

# Review: Linear Algebra, Maximum Likelihood, and Probability

Arin Ghazarian  
Chapman University

# Linear Algebra

# Linear Algebra

- A good understanding of linear algebra is essential for understanding and working with many machine learning algorithms, especially deep learning algorithms

# Matrix

n = number of columns m = number of rows

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

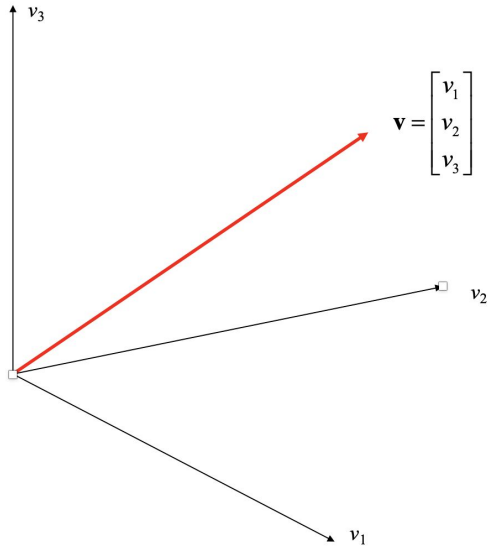
# Design matrix/model matrix/regressor matrix

- Columns are represent the explanatory variables
- Each row represents an individual object/record/sample

		Variables							
Samples		Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable 6	.....	Variable m
	Sample 1	29	31	34	36	39	48	.....	48
	Sample 2	27	29	31	34	36	45	.....	45
	Sample 3	25	27	29	31	34	42	.....	42
	Sample 4	23	25	27	29	31	39	.....	39
	...	...	...	...	...	...	...	.....	...
	Sample n	18	20	21	23	24	26	.....	20

# Vector

- $n \times 1$  matrix
- One point in an  $n$ -dimensional space



$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

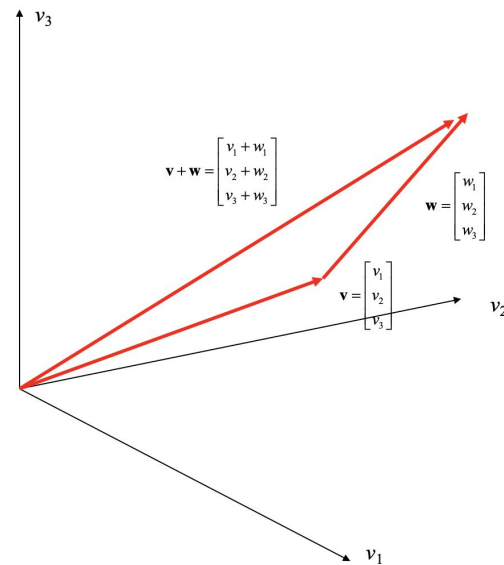
# Tensor

- In the general case, an array of numbers arranged on a regular grid with a variable number of axes is known as a tensor.
- We identify the element of  $A$  at coordinates  $(i, j, k)$  by writing  $A_{i,j,k}$ .

# Matrix Operations: Addition

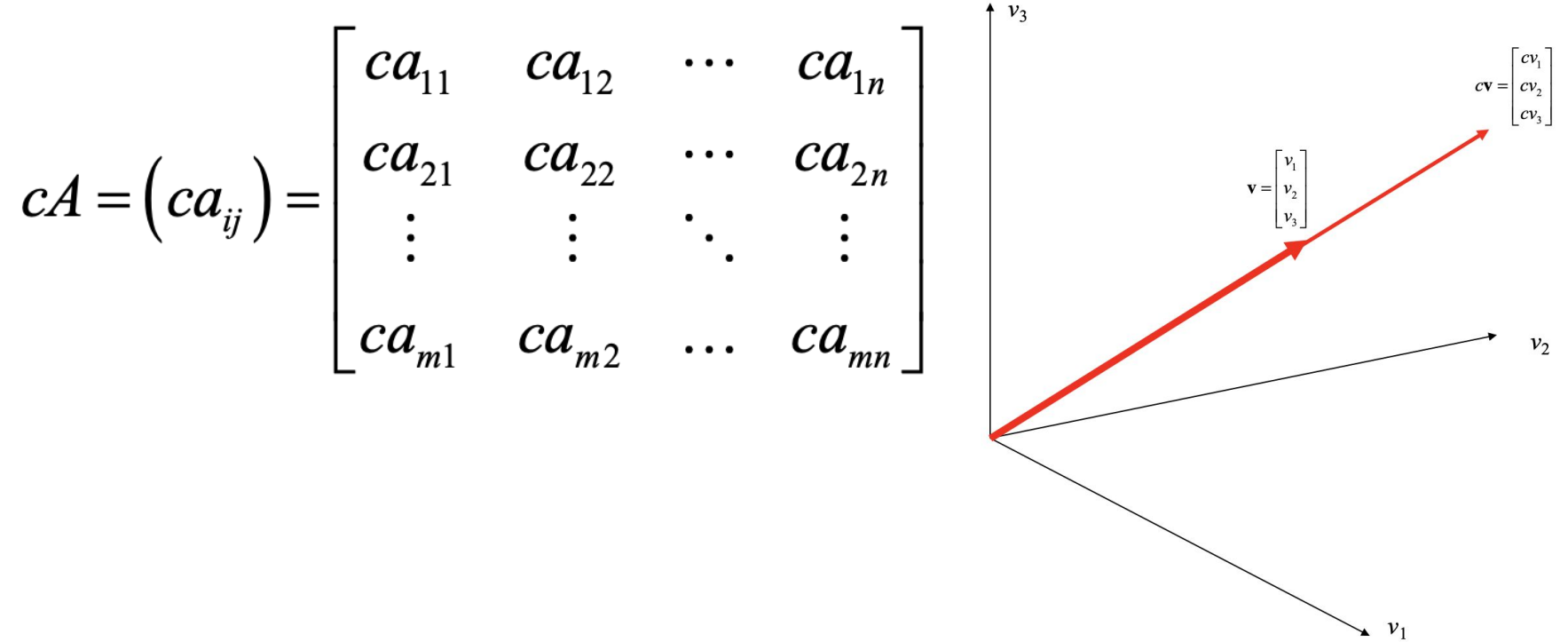
Both A and B are  $n \times m$  matrices (must have the same dimensions)

$$A + B = (a_{ij} + b_{ij}) = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$





# Matrix Operations: Scalar Multiplication



# Matrix Operations: Multiplication

Let A denote an  $n \times m$  matrix and B denote an  $m \times k$  matrix, then  $C=A \cdot B$  will be an  $n \times k$  matrix where

$$c_{il} = \sum_{j=1}^m a_{ij} b_{jl}$$

# Identity Matrix

Square Matrix with diagonal elements equal to 1

$$AI = A$$

$$IA = A$$

$$I = I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

# Transpose of a Matrix ( $A^T$ or $A'$ )

Flips a matrix over its diagonal

**A**

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$$

# Symmetric Matrices

- An  $n \times n$  matrix,  $A$ , is said to be symmetric if:

$$A' = A$$

# Definition: Inverse of a Matrix

- Both A and B are  $n \times n$  matrices, then B is called the inverse of matrix A if:

$$AB = BA = I$$

- If the matrix B exists then A is called invertible (not all matrices are invertible)
- The inverse of Matrix A is denoted by  $A^{-1}$

$$\begin{bmatrix} b & a \\ a & b \end{bmatrix}^{-1} = \frac{1}{b^2 - a^2} \begin{bmatrix} b & -a \\ -a & b \end{bmatrix}$$

# Invertible vs Singular

- $A^{-1}$  exists if  $|A| \neq 0$  and  $A$  is a square matrix.
- A square matrix that is not invertible is called singular or degenerate
- The solution to the system of linear equations  $Ax=b$  will be  $x=A^{-1}b$

# Linear Independence

- A set of vectors is linearly independent if no vector in the set is a linear combination of the other vectors.
- Together, this means that the matrix must be square, that is, we require that  $m = n$  and that all of the columns must be linearly independent.
- A square matrix with linearly dependent columns is known as singular.



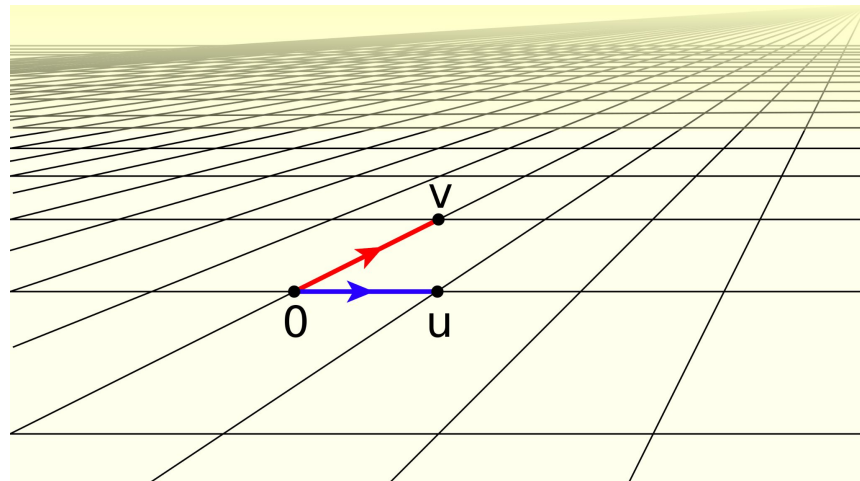
# Linear Combination

- Formally, a linear combination of some set of vectors  $\{v^{(1)}, \dots, v^{(n)}\}$  is given by multiplying each vector  $v_{(i)}$  by a corresponding scalar coefficient and adding the results:

$$\sum_i c_i \mathbf{v}^{(i)}.$$

# Span

- The span of a set of vectors is the set of all points obtainable by linear combination of the original vectors.
- Determining whether  $Ax = b$  has a solution thus amounts to testing whether  $b$  is in the span of the columns of  $A$ . This particular span is known as the column space or the range of  $A$ .



[https://en.wikipedia.org/wiki/Linear\\_span](https://en.wikipedia.org/wiki/Linear_span)

The cross-hatched plane is the linear span of  $u$  and  $v$  in  $\mathbb{R}^3$ .

