Warm-up: Probability Theory Refresher (and a few more things)

Mário A. T. Figueiredo

Instituto Superior Técnico & Instituto de Telecomunicações

Lisboa, Portugal

LxMLS 2017: Lisbon Machine Learning School

June 14, 2018
Probability theory has its roots in games of chance. Great names of science: Bayes, Bernoulli(s), Boltzman, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ... Tool to handle uncertainty, information, knowledge, observations, ... thus also learning, decision making, inference, science,...

Mário A. T. Figueiredo (IST & IT)
LxMLS 2017: Probability Theory
June 14, 2018
Probability theory has its roots in games of chance.

- Inference and learning

- Tool to handle uncertainty, information, knowledge, observations, ...
- Thus also learning, decision making, inference, science,...
Probability theory has its roots in games of chance.

Great names of science: Bayes, Bernoulli(s), Boltzman, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ...

Inference and learning
Probability theory

- Probability theory has its roots in games of chance
- Great names of science: Bayes, Bernoulli(s), Boltzman, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ...
- Tool to handle uncertainty, information, knowledge, observations, ...
Probability theory

- Probability theory has its roots in games of chance
- Great names of science: Bayes, Bernoulli(s), Boltzmann, Cardano, Cauchy, Fermat, Huygens, Kolmogorov, Laplace, Pascal, Poisson, ...
- Tool to handle uncertainty, information, knowledge, observations, ...
- ...thus also learning, decision making, inference, science,...
Still important today?

CONTENTS

3 Probability and Information Theory 51
  3.1 Why Probability? .............................. 52
  3.2 Random Variables ............................. 54
  3.3 Probability Distributions .................... 54
  3.4 Marginal Probability ......................... 56
  3.5 Conditional Probability ...................... 57
  3.6 The Chain Rule of Conditional Probabilities 57
  3.7 Independence and Conditional Independence 58
  3.8 Expectation, Variance and Covariance .......... 58
  3.9 Common Probability Distributions .......... 60
  3.10 Useful Properties of Common Functions ...... 65
  3.11 Bayes’ Rule .................................. 68
  3.12 Technical Details of Continuous Variables .. 68
  3.13 Information Theory ......................... 70
  3.14 Structured Probabilistic Models ............. 74
### Contents

3 Probability and Information Theory 51

3.1 Why Probability? .......................... 52
3.2 Random Variables .......................... 54
3.3 Probability Distributions .................. 54
3.4 Marginal Probability ...................... 56
3.5 Conditional Probability .................... 57
3.6 The Chain Rule of Conditional Probabilities .......... 57
3.7 Independence and Conditional Independence .......... 58
3.8 Expectation, Variance and Covariance .............. 58
3.9 Common Probability Distributions ............... 60
3.10 Useful Properties of Common Functions .......... 65
3.11 Bayes’ Rule ................................ 68
3.12 Technical Details of Continuous Variables .......... 68
3.13 Information Theory .......................... 70
3.14 Structured Probabilistic Models .................. 74

What book is this from?
Do we still need this?
What is probability?

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

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

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

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

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

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

  *Laplace, 1814*
What is probability?

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

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

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

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

  *Laplace, 1814*

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

  ...relative frequency of occurrence of \( A \) in infinite number of trials.
What is probability?

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

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

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

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

  Laplace, 1814

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

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

- **Subjective probability:** \( \mathbb{P}(A) \) is a degree of belief.  
  de Finetti, 1930s

  ...gives meaning to \( \mathbb{P}(\text{“it will rain today”}) \), or \( \mathbb{P}(\text{“I’ll have the flue next winter”}) \).
Key concepts: Sample space and events

- **Sample space** $\mathcal{X} =$ set of possible outcomes of a random experiment.
  
  **Examples:**
  
  - Tossing two coins: $\mathcal{X} = \{HH, TH, HT, TT\}$
  
  - Roulette: $\mathcal{X} = \{1, 2, ..., 36\}$
  
  - Draw a card from a shuffled deck: $\mathcal{X} = \{A\spadesuit, 2\spadesuit, ..., Q\spadesuit, K\spadesuit\}$. 
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: $\mathcal{X} = \{1, 2, \ldots, 36\}$
  
  - Draw a card from a shuffled deck: $\mathcal{X} = \{A\spadesuit, 2\spadesuit, \ldots, Q\diamond, K\diamond\}$.

