§1.1 Floating Point Number System

$$\boxed{f} = 2^{-d}\left[\sum_{k=0}^{d-1} b_{d-k}\cdot 2^{k}\right]$$

$d=3$

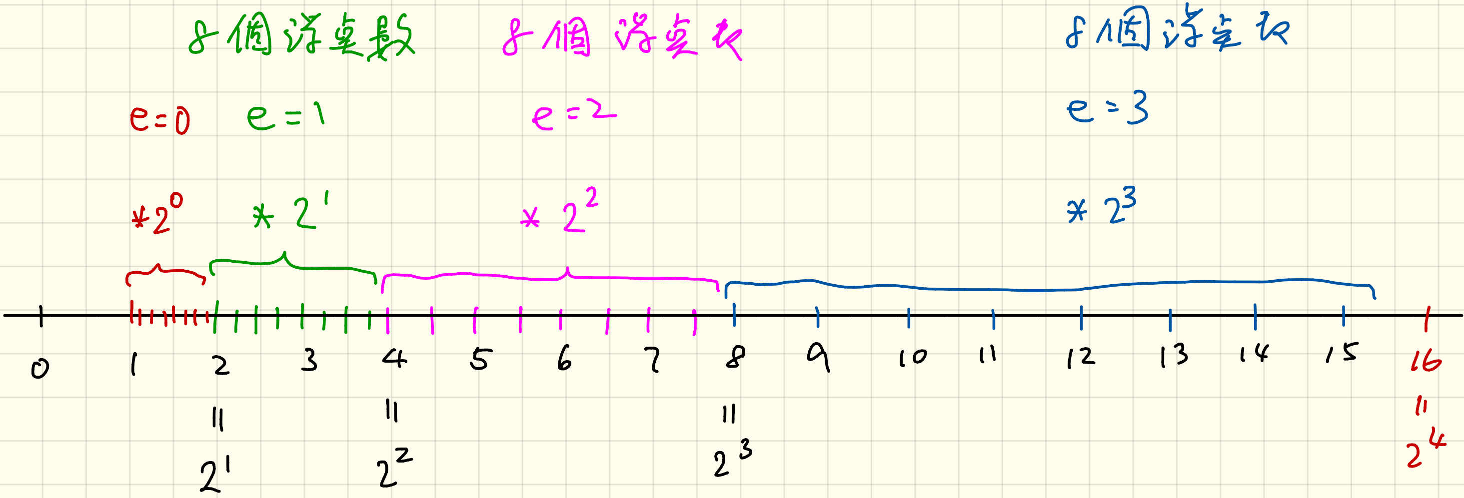
$$= 2^{-3}\left[\sum_{k=0}^{2} b_{3-k}\cdot 2^{k}\right]$$

$$= 2^{-3}\left[\underbrace{b_1}_{0/1} 2^2 + \underbrace{b_2}_{0/1} 2^1 + \underbrace{b_3}_{0/1} 2^0\right]$$

$$\pm(1+\boxed{f})\cdot 2^{e} = \pm\left[1.\boxed{b_1 \quad b_2 \quad b_3}\right] * 2^{e}$$

$z=0$	1.	0	0	0
$z=1$	1.	0	0	1
$z=2$	1.	0	1	0
$\vdots$			$\vdots$	
$z=7$	1.	1	1	1

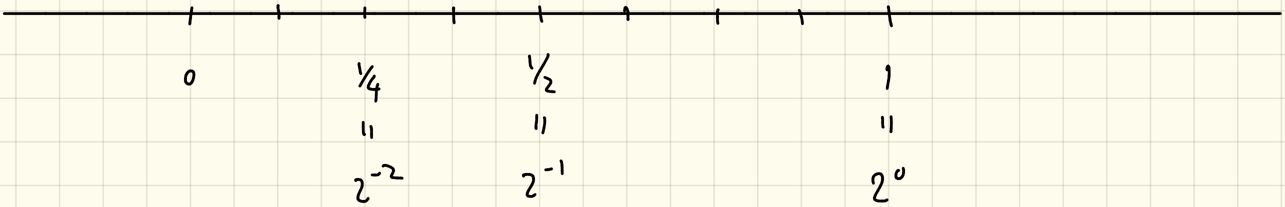
8 個 z



$$\frac{(2^2-2^1)}{8} \xrightarrow{*2} \frac{2^3-2^2}{8} \xrightarrow{*2} \frac{2^4-2^3}{8}$$

