Machine epsilon

Rounding :

$$1.0 = 1.\underbrace{0\ldots0}_{t-1\ 0's} \times b^0$$

Next number after 1.0 is

$$1.\underbrace{0\ldots01}_{t-2\ 0's} \times b^0$$

$$1.0 = 1.\overbrace{0\ldots000}^{t-2} \times b^0$$

$$\epsilon_{mach} = 0.0\ldots005 \times b^0$$

$$1.0\ldots005 \times b^0$$

$$1.0 + \epsilon_{mach} = 1.0\ldots01 \times b^0$$

base 10 (1's if base 2)

round to the next number after 1.0

$$\text{So } \epsilon_{mach} = 5 \times b^{-t} = \frac{1}{2} b^{1-t}$$

— rounding

FORTRAN has built-in (intrinsic) functions

epsilon(1.0) $\approx 10^{-7}$ for REAL

epsilon(1D0) $\approx 10^{-15}$ for DOUBLE PRECISION

Roundoff Error in Multiplication

$$x_1 = \bar{x}_1 + e_1 , \quad x_2 = \bar{x}_2 + e_2$$

$\bar{x}_1, \bar{x}_2$ are the numbers on the ~~comput~~ computer which represent the actual numbers x_1, x_2; e_1, e_2 are the roundoff errors.

Multiply $\bar{x}_1$ and $\bar{x}_2$ on the machine,

$$\bar{x}_1 \bar{x}_2 = \overline{\bar{x}_1 \bar{x}_2} + e_{12}^*$$

(e_{12}^* \ error in $\bar{x}_1 \bar{x}_2$)

| Relative error in $\bar{x}_1 \bar{x}_2$ | $\leq \epsilon_{mach}$ (for the numbers we are computing with)

$$\therefore \frac{|e_{12}^*|}{|\bar{x}_1 \bar{x}_2|} \leq \epsilon_{mach} , \text{ i.e. } |e_{12}^*| \leq \epsilon_{mach} |\bar{x}_1 \bar{x}_2|$$

In practice this is a reasonable estimate not simply an upper bound, i.e.

$$|e_{12}^*| \sim \epsilon_{mach} |\bar{x}_1| |\bar{x}_2|$$

($\sim$ means "roughly equals".)

What is the error in $x_1 x_2$?

$$x_1 x_2 = (\bar{x}_1 + e_1)(\bar{x}_2 + e_2)$$

$$= \bar{x}_1 \bar{x}_2 + e_1 \bar{x}_2 + e_2 \bar{x}_1 + e_1 e_2$$

neglect since very small

$$= \overline{\bar{x}_1 \bar{x}_2} + e_{12}^{\not} + e_1 \bar{x}_2 + e_2 \bar{x}_1$$

$$\sim \overline{\bar{x}_1 \bar{x}_2} \pm |\bar{x}_1||\bar{x}_2| \epsilon_{mach} + e_1 \bar{x}_2 + e_2 \bar{x}_1$$

$$\therefore \left| \frac{x_1 x_2 - \overline{\bar{x}_1 \bar{x}_2}}{x_1 x_2} \right| \sim \left| \frac{\bar{x}_1}{x_1} \right| \left| \frac{\bar{x}_2}{x_2} \right| \epsilon_{mach} + \left| \frac{e_1}{x_2} \right| \left| \frac{\bar{x}_2}{x_2} \right| + \left| \frac{e_2}{x_2} \right| \left| \frac{\bar{x}_1}{x_1} \right|$$

$$\sim \epsilon_{mach} + \left| \frac{e_1}{x_1} \right| + \left| \frac{e_2}{x_2} \right|$$

$$\rightarrow \epsilon_{mach} + RE_1 + RE_2$$

RE_1: relative error in x_1 when entered in computer.

RE_2: relative error in x_2 when entered in computer.

$$\sim 3\,\epsilon_{mach}$$

Similarly for division.

Relative Function Test

Consider a function $f(x)$ of the form

$$f(x) = \sum_{k=1}^{K} t_k(x)$$

ie f is a sum of K terms.

We seek a zero of f, i.e. x^* such that $f(x^*) = 0$. This implies

$$\sum_{k=1}^{K} t_k(x^*) = 0$$

In general we do not expect each $t_k(x^*) = 0$ for $k = 1:K$.

Thus to get 0 there must be cancellation of the $t_k(x^*)$.

