CS6210 - Homework/Assignment-1

Arnabd Das(u1014840)

September 5, 2016

Question-1: Chapter-1: Exercise-1

Given $f(x) = e^{-2x}$: The first order derivative is approximated as:

$$f'(x_0) = \frac{f(x_0+h) - f(x_0)}{h}$$

It has the approximate discretisation error as follows, for very small values of h:

$$\left| f'(x_0) \ - \frac{f(x_0+h) - f(x_0)}{h} \right| \approx \frac{h}{2} \left| f''(x_0) \right|$$

The matlab code used to generate the plot is available as attachment Prob1.m. Plots of the actual error(err) and the expected approximate discretization error(derr) is plotted as shown in the figure below.

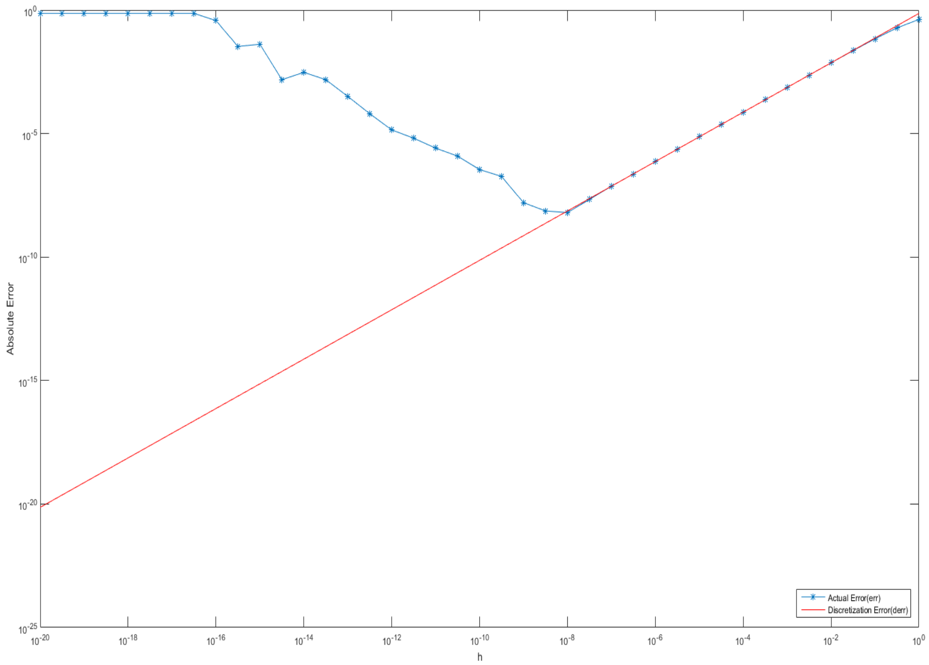

Figure 1: Error Plot for $f'(x) = e^{-2x}$ at $x = 0.5$

Similar to Example-1.3, here also the discretization error dominates the error function as seen by the overlap of the err plot and the derr plot for decreasing values of h. This dominance continues until it reaches h = 10^{-8}, from thereon round-off error starts to dominate. Thus for decreasing values of h from 1 to 10^{-8}, as expected we see the error reduces with smaller values of h linearly as dictated by the approximate discretization function. However, once round-off error takes over, the linear pattern is demolished and the error starts to increase with decreasing values of h. Also in the area of dominance of round-off error, the plot is not very regular/smooth with certain unevenness. In Example-1.3, we see that around the inversion point of the actual error function, the actual error function plot dips below the discretization error plot for a brief period before the inversion takes place. In other words, the minima of the error function happens distinctly below the derr line. However, in this problem, although the pattern is similar , there isn't any sharp dip around the inversion point, and the minima approximately coincides with the derr line without dipping further.

