CS70: Discrete Mathematics and Probability Theory

UC Berkeley

Kelvin Lee

December 2, 2020

Contents

1 Mathematical Notations

1.1 Sets

- $\{\}$: empty set.
- $A \subset B$: A is a **proper subset** of B, i.e. A is strictly contained in B.
- $A \subseteq B$: A is a **subset** of B, i.e. A is strictly contained in B.
- $|A|$: **cardinality** of A, or the size of A.
- $A \cup B$: the **union** of A and B.
- $A \cap B$: the **intersection** of A and B.
- $A \backslash B$: **relative complement,** elements in A but not in B.
- $A \times B$: **Cartesian product**, $\{(a, b) \mid a \in A, b \in B\}$.
- $\mathscr{P}(S)$: the set of all subsets of S, also called **power set** of S.

1.2 Commonly used sets

- $\mathbb{N}$: the set of all natural numbers: $\{0, 1, 2, 3, \ldots\}$.
- $\mathbb{Z}$: the set of all integer numbers: $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$.
- $\mathbb{Q}$: the set of all rational numbers: $\left\{\frac{a}{b} \mid a, b \in \mathbb{Z}, b \neq 0\right\}$.
- $\mathbb{R}$: the set of all real numbers.
- $\mathbb{C}$: the set of all complex numbers.

1.3 Universal and existential quantifiers

- $\forall$: for all.
- $\exists$: there exists.

2 Proofs

2.1 Techniques

- **Direct Proof:** show $P \implies Q$ where P is a given fact and Q is the claim.
- **Contrapositive:** prove $\neg Q \implies \neg P$ if need to show $P \Rightarrow Q$,
- **Contradiction:** to prove claim P, assume $\neg P$ is true and arrive at $R \wedge \neg R$, which is a contradiction. Hence P is true.
- **By cases:** prove P in separate cases, if all cases are true, then P must be true.
- **Induction:** consists of three main components
  1. **Base case:** show that $P(0)$ is true.
  2. **Induction Hypothesis:** Assume $P(k)$ is true for any $k \geq 0$.
  3. **Inductive Step:** prove that $P(k+1)$ is true by showing $P(k) \implies P(k+1)$.

3 Graph Theory

3.1 Basic Terminology

Definition 1 (Graph)**.** A graph G is defined by a set of vertices V and a set of edges E. We write $G = (V, E)$.

- **Directed graph:** with directed edges, i.e., $(u, v) \neq (v, u)$.
- **Undirected graph:** with undirected edges, i.e., $(u, v) = (v, u)$.

Definition 2 (Degree)**.** The degree of a vertex v is defined by the number of edges that are incident to v. A vertex with degree 0 is an *isolated* vertex.

Definition 3 (In-degree)**.** The in-degree of a vertex v is the number of ingoing edges to v.

Definition 4 (Out-degree)**.** The out-degree of a vertex v is the number of outgoing edges from v.

Theorem 5 (Handshaking)**.** Let $G = (V, E)$ be an undirected graph with m edges. Then

$$\sum_{v \in V} \deg(v) = 2m$$

(This applies even if multiple edges and loops are present.)

Theorem 6. An undirected graph has an even number of vertices of odd degree.

Proof. Let V_1 and V_2 be the set of vertices of even degree and the set of vertices of odd degree, respectively, in an undirected graph $G = (V, E)$ with m edges. Then

$$2m = \sum_{v \in V} \deg(v) = \sum_{v \in V_1} \deg(v) + \sum_{v \in V_2} \deg(v)$$

Because $\deg(v)$ is even for $v \in V_1$, the first term in the right-hand side of the last equality is even. Furthermore, the sum of the two terms on the right-hand side of the last equality is even, because this sum is $2m$. Hence, the second term in the sum is also even. Because all the terms in this sum are odd, there must be an even number of such terms. Thus, there are an even number of vertices of odd degree. □

Theorem 7 (Euler's Theorem)**.** An undirected graph $G = (V, E)$ has an Eulerian tour iff G is even degree, and connected (except possibly for isolated vertices).

Proof. See notes. □

Definition 8 (Complete graph)**.** A graph G is called **complete** if each pair of its vertices is connected by an edge. We use K_n to denote a complete graph on n vertices.

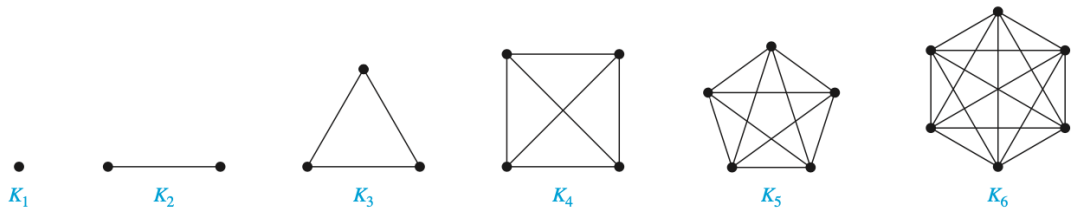


Figure 1: Examples of complete graphs.

3.2 Bipartite Graphs

Definition 9 (Bipartite)**.** A simple graph G is called **bipartite** if its vertex set V can be partitioned into two disjoint sets V_1 and V_2 such that every edge in the graph connects a vertex in V_1 and a vertex in V_2 (so that no edge in G connects either two vertices in V_1 or two vertices in V_2). We use $K_{n,m}$ to denote a complete bipartite graph partitioned into n and m vertices.

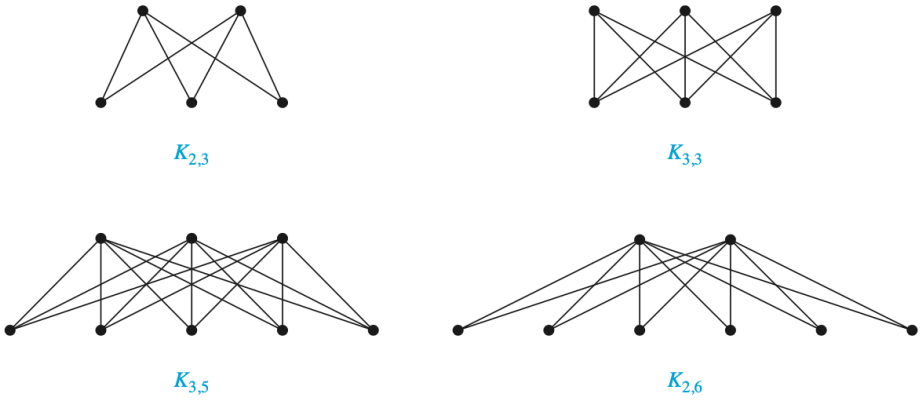


Figure 2: Examples of complete bipartite graphs.

Theorem 10. A simple graph is **bipartite** if and only if it is possible to assign one of two different colors to each vertex of the graph so that no two adjacent vertices are assigned the same color.

Proof. First, suppose that $G = (V, E)$ is a bipartite simple graph. Then $V = V_1 \cup V_2$, where V_1 and V_2 are disjoint sets and every edge in E connects a vertex in V_1 and a vertex in V_2. If we assign one color to each vertex in V_1 and a second color to each vertex in V_2, then no two adjacent vertices are assigned the same color.

Now suppose that it is possible to assign colors to the vertices of the graph using just two colors so that no two adjacent vertices are assigned the same color. Let V_1 be the set of vertices assigned one color and V_2 be the set of vertices assigned the other color. Then, V_1 and V_2 are disjoint and $V = V_1 \cup V_2$. Furthermore, every edge connects a vertex in V_1 and a vertex in V_2 because no two adjacent vertices are either both in V_1 or both in V_2. Consequently, G is bipartite. □

3.3 Connectivity

Definition 11 (Connected)**.** A graph is said to be **connected** if there is a path between any two distinct vertices.

- A **connected/disconnected** graph always consists collection of **connected components**, i.e., sets $V_1, ..., V_k$ of vertices, such that all vertices in a set V_i are connected.

3.4 Planarity

A graph is called **planar** if it can be drawn in the plane without any edges crossing (where a crossing of edges is the intersection of the lines or arcs representing them at a point other than their common endpoint).

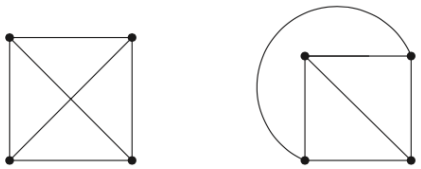


Figure 3: K_4 is planar because it can be drawn without crossing edges.

3.4.1 Euler's Formula

Theorem 12 (Euler's formula)**.** For every **connected planar** graph with v vertices, f faces, and e edges,

$$v + f = e + 2.$$

Proof. By induction on e. It clearly holds when $e = 0$, and $v = f = 1$. Now take any connected planar graph. We consider two cases:

1. If it is a tree, then $f = 1$ (drawing a tree on the plane does not subdivide the plane), and $e = v - 1$ (check homework).
2. If it is not a tree, find a cycle and delete any edge of the cycle. This amounts to reducing both e and f by one. By induction the formula is true in the smaller graph, and so it must be true in the original one.

□

Question. What happens when the graph is disconnected? How does the number of connected components enter the formula?

- Take a planar graph with f faces, and consider one face. It has a number of sides, that is, edges that bound it clockwise.
- Note that an edge may be counted twice if it has the same face on both sides (such edges are called **bridges**).
- Let s_i be the number of sides of face i. If we add the s_i's we are going to get $2e$, because each edge is counted twice.

- We conclude that, in any planar graph,

$$\sum_{i=1}^{f} s_i = 2e$$

- Notice that, since we don't allow parallel edges between the same two nodes, and if we assume that there are at least two edges (so there are at least three vertices), every face has at least three sides, or $s_i \geq 3$ for all i.
- It follows that $3f \leq 2e$.
- Solving for f and plugging into Euler's formula we get

$$e \leq 3v - 6.$$

We have just proved the following corollary:

Corollary 13. If G is a connected planar simple graph with e edges and v vertices, where $v \geq 3$, then $e \leq 3v - 6$.

Remark. This is an important fact.

