Probability Theory Refresher

Mário A. T. Figueiredo

Instituto Superior Técnico & Instituto de Telecomunicações

Lisboa, Portugal

LxMLS 2016: Lisbon Machine Learning School

July 21, 2016
Probability theory

The study of probability has roots in games of chance. Great names of science include Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzmann, de Finetti, and others. Probability theory is a natural tool to model uncertainty, information, knowledge, belief, observations, and thus also learning, decision making, inference, science, and so on.

Mário A. T. Figueiredo (IST & IT)
The study of probability has roots in games of chance
Probability theory

The study of probability has roots in games of chance.

Great names of science: Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzman, de Finetti, ...
Probability theory

- The study of probability has roots in games of chance
- Great names of science: Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzman, de Finetti, ...
- Natural tool to model uncertainty, information, knowledge, belief, observations, ...
The study of probability has roots in games of chance

Great names of science: Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzman, de Finetti, ...

Natural tool to model uncertainty, information, knowledge, belief, observations, ...

...thus also learning, decision making, inference, science,...
What is probability?

- **Classical definition**: \( \mathbb{P}(A) = \frac{N_A}{N} \)

  ...with \( N \) mutually exclusive equally likely outcomes, \( N_A \) of which result in the occurrence of \( A \).

  *Example*: \( \mathbb{P}(\text{randomly drawn card is } \spadesuit) = \frac{13}{52} \).

  *Example*: \( \mathbb{P}(\text{getting 1 in throwing a fair die}) = \frac{1}{6} \).
What is probability?

- **Classical definition:** $\mathbb{P}(A) = \frac{N_A}{N}$

  ...with $N$ mutually exclusive equally likely outcomes, $N_A$ of which result in the occurrence of $A$.  

  **Example:** $\mathbb{P}($randomly drawn card is ♣$) = \frac{13}{52}$.

  **Example:** $\mathbb{P}($getting 1 in throwing a fair die$) = \frac{1}{6}$.

- **Frequentist definition:** $\mathbb{P}(A) = \lim_{N \to \infty} \frac{N_A}{N}$

  ...relative frequency of occurrence of $A$ in infinite number of trials.
What is probability?

- **Classical definition:** \( \mathbb{P}(A) = \frac{N_A}{N} \)

...with \( N \) mutually exclusive equally likely outcomes, \( N_A \) of which result in the occurrence of \( A \).  

*Laplace, 1814*

**Example:** \( \mathbb{P}(\text{randomly drawn card is } \heartsuit) = \frac{13}{52} \).

**Example:** \( \mathbb{P}(\text{getting 1 in throwing a fair die}) = \frac{1}{6} \).

- **Frequentist definition:** \( \mathbb{P}(A) = \lim_{N \to \infty} \frac{N_A}{N} \)

...relative frequency of occurrence of \( A \) in infinite number of trials.

- **Subjective probability:** \( \mathbb{P}(A) \) is a degree of belief.  

*de Finetti, 1930s*

...gives meaning to \( \mathbb{P}(\text{“it will rain tomorrow”}) \).
Key concepts: Sample space and events

- **Sample space** \( \mathcal{X} \) = set of possible outcomes of a random experiment.

Examples:

- Tossing two coins: \( \mathcal{X} = \{HH, TH, HT, TT\} \)
- Roulette: \( \mathcal{X} = \{1, 2, ..., 36\} \)
- Draw a card from a shuffled deck: \( \mathcal{X} = \{A\spadesuit, 2\spadesuit, ..., Q\diamond, K\diamond\} \).
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} =$ set of possible outcomes of a random experiment.

Examples:
- Tossing two coins: $\mathcal{X} = \{HH, TH, HT, TT\}$
- Roulette: $\mathcal{X} = \{1, 2, ..., 36\}$
- Draw a card from a shuffled deck: $\mathcal{X} = \{A\spadesuit, 2\spadesuit, ..., Q\diamond, K\diamond\}$.

- **An event** $A$ is a subset of $\mathcal{X}$: $A \subseteq \mathcal{X}$ (also written $A \in 2^{\mathcal{X}}$).

Examples:
- “exactly one H in 2-coin toss”: $A = \{TH, HT\} \subset \{HH, TH, HT, TT\}$.
- “odd number in the roulette”: $B = \{1, 3, ..., 35\} \subset \{1, 2, ..., 36\}$.
- “drawn a ♥ card”: $C = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\} \subset \{A\spadesuit, ..., K\diamond\}$.
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} =$ set of possible outcomes of a random experiment.

  (More delicate) examples:
  - Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$
Key concepts: Sample space and events

- **Sample space** \( \mathcal{X} \) = set of possible outcomes of a random experiment.

  (More delicate) examples:
  - Distance travelled by tossed die: \( \mathcal{X} = \mathbb{R}_+ \)
  - Location of the next rain drop on a given square tile: \( \mathcal{X} = \mathbb{R}^2 \)

- Properly handling the continuous case requires deeper concepts:
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} =$ set of possible outcomes of a random experiment.
  
  *(More delicate) examples:*
  
  - Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$

- Properly handling the continuous case requires deeper concepts:
  
  - Let $\Sigma$ be collection of subsets of $\mathcal{X}$: $\Sigma \subseteq 2^\mathcal{X}$
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} =$ set of possible outcomes of a random experiment.

(More delicate) examples:
- Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
- Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$

Properly handling the continuous case requires deeper concepts:
- Let $\Sigma$ be collection of subsets of $\mathcal{X}$: $\Sigma \subseteq 2^\mathcal{X}$
- $\Sigma$ is a $\sigma-$algebra if
  - $A \in \Sigma \implies A^c \in \Sigma$
  - $A_1, A_2, \ldots \in \Sigma \implies \bigcup_{i=1}^{\infty} A_i \in \Sigma$
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} = \text{set of possible outcomes of a random experiment.}

  (More delicate) examples:
  - Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$

- **Properly handling the continuous case requires deeper concepts:**
  - Let $\Sigma$ be collection of subsets of $\mathcal{X}$: $\Sigma \subseteq 2^{\mathcal{X}}$
  - $\Sigma$ is a $\sigma$–algebra if
    1. $A \in \Sigma \implies A^c \in \Sigma$
    2. $A_1, A_2, \ldots \in \Sigma \implies \bigcup_{i=1}^{\infty} A_i \in \Sigma$
  - Corollary: if $\Sigma \subseteq 2^{\mathcal{X}}$ is a $\sigma$-algebra, $\emptyset \in \Sigma$ and $\mathcal{X} \in \Sigma$
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} = \text{set of possible outcomes of a random experiment.}

(More delicate) examples:
  - Distance travelled by tossed die: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop on a given square tile: $\mathcal{X} = \mathbb{R}^2$

- Properly handling the continuous case requires deeper concepts:
  - Let $\Sigma$ be collection of subsets of $\mathcal{X}$: $\Sigma \subseteq 2^{\mathcal{X}}$
  - $\Sigma$ is a $\sigma$–algebra if
    - $A \in \Sigma \Rightarrow A^c \in \Sigma$
    - $A_1, A_2, \ldots \in \Sigma \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \Sigma$
  - Corollary: if $\Sigma \subseteq 2^{\mathcal{X}}$ is a $\sigma$-algebra, $\emptyset \in \Sigma$ and $\mathcal{X} \in \Sigma$
  - Example in $\mathbb{R}^n$: collection of Lebesgue-measurable sets is a $\sigma$–algebra.
Kolmogorov’s Axioms for Probability

- Probability is a function that maps events $A$ into the interval $[0, 1]$.

  Kolmogorov’s axioms (1933) for probability $\mathbb{P} : \Sigma \rightarrow [0, 1]$. 

  - $\mathbb{P}(\emptyset) = 0$
  - $C \subset D \implies \mathbb{P}(C) \leq \mathbb{P}(D)$
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (union bound)
Kolmogorov’s Axioms for Probability

- Probability is a function that maps events $A$ into the interval $[0, 1]$.

Kolmogorov’s axioms (1933) for probability $\mathbb{P} : \Sigma \rightarrow [0, 1]$

- For any $A$, $\mathbb{P}(A) \geq 0$
Kolmogorov’s Axioms for Probability

- Probability is a function that maps events $A$ into the interval $[0, 1]$. 

Kolmogorov’s axioms (1933) for probability $\mathbb{P} : \Sigma \rightarrow [0, 1]$

- For any $A$, $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\mathcal{X}) = 1$
Kolmogorov’s Axioms for Probability

Probability is a function that maps events $A$ into the interval $[0, 1]$.  

**Kolmogorov’s axioms (1933) for probability** $\mathbb{P} : \Sigma \rightarrow [0, 1]$

- For any $A$, $\mathbb{P}(A) \geq 0$
- $\mathbb{P}(\mathcal{X}) = 1$
- If $A_1, A_2 \ldots \subseteq \mathcal{X}$ are disjoint events, then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$
Kolmogorov’s Axioms for Probability

- Probability is a function that maps events \( A \) into the interval \([0, 1]\).

**Kolmogorov’s axioms** (1933) for probability \( \mathbb{P} : \Sigma \rightarrow [0, 1] \)

- For any \( A \), \( \mathbb{P}(A) \geq 0 \)
- \( \mathbb{P}(\mathcal{X}) = 1 \)
- If \( A_1, A_2 \ldots \subseteq \mathcal{X} \) are disjoint events, then \( \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i) \)

- From these axioms, many results can be derived. **Examples:**

  - \( \mathbb{P}(\emptyset) = 0 \)
  - \( C \subseteq D \implies \mathbb{P}(C) \leq \mathbb{P}(D) \)
  - \( \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \)
  - \( \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B) \) (union bound)
Conditional Probability and Independence

- If $P(B) > 0$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (conditional prob. of $A$, given $B$)
Conditional Probability and Independence

- If $P(B) > 0$, $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (conditional prob. of $A$, given $B$)

- ...satisfies all of Kolmogorov’s axioms:
  - For any $A \subseteq \mathcal{X}$, $P(A|B) \geq 0$
  - $P(\mathcal{X}|B) = 1$
  - If $A_1, A_2, \ldots \subseteq \mathcal{X}$ are disjoint, then
    $P\left(\bigcup_{i} A_i \bigg| B\right) = \sum_{i} P(A_i | B)$
Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (conditional prob. of $A$, given $B$)

- ...satisfies all of Kolmogorov’s axioms:
  - For any $A \subseteq \mathcal{X}$, $\mathbb{P}(A|B) \geq 0$
  - $\mathbb{P}(\mathcal{X}|B) = 1$
  - If $A_1$, $A_2$, ..., $\subseteq \mathcal{X}$ are disjoint, then $\mathbb{P}\left(\bigcup_i A_i \bigg| B\right) = \sum_i \mathbb{P}(A_i|B)$

- Independence: $A$, $B$ are independent (denoted $A \perp \perp B$) if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$. 

\[
\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B).
\]
Conditional Probability and Independence

- If $\Pr(B) > 0$, $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$
Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

- Events $A, B$ are independent ($A \perp \perp B$) $\iff$ $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$. 