Question-2: Chapter-1: Exercise-4

Given $g(x) = tanh(cx) = \dfrac{exp(cx) - exp(-cx)}{exp(cx) + exp(-cx)}$; Then ,

$$g\prime(x) = \frac{4c}{(e^{cx} + e^{-cx})^2}$$

and

$$g\prime\prime(x) = \frac{-8c^2(e^{cx} - e^{cx})}{(e^{cx} + e^{cx})^3}$$

Thus,
$\forall$ c, $g\prime(x) > 0$
$\forall x, x \in R, x \geq 0, 0 \leq g(x) \leq 1$; Thus upperbound is 1 and strictly increasing in this range
$\forall x, x \in R, x \leq 0, 0 \geq g(x) \geq -1$; Thus lowerbound is -1 and strictly decreasing in this range

However, since $g\prime\prime(x) < 0$ for $c > 0$ and $x > 0$, it indicates a decelerating function, which means although g(x) is strictly increasing, its rate of increase decreases exponentially , bounding the function to $|1|$ for large x.
However, at x=0, $g\prime(0) = c$, thus for large values of c, the slope follows the value of c around x=0, indicating ill conditioning. The figure 2 below shows the plot of $g(x)$ for x$\in [-1, 1]$ for varying c $\in [0, 1000]$ in steps of 100.

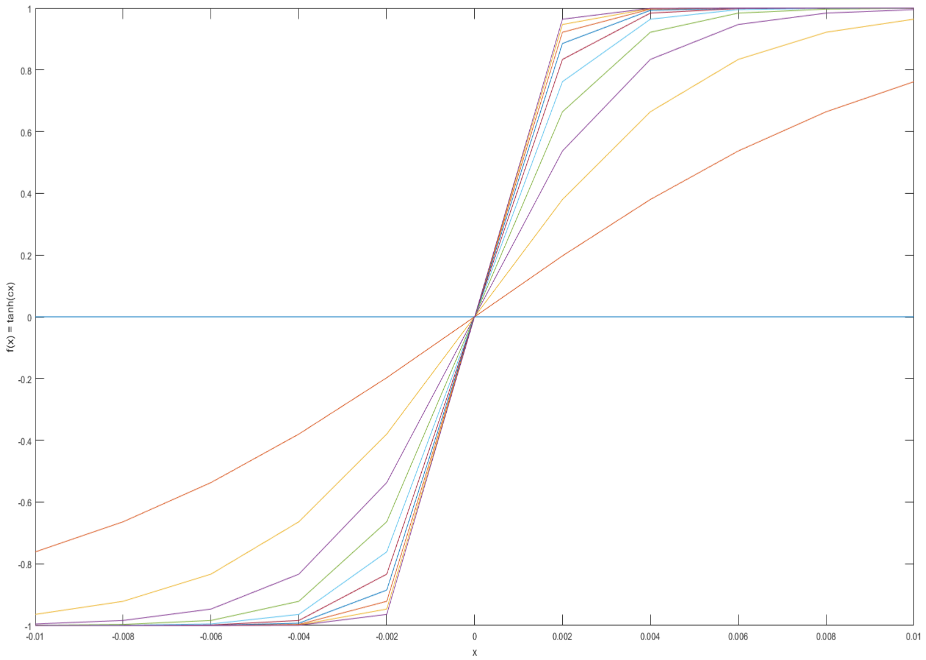

Figure 2: Plot of g(x) for x $\in [-1, 1]$ with parameter c $\in [0, 1000]$

The next figure(Figure-3) indicates the variation of $g\prime(x)$ as explained earlier