- An event $A$ is a subset of $\mathcal{X}$: $A \subseteq \mathcal{X}$ (also written $A \in 2^{\mathcal{X}}$).
  
  Examples:
  
  - “exactly one H in 2-coin toss”: $A = \{TH, HT\}$.
  
  - “odd number in the roulette”: $B = \{1, 3, \ldots, 35\}$.
  
  - “drawn a ♥ card”: $C = \{A♥, 2♥, \ldots, K♥\}$.
Key concepts: Sample space and events

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

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

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

- Properly handling the continuous case requires deeper concepts:
  
  - Sigma algebras
  - Measurable functions
Key concepts: Sample space and events

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

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

- Properly handling the continuous case requires deeper concepts:
  - Sigma algebras
  - Measurable functions

  ...**heavier** stuff, not covered here
Kolmogorov’s Axioms for Probability

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

Kolmogorov’s axioms (1933) for probability
Kolmogorov’s Axioms for Probability

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

  **Kolmogorov’s axioms (1933) for probability**
  
  - 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

- 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

- 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

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

- From these axioms, many results can be derived.
Kolmogorov’s Axioms for Probability

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

**Kolmogorov’s axioms (1933) for probability**

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

- From these axioms, many results can be derived.

**Examples:**

- $\mathbb{P}(\emptyset) = 0$
Kolmogorov’s Axioms for Probability

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

**Kolmogorov’s axioms (1933) for probability**

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

- From these axioms, many results can be derived.

**Examples:**

- $\mathbb{P}(\emptyset) = 0$
- $C \subset D \Rightarrow \mathbb{P}(C) \leq \mathbb{P}(D)$
Kolmogorov’s Axioms for Probability

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

**Kolmogorov’s axioms (1933) for probability**

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

- From these axioms, many results can be derived.

**Examples:**

- $\mathbb{P}(...) = 0$
- $C \subseteq D \Rightarrow \mathbb{P}(C) \leq \mathbb{P}(D)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
Kolmogorov’s Axioms for Probability

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

**Kolmogorov’s axioms (1933) for probability**

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

- From these axioms, many results can be derived.

**Examples:**

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

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

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

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

Independence: $A, B$ are independent ($A \perp \perp B$):

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

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

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

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

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

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

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

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

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

- Relationship with conditional probabilities:

  $$A \perp \perp B \iff \mathbb{P}(A|B) = \mathbb{P}(A)$$
Conditional Probability and Independence

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

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

- Relationship with conditional probabilities:

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

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

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{4\heartsuit\}) = \frac{1}{52}$$
Conditional Probability and Independence

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

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

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

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

  $$\begin{align*}
  \mathbb{P}(A \cap B) &= \mathbb{P}\left(\{4\heartsuit\}\right) = \frac{1}{52} \\
  \mathbb{P}(A) \mathbb{P}(B) &= \frac{1}{13} \frac{1}{4} = \frac{1}{52}
  \end{align*}$$
Conditional Probability and Independence

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

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

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

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

  $$\begin{align*}
  \mathbb{P}(A \cap B) &= \mathbb{P}\{4\heartsuit\} = \frac{1}{52} \\
  \mathbb{P}(A) \mathbb{P}(B) &= \frac{1}{13} \frac{1}{4} = \frac{1}{52} \\
  \mathbb{P}(A|B) &= \mathbb{P}(\text{"4" | "\heartsuit"}) = \frac{1}{13} = \mathbb{P}(A)
  \end{align*}$$
Bayes Theorem

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

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

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

\[
P(B) = \sum_i P(B | A_i) P(A_i) = \sum_i P(B \cap A_i)
\]

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

\[
P(A_i | B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B | A_i) P(A_i)}{P(B)}
\]
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 $[0, 1]$)

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

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

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

![Diagram of random variable](image)
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 $\{0, 1\}$)
Random Variables

