MA633L-Numerical Analysis

Lecture 26 : Numerical Linear Algebra - Norms

Panchatcharam Mariappan[1]

[1]Associate Professor
Department of Mathematics and Statistics
IIT Tirupati, Tirupati

March 10, 2025

Numerical Weather Prediction

Numerical Weather Prediction

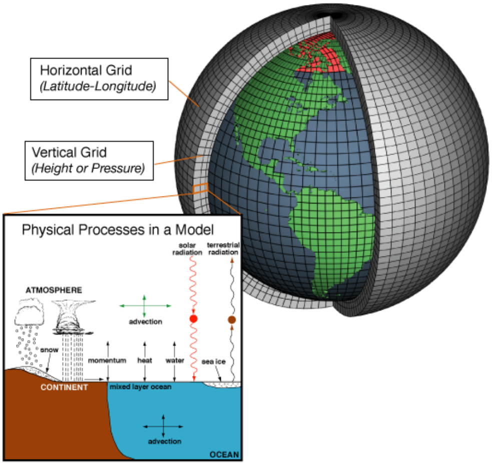


Figure 1: Atmospheric Model Grids, Source: Wikipedia

Numerical Weather Prediction

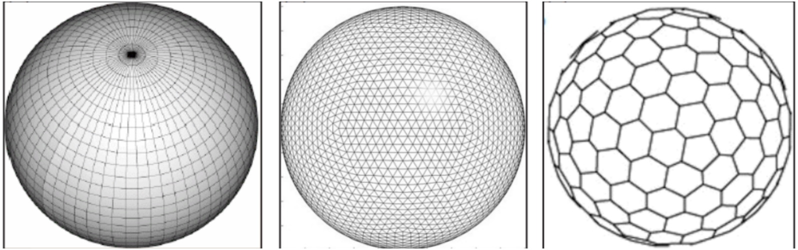


Figure 2: Atmospheric Model Grids Other Pattern, Source: Intech Open

Numerical Weather Prediction

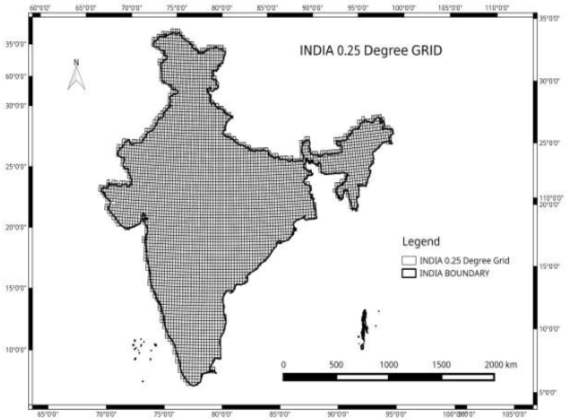


Figure 3: Atmospheric Model Grids: India, Source: Forest Fire Locations in India their spatio temporal patterns

Numerical Weather Prediction

- Weather Prediction of India for the next 10 days
- Area of India is 3.2 million sq.km
- To predict the weather pattern, you should model the atmosphere from sea level to 20 km
- Assume that, we make prediction of weather at each cubical gird with each cube measuring 0.1 km on each side.
- Compute F(lattitude, longitude, elevation, time)=Temperature or Pressure or Rain or ...

Numerical Weather Prediction

- Total number of points = $3.2 \times 10^6 \times 20 \times 10^3 = 6.4 \times 10^{10}$
- If these nodes are connected to each other and influences, the size of the matrix is 10^{10}

Gaussian Elimination

- To solve the NWP matrix, of size 10^{10} using GFLOP machine, it requires, $10^{30}/10^9 = 10^{21}$ seconds, that is, 10^{13} years to compute, but age of earth is 4.5×10^9 years.
- For a tera flops machine (1 TFLOPS $= 10^{12}$ FLOPS), it requires, 10^{10} years.
- For a peta flops machine (1 PFLOPS $= 10^{15}$ FLOPS), it requires, 3.16×10^7 years.
- Researchers use HPC, Supercomputers and sparse linear solvers

Sparse Matrix and Iterative Schemes

Sparse Matrix

Note: there is no strict or standard definition regarding the sparse matrix, but the following is commonly used

Definition 1 (Sparse Matrix)
A sparse matrix is a matrix in which most of the elements are zero. A common criterion is that the number of zeros is roughly equal to the number of rows or column.