# System of Linear Equation

## System of Linear Equation

$$2.0x + 4.0y + 6.0z = 18$$

$$4.0x + 5.0y + 6.0z = 24$$

$$3.0x + 1y - 2.0z = 4$$

## Matrix representation

$$A = \begin{bmatrix} 2.0 & 4.0 & 6.0 \\ 4.0 & 5.0 & 6.0 \\ 3.0 & 1.0 & -2.0 \end{bmatrix} \quad X = \begin{bmatrix} x \\ y \\ z \end{bmatrix} \quad b = \begin{bmatrix} 18.0 \\ 24.0 \\ 4.0 \end{bmatrix}$$

$$X = A^{-1}b$$

<https://www.geeksforgeeks.org/java-program-to-represent-linear-equations-in-matrix-form/>

# Definition: Block Matrices

Partition matrix into submatrices

$$\mathbf{P} = \begin{bmatrix} 1 & 2 & 2 & 7 \\ 1 & 5 & 6 & 2 \\ 3 & 3 & 4 & 5 \\ 3 & 3 & 6 & 7 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{bmatrix}.$$

$$\mathbf{P}_{11} = \begin{bmatrix} 1 & 2 \\ 1 & 5 \end{bmatrix}, \quad \mathbf{P}_{12} = \begin{bmatrix} 2 & 7 \\ 6 & 2 \end{bmatrix}, \quad \mathbf{P}_{21} = \begin{bmatrix} 3 & 3 \\ 3 & 3 \end{bmatrix}, \quad \mathbf{P}_{22} = \begin{bmatrix} 4 & 5 \\ 6 & 7 \end{bmatrix}.$$

## Block Matrices: Multiplication

$$A \cdot B = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}$$

# Inverse of Block Matrices

- Where  $A$  and  $D$  are square of arbitrary size, and  $B$  and  $C$  are conformable (a matrix is conformable if its dimensions are suitable for defining some operation) with them for partitioning. Also,  $A$  and the Schur complement of  $A$  in  $P$ :  $P/A = D - CA^{-1}B$  must be invertible

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix},$$

Some useful relationships in linear algebra

$$(AB)' = B'A'$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(A')^{-1} = (A^{-1})'$$

# Trace of Matrix

the sum of elements on the main diagonal

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}$$

$$A = (a_{ij}) = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$



## Trace Operation Relationships

$$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$$

$$\text{tr}(c\mathbf{A}) = c \text{tr}(\mathbf{A})$$

$$\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^{\top})$$

$$\text{tr}(AB) = \text{tr}(BA)$$

# Determinant ( $\det(A)$ or $|A|$ )

- The determinant is a scalar value that is a function of the entries of a square matrix
- The determinant is nonzero if and only if the matrix is invertible

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

## Properties of Determinant

$$|AB| = |A||B|$$

$$|A^{-1}| = \frac{1}{|A|}$$

$$\det(A^T) = \det(A).$$

# Inner product of vectors (Dot Product)

Sum of the products of the corresponding entries of the two sequences of numbers.

$$\vec{x}' \cdot \vec{y} = \begin{bmatrix} x_1, \dots, x_p \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = x_1 y_1 + \dots + x_p y_p = \sum_{i=1}^p x_i y_i$$

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta,$$

# Angle Between Two Vectors

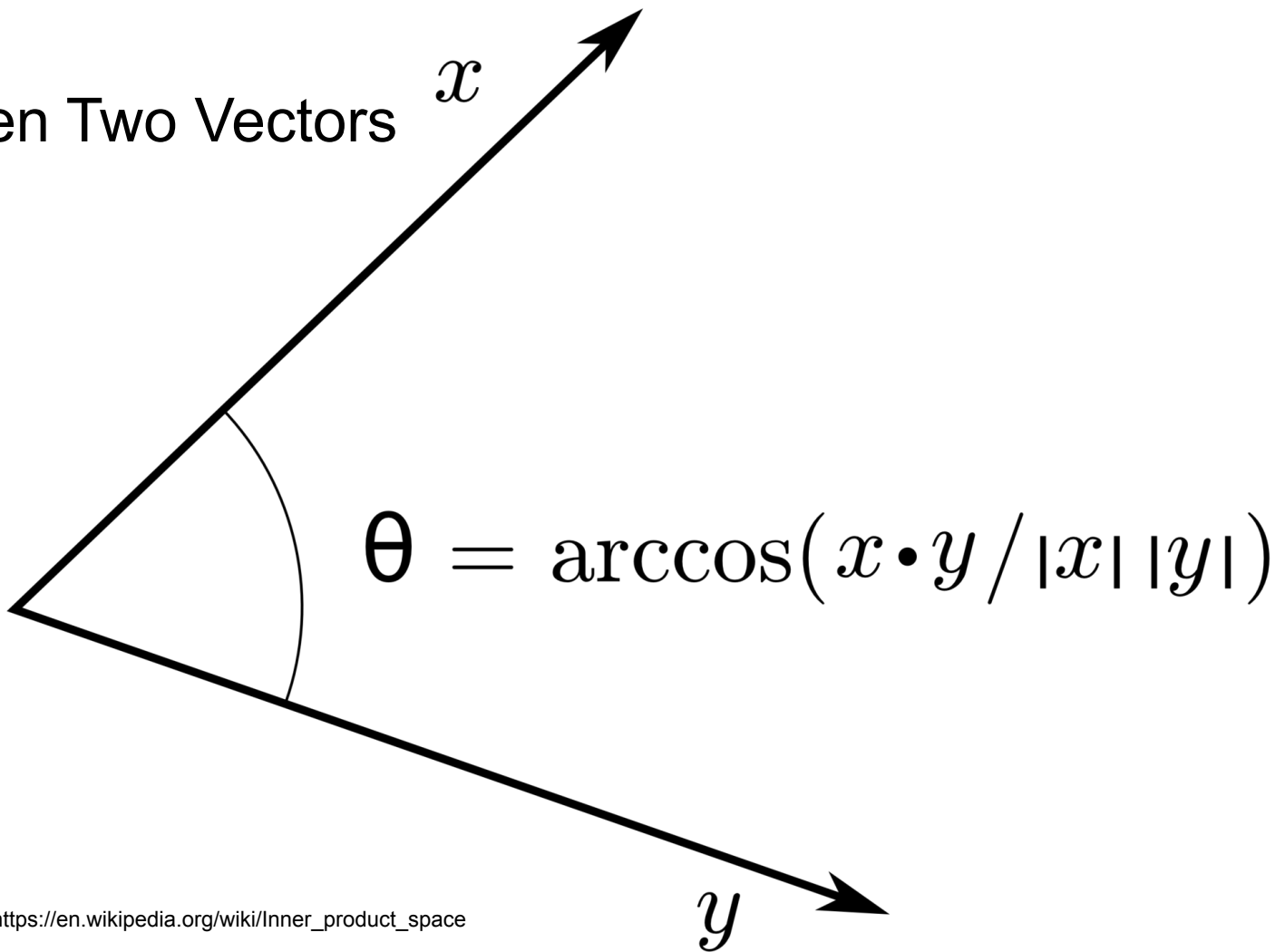
Length of Vector X:

$$\sqrt{\vec{x}' \cdot \vec{x}} = \sqrt{x_1^2 + \cdots + x_p^2} :$$

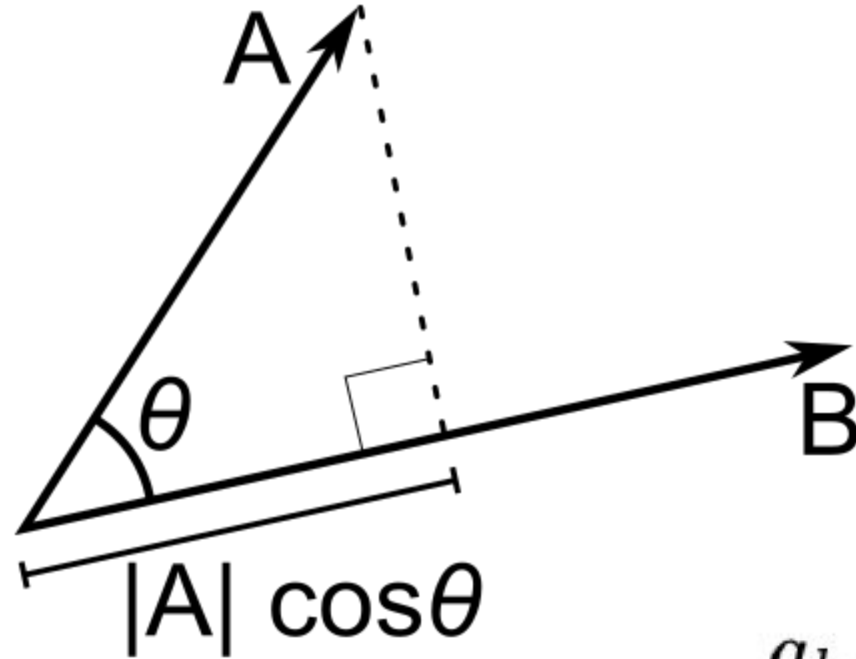
X and Y are p dimensional vectors, the angle between these two vectors:

$$\cos \theta = \frac{\vec{x}' \cdot \vec{y}}{\sqrt{\vec{x}' \cdot \vec{x}} \sqrt{\vec{y}' \cdot \vec{y}}}$$

# Angle Between Two Vectors



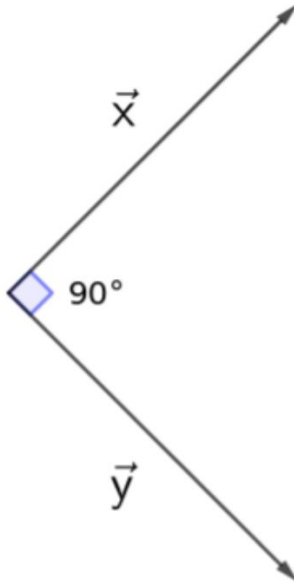
## Scalar Projection



$$a_b = \|\mathbf{a}\| \cos \theta$$

## Orthogonal Vectors

if  $\vec{x}' \cdot \vec{y} = 0$ , then  $\vec{x}$  and  $\vec{y}$  are orthogonal.





