EM 314 - Assignment 1: Solutions

Lecturer: Dr. Janitha Gunatilake

November 19, 2018

Theory

1. For small values of x, the approximation $\sin x \approx x$ is often used. Estimate the error in using this formula with the aid of Taylor's Theorem. For what range of values of x will this approximation give correct results rounded to six decimal places?

Solution: From Taylor's Theorem,

$$\sin x = x + E_3(x) \text{ where } E_3(x) = \frac{x^3}{3!}(-\cos \xi), \quad 0 < |\xi| < |x|.$$

Now, $|\cos(\xi)| \leq 1 \Rightarrow |E_3(x)| \leq \frac{|x|^3}{6}$.
If this approximation is correct rounded to six decimal places,

$$E_3(x) \leq \frac{|x|^3}{6} \leq \frac{10^{-6}}{2} \Rightarrow -0.0144225 \leq x \leq 0.0144225$$

2. Consider the floating point number set $\mathbb{F} \subset \mathbb{R}$ such that $\mathbb{F}(\beta, t, L, U)$. Here β is the base, t is the number of digits in the mantissa and $[L, U]$ is the range of variation of the exponent. Show that the set $\mathbb{F}$ contains precisely $2(\beta - 1)\beta^{t-1}(U - L + 1)$ elements.
Hint: Recall that a floating point number has the form $\pm(.\alpha_1 \dots \alpha_t) \times \beta^E$E.

Solution: Any $x_* \in \mathbb{F}$ can be represented as $x_* = (-1)^s(.\alpha_1 \dots \alpha_t)\beta^E$.

The sign bit can assume 2 values.

Each digit $\alpha_2, \alpha_3, \dots, \alpha_t$ can assume β different values, while α_1 can assume only $\beta - 1$ values. Therefore, the mantissa assumes $(\beta - 1)\beta^{t-1}$ different values.

The exponent can assume $U - L + 1$ different values.

Thus, the set $\mathbb{F}$ contains $2(\beta - 1)\beta^{t-1}(U - L + 1)$ elements.

3. Consider the following approximation f'_h for the derivative f' of a function $f(x)$.

$$f'_h(x) = \frac{1}{h}[f(x+h) - f(x)]$$

Let $E_h(x) = |f'(x) - f'_h(x)|$ be the associated error. Show that $E_h(x) = \mathcal{O}(h)$.

Solution: From Taylor series $f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi)$ where $x < \xi < x+h$. Thus,

$$\begin{aligned} f'(x) &= \frac{f(x+h) - f(x)}{h} - h\frac{f''(\xi)}{2} \\ &= f'_h(x) - h\frac{f''(\xi)}{2}. \end{aligned}$$

Hence,

$$E_h(x) = |f'(x) - f'_h(x)| = \left| h\frac{f''(\xi)}{2} \right|.$$

Now,

$$\lim_{h\to 0} \frac{|E_h(x)|}{|h|} = \frac{|f''(\xi)|}{2} \leq \frac{|f''(x)|_{\max}}{2} \text{ and } E_h = \mathcal{O}(h).$$

Computer Experiments

4. This is a practical experiment on your results in Question 3 above. Use a single Octave/MATLAB script `q5.m` to perform the following tasks.
   (a) Generate pairs of random square matrices A, B, of size n, $n = 500, 1000, 1500, \ldots, 5000$. *Hint*: `rand(n)`.
   (b) For each n, compute AB and measure the CPU time t needed. *Hint*: `cputime()` or `tic()` / `toc()`.
   (c) Plot the points $(\log n, \log t)$ on a graph. *Hint*: `loglog()`.
   (d) Find the best-fit line for the points in part (c). Plot the graph of this line on the same figure. *Hint*: `polyfit()`.
   (e) Let $t = Cn^\alpha$, C is a constant. Estimate α from your results in part (d).
   (f) Compare your experimental results for the cost of matrix multiplication, with your theoretical results in Question 3.

Solution: Discussed in the lab class.

5. Here, we experiment on the approximation f_h' in Question 3, with

$$f(x) = \ln x \text{ and } x = 3.$$

Use a single GNU Octave / MATLAB script `q5.m` to perform the tasks in this question.

(a) Let $N = 10$. Generate a sequence $\{h_k\}$, $k = 1, 2, \ldots N$, such that $h_k = \frac{1}{2^k}$. For each k, compute $f_{h_k}'(x)$ and E_{h_k}.

Solution:

```
format long
H = 1; n = 10;
h = zeros(1,n); e = h;
printf("|  k  |   h  \t |  f'  \t |  E    \n")

for k = 1:n
 H = H/2;
 df = (log(3 + H)-log(3))/H;
 e(1,k) = abs(df-1/3);
 h(1,k) = H;
 printf("|  %d  | %d \t | %d \t | %d \n", k, H, df, e(1,k));
endfor
```

k	h_k	$f_{h_k}'(x)$	E_{h_k}
1	1/2	0.308301	2.5032×10^{-2}
2	1/4	0.320171	1.31625×10^{-2}
3	1/8	0.326576	6.75738×10^{-3}
4	1/16	0.329909	3.42474×10^{-3}
5	1/32	0.331609	1.72415×10^{-3}
6	1/64	0.332468	8.65053×10^{-4}
7	1/128	0.332900	4.33276×10^{-4}
8	1/256	0.333117	2.16826×10^{-4}
9	1/512	0.333225	1.0846×10^{-4}
10	1/1024	0.333279	5.42417×10^{-5}

(b) Observe that $E_{h_k} \to 0$. Assuming $E_h \propto h^\gamma$, use a log-log plot (as in Question 4) to find γ. Do you obtain $E_h = \mathcal{O}(h)$ as expected?