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

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

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

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

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

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

$\mathcal{X}$ 

$\omega$ 

$X(\omega)$ 

$\mathbb{R}$
Random Variables

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

  - **Discrete RV**: range of \( X \) is countable (e.g., \( \mathbb{N} \) or \( \{0, 1\} \))
  - **Continuous RV**: range of \( X \) is uncountable (e.g., \( \mathbb{R} \) or \([0, 1]\))
  - **Example**: number of heads in tossing two coins, \( \mathcal{X} = \{HH, HT, TH, TT\} \), 
    \( X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0. \) 
    Range of \( X = \{0, 1, 2\} \).
Random Variables

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

- **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])

- **Example**: number of heads 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 $X = \mathbb{R}_+$. 

\[\begin{array}{c}
\mathcal{X} \\
\omega \\
X(\omega) \\
\mathbb{R}
\end{array}\]
Random Variables: Distribution Function

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

\[ \mathcal{X} \]

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

\[ X \]

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

\[ x \]

\[ \mathbb{R} \]
Random Variables: Distribution Function

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

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

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

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

- **Probability mass function** (discrete RV): \( f_X(x) = \mathbb{P}(X = x) \),
  \[
  F_X(x) = \sum_{x_i \leq x} f_X(x_i).
  \]
Important Discrete Random Variables

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

  **Example**: a fair roulette $X \in \{1, \ldots, 36\}$, with $f_X(x) = 1/36$

  **Example**: a fair die $X \in \{1, \ldots, 6\}$, with $f_X(x) = 1/6$
Important Discrete Random Variables

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

  *Example*: a fair roulette $X \in \{1, \ldots, 36\}$, with $f_X(x) = 1/36$

  *Example*: a fair die $X \in \{1, \ldots, 6\}$, with $f_X(x) = 1/6$

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

  Compact form: \( f_X(x) = p^x(1 - p)^{1-x} \).

  *Example*: an unfair coin (heads = 0, tails = 1), with $p \neq 1/2$. 
Important Discrete Random Variables

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

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

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

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

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

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

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

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

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

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

**Example**: number of heads in $n$ coin tosses.
Other Important Discrete Random Variables

- **Geometric\( (p) \):** \( X \in \mathbb{N} \), pmf \( f_X(x) = p(1 - p)^{x-1} \).

  **Example:** number of coin tosses until first heads.
Other Important Discrete Random Variables

- **Geometric($p$):** $X \in \mathbb{N}$, pmf $f_X(x) = p(1 - p)^{x-1}$.

  **Example:** number of coin tosses until first heads.

- **Poisson($\lambda$):**

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

  “...probability of the number of independent occurrences in a fixed (time/space) interval, if these occurrences have known average rate”
Other Important Discrete Random Variables

- **Geometric**$(p)$: $X \in \mathbb{N}$, pmf $f_X(x) = p(1 - p)^{x-1}$.

  **Example**: number of coin tosses until first heads.

- **Poisson**$(\lambda)$:

  $$X \in \mathbb{N} \cup \{0\},$$

  $$\text{pmf } f_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

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

  **Examples**: number of rain drops per second on a given area, number of calls per hour in a call center, number of tweets per day by DT, ...
Continuous Random Variables

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

\[
\int_{-\infty}^{\infty} f_X(x) \, dx = 1 \quad \mathbb{P}(X \in [a, b]) = \int_{a}^{b} f_X(x) \, dx
\]
Continuous Random Variables

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

\[
\int_{-\infty}^{\infty} f_X(x) = 1 \quad \mathbb{P}(X \in [a, b]) = \int_{a}^{b} f_X(x) \, dx
\]

- **Notice:** $\mathbb{P}(X = c) = 0$
Important Continuous Random Variables

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- **Example:** Bernoulli variance, \( E(X^2) = E(X) = p \)
Expectation of Functions of RVs

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

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

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

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

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

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

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

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

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

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

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

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

$$\mathbb{P}(X \in A) = \int_A f_X(x) \, dx = \int 1_A(x) f_X(x) \, dx = \mathbb{E}(1_A(X))$$
The importance of the Gaussian
The importance of the Gaussian

Take $n$ independent RVs $X_1, \ldots, X_n$, with $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$.
The importance of the Gaussian

