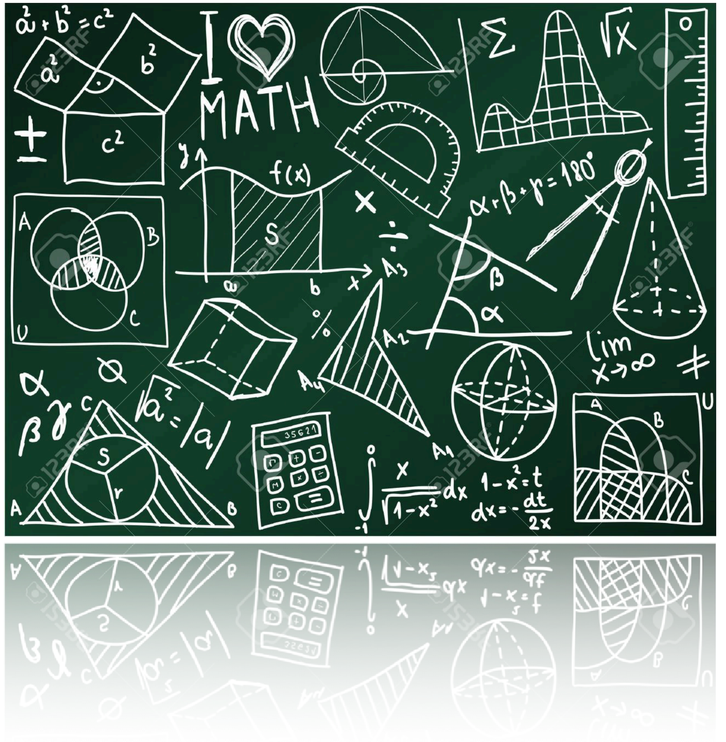
Introduction to Deep Learning (I2DL)

Exercise 2: Math Recap

Overview

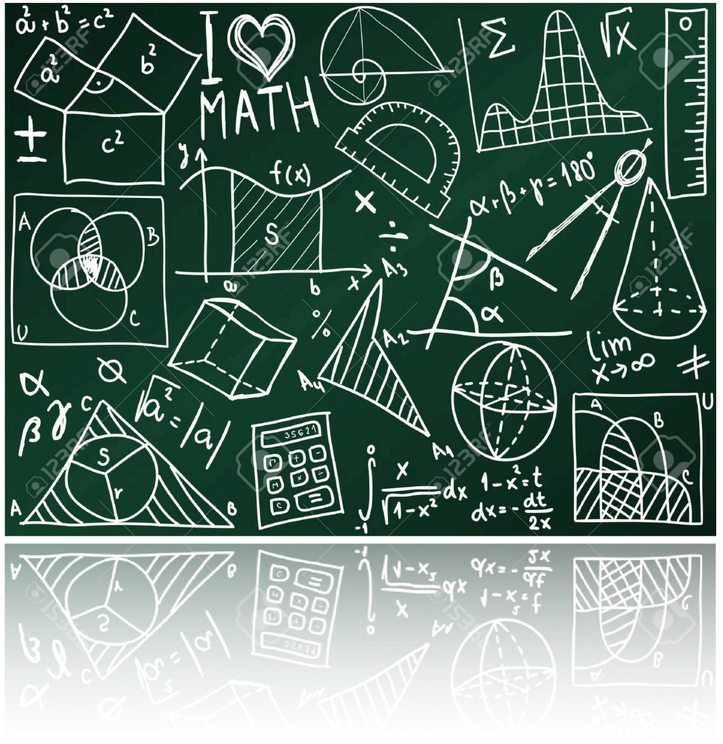
- **Linear Algebra**
  - Vectors and matrices
  - Basic operations on matrices & vectors
  - Tensors
  - Norms, Loss functions
- **Calculus**
  - Scalar derivatives
  - Gradient
  - Jacobian Matrix
  - Chain Rule
- **Probability Theory**
  - Probability space
  - Random variables
  - PMF, PDF, CDF
  - Mean, variance
  - Standard probability distributions



Linear Algebra

Overview

- **Linear Algebra**
  - Vectors and matrices
  - Basic operations on matrices & vectors
  - Tensors
  - Norms, Loss functions
- **Calculus**
  - Scalar derivatives
  - Gradient
  - Jacobian Matrix
  - Chain Rule
- **Probability Theory**
  - Probability space
  - Random variables
  - PMF, PDF, CDF
  - Mean, variance
  - Standard probability distributions



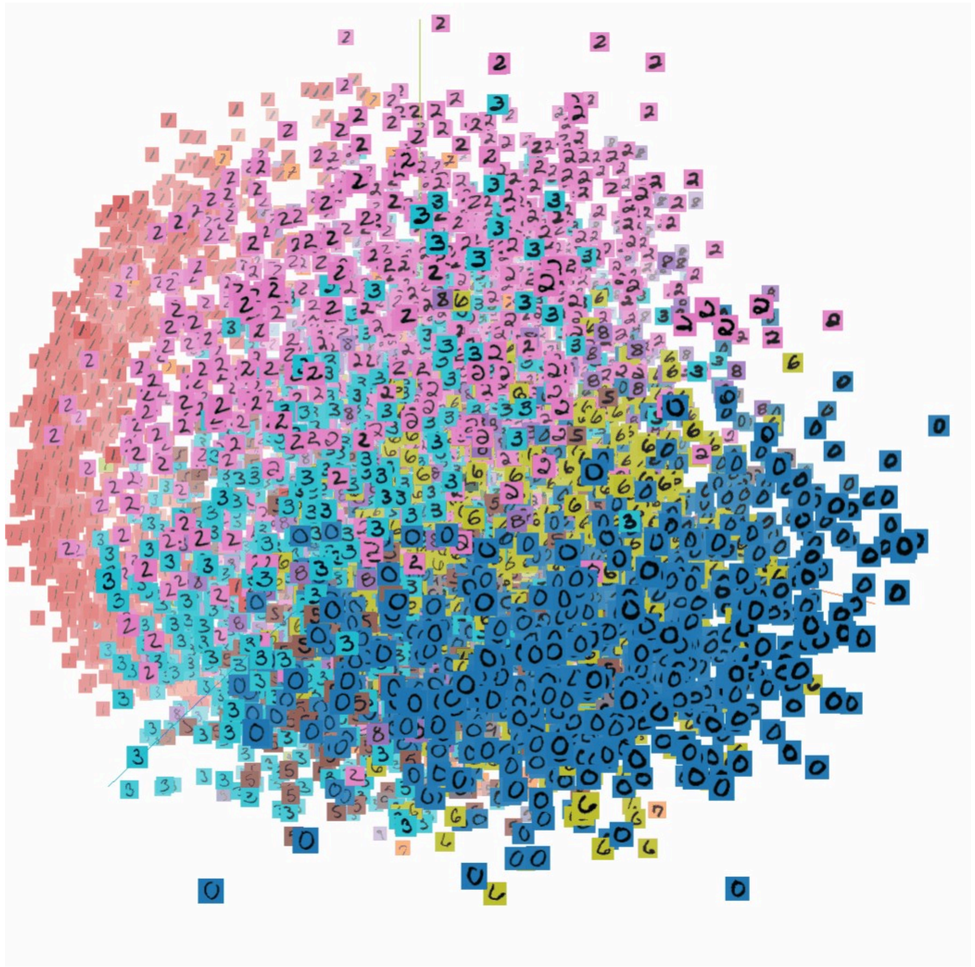
Basic Notation

- **Vector:** An element of $\mathbb{R}^n$ is a vector with n entries.
- **Elements of a vector:** The i-th element of a vector $v \in \mathbb{R}^n$ is denoted by $v_i \in \mathbb{R}$.
- **Matrix:** An element of $\mathbb{R}^{n \times m}$ is a matrix with n rows and m columns.
- **Elements of a matrix:** For $A \in \mathbb{R}^{n \times m}$, the element at the i-th row and j-th column is $A_{ij} \in \mathbb{R}$.
- **Transpose:** Flip rows and columns. For $A \in \mathbb{R}^{n \times m}$, the transpose is $A^T \in \mathbb{R}^{m \times n}$. Vectors can be transposed similarly.

Vector

- An n-dimensional vector describes an element in an n-dimensional space.

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix} \in \mathbb{R}^n$$

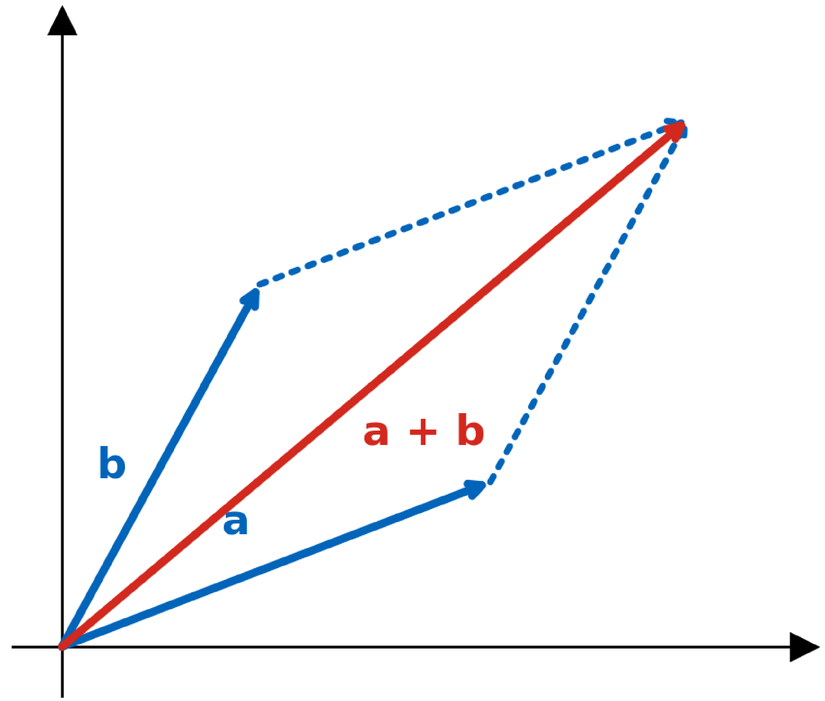


- **Vector operations:**
  - Addition
  - Subtraction
  - Scalar multiplication
  - Dot product

