COMP 3105

Introduction to Machine Learning

Lecture 0.5: Math Review

Junfeng Wen

Fall 2025

School of Computer Science
Carleton University

This Lecture: Math Review

Linear algebra and calculus

- Vectors
- Matrices
- Differentiation

Basic stats

- Probability
- Joint/conditional distribution
- Chain rule, Bayes' rule
- Expectation, variance

This Lecture: Math Review

Linear algebra and calculus

- Vectors
- Matrices
- Differentiation

Basic stats

- Probability
- Joint/conditional distribution
- Chain rule, Bayes' rule
- Expectation, variance

Notations

We call a single number **scalar**: -1, 0, 0.5

- Lower-case letters: a, b, c, x, y, z
- Set of **real numbers** $\mathbb{R}$
- Set of **non-negative** real numbers $\mathbb{R}_+$ or $\mathbb{R}^+$

"Is defined as" $:=$, $\doteq$, $\triangleq$

- E.g., $X \doteq$ observed number of a dice

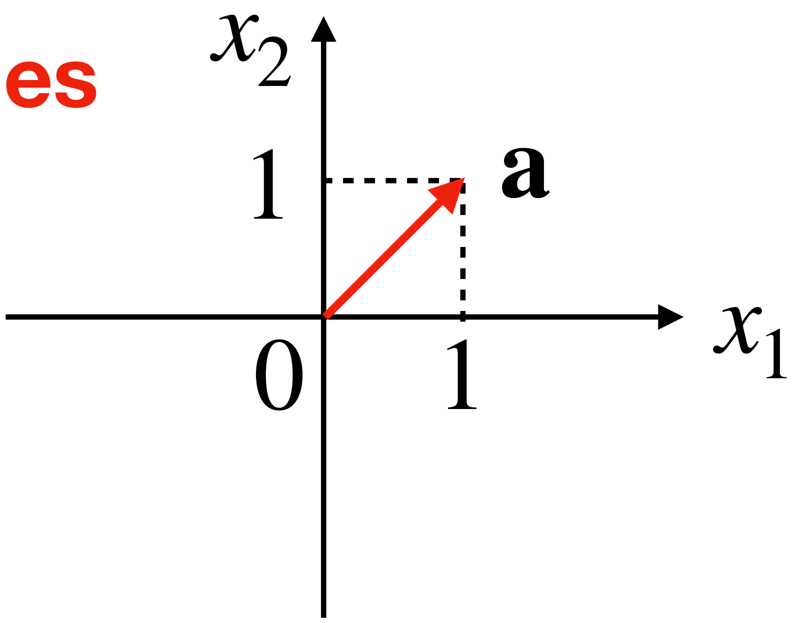
Vectors

Vectors: an array of numbers

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0.5 \\ -1 \end{bmatrix} \in \mathbb{R}^4 \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d$$

- In general, d-dimensional space $\mathbb{R}^d$
- **Bold** lower-case letters $\mathbf{a}, \mathbf{b}, \mathbf{x}, \mathbf{y}$
- By default, a **column** vector
- The numbers are also called **elements** / **entries**
- 2D example

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$



Matrices

Matrices: numbers arranged by rows and columns

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \in \mathbb{R}^{3\times 2} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

- Rows and columns
- Upper-case letters A, B, X, Y
- $X_{i:}$ the ith row of X
- $X_{:j}$ the jth column of X
- Think of colon : as "everything" (in that row / column)
- X_{ij} (or x_{ij}) the element in the ith row & jth column
- Vectors are special matrices

Transpose

Transpose

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0.5 \\ -1 \end{bmatrix} \Rightarrow \mathbf{x}^\top = \begin{bmatrix} 1 & 0 & 0.5 & -1 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \Rightarrow X^\top = \begin{bmatrix} 1 & 3 & 5 \\ 2 & 4 & 6 \end{bmatrix}$$