## Orthogonal matrices

$$P'P = PP' = I \text{ and } P^{-1} = P'$$

The rows of  $P$  have length 1 and are orthogonal to each other (the same applies for the columns)

$$\begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & 0 & -\frac{2}{\sqrt{6}} \end{bmatrix}$$

# Idempotent Matrix

A matrix which, when multiplied by itself, yields itself

$$\begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix}$$

# Differentiation with respect to a vector

$$\frac{df(\vec{x})}{d\vec{x}} = \begin{bmatrix} \frac{df(\vec{x})}{dx_1} \\ \vdots \\ \frac{df(\vec{x})}{dx_p} \end{bmatrix}$$

Differentiation with respect to a vector

$$f(\vec{x}) = \vec{a}'\vec{x} = a_1x_1 + \dots + a_nx_n$$

$$\text{then } \frac{df(\vec{x})}{d\vec{x}} = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\vec{x})}{\partial x_p} \end{bmatrix} = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} = \vec{a}$$

## Gradient of Quadratic Form

$$f(\vec{x}) = \vec{x}' A \vec{x}$$

$$\frac{df(\vec{x})}{d\vec{x}} = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\vec{x})}{\partial x_p} \end{bmatrix} = (\mathbf{A}' + \mathbf{A})\mathbf{x}$$

A is Symmetric

$$f(\vec{x}) = \vec{x}' A \vec{x}$$

$$\frac{df(\vec{x})}{d\vec{x}} = \begin{bmatrix} \frac{\partial f(\vec{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\vec{x})}{\partial x_p} \end{bmatrix} = 2A\vec{x}$$

## Differentiation With Respect to a Matrix

$$\frac{df(X)}{dX} = \left( \frac{\partial f(X)}{\partial x_{ij}} \right) = \begin{bmatrix} \frac{\partial f(X)}{\partial x_{11}} & \dots & \frac{\partial f(X)}{\partial x_{1p}} \\ \vdots & & \vdots \\ \frac{\partial f(X)}{\partial x_{q1}} & \dots & \frac{\partial f(X)}{\partial x_{pp}} \end{bmatrix}$$

# Jacobian

- Suppose  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a function such that each of its first-order partial derivatives exist on  $\mathbb{R}^n$ .
- This function takes a point  $x \in \mathbb{R}^n$  as input and produces the vector  $f(x) \in \mathbb{R}^m$  as output.
- In vector calculus, the Jacobian matrix of a vector-valued function of several variables is the matrix of all its first-order partial derivatives.
- Then the Jacobian matrix of  $f$  is defined to be an  $m \times n$  matrix, denoted by  $J$ , whose  $(i,j)$ th entry is  $J_{ij} = \partial f_i / \partial x_j$ :

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^T f_1 \\ \vdots \\ \nabla^T f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

where  $\nabla^T f_i$  is the transpose (row vector) of the gradient of the  $i$  component.



# Hessian

- The Hessian matrix or Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field. It describes the local curvature of a function of many variables.
- Suppose  $R^m \rightarrow R$  is a function taking as input a vector  $x \in R^n$  and outputting a scalar  $f(x) \in R$ . If all second-order partial derivatives of  $f$  exist, then the Hessian matrix  $H$  of  $f$  is a square  $m \times n$  matrix, usually defined and arranged as

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

# Positive Definite Matrices

- A symmetric matrix  $A$  with real entries is positive-definite if the real number  $\mathbf{x}^T A \mathbf{x}$  is positive for every nonzero real column vector  $\mathbf{x}$ .

$$\vec{x}' A \vec{x} = a_{11}x_1^2 + \cdots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \cdots + 2a_{n-1,n}x_{n-1}x_n > 0$$

- A symmetric matrix,  $A$ , is called positive semi definite if:

$$\vec{x}' A \vec{x} \geq 0 \quad \text{for all } \vec{x} \neq \vec{0}$$

$$\mathbf{z}^T I \mathbf{z} = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = a^2 + b^2$$

# Eigendecomposition

# Eigenvalues and eigenvectors

A is an  $n \times n$  matrix

$$A\vec{x} = \lambda\vec{x} \quad \text{with } \vec{x} \neq \vec{0}$$

Then  $\lambda$  is called an eigenvalue of A and  $\vec{x}$  is called an eigenvector of A.

How to find eigenvalues and eigenvectors?

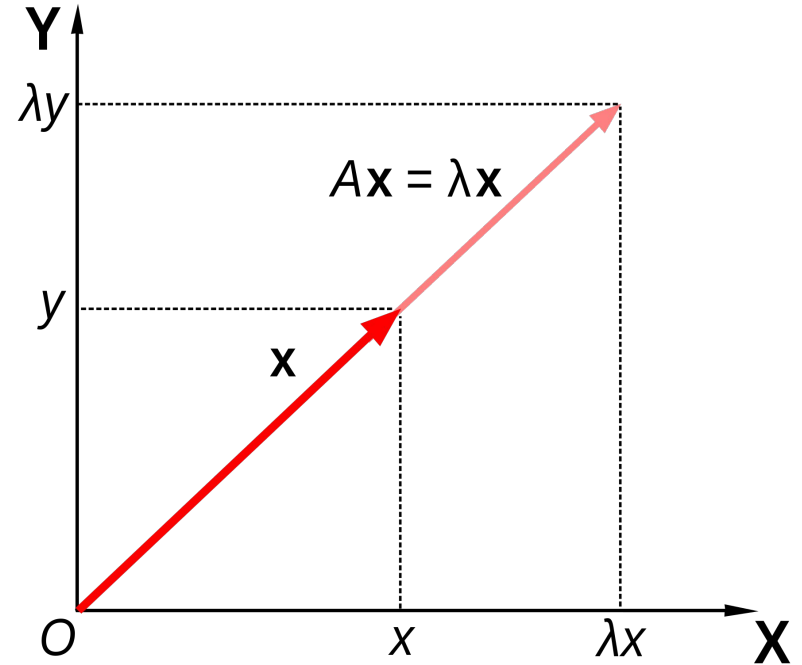
$$(A - \lambda I) \vec{x} = \vec{0}$$

$$|A - \lambda I| = \det \begin{bmatrix} (a_{11} - \lambda) & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & (a_{nn} - \lambda) \end{bmatrix} = 0$$

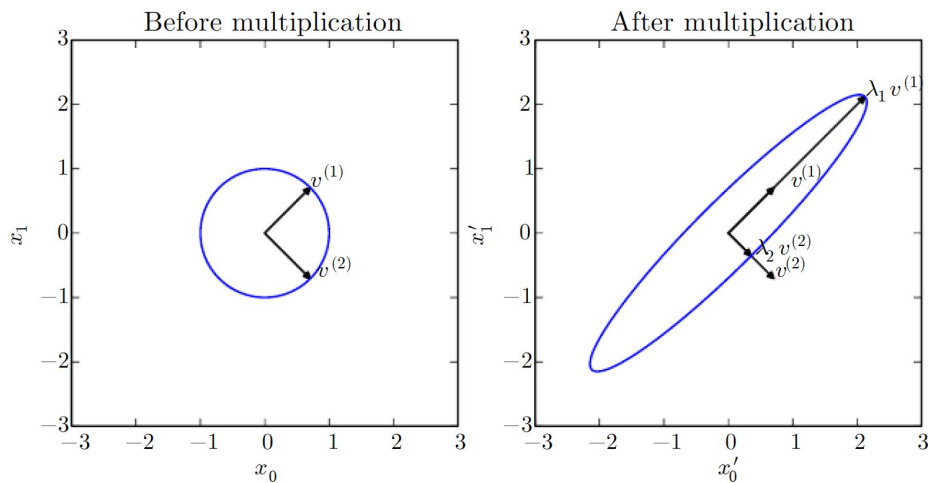
polynomial of degree n in  $\lambda$

# Eigenvectors of a Transformation

Matrix  $A$  acts by stretching the vector  $x$ , not changing its direction, so  $x$  is an eigenvector of  $A$ .



- An example of the effect of eigenvectors and eigenvalues. Here, we have a matrix  $A$  with two orthonormal eigenvectors,  $v^{(1)}$  with eigenvalue  $\lambda_1$  and  $v^{(2)}$  with eigenvalue  $\lambda_2$ .
- (Left) We plot the set of all unit vectors  $u \in \mathbb{R}^2$  as a unit circle. (Right) We plot the set of all points  $Au$ .
- By observing the way that  $A$  distorts the unit circle, we can see that it scales space in direction  $v^{(i)}$  by  $\lambda_i$ .



# Theorem

If the matrix  $A$  is positive definite then the eigenvalues of  $\lambda_1, \lambda_2, \dots, \lambda_n$ , are positive.



# Diagonalization

- $P^{-1}AP=D$  (D is diagonal)
- $A=PDP^{-1}$ 
  - The column vectors of P are the eigenvectors of A
  - The diagonal entries of D are the corresponding eigenvalues of A

$$A = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix}$$

$$\det(A - \lambda I) = \begin{vmatrix} 2 - \lambda & 2 \\ 1 & 3 - \lambda \end{vmatrix} = \lambda^2 - 5\lambda + 4$$

$$\lambda^2 - 5\lambda + 4 = 0 \longrightarrow \begin{cases} \lambda = 4 \\ \lambda = 1 \end{cases}$$

$$(A - I)v = 0$$

$$\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\left. \begin{array}{l} x + 2y = 0 \\ x + 2y = 0 \end{array} \right\} \longrightarrow x = -2y$$

$$v = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$$

$$(A - 4I)v = 0$$

$$\begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\left. \begin{array}{l} -2x + 2y = 0 \\ x - y = 0 \end{array} \right\} \longrightarrow y = x$$

$$v = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

- $A = PDP^{-1}$

$$P = \begin{pmatrix} -2 & 1 \\ 1 & 1 \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \times \begin{bmatrix} -1/3 & 1/3 \\ 1/3 & 2/3 \end{bmatrix}$$

# Singular Value Decomposition (SVD)

# Singular Value Decomposition (SVD) Method

- SVD is a factorization of a real or complex matrix.
- It generalizes the eigendecomposition of a square normal matrix with an orthonormal eigenbasis to any  $M \times N$  matrix.

$$\vec{x} = \vec{U} \cdot \vec{\Sigma} \cdot \vec{V}^T$$

where  $\vec{U}$  and  $\vec{V}$  are unitary matrices—i.e.,  $\vec{U}^T = \vec{U}^{-1}$  and  $\vec{V}^T = \vec{V}^{-1}$ , so  $\vec{V} \cdot \vec{V}^T = \vec{V} \cdot \vec{V}^{-1} = \vec{I}$  and the same for  $\vec{U}$ , hence their columns form orthonormal bases; and  $\vec{\Sigma}$  is diagonal with the “singular values” of  $\vec{X}$  in descending order.

# Singular Value Decomposition (SVD) Method

- The diagonal entries  $\sigma_i = \Sigma_{ii}$  of  $\Sigma$  are uniquely determined by M and are known as the singular values of M.
- The number of non-zero singular values is equal to the rank of x.

$$\vec{x} = \vec{U} \cdot \vec{\Sigma} \cdot \vec{V}^T$$



# SVD

Animated illustration of the SVD of a 2D, real **shearing matrix**  $\mathbf{M}$ . First, we see the **unit disc** in blue together with the two **canonical unit vectors**. We then see the actions of  $\mathbf{M}$ , which distorts the disk to an **ellipse**. The SVD decomposes  $\mathbf{M}$  into three simple transformations: an initial **rotation**  $\mathbf{V}^*$ , a **scaling**  $\mathbf{\Sigma}$  along the coordinate axes, and a final rotation  $\mathbf{U}$ . The lengths  $\sigma_1$  and  $\sigma_2$  of the **semi-axes** of the ellipse are the **singular values** of  $\mathbf{M}$ , namely  $\Sigma_{1,1}$  and  $\Sigma_{2,2}$ .

