LINEAR ALGEBRA

DATA/MSML 603: Principles of Machine Learning

Vector Space

- In this course, we will be dealing with vector space $\mathbb{R}^n \ (n \geq 1)$ with element-wise addition (vector add.) and scalar multiplication (mostly by real num., but sometimes complex num.)
  - **Vector**: $\mathbf{v} \in \mathbb{R}^n$ – an ordered finite array/list of (real) numbers

$$\mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n \qquad \text{Examples}: \mathbf{v} = \begin{bmatrix} 1.5 \\ -2.7 \\ 3.3 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 9.3 \\ -7.8 \\ -2.2 \\ 4.1 \end{bmatrix}$$

  - **Notation**
    - v_i - i-th element of $\mathbf{v}$, $(i = 1, \ldots, n)$
    - (Row) vector: $\mathbf{u} = [u_1 \ u_2 \ \ldots \ u_n]$ or $(u_1, u_2, \ldots, u_n)$
  - **Transpose** of a vector: $\mathbf{v}^T = \begin{bmatrix} 1.5 & -2.7 & 3.3 \end{bmatrix}, \ \mathbf{w}^T = \begin{bmatrix} 9.3 & -7.8 & -2.2 & 4.1 \end{bmatrix}$

Vector Space

- Vector addition (element-wise addition): must have the same dimension

$$\begin{aligned}\mathbf{v}+\mathbf{w} &= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T + \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}^T \\ &= \begin{bmatrix} v_1+w_1 & v_2+w_2 & \cdots & v_n+w_n \end{bmatrix}^T\end{aligned}$$

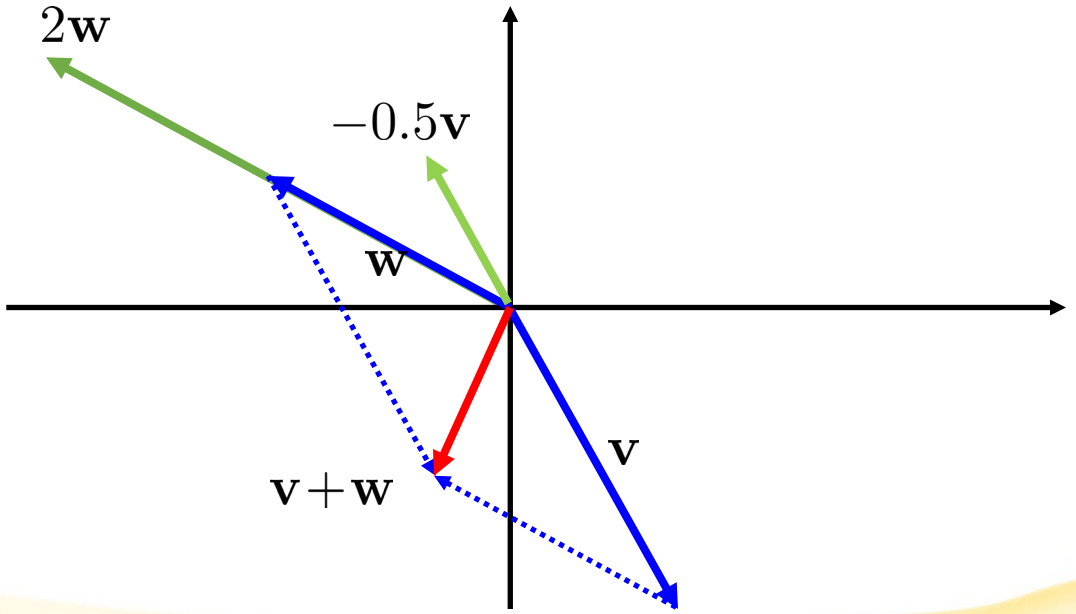
- Scalar multiplication: $r \in \mathbb{C}$

$$r \cdot \mathbf{v} = r \cdot \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^T = \begin{bmatrix} rv_1 & rv_2 & \cdots & rv_n \end{bmatrix}^T$$

$$\text{Example}: \mathbf{v} = \begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 9 \\ -7 \\ -2 \end{bmatrix} \qquad \mathbf{v}+\mathbf{w} = \begin{bmatrix} 10 \\ -9 \\ 1 \end{bmatrix} \quad \text{and} \quad 3 \cdot \mathbf{v} = \begin{bmatrix} 3 \\ -6 \\ 9 \end{bmatrix}$$

Vector Space

- **Geometric picture:** $\mathbf{v} = [1.5 \;\; -2.7]^T, \;\; \mathbf{w} = [-2 \;\; 1]^T$
  1. Vector addition: $\mathbf{v} + \mathbf{w} = [-0.5 \;\; -1.7]^T$
  2. Scalar multiplication: $-0.5\mathbf{v}, \;\; 2\mathbf{w}$



Vector Space

- (Standard) unit vectors: $\mathbf{e}_i$ - zero-one vector whose only non-zero element is the *i*-th element

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \ \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \ \text{and} \ \ \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \qquad (\text{in } \mathbb{R}^3)$$

  - $(\mathbf{e}_i)_j = \delta_{i,j} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

- **Zero** and **one** vectors

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \ \text{and} \ \mathbf{0} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T$$

