Numerical Methods I: Conditioning, Stability of Algorithms, Computer Respresentation of Numbers

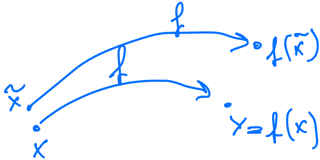
Georg Stadler
Courant Institute, NYU
stadler@cims.nyu.edu

September 14, 2017

Condition numbers

We consider the input-output map

$$x \mapsto f(x)$$



The absolute condition number at x is defined as

$$\kappa_{\text{abs}} = \sup_{\tilde{x} \to x, x \neq \tilde{x}} \frac{\|f(\tilde{x}) - f(x)\|}{\|\tilde{x} - x\|}.$$

The relative condition number at x, for $x \neq 0$, $f(x) \neq 0$, is

$$\kappa_{\text{rel}} = \sup_{\tilde{x} \to x, x \neq \tilde{x}} \frac{\|f(\tilde{x}) - f(x)\| / \|f(x)\|}{\|\tilde{x} - x\| / \|x\|}.$$

- κ small/moderate $\rightarrow$ well conditioned problem
- κ large $\rightarrow$ poorly conditioned problem

Condition numbers and derivatives

Condition numbers are a mathematical concept, they are independent of the algorithm.

If $f : \mathbb{R}^m \to \mathbb{R}^n$ is differentiable, then

$$\kappa_{\text{abs}} = \|Df(x)\|,$$

$$\kappa_{\text{rel}} = \|Df(x)\| \frac{\|x\|}{\|f(x)\|},$$

where $Df(x) \in \mathbb{R}^{m \times n}$ is the Jacobian, and the choice of norms in $\mathbb{R}^m$ and $\mathbb{R}^n$ influences the matrix norm.

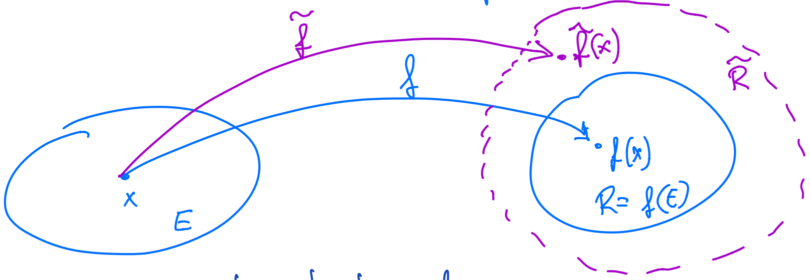
- Condition number of addition/subtraction (numerical cancellation!)
- Solution of $Ax = b$ for perturbations of b
- Solution of $Ax = b$ for perturbations of A

Stability of algorithms

Another source of error is the algorithm, so the question is: **Is $\tilde{f}(x)$ a good approximation to $f(x)$?**

Forward stability:

$\tilde{f}$ is an approximation of f



forward stability tries to bound

$\| f(x) - \tilde{f}(x) \| \lesssim \kappa \| x - \tilde{x} \|$ "forward-stable" if

stability

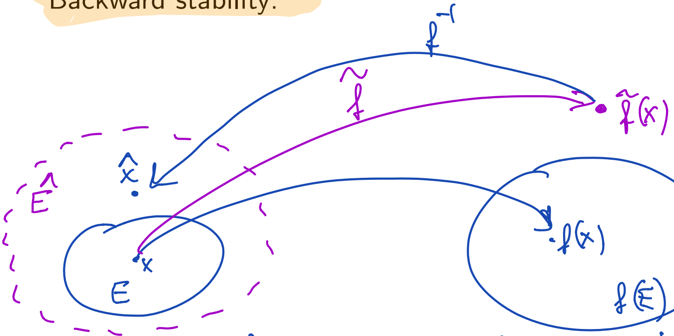
$\tilde{R}$ is not much larger than R

issue: Condition number will play a role

Stability of algorithms

Another source of error is the algorithm, so the question is: Is $\tilde{f}(x)$ a good approximation to $f(x)$?

Backward stability:



algorithm error is mapped back and interpreted as input error. We estimate $\|x - \hat{x}\| \leq \dots$

backward stable if $\hat{E}$ is not much larger than E.

Stability of algorithms

Basic idea: Order basic elementary operations **flops** in the algorithm such that rounding errors have small influence.

- **flops**: floating point operations, i.e., $+, -, \times, /$.
- Roundoff is unavoidable due to finite precission.

A priori versus a posteriori analysis

A priori analysis is performed before a specific solution is computed, i.e., estimates do not depend on a specific numerically computed solution.

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

$\|y - \tilde{y}\| \leq \ldots$ before you've done any computation

A posteriori analysis bounds the error for a specific numerical solution $\hat{x}$ (computed with a specific numerical method), and uses, e.g., residuals for the a posteriori analysis.

estimate $\|y - \tilde{y}\| \leq C(\tilde{y})$ after computed $\tilde{y}$, the numerical solution

$\rightsquigarrow$ better estimates

Notation and other useful concepts

Relative errors:

$$\frac{\|x - x_n\|}{\|x\|} \text{ or } \frac{\|x - x_n\|}{\|x_n\|}$$

Absolute error:

$$\|x - x_n\|$$

- Used for theoretical arguments
- In numerical practice: exact solution is not available, so these errors must be approximated.

Notation and other useful concepts

Speed of convergence

Let $x_n \to x$ in a normed space $X, \|\cdot\|$ for $n \to \infty$.

- Linear convergence:

$$\|x - x_{n+1}\| \leq C\|x - x_n\| \text{ with } 0 \leq C < 1.$$

- Quadratic convergence (only meaningful once $\|x - x_n\| < 1$):

$$\|x - x_{n+1}\| \leq C\|x - x_n\|^2 \text{ with } C > 0.$$

- Superlinear convergence:

$$\|x - x_{n+1}\| \leq c_n\|x - x_n\| \text{ with } c_n \geq 0, \text{ and } c_n \to 0 \text{ for } n \to \infty.$$

- Sublinear convergence:

$$\|x - x_{n+1}\| \leq c_n\|x - x_n\| \text{ with } c_n \geq 0, \text{ and } c_n \to 1 \text{ for } n \to \infty.$$

Notation and other useful concepts

Landau symbols

Let f_n, g_n be sequences in $\mathbb{R}$. Then, for $n \to \infty$:

$$f_n = O(g_n) \quad \Leftrightarrow \quad \exists C > 0, n_0 > 0 : |f_n| \leq C|g_n|,$$
$$f_n = o(g_n) \quad \Leftrightarrow \quad \forall \epsilon \exists n_0 > 0 : |f_n| \leq \epsilon |g_n|$$

Let $f(\cdot), g(\cdot)$ be functions that map to $\mathbb{R}$. Then, for $x \to x_0$:

$$f(x) = O(g(x)) \quad \Leftrightarrow \quad \exists C > 0, U(x_0) : \forall x \in U(x_0) : |f(x)| \leq C|g(x)|,$$
$$f(x) = o(g(x)) \quad \Leftrightarrow \quad \forall \epsilon \exists U(x_0) : \forall x \in U(x_0) : |f(x)| \leq \epsilon |g(x)|$$

$\frac{1}{n^2} = O\left(\frac{1}{n} + \frac{7}{n^2}\right)$ true because $\left|\frac{1}{n^2}\right| \leq C \left|\frac{1}{n} + \frac{7}{n^2}\right|$ as $n \to \infty$