Lecture 1: Introduction

André Martins

Deep Structured Learning Course, Fall 2019

Course Website

`https://andre-martins.github.io/pages/deep-structured-learning-ist-fall-2019.html`

There I'll post:

- Syllabus
- Lecture slides
- Literature pointers
- Homework assignments
- ...

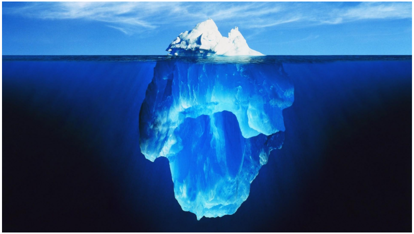
Outline

1. **Introduction**

2. Class Administrativia

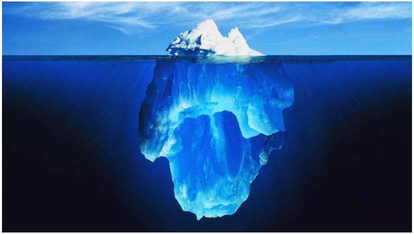
3. Recap
   - Linear Algebra
   - Probability Theory
   - Optimization

What is “‘Deep Learning”?



- Neural networks?
- Neural networks with many hidden layers?
- Anything beyond shallow (linear) models for statistical learning?
- Anything that learns representations?
- A form of learning that is really intense and profound?

What is “‘Deep Learning”?

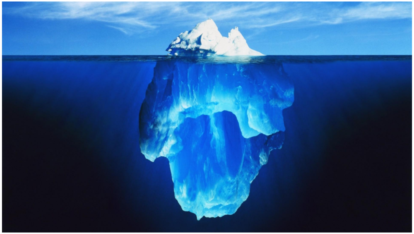


- Neural networks?
- Neural networks with many hidden layers?
- Anything beyond shallow (linear) models for statistical learning?
- Anything that learns representations?
- A form of learning that is really intense and profound?

Where is the "Structure"?

- In the **input** objects (text, graphs, images, ...)
- In the **outputs** we want to predict (parsing, graph labeling, image segmentation, ...)
- In our **model** (convolutional networks, attention mechanisms)
- Related: **latent structure** (typically a way of encoding prior knowledge into the model)

This Course: “Deep Learning + Structure”



Why Did Deep Learning Become Mainstream?

Lots of recent breakthroughs:

- Object recognition
- Speech and language processing
- Chatbots and dialog systems
- Self-driving cars
- Machine translation
- Solving games (Atari, Go)

No signs of slowing down...


smart
On my way

Microsoft's Deep Learning Project Outperforms Humans In Image Recognition

Michael Thomsen, CONTRIBUTOR

I write about tech, video games, science and culture. **FULL BIO**

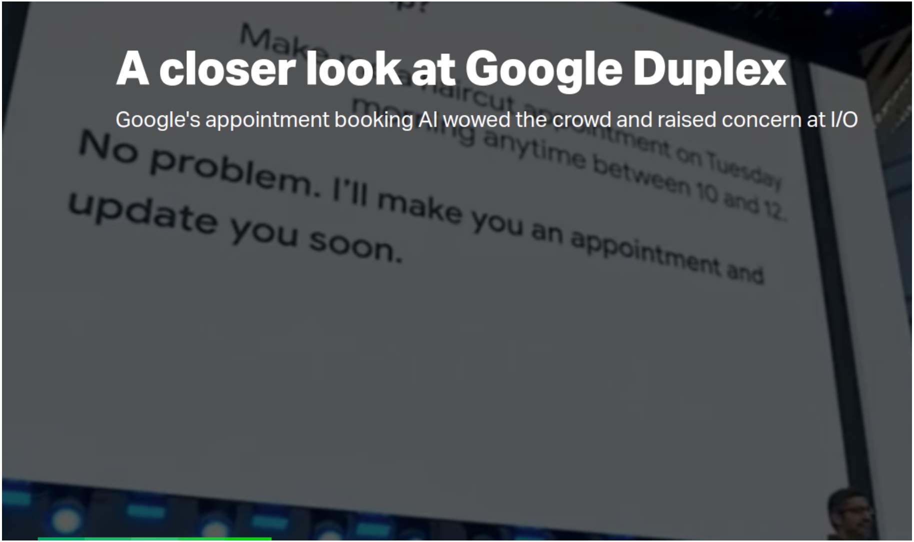
Opinions expressed by Forbes Contributors are their own.

Big Data on AWS

Microsoft's new breakthrough: AI that's as good as humans at listening... on the phone

Microsoft's new speech-recognition record means professional transcribers could be among the first to lose their jobs to artificial intelligence.

By Liam Tung | October 19, 2016 -- 10:10 GMT (11:10 BST) | Topic: Innovation


A closer look at Google Duplex
Google's appointment booking AI wowed the crowd and raised concern at I/O
No problem. I'll make you an appointment and update you soon.


Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

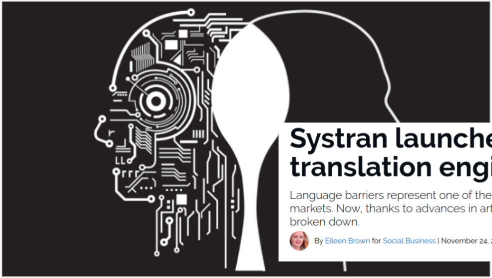
The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

BY GIDEON LEWIS-KRAUS DEC. 14, 2016

Google unleashes deep learning tech on language with Neural Machine Translation

Posted Sep 27, 2016 by Devin Coldewey, Contributor



Systran launches neural machine translation engine in 30 languages

Language barriers represent one of the biggest challenges to develop business strategies among global markets. Now, thanks to advances in artificial intelligence and machine translation, these barriers are being broken down.