Vector Operations

- **Definition:** For $a, b \in \mathbb{R}^n$ we have

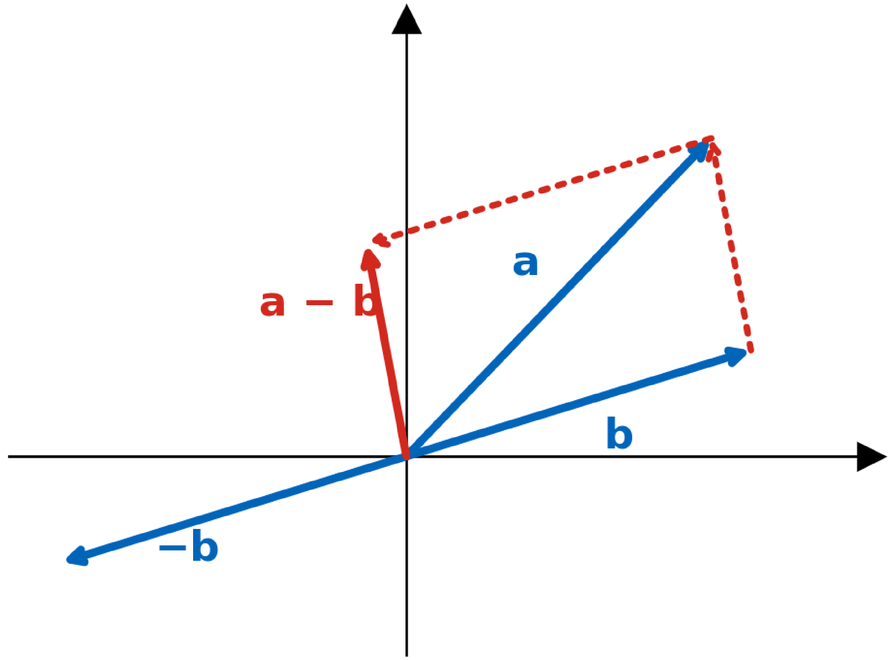
$$a + b = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_n + b_n \end{pmatrix} \in \mathbb{R}^n$$



Vector Operations

- **Definition:** For $a, b \in \mathbb{R}^n$ we have

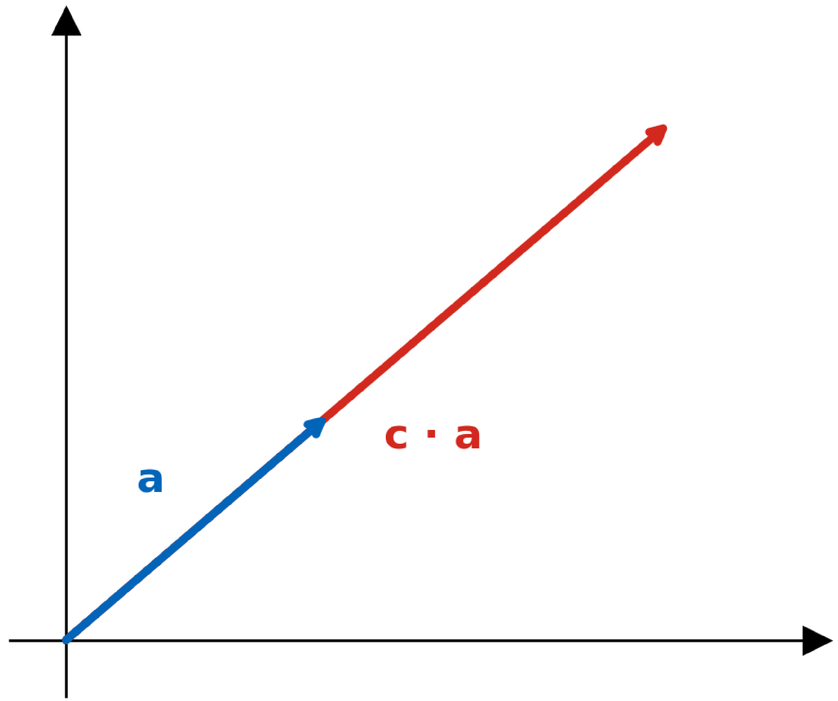
$$a - b = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ \vdots \\ a_n - b_n \end{pmatrix} \in \mathbb{R}^n$$



Vector Operations

- **Definition:** For $a \in \mathbb{R}^n,\ c \in \mathbb{R}$ we have

$$c \cdot a = \begin{pmatrix} c \cdot a_1 \\ c \cdot a_2 \\ \vdots \\ c \cdot a_n \end{pmatrix} \in \mathbb{R}^n$$



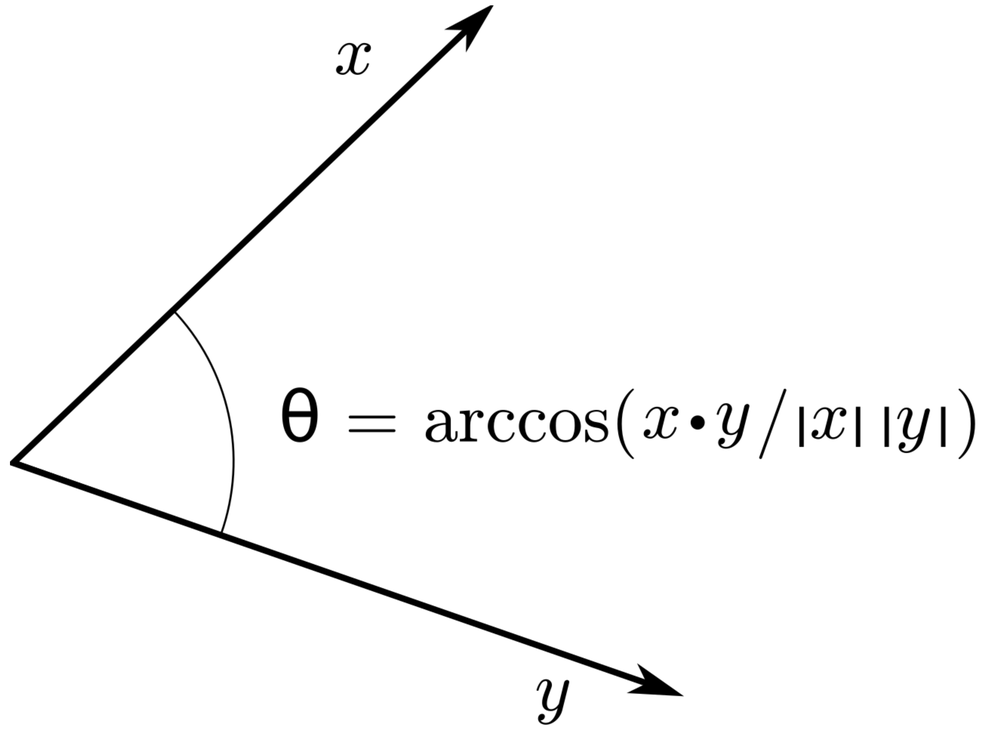
Vector Operations

- **Definition:** For $a, b \in \mathbb{R}^n$, the dot product is defined as follows:

$$a \cdot b = a^T \cdot b = a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n \cdot b_n = \sum_{i=1}^{n} a_i \cdot b_i \in \mathbb{R}$$

Vector Operations

- **Properties:**
  - Commutative: $a \cdot b = b \cdot a$
  - Geometric interpretation:
    $a \cdot b = \|a\| \cdot \|b\| \cdot \cos(\theta)$
  - Orthogonality: Two non-zero vectors are orthogonal to each other
    $\iff a \cdot b = 0$



Matrix

Matrix-vector Multiplication | Matrix-matrix Multiplication | Hadamard Product

- A matrix $A \in \mathbb{R}^{n \times m}$ is denoted as

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

Matrix Operations

Matrix-vector Multiplication | Matrix-matrix Multiplication | Hadamard Product

- Multiplication of matrix with a vector is defined as follows: for $A \in \mathbb{R}^{n \times m},\ b \in \mathbb{R}^m$,

$$A \cdot b = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} a_{11}b_1 + \cdots + a_{1m}b_m \\ a_{21}b_1 + \cdots + a_{2m}b_m \\ \vdots \\ a_{n1}b_1 + \cdots + a_{nm}b_m \end{pmatrix} \in \mathbb{R}^n$$

- **Attention:** The respective dimensions have to fit, otherwise the multiplication is not well-defined.

$$\underbrace{A}_{n \times m} \cdot \underbrace{b}_{m \times 1} = \underbrace{c}_{n \times 1}$$

- **Example:** $A \in \mathbb{R}^{3 \times 2},\ b \in \mathbb{R}^2$ with $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix},\ b = \begin{pmatrix} 2 \\ 3 \end{pmatrix} \implies A \cdot b = \begin{pmatrix} 8 \\ 18 \\ 28 \end{pmatrix}.$

Matrix Operations

Matrix-vector Multiplication | **Matrix-matrix Multiplication** | Hadamard Product

- Similarly, the multiplication of two matrices is defined as follows. For $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{m \times l}$ we have

$$A \cdot B = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & \cdots & b_{1l} \\ \vdots & \ddots & \vdots \\ b_{m1} & \cdots & b_{ml} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1l} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nl} \end{pmatrix} \in \mathbb{R}^{n \times l}$$

- where $c_{ij} = \sum_{k=1}^{m} a_{ik} \cdot b_{kj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{im}b_{mj}$.
- **Attention:** Matrix multiplication is in general **not** commutative, i.e. for $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{m \times n}$ we have $A \cdot B \neq B \cdot A$.