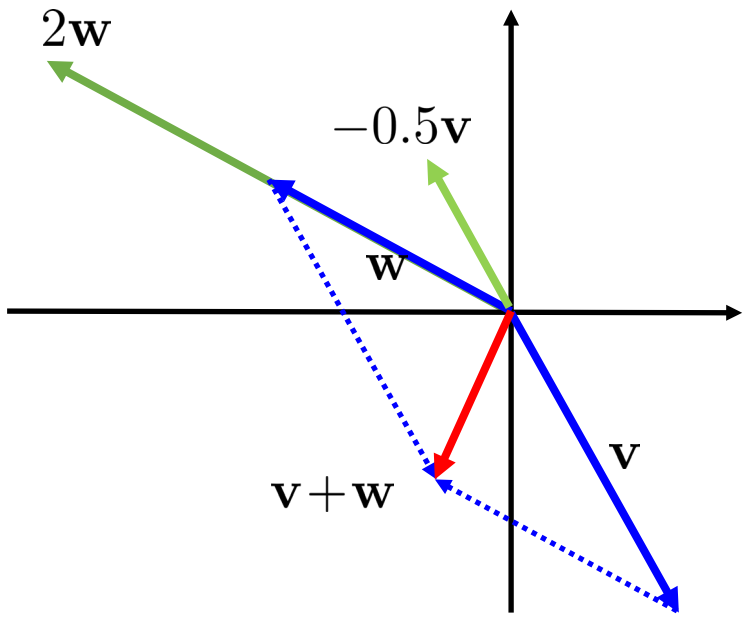
Inner Product

- Inner product between two vectors in $\mathbb{R}^n \ (n \geq 1)$
  - $$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{n} v_i w_i = \langle \mathbf{w}, \mathbf{v} \rangle = ||\mathbf{v}|| \cdot ||\mathbf{w}|| \cos(\theta), \ \ \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$$
    - Also called the dot product $(\theta = \text{angle between } \mathbf{v} \text{ and } \mathbf{w})$
- **Definition:** $\mathbb{R}^n \ (n \geq 1)$ endowed with the dot product is called an **inner product space**
- **Definition:** For $\mathbf{v} \in \mathbb{R}^n$, we define the norm (or length) of $\mathbf{v}$ by

$$||\mathbf{v}|| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle} = \sqrt{\sum_{i=1}^{n} v_i^2}, \ \ \mathbf{v} \in \mathbb{R}^n$$

Inner Product

- Example: $\mathbf{v} = [1.5 \;\; -2.7]^T, \;\; \mathbf{w} = [-2 \;\; 1]^T$



$$\langle \mathbf{v}, \mathbf{w} \rangle = ||\mathbf{v}|| \cdot ||\mathbf{w}|| \cos(\theta), \;\; \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$$

$$\Rightarrow \;\; \theta = \cos^{-1}\left(\frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\|\mathbf{v}\| \cdot \|\mathbf{w}\|} \right)$$

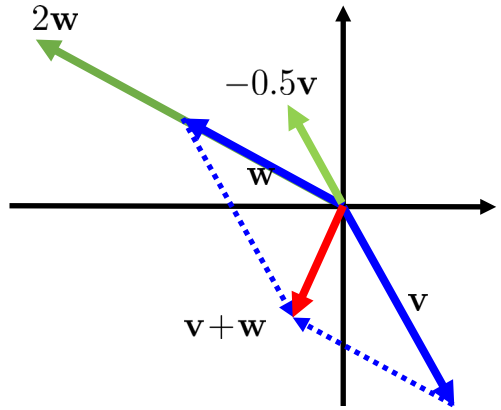
$$\langle \mathbf{v}, \mathbf{w} \rangle = -5.7$$

$$\|\mathbf{v}\| = 3.089, \|\mathbf{w}\| = 2.24$$

$$\theta = \cos^{-1}\left(\frac{-5.7}{3.089 \times 2.34} \right) = 2.54 \text{ rads}$$

$$(\approx 145.5 \text{ degrees})$$

Inner Product



- **Theorem:** For all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and $r \in \mathbb{C}$, the following hold:

  a) $||r\mathbf{v}|| = |r| \cdot ||\mathbf{v}||$

  b) $||\mathbf{v}|| = 0$ if and only if $\mathbf{v} = \mathbf{0}$

  c) **Cauchy-Schwarz inequality:** $|\langle \mathbf{v}, \mathbf{w} \rangle| \leq ||\mathbf{v}|| \cdot ||\mathbf{w}||$ $\quad (\langle \mathbf{v}, \mathbf{w} \rangle = ||\mathbf{v}|| \cdot ||\mathbf{w}|| \cos(\theta))$

  d) Triangle inequality: $||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}||$

- **Definition:** Two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^n$ are said to be orthogonal (or perpendicular) if $\langle \mathbf{v}, \mathbf{w} \rangle = 0$

$$\langle \mathbf{v}, \mathbf{w} \rangle = 0 = ||\mathbf{v}|| \cdot ||\mathbf{w}|| \cos(\theta) \Rightarrow \cos(\theta) = 0 \Leftrightarrow \theta = \frac{\pi}{2}$$

Matrices

- Matrix – a rectangular array of numbers
  - We will (mostly) deal with real-numbered matrices

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & & a_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

m = # of rows
n = # of columns

- $a_{i,j}$ - element of A in the i-th row and j-th column $(1 \leq i \leq m,\ 1 \leq j \leq n)$

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 7 \end{bmatrix} \in \mathbb{R}^{2 \times 3}$$

$$a_{1,2} = 1,\ a_{2,3} = 7$$

Matrix Addition

- Suppose $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ and $\mathbf{C} \in \mathbb{R}^{n \times p}$
- Addition of two matrices of the same dimension

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{1,1} + b_{1,1} & a_{1,2} + b_{1,2} & \cdots & a_{1,n} + b_{1,n} \\ a_{2,1} + b_{2,1} & a_{2,2} + b_{2,2} & & a_{2,n} + b_{2,n} \\ \vdots & & \ddots & \vdots \\ a_{m,1} + b_{m,1} & a_{m,2} + b_{m,2} & \cdots & a_{m,n} + b_{m,n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

