Chapter 1
Machine Arithmetic and Related Matters

Songting Luo

Department of Mathematics
Iowa State University

MATH 561 Numerical Analysis

Condition of a Problem

Consider a problem as map $\mathbf{f}: \mathbf{R}^m \rightarrow \mathbf{R}^n, \ \mathbf{y} = \mathbf{f}(\mathbf{x})$.

How sensitive is $\mathbf{f}$ to small perturbation of $\mathbf{x}$??????

The condition of a problem

- a problem is *well-conditioned if small errors in the data (input) produce small errors in the solution (output).*
- *a problem is* ill-conditioned if small errors in the data (input) may produce large errors in the solution (output).

condition number ...

Condition Number of a Problem

Condition Numbers of **f** at **x**

- $m = 1, n = 1$. $\frac{\Delta y}{y} / \frac{\Delta x}{x} \approx \frac{x f'(x)}{f(x)}$; (relative) condition number of f at x:

$$(cond\ f)(x) = |\frac{x f'(x)}{f(x)}|.$$

- In general. (relative) condition number of **f** at **x**:

$$(cond\ \mathbf{f})(\mathbf{x}) = ||\Gamma(\mathbf{x})||,\ \Gamma(\mathbf{x}) = [\gamma_{\nu\mu}(\mathbf{x})],$$

where $\gamma_{\nu\mu}(\mathbf{x}) = (cond_{\nu\mu}\ \mathbf{f})(\mathbf{x}) = |\frac{x_\mu \frac{\partial f_\nu}{\partial x_\mu}}{f_\nu(\mathbf{x})}|$.

condition number depends on the matrix norm used.

Condition Number of an Algorithm

Consider an algorithm A solving the problem $\mathbf{f}$ as a map

$$\mathbf{f}_A: \ \mathbf{R}^m(t,s) \rightarrow \mathbf{R}^n(t,s), \ \mathbf{y}_A = \mathbf{f}_A(\mathbf{x}).$$

Assume that

$$\forall \mathbf{x} \in \mathbf{R}^m(t,s), \exists \mathbf{x}_A \in \mathbf{R}^m \ s.t. \ \mathbf{f}_A(\mathbf{x}) = \mathbf{f}(\mathbf{x}_A).$$

That is, the computed solution $\mathbf{y}_A$ corresponding to input $\mathbf{x}$ is the exact solution for some different input $\mathbf{x}_A$;
Hope: $\mathbf{x}_A$ is close to $\mathbf{x}$!

Condition Number of A at $\mathbf{x}$

$$(cond\ A)(\mathbf{x}) = \inf_{\mathbf{x}_A} \frac{||\mathbf{x}_A - \mathbf{x}||}{||\mathbf{x}||}/eps.$$

for all $\mathbf{x}_A$ s.t. $\mathbf{y}_A = f(\mathbf{x}_A)$.

Stability

Stability

An algorithm $\mathbf{f}_A$ for a problem $\mathbf{f}$ is said to be *stable if for each* $\mathbf{x}$,

$$\frac{||\mathbf{f}_A(\mathbf{x}) - \mathbf{f}(\mathbf{x}_A)||}{||\mathbf{f}(\mathbf{x}_A)||} = O(eps),$$

for some $\mathbf{x}_A$ *with*

$$\frac{||\mathbf{x}_A - \mathbf{x}||}{||\mathbf{x}||} = O(eps).$$

- An algorithm or numerical process is called stable if small changes in the input only produce small changes in the output.
- An algorithm or numerical process is called unstable if large changes in the output are produced

Some Comments

- Well-/Ill-Conditioned refers to the problem; Stable/Unstable refers to an algorithm or numerical process.
- If the problem is well-conditioned then there is a stable way to solve it.
- If the problem is ill-conditioned then there is no reliable way to solve it in a stable way.
- Mixing roundoff error with an unstable process is a recipe for disaster.
- With exact arithmetic (no roundoff-error), stability is not a concern.

Computer Solution of a Prolem; Overall Error

Given the problem $\mathbf{f}$, and an algorithm $\mathbf{f}_A$ to solve the problem.
$\mathbf{y} = \mathbf{f}(\mathbf{x})$, $\mathbf{x}^* = fl(\mathbf{x})$, $\mathbf{y}_A^* = \mathbf{f}_A(\mathbf{x}^*)$. $\mathbf{y}^* = \mathbf{f}(\mathbf{x}^*)$.
And by assumption:

$$\mathbf{f}_A(\mathbf{x}^*) = \mathbf{f}(\mathbf{x}_A^*), \quad \frac{||\mathbf{x}_A^* - \mathbf{x}^*||}{||\mathbf{x}^*||} = (cond\ A)(\mathbf{x}^*) \cdot eps.$$

Estimate the total error:

$$\frac{||\mathbf{y}_A^* - \mathbf{y}||}{||\mathbf{y}||}.$$

- Total error:

$$\frac{||\mathbf{y}_A^* - \mathbf{y}||}{||\mathbf{y}||} \leqslant \frac{||\mathbf{y}_A^* - \mathbf{y}^*||}{||\mathbf{y}||} + \frac{||\mathbf{y}^* - \mathbf{y}||}{||\mathbf{y}||} \approx \frac{||\mathbf{y}_A^* - \mathbf{y}^*||}{||\mathbf{y}^*||} + \frac{||\mathbf{y}^* - \mathbf{y}||}{||\mathbf{y}||}$$

where $||\mathbf{y}|| \approx ||\mathbf{y}^*||$.