Matrix Operations

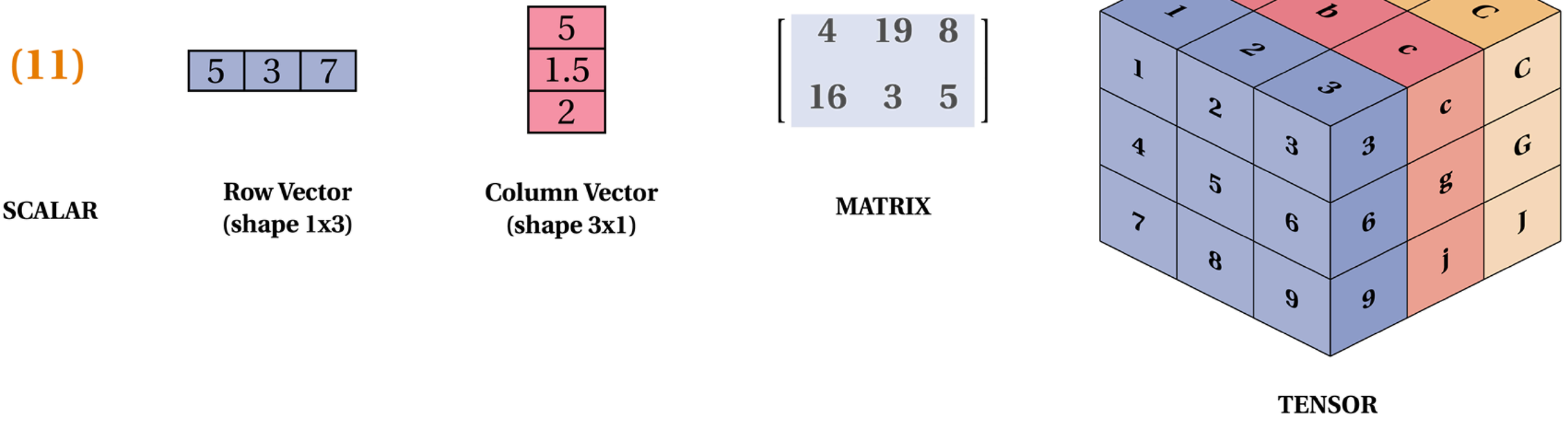
- The Hadamard product is the element-wise product of two matrices. For two matrices of the same dimension $A, B \in \mathbb{R}^{n \times m}$ it is given by

$$A \odot B = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ a_{21} & \cdots & a_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ b_{21} & \cdots & b_{2m} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nm} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1m}b_{1m} \\ a_{21}b_{21} & \cdots & a_{2m}b_{2m} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{n1} & \cdots & a_{nm}b_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m}$$

- $\Rightarrow$ *For all matrix operations, it is important to check the dimensions!*

Tensor

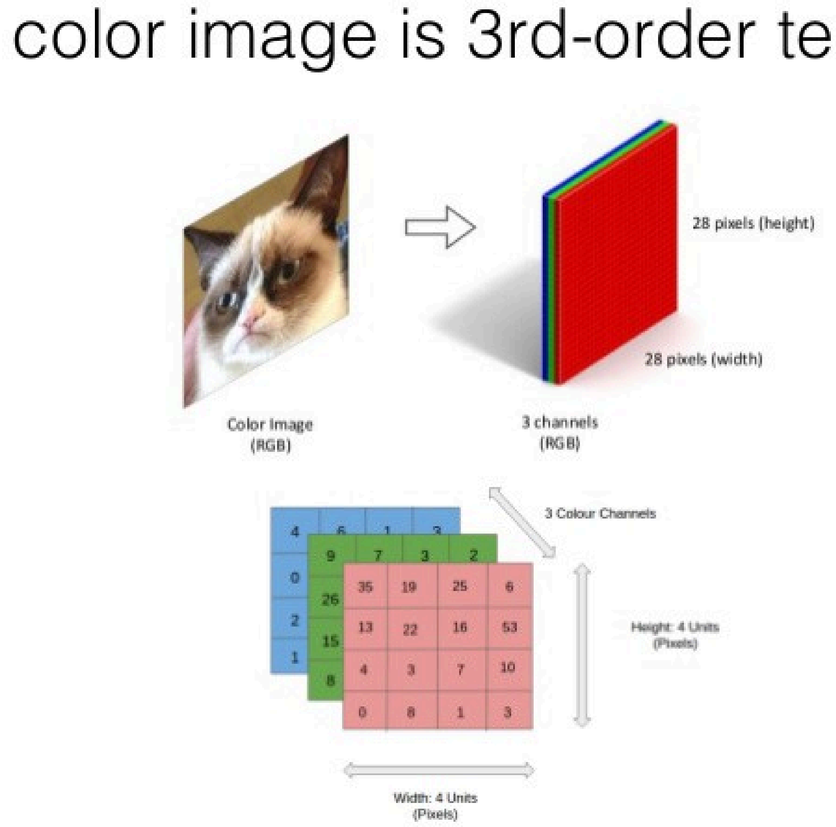
- **Definition:** A tensor is a multidimensional array and a generalization of the concepts of a vector and a matrix.



Tensors in Computer Vision

- *Tensors are used to represent RGB images.*

$$H \times W \times RGB$$



Norm

- **Norm:** measure of the "length" of a vector
- **Definition: A norm is a non-negative function $\|\cdot\| : V \to \mathbb{R}$ which is defined by the following the properties for elements $v, w \in V$:**
  - Triangle inequality: $\|v + w\| \leq \|v\| + \|w\|$
  - $\|a \cdot v\| = a \cdot \|v\|$ for a scalar a
  - $\|v\| = 0$ if and only if $v = 0$
- (V is a vector space over a field $\mathbb{F}$; in our case we have $V = \mathbb{R}^n$)
- **Remark:** Every such function defines a norm on the vector space.
- **Examples:** L1-norm, L2-norm

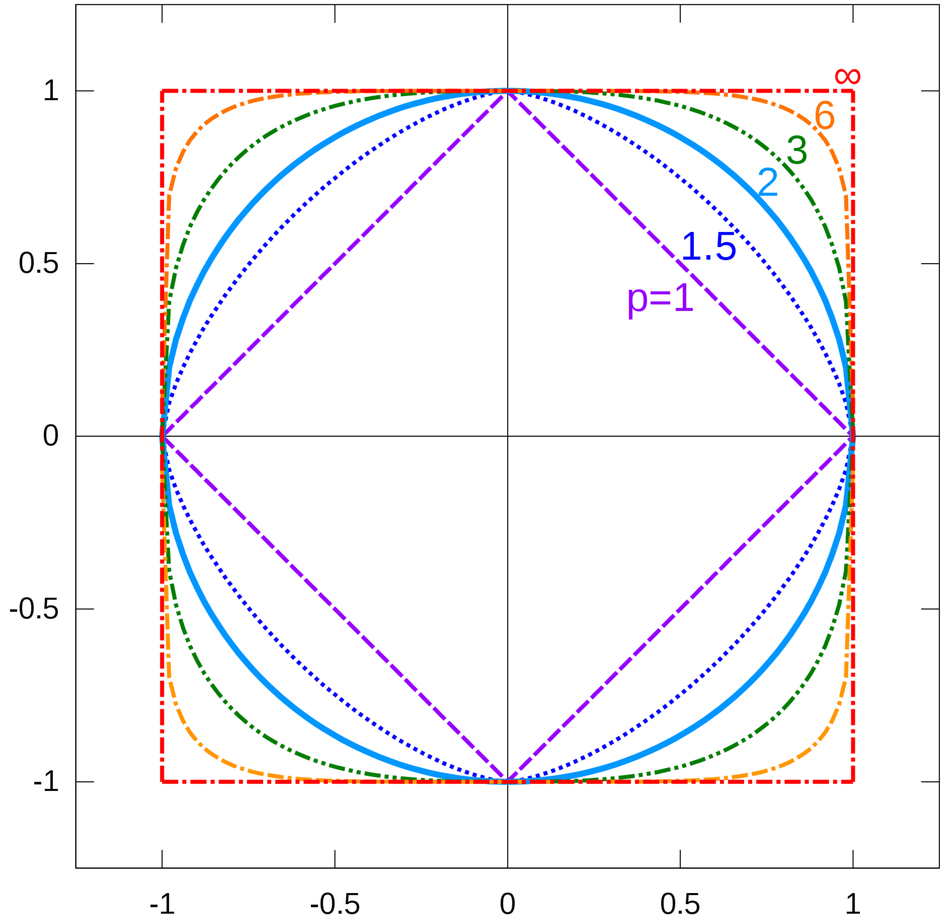
L1-Norm

- **Norm:** measure of the "length" of a vector
- **L1-Norm:** We denote the L1-norm with $\|\cdot\|_1 : \mathbb{R}^n \rightarrow \mathbb{R}$ such that for a vector $v = (v_1, v_2, \ldots, v_n)$

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|$$

- **Example:** Let $v = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix} \in \mathbb{R}^3$, then

$$\|v\|_1 = (1 + 3 + 2) = 6$$



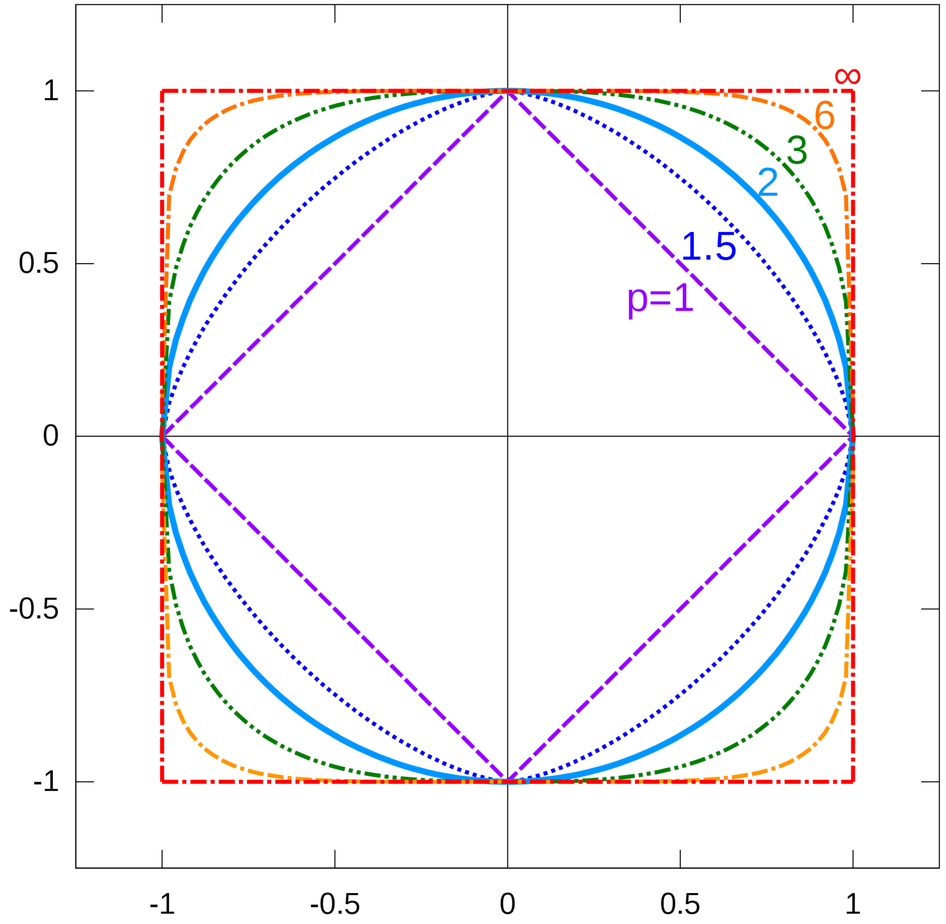
L2-Norm

- **Norm:** measure of the "length" of a vector
- **L2-Norm:** We denote the L2-norm with $\|\cdot\|_2 : \mathbb{R}^n \to \mathbb{R}$ such that for a vector $v = (v_1, v_2, \ldots, v_n)$

