Review of Probability Theory

Mário A. T. Figueiredo

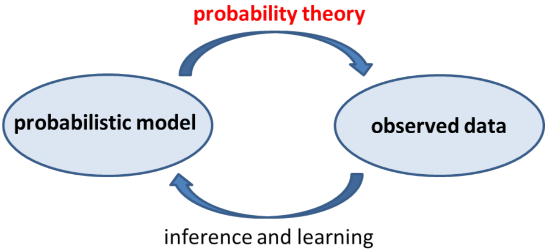
Instituto Superior Técnico & Instituto de Telecomunicações

Lisboa, **Portugal**

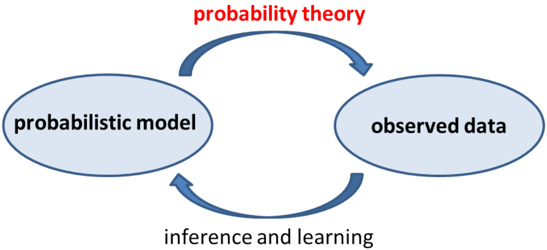
LxMLS 2015: Lisbon Machine Learning School

July 16, 2015

Probability theory

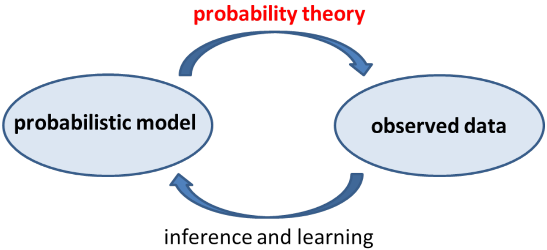


Probability theory



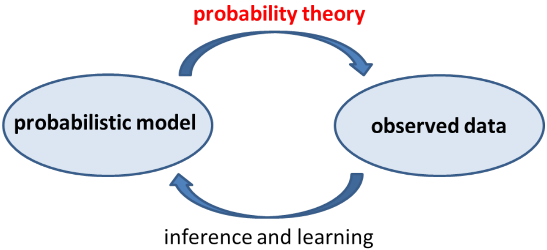
- The study of probability has roots in games of chance (dice, cards, ...)

Probability theory



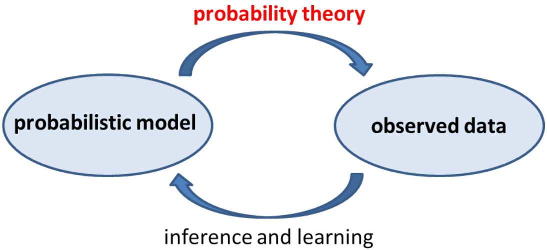
- The study of probability has roots in games of chance (dice, cards, ...)
- Great names of science: Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzman, de Finetti, ...

Probability theory



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

Probability theory



- The study of probability has roots in games of chance (dice, cards, ...)
- Great names of science: Cardano, Fermat, Pascal, Laplace, Kolmogorov, Bernoulli, Poisson, Cauchy, Boltzman, de Finetti, ...
- Natural tool to model uncertainty, information, knowledge, belief, ...
- ...thus also learning, decision making, inference, ...

What is probability?

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

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of A. *Laplace, 1814*

  Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

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

What is probability?

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

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of A. *Laplace, 1814*

  Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

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

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

  ...relative frequency of occurrence of A in infinite number of trials.

What is probability?

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

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of A. *Laplace, 1814*

  Example: $\mathbb{P}(\text{randomly drawn card is } \clubsuit) = 13/52$.

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

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

  ...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: $\mathcal{X} = \{A\clubsuit, 2\clubsuit, ..., Q\diamondsuit, K\diamondsuit\}$.

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: $\mathcal{X} = \{A\clubsuit, 2\clubsuit, ..., Q\diamondsuit, K\diamondsuit\}$.

- 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 $\heartsuit$ card": $C = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\} \subset \{A\clubsuit, ..., K\diamondsuit\}$

Key concepts: Sample space and events

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

  (More delicate) examples:
  - Time until you receive the next email: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop: $\mathcal{X} = \mathbb{R}^2$

Key concepts: Sample space and events

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

  (More delicate) examples:

  - Time until you receive the next email: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop: $\mathcal{X} = \mathbb{R}^2$

- An event A is a measurable subset of $\mathcal{X}$, *i.e.*, $A \in \sigma(F)$, $F \in 2^{\mathcal{X}}$.
  σ-algebra: $\sigma(F)$

  - $A \in \sigma(F) \Rightarrow A^c \in \sigma(F)$
  - $A_1, A_2, ... \in \sigma(F) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \sigma(F)$

Key concepts: Sample space and events

- Sample space $\mathcal{X}$ = set of possible outcomes of a random experiment.
  (More delicate) examples:
  - Time until you receive the next email: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop: $\mathcal{X} = \mathbb{R}^2$
- An event A is a measurable subset of $\mathcal{X}$, *i.e.*, $A \in \sigma(F)$, $F \in 2^{\mathcal{X}}$.
  σ-algebra: $\sigma(F)$
  - $A \in \sigma(F) \Rightarrow A^c \in \sigma(F)$
  - $A_1, A_2, ... \in \sigma(F) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \sigma(F)$
- Classical example in $\mathbb{R}^n$: the collection of Lebesgue measurable sets constitute a $\sigma-$algebra.

Key concepts: Sample space and events

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

  (More delicate) examples:

  - Time until you receive the next email: $\mathcal{X} = \mathbb{R}_+$
  - Location of the next rain drop: $\mathcal{X} = \mathbb{R}^2$

- An event A is a measurable subset of $\mathcal{X}$, *i.e.*, $A \in \sigma(F)$, $F \in 2^{\mathcal{X}}$.
  σ-algebra: $\sigma(F)$

  - $A \in \sigma(F) \Rightarrow A^c \in \sigma(F)$
  - $A_1, A_2, \ldots \in \sigma(F) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \sigma(F)$

- Classical example in $\mathbb{R}^n$: the collection of Lebesgue measurable sets constitute a $\sigma-$algebra.

- For any $F \in 2^{\mathcal{X}}$, $\quad \emptyset \in \sigma(F)$, $\quad \mathcal{X} \in \sigma(F)$.

Kolmogorov's Axioms for Probability

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

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

Kolmogorov's Axioms for Probability

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

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

  - 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 $\mathbb{P} : \sigma(F) \rightarrow [0,\ 1]$

  - 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 $\mathbb{P} : \sigma(F) \rightarrow [0,\ 1]$

  - 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

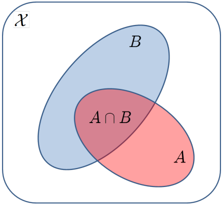
- Probability is a function that maps events A into the interval $[0,1]$.

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

  - 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 $\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i)$

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

  - $\mathbb{P}(\emptyset) = 0$
  - $C \subset D \;\Rightarrow\; \mathbb{P}(C) \leq \mathbb{P}(D)$
  - $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ (union bound)



Conditional Probability and Independence

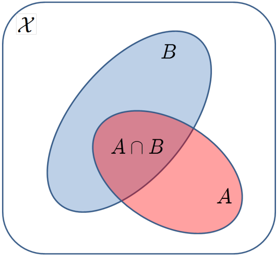
- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (conditional prob. of A, given B)

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (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 $A_1, A_2, \ldots \subseteq \mathcal{X}$ are disjoint, then $\mathbb{P}\Big(\bigcup_i A_i \Big| B\Big) = \sum_i \mathbb{P}(A_i|B)$

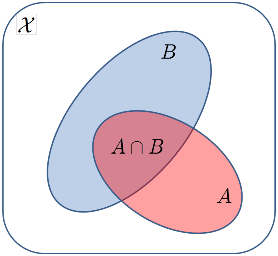


Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (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 $\mathbb{P}\Big(\bigcup_i A_i \Big| B\Big) = \sum_i \mathbb{P}(A_i|B)$



- Independence: A, B are independent (denoted $A \perp\!\!\!\perp B$) if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B).$$

Conditional Probability and Independence

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

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent $(A \perp\!\!\!\perp B) \;\Leftrightarrow\; \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent $(A \perp\!\!\!\perp B) \;\Leftrightarrow\; \mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \;\Leftrightarrow\; \mathbb{P}(A|B) = \mathbb{P}(A)$$

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent ($A \perp\!\!\!\perp B$) $\Leftrightarrow$ $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \;\Leftrightarrow\; \mathbb{P}(A|B) = \mathbb{P}(A)$$

- Example: $\mathcal{X} =$ "52 cards", $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\clubsuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; 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 $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

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

- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \;\Leftrightarrow\; \mathbb{P}(A|B) = \mathbb{P}(A)$$

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

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52} \\ \mathbb{P}(A)\,\mathbb{P}(B) &= \frac{1}{13}\,\frac{1}{4} = \frac{1}{52} \end{aligned}$$