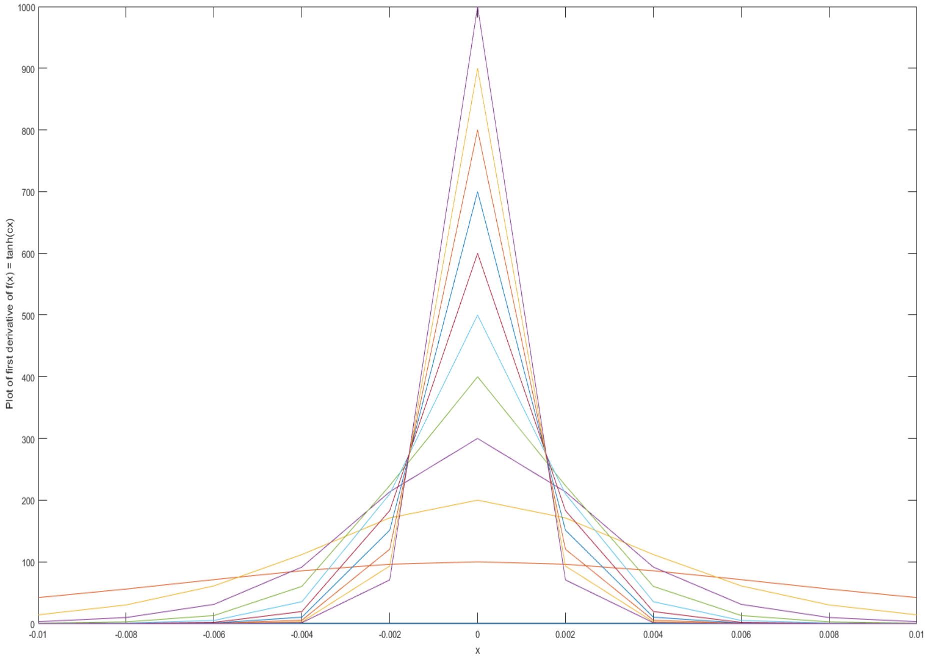

Figure 3: Plot of $g\prime(x)$ for x $\in [-1, 1]$ with parameter c $\in [0, 1000]$

Take the example of c=1000 and x changes from 0 to 0.002, then g(x) changes from 0 to 0.9640, which means a rate of change of 482, sufficiently high to be classified as ill-conditioned. However, if we set c=10, then for the same change of x, g(x) changes from 0 to 0.0200, indicating a relatively well-conditioned problem with a rate of change of 10. Thus, clearly around x=0, the conditioning of the problem is dictated by the choice of 'c' itself.
Note: However, as one moves away from c, due to our earlier argument of a strictly increasing but decelerating function with a positive slope for x $\in [0, \infty)$ and a strictly decreasing but decelerating function with a positive slope in x $\in (\infty, 0]$, the function is well-conditioned for values of x away from 0, and bounded in $[-1, 1]$.

Question-3: Chapter-2: Exercise-2

(a) A transformation, f2, of $f1(x_0, h) = sin(x_0 + h) - sin(x_0)$ can be derived using the trigonometric identity $sin(\phi) - sin(\psi) = 2cos(\frac{\phi + \psi}{2})sin(\frac{\phi - \psi}{2})$, as shown below:

$$f1(x_0, h) = sin(x_0 + h) - sin(x_0)$$

$$=> f1(x_0, h) = 2cos(\frac{x_0 + h + x_0}{2})sin(\frac{x_0 + h - x_0}{2})$$

$$=> f1(x_0, h) = 2cos(x_0 + \frac{h}{2})sin(\frac{h}{2})$$

$$=> f1(x_0, h) = f2(x_0, h)(Derived)$$

where, $f2(x_0, h) = 2cos(x_0 + \frac{h}{2})sin(\frac{h}{2})$
Thus, we have derived an expression f2, which is mathematically equivalent to f1, but do not involve the

cancellation terms.

(b) The formula $\dfrac{f(x_0+h)-f(x_0)}{h}$ introduces significant cancellation errors since $f(x_0+h) - f(x_0)$ are very close for $f(x) = sin(x)$ (well-conditioned problem). Thus instead of using the above form, we can use the form derived in (a) for f2, such that it avoids any form of cancellation errors. Hence, replacing $\dfrac{sin(x_0+h)-sin(x_0)}{h}$ by $\dfrac{2cos(x_0+\dfrac{h}{2})sin(\dfrac{h}{2})}{h}$ to approximate the derivative of $f(x) = sin(x)$ at $x = x_0$. The matlab program is implemented in Prob3.m to compute an approximation of $f\prime(1.2)$ for h $=$ $10^{-20}, 10^{-19}, \ldots 1$.

(c) The plot in Figure-4 shows the error plots using $f1 = \dfrac{sin(x_0+h)-sin(x_0)}{h}$ and $f2 = \dfrac{2cos(x_0+\dfrac{h}{2})sin(\dfrac{h}{2})}{h}$ at $x_0 = 1.2$, depicted by the legends of errf1 and errf2 respectively. Obviously f1 corresponds to the way Example-1.3 was plotted. The derr plot indicates the approximate discretization error function which is linear in h.