By Eileen Brown for Social Business | November 24, 2016 -- 13:49 GMT (13:49 GMT) | Topic: Artificial Intelligence

Siri and Alexa Are Fighting to Be Your Hotel Butler

By **Hui-yong Yu** and **Spencer Soper**
March 22, 2017, 9:00 AM GMT *Updated on* March 22, 2017, 2:13 PM GMT

- Hotels are new frontier for voice-command technologies
- Wynn Las Vegas was first to install Alexa devices in December

AlphaGo Beats Go Human Champ: Godfather Of Deep Learning Tells Us Do Not Be Afraid Of AI

21 March 2016, 10:16 am EDT **By Aaron Mamiit Tech Times**

Last week, Google's artificial intelligence program

Last week, Google's artificial intelligence program AlphaGo dominated its match with South Korean world Go champion Lee Sedol, winning with a 4-1 score.

The achievement stunned artificial intelligence experts, who previously thought that Google's computer program would need at least 10 more years before developing enough to be able to beat a human world champion.

Why Now?

Why does deep learning work now, but not 20 years ago?

Many of the core ideas were there, after all.

But now we have:

- more data
- more computing power
- better software engineering
- a few algorithmic innovations (many layers, ReLUs, better initialization and learning rates, dropout, LSTMs, convolutional nets)

"But It's Non-Convex"

Why does gradient-based optimization work at all in neural nets despite the non-convexity?

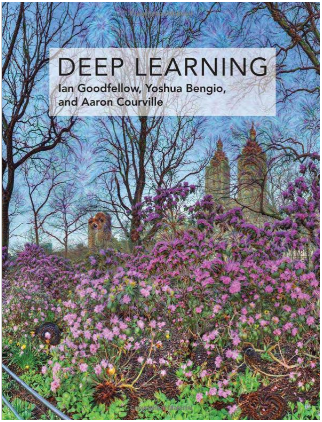
One possible, partial answer:

- there are generally many hidden units
- there are many ways a neural net can approximately implement the desired input-output relationship
- we only need to find one

Recommended Books

Main book:

- **Deep Learning.** Ian Goodfellow, Yoshua Bengio, and Aaron Courville. MIT Press, 2016. Chapters available at http://deeplearningbook.org



Recommended Books

Secondary books:

- **Machine Learning: a Probabilistic Perspective.** Kevin P. Murphy. MIT Press, 2013.
- **Linguistic Structured Prediction.** Noah A. Smith. Morgan & Claypool Synthesis Lectures on Human Language Technologies. 2011.

Tentative Syllabus

Sep 16–20	Introduction and Course Description
Sep 23–27	Linear Classifiers
Sep 30–Oct 4	Feedforward Neural Networks
Oct 7–11	Training Neural Networks
Oct 14–18	Linear Sequence Models
Oct 21–25	Representation Learning and Convolutional Networks
Oct 28–Nov 1	Structured Prediction and Graphical Models
Nov 4–8	Recurrent Neural Networks
Nov 11–15	Sequence-to-Sequence Learning
Nov 18–22	Attention Mechanisms and Neural Memories
Nov 25–29	Deep Reinforcement Learning
Dec 2–6	Deep Generative Models (VAEs, GANs)
Dec 9–13	Final Projects I
Dec 16–20	Final Projects II

Outline

1. Introduction
2. **Class Administrativia**
3. Recap
   - Linear Algebra
   - Probability Theory
   - Optimization

What This Class Is About

- Introduction to deep learning
- Introduction to structured prediction

- **Goal:** after finishing this class, you should be able to:
  - Understand how deep learning works without magic
  - Understand the intuition behind deep structured learning models
  - Apply the learned techniques on a practical problem (NLP, vision, ...)

- **Target audience:**
  - MSc/PhD students with basic background in ML and good programming skills

What This Class Is Not About

It's **not** about:

- Just playing with a deep learning toolkit without learning the fundamental concepts
- Introduction to ML (see Mário Figueiredo's Statistical Learning course and Jorge Marques' Estimation and Classification course)
- Optimization (check João Xavier's Non-Linear Optimization course)
- Natural Language Processing
- ...

Prerequisites

- Calculus and basic linear algebra
- Basic probability theory
- Basic knowledge of machine learning
- Programming (Python & PyTorch preferred)
- Helpful: basic optimization

Course Information

- **Instructors:** André Martins & Vlad Niculae
- **TAs:** Gonçalo Correia & Ben Peters
- **Location:** LT2 (North Tower, 4th floor)
- **Schedule:** Mondays/Fridays 10:00–11:30 (tentative)
- **Communication:**
  piazza.com/tecnico.ulisboa.pt/fall2019/pdeecdsl

Grading

- 4 homework assignments: 60%
  - Theoretical questions & implementation
  - Late days: 10% penalization each late day
- Final project (in groups of 2–3): 40%
  - Final class presentations & poster session (tentative)

Final Project

- **Possible idea:** apply a deep learning technique to a structured problem relevant to your research (NLP, vision, robotics, ...)
- Otherwise, pick a project from a list of suggestions
- Must be finished this semester
- Four evaluation stages: project proposal (10%), midterm report (10%), final report (10%, conference paper format), class presentation (10%)
- List of project suggestions will be made available soon