  - **Elementwise** summation
- Example

$$\begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 7 \end{bmatrix} + \begin{bmatrix} 4 & 3 & 1 \\ 2 & 8 & 5 \end{bmatrix} = \begin{bmatrix} 7 & 4 & 3 \\ 7 & 12 & 12 \end{bmatrix} \in \mathbb{R}^{2 \times 3}$$

Matrix Multiplication

- Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{C} \in \mathbb{R}^{n \times p}$ (# of columns of $\mathbf{A}$ = # of rows of $\mathbf{C}$)
- Multiplication of two matrices

$$\mathbf{AC} = \begin{bmatrix} \sum_{k=1}^{n} a_{1,k}c_{k,1} & \sum_{k=1}^{n} a_{1,k}c_{k,2} & \cdots & \sum_{k=1}^{n} a_{1,k}c_{k,p} \\ \sum_{k=1}^{n} a_{2,k}c_{k,1} & \sum_{k=1}^{n} a_{2,k}c_{k,2} & \cdots & \sum_{k=1}^{n} a_{2,k}c_{k,p} \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^{n} a_{m,k}c_{k,1} & \sum_{k=1}^{n} a_{m,k}c_{k,2} & \cdots & \sum_{k=1}^{n} a_{m,k}c_{k,p} \end{bmatrix} \in \mathbb{R}^{m \times p}$$

  - Element in the i-th row and j-th column given by the product of the i-th row of $\mathbf{A}$ and the j-th column of $\mathbf{C}$

$$\begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 7 \end{bmatrix} \times \begin{bmatrix} 4 & 2 \\ 3 & 5 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 17 & 19 \\ 39 & 58 \end{bmatrix} \in \mathbb{R}^{2 \times 2}$$

- Example

Invertible Matrices and Matrix Inverse

- **Definition:** An $n \times n$ square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be invertible (or nonsingular) if there is another matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_{n \times n}$$

  Moreover, the matrix $\mathbf{B}$ is called the inverse of matrix $\mathbf{A}$ and is denoted by $\mathbf{A}^{-1}$

  - $\mathbf{B}$ is a matrix that undoes what matrix $\mathbf{A}$ does to a vector

- Example:

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{A}^{-1} = \begin{bmatrix} 0.4 & -0.2 \\ -0.1 & 0.3 \end{bmatrix}$$

Determinant

- Determinant of a square matrix
  - 2x2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}, \quad \det(\mathbf{A}) = a_{1,1} \cdot a_{2,2} - a_{1,2} \cdot a_{2,1}$$

  - 3x3 matrix

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}$$

    - Define $\tilde{\mathbf{A}}_{i,j}$ to be the submatrix of $\mathbf{A}$ after deleting the i-th row and the j-th column

$$\det(\mathbf{A}) = \sum_{j=1}^{3} (-1)^{1+j} a_{1,j} \det\left(\tilde{\mathbf{A}}_{1,j}\right)$$

Determinant

- Determinant of a square matrix
  - n x n matrix $\mathbf{A} = [a_{i,j};\ i,j \in \{1,2,\ldots,n\}]$

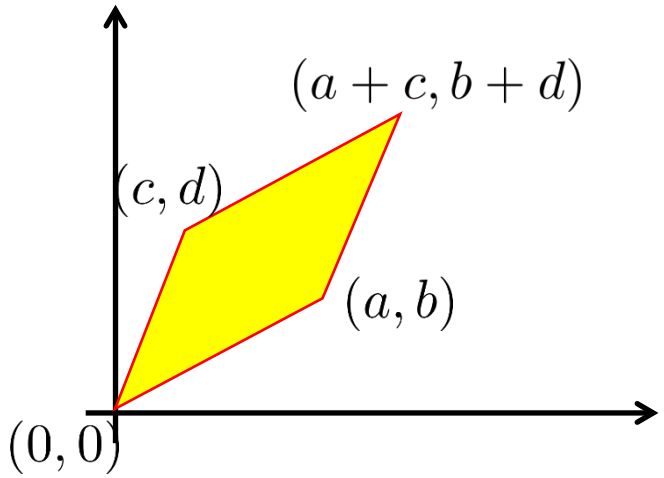
$$\det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{1+j} a_{1,j} \det\left(\tilde{\mathbf{A}}_{1,j}\right)$$

- Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{bmatrix}$$

Determinant

- Theorem: An n x n matrix $\mathbf{A}$ is invertible if and only if its determinant is non-zero
- Special case of 2x2 matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$



area of the parallelogram
$= |\det(\mathbf{A})| = |a \cdot d - b \cdot c|$

Eigenvalues & Eigenvectors

- Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$ is an n x n matrix
- **Definition:** A vector $\mathbf{v} \in \mathbb{R}^n$ is a right eigenvector of $\mathbf{A}$ if there is some constant $\lambda \in \mathbb{C}$ such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$
  - λ is called a (right) eigenvalue of $\mathbf{A}$
- **Definition:** A row vector $\mathbf{w} \in \mathbb{R}^n$ is a left eigenvector of $\mathbf{A}$ if there is some constant $\nu \in \mathbb{C}$ such that $\mathbf{w}\mathbf{A} = \nu\mathbf{w}$
  - ν is called a (left) eigenvalue of $\mathbf{A}$

