CS70: Discrete Math and Probability Theory Notes

1. **Sets**

   (a) In mathematics, a **set** is a well-defined collection of distinct objects. The **cardinality**, or size, of a set A is denoted by $|A|$. If x is an element of A, we denote membership by $x \in A$.

   (b) If every element of a set A is also in set B, then we say that A is a **subset** of B, written $A \subseteq B$.

   (c) Two sets A and B are said to be **equal** if $A \subseteq B$ and $B \subseteq A$.

   (d) A **proper subset** is a set A that is strictly contained in B, written as $A \subset B$, meaning that A excludes at least one element of B.

   (e) The **intersection** of a set A with a set B, written as $A \cap B$, is a set containing all elements which are in both A and B.

   (f) The **union** of a set A with a set B, written as $A \cup B$, is a set of all elements which are in either A or B or both.

   (g) The **relative complement** of A in B, written as $B - A$ or $B \setminus A$, is the set of elements in B, but not in A.

   (h) The **Cartesian product** of A and B, written $A \times B$, is the set of all ordered pairs $\{(a, b) \mid a \in A, b \in B\}$.

2. **Important Sets**

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

3. **Propositional Logic**

   (a) A **proposition** is a declaration that is either true or false. Two propositions P and Q are logically equivalent, denoted $P \equiv Q$, if they share the same truth values.

   (b) Connectives:

   $$\text{And} : P \wedge Q \qquad \text{Or} : P \vee Q \qquad \text{Not} : \neg P$$

   Quantifiers:

   $$\text{There exists} : \exists \qquad \text{For all} : \forall$$

   (c) **Implication:** "If P, then Q", denoted $P \implies Q$. This is logically equivalent to $\neg P \vee Q$. (Note that implication is only false when P is true and Q is false)

   (d) The **contrapositive** of $P \implies Q$ is $\neg Q \implies \neg P$, which is logically equivalent.

   (e) The **converse** of $P \implies Q$ is $Q \implies P$, which is *not* logically equivalent.

   (f) **If and only if** (iff), denoted $P \iff Q$, is equivalent to implication in both directions: $(P \implies Q) \wedge (Q \implies P)$.

   (g) **Distribution Laws:**
   $P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge R)$
   $P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R)$

   (h) **DeMorgan's Laws**
   $\neg(P \vee Q) \iff \neg P \wedge \neg Q$
   $\neg(P \wedge Q) \iff \neg P \vee \neg Q$
   $\neg(\forall x P(x)) \iff \exists x(\neg P(x))$
   $\neg(\exists x P(x)) \iff \forall x(\neg P(x))$

   (i) Equivalencies:
   $\forall x(P(x) \wedge Q(x)) \iff (\forall x, P(x)) \wedge (\forall x, Q(x))$
   $\exists x(P(x) \vee Q(x)) \iff (\exists x, P(x)) \vee (\exists x, Q(x))$

4. **Formal Definitions and Important Lemmas**

   (a) Given integers a and b, we say that a divides b, denoted $a|b$, (or b is divisible by a) iff there exists an integer q such that $b = aq$.

   (b) A prime number is divisible only by 1 and itself.

   (c) An even number is an integer n of the form $n = 2k, k \in \mathbb{Z}$. An odd number is an integer n of the form $n = 2k + 1, k \in \mathbb{Z}$.

   (d) If a^2 is even, then a is even.

   (e) Every natural number greater than one is either prime or has a prime divisor.

   (f) The well-ordering principle states that every nonempty set of positive integers contains a least element.

5. **Proof techniques**

   (a) **Direct Proof:** Goal: Prove $P \implies Q$. Assume P ... therefore Q. Often involves appealing to definitions and using algebraic manipulations.

   (b) **Proof by contraposition:** Goal: Prove $P \implies Q$. Assume $\neg Q$ ... therefore $\neg P$.

   (c) **Proof by contradiction:** Goal: Prove P. Assume $\neg P$. Show that $\neg P \implies R \wedge \neg R$, a contradiction. Therefore, P.

(d) **Proof by cases:** In proving a claim, we don't know which of a set of possible cases is true, but we know that at least one of the cases is true. What we can do then is to prove the result in both cases; then, clearly the general statement must hold.

(e) **Proof by Induction:** Goal: Prove P. Choose an appropriate quantity n to induct on. In the *base case*, show that $P(1)$ is true. In the inductive hypotheses, assume $P(k)$ is true for some $k \geq 1$. In the inductive step, show that the claim holds for $P(k+1)$. Conclude that P holds for all n.

(f) **Strong Induction:** Same as induction, but with stronger inductive hypotheses: Assume $P(j)$ is true for all $1 \leq j \leq k$ for some k. Then show that $P(k+1)$ is true.

6. **Stable Marriage Algorithm:** Each man and woman has an ordered preference list of their n potential partners. The algorithm is as follows:

(a) Each man proposes to the most preferred woman on his list who has not yet rejected him.

(b) Each woman collects all the proposals she received; to the man she likes best, she responds "maybe" (puts him on a string), and to the others, she says "never".

(c) Each rejected man crosses off the woman who rejected him from his list.

Repeat until each women has a man on a string.
Properties of the algorithm:

(a) It outputs a stable pairing, meaning there are no rogue couples (a man and a woman who prefer each other to their current partners). It is not necessarily the only stable pairing.

(b) The algorithm must terminate in at most n^2 days (precisely, $n^2 - 2n + 2$).

(c) (Improvement lemma) If man M proposes to woman W on the k^{th} day, then on every subsequent day W has someone on a string whom she likes at least as much as M.

(d) (Male optimal and Female pessimal) It outputs a pairing such that each man is paired with his optimal woman (the most preferred woman a man can stably end up with).

7. **Graph Theory**

(a) An *undirected* graph G is defined by a set of vertices V and a set of edges E consisting of pairs of vertices. A *directed* graph models one-way relationships, so we have $E \subseteq V \times V$.

(b) The edge $e = \{u, v\}$ is *incident* on vertices u and v, and we say u and v are *adjacent.* If G is undirected, then the *degree* of vertex $u \in V$ is the number of edges incident to u. If G is directed, u has an associated *in-degree* and *out-degree.*

(c) Let G be an undirected graph. A **path** in G is a sequence of edges with no repeated vertices or edges (simple). A **cycle** is a path which starts and ends on the same vertex. A **walk** is a sequence of edges which can have repeated vertices. A **tour** is a walk which starts and ends at the same vertex. **Connected components** are sets $V_1, \ldots, V_k$ of vertices, such that all vertices in a set V_i are connected.

(d) A Eulerian walk is a walk in G that uses each edge exactly once (Eulerian tour if ends at same vertex). *Theorem:* An undirected graph G has a Eulerian tour iff G is even degree and connected (except for isolated vertices).