In practice if $\bar{x}$ is the approximation we compute to x^*, then we get $f(\bar{x}) = \sum_{k=1}^{K} t_k(\bar{x})$. How close to 0 is this? $\neq 0$

The relative function test is the following: if $\{x_i\}$ is an approximation to a zero x^* of $f(x)$, then stop if

$$\left| \frac{f(x_i)}{t_K(x_i)} \right| < \epsilon$$

where ϵ is a given error tolerance & $|t_K(x^*)| \geq |t_{K-1}(x^*)| \geq \dots \geq |t_1(x^*)|$.

Another form for the test is

$$\frac{|f(x_i)|}{\sum_{k=1}^{K} |t_k(x_i)|} = \frac{\left| \sum_{k=1}^{K} t_k(x_i) \right|}{\sum_{k=1}^{K} |t_k(x_i)|} < \epsilon.$$

Eg 1

$t_1(\bar{x}) = 0.\underbrace{x_1x_1x_1x_1x_1x_1x}_{7}\downarrow \text{unreliable after 7th decimal place}$

$t_2(\bar{x}) = 0.x_2x_2\ldots\ldots x_2---$

$\vdots$

$t_K(\bar{x}) = 0.x_Kx_K\ldots\ldots x_K---$

terms are about 1

$0 \approx 0.0000000\_+-$ best outcome possible

$\downarrow$ anything

$\therefore$ the error in the best case will be about 10^{-7}

Eg 2

$t_1(\bar{x}) = \underbrace{x_1x_1\ldots x_1}_{7 \text{ digits before decimal point}}.---$

$\vdots$

$t_K(\bar{x}) = x_Kx_K\ldots x_K.---$

terms are about 10^7

$0 \approx 1.---$

Error Analysis of $\delta_h f(x)$

$$\delta_h f(x) := \frac{f(x+h) - f(x-h)}{2h}$$

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

We can use $\delta_h f(x)$ or $\Delta_h f(x)$ to approximate $\frac{df}{dx}(x) \equiv f'(x)$.

$\delta_h f(x)$ is more accurate.

(b) Truncation error :

$$f(x+h) = f(x) + f'(x)h + f^{(2)}(x)\,h^2/2 + f^{(3)}(x)\,h^3/6 + f^{(4)}(x)\,h^4/24 + f^{(5)}(\eta_+)\,h^5/120$$

$$f(x-h) = f(x) - f'(x)h + f^{(2)}(x)\,h^2/2 - f^{(3)}(x)\,h^3/6 + f^{(4)}(x)\,h^4/24 - f^{(5)}(\eta_-)\,h^5/120$$

$$\delta_h f(x) = f'(x) + f^{(3)}(x)\,h^2/6 + \frac{1}{2}\left[f^{(5)}(\eta_+) + f^{(5)}(\eta_-)\right] h^4/120$$

But if $f^{(5)}$ is continuous there exists η such that

$$\frac{1}{2}\left[f^{(5)}(\eta_+) + f^{(5)}(\eta_-)\right] = f^{(5)}(\eta)$$

using the intermediate value theorem.

$$\delta_h f(x) = f'(x) + \underbrace{f^{3}(x)\,h^2/6 + f^{(5)}(\eta)\,h^4/120}_{\text{truncation error in } \delta_h f(x).}$$

(c) Roundoff error in formula

Assume $\delta_h f(x)$ is calculated exactly (eg in computer registers), then rounded or chopped. Then

$$|\text{roundoff error in } \delta_h f(x)| \leq \epsilon_{mach} \times |\delta_h f(x)|$$

$$\leq \frac{|f(x+h)| + |f(x-h)|}{2h}\,\epsilon_{mach}$$

$$\sim \frac{|f(x)|}{h} \epsilon_{mach}$$

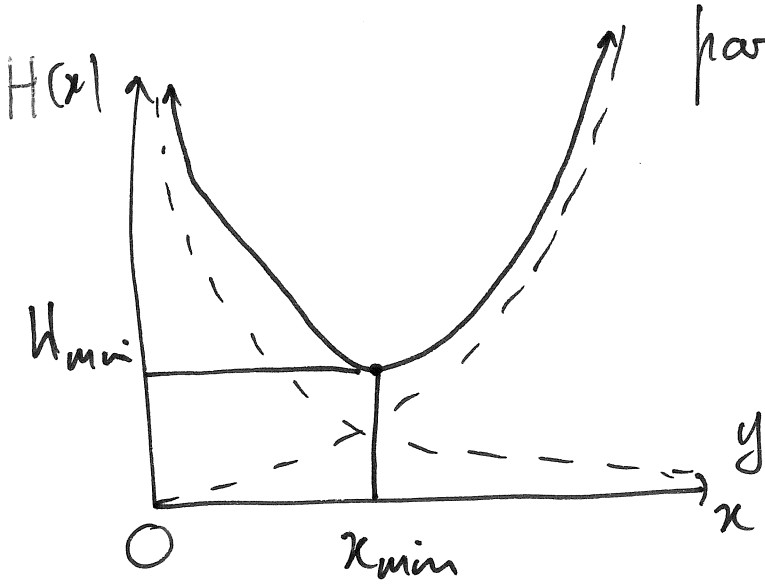
This is a bound on the roundoff error but it is a reasonable estimate usually, i.e.