Definition 2 (Sparsity)
The sparsity of matrix can be represented as

$$\text{Sparsity} = \frac{\text{Number of Zero Elements}}{\text{Total Number of Elements}}$$

Typically, a matrix is considered sparse if sparsity is > 0.5

Iterative Schemes

The underlying principles behind iterative methods to solve $Ax = b$ are as follows:

1. Guess any $x^{(0)}$
2. Compute the residual $r^{(0)} = b - Ax^{(0)}$
3. Compute the residual norm, that is $\|r^{(0)}\|$
4. Use an algorithm to compute $x^{(1)}$ involving A, b and $x^{(0)}$
5. Recompute the residual norm $\|r^{(1)}\| = \|b - Ax^{(1)}\|$
6. For given $x^{(i)}$, iterate this process until you get the residual as zero or an equivalent condition

Vector Norms

Norms

Definition 3 (Vector Norms)

A vector norm, usually denoted by $\|.\|$, is a function from a vector space V to the set of nonnegative real numbers that obeys the following three postulates. $\|.\|: V \to \mathbb{R}_+$ such that

$$
\begin{aligned}
\|\mathbf{x}\| > 0 \quad & \text{if } \mathbf{x} \neq 0, \mathbf{x} \in V \\
\|\alpha \mathbf{x}\| = |\alpha| \|\mathbf{x}\| \quad & \text{if } \alpha \in \mathbb{R}, \mathbf{x} \in V \\
\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\| \quad & \text{if } \mathbf{x}, \mathbf{y} \in V
\end{aligned}
$$

Vector Norms

The most familiar norm on $\mathbb{R}^n$ is the Euclidean norm defined by

$$\|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, \quad \text{where} \quad \mathbf{x} = (x_1, x_2, \cdots, x_n) \tag{1}$$

This is the norm that corresponds to usual concept of length. The other norms are also used. The simplest and easiest norm is ℓ_∞-norm

$$\|\mathbf{x}\|_\infty = \max_{0 \le i \le n} |x_i| \tag{2}$$

Vector Norms

The third important norm is ℓ_1-norm

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i| \tag{3}$$

Example 4

Compute the $\|.\|_1$, $\|.\|_2$ and $\|.\|_\infty$ norms of the following vectors.

$x = (4, 4, -4, 4) \quad v = (0, 2, 2, 2), \quad w = (8, 0, 0, 0)$

	$\|.\|_1$	$\|.\|_2$	$\|.\|_\infty$
x	16	8	4
v	6	$\sqrt{12}$	2
w	8	8	8

Vector Norms

The $p-$ norm or ℓ_p norm of a vector is defined by

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \quad \text{where} \quad \mathbf{x} = (x_1, x_2, \cdots, x_n) \tag{4}$$

Definition 5

If $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ are two vectors in $\mathbb{R}^n$, the ℓ_2 and ℓ_∞ distances between x and y are defined by

