Numerical Linear Algebra Conspectus

Dimitri Tabatadze

December 22, 2023

Contents

1 Week 1 **1**
1.1 Vector Norms 1
1.2 Functions, Convexity 3
1.3 k-means Clustering 3
1.4 Cramer's rule 4

2 Week 2 **4**
2.1 Matrix properties 4
2.2 Spectral decomposition of matrices 5
2.3 Matrix norms 5
2.4 Matrix series 6

3 Week 3 **6**
3.1 Matrix power series 6

4 Week 4 **7**
4.1 Condition number of a matrix 7
4.2 Rgiht perturbation theorem 9
4.3 Left perturbation theorem 11

5 Week 5 **11**
5.1 General perturbation theorem 11

6 Week 7 **13**
6.1 Round-off errors 13

1 Week 1

1.1 Vector Norms

Definition. For $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ to be a vector norm, it should satisfy the following properties:

1. $\|x\| \geq 0, \|x\| = 0$ if and only if $x = 0$ (positivity)
2. $\|\alpha x\| = |\alpha| \cdot \|x\|$ (homogenity)
3. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality or subadditivity)

Properties.

- Inverse Triangle Inequality

$$\|x - y\| \geq |\|x\| - \|y\||$$

- Hoelder Inequality

$$\left| \sum_{i=1}^{n} x_i y_i \right| \leq \|x\|_p \|y\|_q, \frac{1}{p} + \frac{1}{q} = 1, p, q \geq 1$$

- Cauchy-Schwartz Inequality

$$\left| \sum_{i=1}^{n} x_i y_i \right| \leq \|x\|_2 \|y\|_2$$

- Minkowsky Inequality

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p$$

- $$\lim_{p \to \infty} \|x\| p = \|x\|_\infty$$

Examples.

- $\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|$ (the 1-norm)
- $\|x\|_2 = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{\frac{1}{2}}$ (the 2-norm or Euclidean norm)
- $\|x\|_\infty = \max\{|x_1|, |x_2|, \ldots, |x_n|\}$ (the max-norm or infinity norm)
- $\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$ (the p-norm or Hoelder norm)
- $\|x\| = |x_1| + |x_2 - x_1| + \cdots + |x_n - x_{n-1}|$
- $\|x\|_A = \|A^{\frac{1}{2}} x\|_2, A \in \mathbb{R}^{n \times n}, A = A^T > 0$

Equivalence of Vector Norms. For some α and β, there exist constants C_m and C_M such that

$$C_m \|x\|_\alpha \leq \|x\|_\beta \leq C_M \|x\|_\alpha$$

holds true for all vectors x.

Examples.

- $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \cdot \|x\|_2$
- $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n} \cdot \|x\|_\infty$
- $\|x\|_\infty \leq \|x\|_1 \leq n \cdot \|x\|_\infty$
- $\|x\|_\infty \leq \|x\|_p \leq n^{\frac{1}{p}} \cdot \|x\|_\infty, p \geq 1$

Theorem. In $\mathbb{R}^n allvectornormsareequivalent.$

1.2 Functions, Convexity

Definition. f is called convex if

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y), 0 \leq \alpha \leq 1$$

Properties.

- Any vector norm ($\mathbb{R}^n \to \mathbb{R}$) is a uniformly continuous function.
- Any vector norm ($\mathbb{R}^n \to \mathbb{R}$) is a convex function.
- $f(x) = \|x\|_p^p, p > 1, x \in \mathbb{R}^n$ is a strictly convex function.
- Balls $\{\|x\| \leq 1\}$ are convex for any vector norm ($\mathbb{R}^n \to \mathbb{R}$).

Examples.

- The unit ball in the Euclidean is round: $\left\{x \in \mathbb{R}^2, (x_1^2 + x_2^2)^{\frac{1}{2}} \leq 1\right\}$.
- The unit ball in 1-norm is not round: $\left\{x \in \mathbb{R}^2, |x_1| + |x_2| \leq 1\right\}$.