# SVD: EXAMPLE

Singular values can be considered as the importance values of different attributes in the data matrix.

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{5} & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{V}^* = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ -\sqrt{0.2} & 0 & 0 & 0 & -\sqrt{0.8} \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix}$$

## SVD: EXAMPLE

$$\mathbf{V}^* = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ \sqrt{0.4} & 0 & 0 & \sqrt{0.5} & -\sqrt{0.1} \\ -\sqrt{0.4} & 0 & 0 & \sqrt{0.5} & \sqrt{0.1} \end{bmatrix}$$

# SVD

We know that

$$\vec{C} = \vec{V} \vec{\Lambda} \vec{V}^T$$

$$\text{Then } \vec{V} = \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & & v_p \\ | & | & & | \end{bmatrix} \text{ and } \vec{\Lambda} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix},$$

with  $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_p|$ .

# SVD

The columns of V are the principal directions

$$\text{Then } \vec{V} = \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & & v_p \\ | & | & & | \end{bmatrix} \text{ and } \vec{\Lambda} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix},$$

with  $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_p|$ .

$$\vec{C} = \frac{1}{n-1} \vec{x}^T \vec{x}$$

$$= \frac{1}{n-1} \vec{V} \vec{\Sigma} \vec{U}^T \vec{U} \vec{\Sigma} \vec{V}^T$$

$$\vec{U}^T \vec{U} = \vec{I}$$

$$= \frac{1}{n-1} \vec{V} \vec{\Sigma}^2 \vec{V}^T$$

$$\vec{C} = \vec{V} \vec{\Lambda} \vec{V}^T$$

$$= \vec{V} \left( \frac{\vec{\Sigma}^2}{n-1} \right) \vec{V}^T$$

The columns of V are principal directions.

$$\frac{\Sigma_{i,i}^2}{n-1} = \lambda_i$$

Relationship between the eigenvalues of covariance matrix and singular values

# Relationship between SVD and PCA

- Both Dimensionality reduction
- Both PCA and SVD are methods of Matrix Decomposition
- You can calculate the PCA using SVD or using the eigendecomposition of the covariance matrix
- SVD also extracts data in the directions with the highest variances similar to PCA

# Relationship between SVD and PCA

## PCA: Relationship to SVD

Singular value decomposition:

$$\underset{N \times D}{\underline{X}} = \underset{N \times N}{\underline{U}} \underset{N \times D}{\underline{S}} \underset{D \times D}{\underline{V}^T}$$

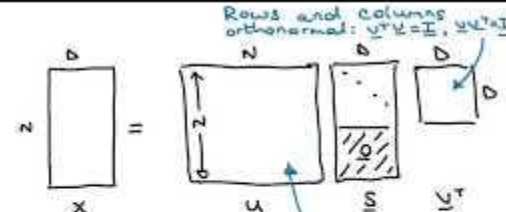
Relationship to PCA:

Take SVD of the design matrix  $\underline{X}$ :

$$\underline{X} = \underline{U} \underline{S} \underline{V}^T$$

$$\begin{aligned} \text{Then } \underline{X}^T \underline{X} &= \underline{V} \underline{S}^T \underline{U}^T \underline{U} \underline{S} \underline{V}^T \\ &= \underline{V} \underline{S}^T \underline{S} \underline{V}^T = \underline{V} \underline{D} \underline{V}^T \end{aligned}$$

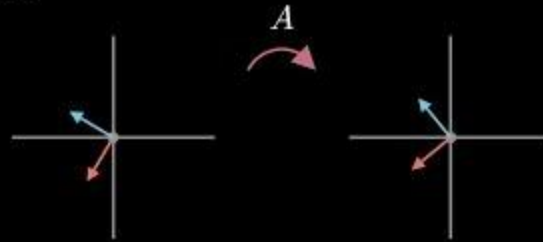
$$(\underline{X}^T \underline{X}) \underline{V} = \underline{V} \underline{D}$$



$\underline{D} = \underline{S}^T \underline{S}$  Diagonal with squares of singular values

$$\underline{X}^T \underline{X} \underline{V} = \underline{V} \underline{D}$$

What is SVD?



$$Av_1 = y_1$$

$$Av_2 = y_2$$

<https://www.youtube.com/watch?v=CpD9XITu3ys>



# SVD Example

[https://www.d.umn.edu/~mhampton/m4326svd\\_example.pdf](https://www.d.umn.edu/~mhampton/m4326svd_example.pdf)

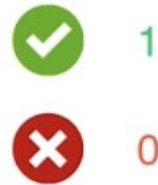
# Movie Recommender System Using SVD



# User/Item Matrix



				
John 	5	1	3	5
Tom 	?	?	?	2
Alice 	4	?	3	?





- For example comedy and action are two aspects to identify the taste of the users in movies
- Factorization discovers these dimensions automatically (latent space), even though does not know the label for these dimensions

## Matrix Factorization



	 Comedy	 Action
A	1	0
B	0	1
C	1	0
D	1	1

	M1	M2	M3	M4	M5
 Comedy	3	1	1	3	1
 Action	1	2	4	1	3

	M1	M2	M3	M4	M5
	3	1	1	3	1
	1	2	4	1	3
	3	1	1	3	1
	4	3	5	4	4

We can fill the missing ratings for each user by factoring the matrix

	M1	M2	M3	M4	M5
A	3		1		1
B	1		4	1	
C	3	1		3	1
D		3		4	4

# Applying SVD on User/Item Matrix

$$M (5 \times 5) = U (5 \times 5) \Sigma (5 \times 5) V^T (5 \times 5)$$

1	1	1	0	0
2	2	2	0	0
1	1	1	0	0
5	5	5	0	0
0	0	0	2	2

0.18	0	0.78	0.5	0.13
0.36	0	0	0.12	0.13
0.18	0	0	0	0
0.9	0	0	0	0
0	0.53	0.01	0	0


 $\times$ 

9.64	0	0	0	0
0	5.29	0	0	0
0	0	0.32	0	0
0	0	0	0.12	0
0	0	0	0	0.1

 $\times$ 

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71
0	0	0.8	0	0
0.6	0.8	0.1	0	0.8
0.54	0.54	0.03	0.2	0.8

$$= U (5 \times 2) \Sigma (2 \times 2) V^T (2 \times 5)$$



User Data

0.18	0
0.36	0
0.18	0
0.90	0
0	0.53


$\Sigma$  (2 x 2)

9.64	0
0	5.29





$V^T$  (2 x 5)

0.58	0.58	0.58	0	0
0	0	0	0.71	0.71

Product Data



# Movie Recommendation Using SVD

	M1	M2	M3	M4	M5
	3	1	1	3	1
	1	2	4	1	3
	3	1	1	3	1
	4	3	5	4	4


Matrix  
Factorization

<https://www.youtube.com/watch?v=ZspR5PZemcs>

# Movie Recommendation Using SVD


Content Based- Movie recommendation

$(9, 0, -6)$



Nikhil

Movies	Reviews Given	Rating
Mission Impossible	✓ ✓	Good
James Bond	✓ ✓	Good
Toy Story	✗ ✓	Bad



Children (1-6)   Animation (0)   Action (9)

Rating System

User likes Action Movies and doesn't like Animation/Children Genre movies

- $(9, 0, -6)$  in order of (Action, Animation, Children)
- Toy Story  $-(0, 1, 1)$  and Star Wars  $-(1, 0, 0)$
- Take the dot product
- Recommend Star Wars

$7.5 + (0, 1, 1) \cdot (-9, 0, -6) = 7.5 - 6 = 1.5$

$5 - W \rightarrow (1, 0, 0) \cdot (-9, 0, -6) = -9$

$(-9, 0, -6) \cdot (0, 1, 1) = -6$

<https://www.youtube.com/watch?v=rFemvJgXY7E>



# Movie Recommendation Using SVD

<https://towardsdatascience.com/beginners-guide-to-creating-an-svd-recommender-system-1fd7326d1f65>

# The Moore-Penrose Pseudoinverse

- Suppose we want to solve

$$Ax=B$$

- Matrix inversion is not defined for matrices that are not square.
- If  $A$  is taller than it is wide, then it is possible for this equation to have no solution. If  $A$  is wider than it is tall, then there could be multiple possible solutions. The Moore-Penrose pseudoinverse allows us to make some headway in these cases.

# The Moore-Penrose Pseudoinverse

- The pseudoinverse of  $A$  is defined as a matrix

$$\mathbf{A}^+ = \lim_{\alpha \searrow 0} (\mathbf{A}^\top \mathbf{A} + \alpha \mathbf{I})^{-1} \mathbf{A}^\top.$$

- Practical algorithms for computing the pseudoinverse are not based on this definition, but rather the formula

$$\mathbf{A}^+ = \mathbf{V} \mathbf{D}^+ \mathbf{U}^\top,$$

- where  $\mathbf{U}$ ,  $\mathbf{D}$  and  $\mathbf{V}$  are the singular value decomposition of  $A$ , and the pseudoinverse  $\mathbf{D}^+$  of a diagonal matrix  $\mathbf{D}$  is obtained by taking the reciprocal of its non-zero elements then taking the transpose of the resulting matrix.