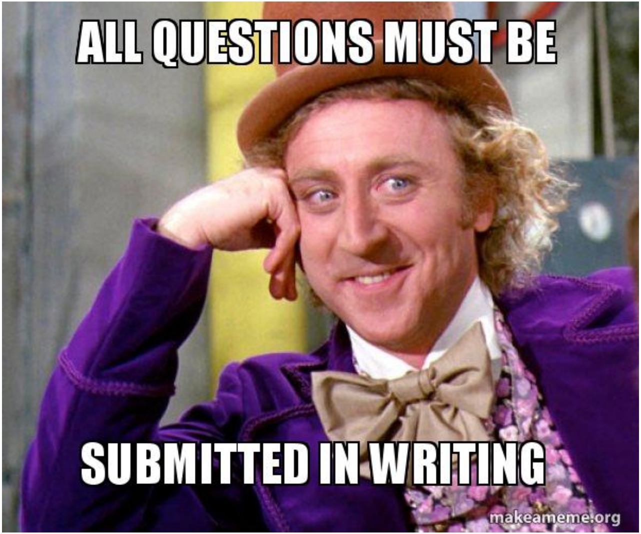
Collaboration Policy

- Assignments are individual
- Students may discuss the questions, as long as they write their own answers and their own code
- If this happens, acknowledge with whom you collaborate!
- Zero tolerance on plagiarism!!
- Always credit your sources!!!

Caveat

- This is the second year I'm teaching this class
- ... which means you're the second batch of students taking it :)
- Constructive feedback will be highly appreciated (and encouraged!)

Questions?



Outline

Quick Background Recap

Slide credits: Prof. Mário Figueiredo (taken from his LxMLS class)

Outline

1. Introduction

2. Class Administrativia

3. Recap

   Linear Algebra

   Probability Theory

   Optimization

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

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

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 $(A^T)_{i,j} = A_{j,i}$.

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

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

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

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

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

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 "magnitude." 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 "magnitude." 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 "magnitude." 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 "magnitude." 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, $\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$.

Norms

- The norm of a vector is (informally) its "magnitude." 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

$$Ax = \lambda x,$$

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

$$Ax = \lambda x,$$

  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

$$Ax = \lambda x,$$

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

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

- Matrix trace:

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

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

$$Ax = \lambda x,$$

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: $|AB| = |A||B|, \quad |A^T| = |A|,$

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: $\det(A^{-1}) = \frac{1}{\det(A)}$.
- Solving system $Ax = b$, if A is invertible: $x = A^{-1}b$.

Matrix Inverse

- Matrix $A \in \mathbb{R}^{n \times n}$ 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: $(A^{-1})^{-1} = A$, $(A^{-1})^T = (A^T)^{-1}$, $(AB)^{-1} = B^{-1}A^{-1}$

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 many 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 $x \in \mathbb{R}^n$, $x^T A x \geq 0$.

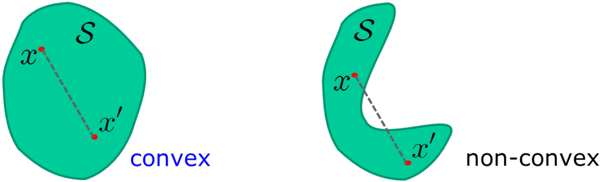
- A symmetric matrix $A \in \mathbb{R}^{n \times n}$ is positive definite (PD) if, for any $x \in \mathbb{R}^n$, $(x \neq 0) \Rightarrow x^T A x > 0$.

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

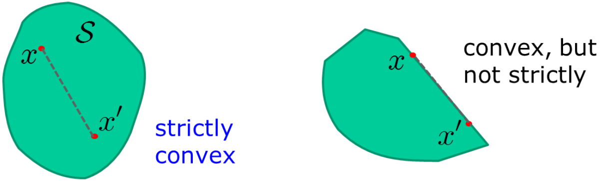
- Matrix $A \in \mathbb{R}^{n \times n}$ is PD $\Leftrightarrow$ all $\lambda_i(A) > 0$.

Convex and strictly convex sets

$\mathcal{S}$ is convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in [0,1], \quad \lambda x + (1-\lambda)x' \in \mathcal{S}$



$\mathcal{S}$ is strictly convex if $x, x' \in \mathcal{S} \Rightarrow \forall \lambda \in (0,1), \quad \lambda x + (1-\lambda)x' \in \mathrm{int}(\mathcal{S})$



Convex and strictly convex functions

Extended real valued function: $f : \mathbb{R}^N \to \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$

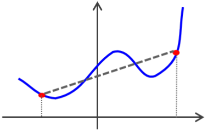
Domain of a function: $\mathrm{dom}(f) = \{x : f(x) \neq +\infty\}$

f is a convex function if

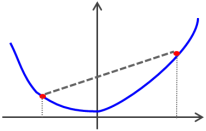
$$\forall \lambda \in [0,1], x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') \leq \lambda f(x) + (1-\lambda) f(x')$$

f is a strictly convex function if

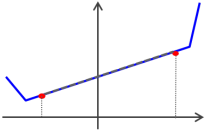
$$\forall \lambda \in (0,1), x, x' \in \mathrm{dom}(f) \;\; f(\lambda x + (1-\lambda)x') < \lambda f(x) + (1-\lambda) f(x')$$