Take \( n \) independent RVs \( X_1, \ldots, X_n \), with \( \mathbb{E}[X_i] = \mu_i \) and \( \text{var}(X_i) = \sigma_i^2 \)

- Their sum, \( Y_n = \sum_{i=1}^{n} X_i \) satisfies:

\[
\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mu_i \equiv \mu
\]
The importance of the Gaussian

Take $n$ independent RVs $X_1, \ldots, X_n$, with $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma^2_i$

- Their sum, $Y_n = \sum_{i=1}^{n} X_i$ satisfies:

  $$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mu_i \equiv \mu$$
  $$\text{var}(Y_n) = \sum_{i} \sigma^2_i \equiv \sigma^2$$

- Let $Z_n = \frac{Y_n - \mu}{\sigma}$, thus $\mathbb{E}[Z_n] = 0$ and $\text{var}(Z_n) = 1$
The importance of the Gaussian

Take $n$ independent RVs $X_1, ..., X_n$, with $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma^2_i$

- Their sum, $Y_n = \sum_{i=1}^{n} X_i$ satisfies:

  $$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mu_i \equiv \mu \quad \text{var}(Y_n) = \sum_{i} \sigma^2_i \equiv \sigma^2$$

- Let $Z_n = \frac{Y_n - \mu}{\sigma}$, thus $\mathbb{E}[Z_n] = 0$ and $\text{var}(Z_n) = 1$

- Central limit theorem: under mild conditions,

  $$\lim_{n \to \infty} Z_n \sim \mathcal{N}(0, 1)$$
Two (or More) Random Variables

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

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

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

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

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

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

  Extends trivially to more than two RVs.

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

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

  Extends trivially to more than two RVs.

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

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

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

  Extends trivially to more than two RVs.

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

- **Independence:**
  
  \( X \independent Y \iff f_{X,Y}(x, y) = f_X(x) f_Y(y) \).
Two (or More) Random Variables

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

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

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

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

• **Conditional pmf** (discrete RVs):

\[
f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}
\]
Conditionals and Bayes’ Theorem

- **Conditional pmf** (discrete RVs):

  \[ f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \]

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

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

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

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

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

- **Conditional pmf** (discrete RVs):

  \[ f_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \]

- **Conditional pdf** (continuous RVs):

  \[ f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \]
  
  ...the meaning is technically delicate.

- **Bayes’ theorem**:

  \[ f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \]  (pdf or pmf).

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

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

<table>
<thead>
<tr>
<th>$f_{X,Y}(x, y)$</th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>1/5</td>
<td>2/5</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>1/10</td>
<td>3/10</td>
</tr>
</tbody>
</table>
Joint, Marginal, and Conditional Probabilities: An Example

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

$$
\begin{array}{|c|c|c|}
\hline
x, y & Y = 0 & Y = 1 \\
\hline
0, 0 & 1/5 & 2/5 \\
0, 1 & 1/10 & 3/10 \\
1, 0 & 1/10 & 3/10 \\
1, 1 & 2/5 & 3/10 \\
\hline
\end{array}
$$

Marginals:

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

$$
f_Y(0) = \frac{1}{5} + \frac{1}{10} = \frac{3}{10}, \quad f_Y(1) = \frac{2}{5} + \frac{3}{10} = \frac{7}{10}.
$$
Joint, Marginal, and Conditional Probabilities: An Example

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

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

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

- **Conditional probabilities:**

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

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

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

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

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

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

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

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

  $\left( \begin{array}{c} n \\ x_1 x_2 \cdots x_K \end{array} \right) = \frac{n!}{x_1! x_2! \cdots x_K!}$

- 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**: \( X = (X_1, ..., X_K), X_i \in \{0, ..., n\}, \) s.t. \( \sum_i X_i = n, \)

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

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

Parameters: \( p_1, ..., p_K \geq 0, \) 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: Multinomial

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

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

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

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

- Generalizes the binomial from binary to \( K \)-classes.