$$\left| \text{roundoff error in } \delta_h f(x) \right| \simeq \frac{|f(x)|}{h} \epsilon_{mach}$$

(d) Total error

$$\left| \text{total error using } \delta_h f(x) \text{ to approximate } f'(x) \right| \leq \underbrace{\left| f^{(3)}(x) h^2/6 \right|}_{\text{truncation}} + \underbrace{\frac{|f(x)|}{h} \epsilon_{mach}}_{\text{roundoff}} + \underbrace{\left| f^{(5)}(x) h^4/120 \right|}_{\text{remainder}} \quad (*)$$

Consider $H(x) = ax^2 + \frac{b}{x}$, $a, b > 0$



Minimise $H(x)$: $\quad H'(x) = 2ax - \frac{b}{x^2} = 0$

$$\Rightarrow \quad x_{min} = \sqrt[3]{\frac{b}{2a}}$$

$$H_{min} = a\left(\frac{b}{2a}\right)^{2/3} + b\left(\frac{b}{2a}\right)^{-\frac{1}{3}}$$

$$= \frac{3}{2}\sqrt[3]{2ab^2}$$

Comparing with the total error bound we see

$$h = x, \quad a = \frac{|f^{(3)}(x)|}{6}, \quad b = |f(x)|\epsilon_{mach}$$

The minimum bound on the total error is occurs when

$$h_{min} = \sqrt[3]{\frac{|f(x)|\epsilon_{mach}}{2|f^{(3)}(x)|/6}} \quad (= x_{min})$$

$$\sim \sqrt[3]{\epsilon_{mach}}\, h \quad (\text{ignore } \sqrt[3]{3})$$

where h is a measure of the length scale of $f(x)$:

$$h^3 = \left| \frac{f(x)}{f^{(3)}(x)} \right|.$$

Minimum bound (& estimate) of the error is

$$\left| \begin{array}{c} \text{total} \\ \text{error} \\ \text{in} \\ \\ \delta_h f(x) \end{array} \right| \sim \frac{f}{h} (\epsilon_{mach})^{2/3}$$

where f is a typical value of $f(x)$.