Example: $X = \{\text{52 cards}\}$, $A = \{\text{3♥}, 3♣, 3♦, 3♣\}$, and $B = \{\text{A♥, 2♥, ..., K♥}\}$; then, $\mathbb{P}(A) = \frac{1}{13}, \mathbb{P}(B) = \frac{1}{4}$, $\mathbb{P}(A \cap B) = \mathbb{P}(\{3♥\}) = \frac{1}{52}$.
Conditional Probability and Independence

- If \( P(B) > 0 \), \( P(A|B) = \frac{P(A \cap B)}{P(B)} \)

- Events \( A, B \) are independent \( (A \perp \perp B) \iff P(A \cap B) = P(A) P(B) \).

- Relationship with conditional probabilities:

\[
A \perp \perp B \iff P(A|B) = P(A)
\]
Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, \[ \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \]

- Events $A$, $B$ are independent ($A \perp \perp B$) $\iff$ $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$.

- Relationship with conditional probabilities:
  \[ A \perp \perp B \iff \mathbb{P}(A|B) = \mathbb{P}(A) \]

- Example: $\mathcal{X} = "$52 cards"$, $A = \{3\heartsuit, 3\diamondsuit, 3\diamond, 3\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, \ldots, K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

\[ \mathbb{P}(A \cap B) = \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52} \]
Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, \[ \mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \]

- Events $A$, $B$ are independent ($A \perp \perp B$) $\iff$ $\mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$.

- Relationship with conditional probabilities:

\[ A \perp \perp B \iff \mathbb{P}(A|B) = \mathbb{P}(A) \]

- Example: $\mathcal{X} =$ “52 cards”, $A = \{3\heartsuit, 3\diamondsuit, 3\spadesuit, 3\clubsuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

\[ \begin{align*}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52} \\
\mathbb{P}(A) \mathbb{P}(B) &= \frac{1}{13} \cdot \frac{1}{4} = \frac{1}{52}
\end{align*} \]
Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

- Events $A$, $B$ are independent ($A \perp \perp B$) $\iff \mathbb{P}(A \cap B) = \mathbb{P}(A) \mathbb{P}(B)$.

- Relationship with conditional probabilities:

\[
A \perp \perp B \iff \mathbb{P}(A|B) = \mathbb{P}(A)
\]

Example: $\mathcal{X} =$ “52 cards”, $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, \ldots, K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

\[
\begin{align*}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52} \\
\mathbb{P}(A) \mathbb{P}(B) &= \frac{1}{13} \cdot \frac{1}{4} = \frac{1}{52} \\
\mathbb{P}(A|B) &= \mathbb{P}(“3”|“\heartsuit”) = \frac{1}{13} = \mathbb{P}(A)
\end{align*}
\]
Bayes Theorem

- Law of total probability: if $A_1, \ldots, A_n$ are a partition of $\mathcal{X}$

\[
\mathbb{P}(B) = \sum_i \mathbb{P}(B | A_i) \mathbb{P}(A_i)
= \sum_i \mathbb{P}(B \cap A_i)
\]
Bayes Theorem

- Law of total probability: if $A_1, \ldots, A_n$ are a partition of $\mathcal{X}$

$$
P(B) = \sum_i P(B|A_i)P(A_i)$$
$$= \sum_i P(B \cap A_i)$$

- Bayes’ theorem: if $\{A_1, \ldots, A_n\}$ is a partition of $\mathcal{X}$

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_i P(B|A_i)P(A_i)}$$
Random Variables

- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$

- Discrete RV: range of $X$ is countable (e.g., $\mathbb{N}$ or $\{0, 1\}$)

- Continuous RV: range of $X$ is uncountable (e.g., $\mathbb{R}$ or $[0, 1]$)

- Example: number of heads in tossing two coins, $X = \{HH, HT, TH, TT\}$, $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.

- Range of $X = \{0, 1, 2\}$.

- Example: distance traveled by a tossed coin; range of $X = \mathbb{R}^+$. 
Random Variables

- A (real) **random variable** (RV) is a function: \( X : \mathcal{X} \rightarrow \mathbb{R} \)

  ▶ **Discrete RV**: range of \( X \) is countable (e.g., \( \mathbb{N} \) or \{0, 1\})

  ▶ **Continuous RV**: range of \( X \) is uncountable (e.g., \( \mathbb{R} \) or \([0, 1]\))

- Example: number of heads in tossing two coins, \( X = \{HH, HT, TH, TT\} \), where:
  - \( X(HH) = 2 \)
  - \( X(HT) = X(TH) = 1 \)
  - \( X(TT) = 0 \)

  Range of \( X \) = \{0, 1, 2\}

- Example: distance traveled by a tossed coin; range of \( X \) = \( \mathbb{R}_+ \).
Random Variables

- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$

  - **Discrete RV**: range of $X$ is countable (e.g., $\mathbb{N}$ or $\{0, 1\}$)
  - **Continuous RV**: range of $X$ is uncountable (e.g., $\mathbb{R}$ or $[0, 1]$)

Example: number of heads in tossing two coins, $X = \{HH, HT, TH, TT\}$, $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.

Range of $X = \{0, 1, 2\}$.

Example: distance traveled by a tossed coin; range of $X = \mathbb{R}^+$. 
Random Variables

- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$

- **Discrete RV**: range of $X$ is countable (e.g., $\mathbb{N}$ or $\{0, 1\}$)

- **Continuous RV**: range of $X$ is uncountable (e.g., $\mathbb{R}$ or $[0, 1]$)

- **Example**: number of head in tossing two coins,
  $\mathcal{X} = \{HH, HT, TH, TT\}$,
  $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.
  Range of $X = \{0, 1, 2\}$. 
Random Variables

- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$

  - Discrete RV: range of $X$ is countable (e.g., $\mathbb{N}$ or $\{0, 1\}$)
  - Continuous RV: range of $X$ is uncountable (e.g., $\mathbb{R}$ or $[0, 1]$)
  - Example: number of head in tossing two coins,
    $\mathcal{X} = \{HH, HT, TH, TT\}$,
    $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.
    Range of $X = \{0, 1, 2\}$.
  - Example: distance traveled by a tossed coin; range of $X = \mathbb{R}_+$. 
Random Variables: Distribution Function

- **Distribution function**: \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

\[
\begin{align*}
\mathcal{X} & \quad \{\omega : X(\omega) \leq x\} \\
\mathbb{R} & \quad x \quad X(\omega) \leq x
\end{align*}
\]
Random Variables: Distribution Function

- **Distribution function:** \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

- **Example:** number of heads in tossing 2 coins; \( \text{range}(X) = \{0, 1, 2\} \).
Random Variables: Distribution Function

- **Distribution function:** \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

- **Example:** number of heads in tossing 2 coins; range(\(X\)) = \{0, 1, 2\).

- **Probability mass function (discrete RV):** \( f_X(x) = \mathbb{P}(X = x) \),
  \[ F_X(x) = \sum_{x_i \leq x} f_X(x_i). \]
Properties of Distribution Functions

\[ F_X : \mathbb{R} \to [0, 1] \] is the distribution function of some r.v. \( X \) iff:

1. It is non-decreasing:
   \[ x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \]
2. \( \lim_{x \to -\infty} F_X(x) = 0 \)
3. \( \lim_{x \to +\infty} F_X(x) = 1 \)
4. It is right semi-continuous:
   \[ \lim_{x \to z^+} F_X(x) = F_X(z) \]

Further properties:

\[ P(X = x) = f_X(x) = F_X(x) - \lim_{z \to x^-} F_X(z) \]
\[ P(z < X \leq y) = F_X(y) - F_X(z) \]
\[ P(X > x) = 1 - F_X(x) \]
Properties of Distribution Functions

\( F_X : \mathbb{R} \rightarrow [0, 1] \) is the distribution function of some r.v. \( X \) iff:

- it is \textbf{non-decreasing}: \( x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \);
Properties of Distribution Functions

$F_X : \mathbb{R} \to [0, 1]$ is the distribution function of some r.v. $X$ iff:

- it is non-decreasing: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- $\lim_{x \to -\infty} F_X(x) = 0$;
Properties of Distribution Functions

$F_X : \mathbb{R} \rightarrow [0, 1]$ is the distribution function of some r.v. $X$ iff:

- it is non-decreasing: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- $\lim_{x \to -\infty} F_X(x) = 0$;
- $\lim_{x \to +\infty} F_X(x) = 1$;
Properties of Distribution Functions

\( F_X : \mathbb{R} \to [0, 1] \) is the distribution function of some r.v. \( X \) iff:

- it is non-decreasing: \( x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \);
- \( \lim_{x \to -\infty} F_X(x) = 0 \);
- \( \lim_{x \to +\infty} F_X(x) = 1 \);
- it is right semi-continuous: \( \lim_{x \to z^+} F_X(x) = F_X(z) \)
Properties of Distribution Functions

\( F_X : \mathbb{R} \to [0, 1] \) is the distribution function of some r.v. \( X \) iff:

- it is non-decreasing: \( x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \);

- \( \lim_{x \to -\infty} F_X(x) = 0 \);

- \( \lim_{x \to +\infty} F_X(x) = 1 \);

- it is right semi-continuous: \( \lim_{x \to z^+} F_X(x) = F_X(z) \)

Further properties:
Properties of Distribution Functions

\( F_X : \mathbb{R} \to [0, 1] \) is the distribution function of some r.v. \( X \) iff:

- it is non-decreasing: \( x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2) \);
- \( \lim_{x \to -\infty} F_X(x) = 0 \);
- \( \lim_{x \to +\infty} F_X(x) = 1 \);
- it is right semi-continuous: \( \lim_{x \to z^+} F_X(x) = F_X(z) \)

Further properties:

- \( \mathbb{P}(X = x) = f_X(x) = F_X(x) - \lim_{z \to x^-} F_X(z) \);
Properties of Distribution Functions

\( F_X : \mathbb{R} \to [0, 1] \) is the distribution function of some r.v. \( X \) iff:

- it is non-decreasing: \( x_1 < x_2 \implies F_X(x_1) \leq F_X(x_2) \);

- \( \lim_{x \to -\infty} F_X(x) = 0 \);

- \( \lim_{x \to +\infty} F_X(x) = 1 \);

- it is right semi-continuous: \( \lim_{x \to z^+} F_X(x) = F_X(z) \)

Further properties:

- \( \mathbb{P}(X = x) = f_X(x) = F_X(x) - \lim_{z \to x^-} F_X(z) \);

- \( \mathbb{P}(z < X \leq y) = F_X(y) - F_X(z) \);
Properties of Distribution Functions

$F_X : \mathbb{R} \rightarrow [0, 1]$ is the distribution function of some r.v. $X$ iff:

- it is **non-decreasing**: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- $\lim_{x \to -\infty} F_X(x) = 0$;
- $\lim_{x \to +\infty} F_X(x) = 1$;
- it is **right semi-continuous**: $\lim_{x \to z^+} F_X(x) = F_X(z)$

Further properties:

- $\mathbb{P}(X = x) = f_X(x) = F_X(x) - \lim_{z \to x^-} F_X(z)$;
- $\mathbb{P}(z < X \leq y) = F_X(y) - F_X(z)$;
- $\mathbb{P}(X > x) = 1 - F_X(x)$. 
Important Discrete Random Variables

- **Uniform**: $X \in \{x_1, \ldots, x_K\}$, pmf $f_X(x_i) = 1/K$.  
  
