COMP 590/776: Computer Vision

Lecture 2: Maths Review

Instructor: Soumyadip (Roni) Sengupta

ULA: Andrea Dunn, William Li, Liujie Zheng

Course Website:
Scan Me!

Overview

- Linear Algebra Review
    - Very important for 3D vision
    - Also important for Deep Neural networks
    - My fav book: “Introduction to Linear Algebra” by Gilbert Strang
- Multivariate Calculus Review
    - Important for image processing
    - Also for designing loss functions in deep neural networks
- Probability Review
    - Forms the core of Machine Learning

Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review

Vector and Matrix Products

Inner Product

$$x^T y \in \mathbb{R} = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \end{array}\right] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

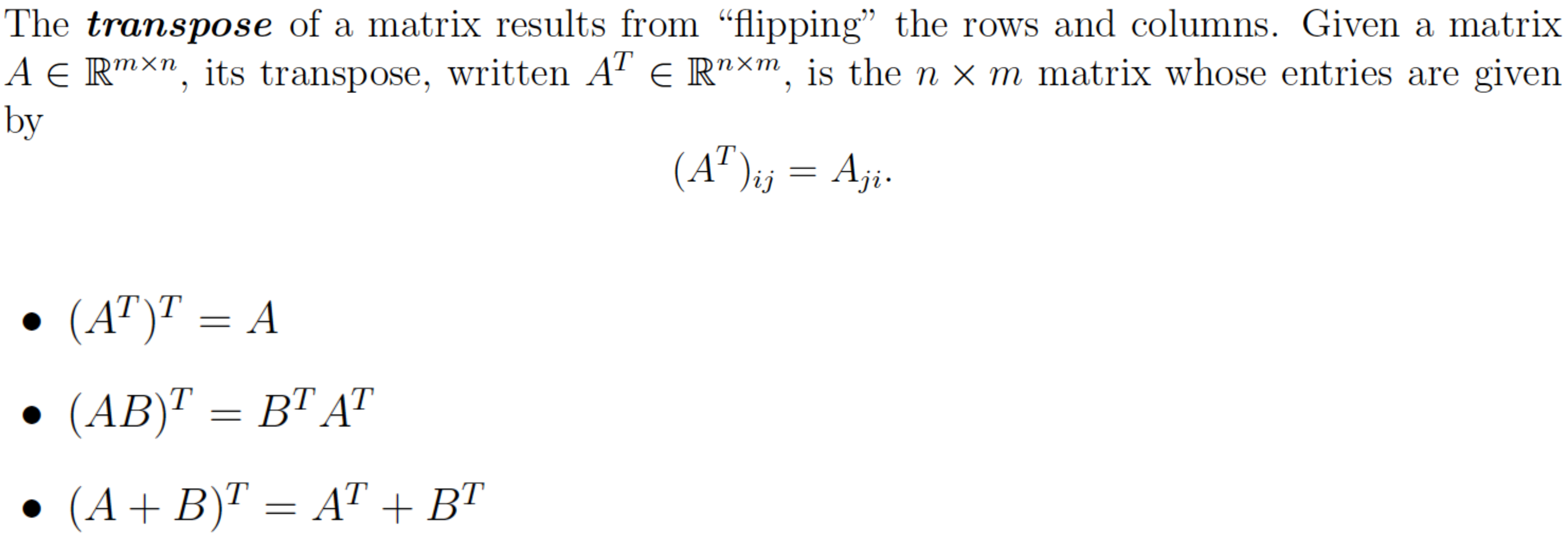
Outer Product

$$xy^T \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \left[\begin{array}{cccc} y_1 & y_2 & \cdots & y_n \end{array}\right] = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}.$$

Vector and Matrix Products

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix}.$$

Transpose

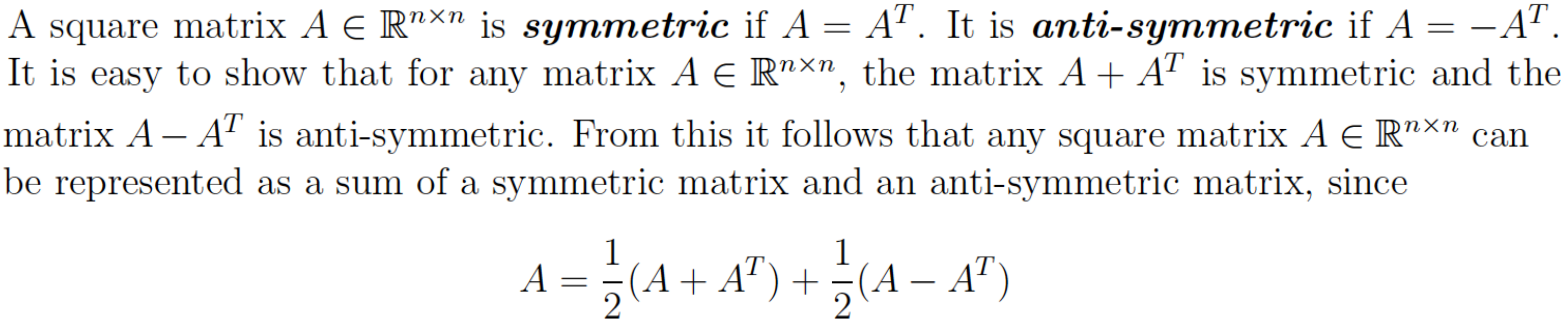


The ***transpose*** of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

Symmetric Matrix



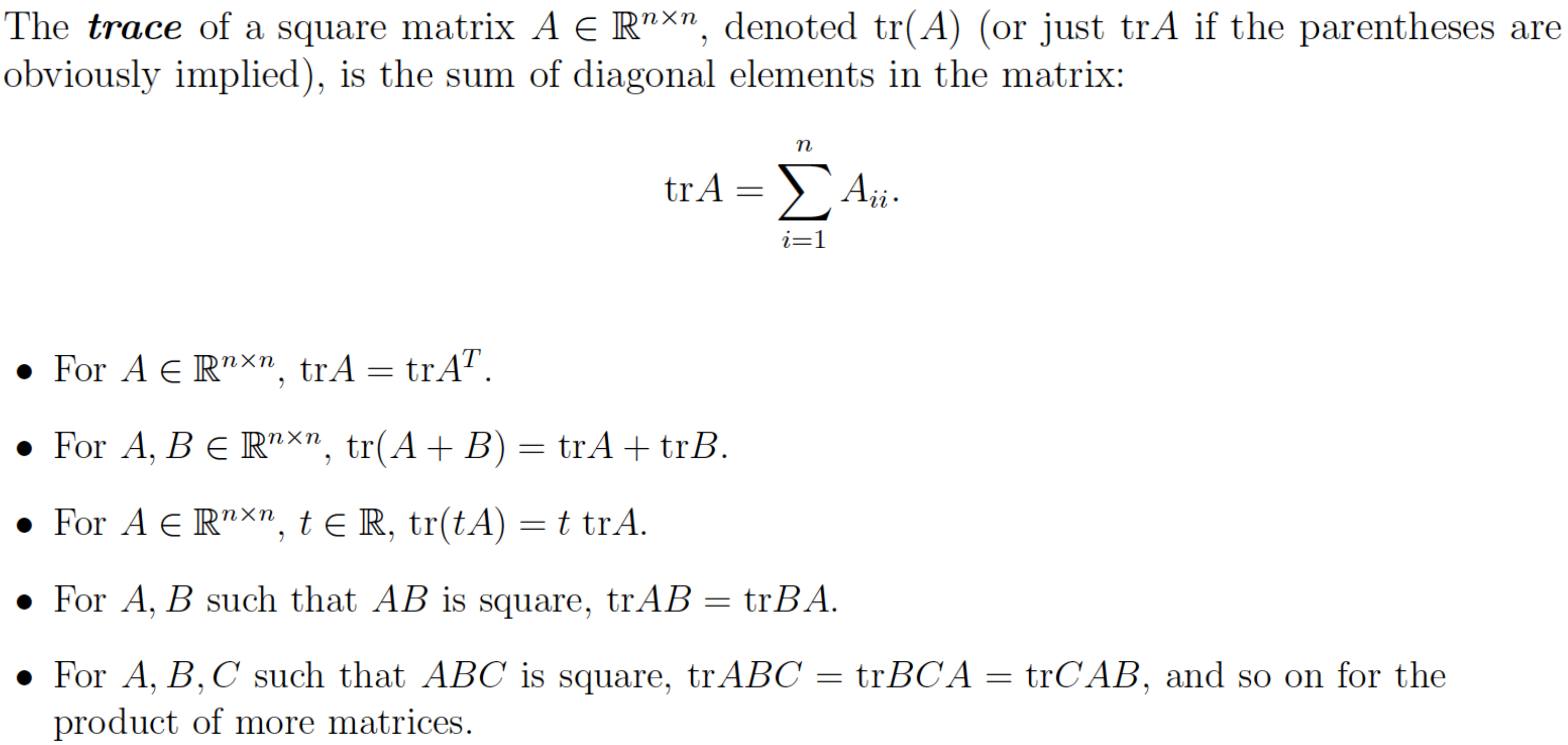
A square matrix $A \in \mathbb{R}^{n\times n}$ is ***symmetric*** if $A = A^T$. It is ***anti-symmetric*** if $A = -A^T$. It is easy to show that for any matrix $A \in \mathbb{R}^{n\times n}$, the matrix $A + A^T$ is symmetric and the matrix $A - A^T$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n\times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

