Catastrophic loss of precision

$$\boxed{\text{Never compute a small number as the difference between two large numbers}}$$

US population (thousands)

x:

Feb	311,189
Mar	311,356

$\rightarrow$ difference $311{,}356 - 311{,}189 = 167$ (reasonable)

in our floating point scheme:

$$fl(311{,}189) = 3.1119 \cdot 10^5$$
$$fl(311{,}356) = 3.1136 \cdot 10^5$$

$$\rightsquigarrow \begin{array}{r} 3.1136 \cdot 10^5 \\ - \; 3.1119 \cdot 10^5 \\ \hline 0.0017 \cdot 10^5 \end{array} = 1.7000 \cdot 10^2$$

$\rightsquigarrow$ relative error of $\approx 2 \cdot 10^{-2}$ but EPSREL $\approx 10^{-5}$???

$\rightarrow$ almost all digits used up to store the "large part" of the numbers", which cancel out.

<u>generally:</u> ~~If~~ for N digits of precision, when the first M digits of two numbers agree, expect only <u>$N - M$</u> digits of precision in the difference.

<u>$M \approx N \longrightarrow$ catastrophic precision loss</u>

Ex: Finite differences

$$f'_{FD}(x) = \frac{f(x+h) - f(x)}{h}$$

$f(x) = x$, want $f'(1) = 1$

a) $h = \frac{2}{3}$, $fl(h) = 0.66667$

$f(x+h) = 1.6667$

$f(x) = 1.0000$

$\rightarrow \quad \frac{f(x+h) - f(x)}{h} = \frac{0.66670}{0.666\underbrace{67}}$

differ in 4th digit.

$\rightarrow \quad f'_{FD}(1) \approx 1 + O(10^{-4})$

) $h = \frac{2}{30}$, $fl(h) = 0.0\underbrace{66667}$

still 5 sign. digits!

<u>but</u> $f(x+h) = 1.0667$

$f(x) = 1.0000$

$$\frac{f(x+h) - f(x)}{h} = \frac{0.066700}{\underbrace{0.066667}_{\text{differ in 3rd digit!}}}$$

$\rightarrow \quad f'_{FD}(1) = 1 + \mathcal{O}(10^{-3})$

→ decreasing h by factor 10 made error <u>worse</u> by factor 10!

→ lost precision because much of mantissa was used to store 1.0000 (the large number)

How to avoid floating point precision loss?

$\to$ Rearrange things <u>creatively</u> to avoid the nasty differences!

<u>Ex:</u> $f(x, \Delta) = \sqrt{x+\Delta} - \sqrt{x}$, $\Delta \ll x$.

for instance: $x = 900$, $\Delta = 4 \cdot 10^{-3}$

$$\to \underbrace{30.0000}_{\text{6 wasted ~~precision~~ digits}}6667 - 30.0000\,0000$$

in our scheme: $fl(f(900, 4 \cdot 10^{-3})) = 0$

here we can use:

$$(\sqrt{x+\Delta} - \sqrt{x})(\sqrt{x+\Delta} + \sqrt{x}) = x + \Delta - x$$
$$= \Delta$$

$$\Rightarrow \sqrt{x+\Delta} - \sqrt{x} = \frac{\Delta}{\sqrt{x+\Delta} + \sqrt{x}}$$

$$\rightsquigarrow \frac{4 \cdot 10^{-3}}{30.0000\,6667 - 30.0000\,0000} \simeq \frac{4 \cdot 10^{-3}}{6.0000 \cdot 10^{2}} \simeq 6.6667 \cdot 10^{-5}$$

here we got all the precision we want!

$$\Rightarrow \quad \boxed{\vec{f}'' = \frac{1}{h^2} A \vec{f}} \qquad \text{(matrix multiplication)}$$

an <u>invert</u> : $\vec{f} = h^2 A^{-1} \vec{f}''$ (finite difference method for PDEs)

<u>ontrivial b.c.s</u>

$$f(x_0) = f_0 \quad , \quad f(x_{N+1}) = f_{N+1}$$

$$\rightarrow \quad f_1'' \simeq \frac{1}{h^2}\left(f_0 - 2 f_1 + f_2 \right)$$

$$f_N'' \simeq \frac{1}{h^2}\left(f_{N-1} - 2 f_N + f_{N+1} \right)$$

$$\Rightarrow \quad \underbrace{\begin{pmatrix} f_1'' - \frac{1}{h^2} f_0 \\ f_2'' \\ \vdots \\ f_{N-1}'' \\ f_N'' - \frac{1}{h^2} f_{N+1} \end{pmatrix}}_{= \vec{f}'' - \frac{1}{h^2} \vec{\Delta}} = h^2 A \vec{f}$$

$\vec{\Delta}$ ~ sparse vector of b.c.s

$$\Leftrightarrow \quad \vec{f} = h^2 A^{-1}\left(\vec{f}'' - \frac{1}{h^2} \vec{\Delta} \right)$$