Determinant and Eigenvalues

- Eigenvalues of $\mathbf{A} \in \mathbb{R}^{n \times n}$ are the roots of its characteristic polynomial given as $\det(\lambda \mathbf{I}_{n \times n} - \mathbf{A})$
- Example:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}$$

$$\begin{aligned} p_{\mathbf{A}}(\lambda) &= \det(\lambda \mathbf{I}_{2 \times 2} - \mathbf{A}) \\ &= (\lambda - a_{1,1})(\lambda - a_{2,2}) - a_{1,2} \cdot a_{2,1} \\ &= \lambda^2 - \underbrace{(a_{1,1} + a_{2,2})}_{= \mathrm{Tr}(\mathbf{A})} \lambda + \underbrace{(a_{1,1} \cdot a_{2,2} - a_{1,2} \cdot a_{2,1})}_{= \det(\mathbf{A})} \end{aligned}$$

Determinant and Eigenvalues

- Properties of determinants: $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$
  - Sum of n eigenvalues of $\mathbf{A}$ is equal to the trace of $\mathbf{A}$, i.e., $\sum_{i=1}^{n} a_{i,i}$
  - Product of n eigenvalues of $\mathbf{A}$ is equal to det($\mathbf{A}$)
  - $\det\left(\mathbf{I}_{n \times n}\right) = 1$
  - $\det\left(\mathbf{A}^T\right) = \det\left(\mathbf{A}\right)$
  - If $\mathbf{A}$ is invertible, $\det\left(\mathbf{A}^{-1}\right) = \left(\det\left(\mathbf{A}\right)\right)^{-1}$
  - $\det\left(\mathbf{AB}\right) = \det\left(\mathbf{A}\right) \times \det\left(\mathbf{B}\right)$
  - For a triangular matrix, $\det\left(\mathbf{A}\right) = \prod_{i=1}^{n} a_{i,i}$

Symmetric Matrices

- **Definition:** A real-valued matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be symmetric if $\mathbf{A} = \mathbf{A}^T$, i.e., $a_{i,j} = a_{j,i}$ for all $i, j \in [1:n]$, where $[1:n] := \{1, 2, \ldots, n\}$

- Fun facts
  - Eigenvalues of a symmetric real matrix are real
  - Eigenvectors of a symmetric real matrix corresponding to two **distinct** eigenvalues are orthogonal

- Example:

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}$$

Symmetric Matrices

- **Definition:** A symmetric real matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be positive semidefinite (or non-negative definite) if

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n$$

- **Definition:** A symmetric real matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be positive definite if

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$$

- Theorem: The eigenvalues of a positive semidefinite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are real and non-negative. Similarly, the eigenvalues of a positive definite matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ are real and positive.

Singular Values and Vectors

- Suppose $\mathbf{A} \in \mathbb{R}^{m \times n}$
- **Definition:** The singular values of $\mathbf{A}$ are the positive square roots of non-zero eigenvalues of either $\mathbf{A}^T\mathbf{A} \in \mathbb{R}^{n \times n}$ or $\mathbf{A}\mathbf{A}^T \in \mathbb{R}^{m \times m}$ (note that both are symmetric)
- **Definition:** The right singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}^T\mathbf{A}$ associated with non-zero eigenvalues. Similarly, the left singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$ associated with non-zero eigenvalues

Matrix Decomposition

- Singular value decomposition: Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix (with rank r). Then, $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

where $\mathbf{U} \in \mathbb{R}^{m \times r}$ satisfies $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{r \times r}$, $\mathbf{V} \in \mathbb{R}^{n \times r}$ satisfies $\mathbf{V}^T\mathbf{V} = \mathbf{I}_{r \times r}$, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_r)$, where $\sigma_i, i = 1, \ldots, r,$ are the singular values of $\mathbf{A}$

  - Columns of matrix $\mathbf{U}$ consist of left singular vectors of $\mathbf{A}$
  - Columns of matrix $\mathbf{V}$ consist of right singular vectors of $\mathbf{A}$

- Example:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

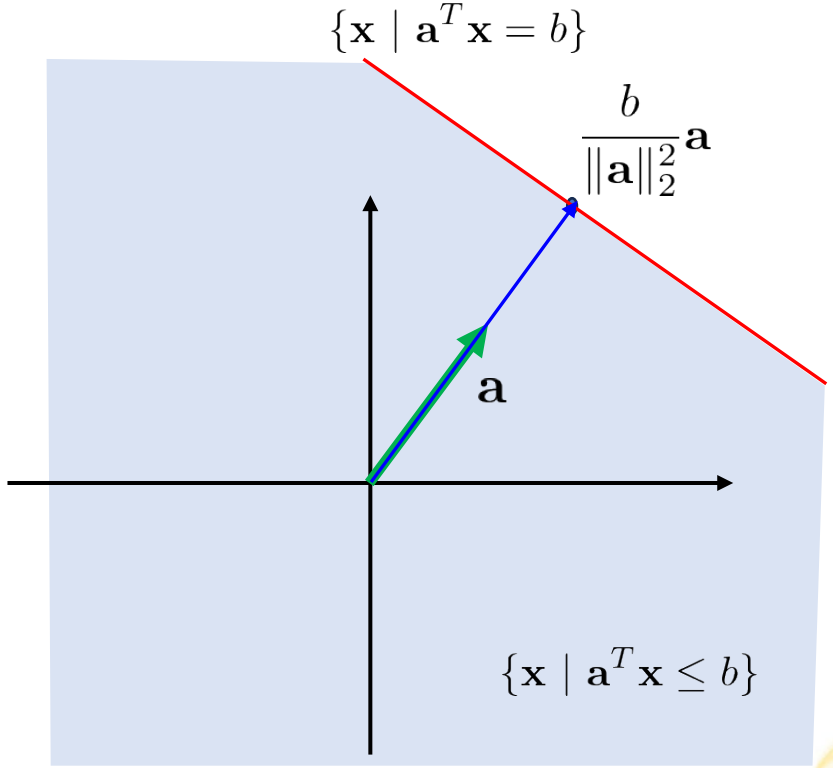
Hyperplane, Halfspace

- Hyperplane

$$\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} = b\},\ \mathbf{a} \neq \mathbf{0}$$