(e) Recursive algorithm for Eulerian Tour:

```
def Euler(G,s):
    T = findTour(G,s)
    return splice(T,Euler(G_1,s_1),...
        EULER(G_k,s_k))
```

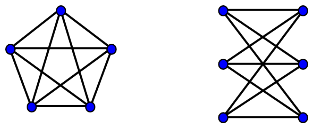
($G_1, \ldots, G_k$ are the connected components when the edges in T are removed from G, and s_i is the first vertex in T that intersects G_i. findTour(G, s), finds any tour from s. splice returns a combined tour obtained by traversing the edges of T, and whenever it reaches a vertex s_i that intersects another tour T_i, it takes a detour to traverse T_i from s_i back to s_i again, and only then it continues traversing T.)

(f) A graph is a **tree** if it is connected and acyclic. Equivalent definitinions include: There are $v - 1$ edges; the removal of any edge disconnects the graph; the addition of any edge creates a cycle. Unique path between any pair of vertices. Note that a tree must contain a degree 1 node (called a leaf).

(g) A graph is **planar** if it can be drawn on the plane without crossings. **Faces** f are the regions into which the graph subdivides the plane.

(h) Euler's Formula: For every connected planar graph, $v + f = e + 2$ (The converse is NOT true). Important consequence: $e \leq 3v - 6$ (Planar graphs are sparse).

(i) A graph is non-planar if and only if it contains K_5 or $K_{3,3}$:



(Contains means that one can identify nodes in the graph which are connected as K_5 or $k_{3,3}$ through paths such that each two of these paths share a vertex.)

(j) A **bipartite graph** is a graph whose vertices can be divided into two disjoint sets (no shared elements) U and V such that *every* edge connects a vertex in U to a vertex in V.

(k) A **Complete Graph** is a graph in which all nodes are connected to all other nodes. Let K_n denote the unique complete graph for n vertices. K_n has $n(n-1)/2$ edges.

(l) Useful facts:

i. By definition of a cycle, removing any edge in the cycle does not disconnect the graph G.

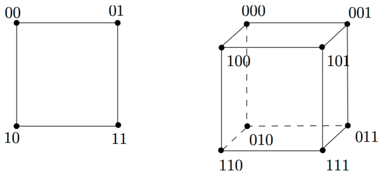
ii. If G is connected, then the addition of any vertex incident on more than one edge will create a cycle.

iii. The sum of the degrees of all vertices in a graph is $2|E|$ (twice the number of edges).

iv. In proofs relating to the maximum degree of a graph, do induction on the number of edges or vertices (with fixed maximum degree).

(m) **Hypercubes:**

The vertex set of the n-dimensional hypercube $G = (V, E)$ is given by $V = \{0,1\}^n$, where $\{0,1\}^n$ denotes the set of all n-bit strings. There are 2^n vertices and $n2^{n-1}$ edges. Two vertices x and y are connected by edge $\{x, y\}$ if and only if x and y differ in exactly one bit position.



Recursive definition of hypercube: Define the 0-subcube and 1-subcube as the $(n-1)$-dimensional hypercube with vertices labeled by $0x$ for $x \in \{0,1\}^{n-1}$ and $1x$ for $x \in \{0,1\}^{n-1}$, respectively. Then, the n-dimensional hypercube is obtained by placing an edge between each pair of vertices $0x$ in the 0-subcube and $1x$ in the 1-subcube.

8. Modular Arithmetic

(a) $x \pmod n = x - (\lfloor \frac{x}{n} \rfloor n)$

(b) We can perform any sequence of arithmetic operations (mod n) and the result is robust — it remains unchanged whether we reduce by (mod n) only once at the end of all operations, or we reduce each intermediate result by (mod n).

(c) For two integers a and b, we define a to be congruent to b modulo n ($a \equiv_n b$) iff:
$\iff a(\text{mod } n) = b(\text{mod } n)$
$\iff n|(a-b)$
$\iff a = b + kn$

(d) For all $n \geq 1$ and $a, b, c, d \in \mathbb{Z}$, if $a \equiv_n b$ and $c \equiv_n d$, then

i. $a + c \equiv_n b + d$

ii. $a - c \equiv_n b - d$

iii. $ac \equiv_n bd$

(e) The **multiplicative inverse** of $x \pmod n$ is y such that $xy \pmod n = 1$. *Theorem:* x has a multiplicative inverse modulo n if and only if $gcd(x, n) = 1$. Moreover, this inverse is unique.

(f) We note that if $x \geq y$, then $gcd(x, y) = gcd(y, x \pmod y)$. This allows to write the following algorithm:
Extended gcd algorithm for $x \geq y \geq 0$ and $x > 0$:

```
def extended-gcd($x, y$):
    if $y == 0$:
        return $(x, 1, 0)$
    $gcd, a, b$ = extended-gcd($y, x(\text{mod } y)$)
    return $(gcd, b, a - \lfloor \frac{x}{y} \rfloor b)$
```

It returns (gcd, a, b) such that gcd is $gcd(x, y)$ and a,b satisfy $gcd = ax + by$. Thus b is the multiplicative inverse of $x \pmod n$.
This algorithm uses $2n$ divisions, where n is the number of bits (an integer x requires $\lfloor log_2(x) \rfloor + 1$) bits).

9. Bijections

(a) A **function** $f : A \mapsto B$ maps values from a domain A to a range B. A function f is **bijective** iff it is **surjective** (onto): $\forall b \in B, \exists a \in A$ such that $f(a) = b$; and **injective** (one-to-one): $\forall a, a' \in A$, if $f(a) = f(a')$, then $a = a'$.

(b) *Lemma:* A function $f : A \mapsto A$ is a bijection iff there is an inverse function $g : A \mapsto A$ such that $f(g(x)) = x$ and $g(f(y)) = y \ \forall x, y \in A$.

(c) **Fermat's Little Theorum:** For any prime p and any $a \in \mathbb{Z}$, $a^p \equiv a \pmod p$. If a is not divisble by p, such as if $a \in \{1, 2, \ldots, p-1\}$, then we have $a^{p-1} \equiv 1 \pmod p$.

10. RSA

(a) RSA is a form of public key encryption which uses an easy-to-compute but difficult to invert (without the key) bijection.

(b) Encryption $E : \{0, \ldots, N-1\} \mapsto \{0, \ldots, N-1\}$ is given by

$$E(x) = x^e \pmod N$$

where $N = pq$ for two large primes p and q and e is relatively prime to $(p-1)(q-1)$.

(c) Decryption is given by

$$D(x) = x^d \pmod N$$

where d is the multiplicative inverse of e $(\text{mod } (p-1)(q-1))$.

