My Notes for Discrete Probability: Detailed Explanations on Concepts and Proofs

Mohamed Syaheer Altaf

September 23, 2025

1 Events and Probability

In this note, we consider only discrete probability spaces. Recall some basic terminologies:

- **Experiment:** any procedure whose outcomes are well-defined and which is repeatable indefinitely.
- **Sample space:** the set of all possible, well-defined outcomes of an experiment.
- **Event:** a subset of the sample space.

1.1 Probability

Definition 1 (Probability Space)**.** *A probability space has three components:*

1. *a sample space Ω.*
2. *a family of sets $\mathcal{F}$ representing* ***events****, where each set in $\mathcal{F}$ is a subset of the sample space Ω (i.e., $\mathcal{F} = 2^{\Omega}$).*
3. *a probability function $Pr : \mathcal{F} \longrightarrow \mathbb{R}$ satisfying Definition* **??***.*

Definition 2 (Probability Axioms)**.** *A probability function is any function $Pr : \mathcal{F} \longrightarrow \mathbb{R}$ that satisfies the following conditions:*

1. *For any event $\mathcal{E} \in \mathcal{F}$, $Pr(\mathcal{E}) \geq 0$.*
2. *$Pr(\Omega) = 1$.*
3. *For any finite or countably infinite sequence of* ***pairwise mutually disjoint*** *events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \ldots,$*

$$Pr\Big(\bigcup_{i \geq 1} \mathcal{E}_i\Big) = \sum_{i \geq 1} Pr(\mathcal{E}_i).$$

Lemma 1 (Inclusion-Exclusion Principle)**.** *Let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ be any n events. Then*

$$Pr\Big(\bigcup_{i=1}^{n} \mathcal{E}_i\Big) = \sum_{i=1}^{n} Pr(\mathcal{E}_i) - \sum_{1\le i<j\le n} Pr(\mathcal{E}_i \cap \mathcal{E}_j) + \sum_{1\le i<j<k\le n} Pr(\mathcal{E}_i \cap \mathcal{E}_j \cap \mathcal{E}_k) - \cdots + (-1)^{n+1} Pr\Big(\bigcap_{i=1}^{n} \mathcal{E}_i\Big).$$

Proof. We prove the **inclusion-exclusion principle** using mathematical induction.

Base Case: For $n = 2$, consider two events $\mathcal{E}_1$ and $\mathcal{E}_2$. Note that:

1. $\Pr(\mathcal{E}_1) = \Pr\Big(\mathcal{E}_1 \setminus (\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Pr(\mathcal{E}_1 \cap \mathcal{E}_2)$, and
2. $\Pr(\mathcal{E}_2) = \Pr\Big(\mathcal{E}_2 \setminus (\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Pr(\mathcal{E}_1 \cap \mathcal{E}_2)$.

Thus,

$$\begin{aligned}
\Pr(\mathcal{E}_1 \cup \mathcal{E}_2) &= \Pr\Big(\mathcal{E}_1 \setminus (\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Pr\Big(\mathcal{E}_2 \setminus (\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Pr(\mathcal{E}_1 \cap \mathcal{E}_2) \\
&= \Big(\Pr(\mathcal{E}_1) - \Pr(\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Big(\Pr(\mathcal{E}_2) - \Pr(\mathcal{E}_1 \cap \mathcal{E}_2)\Big) + \Pr(\mathcal{E}_1 \cap \mathcal{E}_2) \\
&= \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) - \Pr(\mathcal{E}_1 \cap \mathcal{E}_2).
\end{aligned}$$

This concludes the proof for the base case.

Induction Step: Assume the identity holds for all $n = k$. We wish to show it holds for $n = k + 1$. Consider

$$\Pr\Big(\bigcup_{i=1}^{k+1} \mathcal{E}_i\Big) = \Pr\Big(\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) \cup \mathcal{E}_{k+1}\Big).$$

By the base case, we have

$$\begin{aligned}
\Pr\Big(\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) \cup \mathcal{E}_{k+1}\Big) &= \Pr\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) + \Pr(\mathcal{E}_{k+1}) - \Pr\Big(\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) \cap \mathcal{E}_{k+1}\Big) \\
&= \Pr\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) + \Pr(\mathcal{E}_{k+1}) - \Pr\Big((\mathcal{E}_1 \cap \mathcal{E}_{k+1}) \cup \cdots \cup (\mathcal{E}_k \cap \mathcal{E}_{k+1})\Big).
\end{aligned}$$

The final equality is due to the **distributive law** in set theory. Note that both $\Pr\Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big)$ and $\Pr\Big((\mathcal{E}_1 \cap \mathcal{E}_{k+1}) \cup \cdots \cup (\mathcal{E}_k \cap \mathcal{E}_{k+1})\Big)$ can be expressed using the induction hypothesis. Substituting these expressed terms back into the equation above completes the induction.

For clarity, the expression for the last term is given by:

$$\begin{aligned}
\Pr\Big((\mathcal{E}_1 \cap \mathcal{E}_{k+1}) \cup \cdots \cup (\mathcal{E}_k \cap \mathcal{E}_{k+1})\Big) &= \sum_{i=1}^{k} \Pr(\mathcal{E}_i \cap \mathcal{E}_{k+1}) - \sum_{1\le i<j\le k} \Pr(\mathcal{E}_i \cap \mathcal{E}_j \cap \mathcal{E}_{k+1}) \\
&\quad + \cdots + (-1)^{k+1} \Pr\Big(\bigcap_{i=1}^{k+1} \mathcal{E}_i\Big).
\end{aligned}$$

This completes the proof. □

Definition 3 (Independence)**.** *Two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are* ***independent*** *if and only if*

$$Pr(\mathcal{E}_1 \cap \mathcal{E}_2) = Pr(\mathcal{E}_1) \cdot Pr(\mathcal{E}_2).$$

More generally, events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ are ***mutually independent*** *if and only if, for any subset $I \subseteq \{1, 2, \ldots, k\}$,*

$$Pr\Big(\bigcap_{i \in I} \mathcal{E}_i\Big) = \prod_{i \in I} Pr(\mathcal{E}_i).$$

A simple computation shows that the number of unique, unordered subsets (i.e., the number of conditions required to establish mutual independence) for a collection of n events is given by:

$$\binom{n}{2} + \binom{n}{3} + \cdots + \binom{n}{n-1} + \binom{n}{n}.$$

For example, if we have a collection of three events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$, independence amounts to satisfying these four conditions:

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2) = \Pr(\mathcal{E}_1) \cdot \Pr(\mathcal{E}_2),$$

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_3) = \Pr(\mathcal{E}_1) \cdot \Pr(\mathcal{E}_3),$$

$$\Pr(\mathcal{E}_2 \cap \mathcal{E}_3) = \Pr(\mathcal{E}_2) \cdot \Pr(\mathcal{E}_3), \text{ and}$$

$$\Pr(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) = \Pr(\mathcal{E}_1) \cdot \Pr(\mathcal{E}_2) \cdot \Pr(\mathcal{E}_3).$$

1.2 Conditional Probability

Definition 4 (Conditional Probability)**.** *For any two events $\mathcal{E}_1$ and $\mathcal{E}_2$, the* ***conditional probability*** *of $\mathcal{E}_1$ given $\mathcal{E}_2$ is denoted by $Pr(\mathcal{E}_1|\mathcal{E}_2)$ (i.e., the probability of $\mathcal{E}_1$ knowing that $\mathcal{E}_2$ has already occurred). This can be computed by the following definition:*

$$Pr(\mathcal{E}_1|\mathcal{E}_2) = \frac{Pr(\mathcal{E}_1 \cap \mathcal{E}_2)}{Pr(\mathcal{E}_2)}.$$

The conditional probability is well defined if and only if $Pr(\mathcal{E}_2) > 0$.

Lemma 2. *If two events $\mathcal{E}_1$ and $\mathcal{E}_2$ are independent, then*

$$Pr(\mathcal{E}_1|\mathcal{E}_2) = Pr(\mathcal{E}_1).$$

Proof. This result follows immediately from Definitions **??** and **??**. □

Lemma 3 (Multiplication Rule)**.** *For a collection of events $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$, the* ***multiplication rule*** *is described by*

$$Pr\Big(\bigcap_{i=1}^{k} \mathcal{E}_i\Big) = Pr(\mathcal{E}_1) \cdot Pr(\mathcal{E}_2|\mathcal{E}_1) \cdot Pr(\mathcal{E}_3|\mathcal{E}_1 \cap \mathcal{E}_2) \dots Pr\Big(\mathcal{E}_k \Big| \bigcap_{i=1}^{k-1} \mathcal{E}_i\Big).$$

Proof. Observe that

$$\Pr\Big(\bigcap_{i=1}^{k} \mathcal{E}_i\Big) = \Pr(\mathcal{E}_1)\frac{\Pr(\mathcal{E}_1 \cap \mathcal{E}_2)}{\Pr(\mathcal{E}_1)} \frac{\Pr(\mathcal{E}_3 \cap (\mathcal{E}_1 \cap \mathcal{E}_2))}{\Pr(\mathcal{E}_1 \cap \mathcal{E}_2)} \dots \frac{\Pr\Big(\mathcal{E}_k \cap \Big(\bigcap_{i=1}^{k-1} \mathcal{E}_i\Big)\Big)}{\Pr\Big(\bigcap_{i=1}^{k-1} \mathcal{E}_i\Big)} \tag{1}$$

Then applying the definition of conditional probability yields the result. □

This rule is analogous to multiplying along a branch of a **tree-based sequence**.

Lemma 4 (Law of Total Probability)**.** *Let $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$ be* ***pairwise mutually disjoint*** *events in the sample space Ω, and let $\bigcup_{i=1}^{k} \mathcal{E}_i = \Omega$. Then, for some event $F \subseteq \Omega$, its probability can be computed by*

$$Pr(F) = \sum_{i=1}^{k} Pr(F \cap \mathcal{E}_i)$$

$$= \sum_{i=1}^{k} Pr(F|\mathcal{E}_i) Pr(\mathcal{E}_i).$$

Proof. Since $\bigcup_{i=1}^{k} \mathcal{E}_i = \Omega$, for any event $F \subseteq \Omega$ we have

$$F = F \cap \Omega = F \cap \Big(\bigcup_{i=1}^{k} \mathcal{E}_i\Big) \quad \Rightarrow \quad F = \bigcup_{i=1}^{k} (F \cap \mathcal{E}_i)\,.$$

Moreover, the events $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$ are **pairwise mutually disjoint** (i.e., $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ for all $i \neq j$), so the collection $F \cap \mathcal{E}_1, F \cap \mathcal{E}_2, \dots, F \cap \mathcal{E}_k$ is also pairwise mutually disjoint. Thus, by the axioms of probability, the first equality follows. The second equality follows from the definition of conditional probability. □

Theorem 1 (Bayes' Theorem)**.** *Let $\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k$ be* ***pairwise mutually disjoint*** *events in the sample space Ω, and let $\bigcup_{i=1}^{k} \mathcal{E}_i = \Omega$. Then for any event $F \subseteq \Omega$ and any $\mathcal{E}_j \in \{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_k\}$,*

$$Pr(\mathcal{E}_j|F) = \frac{Pr(F|\mathcal{E}_j)Pr(\mathcal{E}_j)}{Pr(F)}$$

$$= \frac{Pr(F|\mathcal{E}_j)Pr(\mathcal{E}_j)}{\sum_{i=1}^{k} Pr(F|\mathcal{E}_i)Pr(\mathcal{E}_i)}.$$

