CME 308: Stochastic Methods in Engineering

Samuel Wong
Department of Statistics
Stanford University

Abstract

The course starts with the application and the proof of the Law of Large Numbers and the Central Limit Theorem. Monte Carlo methods are then introduced as well as convergence and measure-theoretic probability. Statistical models, hypothesis testing, and parameter estimation both in the parametric and non-parametric setting are discussed. Both frequentist and Bayesian methodologies are covered. Lastly, the course moves to discussions about Markov chains/processes. The settings with discrete and continuous time and space are examined with focus on first transition analysis and stochastic control.

Contents

1 Weak Law of Large Numbers

The Weak Law of Large Numbers is statement that if we have random variables $X_1, X_2, ..., X_n$ iid, that $\bar{X}_n$ will converge to $E[X_1]$ in probability.

We can motivate this law with an example. Assume we have an unbiased coin, where flipping heads is denoted as 1 and tails as 0. We can flip this coin n times, where $P(X_i = 1) = \frac{1}{2} = P(X_i = 0)$. We want to prove that the sum, $S_n = X_1 + X_2 + ... + X_n$ converges to np. In other words, $S_{1000} \approx 500$, or more generally, $S_n \approx \frac{n}{2}$. Since $S_n \to \infty$ as $n \to \infty$, it is easier to prove $\bar{X}_n = \frac{S_n}{n} \to \frac{1}{2}$.

We first define convergence in probability as the following. We say that $Z_n \xrightarrow{p} Z_\infty$ if $P(|Z_n - Z_\infty| > \epsilon) \to 0$ as $n \to \infty$.

Therefore, we can state the Weak Law of Large Numbers as $P(|\bar{X}_n - E[X_1]| > \epsilon) \to 0$ as $n \to \infty$.

1.1 Proving the Weak Law of Large Numbers

Let us examine two ways to prove the Weak Law of Large Numbers. We will see why the second approach is much more preferable.

1.1.1 Approach 1

Approach 1 is to directly compute the distribution of S_n.

$P(S_n \leq x) = \int ... \int f(x_1, x_2, ..., x_n) dx_n dx_{n-1} ... dx_1$

The is a full n-dimensional integration! Let us take a closer look at the specific case where the X_i's are independent.

$f_{S_n}(x) = \int_{-\infty}^{\infty} f_{S_n|S_{n-1}}(x|y) f_{S_n-1}(y) dy$. Let's let $S_n = x$ and $S_{n-1} = y$, therefore $X_n = x - y$. We can rewrite as $\int_{-\infty}^{\infty} f_{X_n}(x-y) f_{S_n-1}(y) dy$, which we can see is a convolution. In convolution notation, we can rewrite this as $(f_{X_n} \circledast f_{S_n-1})(x)$. Using recursion, this equals $(f_{X_n} \circledast f_{X_{n-1}} \circledast ... \circledast f_{X_1})(x)$

Most commonly, $f_{S_n}(x)$ cannot be solved in closed form. However, let's provide one example where it can be to show how this approach would work.

Assume $X_1, X_2, ..., X_n$ iid and are distributed exponentially with parameter λ. Then $f_X(x) = \lambda e^{-\lambda x}$, when $x \geq 0$, and 0 otherwise. Therefore, we have that

$f_{S_2}(x) = \int_0^x f_{X_2}(x-y) f_{X_1}(y) dy$
$= \int_0^x \lambda e^{-\lambda(x-y)} \lambda e^{-\lambda y} dy$
$= e^{-\lambda x} \int_0^x \lambda^2 dy$
$= \lambda^2 x e^{-\lambda x}$

By induction, $f_{S_n}(x) = \frac{\lambda(\lambda x)^{n-1}}{(n-1)!} e^{-\lambda x}$, when $x \geq 0$, and 0 otherwise. This is a gamma (or Erlang) random variable with parameters (λ, n).

1.1.2 Approach 2

Approach 2 is to bound the probability, $P(|\bar{X}_n - E(X_1)| > \epsilon)$.

We first need to prove the Markov Inequality, which states that: For any $w \geq 0$, $P(W > w) \leq \frac{E[W]}{w}$.

$P(W > w) = E[\mathbb{1}(W > w)]$
$\leq E[\frac{W}{w} \mathbb{1}(W > w)]$
$\leq \frac{E[W]}{w}$

Now that we have proved the Markov Inequality, we can use the result to prove the Chebyshev Inequality, which states that: For any X with $Var(X) < \infty$, $P(|X - E(X)| > \epsilon) \leq \frac{Var(X)}{\epsilon^2}$

Let $W = |X - E(X)|^2$. Then by Markov's Inequality, we have $P(|X - E(X)|^2 > \epsilon^2) \leq \frac{E[|X-E(X)|^2]}{\epsilon^2}$. We know that $E[|X - E(X)|^2]$ is defined as $Var(X)$. Therefore, $P(|X - E(X)| > \epsilon) \leq \frac{Var(X)}{\epsilon^2}$.

Now, we have Chebyshev's Inequality and can use that to prove the Weak Law of Large Numbers. We can let the X in Chebyshev's Inequality equal to $\bar{X}_n$. Therefore, $P(|\bar{X}_n - E(\bar{X}_n)| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2}$. We can calculate $E(\bar{X}_n) = E(\frac{1}{n}\sum_{i=1}^n X_i) = \frac{1}{n}E(\sum_{i=1}^n X_i) = E(X_1)$. Now let's figure out what $Var(\bar{X}_n)$ is!

Let us assume that $X_1, X_2, ...X_n$ are iid, and $Var(X_1) < \infty$. Then,

$Var(S_n) = E[S_n - E(S_n)])^2$
$= E[\sum_{i=1}^n (X_i - E(X_i))]^2$
$= E[\sum_{i=1}^n \tilde{X}_i]^2$, where $\tilde{X}_i = X_i - E[X_i]$
$= \sum_{i=1}^n E[\tilde{X}_i]^2 + \sum_{i \neq j} E[\tilde{X}_i \tilde{X}_j]$
$= \sum_{i=1}^n Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$
$= nVar(X_1)$, since iid means that $Var(X_1) = ... = Var(X_n)$ and all $Cov(X_i, X_j) = 0$

Therefore, $Var(\bar{X}_n) = Var(\frac{1}{n}S_n) = \frac{1}{n^2}Var(S_n) = \frac{Var(X_1)}{n}$

Plugging this into our Chebyshev Inequality, we get that $P(|\bar{X}_n - E(\bar{X}_n)| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{Var(X_1)}{n\epsilon^2} \to 0$ as $n \to \infty$. Therefore, we have proved the Weak Law of Large Numbers.

1.2 Applications of Weak Law of Large Numbers

1.2.1 Coin Toss

If we consider an unbiased coin flip, and each head has a value of 1 and each tail has a value of 0, then $E[X_1] = \frac{1}{2}$. Therefore, by the Weak Law of Large Numbers (WLLN), we have that $\bar{X}_n \xrightarrow{p} E[X_1] = \frac{1}{2}$.

1.2.2 Gambling

Assume we have two gamblers, who each need to wager w_1 and w_2 on a game. For each game, the winner takes the entire pot, $w_1 + w_2$. Gambler 1 wins each game with a probability of p, and Gambler 2 wins each game with a probability of $1 - p$, also denoted as q. How much will each Gambler wager to play the game?

Using the WLLN, we can see that for Gambler 1, $\bar{X}_n \xrightarrow{p} p(w_1 + w_2) - w_1$, and similarly for Gambler 2, $\bar{X}_n \xrightarrow{p} q(w_1 + w_2) - w_2$. Therefore, Gambler 1 wants to bet such that $p(w_1 + w_2) - w_1 \geq 0$, which can be rewritten as $\frac{w_2}{w_1} \geq \frac{q}{p}$. Gambler 2 want to bet such that $q(w_1 + w_2) - w_2 \geq 0$, which can be rewritten as $\frac{w_2}{w_1} \leq \frac{q}{p}$. Therefore, the equilibrium solution is $\frac{w_2}{w_1} = \frac{q}{p}$, which is how much each player will wager.

Similarly, we have the problem of points, where we have 2 players and the first to win 10 rounds wins all the money. The game is interrupted before either player gets to 10, so how should the pot be divided? We should divide the pot up in terms of the odds of each player winning.

1.2.3 Monte Carlo

If our goal is to compute $\alpha = E[X]$, when it is analytically untractable, we can use the Monte Carlo method. We can simulate a variable X once, calling it X_1. We can repeat the simulation and obtain $X_2, ...X_n$. Using the WLLN, $\bar{X}_n \xrightarrow{p} E[X]$

1.2.4 News Vendor Problem

Assume we are a new vendor that sells newpapers every day. Let x be the order quantity of newspapers for us to order so that we can sell it. We want to find the x^*, the optimal order quantity to have the highest total profit possible. Let us assume the demand, $D_1, D_2, ...$ is a continuous random variable and is distributed iid.

Our profit for an order quantity x can be calculated as $R(x) = r(x \wedge D_i) + s[x - D_i]^+ - cx$, where r is the revenue generated per newspaper for selling one copy, s is the revenue generated from the salvage value of one copy, and c is the cost of ordering one copy.

By WLLN, $\bar{R}_n \xrightarrow{p} E[R_1(x)]$. This means that our optimal order amount can be found by maximizing $E[R_1(x)]$.

$E[R(x)] = rE[min(D,x)] + sE[x-D]^+ - cx$
$= r[xP(D>x) + \int_0^x yf_D(y)dy] + s\int_0^x (x-y)f_D(y)dy - cx$

$\frac{d}{dx} = rP(D>x) + rx(-f_D(x)) + rx(f_D(x)) + \frac{d}{dx}[sx\int_0^x f_D(y)dy - s\int_0^x yf_D(y)dy] - c$
$= rP(D>x) + sP(D\leq x) + sxf_D(x) - sxf_D(x) - c$
$= rP(D>x) + sP(D\leq x) - c = 0$

Therefore, the optimal order quantity, x^* is where $P(D\leq x^*) = \frac{r-c}{r-s}$, where $s<c<r$.

1.2.5 Investment Example

We examine the value of a portfolio at different times, $V_0, V_1, ...$, and we define the return at time i to be $R_i = \frac{V_i}{V_{i-1}} - 1$. Therefore we can rewrite $V_n = V_{n-1}(1+R_n) = V_0(1+R_1)...(1+R_n)$. Let us assume the returns, $R_1, R_2...$ are distributed iid. We can look at two different approaches to maximizing our money.

The first approach is to maximize $E(R_n)$, which is equivalent to maximizing $E(V_n)$. We see this by the following:

$V_n = V_0(1+R_1)...(1+R_n)$
$E[V_n] = E[V_0(1+R_1)...(1+R_n)]$
$= E[V_0]\prod_{i=1}^n E[1+R_i]$
$= E[V_0](1+E[R_1])^n$

We then can rewrite this in log form as the following:
$\frac{1}{n}log(E[V_n]) \to \frac{log(E[V_0])}{n} + log(1+E[R_1])$

The second approach is to maximize expected return based on some variance constraint. We start by first rewriting the formula in terms of logs:

$log(V_n) = log(V_0) + \sum_{i=1}^n log(1+R_i)$

We define the X_i as $1+R_i$. Using the WLLN, we can see that

$\frac{1}{n}log(E_n) \xrightarrow{p} \frac{log(E[V_0])}{n} + E[log(1+R_1)]$, which is different than the first approach!

Assuming log-normal returns, such that $1+R_i = exp(Z_i)$, where the Z_i are normally distributed $N(\gamma, \sigma^2)$, we can see that in the first approach, $E[log(1+R_i)] = E[Z_i] = \gamma$. Using the second approach, $log(1+E[R_i]) = log(E[1+R_i]) = log(E[exp(Z_i)]) = log(exp(\gamma + \frac{\sigma^2}{2}) = \gamma + \frac{\sigma^2}{2}$.

Therefore, if we go by the first approach, we will maximize by choosing the investment with the largest γ, or expected value. If we go with the second approach, we will maximize by choosing the investment with the largest $\gamma + \frac{\sigma^2}{2}$, which factors in the expected value and the risk of the investment.

1.3 Extension of the Weak Law of Large Numbers

We can extend the WLLN from an iid environment to a situation with dependent variables. Let us assume $X_1, X_2, ...$ is a stationary sequence. Therefore, $(X_1, X_2, ...) \stackrel{D}{=} (X_2, X_3, ...) \stackrel{D}{=} ... \stackrel{D}{=} (X_n, X_{n+1}, ...)$

$Var(S_n) = \sum_{i=1}^n Var(X_i) + \sum_{i\neq j} Cov(X_i, X_j)$
$= nVarX_1 + \sum_{i\neq j} Cov(X_0, X_{i-j})$
$= nVarX_1 + 2\sum_{i=1}^{n-1}(n-i)Cov(X_0, X_i)$

Since it is a stationary distribution, $Cov(X_0, X_n) \to 0$ as $n\to\infty$, therefore $Var(S_n) \to nVar(X_1)$, and $Var(\bar{X}_n) \to \frac{Var(X_1)}{n}$

Therefore, $P(|\bar{X}_n - E(X_1)| > \epsilon) \leq \frac{Var(\bar{X}_n)}{\epsilon^2} = \frac{Var(X_1)}{n\epsilon^2} \to 0$ as $n\to\infty$.

2 Central Limit Theorem

Now that we have the WLLN, we are still interested in a few more questions about $\bar{X}_n$ that the WLLN cannot answer. For example, we want to know how fast $\bar{X}_n$ converges to $E[X_1]$. We may also want to know what is $P(|\bar{X}_n - E(X_1)| > z$. These questions can be answered with the Central Limit Theorem.

2.1 Hypothesized Result

Assume we have random variables $X_1, X_2, ..., X_n$ that are distributed iid. Again, let us denote $S_n = X_1 + X_2 + ... + X_n$. Rather than look at this sequence of random variables, let us transform these random variables into a new sequence of random variables $W_1, ..., W_n$. Let us define $\tilde{X}_i = \frac{X_i - E(X_i)}{\sigma(X_i)}$, and $\tilde{S}_n = \sum_{i=1}^{n} \tilde{X}_i$. Now we can define $W_n = \frac{S_n - nE(X_1)}{\sqrt{n}\sigma(X_1)} = \frac{\tilde{S}_n}{\sqrt{n}}$. Using the WLLN, we want to see what happens when W_n converges to W_∞.

From the way we constructed the W_i, we know that $E[W_\infty] = 0$ and $Var[W_\infty] = 1$.

We know that if we look at W_{2n}, it must also converge to W_∞. We can write $W_{2n} = \frac{\tilde{S}_{2n}}{\sqrt{2n}} = \frac{1}{\sqrt{2}} \frac{\sum_{i=1}^{n} \tilde{X}_i}{\sqrt{n}} + \frac{1}{\sqrt{2}} \frac{\sum_{i=n+1}^{2n} \tilde{X}_i}{\sqrt{n}}$. Since each of the sequences on the right must also converge to W_∞, we have that $W_\infty \stackrel{D}{=} \frac{1}{\sqrt{2}} W_\infty + \frac{1}{\sqrt{2}} W_\infty$. Basically, we are stating that if we sum two independent distributions, we get the same family of distributions back in a linear fashion. There is only one distribution that has these 3 properties:

- $E[W_\infty] = 0$
- $Var[W_\infty] = 1$
- Summing two independent distributions of the same family, we get the same family of distributions back in a linear fashion

This special distribution is the $N(0, 1)$. Therefore, we hypothesize that $\frac{S_n - nE(X_1)}{\sqrt{n}\sigma(X_1)}$ converges to a $N(0, 1)$ given that $X_1, X_2, ...$ iid and that $Var(X_1) < \infty$.

In what way does it converge? It converges in distribution. What we are hypothesizing is that $P(\frac{S_n - nE(X_1)}{\sqrt{n}\sigma(X_1)} \leq x) \to \Phi(x)$ as $n \to \infty$.

The definition of convergence in distribution is as follows. Given a sequence $(W_n : n \geq 0)$, we say that the sequence converges in distribution to W_∞ if $P(W_n \leq x) \to P(W_\infty \leq x)$ as $n \to \infty$ at every x which is a continuity point of $P(W_\infty \leq \cdot)$

Therefore we can rewrite CLT as $S_n \approx nE[X_1] + \sqrt{n}\sigma(X_1)N(0, 1)$, where $nE[X_1]$ is where the distribution is concentrated and the rest accounts for the stochastic fluctuations. This is also equivalent to writing $\bar{X}_n \approx E[X_1] + \frac{1}{\sqrt{n}}\sigma(X_1)N(0, 1)$. Therefore, the convergence rate of $\bar{X}_n$ to $E[X_1]$ is $\frac{1}{\sqrt{n}}$.

2.2 Applications

2.2.1 Coin Toss

Assume we have an unbiased coin, and S_n is the number of heads tossed. Then each X_i is Bernoulli distributed, and therefore $E[X_1] = \frac{1}{2}$ and $Var[X_1] = \frac{1}{4}$. Therefore, $S_n \stackrel{D}{\approx} \frac{n}{2} + \sqrt{n}\frac{1}{2}N(0, 1)$.

If we want to find the probability of having more than 210 heads in 400 coin tosses, we can solve $P(S_{400} > 210) \approx P(200 + 20 * \frac{1}{2}N(0, 1) > 210) = P(N(0, 1) > 1) = 1 - \Phi(1) = 0.16$.

2.2.2 Arrival of nth Customer

We want to find out the distribution of the arrival of the nth customer. We can model the X_1 as the time the first customer arrives, X_2 as the time between the first and second customer, etc. Therefore $S_n = X_1 + X_2 + ... + X_n$ is the arrival time of the nth customer. Let us assume the X_i's are exponentially distributed with parameter λ.

$P(S_n > x) \approx P(\frac{n}{\lambda} + \frac{\sqrt{n}}{\lambda}N(0, 1) > x) = P(N(0, 1) > \frac{x - n\lambda}{\sqrt{n}})$

2.2.3 Gamma Random Variables

A gamma random variable is a sum of iid exponentially distributed random variables. $Gamma(\lambda, n) \stackrel{D}{=} \sum_{i=1}^{n} X_i$. From the previous example of the arrival of the nth customer, we saw that the CLT applies with mean $\frac{n}{\lambda}$ and variance $\frac{n}{\lambda^2}$. Therefore, we can see that $Gamma(\lambda, n) \stackrel{D}{\approx} N(\frac{n}{\lambda}, \frac{n}{\lambda^2})$. Although n is an integer in this example, we can generalize this to non-integers as well.

2.3 Rigorous Proof

We are going to prove the Central Limit Theorem using characteristic functions. First, let us understand what a characteristic function is.

2.3.1 Characteristic Functions

The characteristic function $c_X(\theta)$ is defined as $E[e^{i\theta X}]$. This in turn is equal to $E[cos(\theta X) + iE[sin(\theta X)]]$. Three properties of characteristic functions are:

1. $X \stackrel{D}{\approx} Y$ iff $c_X(\cdot) = c_Y(\cdot)$
2. $W_n \Rightarrow W_\infty$ iff $c_{W_n}(\theta) \to c_{W_\infty}(\theta)$ as $n \to \infty$ at each θ
3. If X, Y are independent random variables, then $c_{X+Y}(\theta) = c_X(\theta) + c_Y(\theta)$

2.3.2 Proof of Central Limit Theorem

From property 2 of characteristic functions, we know that $W_n \Rightarrow W_\infty$ iff $c_{W_n}(\theta) \to c_{W_\infty}(\theta)$ as $n \to \infty$ at each θ. Therefore, to prove the Central Limit Theorem, we are going to show that $c_{\frac{\tilde{S}_n}{\sqrt{n}}}(\theta) \to c_{N(0,1)}(\theta)$ as $n \to \infty$, where $\frac{\tilde{S}_n}{\sqrt{n}} = \frac{S_n - nE(X_1)}{\sqrt{n}\sigma(X_1)}$.

First we need to find the characteristic function of a normal distribution. For $z \stackrel{D}{=} N(\mu, \sigma^2)$, $c_Z(\theta) = exp(i\theta\mu - \frac{\theta^2\sigma^2}{2})$. Therefore, $c_{N(0,1)}(\theta) = exp(-\frac{\theta^2}{2})$.
Now we need to find $c_{\frac{\tilde{S}_n}{\sqrt{n}}}(\theta)$.

$c_{\frac{\tilde{S}_n}{\sqrt{n}}}(\theta) = E[e^{i\theta\frac{\tilde{S}_n}{\sqrt{n}}}]$
$= E[e^{\frac{i\theta}{\sqrt{n}}\sum_{j=1}^{n}\tilde{X}_j}]$
$= \prod_{j=1}^{n} c_{\tilde{X}_j}(\frac{\theta}{\sqrt{n}})$
$= [c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}})]^n$

We can see that $\frac{\theta}{\sqrt{n}} \to 0$ as $n \to \infty$. We expand $c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}})$ using a Taylor series expansion at the origin.

$c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}}) = c_{\tilde{X}_1}(0) + \frac{\theta}{\sqrt{n}}c'_{\tilde{X}_1}(0) + \frac{\theta^2}{2n}c''_{\tilde{X}_1}(0) + ...$

To complete the first and second order terms, we calculate
$\frac{d}{d\theta}c_{\tilde{X}_1}(\theta) = \frac{d}{d\theta}E[e^{i\theta\tilde{X}_1}] = E[\frac{d}{d\theta}e^{i\theta\tilde{X}_1}] = E[i\tilde{X}_1 e^{i\theta\tilde{X}_1}]$
$\frac{d^2}{d\theta^2}c_{\tilde{X}_1}(\theta) = \frac{d}{d\theta}E[i\tilde{X}_1 e^{i\theta\tilde{X}_1}] = E[(i\tilde{X}_1)^2 e^{i\theta\tilde{X}_1}] = -E[\tilde{X}_1^2 e^{i\theta\tilde{X}_1}]$

Therefore, $c_{\tilde{X}_1}(0) = 1$, $c'_{\tilde{X}_1}(0) = 0$, and $c''_{\tilde{X}_1}(0) = -1$.

Therefore, $c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}}) = 1 - \frac{\theta^2}{2n} + ...$

$c_{\frac{\tilde{S}_n}{\sqrt{n}}}(\theta) = [c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}})]^n$
$= (1 - \frac{\theta^2}{2n} + ...)^n \to e^{-\frac{\theta^2}{2}}$ since $(1 - \frac{z}{n})^n \to e^{-z}$

Therefore we have shown that the characteristic function of $\frac{\tilde{S}_n}{\sqrt{n}}$ converges to the characteristic function of the $N(0, 1)$.

2.4 Generalizations

2.4.1 CLT implies WLLN

We claim that CLT implies WLLN. In other words, if we know that $\sqrt{n}(\bar{X}_n - E[X_1]) \Rightarrow w$, then we can conclude that $\bar{X}_n \stackrel{P}{\to} E[X_1]$.

Assuming CLT, we can write the definition of convergence in distribution as $P(\sqrt{n}(\bar{X}_n - E[X_1]) > w) \to P(W > w)$, for a given value of w. We want to show that $\bar{X}_n \xrightarrow{P} E[X_1]$, or in other words, that $P(|\bar{X}_n - E[X_1]| > \epsilon) \to 0$ as $n \to \infty$. We can multiply both sides by $\sqrt{n}$ to get that $P(\sqrt{n}(\bar{X}_n - E[X_1]) > \sqrt{n}\epsilon)$.

As n gets large, we can choose any value w such that $\sqrt{n}\epsilon > w$. Therefore, by the definition of convergence stated above, we can write:

$P(\sqrt{n}(\bar{X}_n - E[X_1]) > \sqrt{n}\epsilon) \leq P(\sqrt{n}(\bar{X}_n - E[X_1]) > w)$. Therefore,

$limsup_{n\to\infty} P(\sqrt{n}(\bar{X}_n - E[X_1] > \sqrt{n}\epsilon) \leq limsup_{n\to\infty} P(\sqrt{n}(\bar{X}_n - E[X_1]) > w) = P(W > w) = 0.$

Therefore, CLT implies WLLN.

2.4.2 Random Vectors (Multivariate CLT)

We can generalize the CLT to vectors of random variables. For example, assume we have $Y_1, Y_2, ..., Y_n$ iid, and we want to look at Y and Y^2. Let X be a column vector of $(Y, Y^2)^T$. Then the CLT can be generalized such that $\bar{X}_n \xrightarrow{P} E[X] = (E[Y], E[Y^2])$. The first component of the $\bar{X}_n$ vector converges to the first component of the $E[X]$ vector and the second component converges to the second component.

More formally, in the case where $X_1, X_2..., X_n$ iid is $\mathbb{R}^d$-valued (column vectors), and $E[||X_1||^2] < \infty$, then $\frac{S_n - nE[X_1]}{\sqrt{n}} \Rightarrow N(0, C)$, where N is a multivariate normal distribution and C is the covariance matrix. $C = E([X_1][X_1]^T) - E[X_1]E[X_1]^T$, and consequently $C_{ij} = Cov(X_{1(i)}, X_{1(j)})$

For example, returning back to the case where $X = (Y, Y^2)^T$, then $C = [[Var(Y), Cov(Y, Y^2)][Cov(Y, Y^2), Var(Y^2)]]$, and $Cov(Y, Y^2) = E[Y^3] - E[Y]E[Y]^2$.

2.4.3 Non-Identically Distributed Dependent RV's

We can also generalize this to non-identically distributed dependent random variables. Let $X_1, X_2, ...X_n$ be a sequence of random variables and $S_n = \sum_{i=1}^n X_i$.

Then, $S_n \overset{D}{\approx} N(E[S_n], E([S_n][S_n^T]) - E[S_n]E[S_n]^T)$.

There are however, a few caveats; circumstances where we should NOT expect the CLT to hold:

- One of the X_i's dominates the other X_j's
- $Var(X_i) = \infty$
- Dependence between the X_i's is too strong. We need $Cov(X_i, X_{i+n} \to 0)$ as $n \to \infty$. This is called "asymptotic independence."

2.4.4 Stationary Sequence Example

Let's look at the case where $X_0, X_1, ...$ is a stationary sequence of dependent random variables. Therefore $(X_0, X_1, ...) \overset{D}{=} (X_m, X_{m+1}, ...)$, where $m \geq 1$. By the generalization of the CLT presented above for dependent random variables, we have that:

$S_n \overset{D}{\approx} N(nE[X_0], nVar[X_0] + 2\sum_{i=1}^{n-1}(n-i)Cov(X_i, X_j)) = N(n\mu, v_n)$

In the case where X_i's were iid, the variance $v_n = nVar(X_0)$, so it grew linearly in n. Let's look to see what conditions need to hold in this stationary case for v_n to grow linearly. We take a look at $\frac{v_n}{n}$.

$\frac{v_n}{n} \to Var(X_0) + 2\sum_{i=1}^{\infty} Cov(X_0, X_i) \overset{\triangle}{=} v_\infty$

Therefore, $v_n \approx nv_\infty$. We can see that v_∞ will only converge when the covariance term grows slowly enough that the summation is summable and does not grow to ∞. Therefore, v_n grows linearly when $\sum_{i=1}^{\infty} |Cov(X_0, X_i)| < \infty$ and $Var(X_0) < \infty$. We call this short range dependent.

If the covariance term faster than linearly, then we can write it as $Cov(X_0, X_i) \approx ci^{-\delta}$ for $(0 \leq \delta \leq 1)$. In this case, $v_n \approx \tilde{c}n^{1+p}$, where $p > 0$ and $\tilde{c}$ is some constant. Therefore, $S_n \overset{D}{\approx} N(nE[X_0], \tilde{c}n^{1+p})$.

Then $\frac{S_n - nE[X_0]}{n^{\frac{1+p}{2}}} \Rightarrow N(0, \tilde{c})$. As we can see, we need a normalizing factor of larger than $\sqrt{n}$. In this type of situation, where the covariance terms are going to 0 so slowly, that the sum does not converge, but we still have CLT. However, we call this a "long range dependent sequence."

2.4.5 Return on Investment Example

Let $D = \sum_{n=0}^{\infty} e^{-\alpha n} R_n$, where the discount factor models the time value of money. Let's assume that $R_0, R_1, ...$ are iid. However, D is NOT the sum of identically distributed random variables because each R_i has a discount factor associated with it. Depending on the value of α, the first few terms could dominate the entire sum. In that case, we do not expect the CLT to hold.

However, if $\alpha \approx 0$, then the earlier terms do not dominate and we can expect the Central Limit Theorem to hold. Therefore,

$D \overset{D}{\approx} N(E[D], Var[D]) = N(\frac{E[R_0]}{1-e^{-\alpha}}, \sum_{n=0}^{\infty} e^{-2\alpha n} Var[R_0]) = N(\frac{E[R_0]}{1-e^{-\alpha}}, \frac{Var[R_0]}{1-e^{-2\alpha}})$

3 Monte Carlo Method

The Monte Carlo Method is one application of LLN. Our goal in using the Monte Carlo method is to compute $\alpha = E[X]$. We can do this by repeatedly generating $X_1, X_2, ...$ iid, and from the WLLN, we know that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X]$ as $n \to \infty$. We also know from the CLT, that $\sqrt{n}(\bar{X}_n - E[X]) \Rightarrow \sigma(X)N(0,1)$, therefore $\bar{X}_n \overset{D}{\approx} E[X] + \frac{\sigma(X)}{n}N(0,1)$.

We therefore have the following insight:

- The convergence rate is $n^{-\frac{1}{2}}$, what we call "square root convergence rate"
- Monte Carlo is appropriate for settings in which one will satisfied with 2 or 3 significant figures of accuracy (more accuracy will take exponentially more n)
- Normal approximation can be used in Monte Carlo to deduce the error bars
- $\sigma(X)$ is a measure of the "difficulty" of the problem

Some advantages of Monte Carlo include:

- Flexibility - can be used to compute an enormous variety of different probabilities and expectations of interest
- Relatively easy to code - changing the underlying distribution of the variables could mean only change one line of code
- Look at "sample paths" - can see how the model got to its answer from the randomized path that it took to get there
- Monte Carlo tends to be the "method of choice" in high dimensional integration

3.1 Confidence Intervals

We can asses the error bars that come from the confidence intervals using the CLT (normal distribution) for $\bar{X}_n$. From convergence in distribution, we have that $P(-z \leq \frac{\bar{X}_n - E[X]}{\sigma(X)} \leq z) \to P(-z \leq N(0,1) \leq z)$ as $n \to \infty$.

$P(-z \leq \frac{\bar{X}_n - E[X]}{\sigma(X)} \leq z)$
$= P(\bar{X}_n - \frac{z\sigma(X)}{\sqrt{n}} \leq E[X] \leq \bar{X}_n + \frac{z\sigma(X)}{\sqrt{n}})$
$= P(E[X] \in [\bar{X}_n - \frac{z\sigma(X)}{\sqrt{n}}, \bar{X}_n + \frac{z\sigma(X)}{\sqrt{n}}])$

Therefore, $P(E[X] \in [\bar{X}_n - \frac{z\sigma(X)}{\sqrt{n}}, \bar{X}_n + \frac{z\sigma(X)}{\sqrt{n}}]) \to P(-z \leq N(0,1) \leq z)$ as $n \to \infty$. Practically, we can calculate the left side, and then use a Normal table to determine the confidence interval for $\alpha = E[X]$.

One caveat however is that $\sigma(X)$ is typically unknown. Therefore, we must estimate $\sigma(X)$ through the sample variance. Sample variance is defined as $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We want to show that $s_n^2 \xrightarrow{P} \sigma^2(X)$ as $n \to \infty$.

$s_n^2 = \frac{1}{n-1}[\sum_{i=1}^{n}(X_i)^2 - 2\bar{X}_n\sum_{i=1}^{n}X_i + \sum_{i=1}^{n}\bar{X}_n^2]$
$= \frac{1}{n-1}[\sum_{i=1}^{n}X_i^2 - 2n\bar{X}_n^2 + n\bar{X}_n^2]$
$= \frac{1}{n-1}[\sum_{i=1}^{n}X_i^2 - n\bar{X}_n^2]$
$= \frac{\sum_{i=1}^{n}X_i^2}{n-1} - \frac{n}{n-1}\bar{X}_n^2 \xrightarrow{P} E[X^2] - (E[X])^2 = \sigma^2(X)$

Therefore, we can generate an approximate confidence interval for $E[X]$ with $[\bar{X}_n - \frac{zs_n}{\sqrt{n}}, \bar{X}_n + \frac{zs_n}{\sqrt{n}}]$

We can see that the CLT still holds when we use $s_n(X)$ instead of $\sigma(X)$. We want to argue that $\frac{\sqrt{n}(\bar{X}_n - E[X])}{s(X)} \Rightarrow N(0,1)$. We can rewrite the left side as $\frac{\sigma(X)}{s(X)} \cdot \frac{\sqrt{n}(\bar{X}_n - E[X])}{\sigma(X)}$. We have previously proven that $s_n^2 \xrightarrow{P} \sigma^2(X)$. We also have a theorem that says if $W_n \xrightarrow{P} W_\infty$, and g is continuous, then $g(W_n) \xrightarrow{P} g(W_\infty)$. Therefore, if we choose $g(x) = \sqrt{x}$ and $s_n^2 \xrightarrow{P} \sigma^2(X)$, then $s_n \xrightarrow{P} \sigma(X)$, and $\frac{\sigma(X)}{s(X)} \xrightarrow{P} 1$. From CLT, we already know that $\frac{\sqrt{n}(\bar{X}_n - E[X])}{\sigma(X)} \Rightarrow N(0,1)$. Therefore, we have proved that $\frac{\sqrt{n}(\bar{X}_n - E[X])}{s(X)} \Rightarrow N(0,1)$.

A general result of what we have just proved is this. If $W_n \Rightarrow W_\infty$ and $\beta_n \xrightarrow{P} a$, where a is a deterministic limit, then $\beta_n W_n \Rightarrow aW_\infty$.

3.2 Applying Monte Carlo in Practice

3.2.1 Motivating Example

In project management, we come across critical path planning. Assume that to complete a project we need to get from point A to point B. Between point A and B are many different tasks, call them $X_1, X_2, ...X_n$. Many of these tasks are dependent on each other, for example, say X_4 cannot start until X_2 is complete, or X_8 requires both X_3 and X_7 to start, etc. Let us define the critical path (maximum length) as C through the network. Any delay to C will delay the completion of the project. $C = L(X_1, X_2, ...X_n)$. It is a function of the completion times of the individual tasks.

