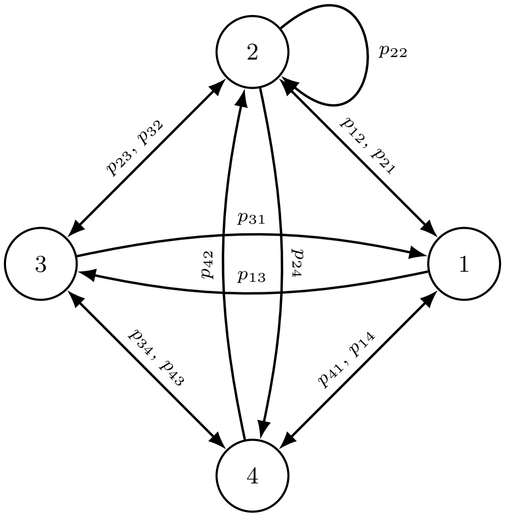
EECS 126: Random Processes

Lecture Notes

Semester Fall 2025



Alexander Lu

Instructor: Preeya Khanna

Contents

1 Bernoulli and Poisson Processes 3

1.1 Bernoulli Process . . . 3

1.1.1 Independent and Memorylessness . . . 4

1.1.2 Interarrival Times . . . 5

1.1.3 The kth Arrival Time . . . 5

1.1.4 Splitting and Mergeing Bernoulli Processes . . . 7

1.1.5 Estimating Binomial with Poisson . . . 7

1.2 The Poisson Process . . . 8

1.2.1 Independence and Memorylessness . . . 10

1.2.2 Interarrival Times and Arrival Times . . . 11

1.2.3 Splitting and Merging Poisson Processes . . . 12

1.3 Bernoulli and Poisson Processes, and Sums of Random Variables . . . 13

1.4 The Random Incidence Paradox . . . 13

2 Markov Chains 15

2.1 Discrete-Time Markov Chains . . . 15

2.1.1 The Probability of a Path . . . 16

2.1.2 n-Step Transition Probabilities . . . 16

2.2 Classification of States . . . 17

2.2.1 Periodicity . . . 19

2.3 Steady-State Behavior . . . 19

2.3.1 Long-Term Frequency Interpretations . . . 21

2.3.2 Birth-Death Processes . . . 21

2.4 Absorption Probabilities and Expected Time to Absorption . . . 22

2.4.1 Mean First Passage and Recurrence Times . . . 23

2.5 Continuous-Time Markov Chains . . . 24

2.5.1 Approximation of CTMC with DTMC . . . 25

2.5.2 Steady-State Behavior of CTMC . . . 26

2.5.3 CTMC Birth-Death Processes . . . 28

2.6 Reversible Processes . . . 28

3 Bayesian Statistical Inference 29

3.1 Bayesian Inference and the Posterior Distribution . . . 30

3.2 Point Estimation, Hypothesis Testing, and MAP . . . 31

3.2.1 Point Estimation . . . 31

3.2.2 Hypothesis Testing . . . 31

3.3 Bayesian Least Mean Squares Estimation (LMS) 32
3.3.1 Estimation Error 33
3.4 Bayesian Linear Least Mean Squares Estimation (LLSE) 33

4 Frequentist Statistical Inference **36**
4.1 Frequentist Parameter Estimation 36
4.2 Maximum Likelihood Estimation (MLE) 36

Chapter 1

Bernoulli and Poisson Processes

Stochastic processes are probabilitistic models that evolve over time steps, generating a sequence of values. Many of the data in real life may be modeled as stochastic processes, such as stock prices, weather patterns, and queuing systems. We may often think of such a process as a sequence of random variables $\{X_n\}_{n=0}^{\infty}$, where X_n is the value of the process at time step n. There are three major things that analysts typically search for:

- **Dependency Relationships:** How does the value of the process at time step n depend on previous time steps? For example, is X_n dependent only on X_{n-1} (a *Markov* process), or does it depend on all previous values?
- **Long-Term Behavior:** As n becomes large, does the process converge to a steady-state distribution? For example, does the average value of X_n converge to some constant?
- **Boundary Events:** What is the probability that the process reaches some critical threshold? For example, what is the probability that X_n exceeds some value a at any time step?

There are two major processes that we consider: the **Arrival-type processes** and the **Markov process**.

1.1 Bernoulli Process

A **Bernoulli process** is a discrete-time stochastic process that models a sequence of independent and identically distributed (i.i.d.) Bernoulli trials. At each time step, an trial occurs with a fixed probability p of success, and a probability $1-p$ of failure. In the context of arrival-time processes, a success may represent the arrival of an event (e.g., a customer arriving at a service point), while a failure represents no arrival. As a recap, here are some key properties of a Bernoulli process:

Definition 1.1.1: Bernoulli Process

A **Bernoulli process** is a discrete-time stochastic process $\{X_n\}_{n=0}^{\infty}$ where each X_n is a Bernoulli random variable with parameter p. The trials are independent, and the probability of success (arrival) at each time step is constant.

- **PMF:**
$$P(X = m) = \binom{n}{m} p^m (1-p)^{n-m}, \quad m = 0, 1, \dots, n$$

- **Expectation and Variance:**
$$\mathbb{E}[X] = np, \quad \text{Var}(X) = np(1-p)$$

1.1.1 Independent and Memorylessness

Two key properties of the Bernoulli process are **independence** and **memorylessness**. Independence assures us that the outcome on one trial does not affect the outcome of any other trials (the arrival of a customer does not influence future arrivals). Memorylessness means that the probability of an arrival occuring in a given time step is independent of past arrivals (the probability of a customer arriving now does not depend on how many customers have arrived in the past).

Independence is a powerful tool. Consider any two random variables X and Y. If they are independent, then any two functions of them are also independent. That is, if f and g are any functions, then $f(X)$ and $g(Y)$ are independent random variables.

Memorylessness gives us the **fresh-start** property: after any time step, the process essentially restarts itself, and remaining trials also form a Bernoulli process with the same parameter p.

Within a Bernoulli process, we may also consider the first arrival time. Let T denote the time of the first arrival (success) in a Bernoulli process with parameter p. Then, T follows a **Geometric distribution** with parameter p. Again, here is a recap of its properties:

Definition 1.1.2: Geometric Distribution

A random variable T follows a **Geometric distribution** with parameter p if it represents the time of the first success in a sequence of independent Bernoulli trials with success probability p.

- **PMF:**
$$P(T = k) = (1-p)^{k-1} p, \quad k = 1, 2, \dots$$

- **Expectation and Variance:**
$$\mathbb{E}[T] = \frac{1}{p}, \quad \text{Var}(T) = \frac{1-p}{p^2}$$

We may now observe how the memorylessness property of the Bernoulli process is a direct consequence of the memorylessness of the Geometric distributin. Suppose that our Bernoulli process has been running for n trials without success. Since we know that the first success time T is Geometric with parameter p, by the memorylessness property of the Geometric distribution, we have:

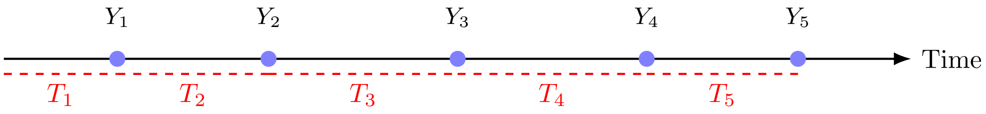
$$\begin{aligned} P(T > n + k \mid T > n) &= \frac{P(T > n + k)}{P(T > n)} \\ &= \frac{(1-p)^{n+k}}{(1-p)^n} \\ &= (1-p)^k \\ &= P(T > k) \end{aligned}$$

This shows that the probability of waiting an additional k trials for the first success does not depend on how many trials have already passed without success, which is exactly the memorylessness property of the Bernoulli process.

1.1.2 Interarrival Times

We introduce two important variables in the context of arrival processes:

- **Arrivial Time Y_k:** The time of the k-th arrival (success).
- **Interarrival Time T_k:** The time between the $(k-1)$-th and k-th arrivals, defined as $X_k = Y_k - Y_{k-1}$ with $Y_0 = 0$.



Intuitively, we notice that the interarrival times $\{T_k\}$ are i.i.d. Geometric random variables with parameter p. This is because over discrete time, the interarrival time is essentially a count of the number of Bernoulli trials until the next success.

1.1.3 The kth Arrival Time

The arrival times $\{Y_k\}$ can be expressed as a cumulative sum of the interarrival times:

$$Y_k = \sum_{i=1}^{k} T_i$$

We may derive some useful properties of the k-th arrival time Y_k. In fact, it's distribution is so important that we give it a name:

Definition 1.1.3: Pascal Distribution, Negative Binomial Distribution

- The k-th arrival time Y_k in a Bernoulli process is modeled by a sum of k

independent Geometric random variables with parameter p.

$$Y_k = \sum_{i=1}^{k} T_i$$

- The mean and variance of Y_k are given by:

$$\mathbf{E}[Y_k] = \mathbf{E}\left[\sum_{i=1}^{k} T_i\right] = \sum_{i=1}^{k} \mathbf{E}[T_i] = k \cdot \frac{1}{p} = \frac{k}{p}$$
$$\operatorname{Var}(Y_k) = \operatorname{Var}\left(\sum_{i=1}^{k} T_i\right) = \sum_{i=1}^{k} \operatorname{Var}(T_i) = k \cdot \frac{1-p}{p^2} = \frac{k(1-p)}{p^2}$$

- The PMF of Y_k is given by:

$$\mathbf{P}(Y_k = t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \quad t = k, k+1, \dots$$

This is known as the **Pascal PMF of order** k, (or Negative Binomial PMF).