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} (v_i)^2}$$

- **Example:** Let $v = \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix} \in \mathbb{R}^3$, then

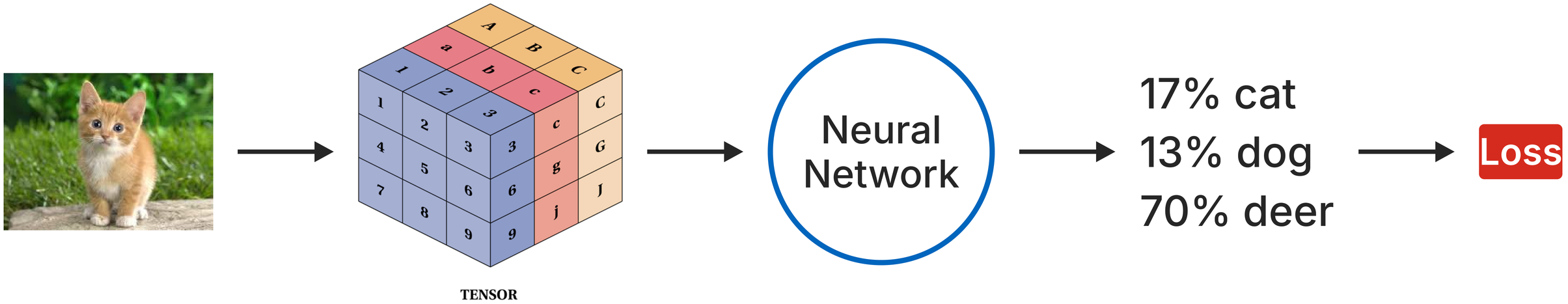
$$\|v\|_2 = \sqrt{1^2 + (-3)^2 + 2^2} = \sqrt{14}$$



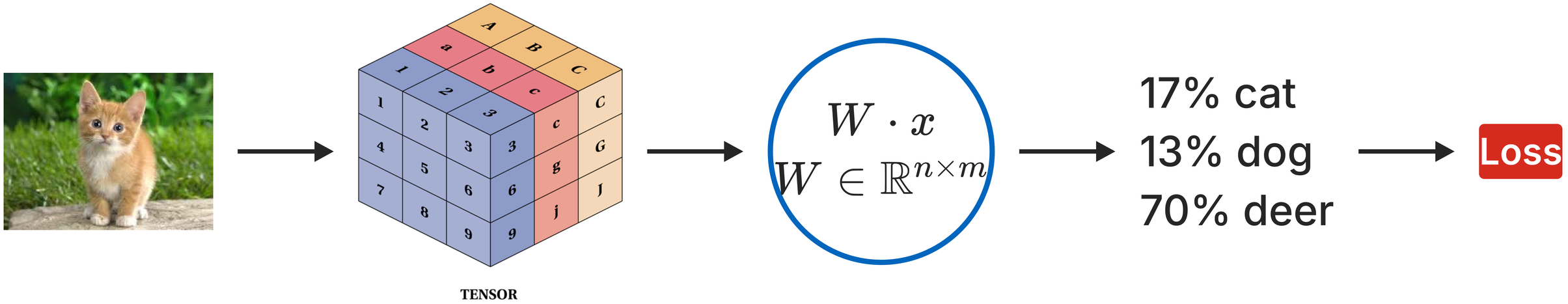
Loss functions

- **A loss function is a function that takes as input two vectors and as output measures the distance between these two**
  - $\rightarrow$ uses a norm to measure the distance
  - $\rightarrow$ L1-Loss uses the L1-norm, L2-Loss uses the L2-norm
- **L1-Loss:** The L1-Loss between two vectors $v, w \in \mathbb{R}^n$ is defined as $L_1(v, w) = \|v - w\|_1 = \sum_{i=1}^{n} |v_i - w_i|$.
- **L2-Loss:** The L2-Loss between two vectors $v, w \in \mathbb{R}^n$ is defined as $L_2(v, w) = \|v - w\|_2 = \sqrt{(v_1 - w_1)^2 + \ldots + (v_n - w_n)^2}$.

Outlook

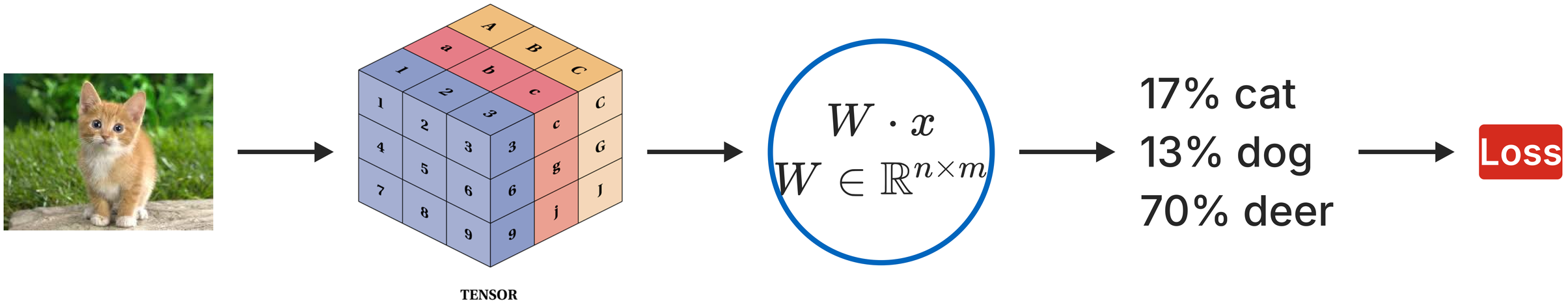


Outlook



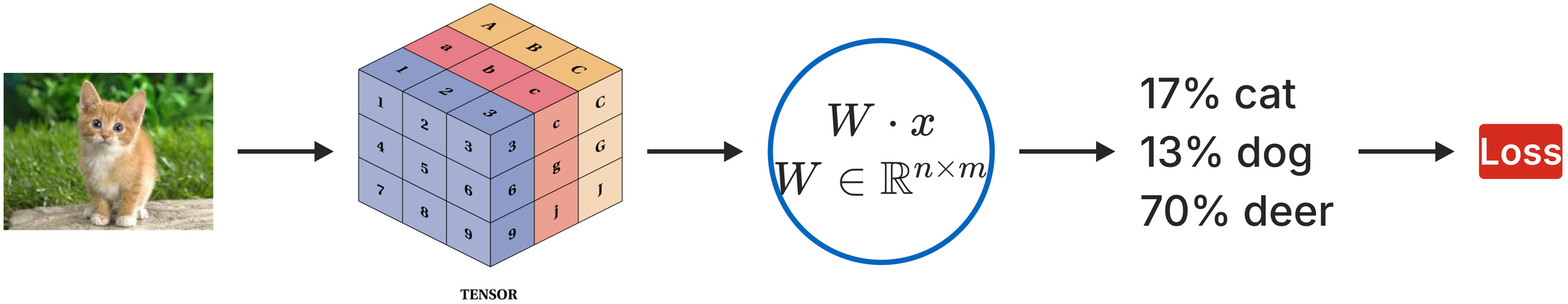
The elements of the matrix W are called weights and they determine the prediction of our network.

Outlook



How can we get an accurate matrix W to minimize the loss?

Outlook

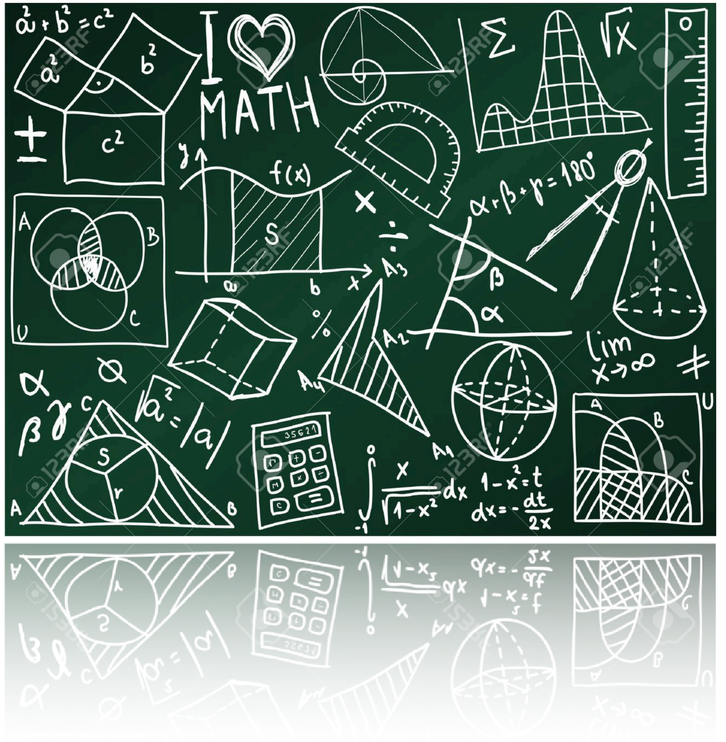


Gradient Descent: Method to approximate the best values for the weights

Calculus

Overview

- **Linear Algebra**
  - Vectors and matrices
  - Basic operations on matrices & vectors
  - Tensors
  - Norms, Loss functions
- **Calculus**
  - Scalar derivatives
  - Gradient
  - Jacobian Matrix
  - Chain Rule
- **Probability Theory**
  - Probability space
  - Random variables
  - PMF, PDF, CDF
  - Mean, variance
  - Standard probability distributions