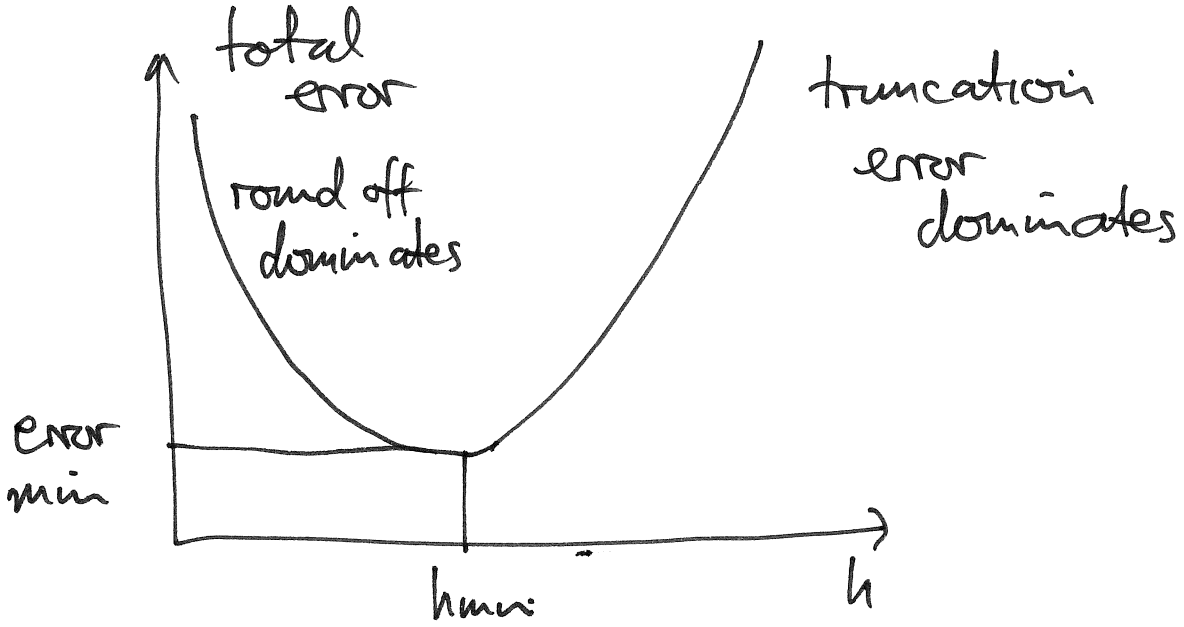
But $f/h \sim f'(x)$ so the

$$\left| \begin{array}{c} \text{minimum} \\ \text{relative} \\ \text{error in} \\ \delta_h f(x) \end{array} \right| \sim (\epsilon_{mach})^{2/3}.$$

In single precision with $\epsilon_{mach} \approx 10^{-7}$

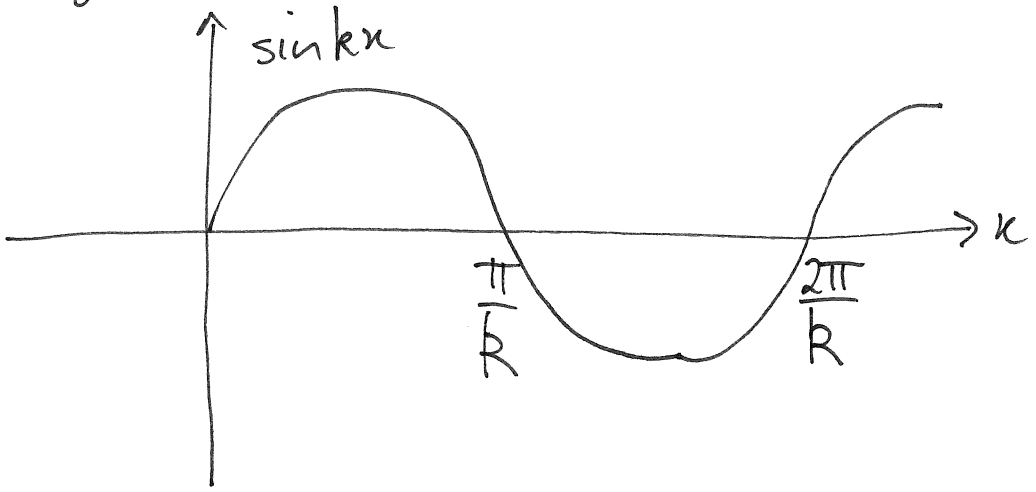
$(\epsilon_{mach})^{2/3} \approx 10^{-4}$ & $(\epsilon_{mach})^{1/3} \approx 10^{-2}$.