$$\begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 5 \end{bmatrix} ?$$

$$\begin{bmatrix} 0 & 1 & -2 \\ -1 & 0 & 3 \\ 2 & -3 & 0 \end{bmatrix} ?$$

Trace

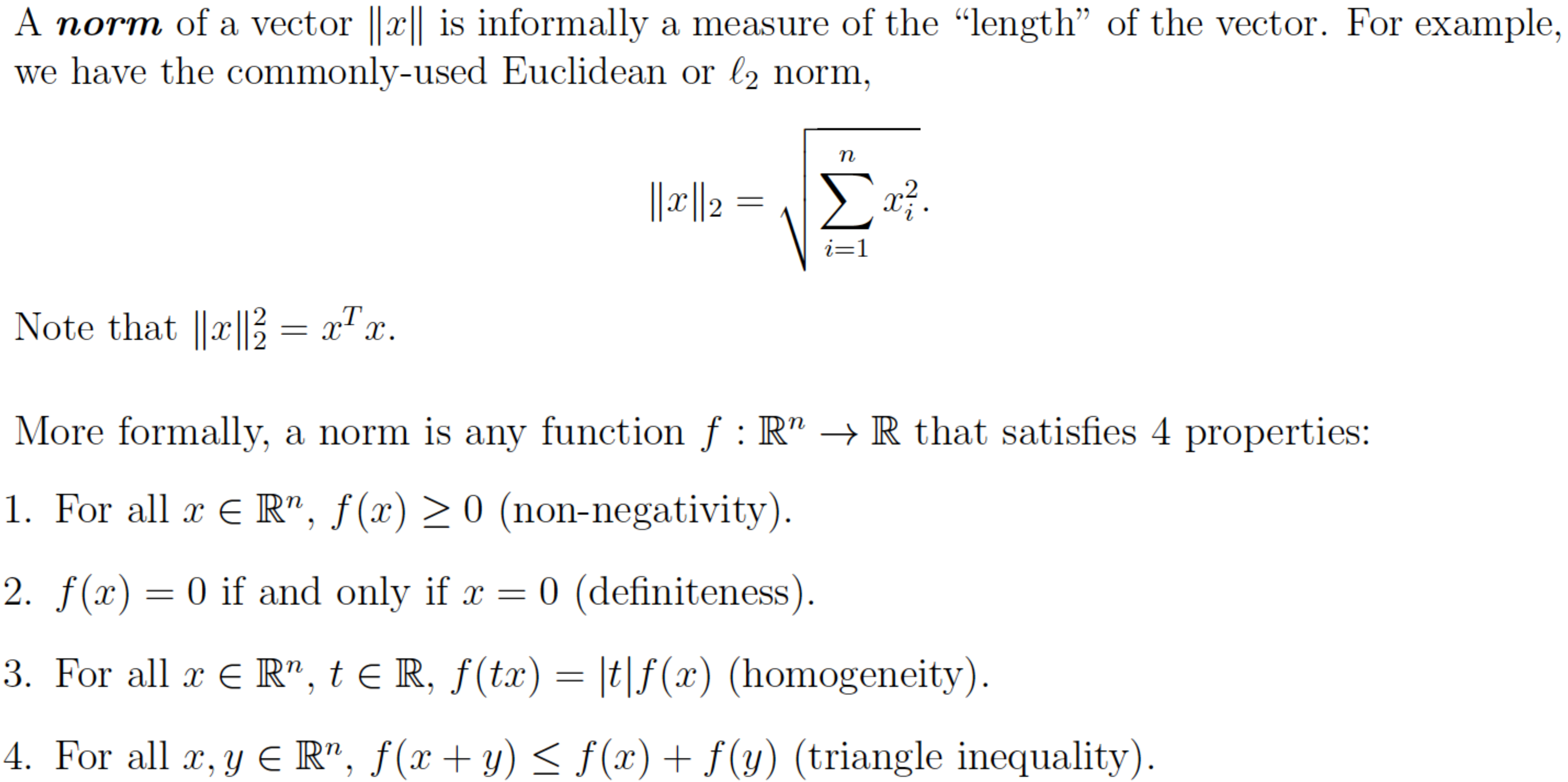


The ***trace*** of a square matrix $A \in \mathbb{R}^{n\times n}$, denoted $\mathrm{tr}(A)$ (or just $\mathrm{tr}A$ if the parentheses are obviously implied), is the sum of diagonal elements in the matrix:

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}.$$

- For $A \in \mathbb{R}^{n\times n}$, $\mathrm{tr}A = \mathrm{tr}A^T$.
- For $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B$.
- For $A \in \mathbb{R}^{n\times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\,\mathrm{tr}A$.
- For A, B such that AB is square, $\mathrm{tr}AB = \mathrm{tr}BA$.
- For A, B, C such that ABC is square, $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on for the product of more matrices.

Norm



A ***norm*** of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Note that $\|x\|_2^2 = x^T x$.

More formally, a norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies 4 properties:

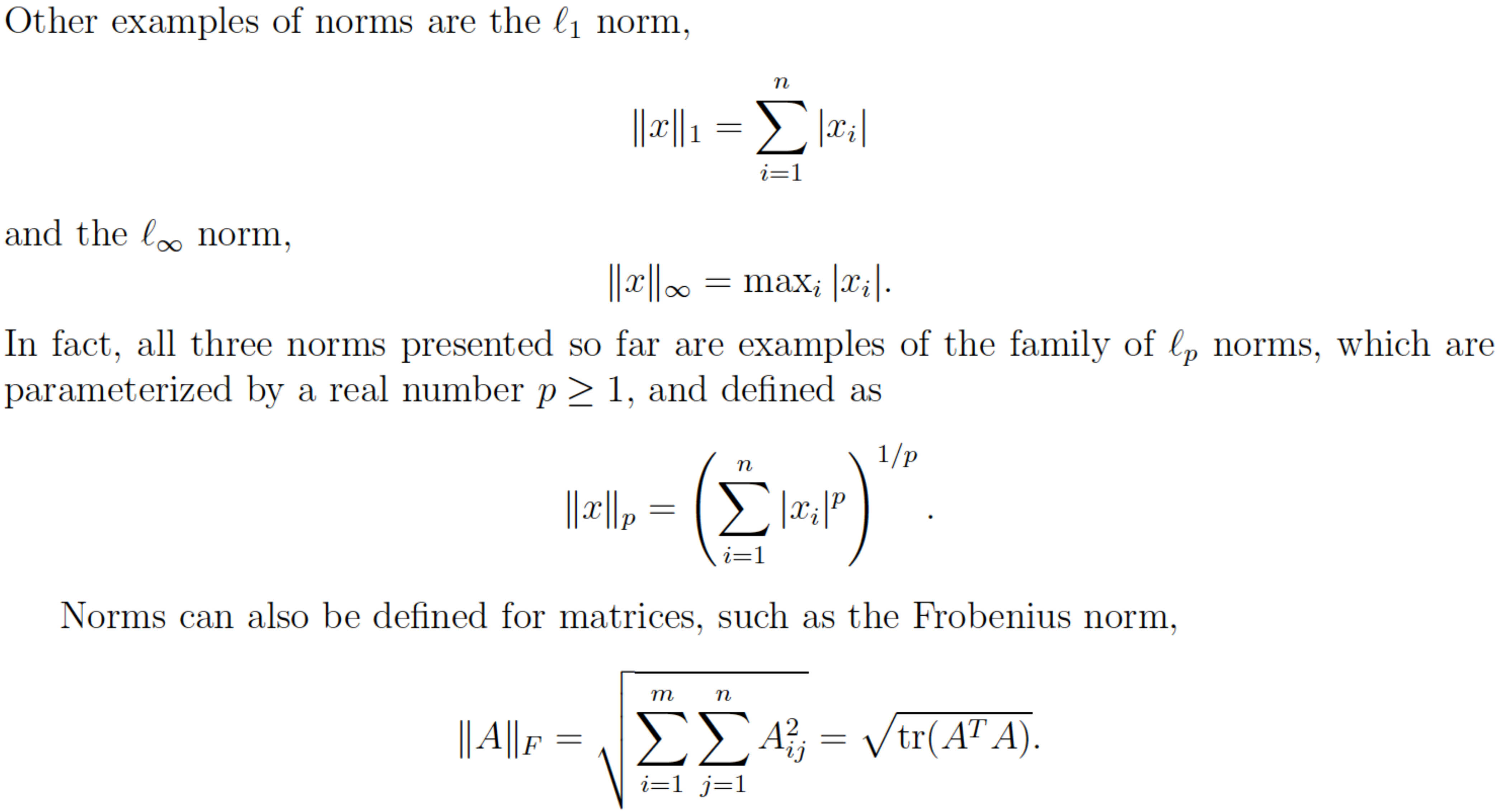
1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).

2. $f(x) = 0$ if and only if $x = 0$ (definiteness).

3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t| f(x)$ (homogeneity).

4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Norm



Other examples of norms are the ℓ_1 norm,

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

and the ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Norms can also be defined for matrices, such as the Frobenius norm,

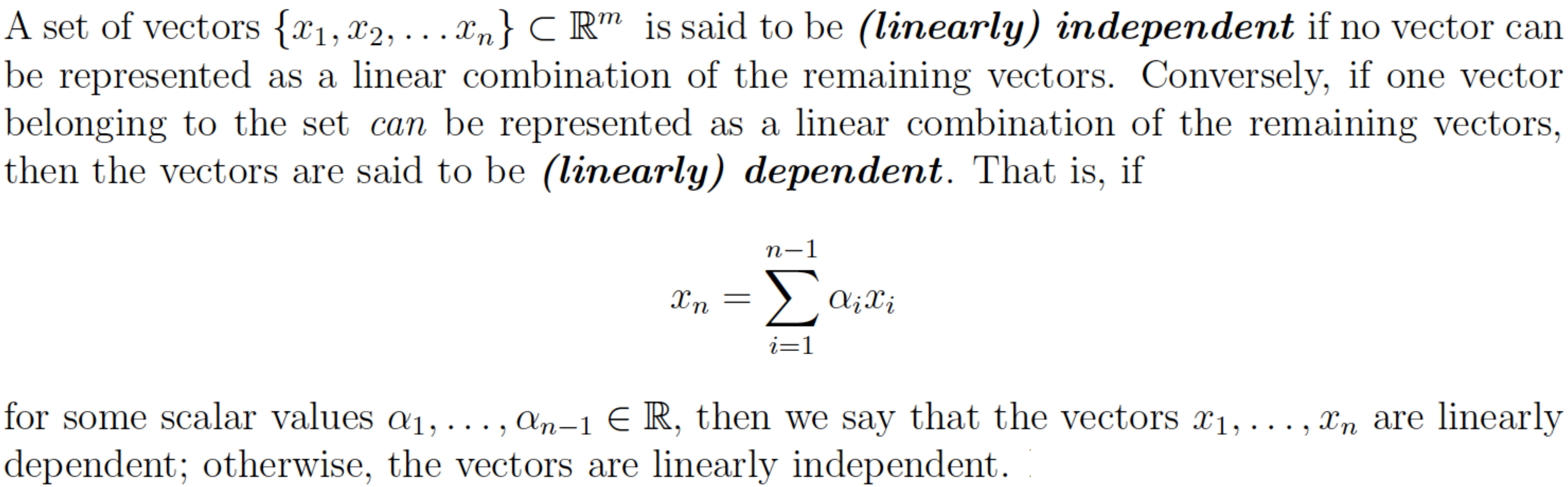
$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{\operatorname{tr}(A^T A)}.$$

Orthogonality

Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

$$U^T U = I = U U^T.$$

Linear Independence



A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be ***(linearly) independent*** if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set *can* be represented as a linear combination of the remaining vectors, then the vectors are said to be ***(linearly) dependent***. That is, if

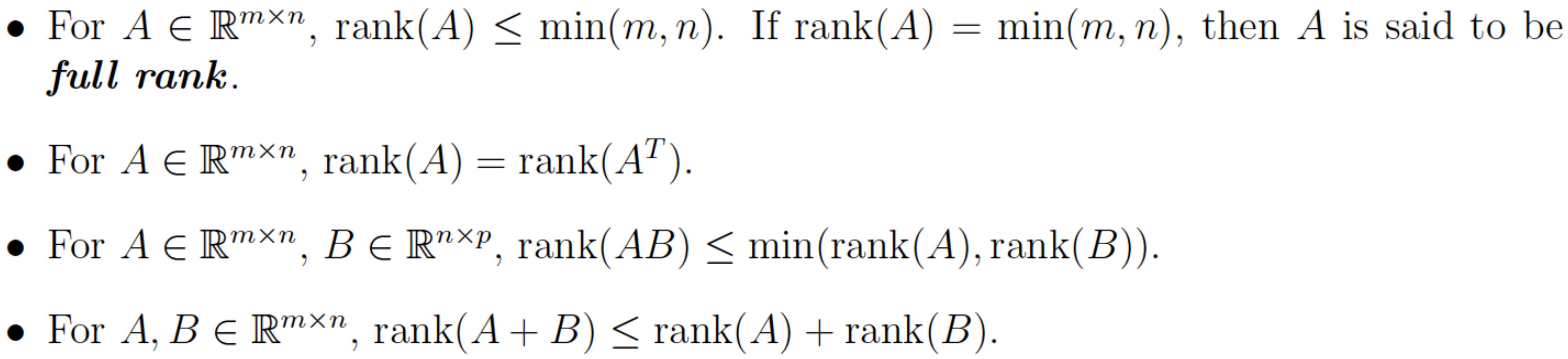
$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $x_1, \ldots, x_n$ are linearly dependent; otherwise, the vectors are linearly independent.