- **Example**: tossing \( n \) independent fair dice, \( p_1 = \cdots = p_6 = 1/6 \).

\[
        x_i = \text{number of outcomes with } i \text{ dots (of course, } \sum_i x_i = n)\]

- **Example**: bag of words (BoW) multinomial model with vocabulary of \( K \) words
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:

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

- **Covariance** between two RVs:

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

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

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

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

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

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

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

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

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

- **Covariance** between two RVs:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Combining the 2-nd and the 4-th facts is called **standardization**
Exponential Families

A pdf or pmf $f_X(x|\eta)$, with parameter(s) $\eta$, for $X \in \mathcal{X}$, is in an exponential family if

$$f_X(x|\eta) = \frac{1}{Z(\eta)} \ h(x) \ \exp(\eta^T \phi(x))$$
Exponential Families

A pdf or pmf \( f_X(x|\eta) \), with parameter(s) \( \eta \), for \( X \in \mathcal{X} \), is in an exponential family if

\[
f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x))
\]

where \( \eta^T \phi(x) = \sum_j \eta_j \phi_j(x) \) and

\[
Z(\theta) = \int_{\mathcal{X}} h(x) \exp(\eta^T \phi(x)) \, dx.
\]
Exponential Families

A pdf or pmf \( f_X(x|\eta) \), with parameter(s) \( \eta \), for \( X \in \mathcal{X} \), is in an exponential family if

\[
f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x))
\]

where \( \eta^T \phi(x) = \sum_j \eta_j \phi_j(x) \) and

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

- Canonical parameter(s): \( \eta \)
- Sufficient statistics: \( \phi(x) \)
- Partition function: \( Z(\eta) \)

Examples: Bernoulli, Poisson, binomial, multinomial, Gaussian, exponential, beta, Dirichlet, Laplacian, log-normal, Wishart, ...
Exponential Families

\[ f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x)) \]
Exponential Families

\[ f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x)) \]

**Example:** Bernoulli pmf \( f_X(x) = p^x (1 - p)^{1-x} \),

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

thus \( \eta = \log \frac{p}{1-p}, \phi(x) = x, Z(\eta) = 1 + e^\eta, \) and \( h(x) = 1. \)
Exponential Families

\[ f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x)) \]

Example: Bernoulli pmf \( f_X(x) = p^x (1 - p)^{1-x} \),

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

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

Notice that \( p = \frac{e^\eta}{1+e^\eta} \)

(logistic transformation)
More on Exponential Families

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

\[ X_1, \ldots, X_m \sim f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x)) \]

then

\[ f_{X_1,\ldots,X_m}(x_1, \ldots, x_m|\eta) = \frac{1}{Z(\eta)^m} \left( \prod_{j=1}^{m} h(x_i) \right) \exp \left( \eta^T \sum_{j=1}^{m} \phi(x_j) \right) \]
More on Exponential Families

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

\[
X_1, \ldots, X_m \sim f_X(x|\eta) = \frac{1}{Z(\eta)} h(x) \exp(\eta^T \phi(x))
\]

then

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

- Expected sufficient statistics:

\[
\frac{d \log Z(\eta)}{d \eta} = \frac{dZ(\eta)}{Z(\eta)} = \frac{1}{Z(\eta)} \int \phi(x) h(x) \exp(\eta^T \phi(x)) dx = \mathbb{E}(\phi(X))
\]
Important Inequalities

- **Markov’s inequality**: 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 inequality:** 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:

$$t\mathbb{P}(X > t) = \int_t^\infty t f_X(x) \, dx \leq \int_t^\infty x f_X(x) \, dx = \mathbb{E}(X) - \int_0^t x f_X(x) \, dx \leq \mathbb{E}(X)$$
Important Inequalities

- **Markov’s inequality:** 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:
  \[ t \mathbb{P}(X > t) = \int_t^\infty t f_X(x) \, dx \leq \int_t^\infty x f_X(x) \, dx = \mathbb{E}(X) - \int_0^t x f_X(x) \, dx \leq \mathbb{E}(X) \geq 0 \]

- **Chebyshev’s inequality:** $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \text{var}(Y)$, then
  \[ \mathbb{P}(|Y - \mu| \geq s) \leq \frac{\sigma^2}{s^2} \]

  ...simple corollary of Markov’s inequality, with $X = |Y - \mu|^2$, $t = \sigma^2$
Important Inequalities

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

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

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

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

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

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

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

Jensen’s inequality: if \( g \) is a real convex function, then

\[
g(\mathbb{E}(X)) \leq \mathbb{E}(g(X))
\]
Important Inequalities

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

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

  ![Diagram showing non-convex and convex functions]

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.
Information, entropy, and all that...