  - Bernoulli RV: $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \end{cases}$  
    
    Can be written compactly as $f_X(x) = px(1-p)^{1-x}$.
  
  - Binomial RV: $X \in \{0, 1, \ldots, n\}$ (sum of $n$ Bernoulli RVs)  
    
    $f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$

Binomial coefficients ("$n$ choose $x$":)  
$\binom{n}{x} = \frac{n!}{x!(n-x)!}$
Important Discrete Random Variables

- **Uniform**: \( X \in \{x_1, \ldots, x_K\} \), pmf \( f_X(x_i) = 1/K \).

- **Bernoulli RV**: \( X \in \{0, 1\} \), pmf \( f_X(x) = \begin{cases} p & \iff x = 1 \\ 1 - p & \iff x = 0 \end{cases} \)

  Can be written compactly as \( f_X(x) = p^x (1 - p)^{1-x} \).
Important Discrete Random Variables

- **Uniform**: $X \in \{x_1, ..., x_K\}$, pmf $f_X(x_i) = 1/K$.

- **Bernoulli RV**: $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \iff x = 1 \\ 1 - p & \iff x = 0 \end{cases}$

  Can be written compactly as $f_X(x) = p^x(1-p)^{1-x}$.

- **Binomial RV**: $X \in \{0, 1, ..., n\}$ (sum of $n$ Bernoulli RVs)

  $f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$
Important Discrete Random Variables

- **Uniform**: $X \in \{x_1, ..., x_K\}$, pmf $f_X(x_i) = 1/K$.

- **Bernoulli RV**: $X \in \{0, 1\}$, pmf $f_X(x) = \begin{cases} p & \iff x = 1 \\ 1 - p & \iff x = 0 \end{cases}$

  Can be written compactly as $f_X(x) = p^x (1 - p)^{1-x}$.

- **Binomial RV**: $X \in \{0, 1, ..., n\}$ (sum of $n$ Bernoulli RVs)

  $f_X(x) = \text{Binomial}(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$

  **Binomial coefficients**
  ("$n$ choose $x$"):

  $\binom{n}{x} = \frac{n!}{(n - x)! x!}$
Other Important Discrete Random Variables

- **Geometric($p$):** $X \in \mathbb{N}$, pmf $f_X(x) = p(1 - p)^{x-1}$.
  
  (e.g., number of trials until the first success).
Other Important Discrete Random Variables

- **Geometric($p$):** $X \in \mathbb{N}$, pmf $f_X(x) = p(1 - p)^{x-1}$.
  (e.g., number of trials until the first success).

- **Poisson($\lambda$):** $X \in \mathbb{N} \cup \{0\}$, pmf $f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}$

Notice that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^\lambda$, thus $\sum_{x=0}^{\infty} f_X(x) = 1$.

“...probability of the number of independent occurrences in a fixed (time/space) interval if these occurrences have known average rate”
Random Variables: Distribution Function

- **Distribution function**: $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$

\[\mathcal{X}\]

\[\{\omega : X(\omega) \leq x\}\]

\[\mathbb{R}\]

\[X\]

\[X(\omega) \leq x\]

\[x\]
Random Variables: Distribution Function

- Distribution function: \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

Example: continuous RV with uniform distribution on \([a, b]\).

Probability density function (pdf, continuous RV):

\[
F_X(x) = \int_{-\infty}^{x} f_X(u) \, du,
\]

\[
P(X \in [c,d]) = \int_{c}^{d} f_X(x) \, dx,
\]

\[
P(X = x) = 0
\]
Random Variables: Distribution Function

- **Distribution function:** \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

- **Example:** continuous RV with uniform distribution on \([a, b]\).

- **Probability density function (pdf, continuous RV):** \( f_X(x) \)
Random Variables: Distribution Function

- **Distribution function**: $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$

- **Example**: continuous RV with uniform distribution on $[a, b]$.

- **Probability density function** (pdf, continuous RV): $f_X(x)$

  \[ F_X(x) = \int_{-\infty}^{x} f_X(u) \, du, \]
Random Variables: Distribution Function

- **Distribution function**: \( F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\}) \)

- **Example**: continuous RV with uniform distribution on \([a, b]\).

- **Probability density function (pdf, continuous RV)**: \( f_X(x) \)

\[
F_X(x) = \int_{-\infty}^{x} f_X(u) \, du, \quad \mathbb{P}(X \in [c, d]) = \int_{c}^{d} f_X(x) \, dx,
\]
Random Variables: Distribution Function

- **Distribution function:** $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$

- **Example:** continuous RV with uniform distribution on $[a, b]$.

![Diagram of distribution function and probability density function for a uniform distribution on [a, b].]

- **Probability density function (pdf, continuous RV):** $f_X(x)$

\[
F_X(x) = \int_{-\infty}^{x} f_X(u) \, du, \quad \mathbb{P}(X \in [c, d]) = \int_{c}^{d} f_X(x) \, dx, \quad \mathbb{P}(X = x) = 0
\]
Important Continuous Random Variables

- **Uniform**: \( f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \iff x \in [a, b] \\ 0 & \iff x \not\in [a, b] \end{cases} \) (previous slide).

- **Gaussian**: \( f_X(x) = \text{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

- **Exponential**: \( f_X(x) = \text{Exp}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \iff x \geq 0 \\ 0 & \iff x < 0 \end{cases} \)
Important Continuous Random Variables

- **Uniform**: 
  \[ f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \iff x \in [a, b] \\ 0 & \iff x \notin [a, b] \end{cases} \]

  (previous slide).

- **Gaussian**: 
  \[ f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Important Continuous Random Variables

- **Uniform:** \( f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \iff x \in [a, b] \\ 0 & \iff x \not\in [a, b] \end{cases} \) (previous slide).

- **Gaussian:** \( f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

- **Exponential:** \( f_X(x) = \text{Exp}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \iff x \geq 0 \\ 0 & \iff x < 0 \end{cases} \)
Expectation of Random Variables

- **Expectation:** 
  \[ E(X) = \begin{cases} 
  \sum_{i} x_i f_X(x_i) & X \in \{x_1, \ldots, x_K\} \subset \mathbb{R} \\
  \int_{-\infty}^{\infty} x f_X(x) \, dx & X \text{ continuous} 
  \end{cases} \]
Expectation of Random Variables

- **Expectation:** \( \mathbb{E}(X) = \begin{cases} \sum_{i} x_i f_X(x_i) & X \in \{x_1, \ldots, x_K\} \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x f_X(x) \, dx & X \text{ continuous} \end{cases} \)

- **Example:** Bernoulli, \( f_X(x) = p^x (1 - p)^{1-x} \), for \( x \in \{0, 1\} \).
  \[ \mathbb{E}(X) = 0 (1 - p) + 1 p = p. \]
Expectation of Random Variables

- **Expectation**: \( \mathbb{E}(X) = \begin{cases} \sum_{i} x_i f_X(x_i) & X \in \{x_1, \ldots, x_K\} \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x f_X(x) \, dx & X \text{ continuous} \end{cases} \)

- **Example**: Bernoulli, \( f_X(x) = p^x (1 - p)^{1-x}, \text{ for } x \in \{0, 1\}. \)
  \[ \mathbb{E}(X) = 0(1-p) + 1 p = p. \]

- **Example**: Binomial, \( f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \text{ for } x \in \{0, \ldots, n\}. \)
  \[ \mathbb{E}(X) = n p. \]
Expectation of Random Variables

- **Expectation**: \( \mathbb{E}(X) = \begin{cases} \sum_{i} x_i f_X(x_i) & X \in \{x_1, \ldots x_K\} \subset \mathbb{R} \\ \int_{-\infty}^{\infty} x f_X(x) \, dx & X \text{ continuous} \end{cases} \)

- **Example**: Bernoulli, \( f_X(x) = p^x (1 - p)^{1-x} \), for \( x \in \{0, 1\} \).
  \[ \mathbb{E}(X) = 0(1 - p) + 1p = p. \]

- **Example**: Binomial, \( f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} \), for \( x \in \{0, \ldots, n\} \).
  \[ \mathbb{E}(X) = np. \]

- **Example**: Gaussian, \( f_X(x) = \mathcal{N}(x; \mu, \sigma^2) \).
  \[ \mathbb{E}(X) = \mu. \]
Expectation of Random Variables

- **Expectation:** 
  \[ E(X) = \begin{cases} 
  \sum_{i} x_i f_X(x_i) & X \in \{x_1, \ldots x_K\} \subset \mathbb{R} \\
  \int_{-\infty}^{\infty} x f_X(x) \, dx & X \text{ continuous} 
  \end{cases} \]

- **Example:** Bernoulli, 
  \[ f_X(x) = p^x (1 - p)^{1-x}, \text{ for } x \in \{0, 1\}. \]
  \[ E(X) = 0(1 - p) + 1p = p. \]

- **Example:** Binomial, 
  \[ f_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, \text{ for } x \in \{0, \ldots, n\}. \]
  \[ E(X) = np. \]

- **Example:** Gaussian, 
  \[ f_X(x) = \mathcal{N}(x; \mu, \sigma^2). \]
  \[ E(X) = \mu. \]