Rank of a matrix

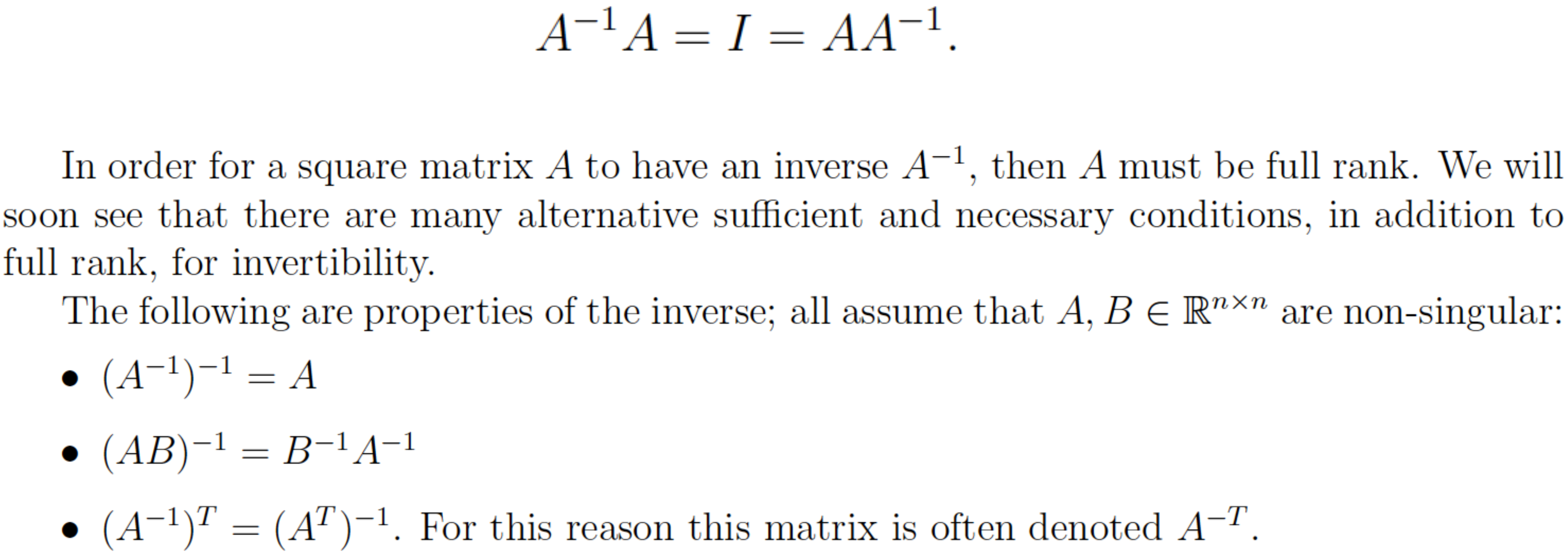
Rank of a matrix A is the maximal number of linearly independent columns or rows.

A matrix is full ranked, if all of its columns/rows are linearly independent.



- For $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) \leq \min(m,n)$. If $\mathrm{rank}(A) = \min(m,n)$, then A is said to be ***full rank***.
- For $A \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A) = \mathrm{rank}(A^T)$.
- For $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, $\mathrm{rank}(AB) \leq \min(\mathrm{rank}(A), \mathrm{rank}(B))$.
- For $A, B \in \mathbb{R}^{m\times n}$, $\mathrm{rank}(A+B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$.

Inverse

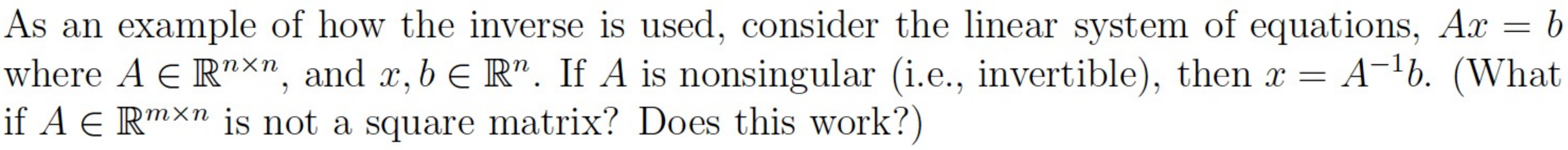


$$A^{-1}A = I = AA^{-1}.$$

In order for a square matrix A to have an inverse A^{-1}, then A must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T}.



As an example of how the inverse is used, consider the linear system of equations, $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, and $x, b \in \mathbb{R}^n$. If A is nonsingular (i.e., invertible), then $x = A^{-1}b$. (What if $A \in \mathbb{R}^{m \times n}$ is not a square matrix? Does this work?)

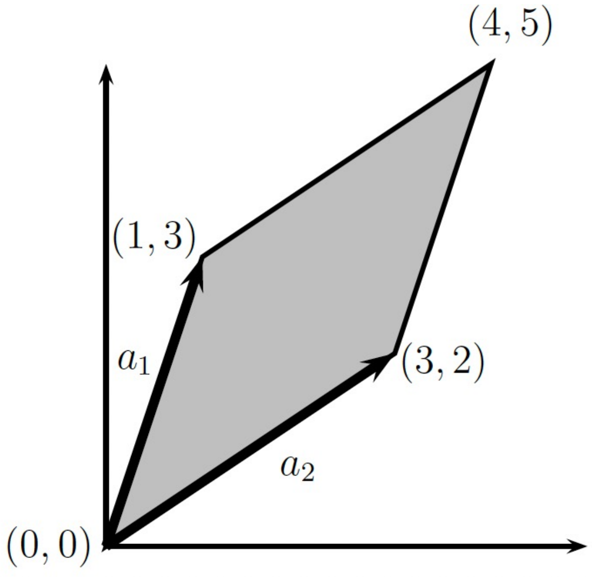
Determinant

How do we find determinant of a nxn matrix?
See Leibniz formula for determinants

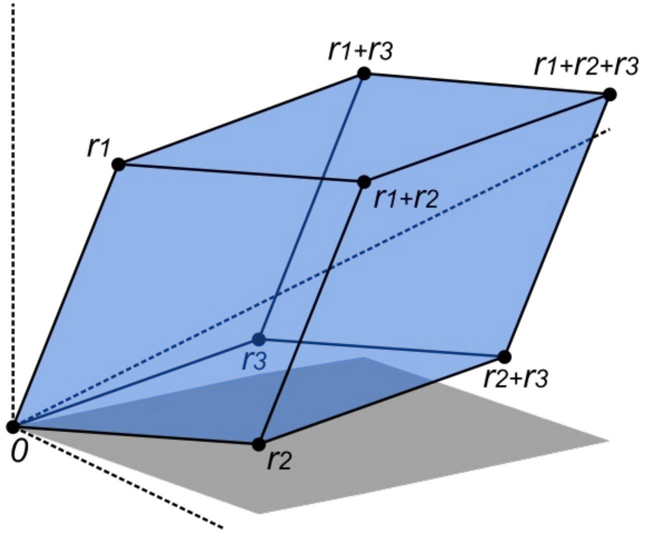
$$\left|\begin{bmatrix} a_{11} \end{bmatrix}\right| = a_{11}$$

$$\left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| = a_{11}a_{22} - a_{12}a_{21}$$

$$\left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| = \begin{array}{l} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ \quad -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{array}$$



- Determinant of 2x2 matrix is the area of the parallelogram formed by the column vectors of the matrix.
- Determinant of 3x3 matrix is the volume of a parallelopiped formed by the 3 column vectors of the matrix
- Sign indicates whether the transformation preserves or reverse orientation.



Eigenvalue and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector***[3] if

$$Ax = \lambda x, \quad x \neq 0.$$

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

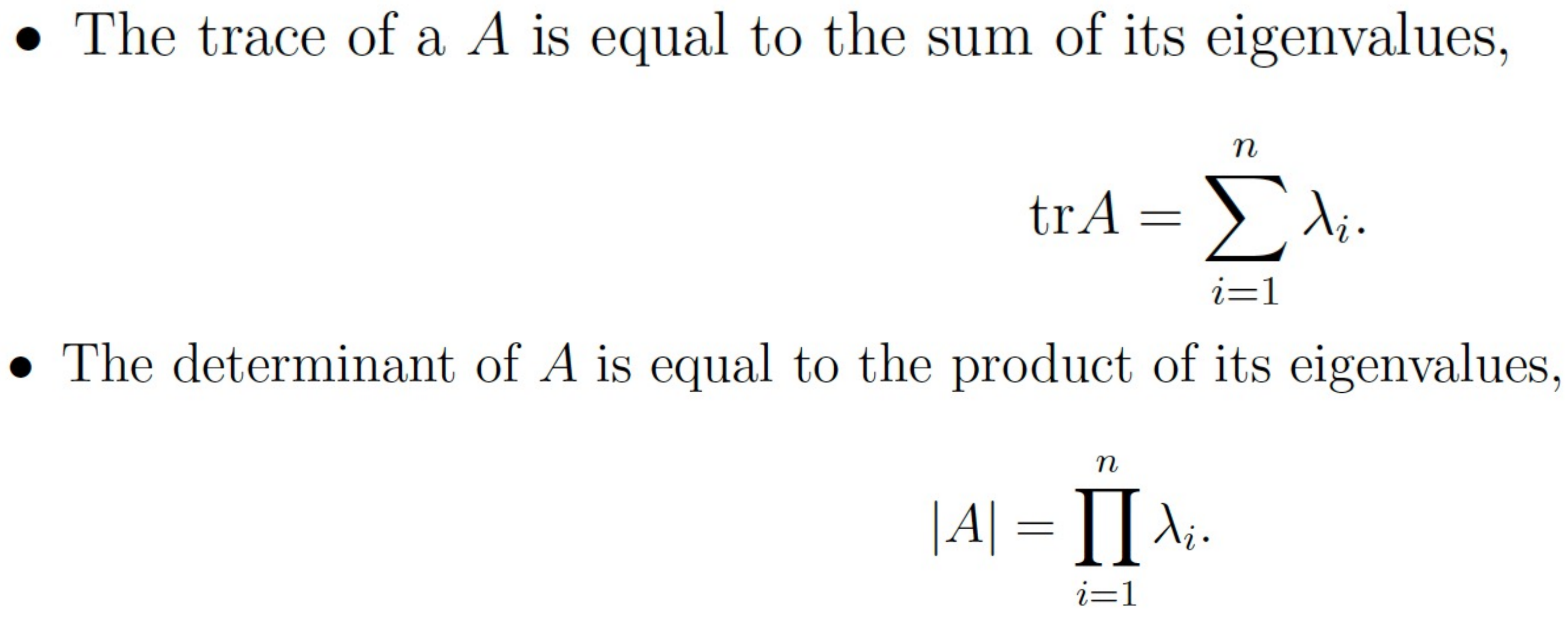
$$|(\lambda I - A)| = 0. \Longrightarrow (\lambda_1 - \lambda)(\lambda_2 - \lambda)\cdots(\lambda_n - \lambda),$$

Characteristic polynomial

$$AX = X\Lambda \Longrightarrow X \in \mathbb{R}^{n\times n} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

If the eigenvectors of A are linearly independent, then the matrix X will be invertible, so $A = X\Lambda X^{-1}$. A matrix that can be written in this form is called ***diagonalizable***.