Basic properties

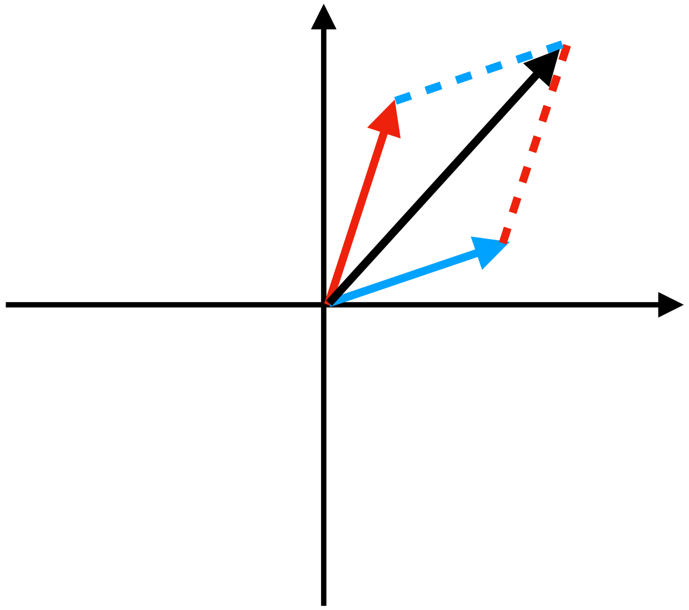
- $(A + B)^\top = A^\top + B^\top$
- $(AB)^\top = B^\top A^\top$
- If $X^\top = X$, symmetric matrix (must be square)

Vector Operations: Sum

Sum of two vectors: sum of individual elements

$$\begin{bmatrix} 1 \\ 0 \\ 0.5 \\ -1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ 0.5 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ -2 \end{bmatrix} \qquad \mathbf{x} + \mathbf{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ x_3 + y_3 \\ \vdots \\ x_d + y_d \end{bmatrix}$$

- Geometric interpretation in 2D



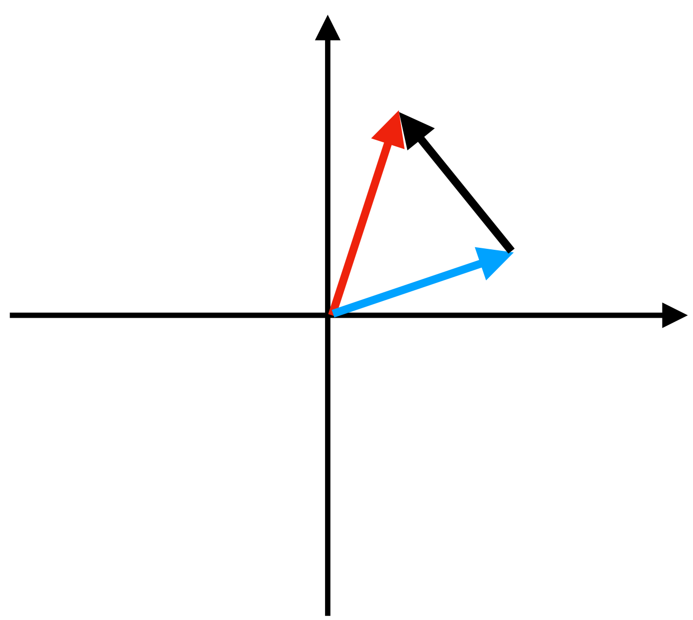
$$\mathbf{s} = \mathbf{x} + \mathbf{y}$$

Vector Operations: Difference

Difference of two vectors: difference of individual elements

$$\begin{bmatrix} 1 \\ 0 \\ 0.5 \\ -1 \end{bmatrix} - \begin{bmatrix} -1 \\ 1 \\ 0.5 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{x} - \mathbf{y} = \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \\ x_3 - y_3 \\ \vdots \\ x_d - y_d \end{bmatrix}$$