- **Linearity of expectation:**
  \[ E(X + Y) = E(X) + E(Y); \quad E(\alpha X) = \alpha E(X), \quad \alpha \in \mathbb{R} \]
Expectation of Functions of Random Variables

\[ E(g(X)) = \begin{cases} 
\sum_i g(x_i)f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x)f_X(x)\,dx & X \text{ continuous}
\end{cases} \]
Expectation of Functions of Random Variables

\[ E(g(X)) = \begin{cases} 
\sum_{i} g(x_i) f_X(x_i) & \text{X discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x) f_X(x) \, dx & \text{X continuous}
\end{cases} \]

- **Example:** variance, \( \text{var}(X) = E\left( (X - E(X))^2 \right) \)
Expectation of Functions of Random Variables

\[ 
\mathbb{E}(g(X)) = \begin{cases} 
\sum_i g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x) f_X(x) \, dx & X \text{ continuous} 
\end{cases} 
\]

- **Example:** variance, \( \text{var}(X) = \mathbb{E} \left( (X - \mathbb{E}(X))^2 \right) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \)

- **Example:** Bernoulli variance, \( \mathbb{E}(X^2) = \mathbb{E}(X) = p \)
Expectation of Functions of Random Variables

\[ \mathbb{E}(g(X)) = \begin{cases} 
\sum_{i} g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x) f_X(x) \, dx & X \text{ continuous} 
\end{cases} \]

- **Example:** variance, \( \text{var}(X) = \mathbb{E}\left( (X - \mathbb{E}(X))^2 \right) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \)
- **Example:** Bernoulli variance, \( \mathbb{E}(X^2) = \mathbb{E}(X) = p \), thus \( \text{var}(X) = p(1 - p) \).
Expectation of Functions of Random Variables

\[ E(g(X)) = \begin{cases} 
\sum_{i} g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x) f_X(x) \, dx & X \text{ continuous} 
\end{cases} \]

- **Example:** variance, \( \text{var}(X) = E\left( (X - E(X))^2 \right) = E(X^2) - E(X)^2 \)

- **Example:** Bernoulli variance, \( E(X^2) = E(X) = p \), thus \( \text{var}(X) = p(1-p) \).

- **Example:** Gaussian variance, \( E((X - \mu)^2) = \sigma^2 \).
Expectation of Functions of Random Variables

\[ \mathbb{E}(g(X)) = \begin{cases} 
\sum_{i} g(x_i) f_X(x_i) & X \text{ discrete, } g(x_i) \in \mathbb{R} \\
\int_{-\infty}^{\infty} g(x) f_X(x) \, dx & X \text{ continuous} 
\end{cases} \]

- Example: variance, \( \text{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \)

- Example: Bernoulli variance, \( \mathbb{E}(X^2) = \mathbb{E}(X) = p \), thus \( \text{var}(X) = p(1 - p) \).

- Example: Gaussian variance, \( \mathbb{E}((X - \mu)^2) = \sigma^2 \).

- Probability as expectation of indicator, \( 1_A(x) = \begin{cases} 1 & \iff x \in A \\
0 & \iff x \notin A \end{cases} \)

\[ \mathbb{P}(X \in A) = \int_{A} f_X(x) \, dx = \int 1_A(x) f_X(x) \, dx = \mathbb{E}(1_A(X)) \]
Two (or More) Random Variables

- Joint pmf of two discrete RVs: \( f_{X,Y}(x, y) = \mathbb{P}(X = x \land Y = y) \).

  Extends trivially to more than two RVs.
Two (or More) Random Variables

- **Joint pmf** of two discrete RVs: \( f_{X,Y}(x, y) = \mathbb{P}(X = x \land Y = y) \).
  
  Extends trivially to more than two RVs.

- **Joint pdf** of two continuous RVs: \( f_{X,Y}(x, y) \), such that
  
  \[
  \mathbb{P}((X, Y) \in A) = \int_A f_{X,Y}(x, y) \, dx \, dy, \quad A \in \sigma(\mathbb{R}^2)
  \]

  Extends trivially to more than two RVs.

Independence:

\( X \perp \perp Y \iff f_{X,Y}(x, y) = f_X(x) f_Y(y) \).
Two (or More) Random Variables

- **Joint pmf** of two discrete RVs: \( f_{X,Y}(x, y) = \mathbb{P}(X = x \land Y = y) \).

  Extends trivially to more than two RVs.

- **Joint pdf** of two continuous RVs: \( f_{X,Y}(x, y) \), such that

  \[
  \mathbb{P}((X, Y) \in A) = \int_A \int f_{X,Y}(x, y) \, dx \, dy, \quad A \in \sigma(\mathbb{R}^2)
  \]

  Extends trivially to more than two RVs.

- **Marginalization**: \( f_Y(y) = \begin{cases} 
  \sum_x f_{X,Y}(x, y), & \text{if } X \text{ is discrete} \\
  \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx, & \text{if } X \text{ continuous}
\end{cases} \)
Two (or More) Random Variables

- **Joint pmf** of two discrete RVs: \( f_{X,Y}(x, y) = \mathbb{P}(X = x \land Y = y) \).

  Extends trivially to more than two RVs.

- **Joint pdf** of two continuous RVs: \( f_{X,Y}(x, y) \), such that

  \[
  \mathbb{P}((X, Y) \in A) = \int \int_A f_{X,Y}(x, y) \, dx \, dy, \quad A \in \sigma(\mathbb{R}^2)
  \]

  Extends trivially to more than two RVs.

- **Marginalization:** \( f_Y(y) = \begin{cases} 
\sum_x f_{X,Y}(x, y), & \text{if } X \text{ is discrete} \\
\int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx, & \text{if } X \text{ continuous}
\end{cases} \)

- **Independence:**

  \[X \independent Y \iff f_{X,Y}(x, y) = f_X(x) f_Y(y)\]
Two (or More) Random Variables

- **Joint pmf** of two discrete RVs: \( f_{X,Y}(x, y) = P(X = x \land Y = y) \).
  
  Extends trivially to more than two RVs.

- **Joint pdf** of two continuous RVs: \( f_{X,Y}(x, y) \), such that

  \[
  P((X, Y) \in A) = \int \int_A f_{X,Y}(x, y) \, dx \, dy, \quad A \in \sigma(\mathbb{R}^2)
  \]

  Extends trivially to more than two RVs.

- **Marginalization**: 

  \[
  f_Y(y) = \begin{cases} 
  \sum_{x} f_{X,Y}(x, y), & \text{if } X \text{ is discrete} \\
  \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx, & \text{if } X \text{ continuous}
  \end{cases}
  \]

- **Independence**: 

  \[
  X \perp Y \iff f_{X,Y}(x, y) = f_X(x) \, f_Y(y) \quad \Rightarrow \quad \mathbb{E}(XY) = \mathbb{E}(X) \, \mathbb{E}(Y).
  \]
Conditionals and Bayes’ Theorem

- **Conditional pmf** (discrete RVs):
  \[ f_{X|Y}(x|y) = \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x \land Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]
Conditionals and Bayes’ Theorem

- **Conditional pmf** (discrete RVs):
  \[ f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x \land Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]

- **Conditional pdf** (continuous RVs):
  \[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \]

  ...the meaning is technically delicate.
Conditionals and Bayes’ Theorem

- **Conditional pmf** (discrete RVs):
  \[ f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x \land Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]

- **Conditional pdf** (continuous RVs):
  \[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]
  ...the meaning is technically delicate.

- **Bayes’ theorem**: \[ f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \] (pdf or pmf).
Conditionals and Bayes’ Theorem

- **Conditional pmf** (discrete RVs):
  \[ f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x \land Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}. \]

- **Conditional pdf** (continuous RVs):
  \[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \]
  ...the meaning is technically delicate.

- **Bayes’ theorem**:
  \[ f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} \] (pdf or pmf).

- Also valid in the mixed case (e.g., \(X\) continuous, \(Y\) discrete).
Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

  \[
  \begin{array}{|c|c|c|}
  \hline
  f_{X,Y}(x, y) & Y = 0 & Y = 1 \\
  \hline
  X = 0 & 1/5 & 2/5 \\
  X = 1 & 1/10 & 3/10 \\
  \hline
  \end{array}
  \]
Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

<table>
<thead>
<tr>
<th>$f_{X,Y}(x, y)$</th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$1/5$</td>
<td>$2/5$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$1/10$</td>
<td>$3/10$</td>
</tr>
</tbody>
</table>

- Marginals:
  
  $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$, \quad $f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$,
  
  $f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}$, \quad $f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}$. 

Joint, Marginal, and Conditional Probabilities: An Example

- A pair of binary variables $X, Y \in \{0, 1\}$, with joint pmf:

<table>
<thead>
<tr>
<th>$f_{X,Y}(x, y)$</th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>1/5</td>
<td>2/5</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>1/10</td>
<td>3/10</td>
</tr>
</tbody>
</table>

- Marginals: $f_X(0) = \frac{1}{5} + \frac{2}{5} = \frac{3}{5}$, $f_X(1) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$,
  $f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}$, $f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}$.

- Conditional probabilities:

| $f_{X|Y}(x|y)$ | $Y = 0$ | $Y = 1$ |
|----------------|---------|---------|
| $X = 0$        | 2/3     | 4/7     |
| $X = 1$        | 1/3     | 3/7     |

| $f_{Y|X}(y|x)$ | $Y = 0$ | $Y = 1$ |
|----------------|---------|---------|
| $X = 0$        | 1/3     | 2/3     |
| $X = 1$        | 1/4     | 3/4     |
An Important Multivariate RV: Multinomial

- **Multinomial**: \( X = (X_1, \ldots, X_K) \), \( X_i \in \{0, ..., n\} \), such that \( \sum_i X_i = n \),

\[
f_X(x_1, \ldots, x_K) = \begin{cases} 
\binom{n}{x_1 \ x_2 \ \cdots \ x_K} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \iff \sum_i x_i = n \\
0 & \iff \sum_i x_i \neq n 
\end{cases}
\]

\[
\binom{n}{x_1 \ x_2 \ \cdots \ x_K} = \frac{n!}{x_1! \ x_2! \ \cdots \ x_K!}
\]

Parameters: \( p_1, \ldots, p_K \geq 0 \), such that \( \sum_i p_i = 1 \).
An Important Multivariate RV: Multinomial

- **Multinomial**: \( X = (X_1, ..., X_K) \), \( X_i \in \{0, ..., n\} \), such that \( \sum_i X_i = n \),

\[
f_X(x_1, ..., x_K) = \begin{cases} 
\left( \begin{array}{c} n \\ x_1 \ x_2 \ ... \ x_K \end{array} \right) p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \iff \sum_i x_i = n \\
0 & \iff \sum_i x_i \neq n
\end{cases}
\]

\[
\left( \begin{array}{c} n \\ x_1 \ x_2 \ ... \ x_K \end{array} \right) = \frac{n!}{x_1! \ x_2! \ ... \ x_K!}
\]

Parameters: \( p_1, ..., p_K \geq 0 \), such that \( \sum_i p_i = 1 \).