non-convex



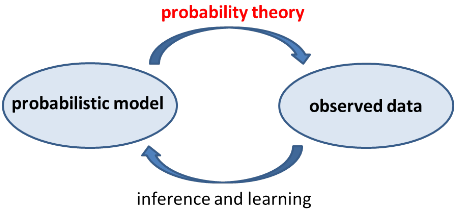
convex
strictly convex



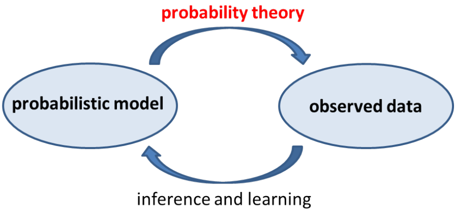
convex, not strictly

Outline

Probability theory

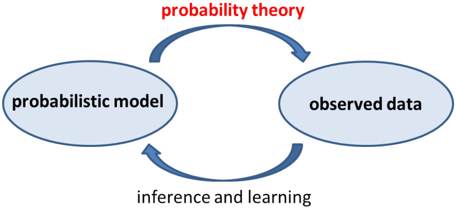


Probability theory



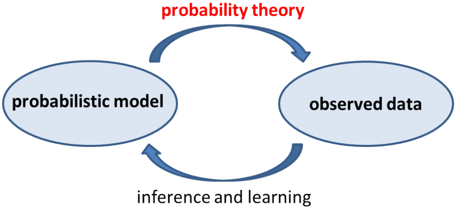
- "Essentially, all models are wrong, but some are useful"; G. Box, 1987

Probability theory



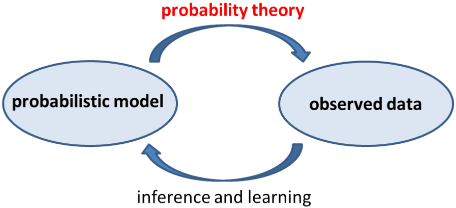
- "Essentially, all models are wrong, but some are useful"; G. Box, 1987
- The study of probability has roots in games of chance (dice, cards, ...)

Probability theory



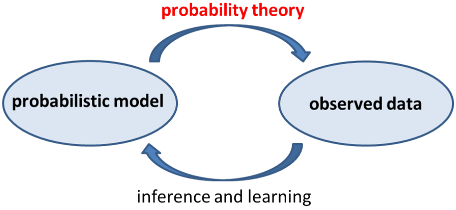
- "Essentially, all models are wrong, but some are useful"; G. Box, 1987
- The study of probability has roots in games of chance (dice, cards, ...)
- Great names in science: Cardano, Fermat, Pascal, Laplace, Gauss, Huygens, Legendre, Poisson, Kolmogorov, Bernoulli, Cauchy, Gibbs, Boltzman, de Finetti, ...

Probability theory



- "Essentially, all models are wrong, but some are useful"; G. Box, 1987
- The study of probability has roots in games of chance (dice, cards, ...)
- Great names in science: Cardano, Fermat, Pascal, Laplace, Gauss, Huygens, Legendre, Poisson, Kolmogorov, Bernoulli, Cauchy, Gibbs, Boltzman, de Finetti, ...
- Natural tool to model uncertainty, information, knowledge, belief, ...

Probability theory



- "Essentially, all models are wrong, but some are useful"; G. Box, 1987
- The study of probability has roots in games of chance (dice, cards, ...)
- Great names in science: Cardano, Fermat, Pascal, Laplace, Gauss, Huygens, Legendre, Poisson, Kolmogorov, Bernoulli, Cauchy, Gibbs, 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 event 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: $\mathbb{P}(A) = \frac{N_A}{N}$

  ...with N mutually exclusive equally likely outcomes,
  N_A of which result in the occurrence of event 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: $\mathbb{P}(A) = \lim_{N \to \infty} \frac{N_A}{N}$

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

What is probability?

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

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

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

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

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

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

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

  ...gives meaning to $\mathbb{P}$("Tomorrow it will rain").

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 $\mathcal{X}$: $A \subseteq \mathcal{X}$.

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

Kolmogorov's Axioms for Probability

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

  Kolmogorov's axioms (1933) for probability $\mathbb{P}$

Kolmogorov's Axioms for Probability

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

  Kolmogorov's axioms (1933) for probability $\mathbb{P}$

  - 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 $\mathbb{P}$

  - 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 $\mathbb{P}$

  - 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

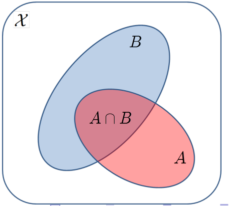
- Probability is a function that maps events A into the interval $[0,1]$.

  Kolmogorov's axioms (1933) for probability $\mathbb{P}$

  - 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

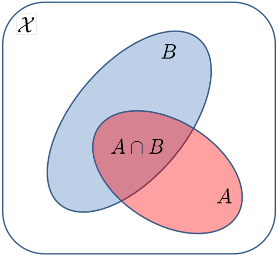
- If $\mathbb{P}(B) > 0$, $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$ (conditional prob. of A given B)

Conditional Probability and Independence

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

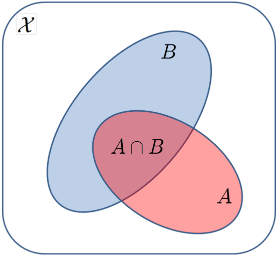


Conditional Probability and Independence

- 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 $A_1, A_2, \ldots \subseteq \mathcal{X}$ are disjoint, then $\mathbb{P}\left(\bigcup_i A_i \middle| B\right) = \sum_i \mathbb{P}(A_i|B)$



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

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

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

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:

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

- Example: $\mathcal{X} =$ "52 cards", $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\spadesuit\}$, 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 $\mathbb{P}(B) > 0$, $\qquad \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$
- 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:

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

- Example: $\mathcal{X} =$ "52 cards", $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\spadesuit\}$, 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:

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

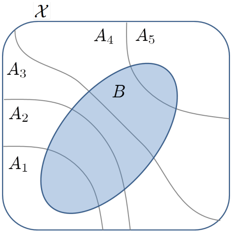
- Example: $\mathcal{X} =$ "52 cards", $A = \{3\heartsuit, 3\clubsuit, 3\diamondsuit, 3\spadesuit\}$, and $B = \{A\heartsuit, 2\heartsuit, ..., K\heartsuit\}$; then, $\mathbb{P}(A) = 1/13$, $\mathbb{P}(B) = 1/4$

