Condition number and limits of accuracy $\boxed{\text{cond}}$

Problems

1. find root r of $f(x)$
2. given $\vec{u}, \vec{v}$ compute $\vec{u}^T \vec{v}$
3. given $A\vec{x} = \vec{b}$, compute solution $\vec{x}$

Model problem: evaluate $y = h(t)$ (scalar function)

1. "$r = h(f)$"
2. $\vec{u}^T \vec{v} = h(\vec{u}, \vec{v})$
3. $\vec{x} = h(A, \vec{b})$

Important to understand inputs/outputs

Each problem has a condition # which measures how sensitively output depends on input. In practice, problems w/ large cond #s are hard to solve numerically.

Model problem

$$y = h(t)$$

$$y + \delta y = h(t + \delta t)$$

so $\delta y = \frac{\delta y}{\delta t}\delta t = \frac{h(t+\delta t) - h(t)}{\delta t}\delta t \simeq h'(t)\,\delta t$

$$\hat{\kappa}(t) = \max_{\delta t} \left| \frac{\delta y}{\delta t} \right| = |h'(t)|$$

if $h(t)$ differentiable

absolute condition #.

if $h(t)$ differentiable

$$\kappa(t) = \max_{\delta t} \frac{|\delta y / y|}{|\delta t / t|} = \left| \frac{t\,h'(t)}{h(t)} \right|$$

relative condition #

BACKWARD SIDE FORWARD SIDE

input $\longrightarrow$ $\boxed{\text{solution process}}$ $\longrightarrow$ output

$t \longrightarrow \boxed{\text{process to evaluate } h(t)} \longrightarrow y$

We speak of <u>forward</u> and <u>backward</u> errors.

We want $y = h(t)$, we get $y_A = h_A(t)$

Accuracy

$$\left|\frac{y_A - y}{y}\right| = O(\varepsilon_{mach})$$

$$\left|\frac{h_A(t) - h(t)}{h(t)}\right|$$

Backward stability

$$h_A(t) = h(t + \delta t) \quad \text{for small } \delta t.$$

$$|\delta t / t| = O(\varepsilon_{mach})$$

Example $h(t) = at$ $\quad a \in \mathbb{R}$ fixed

$$h_A(t) = fl(a) \odot fl(t)$$

$$= a(1+\alpha) \odot t(1+\beta)$$

$\alpha, \beta = O(\varepsilon_{mach})$

$\left|\frac{f(t) - t}{t}\right| \leq \frac{1}{2}\varepsilon_{mach}$

$$= at(1+\alpha)(1+\beta)(1+\gamma)$$

$$\left|\frac{y_A - y}{y}\right| = \left|\frac{at(1+\alpha)(1+\beta)(1+\gamma) - at}{at}\right|$$

accurate

$$= \left|(1+\alpha)(1+\beta)(1+\gamma) - 1\right| = O(\varepsilon_{mach})$$

Also backward stable.

$$h_A(t) = at(1+\alpha)(1+\beta)(1+\gamma)$$

$$= a\Big[\underbrace{t(1+\alpha)(1+\beta)(1+\gamma) - t}_{\delta t} + t\Big]$$

$$= a(t+\delta t)$$

$$= h(t+\delta t)$$

"exact answer for perturbed input"

where $\delta t = t\underbrace{[(1+\alpha)(1+\beta)(1+\gamma) - 1]}_{O(\varepsilon_{mach})}$

so $|\delta t / t| = O(\varepsilon_{mach})$

Backward stable algorithms are accurate

$$\left|\frac{h_A(t)-h(t)}{h(t)}\right| = \left|\frac{h(t+\delta t)-h(t)}{h(t)}\right|$$

$$= |\delta y/y|$$

$$= \underbrace{\frac{|\delta y/y|}{|\delta t/t|}}_{\le \kappa(t)}\overbrace{\left|\frac{\delta t}{t}\right|}^{O(\varepsilon_{mach})}$$

$$= \kappa(t)\, O(\varepsilon_{mach})$$

root finding

Given $f(x)$ find r s.t. $f(r)=0$. Solve $f(x)=0$.

$$r = h(f)$$

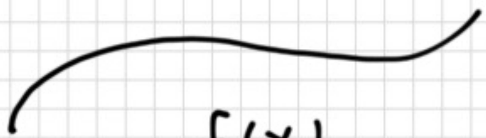
$$r_A = h_A(f) \stackrel{?}{=} h(f+\delta f)$$

always possible

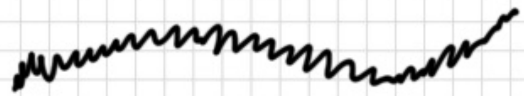
$$f(x)+\delta f(x) = f(x) - \underbrace{f(r_A)}_{\delta f}$$

Better model

$$f(x) + \delta f(x) = f(x) + \varepsilon g(x)$$



$f(x)$



$f(x) + \varepsilon_{mach}\, g(x)$

$$f(r_A) + \varepsilon_{mach}\, g(r_A) = 0 .$$

so we find root ~~of~~ $f + \delta f$

Sensitivity formula for roots

forward error (relative) $\left| \frac{r_A - r}{r} \right|$

backward error $|f(r_A)|$

To compute forward error consider

$$f(r_A) + \varepsilon g(r_A) = 0$$

$$f(r+\delta r) + \varepsilon g(r+\delta r) = 0$$

$$\underbrace{\cancel{f(r)}}_{0} + f'(r)\,\delta r + \varepsilon g(r) + \underbrace{\cancel{\varepsilon g'(r)\,\delta r}}_{\text{small}} + \underbrace{\cancel{O(\delta r^2)}}_{\text{small}} = 0$$

$$\delta r \simeq -\frac{\varepsilon g(r)}{f'(r)}$$

$$\left|\frac{r_A - r}{r}\right| \simeq \varepsilon \left|\frac{g(r)}{r f'(r)}\right|$$

$$\text{error magnification} = \frac{\text{relative forward error}}{\text{backward error}}$$