- Generalizes the binomial from binary to \( K \)-classes.
An Important Multivariate RV: Multinomial

- **Multinomial**: $X = (X_1, ..., X_K)$, $X_i \in \{0, ..., n\}$, such that $\sum_i X_i = n$.

$$f_X(x_1, ..., x_K) = \begin{cases} \binom{n}{x_1 x_2 \cdots x_K} p_1^{x_1} p_2^{x_2} \cdots p_K^{x_K} & \text{if } \sum_i x_i = n \\ 0 & \text{if } \sum_i x_i \neq n \end{cases}$$

$$\binom{n}{x_1 x_2 \cdots x_K} = \frac{n!}{x_1! x_2! \cdots x_K!}$$

Parameters: $p_1, ..., p_K \geq 0$, such that $\sum_i p_i = 1$.

- Generalizes the binomial from binary to $K$-classes.

- **Example**: tossing $n$ independent fair dice, $p_1 = \cdots = p_6 = 1/6$.
  $x_i =$ number of outcomes with $i$ dots. Of course, $\sum_i x_i = n$. 
An Important Multivariate RV: Gaussian

- **Multivariate Gaussian**: $X \in \mathbb{R}^n$, 

$$f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left(-\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu) \right)$$

Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n \times n}$.

Expected value: $E(X) = \mu$. Meaning of $C$: next slide.
An Important Multivariate RV: Gaussian

- Multivariate Gaussian: $X \in \mathbb{R}^n$,

$$f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{\det(2\pi C)}} \exp\left(-\frac{1}{2}(x - \mu)^T C^{-1}(x - \mu)\right)$$

- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n \times n}$.
  Expected value: $\mathbb{E}(X) = \mu$. Meaning of $C$: next slide.
An Important Multivariate RV: Gaussian

- **Multivariate Gaussian**: \( X \in \mathbb{R}^n \),

\[
f_X(x) = \mathcal{N}(x; \mu, C) = \frac{1}{\sqrt{\det(2\pi C)}} \exp \left( -\frac{1}{2} (x - \mu)^T C^{-1} (x - \mu) \right)
\]

- **Parameters**: vector \( \mu \in \mathbb{R}^n \) and matrix \( C \in \mathbb{R}^{n \times n} \).
  Expected value: \( \mathbb{E}(X) = \mu \). Meaning of \( C \): next slide.
Covariance, Correlation, and all that...

**Covariance** between two RVs:

\[
\text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X)) (Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
\]
Covariance, Correlation, and all that...

- **Covariance** between two RVs:
  \[
  \text{cov}(X, Y) = \mathbb{E}\left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
  \]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).
Covariance, Correlation, and all that...

- **Covariance** between two RVs:
  \[
  \text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X)) \left( Y - \mathbb{E}(Y) \right) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
  \]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).

- **Correlation**: \( \text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \in [-1, 1] \)
Covariance, Correlation, and all that...

- **Covariance** between two RVs:

\[
\text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X)) (Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X) \mathbb{E}(Y)
\]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).

- **Correlation**: \( \text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}} \in [-1, 1] \)

- \( X \perp Y \iff f_{X,Y}(x,y) = f_X(x) f_Y(y) \)
Covariance, Correlation, and all that...

- **Covariance** between two RVs:
  \[
  \text{cov}(X, Y) = \mathbb{E}\left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
  \]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).

- **Correlation**: \( \text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \in [-1, 1] \)

- \( X \perp \!\!\!\!\perp Y \iff f_{X,Y}(x,y) = f_X(x)f_Y(y) \Rightarrow \text{cov}(X, Y) = 0. \)
Covariance, Correlation, and all that...

- **Covariance** between two RVs:
  \[
  \text{cov}(X, Y) = \mathbb{E} \left[ (X - \mathbb{E}(X)) (Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
  \]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).

- **Correlation**: 
  \[
  \text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} \in [-1, 1]
  \]

- \( X \perp \perp Y \iff f_{X,Y}(x,y) = f_X(x)f_Y(y) \Rightarrow \text{cov}(X, Y) = 0. \)

- **Covariance matrix** of multivariate RV, \( X \in \mathbb{R}^n \):
  \[
  \text{cov}(X) = \mathbb{E} \left[ (X - \mathbb{E}(X))(X - \mathbb{E}(X))^T \right] = \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T
  \]
Covariance, Correlation, and all that...

- **Covariance** between two RVs:
  \[
  \text{cov}(X, Y) = \mathbb{E}\left[ (X - \mathbb{E}(X))(Y - \mathbb{E}(Y)) \right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)
  \]

- Relationship with variance: \( \text{var}(X) = \text{cov}(X, X) \).

- **Correlation**: \( \text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \in [-1, 1] \)

- \( X \perp \perp Y \iff f_{X,Y}(x, y) = f_X(x)f_Y(y) \Rightarrow \text{cov}(X, Y) = 0 \).

- **Covariance matrix** of multivariate RV, \( X \in \mathbb{R}^n \):
  \[
  \text{cov}(X) = \mathbb{E}\left[ (X - \mathbb{E}(X))(X - \mathbb{E}(X))^T \right] = \mathbb{E}(XX^T) - \mathbb{E}(X)\mathbb{E}(X)^T
  \]

- **Covariance of Gaussian RV**, \( f_X(x) = \mathcal{N}(x; \mu, C) \Rightarrow \text{cov}(X) = C \)
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
More on Expectations and Covariances

Let \( A \in \mathbb{R}^{n \times n} \) be a matrix and \( a \in \mathbb{R}^n \) a vector.

- If \( \mathbb{E}(X) = \mu \) and \( Y = AX \), then \( \mathbb{E}(Y) = A\mu \);

- If \( \mathbb{E}(X) = \mu \) and \( Y = X - \mu \), then \( \mathbb{E}(Y) = 0 \);

- If \( \text{cov}(X) = C \) and \( Y = AX \), then \( \text{cov}(Y) = ACA^T \);

- If \( \text{cov}(X) = C \) and \( Y = a^T X \in \mathbb{R} \), then \( \text{var}(Y) = a^T C a \geq 0 \);

- If \( \text{cov}(X) = C \) and \( Y = C^{-1}/2 \ X \), then \( \text{cov}(Y) = I \);

Combining the 2-nd and the 4-th facts is called standardization.
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- If $\mathbb{E}(X) = \mu$ and $Y = X - \mu$, then $\mathbb{E}(Y) = 0$;
- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = ACA^T$;
- If $\text{cov}(X) = C$ and $Y = a^T X \in \mathbb{R}$, then $\text{var}(Y) = a^T Ca \geq 0$;
- If $\text{cov}(X) = C$ and $Y = C^{-1}/2 X$, then $\text{cov}(Y) = I$;

Combining the 2-nd and the 4-th facts is called standardization.
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- If $\mathbb{E}(X) = \mu$ and $Y = X - \mu$, then $\mathbb{E}(Y) = 0$;
- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = ACA^T$;
- If $\text{cov}(X) = C$ and $Y = a^TX \in \mathbb{R}$, then $\text{var}(Y) = a^TCa \geq 0$;

Combining the 2-nd and the 4-th facts is called standardization.
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- If $\mathbb{E}(X) = \mu$ and $Y = X - \mu$, then $\mathbb{E}(Y) = 0$;
- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = ACA^T$;
- If $\text{cov}(X) = C$ and $Y = a^T X \in \mathbb{R}$, then $\text{var}(Y) = a^T Ca \geq 0$;
- If $\text{cov}(X) = C$ and $Y = C^{-1/2} X$, then $\text{cov}(Y) = I$;
More on Expectations and Covariances

Let $A \in \mathbb{R}^{n \times n}$ be a matrix and $a \in \mathbb{R}^n$ a vector.

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;

- If $\mathbb{E}(X) = \mu$ and $Y = X - \mu$, then $\mathbb{E}(Y) = 0$;

- If $\text{cov}(X) = C$ and $Y = AX$, then $\text{cov}(Y) = AC A^T$;

- If $\text{cov}(X) = C$ and $Y = a^T X \in \mathbb{R}$, then $\text{var}(Y) = a^T Ca \geq 0$;

- If $\text{cov}(X) = C$ and $Y = C^{-1/2} X$, then $\text{cov}(Y) = I$;

Combining the 2-nd and the 4-th facts is called standardization.
Statistical Inference

- **Scenario:** observed RV $Y$, depends on unknown variable(s) $X$.
- **Goal:** given an observation $Y = y$, infer $X$. 

Two main philosophies:

- **Frequentist:** $X = x$ is fixed, but unknown;
- **Bayesian:** $X$ is a RV with pdf/pmf $f_X(x)$ (the prior).

Prior $\Leftrightarrow$ knowledge about $X$.

In both philosophies, a central object is $f_{Y|X}(y|x)$.

Several names: likelihood function, observation model, ...

This is not machine learning!

$f_{Y,X}(y,x)$ is assumed known.

In the Bayesian philosophy, all the knowledge about $X$ is carried by $f_{X|Y}(x|y) = f_{Y|X}(y|x) f_X(x) f_Y(y) = f_{Y,X}(y,x) f_Y(y)$...

...the posterior (or a posteriori) pdf/pmf.
Statistical Inference

- **Scenario**: observed RV $Y$, depends on unknown variable(s) $X$.
- **Goal**: given an observation $Y = y$, infer $X$.

- **Two main philosophies**:
  - **Frequentist**: $X = x$ is fixed, but unknown;
  - **Bayesian**: $X$ is a RV with pdf/pmf $f_X(x)$ (the prior)
    - prior $\Leftrightarrow$ knowledge about $X$
Statistical Inference

- **Scenario**: observed RV $Y$, depends on unknown variable(s) $X$.
  - **Goal**: given an observation $Y = y$, infer $X$.

- **Two main philosophies**:
  - **Frequentist**: $X = x$ is fixed, but unknown;
  - **Bayesian**: $X$ is a RV with pdf/pmf $f_X(x)$ (the prior)
    - prior $\Leftrightarrow$ knowledge about $X$

- In both philosophies, a central object is $f_Y|X(y|x)$
  - several names: likelihood function, observation model,...
Statistical Inference

- **Scenario**: observed RV $Y$, depends on unknown variable(s) $X$.
  - **Goal**: given an observation $Y = y$, infer $X$.