- Geometric interpretations in 2D



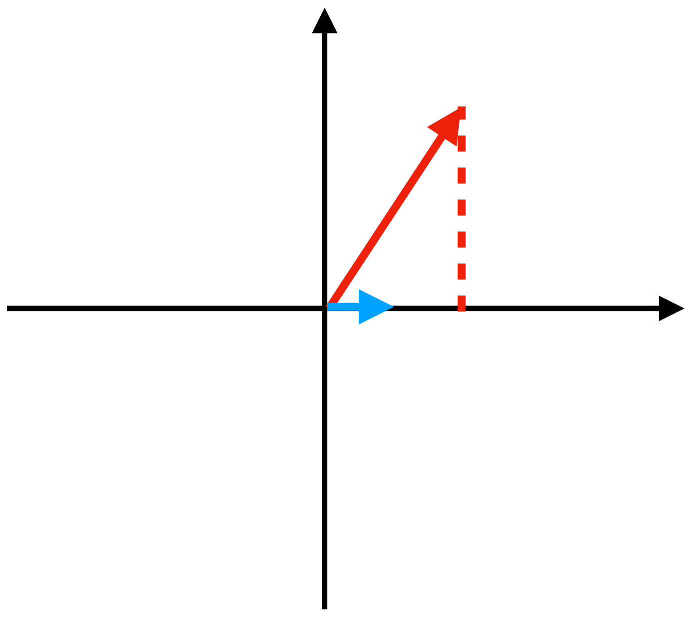
$$\mathbf{d} = \mathbf{x} - \mathbf{y}$$

Vector Operations: Inner Product

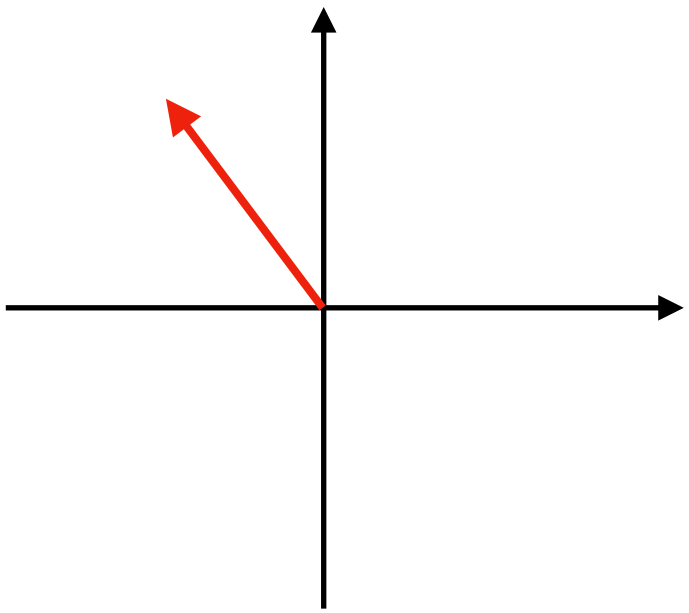
Inner product (aka dot product) of two vectors

$$\mathbf{x}^\top \mathbf{y} = [2 \quad 3] \times \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2 \qquad \mathbf{x} \cdot \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{d} x_i y_i$$

- Geometric interpretation in 2D: “projection”
- Interchangeable: $\mathbf{x}^\top \mathbf{y} = \mathbf{y}^\top \mathbf{x}$



Vector Operations: Norm



Norm of a vector: e.g. $\mathbf{x} = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$

- "Length"
- L_2 norm $\|\mathbf{x}\|_2 = \sqrt{(-3)^2 + 4^2} = 5$
- L_1 norm $\|\mathbf{x}\|_1 = |-3| + |4| = 7$

In general L_p norm: $\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p}$

Relation to inner product: $\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x} = \sum_i x_i^2$