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$

and

$$\|\mathbf{x} - \mathbf{y}\|_\infty = \max_{1 \le i \le n} |x_i - y_i|$$

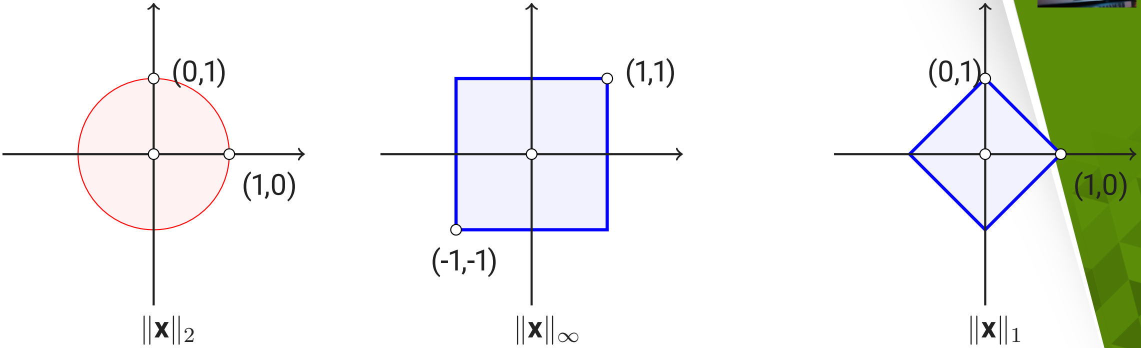
Vector Norms

To understand these norms better, see the visualization of the following picture which gives the sketch of the set

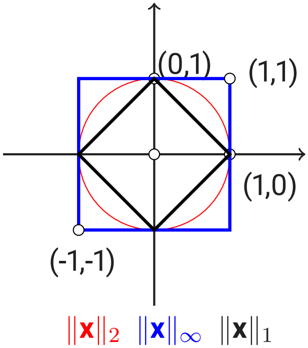
$$\{\mathbf{x} : \mathbf{x} \in \mathbb{R}^2, \|\mathbf{x}\| \leq 1\}$$

This set is called unit cell or the unit ball in two-dimensional vector space.

Vector Norms



Vector Norms

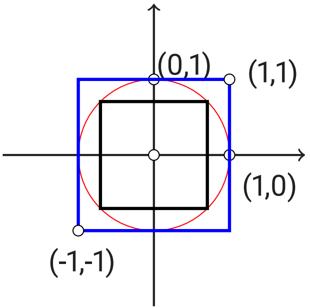


Vector Norms

Theorem 6

For each $x \in \mathbb{R}^n$,

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



$\|\mathbf{x}\|_2$ $\|\mathbf{x}\|_\infty$ $\sqrt{2}\|\mathbf{x}\|_\infty$

Matrix Norm

If $A = (a_{ij})$ is an $n \times n$ matrix and $\|.\|$ is a vector norm in $\mathbb{R}^n$, then the matrix norm is

$$\|A\| = \max_{\|x\|=1} \|Ax\| \qquad \text{(Vector Induced Norm)}$$

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \qquad \text{(Row Sum Norm)}$$

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}| \qquad \text{(Column Sum Norm)}$$

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2} \qquad \text{(Frobenius Norm)}$$

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sigma_{max}(A)$$

Condition Number

If A is an invertible matrix, then its condition number $\kappa(A)$ is defined by

$$\kappa(A) = \|A\|\|A^{-1}\| \tag{5}$$

$$\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)} \qquad (l_2\text{Norm})$$

$$\kappa(A) = \frac{|\lambda_{max}(A)|}{|\lambda_{min}(A)|} \qquad (A * A = AA * \text{A is normal})$$

$$\kappa(A) = 1 \qquad (A * A = AA* = I\text{A is unitary})$$

Iterative Methods

Theorems

Theorem 7

If A is and $n \times n$ matrix such that $\|A\| < 1$, then $I - A$ is invertible, and

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

From this theorem, we can observe that

$$\|(I - A)^{-1}\| \leq \sum_{k=0}^{\infty} \|A^k\| \leq \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|}$$

Theorems

Theorem 8

If A and B are $n \times n$ matrices such that $\|I - AB\| < 1$, then A and B are invertible. Furthermore,

$$A^{-1} = B\sum_{k=0}^{\infty}(I - AB)^k$$

and

$$B^{-1} = \sum_{k=0}^{\infty}(I - AB)^k A$$

Iterative Methods

The general algorithm for solving a system $Ax = b$ is as follows:

1. Choose a nonsingular matrix Q
2. Choose an arbitrary starting vector $x^{(0)}$
3. Generate vectors $x^{(1)}, x^{(2)}, \cdots$ recursively from the equation

$$Qx^{(k)} = (Q - A)x^{(k-1)} + b, \quad k = 1, 2, 3, \cdots \tag{6}$$

Suppose x is the solution of the system $Ax = b$ and $x^{(k)}$ converges to x, as $k \to \infty$, then

$$Qx = (Q - A)x + b$$

Note, that the system (6) should be easy to solve for $x^{(k)}$ when the right hand side is known. Also, Q should be chosen to ensure that $x^{(k)}$ converges to x, no matter, what initial vector is used and the convergence should be rapid.

