

# Quick Tour of Basic Linear Algebra and Probability Theory

CS246: Mining Massive Data Sets  
Winter 2011

# Outline

**1** Basic Linear Algebra

2 Basic Probability Theory

# Matrices and Vectors

- Matrix: A rectangular array of numbers, e.g.,  $A \in \mathbb{R}^{m \times n}$ :

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

# Matrices and Vectors

- Matrix: A rectangular array of numbers, e.g.,  $A \in \mathbb{R}^{m \times n}$ :

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

- Vector: A matrix consisting of only one column (default) or one row, e.g.,  $x \in \mathbb{R}^n$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

# Matrix Multiplication

- If  $A \in \mathbb{R}^{m \times n}$ ,  $B \in \mathbb{R}^{n \times p}$ ,  $C = AB$ , then  $C \in \mathbb{R}^{m \times p}$ :

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

# Matrix Multiplication

- If  $A \in \mathbb{R}^{m \times n}$ ,  $B \in \mathbb{R}^{n \times p}$ ,  $C = AB$ , then  $C \in \mathbb{R}^{m \times p}$ :

$$C_{ij} = \sum_{k=1}^n A_{ik} B_{kj}$$

- Special cases: Matrix-vector product, inner product of two vectors. e.g., with  $x, y \in \mathbb{R}^n$ :

$$x^T y = \sum_{i=1}^n x_i y_i \in \mathbb{R}$$

# Properties of Matrix Multiplication

- Associative:  $(AB)C = A(BC)$

# Properties of Matrix Multiplication

- Associative:  $(AB)C = A(BC)$
- Distributive:  $A(B + C) = AB + AC$

# Properties of Matrix Multiplication

- Associative:  $(AB)C = A(BC)$
- Distributive:  $A(B + C) = AB + AC$
- Non-commutative:  $AB \neq BA$

# Properties of Matrix Multiplication

- Associative:  $(AB)C = A(BC)$
- Distributive:  $A(B + C) = AB + AC$
- Non-commutative:  $AB \neq BA$
- Block multiplication: If  $A = [A_{ik}]$ ,  $B = [B_{kj}]$ , where  $A_{ik}$ 's and  $B_{kj}$ 's are matrix blocks, and the number of columns in  $A_{ik}$  is equal to the number of rows in  $B_{kj}$ , then  $C = AB = [C_{ij}]$  where  $C_{ij} = \sum_k A_{ik} B_{kj}$

# Properties of Matrix Multiplication

- Associative:  $(AB)C = A(BC)$
- Distributive:  $A(B + C) = AB + AC$
- Non-commutative:  $AB \neq BA$
- Block multiplication: If  $A = [A_{ik}]$ ,  $B = [B_{kj}]$ , where  $A_{ik}$ 's and  $B_{kj}$ 's are matrix blocks, and the number of columns in  $A_{ik}$  is equal to the number of rows in  $B_{kj}$ , then  $C = AB = [C_{ij}]$

where  $C_{ij} = \sum_k A_{ik} B_{kj}$

**Example:** If  $\vec{x} \in \mathbb{R}^n$  and  $A = [\vec{a}_1 | \vec{a}_2 | \dots | \vec{a}_n] \in \mathbb{R}^{m \times n}$ ,

$B = [\vec{b}_1 | \vec{b}_2 | \dots | \vec{b}_p] \in \mathbb{R}^{n \times p}$ :

$$A\vec{x} = \sum_{i=1}^n x_i \vec{a}_i$$

$$AB = [A\vec{b}_1 | A\vec{b}_2 | \dots | A\vec{b}_p]$$

# Operators and properties

- Transpose:  $A \in \mathbb{R}^{m \times n}$ , then  $A^T \in \mathbb{R}^{n \times m}$ :  $(A^T)_{ij} = A_{ji}$

# Operators and properties

- Transpose:  $A \in \mathbb{R}^{m \times n}$ , then  $A^T \in \mathbb{R}^{n \times m}$ :  $(A^T)_{ij} = A_{ji}$
- Properties:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$

# Operators and properties

- Transpose:  $A \in \mathbb{R}^{m \times n}$ , then  $A^T \in \mathbb{R}^{n \times m}$ :  $(A^T)_{ij} = A_{ji}$
- Properties:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$
- Trace:  $A \in \mathbb{R}^{n \times n}$ , then:  $tr(A) = \sum_{i=1}^n A_{ii}$