(d) To show that RSA works, we show that $D(E(x)) = x$, that is, $x^{ed} \equiv x \pmod N$, or $x^{ed} - x \equiv 0 \pmod N$ for $x \in \{0, 1, \ldots, N-1\}$. Since $ed \equiv 1 \pmod{(p-1)(q-1)}$, or written another way, $ed = 1 + k(p-1)(q-1)$ for some integer k, we get $x^{ed} - x = x^{1+k(p-1)(q-1)} - x = x(x^{k(p-1)(q-1)})$. If x is a multiple of p, then clearly the claim follows. If not, use Fermat's Little Theorem to see the claim still follows.

(e) When we have to compute $x^y \pmod N$ for some large y, we use a simple technique known as repeated squaring to keep the number of multiplications $O(\log y)$ instead of $O(y)$.

(f) RSA is secure because given N, e, and $E(x)$, there is no efficient algorithm for finding x. (Amounts to factoring N into p and q). p and q are often over 100 digits.

11. **Polynomials**

(a) A single-variable **polynomial** of degree d is a function of the form

$$p(x) = a_d x^d + a_{d-1} x^{d-1} + \ldots + a_0$$

(b) a is a root of $p(x)$ iff $p(a) = 0$.

(c) **Property 1:** A degree d polynomial has at most d roots.
Proof: This follows from the following facts:

- If a is a root of $p(x)$ with degree d, then $p(x) = (x-a)q(x)$ for a polynomial $q(x)$ with degree $d-1$. (We use the result from polynomial division and show $r(x) = 0$).
- A polynomial $p(x)$ of degree d with distinct roots $a_1, \ldots, a_d$ can be written as $p(x) = c(x-a_1)\ldots(x-a_d)$ (This follows from fact 1).

(d) **Property 2:** Given $d+1$ pairs $(x_1, y_1), \ldots, (x_{d+1}, y_{d+1})$ of points, all x_i distinct, there is a unique polynomial of (at most) degree d that passes through all points.
Proof: Existence of a polynomial is proved by lagrange interpolation. Proof of uniqueness: suppose there is another such polynomial, $q(x)$. Consider a polynomial $r(x) = p(x) - q(x)$. $r(x)$ must a non-zero polynomial of degree d with at most d roots. But $r(x_i) = p(x_i) - q(x_i) = 0$ on $d+1$ distinct points. Contradiction.

(e) **Lagrange interpolation** is a method of reconstructing the unique polynomial of degree d given $d+1$ points $(x_1, y_1), \ldots, (x_{d+1}, y_{d+1})$. The lagrange polynomial is given by

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

where

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

This works because when it is evaluated at x_i, d of the $d+1$ terms in the sum go to 0 and the i^{th} term becomes y_i times 1, as desired.

An alternative method is to solve a system of $d+1$ linear equations where the i^{th} equation is $a_d x_i^d + a_{d-1} x_i^{d-1} + \ldots + a_0 = y_i$.

(f) In **polynomial division**, if we have a polynomial of degree d, we can divide it by a polynomial $q(x)$ of degree $\leq d$. The result will be

$$p(x) = q(x)q'(x) + r(x)$$

with $q'(x)$ the quotient and $r(x)$ the remainder with degree $< d$.

(g) **Finite fields**
Properties 1 and 2 do not hold if the coefficients and variable x are restricted to integers. But for numbers modulo a prime, we can add, subtract, multiply, and divide (by any nonzero number modulo m) since all the numbers have a multiplicative inverse mod m. When we work with numbers modulo a prime m, we say that we are working over a finite field, denoted by F_m or $GF(m)$.

(h) **Counting**
For polynomials of degree $\leq d$ over F_m: given $d+1$ points, there is exactly 1 polynomial that passes through all points. Given d points, there are m; given $d-1$ points, there are m^2, and so on, until given 0 points, there are m^{d+1}.

(i) **Secret Sharing**
We have n people and a secret natural number s. We want to devise a scheme whereby any k people can figure out s, but no group of $k-1$ or fewer has any information about s. Let q be a prime larger than n and s. We will now work over F_q. Pick a polynomial $P(x)$ of degree $k-1$ such that $P(0) = s$. Give $P(1)$ to the first person, $P(2)$ to the second, and so on, until you give $P(n)$ to the n^{th} person. For any $k-1$ people, s could be $0 \leq s \leq q-1$, but they already knew that, so they have no information about s.

12. **Error-correcting codes**

(a) When transmitting information on an unreliable channel (such as the Internet), sometimes packets are lost (**erasure errors**) and sometimes packets are corrupted (**general errors**). Error-correcting codes are a method of encoding messages to prevent both of these errors.

(b) **Erasure errors:** We will work in $GF(q)$ where q is a large prime. Assume the information consists of n packets, denoted $m_1, \ldots, m_n$. Let $P(x)$

be the unique polynomial of degree $n-1$ such that $P(i) = m_i$ for all $1 \leq i \leq n$. We can generate k additional packets by evaluating $P(x)$ at $n+1, \ldots, n+k$ (for $n+k \leq q$). The encoded message then consists of $n+k$ packets. We can reconstruct $P(x)$ from any n transmitted packets, then evaluate $P(x)$ at $x = 1, \ldots, n$ to obtain the original message. This scheme can protect against at most k erasure errors.

(c) **General errors:** As before, we work in $GF(q)$. To guard against k general errors, the encoded message must now contain $2k$ additional packets. We must now reconstruct $P(x)$ from $n+2k$ received messages, $r_1, \ldots, r_{n+2k}$. We know $P(i)$ must equal r_i on at least $n+k$ points, but we don't know which points are correct. Here is an efficient way to find $P(x)$:
Define the degree k polynomial $E(x) = (x - e_1) \ldots (x - e_k)$ to be the error-locator polynomial. Define $Q(x) = P(x)E(x)$. These polynomials take the following form:

$$Q(x) = a_{n+k-1}x^{n+k-1} + \ldots + a_1 x + a_0$$
$$E(x) = x^k + b_{k-1}x^{k-1} + \ldots b_1 x + b_0$$

For all $1 \leq i \leq n+2k$, the equation $Q(i) = r_i E(i)$ is a linear equation in the $n+2k$ unknowns ($a_{n+k-1}, \ldots, a_0$ and $b_{k-1}, \ldots, b_0$). So we get a linear system of $n+2k$ equations which we solve to get $Q(x)$ and $E(x)$. Then $P(x) = \frac{Q(x)}{E(x)}$. Remember to do all this mod q.

13. **Infinity and Countability**

(a) The **power set** $\mathcal{P}(S)$ of S is the set of all subsets of S. If $|S| = k$ is finite, $|\mathcal{P}(S)| = 2^k$.