- Halfspace

$$\{\mathbf{x} \mid \mathbf{a}^T\mathbf{x} \leq b\},\ \mathbf{a} \neq \mathbf{0}$$



PROBABILITY

DATA/MSML 603: Principles of Machine Learning

Probabilistic Model

- Before we can talk about the "probability" of any event of interest to us (e.g., {win the jackpot}), we need a (probabilistic) model that **approximates complicated reality often with many assumptions**
- Defining a (probabilistic) model starts with a **random experiment** in which there is **uncertainty in the outcome**
  - A process or trial in which the outcome is unknown or unpredictable
  - **Examples**
    - Powerball lottery – winning numbers
    - **Coin toss** – lands on either head or tail

Probabilistic Model

Definitions/Terminology

- **Outcome**
  - "**Finest grain" result** – cannot be decomposed into other finer results
    - Example: **toss a coin three times** – Outcome: HTT

- **Sample space** (Ω)
  - Set of all possible distinguishable outcomes of a random experiment
  - Example: Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

- **Event** – a set of outcomes of an experiment, i.e., a subset of
  - Example: {outcome of first toss is H} = {HHH, HHT, HTH, HTT}

Ω

Probability

- Probability measure ($\mathbb{P}$) – assigns a real number in [0, 1] to an event
  - For event *A*, $\mathbb{P}(A)$ is the fraction of times the outcome will be in *A* if the experiment is repeated many, many times
  - Often represents the "likelihood" of an event
- Example #1: 3 coin tosses (assuming fair coin)
  - Ω = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
  - A = { first toss results in heads } = { HHH, HHT, HTH, HTT }
  - B = { at least two heads } = { HHH, HHT, HTH, THH }
  - C = { HHH, HHT, HTT}

$$\mathbb{P}(A) = \frac{1}{2},\ \mathbb{P}(B) = \frac{1}{2},\ \mathbb{P}(C) = \frac{3}{8}$$

Axioms of Probability

Notation:

$A \cup B = A \vee B, \; A \cap B = A \wedge B, \; A^c = \neg A$

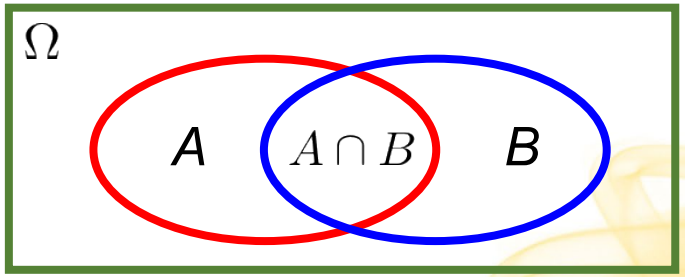
or and not

- The following are trivially true:
  - **Axiom 1:** For any event $A \subset \Omega, \mathbb{P}(A) \geq 0$
  - **Axiom 2:** $\mathbb{P}(\Omega) = 1$
  - **Axiom 3:** For any countable collection $A_1, A_2, \cdots$ of mutually exclusive events, i.e., $A_i \cap A_j = \emptyset, i \neq j$,

$$\mathbb{P}(A_1 \cup A_2 \cup \ldots) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots$$

- Consequences:
  - $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ and $\mathbb{P}(A^c) + \mathbb{P}(A) = 1$
  - For any A and B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  - If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
  - If $A_1, A_2, \ldots, A_n$ are events, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$



Joint Probability

- **Joint probability**: When dealing with multiple events of interest, describes the probability for different combinations of events
- Examples:
  - Joint probability of alarm (A) and burglary (B)

	alarm	no alarm
burglary	0.09	0.01
no burglary	0.1	0.8

Prior probability of burglary:

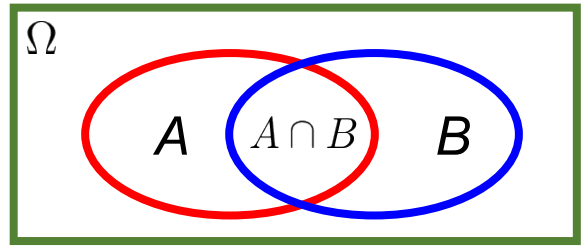
$P(\text{Burglary}) = 0.1$

by marginalization over Alarm

Conditional Probability

Definition: The conditional probability of an event *A* given another event *B* (i.e., given that we know event *B* is true) with $\mathbb{P}(B) > 0$ is

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A \wedge B)}{\mathbb{P}(B)}$$



- Given that event *B* occurred, i.e., the outcome of the experiment lies in *B*, what is the probability that the event *A* also occurred, i.e., the outcome also lies in *A*?
  - In order for the outcome to belong to *A* at the same time, it must belong to $A \cap B$
  - In general, knowing that event *B* took place changes the likelihood of event *A*

$$\mathbb{P}(A) \neq \mathbb{P}(A \mid B)$$

Conditional Probability

- Example: joint probability of alarm (A) and burglary (B) given by

	alarm	no alarm
burglary	0.09	0.01
no burglary	0.1	0.8

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0.09}{0.1} = 0.9$$

$$\mathbb{P}(A|B^c) = \frac{\mathbb{P}(A \cap B^c)}{\mathbb{P}(B^c)} = \frac{0.1}{0.9} = \frac{1}{9}$$

