NUMERICAL ALGORITHMS

COL-726

HOMEWORK-1

Kartik Jain
2019MCS2563

1 ANSWER:

Given, $f: X \to Y$ and $g: Y \to Z$

$h = g \circ f : X \to Z$ and $y = f(x)$

Relative condition number for h, $K_h(x) = \sup_{\delta x} \dfrac{\|\delta h\| / \|h\|}{\|\delta x\| / \|x\|}$

$$K_h(x) = \sup_{\delta x} \frac{\|\delta\, g\circ f(x)\| / \|g\circ f(x)\|}{\|\delta x\| / \|x\|} \quad , \because h(x) = g\circ f(x)$$

$$K_h(x) = \sup_{\delta x} \frac{\|\delta g(y)\| / \|g(y)\|}{\|\delta x\| / \|x\|} \quad \because f(x) = y$$

$$K_h(x) = \sup_{\delta x} \frac{\|\delta g(y)\| / \|g(y)\|}{\|(\delta y)\| / \|y\|} \cdot \frac{\|\delta y\| / \|y\|}{\|\delta x\| / \|x\|}$$

$$K_h(x) = \frac{\|\delta g(y)\| / \|g(y)\|}{\|\delta y\| / \|y\|} \cdot \sup_{\delta x} \frac{\|\delta f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \quad , \because f(x) = y$$

$$K_h(x) \le \sup_{\delta y} \frac{\|\delta g(y)\| / \|y\|}{\|\delta y\| / \|y\|} \cdot \sup_{\delta x} \frac{\|\delta f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|}$$

$$\boxed{K_h(x) \le K_g(y) \cdot K_f(x)}$$

where $K_g(y)$ is the condition number of $g(y)$

and $K_f(x)$ is the condition number of $f(x)$

② ANSWER :

(a) $\tilde{\sin} x = \tilde{f}(x)$ has relative error of $\epsilon_{machine} = \epsilon_m$ (let)

$\Rightarrow \tilde{f}(x) = f(x)(1+\epsilon_m)$

To calculate $\frac{\sin x}{x}$ for real numbers considering relative error of ϵ_m.

(due to division)

$\Rightarrow \tilde{y} = \tilde{f}(x) = \frac{\sin \tilde{x}}{\tilde{x}}(1+\epsilon_1)(1+\epsilon_3)$ $\tilde{x}$ = floating point (x).

$= \frac{\sin[x(1+\epsilon_2)]}{x(1+\epsilon_2)}(1+\epsilon_1)$ $\because fl(x) = x(1+\epsilon)$

Ignoring higher order terms.

$= \frac{1}{x(1+\epsilon_2)}\left[x(1+\epsilon_2) - \frac{x^3}{3!}(1+\epsilon_2)^3 + \dots\right](1+\epsilon_1)$

$= \left[x - \frac{x^3(1+\epsilon_2)^2}{3!}\dots\right]\frac{(1+\epsilon_1)}{x}$

$\Rightarrow$ absolute error $= \frac{\|\tilde{y} - y\|}{\cancel{\|y\|}}$ —— ①

$\tilde{y} - y = \left[x - \frac{x^3(1+\epsilon_2)^2}{3!}\right]\frac{(1+\epsilon_1)}{x} - \frac{\sin x}{x}$

$= \frac{1}{x}\left[x - \frac{x^3(1+\epsilon_2)^2}{3!}\right](1+\epsilon_1) - \frac{1}{x}\left[x - \frac{x^3}{3!} + \dots\right]$

$= \frac{1}{x}\left[x(1+\epsilon_1) - \frac{x^3}{3!}(1+\epsilon_2)^2(1+\epsilon_1) + \dots\right] - \frac{1}{x}\left[x - \frac{x^3}{3!} + \dots\right]$

$= \frac{1}{x}\left[x + x\cdot O(\epsilon) - \frac{x^3}{3!}(1+O\epsilon)) + \dots\right] - \frac{1}{x}\left[x - \frac{x^3}{3!} + \dots\right]$