- It tells us that planar graphs are sparse, they cannot have too many edges.
- It also tells us that K_5 is not planar.
- $K_{3,3}$ has $v = 6, e = 9$ so it satisfies the Euler's formula. However, using the equation above gives us $4f \leq 2e$, and solving for f and plugging into Euler's formula, $e \leq 2v - 4$, which shows that $K_{3,3}$ is non-planar.

So, we have established that K_5 and $K_{3,3}$ are both non-planar. In some sense, these are the only non-planar graphs. This is made precise in the following famous result, due to the Polish mathematician Kuratowski (this is what "K" stands for).

Theorem 14 (Kuratowski's Theorem)**.** A graph is **non-planar** if and only if it contains K_5 or $K_{3,3}$.

Proof. See notes. □

3.5 Trees

If G is a tree, then

- G is connected and contains no cycles.
- G is connected and has $n-1$ edges (where $n = |V|$).
- G is connected, and the removal of any single edge disconnects G.
- G has no cycles, and the addition of any single edge creates a cycle.

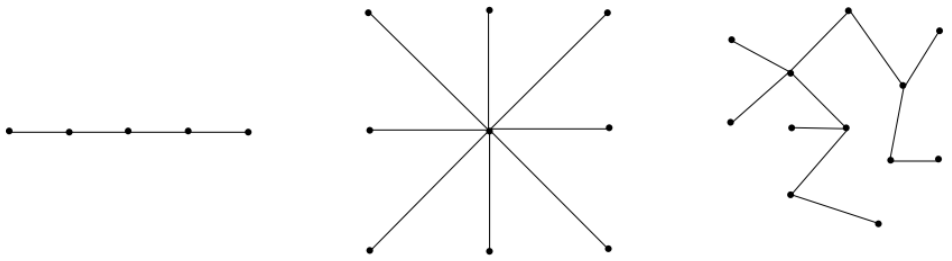


Figure 4: Examples of trees.

Theorem 15. G is connected and contains no cycles is equivalent to G is connected and has $n-1$ edges.

Proof. See notes. □

3.6 Hypercubes

- The vertex set of the n-dimensional hypercube $G = (V, E)$ is given by $V = \{0,1\}^n$, where recall $\{0,1\}^n$ denotes the set of all n-bit strings.
- Each vertex is labeled by a unique n-bit string, such as $00110 \cdots 0100$.
- Two vertices x and y are connected by edge $\{x, y\}$ if and only if x and y differ in exactly one bit position.
- For example, $x = 0000$ and $y = 1000$ are neighbors, but $x = 0000$ and $y = 0011$ are not.
- More formally, $x = x_1x_2 \ldots x_n$ and $y = y_1y_2 \ldots y_n$ are neighbors if and only if there is an $i \in \{1, \ldots, n\}$ such that $x_j = y_j$ for all $j \neq i$, and $x_i \neq y_i$.
- The n-dimensional hypercube has 2^n vertices.

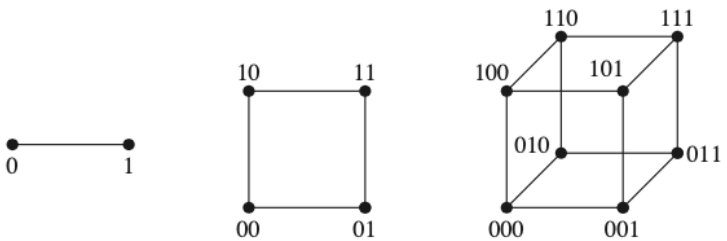


Figure 5: Examples of hypercubes.

Lemma 16. The total number of edges in an n-dimensional hypercube is $n2^{n-1}$.

Proof. The degree of each vertex is n, since n bit positions can be flipped in any $x \in \{0,1\}^n$. since each edge is counted twice, once from each endpoint, this yields a total of $n2^n/2 = n2^{n-1}$ edges. □

4 Modular Arithmetic

4.1 Congruence

Definition 17 (Congruence). x is **congruent** to y modulo m or $x \equiv y \pmod{m}$ if and only if any one of the following is true:

- $(x - y)$ is divisible by m
- x and y have the same remainder w.r.t. m
- $x = y + km$ for some integer k
- In modulo m, only the numbers $\{0, 1, 2, \ldots, m-1\}$ exist.
- Division is not well-defined.

4.2 Multiplicative Inverse

Definition 18 (Multiplicative Inverse). In the modular space, the **multiplicative inverse** of $x \bmod m$ is y if $xy \equiv 1 (\mathrm{mod} m)$.

Theorem 19 (Modular operations). $a \equiv c \bmod m$ and $b \equiv d \bmod m \implies a + b \equiv c + d \pmod{m}$ and $a \cdot b \equiv c \cdot d \pmod{m}$.

Theorem 20 (Existence of multiplicative inverse). $\gcd(x, m) = 1 \implies x$ has a multiplicative inverse modulo m and it is **unique**.

4.3 Euclid's Algorithm

Question. How do we compute gcd of two numbers x and y?

Theorem 21 (Euclid's Algorithm). Let $x \geq y > 0$. Then

$$gcd(x, y) = gcd(y, x \bmod y)$$

Example 4.1. Compute gcd(16,10):

$$\begin{aligned} \gcd(16, 10) &= \gcd(10, 6) \\ &= \gcd(6, 4) \\ &= \gcd(4, 2) \\ &= \gcd(2, 0) \\ &= 2. \end{aligned}$$

4.4 Extended Euclid's algorithm

Question. How to compute the multiplicative inverse?

- Need an algorithm that returns integers a and b such that:

$$\gcd(x, y) = ax + by.$$

Theorem 22 (Bézout's Identity)**.** For nonzero integers x and y, let d be the greatest common divisor such that $d = \gcd(x, y)$. Then, there exist integers a and b such that

$$ax + by = d.$$

- When $\gcd(x, y) = 1$, we can deduce that b is an inverse of $y \bmod x$.
- This uses back substitutions repetitively so that the final expression is in terms of x and y.

4.5 Functions

Definition 23 (Function)**.** Let A and B be nonempty sets. A **function** f from A to B is an assignment of exactly one element of B to each element of A. (vertical line test)

- To denote such a function, we write $f : A \to B$ (f maps A to B).
- A is the **domain** and B is the **co-domain**.
- Pre-image is a **subset** of domain, and the image/range is the **subset** of co-domain.
  - If $f(a) = b$, where $a \in A$ and $b \in B$, then we say that b is the image of a and a is the pre-image of b.

4.6 Bijection

Definition 24 (One-to-one)**.** A function f is said to be **one-to-one** if and only if $f(a) = f(a')$ implies that $a = a'$ for all $a, a' \in A$. A function is said to be **injective** if it is **one-to-one**.

- To show that a function is *one-to-one*, we show that $a \neq a' \implies f(a) \neq f(a')$. (Why?)

Definition 25 (Onto)**.** A function f is called **onto**, or a surjection, if and only if for every element $b \in B$ there is an element $a \in A$ such that $f(a) = b$. We also say that f is **surjective** if it's onto.

- To show that a function is *onto*, choose $a = f^{-1}(b)$ and so $f(f^{-1}(b)) = b$.

Definition 26 (Bijection)**.** A function f is a **bijection** if and only if it is both *one-to-one* and *onto*. We also say that f is bijective.

- If $f : A \to B$ is a bijection, it will have an **inverse** function (a lemma from notes), and $|A| = |B|$.

4.7 Fermat's Little Theorem

Theorem 27 (Fermat's Little Theorem)**.** For any prime p and any $a \in \{1, 2, ..., p-1\}$, we have

$$a^{p-1} \equiv 1 \pmod{p}.$$

Proof. Consider $S = \{1, 2, \ldots, p-1\}$ and $S' = \{a \bmod p, 2a \bmod p, \ldots, (p-1)a \bmod p\}$. They are the same set under mod p (different order).

$$\prod_{k=1}^{p-1} k \equiv \prod_{k=1}^{p-1} ka \pmod{p}$$

$$(p-1)! \equiv a^{p-1}(p-1)! \pmod{p}$$

$$a^{p-1} \equiv 1 \pmod{p}$$

□

4.8 Chinese Remainder Theorem

Theorem 28 (Chinese Remainder Theorem)**.** Let $n_1, n_2, \ldots, n_k$ be positive integers that are coprime to each other. Then, for any integers a_i, the system of simultaneous congruences

$$x \equiv a_1 \pmod{n_1}, x \equiv a_2 \pmod{n_2}, \ \ldots \ , x \equiv a_k \pmod{n_k}$$

has a unique solution

$$x = \left(\sum_{i=1}^{k} a_i b_i \right) \bmod N$$

where $N = \prod_{i=1}^{k} n_i$ and $b_i = \frac{N}{n_i} \left(\frac{N}{n_i} \right)^{-1}_{n_i}$ where $\left(\frac{N}{n_i} \right)^{-1}_{n_i}$ denotes the multiplicative inverse $(\mathrm{mod} n_i)$ of the integer $\frac{N}{n_i}$.

Proof. To see why x is a solution, notice that for each $i = 1, 2, \ldots, k$, we have

$$\begin{aligned} x &\equiv a_1 y_1 z_1 + a_2 y_2 z_2 + \cdots + a_k y_k z_k \pmod{n_i} \\ &\equiv a_i y_i z_i \pmod{n_i} \\ &\equiv a_i \pmod{n_i}. \end{aligned}$$

- The second line follows since $y_j \equiv 0 \bmod n_i$ for each $j \neq i$.
- The third line follows since $y_i z_i \equiv 1 \bmod n_i$.

Now, to prove uniqueness, suppose there are two solutions x and y.

- Then $n_1 \,|(x-y), n_2|\,(x-y), \ldots, n_k \mid (x-y)$.
- Since $n_1, n_2, \ldots, n_k$ are relatively prime, we have that $n_1 n_2 \cdots n_k$ divides $x - y$, or

$$x \equiv y \pmod{N}.$$

Thus, the solution is unique modulo N. □

General construction:

1. Compute $N = n_1 \times n_2 \times \cdots \times n_k$.
2. For each $i = 1, 2, \ldots, k$, compute
$$y_i = \frac{N}{n_i} = n_1 n_2 \cdots n_{i-1} n_{i+1} \cdots n_k.$$
3. For each $i = 1, 2, \ldots, k$, compute $z_i \equiv y_i^{-1} \mod n_i$ (z_i exists since $n_1, n_2, \ldots, n_k$ are pairwise coprime).
4. Compute
$$x = \sum_{i=1}^{k} a_i y_i z_i$$
and $x \mod N$ is the unique solution modulo N.