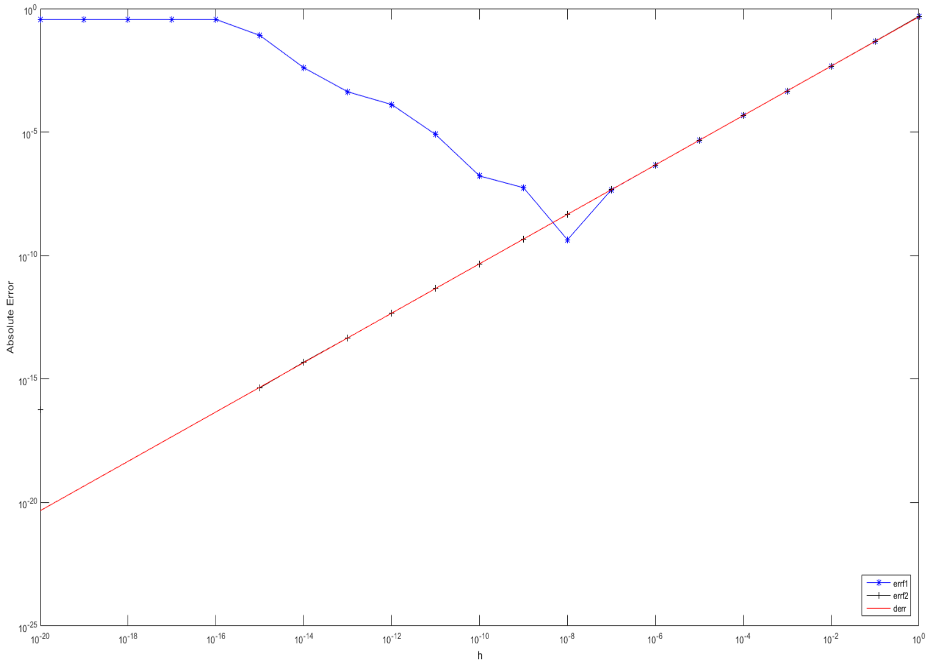

Figure 4: Error plot for f1, f2 and discretization error at $x = 1.2$

The difference in accuracy is clearly visible wherein f2 provides more accurate results than f1 and is predominantly in shape with derr for even smaller values of h, than f1. The errf2 slightly diverges from derr at around $= 10^{-15}$ while for f1 it diverged and inverted around $h = 10^{-8}$. Using f2, beyond $h = 10^{-15}$, the error nearly dies to almost 0 for even smaller values of h indicating the absence of cancellation errors while for f1 the error went up to higher values in these ranges.

Question-4: Chapter-2: Exercise-10

(a) Given : $f1(x, \delta) = cos(x + \delta) - cos(x)$

f1 can be transformed to an equivalent form, f2, using the trigonometric formula , $cos(\phi) - cos(\psi) = -2sin(\frac{\phi+\psi}{2})sin(\frac{\phi-\psi}{2})$, as shown below:

$$f1(x,\delta) = cos(x+\delta) - cos(x)$$
$$=> f1(x,\delta) = -2sin(\frac{x+\delta+x}{2})sin(\frac{x+\delta-x}{2})$$
$$=> f1(x,\delta) = -2sin(x+\frac{\delta}{2})sin(\frac{\delta}{2})$$
$$=> f1(x,\delta) = f2(x,\delta)$$

where $f2(x,\delta) = -2sin(x+\frac{\delta}{2})sin(\frac{\delta}{2})$
Thus,
$\frac{f2(x,\delta)}{\delta} = \frac{-2sin(x+\frac{\delta}{2})sin(\frac{\delta}{2})}{\delta}$
For sufficiently small δ , that is, in the limit of $\delta \to 0$, $sin(\frac{\delta}{2}) \to \frac{\delta}{2}$ and $sin(x+\frac{\delta}{2}) \to sin(x)$. Therefore,
$\frac{f2(x,\delta)}{\delta} = \frac{-2sin(x+\frac{\delta}{2})\frac{\delta}{2}}{\delta}$
$=> \frac{f2(x,\delta)}{\delta} = -sinx$ (derived)

(b) To derive f2, lets start from f1

$$f1(x,\delta) = cos(x+\delta) - cos(x)$$
$$=> f1(x,\delta) = -2sin(\frac{x+\delta+x}{2})sin(\frac{x+\delta-x}{2})$$
$$=> f1(x,\delta) = -2sin(x+\frac{\delta}{2})sin(\frac{\delta}{2})$$
$$=> f1(x,\delta) = f2(x,\delta) \qquad (derived)$$

where $f2(x,\delta) = -2sin(x+\frac{\delta}{2})sin(\frac{\delta}{2})$

(c) The matlab scipt to calculate $g1(x,\delta) = \frac{f1(x,\delta)}{\delta} + sin(x)$ and $g2(x,\delta) = \frac{f2(x,\delta)}{\delta} + sin(x)$ for x=3 and $\delta = 10^{-11}$ is available in Prob4.m. The plots are generated for x=3 and δ varying as $[10^{-20}, 10^{-19}, \ldots, 1]$. From the evaluation, we get $g1(3, 10^{-11}) = 4.406e - 07$ and $g2(3, 10^{-11}) = 4.95e - 12$.