Additionally, $Pr(\mathcal{E}_j)$ and $Pr(\mathcal{E}_j|F)$ are known as the ***prior probability*** *and* ***posterior probability****, respectively.*

Proof. Note that

$$\Pr(\mathcal{E}_j \cap F) = \Pr(F \cap \mathcal{E}_j) = \Pr(F|\mathcal{E}_j)\Pr(\mathcal{E}_j)\,,$$

and then the result follows from the law of total probability. □

2 Discrete Random Variables and Expectation

2.1 Discrete Random Variables

Definition 5 (Random Variable). *A random variable X on a sample space Ω is a real-valued function on Ω; that is, $X : \Omega \longrightarrow \mathbb{R}$. A **discrete random variable** is one that takes on only a finite or countably infinite number of values.*

Definition 6 (Probability of a Discrete Random Variable). *Let X be a discrete random variable and $\alpha \in \mathbb{R}$. Then the probability of the event $X = \alpha$ is*

$$Pr(X = \alpha) = \sum_{\omega \in \Omega \colon X(\omega) = \alpha} Pr(\omega).$$

Lemma 5 (Probability of a Constant). *Suppose we have a constant $a \in \mathbb{R}$ considered as a discrete random variable (i.e., the function $a : \Omega \longrightarrow a$). Then*

$$Pr(a) = 1.$$

Proof. We have

$$\Pr(a) = \sum_{\omega \in \Omega \colon a(\omega) = a} \Pr(\omega) = \sum_{\omega \in \Omega} \Pr(\omega) = \Pr(\Omega) = 1.$$

□

The notion of independence and conditional probability for (discrete) random variables is similar to Definitions **??** and **??**.

Definition 7. *A collection of random variables $X_1, X_2, \ldots, X_k$ is said to be mutually independent if, for any subset $I \subseteq \{1, 2, \ldots, k\}$ and any values $\{x_i\}_{i \in I}$,*

$$Pr\Big(\bigcap_{i \in I} \{X_i = x_i\}\Big) = \prod_{i \in I} Pr(X_i = x_i).$$

Therefore, two random variables X and Y are independent if and only if

$$Pr\Big(\{X = x\} \cap \{Y = y\}\Big) = Pr(X = x) Pr(Y = y)$$

for all values x and y. Conventionally, the joint probability $Pr\Big(\{X = x\} \cap \{Y = y\}\Big)$ is written as $Pr(X = x, Y = y)$.

Definition 8. *The conditional probability of two random variables X and Y is defined as*

$$p_{X|Y}(x|y) \overset{def}{=} Pr(X = x | Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

for all values x and y, provided that $Pr(Y = y) > 0$.

Lemma 6 (Marginal Distribution). *Given two discrete random variables X and Y, then*

$$Pr(X = x) = \sum_{y} Pr(X = x, Y = y).$$

Proof. The sample space Ω for the joint probability distribution $\Pr(X = x, Y = y)$ consists of all possible ordered pairs (x, y). Hence,

$$\sum_{x,y} \Pr(X = x, Y = y) = 1 \, .$$

For instance, if we sum over the y-coordinate first and then over x, we have

$$\sum_{x} \Big(\sum_{y} \Pr(X = x, Y = y) \Big) = 1 \, .$$

We also know that

$$\sum_{x} \Pr(X = x) = 1 \, .$$

Thus,

$$\sum_{x} \Big(\sum_{y} \Pr(X = x, Y = y) \Big) = \sum_{x} \Pr(X = x) \quad \Rightarrow \quad \Pr(X = x) = \sum_{y} \Pr(X = x, Y = y) \, .$$

Similarly, summing over x first yields

$$\Pr(Y = y) = \sum_{x} \Pr(X = x, Y = y) \, .$$

An alternative proof considers the collection of events $\{(X = x, Y = y_1), (X = x, Y = y_2), \dots\}$. Since these events are **pairwise mutually disjoint**, the law of total probability gives

$$\Pr(X = x) = \sum_{y} \Pr(X = x, Y = y).$$

□

2.2 Expectation

Definition 9 (Expectation)**.** *Given a discrete random variable X, its* ***expectation****, denoted by $E[X]$, is given by*

$$E[X] = \sum_{i} i \, Pr(X = i) \, ,$$

where the summation is taken over all values in the range of X. The expectation is a ***weighted sum*** *(with weight $Pr(X = i)$) and represents the average or mean value. Furthermore, the expectation is said to be* ***finite*** *if*

$$\sum_{i} |i| \, Pr(X = i) < \infty \, .$$

Lemma 7 (Expectation of a Constant)**.** *Suppose we have a constant $a \in \mathbb{R}$ as a discrete random variable; then*

$$E[a] = a.$$

Proof. By the definition of expectation and the probability of a constant,

$$E[a] = \sum_{a} a \Pr(a) = a \Pr(a) = a.$$

Note that the second equality reflects that, since the range of the function a is $\{a\}$, there is only one summand. □

Lemma 8. *Given two **independent** discrete random variables X and Y, we have*

$$E[XY] = E[X]E[Y].$$

Proof. By definition of expectation,

$$\begin{aligned} E[XY] &= \sum_{x,y} (xy) \Pr(X = x, Y = y) \\ &= \sum_{x,y} (xy) \Pr(X = x) \Pr(Y = y) \quad \text{(by independence)} \\ &= \Big(\sum_{x} x \Pr(X = x)\Big)\Big(\sum_{y} y \Pr(Y = y)\Big) \\ &= E[X]E[Y]. \end{aligned}$$

□

If X and Y are **not independent**, one must compute the joint probability of the new discrete variable $Z = XY$, whose sample space consists of ordered pairs.

Theorem 2 (Linearity of Expectations)**.** *For any finite collection of discrete random variables $X_1, X_2, \ldots, X_k$ with **finite expectations**,*

$$E\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} E[X_i].$$

Proof. We begin with the base case by considering two random variables X_1 and X_2. By Definition **??**,

$$E[X_1 + X_2] = \sum_{x_1} \sum_{x_2} \Big[(x_1 + x_2) \Pr(X_1 = x_1, X_2 = x_2)\Big].$$

Then, we can write

$$\begin{aligned} E[X_1 + X_2] &= \sum_{x_1} \sum_{x_2} \Big[x_1 \Pr(X_1 = x_1, X_2 = x_2)\Big] + \sum_{x_1} \sum_{x_2} \Big[x_2 \Pr(X_1 = x_1, X_2 = x_2)\Big] \\ &= \sum_{x_1} x_1 \Big(\sum_{x_2} \Pr(X_1 = x_1, X_2 = x_2)\Big) + \sum_{x_2} x_2 \Big(\sum_{x_1} \Pr(X_1 = x_1, X_2 = x_2)\Big) \\ &= \sum_{x_1} x_1 \Pr(X_1 = x_1) + \sum_{x_2} x_2 \Pr(X_2 = x_2) \quad \text{(by Lemma ??)} \\ &= E[X_1] + E[X_2]. \end{aligned}$$

Thus, the base case is proved. Now, assume the theorem holds for $n = k$ random variables. For $n = k + 1$, let $Z = \sum_{i=1}^{k} X_i$. Then

$$E\Big[\sum_{i=1}^{k+1} X_i\Big] = E[Z + X_{k+1}] = E[Z] + E[X_{k+1}]\,,$$

which completes the induction. □

It is worth mentioning that linearity of expectations holds even if the discrete random variables are **not independent**. For example, if $Y = X + X^2$, then

$$E[Y] = E[X + X^2] = E[X] + E[X^2]\,.$$

Lemma 9. *For any constants a and b, and any random variable X,*

$$E[aX + b] = aE[X] + b.$$

Proof. By the linearity of expectation and the expectation of a constant,

$$\begin{aligned} E[aX + b] &= E[aX] + E[b] \\ &= E[aX] + b\,. \end{aligned}$$

Noting that a is a constant, we have

$$E[aX] = E[a]E[X] = aE[X].$$

Thus, the result follows. □

2.3 Conditional Expectation

Definition 10 (Conditional Expectation)**.** *We define the conditional expectation of a random variable X given a random variable Y as*

$$E[X|Y = y] = \sum_{x} x\, Pr(X = x|Y = y)\,,$$

where the summation is over all values of x in the range of X. Similarly, the conditional expectation of X given an event $\mathcal{E}$ is defined as

$$E[X|\mathcal{E}] = \sum_{x} x\, Pr(X = x|\mathcal{E}).$$

Lemma 10. *For any random variables X and Y,*

$$E[X] = \sum_{y} Pr(Y = y)\, E[X|Y = y],$$

where the summation is taken over all values in the range of Y.

Proof. By Definition **??**,

$$\begin{aligned}
\sum_y \Pr(Y=y)\, E[X|Y=y] &= \sum_y \Pr(Y=y) \sum_x x \Pr(X=x|Y=y) \\
&= \sum_x \sum_y x \Pr(X=x|Y=y) \Pr(Y=y) \\
&= \sum_x \sum_y x \Pr(X=x, Y=y) \quad \text{(by Definition ??)} \\
&= \sum_x x \Pr(X=x) \quad \text{(by Lemma ??)} \\
&= E[X].
\end{aligned}$$

□

Lemma 11. *For any collection of discrete random variables $X_1, X_2, \ldots, X_n$ with* ***finite expectations*** *and for any random variable Y,*

$$E\Big[\Big(\sum_{i=1}^{n} X_i\Big)\Big| Y=y\Big] = \sum_{i=1}^{n} E[X_i|Y=y].$$

Proof. The proof is analogous to the proof for the linearity of expectation in Theorem **??**. For the base case with two random variables X_1 and X_2, by Definition **??** we have

$$E[X_1+X_2|Y=y] = \sum_{x_1}\sum_{x_2}\Big[(x_1+x_2)\Pr(X_1=x_1, X_2=x_2|Y=y)\Big],$$

and by the same technique as before one may show that

$$E[X_1+X_2|Y=y] = E[X_1|Y=y] + E[X_2|Y=y].$$

□

Definition 11 (Random Conditional Expectation Variable)**.** *The expression $E[X|Y]$ is defined as the random variable $f(Y)$ that takes on the value $E[X|Y=y]$ when $Y=y$.*

Lemma 12 (Law of Total Expectation)**.**

$$E[X] = E\Big[E[X|Y]\Big].$$

Proof. By Definition **??**, we have $E[X|Y] = f(Y)$, where $f(Y)$ takes on the value $E[X|Y=y]$ when $Y=y$. Hence, by the definition of expectation,

$$\begin{aligned}
E\Big[E[X|Y]\Big] &= \sum_y E[X|Y=y]\Pr(Y=y) \\
&= E[X],
\end{aligned}$$

where the final equality follows from Lemma **??**. □

Theorem 3 (Special Case of Law of Total Expectation)**.** *Let $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_k$ be a **partition** of a sample space. For a random variable X,*

$$E[X] = \sum_i E[X|\mathcal{E}_i]\, Pr(\mathcal{E}_i).$$

Proof.

$$\begin{aligned}
\sum_i E[X|\mathcal{E}_i]\,\Pr(\mathcal{E}_i) &= \sum_i \sum_x x \Pr(X = x|\mathcal{E}_i)\Pr(\mathcal{E}_i) \\
&= \sum_i \sum_x x \Pr(X = x \cap \mathcal{E}_i) && \text{(by Definition ??)} \\
&= \sum_x x \sum_i \Pr(X = x \cap \mathcal{E}_i) && \text{(by factorization)} \\
&= \sum_x x \Pr(X = x) && \text{(by Lemma ??)} \\
&= E[X].
\end{aligned}$$

□

The law of total expectation can be interpreted as: *the unconditional average can be obtained by averaging the conditional averages.* For example, consider the following real-world analogy: *suppose we have two groups A and B with a and b members respectively. The average height in A is h_1 and in B is h_2. If we mix A and B together, the overall average height is obtained as a weighted sum, namely,*

$$\text{average height (i.e., unconditional average)} = h_1 \cdot \frac{a}{a+b} + h_2 \cdot \frac{b}{a+b}\,,$$

where h_1 and h_2 are the conditional averages.