Intuitive way to solve for CRT:

1. Begin with the congruence with the largest modulus, $x \equiv a_k \,(\mathrm{mod} n_k)$.
2. Re-write this modulus as an equation, $x = j_k n_k + a_k$, for some positive integer j_k.
3. Substitute the expression for x into the congruence with the next largest modulus, $x \equiv a_k \,(\mathrm{mod} n_k) \implies j_k n_k + a_k \equiv a_{k-1} \,(\mathrm{mod} n_{k-1})$.
4. Solve this congruence for j_k.
5. Write the solved congruence as an equation, and then substitute this expression for j_k into the equation for x.
6. Continue substituting and solving congruences until the equation for x implies the solution to the system of congruences.

Example 4.2. Solve for the following system of congruences

$$\begin{cases} x \equiv 1 & (\mathrm{mod}\ 3) \\ x \equiv 4 & (\mathrm{mod}\ 5) \\ x \equiv 6 & (\mathrm{mod}\ 7) \end{cases}$$

Solution. Start with mod 7.

1. Write $x = 7k + 6$.
2. Then we have $7k + 6 \equiv 4 \pmod 5 \implies k \equiv 4 \pmod 5$.
3. Then solving for k gives $5j + 4$.
4. Now we have $x = 7k + 6 = 7(5j + 4) + 6 = 35j + 34$.
5. Then $35j + 34 \equiv 1 \pmod 3 \implies j \equiv 0 \pmod 3 \implies j = 3t$.
6. Finally, we have $x = 35(3t) + 34 = 105t + 34 \implies x \equiv \boxed{34} \pmod{105}$.

5 RSA

5.1 Basic Ideas

- Alice and Bob wish to communicate secretly over some (insecure) link, and Eve tries to discover what they are saying.
- Alice transmits a message x (in binary) to Bob by applying her **encryption function** E to x and send the encrypted message $E(x)$ over the link.
- Bob, after receiving $E(x)$, applies his **decryption function** D to it and recover the original message: i.e., $D(E(x)) = x$.
- Since the link is insecure, Eve may know what $E(x)$ is.
- We would like to have an encryption function E such that only knowing $E(x)$ cannot reveal anything about x.
- The idea is that each person has a **public key** known to the whole world and a **private key** known only to him- or herself.
- Alice encodes x using Bob's public key. Bob then decrypts it using his private key, thus retrieving x.

5.2 RSA Scheme

- Let p and q be two large primes, and let $N = pq$ (p and q are not public).
- Treat messages to Bob as numbers modulo N, excluding trivial values 0 and 1.
- Let e be any number that is relatively prime to $(p-1)(q-1)$ (Typically e is a small value).
- Then Bob's public key is the pair of numbers (N, e) and his private key is $d = e^{-1} \pmod{(p-1)(q-1)}$.

5.3 RSA Encryption

- **Encryption**: Alice computes the value $E(x) = x^e \bmod N$ and sends this to Bob.
- **Decryption:** Upon receiving the value $y = E(x)$, Bob computes $D(y) = y^d \bmod N$; this will be equal to the original message x.

Theorem 29. Using the encryption and decryption functions E and D, we have $D(E(x)) = x \pmod{N}$ for every possible message $x \in \{0, 1, ..., N-1\}$.

Proof. This can be proved using Chinese Remainder Theorem or Fermat's Little Theorem. For more details, please refer to notes. □

6 Polynomials

6.1 Properies of polynomials

- **Property 1:** A non-zero polynomial of degree d has at most d roots.
- **Property 2:** A polynomial of degree d is **uniquely** determined by $d+1$ distinct points.

6.2 Polynomial Interpolation

Question. Given $d+1$ distinct points, how do we determine the polynomial?

- We use a method called **Lagrange Interpolation**, which works similarly to the **Chinese Remainder Theorem**.
- Suppose the given points are $(x_1, y_1), \ldots, (x_{d+1}, y_{d+1})$. We want to find a polynomial $p(x)$ such that $p(x_i) = y_i$ for $i = 1, \ldots, d+1$.
- In other words, we want to find polynomials $p_1(x), \ldots, p_{d+1}(x)$ such that

$$\begin{aligned}
&p_1(x) = 1 \text{ at } x_1 \text{ and } p_1(x) = 0 \text{ at } x_2, \ldots, x_{d+1};\\
&p_2(x) = 1 \text{ at } x_2 \text{ and } p_2(x) = 0 \text{ at } x_1, x_3 \ldots, x_{d+1};\\
&p_3(x) = 1 \text{ at } x_3 \text{ and } p_3(x) = 0 \text{ at } x_1, x_2, x_4, \ldots, x_{d+1} \text{ and so on...}
\end{aligned}$$

6.3 Lagrange Interpolation

- Let's start by finding $p_1(x)$.
- Since $p_1(x) = 0$ at $x_2, \ldots, x_{d+1}$, $p_1(x)$ must be a multiple of

$$q_1(x) = (x - x_2)(x - x_3)\ldots(x - x_{d+1}).$$

- We also need $p_1(x) = 1$ at x_1. Notice that

$$q_1(x_1) = (x_1 - x_2)(x_1 - x_3)\ldots(x - x_{d+1}).$$

- Then $p_1(x) = \frac{q_1(x)}{q_1(x_1)}$ is the polynomial we are looking for.
- Similarly for $p_i(x)$, we have $p_i(x) = \frac{q_i(x)}{q_i(x_i)}$.
- After finding $p_1(x), \ldots, p_{d+1}(x)$, we can construct $p(x)$ by scaling up each bit by corresponding y_i:

$$p(x) = \sum_{i=1}^{d+1} y_i \cdot p_i(x)$$

  This should remind you of CRT.

- Now let us define $\Delta_i(x)$ in the following way (think of them as a basis):

$$\Delta_i(x) = \frac{\prod_{i \neq j}(x - x_j)}{\prod_{i \neq j}(x_i - x_j)}.$$

- Then we have an **unique** polynomial

$$\boxed{p(x) = \sum_{i=1}^{d+1} y_i \Delta_i(x).}$$

6.4 Finite Fields

- The properties of a polynomial would not hold if the values are restricted to being natural numbers or integers because dividing two integers does not generally result in an integer.
- However, if we work with numbers modulo m where m is a prime number, then we can add, subtract, multiply and divide.
- Then **Property 1** and **Property 2** hold if the coefficients and the variable x are restricted to take on values modulo m. When we work with numbers modulo m, we are working over a **finite field**, denoted by $GF(m)$ (**Galois Field**).

6.5 Secret Sharing

6.5.1 Basic Ideas

- Suppose there are n people. Let s be the secret number and q be a prime number greater than n and s. We will work over $GF(q)$.
- Pick a random polynomial $P(x)$ of degree $k-1$ such that $P(0) = s$.
- Distribute $P(1), \ldots P(n)$ to each person so that each one receives one value.
- Then in order to know what s is, at least k of the n people must work together so that they can perform **Lagrange interpolation** and find P.
- If there are less than k people, they will learn nothing about s!

7 Error Correcting Codes

7.1 Basic Ideas

- **Goal:** Transmit messages across an **unreliable** communication channel.
- The channel may cause **packets**(parts of the message) to be **lost**, or even **corrupted**.
- **Error correcting code** is an encoding scheme to protect messages against these errors by introducing redundancy.

7.2 Erasure Errors

- **Erasure errors** refer to some packets being **lost** during transmission.
- Suppose that the message consists of n packets and at most k packets are lost during transmission.
- To prevent this error, we encode the initial message into a redundant encoding consisting of $n + k$ packets such that the receiver can reconstruct the message from any n received packets using **Lagrange interpolation**.

7.3 General Errors

- Now suppose the packets are **corrupted** during transmission due to channel noise. Such error is called **general errors**.
- Suppose that k out of n characters are corrupted and we have no idea which k these are.
- To guard against k general errors, we must transmit $n + 2k$ characters.
- To reconstruct the polynomial, we need to find a polynomial $P(x)$ of degree $n - 1$ such that $P(i) = r_i$ for at least $n + k$ values of i.

7.4 Error-locator Polynomial

- To efficiently find the polynomial $P(x)$, we need the locations of the k errors.
- Let $e_1, ..., e_k$ be the k locations at which errors occurred. We don't know where these errors are.
- Guessing where the errors are will take exponential time, which is inefficient, so we use the **error-locator polynomial**:

$$E(x) = (x - e_1)(x - e_2) \dots (x - e_k).$$

- Then we have the following:

$$P(i)E(i) = r_i E(i) \quad \text{for } 1 \leq i \leq n + 2k.$$

  This is known as the **Berlekamp–Welch algorithm**.

7.5 Berlekamp–Welch algorithm

- Define $Q(x) = P(x)E(x)$. We have $n + 2k$ equations with $n + 2k$ unknown coefficients:
$$Q(i) = r_i E(i) \quad \text{for } 1 \leq i \leq n + 2k.$$

- We can solve the systems of linear equations and get $E(x)$ and $Q(x)$.
- Finally we compute $\frac{Q(x)}{E(x)}$ to obtain $P(x)$.

8 Counting

8.1 Counting Rules

Theorem 30 (First Rule of Counting)**.** If there are n ways of doing something, and m ways of doing another thing after that, then there are $n \times m$ ways to perform both of these actions.

- Order matters(permutations).
- Sampling k elements from n items:
  - With replacement: n^k.
  - Without replacement: $\frac{n!}{(n-k)!}$.

Theorem 31 (Second Rule of Counting)**.** If order doesn't matter count ordered objects and then divide by number of orderings.

- Without replacement and ordering doesn't matter (combinations).
- Number of ways of choosing k-element subsets out of a set of size n:

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

8.2 Stars and Bars

Stars and Bars is a technique used to solve for problems that sample with replacement but order doesn't matter by establishing a bijection between the problem and the stars and bars problem.

Problem 1. Consider the equation $a+b+c+d = 12$ where a, b, c, d are non-negative integers. How many solutions are there to this equation?