Conditional Probability and Independence

- If $\mathbb{P}(B) > 0$, $\quad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- Events A, B are independent ($A \perp\!\!\!\perp B$) $\Leftrightarrow$ $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.
- Relationship with conditional probabilities:

$$A \perp\!\!\!\perp B \;\Leftrightarrow\; \mathbb{P}(A|B) = \mathbb{P}(A)$$

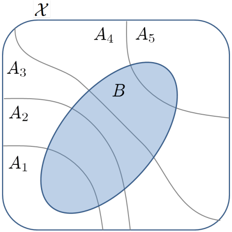
- Example: $\mathcal{X} =$ "52 cards", $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\clubsuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

$$\begin{aligned}
\mathbb{P}(A \cap B) &= \mathbb{P}(\{3\heartsuit\}) = \frac{1}{52} \\
\mathbb{P}(A)\,\mathbb{P}(B) &= \frac{1}{13}\,\frac{1}{4} = \frac{1}{52} \\
\mathbb{P}(A|B) &= \mathbb{P}(\text{"3"}\,|\,\text{"}\heartsuit\text{"}) = \frac{1}{13} = \mathbb{P}(A)
\end{aligned}$$

Bayes Theorem

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

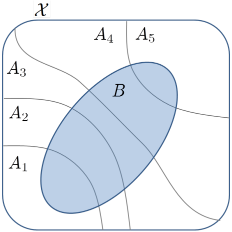
$$\begin{aligned}\mathbb{P}(B) &= \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)\\ &= \sum_i \mathbb{P}(B \cap A_i)\end{aligned}$$



Bayes Theorem

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

$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$
$$= \sum_i \mathbb{P}(B \cap A_i)$$

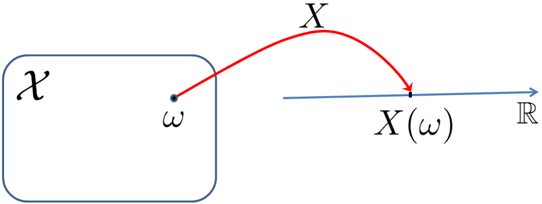


- Bayes' theorem: if $\{A_1, ..., A_n\}$ is a partition of $\mathcal{X}$

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_i)\,\mathbb{P}(A_i)}{\sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

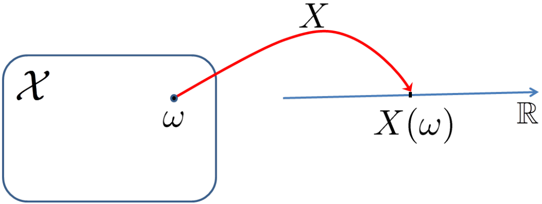
Random Variables

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



Random Variables

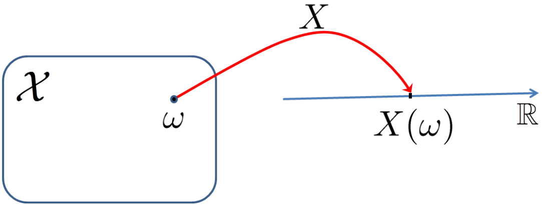
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



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

Random Variables

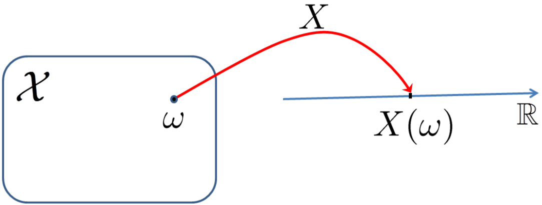
- A (real) random variable (RV) is a function: $X : \mathcal{X} \to \mathbb{R}$



  - 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])

Random Variables

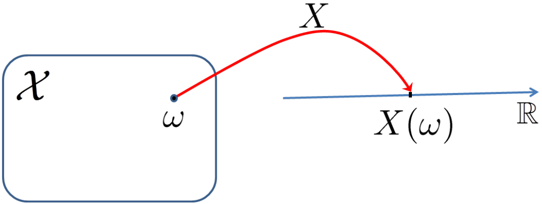
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



  - 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

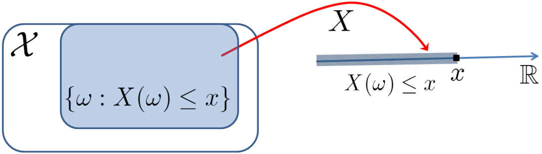
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



  - 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 $X = \mathbb{R}_+$.

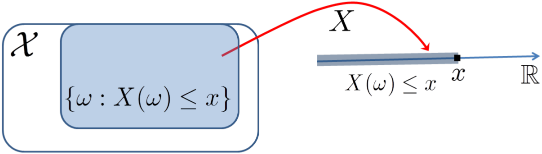
Random Variables: Distribution Function

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

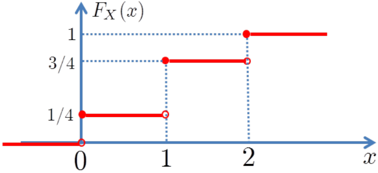


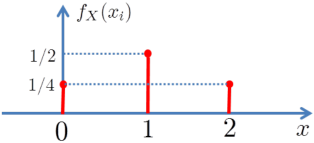
Random Variables: Distribution Function

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



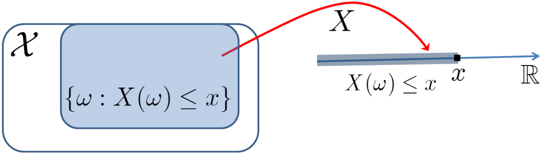
- Example: number of heads in tossing 2 coins; range$(X) = \{0, 1, 2\}$.



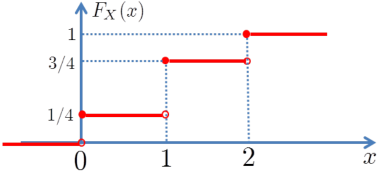


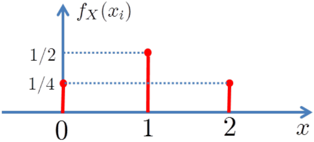
Random Variables: Distribution Function

- 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:

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: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;

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: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- $\lim\limits_{x \to -\infty} F_X(x) = 0$;

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: $x_1 < x_2 \Rightarrow F_X(x_1) \leq F_X(x_2)$;
- $\displaystyle\lim_{x \to -\infty} F_X(x) = 0$;
- $\displaystyle\lim_{x \to +\infty} F_X(x) = 1$;

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

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

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

Further properties:

- $\mathbb{P}(X = x) = f_X(x) = F_X(x) - \lim\limits_{z \to x^-} F_X(z)$;
- $\mathbb{P}(z < X \leq y) = F_X(y) - F_X(z)$;

Properties of Distribution Functions

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

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

Further properties:

- $\mathbb{P}(X = x) = f_X(x) = F_X(x) - \lim\limits_{z \to x^-} F_X(z)$;
- $\mathbb{P}(z < X \leq y) = F_X(y) - F_X(z)$;
- $\mathbb{P}(X > x) = 1 - F_X(x)$.

Important Discrete Random Variables

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

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 $f_X(x) = \left\{ \begin{array}{ccc} p & \Leftarrow & x = 1 \\ 1-p & \Leftarrow & x = 0 \end{array} \right.$

  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 $f_X(x) = \begin{cases} p & \Leftarrow \quad x = 1 \\ 1 - p & \Leftarrow \quad x = 0 \end{cases}$

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

- Binomial RV: $X \in \{0, 1, ..., n\}$ (sum on 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 $f_X(x) = \begin{cases} p & \Leftarrow \quad x = 1 \\ 1-p & \Leftarrow \quad x = 0 \end{cases}$

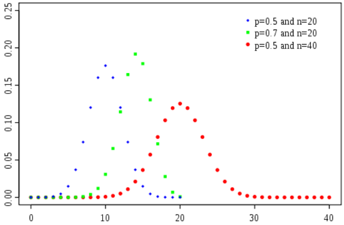
  Can be written compactly as $f_X(x) = p^x(1-p)^{1-x}$.

- Binomial RV: $X \in \{0, 1, ..., n\}$ (sum on 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"):

$$\binom{n}{x} = \frac{n!}{(n-x)!\,x!}$$



More 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).

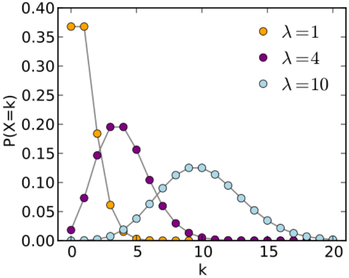
More Important Discrete Random Variables

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

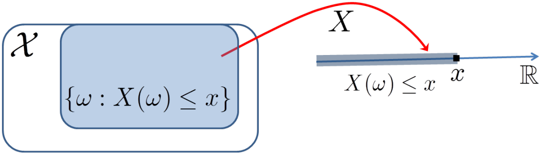
  Notice that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$, thus $\sum_{x=0}^{\infty} f_X(x) = 1$.