# Operators and properties

- Transpose:  $A \in \mathbb{R}^{m \times n}$ , then  $A^T \in \mathbb{R}^{n \times m}$ :  $(A^T)_{ij} = A_{ji}$
- Properties:
  - $(A^T)^T = A$
  - $(AB)^T = B^T A^T$
  - $(A + B)^T = A^T + B^T$
- Trace:  $A \in \mathbb{R}^{n \times n}$ , then:  $tr(A) = \sum_{i=1}^n A_{ii}$
- Properties:
  - $tr(A) = tr(A^T)$
  - $tr(A + B) = tr(A) + tr(B)$
  - $tr(\lambda A) = \lambda tr(A)$
  - If  $AB$  is a square matrix,  $tr(AB) = tr(BA)$

# Special types of matrices

- Identity matrix:  $I = I_n \in \mathbb{R}^{n \times n}$ :

$$I_{ij} = \begin{cases} 1 & i=j, \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall A \in \mathbb{R}^{m \times n}$ :  $AI_n = I_m A = A$

# Special types of matrices

- Identity matrix:  $I = I_n \in \mathbb{R}^{n \times n}$ :

$$I_{ij} = \begin{cases} 1 & i=j, \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall A \in \mathbb{R}^{m \times n}$ :  $AI_n = I_m A = A$
- Diagonal matrix:  $D = \text{diag}(d_1, d_2, \dots, d_n)$ :

$$D_{ij} = \begin{cases} d_i & j=i, \\ 0 & \text{otherwise.} \end{cases}$$

# Special types of matrices

- Identity matrix:  $I = I_n \in \mathbb{R}^{n \times n}$ :

$$I_{ij} = \begin{cases} 1 & i=j, \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall A \in \mathbb{R}^{m \times n}$ :  $AI_n = I_m A = A$
- Diagonal matrix:  $D = \text{diag}(d_1, d_2, \dots, d_n)$ :

$$D_{ij} = \begin{cases} d_i & j=i, \\ 0 & \text{otherwise.} \end{cases}$$

- Symmetric matrices:  $A \in \mathbb{R}^{n \times n}$  is symmetric if  $A = A^T$ .

# Special types of matrices

- Identity matrix:  $I = I_n \in \mathbb{R}^{n \times n}$ :

$$I_{ij} = \begin{cases} 1 & i=j, \\ 0 & \text{otherwise.} \end{cases}$$

- $\forall A \in \mathbb{R}^{m \times n}$ :  $AI_n = I_m A = A$
- Diagonal matrix:  $D = \text{diag}(d_1, d_2, \dots, d_n)$ :

$$D_{ij} = \begin{cases} d_i & j=i, \\ 0 & \text{otherwise.} \end{cases}$$

- Symmetric matrices:  $A \in \mathbb{R}^{n \times n}$  is symmetric if  $A = A^T$ .
- Orthogonal matrices:  $U \in \mathbb{R}^{n \times n}$  is orthogonal if  $UU^T = I = U^T U$

# Linear Independence and Rank

- A set of vectors  $\{x_1, \dots, x_n\}$  is linearly independent if  $\nexists \{\alpha_1, \dots, \alpha_n\}: \sum_{i=1}^n \alpha_i x_i = 0$

# Linear Independence and Rank

- A set of vectors  $\{x_1, \dots, x_n\}$  is linearly independent if  $\nexists \{\alpha_1, \dots, \alpha_n\}: \sum_{i=1}^n \alpha_i x_i = 0$
- Rank:  $A \in \mathbb{R}^{m \times n}$ , then  $\text{rank}(A)$  is the maximum number of linearly independent columns (or equivalently, rows)

# Linear Independence and Rank

- A set of vectors  $\{x_1, \dots, x_n\}$  is linearly independent if  $\nexists \{\alpha_1, \dots, \alpha_n\}: \sum_{i=1}^n \alpha_i x_i = 0$
- Rank:  $A \in \mathbb{R}^{m \times n}$ , then  $\text{rank}(A)$  is the maximum number of linearly independent columns (or equivalently, rows)
- Properties:
  - $\text{rank}(A) \leq \min\{m, n\}$
  - $\text{rank}(A) = \text{rank}(A^T)$
  - $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$
  - $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$

# Matrix Inversion

- If  $A \in \mathbb{R}^{n \times n}$ ,  $\text{rank}(A) = n$ , then the inverse of  $A$ , denoted  $A^{-1}$  is the matrix that:  $AA^{-1} = A^{-1}A = I$
- Properties:
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^{-1})^T = (A^T)^{-1}$