General case

$$\frac{2^{e+1}-2^e}{2^d} = \frac{2^e(2-1)}{2^d} = 2^{e-d}$$




```
>> eps

>> 1 + eps            >> 1 - eps

>> 1+ eps/2           >> 1- eps/2

>> 1+ eps/10
```

	Answer
>> eps	
>> (1+ eps/2) - 1	0
>> 1+ (eps/2 - 1)	1.1102e-16

>> 1+ eps

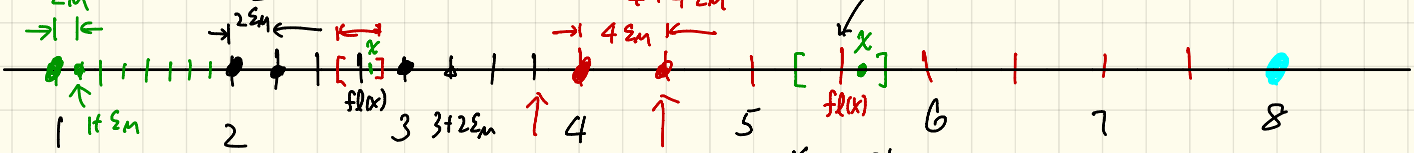
>> 1+ eps/2

>> 1+ eps/10

>> 1- eps

>> 1- eps/2

$$\frac{|(1+\varepsilon_M)-1|}{1} = \varepsilon_M \qquad \frac{|(2+2\varepsilon_M)-2|}{2} = \varepsilon_M \qquad \frac{|(4+4\varepsilon_M)-4|}{(4)} = \varepsilon_M$$



$$\frac{|(3+2\varepsilon_M)-3|}{3}$$

2^{e-d} $\qquad 2^e/2^d$

2^e

$2^{-d} = \varepsilon_M$ machine precision

浮點數表示的絕對誤差

$$\boxed{|fl(x)-x|} \le \frac{1}{2}\left[2^{e-d}\right] = 2^{e-d-1}$$

設 $x>0$, $fl(x)>x$

$$fl(x) - x \le \frac{1}{2}\varepsilon_M x$$

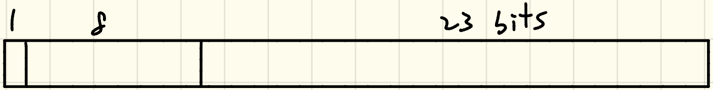
$$fl(x) \le x\left(1+\frac{1}{2}\varepsilon_M\right) = x(1+\varepsilon), \quad \varepsilon \le \frac{1}{2}\varepsilon_M$$

浮點數表示的相對誤差

$$\boxed{\frac{|fl(x)-x|}{|x|}} \le \frac{2^{e-d-1}}{2^e} = 2^{-d-1} = \frac{1}{2}\left(2^{-d}\right) = \frac{1}{2}\varepsilon_M$$

$x \in [2^e, 2^{e+1})$

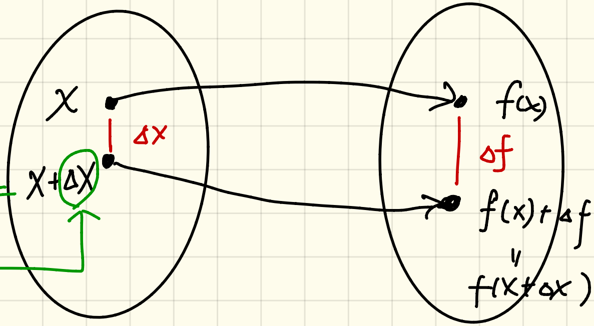
IEEE Single Precision



§ 1.2 Condition Number

In floating number

$$fl(x) = x(1+\varepsilon) = x + \varepsilon x$$



Kappa

$$\kappa = \left|\frac{\Delta f}{\Delta x}\right| = \left|\frac{f(x+\Delta x) - f(x)}{\Delta x}\right| = |f'(x)|$$

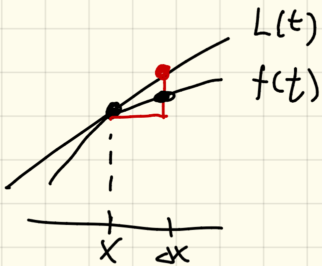
as $\Delta x \to 0$

$$\frac{\text{(relative) change in output}}{\text{(relative) change in input}}$$

$\kappa\ f(x)$ 與 (1) data (x) (2) problem (f) 有關

$$\kappa = \frac{\left|\frac{\Delta f}{f}\right|}{\left|\frac{\Delta x}{x}\right|} = \left|\frac{\frac{\Delta x f'(x)}{f(x)}}{\frac{\Delta x}{x}}\right| = \left|\frac{x f'(x)}{f(x)}\right|$$

as $\Delta x \to 0$



$$\frac{\Delta f}{f(x)} = \frac{f(x+\Delta x) - f(x)}{f(x)} = \frac{\Delta x\, f'(x)}{f(x)} + \cancel{O(\Delta x^2)}$$