Overall Error

- 1st term of RHS

$$\frac{||\mathbf{y}_A^* - \mathbf{y}^*||}{||\mathbf{y}^*||} = \frac{||\mathbf{f}_A(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}^*)||}{||\mathbf{f}(\mathbf{x}^*)||} = \frac{||\mathbf{f}(\mathbf{x}_A^*) - \mathbf{f}(\mathbf{x}^*)||}{||\mathbf{f}(\mathbf{x}^*)||}$$

$$\leqslant (cond\ \mathbf{f})(\mathbf{x}^*) \cdot \frac{||\mathbf{x}_A^* - \mathbf{x}^*||}{||\mathbf{x}^*||} = (cond\ \mathbf{f})(\mathbf{x}^*) \cdot (cond\ A)(\mathbf{x}^*) \cdot eps.$$

- 2nd term of RHS

$$\frac{||\mathbf{y}^* - \mathbf{y}||}{||\mathbf{y}||} = \frac{||\mathbf{f}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x})||}{||\mathbf{f}(\mathbf{x})||} \leqslant (cond\ \mathbf{f})(\mathbf{x}) \cdot \frac{||\mathbf{x}^* - \mathbf{x}||}{||\mathbf{x}||} = (cond\ \mathbf{f})(\mathbf{x}) \cdot \epsilon.$$

  ϵ is the roundoff error.

- Thus

$$\frac{||\mathbf{y}_A^* - \mathbf{y}||}{||\mathbf{y}||} \leqslant (cond\ \mathbf{f})(\mathbf{x})\{\epsilon + (cond\ A)(\mathbf{x}^*) \cdot eps\}.$$

  where $(cond\ \mathbf{f})(\mathbf{x}^*) \approx (cond\ \mathbf{f})(\mathbf{x})$.

Taylor Polynomials

Theorem (Taylor's Theorem)

Suppose $f \in C^n[a,\ b]$, *that* $f^{(n+1)}$ *exists on* $[a,\ b]$, *and* $x_0 \in [a,\ b]$. *For every* $x_0 \in [a,\ b]$, *there exists a number* $\xi(x)$ *between* x_0 *and* x *with*

$$f(x) = P_n(x) + R_n(x),$$

where

$$\begin{aligned} P_n(x) =& f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 \\ & + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n = \sum_{k=0} n \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \end{aligned}$$

and

$$R_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}(x - x_0)^{n+1}.$$

Brief Introduction to Linear Algebra 1

- Matrix-vector product $\mathbf{b} = \mathbf{A}\mathbf{x}$

$$b_i = \sum_{j=1}^{n} a_{ij} x_j.$$

- The map $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$ is linear, for any $\mathbf{x}, \mathbf{y}$ and α,

$$\mathbf{A}(\mathbf{x} + \mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y}$$
$$\mathbf{A}(\alpha\mathbf{x}) = \alpha\mathbf{A}\mathbf{x}.$$

- Conversely, every linear map can be expressed as multiplication by a matrix.

Brief Introduction to Linear Algebra 2

Vector spaces.

- Vector space spanned by a set of vectors is composed of linear combinations of these vectors
- If S_1 and S_2 are two subspaces, then $S_1 \cap S_2$ is a subspace. So is $S_1 + S_2$.
- Two subspaces S_1 and S_2 are complementary subspaces of each other if $S_1 + S_2 = \mathbf{C}^m$ and $S_1 \cap S_2 = \{\mathbf{0}\}$.

Brief Introduction to Linear Algebra 3

Definition

Range and Null Space

- The range of a matrix $\mathbf{A}$, written as range($\mathbf{A}$), is the set of vectors that can be expressed as $\mathbf{Ax}$ for some $\mathbf{x}$.
- The null space of $\mathbf{A}$, written as null($\mathbf{A}$), is the set of vectors $\mathbf{x}$ that satisfy $\mathbf{Ax} = 0$.

Theorem (Rank-Nullity Theorem)

Given $\mathbf{A} \in \mathbf{C}^{m \times n}$,

$$dim(null(\mathbf{A})) + rank(\mathbf{A}) = n.$$

Brief Introduction to Linear Algebra 4

- Transpose $\mathbf{A}^T$; Hermitian conjugate or transpose conjugate $\mathbf{A}^*$ ($\mathbf{A}^H$).
- Symmetric $\mathbf{A} = \mathbf{A}^T$; Hermitian $\mathbf{A} = \mathbf{A}^*$; skew-symmetric $\mathbf{A} = -\mathbf{A}^T$; skew-Hermitian $\mathbf{A} = -\mathbf{A}^*$.
- Diagonal matrix, Upper (Lower) triangular matrix, etc..
- Nonsingular or invertible matrix; Inverse matrix $\mathbf{A}^{-1}$. Unitary matrix $\mathbf{A}^* = \mathbf{A}^{-1}$.