Proof. We prove the formula of the PMF of Y_k. Note that it only makes sense for $t \geq k$, since we need at least k trials to have k arrivals. For $t \geq k$, we may split the event $\{Y_K = t\}$ into two separate events:

- **Event A:** The k-th arrival occurs at time t.
- **Event B:** The first $k-1$ arrivals occur in the first $t-1$ trials.

In order for $\{Y_k = t\}$ to occur, both event A and event B must occur. Since A and B are independent, we note that:

$$\mathbf{P}(Y_k = t) = \mathbf{P}(A \cap B) = \mathbf{P}(A)\,\mathbf{P}(B)$$

We may compute $\mathbf{P}(A)$ and $\mathbf{P}(B)$ separately. First, we have:

$$\mathbf{P}(A) = p$$

since the k-th arrival must occur at time t. Next, we compute $\mathbf{P}(B)$. Note that B is just the binomial distribution counting the number of arrivals in $t-1$ trials. Thus, we have:

$$\mathbf{P}(B) = \binom{t-1}{k-1} p^{k-1} (1-p)^{(t-1)-(k-1)} = \binom{t-1}{k-1} p^{k-1} (1-p)^{t-k}$$

Combining these two results, we obtain:

$$\mathbf{P}(Y_k = t) = \mathbf{P}(A)\,\mathbf{P}(B) = p\binom{t-1}{k-1} p^{k-1} (1-p)^{t-k} = \binom{t-1}{k-1} p^{k} (1-p)^{t-k}$$

□

1.1.4 Splitting and Mergeing Bernoulli Processes

We may consider two operations on Bernoulli processes: **splitting** and **merging**.

Definition 1.1.4: Splitting, Merging

- **Splitting:** Given a Bernoulli process with parameter p, for every arrival, we may choose to keep it with a probability q, or discard it with probability $1-q$. The resulting process of kept arrivals is a Bernoulli process, with parameter pq. The discarded arrivals also form a Bernoulli process with parameter $p(1-q)$.
- **Merging:** Given two independent Bernoulli processes with patameters p and q, we include an arrival in the merged process if it occurs in either of the two original processes. The resulting merged process is a Bernoulli process with parameter $p+q-pq$.

Why would we want to do this? Consider the following examples:

- **Splitting:** Splitting may be used to model a situation where arrivals are randomly filtered. For example, in a network system, packets may arrive according to a Bernoulli process, but some packets may be dropped due to congestion or errors. By splitting the original process, we can analyze the behavior of the successfully transmitted packets separately from the dropped packets.
- **Merging:** Merging may be used to model a situation where arrivals from multiple sources are combined into a single process. For example, in a call center, calls may arrive from different departments according to separate Bernoulli processes. By merging these processes, we can analyze the overall call arrival rate and behavior.

1.1.5 Estimating Binomial with Poisson

In the real world, processes rarely occur at large, discrete time intervals. It may be useful to consider a continuous-time approximation of the Bernoulli process. Consider the scenario where we have a large number of trials n, each with a small probability of success p (Intuitively, this is like dividing a fixed time interval into many small subintervals, where each subinterval has a small chance of an arrival). The mean of our Bernoulli process is np, we increase n and decrease p such that $np = \lambda$ where λ is a fixed constant. We observe that taking the limit of the binomial in this manner reduces to the **Poisson distribution** with rate parameter λ:

$$\lim_{n\to\infty, p\to 0, np=\lambda} \text{Binomial}(n,p) = \text{Poisson}(\lambda)$$

Definition 1.1.5: Poisson Random Variable

- A Poisson random variable X with rate parameter $\lambda > 0$ has the following

PMF over the non-negative integers:

$$\mathbf{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

- The mean and the variance of a Poisson random variable are both equal to λ:

$$\mathbf{E}[X] = \lambda, \quad \mathrm{Var}(X) = \lambda$$

- The limit of the binomial PMF as n increases and p decreases such that $np = \lambda$ simplifies to the Poisson PMF

Proof. We prove the limit of the binomial PMF. Let Y be a binomial random variable with parameters n and p. Let $\lambda = np$. We have that:

$$\begin{aligned}\mathbf{P}(Y = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^n \cdot \left(1 - \frac{\lambda}{n}\right)^{-k}\end{aligned}$$

Note that as n approaches infinity, the $\frac{n-i}{n}$ terms approach 1 for $i = 0, 1, \ldots, k-1$. Additionally:

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}, \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \to 1$$

Hence, when we increase n to infinity, we obtain:

$$\lim_{n\to\infty} \mathbf{P}(Y = k) = 1 \cdot \frac{\lambda^k}{k!} \cdot e^{-\lambda} \cdot 1 = \frac{e^{-\lambda}\lambda^k}{k!}$$

As desired. □

1.2 The Poisson Process

The limitation of the Bernoulli process that we just explored is that it only measures arrivals at discrete time intervals. This is rarely the case in real life, where arrivals may occur at any continuous time. If two arrivals (successes) occurs at the same discrete interval, the Bernoulli process has no way of recording both events. To be able to model such scenarios, we somehow need to encode continuous time interarrival times into our model. The solution is to use the **Poisson Process**, which is a continuous-time analog of the Bernoulli process.

The trick is to shrink the time intervals as small as possible, to the point where we are essentially working over a continuous timeline. Now, any number $t \in \mathbb{R}$ is a valid arrival time. Similarly, for any real interval τ of time, we may count the number of arrivals within that interval. We define:

$$\mathbf{P}(k, \tau) = \mathbf{P}(\text{Exactly } k \text{ arrivals occur in time interval of length } \tau)$$

Definition 1.2.1: Poisson Process

A Poisson Process with rate parameter $\lambda > 0$ satisfies the following properties:

- **Time-Homogeneity:** The probability k, τ of K arrivals is the same for all time intervals of length τ. That is, the process has a constant rate λ of arrivals per unit time.
- **Independent:** The number of arrivals during a particular interval is independent of the number of arrivals during any other non-overlapping interval.
- **Small Interval Probabilities:** For a small time interval, the probability of one arrival is approximately proportional to the length of the interval, while the probability of two or more arrivals is negligible. Specifically, for a small interval of length Δt:

  $$\mathbf{P}(1, \Delta t) \approx \lambda \Delta t, \quad \mathbf{P}(0, \Delta t) \approx 1 - \lambda \Delta t, \quad \mathbf{P}(k \geq 2, \Delta t) \approx 0$$

  This is what makes the Poisson process a continuous-time analog of the Bernoulli process.
- **Expectation and Variance:** The expectation and variance of arrivals in a time interval of length τ are equal and given by:

  $$\mathbf{E}[N_\tau] = \lambda \tau, \quad \mathrm{Var}(N_\tau) = \lambda \tau$$

Consider any interval of length τ. Now imagine evenly splitting this interval into many small subintervals of length δ. We have a total of $n = \frac{\tau}{\delta}$ subintervals. Due to the small interval properties property of the Poisson process, we may approximate this interval as a Bernoulli process with n trials and success probability $p = \lambda\delta$. Thus, the number of arrivals in this interval follows a Binomial distribution:

$$N_\tau \sim \text{Binomial}\left(n = \frac{\tau}{\delta}, p = \lambda\delta\right)$$

Taking the limit as $\delta \to 0$ (and thus $n \to \infty$), we obtain:

$$\mathbf{P}(k, \tau) = e^{-\lambda\tau} \frac{(\lambda\tau)^k}{k!}$$

As we did in the previous section. We see that the expected number of arrivals in an interval of length τ is $\lambda\tau$, and the variance is also $\lambda\tau$.

We now move to derive the interarrival times of the Poisson process. Let T denote the time of the first arrival in a Poisson process with rate λ. To find the distribution of T, we find the CDF first:

$$F_T(t) = \mathbf{P}(T \leq t) = 1 - \mathbf{P}(T > t) = 1 - \mathbf{P}(0, t) = 1 - e^{-\lambda t}$$

Differentiating, we obtain the PDF of T:

$$f_T(t) = \frac{d}{dt} F_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$

This is the PDF of an **Exponential distribution** with rate parameter λ. Hence, the interarrival times in a Poisson process are i.i.d. Exponential random variables with rate parameter λ. Comparing this to the Bernoulli process, we see that the exponential distribution is the continuous-time analog of the geometric distribution. Both distributions model the survival time of a process, which make them ideal for modeling interarrival times. Here is a recap of the Exponential distribution:

Definition 1.2.2: Exponential Distribution

A random variable T follows an **Exponential distribution** with rate parameter $\lambda > 0$ if it represents the time until the first arrival in a Poisson process with rate λ.