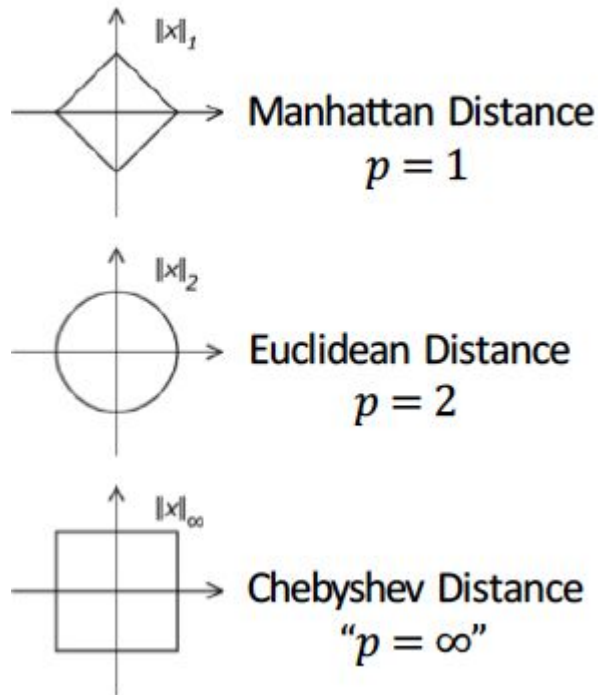
Norms

# Norms

- Norms, including the  $L_p$  norm, are functions mapping vectors to non-negative values. On an intuitive level, the norm of a vector  $x$  measures the distance from the origin to the point  $x$ . More rigorously, a norm is any function  $f$  that satisfies the following properties:
  - $f(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \mathbf{0}$
  - $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$  (the **triangle inequality**)
  - $\forall \alpha \in \mathbb{R}, f(\alpha \mathbf{x}) = |\alpha|f(\mathbf{x})$

# Distance Measure

- The squared  $L^2$  norm is more convenient to work with mathematically and computationally than the  $L^2$  norm itself. For example, the derivatives of the squared  $L^2$  norm with respect to each element of  $x$  each depend only on the corresponding element of  $x$ , while all of the derivatives of the  $L^2$  norm depend on the entire vector. In many contexts, the squared  $L^2$  norm may be undesirable because it increases very slowly near the origin. In several machine learning applications, it is important to discriminate between elements that are exactly zero and elements that are small but nonzero.
- The  $L^1$  norm is commonly used in machine learning when the difference between zero and nonzero elements is very important. Every time an element of  $x$  moves away from 0 by epsilon, the  $L^1$  norm increases by epsilon.



# Euclidean Distance

$$d(\vec{x}, \vec{y})_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

# Manhattan Distance

$$d(\vec{x}, \vec{y})_1 = \|\vec{x} - \vec{y}\|_1 = \sum_{i=1}^n |x_i - y_i|$$






# Chebyshev Distance

$$d_{\max}(\vec{x}, \vec{y}) = \lim_{p \rightarrow \infty} \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}$$
$$= \max_i |x_i - y_i|$$

- The minimum number of moves a king requires to move between two squares.
- This is because a king can move diagonally, so that the jumps to cover the smaller distance parallel to a rank or column is effectively absorbed into the jumps covering the larger.
- In this image you can see the Chebyshev distances of each square from the square f6.

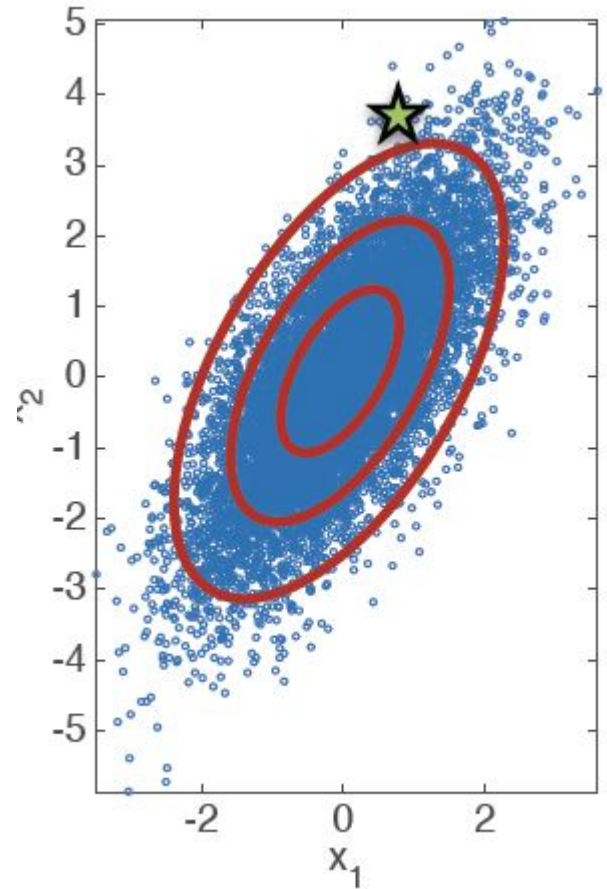
	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

- The dot product of two vectors can be rewritten in terms of norms

$$\mathbf{x}^\top \mathbf{y} = ||\mathbf{x}||_2 ||\mathbf{y}||_2 \cos \theta$$

# Mahalanobis Distance

$$d_{\text{Mahalanobis}}(\vec{x}, \vec{\mu}) = \sqrt{(\vec{x} - \vec{\mu})^T \vec{S}^{-1} (\vec{x} - \vec{\mu})}$$



# Frobenius Norm

- Sometimes we may also wish to measure the size of a matrix. In the context of deep learning, the most common way to do this is with the otherwise obscure Frobenius norm:

$$||A||_F = \sqrt{\sum_{i,j} A_{i,j}^2},$$

- Which is analogous to the L2 norm of a vector.

# Probability

# Basic Statistics Concepts

## Expected Value

$$\mathbf{E}[X] = \sum_{i=1}^k x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k.$$

## Expected Value: Example

- Let  $X$  represent the outcome of a roll of a fair six-sided die:

$$E[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$



# Example: Roulette Game

- 38 numbered pockets around the edge.
- Random variable  $X$  represents the (monetary) outcome of a \$1 bet on a single number ("straight up" bet).
- If the bet wins ( probability  $= 1 / 38$ ), the payoff is \$35; otherwise the player loses the bet.
- Calculate the Expected Profit from this game?



# Example: Roulette Game

- 38 numbered pockets around the edge.
- Random variable  $X$  represents the (monetary) outcome of a \$1 bet on a single number ("straight up" bet).
- If the bet wins ( probability =  $1 / 38$ ), the payoff is \$35; otherwise the player loses the bet.
- Calculate the Expected Profit from this game?



$$E[\text{gain from \$1 bet}] = -\$1 \cdot \frac{37}{38} + \$35 \cdot \frac{1}{38} = -\$ \frac{1}{19}.$$

# Variance

Measure of dispersion

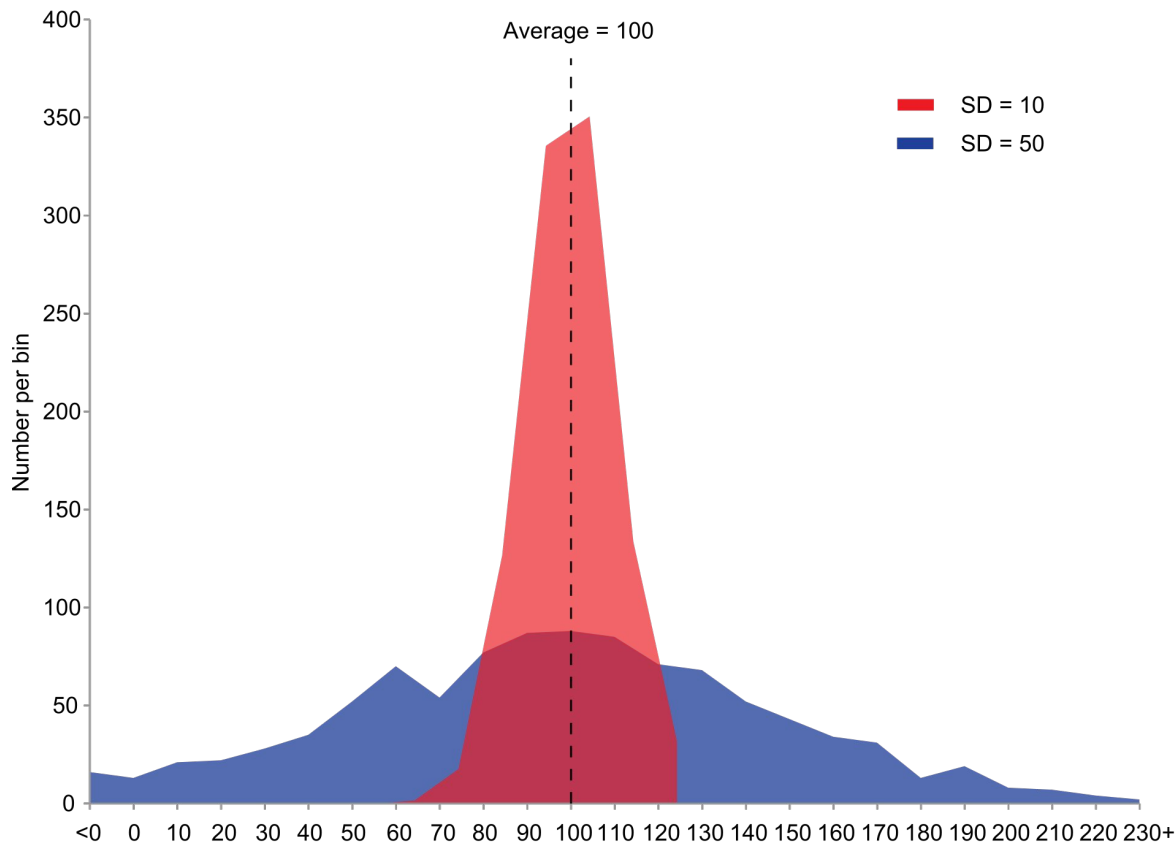
$$\mu = E[X]:$$

$$\text{Var}(X) = E[(X - \mu)^2]$$

If all values are equally likely:

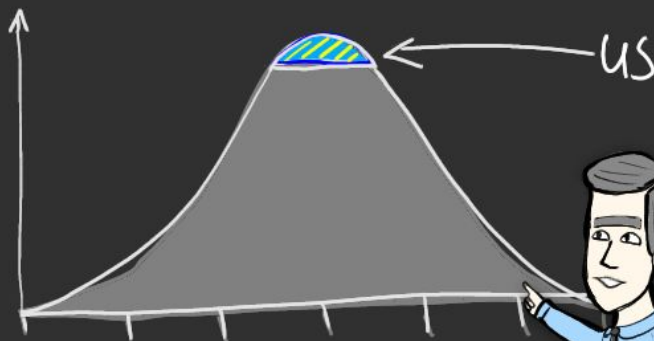
$$\mu = \frac{1}{n} \sum_{i=1}^n x_i.$$

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$



A120K

we are in the top 1 percent



AS YOU CAN  
SEE, WE ARE THE  
BEST OF THE  
BEST

WOW!!