- Let's simplify this problem a little bit. Suppose we have 12 and 3 bars.

$$\star\star\,|\,\star\star|\,\star\star\star\,|\,\star\star\star\star\star$$

- How many ways can we arrange them? $\binom{12+3}{3} = \binom{15}{3}$
- This is the answer to our original problem! Do you see the bijection between the two problems?

Theorem 32 (Stars and Bars)**.** The number of ways to distribute n indistinguishable objects into k distinguishable bins is

$$\binom{n+k-1}{k-1}.$$

- Don't memorize the formula! Try to visualize the problem by connecting it to stars and bars. Draw out the stars and the bars!

- Again, this method is useful for with replacement but order doesn't matter type of problems.

Theorem 33 (Zeroth Rule of Counting:)**.** If a set A has a bijection relationship with a set B, then $|A| = |B|$.

The stars and bars method relies on this counting rule and this is the key to many combinatorial arguments as we will explore further later.

8.3 Binomial Theorem

Theorem 34 (Binomial Theorem)**.** For all $n \in \mathbb{N}$,

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Proof. See notes. □

Corollary 35. For all $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0.$$

Proof. Plug in $a = -1$ and $b = 1$ for the binomial theorem. □

8.4 Combinatorial Proofs

- Intuitive counting arguments. No tedious algebraic manipulation.
- Proofs by stories: same story from multiple perspectives.
- Proving an identity by counting the same thing in two different ways.
- Useful identity:

$$\binom{n}{k} = \binom{n}{n-k}.$$

- Choosing k objects to include is equivalent to choosing $n-k$ objects to exclude.

Example 8.1. Using combinatorial arguments, show that

$$\sum_{i=0}^{n} \binom{n}{i} = 2^n.$$

Proof. We can use binomial theorem by letting $a = b = 1$, however this is not what the question is asking for.
RHS: Total number of subsets of a set of size n.
LHS: The number of ways to choose a subset of size i is $\binom{n}{i}$. To find the total number of subsets, we simply add all the cases when $i = 0, 1, 2, \ldots, n$. □

8.5 Principle of Inclusion-Exclusion

Theorem 36 (Principle of Inclusion-Exclusion(General):)**.** Let $A_1, \ldots, A_n$ be arbitrary subsets of the same finite set A. Then,

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq \{1,\ldots,n\}: |S|=k} \left| \cap_{i \in S} A_i \right|.$$

Proof. See notes. □

Theorem 37 (Principle of Inclusion-Exclusion(Simplified):)**.**

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

8.6 Summary

	with replacement	w/o replacement
order matters	n^k	$\frac{n!}{(n-k)!}$
order doesn't matter	$\binom{n+k-1}{k-1}$	$\binom{n}{k}$

9 Countability

Question. How do we determine if two sets have the same cardinality, or size?

This is obvious for finite sets, but for infinite sets it becomes quite tricky. We'll see how to formulate the question.

9.1 Bijection

- Two finite sets have the same size if and only if their elements can be paired up, so that each element of one set has a unique partner in the other set, and vice versa.
- We formalize this through the concept of a **bijection**, which is discussed in section 4.6.

9.2 Cardinality

- To show that two infinite sets have the same cardinality, we demonstrate a pairing between elements of the two sets, i.e., establish a bijection (one-to-one correspondence) between the two sets.

Problem 2. Are there more natural numbers $\mathbb{N}$ than there are positive integers $\mathbb{Z}^+$?

Answer. It is tempting to answer yes, because every positive integer is also a natural number, and the natural numbers have one extra element 0. However, we can actually define a mapping between the natural numbers and the positive integers as follows:

- Define a function $f : \mathbb{N} \to \mathbb{Z}^+$ such that $f(n) = n+1$. Then we can see there's a one-to-one correspondence in the following figure (try to prove it on your own).

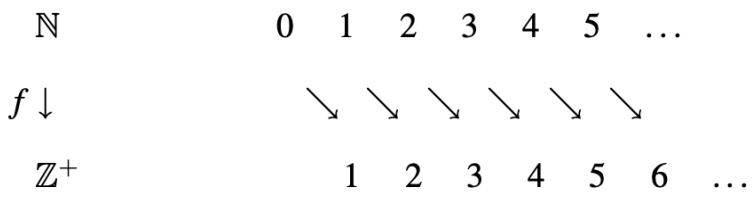


Figure 6: Bijection between $\mathbb{N}$ and $\mathbb{Z}^+$

- Since we have shown a bijection between $\mathbb{N}$ and $\mathbb{Z}^+$, this tells us that there are exactly as many natural numbers as there are positive integers.
- This also implicityly showed the fact that $\infty + 1 = \infty$!

Exercise 1. Show that the cardinality for the set of natural numbers and the set of even natural numbers are the same.

Problem 3. What about $\mathbb{N}$ and $\mathbb{Z}$?

Answer. It may seem obvious that $\mathbb{Z}$ is larger because it includes negative numbers. However, they both actually have the same size! Let's see why, consider the following function f:

$$0 \to 0, 1 \to -1, 2 \to 1, 3 \to -2, 4 \to 2, \ldots, 124 \to 62, \ldots$$

In other words, the function is defined as follows:

$$f(x) = \begin{cases} \frac{x}{2}, & \text{if } x \text{ is even} \\ \frac{-(x+1)}{2}, & \text{if } x \text{ is odd} \end{cases}$$

This function $f : \mathbb{N} \to \mathbb{Z}$ is in fact a bijection, refer to the notes for the details. Thus, the two sets have the same size.

Definition 38 (Countable)**.** A set S is **countable** if there is a bijection between S and $\mathbb{N}$ or some subset of $\mathbb{N}$.

- Intuitively, any finite set S is clearly **countable**. But the actual reasoning behind is because there is a bijection between S and the subset $\{0, 1, 2, ..., m-1\}$, where $m = |S|$ is the size of S.
- The examples we did earlier are countable because they are subsets of $\mathbb{N}$.

Problem 4. Now consider the set of rational numbers $\mathbb{Q}$, is it larger than $\mathbb{N}$? Recall that $\mathbb{Q} = \left\{ \frac{x}{y} \mid x, y \in \mathbb{Z}, y \neq 0 \right\}$.

Answer. The two sets actually have the same cardinality! Let's look at this using a different way by introducing some new definitions and an important theorem.

Definition 39. If there is a injective function $f : A \to B$, then $|A| \leq |B|$.

Definition 40. If there is a surjective function $f : A \to B$, then $|A| \geq |B|$.

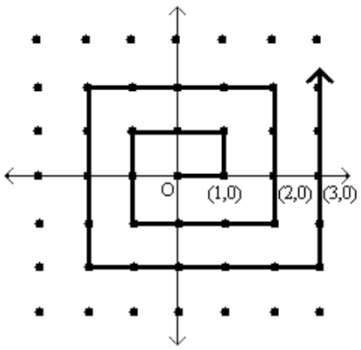
Theorem 41 (Schröder–Bernstein Theorem (Cantor–Bernstein))**.** If A and B are sets with $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$. In other words, if there are injective functions $f : A \to B$ and $g : B \to A$, then there is a bijection h between A and B.

Proof. The proof of this theorem is out of scope for this class. We'll skip that for now. □

Remark. This theorem will be very useful when we want to show a set S is countable. We can give separate injections $f : S \to \mathbb{N}$ and $g : \mathbb{N} \to S$, instead of designing a bijection (which is trickier).

- Now back to our problem. First it is obvious that $|\mathbb{N}| \leq |\mathbb{Q}|$ because $\mathbb{N} \subseteq \mathbb{Q}$.
- Now the theorem comes in handy and all we need to do now is to prove $|\mathbb{Q}| \leq |\mathbb{N}|$.
- Recall the definition, we must exhibit an injection $f : \mathbb{Q} \to \mathbb{N}$.
- Notice that each rational number $\frac{a}{b}$ $(\gcd(a, b) = 1)$ can be represented by the point $(a, b) \in \mathbb{Z} \times \mathbb{Z}$ (the set of all pairs of integers).
- However, not all points are valid, espeically when the corresponding $\frac{a}{b}$ is undefined (except for $(0, 0)$, which is used to represent rational number 0). The points whose rational representation is an unfactored fraction are also invalid.
- Thus, we can actually tell that $|\mathbb{Z} \times \mathbb{Z}| \geq |\mathbb{Q}|$.

- If we are able to come up with an injection from $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{N}$, then this will also be an injection from $\mathbb{Q}$ to $\mathbb{N}$ (why?).



- The idea is to map each pair (a, b) to its position along the spiral, starting at the origin, as shown in the picture above (basically we are indexing throught each point starting from index 0).
- It is clear that this mapping maps every pair of integers injectively to a natural number.
- Thus we have $|\mathbb{Q}| \leq |\mathbb{Z} \times \mathbb{Z}| \leq |\mathbb{N}|$. Remember that $|\mathbb{N}| \leq |\mathbb{Q}|$, then by the Cantor-Bernstein Theorem $|\mathbb{N}| = |\mathbb{Q}|$.

9.3 Cantor's Diagonalization

Now let's consider the set of real numbers, $\mathbb{R}[0, 1]$ specifically. We'll see that it is uncountable using a method called **diagonalization**.

Theorem 42. The real interval $\mathbb{R}[0, 1]$ is uncountable.

Proof. Assume for the sake of contradiction that there is a bijection $f : \mathbb{N} \to \mathbb{R}[0, 1]$. Then, we can enumerate the real numbers in an infinite list $f(0), f(1), f(2), \ldots$ as follows:

$$\begin{aligned} f(0) &= 0.\textcircled{5}2149356\ldots \\ f(1) &= 0.1\textcircled{4}162985\ldots \\ f(2) &= 0.94\textcircled{7}82712\ldots \\ f(3) &= 0.530\textcircled{9}8175\ldots \\ &\vdots \end{aligned}$$

The number circled in the diagonal can be some real number $r = 0.5479\ldots$. Now consider the real number s obtained by modifying every digit of r such that each digit d is replaced with $d + 2 \pmod{10}$, so $s = 0.7691.....$ Then we claim that s is not included in our infinite list of real numbers. Suppose for contradiction that it us, and that it was the n^{th} number in the list, $f(n)$. But by construction s differs from $f(n)$ in the $(n + 1)$th digit, so they cannot be equal! So we have constructed a real number s that is not in the range of f. But this contradicts the assertion that f is a bijection. Thus the real numbers are not countable.