- **Two main philosophies**:
  - **Frequentist**: $X = x$ is fixed, but unknown;
  - **Bayesian**: $X$ is a RV with pdf/pmf $f_X(x)$ (the prior)
    - prior $\Leftrightarrow$ knowledge about $X$

- In both philosophies, a central object is $f_{Y|X}(y|x)$
  - several names: likelihood function, observation model,...

- This in **not** machine learning! $f_{Y,X}(y,x)$ is assumed known.
Statistical Inference

- **Scenario:** observed RV $Y$, depends on unknown variable(s) $X$.
  - **Goal:** given an observation $Y = y$, infer $X$.

- **Two main philosophies:**
  - **Frequentist:** $X = x$ is fixed, but unknown;
  - **Bayesian:** $X$ is a RV with pdf/pmf $f_X(x)$ (the prior)
  - prior $\Leftrightarrow$ knowledge about $X$

- In both philosophies, a central object is $f_{Y|X}(y|x)$
  - several names: likelihood function, observation model,...

- This in **not** machine learning! $f_{Y,X}(y,x)$ is assumed known.

- In the Bayesian philosophy, all the knowledge about $X$ is carried by
  \[
  f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{Y,X}(y,x)}{f_Y(y)}
  \]
  - ...the posterior (or a posteriori) pdf/pmf.
Statistical Inference

- The posterior pdf/pmf $f_{X|Y}(x|y)$ has all the information/knowledge about $X$, given $Y = y$ (conditionality principle).
Statistical Inference

- The posterior pdf/pmf $f_{X|Y}(x|y)$ has all the information/knowledge about $X$, given $Y = y$ (conditionality principle).

- How to make an optimal “guess” $\hat{x}$ about $X$, given this information?
Statistical Inference

- The posterior pdf/pmf $f_{X|Y}(x|y)$ has all the information/knowledge about $X$, given $Y = y$ (conditionality principle).

- How to make an optimal “guess” $\hat{x}$ about $X$, given this information?

- Need to define “optimal”: loss function: $L(\hat{x}, x) \geq 0$ measures “loss”/“cost” incurred by “guessing” $\hat{x}$ if truth is $x$. 
Statistical Inference

- The posterior pdf/pmf $f_{X|Y}(x|y)$ has all the information/knowledge about $X$, given $Y = y$ (conditionality principle).

- How to make an optimal “guess” $\hat{x}$ about $X$, given this information?

- Need to define “optimal”: loss function: $L(\hat{x}, x) \geq 0$ measures “loss”/“cost” incurred by “guessing” $\hat{x}$ if truth is $x$.

- The optimal Bayesian decision minimizes the expected loss:

$$\hat{x}_\text{Bayes} = \arg \min_{\hat{x}} \mathbb{E}[L(\hat{x}, X)|Y = y]$$

where

$$\mathbb{E}[L(\hat{x}, X)|Y = y] = \begin{cases} 
\int L(\hat{x}, x) f_{X|Y}(x|y) \, dx, & \text{continuous (estimation)} \\
\sum_x L(\hat{x}, x) f_{X|Y}(x|y), & \text{discrete (classification)}
\end{cases}$$
Classical Statistical Inference Criteria

- Consider that $X \in \{1, \ldots, K\}$ (discrete/classification problem).
Classical Statistical Inference Criteria

- Consider that $X \in \{1, \ldots, K\}$ (discrete/classification problem).
- Adopt the 0/1 loss: $L(\hat{x}, x) = 0$, if $\hat{x} = x$, and 1 otherwise.
Classical Statistical Inference Criteria

- Consider that $X \in \{1, \ldots, K\}$ (discrete/classification problem).

- Adopt the $0/1$ loss: $L(\hat{x}, x) = 0$, if $\hat{x} = x$, and 1 otherwise.

- Optimal Bayesian decision:

$$
\hat{x}_{\text{Bayes}} = \arg \min_{\hat{x}} \sum_{x=1}^{K} L(\hat{x}, x) f_{X|Y}(x|y)
$$

$$
= \arg \min_{\hat{x}} \left(1 - f_{X|Y}(\hat{x}|y)\right)
$$

$$
= \arg \max_{\hat{x}} f_{X|Y}(\hat{x}|y) \equiv \hat{x}_{\text{MAP}}
$$

MAP = maximum a posteriori criterion.
Classical Statistical Inference Criteria

- Consider that \( X \in \{1, \ldots, K\} \) (discrete/classification problem).

- Adopt the 0/1 loss: \( L(\hat{x}, x) = 0 \), if \( \hat{x} = x \), and 1 otherwise.

- Optimal Bayesian decision:

\[
\hat{x}_{\text{Bayes}} = \arg \min_{\hat{x}} \sum_{x=1}^{K} L(\hat{x}, x) f_{X|Y}(x|y) \\
= \arg \min_{\hat{x}} \left( 1 - f_{X|Y}(\hat{x}|y) \right) \\
= \arg \max_{\hat{x}} f_{X|Y}(\hat{x}|y) \equiv \hat{x}_{\text{MAP}}
\]

MAP = maximum a posteriori criterion.

- Same criterion can be derived for continuous \( X \)
Classical Statistical Inference Criteria

- Consider the MAP criterion \( \hat{x}_{\text{MAP}} = \arg \max_x f_{X|Y}(x|y) \)
Classical Statistical Inference Criteria

- Consider the MAP criterion \( \hat{x}_{\text{MAP}} = \arg \max_x f_{X|Y}(x|y) \)

- From Bayes law:

\[
\hat{x}_{\text{MAP}} = \arg \max_x \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \arg \max_x f_{Y|X}(y|x) f_X(x)
\]

...only need to know posterior \( f_{X|Y}(x|y) \) up to a normalization factor.
Classical Statistical Inference Criteria

- Consider the MAP criterion $\hat{x}_{\text{MAP}} = \arg \max_x f_{X|Y}(x|y)$

- From Bayes law:

$$\hat{x}_{\text{MAP}} = \arg \max_x \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} = \arg \max_x f_{Y|X}(y|x) f_X(x)$$

...only need to know posterior $f_{X|Y}(x|y)$ up to a normalization factor.

- Also common to write:

$$\hat{x}_{\text{MAP}} = \arg \max_x \left( \log f_{Y|X}(y|x) + \log f_X(x) \right)$$
Classical Statistical Inference Criteria

- Consider the MAP criterion $\hat{x}_{\text{MAP}} = \arg \max_x f_{X|Y}(x|y)$

- From Bayes law:

$$\hat{x}_{\text{MAP}} = \arg \max_x \frac{f_Y|X(y|x) f_X(x)}{f_Y(y)} = \arg \max_x f_{Y|X}(y|x) f_X(x)$$

...only need to know posterior $f_{X|Y}(x|y)$ up to a normalization factor.

- Also common to write:

$$\hat{x}_{\text{MAP}} = \arg \max_x \left( \log f_{Y|X}(y|x) + \log f_X(x) \right)$$

- If the prior is flat, $f_X(x) = C$, then,

$$\hat{x}_{\text{MAP}} = \arg \max_x f_{Y|X}(y|x) \equiv \hat{x}_{\text{ML}}$$

ML = maximum likelihood criterion.
Statistical Inference: Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs:
  \[ Y = (Y_1, ..., Y_n), \text{ with } Y_i \in \{0, 1\}. \]
- Common pmf $f_{Y_i|X}(y|x) = x^y(1 - x)^{1-y}$, where $x \in [0, 1]$. 
Statistical Inference: Example

- Observed \( n \) i.i.d. (independent identically distributed) Bernoulli RVs: \( Y = (Y_1, ..., Y_n) \), with \( Y_i \in \{0, 1\} \).

  Common pmf \( f_{Y_i|X}(y|x) = x^y(1 - x)^{1-y} \), where \( x \in [0, 1] \).

- Likelihood function: \( f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} \)

  Log-likelihood function:

  \[
  \log f_{Y|X}(y_1, ..., y_n|x) = n \log(1 - x) + \log \frac{x}{1 - x} \sum_{i=1}^{n} y_i
  \]
Statistical Inference: Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs: $Y = (Y_1, ..., Y_n)$, with $Y_i \in \{0, 1\}$.

  Common pmf $f_{Y_i|X}(y|x) = x^y (1 - x)^{1-y}$, where $x \in [0, 1]$.

- Likelihood function: $f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i}$

  Log-likelihood function:

  $$\log f_{Y|X}(y_1, ..., y_n|x) = n \log(1 - x) + \log \frac{x}{1 - x} \sum_{i=1}^{n} y_i$$

- Maximum likelihood: $\hat{x}_{ML} = \arg \max_x f_{Y|X}(y|x) = \frac{1}{n} \sum_{i=1}^{n} y_i$
Statistical Inference: Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs: $Y = (Y_1, ..., Y_n)$, with $Y_i \in \{0, 1\}$.

- Common pmf $f_{Y_i|X}(y|x) = x^y (1 - x)^{1-y}$, where $x \in [0, 1]$.

- Likelihood function: $f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i}$

- Log-likelihood function:

$$\log f_{Y|X}(y_1, ..., y_n|x) = n \log(1 - x) + \log \frac{x}{1 - x} \sum_{i=1}^{n} y_i$$

- Maximum likelihood: $\hat{x}_{ML} = \arg \max_x f_{Y|X}(y|x) = \frac{1}{n} \sum_{i=1}^{n} y_i$

- Example: $n = 10$, observed $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$, $\hat{x}_{ML} = 7/10$. 
Statistical Inference: Example (Continuation)

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

Likelihood:

$$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x y_i (1 - x)^{1 - y_i}$$

How to express knowledge that (e.g.) $X$ is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

In our case, the Beta pdf

$$f_X(x) \propto x^{\alpha - 1} (1 - x)^{\beta - 1}, \alpha, \beta > 0$$

Posterior:

$$f_{X|Y}(x|y) = x^{\alpha - 1 + \sum_i y_i} (1 - x)^{\beta - 1 + n - \sum_i y_i}$$

MAP:

$$\hat{x}_{MAP} = \frac{\alpha + \sum_i y_i - 1}{\alpha + \beta + n - 2}$$

Example:

$$\alpha = 4, \beta = 4, n = 10, y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$$

$$\hat{x}_{MAP} = 0.625$$

(recall $\hat{x}_{ML} = 0.7$)
Statistical Inference: Example (Continuation)

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:

$$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n-\sum_i y_i}$$
Statistical Inference: Example (Continuation)

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  \[
  f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n - \sum_i y_i}
  \]

- How to express knowledge that (e.g.) $X$ is around $1/2$? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

- Example: $\alpha = 4$, $\beta = 4$, $n = 10$, $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$, $\hat{x}_{\text{MAP}} = 0.625$ (recall $\hat{x}_{\text{ML}} = 0.7$).
Statistical Inference: Example (Continuation)

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  \[
  f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n - \sum_i y_i}
  \]

- How to express knowledge that (e.g.) $X$ is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

  - In our case, the Beta pdf
    \[
    f_X(x) \propto x^{\alpha - 1} (1 - x)^{\beta - 1}, \quad \alpha, \beta > 0
    \]
Observed \( n \) i.i.d. (independent identically distributed) Bernoulli RVs.

