

Felix Juefei Xu & Dipan K. Pal

# Pattern Recognition Theory

**Recitation 1: Linear Algebra**

# The Basics

- Scalar ( $x$ )
  - A real number

- Vector ( $\mathbf{x}$ )

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$$

# The Basics

- Matrix (**X**)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & & x_{nm} \end{bmatrix}$$

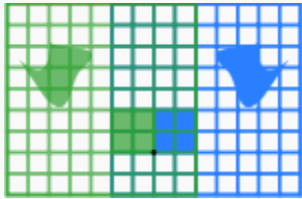

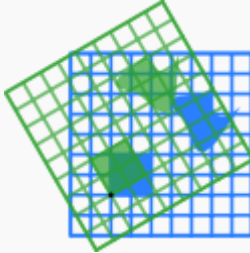
# Types of Matrices

- Square & Rectangular matrices
- Upper & Lower Triangular matrices
- Identity matrix
- Symmetric & Skew-symmetric matrices
- Hermitian matrix
- Singular matrix
- Orthogonal matrix

# Consider a matrix as...

- ... an operator ...
- ... which linearly transforms a vector ...
- ... to a different vector space!

For example:

Horizontal flip	Scaling by a factor of 3/2	Rotation by $\pi/6^R = 30^\circ$
$\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 3/2 & 0 \\ 0 & 3/2 \end{bmatrix}$	$\begin{bmatrix} \cos(\pi/6^R) & -\sin(\pi/6^R) \\ \sin(\pi/6^R) & \cos(\pi/6^R) \end{bmatrix}$
		

# Vector Spaces & Subspaces

For any  $m \times n$  matrix  $\mathbf{A}$  with rank  $r$ :

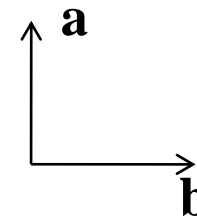
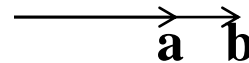
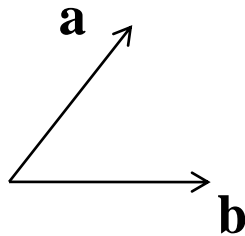
- Column Space
  - Contains all linear combinations of the columns of the matrix
  - Has dimension  $r$
  - $\mathbf{Ax} = \mathbf{b}$  can be solved iff  $\mathbf{b}$  is in the column space of  $\mathbf{A}$ .
- Null Space
  - Contains all vectors  $\mathbf{x}$  such that  $\mathbf{Ax} = \mathbf{0}$
  - Has dimension  $(n-r)$
- Row Space / Left Null Space (equivalent to Column & Null Space on  $\mathbf{A}^T$ )

# Inner Products

- Inner product of  $\mathbf{x}$  and  $\mathbf{y}$ : Sum of element-wise products

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i$$

- $\mathbf{x}$  and  $\mathbf{y}$  must be equal in length
- Result is a **scalar**
  - Test of similarity of two vectors
  - Don't forget to normalize vectors before comparing!



# Outer Products

- Outer product of  $\mathbf{x}$  and  $\mathbf{y}$ :

$$\mathbf{x} \otimes \mathbf{y} = \mathbf{xy}^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & x_1 y_3 \\ x_2 y_1 & x_2 y_2 & x_2 y_3 \\ x_3 y_1 & x_3 y_2 & x_3 y_3 \\ x_4 y_1 & x_4 y_2 & x_4 y_3 \end{bmatrix}$$

- $\mathbf{x}$  and  $\mathbf{y}$  can be of different lengths
- Result is a matrix



# Linear independence

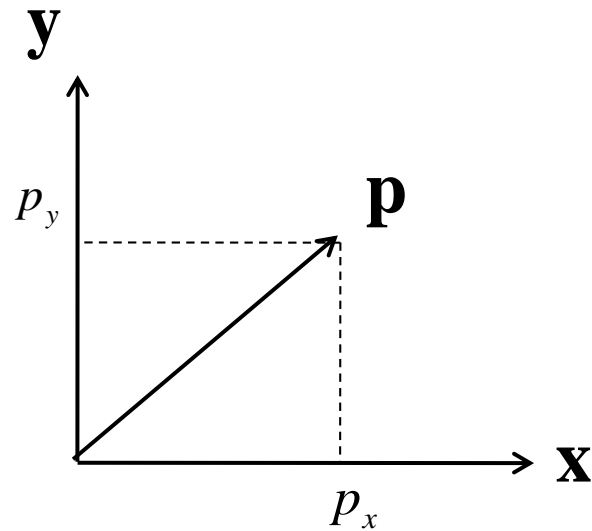
- Non-zero vectors  $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$  are linearly independent only if

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 \neq 0$$

for any non-zero set of constants  $a_n$

- We say a family of vectors is a linearly independent family if none of them can be written as a linear combination of finitely many other vectors in the family.