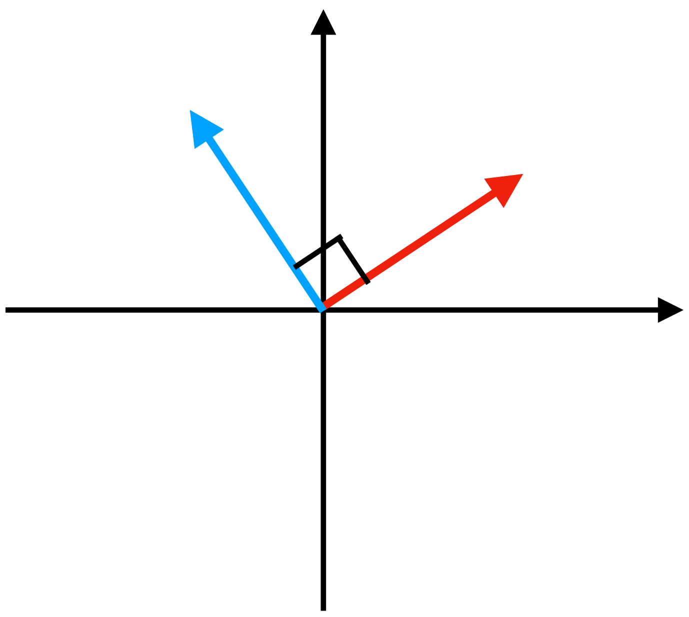
Unit vector: a vector of length 1 (i.e., its norm is 1)

Vector Operations: Orthogonality

Orthogonal vectors if $\mathbf{x}^\top \mathbf{y} = 0$

For example, in 2D $\mathbf{x} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$

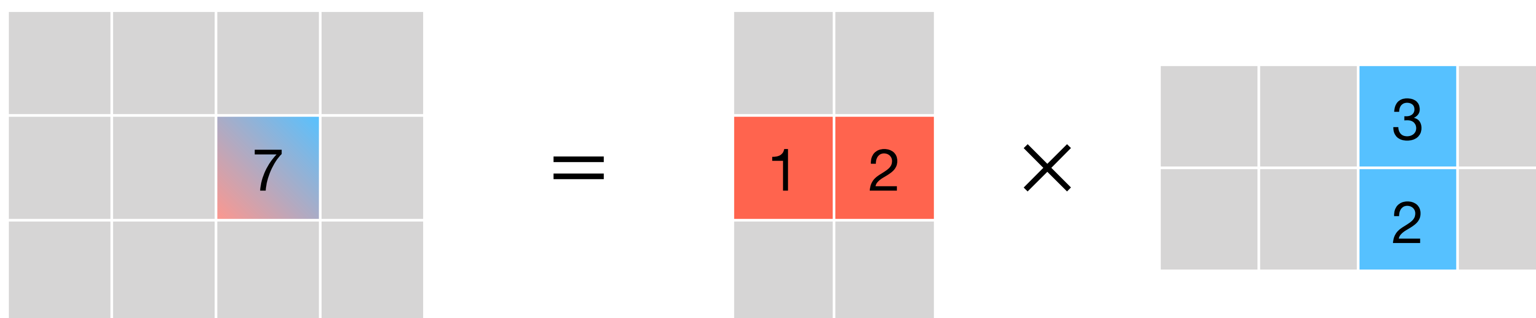
- $\mathbf{x}^\top \mathbf{y} = [3 \times (-2)] + [2 \times 3] = 0$
- Aka perpendicular



Matrix Operations: Multiplication

Matrix multiplications: matrix-matrix multiplications

$$\underset{m\times n}{C} = \underset{m\times \underline{d}}{A} \; \underset{\underline{d}\times n}{B} \qquad c_{ij} = \sum_{k=1}^{d} a_{ik} b_{kj}$$



$$\underset{3\times 4}{C} \qquad \underset{3\times 2}{A} \qquad \underset{2\times 4}{B}$$

Must match (compatible)

- Doing multiple vector inner products
- Chain $\underset{d_1\times d_5}{A} = \underset{d_1\times d_2}{A_1} \; \underset{d_2\times d_3}{A_2} \; \underset{d_3\times d_4}{A_3} \; \underset{d_4\times d_5}{A_4}$

Matrix Operations: Multiplication

Matrix-vector multiplications

- Vector is a one-column matrix

$$\underset{3\times2}{A}\ \underset{2\times1}{\mathbf{x}} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \underset{3\times1}{\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \\ a_{31}x_1 + a_{32}x_2 \end{bmatrix}}$$