1.3 k-means Clustering

k-means algorithm loop. Repeat until convergrence:

1. **Partition.** for each vector $c_i = (s_1, s_2, \ldots, s_n), i = 1, 2, \ldots, m,$ sasign nearest representative $z_j = (s_1, s_2, \ldots, s_n), j = 1, 2, \ldots, k.$
2. **Update.** set z_j to the mean in the group $j = 1, 2, \ldots, k.$

Definition of "nearest representative". For a vector c_i the nearest representative is z_j if

$$\|c_i - z_j\| = \min\{\|c_i - z_1\|, \|c_i - z_2\|, \ldots, \|c_i - z_k\|\}$$

Comparing diferent partitions. The following formula can be used for calculating the cost of the partition

$$\sum_{i=1}^{n} \min_{j=1,\dots,k} \|c_i - z_j\|$$

Facts.

- The algorithm doesn't always converge, it's heuristic.
- The final partition is affected by the set of initial representatives and the norm.
- there are various approaches for updating representatives.

1.4 Cramer's rule

Given a system of linear equations $Ax = b$ where $A = [a_1\ a_2\ \dots\ a_n] \in \mathbb{R}^{n\times n}, \det(A) \neq 0$ and $x = (x_1, x_2, \dots, x_n)^T, b = (b_1, b_2, \dots, b_n)^T$ then

$$x_i = \frac{\det(A_i)}{\det(A)} = \frac{\det([a_1, \dots, a_{i-1}, b, a_{i+1}, \dots, a_n])}{\det(A)}$$

where A_i is the matrix formed by replacing the i-th column of A by the column vector b.

2 Week 2

2.1 Matrix properties

Determinant $A, B \in \mathbb{R}^{n\times n}$

Notation:

$$\det(A) \equiv |A|$$

- $\det(\alpha A) = \alpha^n \det(A)$.
- $\det(A) = \det(A^T)$.
- $\det(AB) = \det(A) \cdot \det(B)$.
- $\det(A) = 0 \impliedby$ any two rows or any two columns are co-linear.
- Interchanging two columns or two rows in a matrix changes sign of its determinant.
- The product of n eigenvalues of A is its determinant.

Invertable matrices $A, B \in \mathbb{R}^{n\times n}$

- $\det(A) \neq 0$.
- $(AB)^{-1} = B^{-1}A^{-1}$.
- $(A^{-1})^{-1} = A$.
- $(\alpha A)^{-1} = \frac{1}{a}A^{-1}$.
- $(A^T)^{-1} = (A^{-1})^T$.
- All eigenvalues of A are nonzero.
- Columns and rows of A are linearly independent.
- The product of two lower or upper triangular matrices is a lower or upper triangular matrix respectively.
- The diagonal entries of a lower/upper triangular matrix are its eigenvalues.
- The inverse of a lower or upper triangular matrices is a lower or upper triangular matrix respectively.

2.2 Spectral decomposition of matrices

Given that $A \in \mathbb{R}^{n\times n}$ is a matrix with n eigenvectors q_i and respective eigenvalues λ_i, we know that

$$Aq_i = \lambda_i q_i$$

and it's also clear that

$$\begin{aligned} AQ &= Q\Lambda \\ A &= Q\Lambda Q^{-1} \end{aligned}$$

Where Q is a square $n \times n$ matrix whose i-th column is the eigenvector q_i of A, and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues. From this,

2.3 Matrix norms

$\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a norm if

1. $\|A\| \geq 0,\ \|A\| = 0 \iff A = 0$
2. $\|\alpha A\| = |\alpha|\|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$

Examples:

- $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$ — one-norm or maximum column sum norm.

- $\|A\|_2 = \sqrt{\rho(A^TA)}$ — the two-norm or spectral norm.
- $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$ — infinity-norm or maximum row sum norm.
- $\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$ — Frobenius norm (sub-multiplicative).
- $\|A\|_{\max} = \max_{1\leq i\leq m, 1\leq j\leq n} |a_{ij}|$ — maximum norm.