# Linear independence & Orthogonality



Inner product of orthogonal vectors is 0

$$\mathbf{x}^T \mathbf{y} = 0$$

# Projections

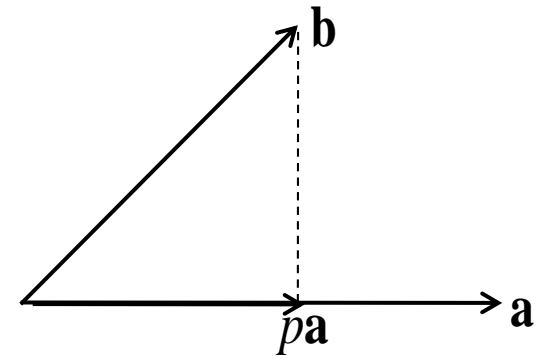
- We want to find the value of  $p$  which minimizes the error  $\|\mathbf{b} - p\mathbf{a}\|$ , i.e.

$$\begin{aligned}\hat{p} &= \arg \min_p \|\mathbf{b} - p\mathbf{a}\| \\ &= \arg \min_p (\mathbf{b} - p\mathbf{a})^T (\mathbf{b} - p\mathbf{a}) \\ &= \arg \min_p (\mathbf{b}^T \mathbf{b} + p^2 \mathbf{a}^T \mathbf{a} - 2p \mathbf{a}^T \mathbf{b})\end{aligned}$$

Taking derivative and setting it to 0:

$$0 = 2\hat{p}\mathbf{a}^T \mathbf{a} - 2\mathbf{a}^T \mathbf{b}$$

$$\hat{p} = \frac{\mathbf{a}^T \mathbf{b}}{\mathbf{a}^T \mathbf{a}}$$



- Hence, if  $\mathbf{a}$  is unit norm, the projection coefficient is equivalent to the inner product of  $\mathbf{a}$  and  $\mathbf{b}$

# Matrix Operations

## 1. Matrix Transpose, Conjugate Transpose

- $(\mathbf{A}^T)^T = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$
- $(r\mathbf{A})^T = r\mathbf{A}^T$

## 2. Matrix Determinant

- Only exists for square matrices
- $\det(a\mathbf{X}) = a^n \det(\mathbf{X})$
- $\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})$
- $\det(\mathbf{A}^T) = \det(\mathbf{A})$
- $\det(\mathbf{A}^{-1}) = (\det(\mathbf{A}))^{-1}$

# Matrix Operations

## 3. Matrix arithmetic:

- Addition & Subtraction: Element-wise operations
- Matrix Multiplication:
  - $\mathbf{AB} \neq \mathbf{BA}$
- Vector Multiplication
- Scalar Multiplication

## 4. Matrix Inverse:

- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$

## 5. Matrix Rank

# Matrix Operations

## 6. (Moore-Penrose) Pseudo-Inverse :

- The pseudoinverse  $A^+$  of a matrix  $A$  is a generalization of the inverse matrix. The most widely known is the Moore-Penrose pseudoinverse.
- The pseudoinverse is defined and unique for all matrices whose entries are real or complex numbers.
- If the matrix  $A$  has dimensions  $m \times n$ , and is full rank, then use the left inverse if  $m > n$ , and the right inverse if  $m < n$ .
- Left inverse:  $A_{\text{left}}^{-1} = (A^T A)^{-1} A^T$ , i.e.  $A_{\text{left}}^{-1} A = I_n$
- Right inverse:  $A_{\text{right}}^{-1} = A^T (A A^T)^{-1}$ , i.e.  $A A_{\text{right}}^{-1} = I_m$

# System of Linear Equations

- Suppose we have a set of linear equations such as the following example:

$$\begin{aligned}a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3\end{aligned}$$

In matrix form, we can write this as:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\mathbf{A} \mathbf{x} = \mathbf{b}$$

# System of Linear Equations $Ax = b$

	Under-determined	Well defined	Over-determined
Equations vs. Unknowns	Less linearly independent equations than unknowns	As many linearly independent equations as unknowns	More linearly independent equations than unknowns
<b>A</b>	<b>A</b> is “fat”	<b>A</b> is square	<b>A</b> is “tall”
# of solutions	Usually infinitely many solutions	Usually one solution	Usually no solution
Typical solution	Minimum Norm solutions $\mathbf{x} = \mathbf{A}^T (\mathbf{A}\mathbf{A}^T)^{-1} \mathbf{b}$	<u>Exact</u> solution: $\mathbf{x} = \mathbf{A}^{-1} \mathbf{b}$	<u>Least squared error</u> solution: $\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}$