(b) A set S is **countable** if there is a bijection between S and $\mathbb{N}$ or some subset of $\mathbb{N}$. By the isomorphism principle, if there is a bijection between two sets, then they have the same cardinality. If $|S| = |\mathbb{N}|$, S is said to be countably infinite.

(c) For example, we show $\mathbb{Z}$ is countably infinite by constructing the bijection $f : \mathbb{N} \to \mathbb{Z}$ defined as

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

(d) Another way to compare cardinality is to show the existence of a one-to-one function $f : A \to B$ and a one-to-one function $g : B \to A$, which implies the existence of a bijection.

(e) Using this idea, we can (surprisingly) show $\mathbb{Q}$ and $\mathbb{N}$ have the same cardinality, using a spiral integer lattice argument.

(f) The set of all binary strings of finite length, $\{0,1\}^*$, also has a bijection to $\mathbb{N}$. The idea is to construct a bijection $f : \{0,1\}^* \to \mathbb{N}$, where we define $f(s)$ to be the index of string s in an enumerated list of all finite-length binary strings enumerated by first listing all strings in increasing order of length, then by lexicographic order.

(g) **Cantor's Diagonalization Proof** shows that $\mathbb{R}$ is uncountable. The idea is as follows: suppose a bijection $f : \mathbb{N} \to \mathbb{R}[0,1]$ exists; enumerate the infinite list. Consider the real number D formed by the diagonal of the list; then modify every digit by say $+2 \pmod{10}$, call this number S. S does not exist in the list (suppose it existed as the i^{th} number, then the i^{th} digit would be the same as D's i^{th} digit), contradicting the fact that f is a bijection.

14. **Computability**

(a) Alan Turing showed that there is no program that can determine whether another program will halt. The idea is as follows: suppose such a program `TestHalt` exists.

$$\texttt{TestHalt(P,x)} = \begin{cases} \text{True} & \text{if P halts on x} \\ \text{False} & \text{if P loops on x} \end{cases}$$

Now consider the program `Turing`:

```
def Turing(P):
    if TestHalt(P,P): loop forever
    else: halt
```

where we treat the bit string representation of P as input to `TestHalt`. Now consider `Turing(Turing)`. Whether it halts or not, it leads to a contradiction like that of the liar's paradox. This paradox shows the assumption that `TestHalt` exists was False.

(b) It turns out that writing a program that tests if P halts on a specific input 0 is also impossible, because if it was, then `TestHalt` would be solvable (consider a program that P' that takes 0 as input and simply returns P(x)).

(c) **Godel's Incompleteness Theorem** shows that arithmetic (or any formal system that formalizes computation) is either *inconsistent* (there are false statements that can be proved) or *incomplete* (there are true statements that cannot be proved). Let S be the statement "P halts on input 0". Since we assume arithmetic is complete and consistent, S is true or false and there is a proof either way. All proofs are finite binary strings, so there are countably many proofs. Consider a program that takes as input the program P, and proceeds to check every possible proof until it finds one that proves either P(0) halts, or P(0) does not halt. This program will terminate in finite time, and it will correctly answer the Easy Halting Problem (contradiction!)

15. **Counting**

    (a) **First rule of counting:** If an object can be made by a succession of k choices where there are n_1 ways of making the first choice, n_2 ways of making the second choice, and so on up to the n_k-th choice, then the total number of distinct objects is the product of $n_1, \ldots, n_k$.

    (b) The number of ways to draw k samples from n items (permutations):

    - Sample with replacement, order matters:

    $$n^k$$

    - Sample without replacement, order matters:

    $$\frac{n!}{(n-k)!}$$

    (c) **Second rule of counting:** The number of ways to select k distinct elements of a set of size n where order is irrelevant (i.e. a hand in poker) is the number of ways if order mattered, divided by the number of possible orderings.

    (d) The number of ways to choose k elements from a set S of size n (combinations):

    - Sample without replacement, order irrelevant:

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

    - Sample with replacement, order irrelevant:

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

    We see this by considering stars and bars: how many ways can you distribute k balls into n bins? A solution is represented by k stars and $n-1$ bars, or separators. Any solution can be represented as choosing k out of $k+n-1$ positions to place stars and filling the remaining spaces with bars.

    (e) To generalize balls and bins:

    - Indistinguishable balls $\implies$ order irrelevant
    - Distinguishable balls $\implies$ order matters
    - One ball per bin $\implies$ without replacement
    - Multiple balls allowed per bin $\implies$ with replacement

16. **Discrete Probability**

    (a) The outcome of a **random experiment** is called a **sample point**, denoted ω. The **sample space**, often denoted by Ω, is the set of all possible outcomes.

    (b) A **probability space** is a sample space along with a probability $P[\omega]$ for each sample point ω such that $\sum_{\omega \in \Omega} P[\omega] = 1$.

    (c) For an **event** $A \subset \Omega$, we define

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

    (i.e., the sum of the probabilities of the sample points in A). If the probability space is uniform, then $P(A) = \frac{|A|}{|\Omega|}$.

    (d) An event B is the **complement** (sometimes denoted $\bar{A}$) of event A if it consists of precisely those sample points which are not in A. So $P(B) = 1 - P(A) = P(\neg A)$.

17. **Conditional Probability**

    (a) $P(A|B)$ denotes the probability of event A given event B. For events A and B in the same probability space, define

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

    where $P(A \cap B)$ (or $P(A, B)$) is the probability of A and B.

    (b) **Bayes' Rule:**

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

    (c) **Total Probability Rule:** If A partitions Ω, then

    $$P(B) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$$

    which can be generalized further if A is not binary.

    (d) Events A and B are

    i. **independent** if $P(A \cap B) = P(A)P(B)$

    ii. **positively correlated** if $P(A \cap B) > P(A)P(B)$

    iii. **negatively correlated** if $P(A \cap B) < P(A)P(B)$

    (e) In general, n events are **mutually independent** if *every* subset of the n events is independent (so pairwise independence does *not* imply mutual independence).

    (f) **Intersections (ands) of Events:** For intersections of possibly *non-independent* events, we use the **product rule:**

    $$P(A_1, \ldots, A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_{n-1}, A_{n-2}, \ldots, A_1)$$

    (g) **Unions (ors) of Events:** If the events are disjoint or **mutually exclusive**, that is, for any pair, they cannot *both* occur, the probability is the sum of the individual probabilities. Otherwise, this is an overestimate (called the **union bound**). Need to subtract out probabilities of pairwise intersections, and add back the probabilities of three-way intersections, and so on.

(h) *Symmetry:* Suppose we draw k cards from a deck. The probability that the i^{th} card drawn is the queen of spades is the same as the probability that the j^{th} card is the queen of spades. Note that this changes as soon as the outcome of another draw is known.

