MA5233 Computational Mathematics

Lecture 2: Error Analysis

Simon Etter

NUS National University of Singapore

2019/2020

Error Analysis

Floating-point numbers

- `+,-,*,/,sqrt` are all approximate.
- From the IEEE standard:
  "[...] every operation shall be performed as if it first produced an intermediate result correct to infinite precision and with unbounded range, and then rounded that result [...]"
- This only refers to single operations. Errors can still accumulate in longer calculations.

Key to understanding error propagation

- Conditioning
- Stability

Error Analysis

Conditioning
A function $f(x)$ is called *well-conditioned* if small relative perturbations in input x lead to small relative perturbations in output $f(x)$.

Condition number

$$\kappa(f,x) := \lim_{\Delta x \to 0} \frac{|f(x+\Delta x) - f(x)|}{|f(x)|} \frac{|x|}{|\Delta x|} = \frac{|f'(x)|}{|f(x)|} |x|$$

Note
Conditioning is a property of the mathematical problem.
Conditioning does not depend on the method used to solve the problem.

Example
Predicting the outcome of a die role is ill-conditioned:
small perturbations in how you role the die may change the outcome.

Error Analysis

Forward and backward error

Let $\tilde{f}(x)$ be a numerical approximation to $f(x)$.

- Relative forward error: $\frac{|\tilde{f}(x)-f(x)|}{|f(x)|}$.
- Relative backward error: $\frac{|\tilde{x}-x|}{|x|}$ where $\tilde{x}$ such that $f(\tilde{x}) = \tilde{f}(x)$.

Forward and backward stability

$\tilde{f}(x)$ is called *forward/backward stable* if the corresponding error is "small" (usually $\mathcal{O}(\texttt{eps})$).

Examples

- IEEE standard guarantees that `+`,`-`,`*`,`/`,`sqrt` are forward stable.
- `sqrt(x)` is backward stable:
  - forward stability guarantees $\texttt{sqrt(x)} = \sqrt{x}\,(1+\varepsilon)$ with $|\varepsilon| \ll 1$;
  - hence `sqrt(x)` is the correct solution for

$$\tilde{x} = x\,(1+\varepsilon)^2 = x\,(1+2\varepsilon) + \mathcal{O}(\varepsilon^2).$$

Error Analysis

Fundamental theorem of error analysis

$$\text{forward error} \approx \text{condition number} \times \text{backward error}$$

Proof.

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

Error Analysis

Composition theorem

If $\tilde{f}(x)$, $\tilde{g}(x)$ are forward-stable approximations of well-conditioned functions $f(x)$, $g(x)$, then $\tilde{f}\big(\tilde{g}(x)\big)$ is forward stable.

By previous theorem, forward may be replaced with backward stability.

Proof.

$$\begin{aligned}
\tilde{f}\big(\tilde{g}(x)\big) &= \tilde{f}\big(g(x)\,(1+\varepsilon)\big) && \text{stability of } \tilde{g}(x) \\
&= f\big(g(x)\,(1+\varepsilon)\big)\,(1+\varepsilon) && \text{stability of } \tilde{f}(x) \\
&= f\big(g(x)\big)\,(1+2\varepsilon) && \text{conditioning of } f(x)
\end{aligned}$$

ε is a generic constant which represents numbers on the order of machine precision. Different occurrences of ε need not have the same value, and $\varepsilon^2 = 0$.

Error Analysis

Summary

- Floating-point arithmetic necessarily involves rounding errors.
- Error analysis is based on conditioning and backward stability.
- Backward error allows us to compare rounding error against other sources of errors, e.g. measurement or previous computations.
- Condition number quantifies error amplification.
- Condition number is independent of algorithm.

Error Analysis

Vector norm ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$)

Map $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ satisfying the following.

- Absolute homogeneity: $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{K}$.
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.
- Nonnegativity: $\|x\| \geq 0$ and $\|x\| = 0$ iff $x = 0$.

Important vector norms

Definitions

1-norm: $\|x\|_1 := \sum_{k=1}^n |x_k|$

2-norm: $\|x\|_2 := \sqrt{\sum_{k=1}^n |x_k|^2}$

Inf-norm: $\|x\|_\infty := \max_k |x_k|$

Relations

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \, \|x\|_2$$

$$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \, \|x\|_\infty$$

$$\|x\|_\infty \leq \|x\|_1 \leq n \, \|x\|_\infty$$

Norm equivalence theorem

For any two norms $\|\cdot\|_a$, $\|\cdot\|_b$ on a finite-dimensional vector space V, there exists a constant C such that

$$\frac{1}{C} \|x\|_a \leq \|x\|_b \leq C \, \|x\|_a \qquad \forall x \in V.$$

Error Analysis

Matrix norm ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$)
Vector norm on $\mathbb{K}^{n \times n}$ which additionally satisfies the following.