# Range and Nullspace of a Matrix

■ Span:  $span(\{x_1, \dots, x_n\}) = \{\sum_{i=1}^n \alpha_i x_i \mid \alpha_i \in \mathbb{R}\}$

# Range and Nullspace of a Matrix

- Span:  $span(\{x_1, \dots, x_n\}) = \{\sum_{i=1}^n \alpha_i x_i \mid \alpha_i \in \mathbb{R}\}$
- Projection:  
 $Proj(y; \{x_i\}_{1 \leq i \leq n}) = \underset{v \in span(\{x_i\}_{1 \leq i \leq n})}{\operatorname{argmin}} \{\|y - v\|_2\}$

# Range and Nullspace of a Matrix

- Span:  $span(\{x_1, \dots, x_n\}) = \{\sum_{i=1}^n \alpha_i x_i \mid \alpha_i \in \mathbb{R}\}$
- Projection:  
 $Proj(y; \{x_i\}_{1 \leq i \leq n}) = \operatorname{argmin}_{v \in span(\{x_i\}_{1 \leq i \leq n})} \{\|y - v\|_2\}$
- Range:  $A \in \mathbb{R}^{m \times n}$ , then  $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$  is the span of the columns of  $A$

# Range and Nullspace of a Matrix

- Span:  $span(\{x_1, \dots, x_n\}) = \{\sum_{i=1}^n \alpha_i x_i \mid \alpha_i \in \mathbb{R}\}$
- Projection:  
 $Proj(y; \{x_i\}_{1 \leq i \leq n}) = \operatorname{argmin}_{v \in span(\{x_i\}_{1 \leq i \leq n})} \{\|y - v\|_2\}$
- Range:  $A \in \mathbb{R}^{m \times n}$ , then  $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$  is the span of the columns of  $A$
- $Proj(y, A) = A(A^T A)^{-1} A^T y$

# Range and Nullspace of a Matrix

- Span:  $span(\{x_1, \dots, x_n\}) = \{\sum_{i=1}^n \alpha_i x_i \mid \alpha_i \in \mathbb{R}\}$
- Projection:  
 $Proj(y; \{x_i\}_{1 \leq i \leq n}) = \operatorname{argmin}_{v \in span(\{x_i\}_{1 \leq i \leq n})} \{\|y - v\|_2\}$
- Range:  $A \in \mathbb{R}^{m \times n}$ , then  $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\}$  is the span of the columns of  $A$
- $Proj(y, A) = A(A^T A)^{-1} A^T y$
- Nullspace:  $null(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$

# Determinant

- $A \in \mathbb{R}^{n \times n}$ ,  $\mathbf{a}_1, \dots, \mathbf{a}_n$  the rows of  $A$ ,  
 $S = \{ \sum_{i=1}^n \alpha_i \mathbf{a}_i \mid 0 \leq \alpha_i \leq 1 \}$ , then  $\det(A)$  is the volume of  $S$ .
- Properties:
  - $\det(I) = 1$
  - $\det(\lambda A) = \lambda \det(A)$
  - $\det(A^T) = \det(A)$
  - $\det(AB) = \det(A)\det(B)$
  - $\det(A) \neq 0$  if and only if  $A$  is invertible.
  - If  $A$  invertible, then  $\det(A^{-1}) = \det(A)^{-1}$

# Quadratic Forms and Positive Semidefinite Matrices

- $A \in \mathbb{R}^{n \times n}$ ,  $x \in \mathbb{R}^n$ ,  $x^T Ax$  is called a quadratic form:

$$x^T Ax = \sum_{1 \leq i, j \leq n} A_{ij} x_i x_j$$

# Quadratic Forms and Positive Semidefinite Matrices

- $A \in \mathbb{R}^{n \times n}$ ,  $x \in \mathbb{R}^n$ ,  $x^T A x$  is called a quadratic form:

$$x^T A x = \sum_{1 \leq i, j \leq n} A_{ij} x_i x_j$$

- $A$  is positive definite if  $\forall x \in \mathbb{R}^n : x^T A x > 0$
- $A$  is positive semidefinite if  $\forall x \in \mathbb{R}^n : x^T A x \geq 0$
- $A$  is negative definite if  $\forall x \in \mathbb{R}^n : x^T A x < 0$
- $A$  is negative semidefinite if  $\forall x \in \mathbb{R}^n : x^T A x \leq 0$