- **PDF:**
$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0$$
- **CDF:**
$$F_T(t) = 1 - e^{-\lambda t}, \quad t \geq 0$$
- **Expectation and Variance:**
$$\mathbf{E}[T] = \frac{1}{\lambda}, \quad \mathrm{Var}(T) = \frac{1}{\lambda^2}$$

Interestingly, the Poisson Process provides an intuitive way of showing that the sum of independent Poisson Random variables is also Poisson. Consider two independent Poisson processes with rate parameters λ_1 and λ_2. We may merge these two processes into a single process by including an arrival if it occurs in either of the two original processes. The resulting merged process is a Poisson process with rate parameter $\lambda_1 + \lambda_2$. Thus, if $X_1 \sim \mathrm{Poisson}(\lambda_1)$ and $X_2 \sim \mathrm{Poisson}(\lambda_2)$ are independent Poisson random variables, then their sum $X = X_1 + X_2$ follows a Poisson distribution with parameter $\lambda_1 + \lambda_2$:

$$X \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$$

1.2.1 Independence and Memorylessness

Since the Poisson Process is derived from the Bernoulli process, it inherits the independence and memorylessness properties. The number of arrivals in non-overlapping intervals are independent, and the probability of an arrival occurring in a given interval is independent of past arrivals.

Definition 1.2.3: Independence and Memorylessness of Poisson Process

- **Independence:** For any time $t > 0$, the history of the process after t is a Poisson process, and is independent of the history before time t.
- **Memorylessness:** Let t be a given time and let $\overline{T}$ be the time of the first arrival after time t. THen, $\overline{T} - t$ has an exponential distribution with paramneter λ, and is independent of the history of the process before time t.

$$\mathbf{P}\big(\overline{T} - t > s\big) = \mathbf{P}(0, s) = e^{-\lambda s} = \mathbf{P}(T > s)$$

Intuition: *Consider a thought experiment to understand memorylessness. Suppose we are in line at a food truck, and there are three separate food trucks that each serve customers according to independent Poisson processes with identical rate parameters λ. Assume that all food trucks are initially busy and no one else is in front of you in line, then the probability that we will be the last to be served is $\frac{1}{3}$. This is because the moment we arrive at the food trucks, the processes essentially restart themselves due to memorylessness, and each food truck has an equal chance of being the last to finish serving.*

1.2.2 Interarrival Times and Arrival Times

We may now consider the k-th interarrival time T_k in a Poisson process with rate parameter λ. Since the time to the first event is exponentially distributed with parameter λ, amd the interarrival times are i.i.d. (due to independent), we have that:

$$T_k \sim \text{Exp}(\lambda)$$

The k-th arrival time Y_k is the same as before: the cumulative sum of the first k interarrival times:

$$Y_k = \sum_{i=1}^{k} T_i$$

The sum of k independent Exponential random variables with parameter λ is so important that we give it a name:

Definition 1.2.4: Erlang (Gamma) Distribution

- The PDF of Y_k is given by:

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad y \geq 0$$

  and is known as the **Erlang PDF of order** k (or Gamma PDF).

- The mean and variance of Y_k are given by:

$$\mathbf{E}[Y_k] = \mathbf{E}\left[\sum_{i=1}^{k} T_i\right] = \sum_{i=1}^{k} \mathbf{E}[T_i] = \frac{k}{\lambda}$$
$$\mathrm{Var}(Y_k) = \mathrm{Var}\left(\sum_{i=1}^{k} T_i\right) = \sum_{i=1}^{k} \mathrm{Var}(T_i) = \frac{k}{\lambda^2}$$

Proof. To obtain the distribution of Y_k, we find the CDF first:

$$\begin{aligned} F_{Y_k}(y) = \mathbf{P}(Y_k \le y) &= \mathbf{P}(\text{there are at least } k \text{ arrivals in time } y) \\ &= 1 - \mathbf{P}(\text{there are at most } k-1 \text{ arrivals in time } y) \\ &= 1 - \sum_{j=0}^{k-1} j, y \\ &= 1 - \sum_{j=0}^{k-1} e^{-\lambda y} \frac{(\lambda y)^j}{j!} \end{aligned}$$

Now we perform differentiation to find the PDF:

$$\begin{aligned} f_{Y_k}(y) = \frac{d}{dy} F_{Y_k}(y) &= \frac{d}{dy}\left(1 - \sum_{j=0}^{k-1} e^{-\lambda y} \frac{(\lambda y)^j}{j!}\right) \\ &= -\sum_{j=0}^{k-1}\left(-\lambda e^{-\lambda y} \frac{(\lambda y)^j}{j!} + e^{-\lambda y} \frac{\lambda^j y^{j-1} j}{j!}\right) \\ &= \lambda e^{-\lambda y}\left(\sum_{j=0}^{k-1} \frac{(\lambda y)^j}{j!} - \sum_{j=1}^{k-1} \frac{\lambda^{j-1} y^{j-1}}{(j-1)!}\right) \\ &= \lambda e^{-\lambda y} \frac{\lambda^{k-1} y^{k-1}}{(k-1)!} \\ &= \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!} \end{aligned}$$

As desired. □

1.2.3 Splitting and Merging Poisson Processes

Splitting a Poisson works the same way as splitting a Bernoulli process. Given a Poisson process with rate parameter λ, for every arrival, we keep it with probability p and discard it with probability $1-p$. The selections are made independently, and the result is a Poisson process with rate parameter $p\lambda$. The discarded arrivals also form a Poisson process with rate parameter $(1-p)\lambda$. Intuitively, this makes sense as we make each arrival ”harder” by a factor of p.

Merging is also similar. Given two independent Poisson processes with rate parameters λ_1 and λ_2, we include an arrival in the merged process if it occurs in either of the two original processes. The resulting merged process is a Poisson process with rate parameter $\lambda_1 + \lambda_2$.

1.3 Bernoulli and Poisson Processes, and Sums of Random Variables

The Bernoulli and Poisson Processes may be used to provide simple explainations for sums of independent random variables. We have the ffollowing properties:

Corollary 1.3.1

Let $N, X_1, X_2, \dots,$ be independent random variables, where N takes nonnegative integer values. Let $Y = X_1 + \cdots + X_N$ for positive values of N, and let $Y = 0$ when $N = 0$.

- If X_i is Bernoulli with parameter p, and N is binomial with parameters m and q, then Y is binomial with parameters m and pq.
- If X_i is Bernoulli with parameter p, and N is Poisson with parameter λ, then Y is Poisson with parameter λp.
- If X_i is geometric with parameter p, and N is geometic with parameter q, then Y is geometric with parameter pq.
- If X_i is exponential with parameter λ, and N is geometric with parameter q, then Y exponential with parameter λq.

1.4 The Random Incidence Paradox

The **Random Incidence Paradox** for Poisson processes states a counter-intuitive result about the interarrival-time lengths of arrivals. Consider a Poisson process with rate parameter λ, and fix a time t. Let S denote the length of the interarrival interval that contains time t. If we consider the interarrival times $\{T_i\}$ of the Poisson process, we know that each T_i is an independent Exponential random variable with parameter λ. Intuitively, one might expect that S also follows an Exponential distribution with parameter λ. However, this is not the case. In fact, S has a distribution of Erlang with order 2.

Consider the interval $[Y_{k-1}, Y_k]$ that contains time t. $S = Y_k - Y_{k-1}$. Moreover, we may express S as:

$$S = (t - Y_{k-1}) + (Y_k - t)$$

Both random variables $t - Y_{k-1}$ and $Y_k - t$ are independent and identically distributed expoenential random variables with parameter λ. This is due to the memorylessness property of the Poisson process. Thus, S is the sum of two independent Exponential random variables with parameter λ, which means that S follows an Erlang distribution with order 2 and rate parameter λ.

How can be this be??? If we just take any interarrival time T_i, it has an average length of $\frac{1}{\lambda}$. However, if we pick a random time t and look at the interarrival interval containing t, the average length of the interval is $\frac{2}{\lambda}$. This paradox arises because longer interarrival intervals are more likely to contain the randomly chosen time t. Thus, when we condition on the interval containing t, we are more likely to select longer intervals, leading to a longer average length.

Chapter 2

Markov Chains

The Bernoulli and Poisson processes were stochastic processes that were memorylessness, meaning that past events did not influence future events. However, many real-world processes do have memory. The opening price of a stock today may depenend on the closing price yesterday. To model these processes, we simplify events as being in one of a finite number of **states**, with transition probabilities between states being **time-invariant**.

2.1 Discrete-Time Markov Chains

Our first attempt at modeling these processes is the **Discrete-Time Markov Chain** (DTMC), where state updates occur at discrete time steps.

Definition 2.1.1: Discrete-Time Markov Chain (DTMC)

- At each time step n, the state of the chain is denoted as X_n, taking on values from a finite **state space** $S = \{1, 2, \ldots, m\}$.
- Transitions between states are governed by **transition probabilities** p_{ij}, which represent the probability of the chain moving from state i to state j in one time step:
$$p_{ij} = \mathbf{P}(X_{+1} = j \mid X_n = i)$$
- The transition probabilites moving out of a state must sum to 1:
$$\sum_{j \in S} p_{ij} = 1$$
- Markov chains satisfy the **Markov Property**, which states that the probability of moving from one state to the next does not depend on the history of previous states.
$$\begin{aligned}\mathbf{P}(X_{n+1}j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1) &= \mathbf{P}(X_{n+1} = j \mid X_n = i)\\ &= p_{ij}\end{aligned}$$

The transition probabilities of a DTMC may be represented in matrix form as the **transition matrix** P, where the element in the i-th row and j-th column is p_{ij}:

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,m} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m,1} & p_{m,2} & \cdots & p_{m,m} \end{bmatrix}$$

To construct a valid markov model, sometimes we need to introduce new states that encode more information about the history of the proces.