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

Bayes Theorem

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

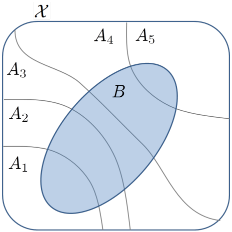
$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$
$$= \sum_i \mathbb{P}(B \cap A_i)$$



Bayes Theorem

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

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

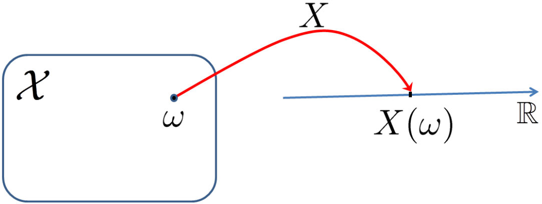


- Bayes' theorem: if $\{A_1, \ldots, A_n\}$ is a partition of $\mathcal{X}$

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

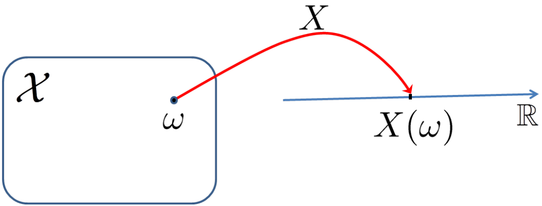
Random Variables

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



Random Variables

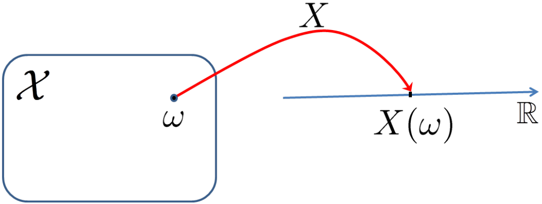
- A (real) random variable (RV) is a function: $X : \mathcal{X} \to \mathbb{R}$



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

Random Variables

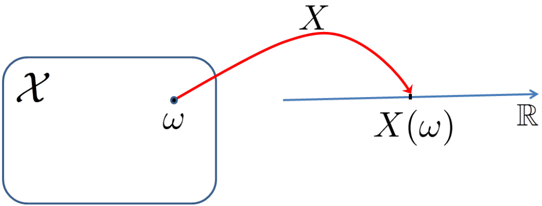
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



  - 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 $[0, 1]$)

Random Variables

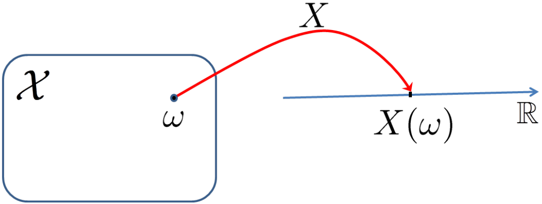
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



- 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

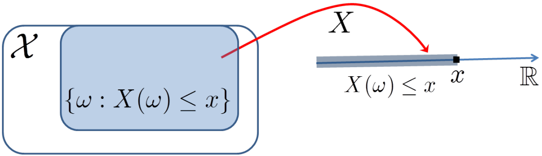
- A (real) random variable (RV) is a function: $X : \mathcal{X} \rightarrow \mathbb{R}$



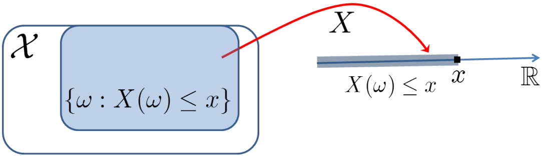
  - Discrete RV: range of X is countable (*e.g.*, $\mathbb{N}$ or $\{0, 1\}$)
  - Continuous RV: range of X is uncountable (*e.g.*, $\mathbb{R}$ or [0, 1])
  - Example: number of head in tossing two coins,
    $\mathcal{X} = \{HH, HT, TH, TT\}$,
    $X(HH) = 2$, $X(HT) = X(TH) = 1$, $X(TT) = 0$.
    Range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

Random Variables: Distribution Function

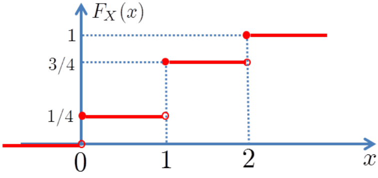
- Distribution function: $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$

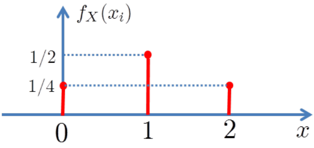


- Distribution function: $F_X(x) = \mathbb{P}(\{\omega \in \mathcal{X} : X(\omega) \leq x\})$



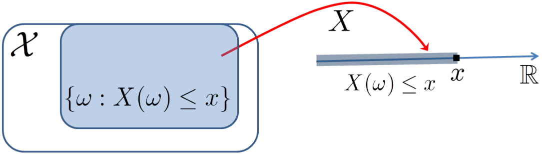
- Example: number of heads in tossing 2 coins; range$(X) = \{0, 1, 2\}$.



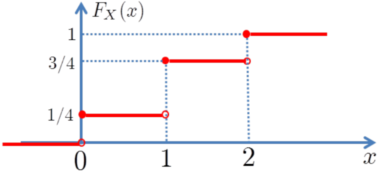


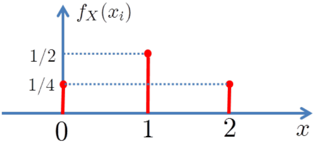
Random Variables: Distribution Function

- 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): $f_X(x) = \mathbb{P}(X = x)$,

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i).$$

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

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

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)

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

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

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

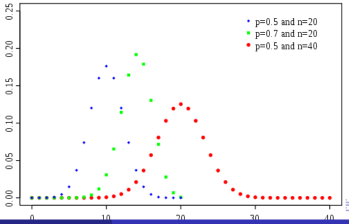
  Can be written compactly as $f_X(x) = p^x(1-p)^{1-x}$.