# Eigenvalues and Eigenvectors

- $A \in \mathbb{R}^{n \times n}$ ,  $\lambda \in \mathbb{C}$  is an eigenvalue of  $A$  with the corresponding eigenvector  $x \in \mathbb{C}^n$  ( $x \neq 0$ ) if:

$$Ax = \lambda x$$

- eigenvalues: the  $n$  possibly complex roots of the polynomial equation  $\det(A - \lambda I) = 0$ , and denoted as  $\lambda_1, \dots, \lambda_n$
- Properties:
  - $\text{tr}(A) = \sum_{i=1}^n \lambda_i$
  - $\det(A) = \prod_{i=1}^n \lambda_i$
  - $\text{rank}(A) = |\{1 \leq i \leq n \mid \lambda_i \neq 0\}|$

# Matrix Eigendecomposition

- $A \in \mathbb{R}^{n \times n}$ ,  $\lambda_1, \dots, \lambda_n$  the eigenvalues, and  $x_1, \dots, x_n$  the eigenvectors.  $X = [x_1 | x_2 | \dots | x_n]$ ,  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ , then  $AX = X\Lambda$ .
- $A$  called diagonalizable if  $X$  invertible:  $A = X\Lambda X^{-1}$
- If  $A$  symmetric, then all eigenvalues real, and  $X$  orthogonal (hence denoted by  $U = [u_1 | u_2 | \dots | u_n]$ ):

$$A = U\Lambda U^T = \sum_{i=1}^n \lambda_i u_i u_i^T$$

- A special case of Singular Value Decomposition

# Outline

1 Basic Linear Algebra

**2 Basic Probability Theory**

# Elements of Probability

- Sample Space  $\Omega$ : Set of all possible outcomes
- Event Space  $\mathcal{F}$ : A family of subsets of  $\Omega$
- Probability Measure: Function  $P : \mathcal{F} \rightarrow \mathbb{R}$  with properties:
  - 1  $P(A) \geq 0$  ( $\forall A \in \mathcal{F}$ )
  - 2  $P(\Omega) = 1$
  - 3  $A_i$ 's disjoint, then  $P(\bigcup_i A_i) = \sum_i P(A_i)$

# Conditional Probability and Independence

- For events  $A, B$ :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- $A, B$  independent if  $P(A|B) = P(A)$  or equivalently:  
 $P(A \cap B) = P(A)P(B)$

# Random Variables and Distributions

- A random variable  $X$  is a function  $X : \Omega \rightarrow \mathbb{R}$   
Example: Number of heads in 20 tosses of a coin
- Probabilities of events associated with random variables defined based on the original probability function. e.g.,  
$$P(X = k) = P(\{\omega \in \Omega \mid X(\omega) = k\})$$
- Cumulative Distribution Function (CDF)  $F_X : \mathbb{R} \rightarrow [0, 1]$ :  
$$F_X(x) = P(X \leq x)$$
- Probability Mass Function (pmf):  $X$  discrete then  
$$p_X(x) = P(X = x)$$
- Probability Density Function (pdf):  $f_X(x) = dF_X(x)/dx$

# Properties of Distribution Functions

## ■ CDF:

- $0 \leq F_X(x) \leq 1$
- $F_X$  monotone increasing, with  $\lim_{x \rightarrow -\infty} F_X(x) = 0$ ,  
 $\lim_{x \rightarrow \infty} F_X(x) = 1$

## ■ pmf:

- $0 \leq p_X(x) \leq 1$
- $\sum_x p_X(x) = 1$
- $\sum_{x \in A} p_X(x) = p_X(A)$

## ■ pdf:

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x) dx = 1$
- $\int_{x \in A} f_X(x) dx = P(X \in A)$

# Expectation and Variance

- Assume random variable  $X$  has pdf  $f_X(x)$ , and  $g : \mathbb{R} \rightarrow \mathbb{R}$ .  
Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

- for discrete  $X$ ,  $E[g(X)] = \sum_x g(x)p_X(x)$

- Properties:

- for any constant  $a \in \mathbb{R}$ ,  $E[a] = a$

- $E[ag(X)] = aE[g(X)]$

- Linearity of Expectation:

$$E[g(X) + h(X)] = E[g(X)] + E[h(X)]$$

- $Var[X] = E[(X - E[X])^2]$

## Some Common Random Variables

- $X \sim \text{Bernoulli}(p)$  ( $0 \leq p \leq 1$ ):

$$p_X(x) = \begin{cases} p & x=1, \\ 1-p & x=0. \end{cases}$$

- $X \sim \text{Geometric}(p)$  ( $0 \leq p \leq 1$ ):  $p_X(x) = p(1-p)^{x-1}$

- $X \sim \text{Uniform}(a, b)$  ( $a < b$ ):