- Probability of **false alarm**

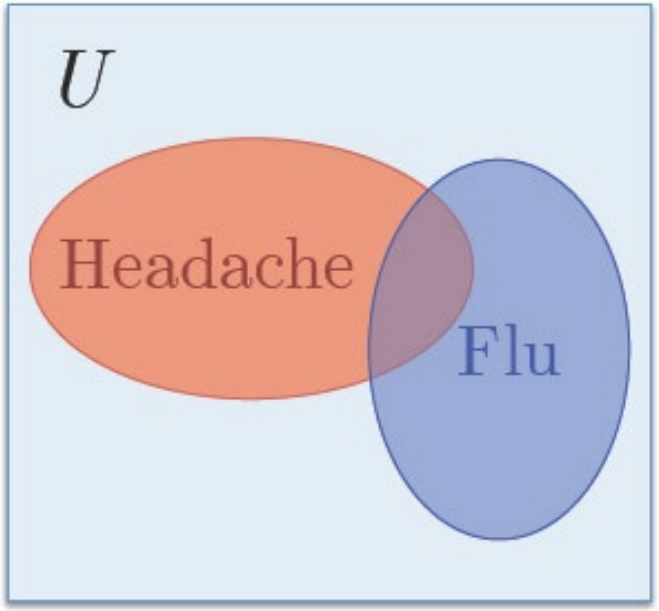
$$\mathbb{P}(A^c|B) = \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} = \frac{0.01}{0.1} = 0.1$$

- Probability of **missed detection**

Conditional Probability

What is it useful for?

- Inference – simple example



Headaches are rare and flu is rarer, but if you are coming down with the flu there is a 50-50 chance you will have a headache.

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

P(flu | headache) = ?

Bayes' Rule

- Bayes' rule: Suppose that A and B are two events with positive probabilities, i.e., $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then,

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\,\mathbb{P}(B)}{\mathbb{P}(A)}$$

  - Important formula in probabilistic machine learning
  - Allows us to reason from evidence to hypotheses

$$\text{P(hypothesis | evidence)} = \frac{\text{P(evidence | hypothesis) x P(hypothesis)}}{\text{P(evidence)}}$$

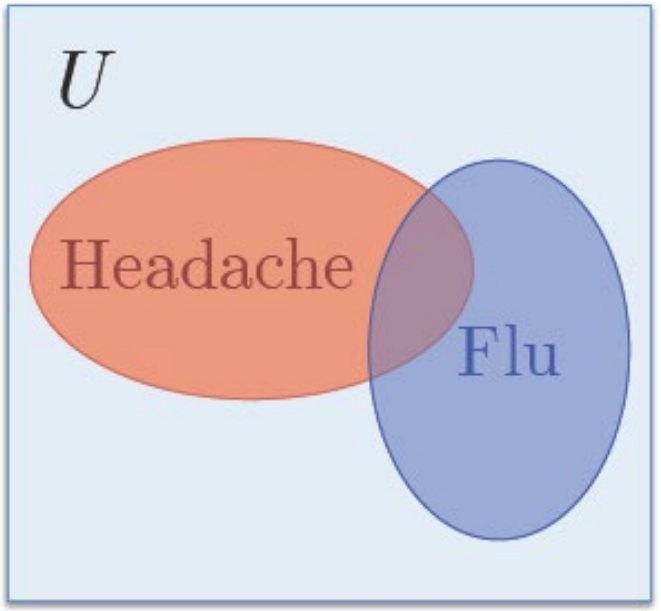
Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Bayes' Rule

- Example: P(headache) = 1/10, P(flu) = 1/40, P(headache | flu) = 1/2
  - Given headache (evidence), what is

    P(flu | headache)

    for hypothesis = {flu} ?



$$P(\text{flu} \mid \text{headache}) = \frac{P(\text{headache} \mid \text{flu}) \times P(\text{flu})}{P(\text{headache})}$$

$$= \frac{1/2 \times 1/40}{1/10} = 1/8$$

Bayes' Rule

- Example: joint probability of alarm (A) and burglary (B) given by

	alarm	no alarm
burglary	0.09	0.01
no burglary	0.1	0.8

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{0.09}{0.1} = 0.9$$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \times \mathbb{P}(B)}{\mathbb{P}(A)} = \frac{0.9 \times 0.1}{0.19} = \frac{9}{19}$$

Independence

Definition: Events A and B are "independent" if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$

- Consequence: $\mathbb{P}(A \mid B) = \mathbb{P}(A) \quad \text{and} \quad \mathbb{P}(B \mid A) = \mathbb{P}(B)$

  (assuming $\mathbb{P}(A) > 0 \quad \text{and} \quad \mathbb{P}(B) > 0$)

  - Knowing that event A or B took place does not change the "likelihood" of the other event

- A collection of events $\{A_i; i \in I\}$ is "independent" if $\mathbb{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathbb{P}(A_i)$ for all finite $J \subset I$
- A collection of events $\{A_i; i \in I\}$ is said to be "pairwise independent" if

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j) \quad \text{for all} \quad i \neq j$$

Example:

- Properties: If *A* and *B* are independent, so are
  - *A* and B^c, and
  - A^c and B^c

- Example: Suppose that there are two coins – coin 1 with P(H) = 0.5 and coin 2 with P(H) = 0.3. We randomly choose one of the two coins with equal probability of 0.5 and toss it twice. Define A = {first toss = heads} and B = {second toss = heads}. Are A and B independent?