Induced norm For any induced matrix norm

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}$$

the following holds true:

- $\rho(A) \leq \|A\|,\ A \in \mathbb{C}^{n\times n}$
- $\rho(A) = \lim_{k\to\infty} \left(\|A^k\|\right)^{\frac{1}{k}},\ A \in \mathbb{C}^{n\times n}$
- $\|A\| = \inf\{\lambda \in \mathbb{R} : \|Ax\| \leq \lambda\|x\|,\ x \in \mathbb{C}^n\}$

2.4 Matrix series

a sequence of matrices $\{A_k\}_1^\infty$.

- $\lim_{k\to\infty} a_{ij}^{(k)} = b_{ij} \implies \lim_{k\to\infty} A_k = B$
- $\rho(A) < 1,\ A \in \mathbb{R}^{n\times n},\ A_k = A^k \implies \lim_{k\to\infty} A_k = 0$

3 Week 3

3.1 Matrix power series

Suppose R is the radius of convergence of absolutely convergent power series $\sum_{k=0}^{\infty} a_k x^k$ then matrix power series $\sum_{k=0}^{\infty} a_k A^k$ converges if $\|A\| < R$ or $\rho(A) < R$.

Proof. Given $\|A\| < R$ where $\|\cdot\|$ is a norm with sub-multiplicativity:

$$\begin{aligned}
\|S_{n+p} - S_n\| = \|\sum_{k=0}^{n+p} a_k A^k - \sum_{k=0}^{n} a_k A^k\| \quad &= \|\sum_{k=n+1}^{n+p} a_k A^k\| \\
\leq \sum_{k=n+1}^{n+p} |a_k|\|A^k\| \quad &\leq \sum_{k=n+1}^{n+p} |a_k|\|A\|^k \\
\leq \sum_{k=n+1}^{\infty} |a_k|\|A\|^k \quad &\leq \sum_{k=n+1}^{\infty} |a_k| R^k
\end{aligned}$$

□

Proof. $\exists \|\cdot\|,\ \delta > 0,\ R > \rho(A) + \delta > \|A\| \implies$

$$\begin{aligned} \|S_{n+p} - S_n\| = \|\sum_{k=0}^{n+p} a_k A^k - \sum_{k=0}^{n} a_k A^k\| &= \|\sum_{k=n+1}^{n+p} a_k A^k\| \\ \le \sum_{k=n+1}^{n+p} |a_k| \|A^k\| &\le \sum_{k=n+1}^{n+p} |a_k| \|A\|^k \\ \le \sum_{k=n+1}^{\infty} |a_k| \|A\|^k &\le \sum_{k=n+1}^{\infty} |a_k| R^k \end{aligned}$$

$\square$

4 Week 4

4.1 Condition number of a matrix

The condition number $\mathrm{cond}_\alpha(A)$ is defined as

$$\mathrm{cond}_\alpha(A) = \|A\|_\alpha \|A^{-1}\|_\alpha.$$

We use condition number to characterize ill-conditioned and well-conditioned problems.

ill-conditioned system of linear equations Example ($Ax = b$)

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4.001 & 2.002 \\ 1 & 2.002 & 2.004 \end{pmatrix},\ b = \begin{pmatrix} 1 \\ 8.0021 \\ 5.006 \end{pmatrix} \implies x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

This system is ill-conditioned because with a *small* relative perturbation ($\frac{\|b-\tilde{b}\|}{\|b\|} = 1.3975 \cdot 10^{-5}$) relative error is *large* ($\frac{\|x-\tilde{x}\|}{\|x\|} = 1.3461$):

$$\tilde{b} = \begin{pmatrix} 1 \\ 8.0020 \\ 5.0061 \end{pmatrix} \implies \tilde{x} = \begin{pmatrix} 3.0850 \\ -0.0436 \\ 1.0022 \end{pmatrix}.$$

The condition number of an ill-conditioned matrix is *large*. In this example:

- $\mathrm{cond}_2(A) \approx 31062.16\ldots$
- $\mathrm{cond}_F(A) \approx 31326.00\ldots$
- $\mathrm{cond}_\infty(A) \approx 48170.06\ldots$
- $\mathrm{cond}_1(A) \approx 48170.06\ldots$

well-conditioned system of linear equations Example ($Ax = b$)

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix},\ b = \begin{pmatrix} 3 \\ 7 \end{pmatrix} \implies x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

This system is ill-conditioned because with a *small* relative perturbation ($\frac{\|b-\tilde{b}\|}{\|b\|} = 1.875 \cdot 10^{-5}$) relative error is also *small* ($\frac{\|x-\tilde{x}\|}{\|x\|} = 10^{-5}$):

$$\tilde{b} = \begin{pmatrix} 3.0001 \\ 7.0001 \end{pmatrix} \implies \tilde{x} = \begin{pmatrix} 0.9999 \\ 1.0001 \end{pmatrix}.$$

The condition number of an ill-conditioned matrix is *small*. In this example:

- $\text{cond}_2(A) \approx 14.93\ldots$
- $\text{cond}_F(A) \approx 15$
- $\text{cond}_\infty(A) \approx 48170.06\ldots$
- $\text{cond}_1(A) \approx 48170.06\ldots$

Porperties

- $\text{cond}(A) \geq 1$ for any induced norm. (via submultiplicativity)
- $\text{cond}(\alpha A) = \text{cond}(A)$ for any norm.
- $\text{cond}(AB) \leq \text{cond}(A)\,\text{cond}(B)$ for any submultiplicative norm.
- $\text{cond}_1(A) = \text{cond}_\infty(A^T)$
- $\text{cond}(A) = \text{cond}(A^{-1})$
- $\text{cond}_2(A) = 1 \iff A^T A = \alpha I,\ \alpha \neq 0$
- $\text{cond}_2(A) = \text{cond}_2(A^T)$
- $\text{cond}_2(A^T A) = (\text{cond}_2(A))^2$

Proof.

$$\begin{aligned}
\mathrm{cond}_2(A^TA) &= \|A^TA\|_2 \cdot \|(A^TA)^{-1}\|_2 \\
&= \sqrt{\rho((A^TA)^T(A^TA))} \cdot \sqrt{\rho(((A^TA)^{-1})^T((A^TA)^{-1}))} \\
&= \sqrt{\rho((A^TA)(A^TA))} \cdot \sqrt{\rho(((A^{-1})^T(A^{-1}))^T((A^{-1})^T(A^{-1})))} \\
&= \sqrt{\rho((A^TA)^2)} \cdot \sqrt{\rho(((A^{-1})^T(A^{-1}))^2)} \\
&= \sqrt{\rho((A^TA)^2)} \cdot \sqrt{\rho(((A^TA)^{-1})^2)} \\
&= \sqrt{\rho(A^TA)^2} \cdot \sqrt{\rho((A^TA)^{-1})^2} \quad \Longleftarrow \ A^TA \text{ is symmetric} \\
&= \left(\sqrt{\rho(A^TA)} \cdot \sqrt{\rho((A^TA)^{-1})}\right)^2 \\
&= \left(\|A\|_2 \cdot \|A^{-1}\|_2\right)^2 \\
&= (\mathrm{cond}_2(A))^2
\end{aligned}$$

□

- $\mathrm{cond}_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \Longleftarrow A$ is a symmetric positive definite matrix with $\lambda_{\max}, \lambda_{\min} \in \mathbb{R}$ eigenvalues of A.
- $\mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(VA) \Longleftarrow U, V$ unitary matrices

J. Hadamard wel posedness

- There exists a solution.
- The solution is unique.
- The solution changes continuously along input.

4.2 Rgiht perturbation theorem

Given an invertible matrix $A \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$ with $Ax = b,\ \delta x = A^{-1}(b + \delta b)$ where $\delta b \in \mathbb{R}^n$

$$\frac{1}{\|A\|\|A^{-1}\|} \cdot \frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}$$
$$\Updownarrow$$
$$\frac{1}{\mathrm{cond}(A)} \cdot \frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}$$

holds true.