- Submultiplicativity: $\|AB\| \leq \|A\| \, \|B\|$.

Induced matrix norm
Given vector norm $\| \cdot \|$, define induced matrix norm through

$$\|A\| := \max_{\|x\|=1} \|Ax\|.$$

Matrix norms $\| \cdot \|_p$ will always refer to norms induced by vector norms $\| \cdot \|_p$.

Frobenius norm

$$\|A\|_F := \sqrt{\sum_{i,j=1}^{n} |A_{ij}|^2}$$

Error Analysis

Theorem

$$\|A\|_1 = \max_{j \in \{1,\ldots,n\}} \sum_{i=1}^{n} |A_{ij}| \qquad \|A\|_2 = \max_{k \in \{1,\ldots,n\}} \sigma_k(A)$$

$$\|A\|_\infty = \max_{i \in \{1,\ldots,n\}} \sum_{j=1}^{n} |A_{ij}| \qquad \|A\|_F = \sqrt{\sum_{k=1}^{n} \sigma_k(A)^2}$$

$\sigma_k(A)$ denote the singular values of A.

Error Analysis

Condition number of matrix

$$\kappa(A) := \|A\| \, \|A^{-1}\|$$

Motivation.

- Consider the function $f(x) := Ax$.
- Its condition number is given by

$$\kappa(f, x) = \frac{\|\nabla f\|}{\|f(x)\|} \|x\| = \frac{\|A\|}{\|Ax\|} \|x\| \leq \|A\| \, \|A^{-1}\|.$$

Why does condition number of Ax depend on A^{-1}?

- Assume A is singular, $x \in \ker(A)$ and $\Delta x \notin \ker(A)$ such that $\|\Delta x\| \ll \|x\|$.
- Then $Ax = 0$ but $A(x + \Delta x) \neq 0$.
- Hence small relative perturbation in x leads to infinitely large relative perturbation in Ax.

Error Analysis

Conditioning of addition

$$\kappa_1\left(+, (x\,y)^T\right) = \frac{\left\|(1\ 1)\right\|_1}{|x+y|} \left\|(x\ y)^T\right\|_1 = \frac{|x|+|y|}{|x+y|}$$

Hence addition is well-conditioned unless $|x+y| \ll |x| + |y|$.

Remarks:

- We used 1-norm, but conclusion holds for all norms due to norm equivalence.
- $\left\|(1\ 1)\right\|_1$ denotes operator 1-norm because $(1\ 1)$ is a row vector.
  Hence, $\left\|(1\ 1)\right\|_1 = \left\|(1\ 1)^T\right\|_\infty = 1$.

Stability of addition

According to IEEE specification, it holds

$$\texttt{x + y} = (x+y)(1+\varepsilon) = x(1+\varepsilon) + y(1+\varepsilon).$$

Hence, `x+y` is exact result for input $\tilde{x} = x(1+\varepsilon)$, $\tilde{y} = y(1+\varepsilon)$.

Error Analysis

Floating-point addition in practice
The ill-conditioning of addition is a real issue!

Example. Set `x = eps()/2`. In FP arithmetic, we then have

```
(1 + 2*x + x^2) - (1 + x)^2 == eps()
```

because `1 + 2x + x^2 -> 1 + eps()` but `1 + x -> 1`.

Error Analysis

Conditioning of multiplication
Assume w.l.o.g. $|x| > |y|$.

$$\kappa_1\big(\times, (x\,y)^T\big) = \frac{\big\|(y\ x)\big\|_1}{|xy|}\,\big\|(x\ y)^T\big\|_1 = \frac{|x|\,(|x|+|y|)}{|xy|} \leq 2\frac{|x|}{|y|}$$

Hence multiplication is well-conditioned unless $|x| \gg |y|$.

Same remarks as on previous slide apply.

Stability of multiplication
According to IEEE specification, it holds

$$\texttt{x * y} = (xy)\,(1+\varepsilon) = x\,\big(y\,(1+\varepsilon)\big).$$

Hence, `x+y` is exact result for input $\tilde{x} = x$, $\tilde{y} = y\,(1+\varepsilon)$.

Error Analysis

Floating-point multiplication in practice

Ill-conditioning of multiplication is almost always harmless.

Example. Consider the computation `1e30 * pi`.

- The previous slide assumed rounding error in `pi` may be $\mathcal{O}\left(\varepsilon\, 10^{30}\right)$.
- However, rounding error in `pi` is $\mathcal{O}(\varepsilon\, \pi)$.

Error Analysis

References and further reading

- N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics (2002), `doi:10.1137/1.9780898718027`