18. **Hashing and Load Balancing**

(a) We model hashing as placing k balls (representing keys) into n bins (representing array indices), assuming a uniform and random hash function.

(b) Let A be the event that there are no collisions. What is the largest value of k such that $P(A)$ is (say) $\geq \frac{1}{2}$?
First enumerate all $m = \binom{k}{2} = \frac{k(k-1)}{2}$ possible pairs of keys. Then the event some collisions occurs $\bar{A} = \cup_{i=1}^{m} A_i$, where A_i denotes the event that the i^{th} pair of keys has a collision. Since $P(A_i) = \frac{1}{n}$, by the union bound we have

$$P(\bar{A}) \leq \sum_{i=1}^{m} P(A_i] = m \times \frac{1}{n} = \frac{k(k-1)}{2n} \approx \frac{k^2}{2n}$$

So $P(A) \geq \frac{k^2}{2n}$ and we should have $k = \sqrt{n}$ for the probability of no collisions to be at least $\frac{1}{2}$. Tighter bounds can be drawn (see below), but the relationship is still $O(\sqrt{n})$.

(c) Advanced: $P(A) \approx e^{-\frac{k^2}{2n}}$. The largest value of k such that $e^{-\frac{k^2}{2n}} \geq \frac{1}{2}$ is $\approx 1.177\sqrt{n}$.

(d) In load balancing, each task selects a processor uniformly at random. let A_k be the event that the load of some processor is at least k. We want to find the lowest value of k such that $P(A_k) \leq \frac{1}{2}$. (Then with probability $\frac{1}{2}$, every processor will have load at most k).

(e) Assume the number of tasks is equal to the number of processors, n. We will find k such that $P(A_k(1)) \leq \frac{1}{2n}$, because then we will know that $\forall i, P(A_k(i)) \leq \frac{1}{2n}$ and hence $P(A_k) \leq n \times \frac{1}{2n} = \frac{1}{2}$ by the union bound. We do this by bounding the probability that bin 1 has at least k balls, then finding k so that this bound is less than $\frac{1}{2n}$. We find that $P(A_k(1)) \leq \binom{n}{k}\frac{1}{n^k}$. Taking $k = \frac{\ln n}{\ln \ln n}$ satisfies $\binom{n}{k}\frac{1}{n^k} \leq \frac{1}{2n}$.

(f) Coupons: There are n different coupons, one at random in each cereal box. You buy m boxes. The probability you miss a specific coupon is $(1-\frac{1}{n})^m$. This is approximately $e^{-\frac{m}{n}}$. By the union bound, the probability you miss at least one coupon is $\leq ne^{-\frac{m}{n}}$.

19. **Random Variables and Expectation**

(a) A **random variable** X for an experiment with sample space Ω is a function $X : \Omega \mapsto \mathbb{R}$ that assigns each sample point $\omega \in \Omega$ a real number $X(\omega)$.

(b) The event $X = x$ is the subset of outcomes ω such that $X(\omega) = x$.

(c) The **distribution** of X is the set $x, P(X = x))$ for all x in the range of X. This forms a partition of Ω, so the sum of $P(X = x)$ for all x is 1.

(d) Two random variables X and Y are **independent** if $\forall x, y\ P(X = x, Y = y) = P(X = x)P(Y = y)$. Functions of independent random variables are also independnent.

(e) The **expectation** of a discrete random variable X (mean or average) is defined as

$$E(X) = \sum_{x} xP(X = x)$$

Equivalently,

$$E(X) = \sum_{\omega \in \Omega} X(\omega) \times P(\omega)$$

(f) *Linearity of expectation:* For any two random variables X and Y in the same probability space, we have

$$E(X + Y) = E(X) + E(Y)$$
$$E(cX) = cE(x)$$

If X and Y are independent, we also have

$$E(XY) = E(X)E(Y)$$

Linearity is very powerful, as demonstrated by the examples below.

(g) Let X be the number of students who get their own homework back after shuffling in a class size of 20 (the number of fixed points). We can write X as $X = X_1 + X_2 + \ldots + X_{20}$, where X_i is 1 if student i receives their own homework, and 0 else. Such a $\{0, 1\}$-valued random variable 1_A is called an **indicator** variable of event A, and is said to have the **Bernoulli** distribution. The expectation is easy to calculate:

$$E(1_A) = P(1_A = 1)$$

(We also have $\text{Var}(1_A) = P(A)(1 - P(A))$). In our example, $P(X_i = 1) = \frac{1}{20}$. Use linearity of expectation to see that $E(X) = 20 \times \frac{1}{20} = 1$. In other words, the expected number of fixed points in a random permutation of n items is always 1, regardless of n.

(h) Hashing examples: Throw k balls into n bins.
The *expected number of balls per bin?*
Let X be the number of balls in a particular bin. Then $X = X_1 + X_2 + \ldots + X_k$, where X_i is 1 if ball i lands in that bin, otherwise 0. $E(X_i) = P(X_i = 1) = \frac{1}{n}$, so $E(X) = \frac{k}{n}$.
The *expected number of empty locations*?
Let Y be the number of empty locations. Then

$Y = Y_1+Y_2+\ldots+Y_n$, where Y_i is 1 if bin i is empty and 0 otherwise. The probability a bin is empty is $P(Y_i = 1) = (1 - \frac{1}{n})^k$, so $E(Y) = n\left(1 - \frac{1}{n}\right)^k$.
The *expected number of collisions*?
A collision occurs when we hash an item to a location that already contains an item. Use the result from above; there are $n - Y$ non-empty locations, so there are $k - (n - Y) = k - n + Y$ collisions. Substitute the value for $E(Y)$ to see there are $k - n + n\left(1 - \frac{1}{n}\right)^k$ expected collisions.

(i) Coupon example: How many boxes do we need to buy before we have all n coupons? Let X denote this number. Then $X = X_1 + X_2 + \ldots + X_n$, where X_i denotes the number of boxes we buy before getting a new coupon (not one of $i - 1$ we already have). X_i has the geometric distribution where probability of success is $1 - \frac{i-1}{n} = \frac{n-i+1}{n}$, so $E(X_i) = \frac{1}{\frac{n-i+1}{n}} = \frac{n}{n-i+1}$. Then $E(X) = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\right) = n\sum_{i=1}^{n}\frac{1}{i} \in O(n \ln n)$.

(j) If X is a random variable and $f : \mathbb{R} \mapsto \mathbb{R}$ is a function, then $f(X)$ is a random variable. We have

$$E(f(X)) = \sum_x f(x)P(X = x)$$

More generally,

$$E(f(X, Y)) = \sum_{x,y} f(x, y)P(X = x, Y = y)$$

20. **Variance**

(a) For a random variable X with expectation $E(X) = \mu$, the **variance** is defined as

$$\mathrm{Var}(X) = E\left((X - \mu)^2\right)$$

The **standard deviation** $\sigma(X)$ is defined as $\sqrt{\mathrm{Var}(X)}$.