Likelihood:
\[
f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n - \sum_i y_i}
\]

How to express knowledge that (e.g.) \( X \) is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

- In our case, the Beta pdf
  \[
f_X(x) \propto x^{\alpha-1} (1 - x)^{\beta-1}, \quad \alpha, \beta > 0
\]
- Posterior: \( f_{X|Y}(x|y) = x^{\alpha-1 + \sum_i y_i} (1 - x)^{\beta-1 + n - \sum_i y_i} \)
Statistical Inference: Example (Continuation)

- Observed \( n \) i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  \[
  f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n-\sum_i y_i}
  \]

- How to express knowledge that (e.g.) \( X \) is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

  - In our case, the Beta pdf
    \[
    f_X(x) \propto x^{\alpha-1} (1 - x)^{\beta-1}, \quad \alpha, \beta > 0
    \]
  
  - Posterior: \( f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i} (1 - x)^{\beta-1+n-\sum_i y_i} \)
  
  - MAP: \( \hat{x}_{MAP} = \frac{\alpha+\sum_i y_i-1}{\alpha+\beta+n-2} \)

\[\text{PDF} \]
\[x \quad 0.0 \quad 0.2 \quad 0.4 \quad 0.6 \quad 0.8 \quad 1.0\]
\[0.0 \quad 0.5 \quad 1.0 \quad 1.5 \quad 2.0 \quad 2.5\]

\[\begin{array}{c}
\alpha=\beta=0.5 \\
\alpha=5,\beta=1 \\
\alpha=1,\beta=3 \\
\alpha=2,\beta=2 \\
\alpha=2,\beta=5 \\
\end{array}\]
Statistical Inference: Example (Continuation)

- Observed \( n \) i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  \[
  f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i}(1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n-\sum y_i}
  \]

- How to express knowledge that (e.g.) \( X \) is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.
  
  ▶ In our case, the Beta pdf
  \[
  f_X(x) \propto x^{\alpha-1}(1 - x)^{\beta-1}, \quad \alpha, \beta > 0
  \]
  
  ▶ Posterior:
  \[
  f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i} (1 - x)^{\beta-1+n-\sum_i y_i}
  \]
  
  ▶ MAP:
  \[
  \hat{x}_{\text{MAP}} = \frac{\alpha+\sum_i y_i}{\alpha+\beta+n-2}
  \]
  
  ▶ Example: \( \alpha = 4, \beta = 4, n = 10, \)
  \( y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1), \)
  \[
  \hat{x}_{\text{MAP}} = 0.625 \quad (\text{recall} \ \hat{x}_{\text{ML}} = 0.7)
  \]
Another Classical Statistical Inference Criterion

- Consider that $X \in \mathbb{R}$ (continuous/estimation problem).

Optimal Bayesian decision:

$$\hat{x}_{\text{Bayes}} = \arg\min_{\hat{x}} E[(\hat{x} - X)^2 | Y = y]$$

$$= \arg\min_{\hat{x}} \hat{x}^2 - 2\hat{x} E[X | Y = y] = E[X | Y = y] \equiv \hat{x}_{\text{MMSE}}$$

MMSE = minimum mean squared error criterion.

Does not apply to classification problems.
Another Classical Statistical Inference Criterion

- Consider that $X \in \mathbb{R}$ (continuous/estimation problem).
- Adopt the squared error loss: $L(\hat{x}, x) = (\hat{x} - x)^2$

Optimal Bayesian decision:

$$\hat{x}_{\text{Bayes}} = \arg\min_{\hat{x}} E[(\hat{x} - X)^2 | Y = y] = \arg\min_{\hat{x}} \hat{x}^2 - 2\hat{x}E[X | Y = y] = E[X | Y = y] \equiv \hat{x}_{\text{MMSE}}$$

MMSE = minimum mean squared error criterion.

Does not apply to classification problems.
Another Classical Statistical Inference Criterion

- Consider that $X \in \mathbb{R}$ (continuous/estimation problem).

- Adopt the squared error loss: $L(\hat{x}, x) = (\hat{x} - x)^2$

- Optimal Bayesian decision:

  $$\hat{x}_{\text{Bayes}} = \arg\min_{\hat{x}} \mathbb{E}[(\hat{x} - X)^2 | Y = y]$$
  $$= \arg\min_{\hat{x}} \hat{x}^2 - 2 \hat{x} \mathbb{E}[X | Y = y]$$
  $$= \mathbb{E}[X | Y = y] \equiv \hat{x}_{\text{MMSE}}$$

MMSE = minimum mean squared error criterion.
Another Classical Statistical Inference Criterion

- Consider that $X \in \mathbb{R}$ (continuous/estimation problem).

- Adopt the squared error loss: $L(\hat{x}, x) = (\hat{x} - x)^2$

- Optimal Bayesian decision:

$$\hat{x}_{\text{Bayes}} = \arg\min_{\hat{x}} \mathbb{E}[(\hat{x} - X)^2|Y = y]$$

$$= \arg\min_{\hat{x}} \hat{x}^2 - 2\hat{x}\mathbb{E}[X|Y = y]$$

$$= \mathbb{E}[X|Y = y] \equiv \hat{x}_{\text{MMSE}}$$

MMSE $= \text{minimum mean squared error criterion}$.

- Does not apply to classification problems.
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood:

$$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x\sum_{i} y_i (1 - x)^{n-\sum_{i} y_i}$$
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:

$$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x \sum_i y_i (1 - x)^{n-\sum_i y_i}$$

- In our case, the Beta pdf

$$f_X(x) \propto x^{\alpha-1} (1 - x)^{\beta-1}, \quad \alpha, \beta > 0$$

In our case, the Beta pdf

$$f_X(x) \propto x^{\alpha-1} (1 - x)^{\beta-1}, \quad \alpha, \beta > 0$$
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  $$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1-x)^{1-y_i} = x^{\sum_i y_i} (1-x)^{n-\sum_i y_i}$$

- In our case, the Beta pdf
  $$f_X(x) \propto x^{\alpha-1} (1-x)^{\beta-1}, \quad \alpha, \beta > 0$$

- Posterior:
  $$f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i} (1-x)^{\beta-1+n-\sum_i y_i}$$

Conjugate prior equivalent to “virtual” counts; often called smoothing in NLP and ML.
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  
  $$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n - \sum_i y_i}$$

- In our case, the Beta pdf
  
  $$f_X(x) \propto x^{\alpha-1} (1 - x)^{\beta-1}, \quad \alpha, \beta > 0$$

- Posterior:
  
  $$f_{X|Y}(x|y) = x^{\alpha - 1 + \sum_i y_i} (1 - x)^{\beta - 1 + n - \sum_i y_i}$$

- MMSE:
  
  $$\hat{x}_{MMSE} = \frac{\alpha + \sum_i y_i}{\alpha + \beta + n}$$
Back to the Bernoulli Example

- Observed $n$ i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:
  $$f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$$

- In our case, the Beta pdf
  $$f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \alpha,\beta > 0$$

- Posterior: $f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i}(1-x)^{\beta-1+n-\sum_i y_i}$

- MMSE: $\hat{x}_{\text{MMSE}} = \frac{\alpha+\sum_i y_i}{\alpha+\beta+n}$

- Example: $\alpha = 4, \beta = 4, n = 10$, $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$,
  $$\hat{x}_{\text{MMSE}} \approx 0.611 \text{ (recall that } \hat{x}_{\text{MAP}} = 0.625, \hat{x}_{\text{ML}} = 0.7)$$
Back to the Bernoulli Example

- Observed \( n \) i.i.d. (independent identically distributed) Bernoulli RVs.

- Likelihood:

\[
f_{Y|X}(y_1, \ldots, y_n|x) = \prod_{i=1}^{n} x^{y_i} (1 - x)^{1-y_i} = x^{\sum_i y_i} (1 - x)^{n-\sum_i y_i}
\]

- In our case, the Beta pdf

\[
f_X(x) \propto x^{\alpha-1}(1 - x)^{\beta-1}, \alpha, \beta > 0
\]

- Posterior: \( f_{X|Y}(x|y) = \\
x^{\alpha-1+\sum_i y_i}(1 - x)^{\beta-1+n-\sum_i y_i}
\]

- MMSE: \( \hat{x}_{\text{MMSE}} = \frac{\alpha+\sum_i y_i}{\alpha+\beta+n} \)

- Example: \( \alpha = 4, \beta = 4, n = 10, \\
y = (1, 1, 1, 0, 1, 0, 1, 1, 1, 1) \\
\hat{x}_{\text{MMSE}} \simeq 0.611 \) (recall that \( \hat{x}_{\text{MAP}} = 0.625, \hat{x}_{\text{ML}} = 0.7 \))

- Conjugate prior equivalent to “virtual” counts; often called \textit{smoothing} in NLP and ML.
The Bernstein-Von Mises Theorem

In the previous example, we had
\[ n = 10, \quad y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1), \quad \text{thus } \sum_i y_i = 7. \]
With a Beta prior with \( \alpha = 4 \) and \( \beta = 4 \), we had
\[
\hat{x}_{\text{ML}} = 0.7, \quad \hat{x}_{\text{MAP}} = \frac{3 + \sum_i y_i}{6 + n} = 0.625, \quad \hat{x}_{\text{MMSE}} = \frac{4 + \sum_i y_i}{8 + n} \approx 0.611.
\]
The Bernstein-Von Mises Theorem

In the previous example, we had
\[ n = 10, \ y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1), \] thus \[ \sum_i y_i = 7. \]
With a Beta prior with \( \alpha = 4 \) and \( \beta = 4 \), we had
\[ \hat{x}_{ML} = 0.7, \quad \hat{x}_{MAP} = \frac{3 + \sum_i y_i}{6 + n} = 0.625, \quad \hat{x}_{MMSE} = \frac{4 + \sum_i y_i}{8 + n} \approx 0.611 \]

Consider \( n = 100 \), and \( \sum_i y_i = 70 \), with the same Beta(4,4) prior
\[ \hat{x}_{ML} = 0.7, \quad \hat{x}_{MAP} = \frac{73}{106} \approx 0.689, \quad \hat{x}_{MMSE} = \frac{74}{108} \approx 0.685 \]

... both Bayesian estimates are much closer to the ML.
The Bernstein-Von Mises Theorem

- In the previous example, we had $n = 10$, $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$, thus $\sum_i y_i = 7$.
  With a Beta prior with $\alpha = 4$ and $\beta = 4$, we had
  
  \[ \hat{x}_{\text{ML}} = 0.7, \quad \hat{x}_{\text{MAP}} = \frac{3 + \sum_i y_i}{6 + n} = 0.625, \quad \hat{x}_{\text{MMSE}} = \frac{4 + \sum_i y_i}{8 + n} \approx 0.611 \]

- Consider $n = 100$, and $\sum_i y_i = 70$, with the same Beta(4,4) prior
  
  \[ \hat{x}_{\text{ML}} = 0.7, \quad \hat{x}_{\text{MAP}} = \frac{73}{106} \approx 0.689, \quad \hat{x}_{\text{MMSE}} = \frac{74}{108} \approx 0.685 \]

... both Bayesian estimates are much closer to the ML.