□

Remark. The reason that we modified each digit by adding 2 (mod 10) instead of adding 1 is that the same real number can have two decimal expansions; for example $0.999\ldots = 1.000\ldots$. But if two real numbers differ by more than 1 in any digit they cannot be equal. Thus our modification is safe. (We can also replace each digit by some different digit chosen from the range $\{1, 2, \ldots, 8\}$.)

Theorem 43. If A and B are countable sets, then $A \cup B$ is also countable.

9.4 Power Sets and Higher Orders of Infinity

Definition 44 (Power Set)**.** Recall that the **power set** of S, denoted by $\mathscr{P}(S)$, is the set of all subsets of S. More formally, it is defined as $\mathscr{P}(S) = \{T : T \subseteq S\}$.

Question. What is the cardinality of $\mathscr{P}(S)$?

Answer. If $|S| = k$ is finite, then $|\mathscr{P}(S)| = 2^k$.(why?)

- For finite sets S, the cardinality of the power set of S is exponentially larger than the cardinality of S.
- What about infinite (countable) sets?
- We claim that there is no bijection from S to $\mathscr{P}(S)$ so $\mathscr{P}(S)$ is not countable.
- For example the set of all subsets of natural numbers is not countable, even though the set of natural numbers itself is countable.

Theorem 45. $|\mathscr{P}(\mathbb{N})| > |\mathbb{N}|$.

Proof. See notes. □

10 Discrete Probability

10.1 Probabilistic Models

A **probabilistic model** is a mathematical description of an uncertain situation. The elements of a probabilistic model includes

- **sample space** Ω : set of all all possible outcomes of an experiment.
- **probability law:** assigns to a set A of possible outcomes (**event**) a nonnegative value $\mathbb{P}(A)$ (probability of A) that encodes the knowledge about the likelihood of the elements of A.

A recap of all basic terminologies:

Definition 46 (Experiment)**.** An **experiment** is a procedure that yields one of a given set of possible outcomes.

Definition 47 (Sample space)**.** The **sample space** of the experiment is the set of possible outcomes.

Definition 48 (Sample point)**.** A **sample point** is an element of the sample space.

Definition 49 (Event)**.** An **event** is a subset of the sample space.

10.2 Probability Space

Definition 50 (Probability Space)**.** The **probability space** is defined by the triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the *sample space*, $\mathcal{F} \subseteq \Omega$ is the *event space* and $\mathbb{P}$ is the *probability function*, satisfying the following axioms:

Probability Axioms (Kolmogorov):

- **Nonnegativity:** for all sample points $\omega \in \Omega$,

$$\mathbb{P}(\omega) \geq 0.$$

- **Additivity:** any countable sequence of **disjoint sets** (mutually exclusive events) $E_1, E_2, \ldots$ satisfies

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}\left(E_i\right).$$

- **Normalization:** the sum of all probabilities must be 1, thus

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = \mathbb{P}(\Omega) = 1.$$

Definition 51 (Probability)**.** For any event $A \subseteq \Omega$, we define the **probability** of A to be

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega).$$

10.2.1 Properties of Probability Laws

- $\mathbb{P}(\emptyset) = 0$.
- $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$, where $\overline{A}$ (or A^c) is the **complement** of A.
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A) - \mathbb{P}(A \cap B)$.
- If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

10.3 Discrete Uniform Probability Space

Theorem 52 (Discrete Uniform Probability Law)**.** In a uniform probability space, all sample points have the same probability $\frac{1}{|\Omega|}$. Thus the probability of an event A is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}.$$

Remark. For uniform spaces, computing probabilities is simply counting sample points.

Example 10.1 (Poker Hands)**.** Consider shuffling a deck of cards and dealing a poker hand. In this case, the sample space $\Omega = \{\text{all possible poker hands}\}$. Hence, $|\Omega| = \binom{52}{5}$. Assuming that the probability of each outcome is equally likely and so we have an uniform probability space.
Let A be the event that the poker hand is a flush (same suit). Since the probability space is uniform, computing $\mathbb{P}(A)$ reduces to simply computing $|A|$, the number of poker hands that are flushes. There are 13 cards in each suit, so the number of flushes in each suit is $\binom{13}{5}$. The total number of flushes is therefore $4 \cdot \binom{13}{5}$. Then we have

$$\mathbb{P}\left(\text{hand is a flush}\right) \approx 0.002.$$

Example 10.2 (Balls and Bins)**.** Consider the experiment of throwing 20 labelled balls into 10 labeled bins. Assume that each ball is equally likely to land in any bin.
The sample space Ω is equal to $\{(b_1, b_2, \ldots, b_{20}) : 1 \leq b_i \leq 10 \text{ for each } i = 1, \ldots, 20\}$, where the component b_i denotes the bin in which ball i lands. Then $|\Omega| = 10^{20}$, since each element b_i in the sequence has 10 possible choices and there are 20 elements in the sequence. In general, throwing m balls into n bins gives a sample space of size n^m.

Let A be the event that bin 1 is empty. Since the probability space is uniform, we simply need to count how many outcomes have this property. This is exactly the number of ways all 20 balls can fall into the remaining nine bins, which is 9^{20}. Hence, $\mathbb{P}(A) = \frac{9^{20}}{10^{20}} = \left(\frac{9}{10}\right)^{20} \approx 0.12$. Let B be the event that bin 1 contains at least one ball. This event is the complement $\overline{A}$ of A. So $\mathbb{P}(B) = 1 - \mathbb{P}(A) \approx 0.88$. More generally, if we throw m balls into n bins, we have:

$$\mathbb{P}(\text{bin 1 is empty}) = \left(\frac{n-1}{n}\right)^m = \left(1 - \frac{1}{n}\right)^m.$$

10.3.1 Birthday Paradox

The **birthday paradox** examines the chances that two people in a group have the same birthday. It is called a "paradox" because it is counter-intuitive. Suppose there are 365 days in a year. Then $S = \{1, \ldots, 365\}$, and the experiment consists of drawing a sample of n elements from S, where the elements are the birth dates of n people in a group. Then $|\Omega| = 365^n$ because there are 365 possible birth dates for each person. Let A be the event that at least a pair of people have the same birthday. If we want to determine $\mathbb{P}(A)$, it is simpler to first compute the probability of the complement of A; i.e., $\mathbb{P}(\overline{A})$, where $\overline{A}$ is the event that no two people have the same birthday.

Since the probability space is uniform, we just need to determine $|\overline{A}|$, the number of ways for no two people to have the same birthday. There are 365 choices for the first person, 364 for the second, $\ldots, 365-n+1$ choices for the n-th person, for a total of $365 \times 364 \times \cdots \times (365-n+1)$ by the First Rule of Counting from previous section; we are sampling without replacement and the order matters. Thus we have

$$\mathbb{P}(\overline{A}) = \frac{|\overline{A}|}{|\Omega|} = \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n},$$

so $\mathbb{P}(A) = 1 - \mathbb{P}(\overline{A}) = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$. Here $\mathbb{P}(A)$ is a function of n. As n increases $\mathbb{P}(A)$ increases. For example, with $n = 23$ people, you should be willing to bet that at least a pair of people have the same birthday, since $\mathbb{P}(A)$ is larger than 50%. For $n = 60$ people, $\mathbb{P}(A)$ is over 99%!

10.4 Conditional Probability

Conditional probability provides a way to reason about the outcome of an experiment, based on partial information. We wish to quantify the likelihood that the outcome also belongs to some other given event A by constructing a new probability law to take into account the available knowledge.

Definition 53 (Conditional Probability)**.** Let B be an event such that $\mathbb{P}(B) > 0$. The **conditional probability** of A given B, denoted by $\mathbb{P}(A|B)$ is defined as

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

10.4.1 Independence

Definition 54 (Independence)**.** Event A and B are **independent** if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \text{ or } \mathbb{P}(A|B) = \mathbb{P}(A).$$

Definition 55 (Mutual Independence)**.** Events $\{A\}_{i=1}^n$ are **mutually independent** if

$$\mathbb{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbb{P}(A_i).$$

Definition 56 (Chain Rule)**.** For any events $A_1, \ldots, A_n$,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}\left(A_1\right) \cdot \mathbb{P}\left(A_2 \mid A_1\right) \cdot \mathbb{P}\left(A_3 \mid A_1 \cap A_2\right) \cdot \cdots \cdot \mathbb{P}\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right).$$

10.4.2 Conditional Independence

Definition 57 (Conditional Independence)**.** Given event C such that $\mathbb{P}(C) > 0$, events A and B are called **conditionally independent** if

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

or

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

10.4.3 Law of Total Probability

We define a partition of an event as follows:

- **(Partition of an event).** We say that an event A is partitioned into n events $A_1, \ldots, A_n$ if
  1. $A = A_1 \cup A_2 \cup \cdots \cup A_n$,
  2. $A_i \cap A_j = \varnothing$ for all $i \neq j$ (i.e., $A_1, \ldots, A_n$ are **mutually exclusive**).

  In simpler terms, each outcome in A belongs to exactly one of the events $A_1, \ldots, A_n$.
- Now, let $A_1, \ldots, A_n$ be a **partition** of the sample space Ω. Then, the **Law of Total Probability** for any event B is as follows:

Theorem 58 (Law of Total Probability)**.**

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}\left(B \cap A_i\right) = \sum_{i=1}^{n} \mathbb{P}\left(B \mid A_i\right) \mathbb{P}\left(A_i\right).$$

10.4.4 Bayes' Rule

Given $\mathbb{P}(B|A)$, how do we compute $\mathbb{P}(A|B)$?

- Using the definition and chain rule, we have

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

- Combining with **Law of Total Probability**, we have the following result:

Theorem 59 (Bayes' Rule)**.** Let $\{A\}_{i=1}^{n}$ be disjoint events that form a partition of the sample space, and that $\mathbb{P}(A_i) > 0$ for all i. Then, for any event B such that $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A_i \cap B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_{j=1}^{n} \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

10.4.5 Inclusion-Exclusion Principle

Theorem 60 (Inclusion-Exclusion Principle)**.** Let $A_1, \ldots, A_n$ be events in some probability space, where $n \geq 2$. Then, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k-1} \sum_{S \subseteq \{1,\ldots,n\}: |S|=k} \mathbb{P}\left(\bigcap_{i \in S} A_i\right).$$

The right hand side is equivalent to

$$\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \ldots + (-1)^{n-1} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n).$$

10.4.6 Union Bound

- Very useful for proving upper bounds for randomized algorithms.