# Sample vs Population Variance

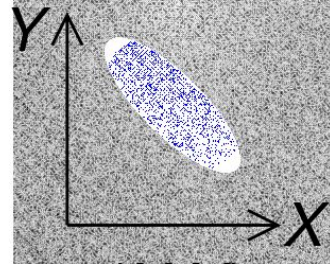
- If all possible observations of the system are present then the calculated variance is called the population variance.
- In practice, only a subset is available, and the variance calculated from this is called the sample variance. The variance calculated from a sample is considered an estimate of the full population variance.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

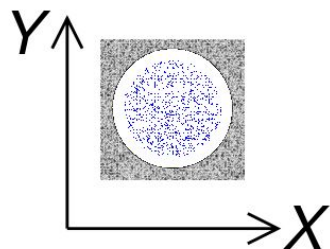
# Covariance

$$\text{cov}(X, Y) = \text{E} [(X - \text{E}[X])(Y - \text{E}[Y])]$$

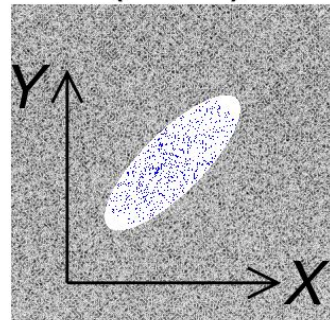
$$\text{cov}(X, X) = \text{var}(X) \equiv \sigma^2(X) \equiv \sigma_X^2.$$



$\text{cov}(X, Y) < 0$



$\text{cov}(X, Y) \approx 0$



$\text{cov}(X, Y) > 0$

# Variance/Covariance Matrix

$$\begin{matrix} & \begin{matrix} x & y \end{matrix} \\ \begin{matrix} x \\ y \end{matrix} & \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(x, y) & \text{var}(y) \end{bmatrix} \end{matrix}$$

$$\begin{matrix} & \begin{matrix} x & y & z \end{matrix} \\ \begin{matrix} x \\ y \\ z \end{matrix} & \begin{bmatrix} \text{var}(x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(x, y) & \text{var}(y) & \text{cov}(y, z) \\ \text{cov}(x, z) & \text{cov}(y, z) & \text{var}(z) \end{bmatrix} \end{matrix}$$



# Variance/Covariance Matrix

$$\begin{aligned}\text{Var}[X] &= E \begin{bmatrix} (X_1 - E[X_1])(X_1 - E[X_1]) & \dots & (X_1 - E[X_1])(X_K - E[X_K]) \\ \vdots & \ddots & \vdots \\ (X_K - E[X_K])(X_1 - E[X_1]) & \dots & (X_K - E[X_K])(X_K - E[X_K]) \end{bmatrix} \\ &= \begin{bmatrix} E[(X_1 - E[X_1])^2] & \dots & E[(X_1 - E[X_1])(X_K - E[X_K])] \\ \vdots & \ddots & \vdots \\ E[(X_K - E[X_K])(X_1 - E[X_1])] & \dots & E[(X_K - E[X_K])^2] \end{bmatrix} \\ &= \begin{bmatrix} \text{Var}[X_1] & \dots & \text{Cov}[X_1, X_K] \\ \vdots & \ddots & \vdots \\ \text{Cov}[X_K, X_1] & \dots & \text{Var}[X_K] \end{bmatrix}\end{aligned}$$



# Momentums: First and Second

- If the function is a probability distribution then
  - The first moment is the expected value
  - The second central moment is the variance
  - The third standardized moment is the skewness
  - The fourth standardized moment is the kurtosis.
- The mathematical concept is closely related to the concept of moment in physics.
- The n-th raw moment (i.e., moment about zero) of a random variable

$$\mu'_n = \langle X^n \rangle \stackrel{\text{def}}{=} \begin{cases} \sum_i x_i^n f(x_i), & \text{discrete distribution} \\ \int x^n f(x) dx, & \text{continuous distribution} \end{cases}$$

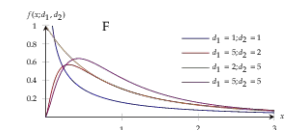
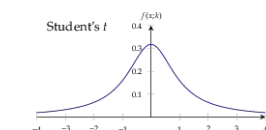
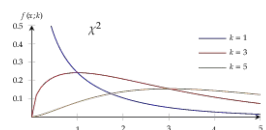
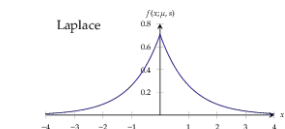
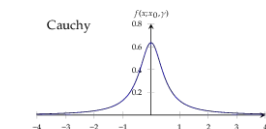
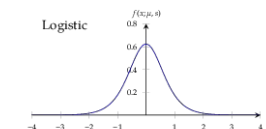
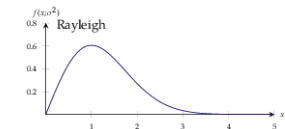
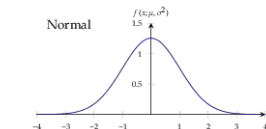
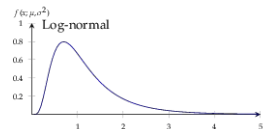
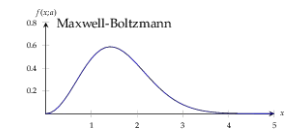
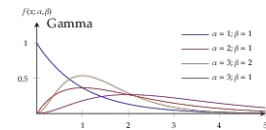
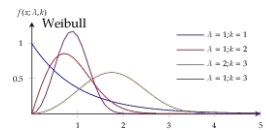
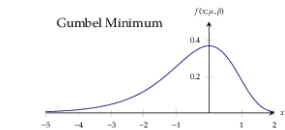
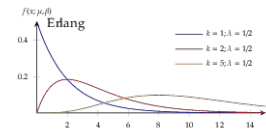
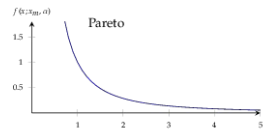
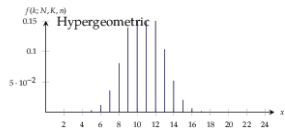
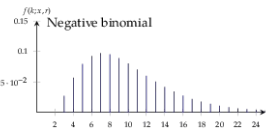
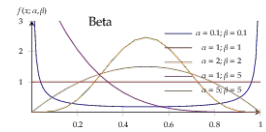
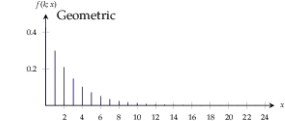
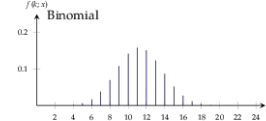
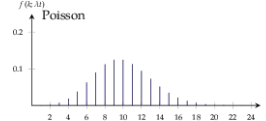
- The n-th moment of a real-valued continuous random variable with density function  $f(x)$