Iterative Methods

Note that the true solution x satisfies the equation

$$x = (I - Q^{-1}A)x + Q^{-1}b \tag{7}$$

Therefore, x is a fixed point of the mapping

$$f(x) = (I - Q^{-1}A)x + Q^{-1}b$$

From (6)

$$Qx^{(k)} = (Q - A)x^{(k-1)} + b, \quad k = 1, 2, 3, \cdots$$

$$\implies x^{(k)} = (I - Q^{-1}A)x^{(k-1)} + Q^{-1}b, \quad k = 1, 2, 3, \cdots$$

Now,

$$x^{(k)} - x = (I - Q^{-1}A)(x^{(k-1)} - x), \quad k = 1, 2, 3, \cdots \tag{8}$$

$$\implies \|x^{(k)} - x\| = \|(I - Q^{-1}A)\| \|(x^{(k-1)}) - x\|$$

$$\implies \|x^{(k)} - x\| = \|(I - Q^{-1}A)\|^{k} \|(x^{(0)}) - x\|$$

Iterative Methods

If $\|I - Q^{-1}A\| < 1$, we can conclude that

$$\lim_{k\to\infty} \|x^{(k)} - x\| = 0$$

By above theorem, it is guaranteed that if $\|I - Q^{-1}A\| < 1$, then both $Q^{-1}A$ and A are invertible.

Theorem 9
If $\|I - Q^{-1}A\| < 1$ for some matrix norm, then the sequence produced by (6) converges to the solution of $Ax = b$ for any initial vector $x^{(0)}$.

Theorem 10
If all eigenvalues of $I - Q^{-1}A$ lies in the open unit disc $|z| < 1$, then the sequence produced by (6) converges to the solution of $Ax = b$ for any initial vector $x^{(0)}$.

Iterative Methods

The above theorem implies that the spectral radius of $I - Q^{-1}A$ must be less than 1, that is,

$$\rho(I - Q^{-1}A) < 1$$

Let $r^{(k)}$ denote the residual vector obtained from $x^{(k)}$ after k iterations, then we get

$$r^{(k)} = b - Ax^{(k)}.$$

By above theorem, if $\|I - Q^{-1}A\| < 1$, then $\|r^{(k)}\| \to 0$.

Theorem

The distance between two matrices $A, B \in M_{n\times n}(\mathbb{R})$ is $\|A - B\|$.

Theorem 11

If $\|.\|$ is a vector norm in $\mathbb{R}^n$, then the matrix norm is

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

Theorems and Proof

Theorem 12

If $||.||$ is a vector norm in $\mathbb{R}^n$, then the matrix norm is

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

Proof: It is enough to prove that $||.||$ is a matrix norm. If $A \neq 0$, at least one of its column is not a zero vector. Let jth column $(A)_j \neq 0$. Then

$$||A|| = \max_{||x||=1} ||Ax|| \geq ||Ae_j|| = ||(A_j)|| > 0$$

$$||\alpha A|| = \max_{||x||=1} ||\alpha Ax|| = \max_{||x||=1} |\alpha|||Ax|| = |\alpha| \max_{||x||=1} ||Ax|| = |\alpha|||Ax||$$

Theorems and Proof

Now,

$$\|A+B\| = \max_{\|x\|=1} \|(A+B)x\| = \max_{\|x\|=1} \|Ax+Bx\| \leq \max_{\|x\|=1} (\|Ax\| + \|Bx\|)$$

$$\max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\| = \|A\| + \|B\|$$

$$\|AB\| = \max_{\|x\|=1} \|ABx\| \leq \max_{\|x\|=1} \|A\|\|Bx\| \leq \max_{\|x\|=1} \|A\|\|B\|\|x\| = \|A\|\|B\|$$

Hence the proof.