2.1.1 The Probability of a Path

The probability of a path occuring in a markov process is just the application of the product rule. Consider a path of states $i_0, i_1, \ldots, i_n$. The probaiblity of this path occuring is given by:

$$\mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \mathbf{P}(X_0 = i_0)\, p_{i_0,i_1} p_{i_1,i_2} \cdots p_{i_{n-1},i_n}$$

We can do this as it is a direct consequence of the Markov property:

$$\begin{aligned} &\mathbb{P}(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) \\ &\qquad = \mathbf{P}(X_n = i_n \mid X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1})\, \mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \\ &\qquad = p_{i_{n-1},i_n} \mathbf{P}(X_0 = i_0, X_1 = i_1, \ldots, X_{n-1} = i_{n-1}) \end{aligned}$$

This can be applied inductively to prove the path probability we have above.

2.1.2 n-Step Transition Probabilities

It's also possible to compute probabilities for future steps beyond one step away from our current state. This probability is denoted by the **n-step transition probabilities** (r_{ij}), which represents the probability that the state after n time-steps will be j, given that the current state is i:

$$r_{ij}(n) = \mathbf{P}(X_n = j \mid X_0 = i)\,.$$

We may apply recursion to solve for $r_{ij}(n)$, using the **Chapman-Kolmogorov equation**:

Theorem 2.1.2: Chapman-Kolmogorov Equation for the n-Step Transition Probabilities

The n-step transition probability is generated recursively as such:

$$r_{ij} = \sum_{k \in S} p_{k,j} r_{i,k}(n-1), \quad \text{for } n > 1, \text{ and all } i, j$$

where the base case is $r_{ij}(1) = p_{ij}$

Proof. This can be verified using total-probability:

$$\begin{aligned}\mathbf{P}(X_n = j \mid X_0 = i) &= \sum_{k \in S} \mathbf{P}(X_{n-1} = k \mid X_0 = i)\, \mathbf{P}(X_n = j \mid X_{n-1} = k, X_0 = i) \\ &= \sum_{k \in S} p_{kj} r_{ik}(n-1)\end{aligned}$$

□

In fact, r_{ij} for another two dimensional matrix called the **n-step transion probability matrix**, which gives us the n-step transition probabilities for all pairs of states.

2.2 Classification of States

States within a markov chain may have different properties depending on their possible transitions, classifying these states will allow us to analyze the long-term frequency of state visits.

A state j is **accessible** from state i if for some positive n, the n-step transition probability $r_{i,j}(n)$ is positive. Equivalently, this means that there exists a possible state sequence $i, i_1, \ldots, i_{n-1}, j$ that starts and i and ends at j, where all intermittent transitions have positive probability. We define:

$$A(i) = \{\text{the set of states that are accessible from } i\}$$

A state i is **recurrent** if for every j that is accessible from i, i is also accessible from j. That is: $\forall j \in A(i), i \in A(j)$. Now we see that if our process ever reaches some recurrent state i, we could only ever visit states that are accessible from i, so no matter how far we are in the future, there is always a probability that we will return to i (which will happen given enough time). For an infinitely long Markov Process, if a recurrent state is visited once, it is guaranteed to be revisited an infinite number of times.
If a state i is not recurrent, it's **transient**. Hence, state i is transient if there exists some state $j \in A(i)$ such that $i \notin A(j)$. If a processes is in state i, then given enough time, it will eventually reach a state j such that $i \notin A(j)$. When this happens, there is a 0 probability of ever visiting i again. Hence, any transient state will only be visited a finite-number of times.
If i happens to be a recurrent state, then $A(i)$, the set of states accessible i, forms a **recurrent class**. Within a recurrent class, all states are accessible from each other, but no states outside of the class are accesssible. We can see that logically, for a set of states, if any one of the states were recurrent, the entire set must form a recurrent class, as otherwise, there would exist some transient state that would contradict the recurrency of our recurrent state.

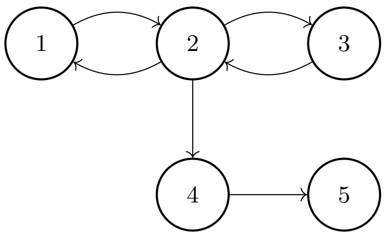


Figure 2.1: In this markov chain, states 1, 2, and 3 form a recurrent class, as they are all accessible from each other. States 4 and 5 are transient states, as once the process reaches state 4, there is no way to return to states 1, 2, or 3.

Similarly, it could also be logically argued that at least one recurrent state must be accessible from any transient state. Suppose that there were no recurrent states that are accessible from some transient state. Then that means that all states that are accessible from the transient state must also be transient, and thus can only be visited a finite number of times. Since there are only a finite set of states, and each transient state can only be visited a finite number of times, this means that starting from the transient state, there is a 0 probability of the process continuing infinitely, which is a contradiction. Hence, at least one recurrent state must be accessible from any transient state. This gives us enough information to decompose a markov chain:

Theorem 2.2.1: Markov Chain Decomposition

- A Markov chain may be decomposed into one or more recurrent classes, and possibly some transient states.
- A recurrent state is accessible from all states in its class, but not accessible from recurrent states in other classes
- A transient state is not accessible from any recurrent state.
- At least one, possibly more, recurrent states are accessible from any transient state.

Decompositions are useful because they allow us to analyze the long-term behavior of Markov processes:

- If a Markov run enters or starts in a recurrent state, it always stays in the recurrent class of that state.
- If a Markov run starts in a transient state, the state trajectory could possibly visit a finite amount of transient states before entering a recurrent class, after which it will stay in that recurrent class forever.

Hence, to fully understand the markov process, we need to consider both long-term behavior within recurrent classes, as well as short-term behavior when entering recurrent classes through transient states.

2.2.1 Periodicity

A recurrent class has an important property known as **periodicity**, which measures the degree of cyclic behavior within the class. A recurrent class is **periodic** if the states can be grouped into $d > 1$ disjoint subsets $S_1, \dots, S_d$ where all transitions from one subset lead to the next subset:

$$\text{if } i \in S_k \text{ and } p_{ij} > 0. \quad \text{then } \begin{cases} j \in S_{k+1}, & \text{if } k < d \\ j \in S_1, & \text{if } k = d. \end{cases}$$

To unpack this definition, we can think of pipelining the recurrent class into d stages, where each stage has some collection of states. From any state in stage k, the only possible transitions are to stages in $k+1$ (or back to stage 1 if $k = d$). This means that no matter how our state-trajectory looks, it can always be grouped into sections of length d, where each section exactly one state from each stage. Thus, if we start in some state i in stage S_k, we can only return to i after a multiple of d time-steps. Hence, the period of state i is d.

A easier method to compute the period of a state is to find the greatest common divisor of the lengths of all paths that start and end at that state. In this case, it helps to have a state transition diagram to visualize all possible paths. If the period of a state is 1, then the state is **aperiodic**. A recurrent class is aperiodic if all states in the class are aperiodic. An ecen stronger statement states that as long as one state in the class has period 1, then the entire class is aperiodic.

Notice that for a periodic recurrent class, a time n, and a state i in the class, there must exist one or more states j such that $r_{ij}(n) = 0$. This is because i belongs in some state S_k, and can only transition to states in $S_{k+n \mod d}$ in n time steps. The converse of this statement gives us a method to check for aperiodicity:

Theorem 2.2.2: Aperiodicity Check

A recurrent class is aperiodic if and only if there exists some time n such that for all states i, j in the class, $r_{ij}(n) > 0$.
Check if there exists some time n and state i such that $r_{ij}(n) > 0$ for all states j in the class.

2.3 Steady-State Behavior

One of our main two goals of analyzing markov chains is to understand their long-term behavior. A common question to ask is: *After a long time, what fraction of time will the markov chain spend in each state?*. Does the markov chain ever converge to a steady-state distribution over the states, regardless of the initial state? We first analyze the behavior of a single recurrent class, along with some transient states. Our results from a single recurrent class would be easily extendable to multiple recurrent classes.
Before analyzing the steady-state behavior, we must first define when such a behavior actually converges for a markov chain. A markov chain is **irreducible** if it consists of a

single recurrent class, meaning that all states are accessible from each other. An irreducible markov chain is **ergodic** if it is aperiodic as well. A non-ergodic markov chain may not converge to a steady-state distribution, as periodic behavior could result in oscillations between states. However, we assert that an ergodic markov chain will always converge to a unique steady-state distribution independent of the initial state i. We introduce the concept of the **stationary distribution**:

Definition 2.3.1: Stationary Distribution π

The stationary distribution π of a markov chain is a probability distribution over the states that describes the long-term fraction of time spent in each state (Or equivalently, the probability of being in each state after a long time):

$$\pi = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \end{bmatrix}$$

where π_j is the **steady-state probability of** j, which is thelong-term fraction of time spent in state j:

$$\pi_j \approx \mathbf{P}(X_n = j) \text{ for large } n$$

Theorem 2.3.2: Steady-State Convergence Theorem

Consider an ergodic markov chain. Then the states converge to a unique stationary distribution π which satisfies the following properties:

- For all states $j \in S$,

$$\lim_{n\to\infty} r_{ij}(n) = \pi_j, \quad \text{for all } i \in S$$

- The stationary distribution satisfies the **balance equations**:

$$\pi = \pi P$$
$$\sum_{j\in S} \pi_j = 1$$

- We have

$$\pi_j = 0, \quad \text{for all transient states } j$$
$$\pi_j > 0, \quad \text{for all recurrent states } j$$

We can see that the balance equations are actally a consequence of the Chapman-Kolmogorov equations. To solve the stationary distribution, we solve the system formed by the balance equations and the normalization condition:

$$\pi_i = \sum_{j\in S} \pi_j p_{ji}, \quad \text{for all } i \in S$$
$$\sum_{j\in S} \pi_j = 1$$

2.3.1 Long-Term Frequency Interpretations

The steady-state probability π_j of state j has an important interpretation related to long-term frequency: π_j is the long-term fraction of time spent in state j.