Eigenvalue and Eigenvectors



- The trace of a A is equal to the sum of its eigenvalues,

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

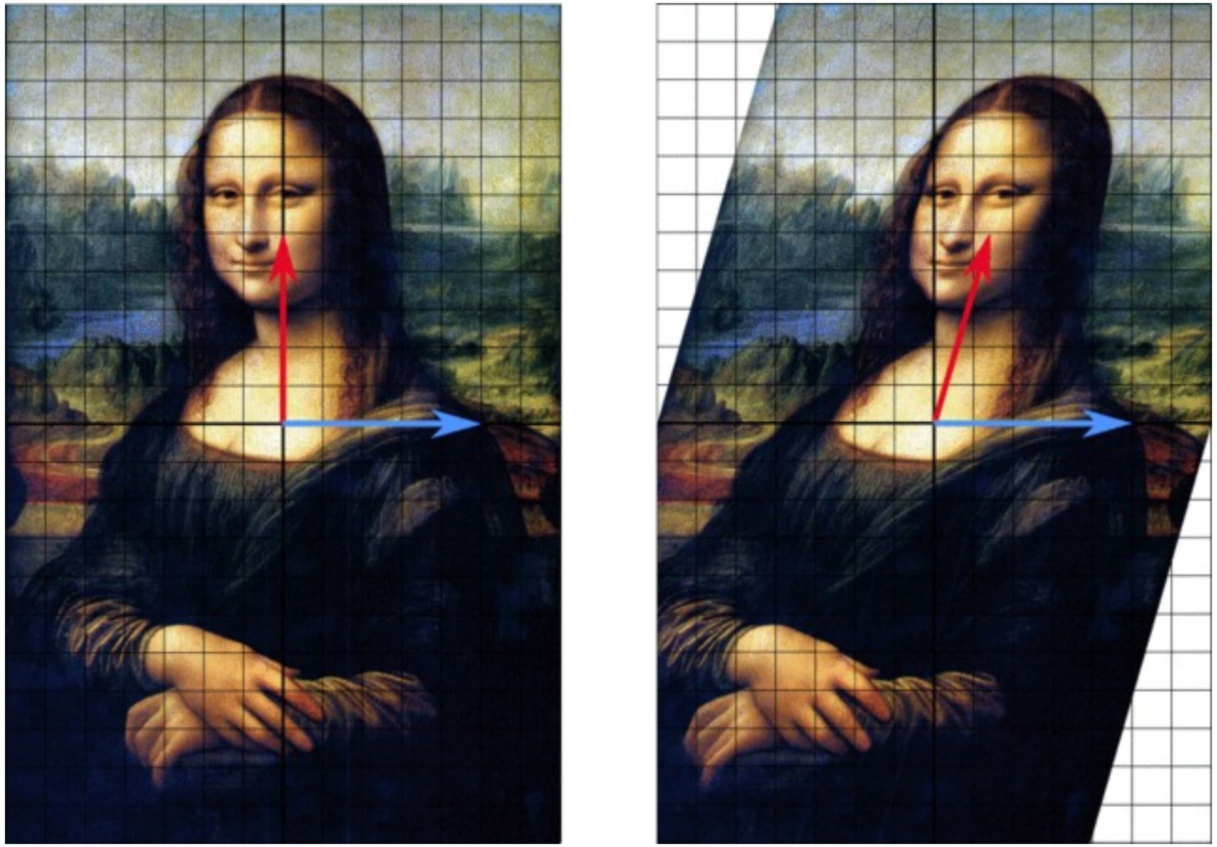
- The determinant of A is equal to the product of its eigenvalues,

$$|A| = \prod_{i=1}^{n} \lambda_i.$$

Eigenvalue and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector*** if

$$Ax = \lambda x, \quad x \neq 0. \qquad \begin{pmatrix} I & M \\ 0 & I \end{pmatrix}.$$



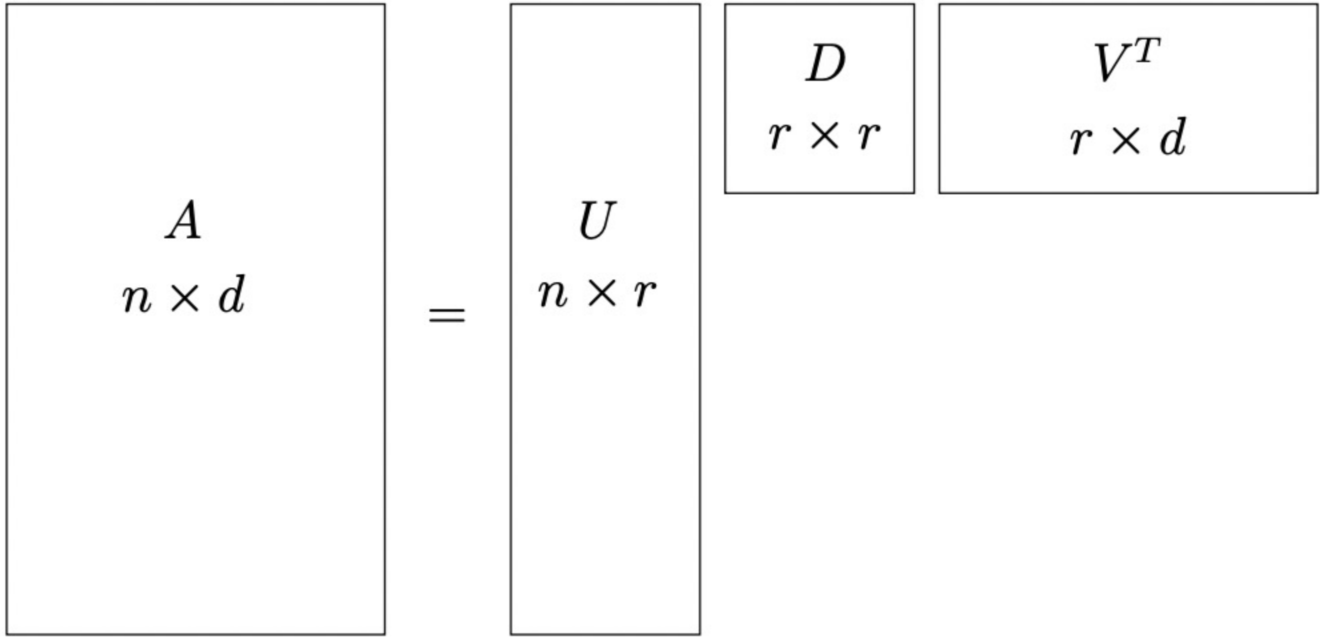
In this shear mapping the red arrow changes direction, but the blue arrow does not. The blue arrow is an eigenvector of this shear mapping because it does not change direction, and since its length is unchanged, its eigenvalue is 1.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x + my \\ y \end{pmatrix} = \begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

[a,0] is an eigenvector for any value of a.

Singular Value Decomposition

$$A = UDV^T \qquad A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$



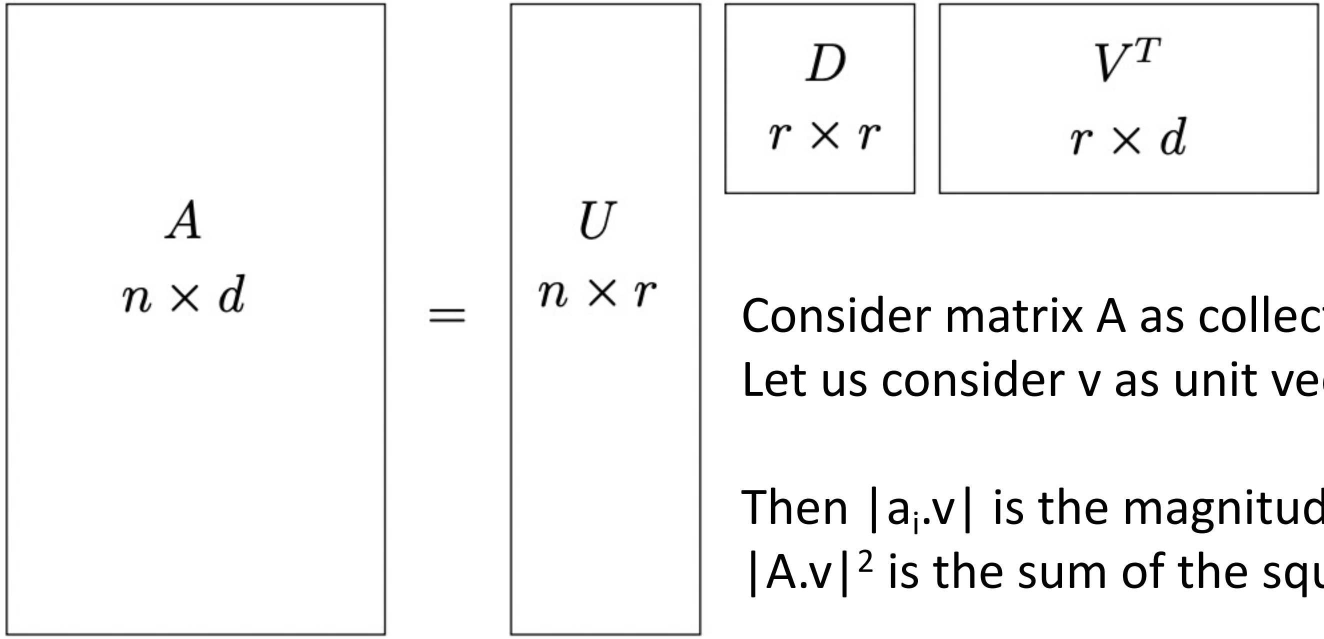
U and V are orthonormal matrices, i.e. $U^TU=I$ and $V^TV=I$

D is a diagonal matrix, where each diagonal element is known as singular values. $D_{ii}=\sigma_i$

r is the rank of the matrix
r <= min (n,d)

Singular Value Decomposition

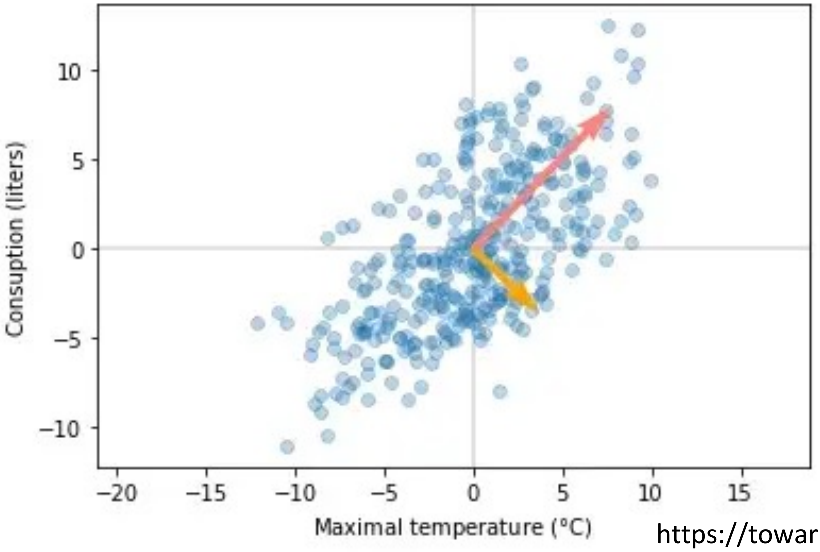
$$A = UDV^T \quad A = \sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T.$$



$$\underset{n \times d}{A} = \underset{n \times r}{U} \; \underset{r \times r}{D} \; \underset{r \times d}{V^T}$$

Consider matrix A as collection of ‘n’ d-dimensional vectors a_i.
Let us consider v as unit vector in d-dimensional space.