$\left[\because (1+\epsilon_2)(1+\epsilon) = 1+\epsilon_1+\epsilon_2+\epsilon_1\epsilon_2 = 1+O(\epsilon)\right]$

ignoring higher order terms$(\epsilon^2 \dots)$

$= \frac{1}{x}\left[x + x\cdot O(\epsilon) - \frac{x^3}{3!}(1+O(\epsilon) + \dots\right] - \frac{1}{x}\left[x - \frac{x^3}{3!} + \dots\right]$

$$= \frac{1}{x}\left[\left(x-\frac{x^3}{3!}+\dots\right)+\left(x-\frac{x^3}{3!}+\dots\right)O(\epsilon)\right]-\frac{1}{x}\left[x-\frac{x^3}{3!}+\dots\right]$$

$$= \cancel{\frac{1}{x}\left(x-\frac{x^3}{3!}+\dots\right)}+\frac{1}{x}\left(x-\frac{x^3}{3!}+\dots\right)O(\epsilon)\cancel{-\frac{1}{x}\left(x-\frac{x^3}{3!}+\dots\right)}$$

$$= \frac{1}{x}\left(x-\frac{x^3}{3!}+\dots\right)O(\epsilon)$$

$$= \frac{\sin x}{x}\,O(\epsilon)$$

$\Rightarrow$ Relative ~~absolute~~ error $= \frac{\|\hat{y}-y\|}{\|y\|}$, absolute error $= \|\hat{y}-y\|$

$$= \frac{(\sin x)/x \cdot O(\epsilon)}{(\sin x)/x} \qquad\qquad = \frac{\sin x}{x}\cdot O(\epsilon)$$

Relative ~~absolute~~ error $= O(\epsilon)$ $\Rightarrow$ absolute error $= \frac{\sin x}{x}\cdot O(\epsilon)$

(b) absolute error $= \frac{\sin x}{x}\cdot O(\epsilon) = f(x)\cdot O(\epsilon)$

and the maximum value of $\frac{\sin x}{x}$ can be 1

$\Rightarrow$ absolute error $= 1\cdot O(\epsilon) = O(\epsilon)$.

(b)

$\rightarrow$ We have relative forward error $= \frac{\|\hat{y}-y\|}{\|y\|}$

& $\hat{y}-y = \left(\frac{\sin x}{x}\right)\cdot(1+\epsilon) - \sin x$

$$= \frac{\sin x}{x}\cdot(O(\epsilon)+1)$$

$\Rightarrow$ relative forward error $= \frac{\left(\frac{\sin x}{x}\right)(1+O(\epsilon))}{\left(\frac{\sin x}{x}\right)} = O(\epsilon)$.

$\Rightarrow$ The algorithm is accurrate.

→ Given that x is small & non zero,

Then,

$$\Rightarrow \left(\widetilde{\frac{\sin x}{x}}\right) = \frac{\sin x}{x}(1+\epsilon_1) = \frac{\sin x(1+\epsilon_2)}{x(1+\epsilon_2)}(1+\epsilon_1)$$

also, for backward stability we should have $f(\tilde{x}) = \tilde{f}(x)$

And for $x = \pi$,

$$f(\tilde{x}) = \frac{\sin x(1+\epsilon)}{x(1+\epsilon)} = \frac{\sin \pi + \pi\epsilon}{\pi + \pi\epsilon} = \frac{-\sin \pi\epsilon}{\pi + \pi\epsilon}$$

$$= \frac{-\epsilon}{\pi + \pi\epsilon} \quad [\because \text{for very small numbers } \sin x \approx x]$$

and $\tilde{f}(x) = \frac{\sin \pi}{\pi} \cdot (1+\epsilon) = \frac{0}{\pi}(1+\epsilon) = \underline{0}$

$\Rightarrow f(\tilde{x}) \neq \tilde{f}(x)$

$\Rightarrow \frac{\sin x}{x}$ is not backward stable.

⇒ Also in general for some smaller value of x

$$\left(\widetilde{\frac{\sin x}{x}}\right) = \frac{\sin x}{x}\cdot(1+\epsilon) \approx \frac{x}{x}(1+\epsilon) = (1+\epsilon)$$

and $$\frac{\sin x(1+\epsilon)}{x(1+\epsilon)} = \frac{\epsilon}{x(1+\epsilon)}$$

Thus $\tilde{f}(x) \neq f(\tilde{x})$

& $\frac{\sin x}{x}$ is not backward stable

Scanned by CamScanner

(3) ANSWER :

Let us assume that the matrix U have inverse even if some diagonal entry is zero.