3 Moments, Variance, and Covariance

Definition 12 (Moment)**.** *The kth moment of a random variable X is defined as*

$$E[X^k]\,.$$

Definition 13 (Variance and Standard Deviation)**.** *The **variance** of a random variable X is defined as*

$$\begin{aligned}
Var[X] &= E\Big[(X - E[X])^2\Big] && \textit{(by definition)} \\
&= E\Big[X^2 - 2X\,E[X] + \big(E[X]\big)^2\Big] \\
&= E[X^2] - 2\,E\Big[X\,E[X]\Big] + \big(E[X]\big)^2 && \textit{(by linearity of expectation)} \\
&= E[X^2] - 2\,E[X]\,E[X] + \big(E[X]\big)^2 && \textit{(since } E[X] \textit{ is a constant)} \\
&= E[X^2] - \big(E[X]\big)^2.
\end{aligned}$$

Note that we used the linearity of expectation and the fact that $E[X]$ is a constant to derive the final equality for variance. Furthermore, the ***standard deviation*** *of X is defined as*

$$\sigma[X] = \sqrt{Var[X]}.$$

Definition 14 (Covariance)**.** *The* ***covariance*** *of two random variables X and Y is defined as*

$$Cov(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big].$$

Theorem 4. *For any two random variables X and Y,*

$$Var[X+Y] = Var[X] + Var[Y] + 2\,Cov(X,Y).$$

Proof.

$$\begin{aligned}
\text{Var}[X+Y] &= E\Big[(X+Y-E[X+Y])^2\Big] && \text{(by definition)}\\
&= E\Big[(X+Y-E[X]-E[Y])^2\Big] && \text{(by linearity)}\\
&= E\Big[(X-E[X])^2 + (Y-E[Y])^2 + 2\,(X-E[X])(Y-E[Y])\Big] &&\\
&= E\Big[(X-E[X])^2\Big] + E\Big[(Y-E[Y])^2\Big] + 2\,E\Big[(X-E[X])(Y-E[Y])\Big] && \text{(by linearity)}\\
&= \text{Var}[X] + \text{Var}[Y] + 2\,\text{Cov}(X,Y).
\end{aligned}$$

□

Lemma 13. *If X and Y are two* ***independent*** *random variables, then*

$$Cov(X,Y) = 0 \quad \Rightarrow \quad Var[X+Y] = Var[X] + Var[Y].$$

Proof.

$$\begin{aligned}
\text{Cov}(X,Y) &= E\Big[(X-E[X])(Y-E[Y])\Big] && \text{(by definition)}\\
&= E\Big[X-E[X]\Big]\,E\Big[Y-E[Y]\Big] && \text{(by independence)}\\
&= \Big(E[X]-E[X]\Big)\,\Big(E[Y]-E[Y]\Big) && \text{(by linearity)}\\
&= 0.
\end{aligned}$$

□

Theorem 5. *Let $X_1, X_2, \cdots, X_k$ be mutually independent random variables. Then,*

$$Var\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} Var[X_i].$$

Proof. We prove this by mathematical induction. The base case for two random variables X_1 and X_2 is established in Lemma **??**. Suppose the identity holds for $n = k$. For $n = k + 1$, let

$$Z = \sum_{i=1}^{k} X_i \, .$$

Then, by Lemma **??**,

$$\mathrm{Var}[Z + X_{k+1}] = \mathrm{Var}[Z] + \mathrm{Var}[X_{k+1}].$$

Thus, by the induction hypothesis, the result holds for $n = k + 1$, and the proof is complete. □

4 Useful Bounds in Probability

4.1 Analytical Inequalities

Lemma 14 (Exponential Bound)**.**

$$1 + x \leq e^x, \quad \textit{for all } x \in \mathbb{R}.$$

Proof. By the Taylor expansion of e^x, we have

$$e^x = \sum_{n \geq 0} \frac{x^n}{n!}.$$

Using this expansion and noting that $e^x > 0$, we obtain

$$e^x - (1 + x) = \left(\frac{x^2}{2!} + \frac{x^3}{3!} \right) + \left(\frac{x^4}{4!} + \frac{x^5}{5!} \right) + \cdots \geq 0.$$

□

Lemma 15 (Stirling's Approximation)**.** *For $n > 0$,*

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e} \right)^n .$$

Definition 15 (Convex Function)**.** *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be convex if for any $x_1, x_2 \in \mathbb{R}$ and $0 \leq \lambda \leq 1$,*

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

A visual interpretation is that the line segment connecting any two points on the graph of f lies above (or on) the graph. In particular, if f is twice differentiable, then f is convex if and only if $f''(x) \geq 0$.

Lemma 16 (Jensen's Inequality)**.** *If f is a convex function, then*

$$E\big[f(X)\big] \geq f\big(E[X]\big).$$

Proof. Assume that f has a Taylor expansion. Let $\mu = E[X]$. Then by Taylor's theorem, there exists a constant c (depending on x) such that

$$\begin{aligned} f(x) &= f(\mu) + f'(\mu)(x-\mu) + \frac{f''(c)(x-\mu)^2}{2} \\ &\geq f(\mu) + f'(\mu)(x-\mu) && \text{(since } f''(c) > 0 \text{ by convexity).} \end{aligned}$$

Taking expectations of both sides yields

$$\begin{aligned} E\big[f(X)\big] &\geq E\Big[f(\mu) + f'(\mu)(X-\mu)\Big] \\ &= E\big[f(\mu)\big] + f'(\mu)E[X-\mu] && \text{(by linearity of expectation)} \\ &= f\big(E[X]\big) + f'\left(E[X]\right) \cdot \left(E[X] - E[X]\right) \\ &= f\big(E[X]\big). \end{aligned}$$

□

4.2 Probability and Moment Bounds

Lemma 17 (Union Bound)**.** *For any finite or countably infinite sequence of events $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \ldots,$*

$$Pr\Big(\bigcup_{i\geq 1} \mathcal{E}_i\Big) \leq \sum_{i \geq 1} Pr(\mathcal{E}_i).$$

*This result is also known as **Boole's Inequality**.*

Proof. This is a direct consequence of the Inclusion-Exclusion Principle (see Lemma **??**). Equality holds when the events are pairwise mutually disjoint. □

Lemma 18 (Markov's Inequality)**.** *Let X be a random variable that takes on only **non-negative** values. Then, for any positive constant a,*

$$Pr(X \geq a) \leq \frac{E[X]}{a}.$$

Proof.

$$\begin{aligned} E[X] &= E[X \mid X < a] \Pr(X < a) + E[X \mid X \geq a] \Pr(X \geq a) && \text{(by the law of total expectation)} \\ &\geq 0 \cdot \Pr(X < a) + a \cdot \Pr(X \geq a). \end{aligned}$$

□

Lemma 19 (Chebyshev's Inequality)**.** *Let X be a random variable. For any positive constant a,*

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}.$$

Proof.

$$\begin{aligned}\Pr(|X - E[X]| \geq a) &= \Pr\Big((X - E[X])^2 \geq a^2\Big) \\ &\leq \frac{E\big[(X - E[X])^2\big]}{a^2} && \text{(by Markov's Inequality)} \\ &= \frac{\mathrm{Var}[X]}{a^2} && \text{(by definition of variance).}\end{aligned}$$

□

Lemma 20 (Chernoff Bound (Right Tail))**.** *Suppose X is a random variable and a is a real constant. Then for any $t > 0$,*

$$Pr(X \geq a) \leq \min_{t>0} \frac{E\big[e^{tX}\big]}{e^{ta}}.$$

Proof.

$$\begin{aligned}\Pr(X \geq a) &= \Pr(tX \geq ta) \\ &= \Pr\Big(e^{tX} \geq e^{ta}\Big) \\ &\leq \frac{E\big[e^{tX}\big]}{e^{ta}} && \text{(by Markov's Inequality, since } e^{tX} \text{ and } e^{ta} \text{ are always positive).}\end{aligned}$$

Thus, taking the minimum over $t > 0$ yields

$$\Pr(X \geq a) \leq \min_{t>0} \frac{E\big[e^{tX}\big]}{e^{ta}}.$$

□

Lemma 21 (Chernoff Bound (Left Tail))**.** *Suppose X is a random variable and a is a real constant. Then for any $t < 0$,*

$$Pr(X \leq a) \leq \min_{t<0} \frac{E\big[e^{tX}\big]}{e^{ta}}.$$

Proof.

$$\begin{aligned}\Pr(X \leq a) &= \Pr(tX \geq ta) && \text{(since } t < 0\text{)} \\ &= \Pr\Big(e^{tX} \geq e^{ta}\Big) \\ &\leq \frac{E\big[e^{tX}\big]}{e^{ta}} && \text{(by Markov's Inequality).}\end{aligned}$$

□

When considering deviation bounds, notice that Chernoff bounds provide much tighter estimates than Markov's and Chebyshev's inequalities because they utilize *all* moments of X (encoded in the moment generating function), a concept that will be discussed in the following section.

5 Moment Generating Function

Definition 16 (Moment Generating Function)**.** *The **moment generating function** (MGF) of a random variable X is defined as*

$$M_X(t) = E\Big[e^{tX}\Big].$$

Theorem 6. *Let X be a random variable with MGF $M_X(t)$, and assume that exchanging the expectation and differentiation is valid. Then for all $n > 1$, we have*

$$E[X^n] = M_X^{(n)}(0)\,,$$

where $M_X^{(n)}(0)$ denotes the nth derivative of $M_X(t)$ evaluated at $t = 0$.

Proof. Under the given assumption,

$$\begin{aligned} M_X^{(n)}(t) &= \frac{d^n}{dt^n} E\Big[e^{tX}\Big] \\ &= E\Big[\frac{d^n}{dt^n} e^{tX}\Big] \\ &= E\Big[X^n e^{tX}\Big]. \end{aligned}$$

Evaluating at $t = 0$ yields

$$E[X^n] = M_X^{(n)}(0)\,,$$

which completes the proof. □

Lemma 22. *If X and Y are two independent random variables, then*

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Proof.

$$\begin{aligned} M_{X+Y}(t) &= E\Big[e^{t(X+Y)}\Big] \\ &= E\Big[e^{tX} \cdot e^{tY}\Big] \\ &= E\Big[e^{tX}\Big] \cdot E\Big[e^{tY}\Big] && \text{(by independence)} \\ &= M_X(t) \cdot M_Y(t). \end{aligned}$$

□

6 Discrete Probability Distributions and Interpretations

In the previous sections, we introduced all the necessary mathematical tools required to study (discrete) probability. Without providing interpretations, we presented the axioms of probability and various definitions such as those for random variables, conditional probabilities, independence, expectations, variance, covariance, and moments. We also proved many important results—stated as theorems and lemmas—using nothing but set theory and elementary algebra. In this section, we will provide interpretations to build strong intuition when analyzing probability with the tools described earlier.