  "...probability of the number of independent occurrences in a fixed (time/space) interval if these occurrences have known average rate"



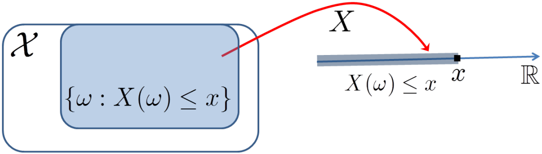
Random Variables: Distribution Function

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

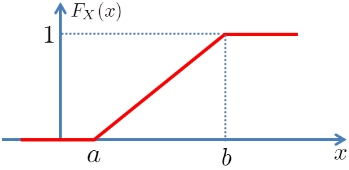


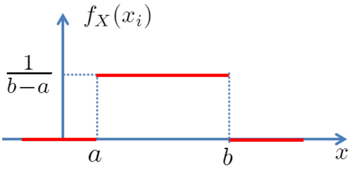
Random Variables: Distribution Function

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



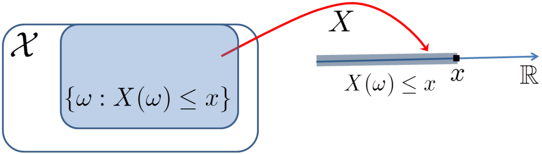
- Example: continuous RV with uniform distribution on $[a, b]$.



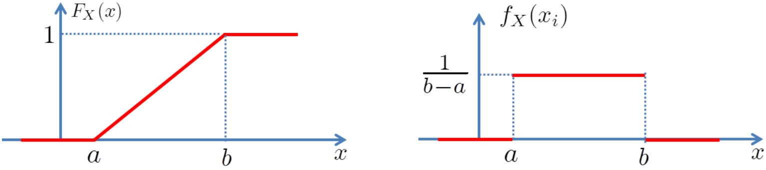


Random Variables: Distribution Function

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



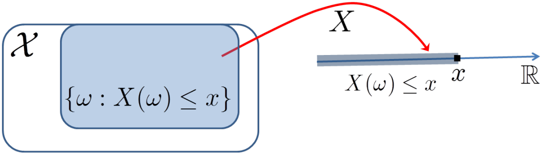
- Example: continuous RV with uniform distribution on $[a, b]$.



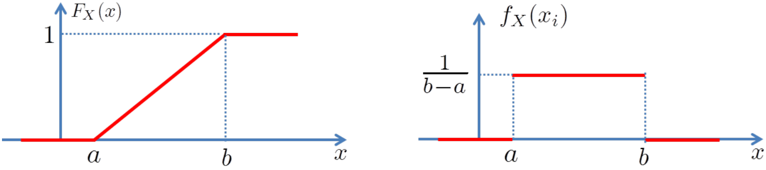
- Probability density function (pdf, continuous RV): $f_X(x)$

Random Variables: Distribution Function

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



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

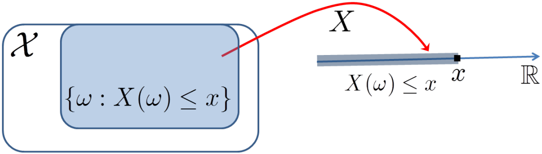


- Probability density function (pdf, continuous RV): $f_X(x)$

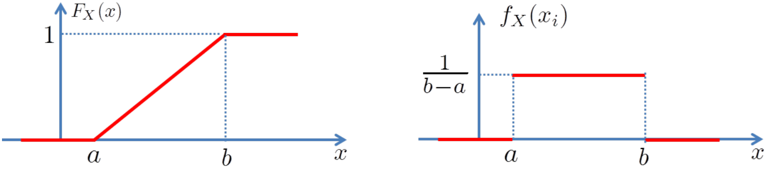
$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du,$$

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 $[a, b]$.

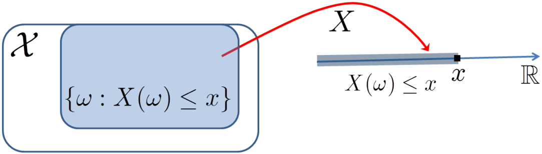


- Probability density function (pdf, continuous RV): $f_X(x)$

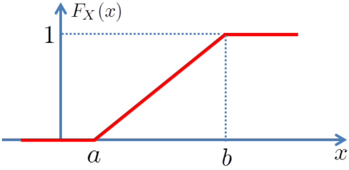
$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du, \quad \mathbb{P}(X \in [c, d]) = \int_{c}^{d} f_X(x)\, dx,$$

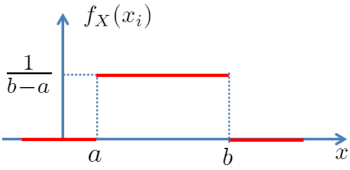
Random Variables: Distribution Function

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



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





- 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

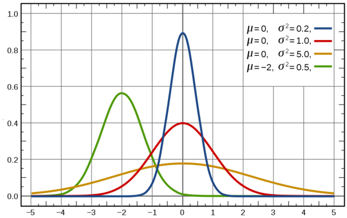
- Uniform: $f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \Leftarrow \quad x \in [a,\ b] \\ 0 & \Leftarrow \quad x \notin [a,\ b] \end{cases}$ (previous slide).

Important Continuous Random Variables

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

  (previous slide).

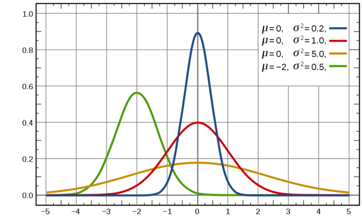
- Gaussian: $f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \dfrac{1}{\sqrt{2\,\pi\,\sigma^2}} e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$

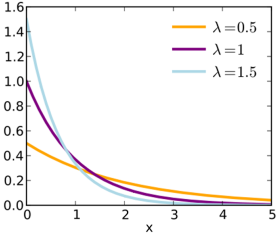


Important Continuous Random Variables

- Uniform: $f_X(x) = \text{Uniform}(x; a, b) = \begin{cases} \frac{1}{b-a} & \Leftarrow \quad x \in [a,\ b] \\ 0 & \Leftarrow \quad x \notin [a,\ b] \end{cases}$
  (previous slide).

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





- Exponential: $f_X(x) = \text{Exp}(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \Leftarrow \quad x \geq 0 \\ 0 & \Leftarrow \quad x < 0 \end{cases}$

Expectation of Random Variables

- Expectation: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i\, f_X(x_i) & X \in \{x_1, ...x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x\, f_X(x)\, dx & X \text{ continuous} \end{cases}$

Expectation of Random Variables

- Expectation: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i \, f_X(x_i) & X \in \{x_1, ...x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x \, f_X(x) \, dx & X \text{ continuous} \end{cases}$

- 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: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i \, f_X(x_i) & X \in \{x_1, ... x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x \, f_X(x) \, dx & X \text{ continuous} \end{cases}$

- Example: Bernoulli, $f_X(x) = p^x \, (1-p)^{1-x}$, for $x \in \{0, 1\}$.

  $\mathbb{E}(X) = 0\,(1-p) + 1\,p = p.$

- Example: Binomial, $f_X(x) = \binom{n}{x} p^x \, (1-p)^{n-x}$, for $x \in \{0, ..., n\}$.

  $\mathbb{E}(X) = n\,p.$

Expectation of Random Variables

- Expectation: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i\, f_X(x_i) & X \in \{x_1, ...x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x\, f_X(x)\, dx & X \text{ continuous} \end{cases}$

- Example: Bernoulli, $f_X(x) = p^x\,(1-p)^{1-x}$, for $x \in \{0,\, 1\}$.

  $\mathbb{E}(X) = 0\,(1-p) + 1\,p = p.$

- Example: Binomial, $f_X(x) = \binom{n}{x} p^x\,(1-p)^{n-x}$, for $x \in \{0, ..., n\}$.

  $\mathbb{E}(X) = n\,p.$

- Example: Gaussian, $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$. $\quad \mathbb{E}(X) = \mu.$

Expectation of Random Variables

- Expectation: $\mathbb{E}(X) = \begin{cases} \displaystyle\sum_i x_i \, f_X(x_i) & X \in \{x_1, ...x_K\} \subset \mathbb{R} \\ \displaystyle\int_{-\infty}^{\infty} x \, f_X(x) \, dx & X \text{ continuous} \end{cases}$

- Example: Bernoulli, $f_X(x) = p^x \, (1-p)^{1-x}$, for $x \in \{0, 1\}$.

  $$\mathbb{E}(X) = 0 \, (1-p) + 1 \, p = p.$$

- Example: Binomial, $f_X(x) = \binom{n}{x} p^x \, (1-p)^{n-x}$, for $x \in \{0, ..., n\}$.

  $$\mathbb{E}(X) = n \, p.$$

- Example: Gaussian, $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$. $\quad \mathbb{E}(X) = \mu$.