Derivatives

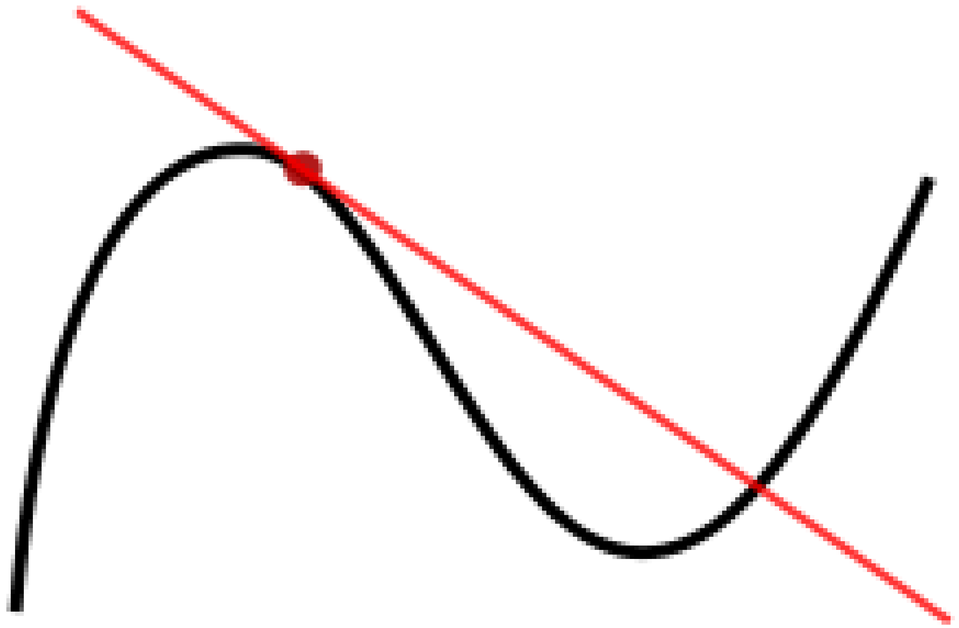
- **Well known:** Scalar derivatives, i.e. derivatives of functions $f : \mathbb{R} \rightarrow \mathbb{R}$
- **Matrix calculus:** Extension of calculus to higher dimensional setting, i.e. functions like $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $f : \mathbb{R} \rightarrow \mathbb{R}^n$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ for $n, m \in \mathbb{N}$
- Actual calculus we use is relatively trivial, but the notation can often make things look much more difficult than they are.

Overview

Setting	Derivative	Notation
$f : \mathbb{R} \to \mathbb{R}$	Scalar derivative	$f'(x)$
$f : \mathbb{R}^n \to \mathbb{R}$	Gradient	$\nabla f(x)$
$f : \mathbb{R}^{n \times m} \to \mathbb{R}$	Gradient	$\nabla f(x)$
$f : \mathbb{R}^n \to \mathbb{R}^m$	Jacobian	J_f

Scalar derivatives

- **Setting:** $f : \mathbb{R} \rightarrow \mathbb{R}$
- **Notation:** $f'(x)$ or $\frac{df}{dx}$
- **Derivative:** Derivative of a function at a chosen input value is the slope of the tangent line to the graph of the function at that point.



Derivation Rules

Common functions	Derivative
$f(x) = c \text{ for } c \in \mathbb{R}$	$f'(x) = 0$
$f(x) = x$	$f'(x) = 1$
$f(x) = x^n \text{ for } n \in \mathbb{N}$	$f'(x) = n \cdot x^{n-1}$
$f(x) = e^x$	$f'(x) = e^x$
$f(x) = \ln(x)$	$f'(x) = \frac{1}{x}$
$f(x) = \sin(x)$	$f'(x) = \cos(x)$
$f(x) = \cos(x)$	$f'(x) = -\sin(x)$

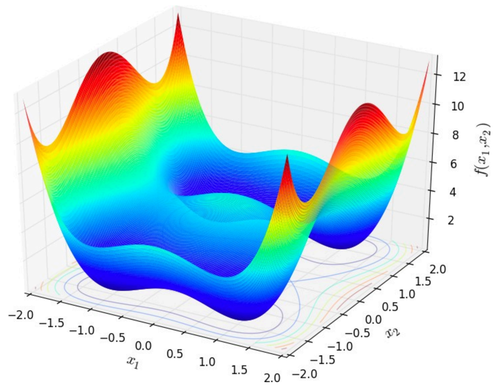
Derivation Rules

Rule	Function	Derivative
Sum rule	$f(x) + g(x)$	$f'(x) + g'(x)$
Difference rule	$f(x) - g(x)$	$f'(x) - g'(x)$
Multiplication by constant	$c \cdot f(x)$	$c \cdot f'(x)$
Product rule	$f(x) \cdot g(x)$	$f'(x) \cdot g(x) + f(x) \cdot g'(x)$
Quotient rule	$\frac{f(x)}{g(x)}$	$\frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{(g(x))^2}$
Chain rule	$f(g(x))$	$f'(g(x)) \cdot g'(x)$

Multivariate functions $f : \mathbb{R}^n \to \mathbb{R}$

Multivariate Function

- $f : \mathbb{R}^n \to \mathbb{R}$



Gradient

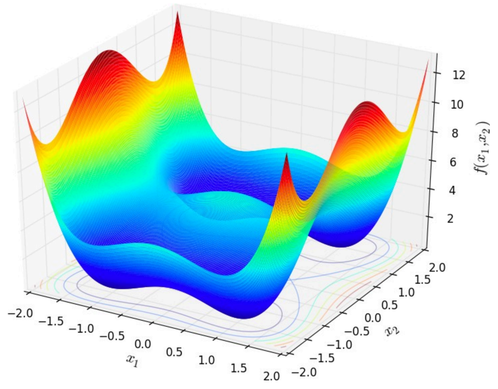
- $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$
- *Partial derivative:* $\frac{\partial f(x)}{\partial x_i}$ is the derivative of f with respect to x_i treating the other variables as constants.

$$\nabla f : x \mapsto \nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{pmatrix}$$

Multivariate functions $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$

Multivariate Function

- $f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$



Gradient

- $\nabla f : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^{n \times m}$

$$\nabla f : x \mapsto \nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_{11}} & \frac{\partial f(x)}{\partial x_{12}} & \cdots & \frac{\partial f(x)}{\partial x_{1m}} \\ \frac{\partial f(x)}{\partial x_{21}} & \frac{\partial f(x)}{\partial x_{22}} & \cdots & \frac{\partial f(x)}{\partial x_{2m}} \\ \vdots & & & \vdots \\ \frac{\partial f(x)}{\partial x_{n1}} & \frac{\partial f(x)}{\partial x_{n2}} & \cdots & \frac{\partial f(x)}{\partial x_{nm}} \end{pmatrix}$$

Gradient – Example 1

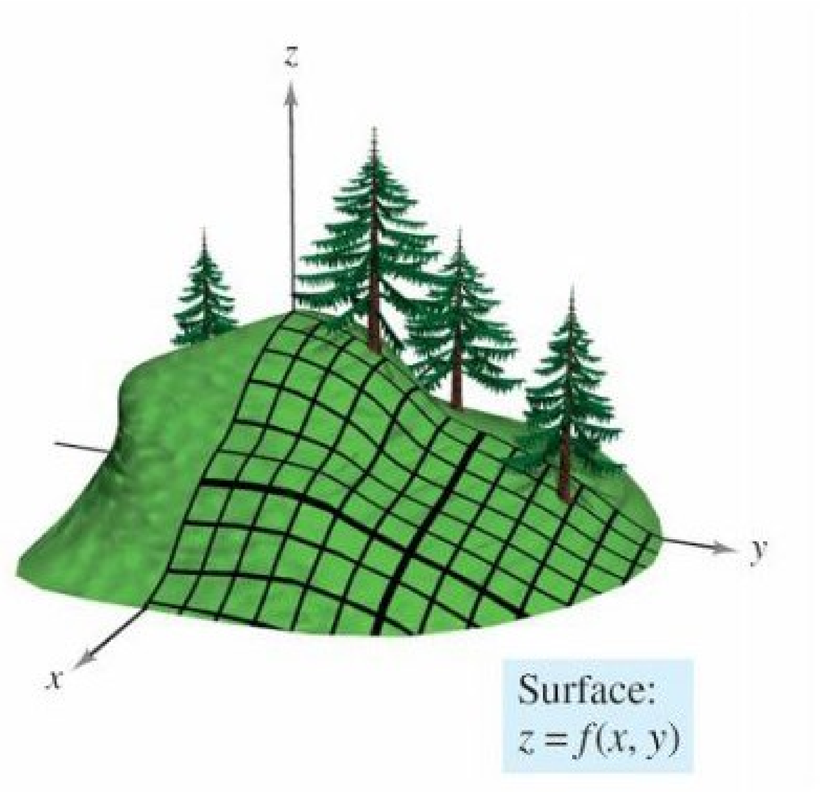
$$f(x, y) = 3x^2y$$

$$\nabla f(x, y) = \left[\frac{\partial f(x,y)}{\partial x}, \quad \frac{\partial f(x,y)}{\partial y} \right]$$

$$\frac{\partial}{\partial x} 3yx^2 = 3y \frac{\partial}{\partial x} x^2 = 3y \cdot 2x = 6yx$$

$$\frac{\partial}{\partial y} 3x^2y = 3x^2 \frac{\partial y}{\partial y} = 3x^2 \times 1 = 3x^2$$