- Result is again a (column) vector

Special Matrices: Identity & Inverse

Identity matrix

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad I_d = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{d \times d}$$

- $I_d \cdot \mathbf{x} = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{R}^d$

Inverse matrix: the inverse matrix of A, denoted A^{-1}

- $AA^{-1} = A^{-1}A = I$ (analogous to $a \cdot \dfrac{1}{a} = 1$ for scalar $a \neq 0$)

Basic properties

- $(A^{-1})^{-1} = A$ (analogous to $\dfrac{1}{1/a} = a$ for scalar $a \neq 0$)
- $(A^\top)^{-1} = (A^{-1})^\top$
- $(AB)^{-1} = B^{-1}A^{-1}$ if both A and B are invertible

Differentiation

Differentiate a **smooth** function $f(x)$, e.g.,

$$f(x) = (x-3)^2 \quad \Longrightarrow \quad f'(x) = 2(x-3)$$

- Chain rule $f(x) = g(h(x)) \Longrightarrow f'(x) = g'(h(x)) \times h'(x)$

Find the **minimum value** of a function $\min_x f(x)$

- In the example above $\min_x f(x) = f(3) = 0$

The **argument** that achieves the minimum value $\operatorname{argmin}_x f(x)$

- In the example above $\operatorname{argmin}_x f(x) = 3$

How about functions with **multiple** arguments?

Differentiation

For a multi-variate function $f(\mathbf{x}) : \mathbb{R}^d \mapsto \mathbb{R}$, e.g.,

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x} = x_1^2 + x_2^2$$

Partial derivative

$$\frac{\partial f(\mathbf{x})}{\partial x_1} = 2x_1 \qquad \frac{\partial f(\mathbf{x})}{\partial x_2} = 2x_2$$

The **gradient**

$$\nabla f(\mathbf{x}) \doteq \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_d} \end{bmatrix} \qquad \nabla f(\mathbf{x}) \doteq \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 2\mathbf{x}$$

- Inner product $\mathbf{x}^\top \mathbf{x}$ is **analogous** to quadratic x^2 in scalar case
- Gradient $\nabla f(\mathbf{x})$ should have the **same shape** as the argument $\mathbf{x}$

This Lecture: Math Review

Linear algebra and calculus

- Vectors
- Matrices
- Differentiation

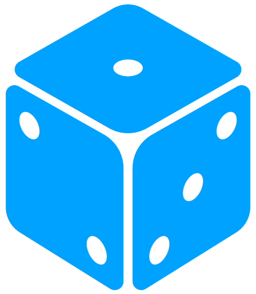
Basic stats

- Probability
- Joint/conditional distribution
- Chain rule, Bayes' rule
- Expectation, variance

Probability

The probability of an **event** $E \subseteq \Omega$ (space of all outcomes)

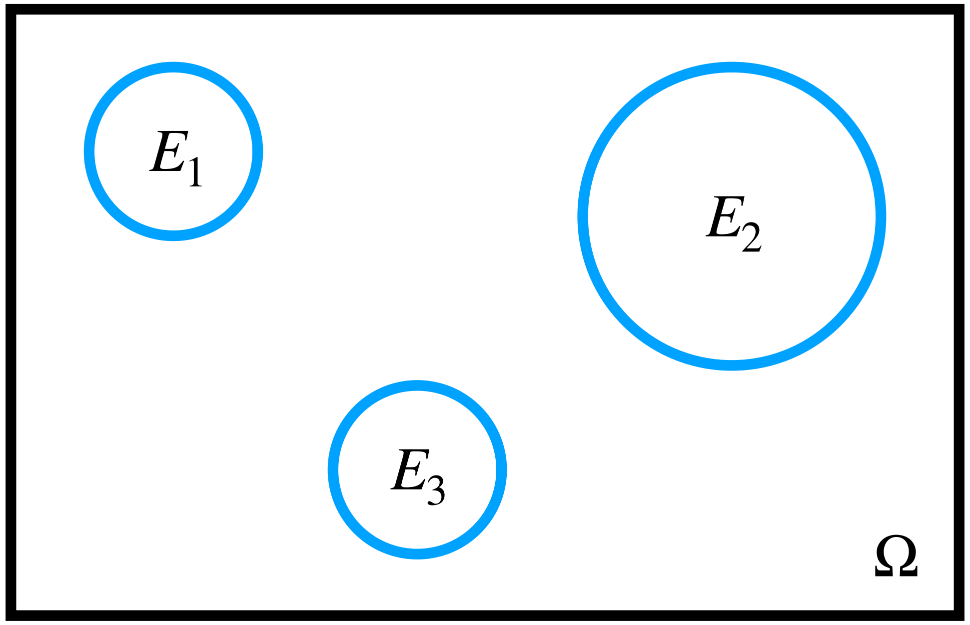
- E.g., the coin toss is heads
- E.g., the dice roll is 3
- E.g., the dice roll is odd