Taylor's

as $\Delta x \to 0$

幾乎沒有 rounding error
關注源自於問題 f 和
資料 x 的影響

In floating point system, we have

$$\tilde{x} = x(1+\varepsilon) \qquad \text{where} \quad |\varepsilon| \le \frac{1}{2}\varepsilon_{mach}$$
$$= x + \varepsilon x$$

$$K = \lim_{\varepsilon \to 0} \frac{\left|\dfrac{f(x) - f(\tilde{x})}{f(x)}\right|}{\left|\dfrac{x - \tilde{x}}{x}\right|} = \lim_{\varepsilon \to 0} \frac{\left|\dfrac{f(x) - f(x(1+\varepsilon))}{f(x)}\right|}{\left|\dfrac{\varepsilon x}{x}\right|} = \lim_{\varepsilon \to 0} \frac{|f(x) - f(x(1+\varepsilon))|}{|\varepsilon f(x)|}$$

$$= \lim_{\varepsilon \to 0} \left| \underbrace{\frac{f(x) - f(\varepsilon x)}{\varepsilon x}}_{f'(x)} \cdot \frac{x}{f(x)} \right| = \left| \frac{x f'(x)}{f(x)} \right|$$

for small ε (not $\to 0$)

$$K \approx \left| \frac{f(x) - f(x+\varepsilon x)}{\varepsilon f(x)} \right|$$

$$\Rightarrow \left| \frac{f(x+\varepsilon x) - f(x)}{f(x)} \right| \approx K \cdot |\varepsilon|$$

When the data x is perturbed by a small amount, we expect the relative change to be magnified by a factor of K.
Large condition number suggests when errors cannot be expected to remain comparable in the size of the round off error.

Example 1:

$f(x) = x - c$

兩個相近數相減，$\kappa \gg 1$

$$\Rightarrow \kappa_f(x) = \left|\frac{x f'(x)}{f(x)}\right| = \left|\frac{x}{x-c}\right| \quad \left[\frac{x}{\ll 1} \to x \approx c\right] \uparrow 大$$

$$pi - \frac{355}{113} \qquad \begin{array}{r} 7 \qquad\qquad 9 \qquad\qquad 7 \\ 3.141592|653589793| \\ -\ 3.141592|920353983| \\ \hline -\ 0.000000|\underbrace{266764189}_{有意義}|\underbrace{4049666}_{亂數!} \end{array} = -2.667\cdots \times 10^{-7}$$

$x - c$

$$\kappa = \left|\frac{3.141592653589793}{-2.667\cdots \times 10^{-7}}\right| \approx 10^{7}$$

$\log_{10}\kappa \approx 7$

Substractive cancelling: 7位有效位數消失！

Try $pi - \frac{103683}{32989}$

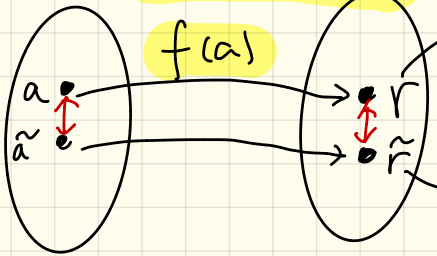
Example 2 :

small perturbation in a

$a + \Delta a = \tilde{a}$

求 ax^2+bx+c 的根

$f(a)$



r: ax^2+bx+c 的根

$\tilde{r}$: $\tilde{a}x^2+bx+c$ 的根

$$\kappa = \left|\frac{x f'(x)}{f(x)}\right| = \left|\frac{a f'(a)}{f(a)}\right| = \left|\frac{a\frac{df(a)}{da}}{f(a)}\right| = \left|\frac{\frac{a r^{\cancel{2}}}{p'(r)}}{\cancel{f(a)}}\right| = \left|\frac{a r}{p'(r)}\right|$$