Conditional Independence

Definition: Suppose that *A*, *B*, and *C* are events with $\mathbb{P}(C) > 0$. Then, events *A* and *B* are said to be "conditionally independent" given event *C* if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$$

Example 1: Suppose that there are two coins – coin 1 with P(H) = 0.5 and coin 2 with P(H) = 0.3. We randomly choose one of the two coins with equal probability of 0.5 and toss it twice. Define A = {first toss = heads}, B = {second toss = heads}, and C = {choose coin 1}. Are A and B conditionally independent given C?

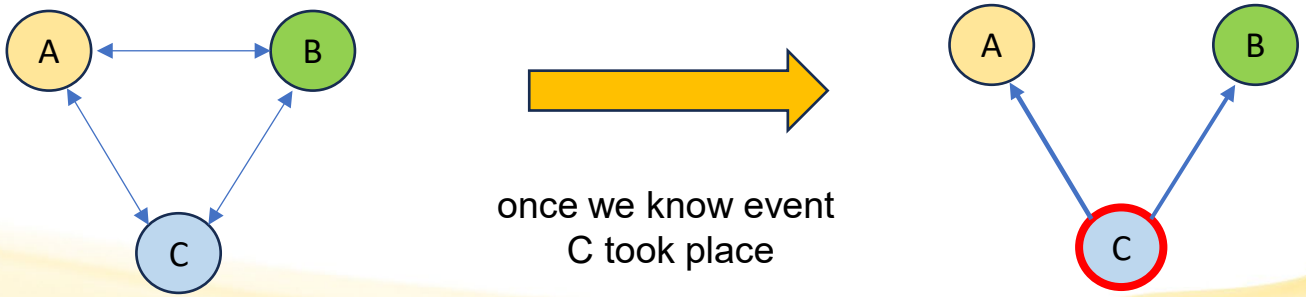
Conditional Independence

- What is conditional independence good for?
  - Often used to simplify the expression for joint probability of multiple events
- Example: In general, we know

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(C)\mathbb{P}(A \cap B \mid C) = \mathbb{P}(C)\mathbb{P}(B \mid C)\mathbb{P}(A \mid B \cap C)$$

  - When A and B are conditionally independent given C, it can be simplified somewhat

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(C)\mathbb{P}(A \cap B \mid C) = \mathbb{P}(C)\mathbb{P}(B \mid C)\mathbb{P}(A \mid C)$$



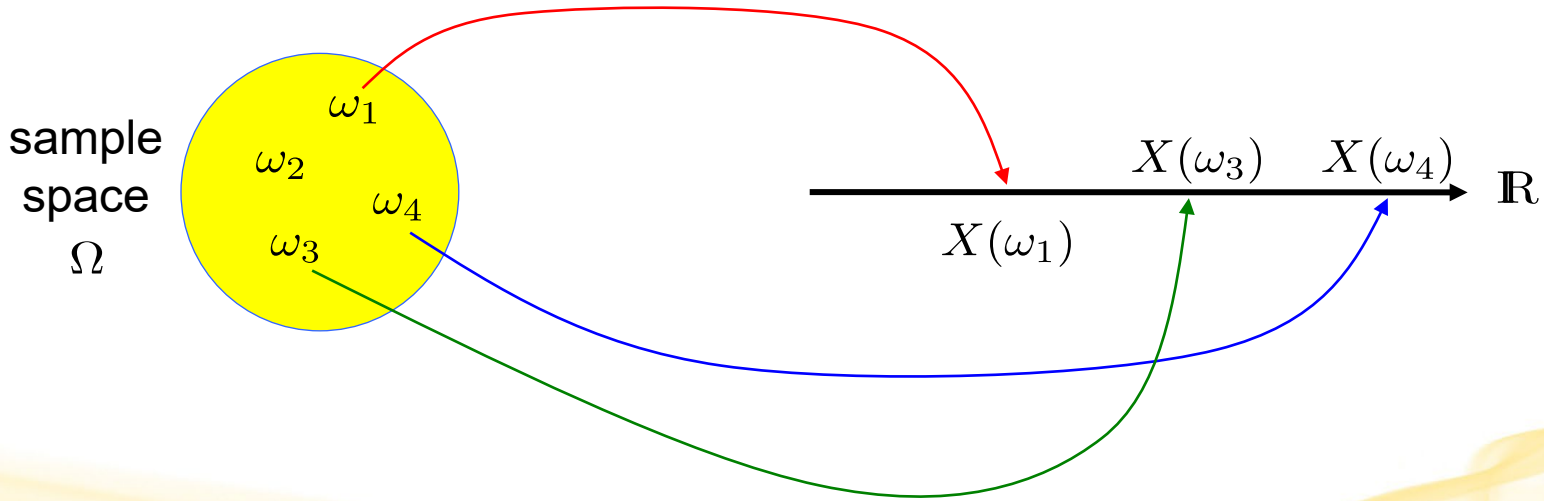
Random Variables

Definition: A (real-valued) **random variable** (RV) X is a function $X : \Omega \to \mathbb{R}$ with the property that

$$\{\omega \in \Omega \mid X(\omega) \leq x\} \in \mathcal{F}$$ ← σ-field

for each $x \in \mathbb{R}$ (X is said to be $\mathcal{F}$-measureable)