When we conduct an experiment, the outcome belongs to some sample space Ω, and we are interested in the occurrence of certain event(s) $\mathcal{E}$. Conventionally, the probability function is defined as the ratio of the number of elements in the event $\mathcal{E}$ to the total number of elements in the sample space Ω; that is,

$$\Pr(\mathcal{E}) = \frac{|\mathcal{E}|}{|\Omega|} .$$

However, it is important to note that this definition assumes all single-element events (i.e., all outcomes) are equally likely. We then introduced random variables, which assign a numerical value (from a discrete set) to each outcome. In this case, the possible numerical values *need not* all have the same probability (e.g., more than one outcome can have the same value). In fact, the case where all numerical values are equally likely corresponds to the so-called *discrete uniform distribution.* When conducting an experiment where numerical values are observed, we are essentially drawing random samples from well-defined *distributions* in which some values may occur more frequently than others. Formally, we write this as

$$X \sim F(\gamma) ,$$

which is read as "the random variable X has distribution $F(\gamma)$," where γ denotes some parameter.

For example, consider the experiment of flipping a coin. If we assign the numerical value $X = 1$ for heads and $X = 0$ for tails, and if heads occurs with probability p (so that, by the axiom of probability, $\Pr(X = 0) = 1-p$), then this is known as the *Bernoulli experiment*, and we say that X follows a *Bernoulli distribution* (i.e., $X \sim \text{Bernoulli}(p)$).

Expectations, variance, covariance, and moments are properties of the distributions of numerical values. The expectation of a random variable represents its long-run average when observed over many trials. Variance quantifies how far the values can deviate from the mean. Understanding these concepts through interpretations gives concrete meaning to Definitions **??** and **??**. Covariance measures the linear relationship between two random variables (for example, how X varies with Y but they *need not* have causal influence with one another), and moments describe the shape of the distribution's graph. These are all valuable tools for analyzing the values drawn from a distribution, revealing order within randomness in the long run—an essential reason why we study statistics!

We will now examine some well-known discrete probability distributions (often described as **probability mass functions** in the literature on discrete random variables) and provide their respective expectations and variances.

6.1 General Probability Mass Function

In general, we can design any distribution we deem appropriate for drawing samples, provided it adheres to the axioms of probability. Suppose the outcomes come from a discrete set X. For each $x_i \in X$, we define the probability mass function (PMF) of X, denoted by $P(X)$, as follows:

$$P(X) = \begin{cases} p_1, & \text{if } X = x_1, \\ \vdots \\ p_i, & \text{if } X = x_i, \end{cases}$$

where $p_i \overset{\text{def}}{=} \Pr(X = x_i)$ and, by the axioms of probability (see Definition **??**), we have $\sum_i p_i = 1$. From this definition, the expectation is given by

$$E[X] = \sum_i x_i p_i,$$

and variance and higher moments follow accordingly.

For example, consider an unfair 3-sided die with the following PMF:

$$P(X) = \begin{cases} \frac{2}{3}, & \text{if } X = 1, \\ \frac{1}{6}, & \text{if } X = 2, \\ \frac{1}{6}, & \text{if } X = 3. \end{cases}$$

Clearly, the expected value is

$$E[X] = 1 \cdot \frac{2}{3} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} = \frac{3}{2}\,.$$

6.2 Discrete Uniform Random Variable

Definition 17 (Discrete Uniform Random Variable)**.** *The* discrete uniform distribution *$X \sim U(\{a, a+1, \ldots, b\})$ has the following PMF:*

$$P(X) = \begin{cases} \frac{1}{b-a+1}, & \textit{if } X = a, a+1, a+2, \ldots, b, \\ 0, & \textit{otherwise,} \end{cases}$$

where a and b are integers with $a < b$.

Theorem 7. *The expectation of a discrete uniform random variable X is*

$$E[X] = \frac{a+b}{2}\,,$$

and its variance is

$$\mathit{Var}[X] = \frac{(b-a+1)^2 - 1}{12}\,.$$

Proof. We can find $E[X]$ by using the sum of an arithmetic progression:

$$\begin{aligned} E[X] &= \sum_x x\,P(x) \\ &= \frac{1}{b-a+1}\Big(a + (a+1) + \cdots + [a + (b-a)]\Big) \\ &= \frac{1}{b-a+1}\Big[a(b-a+1) + S_{b-a}\Big] \\ &= \frac{1}{b-a+1}\left[a(b-a+1) + \frac{(b-a)(b-a+1)}{2}\right] \\ &= \frac{a(b-a+1) + \frac{(b-a)(b-a+1)}{2}}{b-a+1} \\ &= a + \frac{b-a}{2} \\ &= \frac{2a+b-a}{2} \\ &= \frac{a+b}{2}. \end{aligned}$$

To compute the variance, we first need to determine $E[X^2]$. One can compute $E[X^2]$ directly by summing the squares of the outcomes. A routine (though somewhat lengthy) calculation shows that

$$E[X^2] = a^2 + a(b-a) + \frac{(b-a)(2(b-a)+1)}{6}\,.$$

Then, by the definition of variance (see Definition **??**),

$$\begin{aligned}
\mathrm{Var}[X] &= E[X^2] - \Big(E[X]\Big)^2 \\
&= \left(a^2 + a(b-a) + \frac{(b-a)(2(b-a)+1)}{6}\right) - \Big(a + \frac{b-a}{2}\Big)^2 \\
&= \frac{(b-a+1)^2-1}{12}.
\end{aligned}$$

This completes the proof. □

6.3 Bernoulli Random Variable

Definition 18 (Bernoulli Random Variable)**.** *A* Bernoulli *(or* indicator*) random variable $X \sim I(p)$ takes on values $\{0,1\}$ and has the following PMF:*

$$P(X) = \begin{cases} p, & \textit{if } X = 1, \\ 1-p, & \textit{if } X = 0. \end{cases}$$

This distribution is also known as a Bernoulli trial.

Theorem 8. *The expectation of a Bernoulli random variable X is*

$$E[X] = p,$$

and its variance is

$$Var[X] = p(1-p).$$

Proof.

$$\begin{aligned}
E[X] &= 0 \cdot (1-p) + 1 \cdot p \\
&= p.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[X] &= \Big[0^2 \cdot (1-p) + 1^2 \cdot p\Big] - p^2 \\
&= \Big[p\Big] - p^2 \\
&= p(1-p).
\end{aligned}$$

□

6.4 Binomial Random Variable

Definition 19 (Binomial Random Variable)**.** *A* binomial *random variable $X \sim B(n,p)$ has the following distribution on $j = 0, 1, \ldots, n$:*

$$Pr(X = j) = \binom{n}{j} p^j (1-p)^{n-j}.$$

This distribution represents the ***number of successes*** *in a sequence of n independent* Bernoulli trials *that are identically distributed (i.e., each trial has the same parameter p); hence, a Bernoulli random variable is a special case of a binomial random variable when $n = 1$.*

Theorem 9. *The expectation of a binomial random variable X is*

$$E[X] = np,$$

and its variance is

$$Var[X] = np(1-p).$$

Proof. Let $X_1, X_2, \ldots, X_n \sim I(p)$ be independent Bernoulli random variables. Note that

$$X = \sum_{i=1}^{n} X_i,$$

which counts the number of successes. Then,

$$\begin{aligned} E[X] &= E\Big[\sum_{i=1}^{n} X_i\Big] \\ &= \sum_{i=1}^{n} E[X_i] && \text{(by linearity)} \\ &= np. \end{aligned}$$

Similarly, due to independence,

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] \\ &= \sum_{i=1}^{n} \mathrm{Var}[X_i] \\ &= np(1-p). \end{aligned}$$

□

6.5 Poisson Binomial Random Variable

Definition 20 (Poisson Binomial Random Variable)**.** *Let $X_1, X_2, \ldots, X_n$ be a sequence of n independent* Bernoulli trials *that are* ***not*** *identically distributed (i.e., $Pr(X_i = 1) = p_i$). Therefore, a* Poisson binomial random variable *$X = \sum_{i=1}^{n} X_i$ has the following distribution:*

$$Pr(X = k) = \sum_{S \in F_k} \prod_{i \in S} p_i \prod_{j \in S^c} (1 - p_j),$$

where F_k is the set of all subsets of $\{1, 2, 3, \ldots, n\}$ of size k (representing the k successes). Consequently, its mean is

$$\mu = E[X] = \sum_{i=1}^{n} p_i,$$

and its variance is

$$Var[X] = \sum_{i=1}^{n} p_i(1 - p_i).$$

Naturally, a Poisson binomial random variable generalizes a binomial random variable, which is the special case where $p_1 = p_2 = \cdots = p_n$. The following theorem pertains to its MGF.

Theorem 10. *Let X be a Poisson binomial random variable. Then, for any real t,*

$$M_X(t) \leq e^{\mu(e^t - 1)}.$$

Proof.

$$\begin{aligned} M_X(t) &= E\Big[e^{tX}\Big] \\ &= E\Big[e^{t\sum_{i=1}^{n} X_i}\Big] \\ &= \prod_{i=1}^{n} E\Big[e^{tX_i}\Big] \\ &= \prod_{i=1}^{n} \Big[(1 - p_i) + p_i e^t\Big] \\ &= \prod_{i=1}^{n} \Big[1 + (e^t - 1)p_i\Big] \\ &\leq \prod_{i=1}^{n} e^{p_i(e^t - 1)} && \text{(by Lemma ??)} \\ &= e^{(e^t - 1)\sum_{i=1}^{n} p_i} \\ &= e^{\mu(e^t - 1)}. \end{aligned}$$

□

6.6 Poisson Random Variable

Definition 21 (Poisson Random Variable)**.** *A* Poisson *random variable $X \sim Pois(\lambda)$ has the following PMF:*

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where the parameter λ is equal to $E[X]$. This distribution models the number of events occurring in a fixed time interval when the events occur at a known constant mean rate.

The motivation for this distribution will be explained in detail in the Balls-and-Bins Model section.