It is impossible to get $E[C]$ in closed form, but we can use Monte Carlo. We can generate a sequence of $C_1, C_2, ...C_m$ and then calculate $\bar{C}_m$ to estimate $E[C]$. For each $C_1, C_2, ...C_m$, we generate $X_1, X_2, ...X_n$ to calculate each C_i.

What we require are algorithms that can generate X_i's from their associated underlying distributions. Therefore we need non-uniform random number generation. To perform this, we first need to perform uniform random number generation, and then put these random numbers as inputs into the non-uniform random number generator.

3.2.2 Uniform Random Generation

The goal of uniform random generation is to generate iid uniform random variables on [0,1] using an algorithm. A few different methods to do this include:

- Midsquare Method
  - To use this method, we first choose a 4 digit number. We square it, which then will most likely produce an 8 digit number. We select the middle 4 digits of this 8 digit number and repeat the process.
  - A large con of this method is that the algorithm can degenerate if the middle 4 digits end up with 4 0's.
  - This method is no longer commonly used.
- Linear Congruential Generators
  - We generate $x_{n+1} = (ax_n + b) \mod m$
  - Therefore, x_i can take on values between $\{0, 1, ...m-1\}$
  - We then generate our u_i as $\frac{x_i}{m}$ so that u_i is between 0 and 1
  - We want a full period generator (one that will produce all the values between 0 and $m-1$ without missing any, so the sample will look uniform
  - Turns out you can get a full period generator by choosing m as a prime number and a, b accordingly
  - This method is still widely used today.
- Physical Generators
  - Use a physical device to generate a uniform distribution, such as actually flipping a coin and recording 1 if heads and 0 if tails
  - Can use more complex items such as particle emissions to generate the uniform random variables
  - The main con of this approach is that it is expensive and slow (how long does it take to flip a coin and record the output?)
  - Another con is that the reality may not be perfectly unbiased - for example a coin in reality may have a 49.5% chance of heads because of wear and tear of the coin

3.2.3 Non-Uniform Random Generation

The goal of non-uniform random generation is to generate random variables $Z_1, Z_2, ..., Z_n$ iid having a prescribed distribution, given $U_1, U_2, ..., U_n$ as $unif[0, 1]$ as inputs. Two methods performing non-uniofrm random generation are the Inversion Algorithm and the Acceptance-Rejection Algorithm.

The Inversion Algorithm creates an output $Z_i = F_Z^{-1}(U_i)$, where F_Z^{-1} is the function inverse to the CDF $F_Z(\cdot)$. Since F_Z is a monotone increasing function, $F_Z(F_Z^{-1}(x)) = F_Z^{-1}(F_Z(x)) = x$. We can show that this inversion algorithm works by the following:

$P(F_Z^{-1}(U) \leq x) = P(F_Z(F_Z^{-1}(U)) \leq F_Z(x))$
$= P(U \leq F_Z(x))$

$= P(U \leq y) = y$

Therefore, we've established that $P(F_Z^{-1}(U) \leq x) = F_Z(x)$. This Inversion Method works not only for continuous CDFs, but also for discrete random variables as well.

An example of how this method would work is as follows. Suppose we want to generate a sequence of Z_i's that are exponentially distributed with parameter λ. We know that $F_Z(x) = 1 - e^{-\lambda x}$.

$y = F_Z^{-1}(x)$
$F_Z(y) = x$
$1 - e^{-\lambda y} = x$
$1 - x = e^{-\lambda y}$
$y = -\frac{1}{\lambda} log(1 - x)$

Therefore, we can use the uniform random generator to generate the U_i's, and then we can generate the Z_i's by plugging the U_i's in as $Z_i = -\frac{1}{\lambda} log(1 - U_i)$.

The second approach is called the Acceptance-Rejection Algorithm. The advantage of this approach is that while the Inversion approach works well with random variables having an easily computable F_Z^{-1}, this is not always the case. For example, the normal distribution has no closed form representation of F_Z. The Acceptance-Rejection Algorithm works not with the CDF, but rather with the PDF, $f_Z(\cdot)$.

To use this algorithm, we need to find a majorizing density $g(\cdot)$ such that $sup_x \frac{f_Z(x)}{g(x)} < \infty$. We also need to have a fast algorithm available for generating random variables W that have the density $g(\cdot)$. For example, we can use the exponential pdf, $\lambda e^{\lambda x}$, if it is a majorizing density, because we have just shown that we can quickly generate these using the Inversion Algorithm.

When we say that $sup_x \frac{f_Z(x)}{g(x)} < \infty$, we can see that $f_Z(x)$ must therefore go to 0 at a faster rate than $g(x)$ at the tail ends of the distribution, otherwise the sup_x will go to ∞.

Let $c = sup_x \frac{f_Z(x)}{g(x)}$. Then $f_Z(x) \leq cg(x)$. We can write $f_Z(x) = cg(x)\frac{f_Z(x)}{cg(x)}$. We can view $g(x)$ as the likelihood that w takes on the value x, and $\frac{f_Z(x)}{cg(x)}$ as $P(U \leq \frac{f_Z(x)}{cg(x)})$. Notice that $\frac{f_Z(x)}{cg(x)} \leq 1$.

The steps to carry out the algorithm are as follows:

1. Generate w having density $g(\cdot)$.
2. Generate a u from a uniform distribution.
3. If the $u \leq \frac{f_Z(w)}{cg(w)}$, return $z = w$, else go back to step 1.

The Acceptance-Rejection Algorithm generalizes beyond generating scalar random variables. We can use the same idea for generating Z that is $\mathbb{R}^d$-valued. In this case W is also $\mathbb{R}^d$-valued. We now want to satisfy $sup_{x \in \mathbb{R}^d} \frac{f_Z(x)}{g(x)} < \infty$, and the steps remain the same as the ones previously outlined above.

3.3 High Dimensional Monte Carlo

In many stochastic situations, we will be dealing with high dimensional problems. For example, in the project management example where the critical path C is a function of task durations $X_1, X_2, ...X_n$, we could have hundreds or thousands of tasks in real life. If we were to compute this numerically, we would need to solve $\alpha = \int ... \int h(x_1, x_2..., x_n) f(x_1, x_2, ..., x_n) dx_1 dx_2 ... dx_n$. Let's take a closer look at what this would mean.

Let's start with the simple case where there is only 1 dimension, $d = 1$. Let us look at what would happen for integration in the [0,1] region. We can generalize all problems to be integrals in the [0,1] for the following reason. We are looking for $\alpha = E[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$. We know that from the Inversion Algorithm that $X = F^{-1}(U)$. Therefore we can rewrite the formula as $\alpha = E[h(F^{-1}(U))] = \int_0^1 h(F^{-1}(U)) du = \int_0^1 g(x) dx$, for some function g.

If we were to compute the approximate area as a sum of the area of n rectangles, we would calculate $\alpha_n = \sum_{i=0}^{n-1} g(\frac{i}{n})\frac{1}{n}$. The true value of $\alpha = \sum_{i=1}^{n} \int_{\frac{i-1}{n}}^{\frac{i}{n}} g(x) dx$. We want to look at the rate of convergence, so we therefore write $\alpha_n - \alpha = \sum_{i=0}^{n-1} \int_{\frac{i-1}{n}}^{\frac{i}{n}} [g(\frac{i-1}{n}) - g(x)] dx$. We assume that g is continuously differentiable in the interval [0,1], and that the derivate of g is

bounded, i.e. $|g'(x)| \leq m$, for $x \in [0,1]$. We can use a Taylor series to represent $g(\frac{i-1}{n}) - g(x)$ as $g'(\frac{i-1}{n})(\frac{i-1}{n} - x)$+ 2nd order term + ... Therefore, we can bound the difference by the following:

$|\alpha_n - \alpha| \leq m \sum_{i=0}^{n-1} [|x - (\frac{i-1}{n})|] dx$

We can see that this difference converges at the rate of $mO(\frac{1}{n})$.

In higher dimensions, where $d \geq 2$, we can similarly write $\alpha = \int_{[0,1]^d} g(\vec{x}) d\vec{x}$. In this case, we can also write $\alpha_n = \sum_{i=1}^{n} g(\vec{x_i}) \frac{1}{n}$. Generalizing from the 1 dimensional case where we split the area into n rectangles, we can now split the volume into n subhypercubes, denoted H_i of equal volume of n^{-1}. We can choose a point $\vec{x_i} \in H_i$ as the representative point for H_i.

What we see is that the difference $|\alpha_n - \alpha|$ is going to depend on $g(\vec{x}) - g(\vec{x_i}) = \nabla s(\epsilon)(\vec{x} - \vec{x_i})$, which is dependent on $||\vec{x} - \vec{x_i}||$. For each subhypercube, the volume is n^{-1}, and therefore each side $a = n^{-\frac{1}{d}}$. Therefore, the convergence rate of the difference, $|\alpha_n - \alpha| = O(n^{-\frac{1}{d}})$. We can see that there will be a slow convergence rate for large d.

Therefore, performing the integration numerically at low dimensions can make sense, but in higher dimensions, the integration becomes impossibly slow. For example, if we had a problem with 10 dimensions and we wanted to get to the error tolerance of 0.1, or 10%, we would require function evaluations on the order of $10^{10} = 10000000000$! Even with more sophisticated integration rule (Simpson's rule, higher order rule), we still run into curse of dimensionality problems.

In contrast, by Central Limit Theorem, the convergence is $n^{-\frac{1}{2}}$ regardless of dimension. That means that for $\alpha = E[X]$, where $X = g(U_1, ..., U_5)$ or $X = g(U_1, ..., U_5 0)$, they both converge at $n^{-\frac{1}{2}}$.

Therefore, in higher dimensionality problems, we want to use the Monte Carlo method! In 1 dimension, the integration numerically will converge faster than Monte Carlo, and in 2 dimensions it will be roughly the same, but in 3 or more dimensions, the Monte Carlo approach will converge much faster. It is "dimensionality insensitive."

3.4 Variance Reduction

We want to choose an appropriate estimate of $\alpha = E[X]$. Given that we have multiple estimators of $E[X]$, we want to choose the one with the smallest variance, so that we can have faster convergence. If we have $\bar{X}_n \overset{D}{\approx} \frac{\sigma(X)}{\sqrt{n}} N(0,1)$ and also $\bar{Y}_n \overset{D}{\approx} \frac{\sigma(Y)}{\sqrt{n}} N(0,1)$, then we will choose to use $\bar{Y}_n$ over $\bar{X}_n$ if $\sigma^2 Y \leq \sigma^2 X$. We call these methods "variation reduction techniques." We will look at some different variation reduction techniques in detail, Control Variates, Antithetics, and the Method of Common Random Numbers.

3.4.1 Control Variates

We want to find $\alpha = E[X]$. Let us assume there is a variable C that has a joint distribution with X such that $E[C] = 0$. An example of this would be if we define $C_i = X_i - E[X_i]$. Then our "control variate" C_i would have $E[C] = 0$.

We denote a variable $Y = X - \lambda C$. Note that since $E[C] = 0$, that $E[Y] = E[X]$, so Y is another appropriate estimator for $E[X]$. However, if we take a look at the variance of Y, we see that

$Var[Y(\lambda)] = Var[X] - 2\lambda Cov(X, C) + \lambda^2 Var[C]$

When we try to minimize the variance by taking its derivative and setting it equal to 0, we find that the optimal λ is $\lambda^* = \frac{Cov(X,C)}{Var(C)}$.

If we use this optimal λ^*, we see that the variance equals

$Var[Y(\lambda^*)] = Var[X] - 2\frac{Cov(X,C)^2}{Var[C]} + \frac{Cov(X,C)^2}{Var[C]}$
$= Var[X](1 - \frac{Cov(X,C)^2}{Var[X]Var[C]})$
$= Var[X](1 - \rho(X,C)^2)$, where $\rho(X,C)$ is the squared coefficient of correlation between -1 and 1

Therefore, we can see that if we choose a control variate that is highly correlated (positively or negatively) with X, then we can greatly reduce the variance.

We want to generalize this to the vector case. Suppose in the project management example we have $\alpha = E[h(X_1, X_2, ..., X_d)]$, where each of the X_i comes from its own distribution. We can have a different control variate for each X_i depending on the mean of that particular distribution. In this case, we have $C_i = X_i - E[X_i]$, and $Y[\vec{\lambda}] = X - \vec{\lambda}\vec{C}$, where $\vec{\lambda} = (\lambda_1, ..., \lambda_d)$ and $\vec{C} = (C_1, ..., C_d)^T$.

We can generalize our result from the one above to find that the optimizing $\vec{\lambda}$ that minimizes $Var[Y(\vec{\lambda})]$ is:

$\lambda^{*T} = \Sigma_{\vec{C}}^{-1}\Sigma_{X\vec{C}}$, where $\Sigma_{\vec{C}}^{-1}$ is the inverse of the covariance matrix of $\vec{C}$, and $\Sigma_{X\vec{C}}$ is the column vector whose ith entry is $Cov(X, C_i)$

In practice, we don't know $Cov(X, C)$ and $Var(C)$, so we estimate them using the sample data. We use the sample data to estimate $\hat{\lambda}_n$, which asymptotically $\to \lambda^*$ as $n \to \infty$.

3.4.2 Antithetics

We can use antithetics to reduce variance by the following formula for a non-decreasing function f and a symmetric random variable X:

$\hat{\alpha} = \frac{1}{2n}\sum_{i=1}^{n}[f(X_i) + f(-X_i)]$

The intuition behind antithetics is that if we sample an X_i that is far away from the expected value we are estimating, then we will equally use the opposite value, and those two values will approximately cancel out to get us an average close to our expected value. Antithetics can be proven to extend beyond a symmetric random variable X to apply to a uniform random variable on [0, 1]. Therefore, we get:

$\hat{\alpha} = \frac{1}{2n}\sum_{i=1}^{n}[f(U) + f(1 - U)]$

So for a practical example, we can go back to the project management example, where we want to estimate the longest path which equals $h(X_1, X_2, ...X_n)$. Say these X_i come from an exponential. We can first randomly generate n numbers from uniform [0,1], one for each of the X_n. We take each random number u_i and use the inversion method for u_i to get our x_i. We have n of these, one for each X_i. We then plug these x_i into h to get an estimate. Keeping these same u_i's, we do exactly the same, but in the inversion step, we invert $1 - u_i$. Therefore, using one set of u_i, we have generated 2 estimates now, and we take the average. This would be considered 1 random estimate of the longest path. As typical Monte Carlo suggests, we would repeat this a large number of times and take the average of those to get our final estimate $\hat{\alpha}$.

Using this antithetics method, we can have the same expected value of the estimate while reducing the variance.

3.4.3 Method of Common Random Numbers

In the Method of Common Random Numbers, we want to find $\alpha = E[X] - E[Y]$. Practically, this would happen for example if we were comparing two different methods to see which one is better. What we claim is that if $X = h_1(\vec{Z})$ and $Y = h_2(\vec{Z})$, where $\vec{Z} = (Z_1, ..., Z_d)$, then by using the same randomly generated Z_ij's, that we will have a variance reduction. This method works best when h_1 and h_2 are monotone in $\vec{Z}$. Since $\vec{Z}$ is a vector, monotonicity means that if ANY dimension x_i is increasing/decreasing, then $h(x_i)$ is always increasing/decreasing in the appropriate direction. Therefore, what we are requiring is that X and Y move in the same direction in response to changes in $\vec{Z}$.

Let us closely examine the method of common random numbers in 1 dimension and compare it to the standard Monte Carlo method. In standard Monte Carlo, we would generate n instances to estimate X and n more instances independently to estimate Y. Therefore our estimate would be $\frac{1}{n}\sum_{i=1}^{n} h(z_{1i}) - \frac{1}{n}\sum_{i=1}^{n} h(z_{2i})$. The variance of the estimate would be $Var[\hat{\alpha}] = \frac{1}{n}Varh_1(Z) + \frac{1}{n}Varh_2(Z)$.

Using common random numbers, where we generate one set of n instances and use them to estimate both X and Y, we see that the variance is

$\frac{1}{n}Var[h_1(Z) - h_2(Z)]$
$= \frac{1}{n}[Var[h_1(Z)] - 2Cov[h_1(Z), h_2(Z)] + Var[h_2(Z)]]$

Therefore, if $Cov[h_1(Z), h_2(Z)] \geq 0$, then we will have variance reduction! We claim that the $Cov[h_1(Z), h_2(Z)] \geq 0$ when h_1 and h_2 are monotone in the same direction. We can show this by first realizing that if h_1 and h_2 are monotone in the same direction, that when $Z_1 \geq Z_2$, then $h_1(Z_1) - h_1(Z_2) \geq 0$ and $h_2(Z_1) - h_2(Z_2) \geq 0$. It is also true that when $Z_1 < Z_2$, then

$h_1(Z_1) - h_1(Z_2) < 0$ and $h_2(Z_1) - h_2(Z_2) < 0$. Therefore, $h_1(Z_1) - h_1(Z_2)$ always has the same sign as $h_2(Z_1) - h_2(Z_2)$, so $[h_1(Z_1) - h_1(Z_2)][h_2(Z_1) - h_2(Z_2)] \geq 0$. Therefore, the expected value of that is also ≥ 0.

$E([h_1(Z_1) - h_1(Z_2)][h_2(Z_1) - h_2(Z_2)]) \geq 0$
$= 2E[h_1(Z_1)h_2(Z_2)] - 2E[h_1(Z_1)]E[h_2(Z_1)] \geq 0$ since the distribution of Z_i's are the same
$= 2Cov[h_1(Z), h_2(Z)] \geq 0$

This calculation assumes the variance of $h_1(Z)$ and $h_2(Z)$ is finite.

4 Convergence Concepts in Probability

In deterministic mathematics, we can view X as a function that maps say a value in $[0,1] \to$ a $X(t)$ in $\mathbb{R}$. We can look at convergence in these deterministic functions as follows:

- L^P Convergence
  - $x_n \to x_\infty$ iff $\int_a^b |x_n(s) - x_\infty(x)|^p ds \to 0$ as $n \to \infty$
- Convergence in the sup norm (uniformly)
  - $x_n \to x_\infty$ iff $sup_{x \in [a,b]} |x_n(s) - x_\infty| \to 0$ as $n \to \infty$

Now, we want to look at convergence when X is a random variable. We can view random variables as functions that map $\Omega \to \mathbb{R}$, where Ω is the sample space. When we talk about convergence of a random variable, we have different definitions of convergence than in the deterministic case. Below are 4 different definitions of convergence that we use when X is a random variable.

- Convergence with Probability 1 (Almost Sure Convergence)
  - $X_n \overset{a.s.}{\to} X_\infty$ iff $A : \{\omega : X_n(\omega) \to X_\infty(\omega) \text{ as } n \to \infty\}$ with $P(A) = 1$
- L^P Convergence ($p \geq 1$)
  - $X_n \overset{L^P}{\to} X_\infty$ iff $E[|X_n - X_\infty|]^p \to 0$ as $n \to \infty$
- Total Variation Convergence
  - $X_n \overset{t.v.}{\to} X_\infty$ iff $sup_A |P(X_n \in A) - P(X_\infty \in A)| \to 0$ as $n \to \infty$ over all subsets A
- Convergence in Probability
  - $X_n \overset{P}{\to} X_\infty$ iff $\forall \epsilon > 0$, $P(|X_n - X_\infty| > \epsilon) \to 0$ as $n \to \infty$
- Convergence in Distribution (Weak Convergence)
  - $X_n \Rightarrow X_\infty$ iff $P(X_n \leq x) \to P(X_\infty \leq x)$ as $n \to \infty$ at all continuity points of $P(X_\infty \leq \cdot)$

We now look at how Almost Sure Convergence is used in a theorem. We claim that if $X_n \overset{a.s.}{\to} X_\infty$ and $T_n \overset{a.s.}{\to} \infty$, where $T_n \in \mathbb{Z}_+$, then $X_{T_n} \to X_\infty$. Note that this theorem may not hold for all types of convergence. For example if $X_n \overset{P}{\to} X_\infty$, then this result does not necessarily hold.

We can prove this using an path-by-path (omega-by-omega) argument. Since we have almost sure convergence for both X_n and T_n, we know that we have an A and B where $A : \{\omega : X_n(\omega) \to X_\infty(\omega) \text{ as } n \to \infty\}$, where $P(A) = 1$, and $B : \{\omega : T_n(\omega) \to \infty \text{ as } n \to \infty\}$, where $P(B) = 1$. Since $P(A) = 1$ and $P(B) = 1$, we know that $P(A \cap B) = 1$. Let us now select ω's that work for both A and B, therefore $\omega \in A \cap B$. For this set of ω, $X_{T_n(\omega)}(\omega) \to X_\infty(\omega)$. Let us now define $C : \{\omega : X_{T_n(\omega)}(\omega) \to X_\infty(\omega) \text{ as } n \to \infty\}$. We know that $A \cap B \subseteq C$. We also know that $P(A \cap B) = 1$, therefore $P(C) = 1$.

4.1 Convergence Implications

Some notions of convergence are "stronger than" others. In other words, if we know a random variable converges in one sense, then we know that it will also converge in another.

Almost sure convergence implies convergence in probability. L^P convergence implies convergence in probability. Convergence in probability implies convergence in distribution. Total variation convergence implies convergence in distribution.

4.1.1 L^P implies Probability

$P(|X_n - X_\infty| > \epsilon) \leq \frac{E[|X_n - X_\infty|]}{\epsilon}$ by Markov's Inequality
$P(|X_n - X_\infty|^p > \epsilon^p) \leq \frac{E[|X_n - X_\infty|^p]}{\epsilon^p}$

L^P convergence means that $E[|X_n - X_\infty|^p \to 0$ as $n \to \infty$, therefore, $\frac{E[|X_n - X_\infty|^p]}{\epsilon^p} \to 0$ as $n \to \infty$.

Therefore, $P(|X_n - X_\infty| > \epsilon) \to 0$ as $n \to \infty$, so we have convergence in probability.

4.1.2 Total Variance Implies Distribution

Let $A = (-\infty, x]$. Then since we have convergence in total variance, $sup_x|P(X_n \leq x) - P(X_\infty \leq x)| \to 0$ as $n \to \infty$. Therefore, $X_n \Rightarrow X_\infty$ and we have convergence in distribution.

4.1.3 Almost Sure Implies Probability

When we have almost sure convergence, we can say that $\forall \epsilon > 0$, for each ω, there exists $N(\epsilon)$ such that when $n \geq N(\epsilon)$, then $|X_n(\omega) - X_\infty(\omega)| < \epsilon$. For different ω, this could mean different $N(\epsilon)$, but we do know that one exists for each ω.

$P(|X_n - X_\infty| > \epsilon)$
$= P(|X_n - X_\infty| > \epsilon, N(\epsilon) \leq n) + P(|X_n - X_\infty| > \epsilon, N(\epsilon) > n)$
$= 0 + P(|X_n - X_\infty| > \epsilon, N(\epsilon) > n)$
$\leq P(N(\epsilon) > n) \to 0$ as $n \to \infty$

4.1.4 Probability Does Not Imply L^P

Let U be uniform on [0,1] and let $X_n = 0$ if $U > \frac{1}{n}$ and n if $U \leq \frac{1}{n}$.

$P(X_n = 0) = 1 - \frac{1}{n} \to 1$ as $n \to \infty$

Therefore, $X_n \xrightarrow{P} X_\infty$, where $X_\infty = 0$.

However, $E[|X_n - X_\infty|] = E[|X_n - 0|] = E[|X_n|] = n \cdot \frac{1}{n} = 1$. Therefore, we do not have L^1 convergence here. $X_n \overset{L^1}{\not\to} X_\infty$.

4.1.5 Probability Does Not Imply Almost Sure

We will use an example to show a case where there is convergence in probability but not convergence almost surely. We take integers starting from 0 and go to ∞. We group these integers in groups such that the jth group has j consecutive integers. For example, the first group is 0, the second group is 1 and 2, the third group is 3, 4, and 5 etc. For each group, we randomly select one the elements to be equal to 1 and the rest equal to 0. For example, one possible selection would be $Z_0 = 1, Z_1 = 0, Z_2 = 1, Z_3 = 0, Z_4 = 1, Z_5 = 0, ...$

What is the probability that $Z_n = 0$? Well, $P(Z_n = 0) = 1 - \frac{1}{k_{n+1} - k_n}$, which $\to 1$ as $n \to \infty$. Therefore, we have convergence in probability. $Z_n \xrightarrow{P} Z_\infty$, where $Z_\infty = 0$.

However, if we look at the sequence of Z_n, no matter how far down we go, there will always eventually be another 1. Therefore, we have an infinite number of 0's and 1's in our sequence. There is no $n \geq N(\epsilon)$ such that we only have 0's past that point. Therefore, we do not have convergence almost surely. $Z_n \overset{a.s.}{\not\to} Z_\infty$.

This example shows that if $Z_n \xrightarrow{P} 0$, and if we define T_n to be the nth time at which the Z sequence takes on the value 1. Therefore, $T_n \to \infty$. Clearly, $Z_{T_n} = 1$, and therefore does not converge to 0. One might expect a subsequence to converge to what the sequence converges to, but that is only if the sequence converges almost surely, and the subsequence converges almost surely to ∞ as we have proved before.

4.1.6 Distribution Partial Converse to Almost Sure

One special case of Skorohod's Representation Theorem states that if $X_n \Rightarrow X_\infty$ as $n \to \infty$, with the X_i's having positive densities, then $F^{-1}_{X_n}(U) \overset{a.s.}{\to} F^{-1}_{X_\infty}(U)$. Another way of stating this is that if $X_n \Rightarrow X_\infty$ as $n \to \infty$, then there exists a sample space (probability space) supporting random variables $(X'_n : 1 \leq n \leq \infty)$ such that:

- $X'_n \overset{D}{=} X_n, 1 \leq n \leq \infty$
- $X'_n \overset{a.s.}{\to} X'_\infty$

We can see how we can take advantage of this theorem in the following way. If we know that $X_n \Rightarrow X_\infty$, then we know that we can find a sample space such that $X'_n \overset{a.s.}{\to} X'_\infty$. From our knowledge of real-valued limits, we know that if $x_n \to x_\infty$, then for every continuous function h, then $h(x_n) \to h(x_\infty)$. Since we know that $X'_n \overset{a.s.}{\to} X'_\infty$, we can use the path-by-path (omega-by-omega) argument, to show that we have real-valued convergence for every single ω, and therefore $h(X'_n) \overset{a.s.}{\to} h(X'_\infty)$. Since convergence almost surely implies convergence in distribution, then $h(X'_n) \Rightarrow h(X'_\infty)$. We know that $h(X'_n)$ has the

same distribution as $h(X_n)$, and $h(X'_\infty)$ has the same distribution as $h(X_\infty)$, so $h(X_n) \Rightarrow h(X_\infty)$.

Therefore, what we have just shown is that if $X_n \Rightarrow X_\infty$, then $h(X_n) \Rightarrow h(X_\infty)$ for every continuous function $h(\cdot)$. This is a widely used result in statistics and probability and is called the "Continuous Mapping Principle."

5 Measure-Theoretic Probability

Measure-theoretic probability was an advance in probability theory that occurred roughly in the early 20th century that can handle more complex modeling environments than conventional calculus-based probability can. It can be thought of as an extension to the traditional calculus-based probability theory.

For example, let us look at coin tosses where heads is 1 and tails is 0. If we toss a coin m times, then our sample space $\Omega_m = \{(X_1, X_2, ..., X_m) : X_i \in \{0,1\}\} = \{0,1\}^m$. The size of the sample size is 2^m, and therefore, when we want to calculate the probability of an event, we can simply look at $P(X_1 = x_1, ..., X_m = x_m) = 2^{-m}$ for an unbiased coin.

However, sometimes we may want at an infinite number of coin tosses. Now our sample space $\Omega = \{0,1\}^\infty$. An example of this would be if we want to examine $A = \{\omega : \frac{1}{n}\sum_{i=1}^n X_i(\omega) \to \frac{1}{2}\}$ as $n \to \infty$. If we want to show that $\bar{X}_n \overset{a.s.}{\to} \frac{1}{2}$ as $n \to \infty$, we need to show that $P(A) = 1$. Therefore, we need to figure out the probability of an infinite dimensional event, because as $n \to \infty$, we have to calculate probabilities of an infinite amount of coin tosses.

We can see that using the typical calculus-based probability approach won't work well here. For example, if we were looking at a finite number of coin tosses, the probability of our sample space Ω_m can be found by $P(\Omega_M) = \sum_{\omega_m \in \Omega_m} P(\{\omega_m\})$. Basically, to calculate the probability of an event, we take the probability of each event and sum them up to get the total probability. However, if we think of an infinite amount of coin tosses, the probability of an event ω will be $P(\{\omega\}) = \prod_{i=1}^\infty P(X_i = x_i) = (\frac{1}{2})^\infty = 0$. So if we try to calculate probabilities like we normally would, say $P(\Omega) = \sum_{\omega \in \Omega} P(\{\omega\})$, then we would have two problems. The first is that every single $P(\{\omega\}) = 0$. The second is that we actually cannot write the summation because we can only write summations for finite or countably infinite amount of terms, but in this case we have an uncountably infinite number of terms.

Therefore, we need to find a better way to assign probabilities to infinite dimensional events and infinite dimensional sample spaces, which is to use measure-theoretic probabilities!

In measure theoretic integration, for a sample space Ω, we have that:

- $P(A) \geq 0$
- $P(\Omega) = 1$
- $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$ for A_i disjoint

Using these sample space properties, we have that $\int_\Omega \mathbb{1}(A, \omega) P(d\omega) = E[\mathbb{1}(A)] = P(A)$. We can then see that $E[\sum_{i=1}^m c_i \mathbb{1}(A_i)] = \sum_{i=1}^m c_i P(A_i)$. Therefore, we can define $X = lim_{m\to\infty} \sum_{i=1}^\infty c_{im} \mathbb{1}(A_{im})$. Then we can calculated the expected value of X as $E[X] = \int_\Omega X(\omega) P(d\omega) = lim_{m\to\infty} E[\sum_{i=1}^m c_{im} \mathbb{1}(A_{im})]$. This allows us to work with a much more general and abstract sample space.

The one subtlety that arises when we use measure theoretic probability to do assignments is that we may not be able to consistently assign probabilities to every single subset of Ω (non-measurable subsets). There may be some subsets where the probability is not defined.

Let's take a closer look at how we would assign probabilities to infinite dimensional sample spaces. Revisitng the coin toss example, we have $\Omega = \{0,1\}^\infty$. If we take a finite dimensional subset, where $P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m) = 2^{-m}$, we know how to assign a probability to that. As it turns out, there is a theorem that states that if we can assign probabilities to all finite dimensional events, then there is a unique way to extend assignments to infinite dimensional events (Kolmogorov Extension Theorem). In practice then, we can assign probabilities to the finite dimensional events, then this extension theorem tells us that there is a unique probability on the infinite dimensional sample space that is consistent with that finite dimensional assignment.

5.1 Interchange Limit Results

In many settings, we need to check that we can interchange expectations and limits, because this is not always the case.

5.1.1 Bounded Convergence Theorem

The Bounded Convergence Theorem tells us that if $X_n \overset{a.s.}{\to} X_\infty$, and $|X_n(\omega)| \leq c < \infty$ for all ω and $n \geq 1$, then $E[X_n] \to E[X_\infty]$. Proof is below:

$|E[X_n] - E[X_\infty]|$
$= |E[X_n - X_\infty]|$
$\leq E[|X_n - X_\infty|]$
$= E[(|X_n - X_\infty|)\mathbb{1}(|X_n - X_\infty| \leq \epsilon)] + E[(|X_n - X_\infty|)\mathbb{1}(|X_n - X_\infty| > \epsilon)]$
$\leq \epsilon P(|X_n - X_\infty| \leq \epsilon)) + 2cP(|X_n - X_\infty| > \epsilon))$, since X_n and X_∞ are both less than c
$\leq \epsilon$

since the probability in the first term is bounded by 1, and the second term goes to 0 since almost sure convergence implies convergence in probability.

Therefore, $limsup|E[X_n] - E[X_\infty]| \leq \epsilon$. Since ϵ can be made arbitrarily small, $limsup|E[X_n] - E[X_\infty]| = 0$. Therefore, $E[X_n] \to E[X_\infty]$.

5.1.2 Monotone Convergence Theorem

The Monotone Convergence Theorem states that if $X_n(\omega) \geq 0$ and $X_n(\omega)$ increases monotonically to $X_\infty(\omega)$ as $n \to \infty$ for all $\omega \in \Omega$, thne $E[X_n]$ increases monotonically to $E[X_\infty]$. Note that $X_\infty(\omega)$ can be finite or $+\infty$.

5.1.3 Dominated Convergence Theorem

If $X_n \overset{a.s.}{\to} X_\infty$ and $|X_n(\omega)| \leq Y(\omega)$ for all $\omega \in \Omega$, $n \geq 1$, and $E[Y] < \infty$, then $E[X_n] \to E[X_\infty]$ as $n \to \infty$.

5.1.4 Fatou's Lemma

If $X_n \geq 0$, then $E[liminf_{n\to\infty} X_n] \leq liminf_{n\to\infty} E[X_n]$.

5.1.5 Moving Results from Almost Sure to Distribution

We can use Skorohod's Representation Theorem to move these results from requiring almost sure convergence to convergence in distribution. This would be useful in a circumstance where we have $X_n \Rightarrow X_\infty$ and we want to show that $E[X_n] \to E[X_\infty]$. By Skorohod's representation theorem, we know that if $X_n \Rightarrow X_\infty$, then there exists a probability space supporting random variables $(X'_n : n \geq 1)$ such that $X'_n \overset{D}{=} X_n, 1 \leq n \leq \infty$ and that $X'_n \overset{a.s.}{\to} X'_\infty$.

Therefore, if we know that $X_n \Rightarrow X_\infty$, then we know that we can find $X'_n \overset{a.s.}{\to} X'_\infty$ by Skorohod's Representation Theorem. By the Bounded Convergence Theorem, $E[X'_n] \to E[X'_\infty]$. Therefore, $E[X_n] \to E[X_\infty]$ since the X'_n's have the same distribution as the X_n's.

One other important result is as follows. If $X_n \Rightarrow X_\infty$, then again by Skorohod's Representation Theorem, $X'_n \overset{a.s.}{\to} X_\infty$. Then, for any continuous function h, $h(X'_n) \overset{a.s.}{\to} h(X'_\infty)$. Therefore, by the Bounded Convergence Theorem, $E[h(X'_n)] \to E[h(X'_\infty)]$ as $n \to \infty$, and therefore $E[h(X_n)] \to E[h(X_\infty)]$ as $n \to \infty$ for any bounded and continuous function h. The converse also turns out to be true. If $E[h(X_n)] \to E[h(X_\infty)]$ as $n \to \infty$ for all bounded and continuous functions h, then $X_n \Rightarrow X_\infty$.