- Linearity of expectation:
  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$; $\quad \mathbb{E}(\alpha \, X) = \alpha \mathbb{E}(X)$, $\; \alpha \in \mathbb{R}$

Expectation of Functions of Random Variables

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

Expectation of Functions of Random Variables

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

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2 \Big)$

Expectation of Functions of Random Variables

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

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

- Example: Bernoulli variance, $\mathbb{E}(X^2) = \mathbb{E}(X) = p$

Expectation of Functions of Random Variables

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

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

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

Expectation of Functions of Random Variables

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

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

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

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

Expectation of Functions of Random Variables

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

- Example: variance, $\text{var}(X) = \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)^2\Big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$

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

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

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

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, dx = \int \mathbf{1}_A(x)\, f_X(x)\, dx = \mathbb{E}(\mathbf{1}_A(X))$$

Moments of Random Variables

- Non-central moments of order k:

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

Moments of Random Variables

- Non-central moments of order k:

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

- Central moments of order $k \in \mathbb{N}$: $\mathbb{E}\big(|X - \mathbb{E}(X)|^k\big)$.

Moments of Random Variables

- Non-central moments of order k:

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

- Central moments of order $k \in \mathbb{N}$: $\mathbb{E}\big(|X - \mathbb{E}(X)|^k\big)$.
- ...if the integral/sum exits.

Moments of Random Variables

- Non-central moments of order k:

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

- Central moments of order $k \in \mathbb{N}$: $\mathbb{E}\big(|X - \mathbb{E}(X)|^k\big)$.
- ...if the integral/sum exits.
- If k-th moment exists, then the j-th moment exists, for $j \leq k$.

$$\begin{aligned} \int_{-\infty}^{\infty} |x|^j\, f_X(x)\, dx &= \int_{|x|\leq 1} |x|^j\, f_X(x)\, dx + \int_{|x|>1} |x|^j\, f_X(x)\, dx \\ &\leq \int_{|x|\leq 1} f_X(x)\, dx + \int_{|x|>1} |x|^k\, f_X(x)\, dx \leq 1 + \mathbb{E}\big(|X|^k\big) < \infty \end{aligned}$$

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

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

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\, dx\, dy, \qquad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

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

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\,dx\,dy, \qquad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.

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

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

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

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\,dx\,dy, \qquad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.

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

- Independence:

$$X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x,y) = f_X(x)\,f_Y(y) \qquad .$$

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

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

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\, dx\, dy, \qquad A \in \sigma(\mathbb{R}^2)$$

  Extends trivially to more than two RVs.

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

- Independence:

$$X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x,y) = f_X(x)\, f_Y(y) \underset{\not\Leftarrow}{\Rightarrow} \mathbb{E}(X\,Y) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

Conditionals and Bayes' Theorem

- Conditional pmf (discrete RVs):

$$f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x \wedge Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Conditionals and Bayes' Theorem

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

- Conditional pdf (continuous RVs): $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

  ...the meaning is technically delicate.

Conditionals and Bayes' Theorem

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

- Conditional pdf (continuous RVs): $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

  ...the meaning is technically delicate.

- Bayes' theorem: $f_{X|Y}(x|y) = \dfrac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$ (pdf or pmf).

Conditionals and Bayes' Theorem

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

- Conditional pdf (continuous RVs): $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

  ...the meaning is technically delicate.

- Bayes' theorem: $f_{X|Y}(x|y) = \dfrac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$ (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:

$f_{X,Y}(x, y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

Joint, Marginal, and Conditional Probabilities: An Example

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

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

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

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

Joint, Marginal, and Conditional Probabilities: An Example

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

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	1/5	2/5
$X = 1$	1/10	3/10

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

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

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

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

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

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

An Important Multivariate RV: Multinomial

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

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

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

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

- Generalizes the binomial from binary to K-classes.

An Important Multivariate RV: Multinomial

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

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

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

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

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

An Important Multivariate RV: Gaussian

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

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

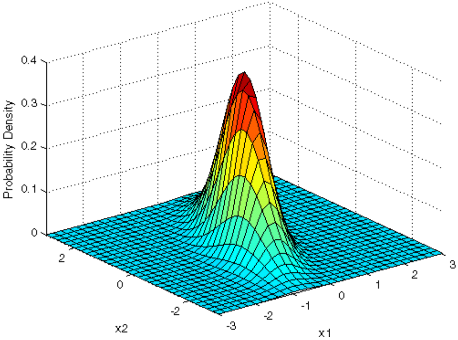
- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n \times n}$.
  Expected value: $\mathbb{E}(X) = \mu$. Meaning of C: next slide.

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:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \; = \; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

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

- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

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

- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y) \;\underset{\not\Leftarrow}{\Rightarrow}\; \text{cov}(X,\, Y) = 0.$

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

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

- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y) \;\substack{\Rightarrow \\ \not\Leftarrow}\; \text{cov}(X,\; Y) = 0.$

- Covariance matrix of multivariate RV, $X \in \mathbb{R}^n$:

$$\text{cov}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^T\Big] = \mathbb{E}(X\,X^T) - \mathbb{E}(X)\mathbb{E}(X)^T$$

Covariance, Correlation, and all that...

- Covariance between two RVs:

$$\text{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\,\big(Y - \mathbb{E}(Y)\big)\Big] \;=\; \mathbb{E}(X\,Y) - \mathbb{E}(X)\,\mathbb{E}(Y)$$

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

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

- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y) \;\underset{\not\Leftarrow}{\Rightarrow}\; \text{cov}(X,\; Y) = 0.$

- Covariance matrix of multivariate RV, $X \in \mathbb{R}^n$:

$$\text{cov}(X) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(X - \mathbb{E}(X)\big)^T\Big] = \mathbb{E}(X\,X^T) - \mathbb{E}(X)\mathbb{E}(X)^T$$

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

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

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 $\text{cov}(X) = C$ 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 $\text{cov}(Y) = I$;

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 $\text{cov}(X) = C$ 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 $\text{cov}(Y) = I$;
- If $f_X(x) = \mathcal{N}(x; 0, I)$ and $Y = \mu + C^{1/2}X$, then $f_Y(y) = \mathcal{N}(y; \mu, C)$;

More on Expectations and Covariances

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

- If $\mathbb{E}(X) = \mu$ and $Y = AX$, then $\mathbb{E}(Y) = A\mu$;
- 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 $\text{cov}(Y) = I$;
- If $f_X(x) = \mathcal{N}(x; 0, I)$ and $Y = \mu + C^{1/2}X$, then $f_Y(y) = \mathcal{N}(y; \mu, C)$;
- If $f_X(x) = \mathcal{N}(x; \mu, C)$ and $Y = C^{-1/2}(X - \mu)$, then $f_Y(y) = \mathcal{N}(y; 0, I)$.

More on Multivariate Gaussians

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

More on Multivariate Gaussians

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

Partition $X = [\underbrace{X_1, ..., X_i}_{X_a}, \underbrace{X_{i+1}, ..., X_n}_{X_b}]^T$ and $\mu = [\underbrace{\mu_1, ..., \mu_i}_{\mu_a}, \underbrace{\mu_{i+1}, ..., \mu_n}_{\mu_b}]^T$;

More on Multivariate Gaussians

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

Partition $X = [\underbrace{X_1, ..., X_i}_{X_a}, \underbrace{X_{i+1}, ..., X_n}_{X_b}]^T$ and $\mu = [\underbrace{\mu_1, ..., \mu_i}_{\mu_a}, \underbrace{\mu_{i+1}, ..., \mu_n}_{\mu_b}]^T$;

Corresponding partition of C: $C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix}$

More on Multivariate Gaussians

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

Partition $X = [\underbrace{X_1, ..., X_i}_{X_a}, \underbrace{X_{i+1}, ..., X_n}_{X_b}]^T$ and $\mu = [\underbrace{\mu_1, ..., \mu_i}_{\mu_a}, \underbrace{\mu_{i+1}, ..., \mu_n}_{\mu_b}]^T$;

Corresponding partition of C: $C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix}$

- $f_{X_a}(\cdot) = \mathcal{N}(\cdot; \mu_a, C_{aa})$;

More on Multivariate Gaussians

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

Partition $X = [\underbrace{X_1, ..., X_i}_{X_a}, \underbrace{X_{i+1}, ..., X_n}_{X_b}]^T$ and $\mu = [\underbrace{\mu_1, ..., \mu_i}_{\mu_a}, \underbrace{\mu_{i+1}, ..., \mu_n}_{\mu_b}]^T$;

Corresponding partition of C: $C = \begin{bmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{bmatrix}$

- $f_{X_a}(\cdot) = \mathcal{N}(\cdot; \mu_a, C_{aa})$;
- $f_{X_b}(\cdot) = \mathcal{N}(\cdot; \mu_b, C_{bb})$;