Then $|a_i.v|$ is the magnitude of project of each data point a_i onto v.
$|A.v|^2$ is the sum of the squared distances of all the data points to the line v.



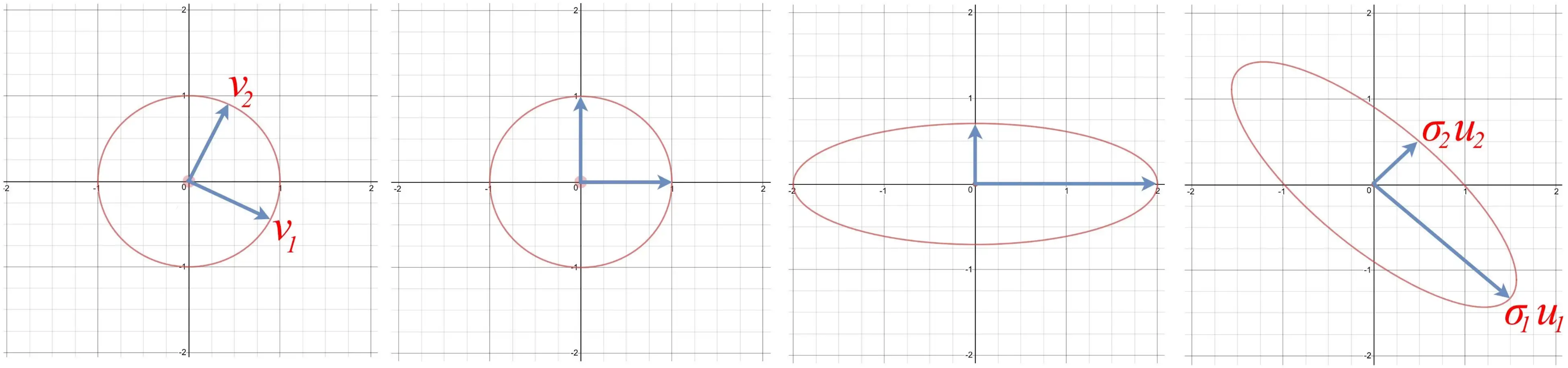
$$\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|.$$

Finding v_1 indicates the direction in which the data is most spread. -> Most informative direction of the data

https://towardsdatascience.com/simple-svd-algorithms-13291ad2eef2

Singular Value Decomposition

$$Ax = USV^T x$$ Change of basis



$\xrightarrow[\text{(rotation)}]{V^T}$ (orthogonal) $\xrightarrow[\text{(scale)}]{S}$ (diagonal) $\xrightarrow[\text{(rotation)}]{U}$ (orthogonal)

Applying A to any vector x can be visualized as…

https://jonathan-hui.medium.com/machine-learning-singular-value-decomposition-svd-principal-component-analysis-pca-1d45e885e491

Eigen-decomposition vs SVD

$A = P.D.P^{-1}$

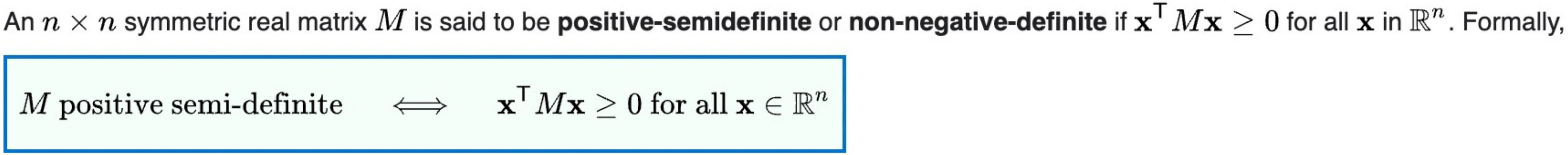
$A = U.D.V^{T}$

- The vectors in the eigen-decomposition matrix P are not necessarily orthogonal, so the change of basis isn't a simple rotation. On the other hand, the vectors in the matrices U and V in the SVD are orthonormal, so they do represent rotations (and possibly flips).

- In the SVD, the nondiagonal matrices U and V are not necessarily the inverse of one another. They are usually not related to each other at all. In the eigen decomposition the nondiagonal matrices P and P^{-1} are inverses of each other.

- The SVD always exists for any sort of rectangular or square matrix, whereas the eigen decomposition can only exists for square matrices, and even among square matrices sometimes it doesn't exist (eigen vectors need to be linearly independent).

They are same when A is positive semi-definite matrix, i.e.



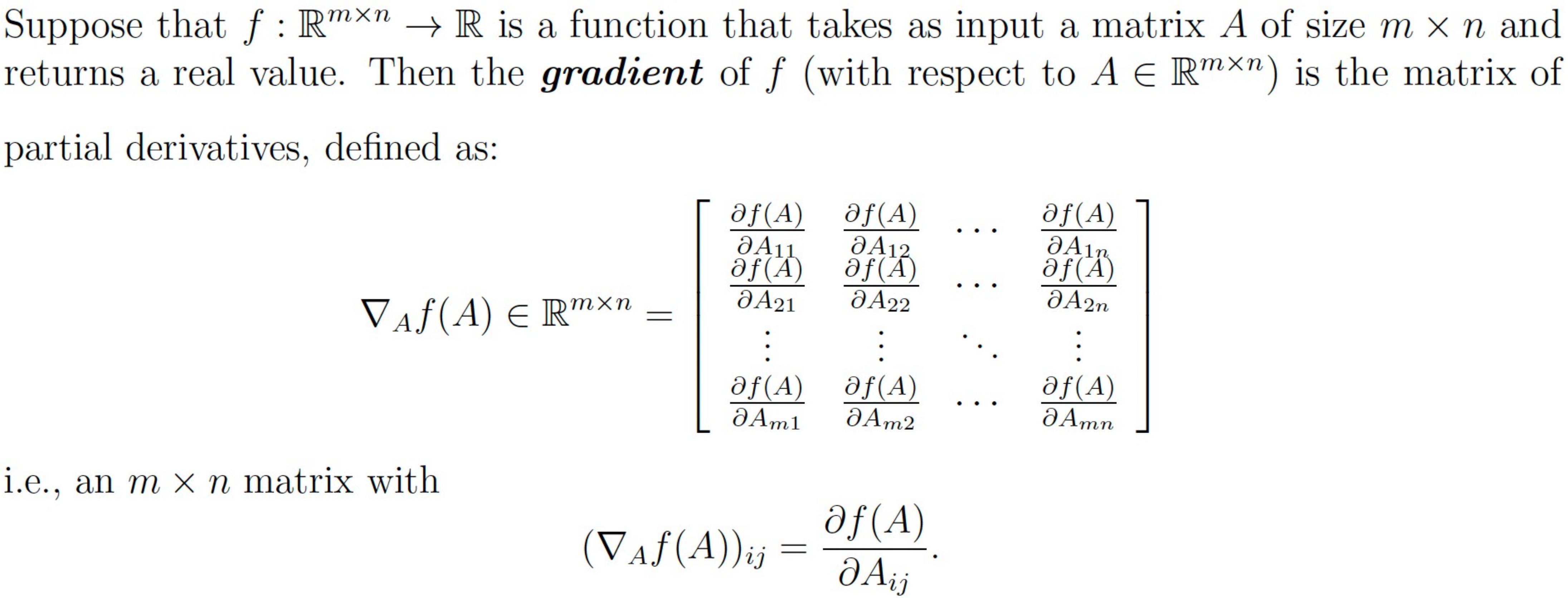
An $n \times n$ symmetric real matrix M is said to be **positive-semidefinite** or **non-negative-definite** if $\mathbf{x}^\top M\mathbf{x} \geq 0$ for all $\mathbf{x}$ in $\mathbb{R}^n$. Formally,

$$M \text{ positive semi-definite} \quad \Longleftrightarrow \quad \mathbf{x}^\top M\mathbf{x} \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n$$

Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review

Gradient



Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m\times n}$) is the matrix of partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

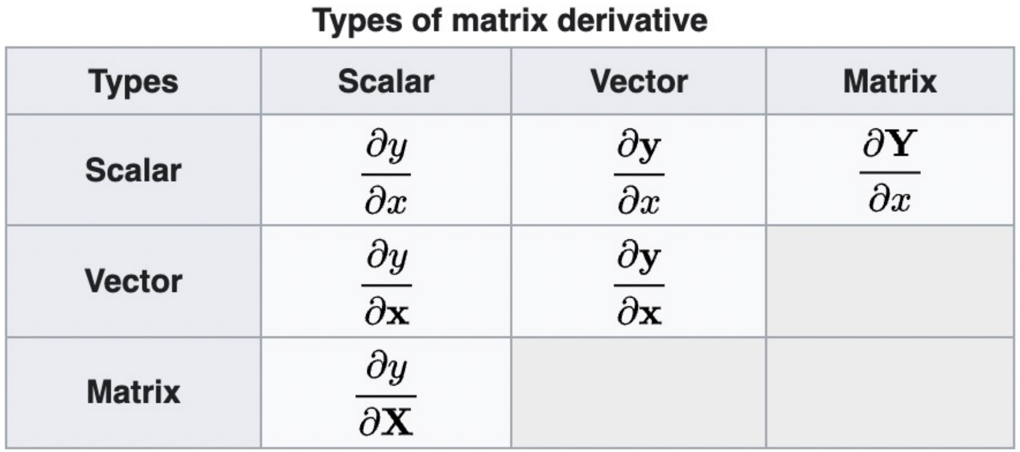
Gradient

$$\frac{\partial y}{\partial \mathbf{x}}$$

y is scalar

x is a nx1 dim vector

→ What is the dimension?



Types of matrix derivative

Types	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$		

$$\frac{\partial \operatorname{tr}(\mathbf{X})}{\partial \mathbf{X}}$$

→ What is the result?

$$\frac{\partial \operatorname{tr}\left(\mathbf{X}^\top \mathbf{A} \mathbf{X}\right)}{\partial \mathbf{X}} =$$

→ What is the result?

Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian*** matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}. \qquad \nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

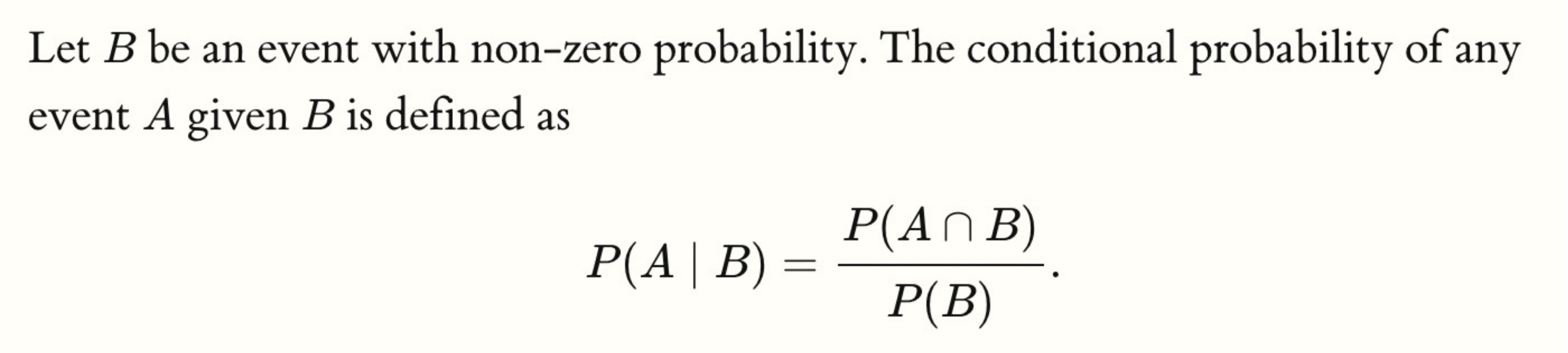
Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

Overview

- Linear Algebra Review
- Multivariate Calculus Review
- Probability Review

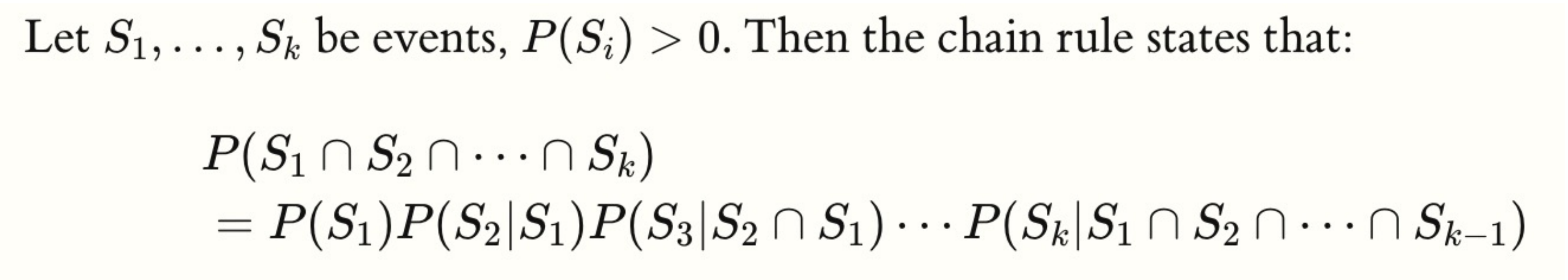
Conditional Probability



Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as

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

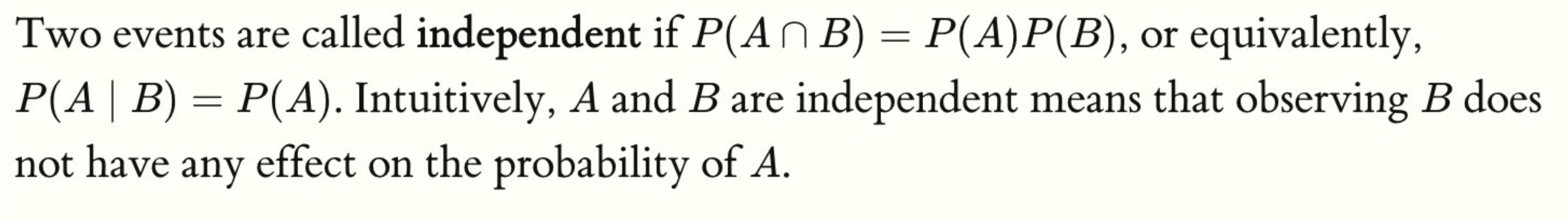
Chain Rule



Let $S_1, \ldots, S_k$ be events, $P(S_i) > 0$. Then the chain rule states that:

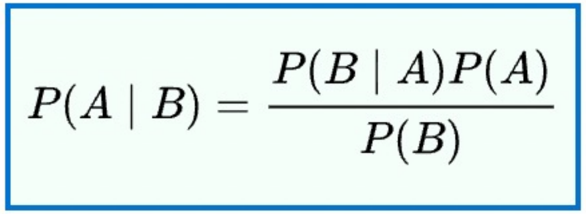
$$\begin{aligned} &P(S_1 \cap S_2 \cap \cdots \cap S_k) \\ &= P(S_1)P(S_2|S_1)P(S_3|S_2 \cap S_1) \cdots P(S_k|S_1 \cap S_2 \cap \cdots \cap S_{k-1}) \end{aligned}$$

Independence



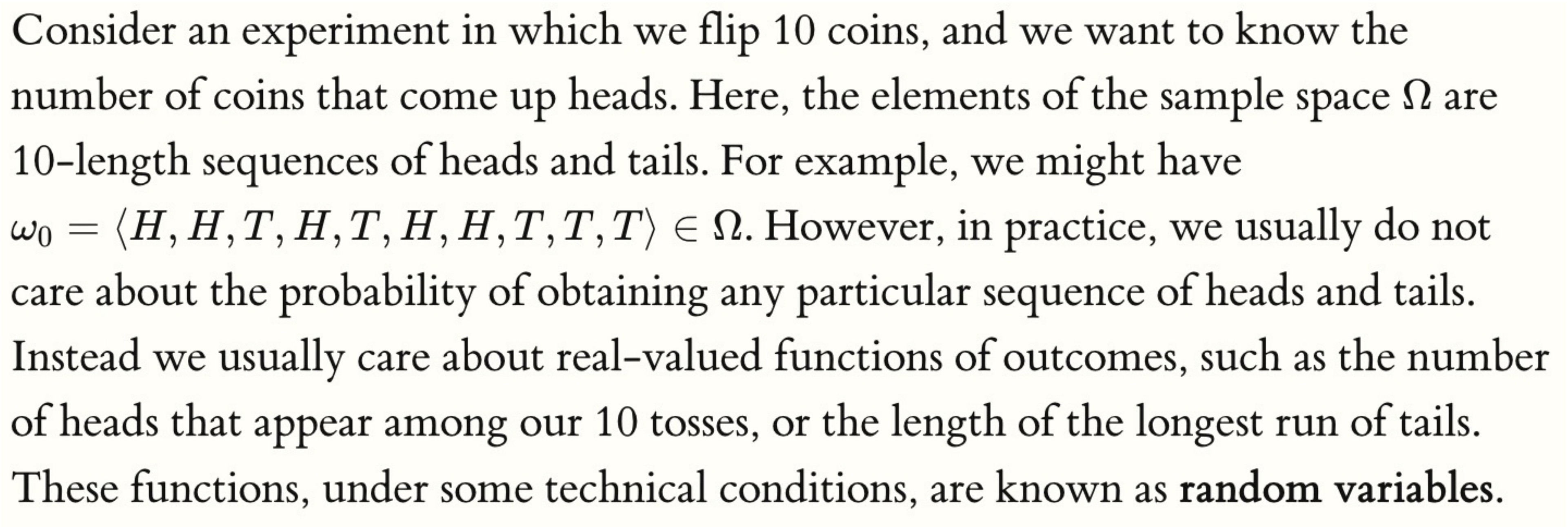
Two events are called **independent** if $P(A \cap B) = P(A)P(B)$, or equivalently, $P(A \mid B) = P(A)$. Intuitively, A and B are independent means that observing B does not have any effect on the probability of A.

Bayes Rule



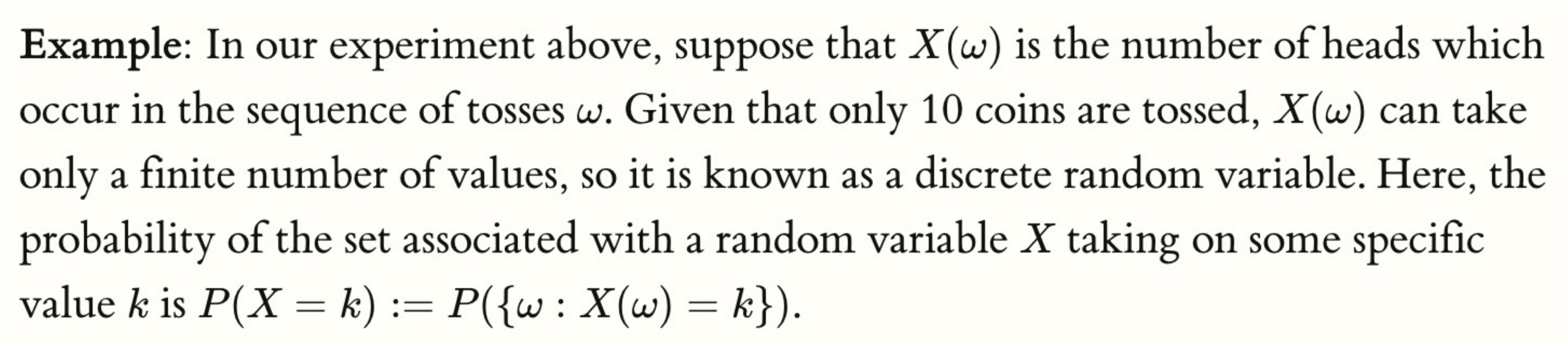
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Random Variables



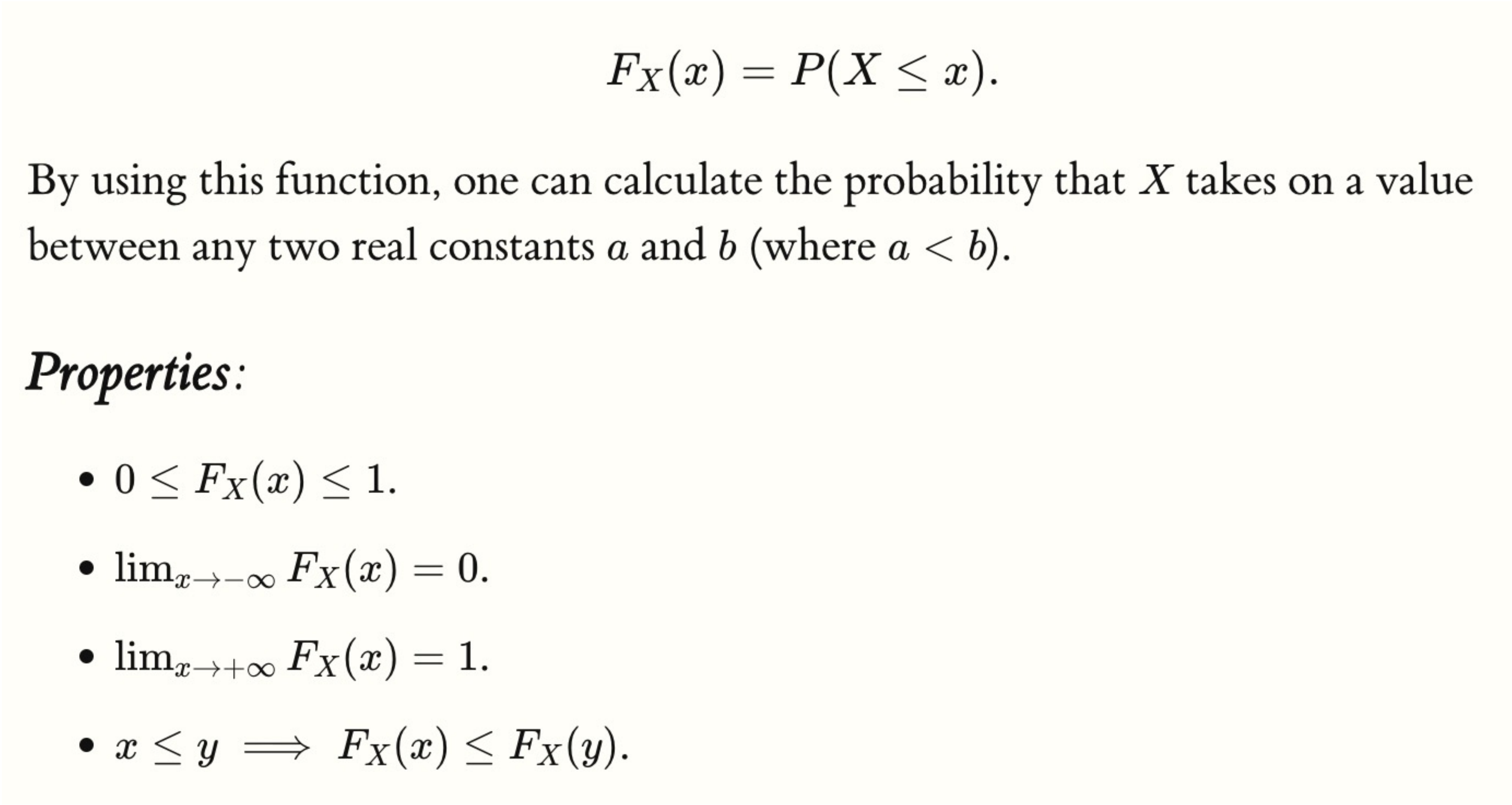
Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

Random Variables



Example: In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a discrete random variable. Here, the probability of the set associated with a random variable X taking on some specific value k is $P(X = k) := P(\{\omega : X(\omega) = k\})$.