Therefore, what we have established is that the following are equivalent:

- $X_n \Rightarrow X_\infty$
- There exists a probability space supporting random variables $(X'_n : n \geq 1)$ such that $X'_n \overset{D}{=} X_n, 1 \leq n \leq \infty$ and that $X'_n \overset{a.s.}{\to} X'_\infty$
- $E[h(X_n)] \to E[h(X_\infty)]$ as $n \to \infty$ for all bounded and continuous functions h

5.2 Proof of Strong Law of Large Numbers

We will now proceed to prove the Strong Law of Large Numbers. We assume $X_1, X_2, ...$ iid with $E[X_1^4] < \infty$. It can actually also be proven under the weaker assumption that $E|X_1| < \infty$. First we will center the X_i's such that $\tilde{X}_i = X_i - E[X_1]$. Denote $\tilde{S_n} = \tilde{X_1} + \tilde{X_2} + ... + \tilde{X_n}$. We want to show that $\frac{\tilde{S_n}}{n} \overset{a.s.}{\to} 0$.

To prove almost sure convergence, we need to show that for every $\epsilon > 0$, there exists a $N(\epsilon)$ such that for any $n \geq N(\epsilon)$, that $|\frac{\tilde{S_n}}{n} - 0| < \epsilon$. If we let $W = \sum_{i=1}^{\infty} \mathbb{1}(|\frac{\tilde{S_n}}{n}| > \epsilon) < \infty$, and we prove that $E[W] < \infty$, that is equivalent to proving $P(W < \infty) = 1$, which is the same as the above for almost sure convergence.

$E[W] = E[\sum_{i=1}^{\infty} \mathbb{1}(|\frac{\tilde{S_n}}{n}| > \epsilon)]$

Fubini's Theorem states that if $Z_n \geq 0$, then $E[\sum_{i=1}^{\infty} Z_n] = \sum_{i=1}^{\infty} E[Z_n]$.

Therefore, since $\frac{\tilde{S_n}}{n}$ is non-negative, we have that

$E[W] = E[\sum_{i=1}^{\infty} \mathbb{1}(|\frac{\tilde{S_n}}{n}| > \epsilon)]$
$= \sum_{i=1}^{\infty} E[\mathbb{1}(|\frac{\tilde{S_n}}{n}| > \epsilon)]$
$= \sum_{i=1}^{\infty} P(|\frac{\tilde{S_n}}{n}| > \epsilon)$

Using Markov's Inequality, $\sum_{i=1}^{\infty} P(|\frac{\tilde{S_n}}{n}|^4 > \epsilon^4) \leq \sum_{i=1}^{\infty} \frac{1}{\epsilon^4} E[[\frac{\tilde{S_n}}{n}]^4]$

Let's take a closer look at the value of $E[[\sum_{i=1}^{n} \tilde{X_i}]^4]$. When we expand out the polynomial, we will get terms of the form $\tilde{X_i}^4$, $\tilde{X_i}^3\tilde{X_j}$, $\tilde{X_i}^2\tilde{X_j}^2$, $\tilde{X_i}\tilde{X_j}\tilde{X_k}^2$, and $\tilde{X_i}\tilde{X_j}\tilde{X_k}\tilde{X_l}$. When we take the expected value of these terms, any term with a $\tilde{X_i}$ in it to the power of 1 will equal 0 since it is mean centered and indepedendent. For example, $E[\tilde{X_i}^3\tilde{X_j}] = E[\tilde{X_i}^3]E[\tilde{X_j}] = E[\tilde{X_i}^3] * 0 = 0$. Therefore, the only terms remaining are that we have $O(n)$ of $\tilde{X_i}^4$ terms, and $O(n^2)$ of $\tilde{X_i}^2\tilde{X_j}^2$ terms.

Therefore, now we have that the right side of the Markov Inequality, $\sum_{i=1}^{\infty} \frac{1}{\epsilon^4} E[[\frac{\tilde{S_n}}{n}]^4] \leq \sum_{i=1}^{\infty} \frac{O(n^2)}{n^4} * \frac{1}{\epsilon^4} < \infty$. Therefore, we know that $\sum_{i=1}^{\infty} P(|\frac{\tilde{S_n}}{n}|^4 < \infty$. This means that $N(\epsilon) < \infty$ with probability 1 for every ϵ.

What remains to be shown is the for every $\epsilon > 0$, our $N(\epsilon) < \infty$. If we let $\epsilon = \frac{1}{m}$, we only need to check this for $m \geq 1$, that $N(\frac{1}{m}) < \infty$. Let $B_m = \{\omega : N(\frac{1}{m}) < \infty\}$. The $P(B_m) = 1$ for $m \geq 1$, and therefore $P(\cap_{i=1}^{\infty} B_m) = 1$. Therefore, if we call the event $A = \{\omega : N(\epsilon, \frac{1}{m}) < \infty, m \geq 1\}$, then $P(A) = 1$. A is equivalent to $\{\omega : \bar{X}_n(\omega) \to E[X_1]$ as $n \to \infty\}$. Therefore, since $P(A) = 1$, we have proved the Strong Law of Large Numbers.

6 Conditional Expectation

Conditional expectation and conditional probability is essential in any study of probability, and is the basis for inference and causation. In the early 20th century, there was development of measure theoretic probability related to conditional expectation to move from the calculus-based approach to one that can account for more complex conditions.

Looking at the calculus-based probability approach, we can calculate a conditional expectation as follows:

$E[Y|\vec{Z} = \vec{z}]$
$= \int_{-\infty}^{\infty} y f_{Y|\vec{Z}}(y|\vec{z})dy$
$= \int_{\infty}^{\infty} y \frac{f_{Y,\vec{Z}}(y,\vec{z})}{f_{\vec{Z}}(\vec{z}}dy$

To perform this calculation, we need to know the joint density of Y and $\vec{Z}$. This works for a finite collection of random variables, but when we have an infinite collection of random variables, we don't have this joint density. A situation where we would have an infinite collection of random variables as $\vec{Z}$, could be a time series. We may want to know what the probability of something is, conditional on all the times prior, such as $P(X(s+u) \in A|X(r) : 0 \leq r \leq s)$.

Let's look at an example that shows us that we cannot use a calculus-based method to get the joint density of an infinite collection of random variables. Let $W_1, W_2, ...$ iid, and let the joint density of them be $f_W(w_1, w_2, ...) = \prod_{i=1}^{\infty} f_W(w_i)$. Suppose $W \in [0, 1]$ and we draw the random variable from a uniform. Therefore, we have $\prod_{i=1}^{\infty} f_W(u_i)$.

Let us take the log and divide my m. We then have $\frac{1}{m} log(\prod_{i=1}^{\infty} f_W(u_i)) = \frac{1}{m} \sum_{i=1}^{m} log f_W(u_i) \overset{a.s.}{\to} E[log f_W(u)]$ by SLLN.

Let us also take a look at Jensen's Inequality, which states that if g is a convex function and R is a random variable, then $g(E[R]) \leq E[g(R)]$. We can prove this by first looking at the property of a convex function. If g is convex, then $g(\sum_{i=1}^{k} p_i x_i) \leq \sum_{i=1}^{k} p_i g(x_i)$, assuming $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. This is saying that the p_i's are a weighted average of the x_i's. If we think about a parabola (convex function), the average of 2 points on the parabola is going to be larger or equal to a point in between these 2 points on the parabola. We can then let the x_i be the R_i's where R_i are iid copies of the random variable R, and $p_i = \frac{1}{n}$. By convexity, $g(\frac{1}{n} \sum_{i=1}^{n} R_i) \leq \frac{1}{n} \sum_{i=1}^{n} g(R_i)$. Then by using the SLLN, $g(E[R]) \leq E[g(R)]$.

Going back to our case, log is a concave function, not convex, so the inequality flips around. Therefore,

$E[log f_W(u)] \leq log E[f_W(u)]$
$= log \int_0^1 f_W(u)du$ since U and W are in [0,1]
$= log(1)$
$= 0$

Actually, unless the variable is degenerate, we can change the inequality to a strict inequality, so $E[log f_W(u)] < 0$. Therefore, $log(\prod_{i=1}^{m} f_W(u)) \to -\infty$ as $m \to \infty$. Therefore, $\prod_{i=1}^{\infty} f_W(u_i) = 0$. This does not make sense since the joint density must integrate to 1 by definition, therefore we have proved that traditional calculus-based approach does not work for this infinite collection. We must have an alternate way of looking at joint densities.

6.1 Prediction Theory

The goal is to predict Y, based on having observed $\vec{Z}$ and a probability distribution jointly describing Y and $\vec{Z}$. Using measure theoretic probability, we can think of this as a probability space supporting Y and $\vec{Z}$ directly. Therefore, having observed a value of $\vec{Z}$, we want to predict Y. In other words, we have a deterministic h, and we want $\hat{Y} = h(\vec{Z})$. We want to find a $\hat{Y}$ that is close to Y, and to do that we need to first define what "closeness" means.

We can define "closeness" by using an L^P norm. An L^P norm is defined as $||w||_p = (E[|w|]^p)^{\frac{1}{p}}$, where $E[|w|]^p < \infty$, and $p \geq 1$. Properties of an L^P norm are:

- $||a \cdot w||_p = |a|||w||_p$
- $||w_1 + w_2||_p \leq ||w_1||_p + ||w_2||_p$ (Minkowski's inequality)

We define the distance between w_1 and w_2 as $||w_1 - w_2||_p$. Therefore, we want to find the $\hat{Y}$ that minimizes $||Y - \hat{Y}||_p$. In other words, we want to find a random variable $\hat{Y} = g^*(\vec{Z})$ which minimizes $||Y - g(\vec{Z})||_p$ over all deterministic $g(\cdot)$ such that $E[|g(\vec{Z})|]^p < \infty$.

In the simple case with no observable $\vec{Z}$, the problem reduces to finding a deterministic constant a^* such that $||Y-a||_p$ is minimized over all possible choices of $a \in \mathbb{R}$.

In the case where $p=2$, we have that $||Y-a||_2 = \sqrt{E[|Y-a|]^2}$. Minimizing this is equivalent to minimizing $E[|Y-a|]^2 = E[Y^2] - 2aE[Y] + a^2$. Taking the derivative and setting it equal to 0 to find the minimum, we get that $-2E[Y]+2a=0$, and therefore $a^* = E[Y]$. Therefore, in the case with observable $\vec{Z}$, we want to use $E[Y|\vec{Z}]$ to minimize the distance.

In the case where $p=1$, we have that $||Y-a||_1 = E[|Y-a|]$.

$E[|Y-a|] = \int_{-\infty}^{\infty} |y-a| f_Y(y)dy$
$= \int_{-\infty}^{a} (a-y) f_Y(y)dy + \int_a^{\infty} (y-a) f_Y(y)dy$

Taking the derivative and set it equal to 0, we get:

$\frac{d}{da}[a\int_{-\infty}^{a} f_Y(y)dy - \int_{-\infty}^{a} y f_Y(y)dy + \int_a^{\infty} y f_Y(y)dy - \int_a^{\infty} a f_Y(y)dy]$
$= \int_{-\infty}^{a} f_Y(y)dy + a f_Y(a) - a f_Y(a) - a f_Y(a) - \int_a^{\infty} f_Y(y)dy + a f_Y(a)$
$= \int_{-\infty}^{a} f_Y(y)dy - \int_a^{\infty} f_Y(y)dy$
$= F_Y(a) - (1 - F_Y(a))$
$= 2F_Y(a) - 1 = 0$

Therefore, $F_Y(a^*) = \frac{1}{2}$. In other words, a^* is the median of the distribution of Y.
We can see that the choice of p matters when assessing the prediction error. When $p=1$, we will penalize more at closer values to the middle, and when $p=2$, we will penalize more at the tail ends.

6.2 Introducing Geometry when p = 2

In the L^2 space, we have special geometric properties that may not be true in other dimensions. For example, we can define $< W_1, W_2 > \stackrel{\Delta}{=} E[W_1 W_2]$ in this space. In addition, we can define a Hilbert space: $\mathscr{H} = \{g(Z) \in L^2 : g \text{ is deterministic}\}$. This Hilbert space we have just defined is a linear subspace of L^2 and is closed. Therefore, when we want to talk about $w_i = g_i(z) \to w_\infty \in L^2$, we can represent both w_i and w_∞ as $g(z) \in \mathscr{H}$. If we let $w \in \mathscr{H}$, then our best prediction $\hat{Y}$ is the point in $\mathscr{H}$, such that $< Y - \hat{Y}, w >= 0$.

The Hilbert Space Projection Theorem states the following. Suppose $\mathscr{H}$ is a closed linear subspace of L^2, then there exists a unique random variable $\hat{Y}$ which minimizes $||Y-w||_2$ over $w \in \mathscr{H}$, and that minimizer $\hat{Y}$ is characterized by $< Y - \hat{Y}, w >= 0$.

Therefore, in L^2, we define $\hat{Y} = E[Y|Z]$, and we only need to assume that $E[Y]^2 < \infty$. This statement holds for any Z, whether it is countable or uncountable. This assumption can actually be extended such that this statement holds if $E[|Y|] < \infty$, $Y \in L^1$. We can show the following.

$E[(Y-\hat{Y})w] = 0$
$E[Yw] = E[\hat{Y}w]$ for every $w \in \mathscr{H}$
$E[Yg(Z)] = E[\hat{Y}g(Z)]$ for every deterministic g such that $g(Z) \in L^2$

Therefore, the expectations are equal and $E[\hat{Y}g(Z)] = g^*(\vec{Z})$. It can be shown that by performing the calculus based method of calculating conditional probability, where $E[Y|\vec{Z} = \vec{z}] = \int_{-\infty}^{\infty} y f_{Y|\vec{Z}}(y|\vec{z})dy$, produces the same $g^*(\vec{Z})$. Therefore, the measure-theoretic way of looking at conditional expectation is a legitimate extension of the calculus-based approach. We can always use the measure-theoretic approach, and we can use the calculus-based approach if we have the joint density.

6.3 Properties of Conditional Expectation in L^2

The following properties are true for the L^2 space.

- $E[Y_1 + Y_2|Z] = E[Y_1|Z] + E[Y_2|Z]$
- If $Y \geq 0$, $E[Y|Z] \geq 0$
- $E[Yg(Z)|Z] = g(Z)E[Y|Z]$
- $E[Y|V] = E[E[Y|Z]|V]$ where $V = g(Z)$ (Tower Property)

7 Large Deviations and Change of Measure

This section will talk about rare event probability theory, specifically a subfield called the theory of large deviations. Say we have $X_1, X_2, ...X_n$ iid and we denote $S_n = X_1 + X_2 + ... + X_n$. S_n, when viewed as a function of n is a random walk. We know from the SLLN that if $E[|X_1|] < \infty$, then $\frac{1}{n}S_n \overset{a.s.}{\to} E[X_1]$. Specifically, if $E[|X_1|] = 0$, then $\frac{S_n}{n} \overset{a.s.}{\to} 0$. In other words, for $a > 0$, $P(\frac{S_n}{n} > a) \to 0$ as $n \to \infty$, and we can rewrite it as $P(S_n > na) \to 0$ as $n \to \infty$. $P(S_n > na)$ is what we call a rare event. As n gets large, the probability of this event occurring becomes rarer and rarer. What we would like to understand is how fast does $P(S_n > na) \to 0$?

When we look at the CLT, all it tells us is that $P(\frac{S_n}{\sqrt{n}\sigma} > \frac{na}{\sqrt{n}\sigma}) - P(N(0,1) > \sqrt{n}\frac{a}{\sigma}) \to 0$, but we already knew that the probability converges from the SLLN. We may hope for a stronger result such that the CLT approximates the rare event at the tail, such as $\frac{P(S_n > na)}{P(N(0,1) > \sqrt{n}\frac{a}{\sigma})} \to 1$, but this actually is not always true. One example where this fails is if we take a coin toss where a head gets us a value of 1 and a tails a -1, both with probability $\frac{1}{2}$. If we select a value of $a > 1$, then $P(S_n > na) \leq P(S_n > n) = 0$, since it obviously cannot happen. However, $P(N(0,1) > \sqrt{n}\frac{a}{\sigma}) > 0$ for all values of n, since the normal distribution has positive densities for all values from $-\infty$ to ∞. Therefore, this ratio is always 0 and never converges to 1. The CLT approximation fails in this setting to approximate the rare events. Therefore, we clearly need a new approximation to handle such large deviation probabilities.

When we look at $P(S_n > n^\delta a)$, we are able to use CLT when $\delta \in [0, \frac{1}{2}]$ to approximate the probability. As we have seen, when $\delta = 1$, we have what we call a large deviations probability, where we cannot use the CLT. When $\delta \in (\frac{1}{2}, 1)$, we are in the moderate deviations setting. It can be proven that when $\delta \in (\frac{1}{2}, \frac{2}{3})$, that the CLT can be used for approximation and that $\frac{P(S_n > n^\delta a)}{P(N(0,1) > n^{-\frac{1}{2}\delta}\frac{a}{\sigma})} \to 1$.

We want to take a closer look at what happens in the large deviations setting if $S_n = na$. Assuming that $a > 0$ and that $E[X_1] = 0$, we want to take a look at $E[X_j | S_n = na]$ for $1 \leq j \leq n$. We have that $\sum_{j=1}^n E[X_j | S_n = na] = na = nE[X_1 | S_n = na]$. Therefore, $a = E[X_1 | S_n = na]$. Therefore, what we are saying is that in this setting, it is really hard for us to even get to na, so when we are at the level of na or higher, we are mostly likely only slightly above na; we are most likely very close. Therefore, $P(S_n > na) \approx P(S_n = na)$. What we are seeing is mostly likely a slow drift up to na, as opposed to any of the X_i making a huge jump.

For a rare event, if we want to estimate $P(S_n > na)$ by Monte Carlo, we will see that it takes an astronomical amount of simulations to do so. In order to see the first rare event, it would take us on average $\frac{1}{p}$ simulations to do so. We know this because we can think about this is a geometric random variable, where success is seeing the rare event.

More precisely, if we were to perform Monte Carlo to generate indicator random variables $\mathbb{1}_1, \mathbb{1}_2, ..., \mathbb{1}_n$, where our probability of interest is $\alpha_n = \frac{1}{n}\sum_{j=1}^n \mathbb{1}_j$, by CLT we would have that $\alpha_n \overset{D}{\approx} \alpha + \sqrt{\frac{\alpha(1-\alpha)}{n}}N(0,1)$. This means that our relative error is distributed $\frac{\alpha_n}{\alpha} - 1 \overset{D}{\approx} \sqrt{\frac{1-\alpha}{\alpha n}}N(0,1)$. Therefore, we converge as $\alpha n \to \infty$, and therefore, we need to pick n very large relative to $\frac{1}{\alpha}$, which would be an astronomical number if α is small. Therefore, instead of crude Monte Carlo, we want to instead turn to a method called importance sampling.

7.1 Importance Sampling

Say we want to obtain the value of $\alpha = \int_{-\infty}^{\infty} r(z)dz$. We are expressing this as an expected value so that we can use Monte Carlo, but in the case of where we want to find a probability, we can just have an indicator function as $r(z)$. For any function $f(z)$, we have that $\int_{-\infty}^{\infty} r(z)dz = \int_{-\infty}^{\infty} \frac{r(z)}{f(z)}f(z)dz = E_F[\frac{r(z)}{f(z)}]$. Therefore, our process is as follows. We can generate $Z_1, Z_2, ..., Z_n$ from the density $f(\cdot)$, and then calculate $\alpha_n = \frac{1}{n}\sum_{i=1}^n \frac{r(z_i)}{f(z_i)}$. Since this process works for any function $f(z)$, we want to select a good choice of f. Good choices of f are functions where $f(z)$ is large where $r(z)$ is large. In other words, we want to sample more from the regions that are more important for the integration of $r(z)$!

We can show that this is true by the following. We claim that the optimal $f^*(x) \propto |r(x)|$. Therefore, $f^*(x) = \frac{|r(x)|}{\int_{-\infty}^{\infty}|r(y)|dy} = \frac{|r(x)|}{\tilde{\alpha}}$. The purpose of the denominator is just a normalizing constant so that the density integrates to 1. Note that if r is non-negative, $\int_{-\infty}^{\infty}|r(y)|dy = \int_{-\infty}^{\infty} r(y)dy = \alpha$. Therefore, if r is non-negative, $\tilde{\alpha} = \alpha$. Our optimal choice is the one that minimizes the variance. Therefore, we want to minimize, $Var_F[\frac{r(z)}{f(z)}] = E_F[\frac{(r(z))^2}{(f(z))^2}] - (E[\frac{r(z)}{f(z)}])^2$. Note that the second term on the right hand side is equal to α^2, which is a constant, therefore minimizing the variance is equivalent to minimizing the second moment, $E_F[\frac{(r(z))^2}{(f(z))^2}]$.

$E_F[\frac{(r(z))^2}{(f(z))^2}] \geq E_F[\frac{|r(z)|}{f(z)}]^2$, since $E[W^2] \geq (E[W])^2$ since variance is always non-negative
$= (\int_{-\infty}^{\infty} \frac{|r(z)|}{f(z)} f(z)dz)^2$
$= (\int_{-\infty}^{\infty} |r(z)|dz)^2$
$= \tilde{\alpha}^2$

However, our choice of $f^*(z) = \frac{r(z)}{\tilde{\alpha}}$. Plugging it into the second moment, we get that $E_F[\frac{(r(z))^2}{(f^*(z))^2}] = \tilde{\alpha}^2$. Therefore, our choice of $f^*(z)$ does indeed give us the lowest bound on the variance and we cannot do any better than this.

Let's look at the scenario where we specifically want to compute $\alpha = P(W > w)$. Let's assume that W takes on the density $g(\cdot)$ and that it is always non-negative. Then we have $P(W > w) = \int_{-\infty}^{\infty} r(z)dz$, where $r(z) = \mathbb{1}(Z \geq w)g(z)$. From above, we choose our function $f^*(z) = \frac{|r(z)|}{\tilde{\alpha}} = \frac{r(z)}{\tilde{\alpha}} = \frac{\mathbb{1}(Z \geq w)g(z)}{P(W \geq w)} = P(W \in dz | W \geq w) = \frac{\mathbb{1}(Z \geq w)g(z)dz}{P(W \geq w)}$. Therefore, in the general case, if we want to compute $\alpha = P(A)$, and we want to use importance sampling using an alternative sampling distribution $\tilde{P}$, then the optimal alternative sampling distribution is $P^*(d\omega) = P(d\omega|A) = \frac{\mathbb{1}(A,\omega)P(d\omega)}{P(A)}$. If we use P^*, then the associated variance is 0 and the corresponding importance sampling estimator is almost surely. Obviously we don't know the distribution, so we cannot select f^* in practice, but we try to use an alternative sampling distribution which we believe is "close" to the conditional distribution.

7.2 Moment Generating Functions

If we have a random variable Γ, we define the moment generating function of Γ as $\varphi(\theta) \triangleq E[e^{\theta\Gamma}]$. If Γ has a density $f(\cdot)$, then $\varphi(\theta) = \int_{-\infty}^{\infty} e^{\theta z} f_\Gamma(z)dz$. When we take derivatives of the moment generating function, we get that $\varphi^{(k)}(\theta) = \frac{d^k}{d\theta^k} E[e^{\theta\Gamma}] = E[\frac{d^k}{d\theta^k} e^{\theta\Gamma}] = E[\Gamma^k e^{\theta\Gamma}]$. When we plug in $\theta = 0$, we get that $\varphi^{(k)}(0) = E[\Gamma^k]$. Therefore, we can generate the moments of Γ using this function and plugging in 0. Note that we can do the limit interchange based on Dominated Convergence Theorem if $\varphi(\theta) < \infty$ in a non-zero neighborhood of 0.

If we have the case where $P(\Gamma > z) \leq e^{-\theta z} E[e^{\theta\Gamma}]$, where $\theta > 0$, we have that the right tail is going to 0 exponentially fast if the expectation is finite. Similarly, when $P(\Gamma < -z) \leq e^{-\theta z} E[e^{\theta\Gamma}]$, where $\theta < 0$, the left tail is going to 0 exponentially fast. When we have left and right tails going to 0 exponentially fast, this is what is called a "light-tailed" random variable. Examples of this include the normal, gamma, exponential, Poisson, and uniform distributions. However, not every random variable is light-tailed. For some variables, $P(\Gamma > z) \approx cz^{-\alpha}$, therefore it goes to 0 slower than exponentially. Instead it goes to 0 algebraically, and these are called "power law tails." These variables are called "heavy-tailed" random variables. The theory of heavy-tailed random variables is quite different than the theory of light-tailed random variables and in many cases, the structural models we observe from these two types of random variables is quite different.

7.3 Upper Bound for Large Deviations

Again we have a large deviations event such that $P(S_n > na)$, where the X_i's are iid with $E[X_1] = 0$. We will try to obtain an upper bound for this probability using an exponential upper bound. An exponential upper bound is useful in a situation where we have a random variable V, and we are looking for $P(V > v)$, and we assume that $E[exp(\theta V)] < \infty$, where $\theta > 0$. We claim that $P(V > v) \leq e^{-\theta v} E[exp(\theta v)]$. Since $E[exp(\theta v)]$ is a constant, we are saying that it equals $ce^{-\theta v}$. Therefore, the right tail of V decays to 0 exponentially fast as a result of this exponential bound. Using Markov's Inequality, we can see that

$P(V > v) = P(\theta V > \theta v)$, since $\theta > 0$
$= P(e^{\theta V} > e^{\theta v})$
$\leq \frac{E[e^{\theta V}]}{e^{\theta v}}$

When we apply the exponential bound to S_n, we see that

$P(S_n > na) \leq e^{-\theta na} E[exp(\theta S_n)]$
$= e^{-\theta na} E[e^{\theta \sum_{i=1}^n X_i}]$
$= e^{-\theta na} E[\prod_{i=1}^n e^{\theta X_i}]$
$= e^{-\theta na} \prod_{i=1}^n E[e^{\theta X_i}]$
$= e^{-\theta na} (E[e^{\theta X_1}])^n$

Therefore, we have $P(S_n > na) \leq e^{-\theta na} (E[e^{\theta X_1}])^n$. We can let $\psi(\theta) = logE[exp(\theta X_1)] = log[\varphi(\theta)]$, where $\varphi(\theta)$ is the moment generating function. Therefore we can write $P(S_n \geq na) \leq exp(-n(\theta a - \psi(\theta)))$ for all $\theta > 0$. We want to pick the θ that minimizes the upper bound, therefore we want to solve $\frac{d}{d\theta}(\theta a - \psi(\theta)) = 0$. θ is a function of a, so the optimal choice is when $\psi'(\theta(a)) = a$.

Recall that $\psi(\theta) = log(\varphi(\theta))$. Therefore,

$\psi'(\theta) = \frac{\varphi'(\theta)}{\varphi(\theta)}$
$= \frac{E[e^{\theta X_1} X_1]}{E[e^{\theta X_1}]}$
$= E[e^{\theta X_1 - \varphi(\theta)} X_1]$
$= \int_{-\infty}^{\infty} x e^{\theta x - \varphi(\theta)} f(x) dx$

We claim that $e^{\theta x - \varphi(\theta)} f(x)$ is a density, in other words it integrates to 1. Denote this quantity as $f_\theta(x)$. Then we have $\int_{-\infty}^{\infty} f_\theta(x) dx = e^{-\varphi(\theta)} \int_{-\infty}^{\infty} e^{\theta x} f_X(x) dx = e^{-\varphi(\theta)} E[e^{\theta x}] = \frac{1}{E[e^{\theta x}]} * E[e^{\theta x}] = 1$. We have proved that it is a density. From before, we have

$\psi'(\theta) = \int_{-\infty}^{\infty} x e^{\theta x - \varphi(\theta)} f(x) dx$
$= E_\theta[X]$.

Therefore, the optimal choice of θ is when $\psi'(\theta(a)) = E_{\theta(a)}[X] = a$. We can see that the tilted distribution $f_{\theta(a)}(\cdot)$ plays a key role in the large deviations event $(S_n > na)$.

Let us try to compute $P(S_n > na)$ using a change of measure in which the increments remain iid but are sampled from $f_{\theta(a)}(\cdot)$. Therefore, when we sample from our alternative distribution $\tilde{P}$, we have that

$\tilde{P}(X_1 \in dx_1, X_2 \in dx_2, ..., X_n \in dx_n)$
$= \prod_{i=1}^n \tilde{P}(X_i \in dx_i)$
$= \prod_{i=1}^n f_{\theta(a)}(x_i) dx_i$
$= \prod_{i=1}^n e^{\theta(a) x_i - \psi(\theta(a))} f_X(x_i) dx_i$
$= E[e^{\theta(a) \sum_{i=1}^n x_i - \psi(\theta(a))} \mathbb{1}(X_1 \in dx_1, ..., X_n \in dx_n)]$

Note that in the middle of our derivation, we had that $\tilde{P}(X_1 \in dx_1, X_2 \in dx_2, ..., X_n \in dx_n) = \prod_{i=1}^n e^{\theta(a) x_i - \psi(\theta(a))} f_X(x_i) dx_i$. Since the exponential is always non-zero, we can divide both sides by that and we have the identity that $\prod_{i=1}^n f_X(x_i) dx_i = \tilde{P}(X_1 \in dx_1, X_2 \in dx_2, ..., X_n \in dx_n) * e^{-\theta(a) \sum_{i=1}^n x_i + n\psi(\theta(a))}$. Now we can get an expression for the original distribution in terms of the alternative distribution, as well as from the alternative to the original.

$P(S_n > na) = \tilde{E}[e^{-\theta(a) S_n + n\psi(\theta(a))} \mathbb{1}(S_n > na)]$

$\tilde{P}(S_n > na) = E[e^{\theta(a) S_n - n\psi(\theta(a))} \mathbb{1}(S_n > na)]$

Now that we have an understanding, we can use importance sampling as follows. We generate W using $X_1, X_2, ..., X_n$ that are iid with a common density $f_{\theta(a)}(\cdot)$. We replicate W a number of times, thereby generating $W_1, W_2, ..., W_m$. We can estimate $P(S_n > na)$ via $\bar{W}_m = \frac{1}{m} \sum_{i=1}^m W_i$. When we simulate these W_i's, we're effectively simulating from the conditional distribution of the X_i's given $S_n > na$.

Therefore, in this light-tailed case, we can put an exponential upper bound on the rare event and we would anticipate seeing iid values where each one is gradually moving towards na. In the heavy tailed case, we would actually expect to see the opposite, where we would have many values close to the mean, and only a few observations be so large that the move the entire total up to na.

8 Introduction to Statistics

In statistics, we are given a set of explanatory variables, $\vec{x_1}, \vec{x_2}, ..., \vec{x_n}$ and we want to predict responses $Y_1, Y_2, ...Y_n$. A typical statistical model can have the form $Y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + ...$ In matrix form, this can take the form $Y = Xa$.

In least squares, we can form an estimate from the data by minimizing $\sum_{i=1}^{n}(Y_i - X_i a)^2$. In ordinary least squares, in matrix form, that is the equaivalent to minimizing $(Y - Xa)^T(Y - Xa)$.

We can generalize this to the form $(Y - Xa)^T H(Y - Xa)$ where H is a SPD matrix. In the ordinary least squares example, $H = I$. In the setting of the least squares problem we want to use statistics for the following:

- Determine the rationale for choosing H
- Attach error bars to our estimate of the vector $\vec{a}$: $\hat{\vec{a}} \pm$ CI
- Hypothesis testing can be used to assess whether we can explain the data equally well through a linear model with fewer variables

If we view $x_1, x_2, ...x_n$ as realized values of random variables from an unknown population distribution F^*, then we can define two types of statistical models of the observed data:

- Non-parametric: Let F^* be general; no structure imposed on F^*
- Parametric: Impose structure on F^*; assume that the population is normally distributed $N(\mu_0, \sigma_0^2)$

8.1 Parametric Statistics

In parametric statistics, we assume that the $X_1, X_2, ..., X_n$ come from a distribution characterized by $\vec{\theta}$. For example, in a normal distribution, $\vec{\theta} = (\mu, \sigma^2)$, and thus $f(\vec{\theta}, x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$. In an exponential, $\vec{\theta} = \lambda$, and thus $f(\vec{\theta}, x) = \lambda e^{-\lambda x}$.

With a frequentist point of view, we want to find an estimator(s) $\hat{\theta}$ for the true value of the parameter(s) θ. This takes the viewpoint that $\hat{\theta}$ is a random variable, since it is a function of $X_1, X_2, ..., X_n$. Since $\hat{\theta}$ is itself a random variable, it has its own pdf.

In order to examine the quality of the estimator, we would like that:

- $E[\hat{\theta}] = \theta_0$. If this is true, we call $\hat{\theta}$ an unbiased estimator
- $E[(\hat{\theta} - \theta_0)^2]$ is small. This is the mean squared error, which equals the variance + bias squared. More formally, $MSE[\hat{\theta}] = Var[\hat{\theta}] + (E[\hat{\theta}] - \theta_0)^2$

We estimate θ_0 from the data. However, there may be multiple estimators we can use, so how do we choose which one we want? For example, in the normal distribution case, $E[N(\theta_0, 1)] = \theta_0$, since by the SLLN, $\bar{X}_n \overset{a.s.}{\to} \theta_0$. Therefore, we can easily think of two different estimators to use: the sample mean and the sample median. What we want to do is to select the estimator with the smaller MSE. In the case where the estimators are unbiased, then there is no bias, so we would want to pick the estimator with the smallest variance.

8.2 Method of Maximum Likelihood

Using the method of maximum likelihood, we look at the observed values of x_i, and estimate the parameter $\hat{\theta}$ in order to maximum the likelihood of seeing those x_i's given that parameter.

Here is an example. Let's say that we want to use the method of maximum likelihood to estimate λ_0 from an exponential distribution after seeing the values $1.2, 2.1, 0.9$.

$L(\lambda) = \lambda e^{-\lambda(1.2)} \lambda e^{-\lambda(2.1)} \lambda e^{-\lambda(0.9)} = \lambda^3 e^{-\lambda(4.2)}$

Maximizing this likelihood is the same as maximizing the log likelihood.