More on Multivariate Gaussians

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

Partition $X = [\underbrace{X_1, ..., X_i}_{X_a}, \underbrace{X_{i+1}, ..., X_n}_{X_b}]^T$ and $\mu = [\underbrace{\mu_1, ..., \mu_i}_{\mu_a}, \underbrace{\mu_{i+1}, ..., \mu_n}_{\mu_b}]^T$;

Corresponding partition of C: $C = \left[\begin{array}{cc} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{array}\right]$

- $f_{X_a}(\cdot) = \mathcal{N}(\cdot; \mu_a, C_{aa})$;
- $f_{X_b}(\cdot) = \mathcal{N}(\cdot; \mu_b, C_{bb})$;
- $f_{X_b|X_a}(x_b|x_a) = \mathcal{N}\big(x_b; \mu_b + C_{ba}C_{aa}^{-1}(x_a - \mu_a), C_{bb} - C_{ba}C_{aa}^{-1}C_{ab}\big)$

Exponential Families

Consider a pdf or pmf $f_X(x|\theta)$, with parameter(s) θ, for $X \in \mathcal{X}^n$

Exponential Families

Consider a pdf or pmf $f_X(x|\theta)$, with parameter(s) θ, for $X \in \mathcal{X}^n$

$f_X(x|\theta)$ is in an exponential family if it has the form

$$f_X(x|\theta) = \frac{1}{Z(\theta)}\, h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)$$

where $A(\theta) = \log Z(\theta)$ and

$$Z(\theta) = \int_{\mathcal{X}^n} h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)\, dx.$$

Exponential Families

Consider a pdf or pmf $f_X(x|\theta)$, with parameter(s) θ, for $X \in \mathcal{X}^n$

$f_X(x|\theta)$ is in an exponential family if it has the form

$$f_X(x|\theta) = \frac{1}{Z(\theta)}\, h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big) = h(x)\, \exp\big(\eta(\theta)^T \phi(x) - A(\theta)\big),$$

where $A(\theta) = \log Z(\theta)$ and

$$Z(\theta) = \int_{\mathcal{X}^n} h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)\, dx.$$

- Canonical parameter(s): $\eta(\theta)$
- Sufficient statistics: $\phi(x)$
- Partition function: $Z(\theta)$
- Curved exponential family: $\dim(\theta) < \dim(\eta(\theta))$

Exponential Families

$$f_X(x|\theta) = \frac{1}{Z(\theta)}\, h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)$$

Examples:

Exponential Families

$$f_X(x|\theta) = \frac{1}{Z(\theta)}\, h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)$$

Examples:

- Bernoulli: $f_X(x) = p^x (1-p)^{1-x}$,

  $f_X(x) = \exp\big(x \log p + (1-x)\log(1-p)\big) = \exp\big(x \log \frac{p}{1-p} + \log(1-p)\big)$,

  thus $\eta(p) = \log \frac{p}{1-p}$, $\phi(x) = x$, $Z(p) = \frac{1}{1-p}$, and $h(x) = 1$.

Exponential Families

$$f_X(x|\theta) = \frac{1}{Z(\theta)}\, h(x)\, \exp\big(\eta(\theta)^T \phi(x)\big)$$

Examples:

- Bernoulli: $f_X(x) = p^x (1-p)^{1-x}$,

  $f_X(x) = \exp\big(x \log p + (1-x)\log(1-p)\big) = \exp\big(x \log \frac{p}{1-p} + \log(1-p)\big)$,

  thus $\eta(p) = \log \frac{p}{1-p}$, $\phi(x) = x$, $Z(p) = \frac{1}{1-p}$, and $h(x) = 1$.

- Gaussian:

$$f_X(x) = \frac{\exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)}{\sqrt{2\pi\sigma^2}} = \frac{\exp\big(-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\big)}{\sqrt{2\pi\sigma^2}},$$

  thus $\eta(\mu, \sigma^2) = [\mu/\sigma^2,\ -1/(2\sigma^2)]^T$, $\phi(x) = [x,\ x^2]^T$,
  $Z(\mu, \sigma^2) = \sqrt{2\pi\sigma^2}\exp\big(\frac{\mu^2}{2\sigma^2}\big)$, and $h(x) = 1$.

More on Exponential Families

- Independent identically distributed (i.i.d.) observations:

$$X_1, ..., X_m \sim f_X(x) = \frac{1}{Z(\eta)} h(x) \exp\left(\eta^T \phi(x)\right)$$

then

$$f_{X_1,...,X_m}(x_1, ..., x_m) = \frac{1}{Z(\eta)^m} \left(\prod_{j=1}^{m} h(x_i) \right) \exp\left(\eta^T \sum_{j=1}^{m} \phi(x_j) \right)$$

More on Exponential Families

- Independent identically distributed (i.i.d.) observations:

$$X_1, ..., X_m \sim f_X(x) = \frac{1}{Z(\eta)} h(x) \exp\left(\eta^T \phi(x)\right)$$

then

$$f_{X_1,...,X_m}(x_1, ..., x_m) = \frac{1}{Z(\eta)^m} \left(\prod_{j=1}^{m} h(x_i)\right) \exp\left(\eta^T \sum_{j=1}^{m} \phi(x_j)\right)$$

- Expected sufficient statistics:

$$\frac{d \log Z(\eta)}{d\eta} = \frac{\frac{dZ(\eta)}{d\eta}}{Z(\eta)} = \frac{1}{Z(\eta)} \int \phi(x) h(x) \exp\left(\eta^T \phi(x)\right) dx = \mathbb{E}\left(\phi(X)\right)$$

Statistical Inference

- Scenario: observed RV Y, depends on unknown variable(s) X.
  Goal: given an observation $Y = y$, infer 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 (not an RV), but unknown;
  Bayesian: X is a random variable with pdf/pmf $f_X(x)$ (the prior);
  this prior expresses/formalizes 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 (not an RV), but unknown;
  Bayesian: X is a random variable with pdf/pmf $f_X(x)$ (the prior); this prior expresses/formalizes 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 (not an RV), but unknown;
  Bayesian: X is a random variable with pdf/pmf $f_X(x)$ (the prior); this prior expresses/formalizes knowledge about X.

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

- This in **not** statistical/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 (not an RV), but unknown;
  Bayesian: X is a random variable with pdf/pmf $f_X(x)$ (the prior); this prior expresses/formalizes knowledge about X.

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

- This in **not** statistical/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" $\widehat{x}$ 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" $\widehat{x}$ about X, given this information?
- Need to define "optimal": loss function: $L(\widehat{x}, x) \in \mathbb{R}_+$ measures "loss"/"cost" incurred by "guessing" $\widehat{x}$ 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" $\widehat{x}$ about X, given this information?
- Need to define "optimal": loss function: $L(\widehat{x}, x) \in \mathbb{R}_+$ measures "loss"/"cost" incurred by "guessing" $\widehat{x}$ if truth is x.
- The optimal Bayesian decision minimizes the expected loss:

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

where

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

Classical Statistical Inference Criteria

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

Classical Statistical Inference Criteria

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

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

Classical Statistical Inference Criteria

- Consider that $X \in \{1, ..., K\}$ (discrete/classification problem).
- Adopt the 0/1 loss: $L(\widehat{x}, x) = 0$, if $\widehat{x} = x$, and 1 otherwise.
- Optimal Bayesian decision:

$$
\begin{aligned}
\widehat{x}_{\text{Bayes}} &= \arg\min_{\widehat{x}} \sum_{x=1}^{K} L(\widehat{x}, x)\, f_{X|Y}(x|y) \\
&= \arg\min_{\widehat{x}} \left(1 - f_{X|Y}(\widehat{x}|y)\right) \\
&= \arg\max_{\widehat{x}} f_{X|Y}(\widehat{x}|y) \;\equiv\; \widehat{x}_{\text{MAP}}
\end{aligned}
$$

MAP = maximum a posteriori criterion.

Classical Statistical Inference Criteria

- Consider that $X \in \{1, ..., K\}$ (discrete/classification problem).
- Adopt the 0/1 loss: $L(\widehat{x}, x) = 0$, if $\widehat{x} = x$, and 1 otherwise.
- Optimal Bayesian decision:

$$\begin{aligned}
\widehat{x}_{\text{Bayes}} &= \arg\min_{\widehat{x}} \sum_{x=1}^{K} L(\widehat{x}, x)\, f_{X|Y}(x|y) \\
&= \arg\min_{\widehat{x}} \left(1 - f_{X|Y}(\widehat{x}|y)\right) \\
&= \arg\max_{\widehat{x}} f_{X|Y}(\widehat{x}|y) \;\equiv\; \widehat{x}_{\text{MAP}}
\end{aligned}$$

MAP = maximum a posteriori criterion.

- Same criterion can be derived for continuous X, using $\lim_{\varepsilon \to 0} L_\varepsilon(\widehat{x}, x)$, where $L_\varepsilon(\widehat{x}, x) = 0$, if $|\widehat{x} - x| < \varepsilon$, and 1 otherwise.