(b) Observations:

i. $\mathrm{Var}(X) = E(X^2) - E(X)^2$

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

iii. $\mathrm{Var}(c + X) = \mathrm{Var}(X)$

(c) For two random variables X and Y, we have

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{cov}(X, Y)$$

For *independent* random variables, this becomes

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

(d) Computing the variance of *dependent* random variables that can be written as a sum of indicator variables is mostly painless. Let $X = 1_{A_1} + \ldots + 1_{A_n}$, where the indicators are identically distributed. Then $E(X) = nP(A_i)$. We have $E(X^2) = \sum_{i=1}^{n}[E(X_i^2) + \sum_{j\neq i} E(X_iX_j)]$. Therefore $\mathrm{Var}(X) = E(X^2) - E(X)^2 = nP(A_i) + n(n - 1)P(A_i \cap A_j) - n^2P(A_i)^2$.

21. **Distributions**

(a) **Binomial distribution:** Models the number of successes in n independent trials. Define X to be the number of heads after flipping a biased coin n times that lands heads with probability p. The probability of the event $X = i$ is the sum of the probabilities of all the sample points with exactly i heads. Any such sample point has probability $p^i(1 - p)^{n-i}$ and there are $\binom{n}{i}$ of these sample points, so

$$P(X = i) = \binom{n}{i}p^i(1 - p)^{n-i}$$

Clearly $E(X) = np$. $\mathrm{Var}(X) = np(1 - p)$.

(b) **Uniform distribution:** where X is uniformly distributed in $\{1, \ldots, n\}$, then $E(X) = \frac{n+1}{2}$. $\mathrm{Var}(X) = \frac{n^2-1}{12}$.

(c) **Geometric distribution:** A random variable X for which

$$P(X = i) = (1 - p)^{i-1}p$$

Usually, X is the number of independent trials *until* a certain event happens with probability p. We have

$$E(X) = \frac{1}{p}$$

To see this, we make use of the Tail Sum formula: *Let X be a random variable that only takes values in $\mathbb{N}$. Then*

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

For a geometric distribution, $P(X \geq i) = (1 - p)^{i-1}$. Using the geometric series formula and the above theorem yield the result.
In addition we have $\mathrm{Var}(X) = \frac{1-p}{p^2}$. We say that the system is memoryless, that is, $P(X > n + m | X > n) = P(X > m)$.

(d) **Negative Binomial Distribution:** Generalization of the geometric distribution: how many trials do we need to get k successes?

$$P(X = i) = \binom{i-1}{k-1}p^k(1 - p)^{i-k}$$

When $k = 1$, this is just the geometric distribution. We have

$$E(X) = \frac{k}{p}$$

In addition, $\mathrm{Var}(X) = \frac{k(1-p)}{p^2}$.

(e) **Poisson distribution:** A random variable X for which

$$P(X = i) = \frac{\lambda^i}{i!}e^{-\lambda}$$

This is the binomial distribution under specific circumstances; when the number of trials n is large and $p = \frac{\lambda}{n}$ for some fixed λ. We have

$$E(X) = \lambda$$

Also, $\mathrm{Var}(X) = \lambda$. If X and Y are Poisson random variables with parameters λ_1 and λ_2, respectively, then $X + Y$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

22. **Inequalities and Confidence Intervals**

(a) **Markov's Inequality:** For any random variable X and any increasing, non-negative function f, we have

$$P(X \geq a) \leq \frac{E(f(X))}{f(a)}$$

for $a \in \mathbb{R}$, $a \neq 0$.

(b) **Chebyshev's Inequality:** For random variable X and $a > 0$,

$$P(|X - E(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

This follows from letting $Y = |X - E(X)|$ and $f(y) = y^2$, and applying Markov's inequality to Y.

(c) **Law of Large Numbers:** Let $X_1, \ldots, X_n$ be pairwise independent with the same distribution and mean μ. Then, for all $\epsilon > 0$, in the limit as $n \to \infty$,

$$P\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0$$

This follows from letting $Y = \frac{X_1+\ldots+X_n}{n}$, and applying Chebyshev's inequality to see $P(|Y - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(X_i)}{n\epsilon^2}$, which tends towards 0 as $n \to \infty$.

(d) **Confidence intervals:** Let $Y = \frac{X_1+\ldots+X_n}{n}$, where $X_1, \ldots, X_n$ are independent and identically distributed random variables. Y can be thought of as a sample average of n samples of X. Each X_i has mean μ and variance σ^2. We are interested in

$$P(|Y - \mu| < a) \geq \delta$$

that is, with confidence δ, the sample average Y is less than a away from the true mean μ.
Observe that $\mathrm{Var}(Y) = \mathrm{Var}(\frac{1}{n}(X_1 + \ldots + X_n)) = \frac{1}{n^2} \cdot n\mathrm{Var}(X_1) = \frac{\sigma^2}{n}$. We have

$$\begin{aligned} P(|Y - \mu| < a) &= 1 - P(|Y - \mu| \geq a] \\ &\geq 1 - \frac{\mathrm{Var}(Y)}{a^2} = 1 - \frac{\sigma^2}{na^2} \end{aligned}$$

Thus $1 - \frac{\sigma^2}{na^2}$ represents our confidence δ. σ^2 can be determined from the situation at hand. If we set $\delta = 0.95$, for example, we can solve for a or n. We find

$$\left(Y - \frac{4.5\sigma}{\sqrt{n}}, Y + \frac{4.5\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for μ. If each X_i is a Bernoulli variable (i.e., X_i is an indicator variable), then $\sigma^2 \leq \frac{1}{4}$, so $\sigma \leq \frac{1}{2}$ and we see

$$\left(Y - \frac{2.25}{\sqrt{n}}, Y + \frac{2.25}{\sqrt{n}}\right)$$

is a 95% confidence interval for p. In practice, not always easy to come up with an upper bound for σ.