Probability Distribution Function (PDF) & Cumulative Distribution function (CDF)

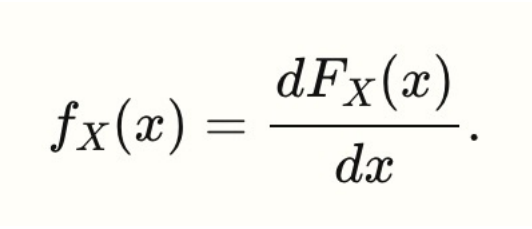


$$F_X(x) = P(X \leq x).$$

By using this function, one can calculate the probability that X takes on a value between any two real constants a and b (where $a < b$).

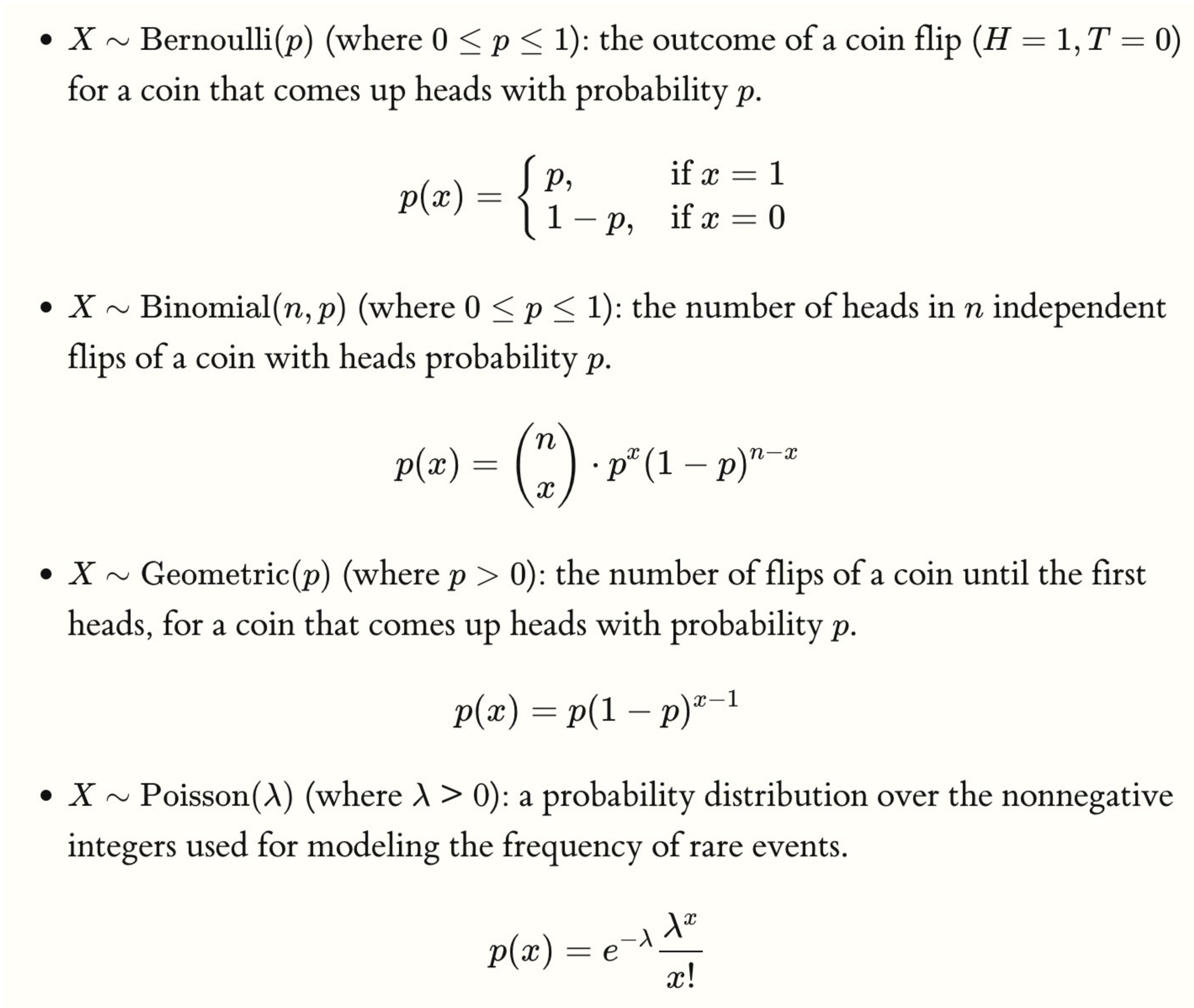
Properties:

- $0 \leq F_X(x) \leq 1.$
- $\lim_{x \to -\infty} F_X(x) = 0.$
- $\lim_{x \to +\infty} F_X(x) = 1.$
- $x \leq y \implies F_X(x) \leq F_X(y).$



$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Discrete random variables



- $X \sim \text{Bernoulli}(p)$ (where $0 \leq p \leq 1$): the outcome of a coin flip ($H = 1, T = 0$) for a coin that comes up heads with probability p.

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

- $X \sim \text{Binomial}(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} \cdot p^x (1 - p)^{n-x}$$

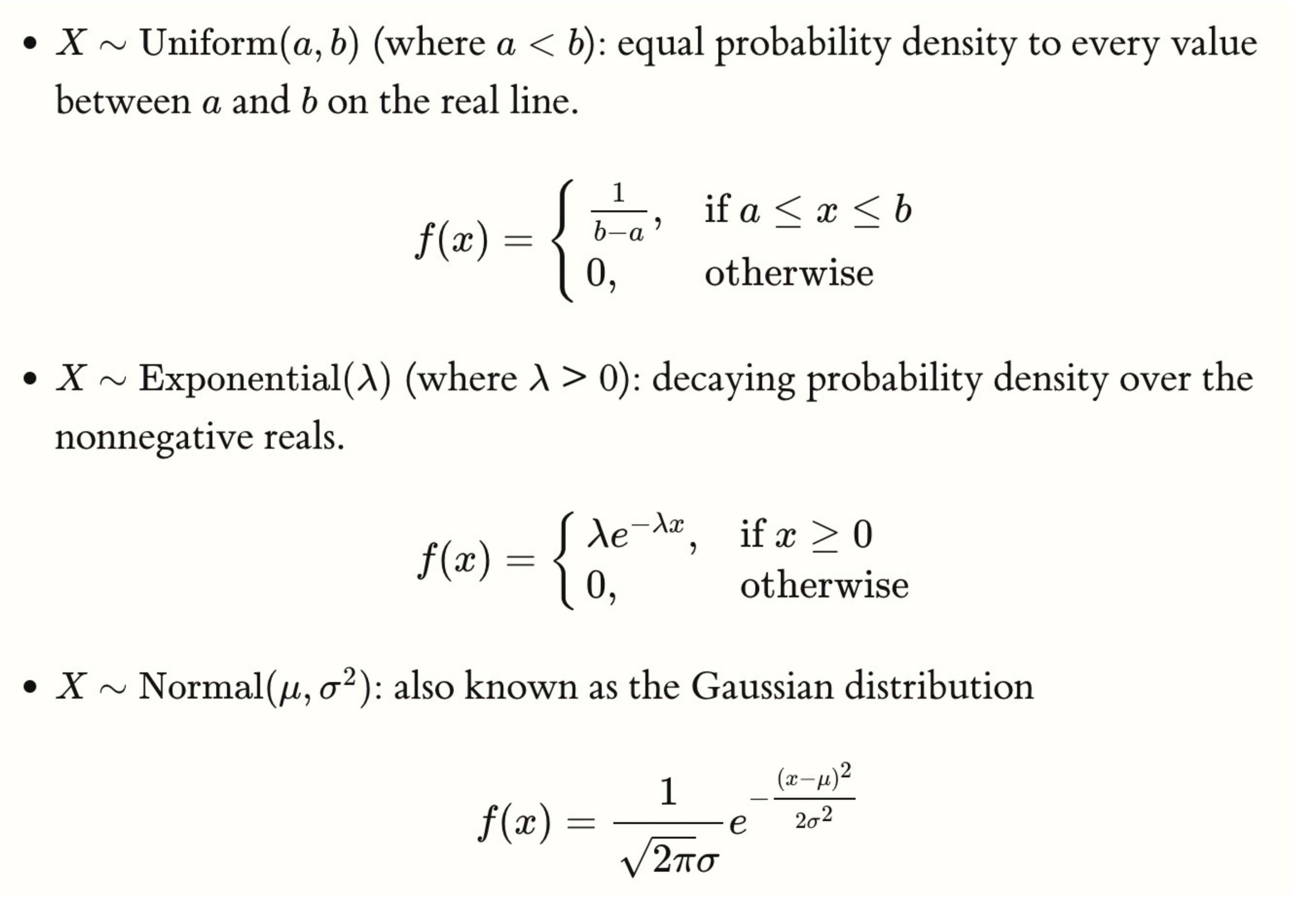
- $X \sim \text{Geometric}(p)$ (where $p > 0$): the number of flips of a coin until the first heads, for a coin that comes up heads with probability p.

$$p(x) = p(1 - p)^{x-1}$$

- $X \sim \text{Poisson}(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous random variables