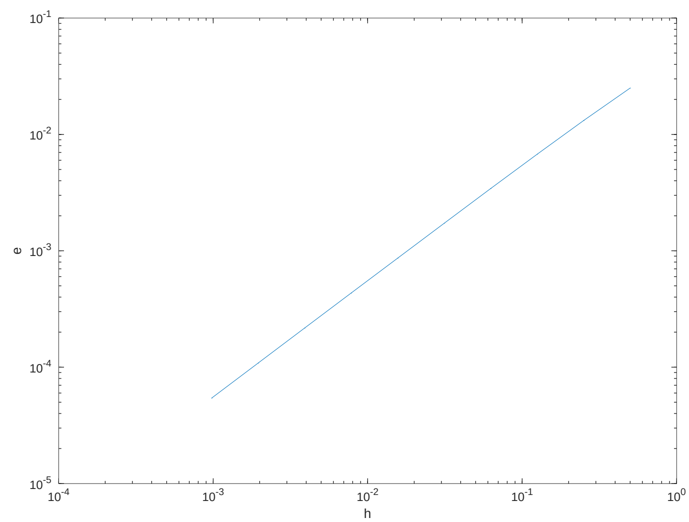
Solution: The best fit line (least-square line) is found using `polyfit()`. I added the following segment.

```
figure(2)
clf
Order = polyfit (log(h), log(e), 1)
plot (log(h), Order(1)*log(h)+Order(2), log(h),log(e),'o' )
xlabel('ln(h)'); ylabel('ln(e)');
```

From which I obtained

```
Order =
   0.987063614379873  -2.960940096915915
```

leading to $\gamma \approx 0.99$. Notice that we didn't get $\gamma = 1$ exactly. However, this practical value is a good estimate considering the round-off errors and recalling that $E_h \propto h^\gamma$ is only an assumption. Theoretically, we have $E_h = \mathcal{O}(h)$, and it describes the limiting behaviour of E_h as $h \to 0$. Thus, we conclude that the experimental results are consistent with theory. Following is the log-log plot of E_h vs h.



(c) Let $N = 40$. Repeat part (a).

(d) Observe that $E_{h_k} \nrightarrow 0$ if $N = 40$. Explain why.

Solution: From the table, we observe that E_{h_k} is decreasing until $k = 24$, remain a constant for $k = 25, 26$, and is again increasing starting $k = 27$. This deviation from the theoretical result is a consequence of working on floating point arithmetic.

k	h_k	$f'_{h_k}(x)$	E_{h_k}
1	0.5	0.308301	0.025032
2	0.25	0.320171	0.0131625
3	0.125	0.326576	0.00675738
4	0.0625	0.329909	0.00342474
5	0.03125	0.331609	0.00172415
6	0.015625	0.332468	0.000865053
7	0.0078125	0.332900	0.000433276
8	0.00390625	0.333117	0.000216826
9	0.00195312	0.333225	0.00010846
10	0.000976562	0.333279	5.42417e-05
11	0.000488281	0.333306	2.71238e-05
12	0.000244141	0.333320	1.35626e-05
13	0.00012207	0.333327	6.78150e-06
14	6.10352e-05	0.333330	3.39080e-06
15	3.05176e-05	0.333332	1.69541e-06
16	1.52588e-05	0.333332	8.47712e-07
17	7.62939e-06	0.333333	4.23858e-07
18	3.81470e-06	0.333333	2.11953e-07
19	1.90735e-06	0.333333	1.06016e-07
20	9.53674e-07	0.333333	5.31630e-08
21	4.76837e-07	0.333333	2.68531e-08
22	2.38419e-07	0.333333	1.33490e-08
23	1.19209e-07	0.333333	6.82970e-09
24	5.96046e-08	0.333333	4.96705e-09
25	2.98023e-08	0.333333	4.96705e-09
26	1.49012e-08	0.333333	4.96705e-09
27	7.45058e-09	0.333333	1.98682e-08
28	3.72529e-09	0.333333	1.98682e-08
29	1.86265e-09	0.333333	7.94729e-08
30	9.31323e-10	0.333333	7.94729e-08
31	4.65661e-10	0.333333	3.17891e-07
32	2.32831e-10	0.333333	3.17891e-07
33	1.16415e-10	0.333332	1.27157e-06
34	5.82077e-11	0.333332	1.27157e-06
35	2.91038e-11	0.333328	5.08626e-06
36	1.45519e-11	0.333328	5.08626e-06
37	7.27596e-12	0.333313	2.03451e-05
38	3.63798e-12	0.333313	2.03451e-05
39	1.81899e-12	0.333252	8.13802e-05
40	9.09495e-13	0.333252	8.13802e-05

(e) Plot $\log(E_{h_k})$ against $\log(h_k)$. Using the graph, estimate the h_k (say h_{min}) that minimizes E_{h_k}.

Solution: From the graph, it seems $h_{\min}$ is slightly greater than 10^{-8}. Also, examining the table, it seems $h_{\min} \approx 1.5 \times 10^{-8}$.