(d) g1 is the error function that contains the cancellation error component. Thus the plot of g1 has the shape similar to Example-1.3, where there is a minima of the error function, before which the discretization error component dominates and after which the cancellation error component dominates for reducing values of δ.
g2 is the error function that is the transformed form of g1, without the cancellation error component. Thus the dominant error component of g2 remains to be primarily the discretization error which varies linearly with δ as indicated in the plot in Figure-5 by the red line.

At $x = 3, d = 10^{-11}$, g1=4.406e-07 and g2=4.95e-12, indicating that at $d = 10^{-11}$, g1 is already under the bad influence of cancellation error and has started to diverge while g2 is still under the influence of

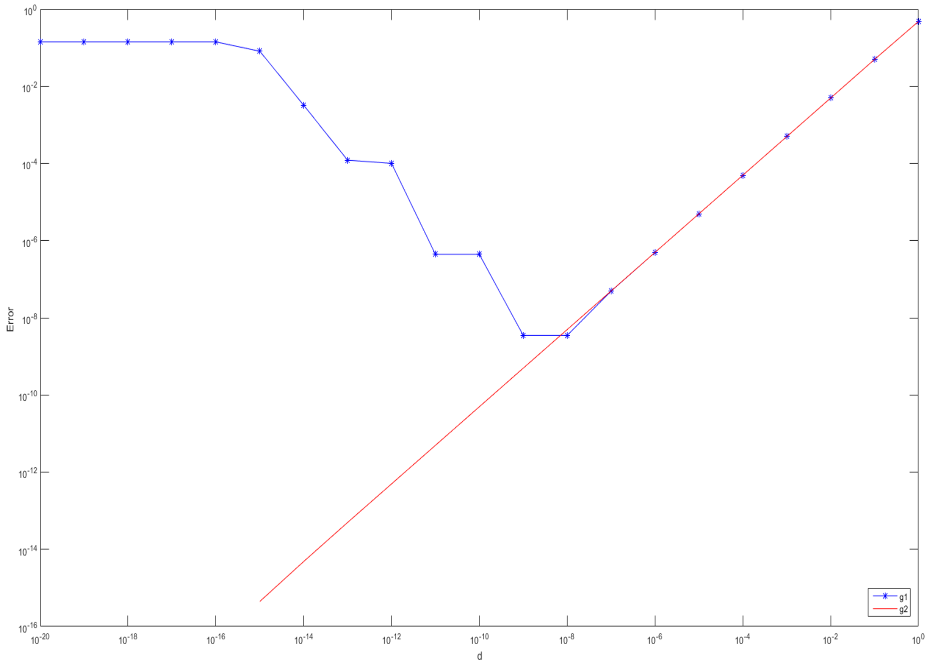

Figure 5: Error Plot for g1(blue) and g2(red)

discretization error only.

The values of g1 and g2 are shown in the table-1 for reference:

Question-5: Chapter-2: Exercise-14

(a) The approximation of the first derivative is given as

$$f\prime(x) = \frac{f(x+h) - f(x)}{h}$$

where, the truncation or discretization error is O(h). The absolute error of the function consists of two components, truncation and rounding error. By simple rearrangement of the Taylor series expansion of function f(x), we obtain,

$$\left| f\prime(x) - \frac{f(x+h) - f(x)}{h} \right| \approx \frac{h}{2} f\prime\prime(x)$$

ignoring higher order terms for very small values of h
If $|f\prime\prime(x)| \leq M$, that is, $|f\prime\prime(x)|$ is upper bounded by M, then, the discretization error component is bounded by $M\dfrac{h}{2}$. Hence,

Absolute Error(E) $= M\dfrac{h}{2}$ + Round-off-Error
The contribution to Round-off error comes from the cancellation terms in the approximation expression of