Proof. Upper bound:

$$\begin{aligned} & A(x+\delta x) = b+\delta b,\ Ax=b \\ \implies \quad & A\delta x = \delta b \\ \implies \quad & \delta x = A^{-1}\delta b \\ \implies \quad & \|\delta x\| = \|A^{-1}\delta b\| \le \|A^{-1}\| \cdot \|\delta b\| \end{aligned} \tag{A}$$

$$\begin{aligned} & Ax = b \\ \implies \quad & b = Ax \\ \implies \quad & \|b\| = \|Ax\| \le \|A\| \cdot \|x\| \end{aligned} \tag{B}$$

now, if we multiply respective sides of A and B

$$\begin{aligned} & \|\delta x\| \cdot \|b\| \le \|A^{-1}\| \cdot \|\delta b\| \cdot \|A\| \cdot \|x\| \\ \implies \quad & \frac{\|\delta x\|}{\|x\|} \le \|A\| \cdot \|A^{-1}\| \cdot \frac{\|\delta b\|}{\|b\|} \end{aligned}$$

Lower bound:

$$\begin{aligned} & A(x+\delta x) = b+\delta b,\ Ax=b \\ \implies \quad & A\delta x = \delta b \\ \implies \quad & \|\delta b\| = \|A\delta x\| \le \|A\| \cdot \|\delta x\| \end{aligned} \tag{C}$$

$$\begin{aligned} & Ax = b \\ \implies \quad & x = A^{-1}b \\ \implies \quad & \|x\| = \|A^{-1}b\| \le \|A^{-1}\| \cdot \|b\| \end{aligned} \tag{D}$$

now we multiply respective sides of C and D

$$\begin{aligned} & \|\delta b\| \cdot \|x\| \le \|A\| \cdot \|\delta x\| \cdot \|A^{-1}\| \cdot \|b\| \\ \implies \quad & \frac{1}{\|A\| \cdot \|A^{-1}\|} \cdot \frac{\|\delta b\|}{\|b\|} \le \frac{\|\delta x\|}{\|x\|} \end{aligned}$$

$\square$

4.3 Left perturbation theorem

Given an invertible matrix $A \in \mathbb{R}^{n\times n}$ satisfying $\|\delta A\| < \frac{1}{\|A^{-1}\|}$ and $Ax = b$ where $x, b \in \mathbb{R}^n$ with $\delta x = (\delta A + A)^{-1}b - x$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\delta A\|}{1 - \|A^{-1}\| \cdot \|\delta A\|}$$
$$\Updownarrow$$
$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}{1 - \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}$$

holds true.

Proof.

$$\begin{aligned}
& (A + \delta A)(x + \delta x) = b \\
\implies \quad & Ax + A\delta x + \delta Ax + \delta A\delta x = b \\
\implies \quad & A\delta x + \delta Ax + \delta A\delta x = 0 \\
\implies \quad & A\delta x = -(\delta Ax + \delta A\delta x) \\
\implies \quad & A\delta x = -\delta A(x + \delta x) \\
\implies \quad & \delta x = A^{-1}(-\delta A(x + \delta x)) \\
\implies \quad & \|\delta x\| = \|A^{-1}\delta A(x + \delta x)\| \\
& \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|(x + \delta x)\| \\
& \leq \|A^{-1}\| \cdot \|\delta A\| \cdot (\|x\| + \|\delta x\|) \\
\implies \quad & (1 - \|A^{-1}\| \cdot \|\delta A\|)\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta A\| \cdot \|x\| \\
\implies \quad & \frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\delta A\|}{1 - \|A^{-1}\| \cdot \|\delta A\|}
\end{aligned}$$

□

5 Week 5

5.1 General perturbation theorem

Given an invertable matrix $A \in \mathbb{R}^{n\times n}$ and nontrivial $b \in \mathbb{R}^n$ satisfying $Ax = b$, perturbations δA and δb in A and b respectively cause perturbation

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right)$$

To prove general perturbation theorem, we first need to prove some theorems. Let $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ be an induced matrix norm, $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$ a vector norm, $M \in \mathbb{C}^{n\times n}$ a matrix with $\|M\| < 1$

1. $(I-M)^{-1}$ exists.