$$\mu_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx.$$

# Probability Distributions

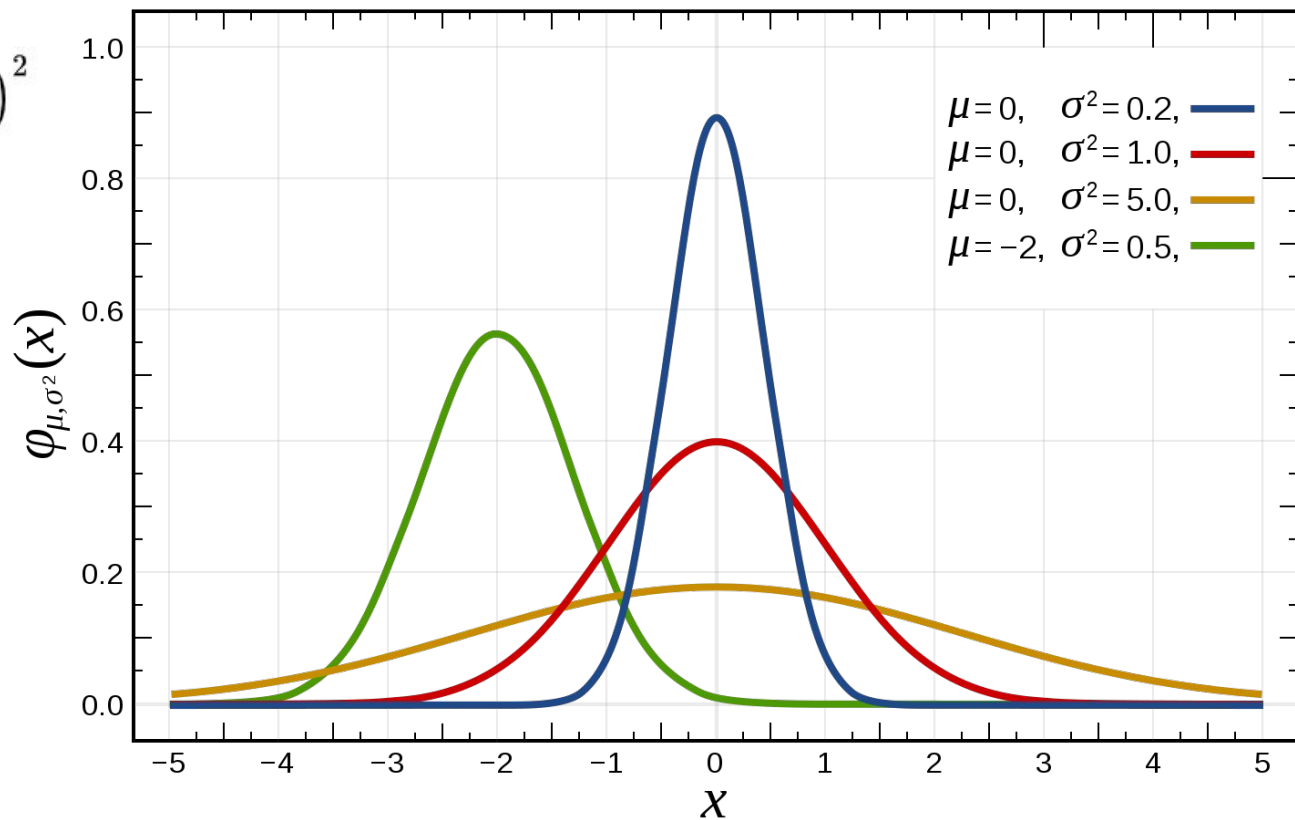
# Probability Distributions

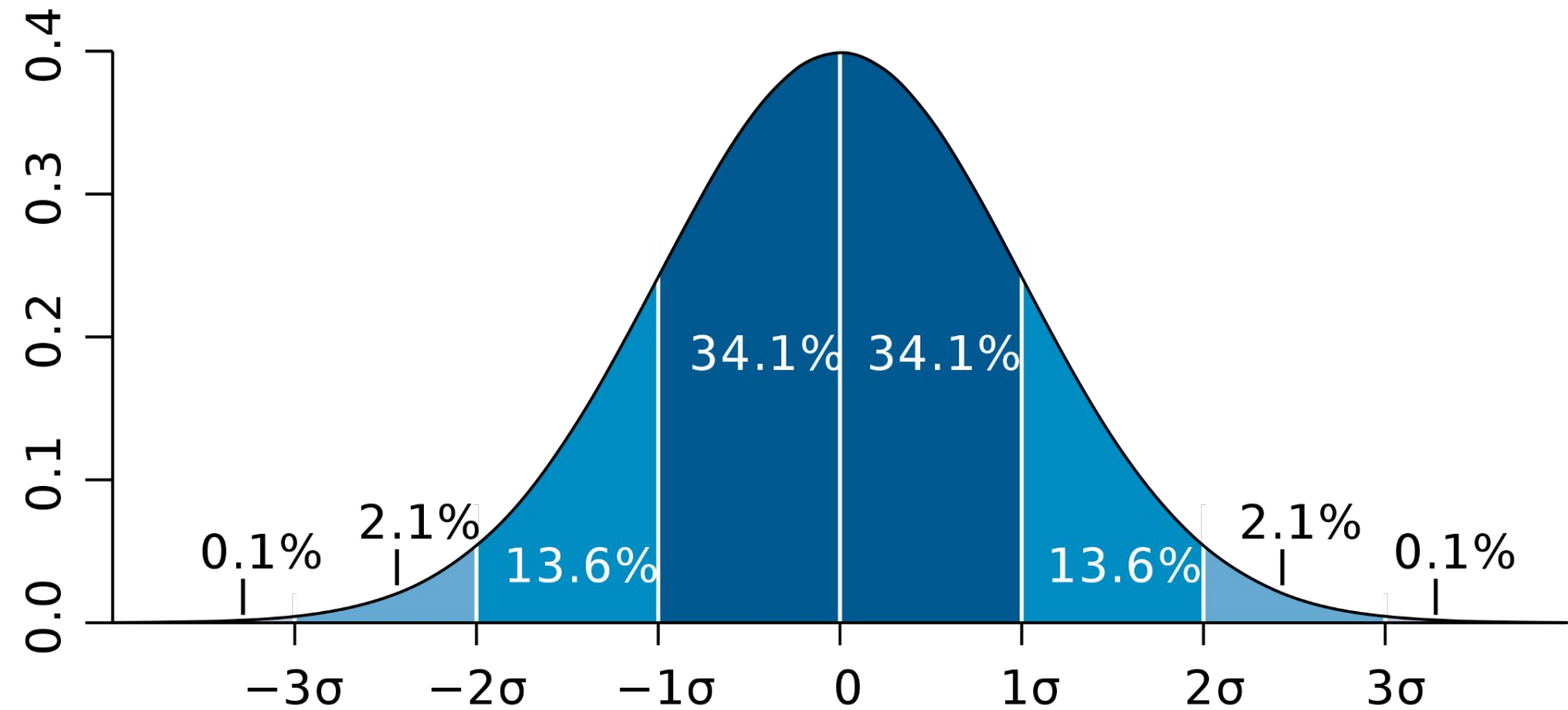
- The mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment

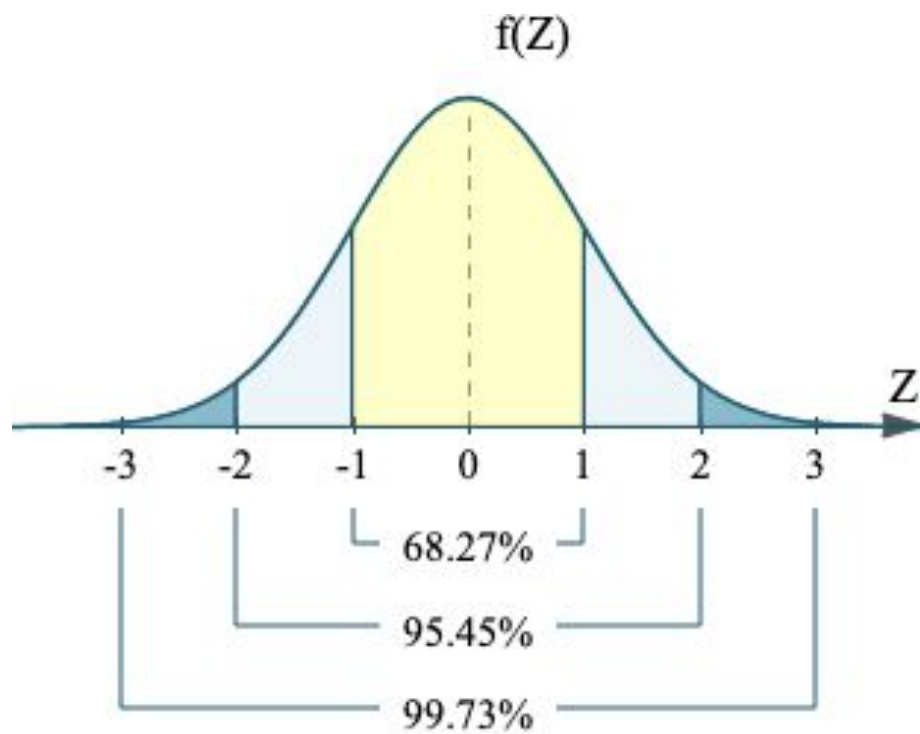


# Normal Distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

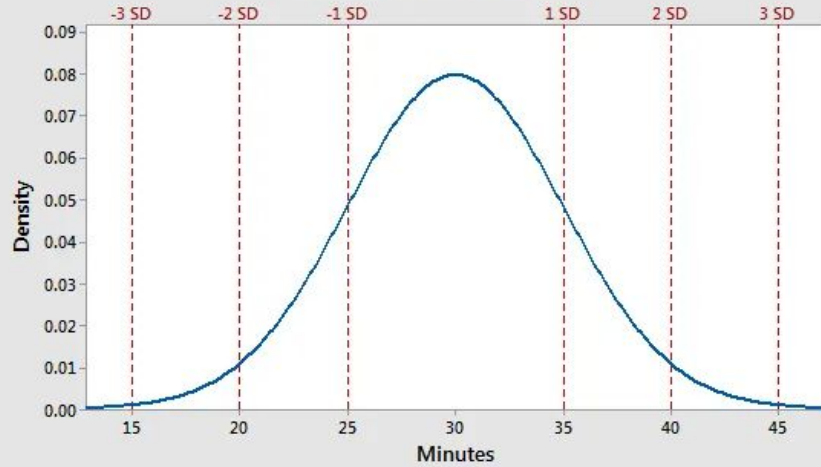






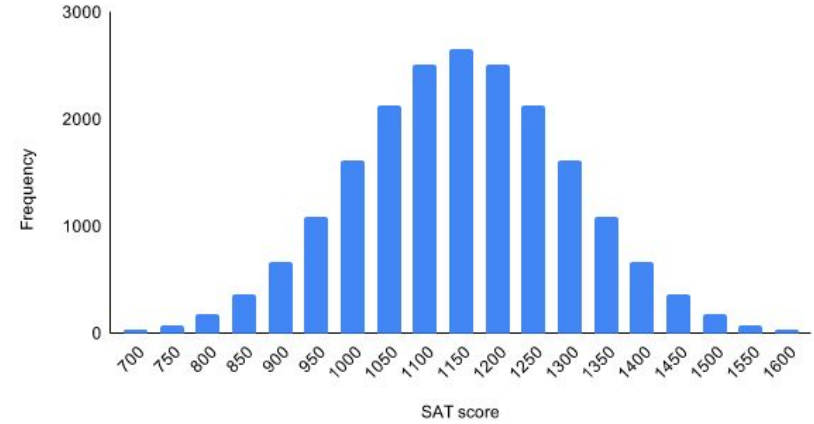
### Distribution of Pizza Delivery Times

Normal, Mean=30, StDev=5



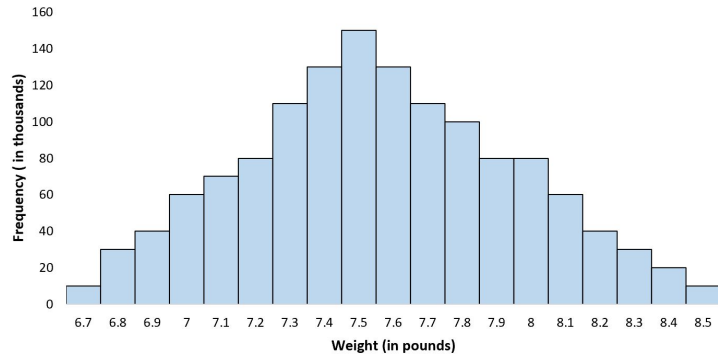
<https://statisticsbyjim.com/basics/normal-distribution/>

### SAT scores in 2020

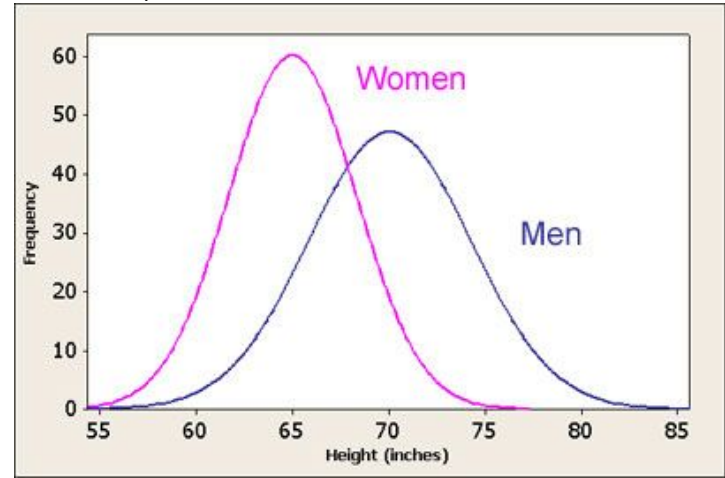


<https://www.scribbr.com/statistics/normal-distribution/>

### Distribution of Newborn Weights



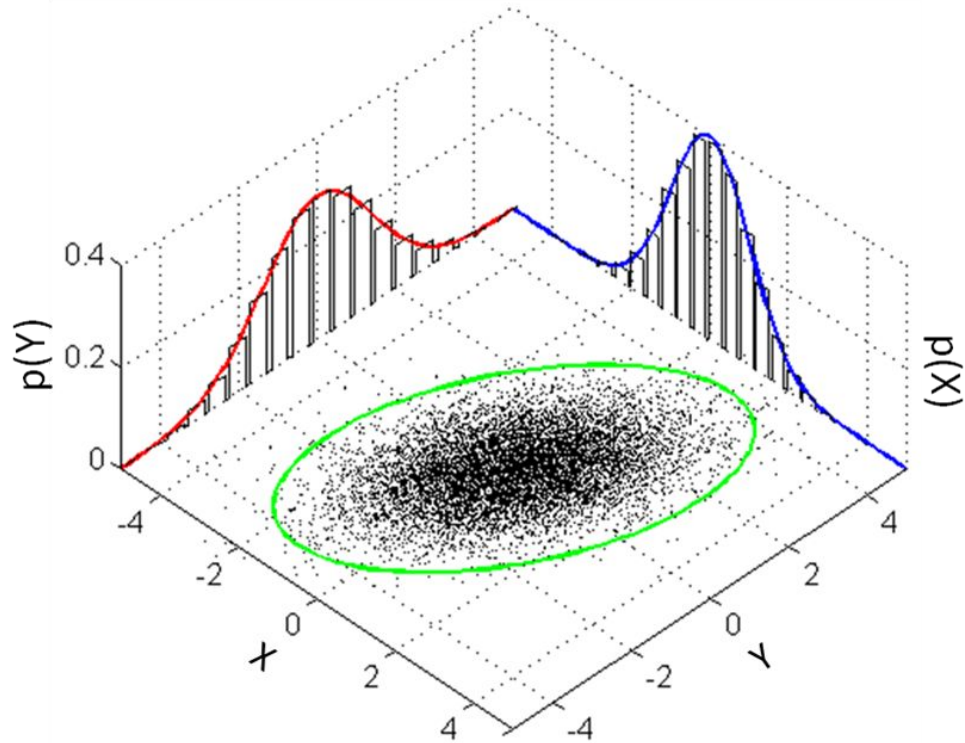
<https://www.statology.org/example-of-normal-distribution/>