Theorem

The following conditions are equivalent: $\mathbf{A} \in \mathbf{C}^{m\times m}$,

(a) $\mathbf{A}$*has an inverse* $\mathbf{A}^{-1}$,

(b) rank($\mathbf{A}$*) is* m,

(c) range($\mathbf{A}$*) is* $\mathbf{C}^m$,

(d) null ($\mathbf{A}$*) is* $\{0\}$,

(e) 0 *is not an eigenvalue of* $\mathbf{A}$,

(f) 0 *is not a singular value of* $\mathbf{A}$,

(g) $det(\mathbf{A}) \neq 0$.

Brief Introduction to Linear Algebra 5

Definition

Orthogonal Vectors

- A pair of vectors are orthogonal if $\mathbf{x}^*\mathbf{y} = 0$.
- Two sets of vectors X and Y are orthogonal if every $\mathbf{x} \in X$ is orthogonal to every $\mathbf{y} \in Y$.
- A set of nonzero vectors S is orthogonal if they are pairwise orthogonal. They are orthonormal if it is orthogonal and in addition each vector has unit Euclidean length.

Theorem

The vectors in an orthogonal set S are linearly independent.

Brief Introduction to Linear Algebra 6

Definition

Vector Norms A norm is a function $\|\cdot\| : \mathbf{C}^m \mapsto \mathbf{R}$ that assigns a real-valued length to each vector. It must satisfy the following conditions:

- $\|\mathbf{x}\| \geqslant 0$, and $\|\mathbf{x}\| = 0$ only if $\mathbf{x} = 0$.
- $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.
- $\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\|$

Definition (p-norms)

p-norm of a vector $\mathbf{x} \in \mathbf{C}^m$:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p}$$

for $1 \leqslant p \leqslant \infty$.

Brief Introduction to Linear Algebra 7

Definition (Induced Matrix Norm)

Given vector norms $\| \cdot \|_{(n)}$ and $\| \cdot \|_{(m)}$ on domain and range of $\mathbf{A} \in \mathbf{C}^{m \times n}$, respectively, the induced matrix norm $\|\mathbf{A}\|_{(m;n)}$ is the smallest number $C \in \mathbf{R}$ for which the following inequality holds for all $\mathbf{x} \in \mathbf{C}^n$:

$$\|\mathbf{A}\mathbf{x}\|_{(m)} \leqslant C\|\mathbf{x}\|_{(n)}.$$

- Induced Matrix norm:

$$\|\mathbf{A}\|_{(m,n)} = \sup_{\mathbf{x} \in \mathbf{C}^n, \mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_{(m)}}{\|\mathbf{x}\|_{(n)}} = \sup_{\mathbf{x} \in \mathbf{C}^n, \|\mathbf{x}\|_{(n)}=1} \|\mathbf{A}\mathbf{x}\|_{(m)}.$$

Brief Introduction to Linear Algebra 7

Some special cases:

- 1-norm: "maximum column sum"

$$\|\mathbf{A}\|_1 = \max_{1 \leqslant j \leqslant n} \|\mathbf{a}_j\|_1.$$

- ∞-norm: "maximum row sum"

$$\|\mathbf{A}\|_1 = \max_{1 \leqslant i \leqslant m} \|\mathbf{a}_i^*\|_1.$$

- 2-norm: largest singular value of **A**.

Brief Introduction to Linear Algebra 8

General Matrix Norms

One can view $m \times n$ matrices as mn-dimensional vectors and obtain general matrix norms which satisfies the three conditions.

Frobenius Norm

One useful norm is Frobenius norm

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\sum_{j=1}^{n}\|\mathbf{a}_j\|_2^2},$$

i.e., 2-norm of mn-vector. Furthermore

$$\|\mathbf{A}\|_F = \sqrt{tr(\mathbf{A}^*\mathbf{A})}.$$

Brief Introduction to Linear Algebra 9

Singular Value Decomposition (SVD)

- SVD is
$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^*$$
where $\mathbf{U} \in \mathbf{C}^{m\times m}$ and $\mathbf{V} \in \mathbf{C}^{n\times n}$ are unitary and $\Sigma \in \mathbf{R}^{m\times n}$ is diagonal.
- Singular values are diagonal entries of Σ, with entries $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n \geqslant 0$.
- Left singular vectors of $\mathbf{A}$ are column vectors of $\mathbf{U}$.
- Right singular vectors of $\mathbf{A}$ are column vectors of $\mathbf{V}$.
- $\mathbf{A}\mathbf{v}_j = \sigma_j\mathbf{u}_j$ for $1 \leqslant j \leqslant n$.