Theorem 61 (Union Bound)**.** Let $A_1, \ldots, A_n$ be events in some probability space. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i).$$

11 Discrete Random variables

Definition 62 (Random variable)**.** A **random variable** X on a sample space Ω is a function $X : \Omega \to \mathbb{R}$ that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.

Remark (Functions of R.V.s are also R.V.s)**.** Let $Y = g(X)$. Then

$$\mathbb{P}(Y = y) = \sum_{x|g(x)=y} \mathbb{P}(X = x).$$

An R.V. itself is a function, and we know that the function of a function is also a function.

Definition 63. The **distribution** of a discrete random variable X is the collection of values $\{(x, \mathbb{P}(X = x)) : x \in \mathcal{X}\}$, where $\mathcal{X}$ is the set of all possible values taken by X.

Definition 64 (Probability Mass Function)**.** The **probability mass function**, or PMF, of a discrete random variable X is a function mapping X's values to their associated probabilities. It is the function $p : \mathbb{R} \to [0, 1]$ defined by

$$p_X(x) = \mathbb{P}(X = x).$$

Definition 65 (Joint Distribution)**.** The **joint distribution** for two discrete random variables X and Y is the collection of values $\{((x, y), \mathbb{P}(X = x, Y = y)) : x \in \mathcal{X}, y \in \mathcal{Y}\}$, where $\mathcal{X}$ is the set of all possible values taken by X and $\mathcal{Y}$ is the set of all possible values taken Y.

Definition 66 (Marginal Distribution)**.** Given the joint distribution for X and Y, the **marginal distribution** for X is as follows:

$$\mathbb{P}(X = x) = \sum_{y \in \mathcal{Y}} \mathbb{P}(X = x, Y = y)$$

Definition 67 (Independence)**.** Random variables X and Y are said to be **independent** if the events $X = x$ and $Y = y$ are independent for all values x, y. Equivalently, the joint distribution of independent R.V's decomposes as

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y), \quad \forall x, y.$$

Definition 68 (Indicator Random variable)**.** $\mathbb{I}_i$, or X_i, denotes the **indicator random variable** that takes on values $\{0, 1\}$ according to whether a specified event occurs or not. Usually $\{\mathbb{I}_i\}_{i=1}^n$ are mutually independent and they are said to be *independent and identically distributed (i.i.d).*

11.1 Expectation

Definition 69 (Expectation)**.** The **expectation** of a discrete random variable X is defined as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x).$$

Alternatively, we also have

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega).$$

11.1.1 Linearity of Expectation

Theorem 70 (Linearity of Expectation)**.** For any two random variables X and Y on the same probability space, we have

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

For any constant a, c, we also have

$$\mathbb{E}[aX + c] = a\mathbb{E}[X] + c.$$

Proof. Let $g(X, Y) = X + Y$. Then we have

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{x,y} (x+y)\mathbb{P}(X=x, Y=y) \\
&= \sum_{x,y} x\mathbb{P}(X=x, Y=y) + \sum_{x,y} y\mathbb{P}(X=x, Y=y) \\
&= \sum_{x} \sum_{y} x\mathbb{P}(X=x, Y=y) + \sum_{y} \sum_{x} y\mathbb{P}(X=x, Y=y) \\
&= \sum_{x} x \sum_{y} \mathbb{P}(X=x, Y=y) + \sum_{y} y \sum_{x} \mathbb{P}(X=x, Y=y) \\
&= \sum_{x} x\mathbb{P}(X=x) + \sum_{y} y\mathbb{P}(Y=y) \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

The proof of the second equality is left as an exercise. □

This is a powerful theorem because this always applies without any assumption about the R.V.s.

Remark. Be careful that this doesn't imply that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, or $\mathbb{E}\left[\frac{1}{X}\right] = \frac{1}{\mathbb{E}[X]}$. These are not true in general.

11.2 Variance

Definition 71 (Variance)**.** The **variance** of a random variable X is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Definition 72 (Standard Deviation)**.** The **standard deviation** of a random variable X

$$\sigma := \sqrt{\mathrm{Var}(X)}.$$

Theorem 73. For a random variable X,

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Proof.

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
&= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right] \\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2.
\end{aligned}$$

Note that $\mathbb{E}[X]$ is a constant. □

Fact. For any constant c and any random variable X, we have

$$\text{Var}(cX) = c^2\text{Var}(X).$$

Theorem 74. For independent random variables X, Y, we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xy \cdot \mathbb{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy \cdot \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y) \\
&= \left(\sum_x x \cdot \mathbb{P}(X = x)\right) \cdot \left(\sum_y y \cdot \mathbb{P}(Y = y)\right) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y]
\end{aligned}$$

where the second line made crucial use of independence. □

Theorem 75. For **independent** random variables X, Y,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Proof.

$$\begin{aligned}
\text{Var}(X + Y) &= \mathbb{E}\left[(X + Y)^2\right] - (\mathbb{E}[X + Y])^2 \\
&= \mathbb{E}\left[X^2\right] + \mathbb{E}\left[Y^2\right] + 2\mathbb{E}[XY] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \left(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2\right) + \left(\mathbb{E}\left[Y^2\right] - \mathbb{E}[Y]^2\right) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]) \\
&= \text{Var}(X) + \text{Var}(Y) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]). \\
&= \text{Var}(X) + \text{Var}(Y).
\end{aligned}$$

□

11.2.1 Covariance

Covariance is a measure of the joint variability of two random variables.

Definition 76 (Covariance)**.** The **covariance** of random variables X and Y, denoted $\text{Cov}(X, Y)$, is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Remark. Some important facts about covariance.

1. If X, Y are independent, then $\text{Cov}(X, Y) = 0$. However, the converse is not true.
2. $\text{Cov}(X, X) = \text{Var}(X)$.
3. *Bilinearity*:
$$\text{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j \text{Cov}(X_i, Y_j).$$
4. For general random variables X and Y,
$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

11.2.2 Correlation

Definition 77 (Correlation)**.** Suppose X, Y are random variables with $\sigma_X, \sigma_Y > 0$. Then the **correlation** of X and Y is

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

and $-1 \leq \rho(X, Y) \leq 1$.

11.3 Discrete Probability Distribution

11.3.1 Bernoulli Distribution

A **Bernoulli** random variable X, denoted as Bernoulli(p), has a PDF of the form

$$\mathbb{P}(X = i) = \begin{cases} p, & \text{if } i = 1 \\ 1 - p, & \text{if } i = 0, \end{cases}$$

where $0 \leq p \leq 1$.
Expectation:

$$\mathbb{E}[X] = p.$$

Variance:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = p - p^2 = p(1 - p).$$

11.3.2 Binomial Distribution

A **binomial** random variable X, denoted as $\text{Bin}(n, p)$, has a PDF of the form

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n.$$

Quick check on normalization:

$$\sum_{i=0}^{n} \mathbb{P}(X = i) = 1 \implies \sum_{i=0}^{n} \binom{n}{i} p^i (1 - p)^{n-i} = 1.$$

A probabilistic proof of the Binomial Theorem for $a = p$ and $b = 1 - p$.

Fact. A binomial random variable is equivalent to sum of n i.i.d Bernoulli variables with parameter p.

Expectation:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathbb{E}[Y_i] = \sum_{i=1}^{n} p = np, \quad \text{where } Y_i \sim \text{Bernoulli}(p).$$

Variance:

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} Y_i\right) = np(1-p), \quad \text{where } Y_i \sim \text{Bernoulli}(p).$$

11.3.3 Hypergeometric Distribution

We are given $N = G + B$ balls, where G balls are good and B balls are bad. Sample n balls *without* replacement and observe k successes. Denoted as Hypergeometric(N, B, n) and has a PDF of the form

$$\mathbb{P}(X = k) = \frac{\binom{G}{k}\binom{B}{n-k}}{\binom{N}{n}}.$$

11.3.4 Geometric Distribution

A **geometric** random variable X, denoted as Geo(p), has a PDF of the form

$$\mathbb{P}(X = k) = (1-p)^{k-1}p, \quad \text{for } i = 1, 2, 3, \ldots.$$

It represents the number of trials until first success, where p is the probability of success.
Quick check on normalization:

$$\sum_{i=1}^{\infty} \mathbb{P}(X = i) = \sum_{i=1}^{\infty} (1-p)^{i-1}p = p\sum_{i=1}^{\infty}(1-p)^{i-1} = p \cdot \frac{1}{1-(1-p)} = 1.$$

Expectation:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i) = \sum_{x=1}^{\infty}(1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p},$$

where the first equality uses the **tail sum formula**, which is on the next page.

Theorem 78 (Tail Sum Formula)**.** Let X be a random variable that takes values in $\{0, 1, 2, \ldots\}$. Then

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i).$$

Proof. We can manipulate the formula for the expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=1}^{\infty} x\mathbb{P}(X = x) \\
&= \sum_{x=1}^{\infty} \sum_{i=1}^{x} \mathbb{P}(X = x) \\
&= \sum_{i=1}^{\infty} \sum_{x=i}^{\infty} \mathbb{P}(X = x) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(X \geq i).
\end{aligned}$$

This is called the *Tail Sum Formula* because we are summing over the tail probabilities of the distribution. □

Remark. Here's a *smarter* way to derive the expectation. Suppose we toss our first coin. There are two possibilities: we get a head with probability p and call it a day, or we get a tail with probability $1 - p$ and we are right back where we just started. In the latter case, we expect $1 + \mathbb{E}[X]$ trials until our first success because we already used one trial. Hence,

$$\mathbb{E}[X] = p \cdot 1 + (1 - p)(1 + \mathbb{E}[X]).$$

This makes use of an important property called the **memoryless property**, which will be covered later.

Variance:

$$\text{Var}(X) = \frac{1 - p}{p^2}.$$

11.3.5 Poisson Distribution

A **Poisson** random variable X, denoted as Poisson(λ), has a PDF of the form

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \text{for } i = 0, 1, 2, \ldots.$$

It is used to model rare events and is an approximation of the limiting case of binomial distribution.