Classical Statistical Inference Criteria

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

Classical Statistical Inference Criteria

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

- From Bayes law:

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

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

Classical Statistical Inference Criteria

- Consider the MAP criterion $\widehat{x}_{\text{MAP}} = \arg\max_x f_{X|Y}(x|y)$
- From Bayes law:

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

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

- Also common to write: $\widehat{x}_{\text{MAP}} = \arg\max_x \left(\log f_{Y|X}(y|x) + \log f_X(x)\right)$

Classical Statistical Inference Criteria

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

- From Bayes law:

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

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

- Also common to write: $\widehat{x}_{\text{MAP}} = \arg\max_x \big(\log f_{Y|X}(y|x) + \log f_X(x)\big)$

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

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

ML = maximum likelihood criterion.

Statistical Inference: Example

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs:
  $Y = (Y_1, ..., Y_n)$, 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: $f_{Y|X}\big(y_1, ..., y_n|x\big) = \displaystyle\prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i}$

  Log-likelihood function:

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

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

  Log-likelihood function:

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

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

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

  Log-likelihood function:

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

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

- Example: $n = 10$, observed $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$, $\widehat{x}_{\mathrm{ML}} = 7/10$.

Statistical Inference: Example (Continuation)

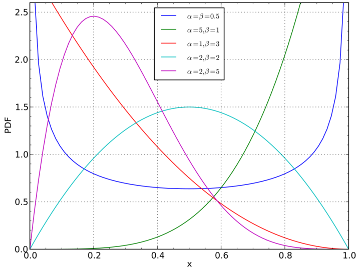
- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.

Statistical Inference: Example (Continuation)

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}\big(y_1, ..., y_n|x\big) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$

Statistical Inference: Example (Continuation)

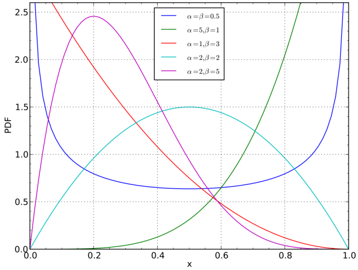
- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}\left(y_1, ..., y_n|x\right) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$
- How to express knowledge that (e.g.) X is around $1/2$? Convenient choice: conjugate prior. Form of the posterior = form of the prior.



Statistical Inference: Example (Continuation)

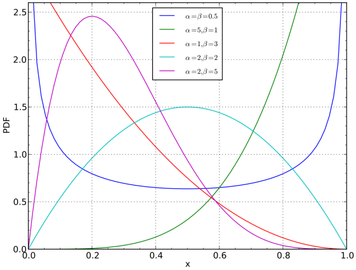
- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$
- How to express knowledge that (e.g.) X is around 1/2? Convenient choice: conjugate prior. Form of the posterior = form of the prior.

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



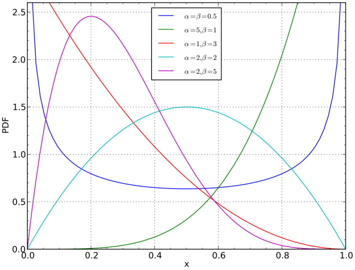
Statistical Inference: Example (Continuation)

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$
- 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 $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \;\; \alpha, \beta > 0$
  - 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: $f_{Y|X}(y_1,...,y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$
- 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
    $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \;\; \alpha, \beta > 0$
  - Posterior:
    $f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i}(1-x)^{\beta-1+n-\sum_i y_i}$
  - MAP: $\widehat{x}_{\text{MAP}} = \frac{\alpha+\sum_i y_i - 1}{\alpha+\beta+n-2}$



Statistical Inference: Example (Continuation)

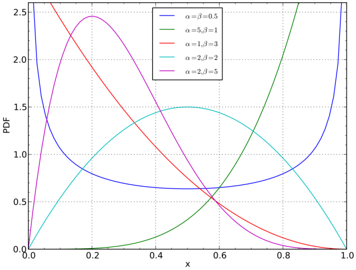
- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.

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

- 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
  $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \;\; \alpha, \beta > 0$
  - Posterior:
  $f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i}(1-x)^{\beta-1+n-\sum_i y_i}$
  - MAP: $\widehat{x}_{\text{MAP}} = \frac{\alpha+\sum_i y_i - 1}{\alpha+\beta+n-2}$
  - Example: $\alpha = 4$, $\beta = 4$, $n = 10$,
  $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$,

$$\widehat{x}_{\text{MAP}} = 0.625 \;\; \left(\text{recall } \widehat{x}_{\text{ML}} = 0.7\right)$$



Another Classical Statistical Inference Criterion

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

Another Classical Statistical Inference Criterion

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

Another Classical Statistical Inference Criterion

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

$$\begin{aligned}\widehat{x}_{\text{Bayes}} &= \arg\min_{\widehat{x}} \mathbb{E}[(\widehat{x} - X)^2 | Y = y] \\ &= \arg\min_{\widehat{x}} \; \widehat{x}^2 - 2\,\widehat{x}\,\mathbb{E}[X|Y = y] \\ &= \mathbb{E}[X|Y = y] \; \equiv \; \widehat{x}_{\text{MMSE}}\end{aligned}$$

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: $L(\widehat{x}, x) = (\widehat{x} - x)^2$
- Optimal Bayesian decision:

$$\begin{aligned} \widehat{x}_{\text{Bayes}} &= \arg\min_{\widehat{x}} \mathbb{E}[(\widehat{x} - X)^2 | Y = y] \\ &= \arg\min_{\widehat{x}} \; \widehat{x}^2 - 2\,\widehat{x}\,\mathbb{E}[X|Y = y] \\ &= \mathbb{E}[X|Y = y] \;\equiv\; \widehat{x}_{\text{MMSE}} \end{aligned}$$

MMSE = 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.

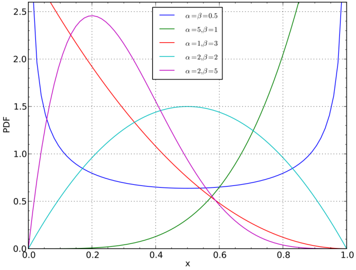
Back to the Bernoulli Example

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}\left(y_1, ..., y_n|x\right) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$

Back to the Bernoulli Example

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}(y_1, ..., y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$

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

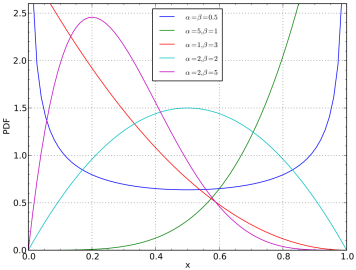


Back to the Bernoulli Example

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

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

  - In our case, the Beta pdf $f_X(x) \propto x^{\alpha-1}(1-x)^{\beta-1}, \;\; \alpha, \beta > 0$
  - Posterior: $f_{X|Y}(x|y) = x^{\alpha-1+\sum_i y_i}(1-x)^{\beta-1+n-\sum_i y_i}$

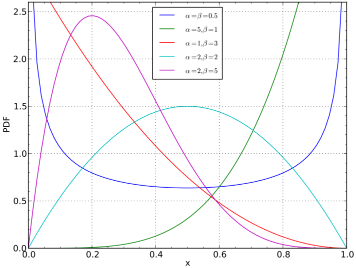


Back to the Bernoulli Example

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

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

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

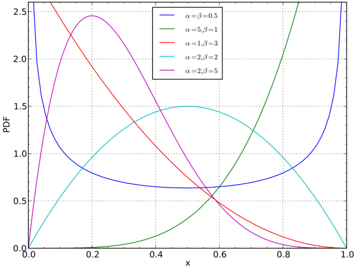


Back to the Bernoulli Example

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

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

  - 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: $\widehat{x}_{\text{MMSE}} = \frac{\alpha+\sum_i y_i}{\alpha+\beta+n}$
  - Example: $\alpha = 4,\ \beta = 4,\ n = 10$, $y = (1, 1, 1, 0, 1, 0, 0, 1, 1, 1)$,

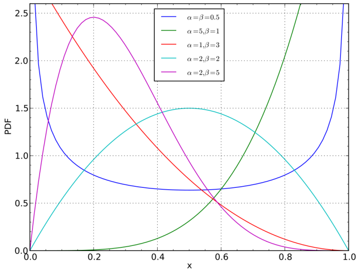


$\widehat{x}_{\text{MMSE}} \simeq 0.611 \;\; \left(\text{recall that } \widehat{x}_{\text{MAP}} = 0.625,\;\; \widehat{x}_{\text{ML}} = 0.7\right)$

Back to the Bernoulli Example

- Observed n i.i.d. (independent identically distributed) Bernoulli RVs.
- Likelihood: $f_{Y|X}(y_1,...,y_n|x) = \prod_{i=1}^{n} x^{y_i}(1-x)^{1-y_i} = x^{\sum_i y_i}(1-x)^{n-\sum_i y_i}$