- 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).

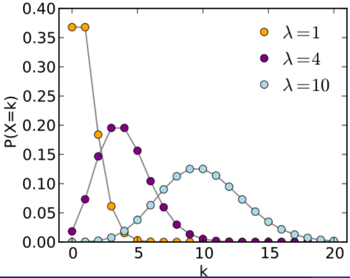
More Important Discrete Random Variables

- 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 $f_X(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

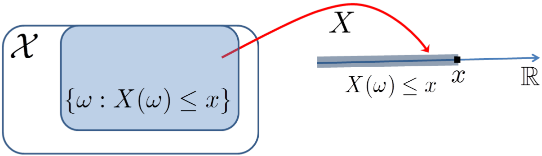
  Notice that $\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$, thus $\sum_{x=0}^{\infty} f_X(x) = 1$.

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



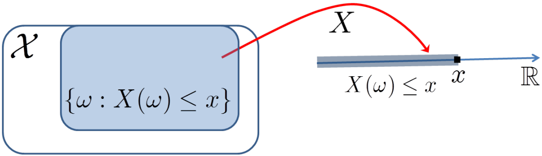
Random Variables: Distribution Function

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

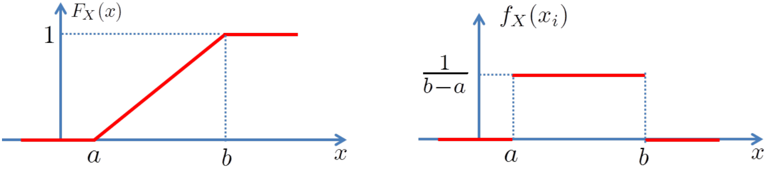


Random Variables: Distribution Function

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

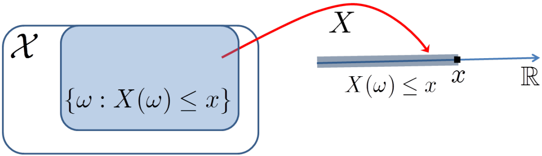


- Example: continuous RV with uniform distribution on $[a, b]$.

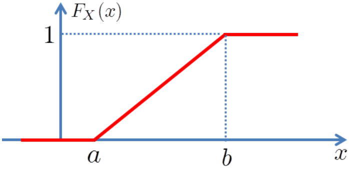


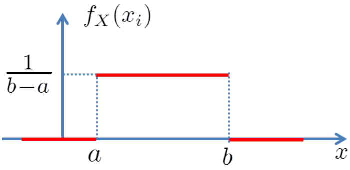
Random Variables: Distribution Function

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



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

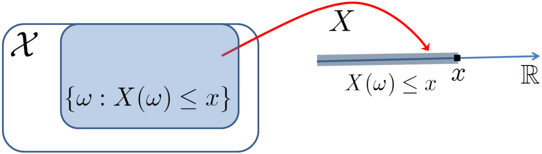




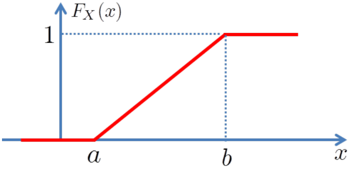
- Probability density function (pdf, continuous RV): $f_X(x)$

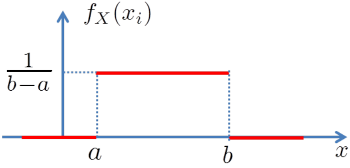
Random Variables: Distribution Function

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



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



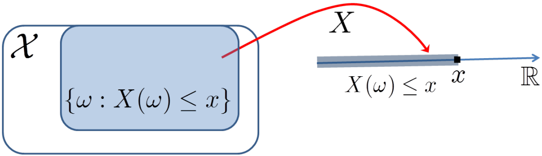


- Probability density function (pdf, continuous RV): $f_X(x)$

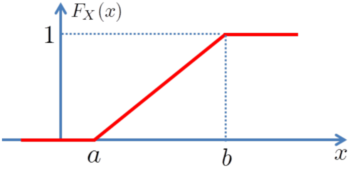
$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du,$$

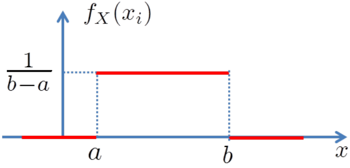
Random Variables: Distribution Function

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



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



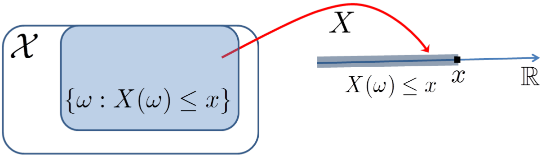


- Probability density function (pdf, continuous RV): $f_X(x)$

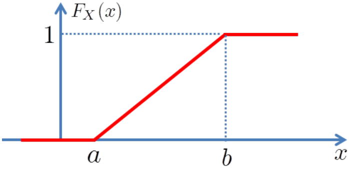
$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du, \quad \mathbb{P}(X \in [c, d]) = \int_{c}^{d} f_X(x)\, dx,$$

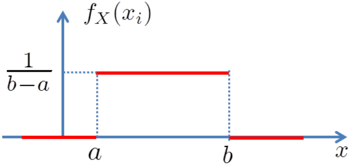
Random Variables: Distribution Function

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



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





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

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

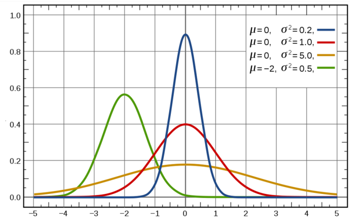
Important Continuous Random Variables

- 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).