The probability of an event $E, P(E)$

- $P(E) \geq 0$
- $P(\Omega) = 1$
- For disjoint events $E_1, E_2, \ldots$

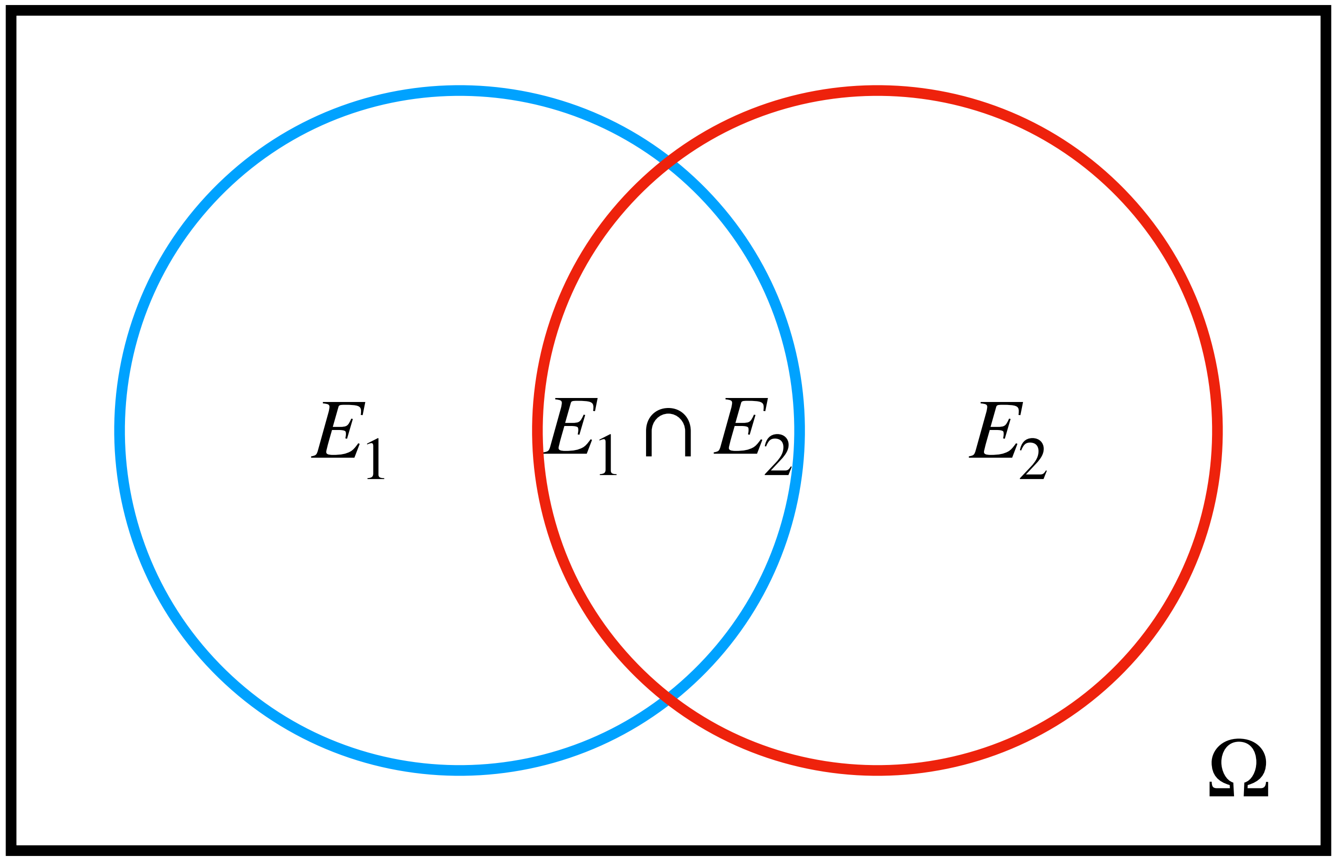
$$P\left(\cup_i E_i\right) = \sum_i P(E_i)$$