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

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

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

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

Measure of uncertainty/randomness of $X$.

With $\log_2$, units are bits/symbol.

Central role in information/coding theory: lower bound on expected number of bits to code $X$.

Widely used: physics, biological sciences (computational biology, neurosciences, ecology, ...), economics, finances, social sciences, ...
Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

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

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

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

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

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

Measure of uncertainty/randomness of $X$ with $\log_2$, units are bits/symbol

Central role in information/coding theory: lower bound on expected number of bits to code $X$

Widely used: physics, biological sciences (computational biology, neurosciences, ecology, ...), economics, finances, social sciences, ...
Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

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

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

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

- **Measure of uncertainty/randomness** of $X$
Information, entropy, and all that...

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

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

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

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

- Measure of uncertainty/randomness of $X$

- With $\log_2$, units are bits/symbol
Entropy of a discrete RV $X \in \{1, \ldots, K\}$:

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

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

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

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

- **With** $\log_2$, **units are bits/symbol**

- Central role in **information/coding theory**: lower bound on expected number of bits to code $X$
Information, entropy, and all that...

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

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

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

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

- Measure of uncertainty/randomness of $X$

- With $\log_2$, units are bits/symbol

- Central role in information/coding theory: lower bound on expected number of bits to code $X$

- Widely used: physics, biological sciences (computational biology, neurosciences, ecology, ...), economics, finances, social sciences, ...
Entropy and all that...

Continuous RV $X$, differential entropy:

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

$\text{Example: for } f_X(x) = \text{Uniform}(x; a, b), \quad h(X) = \log(b-a)$

$\text{Gaussian upper bound: } f_X(x) = \mathcal{N}(x; \mu, \sigma^2), \text{ then } h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$

$\text{For any RV } Y \text{ with } \text{var}(Y) = \sigma^2, \text{ then } h(Y) \leq \frac{1}{2} \log(2\pi e \sigma^2)$

...yet another reason for why the Gaussian is important.
Entropy and all that…

Continuous RV $X$, differential entropy: 

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

- $h(X)$ can be positive or negative (unlike in the discrete case)

Example: for $f_X(x) = \text{Uniform}(x; a, b)$,

$$h(X) = \log(b - a).$$
Entropy and all that...

Continuous RV $X$, differential entropy:

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

- $h(X)$ can be positive or negative (unlike in the discrete case)

  **Example:** for $f_X(x) = \text{Uniform}(x; a, b)$,

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

- **Gaussian upper bound:** $f_X(x) = \mathcal{N}(x; \mu, \sigma^2)$, then

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

  For any RV $Y$ with $\text{var}(Y) = \sigma^2$, then $h(Y) \leq \frac{1}{2} \log(2\pi e \sigma^2)$.

  ...yet another reason for why the Gaussian is important.
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: \( D(f_X \| g_X) \geq 0 \)
\[
D(f_X \| g_X) = 0 \iff f_X(x) = g_X(x), \text{ for } x \in \{1, ..., K\} \]
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:**

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

$$D(f_X \| g_X) = 0 \iff f_X(x) = g_X(x), \text{ for } x \in \{1, \ldots, K\}$$

KLD between two pdf:

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

Kullback-Leibler divergence (KLD) between two pmf:

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

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

KLD between two pdf:

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

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

\[ D(f_X \| g_X) = 0 \iff f_X(x) = g_X(x), \text{ for } x \in \{1, \ldots, K\} \]

KLD between two pdf:

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

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

\[ D(f_X \| g_X) = 0 \iff f_X(x) = g_X(x), \text{ almost everywhere} \]

Issues: not symmetric; \( D(f_X \| g_X) = +\infty \text{ if } g_X(x) = 0 \text{ and } f_X(x) \neq 0 \)
Mutual information (MI) between two random variables:

\[ I(X; Y) = D(f_{X,Y} \| f_X f_Y) \]
Mutual information (MI) between two random variables:

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

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

\( I(X; Y) = 0 \iff X, Y \text{ are independent.} \)
Mutual information (MI) between two random variables:

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

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

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

MI = measure of dependency between two random variables
Mutual information

Mutual information (MI) between two random variables:

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

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

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

MI = measure of dependency between two random variables

MI = number of bits of information that \( X \) has about \( Y \)
Mutual information

Mutual information (MI) between two random variables:

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

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

MI = measure of dependency between two random variables

MI = number of bits of information that \( X \) has about \( Y \)

Bound: \( I(X;Y) \leq \min\{H(X), H(Y)\} \)
Mutual information

Mutual information (MI) between two random variables:

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

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

MI = measure of dependency between two random variables

MI = number of bits of information that \( X \) has about \( Y \)

Bound: \( I(X; Y) \leq \min\{H(X), H(Y)\} \)

Deterministic function: if \( Y = \phi(X) \), then \( I(X; Y) = H(Y) \leq H(X) \)
Recommended Reading (Probability and Statistics)


Part II: Linear Algebra

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

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

\[
\begin{align*}
4x_1 - 5x_2 &= -13 \\
-2x_1 + 3x_2 &= 9
\end{align*}
\]

can be written compactly as \( Ax = b \), where

\[
A = \begin{bmatrix}
4 & -5 \\
-2 & 3
\end{bmatrix}, \quad b = \begin{bmatrix}
-13 \\
9
\end{bmatrix},
\]

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.

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

\[
x = \begin{bmatrix}
x_1 \\
\vdots \\
x_n
\end{bmatrix}.
\]
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 $(A^T)_{i,j} = A_{j,i}$. 

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 $A^T = A$. 

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 $y \in \mathbb{R}^m$: $xy^T \in \mathbb{R}^{n \times m}$, where $(xy^T)_{i,j} = x_i y_j$. 

Mário A. T. Figueiredo (IST & IT) 
LxMLS 2017: Probability Theory 
June 14, 2018 
42/∞
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

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

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

- 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 $(x y^T)_{i,j} = x_i y_j$. 
Given matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, their product is

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

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

- Matrix product is **associative**: $(AB)C = A(BC)$. 
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
  
  \[ C = AB \in \mathbb{R}^{m \times p} \quad \text{where} \quad C_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j} \]

- 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

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

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

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

- Transpose of product: $(AB)^T = B^T A^T$. 

Mário A. T. Figueiredo (IST & IT)  
LxMLS 2017: Probability Theory  
June 14, 2018 43/∞
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

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

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

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

- Transpose of product: $(AB)^T = B^T A^T$.

- Transpose of sum: $(A + B)^T = A^T + B^T$. 
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 \( \ell_p \) 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 \( \ell_p \) 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 \( \ell_1 \) norm, \( \| x \|_1 = \sum_i |x_i| \).
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 \( \ell_p \) 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 \( \ell_1 \) norm, \( \| x \|_1 = \sum_i |x_i| \).

Notable case: the \( \ell_\infty \) norm, \( \| x \|_\infty = \max\{|x_1|, \ldots, |x_n|\} \).
Norms

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

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

- More generally, the \( \ell_p \) 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 \( \ell_1 \) norm, \( \|x\|_1 = \sum_i |x_i| \).

- Notable case: the \( \ell_\infty \) norm, \( \|x\|_\infty = \max \{|x_1|, \ldots, |x_n|\} \).

- Notable case: the \( \ell_0 \) “norm” (not): \( \|x\|_0 = |\{i : x_i \neq 0\}|. \)
The identity matrix $I \in \mathbb{R}^{n \times n}$ is a square matrix such that

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

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

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

- Neutral element of matrix product: $AI = IA = A$.
Special Matrices

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

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


- 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

$$I_{ij} = \begin{cases} 
1 & i = j \\
0 & i \neq j 
\end{cases}$$

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 \end{bmatrix}$$


- **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

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

- Neutral element of matrix product: \( AI = IA = A \).

- **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.