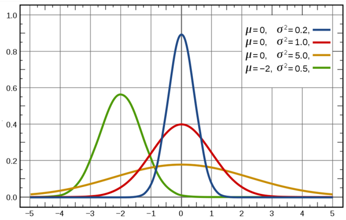
- Gaussian: $f_X(x) = \mathcal{N}(x; \mu, \sigma^2) = \dfrac{1}{\sqrt{2\,\pi\,\sigma^2}} e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}$

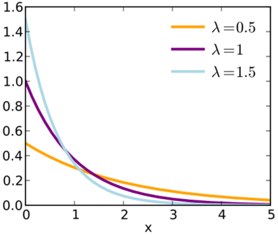


Important Continuous Random Variables

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





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

Expectation of Random Variables

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

Expectation of Random Variables

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

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

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

Expectation of Random Variables

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

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

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

Expectation of Random Variables

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

- 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 $x \in \{0, ..., n\}$.

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

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

Expectation of Random Variables

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

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

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

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

- 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

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

Expectation of Functions of Random Variables

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

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

Expectation of Functions of Random Variables

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

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

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

Expectation of Functions of Random Variables

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

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

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

Expectation of Functions of Random Variables

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

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

- 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

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

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

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

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

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

$$\mathbb{P}\big((X,Y) \in A\big) = \int\int_A f_{X,Y}(x,y)\, dx\, dy.$$

  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

$$\mathbb{P}\big((X,Y) \in A\big) = \int\!\!\int_A f_{X,Y}(x,y)\, dx\, dy.$$

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

  Extends trivially to more than two RVs.

- Joint pdf of two continuous RVs: $f_{X,Y}(x,y)$, such that
$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\, dx\, dy.$$
  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:

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

Two (or More) Random Variables

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

  Extends trivially to more than two RVs.

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

$$\mathbb{P}\big((X,Y) \in A\big) = \iint_A f_{X,Y}(x,y)\,dx\,dy.$$

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

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

Joint, Marginal, and Conditional Probabilities: An Example

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

$f_{X,Y}(x,y)$	$Y = 0$	$Y = 1$
$X = 0$	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: $X = (X_1, \ldots, X_K)$, $X_i \in \{0, \ldots, n\}$, such that $\sum_i X_i = n$,

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

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

Parameters: $p_1, \ldots, p_K \geq 0$, 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 = 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: $X \in \mathbb{R}^n$,

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

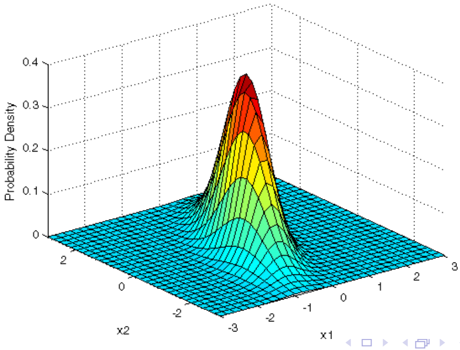
- Parameters: vector $\mu \in \mathbb{R}^n$ and matrix $C \in \mathbb{R}^{n \times n}$.
  Expected value: $\mathbb{E}(X) = \mu$. Meaning of C: next slide.

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:

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

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

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

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

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

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

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

- Relationship with variance: $\mathrm{var}(X) = \mathrm{cov}(X, X)$.
- Correlation: $\mathrm{corr}(X, Y) = \rho(X, Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}} \in [-1, 1]$
- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$

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: $\text{corr}(X, Y) = \rho(X, Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}} \in [-1, 1]$
- $X \perp\!\!\!\perp Y \;\Leftrightarrow\; f_{X,Y}(x, y) = f_X(x)\, f_Y(y) \;\underset{\not\Leftarrow}{\Rightarrow}\; \text{cov}(X,\; Y) = 0$ (example)

Covariance, Correlation, and all that...

- Covariance between two RVs:

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

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

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

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

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

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

Covariance, Correlation, and all that...

- Covariance between two RVs:

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

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

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

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

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

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

- Covariance of Gaussian RV, $f_X(x) = \mathcal{N}(x; \mu, C) \;\Rightarrow\; \mathrm{cov}(X) = C$

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

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 $\text{var}(Y) = a^T Ca \geq 0$;

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 $f_Y(y) = \mathcal{N}(y; 0, I)$.

Central Limit Theorem

Take n independent r.v. $X_1, ..., X_n$ such that $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$

Central Limit Theorem

Take n independent r.v. $X_1, ..., X_n$ such that $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$

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

$$\mathbb{E}[Y_n] = \sum_{i=1}^{n} \mu_i \equiv \mu$$

Central Limit Theorem

Take n independent r.v. $X_1, ..., X_n$ such that $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$

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

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

Central Limit Theorem

Take n independent r.v. $X_1, ..., X_n$ such that $\mathbb{E}[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$

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

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

- ...thus, if $Z_n = \dfrac{Y_n - \mu}{\sigma}$

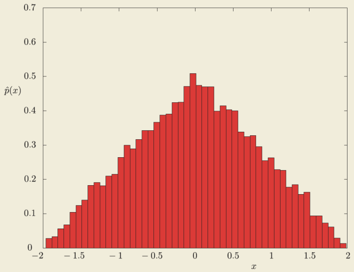
$$\mathbb{E}[Z_n] = 0 \qquad \text{var}(Z_n) = 1$$