General picture

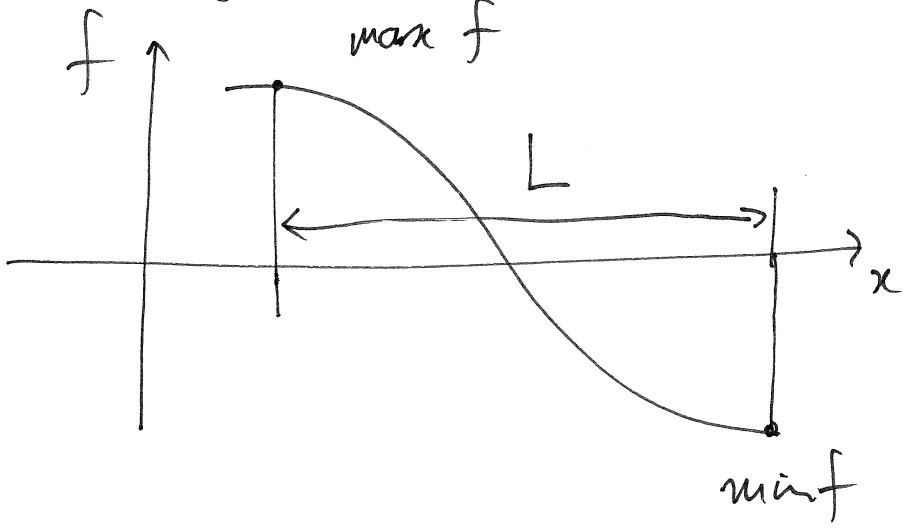


Length scale of a function

Eg $\sin kx$

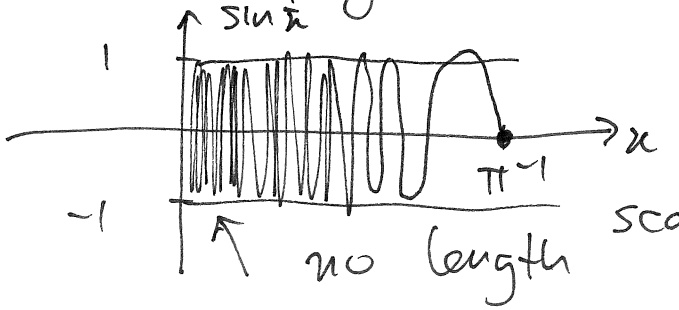


The length scale of $\sin kx$ is $\sim \frac{\pi}{k}$



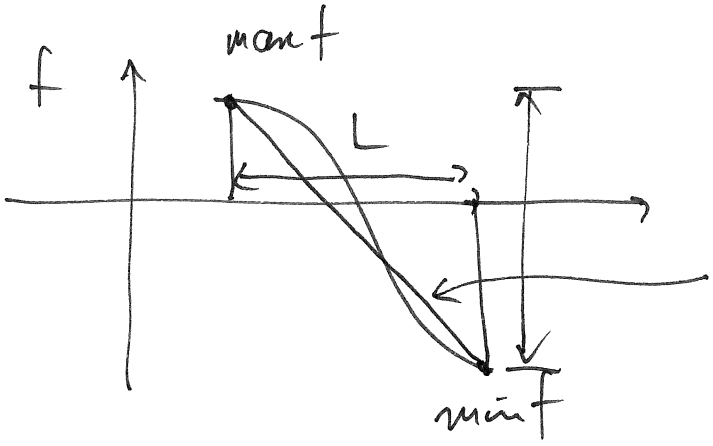
L = length scale of f.

Length scale cannot be defined for every function: eg $\sin \frac{1}{x}$ in $[0, \frac{1}{\pi}]$



↑ no length scale near $x=0$.

We can use L to estimate derivatives of f.



$$\text{slope of line segment} = \frac{\max f - \min f}{L} \sim \frac{f}{L}$$

where f is a typical value of the function f.

This gives

$$f^{(1)}(x) = \text{slope of tangent to } f \text{ at } x \sim \frac{f}{L}$$

For higher derivatives we take

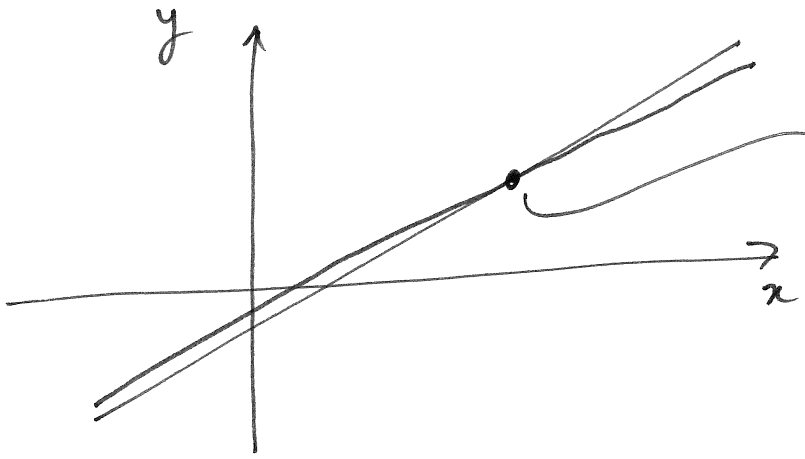
$$f^{(n)}(x) \sim \frac{f}{L^n}$$ ← order of derivative

Eg $f^{(3)}(x) \sim \frac{f(x)}{L^3}$, assuming $f(x)$ is typical

ie $\frac{f^{(3)}(x)}{f(x)} \sim \frac{1}{L^3}$ — we used this in the analysis of $\delta_h f(x)$.

Ill-Conditioned Problems

① Two nearly parallel straight lines:



solution is point of intersection

$$a_1x + b_1y = c_1$$
$$a_2x + b_2y = c_2$$

Here data is : $a_1, b_1, c_1, a_2, b_2, c_2$

Here a slight perturbation in the data shifts the point of intersection a long way.