Joint and Conditional Probability

Joint probability of two events $P(E_1, E_2) \doteq P(E_1 \cap E_2)$

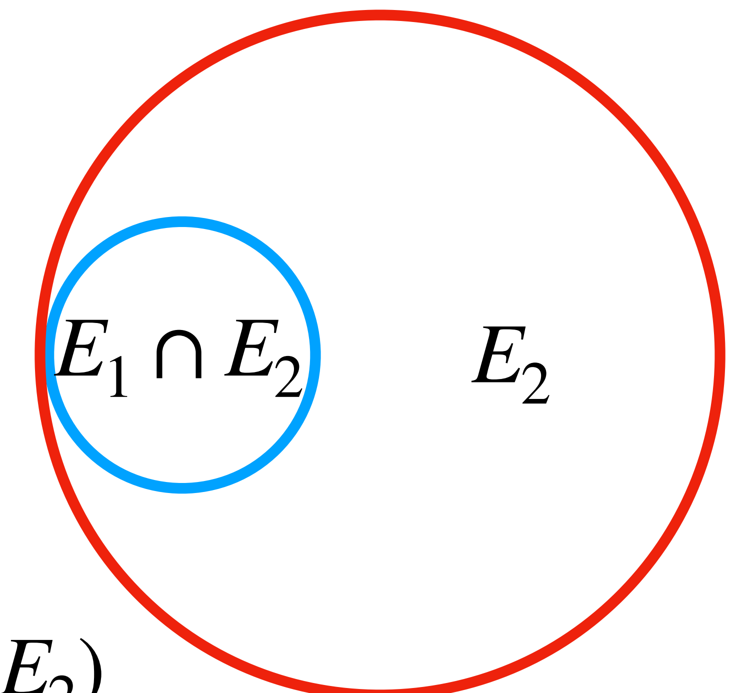
- E_1: It will rain tomorrow
- E_2: The temperature will be below 10C tomorrow
- $E_1 \cap E_2$: Tomorrow will rain **and** below 10C



Joint and Conditional Probability

Conditional probability $P(E_1 \mid E_2) = \dfrac{P(E_1, E_2)}{P(E_2)}$

- Probability of E_1 knowing that E_2 happens, e.g.,
- E_1: The dice roll is 3
- E_2: The dice roll is odd
- $P(E_1 \mid E_2) = 1/3$



Chain rule for events $P(E_1, E_2) = P(E_1 \mid E_2) \times P(E_2)$

- $P(E_2) = 1/2$: 50% to be odd
- $P(E_1 \mid E_2) = 1/3$: 33.3% to be 3 knowing that it's odd
- $P(E_1, E_2) = P(E_1 \mid E_2) \times P(E_2) = 1/3 \times 1/2 = 1/6$

Independent Events / Probabilities

Two events are **independent** if $P(E_1, E_2) = P(E_1)P(E_2)$

Independent example

- E_1: The first coin toss is heads
- E_2: The second coin toss is heads
- $P(E_1, E_2) = 0.5 \times 0.5 = 0.25$

Bayes' Rule

Recall chain rule: $P(E_1, E_2) = P(E_1 \mid E_2) \times P(E_2)$

Bayes' rule: $P(E_1 \mid E_2) = \dfrac{P(E_1, E_2)}{P(E_2)} = \dfrac{P(E_2 \mid E_1)P(E_1)}{P(E_2)}$

Exercise:

Suppose someone is tested positive ($E : T = 1$) for a disease. What is the probability of actually having the disease?