Table 1: δ vs g1 vs g2

δ	g1	g2
1e-20	0.14112	0
1e-19	0.14112	0
1e-18	0.14112	0
1e-17	0.14112	0
1e-16	0.14112	0
1e-15	0.080925	4.4409e-16
1e-14	0.003209	4.8295e-15
1e-13	0.00012168	4.9682e-14
1e-12	0.00010036	4.9502e-13
1e-11	4.406e-07	4.95e-12
1e-10	4.406e-07	4.95e-11
1e-09	3.489e-09	4.95e-10
1e-08	3.489e-09	4.95e-09
1e-07	4.9008e-08	4.95e-08
1e-06	4.9498e-07	4.95e-07
1e-05	4.95e-06	4.95e-06
0.0001	4.95e-05	4.95e-05
0.001	0.00049502	0.00049502
0.01	0.0049523	0.0049523
0.1	0.049693	0.049693
1	0.47747	0.47747

$f\prime(x)$. The bound on the cancellation term is bounded as:

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

Assuming that the absolute errors in evaluating f is bounded by ϵ , then each of the terms in the rhs numerator contributes an ϵ , resulting in round-off error being bounded by $\frac{2\epsilon}{h}$.
Thus,
Absolute Error $\le E = M\frac{h}{2} + 2\frac{\epsilon}{h}$ (derived).

(b) To find the value of h for which the above bound is minimized, we need to find the minima of the bounding function,E, of absolute error.

$$\frac{dE}{dh} = \frac{M}{2} - \frac{2\epsilon}{h^2} = 0$$

Equating first order derivative of E w.r.t h to 0 and checking for $E\prime\prime(h) \ge 0$, we get the value of h for the minima point as

$$h = 2\sqrt{\frac{\epsilon}{M}}$$

(c) Lets discuss and explain Example-1.3 using the results obtained above. The rounding unit employed is 10^{-16}, thus $\eta = 10^{-16}$. Given, f(x)=sin(x) at x = 1.2.
Now, ϵ is the bound on the absolute error for the evaluation of f, thus

$\epsilon = \left|f(x) - fl(f(x))\right|$ and,

$$\eta = \left|\frac{f(x) - fl(f(x))}{f(x)}\right|$$

Thus, $\epsilon = |\eta.f(x)|_{x=1.2} = |\eta.sin(1.2)|$
Since, M is the upperbound on $f\prime\prime(x)$, thus at x=1.2, M $\approx sin(1.2)$.
Hence,

$$\text{h} \quad = \quad 2\sqrt{\frac{\epsilon}{M}} \quad = \quad 2\sqrt{\frac{\eta.sin(1.2)}{sin(1.2)}} \quad = \quad 2\sqrt{10^{-16}} \quad = \quad \text{2e-08}$$

Thus, the minima or inversion for Example-1.3 is expected around 10^{-8}. Referring to the plot of Example-1.3, we get the minima around 10^{-8}, thus confirming our analysis.

(d) The taylor's series expansion can be written with +h and -h as follows:

$$f(x+h) = f(x) + hf\prime(x) + \frac{h^2}{2!}f\prime\prime(x) + \frac{h^3}{3!}f\prime\prime\prime(x) + \cdots$$
$$f(x-h) = f(x) - hf\prime(x) + \frac{h^2}{2!}f\prime\prime(x) - \frac{h^3}{3!}f\prime\prime\prime(x) + \cdots$$

By subtracting the above two equations we get

$$f(x+h) - f(x-h) = 2h[f\prime(x) + \frac{h^2}{3!}f\prime\prime\prime(x) + \frac{h^4}{5!}f\prime\prime\prime\prime\prime(x)] + \cdots$$

With slight rearrangement of the terms and ignoring higher order terms, we see the error terms is of $O(h^2)$, as shown below

$$\left|f\prime(x) - \frac{f(x+h) - f(x-h)}{2h}\right| \approx \frac{h^2}{6}f\prime\prime\prime(x)$$

. The error plot for this approximation is shown in Figure-6, and the matlab code is available in Prob5.m.

In the new set of results, the minima of the error function is reached much earlier, at around $h = 10^{-4}$. This is expected because since the truncation error is reducing at $O(h^2)$ instead of O(h) unlike Example-1.3, the dominance of truncation error will die off earlier, roughly around $\sqrt{(}10^{-8})$, that is 10^{-4}, exactly what the plot in Figure-6 depicts.