- $X \sim \text{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

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

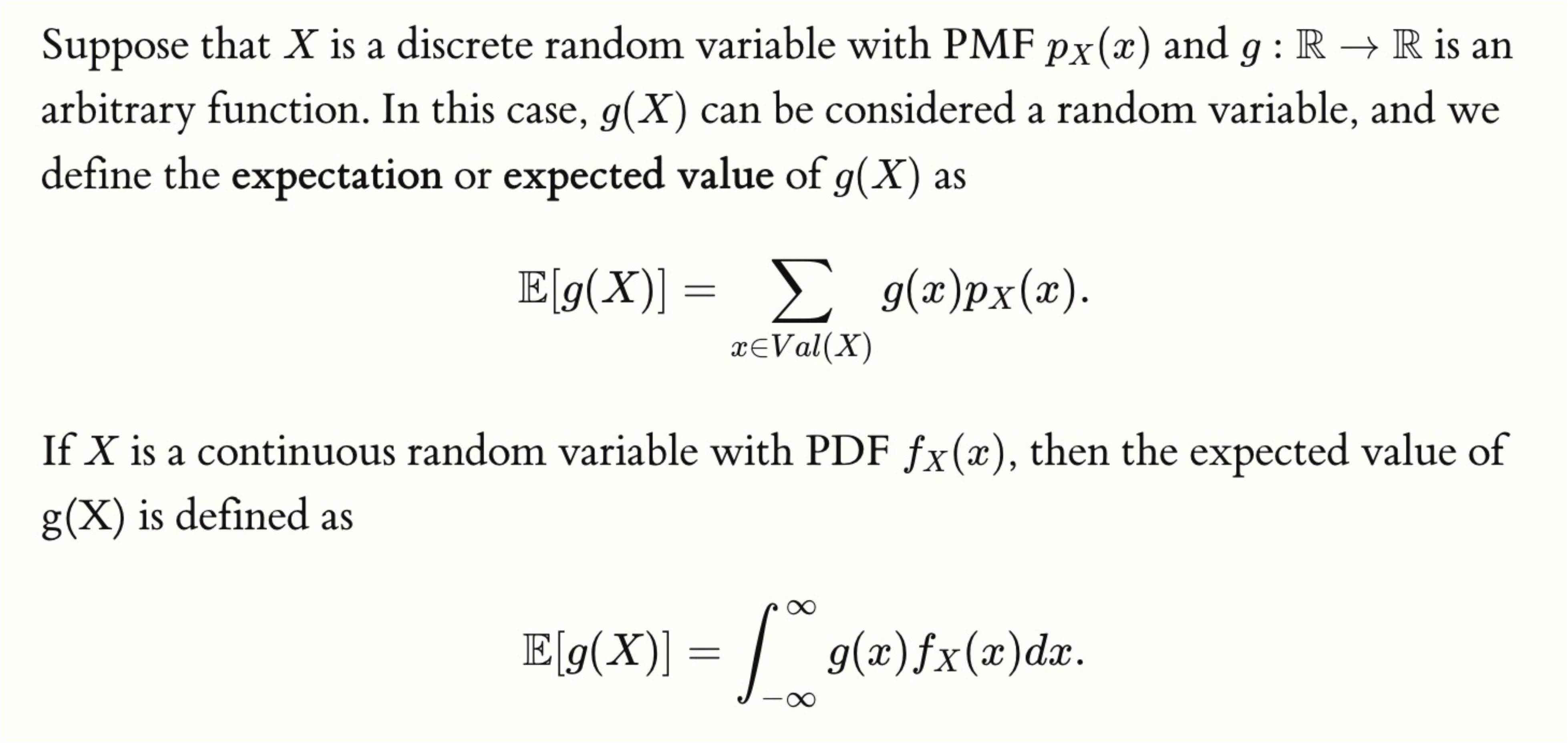
- $X \sim \text{Exponential}(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

- $X \sim \text{Normal}(\mu, \sigma^2)$: also known as the Gaussian distribution

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

Expectation



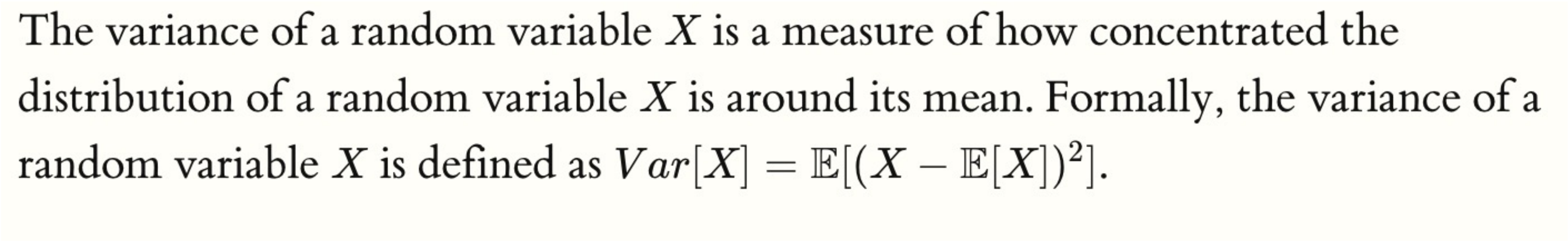
Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the **expectation** or **expected value** of $g(X)$ as

$$\mathbb{E}[g(X)] = \sum_{x \in Val(X)} g(x) p_X(x).$$

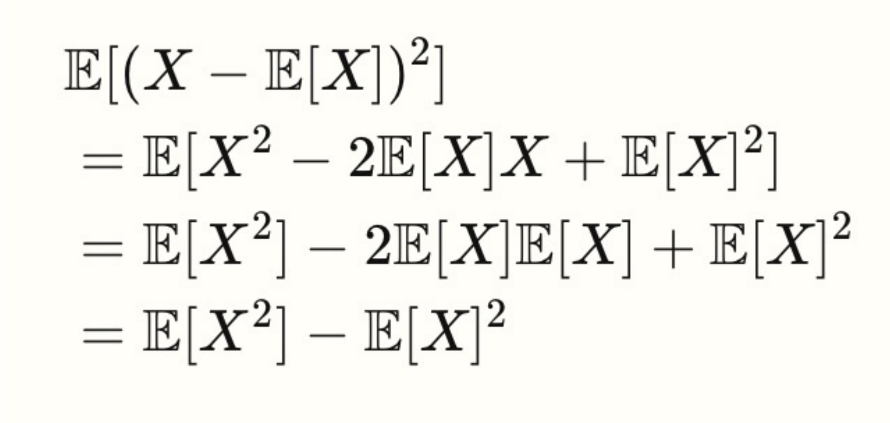
If X is a continuous random variable with PDF $f_X(x)$, then the expected value of g(X) is defined as

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

Variance

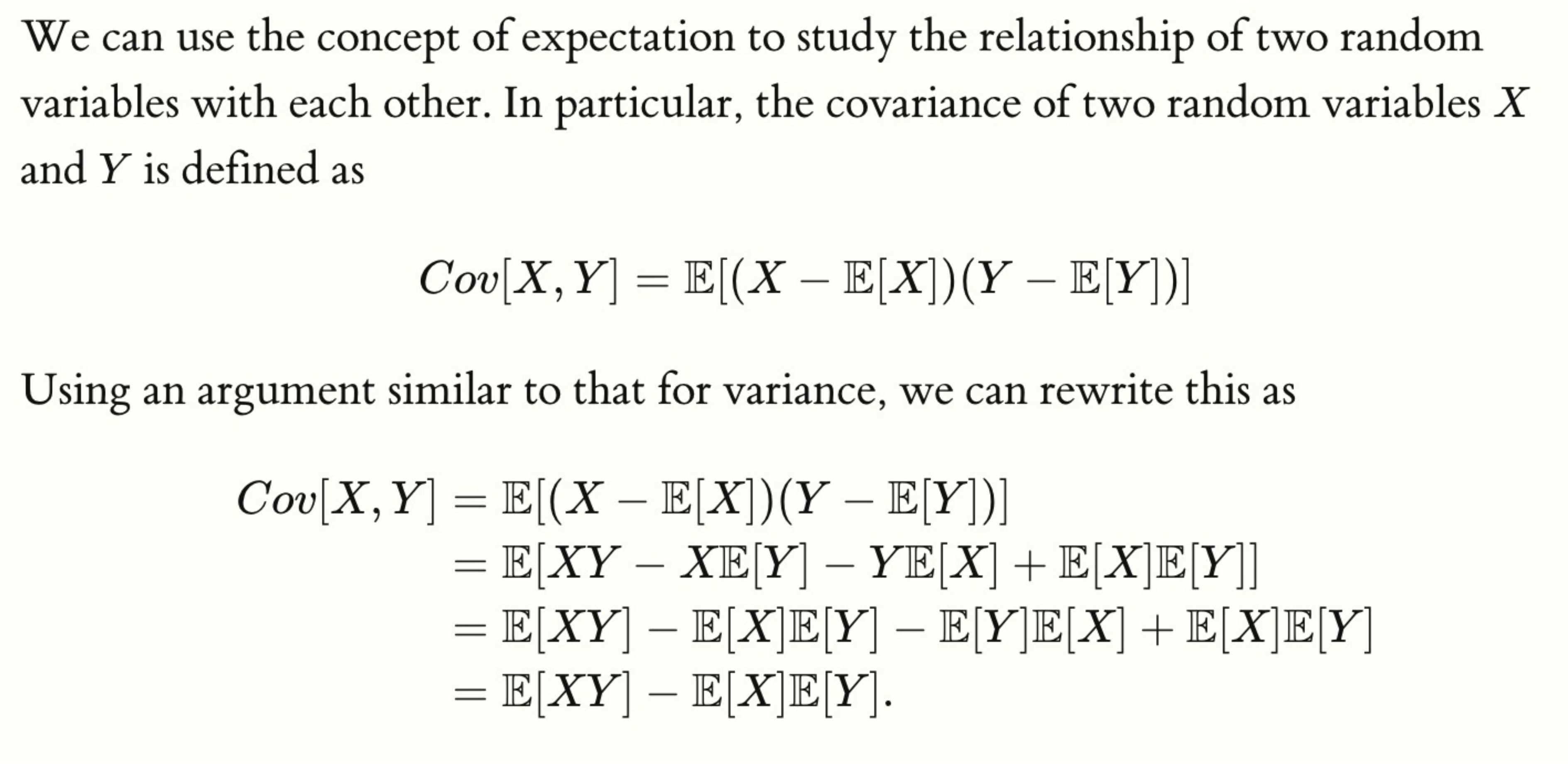


The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as $Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$.



$$\begin{aligned}
&\mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2
\end{aligned}$$

Covariance

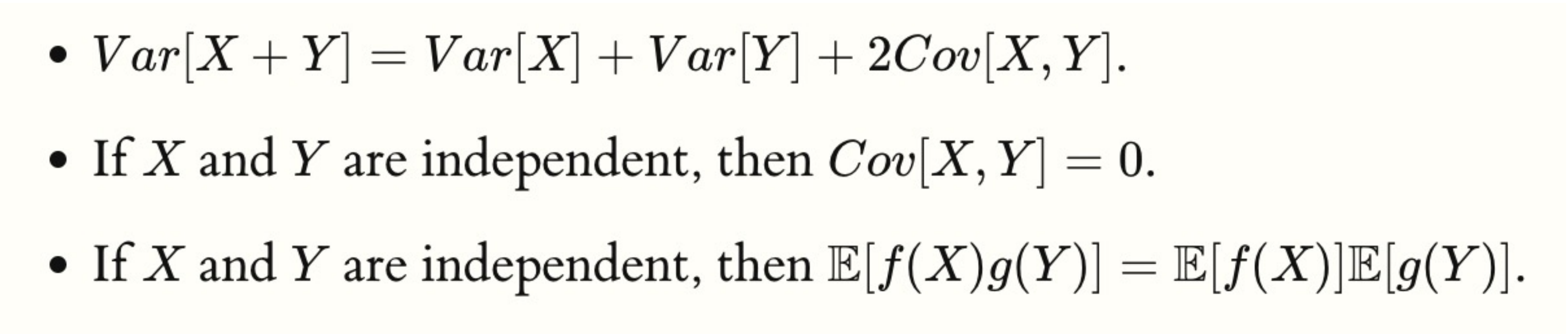


We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the covariance of two random variables X and Y is defined as

$$Cov[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Using an argument similar to that for variance, we can rewrite this as

$$\begin{aligned} Cov[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \end{aligned}$$



- $Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$.
- If X and Y are independent, then $Cov[X, Y] = 0$.
- If X and Y are independent, then $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$.

Slide Credits

- Stanford CS 229, Linear Algebra Review, Zico Kolter.
- Stanford CS 229, Probability Review, Maleki & Do.