Quick check on normalization:

$$\sum_{i=0}^{\infty} \mathbb{P}(X = i) = \sum_{i=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Remark. The second equality uses the **Taylor series expansion**

$$e^x = \sum_{i=1}^{\infty} \frac{x^i}{i!}.$$

Expectation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=0}^{\infty} i \cdot \mathbb{P}(X = i) \\
&= \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{i!} \\
&= \lambda e^{-\lambda} \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} \\
&= \lambda e^{-\lambda} e^{\lambda} \quad (e^{\lambda} = \sum_{j=1}^{\infty} \frac{\lambda^j}{j!} \text{ with } j = i-1) \\
&= \lambda.
\end{aligned}$$

Variance:

Similarly, we can calculate $\mathbb{E}[X(X-1)]$ as follows:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{i=0}^{\infty} i(i-1) \cdot \mathbb{P}(X = i) \\
&= \sum_{i=2}^{\infty} i(i-1) \frac{\lambda^i}{i!} \mathrm{e}^{-\lambda} \quad \text{i=0 and i=1 terms are equal to 0)} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \sum_{i=2}^{\infty} \frac{\lambda^{i-2}}{(i-2)!} \\
&= \lambda^2 \mathrm{e}^{-\lambda} \mathrm{e}^{\lambda} \quad (\text{ since } \mathrm{e}^{\lambda} = \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \text{ with } j = i-2) \\
&= \lambda^2
\end{aligned}$$

Therefore,

$$\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

Theorem 79. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent Poisson random variables. Then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

Proof. For all $k = 0, 1, 2, \ldots$, we have

$$\begin{aligned}
\mathbb{P}(X + Y = k) &= \sum_{j=0}^{k} \mathbb{P}(X = j, Y = k - j) \\
&= \sum_{j=0}^{k} \mathbb{P}(X = j)\mathbb{P}(Y = k - j) \\
&= \sum_{j=0}^{k} \frac{\lambda^j}{j!} e^{-\lambda} \frac{\mu^{k-j}}{(k-j)!} e^{-\mu} \\
&= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \lambda^j \mu^{k-j} \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}
\end{aligned}$$

where the second equality follows from independence, and the last equality from the binomial theorem. □

Theorem 80. If $X_1, X_2, ..., X_n$ are independent Poisson random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively, then

$$X_1 + X_2 + \ldots + X_n \sim \text{Poisson}(\lambda_1 + \lambda_2 + \ldots + \lambda_n).$$

Proof. This can be shown by induction. □

12 Concentration Inequalities and the Laws of Large Numbers

12.1 Markov's Inequality

Theorem 81 (Markov's Inequality)**.** For a **non-negative** random variable X with finite mean,

$$\mathbb{P}[X \geq c] \leq \frac{\mathbb{E}[X]}{c}$$

for any positive constant c.

Proof. Let $\mathcal{X}$ denote the range of X and consider any constant $c \in \mathcal{X}$. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x) \\
&\geq \sum_{x \geq c} x \cdot \mathbb{P}(X = x) \\
&\geq \sum_{x \geq c} c \cdot \mathbb{P}(X = x) \\
&= c \sum_{x \geq c} \mathbb{P}(X = x) \\
&= c\, \mathbb{P}[X \geq c].
\end{aligned}$$

□

Here's a smarter way to prove this inequality.

Proof. Since X is a non-negative and $c > 0$, then for all $\omega \in \Omega$

$$X(\omega) \geq \mathbb{I}\{X(\omega) \geq c\}.$$

The RHS is 0 if $X(\omega) < c$ and is c if $X(\omega) \geq c$ implied by the indicator function. Taking expectations of both sides gives

$$\mathbb{E}[X] \geq c\, \mathbb{E}[\mathbb{I}\{X \geq c\}] = c\, \mathbb{P}(X \geq c).$$

□

What if X can be negative? We'll have the following result.

Theorem 82 (Generalized Markov's Inequality)**.** Let X be an arbitrary random variable with finite mean. Then, for any positive constants c and r,

$$\mathbb{P}(|X| \geq c) \leq \frac{\mathbb{E}\left(|X|^r\right)}{c^r}.$$

Proof. For $c > 0$ and $r > 0$, we have

$$|X|^r \geq |X|^r \mathbb{I}\{|X| \geq c\} \geq c^r \mathbb{I}\{|X| \geq c\}$$

(Note that the last inequality would not hold if r were negative.) Taking expectations of both sides gives

$$\mathbb{E}\left[|X|^r\right] \geq c^r \mathbb{E}[\mathbb{I}\{|X| \geq c\}] = c^r \mathbb{P}(|X| \geq c).$$

□

12.2 Chebyshev's Inequality

We have seen that the variance (or, more correctly the standard deviation) is a measure of *spread*, or deviation from the mean. We can now make this intuition quantitatively precise:

Theorem 83 (Chebyshev's Inequality)**.** For a random variable X with finite expectation $\mathbb{E}[X] = \mu$,

$$\mathbb{P}[|X - \mu| \geq c] \leq \frac{\text{Var}(X)}{c^2}$$

and for any positive constant c.

Proof. Define $Y = (X-\mu)^2$ and note that $\mathbb{E}[Y] = \mathbb{E}\left[(X-\mu)^2\right] = \text{Var}(X)$. Also, notice that the event that we are interested in, $|X-\mu| \geq c$, is exactly the same as the event $Y = (X-\mu)^2 \geq c^2$. Therefore, $\mathbb{P}[|X-\mu| \geq c] = \mathbb{P}\left[Y \geq c^2\right]$. Moreover, Y is obviously nonnegative, so we can apply Markov's inequality in Theorem 17.1 to get

$$\mathbb{P}[|X - \mu| \geq c] = \mathbb{P}\left[Y \geq c^2\right] \leq \frac{\mathbb{E}[Y]}{c^2} = \frac{\text{Var}(X)}{c^2}$$

This completes the proof. □

12.3 Law of Large Numbers

Theorem 84 (Law of Large Numbers)**.** Let $X_1, X_2, \ldots,$ be a sequence of i.i.d. random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ for all i. Then $S_n = X_1 + X_2 + \cdots + X_n$ satisfies

$$\mathbb{P}\left(\left|\frac{1}{n}S_n - \mu\right| < \varepsilon\right) \to 1 \quad \text{as } n \to \infty$$

for every $\varepsilon > 0$, however small.

Proof. Let $Y_n = \frac{X_1 + \cdots + X_n}{n}$. Then

$$\begin{aligned}
\mathbb{P}\left(|Y_n - \mu| \geq \varepsilon\right) &\leq \frac{\text{Var}\left(Y_n\right)}{\varepsilon^2} \\
&= \frac{\text{Var}\left(X_1 + \cdots + X_n\right)}{n^2\varepsilon^2} \\
&= \frac{n\,\text{Var}\left(X_1\right)}{n^2\varepsilon^2} \\
&= \frac{\text{Var}\left(X_1\right)}{n\varepsilon^2} \to 0, \text{ as } n \to \infty
\end{aligned}$$

□

Remark. The **Law of Large Numbers** says that the probability of any deviation ε from the mean, however small, tends to zero as the number of observations n in our average tends to infinity. Thus, by taking n large enough, we can make the probability of any given deviation as small as we like.

13 LLSE, MMSE, and Conditional Expectation

13.1 LLSE

Definition 85 (Least Linear Squares Estimate)**.** Let X and Y be random variables. The least linear squares estimate (LLSE) of Y given X is defined as

$$L(Y \mid X) := E(Y) + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))$$

Observe that the LLSE is a random variable: in fact, it is a function of X.

Theorem 86 (Projection Property of LLSE)**.** The LLSE satisfies

$$E(Y - L(Y \mid X)) = 0$$

$$E((Y - L(Y \mid X))X) = 0.$$

Proof. The proofs are actually relatively straightforward using linearity. Proof of first equation:

$$\begin{aligned} E(Y - L(Y \mid X)) &= E\left(Y - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))\right) \\ &= E(Y) - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(E(X) - E(X)) \\ &= 0 \end{aligned}$$

Proof of second equation:

$$\begin{aligned} E((Y - L(Y \mid X))X) &= E\left(X\left(Y - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))\right)\right) \\ &= E(XY) - E(X)E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}\left(E\left(X^2\right) - E(X)^2\right) \\ &= \text{Cov}(X,Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \cdot \text{Var}(X) \\ &= \text{Cov}(X,Y) - \text{Cov}(X,Y) \\ &= 0. \end{aligned}$$

□

13.2 MMSE

13.3 Conditional Expectation

14 Continuous Probability

14.1 Continuous Random Variables

Definition 87 (Probability Density Function)**.** A **probability density function**, or **PDF**, for a real-valued random variable X is a function $f : \mathbb{R} \to \mathbb{R}$ satisfying:

1. **Non-negativity:** $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. **Normalization:**
$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

The probability that the value of X falls within an interval is

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx \quad \text{for all } a < b.$$

For an interval $[x, x + dx]$ with very small length dx, we have

$$\mathbb{P}(x \leq X \leq x + dx) = \int_x^{x+dx} f(t)dt \approx f(x)dx.$$

Remark. $f(x)$ doesn't correspond to the probability of anything! In particular, $f(x)$ does not have to be bounded by 1. For example, the density of the uniform distribution on the interval $[0, \ell]$ with $\ell = \frac{1}{2}$ is equal to $f(x) = 1/\left(\frac{1}{2}\right) = 2$ for $0 \leq x \leq \frac{1}{2}$, which is greater than 1. To connect density $f(x)$ with probabilities, we need to look at a very small interval $[x, x+dx]$ close to x like we did above. Therefore, we can interpret $f(x)$ as the *probability per unit length* in the vicinity of x.