Then consider two cases :

i) The last element of last diagonal U_{mm} is zero then the entire last row of matrix becomes 0.

which means the rank of U now will be $(m-1)$ and the det(A

determinant(U) = 0 (U is singular)

⇒ The matrix U does not have any inverse.

$$U = \begin{bmatrix} U_{11} & U_{12} & \cdots & U_{1m} \\ & U_{22} & \cdots & U_{2m} \\ & & \ddots & \vdots \\ & & & U_{mm} \end{bmatrix}$$

(ii) Consider some other element $U_{ii} = 0$, then we have

$$U = \begin{bmatrix} U_{11} & \cdots & \cdots & U_{1m} \\ & a_{ii} & & \vdots \\ & & U_{i+1\,i+1} & \vdots \\ & & \ddots & U_{mm} \end{bmatrix}$$

⇒ Thus, now the i^{th} row can be written as the linear combination of the rows $(i+1)$ to (m) ⇒ (All rows are not independent)

which still reduces the rank$(U) = (m-1)$ and $\det(U) = 0$

⇒ A becomes singular.

⇒ Matrix U does not have any inverse

⇒ Thus, in both cases our initial assumption that U is invertible even if some diagonal entry is zero is false.

⇒ Every diagonal entry u_{ii} is non zero.

④ ANSWER :

(a)

$\rightarrow \|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty$

$\Rightarrow$ we know $\|x\|_1 = \sum_i |x_i| = |x_1| + |x_2| + |x_3| + \ldots$

$= \sqrt{(x_1 + x_2 + x_3 + \ldots)^2}$

$\|x\|_1 = \sqrt{x_1^2 + x_2^2 + x_3^2 + \ldots + x_m^2 + \sum_{i \neq j} \sum_j \alpha x_i x_j}$ α = some constant

$\|x\|_1 \geq \sqrt{x_1^2 + x_2^2 + \ldots + x_m^2}$

$\|x\|_1 \geq \|x\|_2$ $\left(\because \sqrt{x_1^2 + x_2^2 + \ldots} = \|x\|_2\right)$

and $\sqrt{x_1^2 + x_2^2 + \ldots + x_m^2} = \sqrt{x_j^2 + \sum_{i \neq j} x_i^2}$

$\|x\|_2 = \sqrt{x_j^2 + \sum_{i \neq j} x_i^2}$ where $|x_j|$ = maximum value in all of x_is, thus x_j^2 will also be maximum then all of x_is

$\|x\|_2 \geq \sqrt{x_j^2}$

$\|x\|_2 \geq x_j$

$\|x\|_2 \geq \max |x_i| \quad \forall i = 1 \ldots m$

$\|x\|_2 \geq \|x\|_\infty$

$$\boxed{\Rightarrow \quad \|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty}$$

(b) $\rightarrow \frac{1}{\sqrt{m}} \|x\|_1 \leq \|x\|_2 \leq \sqrt{m} \|x\|_\infty$

we know $\|x\|_1 = |x_1| + |x_2| + \ldots$

$= \sum_i x_i = \left(\sum_i x_i \cdot 1\right)$

From Cauchy-Schwarz Inequality, we know

$$\sum_i a_i b_i \leq \left(\sum_i a_i^2\right)^{1/2} \left(\sum_i b_i^2\right)^{1/2}$$

$$\Rightarrow \sum_i x_i \cdot 1 \le \left(\sum_i x_i^2\right)^{1/2} \left(\sum_i 1^2\right)^{1/2}$$

$$\Rightarrow \|x\|_1 \le \sqrt{\sum_i x_i^2} \cdot \left(\sum_i 1\right)^{1/2}$$

$$\Rightarrow \|x\|_1 \le \sqrt{\sum_i x_i^2} \cdot (m)^{1/2}$$

$$\Rightarrow \frac{1}{\sqrt{m}} \|x\|_1 \le \|x\|_2$$

Also, $\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_m^2}$

$\|x\|_2 = \sqrt{x_j^2 + \sum_{i \ne j} x_i^2}$ where $x_j = \max |x_i|$

$\|x\|_2 \le \sqrt{x_j^2 + \sum_{i \ne j} x_j^2}$ (Replacing all x_i's with x_j's)