Matrix trace:

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

Matrix determinant:

\[
|A| = \det(A) = \prod_i \lambda_i
\]

Properties:

\[
|AB| = |A||B|,
\]

\[
|A^T| = |A|
\]

\[
|\alpha A| = \alpha^n |A|
\]
Eigenvalues, eigenvectors, determinant, trace

- A vector \( \mathbf{x} \in \mathbb{R}^n \) is an **eigenvector** of matrix \( A \in \mathbb{R}^{n \times n} \) if

\[
A \mathbf{x} = \lambda \mathbf{x},
\]

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

$$Ax = \lambda x,$$

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
  \[
  Ax = \lambda x,
  \]
  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 vector $x \in \mathbb{R}^n$ is an eigenvector of matrix $A \in \mathbb{R}^{n \times n}$ if

$$Ax = \lambda x,$$

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

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

Matrix trace:

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

Matrix determinant:

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

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

\[
Ax = \lambda x,
\]

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

\[ AB = BA = I. \]
Matrix Inverse

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

- The matrix $B$, such that $AB = BA = I$, is denoted $B = A^{-1}$.

- Matrix $A \in \mathbb{R}^{n \times n}$ is invertible $\iff \det(A) \neq 0$.
Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ is 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 $\iff \det(A) \neq 0$.

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.
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 $\iff \det(A) \neq 0$.

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.

- Solving system $Ax = b$, if $A$ is invertible: $x = A^{-1}b$. 
Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ is 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 $\iff$ $\det(A) \neq 0$.

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.

- Solving system $Ax = b$, if $A$ is invertible: $x = A^{-1}b$.

- Properties: $(A^{-1})^{-1} = A$, There are several algorithms to compute $A^{-1}$; general case, computational cost $O(n^3)$.
Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ is 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 $\iff$ \( \det(A) \neq 0 \).

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.

- 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 $\iff$ $\det(A) \neq 0$.

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.

- 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 $A \in \mathbb{R}^{n \times n}$ is 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 $\iff \det(A) \neq 0$.

- Determinant of inverse: $\det(A^{-1}) = \frac{1}{\det(A)}$.

- 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 $\iff$ 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 $\iff$ all $\lambda_i(A) \geq 0$.

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

• Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as

$$A = U \Sigma V^T$$
Singular Value Decomposition

- Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as
  
  $$A = U \Sigma V^T$$

- $\Sigma \in \mathbb{R}^{m \times n}$: singular values
- $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$

- Columns of $U$ are a basis for $\mathbb{R}^m$
- Columns of $V$ are a basis for $\mathbb{R}^n$

...arguably, the most important tool in linear algebra!
Singular Value Decomposition

- Any matrix \( A \in \mathbb{R}^{m \times n} \) can be factored as
  \[
  A = U \Sigma V^T
  \]
- \( \Sigma \in \mathbb{R}^{m \times n} \): singular values
- \( \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}) \)
- \( \sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(A A^T)} \)
  (non-zero eigenvalues)
Singular Value Decomposition

- Any matrix \( A \in \mathbb{R}^{m \times n} \) can be factored as
  \[
  A = U \Sigma V^T
  \]

- \( \Sigma \in \mathbb{R}^{m \times n} \): singular values
- \( \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}}) \)
- \( \sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(A A^T)} \)
  (non-zero eigenvalues)
- Columns of \( U \) are a basis for \( \mathbb{R}^m \)

![Matrix Diagram]
Singular Value Decomposition

- Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as
  \[ A = U\Sigma V^T \]
- $\Sigma \in \mathbb{R}^{m \times n}$: singular values
- $\Sigma = \text{diag}(\sigma_1, ..., \sigma_{\min\{m,n\}})$
- $\sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(AA^T)}$ (non-zero eigenvalues)
- Columns of $U$ are a basis for $\mathbb{R}^m$
- Columns of $V$ are a basis for $\mathbb{R}^n$
Singular Value Decomposition

Any matrix $A \in \mathbb{R}^{m \times n}$ can be factored as

$$A = U \Sigma V^T$$

- $\Sigma \in \mathbb{R}^{m \times n}$: singular values
- $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{\min\{m,n\}})$
- $\sigma_i = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(A A^T)}$
  (non-zero eigenvalues)
- Columns of $U$ are a basis for $\mathbb{R}^m$
- Columns of $V$ are a basis for $\mathbb{R}^n$

... arguably, the most important tool in linear algebra!
Concluding...

Enjoy LxMLS 2018!