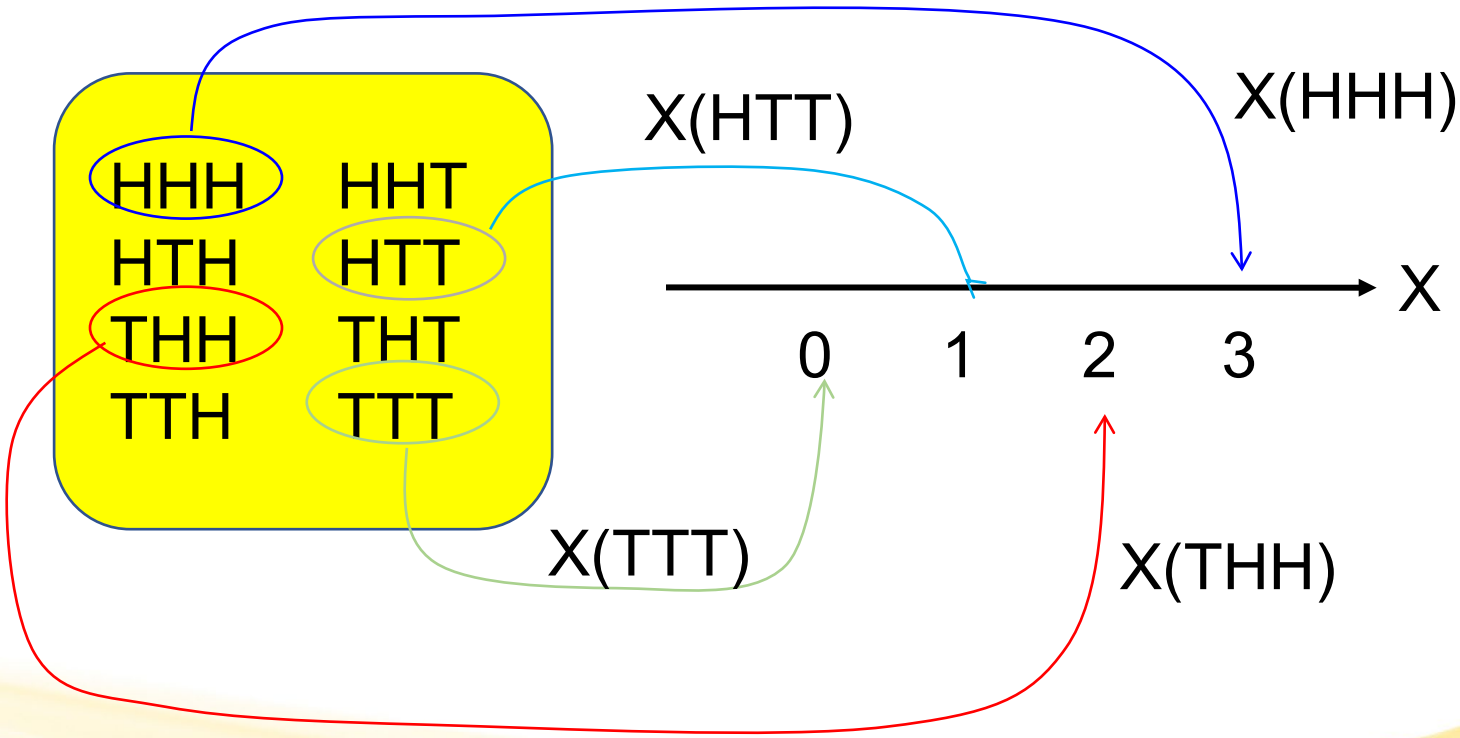
Definition: $S_X = \{X(\omega) \mid \omega \in \Omega\}$ is called the "range" of RV X



Random Variables

Example: 3 coin tosses

- X = number of heads in the outcome
- Ω = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}



X(TTT) = 0
X(TTH) = 1
X(THT) = 1
X(HTT) = 1
X(HHT) = 2
X(HTH) = 2
X(THH) = 2
X(HHH) = 3

Discrete Random Variables

Definition: A RV X is **discrete** if the range S_X of X is countable

Example #1: X = number of heads from 3 coin tosses

$$S_X = \{0, 1, 2, 3\}$$

Example #2: Y = number of tosses till the first heads

$$S_Y = \{1, 2, 3, \ldots\}$$

- e.g., Y(TTH) = 3
- What is the sample space (of the random experiment)?

Probability Mass Function

Definition: The probability mass function (PMF) of a discrete RV X is a function $p_X : \mathbb{R} \to [0, 1]$, where

$$p_X(x) = \mathbb{P}(X = x) \quad \text{for all} \quad x \in \mathbb{R}$$

Example #1: 3 coin tosses (fair coin), X = number of heads

$$\begin{aligned}
\mathbb{P}(X = 0) &= \mathbb{P}(\{TTT\}) = 1/8 \\
\mathbb{P}(X = 1) &= \mathbb{P}(\{HTT, THT, TTH\}) = 3/8 \\
\mathbb{P}(X = 2) &= \mathbb{P}(\{HHT, HTH, THH\}) = 3/8 \\
\mathbb{P}(X = 3) &= \mathbb{P}(\{HHH\}) = 1/8 \\
&\vdots
\end{aligned}$$

list all values

$$p_X(x) = \begin{cases} 1/8 & x = 0 \\ 3/8 & x = 1 \\ 3/8 & x = 2 \\ 1/8 & x = 3 \\ 0 & \text{otherwise} \end{cases}$$

Continuous Random Variable

- Recall that a discrete RV has a countable range
- Some RVs, however, can take on any value in an interval
  - Examples:
    - X = Waiting time for a shuttle bus
    - W = Water level at a reservoir
    - T = Traffic volume over the American region bridge in the morning
- Suppose that we drop a ball on a unit interval [0, 1]. What is the probability that it will land in the middle (i.e., 0.5), assuming it is equally likely to land anywhere on the interval?

Unlike with discrete RVs, we cannot identify a probability for each possible value in the range

Instead, we talk about the probability that the RV falls in some “range”