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

- $X \sim \text{Normal}(\mu, \sigma^2)$ :

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

# Multiple Random Variables and Joint Distributions

$X_1, \dots, X_n$  random variables

■ Joint CDF:  $F_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$

■ Joint pdf:  $f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{\partial^n F_{X_1, \dots, X_n}(x_1, \dots, x_n)}{\partial x_1 \dots \partial x_n}$

■ Marginalization:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_2 \dots dx_n$$

■ Conditioning:  $f_{X_1|X_2, \dots, X_n}(x_1|x_2, \dots, x_n) = \frac{f_{X_1, \dots, X_n}(x_1, \dots, x_n)}{f_{X_2, \dots, X_n}(x_2, \dots, x_n)}$

■ Chain Rule:  $f(x_1, \dots, x_n) = f(x_1) \prod_{i=2}^n f(x_i|x_1, \dots, x_{i-1})$

■ Independence:  $f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$ .

■ More generally, events  $A_1, \dots, A_n$  independent if

$$P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i) \quad (\forall S \subseteq \{1, \dots, n\}).$$

# Random Vectors

$X_1, \dots, X_n$  random variables.  $X = [X_1 X_2 \dots X_n]^T$  random vector.

- If  $g : \mathbb{R}^n \rightarrow \mathbb{R}$ , then

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_{X_1, \dots, X_n}(x_1, \dots, x_n) dx_1 \dots dx_n$$

- if  $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ ,  $g = [g_1 \dots g_m]^T$ , then

$$E[g(X)] = [E[g_1(X)] \dots E[g_m(X)]]^T$$

- Covariance Matrix:

$$\Sigma = \text{Cov}(X) = E[(X - E[X])(X - E[X])^T]$$

- Properties of Covariance Matrix:

- $\Sigma_{ij} = \text{Cov}[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]$
- $\Sigma$  symmetric, positive semidefinite

# Multivariate Gaussian Distribution

$\mu \in \mathbb{R}^n$ ,  $\Sigma \in \mathbb{R}^{n \times n}$  symmetric, positive semidefinite  
 $X \sim \mathcal{N}(\mu, \Sigma)$   $n$ -dimensional Gaussian distribution:

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

- $E[X] = \mu$
- $Cov(X) = \Sigma$

# Parameter Estimation: Maximum Likelihood

Parametrized distribution  $f_X(x; \theta)$  with parameter(s)  $\theta$  unknown.

i.i.d. samples  $x_1, \dots, x_n$  observed.

Goal: Estimate  $\theta$

MLE:  $\hat{\theta} = \operatorname{argmax}_{\theta} \{f(x_1, \dots, x_n; \theta)\}$

# MLE Example

$X \sim \text{Gaussian}(\mu, \sigma^2)$ .  $\theta = (\mu, \sigma^2)$  unknown. Samples  $x_1, \dots, x_n$ .  
Then:

$$f(x_1, \dots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right)$$

Setting:  $\frac{\partial \log f}{\partial \mu} = 0$  and  $\frac{\partial \log f}{\partial \sigma} = 0$

Gives:

$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^n x_i}{n}, \quad \hat{\sigma}_{MLE}^2 = \frac{\sum_{i=1}^n (x_i - \hat{\mu})^2}{n}$$

If not possible to find the optimal point in closed form, iterative methods such as gradient decent can be used.

## Some Useful Inequalities

- Markov's Inequality:  $X$  random variable, and  $a > 0$ . Then:

$$P(|X| \geq a) \leq \frac{E[|X|]}{a}$$

- Chebyshev's Inequality: If  $E[X] = \mu$ ,  $\text{Var}(X) = \sigma^2$ ,  $k > 0$ , then:

$$\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- Chernoff bound:  $X_1, \dots, X_n$  iid random variables, with  $E[X_i] = \mu$ ,  $X_i \in \{0, 1\}$  ( $\forall 1 \leq i \leq n$ ). Then:

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| \geq \epsilon\right) \leq 2 \exp(-2n\epsilon^2)$$

- Multiple variants of Chernoff-type bounds exist, which can be useful in different settings

# References

- 1 CS229 notes on basic linear algebra and probability theory
- 2 Wikipedia!