Theorem

Theorem 13

If $A = (a_{ij})$ is an $n \times n$ matrix, then

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|$$

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|$$

The $\|A\|_1$ norm is also called as the column sum norm as it computes the maximum absolute column sum of the matrix. The $\|A\|_\infty$ norm is also called as the row sum norm as it computes the maximum absolute row sum of the matrix.

Theorems and Proof

Theorem 14

If $A = (a_{ij})$ is an $n \times n$ matrix, then

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

Proof: Observe that when $Ax = b$

$$b_i = \sum_{j=1}^{n} a_{ij} x_j \implies \|b\|_\infty = \max_{1\le i\le n} |b_i| = \max_{1\le i\le n} \left| \sum_{j=1}^{n} a_{ij} x_j \right|$$

Theorems and Proof

$$\|A\|_\infty = \max_{\|x\|_\infty=1} \|Ax\| = \max_{\|x\|_\infty=1} \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}x_j| \le \max_{\|x\|_\infty=1} \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}||x_j|$$

$$\implies \|A\|_\infty \le \left(\max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \right) \max_{\|x\|_\infty=1} |x_j| = \left(\max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}| \right) \max_{\|x\|_\infty=1} \|x\|$$

$$\implies \|A\|_\infty \le \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

It is enough to prove that the equality is achieved for some x. Choose your x such that $x_j = sign(A_{ij})$.

Theorems and Proof

Then the proof follows for row sum follows

$$\|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|$$

For column sum, consider the columns of A as $A_1, A_2, \cdots, A_n$

$$\|Ax\|_1 = \left\|\sum_{i=1}^{n} A_i x\right\|_1 \le \sum_{i=1}^{n} \|A_i x\|_1 = \sum_{i=1}^{n} |x| \|A_i\|_1 \le \max_{1\le j\le n} \|x\|_1 \sum_{i=1}^{n} |a_{ij}|$$

$$\implies \|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 \le \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

The equality and the proof follows by taking the standard basis vector for x.

Theorem

The Frobenius norm is given by $\|A\|_F$

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

Theorems and Proof

Theorem 15

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 = \textbf{trace}(A^T A)$$

$$\|A\|_F^2 = \textbf{trace}(A^T A)$$

Proof: If $A = (a_{ij})$ and $B = (b_{ij})$, then $C = AB = (c_{ij})$ is given by, then

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \implies c_{ii} = \sum_{k=1}^{n} a_{ik} b_{ki}$$

Theorems and Proof

When $C = A^T A$, we have

$$c_{ii} = \sum_{k=1}^{n} a_{ki} a_{ki} = \sum_{k=1}^{n} a_{ki}^2$$

$$\textbf{trace}(A^T A) = \sum_{i=1}^{n} c_{ii} = \sum_{i=1}^{n} \sum_{k=1}^{n} a_{ki}^2 = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2$$

Since

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2$$

the proof follows.

Theorems and Proof

Theorem 16

Frobenius norm is a norm.

Proof: If $A \neq 0$, at least one element is not a zero element. Let it be a_{pk}. Hence

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2} \geq |a_{pk}| > 0$$

$$\|\alpha A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |\alpha a_{ij}|^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |\alpha||a_{ij}|^2} = |\alpha|\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2} = |\alpha|\|A\|_F$$