$\mathcal{L}(\lambda) = logL(\lambda) = 3log(\lambda) - \lambda(4.2)$
$\mathcal{L}'(\lambda) = \frac{3}{\lambda} - 4.2 = 0$
$\hat{\lambda} = \frac{3}{4.2} = \frac{1}{1.4}$

Another example is to look at normally distributed iid random variables $X_1, X_2, ..., X_n$ with $N(\mu_0, \sigma_0^2)$.

$L(\vec{\theta}) = L(\mu, \sigma^2) = \prod_{i=1}^n f(\mu, \sigma^2, X_i)$
$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(X_i-\mu)^2}{2\sigma^2}}$
$= (2\pi\sigma^2)^{\frac{-n}{2}} e^{\frac{-\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}}$

$\mathcal{L}(\mu, \sigma^2) = -\frac{n}{2}[log(2\pi) + log(\sigma^2)] - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2\sigma^2}$

Taking the partial derivative with respect to μ and σ, we get that:

$\frac{d\mathcal{L}}{d\mu} = \frac{2\sum_{i=1}^n (X_i-\mu)}{2\sigma^2} = 0$
$\sum_{i=1}^n X_i - n\hat{\mu} = 0$
$\hat{\mu} = \bar{X}_n$, which is the sample mean!

$\frac{d\mathcal{L}}{d\sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2$
$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu})^2$

When $\hat{\theta} \xrightarrow{P} \theta$, we call that estimator a consistent estimator. In this normal case, $\hat{\mu} \xrightarrow{a.s.} \mu$ and $\hat{\sigma^2} \xrightarrow{a.s.} \sigma^2$, so these two estimators are "almost surely" consistent.

Typically, the preferred estimator for σ^2 is the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, since it is unbiased as $E[s^2] = \sigma_0^2$.

8.3 Transformation Invariance of MLE

One important property of MLE is that it is transformation invariant. What that means is that we can reparametrize the distribution and our new MLE is the corresponding reparameterized MLE. So for example if we reparameterize $N(\mu, \sigma^2)$ to $N(\mu, \sigma)$, then the MLE under the original parameterization is $\hat{\mu}, \hat{\sigma^2}$ and the MLE under the new parameterization is $\hat{\mu}, \hat{\sigma}$. More generally, if we reparameterize from θ to $g(\theta)$, then our MLE goes from $\hat{\theta}$ to $g(\hat{\theta})$.

We can show this as follows. Let our parameterization go from θ to $g(\theta) = \gamma$. Then the MLE for θ_0 is obtained by $max_\theta \mathcal{L}(\theta) = \mathcal{L}(\hat{\theta})$. But $max_\theta \mathcal{L}(\theta) = max_\gamma \mathcal{L}(g^{-1}(\gamma))$, which equals $\mathcal{L}(g^{-1}(\hat{\gamma}))$. Therefore, $\mathcal{L}(\hat{\theta}) = \mathcal{L}(g^{-1}(g(\hat{\gamma})))$. Therefore, $\hat{\gamma} = g^{-1}(\hat{\theta})$.

Let's look at an example of this. Assume we have $X_1, X_2, ..., X_n$ iid distributed with $N(\mu_0, 1)$, and we are interested in $P(X_1 > x)$.

$P(X_1 > x) = P(N(\mu, 1) > x)$
$= P(N(0, 1) + \mu > x)$
$= \bar{\Phi}(x - \mu)$

So, we want to estimate $g(\mu)$ where the function $g(x) = \bar{\Phi}(x - \mu)$. Then, by transformation invariance, we can estimate this probability by plugging in the MLE and therefore $\bar{\Phi}(x - \hat{\mu}) = \bar{\Phi}(x - \bar{X}_n)$.

This discussion tells us that "plug-in" estimators in which we just plug in the MLE in to the formula in question yields an estimator that achieves maximal efficiency.

9 Delta Method

The Delta Method is a fundamental accompaniment to the Central Limit Theorem that helps us analyze the efficiency of different estimators. As we have seen before, in the world of parametric statistics, we can have multiple possible estimators for our parameter θ. For example, if we are estimating μ coming from the normal distribution, two reasonable estimators to use would be the sample mean and the sample median. We want to choose the estimator that has the smallest MSE.

We can see that the MSE and the Central Limit Theorem are intimately connected. Assuming we have a CLT for the estimators that we want to compare, $\hat{\mu_i}$, we can write them down as $n^{\frac{1}{2}}(\hat{\mu_i} - \mu_0) \Rightarrow \eta_i N(0,1)$. We can then square both sides and keep the convergence in distribution by the Continuous Mapping Principle, so $n(\hat{\mu_i} - \mu_0)^2 \Rightarrow \eta_i^2[N(0,1)]^2$. We can look at the expected value of both sides, by our limit interchange theorems, and therefore we have

$nE[(\hat{\mu_i} - \mu_0)^2] \to \eta_i^2 E([N(0,1)]^2) = \eta_i^2$
$E[(\hat{\mu_i} - \mu_0)^2] \to \frac{\eta_i^2}{n}$

Therefore, $MSE(\mu_i) \approx \frac{\eta_i^2}{n}$ as $n \to \infty$. We can compare the MSE's for the different estimators and select the estimator with the lowest one.

However, the estimators are not always in a form where is it a direct sum of random variables and therefore the plain CLT cannot be used. For example, the MLE of an exponential random variable is $\hat{\lambda} = \frac{1}{\bar{X_n}}$ and the MLE of the standard deviation parameter of a normal is $\sqrt{\hat{\sigma^2}} = \sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$. Hence we need to find a way to use a type of a CLT in order to compare the efficiency of these kinds of estimators, which is why we need the Delta Method.

9.1 Delta Method

Assume that we have $X_1, X_2, ...$ iid, where they can be $\mathbb{R}^d$-valued vectors. We also assume that $E([X_1]^2) < \infty$. We have a continuous, smooth function g in the neighborhood of $E[X_1]$ and we want to examine what happens to the quantity $g(\bar{X}_n) - g(E[X_1])$. Using the smooth property, we can look at the first order Taylor expansion terms. By the Mean Value Theorem, $g(\bar{X}_n) - g(E[X_1]) \approx \nabla g(\xi_n)(\bar{X}_n - E[X_1])$, where ξ_n is a value between $\bar{X}_n$ and $E[X_1]$. We know the $\xi_n \overset{a.s.}{\to} E[X_1]$, so therefore $\nabla g(\xi_n) \overset{a.s.}{\to} \nabla g(E[X_1])$. Meanwhile by CLT, the second piece, multiplied by $n^{\frac{1}{2}}$, is converging in distribution such that $n^{\frac{1}{2}}(\bar{X}_n - E[X_1]) \Rightarrow N(0, C)$, where C is the covariance matrix create by $C = E[X_1 X_1^T] - E[X_1]E[X_1]^T$.

Therefore, putting these two pieces together, what the Delta Method states is that if g is continuously differentiable in a neighborhood of $E[X_1]$, then $n^{\frac{1}{2}}(g(\bar{X}_n) - g(E[X_1])) \Rightarrow \nabla g(E[X_1])N(0, C)$.

Note that $\nabla g(E[X_1])$ is a row vector and $N(0, C)$ is a column vector, so multiplying them together results in a scalar, call it η_i, which takes us back to the original way that we would compare the MSE of estimators by the CLT.

If we take the Delta Method and perform the Taylor expansion to include the second order terms, we can glean some insight into the bias of the problem. Now we write $g(\bar{X}_n) - g(E[X_1]) = \nabla g(E[X_1])(\bar{X}_n - E[X_1]) + \frac{1}{2}\sum_{i,j} \frac{d^2 g}{dx_i dx_j}(E[X_1])(\bar{X}_n(i) - E[X(i)])(\bar{X}_n(j) - E[X(j)])$. If we take the expected value of both sides, we notice that

$E[g(\bar{X}_n)] - g(E[X_1]) = 0 + \frac{1}{2}\sum_{i,j} \frac{d^2 g}{dx_i dx_j}(E[X_1])E[(\bar{X}_n(i) - E[X(i)])(\bar{X}_n(j) - E[X(j)])]$
$= \frac{1}{2n}\sum_{i,j} \frac{d^2 g}{dx_i dx_j}(E[X_1])Cov(X(i), X(j)) + O(\frac{1}{n})$

The left hand side is simply the formula for bias. What we can see from this equation is that the bias is strongly dependent on the second-order terms (or the non-linear) structure of g. This is what we would expect because if g is linear, then there is no bias.

We can also have an extension of this to allow for dependent X_i's. If $X_1, X_2, ..., X_n$ are dependent, but $n^{\frac{1}{2}}(\bar{X}_n - \beta) \Rightarrow N(0, C)$, then $n^{\frac{1}{2}}(g(\bar{X}_n) - g(\beta)) \Rightarrow \nabla g(\beta)N(0, C)$.

9.2 Examples

First let us look at applying the Delta Method to the MLE estimator for the exponential distribution, $\hat{\lambda} = \frac{1}{\bar{X_n}} = g(\bar{X}_n)$. Therefore, $g(x) = \frac{1}{x}$, and $g'(X) = -\frac{1}{x^2}$. By the Delta Method, we have that

$\hat{\lambda} - \lambda_0 \approx g'(E[X_1])(\bar{X}_n - E[X_1])$
$= -\frac{1}{(E[X_1])^2}(\bar{X}_n - E[X_1])$

Therefore, we can let $W_i = -\frac{1}{(E[X_1])^2}(X_i - E[X_1])$, and then $\hat{\lambda} - \lambda_0 \approx \bar{W}_n$. In order words, $n^{\frac{1}{2}}(\hat{\lambda} - \lambda_0) \approx n^{\frac{1}{2}}\bar{W}_n \Rightarrow \eta N(0,1)$, where $\eta^2 = Var[W_1]$.

Let us now take a look at applying the Delta Method to the sample standard deviation as an estimator for σ_0 in a normal distribution. We have that $s_n = \sqrt{\frac{1}{n-1}(\sum_{i=1}^n X_i - \bar{X}_n)^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}_n^2} \approx g(\frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n X_i)$. In this case, our function g is $g(x_1, x_2) = \sqrt{x_1 - x_2^2}$. Taking the partials with respect to x_1 and also x_2, we get:

$\frac{dg}{dx_1}(E[X_1^2], E[X_1]) = \frac{1}{2\sigma_0}$
$\frac{dg}{dx_2}(E[X_1^2], E[X_1]) = -\frac{\mu_0}{\sigma_0}$

$s_n - \sigma_0 \approx \frac{1}{2\sigma_0}(\frac{1}{n}\sum_{i=1}^n X_i^2 - E[X_1]^2) - \frac{\mu_0}{\sigma_0}(\frac{1}{n}\sum_{i=1}^n X_i - E[X_1])$.

Therefore, let $W_i = \frac{1}{2\sigma_0}(X_i^2 - E[X_1]^2) - \frac{\mu_0}{\sigma_0}(X_i - E[X_1])$, and $n^{\frac{1}{2}}(s_n - \sigma_n) \Rightarrow \sqrt{Var(W_1)}N(0,1)$.

One other key advantage of these CLTs is that they allow us to get confidence intervals for the unknown statistical parameters. From the Delta Method, we have that $n^{\frac{1}{2}}(\hat{\mu} - \mu_0) \Rightarrow \eta N(0,1)$, and we have an estimate of η, such that $\hat{\eta}^2 \xrightarrow{P} \eta^2$ as $n \to \infty$, then $\frac{n^{\frac{1}{2}}(\hat{\mu} - \mu_0)}{\hat{\eta}} \Rightarrow N(0,1)$ as $n \to \infty$.

To create a confidence interval, we can choose z such that $P(-z \leq N(0,1) \leq z) = 1 - \delta$, then $[\hat{\mu} - \frac{z\hat{\eta}}{\sqrt{n}}, \hat{\mu} + \frac{z\hat{\eta}}{\sqrt{n}}]$ is an approximate $100(1-\delta)\%$ confidence interval for μ_0.

10 Method of Moments

Many times, we would like to use the MLE to get our parameter estimate but it involves a significant amount of numerical computation. We can look to an alternate estimation approach in these cases, the method of moments. A motivating example is the Gamma distribution.

Let us assume that we have $X_1, X_2, ..., X_n$ iid distributed from a Gamma distribution with parameters λ_0 and α_0. We know that for a Gamma distribution, it is only supported at $x \geq 0$ and the pdf is $f_X(x) = \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$. If we were to use maximum likelihood estimation, we would have $L(\lambda, \alpha) = \prod_{i=1}^{n} \frac{\lambda(\lambda x)^{\alpha-1}e^{-\lambda X_i}}{\Gamma(\alpha)}$. Therefore, $\mathcal{L}(\lambda, \alpha) = logL(\lambda, \alpha) = n\alpha log(\lambda) + (\alpha - 1)\sum_{i=1}^{n} log(X_i) - \lambda \sum_{i=1}^{n} X_i - nlog\Gamma(\alpha)$. To maximize this likelihood, we take the derivative and set it equal to 0. So we have:

$\frac{d}{d\lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} X_i = 0$ and
$\frac{d}{d\alpha} = nlog(\lambda) + \sum_{i=1}^{n} log(X_i) - \frac{n}{\Gamma(\alpha)}\Gamma'(\alpha) = 0$

We can solve for $\hat{\lambda}$ in closed form. $\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}_n}$. However, $\hat{\alpha}$ cannot be obtained in closed form, and it must be solved numerically, either by numerical root-finding or by numerical optimization of the log likelihood $\mathcal{L}(\cdot)$

Therefore, instead of using the MLE approach, we can use the method of moments approach. For this Gamma example, let $\vec{\theta_0} = (\lambda_0, \alpha_0) \in \mathbb{R}^2$. We know that $E_{\theta_0}[X_1^k]$ is available in closed form. Specifically, we see that $E_{\theta_0}[X_1] = \frac{\alpha}{\lambda}$ and $Var_{\theta_0}[X_1] = \frac{\alpha}{\lambda^2}$. The SLLN tells us that $\bar{X}_n \overset{a.s.}{\to} E_{\theta_0}[X_1]$ and that $s_n^2 \overset{a.s.}{\to} Var_{\theta_0}[X_1]$.

Using this knowledge, we have that $\bar{X}_n = \frac{\hat{\alpha}}{\hat{\lambda}}$ and $s_n^2 = \frac{\hat{\alpha}}{\hat{\lambda}^2}$. Therefore, we can easily rearrange to solve for our parameter estimates. $\hat{\alpha} = \frac{\bar{X}_n^2}{s_n^2}$ and $\hat{\lambda} = \frac{\bar{X}_n}{s_n^2}$.

Therefore, we reiterate that using the MLE method has higher statistical efficiency, but lower computational tractability, whereas the method of moments has higher computational tractability, but lower statistical efficiency.

We can also use the method of moments to compute quantities outside of moments to do parameter estimation. For example, assume we have $X_1, X_2, ..., X_n$ iid distributed from an Exponential distribution with parameter λ_0. Let us define $k(x) = \mathbb{1}(X \geq 5)$. Therefore, $E_\theta[k(X_1)] = P_\theta[X_1 \geq 5] = P_\lambda(X_1 \geq 5) = e^{-5\lambda}$.

$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i) \geq 5 = e^{-5\hat{\lambda}}$
$-5\hat{\lambda} = log(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq 5))$
$\hat{\lambda} = -\frac{1}{5}log(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \geq 5))$

Therefore, we can look at the indicator function and see how many of the X_i's in our sample are above a threshold and use that in our equation to estimate λ_0. (Note that this is just an example to demonstrate how this would work. In practice, for this particular situation, the MLE would be preferred since it is computable in closed form).

If we did want to compute an MLE numerically by optimizing $\mathcal{L}(\cdot)$, we could use Newton methods, which are iterative methods. What we could select as the starting point for the iterative method is the method of moments estimator, $\hat{\theta}$. We know from the CLT that $\hat{\theta} - \theta_0 = O_P(n^{-\frac{1}{2}})$. In order words, it is converging at a square root n rate to the true parameter. Similarly, the MLE estimate, $\hat{\theta}_{MLE}$ is also converging to θ_0 at a square root n rate. Therefore, the consequence of this is that $\hat{\theta} - \hat{\theta}_{MLE}$ are converging to each other at a square root n rate, meaning that the method of moments estimator is already quite close to the MLE estimator. Since the Newton method converges at a rate of n, in many settings, we only need to do one iteration from the method of moments estimator to get close to the MLE estimator.

11 Estimating Equations

Estimating equations is an approach that allows us to take a unifying perspective on the maximum likelihood estimation and method of moments. Assume we have $X_1, X_2, ...X_n$ iid from a population P_{θ_0}, where $\theta_0 \in \mathbb{R}^d$. Assuming each X_i is m-dimensional, we have a function g that maps $\mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$. We say that $E_{\theta_1} g(\theta_2, X_1) = 0$ iff $\theta_1 = \theta_2$. Therefore, by the SLLN, $\frac{1}{n}\sum_{i=1}^n g(\theta, X_i) \overset{a.s.}{\to} E_{\theta_0}[g(\theta_1, X_1)] = 0$ iff $\theta = \theta_0$.

Therefore, we have a parameter estimate $\hat{\theta}$ for θ_0, which is $\frac{1}{n}\sum_{i=1}^n g(\hat{\theta}, X_i) = 0$. This is called the estimating equation.

We now want to show that the MLE is a special case of this estimating equation. In the MLE approach, we solve for the parameter estimates by taking the gradient of the log-likelihood function and setting it equal to 0. $\mathcal{L}(\theta) = \sum_{i=1}^n log f(\theta, X_i)$. Therefore,

$\nabla_\theta \mathcal{L}(\theta) = \sum_{i=1}^n \frac{\nabla_\theta f(\theta, X_i)}{f(\theta, X_i)} = 0$

To fulfill the estimating equations formula, we let $g(\theta, X_i) = \frac{\nabla_\theta f(\theta, X_i)}{f(\theta, X_i)}$ and we want to show that the expected value of this quantity under θ_0 is 0.

$E_{\theta_0}[\frac{\nabla_\theta f(\theta_0, X_i)}{f(\theta_0, X_i)}] = \int_{\mathbb{R}^m} \frac{\nabla_\theta f(\theta_0, x)}{f(\theta_0, x)} * f(\theta_0, x) dx$
$= \int_{\mathbb{R}^m} \nabla_\theta f(\theta_0, x) dx$
$= \nabla_\theta \int_{\mathbb{R}^m} f(\theta_0, x) dx$
$= \nabla_\theta * 1 = 0$

Therefore, we have proved that the maximum likelihood estimate is a special case of estimating equations.

For the method of moments, we have previously assumed that for the method, we defined a $k(x)$ such that $E_{\theta_1} k(X_1) = E_{\theta_2} k(X_1)$ iff $\theta_1 = \theta_2$. Therefore, we can define our $g(\theta, x) = k(x) - E_\theta k(X_1)$, and it is easy to see that the expected value of that quantity is 0. Therefore, method of moments is shown to also be a special case of estimating equations. In other words, estimating equations subsumes both MLE's and Method of Moments estimators.

11.1 CLT for Estimating Equations

We want to derive a general CLT for estimating equations. We start with our estimating equation: $\frac{1}{n}\sum_{i=1}^n g(\hat{\theta}, X_i) = 0$. Assuming that $\hat{\theta} \overset{a.s.}{\to} \theta_0$, we want to then subtract $\frac{1}{n}\sum_{i=1}^n g(\theta_0, X_i)$ from both sides of the equation. Therefore, we have

$\frac{1}{n}\sum_{i=1}^n g(\hat{\theta}, X_i) - \frac{1}{n}\sum_{i=1}^n g(\theta_0, X_i) = -\frac{1}{n}\sum_{i=1}^n g(\theta_0, X_i)$

If we look at the left hand side of the equation, we can use the Taylor series and Mean Value Theorem to expand that out such that the left hand side equals $\frac{1}{n}\sum_{i=1}^n H(\xi_n, X_i)(\hat{\theta} - \theta_0)$, where ξ_n lies between $\hat{\theta}$ and θ_0. Since we had assumed that $\hat{\theta} \overset{a.s.}{\to} \theta_0$, therefore, $\frac{1}{n}\sum_{i=1}^n H(\xi_n, X_i) \overset{a.s.}{\to} E_{\theta_0}[H(\theta_0, X_1)]$. Therefore, we have

$E_{\theta_0}[H(\theta_0, X_1)](\hat{\theta} - \theta_0) = -\frac{1}{n}\sum_{i=1}^n g(\theta_0, X_i)$
$E_{\theta_0}[H(\theta_0, X_1)](\hat{\theta} - \theta_0)\sqrt{n} = -\frac{1}{\sqrt{n}}\sum_{i=1}^n g(\theta_0, X_i)$
$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{1}{\sqrt{n}}\sum_{i=1}^n (E_{\theta_0}[H(\theta_0, X_1)])^{-1} g(\theta_0, X_1)$

Let $W_i = (E_{\theta_0}[H(\theta_0, X_1)])^{-1} g(\theta_0, X_1)$.

Then, $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \Rightarrow N(0, C)$, where $C = E[WW^T]$

Therefore, we now have a general CLT for estimators derived from estimating equations, that gives us a general vehicle to create CLT's for MLE's and Method of Moments estimators. Using CLT, we can then create confidence intervals for θ_0.

12 Cramer-Rao Bound

The Cramer-Rao bound provides a lower bound of the variance of unbiased parametric estimators. What we can show is that the MLE achieves this Cramer-Rao lower bound and is the most efficient estimator (lowest MSE and amongst unbiased estimators, therefore lowest variance). How we are going to do this is by using estimating equations to provide an expression for the MSE of the MLE, and then determine the Cramer-Rao bound, and show that the MLE equals this lower bound.

Assume we have $X_1, X_2, ..., X_n$ iid from a population P_{θ_0}. Using MLE, the log likelihood that we want to maximize is $\mathcal{L}(\theta) = \sum_{i=1}^n logf(\theta, X_i)$. Therefore, using estimating equations, $\hat{\theta}$ should satisfy $\mathcal{L}'(\hat{\theta}) = \sum_{i=1}^n \frac{f'(\hat{\theta},X_i)}{f(\hat{\theta},X_i)} = 0$. We can then multiply both sides by $\frac{1}{n}$ and add $-\frac{1}{n}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)}$. Therefore,

$\frac{1}{n}\sum_{i=1}^n \frac{f'(\hat{\theta},X_i)}{f(\hat{\theta},X_i)} - \frac{1}{n}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)} = -\frac{1}{n}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)}$

In order to prepare us for the Taylor expansion, we find $\frac{d}{d\theta} = \frac{f'(\theta,x)}{f(\theta,x)} = \frac{f''(\theta,x)}{f(\theta,x)} - \frac{(f'(\theta,x))^2}{(f(\theta,x))^2}$. Now let us Taylor expand the left side of the equation and using the mean value theorem, approximate the first order term. We get that

$\frac{1}{n}\sum_{i=1}^n [\frac{f''(\xi,X_i)}{f(\xi,X_i)} - \frac{(f'(\xi,X_i))^2}{(f(\xi,X_i))^2}](\hat{\theta} - \theta_0) = -\frac{1}{n}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)}$, for a value ξ in between $\hat{\theta}$ and θ_0

Multiplying both sides by $\sqrt{n}$, we get that

$\frac{1}{n}\sum_{i=1}^n [\frac{f''(\xi,X_i)}{f(\xi,X_i)} - \frac{(f'(\xi,X_i))^2}{(f(\xi,X_i))^2}](\hat{\theta} - \theta_0)\sqrt{n} = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)}$

On the left hand side of the equation, since $\hat{\theta} \overset{a.s.}{\to} \theta_0$, therefore $\xi \overset{a.s.}{\to} \theta_0$, and therefore, $\frac{1}{n}\sum_{i=1}^n [\frac{f''(\xi,X_i)}{f(\xi,X_i)} - \frac{(f'(\xi,X_i))^2}{(f(\xi,X_i))^2}] \to E_{\theta_0}[\frac{f''(\theta_0,X_1)}{f(\theta_0,X_1)} - \frac{(f'(\theta_0,X_1))^2}{(f(\theta_0,X_1))^2}]$. On the right hand side of the equation, by CLT, $-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)} \Rightarrow N(0, Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]$. Therefore we have that

$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \frac{N(0, Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]}{E_{\theta_0}[\frac{f''(\theta_0,X_1)}{f(\theta_0,X_1)} - \frac{(f'(\theta_0,X_1))^2}{(f(\theta_0,X_1))^2}]}$

We have that $E_{\theta_0}[\frac{f''(\theta_0,X_1)}{f(\theta_0,X_1)}] = \int_{\mathbb{R}^m} \frac{f''(\theta_0,x)}{f(\theta_0,x)} f(\theta_0,x)dx = \int_{\mathbb{R}^m} f''(\theta_0,x) = \int_{\mathbb{R}^m} \frac{d^2}{d\theta^2} f(\theta,x)|_{\theta=\theta_0} dx = \frac{d^2}{d\theta^2}\int_{\mathbb{R}^m} f(\theta,x)dx|_{\theta=\theta_0} = \frac{d^2}{d\theta^2}1|_{\theta=\theta_0} = 0$. Since we know that $-\frac{1}{n}\sum_{i=1}^n \frac{f'(\theta_0,X_i)}{f(\theta_0,X_i)}$ converges to mean zero, then the denominator simplifies to $E_{\theta_0}[\frac{f''(\theta_0,X_1)}{f(\theta_0,X_1)} - \frac{(f'(\theta_0,X_1))^2}{(f(\theta_0,X_1))^2}] = -Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]$. Therefore, we now have

$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow -\frac{N(0, Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}])}{Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]} = -\frac{N(0,1)}{\sqrt{Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]}}$

Squaring both sides and taking the expected value, we have that

$nE_{\theta_0}([\hat{\theta}_n - \theta_0]^2) \to \frac{1}{Var_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]}$
$E_{\theta_0}([\hat{\theta}_n - \theta_0]^2) \to \frac{1}{nVar_{\theta_0}[\frac{f'(\theta_0,X_1)}{f(\theta_0,X_1)}]}$

Now let us create the expression for the Cramer-Rao bound for any unbiased estimator. Suppose that we have an unbiased estimator $\hat{\theta}$. Then $E_\theta[\hat{\theta}] = \theta$. Suppose that we have $X_1, X_2, ..., X_n$ iid distributed from P_θ, and our estimator $\hat{\theta}$ comes from the what we have observed in the data, so $\hat{\theta} = w(X_1, X_2, ..., X_n)$. Then,

$E_\theta[w(X_1, X_2, ..., X_n)] = \theta$
$\int_{\mathbb{R}^{m\times n}} w(X_1, X_2, ..., X_n)\prod_{i=1}^n f(\theta, X_i)dX_i = \theta$
$\frac{d}{d\theta}\int_{\mathbb{R}^{m\times n}} w(X_1, X_2, ..., X_n)\prod_{i=1}^n f(\theta, X_i)dX_i = \frac{d}{d\theta}\theta$
$\int_{\mathbb{R}^{m\times n}} w(X_1, X_2, ..., X_n)\frac{d}{d\theta}\prod_{i=1}^n f(\theta, X_i)dX_i = 1$
$\int_{\mathbb{R}^{m\times n}} w(X_1, X_2, ..., X_n)\sum_{i=1}^n f'(\theta, X_i)\prod_{j\neq i} f(\theta, X_j)dX_i = 1$
$\int_{\mathbb{R}^{m\times n}} w(X_1, X_2, ..., X_n)\sum_{i=1}^n \frac{f'(\theta,X_i)}{f(\theta,X_i)}\prod_{j=1}^n f(\theta, X_j)dX_i = 1$
$E_\theta[w(X_1, X_2, ..., X_n)\sum_{i=1}^n \frac{f'(\theta,X_i)}{f(\theta,X_i)}] = E_\theta[1]$
$E_\theta[\sum_{i=1}^n \frac{f'(\theta,X_i)}{f(\theta,X_i)}w(X_1, X_2, ..., X_n)] = 1$

Let $Y = \frac{f'(\theta,X_i)}{f(\theta,X_i)}$. Since they are mean zero random variables, then $E[Y] = 0$. We know the covariance formula that $Cov[Y, Z] = E[YZ] - E[Y]E[Z]$, so if $E[Y] = 0$ and $E[YZ] = 1$, then $Cov[Y, Z] = 1$. In this specific case,

$Cov[Y, Z] = Cov[\sum_{i=1}^{n} \frac{f'(\theta, X_i)}{f(\theta, X_i)}, w(X_1, X_2, ..., X_n)] = 1.$

We use the Cauchy-Schwarz Inequality which states that for $E[Y^2] < \infty$ and $E[Z^2] < \infty$, that $(E[|YZ|])^2 \leq E[Y^2]E[Z^2]$.

Therefore, we have that $(E[|YZ|])^2 = (Cov[\sum_{i=1}^{n} \frac{f'(\theta_0, X_i)}{f(\theta_0, X_i)}, w(X_1, X_2, ..., X_n)])^2 = 1^2 \leq Var_\theta(\sum_{i=1}^{n} \frac{f'(\theta_0, X_i)}{f(\theta_0, X_i)}) Var_\theta(\hat{\theta})$

$1 \leq Var_\theta(\sum_{i=1}^{n} \frac{f'(\theta_0, X_i)}{f(\theta_0, X_i)}) Var_\theta(\hat{\theta})$
$= (nVar_\theta[\frac{f'(\theta, X_i)}{f(\theta, X_i)}]) Var_\theta(\hat{\theta})$

Therefore, we have $Var_\theta(\hat{\theta}) \geq \frac{1}{nVar_\theta[\frac{f'(\theta, X_i)}{f(\theta, X_i)}]}$

As shown above, this lower bound is exactly the variance achieved by the MLE asymptotically, and therefore the MLE is proven to be the most efficient unbiased estimator (lowest MSE). Although it is not proven here, it can be shown that even if we relax the unbiasedness assumption, we can still prove the MLE is the most efficient amongst all estimators.

13 Hypothesis Testing

In hypothesis testing, we want to find out if there is just noise in our data, or if there is actually some new effect in the data. One practical example of this is if there is a new drug in development for treating blood pressure, and the FDA will only add the drug to the formulary if it reduces blood pressure by more than 10%. Therefore, what we would do is to run trials $X_1, X_2, ..., X_n$ to perform the hypothesis test. We would develop a null hypothesis, $H_0 : \mu \leq 0.1$ vs. an alternative hypothesis, $H_1 : \mu > 0.1$.

In hypothesis testing, the null hypothesis is our default hypothesis, and we reject H_0 only in the presence of strong statistical evidence to the contrary. Otherwise, with small deviations, we accept (do not reject the null hypothesis. We define the critical region $\mathcal{C}$ to be the region where we would reject the null hypothesis in favor of the alternative hypothesis. We define two types of error:

1. Type I Error: Reject H_0 when H_0 is actually true (false positive)
2. Type II Error: Accept H_0 when H_1 is actually true (false negative)

An example of this would be if our null hypothesis was that a virus is not present, and the alternative hypothesis was that a virus is present, then a Type I Error would be determining that a virus is present when it is not, and a Type II Error would be determining that not virus is present when it truly is present.

When we form a statistical test, for example if we have $H_0 : N(\theta, 1); \theta \leq 0$ vs. $H_1 : N(\theta, 1); \theta > 0$, we are testing if P_θ, the normal distribution in this case with distribution $N(\theta, 1)$, is coming from $\theta \in \Lambda_0$ or Λ_1. If Λ_i is a singleton (a single distribution), we call it a simple hypothesis. If Λ_i is not a singleton (multiple distributions), we call it a composite hypothesis.

13.1 Simple Hypothesis Testing

In simple hypothesis testing, we are testing $H_0 : P_0$ vs. $H_1 : P_1$, where P_0 is the pdf f_0, and P_1 is the pdf f_1. Prior to performing the test, we want to set the significance level: $\alpha = P_0(\mathcal{C})$, or the level at which we are comfortable with making a Type I Error. For all the critical regions $\mathcal{C}$ that fulfill this significance level, we want to find the $\mathcal{C}$ which makes the Type II Error as small as possible. In other words, for every event A such that $\alpha = P_0(A)$, $P_1(\mathcal{C}^C) \leq P_1(A)$, or equivalently, $P_1(\mathcal{C}) \geq P_1(A)$.

By the Neyman-Pearson Lemma, we know that there exists such a region $\mathcal{C}$ and it takes the following form assuming X_i are iid:

$P_0(\{\frac{\prod_{i=1}^n f_0(X_i)}{\prod_{i=1}^n f_1(X_i)} \leq c\}) = \alpha$

What we are saying is that we want to take the ratio of the likelihood of the null hypothesis divided by the likelihood of the alternative hypothesis, and if that ratio is less than some constant under the pdf of the null hypothesis, then we reject the null hypothesis because it is unlikely.

Let's look at a concrete example on how to perform a simple hypothesis test. Let's assume that we are testing $H_0 : f_0 = N(0, 1)$ vs. $H_1 : f_1 = N(1, 1)$. Then we have that $\frac{\prod_{i=1}^n f_0(X_i)}{\prod_{i=1}^n f_1(X_i)} = exp(-\frac{\sum_{i=1}^n X_i^2}{2} + \frac{\sum_{i=1}^n (X_i-1)^2}{2}) = exp(-\sum_{i=1}^n X_i + \frac{n}{2})$.

$P_0(exp(-\sum_{i=1}^n X_i + \frac{n}{2}) \leq c_1)$
$= P_0(\frac{\sum_{i=1}^n X_i}{n} \geq c_2)$
$= P_0(\sqrt{n}\bar{X}_n \geq \sqrt{n}c_2)$
$= P_0(N(0, 1) \geq c_2\sqrt{n}$

Therefore, say we want to set $\alpha = 0.05$, we can then go to the normal table and obtain the value, which in this case is 1.645. Therefore, $c_2 = \frac{1.645}{\sqrt{n}}$, and our critical region is $\mathcal{C} = \{\bar{X}_n \geq \frac{1.645}{\sqrt{n}}\}$. If we take a sample and calculate the sample average and we land in this region, we will reject the null hypothesis.

13.2 Composite Hypothesis Testing

The first case of composite hypothesis testing we will look at is a simple vs. composite test, where H_0 is a simple hypothesis and H_1 is a composite hypothesis. More specifically, $H_0 : P_0$ is simple and $H_1 : P_\theta : \theta \in \Lambda$, where there is more than one value that θ can take.

We call a test a uniformly most powerful test if there is a single $\mathcal{C}_\theta$ that gives the most power for every simple hypothesis test H_0 vs. H_θ for all the $\theta \in \Lambda$. A uniformly most powerful test is not required to always exist.