$$\nabla f(x, y) = \left[6yx, \quad 3x^2 \right]$$



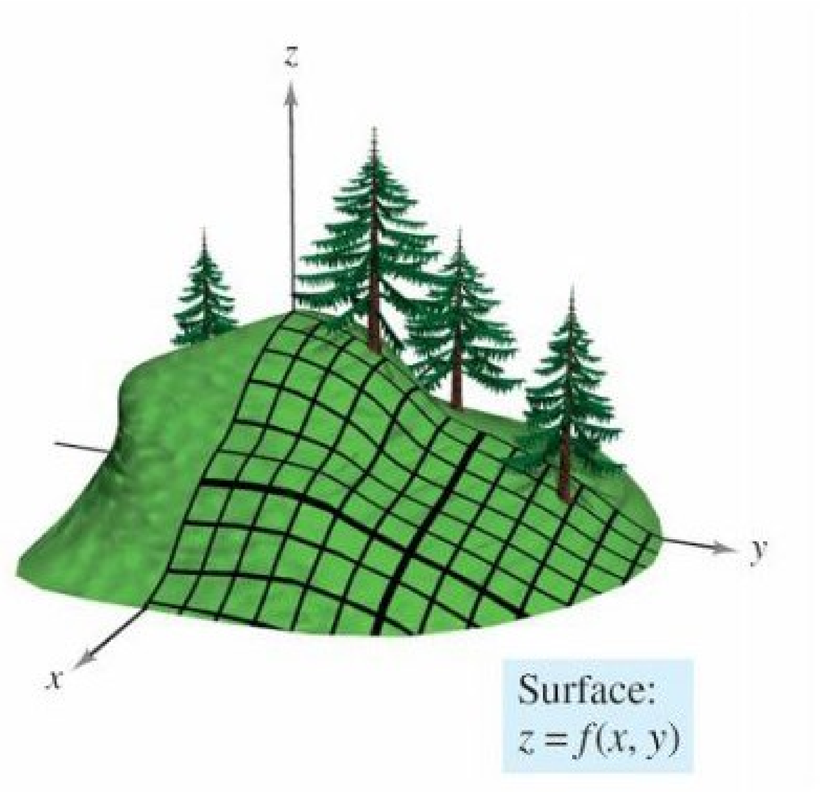
Gradient – Example 2

$$g(x, y) = 2x + y^8$$

$$\frac{\partial g(x,y)}{\partial x} = \frac{\partial\, 2x}{\partial x} + \frac{\partial\, y^8}{\partial x} = 2\,\frac{\partial x}{\partial x} + 0 = 2 \times 1 = 2$$

$$\frac{\partial g(x,y)}{\partial y} = \frac{\partial\, 2x}{\partial y} + \frac{\partial\, y^8}{\partial y} = 0 + 8y^7 = 8y^7$$

$$\nabla g(x, y) = \left[\, 2,\ \ 8y^7 \,\right]$$



Vector-valued functions

Vector-Valued function

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^m$$

$$f : x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \longrightarrow \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}$$

Jacobian Matrix

$$J_f : \mathbb{R}^n \rightarrow \mathbb{R}^{m \times n}$$

$$x \mapsto J_f(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix}$$

Jacobian Matrix – Example 3

- **Assume** $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with $f(x, y) = \begin{pmatrix} f_1(x, y) \\ f_2(x, y) \end{pmatrix}$ where $f_1(x, y) = 3x^2y$ and $f_2(x, y) = 2x + y^8$.
- **Calculate Jacobian matrix:**

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1(x,y)}{\partial x} & \frac{\partial f_1(x,y)}{\partial y} \\ \frac{\partial f_2(x,y)}{\partial x} & \frac{\partial f_2(x,y)}{\partial y} \end{pmatrix} = \begin{pmatrix} 6xy & 3x^2 \\ 2 & 8y^7 \end{pmatrix}$$

Single Variable Chain Rule

- **Setting:** We are given the function $h(x) = f(g(x))$.
- **Task:** Compute the derivative of this function with chain rule.
- **1. Introduce the intermediate variable:** Let $u = g(x)$ be the intermediate variable.
- **2. Compute individual derivatives:** $\frac{\mathrm{d}f}{\mathrm{d}u}$ and $\frac{\mathrm{d}g}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}x}$
- **3. Chain rule:** $\frac{\mathrm{d}h}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}$
- **4. Substitute intermediate variables back**

Single Variable Chain Rule: Example

- **Example:** Let $h(x) = \sin(x^2)$.
- **Task:** Compute the derivative of this function with chain rule.
- **Observation:** Here, $h(x) = f(g(x))$ with $f(x) = \sin(x)$ and $g(x) = x^2$.
- **1. Introduce the intermediate variable:** Let $u = x^2$ be the intermediate variable.
- **2. Compute individual derivatives:** $\frac{\mathrm{d}f}{\mathrm{d}u} = \cos(u)$ and $\frac{\mathrm{d}g}{\mathrm{d}x} = \frac{\mathrm{d}u}{\mathrm{d}x} = 2x$
- **3. Chain rule:** $\frac{\mathrm{d}h}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x} = \cos(u) \cdot 2x$
- **4. Substitute intermediate variables back:** $\frac{\mathrm{d}h}{\mathrm{d}x} = \cos(u) \cdot 2x = \cos(x^2) \cdot 2x$

Total Derivative Chain Rule

- General Formalism:

$$\begin{aligned}\frac{\partial f(x,\, u_1(x),\, \ldots,\, u_n(x))}{\partial x} &= \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u_1}\frac{\partial u_1}{\partial x} + \frac{\partial f}{\partial u_2}\frac{\partial u_2}{\partial x} + \cdots + \frac{\partial f}{\partial u_n}\frac{\partial u_n}{\partial x} \\ &= \frac{\partial f}{\partial x} + \sum_{i=1}^{n} \frac{\partial f}{\partial u_i}\frac{\partial u_i}{\partial x}\end{aligned}$$

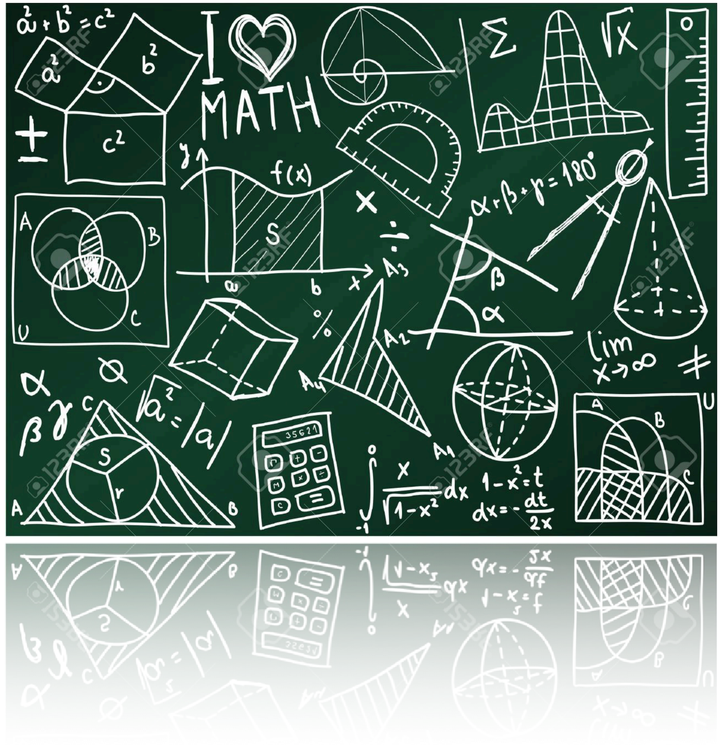
References

- https://en.wikipedia.org/wiki/Matrix_calculus
- http://parrt.cs.usfca.edu/doc/matrix-calculus/index.html
- https://arxiv.org/pdf/1802.01528.pdf
- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives
- https://explained.ai/matrix-calculus/
- http://www.deeplearningbook.org/contents/part_basics.html
- https://towardsdatascience.com/calculating-gradient-descent-manually-6d9bee09aa0b

Probability Theory

Overview

- **Linear Algebra**
  - Vectors and matrices
  - Basic operations on matrices & vectors
  - Tensors
  - Norms, Loss functions
- **Calculus**
  - Scalar derivatives
  - Gradient
  - Jacobian Matrix
  - Chain Rule
- **Probability Theory**
  - Probability space
  - Random variables
  - PMF, PDF, CDF
  - Mean, variance
  - Standard probability distributions



Probability space $(\Omega, \mathcal{F}, \mathbb{P})$

- A probability space consists of three elements $(\Omega,\ \mathcal{F},\ \mathbb{P})$:
- **Sample space Ω:** The set of all outcomes of a random experiment.
- **Event Space $\mathcal{F}$:** A set whose elements $A \in \mathcal{F}$ (called *events*) are subsets of Ω.
- **Probability measure $\mathbb{P}$: A function $\mathbb{P} : \mathcal{F} \to [0,\ 1]$ that satisfies the following three properties:**
  - $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$
  - $\mathbb{P}(\Omega) = 1$
  - $\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$ for $n \in \mathbb{N}$ and disjoint events $A_1, A_2, \ldots, A_n \in \mathcal{F}$
- *The probability space provides a formal model of a random experiment.*

Probability space: Example

- A probability space consists of three elements: $(\Omega,\ \mathcal{F},\ \mathbb{P})$ *(recap)*
- **Example: Tossing a six-sided die**
- **Sample space:** $\Omega = \{1, 2, 3, 4, 5, 6\}$
- **Event space:** $\mathcal{F}_1 = \{\emptyset,\ \Omega\},\ \ \mathcal{F}_2 = \mathcal{P}(\Omega),$
- $\mathcal{F}_3 = \{\emptyset,\ A_1 = \{1, 3, 5\},\ A_2 = \{2, 4, 6\},\ \Omega = \{1, 2, 3, 4, 5, 6\}\}$
- **Probability measure** $\mathbb{P}$**:** $\mathcal{F} \to \mathbb{R}$ with $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$, and in the case of $\mathcal{F}_3$ we know that $\mathbb{P}(A_1) + \mathbb{P}(A_2) = 1.$
- **Example event space** $\mathcal{F}_3$**: Possible probability measures are**
  - $\mathbb{P}_1(A_1) = \frac{1}{2} = \mathbb{P}_1(A_2)$
  - $\mathbb{P}_2(A_1) = \frac{1}{4}$ and $\mathbb{P}_2(A_2) = \frac{3}{4}.$



Random variable

- A random variable is a function defined on the probability space which maps from the sample space to the real numbers, i.e.

$$X : \Omega \to \mathbb{R}.$$

- We distinguish between **discrete** and **continuous** random variables.