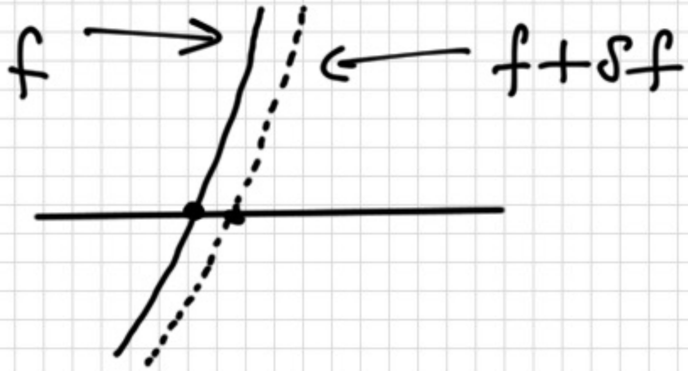
$$f(r_A) + \varepsilon g(r_A) = 0$$

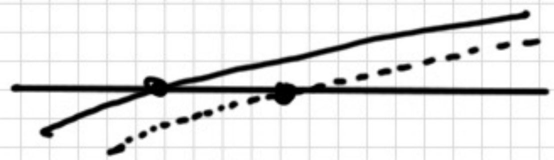
so $|f(r_A)| = \varepsilon |g(r_A)|$

$$\simeq \varepsilon |g(r)|$$

$$= \left| \frac{1}{r f'(r)} \right|$$

result makes sense w/ pictures





Ex $f(x) = (x-1)(x-2)(x-3)$

$\delta f(x) = \varepsilon x^4$ where $\varepsilon = 10^{-7}$

formula ok for any ε but $\varepsilon_{mach} \simeq 2\times 10^{-16}$

What happens to $r=2$.

$$f'(x) = \frac{d}{dx}\left[(x-1)(x-2)(x-3)\right]$$

$$f'(2) = -1 \; ; \; g(x) = x^4, \; g(2) = 16$$

$$\delta r = -\frac{\varepsilon g(2)}{f'(2)} = 1.6 \times 10^{-6}$$

$$r_A = r + \delta r \simeq 2.0000016$$

error mag

$$= \left|\frac{1}{2f'(2)}\right| = \frac{1}{2}$$

Problem: find root r of $f(x) = 0$

$$r = h(f)$$ "function point of view"

output ↑ ↑ input

Actually compute $r_A = h_A(f)$

↑ alogorithm to evaluate h (solve EQN)

(relative) forward error

$$\left| \frac{r_A - r}{r} \right|$$ ← what we want small

backward error $|f(r_A)|$ ← what we can test

<u>Ex</u> $f(x) = (x - 4/3)^3$

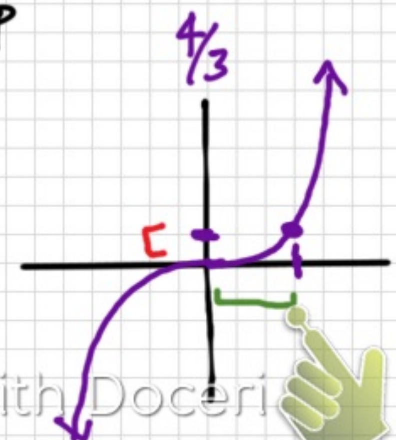
say $r_A = \frac{4}{3} + 10^{-p}$

$$\left(\begin{array}{l} \text{so } r_A = 4/3 + 0.1 \\ \quad\;\; = 4/3 + 0.01 \\ \quad\;\; \text{etc} \end{array} \right)$$

Backward error $|f(r_A)| = 10^{-3p}$

Forward error (relative) $\dfrac{|10^{-p}|}{|4/3|} = \dfrac{3}{4}\, 10^{-p}$

error mag $= \frac{3}{4}\, 10^{2p}$ $\left(\begin{array}{l} \text{so } \frac{3}{4}\, 100 \\ 3/4\; 10{,}000 \end{array} \right)$



Numerically

$$\underbrace{x^3 - 4x^2 + \tfrac{16}{3}x - \tfrac{64}{27}}$$

$$[r_A \;\; k] = \text{bisection}\big(@(x)(x-4/3)^3, \underset{a}{1}, \underset{b}{2}, \underset{tol}{1e\text{-}8}\big)$$

$$= 1.3333333335581686$$

$$\text{abs}((r_A - r)/r) = 1.8626\text{e}{-9}$$

$$|(r_A - 4/3)^3| = 1.5318\text{e}{-26}$$

$$\text{error mag} = 1.2160\text{e}{+17}$$

Numerically

$[r_A, k] =$ bisection(@sin, 3pi/4, 3pi/2, 1e-8)

$\frac{3\pi}{4}$ $\frac{3\pi}{2}$ tol

$= 3.14159265066395 6$

abs((r_A - pi)/pi) = 9.3132e-10

abs(sin(r_A)) = 2.9258e-9

error mag = 0.3183

The approximate root $r_A = h_A(f)$

$$= h(\underbrace{f + \delta f}_{})$$

result from two sources ① round-off errors in function evaluations ② details and roundoff errors in algorithm. → Perturbation of function

$$\underbrace{f + \delta f_{①}}_{\tilde{f}} + \delta f_{②}$$

Can't really see ①. Backward error $|f(r_A)|$

will be nonzero due to ② really $|\tilde{f}(r_A)|$