- $P(T = 1 \mid D = 1) = 1.0$ (Always identifiable)
- $P(T = 1 \mid D = 0) = 0.1$ (Sometimes false alarm)
- $P(D = 1) = 0.1$ (Generally 10% population has it)
- What is $P(D = 1 \mid T = 1)$? (How likely to be real?)

Bayes' Rule

- $P(T = 1 \mid D = 1) = 1.0$
- $P(T = 1 \mid D = 0) = 0.1$
- $P(D = 1) = 0.1$
- What is $P(D = 1 \mid T = 1)$?

First, we need

$$\begin{aligned} P(T = 1) = {} & P(T = 1 \mid D = 1)P(D = 1) \\ & + P(T = 1 \mid D = 0)P(D = 0) \\ = {} & 1.0 \times 0.1 + 0.1 \times 0.9 = 0.19 \end{aligned}$$

Then, by Bayes' rule

$$\begin{aligned} P(D = 1 \mid T = 1) &= \frac{P(T = 1 \mid D = 1)P(D = 1)}{P(T = 1)} \\ &= \frac{1.0 \times 0.1}{0.19} = 0.526 \end{aligned}$$

Random Variable

A **random variable** is a mapping from outcome to real number

$$X : \Omega \mapsto \mathbb{R}$$

- A coin is tossed 10 times
- Let X be the r.v. of number of heads in the sequence
- Observe the outcome {H, T, H, H, H, T, T, H, T, H} $\in \Omega$
- Then $X = 6$ for this outcome
- Use <u>capital letters</u> to represent R.V.: X, Y
- Use <u>lower-case letters</u> to represent its realization/value: x, y
- $P(X = x)$

Discrete & Continuous R.V.

Discrete random variable

- Take countably many values, e.g., number of a dice roll
- Distribution defined by **probability mass function** (PMF)
- Marginalization $P(X = x) = \sum_{y} P(X = x, Y = y)$

Continuous random variable

- Take uncountably many values, e.g., wait time for bus
- Distribution defined by **probability density function** (PDF)
- Marginalization $p(x) = \int_{y} p(x, y) \; dy$

Expectation

A R.V. takes various values. What's the "average" outcome?

- E.g., what's the average number of heads for 10 tosses?

The **expectation** of a r.v., denoted $\mathbb{E}[X]$, is

$$\mathbb{E}[X] = \sum_x x \cdot P(X = x)$$

For example, the expected value of a fair dice

$$\mathbb{E}[X] = \sum_{x=1}^{6} x \cdot \frac{1}{6} = 3.5$$

Some properties

- Linearity $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ for $a, b \in \mathbb{R}$
- $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ in general

Expectation Exercise

What's the expectation of the sum of two (fair) dice rolls?

X_1 value of the first roll

X_2 value of the second roll

$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 3.5 + 3.5 = 7$

Much more convenient than $2 \times \frac{1}{36} + 3 \times \frac{2}{36} + \ldots$

Variance

The **variance** of a r.v. characterizes how varied the outcome can be ($\mu \doteq \mathbb{E}[X]$)

$$\mathbb{V}[X] = \mathbb{E}[(X - \mu)^2]$$

- One useful expression

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2\mu X] + \mathbb{E}[\mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2\end{aligned}$$

Some properties

- $\mathbb{V}[X + a] = \mathbb{V}[X]$ for $a \in \mathbb{R}$
- $\mathbb{V}[aX] = a^2\mathbb{V}[X]$ for $a \in \mathbb{R}$
- $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$ if X, Y uncorrelated