Find $\boxed{r = f(a)}$ s.t. $p(r) = \underbrace{ar^2 + br + c}$ $= 0$

implicit differentiation

$$\Rightarrow \frac{d}{da}\left[\quad\right] = \frac{d}{da}(0)$$

$$\Rightarrow r^2 + a \cdot 2r \cdot \frac{df}{da} + b\frac{df}{da} = 0$$

$$\Rightarrow \frac{df}{da} = \frac{-r^2}{2ar+b} = \frac{-r^2}{p'(r)}$$

$$= \frac{dp(r)}{dr}$$

$$\kappa = \left|\frac{ar}{p'(r)}\right| = \left|\frac{ar}{2ar+b}\right| = \left|\frac{ar}{\pm\sqrt{b^2-4ac}}\right| = \left|\frac{r}{\boxed{r_1-r_2}}\right|$$

$$r = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

$$\Rightarrow \quad 2ar + b = \pm\sqrt{b^2-4ac}$$

$$r_1 - r_2 = \frac{-b+\sqrt{\ }}{2a} - \frac{-b-\sqrt{\ }}{2a} = \frac{\sqrt{b^2-4ac}}{a}$$

$$\Rightarrow \sqrt{b^2-4ac} = a(r_1-r_2)$$

Note: $p'(r) = 0 \iff r$ is a double root. $\kappa \to \infty$

$\quad = 2ar+b = \pm\sqrt{b^2-4ac}$

$p'(r) \approx 0 \iff$ $\boxed{r \text{ is nearly a double root}}$

$\kappa \gg 1$ $\quad 1.01x^2 - 6x + 9$ 條件數大

誤差大!

§ 1.2 Stability

When error in the result of an algorithm exceeds what conditioning can explain, the algorithm is unstable.

The sensitivity of an algorithm depends on the condition numbers of all of its steps.

Example 1.3.3

$$P(x) = (x - 10^6)(x - 10^{-6}) = \underbrace{1}_{a}\, x^2 - \underbrace{(10^6 + 10^{-6})}_{b}\, x + \underbrace{1}_{c}$$

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

computed $\boxed{x_1} = 1000000$ $\qquad$ $\boxed{x_2} = 1.00000761449337 0 \times 10^{-6}$

rel. err. ϕ $\qquad$ 10^{-5}

Calculation	Result	κ
$u_1 = b^2$	$1.000000000002000 \times 10^{12}$	2
$u_2 = u_1 - 4$	$9.999999999980000 \times 10^{11}$	$\lvert u_1\rvert/\lvert u_2\rvert \approx 1.00$
$u_3 = \sqrt{u_2}$	999999.9999990000	1/2
$u_4 = u_3 - b$	2000000	$\lvert u_3\rvert/\lvert u_4\rvert \approx 0.500$
$u_5 = u_4/2$	1000000	1

$u_4 = -u_3 - b,$

$= -999999.9999990000$
$+ 1000000.000001$

两相近数相减 ⊜ 10^{-11}

Subtractive cancelling

Big cond. #

$$\kappa = \frac{|u_3|}{|u_4|} \approx \frac{0.5}{10^{-11}} \approx 10^{11}$$

$$\kappa \cdot \varepsilon \approx 10^{11} \times 10^{-16} \approx 10^{-5}$$

Function	Condition number
$f(x) = x + c$	$\kappa_f(x) = \frac{\lvert x\rvert}{\lvert x+c\rvert}$
$f(x) = cx$	$\kappa_f(x) = 1$
$f(x) = x^p$	$\kappa_f(x) = \lvert p\rvert$
$f(x) = e^x$	$\kappa_f(x) = \lvert x\rvert$
$f(x) = \sin(x)$	$\kappa_f(x) = \lvert x\cot(x)\rvert$
$f(x) = \cos(x)$	$\kappa_f(x) = \lvert x\tan(x)\rvert$
$f(x) = \log(x)$	$\kappa_f(x) = 1/\lvert\log(x)\rvert$

2相近數相減，會有問題

Solution:

$$x_2 = \frac{-b - \sqrt{b^2-4ac}}{2a} = \frac{(-b-\sqrt{\ \ })(-b+\sqrt{\ \ })}{2a(-b+\sqrt{\ \ })}$$

$$= \frac{\cancel{b^2} - (\cancel{b^2} - \overset{2}{\cancel{4}}\cancel{a}c)}{\cancel{2}\cancel{a}(-b+\sqrt{\ \ })} = \frac{2c}{-b+\sqrt{\ \ }} = \frac{\cancel{2}c}{\cancel{2}ax_1}$$

2相近數相加，沒問題

$$\boxed{x_1 = \frac{-b+\sqrt{\ \ }}{2a}}$$

$$= 1.00000000000000 \times 10^{-6}$$

由於計算造成的影響

How close to the true answer ($f(x)$) is your answer ($\tilde{f}(x)$)?