$$\|x\|_2 \le \sqrt{m x_j^2}$$

$$\|x\|_2 \le \sqrt{m} \sqrt{x_j^2}$$

$$\|x\|_2 \le \sqrt{m}\, x_j$$

$$\|x\|_2 \le \sqrt{m} \max_i |x_i|$$

$$\|x\|_2 \le \sqrt{m} \|x\|_\infty$$

$$\boxed{\Rightarrow \frac{1}{\sqrt{m}} \|x\|_1 \le \|x\|_2 \le \sqrt{m} \|x\|_\infty}$$

(b) For any matrix, A of size $(m \times n)$ and x be a vector of size n

$\Rightarrow Ax$ will be a vector of size m.

$\rightarrow \|A\|_2 = \frac{\|Ax\|_2}{\|x\|_2} \leq \frac{\|Ax\|_1}{\|x\|_2}$ $\quad (\because \|x\|_1 \geq \|x\|_2)$

$\|A\|_2 \leq \frac{\|Ax\|_1}{\frac{1}{\sqrt{n}}\|x\|_1}$ $\quad (\because \frac{1}{\sqrt{n}}\|x_1\| \leq \|x\|_2 \Rightarrow$ Result of division will be further greater$)$

$\|A\|_2 \leq \sqrt{n}\,\|A_1\|$

$\rightarrow \|A\|_2 = \frac{\|Ax\|_2}{\|x\|_2} \geq \frac{\frac{1}{\sqrt{m}}\|Ax\|_1}{\|x\|_1}$ $\quad (\because \frac{1}{\sqrt{m}}\|x\|_1 \leq \|x\|_2$ & $\|x\|_1 \geq \|x\|_2)$

$\|A_2\| \geq \frac{1}{\sqrt{m}}\|A\|_1$

$$\boxed{\Rightarrow \frac{1}{\sqrt{m}}\|A_1\| \leq \|A\|_2 \leq \sqrt{n}\,\|A\|_1}$$

$\rightarrow \|A\|_\infty = \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \frac{\|Ax\|_2}{\|x\|_\infty}$ $\quad (\because \|x\|_\infty \leq \|x\|_2)$

$\|A\|_\infty \leq \frac{\|Ax\|_2}{\frac{1}{\sqrt{m}}\|x\|_2}$ $\quad (\because \|x\|_2 \leq \sqrt{m}\,\|x\|_\infty)$

$\Rightarrow \frac{1}{\sqrt{m}}\|x\|_2 \leq \|x\|_\infty$

$\|A\|_\infty \leq \sqrt{n}\,\|A\|_2$

Scanned by CamScanner

$\Rightarrow \because \|A\|_\infty \leq \sqrt{n}\, \|A\|_2$ (Increasing the numerator, decreasing the denominator, Increases the overall result of division)

$\Rightarrow \frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2$

$\rightarrow \|A\|_\infty = \frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \frac{\|Ax\|_\infty}{\|x\|_2}$ $(\because \|x\|_2 \geq \|x\|_\infty)$

$\Rightarrow$ Numerator will be increased

$$\|A\|_\infty \geq \frac{\|Ax\|_\infty}{\|x\|_2} \qquad \left(\|x\|_\infty \geq \frac{1}{\sqrt{m}} \|x\|_2 \right)$$

$$\|A\|_\infty \geq \frac{\frac{1}{\sqrt{m}} \|Ax\|_2}{\|x\|_2}$$

$$\Rightarrow \|A\|_\infty \geq \frac{1}{\sqrt{m}} \|A\|_2$$

$$\boxed{\Rightarrow \quad \frac{1}{\sqrt{m}} \|A\|_2 \leq \|A\|_\infty \leq \sqrt{n} \|A\|_2}$$

(5) ANSWER:

Given A has a singular value decomposition:

condition number, $K(x) = \sup\limits_{\delta b} \frac{\|\delta x\| / \|x\|}{\|\delta b\| / \|b\|}$

for $Ax = b \Rightarrow x = A^{-1} b$

$$K(x) = \sup_{\delta b, b} \frac{\|\delta A^{-1} b\| / \|A^{-1} b\|}{\|\delta b\| / \|b\|}$$

$$K(x) = \sup_{\delta b, b} \frac{\|\delta A^{-1} b\|}{\|\delta b\|} \cdot \frac{\|b\|}{\|A^{-1} b\|}$$

$$K(x) = \|A^{-1}\| \cdot \frac{\|b\|}{\|A^{-1} b\|}$$

$$K(x) = \|A^{-1}\| \sup_{x} \frac{\|Ax\|}{\|x\|} = \|A^{-1}\| \cdot \|A\|$$

$$\boxed{\Rightarrow\ K(x) = \|A\| \cdot \|A^{-1}\|} \quad \text{———} \ (1)$$

And using 2-norm for calculating norms of the matrix:

$\Rightarrow \|A\|_2 = \|U \Sigma V^*\|_2$

$\|A\|_2 \leq \|U\|_2 \cdot \|\Sigma\|_2 \cdot \|V^*\|_2$

$\|A\|_2 \leq \|\Sigma\|_2$ (since $\|U\| = \|V^*\| = I$, since U, V are unitary matrices)

$\|A\|_2 = \|\Sigma\|_2$ → (Taking max. values of each column)

$\|A\|_2 = \underline{\sigma_1}$ (where σ_1 is the max. column value)

and $\Rightarrow \|A^{-1}\|_2 = \|(U \Sigma V^*)^{-1}\|_2$

$\|A^{-1}\|_2 = \| V \Sigma^{-1} U^* \|_2$

$\|A^{-1}\|_2 = \|V\|_2 \, \|\Sigma^{-1}\|_2 \, \|U^*\|$

$\|A^{-1}\|_2 = \|\Sigma^{-1}\|_2$

$\|A^{-1}\|_2 = 1/\sigma_n$ ($\because \Sigma^{-1}$ consists of reciprocating the diagonal elements and then $1/\sigma_n$ becomes the greatest value available).

$\Rightarrow K(x) = \|A\|_2 \cdot \|A^{-1}\|_2$

$$\boxed{K(x) = \frac{\sigma_1}{\sigma_n}}$$

where σ_1 = max value at diagonal of Σ

σ_n = min value at diagonal of Σ

Example:

Consider the Matrix $A = \begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix}$, then

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = U \Sigma V^* \text{ as the SVD of } A$$

then $\sigma_1 = 10$, $\sigma_2 = 0.1$

$$K(A) = \frac{10}{0.1} = \underline{100}.$$

Example

Given $A = \begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}$

$$\Rightarrow K(b) = \sup_{\delta b} \frac{\|\delta A^{-1}b\| / \|\delta A^{-1}b\|}{\|\delta b\| / \|b\|}$$

$$= \sup_{\delta b, b} \frac{\|A^{-1}\delta b\| / \|\delta b\|}{\|A^{-1}b\| / \|b\|}$$

let $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, $\delta b = \begin{bmatrix} \delta b_1 \\ \delta b_2 \end{bmatrix}$

$$A^{-1}\delta b = \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}\begin{bmatrix} \delta b_1 \\ \delta b_2 \end{bmatrix} = \begin{Bmatrix} 0.1\,\delta b_1 \\ 10\,\delta b_2 \end{Bmatrix} = \begin{bmatrix} 0 \\ 10\delta \end{bmatrix}, \quad \text{let } \delta b_1 = 0,\ \delta b_2 = \delta$$

$$\Rightarrow \frac{\|A^{-1}\delta b\|}{\|\delta b\|} = \frac{10\delta}{\delta} = \underline{\underline{10}}.$$

& $A^{-1}b = \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{pmatrix} 0.1 \\ 0 \end{pmatrix}$ let $b = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

$$\Rightarrow \frac{\|A^{-1}b\|}{\|b\|} = \frac{0.1}{1} = 0.1$$

$$\Rightarrow K(b) = \sup_{\delta b, b} \frac{\|A^{-1}\delta b\| / \|\delta b\|}{\|A^{-1}b\| / \|b\|} = \frac{10}{0.1} = \underline{\underline{100}}$$

$\Rightarrow$ Thus, the value obtained achieved the expected bound.