  - 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: $\widehat{x}_{\text{MMSE}} = \frac{\alpha+\sum_i y_i}{\alpha+\beta+n}$
  - Example: $\alpha = 4$, $\beta = 4$, $n = 10$, $y = (1,1,1,0,1,0,0,1,1,1)$,



$$\widehat{x}_{\text{MMSE}} \simeq 0.611 \;\; \left(\text{recall that } \;\widehat{x}_{\text{MAP}} = 0.625, \;\; \widehat{x}_{\text{ML}} = 0.7\right)$$

- Conjugate prior equivalent to "virtual" counts; often called "smoothing" in NLP and 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

$$\widehat{x}_{\text{ML}} = 0.7, \quad \widehat{x}_{\text{MAP}} = \frac{3 + \sum_i y_i}{6 + n} = 0.625, \quad \widehat{x}_{\text{MMSE}} = \frac{4 + \sum_i y_i}{8 + n} \simeq 0.611$$

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

$$\widehat{x}_{\text{ML}} = 0.7, \quad \widehat{x}_{\text{MAP}} = \frac{3 + \sum_i y_i}{6+n} = 0.625, \quad \widehat{x}_{\text{MMSE}} = \frac{4 + \sum_i y_i}{8+n} \simeq 0.611$$

- Consider $n = 100$, and $\sum_i y_i = 70$, with the same Beta(4,4) prior

$$\widehat{x}_{\text{ML}} = 0.7, \quad \widehat{x}_{\text{MAP}} = \frac{73}{106} \simeq 0.689, \quad \widehat{x}_{\text{MMSE}} = \frac{74}{108} \simeq 0.685$$

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

The Bernstein-Von Mises Theorem

- In the previous example, we had
  $n = 10, \;\; y = (1,1,1,0,1,0,0,1,1,1), \;\; \text{thus } \sum_i y_i = 7.$
  With a Beta prior with $\alpha = 4$ and $\beta = 4$, we had

$$\widehat{x}_{\text{ML}} = 0.7, \quad \widehat{x}_{\text{MAP}} = \frac{3 + \sum_i y_i}{6 + n} = 0.625, \quad \widehat{x}_{\text{MMSE}} = \frac{4 + \sum_i y_i}{8 + n} \simeq 0.611$$

- Consider $n = 100$, and $\sum_i y_i = 70$, with the same Beta(4,4) prior

$$\widehat{x}_{\text{ML}} = 0.7, \quad \widehat{x}_{\text{MAP}} = \frac{73}{106} \simeq 0.689, \quad \widehat{x}_{\text{MMSE}} = \frac{74}{108} \simeq 0.685$$

  ... 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,

$$\lim_{n\to\infty} \widehat{x}_{\text{MAP}} = \lim_{n\to\infty} \widehat{x}_{\text{MMSE}} = \widehat{x}_{\text{ML}}$$

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

Important Inequalities

- Markov's ineqality: if $X \geq 0$ is an RV with expectation $\mathbb{E}(X)$, then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}$$

Important Inequalities

- Markov's ineqality: if $X \geq 0$ is an RV with expectation $\mathbb{E}(X)$, then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}$$

Simple proof:

$$t\,\mathbb{P}(X > t) = \int_t^\infty t\, f_X(x)\, dx \leq \int_t^\infty x\, f_X(x)\, dx = \mathbb{E}(X) - \underbrace{\int_0^t x\, f_X(x)\, dx}_{\geq 0} \leq \mathbb{E}(X)$$

Important Inequalities

- Markov's ineqality: if $X \geq 0$ is an RV with expectation $\mathbb{E}(X)$, then

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t}$$

Simple proof:

$$t\,\mathbb{P}(X > t) = \int_t^\infty t\, f_X(x)\, dx \leq \int_t^\infty x\, f_X(x)\, dx = \mathbb{E}(X) - \underbrace{\int_0^t x\, f_X(x)\, dx}_{\geq 0} \leq \mathbb{E}(X)$$

- Chebyshev's inequality: $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \mathrm{var}(Y)$, then

$$\mathbb{P}(|X - \mu| \geq s) \leq \frac{\sigma^2}{s^2}$$

...simple corollary of Markov's inequality, with $X = |Y - \mu|^2, \;\; t = s^2$

Other Important Inequalities

- Cauchy-Schwartz's inequality for RVs:

$$\mathbb{E}(|X\,Y|) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$$

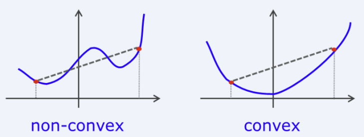
Other Important Inequalities

- 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 $\alpha \in [0, 1]$

$$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y)$$



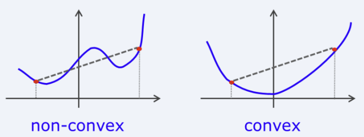
Other Important Inequalities

- 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

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

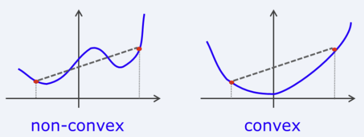
Other Important Inequalities

- 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

$$g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))$$

Examples: $\mathbb{E}(X)^2 \leq \mathbb{E}(X^2) \;\Rightarrow\; \text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0.$
$\mathbb{E}(\log X) \leq \log \mathbb{E}(X)$, for X a positive RV.

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

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

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$
- Measure of uncertainty/randomness of X

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$
- Measure of uncertainty/randomness of X

Continuous RV X, differential entropy: $\boxed{h(X) = -\int f_X(x) \log f_X(x)\, dx}$

Entropy and all that...

Entropy of a discrete RV $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$
- Measure of uncertainty/randomness of X

Continuous RV X, differential entropy: $\boxed{h(X) = -\int f_X(x) \log f_X(x)\, dx}$

- $h(X)$ 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 $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$
- Measure of uncertainty/randomness of X

Continuous RV X, differential entropy: $\boxed{h(X) = -\int f_X(x) \log f_X(x)\, dx}$

- $h(X)$ 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 $X \in \{1, ..., K\}$: $\boxed{H(X) = -\sum_{x=1}^{K} f_X(x) \log f_X(x)}$

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \;\Leftrightarrow\; f_X(i) = 1$, for exactly one $i \in \{1, ..., K\}$.
- Upper bound: $H(X) \leq \log K$;
  $H(X) = \log K \;\Leftrightarrow\; f_X(x) = 1/k$, for all $x \in \{1, ..., K\}$
- Measure of uncertainty/randomness of X

Continuous RV X, differential entropy: $\boxed{h(X) = -\int f_X(x) \log f_X(x)\, dx}$

- $h(X)$ 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 $var(Y) = \sigma^2$, then $h(Y) \leq \frac{1}{2}\log(2\pi e \sigma^2)$

Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) between two pmf:

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

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: $D(f_X \| g_X) \geq 0$

$D(f_X \| g_X) = 0 \;\Leftrightarrow\; f_X(x) = g_X(x), \text{ for } x \in \{1, ..., K\}$

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: $D(f_X \| g_X) \geq 0$

$D(f_X \| g_X) = 0 \;\Leftrightarrow\; f_X(x) = g_X(x), \text{ for } x \in \{1, ..., K\}$

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:

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

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

KLD between two pdf:

$$D(f_X \| g_X) = \int f_X(x) \log \frac{f_X(x)}{g_X(x)}\, dx$$

Positivity: $D(f_X \| g_X) \geq 0$
$D(f_X \| g_X) = 0 \;\Leftrightarrow\; f_X(x) = g_X(x), \text{ almost everywhere}$

Mutual information

Mutual information (MI) between two random variables:

$$I(X;Y) = D\big(f_{X,Y} \| f_X \, f_Y\big)$$

Mutual information

Mutual information (MI) between two random variables:

$$I(X;Y) = D\big(f_{X,Y} \| f_X \, f_Y\big)$$

Positivity: $I(X;Y) \geq 0$
$I(X;Y) = 0 \;\Leftrightarrow\; X, Y$ are independent.

Mutual information

Mutual information (MI) between two random variables:

$$I(X;Y) = D\big(f_{X,Y} \| f_X\, f_Y\big)$$

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

$I(X;Y) = 0 \;\Leftrightarrow\; X, Y$ are independent.

MI is a measure of dependency between two random variables

Recommended Reading (Probability and Statistics)

- K. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012 (Chapter 2).
- L. Wasserman, "All of Statistics: A Concise Course in Statistical Inference", Springer, 2004.

Linear Algebra

- Linear algebra provides (among many other things) a compact way of representing, studying, and solving linear systems of equations

Linear Algebra

- Linear algebra provides (among many other things) a compact way of representing, studying, and solving linear systems of equations
- Example: the system

$$\begin{aligned} 4\,x_1 - 5\,x_2 &= -13 \\ -2\,x_1 + 3\,x_2 &= 9 \end{aligned}$$

can be written compactly as $\boxed{Ax = b}$, where

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \; b = \begin{bmatrix} -13 \\ 9 \end{bmatrix},$$

and can be solved as

$$x = A^{-1}b = \begin{bmatrix} 1.5 & 2.5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -13 \\ 9 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m\times n}$ is a matrix with m rows and n columns.

$$A = \left[\begin{array}{ccc} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{array}\right].$$

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m \times n}$ is a matrix with m rows and n columns.