- This illustrates an important result in Bayesian inference: the Bernstein-Von Mises theorem; under (mild) conditions,
  
  \[ \lim_{n \to \infty} \hat{x}_{\text{MAP}} = \lim_{n \to \infty} \hat{x}_{\text{MMSE}} = \hat{x}_{\text{ML}} \]

message: if you have a lot of data, priors don’t matter much.
Important Inequalities

- **Cauchy-Schwartz’s inequality for RVs:**

\[ \mathbb{E}(|X Y|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \]
Important Inequalities

- **Cauchy-Schwartz’s inequality** for RVs:

\[ \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \]

- Recall that a real function \( g \) is convex if, for any \( x, y, \) and \( \alpha \in [0, 1] \)

\[ g(\alpha x + (1 - \alpha) y) \leq \alpha g(x) + (1 - \alpha) g(y) \]
Important Inequalities

- **Cauchy-Schwartz’s inequality** for RVs:
\[
\mathbb{E}(|X Y|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}
\]

- Recall that a real function \( g \) is convex if, for any \( x, y, \) and \( \alpha \in [0, 1] \)
\[
g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y)
\]

**Jensen’s inequality**: if \( g \) is a real convex function, then
\[
\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))
\]

![Graphs showing non-convex and convex functions](image-url)
Important Inequalities

- **Cauchy-Schwartz’s inequality** for RVs:
  \[ \mathbb{E}(|XY|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \]

- Recall that a real function \( g \) is convex if, for any \( x, y, \) and \( \alpha \in [0, 1] \)
  \[ g(\alpha x + (1 - \alpha)y) \leq \alpha g(x) + (1 - \alpha)g(y) \]

\[ \text{Jensen’s inequality: if } g \text{ is a real convex function, then} \]
\[ \mathbb{E}(g(X)) \geq g(\mathbb{E}(X)) \]

Examples: \( \mathbb{E}(X)^2 \leq \mathbb{E}(X^2) \Rightarrow \text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0. \)
\[ \mathbb{E}(\log X) \leq \log \mathbb{E}(X), \text{ for } X \text{ a positive RV.} \]
Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)$$

Positivity:

$H(X) \geq 0$;

$H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, \ldots, K\}$.

Upper bound:

$H(X) \leq \log K$;

$H(X) = \log K \iff f_X(x) = 1/K$, for all $x \in \{1, \ldots, K\}$.

Measure of uncertainty/randomness of $X$

Continuous RV $X$, differential entropy:

$h(X) = - \int f_X(x) \log f_X(x) \, dx$

$h(X)$ can be positive or negative. Example, if $f_X(x) = \text{Uniform}(x; a, b)$,

$h(X) = \log(b-a)$.

If $f_X(x) = N(x; \mu, \sigma^2)$, then

$h(X) = \frac{1}{2} \log(2 \pi e \sigma^2)$.

If $\text{var}(Y) = \sigma^2$, then

$h(Y) \leq \frac{1}{2} \log(2 \pi e \sigma^2)$.
Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$:

\[
H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)
\]

- **Positivity:** $H(X) \geq 0$;
  
  $H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$. 

Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity:** $H(X) \geq 0$;
  $$H(X) = 0 \iff f_X(i) = 1, \text{ for exactly one } i \in \{1, \ldots, K\}.$$

- **Upper bound:** $H(X) \leq \log K$;
  $$H(X) = \log K \iff f_X(x) = 1/k, \text{ for all } x \in \{1, \ldots, K\}.$$
Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity:** $H(X) \geq 0$;
  $$H(X) = 0 \iff f_X(i) = 1, \text{ for exactly one } i \in \{1, \ldots, K\}.$$  

- **Upper bound:** $H(X) \leq \log K$;
  $$H(X) = \log K \iff f_X(x) = 1/k, \text{ for all } x \in \{1, \ldots, K\}$$

- **Measure of uncertainty/randomness** of $X$
Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity:** $H(X) \geq 0$;
  $H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, \ldots, K\}$.

- **Upper bound:** $H(X) \leq \log K$;
  $H(X) = \log K \iff f_X(x) = 1/k$, for all $x \in \{1, \ldots, K\}$

- Measure of uncertainty/randomness of $X$

Continuous RV $X$, differential entropy:

$$h(X) = - \int f_X(x) \log f_X(x) \, dx$$
Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity**: $H(X) \geq 0$;
  
  $H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, \ldots, K\}$.

- **Upper bound**: $H(X) \leq \log K$;
  
  $H(X) = \log K \iff f_X(x) = 1/k$, for all $x \in \{1, \ldots, K\}$

- **Measure of uncertainty/randomness of $X$**

Continuous RV $X$, differential entropy:

$$h(X) = - \int f_X(x) \log f_X(x) \, dx$$

- $h(X)$ can be positive or negative. Example, if $f_X(x) = \text{Uniform}(x; a, b)$, $h(X) = \log(b - a)$. 


Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity:** $H(X) \geq 0$; $H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, \ldots, K\}$.

- **Upper bound:** $H(X) \leq \log K$; $H(X) = \log K \iff f_X(x) = 1/k$, for all $x \in \{1, \ldots, K\}$

- Measure of uncertainty/randomness of $X$

Continuous RV $X$, differential entropy:

$$h(X) = -\int f_X(x) \log f_X(x) \, dx$$

- $h(X)$ can be positive or negative. Example, if $f_X(x) = \text{Uniform}(x; a, b)$, $h(X) = \log(b - a)$.

- If $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$, then $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$. 
Entropy and all that...

Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

$$H(X) = - \sum_{x=1}^{K} f_X(x) \log f_X(x)$$

- **Positivity:** $H(X) \geq 0$;
  
  $H(X) = 0 \iff f_X(i) = 1$, for exactly one $i \in \{1, \ldots, K\}$.

- **Upper bound:** $H(X) \leq \log K$;
  
  $H(X) = \log K \iff f_X(x) = 1/k$, for all $x \in \{1, \ldots, K\}$

- **Measure of uncertainty/randomness** of $X$

Continuous RV $X$, differential entropy:

$$h(X) = - \int f_X(x) \log f_X(x) \, dx$$

- $h(X)$ can be positive or negative. Example, if $f_X(x) = \text{Uniform}(x; a, b)$, $h(X) = \log(b - a)$.

- If $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$, then $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$.

- If $\text{var}(Y) = \sigma^2$, then $h(Y) \leq \frac{1}{2} \log(2\pi e \sigma^2)$.
Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

\[
D(f_X \parallel g_X) = \sum_{x=1}^{K} f_X(x) \log \frac{f_X(x)}{g_X(x)}
\]

Positivity:

\[ D(f_X \parallel g_X) \geq 0 \]

\[ D(f_X \parallel g_X) = 0 \iff f_X(x) = g_X(x), \text{ for } x \in \{1, \ldots, K\} \]
Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

\[
D(f_X \parallel g_X) = \sum_{x=1}^{K} f_X(x) \log \frac{f_X(x)}{g_X(x)}
\]

Positivity:

\[
D(f_X \parallel g_X) \geq 0 \\
D(f_X \parallel g_X) = 0 \iff f_X(x) = g_X(x), \text{ for } x \in \{1, \ldots, K\}
\]
Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

\[
D(f_X \| g_X) = \sum_{x=1}^{K} f_X(x) \log \frac{f_X(x)}{g_X(x)}
\]

Positivity: \( D(f_X \| g_X) \geq 0 \)
\( D(f_X \| g_X) = 0 \Leftrightarrow f_X(x) = g_X(x), \text{ for } x \in \{1, \ldots, K\} \)

KLD between two pdf:

\[
D(f_X \| g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} \, dx
\]
Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

\[
D(f_X \parallel g_X) = \sum_{x=1}^{K} f_X(x) \log \frac{f_X(x)}{g_X(x)}
\]

Positivity: 
\[
D(f_X \parallel g_X) \geq 0
\]
\[
D(f_X \parallel g_X) = 0 \iff f_X(x) = g_X(x), \quad \text{for } x \in \{1, \ldots, K\}
\]

KLD between two pdf:

\[
D(f_X \parallel g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)} \, dx
\]

Positivity: 
\[
D(f_X \parallel g_X) \geq 0
\]
\[
D(f_X \parallel g_X) = 0 \iff f_X(x) = g_X(x), \quad \text{almost everywhere}
\]
Mutual information

Mutual information (MI) between two random variables:

\[ I(X;Y) = D(f_{X,Y} \parallel f_X f_Y) \]
Mutual information

Mutual information (MI) between two random variables:

\[ I(X; Y) = D(f_{X,Y} \parallel f_X f_Y) \]

Positivity: \[ I(X; Y) \geq 0 \]
\[ I(X; Y) = 0 \iff X, Y \text{ are independent.} \]
Mutual information

Mutual information (MI) between two random variables:

\[ I(X; Y) = D(f_{X,Y} \parallel f_X f_Y) \]

Positivity: \( I(X; Y) \geq 0 \)

\( I(X; Y) = 0 \iff X, Y \text{ are independent.} \)

MI is a measure of dependency between two random variables.
Note covered, but also very important for machine learning:

- Exponential families,
- Basic inequalities (Markov, Chebyshev, etc...)
- Stochastic processes (Markov chains, hidden Markov models,...)