(6) ANSWER:

(a) The formula may lead to loss of accuracy due to:

- cancellation when subtracting b^2 and $4ac$
- cancellation when adding $-b$ and $\sqrt{b^2-4ac}$.

$\Rightarrow$ Also, these both can ~~lead to~~ be the result of the ill conditioning of the problem as a small perturbation might lead to cancellation errors.

⑥

(b)

let $a(h) = 1/4$

$b(h) = 100 + 1.11h$

$c(h) = (100 + 1.11h)^2 - (100 + 1.1h)^2$ [using $a^2 - b^2 = (a+b)(a-b)$]

$= 0.01 * h * (200 + 2.21 * h)$

Then, using the quadroots formula, the smaller magnitude root will be:

$$\frac{-b(h) + \sqrt{b^2 - 4ac}}{2a}$$

$$= \frac{-(100 + 1.11h) + \sqrt{(100+1.11h)^2 - [(100+1.11h)^2 - (100+1.1h)^2] \cdot 1 \cdot \cancel{4} \cdot \frac{1}{\cancel{4}}}}{\cancel{2} \cdot \frac{1}{\cancel{4}_2}}$$

$$= \frac{-(100+1.11h) + \sqrt{\cancel{(100+1.11h)^2} - \cancel{(100+1.11h)^2} + (100 + 1.1h)^2}}{1/2}$$

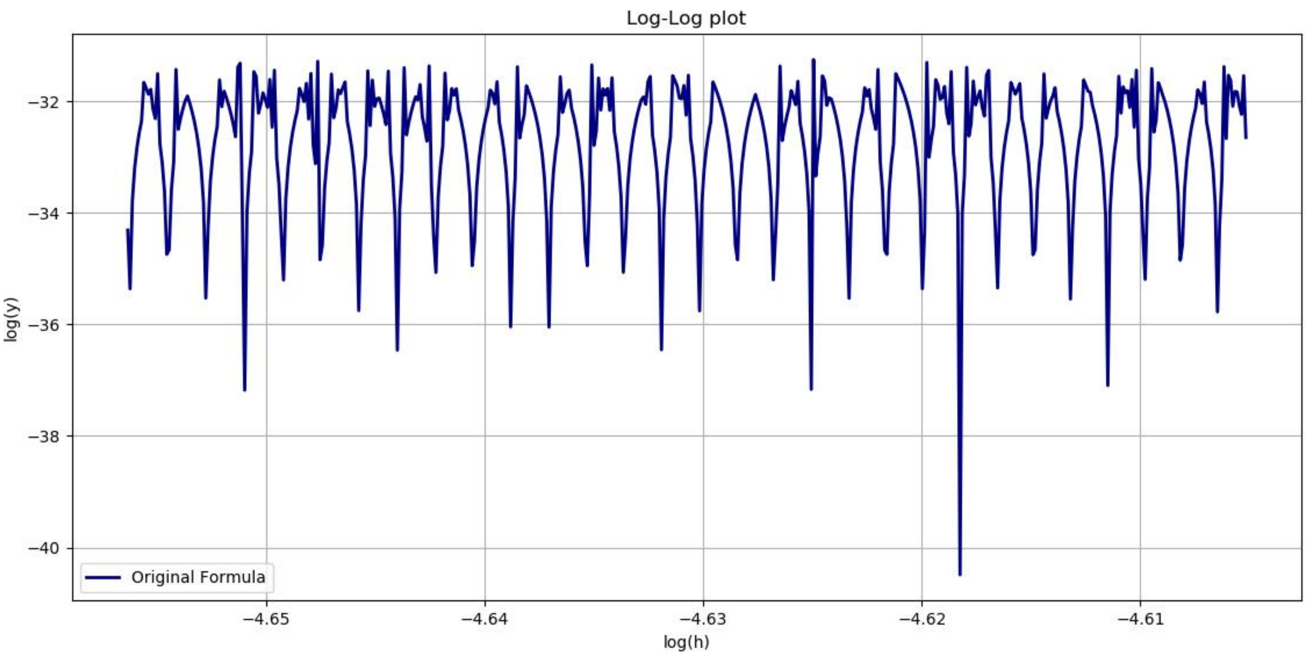
$$= \frac{-(100 + 1.11h) + (100 + 1.1h)}{1/2} \quad \text{—— ①}$$

$= -2 * 0.01h$

$$\boxed{x^* = -0.02h}$$

⇒ In ① the numerator involves a subtraction where values will be almost equal and thus cancellation errors will be there while calculating this subtraction.