Theorem 2.3.3: Steay-State Probabilities as Expected State Frequencies

For an ergodic Markov chain, the steady state probabilities π_j satisfy:

$$\pi_j = \lim_{n\to\infty} \frac{v_{i,j}(n)}{(n)}$$

where $v_{i,j}(n)$ is the expected value of the number of visits to state j within the first n transitions, given that the chain starts in state i.

An additional interpretations arises from the from the balance equations. We know that $\pi_j = \sum_{i\in S} \pi_i p_{i,j}$. This means that the stable-state probability describes the sum of the expected frequencies of arrivals to j. In other words, the expected frequency of being in j is equal to the total incoming frequency flow from all other states.

2.3.2 Birth-Death Processes

We csondier a special class of markov chains known as **Birth-Death Processes**. In these processes, the states are linearly ordered as $S = \{0, 1, 2, \dots, \}$, and transitions are only allowed between neighboring states. From state i, we may either a **birth-process**, where state i transitions to $i+1$ with probability b_i, or a **death-process**, where state i transitions to $i-1$ with probability d_i. The remaining probability $1 - b_i - d_i$ is the probability of staying in state i.

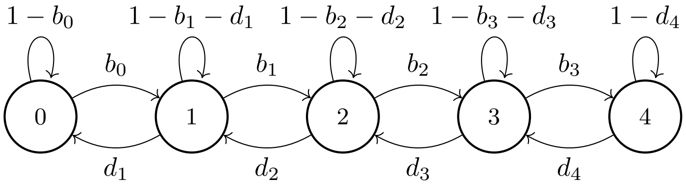


Figure 2.2: A birth-death process with states 0 through 4. From each state i, the process may either transition to state $i+1$ with probability b_i, or to state $i-1$ with probability d_i.

The balance equations for a birth-death process simplify nicely due to the linear structure, as each state i only has incoming transitions from states $i-1$ and $i+1$. Thus, the balance equations become:

$$\begin{aligned}
\pi_0 &= \pi_0(1 - b_0) + \pi_1 d_1 \\
\pi_m &= \pi_m(1 - d_0) + \pi_{m-1} b_{m-1} \\
\pi_i &= \pi_{i-1} b_{i-1} + \pi_i(1 - b_i - d_i) + \pi_{i+1} d_{i+1}, \quad i \geq 1
\end{aligned}$$

Now note, for two neighboring states i and $i+1$, a transition from $i \to i+1$ must be followed by a transition from $i+1 \to i$ before another transition from $i \to i+1$ can occur. Hence, in the long-term, the expected frequency of transitions from i to $i+1$ must equal the expected frequency of transitions from $i+1$ to i. This provides the local balance equations:

$$\pi_i b_i = \pi_{i+1} d_{i+1}, \quad i \geq 0$$

Now we analyze this for increasing values of i:

$$\begin{aligned}
\pi_0 b_0 = \pi_1 d_1 &\implies \pi_1 = \frac{b_0}{d_1}\pi_0 \\
\pi_1 b_1 = \pi_2 d_2 &\implies \pi_2 = \frac{b_1}{d_2}\pi_1 = \frac{b_0 b_1}{d_1 d_2}\pi_0 \\
\pi_2 b_2 = \pi_3 d_3 &\implies \pi_3 = \frac{b_0 b_1 b_2}{d_1 d_2 d_3}\pi_0 \\
&\vdots \\
& \pi_n = \frac{b_0 b_1 \cdots b_{n-1}}{d_1 d_2 \cdots d_n}\pi_0
\end{aligned}$$

We see that generally:

$$\pi_i = \frac{\prod_{k=0}^{i-1} b_k}{\prod_{k=1}^{i} d_k}\pi_0, \quad i \geq 1$$

Combined with the normalization condition, we have:

$$\pi_0 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} b_k}{\prod_{k=1}^{i} d_k}\pi_0 = 1 \implies \pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} b_k}{\prod_{k=1}^{i} d_k}}$$

And we may use this to solve for all steady-state probabilities π_i.

2.4 Absorption Probabilities and Expected Time to Absorption

Having studied the long-term behavior of ergodic markov chains, we now turn our attention to analyzing the short-term behavior of markov chains with transient states. Our goal is to consider the probability of entering a particular recurrent class, as well as the expected time to enter it. Knowing this information will allow us to generalize the steady-state behavior of markov chains with transient states with multiple recurrent classes.

One a chain enters a recurrent state, it will remain in that state forever, so we define a recurrent state k as **absorbing**. With only one unique absorbing state, all other states that are transient will have steady-state probability 0 and eventually lead into state k. With multiple absorbing states, the probability that any will be reached is 1, but the probability of reaching a particular absorbing state depends on the starting transient state. To address this, we define the **absoption probability** $a_{i,k}$ as the probability that starting from transient state i, the markov chain will eventually be absorved into absorbing state k.

Definition 2.4.1: Absorption Probability $a_{i,k}$

Take a markov chain with multiple transient and absorbing states. The absorption probability of absorbing state k starting from transient state i is:

$$\begin{aligned}
a_{k,k} &= 1, \\
a_{i,k} &= \sum_{j\in S} p_{i,j}a_{j,k}, \quad \text{for all transient states } i \\
a_{i,k} &= 0, \quad \text{for all absorbing states } i \neq k
\end{aligned}$$

We may now calculate the probability of entering a given recurrent class. To do this, we take all states within a recurrent class and combine them into a single absorbing state. We can then calculate the absorption probabilities for this absorbing state from all transient states. Note that depending on the initial transient state we start at, the absorption probability into a particular recurrent class may vary.

Similarlym we may also calculate the **expected time to absorption** starting from a particular transient state i. This is defined as μ_i:

$$\begin{aligned}
\mu_i &= \mathbf{E}[\text{Number of transitions until absorption} \mid X_0 = i] \\
&= \mathbf{E}[\min\{n \geq 0 \mid X_n \text{ is absorbing}\} \mid X_0 = i]
\end{aligned}$$

Equations for calculating μ_i may be derived using total expectation. Note that the expected time to absorption from a state i is equal to 1 plus the time to absorbtion from the next state j, with probability p_{ij}. Hence, we have:

$$\begin{aligned}
\mu_i &= 0, \quad \text{if } i \text{ is an sbsorbing state} \\
\mu_i &= 1 + \sum_{i\in S} p_{i,j}\mu_j, \quad \text{if } i \text{ is a transient state}
\end{aligned}$$

2.4.1 Mean First Passage and Recurrence Times

Calculating the expected time to absorption is a problem that can be generalized to calculate the expected time to reach any recurrent state from any other state. Consider a markov chain with a single recurrent class. We define this as $\mu(i, s)$ where i is the starting state and s is the target recurrent state. Similar to before:

$$\begin{aligned}
\mu(i,s) &= \mathbf{E}[\text{Number of transitions to reach } s \text{ for the first time} \mid X_0 = i]] \\
&= \mathbf{E}[\min\{n \geq 0 \mid X_n = s\} \mid X_0 = i]
\end{aligned}$$

Since we only care about the transitions until we reach s for the first time, we modify the markov chain by making state s absorbing, that is, removing all outgoing transitions from s. This effectively turns the other states into transient states, and now the expected time to reach s from i is equivalent to the expected time to absorbtion from i in the modified chain. This is called the **Mean First Passage Time**.

A special case of the mean first passage time is the **Mean Recurrence Time**, which is the expected time to return to a recurrent state s starting from s itself. We denote the mean recurrence time is denoted as $\mu^*(s)$. This may be calculated with the mean first passage times:

$$\mu^*(s) = 1 + \sum_{j \in S} p(s, j)\mu(j, s)$$

2.5 Continuous-Time Markov Chains

Discrete Markov Chains are restricted to state transitions occuring at discrete time steps, which rarely occur in real life. To model processes that change in continuous time, we have to incorporate continuous interarrival times between state transitions. This is what the **Continuous-Time Markov Chain** (CTMC) does. We are concerned with three random variables:

X_n : the state after the nth transition

Y_n : the time of the nth transition

T_n : the time elapsed between the $(n-1)$st and the nth transition

The initial state is X_0, and similarly, $Y_0 = 0$. We make the following assumptions about the CTMC:

Definition 2.5.1: Continuous-Time Markov Chain (CTMC)

- If the current state is i, the time unitl the next transition is expoentially distributed with a given rate parameter ν_i independent of the past history of the process and of the next state.
- If the current state is i, the next state will be j with a given probability p_{ij}, independent of the past history of the process and of the time until the next transition.