Theorem 11. *Let $X \sim$ Pois(λ); then, Var$[X] = \lambda$.*

Proof. We know $E[X] = \lambda$. Consider the expectation

$$\begin{aligned} E[X(X-1)] &= \sum_{k\geq 0} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda^2 e^{-\lambda} \sum_{k\geq 2} \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 e^{-\lambda} \cdot e^{\lambda} && \text{(by Taylor's expansion)} \\ &= \lambda^2 . \end{aligned}$$

Since

$$E[X(X-1)] = E[X^2] - E[X],$$

we obtain

$$E[X^2] = \lambda^2 + \lambda .$$

Thus, by Definition **??**,

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

□

6.7 Geometric Random Variable

Definition 22 (Geometric Random Variable)**.** *Suppose $X_1, X_2, \ldots$ is a sequence of independent and identically distributed Bernoulli trials with parameter p. The geometric random variable $X \sim$ Geom(p) is defined as the trial number on which the first success occurs; that is, all preceding trials result in failure. Its PMF is given by*

$$\Pr(X = k) = \Pr(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0,\ X_k = 1) = (1-p)^{k-1}p .$$

Theorem 12. *Let $X \sim$ Geom(p). Then,*

$$E[X] = \frac{1}{p}$$

and

$$\mathit{Var}[X] = \frac{1-p}{p^2} .$$

Proof. Using the fact from the geometric series that

$$\sum_{k\geq 1} k\, x^{k-1} = \frac{1}{(1-x)^2} ,$$

we have

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} k\, p\, (1-p)^{k-1} \\ &= p \cdot \frac{1}{\big(1-(1-p)\big)^2} \\ &= \frac{1}{p} . \end{aligned}$$

To find the variance, we first compute the second moment $E[X^2]$. Using the fact that

$$\sum_{k\geq 1} k(k-1)\, x^{k-2} = \frac{2}{(1-x)^3}\,,$$

we obtain

$$\begin{aligned}
E[X^2] &= \sum_{k=1}^{\infty} k^2\, p\,(1-p)^{k-1} \\
&= (1-p)p\sum_{k=1}^{\infty} k(k-1)\,(1-p)^{k-2} + (1-p)p\sum_{k=1}^{\infty} k\,(1-p)^{k-2} \\
&\quad \text{(splitting } k^2 \text{ as } k(k-1)+k\text{; note that the second summation equals } E[X] = \frac{1}{p}\text{)} \\
&= (1-p)p\cdot\frac{2}{\big(1-(1-p)\big)^3} + \frac{1-p}{p} \\
&= (1-p)p\cdot\frac{2}{p^3} + \frac{1-p}{p} \\
&= \frac{2(1-p)}{p^2} + \frac{1}{p} \\
&= \frac{2-p}{p^2}\,.
\end{aligned}$$

Thus, the variance is computed as

$$\mathrm{Var}[X] = E[X^2] - \big(E[X]\big)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}\,.$$

□

6.8 Multinomial Random Variable

Definition 23 (Multinomial Random Variable)**.** *Let n be the number of independent trials, and let $k > 0$ be a fixed, finite integer representing k mutually exclusive outcomes with corresponding event probabilities $p_1, \ldots, p_k$ such that $\sum_{i=1}^{k} p_i = 1$. The multinomial random variable $X \sim M_k(n; p_1, ..., p_k)$ has the following PMF:*

$$Pr(X_1 = x_1, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_1!\cdots x_k!}\cdot p_1^{x_1}\cdots p_k^{x_k}, & \text{if } \sum_{i=1}^{k} x_i = n, \\ 0, & \text{otherwise,} \end{cases}$$

for non-negative integers $x_1, \ldots, x_k$. Moreover, its mean *and* variance *are given by $E[X_i] = np_i$ and $\mathrm{Var}[X_i] = np_i(1-p_i)$, respectively.*

The multinomial random variable generalizes the binomial random variable (i.e., a binomial random variable is a 2-category multinomial random variable, where the categories are often denoted as $\{0, 1\}$—meaning $B(n, p) \sim M_2(n; p, 1-p)$).

To derive the PMF above, suppose the k categories are mutually exclusive, and outcomes across trials are independent. Assume that the vector $(x_1, \ldots, x_k)$ forms a **partition** of n. Then, the joint probability follows:

$$\begin{aligned}\Pr(X_1 = x_1, \ldots, X_k = x_k) &= \binom{n}{x_1} p_1^{x_1} \cdot \binom{n - x_1}{x_2} p_2^{x_2} \cdots \binom{n - \sum_{i=1}^{k-1} x_i}{x_k} p_k^{x_k} \\ &= \frac{n!}{x_1! \cdots x_k!} \cdot p_1^{x_1} \cdots p_k^{x_k} \left[\frac{1}{(n - x_1)!} \cdot \frac{(n - x_1)!}{(n - x_1 - x_2)!} \cdots \frac{(n - \sum_{i=1}^{k-1} x_i)!}{(n - \sum_{i=1}^{k} x_i)!} \right]\end{aligned}$$

Note that the term

$$\frac{1}{(n - x_1)!} \cdot \frac{(n - x_1)!}{(n - x_1 - x_2)!} \cdots \frac{(n - \sum_{i=1}^{k-1} x_i)!}{(n - \sum_{i=1}^{k} x_i)!}$$

simplifies to 1, which confirms the multinomial PMF.

6.9 Hypergeometric Random Variable

Definition 24 (Hypergeometric Random Variable)**.** *A hypergeometric random variable $X \sim H(N, K, n)$ models the number of successes drawn uniformly at random from a population without replacement, where there are only two categories. That is, we can think of each draw as resulting in a binary outcome $X_i \in \{0, 1\}$, and define the total number of observed successes as $X = \sum_{i=1}^{n} X_i$, where n is the number of draws. Note that the X_i are not independent due to **sampling without replacement**.*

*This differs from the binomial random variable, which is often interpreted (we will return to this interpretation later) as modeling **sampling with replacement**—that is, drawing from a population of mutually exclusive 2-category outcomes and counting the number of successes.*

The PMF of X is given by:

$$Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where

- *N is the population size,*
- *K is the number of success states in the population,*
- *n is the number of draws without replacement (with $n \leq N$), and*
- *k is the number of* observed *successes (with $k \leq n$).*

Moreover, we can compute the mean using the definition of expectation:

$$E[X] = \sum_{k=0}^{n} k \cdot \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

To simplify this, we use the identity $k\binom{K}{k} = K\binom{K-1}{k-1}$, along with the identity $\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1}$. Applying these to the expectation:

$$E[X] = \sum_{k=0}^{n} K \cdot \frac{\binom{K-1}{k-1}\binom{N-K}{n-k}}{\frac{N}{n}\binom{N-1}{n-1}}.$$

Factor out the constants:

$$E[X] = \frac{Kn}{N} \cdot \sum_{k=0}^{n} \frac{\binom{K-1}{k-1}\binom{N-K}{n-k}}{\binom{N-1}{n-1}}.$$

Now observe that this summation is the PMF of a hypergeometric distribution with parameters $(N-1, K-1, n-1)$, and hence the entire sum evaluates to:

$$\sum_{k=0}^{n} \frac{\binom{K-1}{k-1}\binom{N-K}{n-k}}{\binom{N-1}{n-1}} = 1.$$

Thus, the expected number of successes in a hypergeometric distribution is:

$$\boxed{E[X] = n \cdot \frac{K}{N}}.$$

This result makes intuitive sense: since we're sampling without replacement, the expected number of successes in n draws is proportional to the proportion of successes in the entire population.

In a similar fashion to how the multinomial random variable generalizes the binomial random variable, we can extend the idea of drawing samples without replacement to more than two mutually exclusive outcomes, each with their own number of elements in the population. This leads to the *multivariate hypergeometric random variable.*

Definition 25 (Multivariate Hypergeometric Random Variable)**.** *Suppose we have m mutually exclusive categories, where the number of elements in each category are $K_1, K_2, \ldots, K_m$, making the total population size $N = \sum_{i=1}^{m} K_i$. The multivariate hypergeometric random variable has the following PMF:*

$$Pr(X_1 = k_1, \ldots, X_m = k_m) = \frac{\prod_{i=1}^{m} \binom{K_i}{k_i}}{\binom{N}{n}}$$

where n is the number of draws without replacement, and naturally $n = \sum_{i=1}^{m} k_i$.

Now expounding on the interpretation of how binomial, multinomial, hypergeometric, and multivariate hypergeometric random variables model drawing samples *with or without replacement* from a given population of size N. This abstraction is commonly known as the **urn problem**.

For example, suppose we have two colored marbles in an urn: blue and red. There are m blue marbles and n red marbles, so the total population is $N = m + n$. We wish to draw a marble uniformly at random, observe its color, and return it to the urn (i.e., sampling with replacement). We can repeat this experiment as many times as we like (say, k times), and mathematically, the number of blue marbles drawn (treating blue as a "success") follows a binomial distribution: $B(k, p = \frac{m}{N})$. This gives us another way to interpret flipping a p-weighted coin k times. The only caveat in this interpretation is that $p \in \mathbb{Q}$, which need not hold in more general settings.

We can extend this urn-based abstraction to the other random variables mentioned above. For instance, if we have more than two colored marbles, the setting naturally generalizes to the multinomial case. Interestingly, the urn problem interpretation for the multinomial random variable (now with more than two categories) serves as a precursor to the **balls-and-bins** model—an abstraction that, as we've hinted before, helps illuminate the origins and applications of the Poisson distribution.

7 Balls-and-Bins Model

Definition 26 (Balls-and-Bins Model). *A* ***random process*** *known as the balls-and-bins model proceeds as follows: m balls are thrown uniformly and independently at random into n bins. This abstraction can be modeled using the multinomial random variable; that is, the balls-and-bins model with m balls and n bins follows the distribution $M_n(m; p_1 = \frac{1}{n}, \ldots, p_n = \frac{1}{n})$, since each bin is chosen uniformly at random. We can think of the bins as categories in the multinomial setting.*

Therefore, the probability of observing a specific configuration of balls across the bins, say $(x_1, \ldots, x_n)$ with $\sum_{i=1}^{n} x_i = m$, is given by:

$$Pr(X_1 = x_1, \ldots, X_n = x_n) = \binom{m}{x_1; x_2; \ldots; x_n} \cdot \prod_{i=1}^{n} p_i^{x_i} = \frac{\binom{m}{x_1;x_2;\ldots;x_n}}{n^m},$$

where the multinomial coefficient is defined as $\binom{m}{x_1;x_2;\ldots;x_n} = \frac{m!}{x_1! x_2! \cdots x_n!}$.

There are many interesting questions that naturally arise from this random process. For example, one might ask: What is the expected maximum number of balls in any bin (i.e., the *maximum load*)? How many bins are expected to remain empty? Surprisingly, these questions have real-world implications, especially in the design and analysis of algorithms and data structures.

Before diving into such analyses, we first introduce more results on the Poisson random variable.

7.1 Poisson Random Variable and Its Connection to Balls-and-Bins

We begin by noting that the Poisson random variable arises as a *limit of the binomial distribution.*

Theorem 13. *Suppose $X_n \sim B(n, p)$ where p depends on n such that $\lim_{n\to\infty} np = \lambda$ for some constant λ independent of n. Then, for any fixed k, we have*

$$\lim_{n\to\infty} Pr(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

Proof.

$$\begin{aligned}
\Pr(X_n = k) &= \binom{n}{k} p^k (1-p)^{n-k} && \text{(by Definition \textbf{??})} \\
&\le \frac{n^k}{k!} p^k \cdot \frac{(1-p)^n}{(1-p)^k} \\
&\le \frac{(np)^k}{k!} \cdot \frac{e^{-pn}}{1-pk} && \text{(by Lemma \textbf{??} and \textbf{Bernoulli's Inequality})} \\
&= \frac{e^{-np}(np)^k}{k!} \cdot \frac{1}{1-pk}.
\end{aligned}$$

On the other hand:

$$\begin{aligned}\Pr(X_n = k) &\geq \frac{(n-k+1)^k}{k!} p^k (1-p)^n \\ &\geq \frac{((n-k+1)p)^k}{k!} \cdot e^{-pn}(1-p^2)^n \quad (\text{using } e^x(1-x^2) \leq 1+x \leq e^x \text{ for } |x| \leq 1) \\ &\geq \frac{e^{-np}((n-k+1)p)^k}{k!} \cdot (1-p^2)^n.\end{aligned}$$

Putting both bounds together:

$$\frac{e^{-np}(np-kp+p)^k}{k!}(1-p^2)^n \leq \Pr(X_n = k) \leq \frac{e^{-np}(np)^k}{k!} \cdot \frac{1}{1-pk}.$$

Now, since $\lambda = \lim_{n\to\infty} np$ is finite and independent of n, it follows that $p = \frac{\lambda}{n} \to 0$ as $n \to \infty$. Thus, by the squeeze theorem, we conclude:

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

□

From the theorem above, we see that the Poisson random variable is well-suited to model the number of *rare events*—situations where each event either occurs or does not, with a large number of independent trials n and small success probability p, such that the expected number of successes $\lambda = np$ remains finite. In such settings, rather than using a binomial distribution, which becomes cumbersome to compute and may overestimate the probability of rare events, we approximate it using a Poisson distribution with parameter $\lambda \in (0, \infty)$, often obtained statistically or from prior domain knowledge.