② $y'' - y = 0, \quad y(0) = 1, \quad y'(0) = -1$

$$y = c_1e^{-x} + c_2e^{x}$$

$$\left.\begin{aligned} y(0) = 1 &\Rightarrow c_1 + c_2 = 1 \\ y'(0) = -1 &\Rightarrow -c_1 + c_2 = -1 \end{aligned}\right\} \Rightarrow$$

$$c_1 = 1, \quad c_2 = 0$$

Try $y = e^{px}$

$p^2 - 1 = 0$

$\Rightarrow p = \pm 1$

$\Rightarrow e^{x}, e^{-x}$

$\therefore \quad y = e^{-x} \to 0 \quad \text{as} \quad x \to \infty$

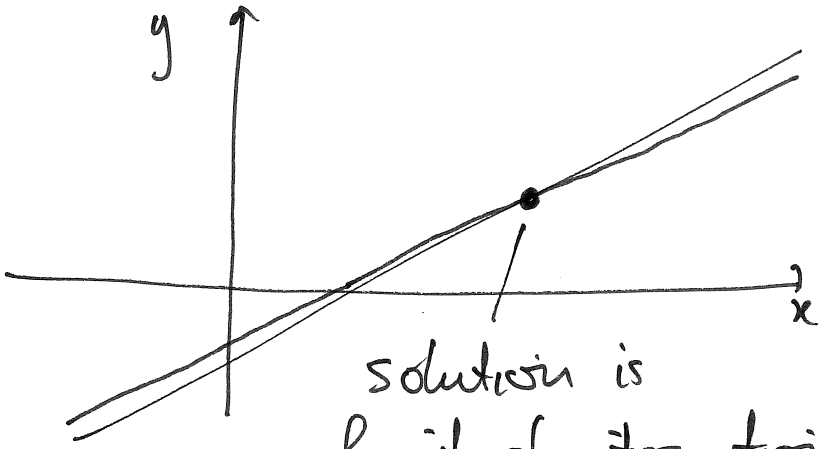
If the data is perturbed so that

$y(0) = 1 + \delta_1, \quad y'(0) = -1 + \delta_2 \Rightarrow \quad c_1 \approx 1, \quad c_2 \neq 0$ but small

Then $y \approx e^{-x} + c_2e^{x} \to +\infty$ as $x \to \infty$ if $c_2 > 0$ if δ_1, δ_2 small

Ill-Conditioned Problems

① Two nearly parallel straight lines:



$$a_1x + b_1y = c_1$$
$$a_2x + b_2y = c_2$$

Here the data is $\{a_1, b_1, c_2\}$, $\{a_1, b_2, c_2\}$.

Here a slight perturbation to the data can shift the point of intersection a long way.

② $y'' - y = 0, \quad y(0)=1, \quad y'(0)=-1.$

y'' = $\frac{d^2y}{dx^2}$

Problem: determine $y(x)$.

$$y = c_1e^x + c_2e^{-x}$$

$$\left.\begin{aligned} y(0) &= c_1 + c_2 = 1 \\ y'(0) &= c_1 - c_2 = -1 \end{aligned}\right\} \Rightarrow c_1 = \quad , \; c_2 = \quad 1$$

$$\Rightarrow y = e^{-x}$$

As $x \to \infty$, $y \to 0$

If the data are perturbed, eg the initial conditions at $x=0$,

$$y(0) = 1 + \delta_1 \qquad y'(0) = -1 + \delta_2$$

where δ_1, δ_2 are small,

then $C_1 \neq 0, \quad C_2 \approx 1$

The solution $y = C_1 e^{+x} + C_2 e^{-x}$

$\neq 0$

For x small, $y \approx e^{-x}$.

But as $x \to \infty$, $y \to \begin{cases} +\infty & C_1 > 0 \\ -\infty & C_1 < 0 \end{cases}$.

(3) Set 1 q8 – 3 term recurrence relation

$$\phi_{k+1}(x) = \frac{2k}{x}\phi_k(x) - \phi_{k-1}(x)$$

$$\phi_k(x) = C_1 J_k(x) + C_2 Y_k(x)$$

$$\left.\begin{aligned} \phi_0(x) &= J_0(x) \\ \phi_1(x) &= J_1(x) \end{aligned}\right\} \Rightarrow \begin{aligned} C_1 &= 1 \\ C_2 &= 0 \end{aligned} \Rightarrow \phi_k(x) = J_k(x).$$