Random variable

- A random variable is a function defined on the probability space which maps from the sample space to the real numbers, i.e. $X : \Omega \to \mathbb{R}$. *(recap)*
- **Example: Tossing a fair six-sided die**
- **Underlying experiment:** $\Omega = \{1, 2, 3, 4, 5, 6\},\ \mathcal{F} = \mathcal{P}(\Omega),\ \mathbb{P}(\{x\}) = \frac{1}{6}\ \forall x \in \Omega$
- **Random variable:** Number that appears on the die, $X : \Omega \to \{1, 2, 3, 4, 5, 6\}\ \Rightarrow$ *discrete* random variable
- **Example:** One element in Ω is $\omega = 4$. Then $X(\omega) = 4$.
- **Probability measure** $\mathbb{P}$**:** $\mathbb{P}(X = 4) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = \omega = 4\}) = \mathbb{P}(\{4\}) = \frac{1}{6}$



Random variable

- A random variable is a function defined on the probability space which maps from the sample space to the real numbers, i.e. $X : \Omega \to \mathbb{R}$. *(recap)*
- **Example: Flipping a fair coin two times**
- **Underlying experiment:** $\Omega = \{(H, H),\ (H, T),\ (T, H),\ (T, T)\},\ \mathcal{F} = \mathcal{P}(\Omega)$ and $\mathbb{P}(\{\omega\}) = \frac{1}{4}\ \forall \omega \in \Omega$
- **Random variable:** number of heads that appeared in the two flips, $X : \Omega \to \{0, 1, 2\}\ \Rightarrow$ *discrete* random variable
- **Example:** One element in Ω is $\omega = (T, H)$. Then $X(\omega) = 1$.
- **Probability measure** $\mathbb{P}$**:** $\mathbb{P}(X = 1) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = 1\}) = \mathbb{P}(\{(H, T),\ (T, H)\}) = \frac{1}{2}$



Random variable

- A random variable is a function defined on the probability space which maps from the sample space to the real numbers, i.e. $X : \Omega \to \mathbb{R}$. *(recap)*
- **Example: radioactive decay**
- **Underlying experiment:** $\Omega = \mathbb{R}_{\geq 0},\ \mathcal{F} = \mathcal{B}(\Omega),\ \mathbb{P} = \lambda$ is the Lebesgue measure
- **Random variable:** indicating amount of time that it takes for a radioactive particle to decay, $X : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \implies$ *continuous* random variable
- **Probability measure** $\mathbb{P}$**:** is defined on the set of events $\mathcal{F}$ and is now used for random variables as follows:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(\{\omega \in \Omega : a \leq X(\omega) \leq b\})$$

Probability measures

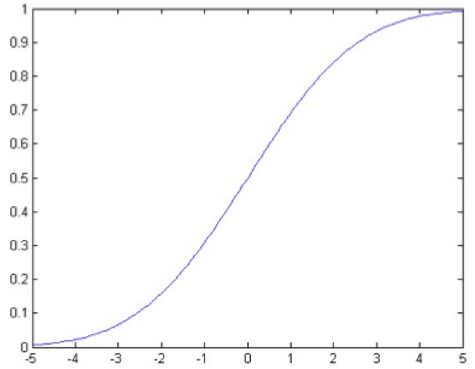
Random variable	Cumulative distribution function (CDF)	PMF / PDF
Discrete	$F_X(x) = \mathbb{P}(X \leq x)$	Probability mass function (PMF), $p_X(x) = \mathbb{P}(X = x)$
Continuous	$F_X(x) = \mathbb{P}(X \leq x)$	Probability density function (PDF)

Cumulative Distribution Function

- A **cumulative distribution function** (CDF) of a random variable X is a function $F_X : \mathbb{R} \rightarrow [0, 1]$ which is defined as

$$F_X(x) = \mathbb{P}(X \leq x).$$

- **Properties: Per definition, it satisfies the following properties:**
  - $0 \leq F_X(x) \leq 1$
  - $\lim_{x \to -\infty} F_X(x) = 0$
  - $\lim_{x \to \infty} F_X(x) = 1$
  - $\forall x \leq y \;\Rightarrow\; F_X(x) \leq F_X(y)$

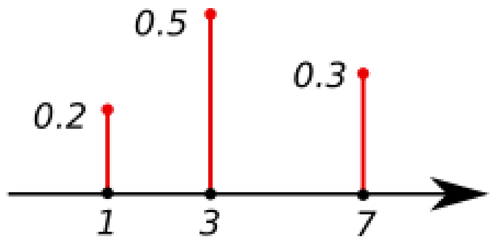


Discrete Case: Probability Mass Function

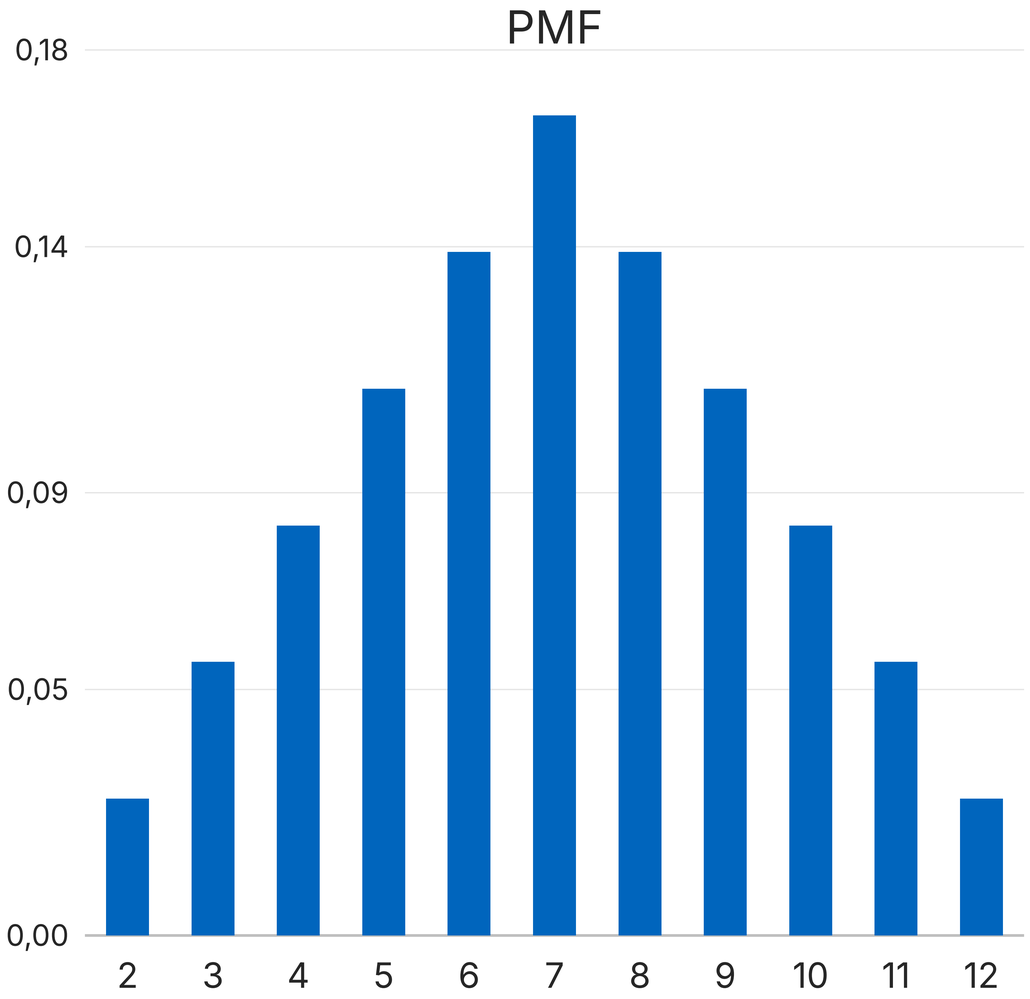
- The **probability mass function** of a random variable is a function $p_X : \Omega \to \mathbb{R}$ defined as

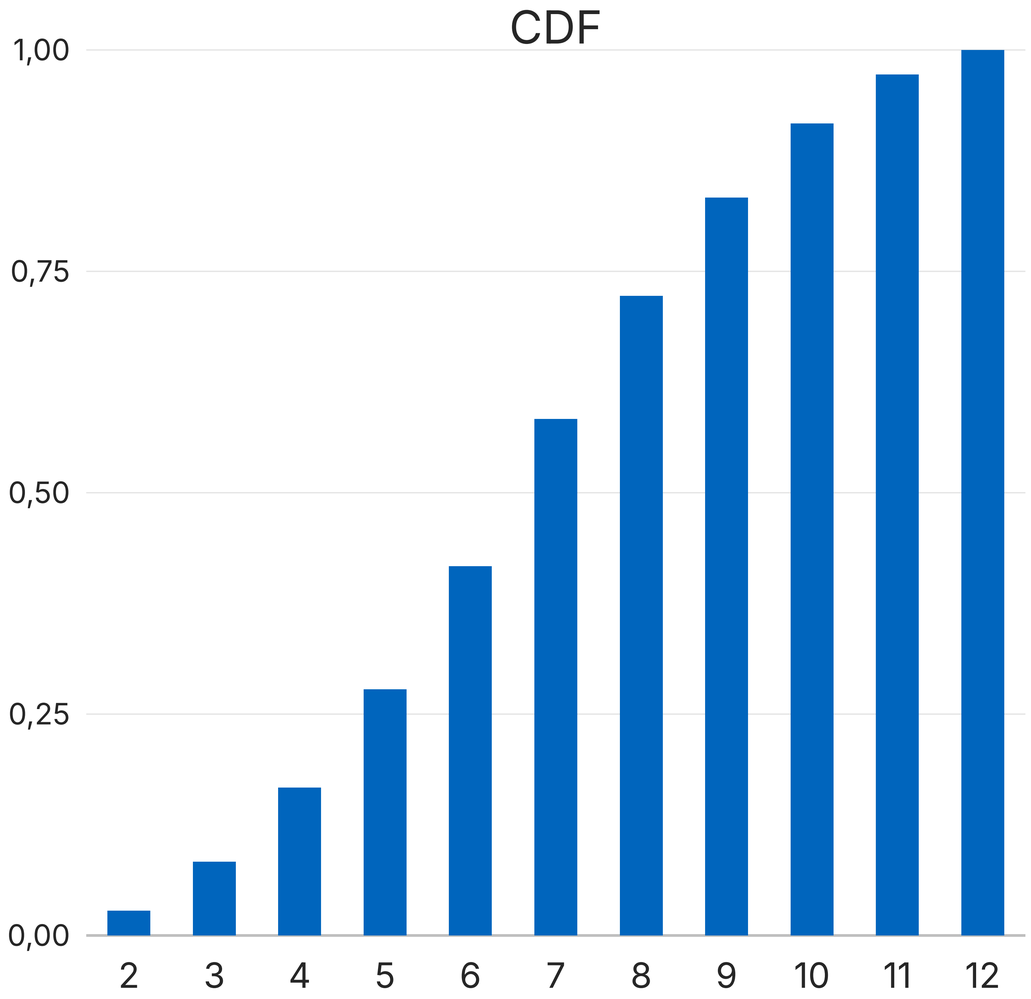
$$p_X(x) = \mathbb{P}(X = x).$$

- **Properties: Again, we can derive some properties:**
  - $0 \leq p_X(x) \leq 1$
  - $\sum_{x \in \Omega} p_X(x) = 1$