# Multivariate Normal Distribution

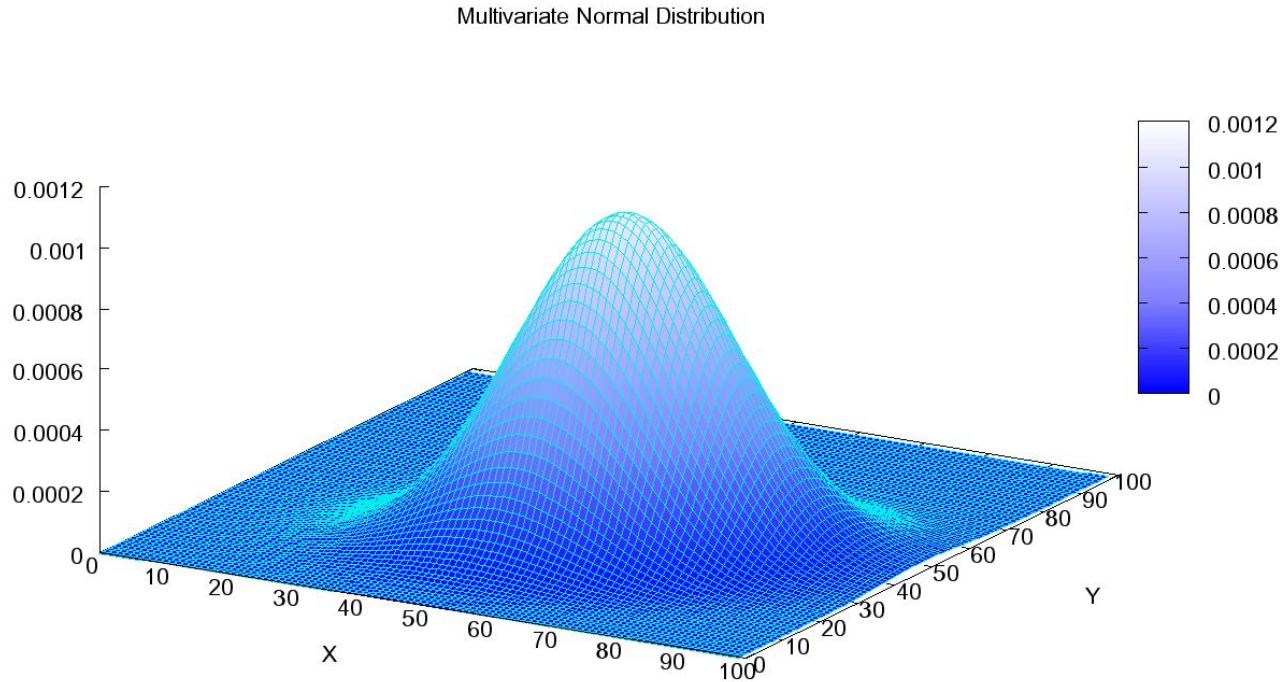
- Generalization of the one-dimensional (univariate) normal distribution to higher dimensions.

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix},$$





# Multivariate Normal Distribution



# Multivariate Normal Distribution

$$N(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

Parameter Estimation: Maximum Likelihood

# Parameter Estimation

- Two methods
  - Maximum Likelihood
    - the parameters values are fixed but unknown.
    - The best estimate of their value is defined to be the one that maximizes the probability of obtaining the samples actually observed.
  - Bayesian
    - The parameters are considered as random variables having some known a priori distribution.
    - Observation of the samples converts this to a posterior density, thereby revising our opinion about the true values of the parameters (Bayesian learning)
- The results from both are nearly identical but different approaches

# Maximum Likelihood Estimation

- Simpler than other methods
- Good convergence as the number of samples increase

# Maximum Likelihood

- Suppose that  $\mathcal{D}$  contains  $n$  samples,  $\mathbf{x}_1, \dots, \mathbf{x}_n$ . Then, since the samples were drawn independently, we have

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k|\boldsymbol{\theta}).$$

- $p(\mathcal{D}|\boldsymbol{\theta})$  is called the likelihood of  $\boldsymbol{\theta}$  with respect to the set of samples.
- The maximum likelihood estimate of  $\boldsymbol{\theta}$  is, by definition, the value  $\hat{\boldsymbol{\theta}}$  that maximizes  $p(\mathcal{D}|\boldsymbol{\theta})$ .
  - This estimate corresponds to the value of  $\boldsymbol{\theta}$  that in some sense best agrees with or supports the actually observed training samples

# Likelihood Example

- Flipping a coin multiple time, what is the likelihood of observing  $x_1, x_2, \dots, x_n$
- Bernoulli distribution  $p_{\theta}(x) = p_p(x) = p^x(1 - p)^{(1-x)}$

$$\mathcal{L}_n(X_1, X_2, \dots, X_n, \theta) = \prod_{i=1}^n p_{\theta}(x_i)$$

$$\Rightarrow \mathcal{L}_n(X_1, X_2, \dots, X_n, p) = \prod_{i=1}^n p_p(x_i) = \prod_{i=1}^n p^{x_i}(1 - p)^{(1-x_i)}$$

$$\Rightarrow \mathcal{L}_n(X_1, X_2, \dots, X_n, p) = p^{\sum x_i}(1 - p)^{(n - \sum x_i)}$$

- Example biased coin  $\theta=0.7$ , likelihood of observing 1,1,1,0:

$$L=0.7^3 \times 0.3^1$$

# Log likelihood

- It is usually easier to work with the logarithm of the likelihood than with the likelihood itself.
- Since the logarithm is monotonically increasing, the  $\hat{\theta}$  that maximizes the log-likelihood also maximizes the likelihood.
- If  $p(D|\theta)$  is a well behaved, differentiable function of  $\theta$ ,  $\hat{\theta}$  can be found using the differential calculus.



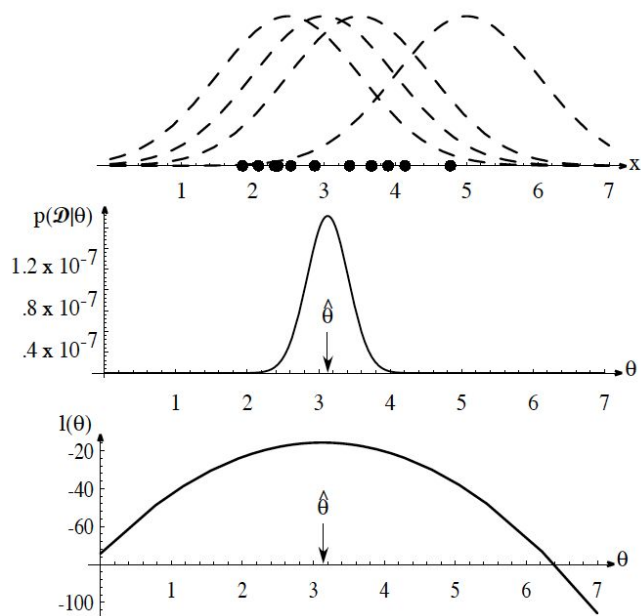


Figure 3.1: The top graph shows several training points in one dimension, known or assumed to be drawn from a Gaussian of a particular variance, but unknown mean. Four of the infinite number of candidate source distributions are shown in dashed lines. The middle figures shows the likelihood  $p(\mathcal{D}|\theta)$  as a function of the mean. If we had a very large number of training points, this likelihood would be very narrow. The value that maximizes the likelihood is marked  $\hat{\theta}$ ; it also maximizes the logarithm of the likelihood — i.e., the log-likelihood  $l(\theta)$ , shown at the bottom. Note especially that the likelihood lies in a different space from  $p(x|\hat{\theta})$ , and the two can have different functional forms.

For analytical purposes, it is usually easier to work with the logarithm of the likelihood than with the likelihood itself. Since the logarithm is monotonically increasing, the  $\hat{\theta}$  that maximizes the log-likelihood also maximizes the likelihood. If  $p(\mathcal{D}|\theta)$  is a well behaved, differentiable function of  $\theta$ ,  $\hat{\theta}$  can be found by the standard methods of differential calculus. If the number of parameters to be set is  $p$ , then we let  $\theta$  denote

# Maximum Likelihood Estimation

- if the number of parameters to be set is  $p$ , then we let  $\theta$  denote the  $p$ -component vector  $\theta = (\theta_1, \dots, \theta_p)^t$  and  $\nabla_{\theta}$  be the gradient operator

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix} .$$

$$l(\boldsymbol{\theta}) \equiv \ln p(\mathcal{D}|\boldsymbol{\theta}).$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^n p(\mathbf{x}_k|\boldsymbol{\theta}). \quad \Rightarrow \quad l(\boldsymbol{\theta}) = \sum_{k=1}^n \ln p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\boxed{\nabla_{\boldsymbol{\theta}} l = \mathbf{0}.}$$

- A solution  $\hat{\theta}$  estimated this way could represent a true global maximum, a local maximum or minimum, or (rarely) an inflection point of  $l(\theta)$ .
- We have to check if the extremum occurs at a boundary of the parameter space
- If all solutions are found, we are guaranteed that one represents the true maximum, though we might have to check each solution individually (or calculate second derivatives) to identify which is the global optimum.
- $\hat{\theta}$  is an estimate
  - it is only in the limit of an infinitely large number of training points that we can expect that our estimate will equal to the true value of the generating function