Theorems and Proof

$$\|A+B\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}+b_{ij}|^2$$

We know

$$|a+b|^2 \leq |a|^2 + |b|^2 + 2|a||b|$$

Hence

$$\|A+B\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} |b_{ij}|^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}||b_{ij}|$$

$$\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}||b_{ij}|$$

Theorems and Proof

By Cauchy-Schwarz inequality

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||b_{ij}| \leq \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}|b_{ij}|^2\right)^{1/2} = \|A\|_F\|B\|_F$$

$$\|A+B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\|A\|_F\|B\|_F \implies \|A+B\|_F \leq \|A\|_F + \|B\|_F$$

$$\|AB\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\sum_{k=1}^{n} a_{ik}b_{kj}\right|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}|^2\sum_{k=1}^{n}|b_{kj}|^2\right)$$

$$\|AB\|_F^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}|^2\sum_{l=1}^{n}|b_{lj}|^2\right) = \sum_{i=1}^{n}\sum_{k=1}^{n}|a_{ik}|^2\sum_{l=1}^{n}\sum_{j=1}^{n}|b_{lj}|^2 = \|A\|_F^2\|B\|_F^2$$

Theorems and Proof

Theorem 17

If U and V are orthogonal, then

$$\|UA\|_F = \|AV\|_F = \|A\|_F$$

Proof:

$$\|UA\|_F^2 = \textbf{trace}((UA)^T(UA)) = \textbf{trace}(A^TU^TUA) = \textbf{trace}(A^TA) = \|A\|_F^2$$

Since $\textbf{trace}(A^TA) = \textbf{trace}(AA^T)$, we get

$$\|AV\|_F^2 = \textbf{trace}((AV)^T(AV)) = \textbf{trace}((AV)(AV)^T)$$

$$= \textbf{trace}(AVV^TA^T) = \textbf{trace}(AA^T) = \textbf{trace}(A^TA) = \|A\|_F^2$$

Example

Example 18

Compute the $\|.\|_1$, $\|.\|_F$ and $\|.\|_\infty$ norms of the following matrix.

$$\begin{pmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{pmatrix}$$

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| = \max\{27, 27, 18, 35\} = 35$$

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| = \max\{14, 36, 28, 29\} = 36$$

$$\|A\|_F = 32.3883$$

Eigenvalues

If I denotes the identity matrix, then $\|I\| = 1$. Also,

$$\|AB\| \leq \|A\|\|B\|$$

Definition 19 (Characteristic Polynomial:)
If A is a square matrix, the characteristic polynomial of A is defined by

$$p(\lambda) = \det(A - \lambda I)$$

Definition 20 (Eigenvalue and Eigenvector)
If p is a characteristic polynomial of A, the roots of p are called eigenvalues or characteristic values of A. If λ is an eigenvalue of A and $x \neq 0$ such that $A\mathbf{x} = \lambda\mathbf{x}$, then $\mathbf{x}$ is an eigenvector of A.

Spectral Radius

Usually, the matrix is considered over the field C though the matrix entries can be real.

Definition 21 (Singular Values)

The singular values σ of an $m \times n$ matrix A are the positive square roots of the nonzero eigenvalues of the $n \times n$ symmetric matrix $A^T A$.

$$\sigma_i(A) = \sqrt{\lambda_i(A^T A)}$$

Definition 22 (Spectral Radius)

The spectral radius $\rho(A)$ of a matrix A is defined by

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

This is also denoted as $\lambda_{max}(A)$.

Spectral Radius

The matrix norm induced by the Euclidean norm (ℓ_2 norm) is the spectral norm, and is given by

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sigma_{max}(A)$$

This is the minimum norm and provides the tightest measure of size.

Theorem 23

If $A = (a_{ij})$ is an $n \times n$ matrix, then

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\|A\|_2,$$

$$\|A\|_F = \sqrt{\text{trace}(A^T A)}$$

$$\rho(A) \leq \|A\|$$

for any natural norm $\|.\|$.

Thanks

Doubts and Suggestions

panch.m@iittp.ac.in

MA633L-Numerical Analysis

Lecture 26 : Numerical Linear Algebra - Norms

Panchatcharam Mariappan[1]

[1]Associate Professor
Department of Mathematics and Statistics
IIT Tirupati, Tirupati

March 10, 2025

भारतीय प्रौद्योगिकी संस्थान तिरुपति
TIRUPATI