Proof.

$$\begin{aligned}
\|(I-M)x\| &= \|x - Mx\| \\
&\geq |\|x\| - \|Mx\|| \\
&= \left|\left(1 - \frac{\|Mx\|}{\|x\|}\right)\|x\|\right| \\
\implies \quad 1 - \|M\| &\leq \left|1 - \frac{\|Mx\|}{\|x\|}\right| \\
\implies \quad \|(I-M)x\| &\geq (1-\|M\|)\|x\| \\
\implies \quad \|(I-M)x\| = 0 &\Rightarrow (1-\|M\|)\|x\| = 0 \Rightarrow \|x\| = 0 \\
\implies \quad (I-M)x = 0 &\Rightarrow x = 0
\end{aligned}$$

Since $(I-M)x = 0$ only when $x = 0$, the kernel of $(I-M) = \{0\}$ meaning that it's invertible.

$\square$

2. $\|(I-M)^{-1}\| \leq \frac{1}{1-\|M\|}$.

Proof.

$$\begin{aligned}
1 = \|I\| &= \|(I-M)(I-M)^{-1}\| \\
&= \|(I-M)^{-1} - M(I-M)^{-1}\| \\
&\geq \left|\|(I-M)^{-1}\| - \|M(I-M)^{-1}\|\right| \\
&\geq \left|\|(I-M)^{-1}\| - \|M\| \cdot \|(I-M)^{-1}\|\right| \\
&\geq (1-\|M\|)\|(I-M)^{-1}\| \\
\implies \quad \|(I-M)^{-1}\| &\leq \frac{1}{1-\|M\|}
\end{aligned}$$

$\square$

3. $(I-M)^{-1} = \sum_{k=0}^{\infty} M^k$

Proof.

$$\begin{aligned}
S_j &= \sum_{k=0}^{j} M^k \\
S_j(I-M) &= \sum_{k=0}^{j} M^k - \sum_{k=0}^{j} M^{k+1} \\
&= I - M^{j+1}
\end{aligned}$$

$\square$

Example Given the condition number $\text{cond}(A) = 10^c$, relative perturbation $\frac{\|\delta b\|}{\|b\|} = 10^{-p}$ and the required accuracy 10^{-r}, for which c and p is the system well conditioned?

Solution. By right perturbation theorem,

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \text{cond}(A)\frac{\|\delta b\|}{\|b\|} && \Longrightarrow \\ 10^{-r} &\leq 10^c \cdot 10^{-p} && \Longrightarrow \\ -r &\leq c - p \end{aligned}$$

6 Week 7

6.1 Round-off errors

m-digit arithmetic. A number $(-1)^s \beta^e \sum_{i=1}^{m} d_{-i}\beta^{-i}$ in base $\beta \in \mathbb{N}$ where digits $d_{-i} \in \{0, \dots, \beta\}$ and the exponent $e \in \{e_{\min}, \dots, e_{\max}\}$ is in m-digit arithmetic.

Overflow. If an operation results in a number with exponent $e > e_{\max}$ overflow will be caused.

Underflow. If an operation results in a number with exponent $e < e_{\min}$ underflow will be caused.

Chopping. The m-digit result of chopping $\tilde{x}$ of the k-digit number x, where $k > m$, can be written as such:

$$\tilde{x} = (-1)^s \beta^e \sum_{i=1}^{\tilde{m}} d_{-i}\beta^{-i}$$

Chopping. The m-digit result of rounding $\tilde{x}$ of the k-digit number x, where $k > m$, can be written as such:

$$\tilde{x} = \begin{cases} (-1)^s \beta^e \sum_{i=1}^{\tilde{m}} d_{-i}\beta^{-i}, & \text{if } d_{-(\tilde{m}+1)} < \frac{\beta}{2} \\ (-1)^s \beta^e \left(\sum_{i=1}^{\mathbb{R}} d_{-i}\beta^{-i} + \beta^{-m} \right), & \text{if } d_{-(\tilde{m}+1)} \geq \frac{\beta}{2} \end{cases}$$

Definitions

Kernel of a matrix
: a set of solutions to the equation $AX = 0$.