For example, suppose the rate of spelling errors in a book is known to be $\lambda' = \frac{1}{1000}$, meaning there is on average 1 error per 1000 words, and errors occur independently. This implies that each word has a small probability p of containing an error (in this case, we know p is small from the estimated probability $p = \lambda' = \frac{1}{1000}$), and the total number of errors in a book with $n = 100{,}000$ words is modeled by $B(n, p)$. Since n is large and p is small, this binomial distribution can be accurately approximated by a Poisson distribution.

To find the probability of observing exactly 3 spelling errors in such a book, we compute the expected number of errors:

$$\lambda = n \cdot \lambda' = 100{,}000 \cdot \frac{1}{1000} = 100.$$

Using the Poisson PMF:

$$\Pr(X = 3) = \frac{e^{-100} \cdot 100^3}{3!} = \frac{e^{-100} \cdot 1{,}000{,}000}{6}.$$

Although this value is extremely small (as expected for such a low count when the expected number is high), it illustrates the use of the Poisson distribution in modeling rare events across a large number of trials.

Now, how does this relate to the balls-and-bins model? Suppose we want to compute the number of balls in the ith bin (recall that we have m balls and n bins); clearly,

$X_i \sim B(m, \frac{1}{n})$. Therefore, the probability of finding exactly r balls in the ith bin is given by

$$p_r = \binom{m}{r}\left(\frac{1}{n}\right)^r\left(1-\frac{1}{n}\right)^{m-r} = \frac{1}{r!}\cdot\frac{m(m-1)\cdots(m-r+1)}{n^r}\left(1-\frac{1}{n}\right)^{m-r}.$$

When both m and n are large compared to r (i.e., when m *swamps* r), we can apply the following approximation:

$$p_r \approx \frac{1}{r!}\cdot\frac{m^r}{n^r}\left(1-\frac{1}{n}\right)^m \approx \frac{e^{-m/n}(m/n)^r}{r!},$$

which matches the definition of the Poisson distribution with parameter $\lambda = \frac{m}{n}$—precisely the average number of balls one would expect per bin.

However, analyzing the number of balls distributed among bins can be surprisingly elusive (even though the distribution follows a multinomial random variable). The difficulty arises from the *dependencies* inherent in the random process: if we know the number of balls in $n-1$ bins, the count in the final nth bin is fully determined. In other words, the *loads* of the bins are not independent.

To sidestep this dependency while still capturing the essential probabilistic behavior, we use the **Poisson approximation**, which treats the number of balls in each bin as an independent Poisson random variable.

Theorem 14 (Poisson Approximation). *Let $Y_1^{(m)}, \ldots, Y_n^{(m)}$ be independent Poisson random variables with parameter $\lambda = \frac{m}{n}$, and let $X_1^{(k)}, \ldots, X_n^{(k)}$ represent the joint distribution of k balls uniformly thrown into n bins (i.e., the balls-and-bins model). Then,*

$$\left(Y_1^{(m)}, \ldots, Y_n^{(m)} \mid \sum_{i=1}^n Y_i^{(m)} = k\right) \sim \left(X_1^{(k)}, \ldots, X_n^{(k)}\right).$$

Proof. The probability of observing a particular configuration $(k_1, \ldots, k_n)$ in the balls-and-bins model (by Definition **??**):

$$\Pr(X_1^{(k)} = k_1, \ldots, X_n^{(k)} = k_n) = \frac{k!}{k_1!\cdots k_n!}\cdot\left(\frac{1}{n}\right)^k.$$

On the other hand, the joint probability of the independent Poisson random variables $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ taking the same values is:

$$\Pr(Y_1^{(m)} = k_1, \ldots, Y_n^{(m)} = k_n) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{k_i}}{k_i!} = e^{-n\lambda}\cdot\prod_{i=1}^n \frac{\lambda^{k_i}}{k_i!}.$$

Now consider the conditional distribution of $(Y_1^{(m)}, \ldots, Y_n^{(m)})$ given that $\sum_{i=1}^n Y_i^{(m)} = k$:

$$\Pr\left(Y_1^{(m)} = k_1, \ldots, Y_n^{(m)} = k_n \mid \sum_{i=1}^n Y_i^{(m)} = k\right) = \frac{\Pr(Y_1^{(m)} = k_1, \ldots, Y_n^{(m)} = k_n)}{\Pr\left(\sum_{i=1}^n Y_i^{(m)} = k\right)}.$$

Using the fact that $X + Y \sim Pois(\lambda_1 + \lambda_2)$ for independent random variables $X \sim Pois(\lambda_1)$ and $Y \sim Pois(\lambda_2)$ (see proof in Lemma **??**), along with $\lambda = \frac{m}{n}$ we get:

$$= \frac{e^{-m} \cdot \frac{(\frac{m}{n})^k}{k_1! \cdots k_n!}}{e^{-m} \cdot \frac{m^k}{k!}} = \frac{k!}{k_1! \cdots k_n!} \cdot \left(\frac{1}{n}\right)^k .$$

□

Lemma 23. *For any two independent Poisson random variables $X \sim Pois(\lambda_1)$ and $Y \sim Pois(\lambda_2)$, we have $X + Y \sim Pois(\lambda_1 + \lambda_2)$.*

Proof.

$$\begin{aligned}
\Pr(X+Y=j) &= \sum_{k=0}^{j} \Pr(X=k) \cdot \Pr(Y=j-k) && \text{(by independence)} \\
&= \sum_{k=0}^{j} \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{j-k}}{(j-k)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{j!} \sum_{k=0}^{j} \frac{j!}{k!(j-k)!} \lambda_1^k \lambda_2^{j-k} && \text{(factoring out constants)} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{j!} \sum_{k=0}^{j} \binom{j}{k} \lambda_1^k \lambda_2^{j-k} && \text{(by the definition of binomial coefficient)} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^j}{j!} && \text{(by the binomial expansion theorem).}
\end{aligned}$$

□

The Poisson approximation helps derive powerful results about *any* function on the loads of the bins (e.g., the maximum load). As brief examples, we provide a theorem and its corollary.

Theorem 15. *Let $f(x_1, \ldots, x_n)$ be a non-negative function. Then,*

$$E\left[f\left(X_1^{(m)}, \ldots, X_n^{(m)}\right)\right] \leq \sqrt{2m\pi} \cdot E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right)\right].$$

Proof.

$$\begin{aligned}
E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right)\right] &= \sum_{k=0}^{\infty} E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right) \,\middle|\, \sum_{i=1}^{n} Y_i^{(m)} = k\right] \cdot \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = k\right) \\
&\text{(by Lemma ??)} \\
&\geq E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right) \,\middle|\, \sum_{i=1}^{n} Y_i^{(m)} = m\right] \cdot \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = m\right) \\
&= E\left[f\left(X_1^{(m)}, \ldots, X_n^{(m)}\right)\right] \cdot \Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = m\right) \\
&\text{(by Poisson approximation).}
\end{aligned}$$

Moreover, by Lemma **??**, we know

$$\Pr\left(\sum_{i=1}^{n} Y_i^{(m)} = m\right) = \frac{m^m e^{-m}}{m!}.$$

Using Stirling's approximation (see Lemma **??**), we have:

$$\begin{aligned} E\left[f\left(X_1^{(m)}, \ldots, X_n^{(m)}\right)\right] &\leq \frac{\sqrt{2m\pi}(me^{-1})^m}{(me^{-1})^m} \cdot E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right)\right] \\ &\leq \sqrt{2m\pi} \cdot E\left[f\left(Y_1^{(m)}, \ldots, Y_n^{(m)}\right)\right]. \end{aligned}$$

□

Following this result, we derive a useful corollary.

Theorem 16. *Suppose an event $\mathcal{E}$ occurs with probability p in the Poisson case. Then, the probability of the same event occurring in the exact balls-and-bins model is bounded above by $\sqrt{2m\pi} \cdot p$.*

Proof. Let f be the indicator function for the event $\mathcal{E}$, so $E[f] = p$. Theorem **??** then directly yields the result. □

To demonstrate how useful this bound is, suppose f counts the total number of empty bins, and let g be an indicator function for whether a single bin is empty. Then $f(Y^{(m)}) = \sum_i g(Y_i^{(m)})$. Since each $Y_i^{(m)} \sim \text{Pois}\left(\frac{m}{n}\right)$, we know that

$$E[g(Y_i^{(m)})] = p_0 = e^{-\frac{m}{n}}.$$

Applying Theorem **??** and using linearity of expectation:

$$E[f(X^{(m)})] \leq \sqrt{2m\pi} \cdot ne^{-\frac{m}{n}}.$$

Thus, we have successfully bounded the *expected number of empty bins* when placing m balls into n bins. This can be empirically verified through simulation! **Note.** In order for this bound to be fully useful (i.e., not overly loose and exhibiting the exponential decay behavior effectively), $m \gg n$.

8 Discrete Stochastic Process: Markov Chain

Definition 27 (Stochastic Process). *A **stochastic process** is a collection of random variables $\tilde{X} = \{X(t) : t \in T\}$, where the index t typically represents time. Thus, $\tilde{X}$ models the evolution of a random variable X over time through some random process.*

We refer to $X(t)$ (also often written as X_t) as the **state** of the process at time t. Two key distinctions characterize the type of process:

- If X_t takes values in a **countable** (possibly infinite) set, then $\tilde{X}$ is called a **discrete-state** process; otherwise, it is a **continuous-state** process.
- If T is a **countable** set, we say $\tilde{X}$ is a **discrete-time** process; otherwise, it is a **continuous-time** process.

In this note, we focus specifically on a type of **discrete-time** stochastic process $X_0, X_1, \dots$ in which the value of X_t depends only on the value of X_{t-1}, and not on the sequence of states that preceded it. This is known as a **Markov chain**. More formally:

Definition 28 (Markov Chain). *A discrete-time stochastic process $\tilde{X} = \{X_t : t \in T\}$ is a **Markov chain** if for all $t \in T$ and all states $a_0, a_1, \dots, a_t$,*

$$\Pr(X_t = a_t \mid X_{t-1} = a_{t-1}, X_{t-2} = a_{t-2}, \dots, X_0 = a_0) = \Pr(X_t = a_t \mid X_{t-1} = a_{t-1}).$$

Moreover, the following notation is convenient:

$$P_{a_{t-1},a_t} \overset{def}{=} \Pr(X_t = a_t \mid X_{t-1} = a_{t-1}).$$

It is important to emphasize that the **Markov property**—i.e., the *memorylessness* of the sequence of the chain—does not imply that X_t is independent of $X_0, X_1, \dots, X_{t-2}$. Rather, it means that any statistical dependence on the past is fully captured by the most recent state X_{t-1}. Also, we only consider discrete-state (and discrete-time) Markov chain for all purposes.

