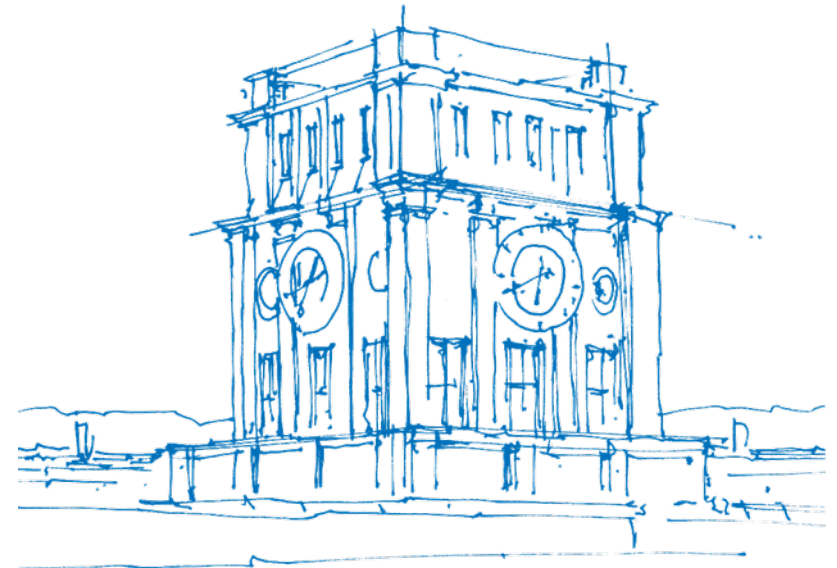




# A Crash Course on (Discrete) Probability Theory

Tao Wu, Yuesong Shen, Zhenzhang Ye

Computer Vision & Artificial Intelligence  
Technical University of Munich



*TUM Uhrenturm*

# Probability Space

A **probability space** is a triplet  $(\Omega, \mathcal{F}, P)$  consisting of:

- a **sample space**  $\Omega$  — an arbitrary non-empty set.
- a  **$\sigma$ -algebra** — a set of subsets (*events*) of  $\Omega$  s.t.
  - $\Omega \in \mathcal{F}$ .
  - $\mathcal{F}$  is closed under complements:

$$A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}.$$

- $\mathcal{F}$  is closed under countable unions:

$$(A_i)_{i \in \mathbb{N}} \in \mathcal{F} \Rightarrow \bigcup_{i \in \mathbb{N}} A_i \in \mathcal{F}.$$

- a **probability measure**  $P : \mathcal{F} \rightarrow [0, 1]$  s.t.
  - $P$  is  $\sigma$ -additive: if  $(A_i)_{i \in \mathbb{N}} \in \Omega$  is a countable collection of *pairwise disjoint* sets, then

$$P\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} P(A_i).$$

- $P(\Omega) = 1$ .



# Immediate Properties (as exercises)

- $\emptyset \in \mathcal{F}$  and  $P(\emptyset) = 0$ .
- $A \subset B \Rightarrow P(A) \leq P(B)$ .
- $P(A \cap B) \leq \min(P(A), P(B))$ .
- $P(A \cup B) \leq P(A) + P(B)$ .
- $P(\Omega \setminus A) = 1 - P(A)$ .
- If  $(A_i)_{i \in \mathbb{N}} \in \mathcal{F}$  is set of pairwise disjoint events s.t.  $\bigcup_{i \in \mathbb{N}} A_i = \Omega$ , then  $\sum_{i \in \mathbb{N}} P(A_i) = 1$ .



# Conditional Probability

Assume  $P(B) \neq 0$ . The **conditional probability** of event  $A$  given  $B$  is defined as

$$P(A|B) := \frac{P(A \cap B)}{P(B)}. \quad (\dagger)$$

- Chain rule:

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) \dots P(A_k|A_{k-1} \cap \dots \cap A_1).$$

Proof: Recursively apply  $(\dagger)$ .

