Midterm 1 review

Wednesday, October 8, 2025 7:24 AM

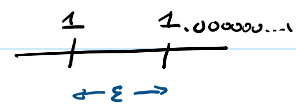
Floating point numbers

for double precision

important concepts: overflow (if $|x| > 10^{308}$ or so)
underflow (if $|x| < 10^{-308}$ or so) } fairly infrequent

roundoff   machine epsilon, ε_m

two different definitions:



① distance from 1 to next floating pt. number
about $2.22 \cdot 10^{-16}$ in double precision

② rounding error at 1
so half of ↑, i.e. $1.11 \cdot 10^{-16}$

All that really matters is $\varepsilon \approx 10^{-16}$

When you represent a real number (π or $1/3$ ...) as a float,
there's rounding error. The size of this error is relative

Ex: $x = \pi$, $fl(x)$ is floating pt. representation

$$|\pi - fl(\pi)| \approx 10^{-16}$$

Ex: $x = 1000 + \pi$,

$$|x - fl(x)| \approx 10^{-13}$$ since relative error is $$\frac{|x - fl(x)|}{|x|} \approx \varepsilon_m$$

Ex $x = 10^{-30}$

$$|x - fl(x)| \approx 10^{-46}$$

Don't confuse roundoff error (always present)
with underflow (fairly rare)

Condition Number

K ("Kappa"), high value = math problem is fundamentally difficult in the sense that it's sensitive to perturbations

"predicting the weather 2 weeks from now is ill-conditioned"

For numerical algorithms, we expect our inputs to have a fixed amount of relative error (why? because floating pt numbers have relative error as we just discussed)

so we look at the relative condition number $K_f(x) := \left| \frac{x}{f(x)} \cdot f'(x) \right|$

Stability

An algo is unstable if any single step has conditioning much worse than the condition number of the problem (not a precise, mathematical definition)

Ex: $f(x) = x$, $K_f(x) = \left| \frac{x}{x} \cdot 1 \right| = 1$

Algo 1: return x. Stable ✓.

Algo 2:
$y = x - 1$ $\longrightarrow$ $K_y(x) = \left| \frac{x}{x-1} \cdot 1 \right|$

if $x \approx 1$ (i.e. $x = 1 + \varepsilon$) then $K_y(1+\varepsilon) = O(1/\varepsilon)$. Bad: UNSTABLE

$z = y + 1$ $\longrightarrow$ $K_z(y) = \left| \frac{y}{y+1} \cdot 1 \right|$

return z

if $y \approx -1$, unstable.

Common mistake: evaluating at x. Evaluate at y

working backwards, that means if $x \approx 0$ is bad.

e.g. $x = \varepsilon$, $y = \varepsilon - 1$,

$y + 1 = \varepsilon$ so $K_z = O(1/\varepsilon)$

CONCLUSION

> Algo 2 is unstable in the regime $x \approx 0$ or $x \approx 1$. Anywhere else it's stable.

(... and it's silly everywhere...)

Big-O

make sure to distinguish $x \to 0$ vs $x \to \infty$

As $x \to 0$, $\frac{1+x}{x} = O(1/x)$

As $x \to \infty$, $\frac{1+x}{x} = O(1)$