The CTMC is no different than the DTMC, except that we introduce exponential interarrival times between state transitions. Hence, we if disregard the interarrival times and just focus on the sequence of states $\{X_n\}$, the sequence forms a DTMC with transition probabilies p_{ij}, called the **embedded DTMC**. A state trajectory for a CTMC may be represented as:

$$A = \{T_1 = t_1, \dots, T_n = t_n, X_0 = i_0, X_1 = i_1, \dots, X_n = i_n\}$$

Hence, the probability of the next transition occuring after time t and going to state j is given by:

$$\begin{aligned} \mathbf{P}(T_{n+1} > t, X_{n+1} = j \mid A) &= \mathbf{P}(T_{n+1} > t, X_{n+1} = j \mid X_n = i_n) \\ &= \mathbf{P}(T_{n+1} > t \mid X_n = i_n)\, \mathbf{P}(X_{n+1} = j \mid X_n = i_n) \\ &\quad \text{Use CDF of Exp} \\ &= e^{-\nu_{i_n} t} p_{i_n j} \end{aligned}$$

We may also calculate the expected time to the next transition given a current state i:

$$\mathbf{E}[T_{n+1} \mid X_n = i] = \frac{1}{\nu_i}$$

Observe that the time until the next transition is independent of the next state, as we choose both separately. Moreover, ν_i serves as the transition rate out of state i, as it denotes the average number of transitions out of state i per unit time. Combining ν_i and p_{ij}, we define the **transition rate from i to j, q_{ij}** as:

$$q_{ij} = \nu_i p_{ij} \Longleftrightarrow p_{ij} = \frac{q_{ij}}{\nu_i}$$

Due to the normalization condition on p_{ij}, $\nu_i = \sum_{j \in S} q_{ij}$. Note that also, we may remove self-cycles in a CTMC as transitioning from a state back to itself does not alter the trajectory, and due to the memorylessness of the exponential distribution, the time until the next transition remains exponentially distributed with rate ν_i. Now for q_{ii}, we define it as $q_{ii} = -\nu_i$. The reason for this is that the total rate of leaving state i must be equal to the negative of the rate of staying in state i. Using these values, we can create a **generator matrix** that parallels the transition matrix for a DTMC.

$$Q = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & q_{mm} \end{bmatrix} = \begin{bmatrix} -\nu_1 & \nu_1 p_{12} & \cdots & \nu_1 p_{1m} \\ \nu_2 p_{21} & -\nu_2 & \cdots & \nu_2 p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \nu_m p_{m1} & \nu_m p_{m2} & \cdots & -\nu_m \end{bmatrix}$$

This matrix will prove useful later when we solve for the stationary distribution. Note that because the interarrival times of the CTMC are independent, the CTMC also satisfies the Markov property.

2.5.1 Approximation of CTMC with DTMC

A CTMC may be approximated with a DTMC with sufficiently small time steps. Let the time-steps be a small positive δ and consider the DTMC Z_n made by observing a CTMC $X(t)$ every δ time units. Since the Markov-property is satisfied for X, it is also satisfied for Z_n. Let $\overline{p_{ij}}$ denote the transition probabilities of Z_n. To compute $\overline{p_{ij}}$ for $j \neq i$, we consider several factors. First, the probability of there being a transition between $n\delta$ and $(n+1)\delta$ is approximately $\nu_i \delta$ for small δ. Given that a transition occurs, the probability that the next state is j is p_{ij}. Hence, the probability of transitioning from i to j in one time-step is approximately:

$$\overline{p_{ij}} = \mathbb{P}(Z_{n+1} = j \mid Z_n = i) = \nu_i \delta p_{ij} + o(\delta) = q_{ij}\delta + o(\delta)$$

The $o(\delta)$ term accounts for the possibility of multiple transitions occuring within the single time-step of length δ, which has probabilty on the order of δ^2 of happening. Hence, as $\delta \to 0$, the $o(\delta)$ is vanishing. Finally, the probability of remaining at state i is:

$$\overline{p_{ii}} = \mathbf{P}(Z_{n+1} = i \mid Z_n = i) = 1 - \sum_{j \neq i} \overline{p_{ij}}$$

Hence, we have approximated the CTMC with a DTMCS with the following transition probabilities:

$$\overline{p_{ij}} = \begin{cases} q_{ij}\delta + o(\delta), & j \neq i \\ 1 - \sum_{j \neq i} q_{ij}\delta + o(\delta), & j = i \end{cases}$$

This is familiar to approximating a Poisson process with a Bernoulli process, where the rate parameter λ of the Poisson process is analogous to the transition rate ν_i of the CTMC.

2.5.2 Steady-State Behavior of CTMC

Now that we have drawn parallels between the CTMC and the DTMC, we may analyze the steady-state behavior of the CTMC using a corresponding DTMC. Consider a CTMC $X(t)$, define a DTMC Z_n as $Z_n = X(n\delta)$ for some small $\delta > 0$. Intuitively, we see that the stable distribution of Z_n should approximate the stable distribution of $X(t)$ as $\delta \to 0$. Let π denote the stationary distribution of Z_n. Note that Z_n is also automatically aperiodic. This is because for any state i, there is a non-negative self-transition probability as:

$$\overline{p_{ii}} = 1 - \delta \sum_{j \neq i} q_{ij} + o(\delta) > 0$$

From before, we know that the balance conditions for Z_n are:

$$\pi_j = \sum_{i \in S} \pi_i \overline{p_{ij}}, \quad \text{for all } j \in S$$

We can expand this by bring out the self-transition term to get:

$$\begin{aligned} \pi_j &= \pi_j \overline{p_{jj}} + \sum_{k \neq j} \pi_k \overline{p_{kj}} \\ &= \pi_j \left(1 - \delta \sum_{k \neq j} q_{kj} + o(\delta) \right) + \sum_{k \neq j} \pi_k (q_{kj}\delta + o(\delta)) \end{aligned}$$

Taking the limit of $\delta \to 0$, and rearranging, we have the **balance equations for CTMC**:

$$\underbrace{\pi_j \sum_{k \neq j} q_{kj}}_{\text{outflow}} = \underbrace{\sum_{k \neq j} \pi_k q_{jk}}_{\text{inflow}}$$

Theorem 2.5.2: CTMC Steady-State Convergence

Consider an ergodic CTMC. The states j concerge to a unique steady-state distribution π that satisfies the following properties:

- For each j, we have:

$$\lim_{t\to\infty} \mathbf{P}(X(t) = j \mid X(0) = i) = \pi_j, \quad \text{for all } i$$

- The π_j are the unique solution to the system of equations below:

$$\begin{aligned} \pi_j \sum_{k\neq j} q_{jk} &= \sum_{k\neq j} \pi_k q_{kj} \\ 1 &= \sum_{k=1}^{m} \pi_k \end{aligned}$$

- The steady state probabilities for the transient states are 0, while those for recurrent states are positive.

The balance equations for CTMC have a nice interpretation. The left-hand side represents the total inflow rate into state j from all other states, while the right-hand side represents the total outflow rate from state j to all other states. In steady-state, the inflow and outflow rates must be equal for each state. This could be understood as the fact that the frequency of transitions out of a state j must be equal to the frequency of transitions into state j. If the two flows were not equal, then the probability of being in state j would either increase or decrease over time, contradicting the assumption of steady-state.

In fact, this nice interpretation allows us to solve for the steady-state probabilities in a similar manner to before. Recall the generator matrix Q of the CTMC. The balance equations may be rewritten in matrix form as:

$$\pi Q = 0$$

To see this, we may take the our current balance equations and rearrange them:

$$\begin{aligned} \pi_j \sum_{k\neq j} q_{jk} &= \sum_{k\neq j} \pi_k q_{kj} \\ \sum_{k\neq j} \pi_k q_{kj} - \pi_j \sum_{k\neq j} q_{jk} &= 0 \\ \text{Note that } \sum_{k\neq j} q_{jk} &= -q_{jj} \\ \sum_{k\neq j} \pi_k q_{kj} + \pi_j q_{jj} &= 0 \\ \sum_{k\in S} \pi_k q_{kj} &= 0 \\ (\pi Q)_j &= 0 \end{aligned}$$

Hence, we have $\pi Q = 0$. Combined with the normalization condition, we may solve for the steady-state probabilities of the CTMC.

2.5.3 CTMC Birth-Death Processes

For a CTMC birth-death process, the balance equations simplify similarly to the discrete case. Moreover, we note that for each state i, the total inflow rate is just $\pi_{i-1}q_{i-1,i}$, while the total outflow rate is $\pi_i q_{i,i+1}$. Hence, in the steady-state, we have the **local balance equations**:

$$\pi_j q_{ji} = \pi_i q_{ij}$$

From here, we may solve for the steady-state probabilities similarly to before:

$$\begin{aligned} \pi_1 &= \frac{q_{01}}{q_{10}}\pi_0 \\ \pi_2 &= \frac{q_{01}q_{12}}{q_{10}q_{21}}\pi_0 \\ &\vdots \\ \pi_n &= \frac{\prod_{k=0}^{n-1} q_{k,k+1}}{\prod_{k=1}^{n} q_{k,k-1}}\pi_0 \end{aligned}$$

Combined with the normalization condition, we have:

$$\pi_0 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} q_{k,k+1}}{\prod_{k=1}^{i} q_{k,k-1}}\pi_0 = 1 \implies \pi_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \frac{\prod_{k=0}^{i-1} q_{k,k+1}}{\prod_{k=1}^{i} q_{k,k-1}}}$$

2.6 Reversible Processes

A special class of Markov Chains are **reversible processes**, which have the property that the process looks the same when time is reversed. Formally, a CTMC is reversible if for all states i, j:

$$\pi_j p_{ji} = \pi_i p_{ij}$$

This condition is known as the **detailed balance equations**, and is a stronger condition than the regular balance equations. Reversible processes have the nice property that their steady-state distributions may be solved using the detailed balance equations, which are often easier to work with. To show that the detailed balance equations satisfy the regular balance equations, sum both sides over all states j:

$$\sum_{j \in S} \pi_j p_{ji} = \sum_{j \in S} \pi_i p_{ij} = \pi_i$$

Hence:

$$\pi_i = \sum_{j \in S} \pi_j p_{ji}$$

Repeated for all i, we see that this is exactly $\pi = \pi P$, satisfying the balance equations.