The second case of composite hypothesis testing is a composite vs. composite test. We will use the Likelihood Ratio Test for performing composite hypothesis testing, which takes the form:

$P_\theta(\frac{sup_{\theta\in\Lambda_0}\prod_{i=1}^n f(\theta,X_i)}{sup_{\theta\in\Lambda_1}\prod_{i=1}^n f(\theta,X_i)} \leq c) \leq \alpha$

We can see that is the very similar to the simple vs. simple case. The difference is that we select the best θ out of all the θ in the null hypothesis and compare it to the best θ out of all the θ in the alternative hypothesis.

Let us look at a concrete example of how to apply a composite hypothesis test. Let us test $H_0 : N(\theta, 1), \theta \leq 0$ vs. $H_1 : N(\theta, 1), \theta > 0$. We know from our MLE calculation that $\hat{\theta} = \bar{X}_n$. Looking at the null hypothesis, we can see that $sup_{\theta\in\Lambda_0}\prod_{i=1}^n f(\theta, X_i)$ equals

$$\begin{cases}\prod_{i=1}^n f(\hat{\theta}, X_i); \hat{\theta} \leq 0\\ \prod_{i=1}^n f(0, X_i); \hat{\theta} > 0\end{cases}$$

Similarly, looking at the alternative hypothesis, we can see that $sup_{\theta\in\Lambda_1}\prod_{i=1}^n f(\theta, X_i)$ equals

$$\begin{cases}\prod_{i=1}^n f(0, X_i); \hat{\theta} \leq 0\\ \prod_{i=1}^n f(\hat{\theta}, X_i); \hat{\theta} > 0\end{cases}$$

$P_\theta(\mathcal{C})$
$= P_\theta(\bar{X}_n \leq 0, exp(\frac{-\sum_{i=1}^n (X_i-\bar{X}_n)^2}{2} + \frac{\sum_{i=1}^n X_i^2}{2}) \leq c) + P_\theta(\bar{X}_n > 0, exp(-\frac{\sum_{i=1}^n X_i^2}{2} + \frac{\sum_{i=1}^n (X_i-\bar{X}_n)^2}{2}) \leq c)$
$= P_\theta(\bar{X}_n \leq 0, exp(\frac{n}{2}\bar{X}_n^2) \leq c) + P_\theta(\bar{X}_n > 0, exp(-\frac{n}{2}\bar{X}_n^2) \leq c)$

Let us first examine the case where $c \geq 1$. From the first term in the 2 term sum above, we can see that $\bar{X}_n^2 \leq c'$ and $\bar{X}_n \leq 0$, so we have that $P_\theta(-\sqrt{c'} \leq \bar{X}_n \leq 0) + P_\theta(\bar{X}_n > 0)$. This is monotonically increasing in θ on $(-\infty, 0]$, therefore it is maximized when $\theta = 0$. At $\theta = 0$, the probability with be at least 0.5, so it would never be significant under any test, where we usually look for values such as $0.01, 0.05$, or 0.1.

Let us now examine the case where $c \leq 1$. We then have

$= 0 + P_\theta(\bar{X}_n > 0, exp(-\frac{n}{2}\bar{X}_n^2) \leq c)$
$= P_\theta(\bar{X}_n > 0, \frac{n}{2}\bar{X}_n^2 \leq log(c^{-1}))$
$= P_\theta(\bar{X}_n \geq \sqrt{\frac{2}{n}log(c^{-1})})$

Again, this is monotonically increasing in θ on $(-\infty, 0]$, therefore it is maximized when $\theta = 0$. Therefore, when we determine the significance level, say at 0.05, and we have $P_0(\bar{X}_n \geq c) = 0.05$, then $P_\theta(\bar{X}_n \geq c) \leq 0.05$ for all θ. Therefore, we have

$P_0(\bar{X}_n \geq c) = 0.05$
$P_0(N(0, 1) \geq \sqrt{n}c) = 0.05$
$c = \frac{1.645}{\sqrt{n}}$

Therefore, our critical region is $\mathcal{C} = \{\bar{X}_n \geq \frac{1.645}{\sqrt{n}}\}$. If we take a sample and calculate the sample average and we land in this region, we will reject the null hypothesis.

14 Non-Parametric Statistics

In the world of non-parametric statistics, we are not assuming that the data comes from a parametric statistical family, but instead we're making much more modest assumptions about the statistical model. In a non-parametric setting, we observe $X_1, X_2, ..., X_n$ iid from a common distribution F_0 (we are not making assumptions about this distribution). We may be interested in estimating $F_0(\cdot)$ or possibly a functional such as $T(F_0) = F_0(x)$, $T(F_0) = \int_{\mathbb{R}} g(s)F_0(dx) = E_0[g(x)]$, or $T(F_0) = F_0^{-1}(p)$ (a quantile).

To begin, let's look at an example. Say we observe 3 different points, $4.1, 2.3, 1.8$. Since we have observed these 3 points we know that the distribution is supported at these three points, and we also do not know whether or not the distribution is supported at points outside of these three points. Therefore, our estimate should only be at these 3 points, p_1, p_2, p_3. If was maximize the likelihood with constraints that $p_1+p_2+p_3 = 1$ and that $p_1 \geq 0, p_2 \geq 0, p_3 \geq 0$, then we obtain that $\hat{p_i} = \frac{1}{3}$.

More generally, if we have observed $X_1, X_2, ..., X_n$ iid distinct points, and we try to maximize the likelihood $maxL(p_1, p_2, ..., p_n) = \prod_{i=1}^n p_i$ with the constraints that $\sum_{i=1}^n p_i = 1$, and that $p_i \geq 0$, then we obtain that $\hat{p_i} = \frac{1}{n}$. Therefore, we have that $\hat{P}_n(\cdot) = \sum_{i=1}^n \frac{1}{n}\delta_{X_i}(\cdot)$, where $\delta_X(A) = 1$ if $x \in A$ and 0 otherwise. We also then have our empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x)$. By the SLLN, we have that $\hat{F}_n(X) \overset{a.s.}{\to} F_0(x)$ as $n \to \infty$.

If we want to estimate the functional $T(F_0) = E[g(X_1)]$ from the data, we can simply use the plug in estimate $T(\hat{F}_n) = \int_{\mathbb{R}} g(x)\hat{F}_n(dx) = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

To get confidence intervals for $T(F_0)$, we have that $n^{\frac{1}{2}}(T(\hat{F}_n) - T(F_0)) = n^{\frac{1}{2}}(\frac{1}{n}\sum_{i=1}^n g(X_i) - E[g(X_1)]) \Rightarrow \sigma_g N(0,1)$, where $\sigma_g^2 = Var[g(X_1)]$. We can then estimate the variance using the sample variance, and $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n [g(X_i) - \frac{1}{n}\sum_{i=1}^n g(X_i)] \overset{a.s.}{\to} \sigma_g^2$. Therefore, we can then create our confidence interval as $[T(\hat{F}_n) - \frac{zs_n}{\sqrt{n}}, T(\hat{F}_n) + \frac{zs_n}{\sqrt{n}}]$.

In order to estimate the pth quantile of F, $T(F) = F^{-1}(p)$, we again would use the plug in principle, and therefore our estimate would be $\hat{F}_n^{-1}(p)$. However, when finding a confidence interval through CLT, we have an issue. Typically, we would have used estimating equations as follows:

$\hat{F}_n(\hat{F}_n^{-1}(p)) = p$
$\hat{F}_n(\hat{F}_n^{-1}(p)) - \hat{F}_n(F_0^{-1}(p)) = p - \hat{F}_n(F_0^{-1}(p))$

However, since $\hat{F}_n$ is not a continuous function, we cannot perform the Taylor expansion on the left hand side and use the Mean Value Theorem to get the first order term. For now, we can assume that for large n, $\hat{F}_n(\cdot) \approx F_0(\cdot)$, and assuming that $f_0(\cdot)$ is continuous at the point $q \overset{\Delta}{=} F_0^{-1}(p)$, we can continue with our estimating equations methodology. $\hat{f}_n(\cdot)$ is close to $f_0(q)$, so we write

$f_0(q)(\hat{F}_n(p) - F_0^{-1}(p)) = p - \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq q)$
$p - \hat{F}_n(q) = p - \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq q)$
$p - \hat{F}_n(q) = \frac{1}{n}\sum_{i=1}^n (p - \mathbb{1}(X_i \leq q))$

Since the right hand side is a sum of mean zero random variables, we can use the CLT and therefore

$\sqrt{n}f_0(q)(\hat{F}_n^{-1}(p) - q) = \sqrt{n}\frac{1}{n}\sum_{i=1}^n (p - \mathbb{1}(X_i \leq q)) \Rightarrow N(0, Var[\mathbb{1}(X_1 \leq q)]) = N(0, p(1-p))$, since Bernoulli
$\sqrt{n}(\hat{F}_n^{-1}(p) - q) \Rightarrow \frac{\sqrt{p(1-p)}}{f_0(q)}N(0,1)$

Through our analysis, we have determined a CLT for the empirical quantile, but two issues remain. The first is that we have not rigorously shown that our calculations make sense since $\hat{F}_n(\cdot)$ is not continuous, and the second is that is order to use our CLT, we need to have the value $f_0(q)$, where we know neither $f_0(\cdot)$, nor q, so we will have to use density estimation.

14.1 Weak Convergence of Stochastic Processes

From our previous discussion, we know that $\hat{F}_n(x) \overset{a.s.}{\to} F_0(x)$ and that

$n^{\frac{1}{2}}(\hat{F}_n(x) - F_0(x)) = n^{\frac{1}{2}}(\frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x) - P(X_i \leq x)) \Rightarrow \sqrt{F_0(x)(1 - F_0(x))}N(0,1)$

We want to generalize the notion of weak convergence from scalar random variables to random functions/stochastic processes. We can see that our CLT values in this case are functions, and we claim that we have

$n^{\frac{1}{2}}(\hat{F}_n(\cdot) - F_0(\cdot)) \Rightarrow Z(\cdot)$

where Z is a Gaussian stochastic process that has mean $E[Z(x)] = 0$ for any x and covariance matrix $Cov(Z(x), Z(y)) = Cov(\mathbb{1}(X_1 \leq x), \mathbb{1}(X_1 \leq y))$. One other nice property that this stochastic process has, which is that if $F_0(\cdot)$ is a continuous distribution, then $Z(\cdot)$ is a continuous process. In other words, if $x_n \to x_\infty$, then $Z(x_n) \to Z(x_\infty)$ along that sequence.

By the Skorohod Representation Theorem we have that the following 3 are equivalent:

- $Y_n \Rightarrow Y_\infty$
- There exists a probability space supporting $Y_n' : 1 \leq n \leq \infty$ such that $Y_n \overset{D}{=} Y_n'$ for $1 \leq n \leq \infty$ and that $Y_n' \overset{a.s.}{\to} Y_\infty'$
- For $g : \mathbb{R} \to \mathbb{R}$ such that g is bounded and continuous, $E[g(Y_n)] \to E[g(Y_\infty)]$

When we are looking at weak convergence of stochastic processes, namely in this case, they are in the space $D(-\infty, \infty)$, the space of discontinuous functions that are right-continuous with left limits everywhere. In this space, we no longer can look at weak convergence through CDFs, as this does not have a meaning, so we will use the the third definition above. We have weak convergence if $E[g(Y_n)] \to E[g(Y_\infty)]$ for every bounded and continuous g where $g : D(-\infty, \infty) \to \mathbb{R}$. To define what is means for g to be continuous, we first define that there is a metric on the space $D(-\infty, \infty)$, where the definition of a metric space is that is satisfies:

- $d(u, v) \leq d(u, w) + d(w, v)$
- $d(u, v) = d(v, u)$
- $d(u, u) = 0$ and $d(u, v) = 0$ only if $u = v$

We say that g continuous means that if $d(u_n, u_\infty) \to 0$, then $g(u_n) \to g(u_\infty)$.

Using this knowledge, we can return back to our problem where we have $n^{\frac{1}{2}}(\hat{F}_n(\cdot) - F_0(\cdot)) \Rightarrow Z(\cdot)$ in $D(-\infty, \infty)$. Using weak convergence in stochastic processes, we know that we can find a probability space supporting $\hat{F}_n'(\cdot)$ and $Z'(\cdot)$ such that $d(n^{\frac{1}{2}}(\hat{F}_n'(\cdot) - F_0(\cdot)), Z'(\cdot)) \overset{a.s.}{\to} 0$, and $\hat{F}_n'(\cdot) \overset{D}{=} \hat{F}_n(\cdot)$ and $Z'(\cdot) \overset{D}{=} Z(\cdot)$.

Returning to our estimating equations formula, we now can replace the $\hat{F}_n(\cdot)$ with $\hat{F}_n'(\cdot)$, so we have

$n^{\frac{1}{2}}(\hat{F}_n'(\hat{F}_n'(p)) - F_0(\hat{F}_n'^{-1}(p))) \Rightarrow Z'(\hat{F}_n'^{-1}(p)))$
$n^{\frac{1}{2}}(p - F_0(\hat{F}_n'^{-1}(p))) \Rightarrow Z'(\hat{F}_n'^{-1}(p)))$
$n^{\frac{1}{2}}(p - F_0(\hat{F}_n'^{-1}(p))) \Rightarrow Z'(q)$, since $Z'(\cdot)$ is continuous
$n^{\frac{1}{2}}(F_0(F_0^{-1}(p)) - F_0(\hat{F}_n'^{-1}(p))) \Rightarrow Z'(q)$
$n^{\frac{1}{2}}[f_0(q)(F_0^{-1}(p) - \hat{F}_n'^{-1}(p))] \Rightarrow Z'(q)$, since F_0 is smooth
$n^{\frac{1}{2}}(q - \hat{F}_n'^{-1}(p))) \Rightarrow \frac{Z'(q)}{f_0(q)}$
$n^{\frac{1}{2}}(q - \hat{F}_n^{-1}(p))) \Rightarrow \frac{Z(q)}{f_0(q)}$, where $Z(q)$ is $N(0, p(1-p))$

14.2 Density Estimation

In a non-parametric approach to density estimation, we have observed $X_1, X_2, ..., X_n$ iid from F_0, and all we know is that the distribution is supported at these points. We do not know if it is supported outside of these points and therefore we cannot use a plug in estimator for the density. We don't even know if the underlying population is continuous or discrete.

One simple way to do density estimation is to bin the data into a histogram with bins of width h, and then to fit a density curve to the bins of the histogram. We hope that as $n \to \infty$ and $h \to 0$, that we have a good estimate of the density. However, this is not the most common approach.

The more common approach is to use a kernel density estimator. How this is works it that we assume a continuous distribution around each of the data points that we have observed (say a normal distribution around each point). Then we take a mixture of these distributions as our density estimation. As n gets larger and larger, we will have more and more of these "normal" distributions around each point, and we will have to make them narrower and narrower. By going through this process, the kernels "smooth" the underlying discrete sample through a mixture of normal densities with the mean at each data point.

Let's suppose that we look at a normal density with mean X_i and variance h. Let us denote $\phi(x)$ as the density of a $N(0, 1)$. Then at every observed data point X_i we can create a distribution around that point as $\phi(\frac{x - X_i}{h}) * \frac{1}{h}$, essentially a normal density

centered at that data point. Then our density estimator is the average of all of these densities. $\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\phi(\frac{x-X_i}{h})$. Now that we have an estimator, we want to be able to select the appropriate width, h in order to have the smallest MSE. MSE in this case is defined as $MSE[\hat{f}_n(x)] = Var[\hat{f}_n(x)] + [E[\hat{f}_n(x)] - f_0(x)]^2$. We recognize that we have a bias-variance tradeoff: if we pick a very small h, we will have low bias, but high variance. If we pick a very large h, we will have low variance, but high bias. Therefore, we want to optimally select an h in between.

Let's first look at the bias. We can calculate that

$E[\hat{f}_n(x)] = E[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\phi(\frac{x-X_i}{h})]$
$= \frac{1}{h}E[\phi(\frac{x-X_i}{h})]$
$= \frac{1}{h}\int_{-\infty}^{\infty} \phi(\frac{x-z}{h})f_0(z)dz$
Let $w = \frac{x-z}{h}$, therefore $dw = -\frac{dz}{dh}$, and $z = x - wh$
$= \int_{-\infty}^{\infty} \phi(w)f_0(x-wh)dw$
$= \int_{-\infty}^{\infty} [f_0(x) - whf_0'(x) + \frac{w^2h^2}{2}f_0''(x) + O(h^3)]dw$ from Taylor expansion
$= f_0(x) - hf_0'(x)\int_{-\infty}^{\infty} w\phi(w)dw + \frac{h^2}{2}f_0''(x)\int_{-\infty}^{\infty} w^2\phi(w)dw + O(h^3)$
We have that $E[N(0,1)] = 0$ and $E[(N(0,1))^2] = 1$
$= f_0(x) + \frac{h^2}{2}f_0''(x) + O(h^3)$

Therefore, the bias is $\frac{h^2}{2}f_0''(x) + O(h^3)$.

Now let's look at the variance term.

$Var[\hat{f}_n(x)] = Var[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}\phi(\frac{x-X_i}{h})]$
$= \frac{1}{n}Var[\frac{1}{h}\phi(\frac{x-X_i}{h})]$
$= \frac{1}{nh^2}Var[\phi(\frac{x-X_i}{h})]$
$= \frac{1}{nh^2}[E[(\phi(\frac{x-X_i}{h}))^2] - (E[\phi(\frac{x-X_i}{h})])^2]$
The second term is small, so we can neglect it.
$\approx \frac{1}{nh^2}\int_{-\infty}^{\infty} (\phi(\frac{x-z}{h}))^2 f_0(z)dz$
Let $w = \frac{x-z}{h}$, therefore $dw = -\frac{dz}{dh}$, and $z = x - wh$
$= \frac{1}{nh}\int_{-\infty}^{\infty} (\phi(w))^2 f_0(x-wh)dw$
$= \frac{1}{nh}\int_{-\infty}^{\infty} (\phi(w))^2 dw f_0(x)$, since for h small, $f_0(x-wh) \approx f_0(x)$

Therefore, if we look at the MSE as a whole, variance + bias squared, we get

$MSE = \frac{1}{nh}\int_{-\infty}^{\infty} (\phi(w))^2 dw f_0(x) + \frac{h^4}{4}(f_0''(x))^2$ + smaller terms

To have the optimal bias variance tradeoff, we're going to want the magnitude of the variance to be on the same order as the magnitude of the bias squared. Therefore we want $h^4 \approx \frac{1}{nh}$, and therefore we want to pick a value of h where $h^5 \approx \frac{1}{n}$.

When we select the optimal h where $h \approx n^{-\frac{1}{5}}$, we see that the MSE is on the order of $n^{-\frac{4}{5}}$, and therefore we RMSE is on the order of $n^{-\frac{2}{5}}$. We see that when we estimate densities, we have a degredation from most estimates, which converge at the rate of $n^{-\frac{1}{2}}$ to $n^{-\frac{2}{5}}$.

15 Censored Data Methods

Often times when working with data, the data is censored, or cut off at certain points. For example, if a company is wanting to offer a new product and wants to see if it should offer a warranty of 5 years, the company can run prototypes to test the duration of the product. However, the company cannot run the test until failure, or it would not launch the product for a long time. Hence, it may run the test until the minimum of the failure time or the end of the test period. This is an example where if the prototype lasts past the test period, the data is censored. We know that it did not fail before a certain point, but do not know the exact time of failure.

In this example, the full failure times T_i are not observed after a certain point, and therefore the data is "right censored." We can also have "left censored" data, for example if all observations prior to a certain point in time are truncated. Another time of censoring can happen for example if we round all failures to the nearest integer. Therefore, we do not have the exact failure time T_i, but rather these failures are binned. For example, we know this failure occured between T_2 and T_3, but not the exact time. This is called "interval censoring."

We can take a look at how to do parametric estimation and non-parametric estimation with censored data.

15.1 Parametric Model with Censoring

Let us assume that we have failure times $T_1, T_2, ..., T_n$ iid from an Exponential with parameter λ_0. However, the data is right censored so we do not observe the T_i after the point in time, t. In order to estimate λ_0 using MLE, we would now have:

$L(\lambda) = \prod_{i=1}^{n}(\mathbb{1}(T_i \leq t)\lambda e^{-\lambda T_i} + \mathbb{1}(T_i > t)e^{-\lambda t})$

For the scenario where the failure point is past the censored point, t, we can use the survival function $(1 - F(x))$ in the likelihood function. Therefore, the log likelihood is:

$\mathcal{L}(\lambda) = \sum_{i=1}^{n} \mathbb{1}(T_i \leq t)(log(\lambda) - \lambda T_i) + \sum_{i=1}^{n} \mathbb{1}(T_i > t)(-\lambda t)$
$= \sum_{i=1}^{n} \mathbb{1}(log(\lambda)) - \lambda \sum_{i=1}^{n} min(T_i, t)$

$\mathcal{L}'(\lambda) = \sum_{i=1}^{n} \mathbb{1}(T_i \leq t)\frac{1}{\lambda} - \sum_{i=1}^{n} min(T_i, t) = 0$

$\hat{\lambda} = \frac{\sum_{i=1}^{n} \mathbb{1}(T_i \leq t}{\sum_{i=1}^{n} min(T_i, t)}$

To get a confidence interval for λ_0, we could work out a CLT for $\hat{\lambda}$. We can take our estimate of $\hat{\lambda}$ and multiply both the numerator and denominator by $\frac{1}{n}$, and then we would have a ratio of two sample averages, exactly the form where we can use the Delta method to get the CLT.

In this example, our censored point t, was a constant. What happens if it is a variable. Take an example in the medical field, where we are performing a clinical trial and we denote T_i as the time of death from our disease of interest and V_i are the time of death from other causes. Now our censored point, V_i is a random variable.

Again, let's assume that T_i is iid from an Exponential with parameter λ_0. Let's assume that V_i is iid from an Exponential with parameter μ_0, and the T_i's are independent from the V_i's. Our likelihood function is:

$L(\lambda, \mu) = \prod_{i=1}^{n}(\mathbb{1}(T_i \leq V_i)\lambda e^{-\lambda T_i} e^{-\mu T_i} + \mathbb{1}(T_i > V_i)\mu e^{-\mu V_i} e^{-\lambda T_i})$
$= \prod_{i=1}^{n} e^{(\lambda+\mu)(min(T_i, V_i))} \lambda^{\sum_{i=1}^{n} \mathbb{1}(T_i \leq V_i)} \mu^{\sum_{i=1}^{n} \mathbb{1}(T_i > V_i)}$

We can then proceed to optimize the log-likelihood, and solve for the parameter estimates in closed form.

15.2 Non-Parametric Model with Censoring

Let's look at the case where we assume a non-parametric model. In a similar scenario where we are looking at deaths from a disease of interest, let us denote x where a patient dies from our disease of interest, and o where a patient dies from other causes. Let's say we observe $x, o, o, 2x, x, o, x, o$, where $2x$ means that 2 patients died simultaneously from the disease of interest. We then denote t_i to be the time of death of a patient only from the disease of interest. For example, the first x is t_1, the $2x$ is t_2, the next x is t_3, and the final x is at time t_4.

Let's say that we want to estimate non-parametrically the survival function, $\bar{F}(t) = P(T > t)$. We also know that $\bar{F}(0) = 1$, that all patients were alive at the start of the trial.

Since this is non-parametric, we only assume probability mass at the points we observed, and the mass at t_j is $\frac{P(T>t_j)}{P(T>t_{j-1})} = \frac{P(T>t_{j-1})-P(T\in(t_{j-1},t_j])}{P(T>t_{j-1})} = 1 - \frac{P(T\in(t_{j-1},t_j])}{P(T>t_{j-1})}$.

Therefore, $\hat{\hat{F}}(t_i) = \prod_{j=1}^{i}[1-\frac{\hat{P}(T\in(t_{j-1},t_j])}{\hat{P}(T>t_{j-1})}]$. Therefore, we can calculate the probability mass at all four points t_1 to t_4 as follows:

$\hat{\hat{F}}(t_1) = 1 - \frac{1}{9}$
$\hat{\hat{F}}(t_2) = (1 - \frac{1}{9})(1 - \frac{2}{8})$
$\hat{\hat{F}}(t_3) = (1 - \frac{1}{9})(1 - \frac{2}{8})(1 - \frac{1}{4})$
$\hat{\hat{F}}(t_4) = (1 - \frac{1}{9})(1 - \frac{2}{8})(1 - \frac{1}{4})(1 - \frac{1}{3})$

Surprisingly, these 4 probabilities do not sum to 1. That is because there is some mass left over for estimation for survival past time t_4. In order words, the remainder of the mass is at t_∞. This estimator that we just calculated is called the Kaplan-Meier non-parametric estimator for censored data.

16 The Bootstrap

Assume we have $X_1, X_2, ...X_n$ iid from P_{θ_0}, and say we want an estimator and confidence interval for $\alpha_{\theta_0} = E_{\theta_0}[g(X_1)]$. To do this from the methods previously discussed, we could use the delta method, but that requires us to calculate $\nabla_\theta \alpha(\hat{\theta})$. The delta method can be difficult to implement when $\nabla_\theta \alpha(\cdot)$ is challenging to compute. Hence we will look toward a method called bootstrapping which can give us an approximate confidence interval of the parameter of interest.

16.1 Parametric Bootstrap

Assume that we have $X_1, X_2, ..., X_n$ iid from a $N(\mu_0, \sigma_0^2)$, and our goal is to produce a confidence interval for the population quantile $q = F_{\theta_0}^{-1}(p)$. We then calcluate:

$F_{\theta_0}(q) = p$
$P(N(\mu_0, \sigma_0^2) \leq q) = p$
$P(\mu_0 + \sigma_0 N(0,1) \leq q) = p$
$P(N(0,1) \leq \frac{q-\mu_0}{\sigma_0}) = p$
$\Phi(\frac{q-\mu_0}{\sigma_0}) = p$
$\frac{q-\mu_0}{\sigma_0} = \Phi^{-1}(p)$
$q = \mu_0 + \sigma_0 \Phi^{-1}(p)$

Therefore, using plug in estimators, we could find the MLEs as $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$, and use these MLE to obtain $\hat{q} = \hat{\mu} + \hat{\sigma}\Phi^{-1}(p)$. That is our estimate of q. But how do we get a confidence interval for q?

First off, we note that if we we knew P_{θ_0}, we could get an exact confidence interval for $\hat{q} - q$. We would simply choose z_1, z_2 such that $P_{\theta_0}(z_1 \leq \hat{q} - q \leq z_2) = \alpha$. In other words, $P_{\theta_0}(\hat{q} - z_2 \leq q \leq \hat{q} - z_1) = \alpha$. Therefore, our exact confidence interval for q is $[\hat{q} - z_2, \hat{q} - z_1]$. However, we do not know the cdf of $\hat{q} - q$ under P_{θ_0}, but we can estimate it using Monte Carlo!

The idea is the same as Monte Carlo, where if we draw enough samples and calculate the estimate, we will get a confidence interval due to the convergence. In this scenario, we would do the following to obtain the confidence interval, which is called bootstrapping.

1. Use the sample $X_1, X_2, ..., X_n$ to obtain $\hat{\mu}$ and $\hat{\sigma}$
2. Using a normal distribution with $\hat{\mu}$ and $\hat{\sigma}$ as the parameters, draw an iid sample $X_{11}^*, X_{12}^*, ..., X_{1n}^*$ (note this synthetic sample should be the same size as the original sample)
3. Do this sampling m times.
4. For each synthetic sample j, calculate $\hat{\mu}_j^*$ and $\hat{\sigma}_j^*$ and use these to calculate $\hat{q}_j^*$.
5. We now have m estimates of q, $\hat{q}_1^*, \hat{q}_2^*, ..., \hat{q}_m^*$
6. For each estimate of q obtained from a synthetic sample, calculate $\hat{q}_j^* - \hat{q}$, where $\hat{q}$ is from the original sample (not the synthetic sample)
7. Use this distribution of $\hat{q}_j^*$ to obtain z_1 and z_2.

Our parametric bootstrap confidence interval is $[\hat{q} - z_2, \hat{q} - z_1]$.

16.2 Non-Parametric Bootstrap

In the non-parametric setting, we assume $X_1, X_2, ..., X_n$ iid from a distribution P_0. Again, assume we want to estimate $q = F_0^{-1}(p)$. We follow the same procedure as in the parametric bootstrap. There are a few differences though, which arise because we don't have a parametric distribution. The first is that to estimate $\hat{q}$ from the original sample, we simply look at the quantiles of the sample, we do not need to calculate the MLEs. The second, is that because we don't have MLEs, when we generate the synthetic samples, we don't need to generate data from a distribution. All we do is we sample with replacement the values in the original sample.

Following the same procedure as in the parametric case, we have our confidence interval $[\hat{F}^{-1}(p) - z_2, \hat{F}^{-1}(p) - z_1]$, where $\hat{q} = \hat{F}^{-1}(p)$.

16.3 Hypothesis Testing with Bootstrap

Bootstrap can also be used to determine the definition of the critical region to be used for hypothesis testing so as to set a givne level of significance. Remember that in hypothesis testing, we reject the null if $P_0(G \leq c) \leq \alpha$. If we cannot compute c in closed form, then one way of computing c is to use the bootstrap method.

17 Bayesian Methods

There are times when taking a Bayesian point of view rather than a frequentist point of view produces a more sensible answer. With a Bayesian perspective, we assume that the parameter itself is a random variable and not a constant. Let's look at an example where a Bayesian approach would produce a more sensible answer than a frequentist approach.

Let's pretend that we have launched a rocket 5 times and all have been successes. We want to know what the probability is of the next launch being a success. Using a frequentist methodology, we would have $X_1, X_2, ..., X_5$ iid Bernoulli with parameter p_0. We would calculate the MLE, which for a Bernoulli, we have $\hat{p} = \bar{X}_n$. In this case, we only have seen successes, so $\hat{p} = \bar{X}_5 = 1$. Logically, this is a nonsensical answer as practically, we wouldn't expect to be 100% confident that the launch would be successful given only seeing 5 previous launches.

With a Bayesian approach, we have $X_1, X_2, ..., X_5$ iid Bernoulli, but this time with parameter p_*, where p_* is itself a random variable. In the absence of having any data, we can see that p_* is uniform from 0 to 1. In other words, having seen no data, we have no reason to believe otherwise than that there is an equal probability of p being any value between 0 and 1. This is called the "prior" on p_*. We then want to look at the distribution of p_* given observed data $X_1, X_2, ..., X_5$, which is called the "posterior". Mathematically, the posterior distribution can be written as $f(p|data)$.

Bayes Rule tells us that $P(A|B) \propto P(A \cap B) \propto P(B|A)P(A)$. Therefore, we can write $f(p|data) \propto L(data|p_* = p)f(p)$. In this case, we have that $f(p|data) \propto p^5 * 1$, since $f(p)$, our prior, is a uniform from 0 to 1. Our posterior distribution must integrate to 1 to be a proper distribution, so we need to scale it by a constant. We find that constant by $\int_0^1 p^5 dp = \frac{1}{6}$. Therefore, our posterior distribution $f(p|data) = 6p^5$.

Notice that the mode of this posterior distribution is when $p = 1$, and we get a value of 6. Notice that the value 1 is exactly the same in this case as the frequentist MLE estimate. Generally, the Bayesian mode of posterior density will lie very close to the frequentist MLE. When we look at the posterior mean, we get that the estimate is $\int_0^1 pf(p|data)dp = \int_0^1 6p^6 = \frac{6}{7}$. This estimate makes much more sense in this case than the frequentist estimate of 1.

17.1 Easy to Compute Bayesian Examples

The first example is a generalization of the previous Bayesian problem. Again, we have a prior which is uniform on [0, 1]. We have $X_1, X_2, ..., X_n$ which are iid Bernoulli random variables with parameter p_*. Therefore, the sum of these random variables is a binomial with parameters p_* and n. Also, since it is Bernoulli, $S_n = \sum_{i=1}^n X_i$. Therefore, we have:
$f(p|data) \propto L(data|p_* = p)f(p)$
$= \binom{n}{S_n} p^{S_n}(1-p)^{n-S_n} * 1$
$\propto p^{S_n}(1-p)^{n-S_n}$

We take note of a special function called the Beta function. The Beta function is defined as $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}$, where $a, b > 0$. When a, b are integers, then we can express $B(a,b)$ in closed form where $B(a,b) = \frac{(m-1)!(n-1)!}{(m+n-1)!}$. Therefore, we note that the normalizing constant for our posterior distribution then is a Beta function, which is $B(S_n + 1, n + 1 - S_n)$.

Therefore, $f(p|data) = \frac{p^{S_n}(1-p)^{n-S_n}}{B(S_n+1, n+1-S_n)}$. Therefore, we find our posterior mean to be $\int_0^1 pf(p|data)dp = \frac{S_n+1}{n+2}$.

The second example is where we have $X_1, X_2, ...X_n$ iid exponentially distributed with parameter λ_*. We have a prior for λ_* that is Gamma distributed with scale parameter r and shape parameter α. The Gamma pdf is supported on values ≥ 0, and the density takes the form $f(\lambda) = \frac{r(r\lambda)^{\alpha-1}e^{-r\lambda}}{\Gamma(\alpha)}$. Therefore, we have

$f(\lambda|data) \propto L(data|\lambda_* = \lambda)f(\lambda)$
$= \prod_{i=1}^n \lambda e^{-\lambda X_i} \frac{r(r\lambda)^{\alpha-1}e^{-r\lambda}}{\Gamma(\alpha)}$
$\propto \lambda^{n+\alpha-1} e^{-\lambda \sum_{i=1}^n X_i - \lambda r}$

We see that the resulting posterior density is another Gamma density in λ. Therefore, we can read from the density that it has scale parameter $\sum_{i=1}^n X_i + r$ and shape parameter $n + \alpha$. To calculate the posterior mean, we already know that the expected value of a gamma random variable is the shape parameter divided by the scale parameter. Therefore, the posterior mean is $\frac{n+\alpha}{\sum_{i=1}^n X_i + r}$.