Discrete Example: Sum of 2 Dice Rolls





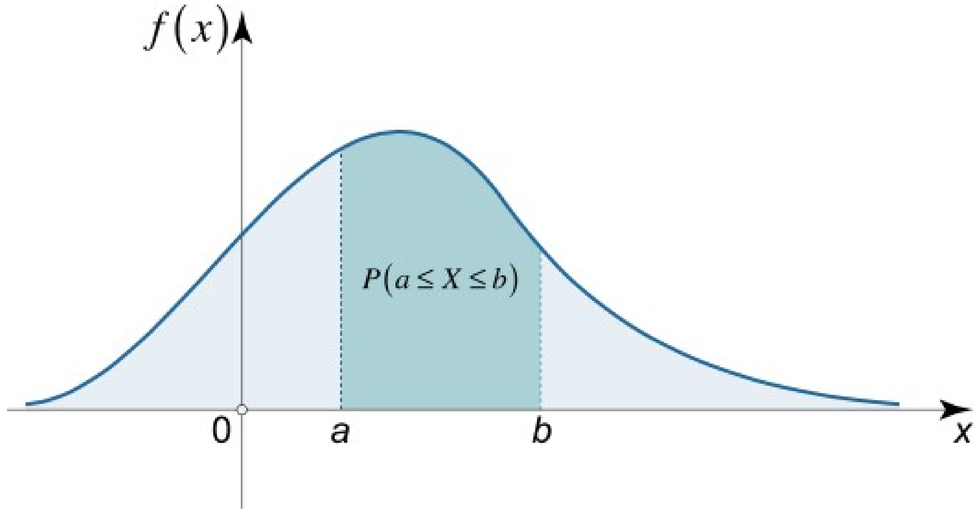
Continuous case: Probability Density Function

- **Continuous case:** For some continuous random variables, the CDF $F_X(x)$ is differentiable everywhere. Then we define the probability density function as the function $f_X(x) : \Omega \to \mathbb{R}$ with

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

- **Properties:**
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
  - $\int_a^b f_X(x)\, dx = F_X(b) - F_X(a)$

*Note: the value of a PDF at any given point x is **not** the probability of that event!*



Expectation of a random variable

- **Idea:** "weighted average" of the values that the random variable can take on
- **Discrete setting:** Assume that X is a discrete random variable with PMF $p_X(x)$. Then the expectation of X is given by

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

- **Continuous setting:** Assume that X is a continuous random variable with PDF $f_X(x)$. Then the expectation of X is given by

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

Expectation: Example

- **Discrete setting:** Assume that X is a discrete random variable with PMF $p_X(x)$. Then the expectation of X is given by $\mathbb{E}[X] = \sum_{x \in \Omega} x \cdot p_X(x)$. *(recap)*
- **Example: Tossing a six-sided die**
- $\Omega = \{1, 2, 3, 4, 5, 6\}$, X: represents the outcome of the toss
- $p_X(x) = \mathbb{P}(X = x) = \frac{1}{6} \; \forall x \in \Omega$
- $\mathbb{E}[X] = \sum_{x \in \Omega} x \cdot p_X(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$



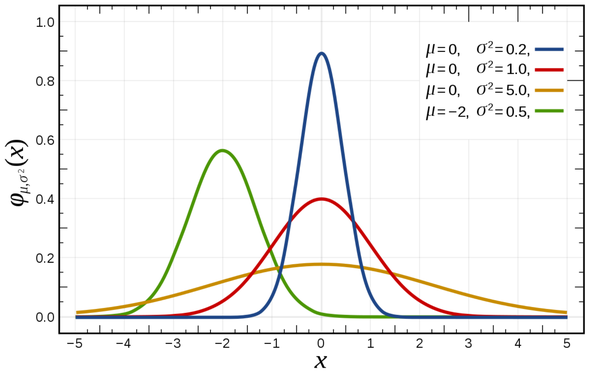
Expectation of a random variable

- **Properties: We encounter several important properties for the expectation, i.e.**
  - $\mathbb{E}[a] = a$ for any constant $a \in \mathbb{R}$
  - **Linearity:** $\mathbb{E}[aX + bY] = a \cdot \mathbb{E}[X] + b \cdot \mathbb{E}[Y]$ for any constants $a, b \in \mathbb{R}$

Variance of a random variable

- **Idea:** The variance of a random variable is a measure how concentrated the distribution of a random variable is around its mean.
- **Definition:** The variance is defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$



Variance of a random variable

- **Definition:** The variance is defined as $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

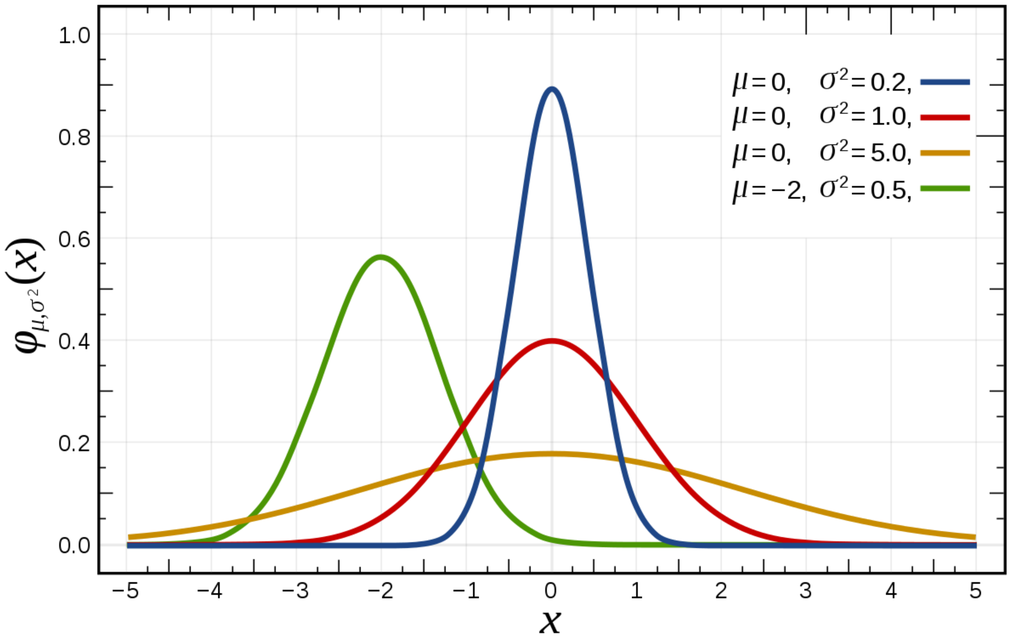
(recap)

- **Example: Tossing a fair six-sided die**
- $\Omega = \{1, 2, 3, 4, 5, 6\}$, X: represents the outcome of the toss
- $p_X(x) = \mathbb{P}(X = x) = \frac{1}{6} \; \forall x \in \Omega, \quad \mathbb{E}[X] = 3.5, \; \mathbb{E}[X]^2 = 12\frac{1}{4}$
- $\mathbb{E}[X^2] = \sum_{x \in \Omega} x^2 \cdot p_X(x) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \ldots + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = 15\frac{1}{6}$
- $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 15\frac{1}{6} - 12\frac{1}{4} = \frac{35}{12} \approx 2.91$

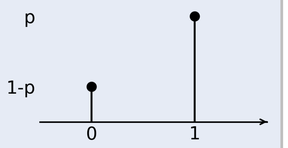
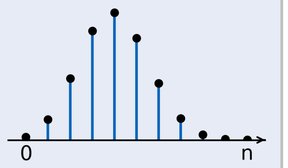
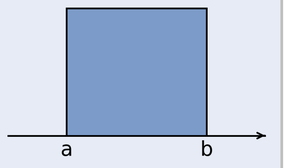
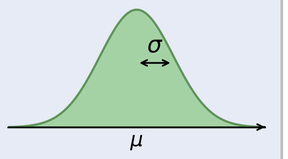


Variance of a random variable

- **Properties: The variance has the following properties, i.e.**
  - $\mathrm{Var}(a) = 0$ for any constant $a \in \mathbb{R}$
  - $\mathrm{Var}(a \cdot X + b) = a^2 \cdot \mathrm{Var}(X)$



Standard Probability Distributions

Distribution	Notation & parameters	PDF or PMF	Mean	Variance	Illustration
Bernoulli *(Discrete)*	$X \sim \text{Ber}(p)$, $0 \leq p \leq 1$	$p_X(k) = p^k(1-p)^{1-k}$	p	$p(1-p)$	
Binomial *(Discrete)*	$X \sim \text{Bin}(n, p)$, $n \in \mathbb{N},\ p \in [0,1]$	$p_X(k) = \binom{n}{k} p^k(1-p)^{n-k}$	$n \cdot p$	$np(1-p)$	
Uniform *(Continuous)*	$X \sim U(a, b)$, $-\infty < a < b < \infty$	$f_X(x) = \frac{1}{b-a}$ if $x \in [a, b]$, else 0	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	
Normal *(Continuous)*	$X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R},\ \sigma^2 \geq 0$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2	

References

- https://cs229.stanford.edu/section/cs229-prob.pdf
  - Comprehensive Probability Review — **recommended!**
- https://stanford.edu/~shervine/teaching/cme-106/cheatsheet-probability
  - Quick Overview
- https://www.deeplearningbook.org/contents/prob.html
  - Another great resource. Also covers information theory basics.

Thank you!