Chapter 3

Bayesian Statistical Inference

Interence is the process of drawing conclusions about an unknown distribution based on observed data. In inference, we typically denote parameters or the unknown distribution as Θ, and the observed data as X.What inference allows us to do is to estimate the parameters of an unknown distribution, test hypothesis about the distribution, and make predictions. There are two schools of thought for statistical inference:

- **Frequentist Inference**: The unknown distribution is constant but unknown. We draw conclusions about the distribution based on the data.
- **Bayesian Inference**: The unknown distribution is a random variable with a **prior** distribution ($p_\Theta(\theta)$). We draw conclusions about the distribution based on posterior distribution $\mathbb{P}_{\Theta|X}(\theta \mid x)$ derived from the data and the prior distribution.

there exists a plethora of problems and methods related to Bayesian inference:

Fact 3.0.1

- **Bayesian statistics** assumes parameters are random variables with known priors.
- **Parameter estimation** involves estimating the parameters of an unknown distribution based on observed data.
- **Hypothesis testing** involves an unknown parameter with finite possible values (hypotheses), and we wish to determine which hypothesis is true (causes least error) based on observed data.
- Primary Bayesian Inference methods include:
  - **Maximum a posteriori probability (MAP):** Out of possible parameter values, select the parameter with the maximum posterior probability given the data:

$$\mathbf{P}(\Theta = \theta \mid X = x) \propto \mathbf{P}(X = x \mid \Theta = \theta)\,\mathbf{P}(\Theta = \theta)$$

  - **Least mean squares (LMS):** Select an estimator of the data that minimizes the mean squared error between the parameter and estimate.

- **Linear least mean squares:** Pick a estimator that is a linear function of the data and minimizes the mean squared error between the parameter and its estimate. Less accurate than LMS, but easier to compute.

3.1 Bayesian Inference and the Posterior Distribution

For Bayesian Inference, we may several assumptions that we know the joint distribution of the data X and parameters Θ. We also assume that we know the prior distribution of the parameters $\mathbf{P}(\Theta = \theta)$ and the conditional distribution of the data given the parameters $\mathbf{P}(X = x \mid \Theta = \theta)$. With this information, we rely on Baye's rule to compute the posterior distribution of the parameters given the data:

Theorem 3.1.1: Bayes' Rule for Bayesian Inference

Given observed data $X = x$, the posterior distribution of the parameters Θ is given by:

$$\mathbf{P}(\Theta = \theta \mid X = x) = \frac{\mathbf{P}(X = x \mid \Theta = \theta)\,\mathbf{P}(\Theta = \theta)}{\sum_{\theta'} \mathbf{P}(X = x \mid \Theta = \theta')\,\mathbf{P}(\Theta = \theta')}$$

In most cases, the denominator is a normalizing constant because it does not depend on θ, so it's often enough to compute the unnormalized posterior:

$$\mathbf{P}(\Theta = \theta \mid X = x) \propto \mathbf{P}(X = x \mid \Theta = \theta)\,\mathbf{P}(\Theta = \theta)$$

Note that practically, the prior distribution $\mathbf{P}(\Theta = \theta)$ represents our beliefs about the parameters before observing any data, while the posterior distribution $\mathbf{P}(\Theta = \theta \mid X = x)$ represents our updated beliefs after observing data $X = x$. The likelihood function $\mathbf{P}(X = x \mid \Theta = \theta)$ quantifies how likely the observed data is given a particular parameter value. By combining the prior and likelihood using Bayes' rule, we obtain the posterior distribution, which reflects both our prior beliefs and the evidence provided by the data. This means that the prior distribution may be incorrect, but as we observe more data, the posterior distribution will converge to the true parameter values.

The data and parmeters may be either discrete or continuous random variables, so we substitute with the apprpriate PMFs or PDFs as needed. However, should our parameter be continuous, the summation in the denominator becomes an integral.

A common distribution used for priors are the **conjugate priors**, which are priors that result in posterior distributions of the same family. The nice thing about conjugate priors is that they simplify the process of calculating the posterior distribution. This because:

- The posterior distribution is in the same family as the prior distribution.
- The parameters of the posterior distribution can be easily updated based on new data. This is because the posterior distribution can be expressed in terms of the prior parameters and the observed data.

3.2 Point Estimation, Hypothesis Testing, and MAP

The **Maximum a Posteriori probability (MAP)** rule is one example of a Bayesian inference method. Given observed data $X = x$, the MAP estimate of the parameters Θ is the value of θ that maximizes the posterior distribution:

$$\hat{\theta} = \text{argmax}_\theta \, p_{\Theta|X}(\theta \mid x), \quad (\Theta \text{ discrete})$$
$$\hat{\theta} = \text{argmax}_\theta \, f_{\Theta|X}(\theta \mid x), \quad (\Theta \text{ continuous})$$

Definition 3.2.1: MAP Rule

Given the observation value x, MAP selects $\hat{\theta}$ such that it maximizes the posterior distribution $p_{\Theta|X}(\theta \mid x)$ if Θ is discrete, or $f_{\Theta|X}(\theta \mid x)$ if Θ is continuous.
If Θ has a discrete distribution, then the MAP rule minimizes the probaiblity of selecting an incorrect hypothesis. This is because the posterior distribution gives the probability of each hypothesis being true given the observed data, so selecting the hypothesis with the highest posterior probability minimizes the chance of error.

3.2.1 Point Estimation

The MAP rule may be used for **point estimation**, which is when given a set of data about our desired distribution, we create a single ”best guess” estimate of Θ. In point estimation, we have an **estimate** $\hat{\theta}$, and an **estimator** function $\hat{\Theta}(X) = g(X)$ that maps the data to the estimate. The estimator is a random variable as it depends on the random observation X.

Definition 3.2.2: Point Estimators

- **Maximum a Posteriori Probability (MAP) Estimator:** Given observed data $X = x$, the MAP estimator picks $\hat{\theta}$ that maximizes the posterior distribution.
- **Least Mean Squares (LMS) Estimator:** Given observed data $X = x$, the LMS estimator picks $\hat{=}\mathbf{E}[[]\,\Theta \mid X = x]$ that minimizes the mean squared error betwee the parameter and estimate.

3.2.2 Hypothesis Testing

Hypothesis testing is a special case of point estimation where the parameter Θ can only take on a finite set of values, each representing a different hypothesis. The goal is to select the hypothesis that is most likely to be true based on the observed data. The MAP rule is particularly useful for hypothesis testing, as it selects the hypothesis with the highest posterior probability given the data, thereby minimizing the probability of making an incorrect decision.
In hypothesis testing, Θ can take on up to m values $\theta_1, \ldots, \theta_m$, each representing a different hypothesis $H_1, \ldots, H_m$. Once we find the posterior distribution for each hypothesis, we can

caluclate the overall probability of being right by summing the posterior probabilities of the selected hypotheses. The MAP rule then selects the hypothesis with the highest posterior probability.

Definition 3.2.3: Hypothesis Testing with MAP

Given observed data $X = x$, the MAP hypothesis testing rule selects hypothesis H_k where:

$$k = \text{argmax}_{1 \leq i \leq m} \mathbf{P}(\Theta = \theta_i \mid X = x)$$

This minimizes the probability of selecting an incorrect hypothesis.

3.3 Bayesian Least Mean Squares Estimation (LMS)

Another method for Bayesian inference is the **Least Mean Squares (LMS)** estimator, also known as the Conditional Expectation estimator. The LMS estimator aims to minimize the mean squared error (MSE) between the true parameter value and the estimate, defined as:

Definition 3.3.1: Mean Squared Error (MSE)

The mean squared error between the parameter Θ and its estimate $\hat{\Theta}$ is:

$$\text{MSE} = \mathbf{E}\Big[(\Theta - \hat{\Theta})^2\Big]$$

We consider a simplified case where there is no observation. We may treat the error $(\Theta - \hat{\theta})^2$ as a random variable, which can be minimized by taking the derivative with respect to $\hat{\theta}$ and setting it to 0:

$$\begin{aligned}\frac{d}{d\hat{\theta}}\mathbf{E}\Big[(\Theta - \hat{\theta})^2\Big] &= \frac{d}{d\hat{\theta}}\sum_{\theta}(\theta - \hat{\theta})^2\mathbf{P}(\Theta = \theta) \\ &= \sum_{\theta} 2(\hat{\theta} - \theta)\mathbf{P}(\Theta = \theta) = 0\end{aligned}$$

Now expanding and rearranging, we have:

$$\begin{aligned}\sum_{\theta} 2(\hat{\theta} - \theta)\mathbf{P}(\Theta = \theta) = 2\hat{\theta}\sum_{\theta}\mathbf{P}(\Theta = \theta) - 2\sum_{\theta}\theta\mathbf{P}(\Theta = \theta) = 0 \\ \implies \hat{\theta} = \mathbf{E}[\Theta]\end{aligned}$$

This same intuition carries over when we condition Θ on observed data $X = x$, in which the LMS estimator becomes:

$$\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$$

In fact, the LMS estimator minimizes the mean squared error between the parameter and its estimate for both discrete and continuous random variables. This is because for both cases, the mean squared error is a quadratic function of $\hat{\theta}$, and taking the derivative and setting it to 0 yields the same result.