# Eigen decomposition

- Any matrix **A** represents a transformation operation on a vector

$$\mathbf{Ax} = \mathbf{x}'$$

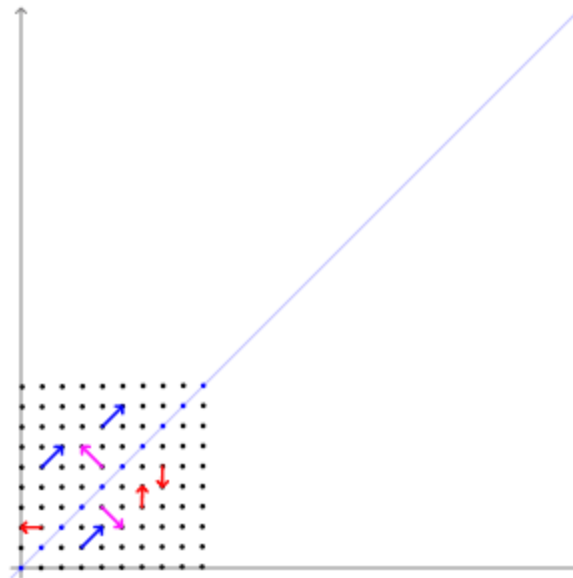
- For certain vectors, the transformation is merely a scale change

$$\mathbf{Ax} = \lambda \mathbf{x}$$

# Eigen decomposition

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

- $\mathbf{x}$  is the eigen vector of transformation  $\mathbf{A}$
- $\lambda$  is the corresponding eigen value



# Determining the eigen decomposition

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

For a non-trivial solution

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

This equation is called the characteristic equation.  
Solve for the eigen values  $\lambda$  and hence obtain 'x's

# Some properties of eigen decomposition

- Rank of  $\mathbf{A}$  = Number of non-zero eigen values of  $\mathbf{A}$  = Number of linearly independent eigen vectors
- Determinant of  $\mathbf{A}$ ,  $|\mathbf{A}| = \prod_{i=1}^{i=n} \lambda_i$
- Trace of  $\mathbf{A} = \sum_{i=1}^{i=n} \lambda_i$

# Some properties of eigen decomposition

- If  $\mathbf{A}$  is symmetric, then the eigen vectors are orthogonal
- If  $\lambda$  is an eigen value of  $\mathbf{A}$ ,
  - then  $\frac{1}{\lambda}$  is an eigen value of  $\mathbf{A}^{-1}$
  - $\lambda - k$  is an eigen value of  $\mathbf{A} - k\mathbf{I}$
  - $\lambda^m$  is an eigen value of  $\mathbf{A}^m$

# Eigen decomposition

- A matrix is
  - positive definite if all its eigen values

$$\lambda_i > 0$$

- positive semi-definite if

$$\lambda_i \geq 0$$

- negative definite if

$$\lambda_i < 0$$

- negative semi-definite if

$$\lambda_i \leq 0$$

# Calculus

$$\nabla f(\mathbf{x}) = \text{gradient}(f(\mathbf{x})) = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

$$\begin{aligned} \frac{\partial \mathbf{x}^T \mathbf{c}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{c}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{c} \\ \frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} &= \mathbf{A} \end{aligned}$$

$$\text{Product Rule: } \frac{\partial (g(\mathbf{x}))^T h(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (h(\mathbf{x}))^T}{\partial \mathbf{x}} g(\mathbf{x}) + \frac{\partial (g(\mathbf{x}))^T}{\partial \mathbf{x}} h(\mathbf{x})$$

$$\text{e.g.: } \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A}^T + \mathbf{A}) \mathbf{x} \quad \text{In this case } (g(\mathbf{x}) = \mathbf{x}, h(\mathbf{x}) = \mathbf{A} \mathbf{x})$$

$$\text{Chain Rule: } \frac{\partial g(h(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial g}{\partial h} \cdot \frac{\partial h}{\partial \mathbf{x}}$$

# Highly Recommended

- “Linear Algebra and its applications” – 4<sup>th</sup> Ed., Gilbert Strang
- The Matrix Cookbook
  - <http://matrixcookbook.com/>
- Gilbert Strang MIT video lectures
  - <http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/video-lectures/>