Inner product
: $a^T b = \sum_{i=1}^{n} a_i b_i,\ a, b \in \mathbb{R}^n$.

Outer product
$ab^T = (b_1 a, b_2 a, \ldots, b_n a) \in \mathbb{R}^{m\times n},\ a \in \mathbb{R}^m, b \in \mathbb{R}^n.$

Determinant, Laplace expansion
$\det(A) = (-1)^{i+1} a_{i1} \det(A_{i1}) + (-1)^{i+2} a_{i2} \det(A_{i2}) + \cdots + (-1)^{i+n} a_{in} \det(A_{in}).$

Spectral radius
$\rho(A) = \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\},\ Ax_i = \lambda_i x,\ x \neq 0,\ A \in \mathbb{R}^{n\times n},\ x \in \mathbb{R}^n.$

Similarity transformation
$A, B \in \mathbb{R}^{n\times n} \implies A$ and $B^{-1}AB$ are similar matrices.

Equivalence of matrix norms
$\exists C_m, C_M > 0,\ C_m\|A\|_\alpha \leq \|A\|_\beta \leq C_M\|A\|_\alpha,\ \forall A \in \mathbb{R}^{m\times n}$. All matrix norms are equivalent.

Sub-multiplicative matrix norm
$\|AB\| \leq \|A\|\|B\|,\ A \in \mathbb{R}^{m\times n},\ B \in \mathbb{R}^{n\times k}.$

Proof. $\|AB\| = \sup_{x\neq 0} \frac{\|ABx\|}{\|x\|} = \sup_{Bx\neq 0} \frac{\|ABx\|}{\|x\|} =$ $\square$

Compatible matrix and vector norms
$\|Ax\|_\alpha \leq \|A\|_\beta \|x\|_\alpha,\ A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n \implies \alpha$ and β are compatible. (Ex. atrix one-norm and vector one-norm are compatible).

Subordinate matrix norm
$\|Ax\|_\alpha \leq \|A\|_\beta \|x\|_\gamma,\ A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n \implies \|\cdot\|_\beta$ is subordinate to vector norms $\|\cdot\|_\alpha$ and $\|\cdot\|_\gamma$.

Induced norm
Matrix norm $\|\cdot\|$ is induced by vector norm $\|\cdot\|_\alpha$

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha} = \sup_{\|x\|_\alpha = 1} \|Ax\|_\alpha.$$

Operator norm

$$\|A\|_{\alpha,\beta} = \sup_{x\neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\beta}.$$

Scalar error
absolute: $|x - \tilde{x}|$, relative: $\frac{|x-\tilde{x}|}{|x|}$ where $\tilde{x} \in \mathbb{R}$ is an approximation of $x \in \mathbb{R}$.

Vector error
absolute: $\|x - \tilde{x}\|$, relative: $\frac{\|x-\tilde{x}\|}{\|x\|}$ where $\tilde{x} \in \mathbb{R}^n$ is an approximation of $x \in \mathbb{R}^n$.

Positive definite matrix
is a matrix M if $x^T M x > 0 \; \forall \; x \in \mathbb{R}^n \setminus \{0\}$.

Positive semi-definite matrix
is a matrix M if $x^T M x \geq 0 \; \forall \; x \in \mathbb{R}^n \setminus \{0\}$.

Hermitian matrix
$a_{ij} = \overline{a_{ji}}$ for all $i, j <= n$ where $A \in \mathbb{C}^{n \times n}$

Unitary matrix
Matrix U is unitary when $U^* U = UU^* = UU^{-1} = I$

Orthogonal matrix
An *unitary* real square matrix $Q \in \mathbb{R}^{n \times n}$ is *orthogonal*. $Q^T = Q^* = Q^{-1}$