- Independence: Two events  $A$  and  $B$  are **independent** ( $A \perp B$ ) iff

$$P(A \cap B) = P(A)P(B) \Leftrightarrow P(A|B) = P(A).$$

# Discrete vs. Continuous Probability Space

- Discrete case (sufficient for this course):
  - $\Omega$  contains at most countably many elements.
  - Typically, we set  $\mathcal{F} = 2^\Omega$  (the power set of  $\Omega$ ).
  - $P$  is characterized by **probability mass function**  $p : \Omega \rightarrow [0, 1]$  s.t.  $\sum_{\omega \in \Omega} p(\omega) = 1$ . This means:

$$\forall A \in \mathcal{F} : P(A) = \sum_{\omega \in A} p(\omega).$$

- Continuous case:
  - $\Omega$  is a (possibly uncountable) *measurable space*.
  - $P$  is a *probability measure*.
  - $P$  often admits a **probability density function**  $p : \Omega \rightarrow [0, 1]$  ( $\Omega \subset \mathbb{R}^n$ ), i.e.,

$$\forall A \in \mathcal{F} : P(A) = \int_A p(\omega) d\omega.$$

Formally,  $p = \frac{dP}{d\omega}$  is the *Radon-Nikodym derivative* of  $P$  w.r.t. the Lebesgue measure.

# Random Variable

A **random variable**  $X$  on  $(\Omega, \mathcal{F}, P)$  is a *measurable function*  $X : \Omega \rightarrow \mathcal{X}$  s.t.

$$\forall A \subset \mathcal{X} \text{ measurable} : P(X \in A) := P(X^{-1}(A)) = P(\{\omega \in \Omega : X(\omega) \in A\}).$$

- **Expectation:**

$$\begin{aligned} \mathbb{E}[X] &:= \int_{\Omega} X(\omega) dP(\omega) && \text{(continuous)} \\ &= \sum_{\omega \in \Omega} X(\omega) p(\omega). && \text{(discrete)} \end{aligned}$$

- **Covariance** ( $X, Y$  are both real-valued RV):

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

- **Variance:**  $\text{Var}[X] := \text{Cov}[X, X]$ .

For discrete RV, it's convenient to directly work on the (discrete) output space  $\mathcal{X}$  :

$$p(x) := P(X = x) \quad (\text{so that } \mathbb{E}[X] = \sum_{x \in \mathcal{X}} xp(x), \text{ etc.})$$



# Simple Properties (as exercises)

Expectation:

- $\mathbb{E}[a] = a$  for any constant  $a$ .
- Linearity:

$$\begin{aligned}\mathbb{E}[X + Y] &= \mathbb{E}[X] + \mathbb{E}[Y], \\ \mathbb{E}[\alpha X] &= \alpha \mathbb{E}[X]. \quad (\alpha \text{ is a constant scalar})\end{aligned}$$

Covariance:

- $\text{Cov}[X, a] = \text{Cov}[a, X] = 0$  for any constant  $a$ .
- Homogeneity:

$$\text{Cov}[\alpha X, \beta Y] = \alpha \beta \text{Cov}[X, Y]. \quad (\alpha, \beta \text{ are constant scalars})$$

- Covariance under independence  $X \perp Y$ :

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \quad \Leftrightarrow \quad \text{Cov}[X, Y] = 0.$$

# Joint, Marginal, and Conditional Probability

Let two discrete random variables  $X, Y$  be given.

**Joint probability:**

$$p(x, y) = P(X = x, Y = y).$$

**Marginal probability:**

$$p(x) = \sum_y p(x, y).$$

**Conditional probability:**

$$p(x|y) = P(X = x|Y = y) = \frac{p(x, y)}{p(y)} = \frac{p(x, y)}{\sum_x p(x, y)}.$$

**Independence** between  $X, Y$  (denoted by  $X \perp Y$ ) iff

$$p(x, y) = p(x)p(y) \Leftrightarrow p(x|y) = p(x).$$

**Bayes' rule:**

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$



# Example on binary segmentation

Consider a simplified binary segmentation problem, where  $X \in \{\text{Dark, Bright}\}$ ,  $Y \in \{\text{Foreground, Background}\}$ .

$p(x, y)$	Dark	Bright	$p(y)$
Foreground	0.163	0.006	0.169
Background	0.116	0.715	0.831
$p(x)$	0.279	0.721	

Figure: Probability Table for Binary Segmentation.

- Conditional probability via joint probability:

$$P(Y = \text{Foregr.} | X = \text{Dark}) = \frac{P(X = \text{Dark}, Y = \text{Foregr.})}{P(X = \text{Dark})} = \frac{0.163}{0.279} = 0.584.$$

- Conditional probability via Bayes' rule:

$$\begin{aligned} P(X = \text{Dark} | Y = \text{Foregr.}) &= \frac{P(Y = \text{Foregr.} | X = \text{Dark})P(X = \text{Dark})}{P(Y = \text{Foregr.})} \\ &= \frac{0.584 \cdot 0.279}{0.169} = 0.964. \end{aligned}$$