23. **Linear Regression and Conditional Expectation**

(a) **Covariance:** The covariance of two random variables X and Y is defined as

$$\mathrm{cov}(X, Y) = E\left[(X - E(X))\,(Y - E(Y))\right]$$

(b) For variables A and B, we have that A and B are

i. $\mathrm{cov}(A, B) > 0$: *positively correlated*

ii. $\mathrm{cov}(A, B) < 0$: *negatively correlated*

iii. $\mathrm{cov}(A, B) = 0$: *uncorrelated* (does *not* imply independence, but independence *does* imply uncorrelated)

(c) Observations:

i. $\mathrm{cov}(X, Y) = E(XY) - E(X)E(Y)$

ii. $\mathrm{Var}(X) = \mathrm{cov}(X, X)$

iii. $\mathrm{cov}(X, Y_1 + Y_2) = \mathrm{cov}(X, Y_1) + \mathrm{cov}(X, Y_2)$

iv. $\mathrm{cov}(aX, Y) = a\ \mathrm{cov}(X, Y)$

(d) How do we compute $E(XY)$? Try to write out X and Y as sums of indicator variables. Then

$$\begin{aligned} &E[(X_1 + \ldots + X_n)(Y_1 + \ldots + Y_n)] \\ &= \sum_{i=1}^{n}\left[E[X_iY_i] + \sum_{j\neq i} E[X_iY_j]\right] \\ &= nP(X_i, Y_i) + n(n-1)P(X_i, Y_j) \end{aligned}$$

(e) The **linear least squares estimate (LLSE)** of Y given X is defined as

$$L(Y|X) = E(Y) + \frac{\mathrm{cov}(X, Y)}{\mathrm{Var}(X)}(X - E(X))$$

(f) Projection Property: The LLSE satisfies

$$\begin{aligned} E\left(Y - L(Y|X)\right) &= 0 \\ E\left((Y - L(Y|X))X\right) &= 0 \end{aligned}$$

(g) Least-Squares Property: Define $L(X) = \{aX + b : a, b \in \mathbb{R}\}$ to be the set of linear functions of X. Then for any $aX + b \in L(X)$,

$$E\left((Y - L(Y|X))^2\right) \leq E\left((Y - aX - b)^2\right)$$

That is, $L(Y|X)$ is the linear function that minimizes the mean-squared error. The mean-squared error is $E\left((Y - L(Y|X))^2\right) = \text{Var}(Y) - \frac{\text{cov}(X,Y)^2}{\text{Var}(X)}$.

(h) *Linear Algebra Interpretation*: The space of random variables over a probability space is a vector space. $\langle X, Y \rangle := E(XY)$ is an inner product, so X and Y are orthogonal if $E(XY) = 0$. The projection property states that $Y - \hat{Y}$ is orthogonal to 1 and X. 1 and X form a basis for $L(Y|X)$, so $\hat{Y}$ is the projection of Y onto $L(X)$.

(i) We can use this interpretation to develop the theory for quadratic functions. Define $Q(X) = \{aX^2 + bX + c : a, b, c \in \mathbb{R}\}$ to be the set of quadratic functions of X. We want to find the projection of Y onto $Q(X)$. We must have $Y - \hat{Y} = Y - aX^2 - bX - c$ be orthogonal to a basis of $Q(X)$, say $1, X$ and X^2. That is,

$$\begin{aligned} E(Y - aX^2 - bX - c) &= 0 \\ E((Y - aX^2 - bX - c)X) &= 0 \\ E((Y - aX^2 - bX - c)X^2) &= 0 \end{aligned}$$

We can solve these equations to find a, b, c.

(j) We have assumed that the distributions of X and Y are known (Bayesian linear regression). In practice, we often observe samples instead. We can still work with the LLSE equation if we assume (X, Y) is uniform on the set of observed samples. The expectations, covariance, and variance are computed from the samples with this assumption in mind.

(k) **Conditional Expectation:**

$$E(X|Y = y) = \sum_x xP(X = x|Y = y)$$

The random variable $E(X|Y)$ is a function of Y, called the *conditional expectation of X given Y*. It takes on value $E(X|Y = y)$ with probability $P(Y = y)$.
Properties:

i. $E(aY_1 + bY_2|X) = aE(Y_1|X) + bE(Y_2|X)$
ii. $E(h(X)Y|X) = h(X)E(Y|X)$
iii. $E(E(X|Y)) = E(X)$ (called the **law of iterated expectation**)

It is useful to keep in mind that $E(X|Y)$ becomes a constant when conditioned on Y.

(l) **MMSE:** Now we return to the task of finding the best estimator of Y given X in more generality. Suppose we wish to find the function $f(X)$ that minimizes

$$E((Y - g(X))^2)$$

for *all* (not necessarily linear or quadratic) possible functions $g(\cdot)$. This function $f(X)$ is called the minimum mean square error (MMSE) estimator of Y given X, and is given by $E(Y|X)$. The proof is based on the **Orthogonality Property**:

$$E((Y - E(Y|X))\phi(X)) = 0$$

This property, analagous to the projection property, shows that $E(Y|X)$ is the projection of Y onto the space of *all* functions of X, $\phi(X)$. That is, the best estimator of Y given X is simply the expected value of Y given X.

24. **Markov Chains**

(a) A **Markov chain** models how processes evolve over time. The random variable X_n denotes the state of the system at time n. It obeys the Markov property, i.e., the past is conditionally independent of the future given the present. Explicitly, a (finite) Markov chain consists of:

i. A set of states $\{1, \ldots, N\}$.
ii. An initial probability distribution over the states π_0, where $\pi_0(i) = P(X_0 = i)$.
iii. Transition probabilities $P(i, j)$, where $P(i, j) = P(X_n = j|X_{n-1} = i)$, i.e., the probability of moving from state i to state j.
iv. The distribution of X_n, which we will denote as $\pi_n(i) = P(X_n = i)$.

The N^2 transition probabilities can be organized into a matrix P such that $P(i, j)$ is the entry in the i^{th} row and j^{th} column. We use the convention that the rows must sum to 1 (although it is more common to have the columns sum to 1).

(b) Writing π_n as a row vector, we have

$$\pi_n = \pi_{n-1}P$$

which leads to

$$\pi_n = \pi_0 P^n$$

(c) Suppose we start in state s and want to find the expected number of steps until reaching another state s'. Define $\beta(i)$ to be the expected time to reach state s', starting at state i (thus $\beta(s') = 0$). We write the **first step equations**:

$$\beta(i) = 1 + \sum_{j=1}^{N} P(i, j)\beta(j)$$

That is, we take 1 step and add the expected number of steps to s' from every state j, weighted by the probability of transitioning to j. Calculating $\beta(i)$ for each state yields N linear equations which we can solve to find $\beta(s)$, the desired quantity.