The third example is where we have $X_1, X_2, ...X_n$ iid normally distributed with mean μ_* and variance, 1. We have a prior for μ_*, which is $N(0,1)$. Therefore, $f(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}}$. We can calculate

$f(\mu|data) \propto L(data|\mu_* = \mu)f(\mu)$
$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(X_i-\mu)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}}$
$\propto exp(-\frac{\mu^2}{2} - \frac{\sum_{i=1}^n (X_i-\mu)^2}{2})$
$\propto exp(-\frac{\mu^2}{2} - \frac{2\mu\sum_{i=1}^n X_i}{2} - \frac{n\mu^2}{2})$
$= exp(\frac{n(n+1)}{2}\mu^2 + \mu\sum_{i=1}^n X_i)$
$\propto exp(-\frac{n+1}{2}(\mu^2 - \frac{2\mu\sum_{i=1}^n X_i}{n+1} + (\frac{\sum_{i=1}^n X_i}{n+1})^2))$
$= exp(-\frac{1}{2}(\mu - \frac{\sum_{i=1}^n X_i}{n+1})^2 * \frac{1}{\frac{1}{n+1}})$

Therefore, we can see that the posterior distribution is another normal distribution with mean $\frac{\sum_{i=1}^n X_i}{n+1}$ and variance $\frac{1}{n+1}$.

17.2 Bayesian Methods for Predictions

Assume that after observing $X_1, X_2, ..., X_n$ iid from a normal distribution with mean μ_* and variance 1, and we want to predict $P(X_{n+1} > z)$. Using the frequentist approach, we would estimate μ_* through the MLE, which in this case is $\bar{X}_n$, and then plug it in to estimate $P(X_{n+1} > z)$. When we obtain the plug in estimate this way, the estimate is not affected by how precise the MLE estimate is of the underlying parameter. In other words, the plug in doesn't know if the $\bar{X}_n$ is based on a sample size of 5 or of 500. It has decoupled the estimation problem in the MLE from the prediction problem in the plug in.

In the Bayesian approach, we want to estimate $P(X_{n+1} > z|X_1, ..., X_n)$. We can rewrite $X_{n+1} \stackrel{D}{=} \mu_* + Z$, where Z is $N(0, 1)$. From the previous section, we found the posterior of μ_* to be a normal with mean $\frac{\sum_{i=1}^n X_i}{n+1}$ and variance $\frac{1}{n+1}$. Therefore, we can write $\mu_* + Z \stackrel{D}{=} \frac{\sum_{i=1}^n X_i}{n+1} + \sqrt{\frac{1}{n+1}}Z' + Z$, where $Z' = N(0, 1)$. We can rewrite the last statement as $\frac{\sum_{i=1}^n X_i}{n+1} + \sqrt{1 + \frac{1}{n+1}}N(0, 1)$. Now we have a nice integrated solution where the probability of $X_{n+1} > z$ is dependent on the number n of X_i's that we have seen. We can also read off how much the uncertainty in the distribution of the mean, μ_* is impacting our probability assessment that $P(X_{n+1} > z)$.

18 Linear Regression

In linear regression, we try to use a linear model to predict a response variable given a set of explanatory variables. The general form of the model is $Y_i = x_i a_0 + \epsilon_i$, where a_0 is the row vector of coefficients and ϵ_i is the column vector of errors. Under a standard scenario, we may assume that ϵ_i is iid from $N(0, \sigma_0^2)$. In matrix form, we have that $Y = Xa_0 + \epsilon$, where $\epsilon \stackrel{D}{=} N(0, \sigma_0^2 I)$. This is a homoscedastic assumption, as shown by the identity matrix. This is a parametric statistical model, and therefore we can solve MLEs for our unknown parameters, a_0 and σ_0^2. We obtain the MLEs to be $\hat{a} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - x_i\hat{a})^2$, or the estimated residual sum of squares. In this scenario, we are able to construct exact confidence intervals for a_0 and σ_0^2.

One change we can make to the model is when we assume the ϵ_i's to be non-Gaussian. In other words, we still have ϵ_i iid with $E[\epsilon_i] = 0$, but we do not have a parametric distribution from which they come from. Therefore, the ϵ_i is non-parametric. However, we still have that $Y = Xa_0 + \epsilon$, and since Y is still linear in X, we have that this model is semi-parametric. To find the best linear unbiased estimator (BLUE), we obtain $\hat{a} = (X^T X)^{-1} X^T Y$ and $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - x_i\hat{a})^2$. However, now to find a confidence interval for the coefficients, $\hat{a}$, we need to use the bootstrap method. We do this by reampling ϵ, since we assume that it is random and the X is fixed. Therefore we generate $\hat{\epsilon_1}, ..., \hat{\epsilon_n}$, by calculating $\hat{\epsilon}_i = Y_i - x_i\hat{a}$ (we use $\hat{a}$ since we don't know the true a_0). We then generate bootstrap samples $Y_{1j}^* = x_j\hat{a} + \hat{\epsilon_1}^*$ for $1 \le j \le n$, where these ϵ_{1j}^* are sampled randomly uniformly from $\hat{\epsilon_1}, ..., \hat{\epsilon_n}$. We then are able to calculate $\hat{a_1}^* = (X^T X)^{-1} X^T Y_1^*$. We repeat this m times so that we have $\hat{a_1}^*, \hat{a_2}^*, ..., \hat{a_m}^*$. We now have a distribution where we can find a z_1 and a z_2 to make a confidence interval, which is $[\hat{a} - z_2, \hat{a} - z_1]$.

We can also alter the linear regression model, where we now view not only the ϵ_i as random, but the X_i as well. Therefore, we still have $Y_i = X_i a_0 + \epsilon_i$, where ϵ_i is iid and $E[\epsilon_i] = 0$ (non-parametric), but we also assume that the ϵ_i are independent of the X_i. Using our OLS estimator, we again get that $\hat{a} = (X^T X)^{-1} X^T Y$. For confidence intervals for the coefficients, now we need to bootstrap not only the ϵ_i, but also the X_i. We sample X_{ij}^* from the empirical distribution that we saw in our sample. We also sample our $\hat{\epsilon_{ij}}^*$ from the empirical distribution from our sample. We now can compute $Y_{ij}^* = x_{ij}^*\hat{a} = \hat{\epsilon_{ij}}^*$. We can compute our OLS estimator by $\hat{a}_i^* = (X_i^{T*} X_i^*)^{-1} X_i^{T*} Y_i^*$, and we do this m times. Now we have $\hat{a_1}^*, \hat{a_2}^*, ..., \hat{a_m}^*$. Again, we can find a z_1 and a z_2 to make a confidence interval, which is $[\hat{a} - z_2, \hat{a} - z_1]$.

So far we have assumed homoscedasticity in our ϵ_i. In practice, often the variance can scale in proportion to our X_i. They are not homoscedastic, but rather heteroscedastic. Therefore, we can alter our model to have a w_i which is a function of X_i. Now our model is $Y_i = X_i a_0 + w_i\epsilon_i$. To solve this problem, we want to find $min_a \frac{1}{n}\frac{\sum_{i=1}^n (Y_i - X_i a)^2}{w_i^2}$. This is a weighted least squares problem. Note that now we have that $\hat{\epsilon_i} = \frac{Y_i - x_i\hat{a}}{w_i}$. When computing confidence intervals, we now also bootstrap w_i^*, which we will obtain from the empirical distribution of the sample. When we bootstrap, we now have $Y_{ij}^* = x_{ij}^*\hat{a} + w_{ij}^*\epsilon_{ij}^*$. Again, we can calculate $\hat{a_1}^*, \hat{a_2}^*, ..., \hat{a_m}^*$. Again, we can find a z_1 and a z_2 to make a confidence interval, which is $[\hat{a} - z_2, \hat{a} - z_1]$.

We have looked at models that assume homoscedasticity and ones that assume heteroscedasticity. If we make the incorrect assumption for what the true model is, for large enough n, we still converge to the correct $\hat{a}$. In other words, our $\hat{a} \stackrel{P}{\to} a_0$. This is great, but we still do get a cost of mis-specifying the model, which is that our estimator $\hat{a}$ will be worse in terms of MSE.

Let us now return back to the parametric setting, except now ϵ is distributed $N(0, \sigma_0^2 C)$. Our MLE is now solved by $min_a (Y - Xa)^T C^{-1} (Y - Xa)$. We obtain $\hat{a} = (X^T C^{-1} X)^{-1} X^T C^{-1} Y$.

We can also take a Bayesian point of view for linear regression, where the parameter a_0 is it itself a random variable. Let us call it now a_*. Therefore, we have $Y = Xa_* + \epsilon$, where $\epsilon \stackrel{D}{=} N(0, \sigma_0^2 C)$. We now can select a prior for a_*. Let us pick $a_* \stackrel{D}{=} N(0, \sigma_0^2 \Lambda_0)$. Therefore $f(a) \propto exp(-\frac{1}{2\sigma^2} a^T \Lambda_0^{-1} a)$. Assuming that $\epsilon \stackrel{D}{=} N(0, \sigma_0^2 I)$, our likelihood function is $L(data|a_* = a) \propto exp(-\frac{1}{2\sigma^2}(Y - Xa)^T (Y - Xa))$. Therefore, we have that

$f(a|data) \propto L(data|a_* = a) f(a)$
$\propto exp(-(a - \mu)^T (\frac{X^T X + \Lambda_0^{-1}}{2\sigma^2})(a - \mu))$

Therefore, the mean of the posterior is $\mu = (X^T X + \Lambda_0^{-1})^{-1} X^T X\hat{a}$. The covariance matrix of the posterior is $C = \sigma^2 (X^T X + \Lambda_0^{-1})^{-1}$. We can see that the effect of the prior is to shrink the posterior towards the mean. We can see that the posterior mean provides the solution that we obtain through ridge regression.

19 Gaussian Random Vectors and Fields

A Gaussian random vector $z \in \mathbb{R}^d$ is a vector that is characterized by its mean vector μ and by its covariance (symmetric) matrix C. For a Gaussian vector, the pdf, $f(z)$, which is the pdf of Z evaluated at z, is

$f(z) = (2\pi)^{-\frac{d}{2}}|det(C^{-1})|exp(-\frac{1}{2}(z-\mu)^T C^{-1}(z-\mu))$

The first property of Gaussian vectors is the following. If we have a Gaussian random vector $Z \stackrel{D}{=} N(\mu, C)$, and the take a linear combination $w = BZ + \nu$, where B is a deterministic matrix, then w is also a Gaussian vector characterized by $w \stackrel{D}{=} N(B\mu + \nu, BCB^T)$. This will be helpful for how we simulate Gaussian random vectors.

19.1 Simulating Gaussian Random Vectors

It is difficult for us to simulate Gaussian random vectors directly by the inversion method, since there is no closed form CDF of the Gaussian distribution. So we must change our approach in order to simulate $w \stackrel{D}{=} N(\mu, C)$.

Let us first assume that we already have a simulated random variable that is $N(0,1)$. We will later discuss how to simulate this random variable, but for now, let us assume we already have it. Let us first start with what happens when $d = 1$. When $d = 1$, $w \stackrel{D}{=} N(\mu, \sigma^2) \stackrel{D}{=} \mu + \sigma N(0,1)$. Therefore, we just need to take the square root of σ, multiply it by the $N(0,1)$ random variable, add μ, and then we have w.

In the general case now, where $d \geq 1$, we have a simulated random variable already of $Z \stackrel{D}{=} N(0, I)$. By the property mentioned above, we have that $w \stackrel{D}{=} N(\mu, C) = \mu + BZ$. Therefore, if we let $w = \mu + BZ$, then we have a $N(\mu, BB^T)$, so we must select the B such that $BB^T = C$. As it turns out, there can be multiple B that satisfy $BB^T = C$, but we want to find the B that is the most tractable. Therefore, we select B such that it is lower triangular, and we do this by using the Cholesky factorization.

An example of how to do this factorization in $d = 2$ is here. Let
$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; C = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$
Then we are trying to solve the system of equations where $w_1 = \mu_1 + \sigma_1 Z_1$ and $w_2 = \mu_2 + az_1 + bz_2$. Therefore, we have $Var(w_2) = \sigma_2^2 = a^2 + b^2$ and $Cov(w_1, w_2) = \rho\sigma_1\sigma_2 = \sigma_1 a$. Therefore, we can solve that $a = \rho\sigma_2$ and $b = \sigma_2\sqrt{1-\rho^2}$. We can generalize this recursive solve of system of equations to higher d.

Now that we understand how to create w from a linear combination of Z normal variables with mean 0 and covariance matrix I, we need to understand how to simulate Z. The easiest way is to simulate two $N(0,1)$ variables at once.

We start by realizing that the CDF, $\Phi(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} exp(-\frac{-x^2}{2})dx = 1$. Therefore, we have that $(\int_{-\infty}^{\infty} exp(-\frac{-x^2}{2})dx)^2 = 2\pi$. We then can see that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} exp(-\frac{(x^2+y^2)}{2})dxdy = 2\pi$. Therefore, we have the joint distribution, $f(z_1, z_2) = \frac{1}{2\pi} exp(-\frac{(z_1^2+z_2^2)}{2})$.

We recognize the radial symmetry and rewrite this in terms of polar coordinates. Therefore, $z_1 = Rcos\theta$ and $z_2 = Rsin\theta$. Since we can see that all angles are evenly distributed, then θ is uniformly distributed over $[0, 2\pi]$. Therefore, we can simulate a random θ, by selecting a random uniform on $[0,1]$ and multiplying it by 2π. We then look at $R = \sqrt{Z_1^2 + Z_2^2}$ and therefore $R^2 = Z_1^2 + Z_2^2$. We recognize that R^2 is the sum of the squares of two independent normals and therefore R^2 is a chi-squared random variable with 2 degrees of freedom. We then represent a chi-squared random variable with 2 degrees of freedom as 2 times an exponential random variable with mean 1. We can then generate this exponential using the inversion method.

To recap, to simulate 2 standard normal Gaussians, $N(0,1)$ at once, we first generate 2 random uniform numbers from $[0,1]$. We use one for an inversion method for an exponential with mean 1, and then multiply it by 2 to get R. We take the other and multiply it by 2π to get θ. We then perform $Rcos\theta$ and $R\sin\theta$ to get two independent standard normal random variables, Z_1 and Z_2.

One other property of Gaussian random vectors is that the conditional distribution of a Gaussian on another Gaussian is also Gaussian, so it can be characterized by its mean and covariance. Namely, assume we have $Z \in \mathbb{R}^d$ with $N(E[Z], C)$. We want to characterize the conditional distribution of Z_1, given Z_2.
$Z = \begin{bmatrix} Z_1 \\ \dots \\ Z_2 \end{bmatrix}; \tilde{Z} = Z - E[Z] = \begin{bmatrix} \tilde{Z_1} \\ \dots \\ \tilde{Z_2} \end{bmatrix}; C = \begin{bmatrix} E[\tilde{Z_1}\tilde{Z_1}^T] & \vdots & E[\tilde{Z_1}\tilde{Z_2}^T] \\ \dots & \dots & \dots \\ E[\tilde{Z_2}\tilde{Z_1}^T] & \vdots & E[\tilde{Z_2}\tilde{Z_2}^T] \end{bmatrix}$

Then the conditional mean is $E[Z_1|Z_2] = E[Z_1] + E[\tilde{Z}_1 \tilde{Z}_2^T](E[\tilde{Z}_2 \tilde{Z}_2^T])^{-1}[Z_2 - E[Z_2]]$, which we see is affine in Z_2, the item it was conditioned on. The covariance matrix of Z_1 given Z_2 is $E[\tilde{Z}_1 \tilde{Z}_1^T] - E[\tilde{Z}_1 \tilde{Z}_2^T](E[\tilde{Z}_2 \tilde{Z}_2^T])^{-1} E[\tilde{Z}_2 \tilde{Z}_1^T]$, which we notice is independent of the value of Z_2.

19.2 Gaussian Random Processes and Random Fields

Gaussian random processes and random fields are usually indexed by either time or space. If it is indexed by time, it is called a Gaussian random process. If it is indexed by space, it is called a Gaussian random field. For example, a Gaussian random process is a collection of Gaussian random variables at different points in time, such as $Z(t) : t \in [0, 1]$. An example of a Gaussian random field would be $Z(x, y) : (x, y) \in \mathbb{R}^2$. Therefore, more generally, we have $Z = Z(\lambda) : \lambda \in \Lambda$. We say that Z is a Gaussian random process/random field is all finite-dimensional distributions of Z are jointly Gaussian. In other words, if we look at any finite combination of $\lambda_1, \lambda_2, ..., \lambda_n \in \Lambda$, then the vector $(Z(\lambda_1), Z(\lambda_2), ..., Z(\lambda_n))^T$ is an n-dimensional Gaussian random vector. Such Gaussian random objects are characterized by their mean: $m(\lambda) = E[Z(\lambda)]$, and their covariance structure $c(\lambda_1, \lambda_2) = Cov(Z(\lambda_1), Z(\lambda_2))$.

In almost all applications, $\Lambda \subseteq \mathbb{R}^d$. We then say that a Gaussian random field/process is stationary if by changing our origin position in $\mathbb{R}^d$, we do not change the basic statistics of the random field/process. For example, no matter where we are, the mean is the same, $m(\lambda) = m, \lambda \in \Lambda$. Similarly, the covariance, $c(\lambda_1, \lambda_2) = c(\lambda_1 - \lambda_2), \lambda_1, \lambda_2 \in \Lambda$.

Many times, an assumption beyond stationary is used, namely that the Gaussian random field/process is isotropic. This stronger condition is as follows: a stationary Gaussian random field/process is said to be isotropic if $c(\lambda_1, \lambda_2) = c(||\lambda_1 - \lambda_2||)$, where $||x||$ is the length of x. In the stationary but not isotropic case, it still allows for preference in the covariance structure in some directions than in others. If we also add the assumption that it is isotropic, the covariance depends only on the distance between λ_1 and λ_2.

20 Markov Chains: Definitions and Examples

Assume we have a stochastic process on discrete time. We have $X = (X_n : n \geq 0)$. What the Markov property tells us is that $P(X_{n+1} \in \cdot | X_0, ..., X_n) = P(X_{n+1} \in \cdot | X_n)$. In other words, by conditioning on the previous X_i, we have all the information up to that point already, and conditioning on all the events prior to that one gives us no additional information. For a discrete state space, we can write $P(X_{n+1} = y | X_0, ..., X_n) = P(n+1, X_n, y)$. For a continuous state space, A, we can write $P(X_{n+1} \in A | X_0, ..., X_n) = \int_A p(n+1, X_n, y)dy$, where $p(n+1, X_n, y)$ is called the transition density.

A Markov chain is defined to have a stationary transition probability of $P(n) = P$. Given the Markov property, we can calculate the probability of a path, as $P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = \mu(x_0)P(1, x_0, x_1)P(2, x_1, x_2)...P(n, x_{n-1}, x_n)$, where $\mu(x_0)$ is the initial distribution. This way of assigning probabilities to the path sequence is called a "consistent specification of probabilities." By extension theorem, there exists a unique extension P to defining probabilities on $\Omega = S^\infty$.

We can look at the stochastic process through a stochastic recursion. $X_{n+1} = f(X_n, Z_{n+1})$. If the Z_i's are independent random variables, then it is equivalent to saying that X_{n+1} is a Markov chain and enjoys the Markov property.

20.1 Examples

This section provides examples of Markov chains. The first example is from inventory management, called (s, S) policies. As a business, we have inventory and once we sell until our inventory drops to the level s, we then reorder back up to the level S. Let $D_1, D_2, ...$ be iid. Then we can write $X_{n+1} = X_n - D_{n+1}$ when $X_n - D_{n+1} \geq s$, and otherwise $X_{n+1} = S$ when $X_n - D_{n+1} < s$. Our state space in this example is $S = \{s, s+1, s+2, ..., S\}$.

The next example is with slotted time queueing models. Assume that there is a line and we can only serve at most 1 customer at a time. Let X_n be the number of customers in the system at time n, and let Z_{n+1} be the number of new customers that come into the line at time $n+1$. Therefore, we can write $X_{n+1} = X_n + Z_{n+1} - 1$ when $X_n + Z_{n+1} \geq 1$, and $X_{n+1} = 0$, when $X_n + Z_{n+1} = 0$ (no customers waiting in line). In other words, $X_{n+1} = [X_n + Z_{n+1-1}]^+$. One caveat is that in this model, we assume that the customers come close to the beginning of the time period, that Z_{n+1} is close to n and not $n+1$. If we assume that the customers come clsoe to the end of the time period, that Z_{n+1} is close to $n+1$ and not n, then we instead write $X_{n+1} = [X_n - 1]^+ + Z_{n+1}$. Our state space in these Markov chains is $S = \{0, 1, 2, ...\}$

The next is an example with a continuous state space that comes from storage theory/hydrology. Suppose that a reservoir has water that at each time, has water coming in (input) and water being taken away (output). Then we can write the supply as $S_{n+1} = S_n + Z_{n+1} - O_{n+1}$. Let us assume that the output is a function of the supply, that $O_{n+1} = aS_{n+1}^b$, where $a, b > 0$. Then we have that $S_{n+1} + O_{n+1} = S_n + Z_{n+1}$, and then $S_{n+1} + aS_{n+1}^b = S_n + Z_{n+1}$. Therefore, we have that $S_{n+1} = v^{-1}(S_n + Z_{n+1})$, where $v(x) = x + ax^b$.

Notice that autoregressive sequences, which are defined as $X_{n+1} = \rho X_n + Z_{n+1}$, where Z_i is iid. Note that this is a specific case of the hydrology example above, when $b = 1$.

In the hydrology example, we may want to find the probability of a certain amount of supply at time $n+1$ given the supply at time n. In other words, we want to calculate $P(S_{n+1} \in A | S_n) = \int_A p(S_n, y)$. To compute this, we need to find the transition density, $p(S_n, y)$. We can do this by taking the CDF and then differentiating it to get the PDF.

$P_X(S_1 \leq y) = P_X(v(S_1) \leq v(y))$, since v is monotone
$= P_X(S_1 + aS_1^b \leq y + ay^b)$
$= P_X(S_0 + Z_1 \leq y + ay^b)$
$= P_X(x + Z_1 \leq y + ay^b)$
$= F_{Z_1}(y + ay^b - x)$

When the take the derivative, we get $\frac{d}{dy} P_X(S_1 \leq y) = f_{Z_1}(y + ay^b - x)(1 + aby^{b-1})$. Therefore, $P_X(S_1 \in A) = \int_A f_{Z_1}(y + ay^b - x)(1 + aby^{b-1})dy$.

20.2 Markov Chains as Matrices

We can write Markov Chains as matrices. Assume in a simple case where we have two states, 1 and 2. The state space is $S = \{1, 2\}$. If you are in state 1, you will move to state 2 with a probability of $\frac{2}{3}$, and will stay in 1 with a probability of $\frac{1}{3}$. If you are in state 2, there is a probability of $\frac{1}{2}$ that you will go to 1, and a probability of $\frac{1}{2}$ that you will remain in 2. We can write this in a matrix as follows:

$$P = \begin{bmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Note that each row in the matrix sums to 1. The two properties of a stochastic matrix are that $\sum_y P(x,y) = 1$ and that $P(x,y) \geq 0$. Note also that P times a column vector of 1's will always return a column vector of 1's because each row sums to 1. We will denote a column vector of all 1's as e. Therefore, $Pe = e$, and we can see that 1 is an eigenvalue of P.

Let us return back to the example of the slotted time queueing model. Our state space is the set of non-negative integers, $S = \mathbb{Z}_+$. Our transition dynamics now are going to be nearest neighbor dynamics. In other words, if we are in state i, we can only go to state $i-1, i$, or $i+1$. We can't jump from say state 3 to 1, we have to go from 3 to 2, then 2 to 1. Remember that in this system, we can only serve 1 customer per time period. We had written down $X_{n+1} = [X_n + Z_{n+1} - 1]^+$ as our model. Let us denote the probabilities, $P(Z_i = 0) = q$, $P(Z_i = 1) = r$, and $P(Z_i = 2) = p$, where $q+r+p = 1$. Then we can write our matrix as

$$P = \begin{bmatrix} 1-p_0 & p_0 & 0 & 0 \\ q_1 & r_1 & p_1 & \ddots \\ 0 & q_2 & r_2 & p_2 \\ 0 & \ddots & \ddots & \ddots \end{bmatrix}$$

This is a tridiagonal matrix.

Another popular graphical use of Markov chains is with Google's Page Rank algorithm, where transitions between nodes (webpages) are links between them.

21 Discrete Markov Chains

For this section, we will examine Markov chains with a discrete state space, S. In other words, $|S| < \infty$. We will look at both transient quantities as well as equilibrium. Transient quantities are influenced by the choice of the initial condition. In a Markov chain, it can we written as $X_{n+1} = f(X_n, Z_{n+1})$, where $P(X_0 \in A) = \mu(A)$ is the initial distribution. Equilibrium behavior on the other hand is not influenced by the choice of the initial condition. It relates to the "long term" behavior of the system after the influence of the initial condition has dissipated.

21.1 Transient Analysis

In transient analysis, we want to look at the two different sets of problems

- Distribution or expectation of a given time n in the future ($E_X[r(X_n)]$)
- Distribution of random variables that depend on the evolution of the path of X depending on the choice of $X_0 = x \in S$. For example the hitting time ($E_X[T]$)

If we have a Markov chain $X = (X_n : n \geq 0)$, and the transition dynamics of the system $P(X_{n+1} = y|X_n = x) = P(n+1, x, y)$, then we can denote a stochastic matrix $P(n+1) = (P(n+1, x, y) : x, y \in S)$. Let's say that we want to compute the probability at time n of the state being y given that we started at x. We have that $P(X_n = y \; X_0 = x) = \sum_{z_1, z_2, ..., z_{n-1}} P(1, x, z_1)P(2, z_1, z_2)...P(n, z_{n-1}, y)$.

If we calculate this product of matrices, we find that $P(X_n = y|X_0 = x) = (P(1)P(2)...P(n))(x, y)$. In other words, if we multiply all the intermediate matrices and then take the entry x, y of that new matrix, we will get the probability at time n of the state being y given that we started at x. More generally, if we want to find the probability of being in state y at time l when starting in state x at k, where $k < l$, then $P(k, l) = P(k+1)P(k+2)...P(l-1)P(l)$, and then we take the x, y entry of $P(k, l)$.

Computationally, this is quite powerful. In a general stochastic process if we were to calculate $P(X_n = y)$ without matrices and this property, we would have to do $\sum_{z_1, z_2, ..., z_{n-1}} P(X_0 = z_1, X_1 = z_2, ..., X_n = y)$, which requires summing all the different combinations. There are $|S|^n$ paths! Therefore, the number of terms in the sum is increasing exponentially in n. However, using the Markov chain matrix approach described above, we would do $P(0, n) = P(1)P(2)...P(n)$. Assuming each matrix is a square matrix with d rows/columns, each matrix product is $O(d^3)$, and therefore multiplying n of them is $O(nd^3)$, which is linear in n! We can even increase speed when the matrices are sparse.

21.1.1 Distribution or Expectation in the Future

Let's look at what happens if we take this transition dynamic matrix and right multiply by a vector r. Let's say for example that we have our matrix $P(k, l) = P(k+1)P(k+2)...P(l)$. We know from above that this is $P(X_l = y|X_k = x)$. If we right multiply by a vector r, we get that $P(k, l)r = \sum_y P(X_l = y|X_k = x)r(y) = E[r(X_l)|X_k = x]$. If we denote $r(y)$ as the "reward" associated with spending 1 unit of time in $y \in S$, and $r(X_l)$ as the random reward earned by the chain at time l, then $E[r(X_l)|X_k = x]$ can be viewed as the expected reward earned at time l, given that $X_k = x$.

To summarize, let's say we want to find the reward $E[r(X_n)|X_0 = x]$. Denote $u(n, x) = E[r(X_n)|X_n = x]$. Then we have trivially that $E[r(X_n)|X_n = x] = r(x)$. Going backwards, we then have $u(n-1, x) = E[r(X_n)|X_{n-1} = x] = \sum_y r(y)P(n, x, y) = (P(n)r)(x)$, the xth entry of that column vector. Going backwards again, we then have that $u(n-2, x) = E[r(X_n)|X_{n-1} = x] = \sum_y P(n-1, x, y)E[r(X_n)|X_{n-1} = y]$. We previously just computed that $E[r(X_n)|X_{n-1} = y] = (P(n)r)(y)$, so $u(n-2, x) = P(n-1)P(n)(y)$. Therefore, the pattern we see when we recurse backwards is that $u(n-j, x) = E[r(X_n)|X_{n-j} = x] = P(n-j+1)u(n-j+1) = P(n-j+1)...P(n)r$.

Let's look at what happens if we multiply $P(X_l = y|X_k = x)$ by a vector μ on the left. We get that $\mu P(X_l = y|X_k = x) = \sum_x \mu(x)P(X_l = y|X_k = x)$. If we view $\mu(x)$ as $P(X_k = x)$, the probability that the Markov chain was at state x at time k, then $\sum_x \mu(x)P(X_l = y|X_k = x) = P(X_l = y)$. In other words, we just get the probability that at time l, we are in state y.

To summarize, let's suppose that we want to compute $P(X_n = y)$ and that we are given $P(X_0 = x) = \mu(x)$. Then $\mu(j) = \mu(j-1)P(j)$.

Note that in Markov chains with stationary transition probabilities, we have that $u(n) = P(1)P(2)...P(n) = P^n r$, with $u(0) = r$, and $\mu P(1)P(2)...P(n) = \mu P^n$, with $\mu(0) = \mu$. For example, assume we have a 2-state Markov chain, where the probability of going from state 1 to state 2 is α, and from 2 to 1 is β. Then the matrix is

$$P = \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix}$$

We can find P's eigenvalues as 1 and $1-\alpha-\beta$. Therefore, if we calculate P^n, we get

$$P^n = \begin{bmatrix} \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \\ \frac{\beta}{\alpha+\beta} & \frac{\alpha}{\alpha+\beta} \end{bmatrix} + (1-\alpha-\beta)^n \begin{bmatrix} \frac{\alpha}{\alpha+\beta} & -\frac{\alpha}{\alpha+\beta} \\ -\frac{\beta}{\alpha+\beta} & \frac{\beta}{\alpha+\beta} \end{bmatrix}$$

We can see in the case where $|(1-\alpha-\beta)| < 1$, then as $n \to \infty$, the second piece goes to 0, and therefore P^n equals only the first piece. Notice that the rows are identical, and therefore $P^n(x,y) \to \pi(y)$. In this case, we see that the P^n value is independent of the row x, which means that the value is not dependent on the initial condition. We see that X_0 and X_n are asymptotically independent, and therefore, this is equilibrium behavior. In the case where $1-\alpha-\beta = 1$, then $\alpha = \beta = 0$, and then P is the identity matrix. This means that if I start off in state 1, I stay there forever, and if I start off in state 2, I stay there forever. This is called a reducible Markov chain, because it can be analyzed as two separate chains. Lastly, if $1-\alpha-\beta = -1$, then $\alpha = \beta = 1$. In this case, if I am in state 1, I will always go to state 2 in the next time, and if I am in state 2, I will always to state 1 next time. This is called a periodic Markov chain. In this case, the period number is 2.

21.2 First Transition Analysis

Denote $T = inf\{n \geq 0 : X_n \in C^C\}$. This is the hitting time of C^C. One example of where this would be used is if we want to find the first time that a system becomes unavailable, where C^C is the set of states where the system is unavailable. What we want to calculate is $E_X[T]$, which clearly depends heavily on our starting state x. Denote $u^*(x) = E_X[T]$. Obviously, if $u(x) = 0$, then $x \in C^C$ (we start off in the set where the system is unavailable). If $x \in C$, then we are solving a system of equations. We have that $u(x) = 1 + \sum_{y \in C} P(x,y)u(y)$, since we take one step, hence the 1, and then we then sum over all the states that are not in C^C (the states in C^C would just be adding zeros). If we denote the matrix B as the part of the transition matrix P, where we go from a state in C to another state in C, then we have that $u = e + Bu$. We don't need to consider when we go to C^C because then the chain stops, so it is just 0. Recall that e is the vector of all 1's. Note that one solution to this is $\infty = \infty$.

Imagine another example where we want to look at the expected reward of an investor at time n. In order words, we want to find $E_X[\sum_{n=1}^{\infty} e^{-\alpha n} r(X_n)]$. Therefore, denote $u^*(x) = E_X[\sum_{n=1}^{\infty} e^{-\alpha n} r(X_n)]$. Using first transition analysis, we have the system of equations that $u = r(x) + e^{-\alpha} \sum_{y \in S} P(x,y)u(y)$. In vector form, we have $u = r + e^{-\alpha} Pu$.

Another example is the Gambler's Ruin problem, where a gambler have initial wealth x and is gambling with another person with initial wealth y. The game ends when the gambler reaches wealth of 0 dollars (goes broke) or a wealth of $x+y$ dollars (wins all of the other person's money). Each game he either wins a dollar or loses a dollar. Say we want to find the probability that our gambler in ruined (the scenario where he ends up with 0 dollars). Denote the hitting time $T = inf\{n \geq 0 : X_n : \{0, x+y\}\}$. We want to find $P_X(X_T = 0)$. Note that in this case, the game ends in two ways, so C^C is two different states. However, we want to calculate the probability that the game ends when the gambler hits 0 dollars. Call this $A \subset C^C$. In this case, we have that $u^*(x) = P_X[X_T \in A, T < \infty]$. We have that $u(x) = 1$ if $x \in A$, and $u(x) = 0$ if $x \in C^C \backslash A$. For $x \in C$, we have the system of equations that $u(x) = \sum_{y \in A} P(x,y) + \sum_{y \in C} P(x,y)u(y)$. If we let the vector $f = (f(x) : x \in C)$, then we can write the equation out in matrix form as $u = f + Bu$, where B again is the submatrix of P that goes from C to C (not C^C).

What if we want to compute moments of the hitting time T? We are now interested in $u_k^*(x) = E_X[T^k]$. In order to compute u^*k, we need to first have computed $u_1^*, u_2^*, ...u_{k-1}^*$. Let's look at the second moment and assume that we have already computed u_1^*. Remember that u_1^* satisfies the linear system $u_1 = e + Bu_1$. We have

$u_2(x) = E_X[T^2]$
$= E_X([1+T']^2)$
$= E_X(1 + 2T' + T'^2)$
$= 1 + 2E_X[T'] + E_X[T'^2]$
$= 1 + 2\sum_{y \in C} P(x,y)u_1(y) + \sum_{y \in C} P(x,y)u_2(y)$

If we let $f(x) = 1 + 2\sum_{y \in C} P(x,y)u_1(y)$, then we can write this equation in matrix vector form as $u_2 = f + Bu_2$.

21.2.1 Conditions for Finite Solutions

Looking at the first transition analysis equations, we have been able to express $u = f + Gu$, where G is a non-negative matrix. We can rewrite this as $(I-G)u = f$, and then left multiplying by the inverse, we have $u = (I-G)^{-1}f$. However, this does not hold all the time. As we have seen, $\infty = \infty$ is sometimes the correct solution, and therefore there is no inverse in that case. We need to figure out when there is a non-infinite solution!