14.1.1 Cumulative Distribution Function

Definition 88. For a continuous random variable X, the **cumulative distribution function**, or **CDF**, is the function as follows:

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(z)dz.$$

It is closely related to the PDF for X:

$$f(x) = \frac{dF(x)}{dx}.$$

14.2 Expectation and Variance

Definition 89 (Expectation)**.** The **expectation** of a continuous random variable X with PDF f is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx.$$

Definition 90 (Variance)**.** The **variance** of a continuous random variable X with PDF f is

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \left(\int_{-\infty}^{\infty} xf(x)dx\right)^2.$$

14.2.1 Exponential Random Variable

An **exponential** random variable X, denoted as $\text{Exp}(\lambda)$, has a PDF of the form

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Quick check on normalization:

$$\int_{-\infty}^{\infty} f(x)dx = \int_{0}^{\infty} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^{\infty} = 1.$$

Expectation:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} \lambda x e^{-\lambda x}dx = -\,xe^{-\lambda x}\Big|_0^{\infty} + \int_{0}^{\infty} e^{-\lambda x}dx = 0 + \left(-\frac{e^{-\lambda x}}{\lambda}\right)\Big|_0^{\infty} = \frac{1}{\lambda}.$$

Variance:

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{0}^{\infty} \lambda x^2 e^{-\lambda x}dx = -\,x^2 e^{-\lambda x}\Big|_0^{\infty} + \int_{0}^{\infty} 2xe^{-\lambda x}dx = 0 + \frac{2}{\lambda}\mathbb{E}[X] = \frac{2}{\lambda^2}.$$

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Theorem 91 (Minimum of Exponential Random Variables)**.** Let $X_1, \ldots, X_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$ respectively. Then the minimum of the random variables is also exponentially distributed:

$$\min\{X_1, \ldots, X_n\} \sim \text{Exp}\left(\lambda_1 + \cdots + \lambda_n\right).$$

Proof.

$$\begin{aligned} \mathbb{P}\left(\min\{X_1, \ldots, X_n\} > t\right) &= \mathbb{P}\left(X_1 > t, \ldots, X_n > t\right) \\ &= \prod_{i=1}^{n} \mathbb{P}\left(X_i > t\right) \\ &= \prod_{i=1}^{n} e^{-\lambda_i t} \\ &= e^{-\left(\sum_{i=1}^{n} \lambda_i\right)t}. \end{aligned}$$

□

14.3 Normal Random Variables

Definition 92 (Normal/Gaussian RV)**.** A **normal** or **Gaussian** random variable X, denoted by $\mathcal{N}(\mu, \sigma^2)$ where μ is the mean and σ^2 is the variance, has a PDF of the form

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2},$$

Let's verify that

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

Proof. We can show this for $\mu = 0$ and $\sigma^2 = 1$ and this will show for the general case. The trick is to show that

$$\left(\int_{-\infty}^{\infty} f_X(x) dx \right)^2 = 1$$

We have

$$\begin{aligned}
\left(\int_{-\infty}^{\infty} f_X(x) dx \right)^2 &= \left(\int_{-\infty}^{\infty} f_X(x) dx \right) \left(\int_{-\infty}^{\infty} f_Y(y) dy \right) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_X(x) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-(x^2+y^2)/2} dx dy.
\end{aligned}$$

Using polar integration, we have $dydx = rdrd\theta$. Then

$$\begin{aligned}
&\int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} e^{-r^2/2} r dr d\theta \\
&\quad = \int_0^{\infty} e^{-r^2/2} r dr \\
&\quad = \int_{-\infty}^{0} e^s ds \\
&\quad = 1.
\end{aligned}$$

□

Definition 93 (Standard Normal RV)**.** The PDF of the *standard normal* distribution $\mathcal{N}(0, 1)$ (with mean 0 and variance 1) is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Since its CDF cannot be expressed in elementary functions, the CDF is denoted by Φ

$$\Phi(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X < x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

Remark. The CDF of a normal random variable is symmetrical, so

$$\Phi(-x) = 1 - \Phi(x).$$

Theorem 94. If $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $Y = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. Equivalently, if $Y \sim \mathcal{N}(0,1)$, then

$$X = \sigma Y + \mu \sim \mathcal{N}\left(\mu, \sigma^2\right).$$

Proof. Given that $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, we can calculate the distribution of $Y = \frac{X-\mu}{\sigma}$ as:

$$\mathbb{P}(a \leq Y \leq b) = \mathbb{P}(\sigma a + \mu \leq X \leq \sigma b + \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\sigma a+\mu}^{\sigma b+\mu} e^{-(x-\mu)^2/\left(2\sigma^2\right)} dx = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-y^2/2} dy.$$

Hence Y is standard normal, which is obtained from X by shifting the origin to μ and scaling by σ. □

Theorem 95 (Sum of Independent Standard Normal RVs)**.** Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be independent standard normal random variables, and suppose $a, b \in \mathbb{R}$ are constants. Then $Z = aX + bY \sim \mathcal{N}\left(0, a^2 + b^2\right)$.

Proof. Since X and Y are independent, the joint density is

$$f(x,y) = f(x) \cdot f(y) = \frac{1}{2\pi} e^{-\left(x^2+y^2\right)/2}$$

The key observation is that $f(x,y)$ is *rotationally symmetric* around the origin, i.e., $f(x,y)$ only depends on the value $x^2 + y^2$, the distance of the point (x,y) from the origin $(0,0)$.

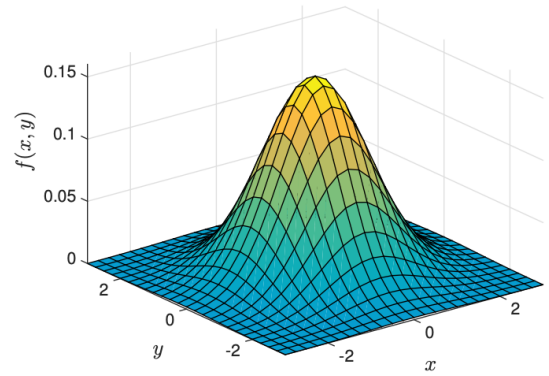


Figure 7: The joint density function $f(x,y) = \frac{1}{2\pi} e^{-\left(x^2+y^2\right)/2}$ is rotationally symmetric.

Thus, $f(T(x,y)) = f(x,y)$ where T is any rotation of the plane $\mathbb{R}^2$ about the origin. It follows that for any set $A \subseteq \mathbb{R}^2$

$$\mathbb{P}[(X,Y) \in A] = \mathbb{P}[(X,Y) \in T(A)].$$

Now given any $t \in \mathbb{R}$, we have

$$\mathbb{P}(Z \leq t) = \mathbb{P}(aX + bY \leq t) = \mathbb{P}((X,Y) \in A)$$

where A is the half plane $\{(x,y) \mid ax + by \leq t\}$. The boundary line $ax + by = t$ lies at a distance $d = \frac{t}{\sqrt{a^2+b^2}}$ from the origin. Therefore, as illustrated in the figure the set A can be rotated into the set

$$T(A) = \left\{(x,y) \mid x \leq \frac{t}{\sqrt{a^2+b^2}}\right\}$$

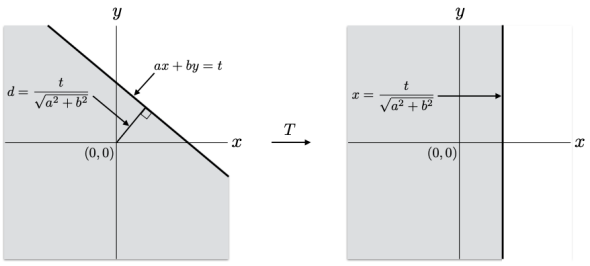


Figure 8: The half plane $ax + by \leq t$ is rotated into the half plane $x \leq \frac{t}{\sqrt{a^2+b^2}}$.

This rotation does not change the probability:

$$\mathbb{P}[Z \leq t] = \mathbb{P}[(X,Y) \in A] = \mathbb{P}[(X,Y) \in T(A)] = \mathbb{P}\left[X \leq \frac{t}{\sqrt{a^2+b^2}}\right] = \mathbb{P}\left[\sqrt{a^2+b^2}X \leq t\right]$$

since the equation above holds for all $t \in \mathbb{R}$, we conclude that Z has the same distribution as $\sqrt{a^2+b^2}X$. Since X is standard normal, we know that $\sqrt{a^2+b^2}X \sim \mathcal{N}(0, a^2+b^2)$. Hence $Z = aX + bY \sim \mathcal{N}(0, a^2+b^2)$. □

Theorem 96 (Sum of Independent Normal RVs)**.** Let $X \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right)$ and $Y \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right)$ be independent normal random variables. Then for any constants $a, b \in \mathbb{R}$,

$$Z = aX + bY \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2).$$

Proof. $Z_1 = \frac{X-\mu_X}{\sigma_X}$ and $Z_2 = \frac{Y-\mu_Y}{\sigma_Y}$ are independent standard normal random variables. Then

$$Z = aX + bY = a\left(\mu_X + \sigma_X Z_1\right) + b\left(\mu_Y + \sigma_Y Z_2\right) = \left(a\mu_X + b\mu_Y\right) + \left(a\sigma_X Z_1 + b\sigma_Y Z_2\right)$$

By Theorem 95, $Z' = a\sigma_X Z_1 + b\sigma_Y Z_2 \sim \mathcal{N}(0, a^2\sigma_X^2 + b^2\sigma_Y^2)$ since $a\mu_X + b\mu_Y$ is a constant. Then by Theorem 94, $Z = \mu + Z' \sim \mathcal{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$, as desired. □

14.4 Central Limit Theorem

Here comes the most important theorem in this class:

Theorem 97 (Central Limit Theorem)**.** Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with common finite expectation $\mathbb{E}[X_i] = \mu$ and finite variance $\mathrm{Var}(X_i) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$. Then, the distribution of $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to $\mathcal{N}(0,1)$ as $n \to \infty$. In other words, for any constant $c \in \mathbb{R}$

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq c\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-x^2/2} dx \quad \text{as } n \to \infty$$

Proof. Out of scope. □

Remark. The CLT is a stronger claim that WLLN. It states that the distribution of the sample average S_n/n for large enough n converges to normal distribution with mean and variance both equal to those of the sample mean. Thus all trace of the distribution of X (no matter how complex) disappears as n gets large.

15 Finite Markov Chains

Definition 98 (Invariant/Stationary Distribution)**.** A distribution π is **invariant** for the transition probability matrix P if it satisfies the following balance equations:

$$\pi = \pi P.$$

Definition 99 (Irreducible)**.** A Markov chain is **irreducible** if it can go from every state i to every other state j, possibly in finite steps.

15.1 Hitting Time

References

[1] UC Berkeley *EECS70 Notes*.