(d) To generalize, suppose each state i gives a reward $R(i)$ for visiting, and we want to know the expected reward we accumulate before reaching a set of states S' (so above, we had $R(i) = 1$ and $S' = \{s'\}$). This can be useful for counting the number of times we expect to hit a certain state, for example. Define $\gamma(i)$ to be the expected sum of rewards starting from i until hitting a state in S'. Then, as before, we can solve

$$\gamma(i) = R(i) + \sum_{j=1}^{N} P(i,j)\gamma(j)$$

(e) Define $\alpha(i)$ to be the probability of reaching S before S' starting from state i, where S and S' are sets of states. Then $\alpha(s) = 0 \; \forall s \in S'$ and $\alpha(s) = 1$ $\forall s \in S$. Then, as before, we can solve

$$\alpha(i) = \sum_{j=1}^{N} P(i,j)\alpha(j)$$

(f) A Markov chain is **irreducible** if from any state we can reach any other state (possibly in multiple steps). A distribution π is an **invariant** distribution if $\pi = \pi P$. Hence if we start with an invariant distribution π_0, then $\pi_0 = \pi_n \; \forall n$.

(g) Theorem: *A finite, irreducible Markov chain has a unique invariant distribution. Furthermore, the fraction of time spent in state i after n steps approaches $\pi(i)$ as $n \to \infty$.*

(h) The invariant distribution satisfies $\pi = \pi P$, which is equivalent to $\pi(P - I) = 0$. This matrix-vector multiplication can be written out as a linear system of N equations, called the **balance equations**. (In linear algebra terms, this is finding a basis of eigenvectors corresponding to eigenvalue 1). Add in the condition that $\sum_{i=1}^{N} \pi(i) = 1$ to obtain the unique eigenvector (invariant distribution).

(i) Convergence to the unique invariant distribution is not guaranteed, as the initial distribution can sometimes matter. Define the **period** of a state i to be the largest integer d, $d(i) \geq 1$, such that the number of (all) time steps required to return to i is a multiple of d. In other words, $d(i)$ is the GCD of the set of integers $n > 0$ where we can return to i in n steps. A Markov chain is **aperiodic** if the period of every state is 1.

(j) Theorem: *For a finite, irreducible Markov chain, the period of every state is the same.*

(k) Theorem: *For a finite, irreducible, aperiodic Markov chain, the distribution converges to the limiting distribution (regardless of initial distribution).*

25. **Continuous Probability**

(a) A *continuous* random variable X can take on infinitely many values; the probability that $X = a$ is 0 for all $a \in \mathbb{R}$. So we must instead consider the probability of the event that X lies within some interval of values.

(b) The **probability density function** (PDF) of a continuous random variable X is a real-valued function $f_X(x)$ such that

i. f_X is non-negative: $\forall x \in \mathbb{R}, f_X(x) \geq 0$.

ii. f_X is normalized: $\int_{\mathbb{R}} f_X(x)dx = 1$.

We can interpret $f_X(x)$ as the probability per unit length near x.

(c) We define the probability that X lies in some interval $[a, b]$ as

$$P(X \in [a,b]) := \int_a^b f_X(x)dx$$

(d) The **cumulative distribution function** (CDF) is defined as

$$F_X(a) := P(X \leq a) = \int_{-\infty}^{a} f_X(x)dx$$

Notice $f_X(x) = \frac{d}{dx}F_X(x)$. Differentiating the CDF is often a quick way to find the PDF.

(e) The discrete results carry over to continuous random variables by replacing summations over x with integration over $\mathbb{R}$. For example, the expectation of a function g of X is

$$E(g(X)) := \int_{\mathbb{R}} g(x)f_X(x)dx$$

(f) The joint distribution of X and Y is $f_{X,Y}(x,y)$. X and Y are independent iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

(g) The **conditional density** of Y given $X = x$ is

$$f_{Y|X=x}(y) := \frac{f_{X,Y}(x,y)}{f_X(x)}$$

Important Continuous Distributions

(h) **Uniform Distribution:** The PDF of a random variable X with the uniform distribution on $[a, b]$ is

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

The CDF is

$$F_X(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x < b \\ 1 & x \geq b \end{cases}$$

We find $E(X) = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$.

(i) **Exponential Distribution:** The exponential distribution is the continuous version of the geometric distribution. An exponential random variable X with parameter $\lambda > 0$ has the PDF

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

for $x \geq 0$. The CDF is

$$F_X(x) = 1 - e^{-\lambda x}$$

We can interpret λ to be the success rate per unit time. We have

$$E(X) = \frac{1}{\lambda} \qquad \text{Var}(X) = \frac{1}{\lambda^2}$$

We also have that the minimum of n independent exponential variables with parameters $\lambda_1, \ldots, \lambda_n$, respectively, is also exponentially distributed with parameter $\lambda_1 + \ldots + \lambda_n$ (In general, these properties are easily derived using the CDF).

(j) **Normal (Gaussian) Distribution:** A normal random variable X with parameters μ and $\sigma > 0$ has the PDF

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

If $\mu = 0$ and $\sigma = 1$, then X has the standard normal distribution. If X is a normal variable, then $Y = \frac{X-\mu}{\sigma}$ is a standard normal variable. We have

$$E(X) = \mu \qquad \text{Var}(X) = \sigma^2$$

We also have that the sum of n independent random variables with parameters μ_i and σ_i is also normally distributed with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i$.

(k) **Central limit theorem:** Recall that the law of large numbers states that the probability of *any* deviation of the sample average of n independent random variables from the true mean tends to 0 as $n \to \infty$.

The **central limit theorem** states something stronger: the distribution of the sample average of n observations of an arbitrary random variable X converges to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Precisely, let $Y = \frac{X_1 + \ldots + X_n}{n}$, where $X_1, \ldots, X_n$ are independent and identically distributed random variables each with mean μ and variance σ^2. Recall that $\text{Var}(Y) = \frac{\sigma^2}{n}$. Define $Z = \frac{(Y-\mu)\sqrt{n}}{\sigma}$. The CLT says Z approaches the standard normal distribution as $n \to \infty$.

(l) **Confidence Intervals revisited:** Using the same definitions provided above, we are interested in

$$P(|Y - \mu| < a) \geq \delta$$

that is, with confidence δ, the sample average Y is less than a away from the true mean μ. We have

$$\begin{aligned} P(|Y - \mu| < a) &= P\left(\left|\frac{(Y-\mu)\sqrt{n}}{\sigma}\right| < \frac{a\sqrt{n}}{\sigma}\right) \\ &= P\left(Z < \frac{a\sqrt{n}}{\sigma}\right) \\ &\approx \int_{-a\sqrt{n}/\sigma}^{a\sqrt{n}/\sigma} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \end{aligned}$$

In the last line, we use the CLT to justify approximating the expression as the CDF of the standard normal distribution. Setting this expression to be ≥ 0.95, we solve for a to find

$$\left(Y - \frac{1.96\sigma}{\sqrt{n}}, Y + \frac{1.96\sigma}{\sqrt{n}}\right)$$

is a 95% confidence interval for μ. Observe that this is much tighter than that derived from Chebyshev's inequality earlier.