Definition 3.3.2: LMS Estimator

Given observed data $X = x$, the LMS estimator picks $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$ that minimizes the mean squared error between the parameter and estimate.

3.3.1 Estimation Error

We commonly denote the LMS estmiator as $\hat{\Theta} = \mathbf{E}[\Theta \mid X]$, and the error as $\tilde{\Theta} = \hat{\Theta} - \Theta$. There are several properties of the estimation error:

Theorem 3.3.3: Properties of the Estimation Error

- The estimation error $\tilde{\Theta}$ is **unbiased**, so it has 0 mean:

$$\mathbf{E}\left[\tilde{T}\right] = 0, \quad \mathbf{E}\left[\tilde{\Theta} \mid X = x\right] = 0$$

- The estimation error $\tilde{\Theta}$ is **uncorrelated** with the estimate $\hat{\Theta}$

$$\text{Cov}(\tilde{\Theta}, \hat{}) = 0$$

- The variance of Θ can be decomposed as:

$$\text{Var}(\Theta) = \text{Var}\left(\hat{\Theta}\right) + \text{Var}\left(\tilde{\Theta}\right)$$

3.4 Bayesian Linear Least Mean Squares Estimation (LLSE)

While the LMS estimation results in a low mean squared error, it is rarely pratical or feasible to compute in real-life, especially with an unknown distribution. Hence, require a class of estimators that are restricted to linear functions. These estimators may result in higher mean squared error, but they are a lot more practical to compute. A linear estimator of a random variable Θ based on observed data X is of the form:

$$\hat{\Theta} = aX + b$$

Hence, the mean squared error of the linear estimator is:

$$\mathbf{E}\left[(\Theta - (aX + b))^2\right]$$

We are interested in finding the values of a and b that minimizes the squared error. We repeat the same process as before, taking the derivative with respect to a and b and setting them to 0. We first expad the expression to get:

$$\begin{aligned}
\mathbf{E}\left[(\Theta - (aX + b))^2\right] &= \mathbf{E}\left[\Theta^2 - 2a\Theta X - 2b\Theta + a^2X^2 + 2abX + b^2\right] \\
&= \mathbf{E}\left[\Theta^2\right] - 2a\mathbf{E}[\Theta X] - 2b\mathbf{E}[\Theta] + a^2\mathbf{E}\left[X^2\right] + 2ab\mathbf{E}[X] + b^2
\end{aligned}$$

Now we have:

$$\frac{\partial}{\partial a}\mathbf{E}\left[(\Theta - (aX+b))^2\right] = -2\mathbf{E}[\Theta X] + 2a\mathbf{E}\left[X^2\right] + 2b\mathbf{E}[X] = 0$$

$$a = \frac{\mathbf{E}[\Theta X] - b\mathbf{E}[X]}{\mathbf{E}[X^2]}$$

Simlarly, we have:

$$\frac{\partial}{\partial b}\mathbf{E}\left[(\Theta - (aX+b))^2\right] = -2\mathbf{E}[\Theta] + 2a\mathbf{E}[X] + 2b = 0$$

$$b = \mathbf{E}[\Theta] - 2a\mathbf{E}[X]$$

Now we may substitute b into the equation for a to get:

$$\begin{aligned}
a &= \frac{\mathbf{E}[\Theta X] - \mathbf{E}[\Theta]\,\mathbf{E}[X] + 2a\mathbf{E}[X]^2}{\mathbf{E}[X^2]} \\
a\left(1 - \frac{2\mathbf{E}[X]^2}{\mathbf{E}[X^2]}\right) &= \frac{\mathbf{E}[\Theta X] - \mathbf{E}[\Theta]\,\mathbf{E}[X]}{\mathbf{E}[X^2]} \\
a(\mathbf{E}\left[X^2\right] - 2\mathbf{E}[X]^2) &= \mathbf{E}[\Theta X] - \mathbf{E}[\Theta]\,\mathbf{E}[X] \\
a &= \frac{\mathbf{E}[\Theta X] - \mathbf{E}[\Theta]\,\mathbf{E}[X]}{\mathbf{E}[X^2] - \mathbf{E}[X]^2} \\
a &= \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{Var}(X)}
\end{aligned}$$

Finally, substituting a back into the equation for b, we have:

$$b = \mathbf{E}[\Theta] - \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{Var}(X)}\mathbf{E}[X]$$

Hence, we have:

Definition 3.4.1: Linear LMS Estimator (LLSE)

- The linear LMS estimator of Θ based on observed X is:

$$\hat{\Theta} = \frac{\mathrm{Cov}(\Theta, X)}{\mathrm{Var}(X)}(X - \mathbf{E}[X]) + \mathbf{E}[\Theta] = \rho\frac{\sigma_\Theta}{\sigma_X}(X - \mathbf{E}[X]) + \mathbf{E}[\Theta]$$

Where ρ is the correlation coefficient between Θ and X, defined as:

$$\rho = \frac{\mathrm{Cov}(\Theta, X)}{\sigma_\Theta \sigma_X}$$

the LLSE error can be calculated as:

$$\begin{aligned}
\mathbf{E}\Big[(\Theta - \hat{\Theta})^2\Big] &= \mathrm{Var}\Big(\Theta - \hat{\Theta}\Big) \\
&= \mathrm{Var}(\Theta) + \mathrm{Var}\Big(\hat{\Theta}\Big) \\
&= \sigma_\Theta^2 + \mathrm{Var}(aX + b) \\
&= \sigma_\Theta^2 + a^2 \mathrm{Var}(X) \\
&= \sigma_\Theta^2 + \frac{\mathrm{Cov}(\Theta, X)^2}{\mathrm{Var}(X)^2} \mathrm{Var}(X) \\
&= \sigma_\Theta^2 - \frac{\mathrm{Cov}(\Theta, X)^2}{\mathrm{Var}(X)} \\
&= \sigma_\Theta^2 (1 - \rho^2)
\end{aligned}$$

Chapter 4

Frequentist Statistical Inference

The frequentist approach to statistical inference treats the unknown parameter as a fixed but unknown value. No assumptionsare made about the prior distribution of the parameter, instead, we draw conclusions sole based on observed data.

4.1 Frequentist Parameter Estimation

The estimator $\hat{\Theta}$ and estimate $\hat{\theta}$ have a couple of impportant properties in frequentist inference:

Definition 4.1.1: Frequentist Estimator Properties

- The **estimation error** $\tilde{\Theta}_n$ is defined as $\tilde{\Theta}_n = \hat{\Theta}_n - \theta$, where $\hat{\Theta}_n$ is the estimator based on n observations, and θ is the true parameter value.
- The **bias** of an estimator $b_\theta(\hat{\Theta}_n)$ is the expected value of the error:

$$b_\theta(\Theta_n) = \mathbf{E}\left[\tilde{\Theta}_n\right] = \mathbf{E}\left[\hat{\Theta}_n\right] - \theta$$

- $\hat{\Theta}_n$ is **unbiased** if $b_\theta(\hat{\Theta}_n) = 0$ for all θ.
- $\hat{\Theta}_n$ is **asymptotically unbiased** if $\lim_{n\to\infty} b_\theta(\hat{\Theta}_n) = 0$ for all θ.
- $\hat{\Theta}_n$ is **consistent** if $\hat{\Theta}_n$ converges in probability to θ for all θ.

The estimation error is cahracterized by the mean squared error (MSE):

$$\mathbb{E}_\theta[\tilde{\Theta}_n^2] = b_\theta^2(\hat{\Theta_n}) + \text{var}_\theta(\hat{\Theta}_n)$$

4.2 Maximum Likelihood Estimation (MLE)

We can think of the **Maximum Likelihood Estimation (MLE)** as the estimator that maximizes the likelihood of ovserving the data given the parameter. Given an observation of $X = (X_1 \dots X_n)$, the MLE estimate $\hat{\theta}$ is given by:

$$\hat{\theta} = \text{argmax}_\theta \, \mathbf{P}(X = (x_1 \dots x_n) \mid \Theta = \theta)$$

When the observed data is continuous, we have:

$$\hat{\theta} = \text{argmax}_\theta \, fX = (x_1 \dots x_n) \mid \Theta = \theta$$

Note that because θ is not a random variable, the quantites $\mathbf{P}(X = x \mid \Theta = \theta)$ and $f_{X=x|\Theta=\theta}$ are not PMFs or PDFs, but rather **likelihood functions**, which are functions of θ for fixed x. However, it may be described as:

$$\mathbf{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n \mid \Theta = \theta) = \prod_{i=1}^{n} \mathbf{P}(X_i = x_i \mid \Theta = \theta)$$

However, it can be difficult to maximize the likelihood function directly. Hence, consider the **log-likelihood function**, where we take the logarithm of the likelihood function. This converts the product into a sum, and finding the maximum of the sum is equivalent to finding the maximum of the product as the logarithm is a monotonically increasing function. Hence, the MLE estimate may be rewritten as:

$$\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{n} \log \mathbf{P}(X_i = x_i \mid \Theta = \theta)$$

Similarly for if X was continuous:

$$\hat{\theta} = \text{argmax}_\theta \sum_{i=1}^{n} \log f_{X_i=x_i|\Theta=\theta}$$