.40000E+01	3.81290E+01	8.74840E+01
5	5.03300E+03	1.20000E+02	4.19420E+01	1.29430E+02
6	2.76820E+04	7.20000E+02	3.84470E+01	1.67880E+02
7	1.52250E+05	5.04000E+03	3.02080E+01	1.98090E+02
8	8.37380E+05	4.03200E+04	2.07680E+01	2.18860E+02
9	4.60560E+06	3.62880E+05	1.26920E+01	2.31550E+02
10	2.53310E+07	3.62880E+06	6.98050E+00	2.38530E+02
11	1.39320E+08	3.99170E+07	3.49020E+00	2.42020E+02
12	7.66260E+08	4.79000E+08	1.59970E+00	2.43620E+02
13	4.21440E+09	6.22700E+09	6.76790E-01	2.44300E+02
14	2.31790E+10	8.71780E+10	2.65880E-01	2.44570E+02
15	1.27480E+11	1.30770E+12	9.74840E-02	2.44670E+02
16	7.01140E+11	2.09230E+13	3.35100E-02	2.44700E+02
17	3.85630E+12	3.55690E+14	1.08420E-02	2.44710E+02
18	2.12100E+13	6.40240E+15	3.31280E-03	2.44710E+02
19	1.16660E+14	1.21650E+17	9.58980E-04	2.44710E+02
20	6.41630E+14	2.43300E+18	2.63720E-04	2.44710E+02
21	3.52900E+15	5.10930E+19	6.90700E-05	2.44710E+02
22	1.94100E+16	1.12400E+21	1.72690E-05	2.44710E+02
23	1.06760E+17	2.58520E+22	4.12970E-06	2.44710E+02
24	5.87180E+17	6.20450E+23	9.46380E-07	2.44710E+02
25	3.22950E+18	1.55110E+25	2.08210E-07	2.44710E+02
26	1.77620E+19	4.03290E+26	4.40430E-08	2.44710E+02
27	9.76910E+19	1.08890E+28	8.97150E-09	2.44710E+02
28	5.37300E+20	3.04890E+29	1.76230E-09	2.44710E+02
29	2.95510E+21	8.84180E+30	3.34220E-10	2.44710E+02
30	1.62530E+22	2.65250E+32	6.12740E-11	2.44710E+02

(b) Compute the $e^{5.5}$ using partial sums

- The value of $e^{5.5}$ converges to 5-significant digits at $k = 18$

(c)

(d)

3 Recurrence in reverse

(a) The reverse recurrence relation is given by:

$$y_{n-1} = \frac{e - y_n}{n}$$

Computing for a few terms down the chain we obtain:

$$\begin{aligned} y_{n-2} &= \frac{e - y_{n-1}}{n-1} \\ &= \frac{ne - e + y_n}{n(n-1)} \\ y_{n-3} &= \frac{e - y_{n-2}}{n-2} \\ &= \frac{n(n-1)e - ne + e - y_n}{n(n-1)(n-2)} \end{aligned}$$

One can denote the pattern as:

$$\begin{aligned} y_{n-p} = (-1)^p \frac{y_n}{n(n-1)(n-2)\dots(n-p+1)} &+ e\left[\frac{1}{n-(p-1)} - \frac{1}{(n-(p-1))(n-(p-2))}\right. \\ &\left. + \frac{1}{(n-(p-1))(n-(p-2))(n-(p-3))} + \dots\right] \end{aligned}$$

To obtain the value of y_k in terms of y_N we replace $n-p$ with k and simplify:

$$\begin{aligned} y_k &= (-1)^{n-k} \frac{y_n}{n(n-1)(n-2)\dots(k+1)} + \text{exponent terms} \\ &= (-1)^{n-k} \frac{y_n k!}{n!} + \text{exponent terms} \end{aligned}$$

The condition number is given as:

$$\begin{aligned} (\text{cond } g_k)(y_k) &= \left|\frac{y_N g'(y_N)}{y_k}\right| \\ &= \left|\frac{y_N \frac{k!}{N!}}{y_k}\right| \end{aligned}$$

Since the k is less than N, y_k is greater than y_N, the upper bound on the condition number, $(\text{cond } g_k)(y_k)$, obtained as:

$$(\text{cond } g_k)(y_k) \leq \frac{k!}{N!}$$

as $\frac{y_N}{y_k} \leq 1$

(b) We know the condition number represents:

$$\begin{aligned} \varepsilon_y &= (\text{cond } g_k)\, \varepsilon_x \\ \frac{\Delta y_k}{y_k} &\leq \frac{k!}{N!} \quad \leq \varepsilon \\ N! &\geq \frac{k!}{\varepsilon} \end{aligned}$$

Here, we have assumed $\varepsilon_x = 1$ and ε is a predefined target error in y_k.

(c) For `python3` the machine epsilon for float is $1.0e^{-15}$(Obtained using numpy.finfo(float)). Using this machine epsilon the value of N obtained is 31. [Check code]

(d) The computed value of y_{20} is 0.123803830762570 and the value of y_{20} directly by integration is 0.123803830762570. [Check code]