Next, we illustrate two standard representations of a Markov chain: (1) the *transition matrix* and (2) the *graph representation*.

8.1 Representations of Markov Chain

8.1.1 Transition Matrix

Definition 29 (One-step Transition Matrix). *Without loss of generality, we assume that the discrete state space of the Markov chain is $\{0, 1, 2, \dots, n\}$ (or $\{0, 1, 2, \dots\}$ if it is countably infinite). The **transition probability** is defined as $P_{i,j} = \Pr(X_t = j \mid X_{t-1} = i)$, as given in Definition **??**. This quantity denotes the probability that the process transitions from state i to state j in one step. Therefore, the Markov chain is completely characterized by the **one-step transition matrix**:*

$$\vec{P} = \begin{bmatrix} P_{0,0} & P_{0,1} & \dots & P_{0,j} & \dots \\ P_{1,0} & P_{1,1} & \dots & P_{1,j} & \dots \\ \vdots & \vdots & \ddots & \vdots & \dots \\ P_{i,0} & P_{i,1} & \dots & P_{i,j} & \dots \\ \vdots & \vdots & \ddots & \vdots & \ddots \end{bmatrix}$$

To verify that this matrix defines a valid probability distribution, we observe the following theorem:

Theorem 17. *For all i, $\sum_{j \geq 0} P_{i,j} = 1$.*

Proof.

$$\begin{aligned}\sum_{j\geq 0} P_{i,j} &= \sum_{j\geq 0} \Pr(X_t = j \mid X_{t-1} = i)\\ &= \sum_{j\geq 0} \frac{\Pr(X_t = j \cap X_{t-1} = i)}{\Pr(X_{t-1} = i)}\\ &= \frac{\Pr(X_{t-1} = i)}{\Pr(X_{t-1} = i)}\\ &= 1.\end{aligned}$$

Since $\sum_{j\geq 0} \Pr(X_t = j \cap X_{t-1} = i) = \Pr(X_{t-1} = i)$ by the law of total probability (see Lemma **??**). □

This matrix representation of a Markov chain is convenient for computing the distribution of future states; that is, the probability of the process being in state i at time t.

Definition 30 (Probability Distribution of States at Time t)**.** *Let $p_i(t)$ denote the probability that the process is in state i at time t, and let $\overline{p}(t) = (p_0(t), p_1(t), \dots)$ be the state distribution vector at time t. Then,*

$$p_i(t) = \sum_{j\geq 0} p_j(t-1)P_{j,i} \quad \textit{(by the law of total probability)},$$

which is equivalently expressed in vector form as:

$$\overline{p}(t) = \overline{p}(t-1) \cdot \vec{P}.$$

Therefore, knowing $\overline{p}(t)$ allows us to compute the state distribution at time $t+1$, and so on.

Theorem 18 (m-step Transition Matrix)**.** *Let $\vec{P}^{(m)}$ be the m-step transition matrix, where the entry $P_{i,j}^{(m)}$ denotes the probability that the chain moves from state i to state j in exactly m steps. Then,*

$$\vec{P}^{(m)} = \vec{P}^m.$$

Proof. The base case $m = 1$ is trivially true. Assume the statement holds for $m = k$, i.e., $\vec{P}^{(k)} = \vec{P}^k$. Then for $m = k + 1$:

$$\begin{aligned}\vec{P}^{(k+1)} &= \vec{P}^{(k)} \cdot \vec{P} && \text{(the next step must follow the one-step transition)}\\ &= \vec{P}^k \cdot \vec{P} && \text{(by the induction hypothesis)}\\ &= \vec{P}^{k+1}.\end{aligned}$$

For convenience, we write $\vec{P}^m$ in place of $(\vec{P})^m$. The entries of $\vec{P}^m$ are given by:

$$P_{i,j}^{(m)} = \sum_{k\geq 0} P_{i,k}^{(m-1)} \cdot P_{k,j},$$

which follows from conditioning on the intermediate state k and applying the law of total probability. □

An important corollary of the theorem above is that for all $t \geq 0$ and $m \geq 1$, we have:

$$\overline{p}(t+m) = \overline{p}(t)\vec{P}^m.$$

This result allows us to predict future distributions. Additionally, another important concept is the notion of a **stationary distribution**.

Definition 31 (Stationary Distribution). *A* ***stationary distribution*** *(also called an equilibrium distribution) of a Markov chain is a probability distribution $\overline{p}$ such that:*

$$\overline{p} \cdot \vec{P} = \overline{p}.$$

This special distribution is typically denoted by $\overline{\pi} = (\pi_0, \pi_1, \dots)$.

As for interpretation, suppose we define the initial distribution $\overline{p}$ at time $t = 0$ arbitrarily—that is, before the stochastic process begins, we randomly choose the initial state according to the probability distribution specified by $\overline{p}$. This initial behavior then propagates forward in time via the one-step transition matrix, determining the probability of being in each state at subsequent time steps. However, if the initial distribution $\overline{p}$ happens to be the stationary distribution $\overline{\pi}$, then the application of the one-step transition matrix leaves the distribution unchanged. In other words, the distribution over states remains $\overline{\pi}$ at every time step, as though the transition dynamics have no net effect.

We will return to this remarkable property shortly in the form of a formal theorem. However, note that one can compute the stationary distribution by solving the linear system $\overline{\pi} - \overline{\pi} \cdot \vec{P} = 0$ (this follows from rearranging the terms in Definition **??**), together with the additional constraint $\sum_i \pi_i = 1$, since $\overline{\pi}$ must be a probability distribution.

We now proceed to the graph representation of a Markov chain.

8.1.2 Graph Representation

Definition 32 (Graph Representation of a Markov Chain). *A Markov chain can be represented as a* ***directed, weighted graph*** *$G = (V, E, w)$ derived from its one-step transition matrix. Its construction is as follows:*

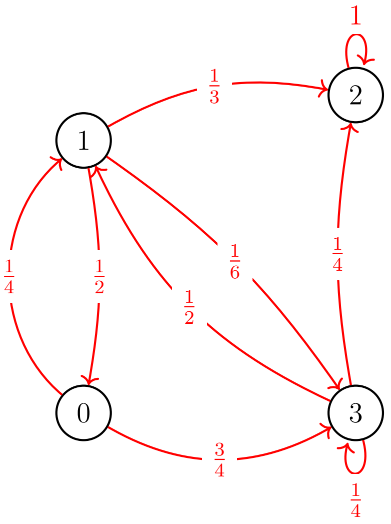
1. *The set of vertices V corresponds to all states of the chain.*
2. *There exists a directed edge $(i, j) \in E$ if and only if $P_{i,j} > 0$.*
3. *If an edge (i, j) exists, then its weight is given by $w(i, j) = P_{i,j}$.*

Self-loops are permitted (i.e., a vertex may have a directed edge to itself), which occurs whenever $P_{i,i} > 0$, implying $w(i, i) > 0$. The total weight of all outgoing edges from any vertex i satisfies:

$$\sum_{j:(i,j)\in E} w(i, j) = 1,$$

as expected for a valid probability distribution.

For example:



This graph corresponds to the following one-step transition matrix:

$$\vec{P} = \begin{bmatrix} 0 & \frac{1}{4} & 0 & \frac{3}{4} \\ \frac{1}{2} & 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

8.2 Classification of States

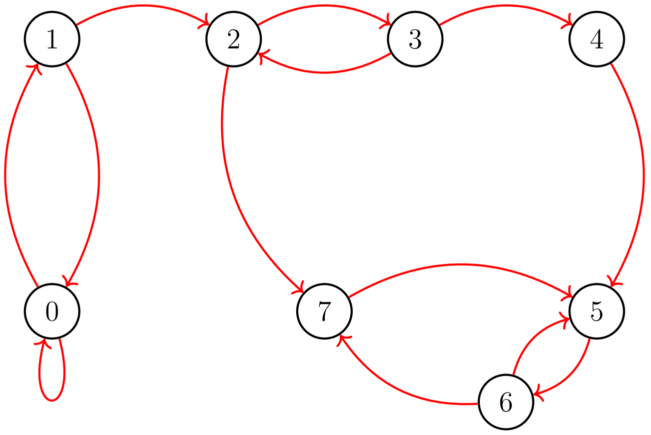
To analyze the *long-term behavior* of a Markov chain, we must classify its states.

Definition 33 (Accessibility and Communicating Relation)**.** *For $n \geq 0$, we say that state j is **accessible** from state i if and only if $P_{i,j}^n > 0$. Moreover, it is trivially true that $P_{i,i}^0 = 1$, which means every state is accessible from itself. Additionally, if two states i and j are accessible from each other (i.e., state j is accessible from state i and state i is accessible from state j), we say that they **communicate** and denote it as $i \leftrightarrow j$.*

In the graph representation (see Definition **??***), it is illustratively clear that $i \leftrightarrow j$ if and only if there exist directed paths connecting i to j and vice versa. If two states communicate, they form an equivalence relation known as the **communicating relation**, which satisfies:*

1. ***Reflexivity**: for any state i, $i \leftrightarrow i$;*
2. ***Symmetry**: if $i \leftrightarrow j$, then $j \leftrightarrow i$;*
3. ***Transitivity**: if $i \leftrightarrow j$ and $j \leftrightarrow k$, then $i \leftrightarrow k$.*

Thus, the states of a Markov chain can be *partitioned* into **communicating classes**—i.e., only members of the same class communicate with each other (courtesy of transitivity). For example, observe the following graph (assuming all weights $P_{i,j} > 0$):



We have the following communicating classes:

- **Class 1**: $\{0, 1\}$
- **Class 2**: $\{2, 3\}$
- **Class 3**: $\{4\}$
- **Class 4**: $\{5, 6, 7\}$

Definition 34 (Irreducibility)**.** *A Markov chain is **irreducible** if all states belong to a single communicating class; that is, for every pair of states (i, j), $i \leftrightarrow j$.*

Lemma 24. *A finite Markov chain (i.e., one with a finite number of states) is irreducible if and only if its graph representation is strongly connected; that is, every pair of vertices is reachable from each other.*

Proof. In the underlying digraph we draw an arc $i \to j$ iff $P_{ij} > 0$.

($\Rightarrow$) *Contrapositive.* Suppose the digraph is **not** strongly connected. Then there exist vertices i and j with no directed path $i \to j$. Hence $P_{ij}^m = 0$ for every $m \geq 1$, so i does not reach j and the chain has more than one communicating class; it is therefore not irreducible.

($\Leftarrow$) Assume the digraph is strongly connected. For any states i, j there is a directed path $i = v_0 \to v_1 \to \cdots \to v_k = j$ with $P_{v_{\ell-1}v_\ell} > 0$ for each ℓ. Multiplying those positive probabilities gives $P_{ij}^k = P_{v_0v_1} \cdots P_{v_{k-1}v_k} > 0$, so i reaches j. By symmetry, all states communicate; the chain is irreducible. □

Definition 35 (Aperiodicity)**.** *For a discrete-time Markov chain with transition matrix $\vec{P}$, define the* period *of state i as*

$$\Delta(i) \;=\; \gcd\{\, m \geq 1 : P_{i,i}^m > 0\}.$$

*The state i is **periodic** if $\Delta(i) > 1$ and **aperiodic** if $\Delta(i) = 1$.*

If a state is periodic, returns to i occur only at multiples of $\Delta(i)$; equivalently, $P_{i,i}^m = 0$ whenever $\Delta(i) \nmid m$.