- Central limit theorem (CLT): under some mild conditions on $X_1, ..., X_n$

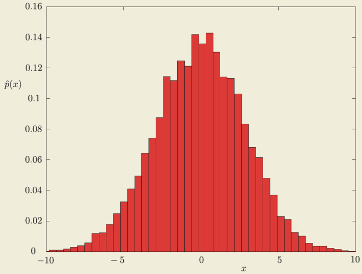
$$\lim_{n \to \infty} Z_n \sim \mathcal{N}(0, 1)$$

Central Limit Theorem

Illustration



Sum of two i.i.d variables from a uniform in [-1,1]



Sum of twenty five i.i.d r.v from a uniform in [-1,1]

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

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

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

Other Important Inequalities: Cauchy-Schwartz

- 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

- Cauchy-Schwartz's inequality for RVs:

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

...why? Because $\langle X, Y\rangle \equiv \mathbb{E}[XY]$ is a valid inner product

Other Important Inequalities: Cauchy-Schwartz

- Cauchy-Schwartz's inequality for RVs:

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

...why? Because $\langle X, Y \rangle \equiv \mathbb{E}[XY]$ is a valid inner product

- Important corollary: let $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X] = \nu$,

Other Important Inequalities: Cauchy-Schwartz

- Cauchy-Schwartz's inequality for RVs:

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

...why? Because $\langle X, Y\rangle \equiv \mathbb{E}[XY]$ is a valid inner product

- Important corollary: let $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X] = \nu$,

$$\begin{aligned}|\text{cov}(X, Y)| &= \mathbb{E}[(X - \mu)(Y - \nu)] \\ &\leq \sqrt{\mathbb{E}[(X - \mu)^2]\,\mathbb{E}[(Y - \nu)^2]} \\ &= \sqrt{\text{var}(X)\,\text{var}(Y)}\end{aligned}$$

Other Important Inequalities: Cauchy-Schwartz

- Cauchy-Schwartz's inequality for RVs:

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

...why? Because $\langle X, Y \rangle \equiv \mathbb{E}[XY]$ is a valid inner product

- Important corollary: let $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X] = \nu$,

$$\begin{aligned} |\text{cov}(X, Y)| &= \mathbb{E}[(X - \mu)(Y - \nu)] \\ &\leq \sqrt{\mathbb{E}[(X - \mu)^2]\,\mathbb{E}[(Y - \nu)^2]} \\ &= \sqrt{\text{var}(X)\,\text{var}(Y)} \end{aligned}$$

- Implication for correlation:

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

Other Important Inequalities: Jensen

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

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

- 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: Jensen

- 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: $\mathbb{E}(X)^2 \leq \mathbb{E}(X^2) \Rightarrow \text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0.$

$$\mathbb{E}(\log X) \leq \log \mathbb{E}(X), \quad \text{for } X \text{ a positive RV.}$$

Entropy and all that...

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

Entropy and all that...

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

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

Entropy and all that...

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

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

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

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

Entropy and all that...

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

- Positivity: $H(X) \geq 0$;
  $H(X) = 0 \Leftrightarrow f_X(i) = 1$, 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: $$h(X) = -\int f_X(x) \log f_X(x) \, dx$$

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

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

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

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

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

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

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

Outline

Minimizing a function

- We are given a function $f : \mathbb{R}^n \to \mathbb{R}$.
- Goal: find x^* that minimizes $f : \mathbb{R}^n \to \mathbb{R}$.
- Global minimum: for any $x \in \mathbb{R}^n$, $f(x^*) \leq f(x)$.
- Local minimum: for any $\|x - x^*\| \leq \delta \Rightarrow f(x^*) \leq f(x)$.

Are these global minima ?

- No, (local minima, saddle points, ...)

Iterative descent methods

Goal: find the minimum/minimizer of $f : \mathbb{R}^d \to \mathbb{R}$

- Proceed in **small steps** in the **optimal direction** till a **stopping criterion** is met.
- **Gradient descent**: updates of the form: $x^{(t+1)} \leftarrow x^{(t)} - \eta_{(t)} \nabla f(x^{(t)})$

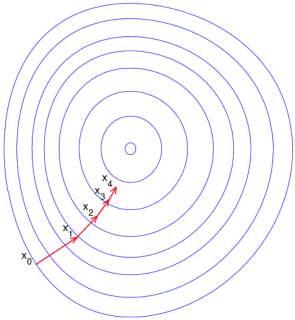


Figure: Illustration of gradient descent.The blue circles correspond to the function values at different points, while the red lines correspond to steps taken in

Convex functions

Pro: Guarantee of a global minima ✓

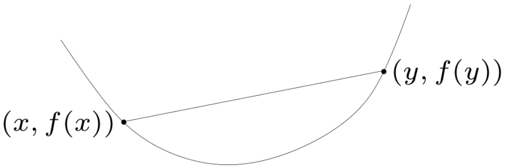


Figure: Illustration of a convex function. The line segment between any two points on the graph lies entirely above the curve.

Pro: No guarantee of a global minima ✗

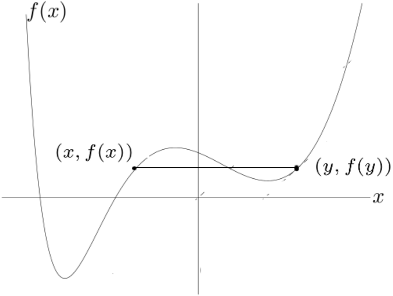


Figure: Illustration of a non-convex function. Note the line segment intersecting the curve.

Thank you!

Questions?

References I