⇒ This is also depicted by the graph below.



Log-Log plot
−32
−34
−36
−38
−40
log(y)
Original Formula
−4.65
−4.64
−4.63
−4.62
−4.61
log(h)

⑥

(c)

→ Implemented quadroots and added it on the same graph for the smaller root.

We found that the log(y) values for new root are more negative (in general) than previous ones which is on the log scale.

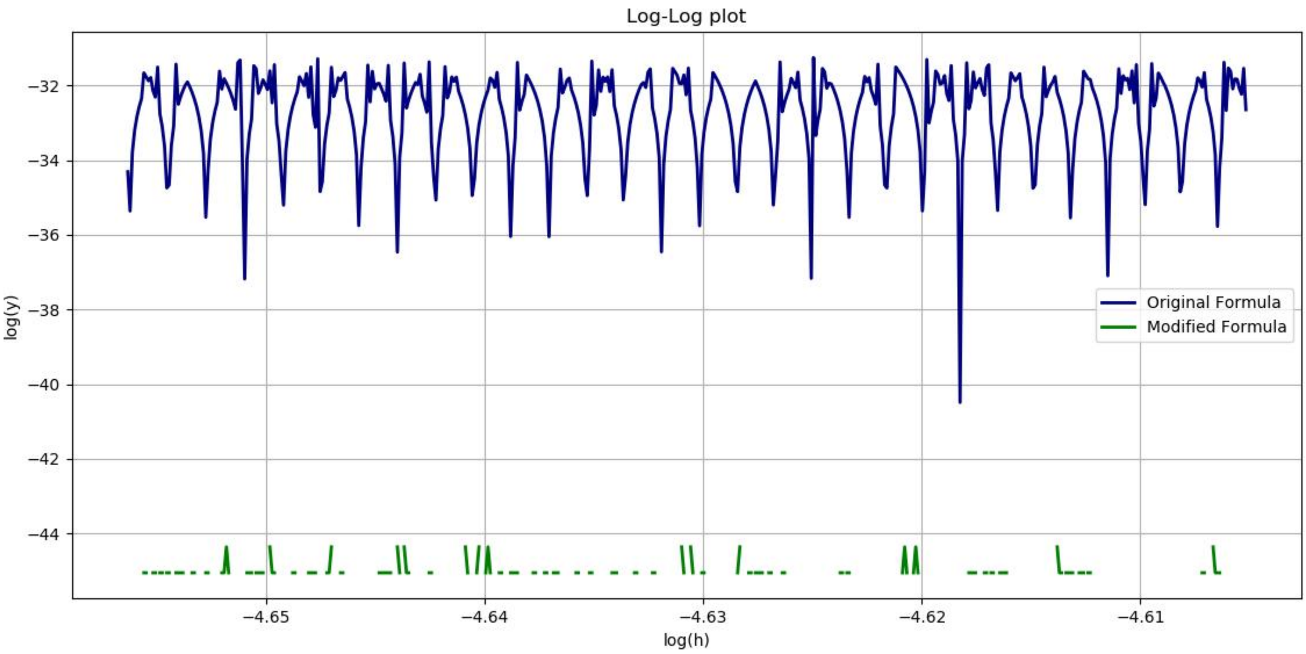
Thus we can infer that the new formula is giving better results.

⇒ But still there are other errors in the new formula which might include truncation errors, rounding errors etc, but as compared to the values calculated by the previous function, they are much better. (Gaps in the line denotes no error).

NOTE: Since it is a log scale, the less the error the more negative the value gets.

⇒ for a small number < 0.1 and > 0, log gives a negative number whose magnitude increases as the number gets smaller.

⇒ Smaller error gives higher negative value (magnitude)



Log-Log plot
log(y)
log(h)
−32
−34
−36
−38
−40
−42
−44
−4.65
−4.64
−4.63
−4.62
−4.61
Original Formula
Modified Formula