To determine if a Markov chain is aperiodic, we apply the following procedure, which can be carried out manually for simple graphs:

1. Partition the chain into its communicating classes.

2. For each class, select an arbitrary state i. Determine whether state i is aperiodic by identifying two or more return paths from i to itself in the graph representation. Let the lengths of these paths be $\ell_1, \ell_2, \ldots, \ell_m$ (this means that $P_{i,i}^{\ell} > 0$). Then state i is aperiodic if and only if

$$\gcd(\ell_1, \ell_2, \ldots, \ell_m) = 1.$$

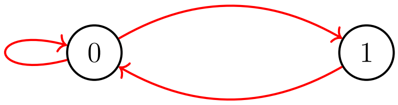
   Ensure that the set $\{\ell_1, \ell_2, \ldots, \ell_m\}$ is exhaustive unless one of the values of ℓ immediately yields 1 when computing the gcd.

3. If a single state within a communicating class is found to be aperiodic, then all states in that class are aperiodic. Otherwise, the entire class is periodic, and the chain is not aperiodic.

4. Repeat this check for all communicating classes to conclude whether the entire chain is aperiodic.

   Quick tip: If the Markov chain is *irreducible* and contains a self-loop (i.e., $P_{i,i} > 0$). Then we can immediately conclude the chain is aperiodic—a direct consequence of the procedure above.

As a simple example:



We can determine that the chain above is aperiodic because:

1. $\Delta(0)$: $\gcd(1, 2) = 1$, (Paths: $0 \to 0$ and $0 \to 1 \to 0$).

2. $\Delta(1)$: $\gcd(2, 3) = 1$, (Paths: $1 \to 0 \to 1$ and $1 \to 0 \to 0 \to 1$).

We continue with definitions and interesting results pertaining to stationary distributions.

Definition 36 (Recurrent States and Transient States)**.** *For $t \geq 1$, denote*

$$r_{i,j}^t \;=\; \Pr\big(X_t = j \textit{ and } X_k \neq j \textit{ for } 1 \leq k \leq t-1 \mid X_0 = i\big),$$

i.e., the probability that the first *transition to state j from state i occurs at time t. We say the state i is **recurrent** if $\sum_{t \geq 1} r_{i,i}^t = 1$; otherwise, if*

$$p \;:=\; \sum_{t \geq 1} r_{i,i}^t \;<\; 1,$$

*then i is **transient**. Thus, if i is recurrent, with probability 1 the chain returns to i (in fact, infinitely often) when started at i; otherwise it returns only with probability $p < 1$. In the transient case, the number of* returns *to i (after time 0) is geometric random variable.*

*Moreover, the **expected time** to first reach j from i (the hitting time) is*

$$h_{i,j} \;=\; \sum_{t \geq 1} t\, r_{i,j}^t.$$

Note that if one state in a communicating class is recurrent, then all states in that class are recurrent; likewise for transience. Therefore, a Markov chain is recurrent if every state in the chain is recurrent.

Definition 37 (Positive Recurrent State)**.** *If a state i is recurrent and $h_{i,i} < \infty$, then i is* ***positive recurrent****; otherwise (recurrent with $h_{i,i} = \infty$) it is* ***null recurrent****.*

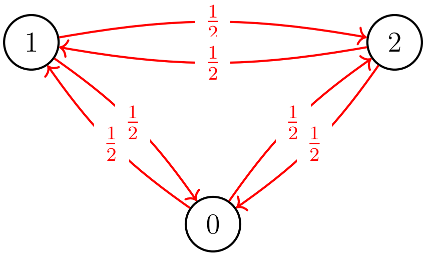
Definition 38 (Ergodic States)**.** *A state is* ***ergodic*** *if it is aperiodic and positive recurrent. Similarly, a Markov chain is ergodic if all of its states are ergodic.*

Theorem 19. *Any finite, irreducible, and ergodic Markov chain has a unique* ***stationary distribution*** *$\overline{\pi}$. Moreover, for each state i,*

$$\pi_i = \lim_{t\to\infty} P^t_{j,i} = \frac{1}{h_{i,i}},$$

so π_i is the limiting probability—independent of the initial state j—that the chain is in state i. It is as if the chain forgets its initial state far out in the future.

The proof of this remarkable theorem is out of scope for this note; however, one may enjoy this result with a concrete example.
Consider the following.



One may easily verify that this chain is finite, irreducible, and ergodic. The straightforward approach is to solve $\overline{\pi}\vec{P} - \overline{\pi} = \vec{0}$ with the additional constraint $\sum_i \pi_i = 1$—this method is of course general and applies to any Markov chain. Thus,

$$\begin{bmatrix}\pi_0 & \pi_1 & \pi_2\end{bmatrix}\begin{bmatrix}0 & \frac{1}{2} & \frac{1}{2}\\ \frac{1}{2} & 0 & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{2} & 0\end{bmatrix} - \begin{bmatrix}\pi_0 & \pi_1 & \pi_2\end{bmatrix} = \begin{bmatrix}0 & 0 & 0\end{bmatrix}.$$

Therefore we obtain the system

$$\begin{aligned}-\pi_0 + \tfrac{1}{2}\pi_1 + \tfrac{1}{2}\pi_2 &= 0,\\ \tfrac{1}{2}\pi_0 - \pi_1 + \tfrac{1}{2}\pi_2 &= 0,\\ \tfrac{1}{2}\pi_0 + \tfrac{1}{2}\pi_1 - \pi_2 &= 0,\\ \pi_0 + \pi_1 + \pi_2 &= 1.\end{aligned}$$

Solving yields $\overline{\pi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$.
Now we apply the theorem above to see if we obtain the same result. Immediately, by symmetry we deduce that $\pi_0 = \pi_1 = \pi_2$, since the hitting time $h_{i,i}$ is the same for every i. Thus it suffices to compute $h_{0,0}$. For convenience, let the random variable X_j denote

the expected time to reach state 0 from state j. Then we have the following recurrences (which can be solved as a linear system):

$$\begin{aligned} X_0 &= \tfrac{1}{2}\big(1 + X_1\big) + \tfrac{1}{2}\big(1 + X_2\big), \\ X_1 &= \tfrac{1}{2} \cdot 1 + \tfrac{1}{2}\big(1 + X_2\big), \\ X_2 &= \tfrac{1}{2} \cdot 1 + \tfrac{1}{2}\big(1 + X_1\big). \end{aligned}$$

Solving gives $X_0 = 3$, hence $\pi_0 = \frac{1}{h_{0,0}} = \frac{1}{3}$, concluding our result spectacularly!

In fact, the chain above is a special type of Markov chain known as a *random walk*, which is often used to analyze algorithms.

Definition 39 (Random walk on undirected graphs)**.** *A random walk on an undirected graph $G = (V, E)$ is a process in which a particle at vertex i moves at each step to a neighboring vertex along an incident edge, choosing each neighbor with probability* $1/\deg(i)$*. Here* $\deg(i)$ *denotes the degree of i.*

9 Counting

Many of the analyses shown here in discrete probability are riddled with *counting*; for example, counting the number of successes to find the probability of success or counting the number of possible outcomes to establish a sample space (or a conditional sample space). In this segment, we aim to cover the mathematics of counting to make sense of many results shown definitively.

9.1 The Counting Principle

This is a mathematical shortcut to find the total number of possible outcomes from a series of collections. Suppose we have r collections, and each collection C_i has n_i elements. To find *all* possible outcomes, we simply compute the *Cartesian product* to obtain the **ordered** r-tuples. Thus,

$$|C_1 \times \cdots \times C_r| \;=\; \prod_{i=1}^{r} n_i,$$

and the result tells us how many possible outcomes there are. We can see this fact is correct because in the first collection there are n_1 elements to choose from, in the second n_2 elements, and so on.

For example, suppose Alice has three shirts, two pairs of pants, and three caps; the number of ways she can combine them is therefore $3 \times 2 \times 3 = 18$ by the counting principle. Likewise, the number of legitimate telephone numbers of a 9-digit sequence where the first digit must be 1 or 9 is $2 \times 10 \times \cdots \times 10 = 2 \cdot 10^8$ by the counting principle.

9.2 k-Permutations

We are interested in choosing k elements from a collection of n *distinct* elements (with $k \leq n$) and then arranging them in a sequence—how many ways can we do this? By the counting principle, we have n choices for the first element, then $n - 1$ for the second, and so on, down to $n - k + 1$ for the kth element. (Note that we might select the same set of k elements more than once, but in different **orders**; this reflects the "arrangement" aspect of k-permutations.) Thus, the number of ways is

$$n\,(n-1)\,\ldots\,(n-k+1) = \frac{n!}{(n-k)!}.$$

It is straightforward to verify that this equality holds. In particular, the number of ways to arrange all n distinct objects—the case of n-permutations—is obtained by setting $k = n$ in the formula above, yielding $n!$.

9.3 Combinations

We are again interested in choosing k elements from a collection of n *distinct* elements ($k \leq n$), **but** now the order (i.e., arrangement) in which they are chosen does not matter. First note that there are $k!$ ways to order any fixed set of k selected elements; accordingly, the number of k-permutations equals $k!$ times the number of k-combinations. Hence the number of combinations is

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

This fact underlies the *binomial distribution*: when counting the number of successes in n Bernoulli trials, we need $\binom{n}{k}$ because we do not care about the order in which the k successes occur, which greatly simplifies the computation.

9.4 Partitions

Suppose we have a collection of n *distinct* elements. We wish to *partition* them into r groups: the first group having n_1 elements, the second n_2 elements, and so on, with $\sum_{i=1}^{r} n_i = n$. Because the order in which the elements are chosen does not matter in each group, we obtain, by repeated use of combinations and the counting principle, that the number of ways to form a partition of type $(n_1, n_2, \ldots, n_r)$ is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{r-1}}{n_r}.$$

Writing each binomial in factorial form,

$$\frac{n!}{n_1!(n-n_1)!}\,\frac{(n-n_1)!}{n_2!(n-n_1-n_2)!}\cdots\frac{(n-n_1-\cdots-n_{r-1})!}{n_r!\,(n-n_1-\cdots-n_{r-1}-n_r)!},$$

and canceling the telescoping terms (with $(n-n_1-\cdots-n_r)! = 0! = 1$) yields

$$\frac{n!}{n_1!\,n_2!\,\ldots\,n_r!} = \binom{n}{n_1; n_2; \ldots; n_r}.$$

Example. To find the number of anagrams of the word `COLOSSAL`, note that there are eight *unique* positions which can be partitioned into five groups according to repeated letters: $G_C = 1$, $G_O = 2$, $G_L = 2$, $G_S = 2$, and $G_A = 1$. Hence the number of anagrams is

$$\frac{8!}{1!\,2!\,2!\,2!\,1!} = 5040.$$

Acknowledgement

I would like to acknowledge *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*, 2nd edition, by Michael Mitzenmacher and Eli Upfal, as the primary source of reference for these notes.