Question-6: Chapter-2: Exercise-19

(a) Given the following formulae for variance-standard deviations and mean

Formula(F1): $\quad s^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Formula(F2): $\quad s^2 \quad = \quad \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2$

where, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$

For Formula F1:

Number of Subtractions = n .
Number of multiplications = n .

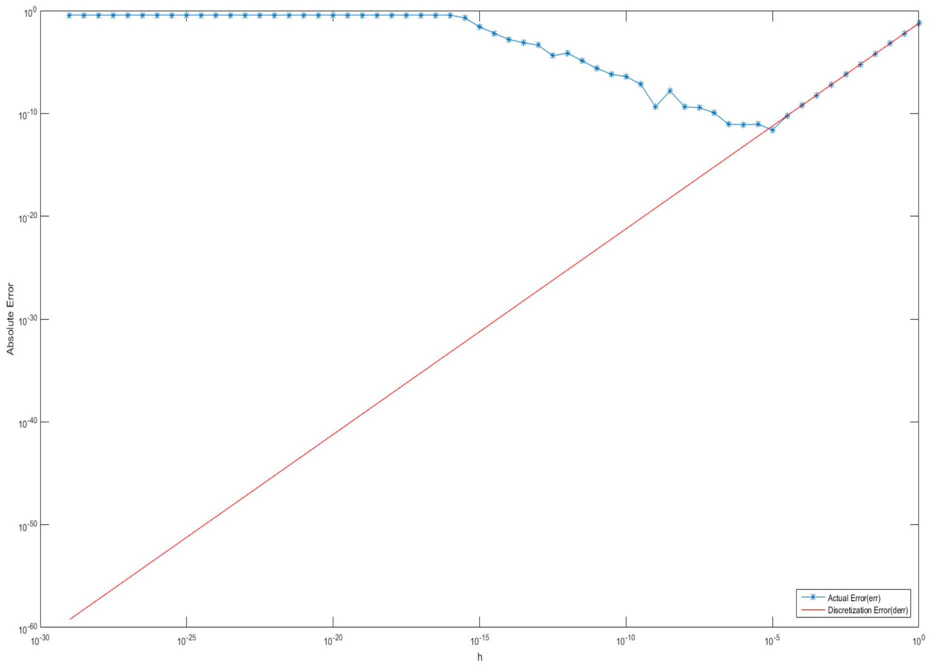

Figure 6: Error Plot for truncation error in $O(h^2)$

Number of Additions = n-1.
Number of Divisions = 1 .

Total Computation cost = 3n

For Formula F2:

Number of Subtractions = 1 .
Number of multiplications = n+1 .
Number of Additions = n-1.
Number of Divisions = 1 .

Total Computation cost = 2n+2

Thus F2 is cheaper than F1 by (n-2) steps.

(b) At first glance F2 appears to be more accurate since it contains just 1 subtraction term, which is order n less than that of F1, which means chances of F2 incurring cancellation errors is less. However, our experimental results suggests otherwise. A careful investigation of the formulas reveal that accuracy of F2 is worse than F1. The sum of squares method(F2) is computationally cheaper since it is computed in one pass with a single subtraction term. However, this subtraction occurs between two relatively close and large positive numbers, and therefore has the possibility of sufferring large cancellation errors resulting in the final result being dominated by round-off. The formula of F1 is well protected against such cases inherently and hence more accurate.

(c) Below table shows the values used for the small experiment. Listing below only one of the 5 experiments done to confirm the results. However, the results from the other non-listed experiments also conform with results.

Table 2: Experimental data for methods of Variance computation

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i^2
10.47	3.18	10.11	109.62
15.65	-2	4	244.92
12.14	1.51	2.28	147.38
13.78	-0.13	0.02	189.89
16.23	-2.58	6.66	263.41
17.17	-3.52	12.39	294.81
11.23	2.42	5.86	126.11
13.42	0.23	0.05	180.16
14.47	-0.82	0.67	209.38
11.97	1.68	2.82	143.28

$\bar{x} = 13.65$
$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 44.86$
F1 $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{44.86}{10} = 4.486$
$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 190.89$
F2 $= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2 =$ 190.89 - 186.32 $=$ 4.57
The same experiment on matlab using double precision gives a result of 4.48602 .
Thus, our small scale experiment also confirms the fact that F1 is more accurate than F2.