First note that we can expand $u = f + Gu$ as follows

$u = f + Gu$
$= f + G(f + Gu)$
$= f + Gf + G^2u$
$= f + Gf + G^2(f + Gu)$
$= f + Gf + ... + G^n + G^{n+1}u$

Therefore, we have that $u^* = \sum_{n=0}^{\infty} G^n f$. Therefore, if this sum is convergent, then we can have $(I - G)^{-1} f$, otherwise we cannot. Let us look at one specific case, the first moment of the hitting time T. We are interested in $u^*(x) = E_X[T]$, and we had seen previously that we can write out the equations as $u = e + Bu$. Therefore, we want to look at $\sum_{n=0}^{\infty} B^n e$. We have the following

$u^*(x) = E_X[T]$
$= E_X[\sum_{j=0}^{T-1} 1]$
$= E_X[\sum_{j=0}^{\infty} \mathbb{1}(T > j)]$
$= \sum_{j=0}^{\infty} E_X[\mathbb{1}(T > j)]$, we can do the interchange by Fubini's Theorem
$= \sum_{j=0}^{\infty} P_X[T > j]$
$= \sum_{j=0}^{\infty} \sum_{z_1 \in C, z_2 \in C, ..., z_j \in C} P(x, z_1)P(z_1, z_2)...P(z_{j-1}, z_j)$
$= \sum_{j=0}^{\infty} (B^j e)(x)$

Therefore, we have shown that $u^* = \sum_{j=0}^{\infty} B^j e$.

Now that we have in the general case that $u^* = \sum_{j=0}^{\infty} G^j f$, we examine more closely the equation $u = (I - G)^{-1} f$. Under what conditions does $(I - G)^{-1} = \sum_{j=0}^{\infty} G^j$? We have the theorem that the following are equivalent assuming non-negative G (every single entry of G is non-negative):

1. $(I - G)^{-1}$ exists and is non-negative
2. $\sum_{n=0}^{\infty} G^n < \infty$
3. $G^n \to 0$ as $n \to \infty$

We can start by proving that item 2) iff item 3). If $\sum_{n=0}^{\infty} G^n < \infty$, then $G^n \to 0$ as $n \to \infty$ is obvious. Thus 2) implies 3). Looking in the other direction, we first recall a few definitions. We recall that the infinity norm of a square matrix A is $||A|| = max_{x \in S} \sum_y |A(x, y)|$ (max of the absolute row sums). We recall that this norm satisfies the property that $||A + B|| \leq ||A|| + ||B||$ as well as $||AB|| \leq ||A||||B||$. Let us now assume that $G^n \to 0$ as $n \to \infty$. Then the infinity norm, $||G^n|| \to 0$ as $n \to \infty$. Thus, there exists an $m \geq 1$ such that $||G^m|| < 1$.

By the property of the norm, we have that $||G^{2m}|| = ||G^m * G^m|| \leq ||G^m||||G^m|| = ||G^m||^2$. More generally, we have that $||G^{nm}|| \leq ||G^m||^n$. Now let's look at the norm of G^{nm+k}. We have that $||G^{nm+k}|| \leq ||G^{nm}|| * ||G^k|| \leq ||G^m||^n * max_{0 \leq k \leq m-1} ||G^k||$. The value of $max_{0 \leq k \leq m-1} ||G^k||$ is just a constant. Therefore, $||G^{nm+k}||$ is going to 0 geometrically fast in n since $||G^m|| < 1$. Therefore, $\sum_{n=0}^{\infty} ||G^n|| < \infty$. As a result, $\sum_{n=0}^{\infty} G^n < \infty$. We have proved that 3) implies 2).

2) iff 3) actually holds even if G is negative, however, we will need the non-negativity of G in the proof regarding 1). Now we can prove that 1) implies 2). Suppose $(I - G)^{-1}$ exists and is non-negative. Note that $(I + G + ... + G^n)(I - G) = I - G^{n+1}$. We can see this when we expand out the expression and cancel the terms. Therefore, since G is non-negative,

$(I + G + ... + G^n)(I - G) \leq I$
$(I + G + ... + G^n)(I - G)(I - G)^{-1} \leq (I - G)^{-1}$; the inequality is preserved since G is non-negative
$I + G + ... + G^n \leq (I - G)^{-1}$
$0 \leq \sum_{n=0}^{\infty} G^n \leq (I - G)^{-1}$

Since our assumption was that $(I - G)^{-1}$ exists, then $\sum_{n=0}^{\infty} G^n < \infty$. Therefore 1) implies 2).

Now assume that 2) and 3) hold. We again note that $(I + G + ... + G^n)(I - G) = I - G^{n+1}$. From 3), we can see that $G^{n+1} \to 0$ as $n \to \infty$. Therefore, $\sum_{n=0}^{\infty} G^n(I - G) = I$. Then $(I - G)^{-1} = \sum_{n=0}^{\infty} G^n \geq 0$. Therefore, 2) and 3) together imply 1).

To summarize, our theorem now provides us a test. If we see that $(I - G)^{-1} \geq 0$, then $\sum_{n=0}^{\infty} G^n < \infty$. If we see that $(I - G)$ is singular, then at least one entry of $\sum_{n=0}^{\infty} G^n = \infty$. Similarly, if we see $(I - G)^{-1}$ exists, but it has at least 1 negative entry,

then at least one entry of $\sum_{n=0}^{\infty} G^n$.

Let's look at a specific example of this, the investment example that we had previously looked at where $u^*(x) = E_X[\sum_{n=0}^{\infty} e^{-\alpha n} r(X_n)]$. We have that $G = e^{-\alpha} P$. Assume now that $|S| = 2$, and the transition matrix is

$$P = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

In our first case, let $e^{-\alpha} = \frac{3}{4} < 1$, then

$$(I - G)^{-1} = \frac{1}{\frac{5}{16} - \frac{3}{32}} \begin{bmatrix} \frac{5}{8} & \frac{1}{4} \\ \frac{3}{8} & \frac{1}{2} \end{bmatrix} \geq 0$$

Therefore, $u^* = (I - G)^{-1} r$.

In our second case, let $e^{-\alpha} = 1$, then $u^* = \infty$. We can see that

$$(I - G)^{-1} = \begin{bmatrix} \frac{1}{3} & -\frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

We can see that $(I - G)$ is singular in this case.

In our third case, let $e^{-\alpha} > 1$, then $u^* = \infty$. We can see that

$$(I - G)^{-1} = \frac{1}{\frac{1}{3} - \frac{4}{3}} \begin{bmatrix} -1 & \frac{2}{3} \\ 2 & -\frac{1}{3} \end{bmatrix}$$

We can see that the matrix is non-negative.

21.2.2 More Rigorous Derivation

Let us take a look at the investment example where $u^*(x) = E_X[\sum_{j=0}^{\infty} e^{-\alpha j} r(X_j)]$. We can write this using a functional w, that takes in all the random sequence of X_j's as inputs. We define $w(x_0, x_1, x_2, ...) = \sum_{j=0}^{\infty} e^{-\alpha j} r(x_j)$. Therefore, we have that $w(x_0, x_1, x_2, ...) = r(x_0) + e^{-\alpha} \sum_{j=0}^{\infty} e^{-\alpha j} r(x_{j+1})$. Note that ths second term is just w applied to a shifted sequence (one that starts with x_1 instead of x_0). Therefore, we can write $w(x) = r(x_0) + e^{-\alpha} w(\theta \circ x)$, where $\theta \circ x$ is defined as the shifted sequence $(x_1, x_2, x_3...)$.

$u^*(x) = E_X[w(X)]$, where $X = (X_j : j \geq 0)$
$= E_X[r(X_0) + e^{-\alpha} w(\theta \circ X)]$
$= r(x) + e^{-\alpha} E_X[w(X_1, X_2, X_3, ...)]$
$= r(x) + e^{-\alpha} \sum_{y \in S} P(x, y) E_Y[w(X_0, X_1, X_2, ...)]$
$= r(x) + e^{-\alpha} \sum_{y \in S} P(x, y) u^*(y)$

This result is exactly what we got previously from First Transition Analysis, just more rigorously derived.

An example of this would be to apply it to the hitting time T. $T = inf\{n \geq 0 : X_n \in C^C\} = w(X)$. We also know that $w(X) = 1 + \mathbb{1}(X_1 \in C) w(\theta \circ X)$. Therefore, we showed that we again obtain our equation $u^*(x) = 1 + \sum_{y \in C} P(x, y) u^*(y)$.

21.3 Equilibrium Analysis

We can write our Markov chain as $X_{n+1} = f(X_n, Z_{n+1})$ where the Z_i's are iid. Let us assume in this case that we have stationary transition probabilities. Then we have statistical equilibrium when $X_n \Rightarrow X_\infty$. In the 2 state Markov chain, where the probability of going from state 1 to state 2 is $1 - \alpha$, and the probability of going from state 2 to state 1 is $1 - \beta$, we have seen that $P(X_n = y | X_0 = x) \to \pi(y)$ as $n \to \infty$, provided that $|1 - \alpha - \beta| < 1$. If we look at the matrix as a whole, $P^n = \Pi$, where

$$\Pi = \begin{bmatrix} \pi \\ \dots \\ \pi \\ \dots \\ \pi \\ \dots \\ \pi \end{bmatrix}$$, where each π is the same row vector.

What this is saying is that no matter which state we start at, when n is large, we will have some equilibrium probability distribution of being in state y. This is called "loss of memory" or "asymptotic independence." Therefore Pi is a rank one matrix that only depends on y.

Doing more analysis, we see that $P^n \to \Pi$, therefore $P^{n+1} \to \Pi$ as well. Therefore, we can write $P^{n+1} = P^n * P \to \Pi P$, as well as $P^{n+1} = P * P^n \to P\Pi$. Therefore, we have that $\Pi = \Pi P = P\Pi$. The row vector π must satisfy $\pi = \pi P$ such that $\sum_x \pi(x) = 1$ and $\pi(x) \geq 0$. This is called the "equilibrium distribution" or the "stationary distribution" or the "steady-state distribution."

We noted that this 2 state Markov chain only reaches the equilibrium distribution in the case where $|1 - \alpha - \beta| < 1$. In the case where $1 - \alpha - \beta = 1$, we have that $\alpha = \beta = 0$, and therefore the state changes every single time period and never settles down. As stated previously, this is called a periodic Markov chain. We have different behavior of P^n at n where n is even vs. when n is odd. Therefore, P^n does not converge.

In the case where $1 - \alpha - \beta = -1$, we have that $\alpha = \beta = 1$. In this case, we always stay in the state in which we started. If we started in state 1, we will always remain in state 1, and if we started in state 2 then we will always remain in state 2. Therefore, we have different behaviors of $P^n(x, y)$ depending on x. Therefore, there is no independence from the initial state, and we do not have $P^n \to \Pi$.

In that scenario, we could view the Markov chain as two separate Markov chains. We have the following definition: the Markov chain is called irreducible if for every $x, y \in S$, there exists $n = n(x, y)$ such that $P^n(x, y) > 0$. In our scenario, if we look at the Markov chain as a whole, it it NOT irreducible since there is no path from state 1 to state 2. However, since we can view it as two separate Markov chains, we can divide it into two different irreducible subclasses.

More generally, we have always decompose the state space S into $\mathcal{C}_1 \cup \mathcal{C}_2 \cup ... \cup \mathcal{C}_l \cup \mathcal{T}$, where $\mathcal{C}_i$ are the irreducible subclasses (the ones where when we get to that subclass it will always stay there) and $\mathcal{T}$ are the transient states (the ones that bounce back and forth finitely, but do not end up staying there eventually but rather eventually goes into one of the irreducible subclasses). We can then write our matrix P as a block matrix indexed by $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_l, \mathcal{T}$.

$$P = \begin{bmatrix} X & 0 & 0 & \dots & 0 & 0 \\ 0 & X & \ddots & \dots & 0 & 0 \\ \vdots & \ddots & X & \dots & 0 & 0 \\ \vdots & \ddots & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \dots & X & 0 \\ X & X & X & \dots & X & X \end{bmatrix}$$

The block of transient states is the last row. At $P^n(x, y)$, we can analyze this last row as the mixture of the equilibrium distributions for $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_l$. The mixture probabilities are $P_X(X_{\mathcal{T}} \in \mathcal{C}_i)$.

Going back to the case where we have a periodic Markov chain, we can partition the state space S into $S = P_1 \cup P_2 \cup ... \cup P_p$. We can then write our matrix P as a block matrix

$$P = \begin{bmatrix} 0 & X & 0 & \dots & 0 & 0 \\ \vdots & 0 & X & 0 & \dots & 0 \\ \vdots & \ddots & 0 & X & 0 & 0 \\ \vdots & \dots & \dots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \dots & 0 & X \\ X & 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

Note that all the diagonals are 0. If there exists x such that $P(x, x) > 0$ (the diagonal was not 0), then the Markov chain has a period $= 1$ and is therefore aperiodic.

We claim that if P is irreducible and aperiodic, then there exists $m \geq 1$ such that $P^m(x, y) > 0$ for every $x, y \in S$. We also claim that there exists an $n = n(x, y)$ such that $P^l(x, y) > 0$ for every $x, y \in S$ for all $l \geq n$. In other words, after a certain point n, we have a path between every single x and y.

An example of this is if we assume there exists $x \in S$ such that $P^3(x, x) > 0$ and $P^4(x, x) > 0$. To check aperiodicity, we look at the greatest common denominator to see if it is 1. If $gcd\{n \geq 1 : P^n(x, y) > 0\} = 1$, then it is aperiodic. In this case,

$gcd\{3,4\} = 1$, so it is aperiodic. If we write out time steps from 1 onwards, we can see that we can reach state x starting from x in time step 3, 6, 9, etc. since $P^3(x,x) > 0$. We can also reach x starting from x in time step 4, 8, etc. since $P^4(x,x) > 0$. We can reach time step 7 from a path of 3 and then a path of 4, and then 10 from two paths of 3 then one path of 4. In this case, $P^l(x,x) > 0$, where the magic number $n = 6$. For time steps greater or equal to 6, we can reach state x starting from x.

We note that when we prove theorems, we typically prove them for when Markov chains are finite irreducible and aperiodic. All theorems that are true for irreducible and aperiodic Markov chains are also true for irreducible and periodic Markov chains. This is because for a periodic Markov chain with period p, we can move along the cycle until some power of P has the property that $P(x,x) > 0$, in which case the matrix is aperiodic. Therefore, we can focus our attention on proving properties of irreducible, aperiodic Markov chains and know that this applies to irreducible, periodic Markov chains as well.

21.3.1 Loss of Memory Property

The loss of memory property, as will be shown below, shows that given certain conditions, the Markov chain will reach an equilibrium state and regardless of what its initial state was. We will now prove the loss of memory property assuming P is irreducible and with P having all its elements > 0. We can generalize this to P^n once we have proved it for the case of P.

Since P has every entry greater than 0, we can lower bound P by a constant c. Therefore, $P(x,y) \geq c > 0$ for all $x, y \in S$. Then, $P(x,y) \geq \frac{1}{d}(dc)$ for all $x, y \in S$. We denote $\delta = dc$, which is a positive scalar. We then denote $\frac{1}{d}$ as $\lambda(y)$, the uniform distribution on $|S|$. Therefore, we have that $P \geq \delta\Lambda$, where Λ is a rank 1 matrix where each row is $\lambda(y)$.

Since $P \geq \delta\Lambda$, we can write $P = \delta\Lambda + (1-\delta)Q$. Since $Q = \frac{(P-\delta\Lambda)}{1-\delta}$, Q is also a stochastic matrix. For a rank one stochastic matrix such as Λ, that we have in this case, for any matrix R, $R\Lambda = \Lambda$. Let's see what happens when we use this property and take powers of P.

$P^2 = P(\delta\Lambda + (1-\delta)Q)$
$= \delta\Lambda + (1-\delta)PQ$
$= \delta\Lambda + (1-\delta)(\delta\Lambda + (1-\delta)Q)Q$
$= \delta\Lambda + (1-\delta)\delta\Lambda Q + (1-\delta)^2Q^2$

$P^3 = P * P^2$
$= P[\delta\Lambda + (1-\delta)\delta\Lambda Q + (1-\delta)^2Q^2]$
$= \delta\Lambda + \delta(1-\delta)\Lambda Q + (1-\delta)^2PQ^2$
$= \delta\Lambda + \delta(1-\delta)\Lambda Q + (1-\delta)^2(\delta\Lambda + (1-\delta)Q)Q^2$
$= \delta\Lambda + \delta(1-\delta)\Lambda Q + (1-\delta)^2\delta\Lambda Q^2 + (1-\delta)^3Q^3$

Therefore, looking at the recursion, we see that $P^n = \sum_{j=0}^{n-1} \delta(1-\delta)^j\Lambda Q^j + (1-\delta)^nQ^n$. If we look at the infinity norm of any stochastic matrix, the infinity norm is 1. When we take the infinity norm of P^n, we get that

$||P^n|| = \sum_{j=0}^{n-1} \delta(1-\delta)^j||\Lambda Q^j|| + (1-\delta)^n||Q^n||$

As $n \to \infty$, the second term goes to 0. Therefore, $||P^n|| \to \sum_{j=0}^{n-1} \delta(1-\delta)^j < \infty$. Therefore, the first term converges absolutely and $P^n \to \sum_{j=0}^{n-1} \delta(1-\delta)^j\Lambda Q^j$ as $n \to \infty$, which we denote as Π.

The theorem that we have just proved states that if we have a Markov chain X that is irreducible and aperiodic with $|S| < \infty$, then $P^n \to \Pi$, where $\pi = \pi P$. The $\sum_x \pi(x) = 1$ and $\pi(x) \geq 0$ and $\pi(x) > 0$ for every $x \in S$.

One implication of this theorem is that if we have a reward function r, and we look at the expected value of the reward at time n, and we have an irreducible, aperiodic Markov chain, then I can write the following:

$E_x[r(X_n)] = \sum_y r(y)P_x(X_n = y) \to \sum_y \pi(y)r(y)$

What this means is that if we want to find the expected reward at time n, we can just multiply the vectors π and r.

One question we may want to ask if that if a reward at time n is πr, then does the sum of the rewards for the n times periods as $n \to \infty$ goes to $n * \pi r$? We want to know if $\sum_{j=0}^{n-1} r(X_j) \to n\pi r$. Dividing both sides by n, that is equivalent to asking if $\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) \to \pi r$. This is asking if the average reward is πr as n gets large. To prove this weak convergence, we assume a stationary stochastic process and use Chebyshev's Inequality. We also mean center our random variable $r(X_j)$ by defining a new random variable $\tilde{r}(X_j) = r(X_j) - \pi r$. Thus, we are trying to prove that $\frac{1}{n}\sum_{j=0}^{n-1} \tilde{r}(X_j) \xrightarrow{P} 0$.

Using Chebyshev's Inequality, we have that

$P_\pi(|\frac{1}{n}\sum_{j=0}^{n-1}\tilde{r}(X_j)| > \epsilon) \leq \frac{Var_\pi(\frac{1}{n}\sum_{j=0}^{n-1}\tilde{r}(X_j))}{\epsilon^2}$

Thus, we want to prove that the variance goes to 0 as $n \to \infty$. Let us look more closely at that term:

$Var_\pi(\frac{1}{n}\sum_{j=0}^{n-1}\tilde{r}(X_j)) = \frac{1}{n^2}(\sum_{j=0}^{n-1} Var_\pi\tilde{r}(X_j) + 2\sum_{i<j} Cov_\pi(\tilde{r}(X_i), \tilde{r}(X_j)))$
$= \frac{1}{n^2}(nVar_\pi\tilde{r}(X_0) + 2\sum_{i=1}^{n-1}(n-i)Cov_\pi(\tilde{r}(X_0), \tilde{r}(X_i)))$

Looking more closely at the covariance term:

$Cov_\pi(\tilde{r}(X_0), \tilde{r}(X_i)) = E_\pi[\tilde{r}(X_0), \tilde{r}(X_i)]$, since the $\tilde{r}(X_i)$ are mean 0 random variables
$= \sum_x \pi(x)\tilde{r}(x)E[\tilde{r}(X_i)|X_0 = x]$
$= \sum_x \pi(x)\tilde{r}(x)(p^i\tilde{r})(x)$

We know that $p^i\tilde{r} \to 0$ as $i \to \infty$. Therefore, the covariance terms go to 0. Therefore, $Var_\pi(\frac{1}{n}\sum_{j=0}^{n-1}\tilde{r}(X_j))$ goes to 0 and thus we have convergence in probability. We have proved that $\frac{1}{n}\sum_{j=0}^{n-1}\tilde{r}(X_j) \xrightarrow{P} 0$. We can also prove almost sure convergence as well.

Therefore, what we now have proved is that for a Markov chain X that is irreducible and aperiodic with $|S| < \infty$, there exists an equilibrium distribution $\pi = (\pi(y) : y > 0)$ such that $\pi(y) > 0$ for $y \in S$ and satisfies $\pi = \pi P$, subject to $\sum_x \pi(x) = 1$ and $\pi(x) \geq 0$. Furthermore, $P^n(x, y) \to \pi(y)$ as $n \to \infty$. Also, for any reward function r, we have $\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) \xrightarrow{a.s.} \pi r$ as $nto\infty$. Note that while the loss of memory property only holds in the aperiodic case, the almost sure LLN can be generalized to periodic Markov chains.

Let's look at an example of an Ehrenfest Markov chain. In this setting, we have two sides of a room, call it side A and side B. We have n molecules total in the room, and k are in side A of the room and $n-k$ are in side B of the room. We are interested in X_n, the number of molecules on side A of the room. At each time step, we select one molecule at random and move it to the other side of the room. Say we have j molecules at the current time. Our transition probability of going from j to $j+1$ is $\frac{n-j}{n}$. We can write a general birth death chain as follows:

$\pi(y) = \pi(y-1)p_{y-1} + \pi(y)r_y + \pi(y+1)q_{y+1}$; for $1 \leq y \leq n-1$

For the endpoints, we have different equations. We have $\pi(0) = \pi(0)r_0 + \pi(1)q_1$ and $\pi(n) = \pi(n)r_n + \pi(n-1)p_{n-1}$. We can solve this system of equations to get that $\pi(y) \propto \pi(0)\frac{p_0p_1\ldots p_{y-1}}{q_1q_2\ldots q_y}$. Since the $\pi(y)$ must sum to 1, we get that $1 = \sum_{y=0}^n \pi(y) = \pi(0)[1 + \sum_{y=1}^n \frac{p_0p_1\ldots p_{y-1}}{q_1q_2\ldots q_y}]$. Therefore, $\pi(0) = \frac{1}{1+\sum_{y=1}^n \frac{p_0p_1\cdots p_{y-1}}{q_1q_2\cdots q_y}}$.

In the Ehrenfest case, $\pi(y) = \binom{n}{y}2^{-n}$, since the distribution is binomial(n, $\frac{1}{2}$), where n is typically enormous, such as in the order of 10^{28}. Therefore, the mean is on the order of 10^{28} and the standard deviation is on the order of 10^{14}. Therefore, the number of molecules per side of the room remain roughly the same given that the mean is so much larger than the standard deviation.

Recall our SLLN for Markov chains $\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) \xrightarrow{a.s.} \pi r$. Let us specify our reward function r as 1 if $x = z$ and 0 otherwise. Then we have that $\frac{1}{n}\sum_{j=0}^{n-1} \mathbb{1}(X_j = z) \xrightarrow{a.s.} \pi(z) > 0$. We denote the nth time we have visited z as $T_n(z) < \infty$. We have that $T_n(z) \xrightarrow{a.s.} \infty$ as $n \to \infty$. Therefore, since we have two convergences almost surely, the subsequence also converges almost surely, so we can write $\frac{1}{T_n(z)}\sum_{j=0}^{T_n(z)-1} \mathbb{1}(X_j = z) \xrightarrow{a.s.} \pi(z) > 0$. Therefore, $\frac{n}{T_n(z)} \xrightarrow{a.s.} \pi(z)$, and hence we have that $\frac{T_n(z)}{n} \xrightarrow{a.s.} \frac{1}{\pi(z)}$. Let $T_n(z) = \tau_1(z) + \tau_2(z) + \ldots + \tau_n(z)$, where $\tau_i(z)$ is the time between visiting z for the ith time and the $(i+1)$th time. Each of these $\tau_i(z)$ is idd and we can invoke the SLLN to conclude that $E_z[\tau(z)] = \frac{1}{\pi(z)}$.

Therefore, what we have just proved is that if we want to find the expected number of times it takes to reach a certain state, we can just take the reciprocal of the equilibrium probability of reaching that state. For example, if in equilibrium, we visit that state with probability $\frac{1}{4}$, then we expect that it takes 4 time steps to reach that state one time.

We can apply this to the Ehrenfest model to figure out when one side of the room will have no molecules and the other side will have all the molecules. We recall that it is binomial, and therefore $P(Bin(n, \frac{1}{2}) = 0) = 2^{-n}$. Using the formula we just derived, $E_0[\tau(0)] = \frac{1}{\pi(0)} = 2^n$. Assuming that we have on the order of 10^{28} molecules, we will see this happen for the first time at $2^{10^{28}}$ time steps, which for all practical purposes, is never.

22 Stochastic Control

Stochastic control or stochastic optimal control is a sub field of control theory that deals with the existence of uncertainty either in observations or in the noise that drives the evolution of the system. We want to look at how to do static optimization, as well as in the case where we make a decision/action at each time step.

22.1 Static Optimization

Let's assume that we have a Markov chain $X = (X_n : n \geq 0)$ that has a dependence on a parameter θ. We first want to examine static optimization, which means that once we fix the θ, we cannot change it during the course of the Markov chain. Let $\theta \in \mathbb{R}^d$. Depending on how θ is set, it will affect the transition matrix, so $P(\theta) = (P(\theta, x, y) : x, y \in S)$. We have that reward function $r(\theta, x)$ and we want to maximize the total rewards $\sum_{i=0}^{n-1} r(\theta, X_i) \approx n \sum_x \pi(\theta, x) r(\theta, x)$. In other words, we want to find the $max_\theta \pi(\theta) r(\theta)$. Let us denote $\alpha(\theta) = \pi(\theta) r(\theta)$. Therefore, to optimize, we want to find the θ^* where $\nabla \alpha(\theta^*) = 0$. If we want to calculate this numerically, even a first order method would require $\nabla \alpha(\theta) = \nabla \pi(\theta) r(\theta) + \pi(\theta) \nabla r(\theta)$. We can compute $\nabla r(\theta)$ easily, but how do we compute $\nabla \pi(\theta)$?

We assume that $P(\theta)$ is irreducible and we start with the equilibrium formula where $\pi(\theta) = \pi(\theta) P(\theta)$. We can specify to the case where $d = 1$ to make the calculation simpler. Therefore, we have $\pi'(\theta) = \pi'(\theta) P(\theta) + \pi(\theta) P'(\theta)$. We rewrite this as $\pi'(\theta)(I - P(\theta)) = \pi(\theta) P'(\theta)$. However, $(I - P(\theta))$ is a singular matrix since P is a stochastic matrix and therefore has 1 as an eigenvalue. We need to use a trick and change the equation to $\pi'(\theta)(I - P(\theta) + \Pi(\theta)) = \pi(\theta) P'(\theta)$, where

$$\Pi(\theta) = \begin{bmatrix} \pi(\theta) \\ \pi(\theta) \\ \vdots \\ \pi(\theta) \end{bmatrix}$$

For this trick to be valid, we need to show that $\pi'(\theta)\Pi(\theta) = 0$. We have that $\sum_x \pi(\theta, x) = 1$ for any θ, therefore $\frac{d}{d\theta} \sum_x \pi(\theta, x) = 0$. Since $\sum_x \pi'(\theta, x) = 0$, then we have $\pi'(\theta)\Pi(\theta) = 0$, and the trick is valid.

Now, we have that $(I - P(\theta) + \Pi(\theta))$ is non-singular, so $\pi'(\theta) = \pi(\theta) P'(\theta)(I - P(\theta) + \Pi(\theta))^{-1}$, where $(I - P(\theta) + \Pi(\theta))^{-1}$ is called the "fundamental matrix."

22.2 Stochastic Control/Markov Decision Chain/Markov Decision Process

Assume we are in a scenario where we have customers from classes 1 to M, which we want to schedule. How we schedule these customers is going to depend on factors such as the revenue/customer, how long it takes to serve a customer, how patient the customer is, etc. If we schedule one customer in one class, it will clearly affect the waiting time for the customer from another class. Therefore, we don't want to make decisions based on just the next customer's revenue, but rather we have to think about impacts for the future as well to maximize the total revenue.

Assume we have a state space S and a set $a(x)$, which is a set of actions available for $x \in S$. Let $r(x, a)$ denote the reward or cost for using action a in state x. Then we have $P(X_n = y | X_0, A_0, ..., X_n, A_n) = P_{A_n}(X_n, y)$. The probability of ending up in state y at time n depends on the previous action and the previous state. We also assume that the decision maker is not able to look into the future to use information there to make a decision currently. This assumption is called "adaptedness."

What we want to do is to find a policy (a set of actions) that maximizes our total reward up to time n. We want to find the $sup_{(A_j : 0 \leq j \leq n)} E[\sum_{j=0}^n r(X_j, A_j) | X_0 = x]$. To solve this, we will use the principle of dynamic programming, specifically using backwards recursion. We start at time n, and note that we will simply take the action that maximizes our reward. Therefore, $v_n^*(x) = max_{a \in a(x)} r(x, a)$. How about at time $n-1$? At $n-1$, we can write $v_{n-1}^*(x) = max_{a \in a(x)} [r(x, a) + \sum_y P_a(x, y) v_n^*(y)]$. Intuitively, we would take the action the maximizes the reward of taking that action plus the expected value of the best choice at time n. We can recurse backwards like this and see that the general form is $v_i^*(x) = max_{a \in a(x)} [r(x, a) + \sum_y P_a(x, y) v_{i+1}^*(y)]$ for $0 \leq i \leq n$. A very nice feature of this approach is that the optimal control can be easily read off the value function!

This example has shown us what it would look like a finite time horizon, namely from time 0 to time n. Let's see what this would look like under an infinite time horizon. Let us look at the time value of money example, where we want to maximize the expected value of our money discounted by the time value of money. Now we have $v^*(x) = sup_{(A_n : n \geq 0)} E[\sum_{j=0}^\infty e^{-\alpha j} r(X_j, A_j) | X_0 = x]$. Using a first transition analysis argument, we have that $v^*(x) = max_{a \in a(x)} [r(x, a) + e^{-\alpha} \sum_y P_a(x, y) v^*(y)]$. This non-linear equation is called the "Optimality Equation," or "Bellman's Equation," or the "Hamilton-Jacobi-Bellman (HJB) Equation."

The v^* satisfies $v = T(v)$, where we have the operator $T(w)(x) = max_{a \in a(x)} [r(x, a) + e^{-\alpha} \sum_y P_a(x, y) w(y)]$. The operator T is a contraction. What the means is that assume we have two vectors w_1 and w_2. We define $||w_1 - w_2|| = max_{x \in S} |w_1(x) - w_2(x)|$.

We have a contraction since $||T(w_1) - T(w_2)|| \leq e^{-\alpha}||w_1 - w_2||$. Basically, the distance between w_1 and w_2, when applied to the operator T, gets shrunk.

If we apply successive approximations to our initial guess v_0, such that $v_1 = T(v_0)$, $v_2 = T(v_1)$ and so on, we have that $v_n = T(v_{n-1})$. We have that $v_n \to v_\infty$ as $n \to \infty$, and that $T(v_{n-1}) \to T(v_\infty)$, where v_∞ is a fixed point of T. Furthermore, since T is a strict contraction, it has a unique fixed point. We can show this by letting $w_1 = T(w_1)$ and $w_2 = T(w_2)$. Then by the contraction formula, $||T(w_1) - T(w_2)|| \leq e^{-\alpha}||w_1 - w_2||$, but $||T(w_1) - T(w_2)|| = ||w_1 - w_2||$ in this case, therefore, w_1 must equal w_2 and that is the unique fixed point. Therefore, we have concluded that $v_n \to v_\infty$ as $n \to \infty$. This method is called "value iteration." In practice, if we let these successive approximations run to a large n, and then we just take the action a at that time n such that $max_{a \in a(x)}[r(x,a) + \sum_y P_a(x,y)v_n(y)]$, that will be approximately optimal stochastic control.

Another method is to re-examine our value function, $v^*(x) = max_{a \in a(x)}[r(x,a) + e^{-\alpha} \sum_y P_a(x,y)v^*(y)]$ Since $v^*(x)$ is the max, we can write $v^*(x) \geq [r(x,a) + e^{-\alpha} \sum_y P_a(x,y)v^*(y)]$ for every $a \in a(x)$. We have an objective function $min \sum_x v^*(x)$. If we compute the solution, then we have a "linear programming" approach for computing the optimal control. However, this method is not always feasible if we have a large S.

Let's now take a look at average reward/average cost. We want to maximize our average reward, so we want to find the $sup_{(A_j : j \geq 0)} limsup_{n \to \infty} E[\frac{1}{n} \sum_{j=0}^{n-1} r(X_j, A_j) | X_0 = x]$. As it turns out, the optimality equation in this case is $v^*(x) + c = max_{a \in a(x)}[r(x,a) + \sum_y P_a(x,y)v^*(y)]$, where c is the equilibrium reward per unit time under the optimal control. We now have an extra unknown c to find on top of all the other unknowns, $(v^*(x) : x \in S)$. Note that there is no contraction here since there is no discounting. Therefore, we will use linear programming.

We will solve the system of the equations of $v^*(x) + c \geq [r(x,a) + \sum_y P_a(x,y)v^*(y)]$ and have the additional equation $min \sum_y v^*(y)$. All linear programming problems have duals, and so we will look at the dual of this LP, which is the $max \sum_{x,a} \pi(x,a)r(x,a)$ subject to $\sum_{a \in a(x)} \pi(y,a) = \sum_x \sum_{a \in a(x)} \pi(x,a)P_a(x,y)$, where $\sum_{a \in a(x)} \pi(y,a)$ is the equilibrium probability of being in state y and choosing action $a \in a(y)$ under the optimal control. As usual, it must hold that $\pi(x,a) \geq 0$ for every x and for every $a \in a(x)$ and that $\sum_{x,a} a(x,a) = 1$. Using this strategy, we can compute the $\pi(y,a)$. Once we have done that, we can calculate $\pi(a|x) = \frac{\pi(x,a)}{\sum_{a'} \pi(x,a')}$.