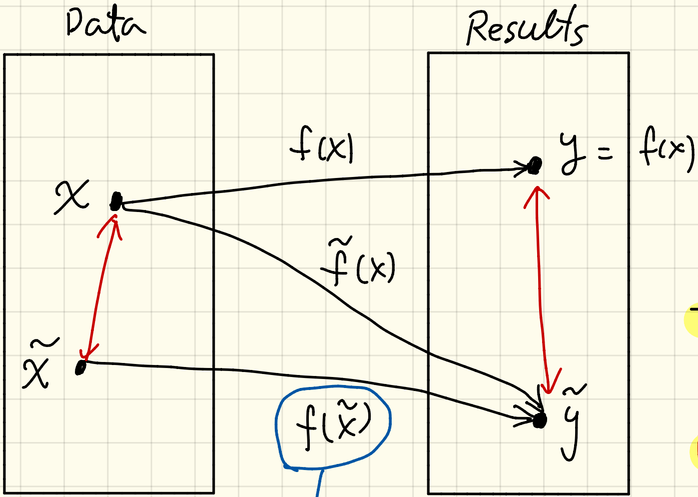
如果 $\exists\, \tilde{x}$, s.t.

$$f(\tilde{x}) = \tilde{f}(x)$$

Backward error

$$\frac{|\tilde{x} - x|}{|x|}$$

small B. error $\Rightarrow$



x 沒有 round-off error

是"計算" $\tilde{f}$ 造成的結果的誤差

$$\frac{|\tilde{f}(x) - f(x)|}{|f(x)|}$$

Forward error

the algorithm ($\tilde{f}(x)$) gives $\boxed{\text{the correct answer to nearly the right problem } \tilde{x}}$

How close to the true question (x) is the question you answered ($\tilde{x}$)?

$r = [-2\ \ -1\ \ \boxed{1\ \ 1}\ \ 3\ \ 6]'$ ⇒ true roots

$p = \text{poly}(r) = [1\ \ -8\ \ 6\ \ 44\ \ -43\ \ -36\ \ 36]$

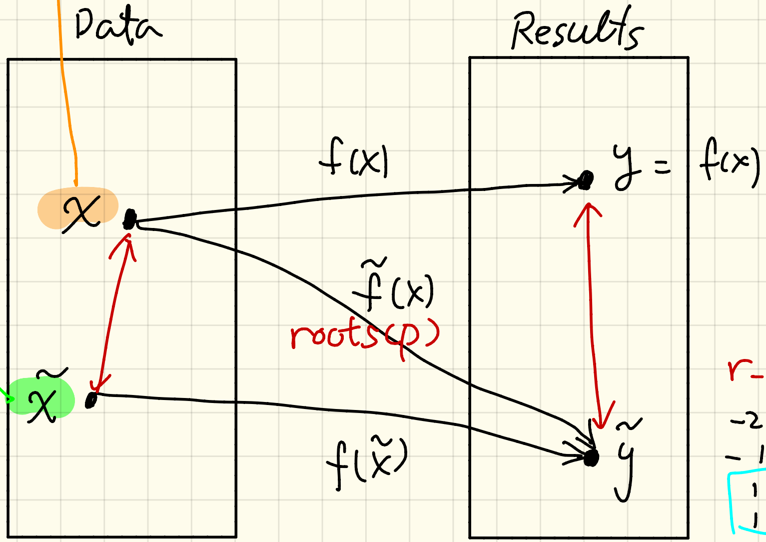
Backward Error

(p_computed - p)./p

-1×10^{-15}
-4×10^{-15}
-4×10^{-16}
-6×10^{-16}
-2×10^{-15}
-1×10^{-15}

$$\text{Forward error} = \frac{|r - r\_computed|}{|r|} =$$

-6×10^{-16}
-9×10^{-16}
9×10^{-5}
4×10^{-9}
1×10^{-16}
1×10^{-15}



r

-2
-1
1
1
3
6

nearly double roots.
ill-conditioned

poly(r_computed)
||
p_computed

1. ...
-8. ...
6. ...
44. ...
-43. ...
-36. ...
36. ...

r_computed

-2, ...
-1. ...
1. ...
1. ...
3, ...
6. ...

If an algorithm always produces small backward errors, then it is stable.