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}.$$

- $x \in \mathbb{R}^n$ is a vector with n components,

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m \times n}$ is a matrix with m rows and n columns.

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}.$$

- $x \in \mathbb{R}^n$ is a vector with n components,

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

- A (column) vector is a matrix with n rows and 1 column.

Notation: Matrices and Vectors

- $A \in \mathbb{R}^{m\times n}$ is a matrix with m rows and n columns.

$$A = \begin{bmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{bmatrix}.$$

- $x \in \mathbb{R}^n$ is a vector with n components,

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

- A (column) vector is a matrix with n rows and 1 column.

- A matrix with 1 row and n columns is called a row vector.

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m \times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m \times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- A matrix A is symmetric if $A^T = A$.

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m \times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- A matrix A is symmetric if $A^T = A$.
- Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m \times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m\times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- A matrix A is symmetric if $A^T = A$.
- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Inner product between vectors $x, y \in \mathbb{R}^n$:

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

Matrix Transpose and Products

- Given matrix $A \in \mathbb{R}^{m \times n}$, its transpose A^T is such that $(A^T)_{i,j} = A_{j,i}$.
- A matrix A is symmetric if $A^T = A$.
- Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m \times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Inner product between vectors $x, y \in \mathbb{R}^n$:

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

- Outer product between vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$: $x\,y^T \in \mathbb{R}^{n \times m}$, where $(x\,y^T)_{i,j} = x_i\, y_j$.

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m \times p} \quad \text{where} \;\; C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \;\; C_{i,j} = \sum_{k=1}^{n} A_{i,k}\,B_{k,j}$$

- Matrix product is associative: $(AB)C = A(BC)$.

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Matrix product is associative: $(AB)C = A(BC)$.

- In general, matrix product is not commutative: $AB \neq BA$.

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Matrix product is associative: $(AB)C = A(BC)$.
- In general, matrix product is not commutative: $AB \neq BA$.
- Transpose of product: $(A\,B)^T = B^T A^T$.

Properties of Matrix Products and Transposes

- Given matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$, their product is

$$C = A\,B \;\in \mathbb{R}^{m\times p} \quad \text{where} \;\; C_{i,j} = \sum_{k=1}^{n} A_{i,k}\, B_{k,j}$$

- Matrix product is associative: $(AB)C = A(BC)$.
- In general, matrix product is not commutative: $AB \neq BA$.
- Transpose of product: $(A\,B)^T = B^T A^T$.
- Transpose of sum: $(A+B)^T = A^T + B^T$.

Norms

- The norm of a vector is (informally) its "length". Euclidean norm:

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

Norms

- The norm of a vector is (informally) its "length". Euclidean norm:

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

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

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

Norms

- The norm of a vector is (informally) its "length". Euclidean norm:

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

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

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

- Notable case: the ℓ_1 norm, $\|x\|_1 = \sum_i |x_i|$.

Norms

- The norm of a vector is (informally) its "length". Euclidean norm:

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

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

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

- Notable case: the ℓ_1 norm, $\|x\|_1 = \sum_i |x_i|$.
- Notable case: the ℓ_∞ norm, $\|x\|_\infty = \max\{|x_1|, ..., |x_n|\}$.

Norms

- The norm of a vector is (informally) its "length". Euclidean norm:

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

- More generally, the ℓ_p norm of a vector $x \in \mathbb{R}^n$, where $p \geq 1$,

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

- Notable case: the ℓ_1 norm, $\|x\|_1 = \sum_i |x_i|$.

- Notable case: the ℓ_∞ norm, $\|x\|_\infty = \max\{|x_1|, ..., |x_n|\}$.

- Notable case: the ℓ_0 "norm" (not): $\|x\|_0 = |\{i : x_i \neq 0\}|$.

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n \times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n\times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A\,I = I\,A = A$.

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n\times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A\,I = I\,A = A$.
- Diagonal matrix: $A \in \mathbb{R}^{n\times n}$ is diagonal if $(i \neq j) \Rightarrow A_{i,j} = 0$.

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n\times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A\,I = I\,A = A$.
- Diagonal matrix: $A \in \mathbb{R}^{n\times n}$ is diagonal if $(i \neq j) \Rightarrow A_{i,j} = 0$.
- Upper triangular matrix: $(j < i) \Rightarrow A_{i,j} = 0$.

Special Matrices

- The identity matrix $I \in \mathbb{R}^{n\times n}$ is a square matrix such that

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

- Neutral element of matrix product: $A\,I = I\,A = A$.
- Diagonal matrix: $A \in \mathbb{R}^{n\times n}$ is diagonal if $(i \neq j) \Rightarrow A_{i,j} = 0$.
- Upper triangular matrix: $(j < i) \Rightarrow A_{i,j} = 0$.
- Lower triangular matrix: $(j > i) \Rightarrow A_{i,j} = 0$.

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

- Matrix trace:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

- Matrix trace:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

- Matrix determinant:

$$|A| = \det(A) = \prod_i \lambda_i$$

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n\times n}$ if

$$A\,x = \lambda\,x,$$

where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

- Matrix trace:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

- Matrix determinant:

$$|A| = \det(A) = \prod_i \lambda_i$$

- Properties: $|AB| = |A||B|$,

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n\times n}$ if

$$A\,x = \lambda\,x,$$

where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

- Matrix trace:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

- Matrix determinant:

$$|A| = \det(A) = \prod_i \lambda_i$$

- Properties: $|AB| = |A||B|, \;\; |A^T| = |A|,$

Eigenvalues, eigenvectors, determinant, trace

- A vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n\times n}$ if

$$A\,x = \lambda\,x,$$

  where $\lambda \in \mathbb{R}$ is the corresponding eigenvalue.

- The eigenvalues of a diagonal matrix are the elements in the diagonal.

- Matrix trace:

$$\text{trace}(A) = \sum_i A_{i,i} = \sum_i \lambda_i$$

- Matrix determinant:

$$|A| = \det(A) = \prod_i \lambda_i$$

- Properties: $|AB| = |A||B|, \quad |A^T| = |A|, \quad |\alpha\,A| = \alpha^n |A|$

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ in invertible if there is $B \in \mathbb{R}^{n \times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ in invertible if there is $B \in \mathbb{R}^{n \times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n \times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A$,

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ in invertible if there is $B \in \mathbb{R}^{n \times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n \times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A, \;\; (A^{-1})^T = (A^T)^{-1}$,

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A, \;\; (A^{-1})^T = (A^T)^{-1}, \;\; (A\,B)^{-1} = B^{-1}A^{-1}$

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n\times n}$ in invertible if there is $B \in \mathbb{R}^{n\times n}$ s.t. $AB = BA = I$.
- ...matrix B, such that $AB = BA = I$, denoted $B = A^{-1}$.
- Matrix $A \in \mathbb{R}^{n\times n}$ is invertible $\Leftrightarrow \det(A) \neq 0$.
- Determinant of inverse: $\det(A^{-1}) = \dfrac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.
- Properties: $(A^{-1})^{-1} = A$, $(A^{-1})^T = (A^T)^{-1}$, $(A\,B)^{-1} = B^{-1}A^{-1}$
- There are several algorithms to compute A^{-1}; general case, computational cost $O(n^3)$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \ \in \ \mathbb{R}$$

is called a quadratic form.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \;\in\; \mathbb{R}$$

  is called a quadratic form.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \ \in \ \mathbb{R}$$

  is called a quadratic form.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite (PD) if, for any $x \in \mathbb{R}^n$, $(x \neq 0) \Rightarrow x^T A x > 0$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n \times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \;\in\; \mathbb{R}$$

  is called a quadratic form.

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive semi-definite (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite (PD) if, for any $x \in \mathbb{R}^n$, $(x \neq 0) \Rightarrow x^T A x > 0$.

- Matrix $A \in \mathbb{R}^{n \times n}$ is PSD $\Leftrightarrow$ all $\lambda_i(A) \geq 0$.

Quadratic Forms and Positive (Semi-)Definite Matrices

- Given matrix $A \in \mathbb{R}^{n\times n}$ and vector $x \in \mathbb{R}^n$,

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j}\, x_i\, x_j \;\in\; \mathbb{R}$$

  is called a quadratic form.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive semi-definite (PSD) if, for any $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

- A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite (PD) if, for any $x \in \mathbb{R}^n$, $(x \neq 0) \Rightarrow x^T A x > 0$.

- Matrix $A \in \mathbb{R}^{n\times n}$ is PSD $\Leftrightarrow$ all $\lambda_i(A) \geq 0$.

- Matrix $A \in \mathbb{R}^{n\times n}$ is PD $\Leftrightarrow$ all $\lambda_i(A) > 0$.