One example of stochastic control is the the optimal stopping problem. In this example, we may have an American option where the expiration time is time n. We can exercise that option any time between the current time and time n. Let $r(x)$ be the reward associated with stopping the system in state x, or in other words, exercising the option at price x. What we want to do is $max_T E_x[r(X_T)]$. Using backwards recursion, we have that at time n, we just get whatever the reward is at the time, in other words, $v_n^*(x) = r(x)$. At time $n-1$, we have that $v_{n-1}^*(x) = max[r(x), \sum_y P(x,y)v_n^*(y)]$. Our general optimality equation is that at time i, we have that $v_i^*(x) = max[r(x), \sum_y P(x,y)v_{i+1}^*(y)]$, where $0 \leq i < n$.

If we want to look at a discounted version of the this problem, we would now be finding $max_T E_x[e^{-\alpha T} r(X_T)]$. Thus, our optimality equation in this case is $v^*(x) = max[r(x), e^{-\alpha} \sum_y P(x,y)v^*(y)]$.

23 Markov Chains: Statistics and Filtering

Assume we observe a sequence of $X_0, X_1, X_2, ..., X_n$. Our goal is to fit a Markov chain model to this observed data. Using parameter estimation and CLT techniques, we are able to apply similar methods to Markov chains.

23.1 Parameter Estimation

We perform parameter estimation for Markov chains very similarly to how parameter estimation is done for a standard MLE parameter. Let us start with an example, the (s, S) policy. Recall that in this policy, we are selling a good and we are trying to model inventory for that good. We keep selling the good until we drop to the level s, and then we immediately order back up to the level S. From a formulaic standpoint, we have $X_{n+1} = X_n - D_{n+1}$ when $X_n - D_{n+1} \geq s$, and $X_{n+1} = S$ when $X_n - D_{n+1} < s$. If we assume linear costs on the number of items ordered, fixed cost on each order that is delivered, linear cost for holding items per period, and iid demand, then this is the optimal policy.

Let us assume that we have $X_0, X_1, ..., X_n$ observed inventory levels, and that our demand $D_1, D_2, ...$ is iid Poission with parameter λ. As a reminder, the pmf of a Poisson is $P(D_j = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, where $k \geq 0$. Let us assume that $s = 3$ and $S = 5$. We can write out our transition matrix as the following:

$$P(\lambda) = \begin{bmatrix} e^{-\lambda} & 0 & 1 - e^{-\lambda} \\ \lambda e^{-\lambda} & e^{-\lambda} & 1 - \lambda e^{-\lambda} - e^{-\lambda} \\ \lambda^2 \frac{e^{-\lambda}}{2} & \lambda e^{-\lambda} & 1 - \lambda^2 \frac{e^{-\lambda}}{2} - \lambda e^{-\lambda} \end{bmatrix}$$

A quick explanation of the matrix. We have three possible states, which are 3, 4, 5, hence the 3 columns and 3 rows. The 3,3 entry and 4,4 entry are $e^{-\lambda}$ since $P(D_j = 0) = e^{-\lambda}$. The right column entries are determined since for a stochastic matrix, each row must sum to 1. Note that there is positive probability going from 3 to 5 or 4 to 5 or 5 to 5 since there can be so much demand that it drops below s and we have to reorder back to S. Assume for 4 time steps we see inventory of 3, 5, 5, 4. We can write our likelihood function as the following (obtained from the matrix): $L(\lambda) = (1-e^{-\lambda})(1-\lambda^2\frac{e^{-\lambda}}{2} - \lambda e^{-\lambda})(\lambda e^{-\lambda})$.

More generally, if we denote $P(\theta) = (P(\theta, x, y) : x, y \in S)$, then the $L_n(\theta) = \prod_{j=1}^n P(\theta, X_{j-1}, X_j)$. To calculate the MLE, we can estimate the true parameter θ_0 via the maximizer of the likelihood $L_n(\cdot)$. The MLE $\hat{\theta}$ should satisfy $\nabla L_n(\hat{\theta}) = 0$, or equivalently, $\nabla \mathcal{L}_n(\hat{\theta}) = 0$, where $\mathcal{L}_n(\cdot)$ is the log likelihood.

Let us now develop a CI for our MLE estimate $\hat{\theta}$. We can use very similar ideas to estimating equations in order to do so. For this derivation, assume that θ is 1-dimensional and we can generalize to higher dimensions. By definition, we have the log-likelihood is $\mathcal{L}_n(\hat{\theta}) = \sum_{j=1}^n logP(\hat{\theta}, X_{j-1}, X_j)$. Therefore the derivative, $\mathcal{L}_n'(\hat{\theta}) = \sum_{j=1}^n \frac{P'(\hat{\theta}, X_{j-1}, X_j)}{P(\hat{\theta}, X_{j-1}, X_j)} = 0$. If we subtract from both sides, we get

$$\sum_{j=1}^n \frac{P'(\hat{\theta}, X_{j-1}, X_j)}{P(\hat{\theta}, X_{j-1}, X_j)} - \sum_{j=1}^n \frac{P'(\theta_0, X_{j-1}, X_j)}{P(\theta_0, X_{j-1}, X_j)} = -\sum_{j=1}^n \frac{P'(\theta_0, X_{j-1}, X_j)}{P(\theta_0, X_{j-1}, X_j)}$$

Although the right hand side is not a mean zero random variable as usual, it is very close and can provide us with a similar property. We denote $\frac{P'(\theta_0, X_{j-1}, X_j)}{P(\theta_0, X_{j-1}, X_j)} = D_j$ and claim that D_j is a Martingale difference. Therefore, we have

$E[D_j | X_0, ..., X_{j-1}] = E[\frac{P'(\theta_0, X_{j-1}, X_j)}{P(\theta_0, X_{j-1}, X_j)} | X_0, ..., X_{j-1}]$
$= E[\frac{P'(\theta_0, X_{j-1}, X_j)}{P(\theta_0, X_{j-1}, X_j)} | X_{j-1}]$ by the Markov property
$= \sum_y \frac{P'(\theta_0, X_{j-1}, y)}{P(\theta_0, X_{j-1}, y)} * P(\theta_0, X_{j-1}, y)$
$= \sum_y P'(\theta_0, X_{j-1}, y)$
$= \sum_y \frac{d}{d\theta} P(\theta_0, X_{j-1}, y)|_{\theta=\theta_0}$
$= \frac{d}{d\theta} \sum_y P(\theta_0, X_{j-1}, y)|_{\theta=\theta_0}$ but the row sums up to 1, which is a constant
$= \frac{d}{d\theta} c$
$= 0$

Therefore, we have a Martingale difference and therefore we can invoke the Martingale CLT, so that $\frac{\sum_{j=1}^n D_j}{\sqrt{n}} \Rightarrow \sigma N(0, 1)$. For the Martingale CLT, $\sigma^2 = lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n D_j^2 \xrightarrow{P} E_\pi[D_j^2] = Var_\pi[D_j]$ since the mean is 0.

For the left hand side of the equation, we can use a similar methodology to estimating equations. We divide the left hand side by $\frac{1}{\sqrt{n}}$, therefore by Taylor expansion we have $\frac{1}{n} \sum_{j=1}^n \frac{d^2}{d\theta^2}[logP(\theta, X_{j-1}, X_j)]|_{\theta=\xi_n} * (\hat{\theta} - \theta_0) * \sqrt{n}$. The ξ_n lies in between $\hat{\theta}$ and θ_0 and so by LLN the left hand side converges to $E_\pi[\frac{d^2}{d\theta^2} logP(\theta, X_{j-1}, X_j)] * sqrtn(\hat{\theta} - \theta_0)$. Therefore, we have our CLT

for our MLE $\hat{\theta}$, which is

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow -\frac{1}{E_\pi[\frac{d^2}{d\theta^2} log P(\theta, X_{j-1}, X_j)]} * \sigma N(0,1)$$

Note that we can complete the analysis even if we don't have every X_j. For example, if we had every 7th X_j, then we would have a likelihood function of $L_n(\theta) = \prod_{j=1}^{n} P^7(\theta, X_{j-1}, X_j)$. However, in this case we would need to compute $P^7(\theta, x, y)$, which can be computationally expensive. If our state space $|S|$ is small, it is computationally tractable, but if our state space is larger, we cannot do it practically.

23.2 Filtering for Markov Chains

Filter arises in settings where we want to do real time control on a system, but we do not get to observe the actual underlying state of the system being controlled. For example, if we return to our (s, S) inventory model and we do not get to directly observe the inventory level X_i, but rather only get to observe W_i (when the system is reordering inventory). Assuming that $s = 3$ and $S = 3$, then our $W_i = \mathbb{1}(X_i = 5)$, and we would observe a sequence of 0's and 1's. In the filtering problem, we want to find $\mu_n = P(X_n = y | W_0, ..., W_n)$.

We can think of w_i as a function of X_i. Therefore, $w_i = f(X_i)$ so $X_i \in f^{-1}(w_i)$, where X_i is the set of all states that map to w_i (not one-to-one). Therefore, we have

$\mu_n(y) = P(X_n = y | W_0 = w_0, ..., W_n = w_n)$
$\propto P(W_0 = w_0, ..., W_n = w_n, X_n = y)$ from Bayes Rule
$= P(X_0 \in f^{-1}(w_0), X_1 \in f^{-1}(w_1), ..., X_n \in f^{-1}(w_n), X_n = y)$, which equals 0 if $y \notin f^{-1}(w_n)$
$= \sum_x P(X_0 \in f^{-1}(w_0), X_1 \in f^{-1}(w_1), ..., X_{n-1} \in f^{-1}(w_{n-1}, X_{n-1} \in x) P(x, y) \mathbb{1}(y \in f^{-1}(w_n))$

We note that the first part is proportional to the definition of μ_{n-1}, so we have that

$= \sum_x \mu_{n-1}(x) P(x, y) \mathbb{1}(y \in f^{-1}(w_n))$

Normalizing the expression to get it to sum to one we have

$$\mu_n(y) = \frac{\sum_x \mu_{n-1}(x) P(x,y) \mathbb{1}(y \in f^{-1}(w_n))}{\sum_{x,z} \mu_{n-1}(x) P(x,z) \mathbb{1}(z \in f^{-1}(w_n))}$$

Therefore, we have a recursive filtering update equation for μ_n. Therefore, we can compute filters easily for Markov chains by performing the recursion. However, if the number of states of the Markov chain is very large, it can be cumbersome to compute this recursion.

All the previous modeling had been done assuming w_i was a function of only X_i. However, in many settings in practice, we want to modeling an iid noise, so we want $w_i = f(X_i) + Z_i$.

In one setting, we can have easy calculation of the filtering even if there is a large number of states, which is in the Gaussian state space models. In this model, we have $X_{n+1} = FX_n + Z_{n+1}$, where the Z_i's are iid with $N(0, \epsilon_1)$. However, we do not get to observe the X_i's directly, but rather we instead observe the W_i's, where $W_{n+1} = GX_n + Y_{n+1}$ where Y_i's are iid with $N(0, \epsilon_2)$. We want to find $P(X_n \in \cdot | w_0, w_1, ..., w_n)$. We know that this conditional distribution is also Gaussian and therefore is completely characterized by its conditional mean and conditional variance. Therefore, all that is needed is that we need an updating rule for the conditional mean and an updating rule for the conditional variance. This can be done recursively very efficiently using what is called a Kalman filter. It is similar to our Markov filter previously discussed by specialized to this form of conditional Gaussians.

24 Recurrence and Transience

We want to be able to extend our ideas from finite state Markov chains to countably infinite state Markov chains. Whereas in the irreducible Markov chain with finite states, we were able to always find an equilibrium, in the countably infinite state Markov chain we may not always be able to do so. The distinction of when we can and cannot comes from the idea of recurrence and transience.

A motivating example to illustrate this is with the Slotted Queueing example. In this case, as previously discussed, there is a queue for a service that can serve at most 1 customer per time period. Therefore, the number of customers in the queue can be written as $X_{n+1} = [X_n + Z_{n+1} - 1]^+$, where Z_n is the number of customers to arrive in the time n. Let us specify to the scenario where we can have at most 2 customers arrive per time period. Let us denote $P(Z_n = 0) = q, P(Z_n = 1) = p, P(Z_n = 2) = r$. This is a birth-death chain that has a countably infinite amount of states because our states are integer values from 0 to ∞. Intuitively, it is clear that if $p > q$, the system is unstable and has no equilibrium because more people are lining up in the queue than the system can serve and therefore the system is going to ∞. It is also intuitive that if $p < q$, the system is stable and has an equilibrium since the system is serving customers faster than they are lining up. However, what happens when $p = q$?

Using our finite state Markov chain formula, we have that in equilibrium, $\pi = \pi P$. Therefore, $\pi(x) = \pi(0) * \frac{p_0 p_1 \cdots p_{x-1}}{q_1 q_2 \cdots q_x} = (\frac{p}{q})^x \pi(0)$. We know that $\pi(x)$ is a probability distribution so it must sum to 1. Therefore, we have that $1 = \sum_x \pi(x) = \pi(0)(1 + \sum_{x=1}^{\infty} (\frac{p}{q})^x)$. Looking specifically at $\sum_{x=1}^{\infty} (\frac{p}{q})^x$, we can see that this will only be finite when $p < q$, and otherwise it will be infinite when $p \geq q$. Therefore, we only have equilibrium when $p < q$ and we do not have an equilibrium when $p \geq q$, even when $p = q$.

We can take a closer look at when $p = q$. We have that $X_{n+1} = [X_n + Z_{n+1} - 1]^+$, so when $X_n \geq 1$, we have that $X_{n+1} = X_n + Z_{n+1} - 1$. We can rewrite the right hand side as $\sum_{j=1}^{n+1} (Z_j - 1) + X_0$. When we have $p = q$, then $E[Z_j] = 1$, so we note that $(Z_j - 1)$ terms are mean zero random variables. Therefore, the sum of these mean zero random variables can be approximated by a normal distribution of $\sqrt{n}\sigma(Z_1)N(0, 1)$. Therefore, when $p = q$, $\frac{X_n}{\sqrt{n}} \Rightarrow \Gamma$ as $n \to \infty$, but when $p > q$, $\frac{X_n}{\sqrt{n}} \to E[Z_1] - 1$ as $n \to \infty$. When $p < q$, $X_n \Rightarrow X_\infty$. The $p = q$ case is an example of a "null recurrent" chain; the $p > q$ is an example of a "transient" chain; the $p < q$ is an example of a "positive recurrent" chain. To understand these definitions, we first need to understand regeneration.

24.1 Regeneration

If we take the previous example of the slotted queueing and we look at the X_i's over time, we can see a random walk that starts at $X_0 = 0$, then moves up for a while, then maybe at X_5 it comes back down to 0, and so on. What we notice is that at every point in the future where the $X_i = 0$, it is equivalent to the point at which we started at time 0, when $X_0 = 0$. Let us denote these times when $X_i = 0$ as $T_0, T_1, ...$ Therefore, we can split up our sequence of X_n into cycles, where each cycle are the observations between two values of T_i. For example, one cycle is between T_0 and T_1, another is between T_1 and T_2, and so on. We note that these cycles are iid cycles, and therefore we would like to generate some LLN or CLT for this iid structure. However, this depends on knowing if we have an infinite amount of these T_i's or not. Let us denote P_0 as the probability that X_n visits state 0 infinitely often. More generally, let us denote P_z as the probability that X_n visits state z infinitely often. Note that while we had denoted our T_i's as the times at which $X_i = 0$, we could perform the exact same analysis for a different point outside of 0, namely z. Denote $\tau(z) = inf\{n \geq 1 : X_n = z\}$, which is the time it takes for X_n to visit state z. Then, for us to be able to develop a LLN or CLT for this iid structure, we need to have that $P_z(\tau(z) < \infty) = 1$.

Our definition of recurrence if as follows. State z is a recurrent state if $P_z(\tau(z) < \infty) = 1$. Otherwise, we will say that state Z is a transient state. If X is irreducible, then recurrence/transience are class properties. What the means is that if one state z is recurrent, then all the states are recurrent. If one state z is transient, then all the states are transient.

Let us denote $N(z) = \sum_{i=1}^{n} \mathbb{1}(X_n = z)$, the total number of visits to state z. We have already established that $N(z) = \infty$ almost surely iff $P_z(\tau(z) < \infty) = 1$. As it turns out, $N(z) = \infty$ iff $E_z[N(z) = \infty]$. It is obvious that if $N(z) = \infty$ implies that $E_z[N(z) = \infty]$, but let us prove the converse is also true.

We can write $P_z(N(z) = k) = P_z(\tau(z) < \infty)^k P_z(\tau(z) = \infty) = P_z(\tau(z) < \infty)^k [1 - P_z(\tau(z) < \infty)^k]$. Therefore, $P_z(N(z) = k)$ is a geometric random variable and its expectation is $\frac{1}{p}$. Therefore, we have that $E_z[N(z)] = \frac{1}{P_z(\tau(z) = \infty)}$. We know that $P_z(\tau(z) = \infty) = 0$ if z is recurrent, and therefore $E_z[N(z)] = \infty$. If z is transient, then $P_z(\tau(z) = \infty)$ is a positive probability and $E_z[N(z)]$ equals a positive scalar.

If we take expectations of both sides, we get that $E[E_z[N(z)]] = \sum_{n=1}^{\infty} E_z[\mathbb{1}(X_n = z)]$. Therefore, $E_z[N(z)] = \sum_{n=1}^{\infty} P^n(z, z)$. If this sum is finite, then $E_z[N(z)]$ is finite. If it is infinite, then $E_z[N(z)]$ is infinite. Therefore, we have the following theorem.

For X irreducible, X is recurrent iff $\sum_{n=1}^{\infty} P^n(z,z) = \infty$. X is transient iff $\sum_{n=1}^{\infty} P^n(z,z) < \infty$.

Let us look at an example of how this is applied. Assume we have a setting where $X_{n+1} = X_n + W_{n+1}$, where W_i's are iid, so a random walk. W_i takes on the value 1 with probability $\frac{1}{2}$, and takes on the value -1 with probability $\frac{1}{2}$. As shown before, in a birth death chain when $p = q$ as in this case, we have no equilibrium but the chain is recurrent.

To prove this, we have to show that $\sum_{i=1}^{\infty} P^n(z,z) = \infty$. Let us specify to the case where $z = 0$. Since this chain is periodic with period = 2, we can look at only the even time periods, and $\sum_{i=1}^{\infty} P^n(0,0) = \sum_{i=1}^{\infty} P^{2n}(0,0)$. Of the $2n$ increments, it must be that n of them are of the kind 1 and n of them are of the kind -1. Therefore, we have a binomial distribution. We have that

$\sum_{i=1}^{\infty} P^{2n}(0,0) = \sum_{n=1}^{\infty} \binom{2n}{n} 2^{-n} * 2^{-n}$
$= \sum_{n=1}^{\infty} \frac{(2n)!}{n!n!} 2^{-n} * 2^{-n}$

By Stirling's approximation, $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$ as $n \to \infty$. Therefore, we have that the expression above $\sim \sum_{n=1}^{\infty} \frac{c}{\sqrt{n}}$, where c is some constant. This summation then equals ∞ as $n \to \infty$. Therefore, we have shown that this is a recurrent Markov chain.

Let's see what happens when we have the same problem, but in 2 dimensions. Denote $\tilde{X}_n = (X_n^1, X_n^2)$, where X_n^1, X_n^2 are independent one dimensional mean zero random walks on z. Let us again specify to the state 0 case, so $z = \vec{0} = (0,0)$. In this case, we have that $P_{\vec{0}}(\tilde{X}_{2n} = \vec{0}) = P_0(X_{2n}^1 = 0)P_0(X_{2n}^2 = 0)$. Note that we are again examining when we have $2n$ since the period = 2. From our calculations before, $P_0(X_{2n} = 0) \sim \frac{c}{\sqrt{n}}$, and therefore our expression simplifies to $\frac{c_1}{n} = \frac{c_2}{\sqrt{n}} \frac{c_2}{\sqrt{n}}$, where c_1, c_2 are some constants. Therefore, $\sum_{n=1}^{\infty} \frac{c}{n} = \infty$ and in this 2 dimensional setting, it is also recurrent.

More generally, if we have d dimensions, we can write $\tilde{X}_n = (X_n^1, X_n^2, ..., X_n^d)$, where each element in the vector is independent. Then we have that $P_{\vec{0}}(\tilde{X}_{2n} = \vec{0}) \sim (\frac{c}{\sqrt{n}})^d = c^d n^{-\frac{1}{2}}$. Therefore, $\sum_{n=1}^{\infty} P^n(z,z) < \infty$ for $d \geq 3$! This interesting result shows that for our nearest neighbor random walk, we have recurrence in 1 or 2 dimensions, but transience in 3 or more dimensions!

24.2 SLLN for Recurrence

We now know that if X is irreducible and recurrent, then $P_z(\tau(z) < \infty) = 1$, and that by regeneration, we have iid z-cycles. We want to examine the average amount of time that Markov chain spends in state y, in order words, we need to create a LLN for $\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}(X_j = y)$. Let's assume that we start at $X_0 = x$. When we are at time n, we may be partly through a z-cycle, so denote $l(n)$ the number of cycles that we have fully completed by time n. When n is large, we say that the average amount of time spent at state y is very similar at time n and at time $l(n)$. Therefore we can write

$\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}(X_j = y) \approx \frac{1}{T_{l(n)}} \sum_{j=0}^{T_{l(n)-1}} \mathbb{1}(X_j = y)$

Therefore, using a path-by-path argument, we have that $\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}(X_j = y) - \frac{1}{T_{l(n)}} \sum_{j=0}^{T_{l(n)-1}} \mathbb{1}(X_j = y) \overset{a.s.}{\to} 0$ as $n \to \infty$. Therefore, we can look at $\frac{1}{T_{l(n)}} \sum_{j=0}^{T_{l(n)-1}} \mathbb{1}(X_j = y)$ and what it converges to to obtain our answer. We have

$\frac{1}{T_{l(n)}} \sum_{j=0}^{T_{l(n)-1}} \mathbb{1}(X_j = y) = \frac{\sum_{i=1}^{l(n)} \sum_{j=T_{i-1}}^{T_i - 1} \mathbb{1}(X_j = y)}{\sum_{i=1}^{l(n)} \tau_i(z)}$
$= \frac{\frac{\sum_{i=1}^{l(n)} \sum_{j=T_{i-1}}^{T_i - 1} \mathbb{1}(X_j = y)}{l(n)}}{\frac{\sum_{i=1}^{l(n)} \tau_i(z)}{l(n)}}$

By the SLLN, this converges almost surely to $\frac{E_z[\sum_{j=0}^{\tau(z)-1} \mathbb{1}(X_j = y)]}{E_z[\tau(z)]}$. We know the the denominator can equal ∞, so we need to check to make sure the numerator does not equal ∞, or else convergence does not make sense. The numerator is $E_z[\sum_{j=0}^{\tau(z)-1} \mathbb{1}(X_j = y)]$, which is the number of visits to state y inside of one z-cycle. The probability of visiting it k times is $P_y(\tau(y) < \tau(z))^{k-1} P_y(\tau(z) < \tau(y))$, which we notice is a geometric random variable. Therefore, the expected value of this geometric random variable, $E_z[\sum_{j=0}^{\tau(z)-1} \mathbb{1}(X_j = y)]$ is always $< \infty$.

Therefore, we have that if X irreducible and recurrent then $\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}(X_j = y) \overset{a.s.}{\to} \frac{E_z[\sum_{j=0}^{\tau(z)-1} \mathbb{1}(X_j = y)]}{E_z[\tau(z)]}$. Note that if $E_z[\tau(z)]$ is infinite, then we converge to 0 almost surely. This is the scenario where our Markov chain is null recurrent. When $E_z[\tau(z)]$ is finite, then we converge to a positive probability, call it $\pi(y)$, which is what is called a positive recurrent Markov chain.

Note what happens when we take the expectation of both sides. We get that $E_x[\frac{1}{n} \sum_{j=0}^{n-1} \mathbb{1}(X_j = y)] \to \frac{E_z[\sum_{j=0}^{\tau(z)-1} \mathbb{1}(X_j = y)]}{E_z[\tau(z)]}$. The right hand side just remains the same. Let us denote the scalar value of the right hand side as $\pi(y)$. Then we have that

$E_x[\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{1}(X_j = y)] \to \pi(y)$
$\frac{1}{n}\sum_{j=0}^{n-1}E_x[\mathbb{1}(X_j = y)] \to \pi(y)$
$\frac{1}{n}\sum_{j=0}^{n-1}P_x[X_j = y] \to \pi(y)$
$\frac{1}{n}\sum_{j=0}^{n-1}P^j(x,y) \to \pi(y)$

Note that there is no z-dependence on the left hand side, and thus no dependence on the right hand side. Therefore what we converge to, $\pi(y)$ is independent of z.

What we want to do is show that this corresponds to $\pi = \pi P$, the equation in the finite Markov chain setting. We have

$\frac{1}{n}\sum_{j=0}^{n-1}P^j(x,y) \to \pi(y)$
$\frac{1}{n}\sum_{j=1}^{n}P^j(x,y) \to \pi(y)$ since $j = 0$ is asymptotically insignificant
$\frac{1}{n}\sum_{j=0}^{n-1}\sum_z P^{j-1}(x,z)P(z,y) \to \pi(y)$
$\sum_z \frac{1}{n}\sum_{j=0}^{n-1}P^{j-1}(x,z)P(z,y) \to \pi(y)$

Note that $\sum_{j=0}^{n-1}P^{j-1}(x,z) = \pi(z)$ so we have

$\sum_z \pi(z)P(z,y) = \pi(y)$
$\pi = \pi P$

Therefore, what we have concluded is that when X is irreducible and positive recurrent, then there exist a probability distribution π such that $\pi = \pi P$ and $\frac{1}{n}\sum_{j=0}^{n-1}\mathbb{1}(X_j = y) \overset{a.s.}{\to} \pi(y)$ as $n \to \infty$. More generally, $\frac{1}{n}\sum_{j=0}^{n-1}r(X_j) \overset{a.s.}{\to} \sum_z r(z)\pi(z)$. Conversely, if there exists a probability solution π to $\pi = \pi P$, then X is positive recurrent.

25 First Transition Analysis for Infinite State Space

Previously, we had looked at first transition analysis for a finite state space. We want to take a closer look at the subtleties behind first transition analysis for an infinite state space. Let us look at the first transition analysis for the hitting time T. We want to find $u^*(x) = E_x[T]$, where $T = inf\{n \geq 0 : X_n \in C^C\}$. Let's apply this to the slotted queueing example, where we denote $C^C = \{0\}$. Therefore the hitting time T is the first time that we have 0 customers. $C = \{1, 2, 3...\}$. Therefore, we have that for $i \geq 1$, $u(i) = 1 + pu(i+1) + qu(i-1)$ and $u(0) = 0$. What is different about this case versus the finite state space case is that we have only one boundary condition since there is no upper bound and therefore, we do not have a unique solution to this system of equations.

From the finite state space derivation, we had that $u = f + Gu$ where G is non-negative. We were able to solve $u^* = \sum_{n=0}^{\infty} G^n f$. If we have f as non-negative, which is often the case, then u^* is non-negative. We claim that u^* is the minimal non-negative solution of $u = f + Gu$. We can show this by expanding

$u = f + Gu$
$= f + G(f + Gu)$
$= f + Gf + G^2u$
$= f + Gf + ... + G^nf + G^{n+1}u$

Therefore, we have that if u is non-negative, then $G^{n+1}u$ is non-negative as well since G is a non-negative matrix. Therefore, $u \geq f + Gf + ... + G^nf$. For $n \to \infty$, $f + Gf + ... + G^nf = \sum_{n=0}^{\infty} G^n f = u^*$. Therefore, we have that $u \geq u^*$, and we have proved that u^* is the minimal non-negative solution. Therefore, we can identify a unique u^* that is probabilistically meaningful.

Another aspect that we were concerned about in the finite state space setting was when we could write $(I-G)^{-1} = \sum_{n=0}^{\infty} G^n f$, and that was namely when $\sum_{n=0}^{\infty} G^n f$ converged. We want to perform a similar analysis in the infinite state space scenario.

We define $h = (h(x) : x \in S)$ as a column vector and $G = (G(x, y) : x, y \in S)$. We define an h-norm as $||h||_w = sup_{x \in S} \frac{|h(x)|}{w(x)}$. If we choose $w(x) = 1$ for $x \in S$, then this is analogous to the infinity norm of h. We note that $w(x)$ need not be constant and can be a function of x. Therefore, we have an induced matrix norm $||A||_w = sup_{h \neq 0} \frac{||Ah||_w}{||h||_w} = sup_x \sum_y \frac{|A(x,y)|w(y)}{w(x)}$.

Therefore, if we can show that there exists w such that $||G||_w < 1$ and $||f||_w < \infty$, then $\sum_{n=0}^{\infty} ||G^n f||_w \leq \sum_{n=0}^{\infty} ||G||_w^n ||f||_w = (1 - ||G||_w)^{-1} ||f||_w < \infty$. Then, $\sum_{n=0}^{\infty} G^n f = (I - G)^{-1} f$ where $(I - G)^{-1}$ is the inverse operator to $(I - G)$ on $h_w = \{h : ||h||_w < \infty\}$.

26 Markov Chains on Continuous State Space

Let's assume that we have a Markov chain $X = (X_n : n \geq 0)$ with state state S. However, in this setting, S is no longer discrete, but rather is continuous and $S \subseteq \mathbb{R}^d$. Now, instead of a transition matrix, we have a transition kernel, where x goes to B instead of a single state y. We write $P_x[X_1 \in B] = P(x, B) = \int_B p(x, y)dy$, where $p(x, y)$ is called the transition density. Let $\mu(B) = P(X_0 \in B)$ denote the initial distribution. We then denote $P_\mu(\cdot) = \int_S \mu(dx)P_x(\cdot)$ as the probability that we start in state x and go to state $\cdot$. In the continuous case, we still enforce the Markov property, so $P(X_{n+1} \in B|X_0, ..., X_n) = P(X_n, B)$.

In the continuous setting, we want to find the analog to $\pi = \pi P$ in the discrete case. This analog in the continuous setting can be written as $\pi(y) = \int_S \pi(x)p(x, y)dx$ for all $y \in S$.

In the continuous setting, we can either be in a compact state space or a non-compact state space. If we are in a compact state space, then this is analogous to the finite state space in the discrete setting. If we are in a non-compact state space, then this is analogous to the countably infinite state space in the discrete setting.

In the compact state space setting, we have $S \subseteq \mathbb{R}^d$, where S is compact (meaning that it is closed and bounded). In this setting, we have that $p(x, y) > 0$ for every $x, y \in S$. Furthermore, we have that starting from any state, $p(\cdot, y)$ is continuous on the state space S. Therefore, since we have a continuous function on a compact set, then $inf_{x \in S} p(x, y) > 0$ for every $y \in S$. This means that the transition probabilities are all bounded below by some constant c. With these conditions, we have that

- There exists an equilibrium distribution π on S such that $\pi(y) = \int_S \pi(x)p(x, y)dx$ for every $y \in S$
- $P_x[X_n \in B] \to \int_B \pi(y)dy$ as $n \to \infty$
- $\frac{1}{n} \sum_{j=0}^{n-1} r(X_j) \overset{a.s.}{\to} \int_S r(y)\pi(y)dy$ for all non-negative functions r

As a note, computing $\pi(y) = \int_S \pi(x)p(x, y)dx$ cannot typically be computed in closed form, so it typically must be done numerically. Two common methods to solve this are to use Monte Carlo and to discretize the linear integral equation.

26.1 Non-Compact State Space

In the non-compact state space setting, we do not know if there is stability in the system at all (for example it could go to ∞). Therefore, we need to settle the question of stability. Intuitively for stability to occur, we would imagine that within our state space S there is a compact state K, and that the movement from any point in S would be going towards K and staying there. In other words, K should be a recurrent subset. Therefore, for us to have stability, one of the conditions that needs to hold is that $P_x(T_K < \infty) = 1$ for $x \in K^C$, where $T_K = \inf\{n \geq 1 : X_n \in K\}$. This means that if we start outside of K, then in finite time we end up in K with probability 1.

Once we get to K, we need to make sure that K acts like a recurrent subset. Therefore, another condition that needs to hold for us to have stability is that $P_x(X_1 \in B) \geq \lambda\varphi(B)$ for every $x \in K$, where $\lambda > 0$ and $\varphi(\cdot)$ is a probability on S. Why does this statement guarantee that K will behave nicely? We can write $P_x(X_1 \in B) = \lambda\varphi(B) + (1 - \lambda)Q(x, B)$. Therefore $Q(x, B) = P_x(X_1 \in B) - \frac{\lambda\varphi(B)}{1-\lambda}$, which show that $Q(x, B)$ is a transition kernel. Therefore, we can view P as a mixture that goes to $\varphi(\cdot)$ with a probability of λ and to $Q(x, \cdot)$ with probability $(1 - \lambda)$. This brings back the idea of regeneration. Essentially, we regenerate every time we get to $\varphi(\cdot)$. We will not regenerate with probability $(1 - \lambda)$ and so the probability of not regenerating after n time steps is $(1-\lambda)^n$. As $n \to \infty$, this goes to 0, and therefore we regenerate with probability 1.

How would be actually validate these conditions in practice? Let's start by looking at how to validate the second condition (which is called a minorization condition), that $P_x(X_1 \in B) \geq \lambda\varphi(B)$ for every $x \in K$. We can write this as $\int_B p(x, y)dy \geq \lambda \int_B \phi(y)dy$. Therefore, this requires that $p(x, y) \geq \lambda\phi(y)$ for every $x \in K$ and for every $y \in S$. Therefore, $\lambda\phi(y) \leq inf_{x \in K} p(x, y)$. Let's choose the value the regenerates the most often, so we choose the upper bound, where $\lambda\phi(y) = inf_{x \in K} p(x, y)$. If we integrate both sides, we have that $\lambda \int_S \phi(y)dy = \int_S inf_{x \in K} p(x, y)dy$. Since $\phi(y)$ is a probability distribution, then $\int_S \phi(y)dy = 1$, so we have that $\lambda = \int_S inf_{x \in K} p(x, y)dy$. Therefore $\phi(y) = \frac{inf_{x \in K} p(x,y)}{\lambda}$, and we require $\lambda > 0$. We have that $\lambda > 0$ when $p(x, y) > 0$, $p(\cdot, y)$ is continuous, and that K is compact. In that case, $inf_{x \in K} p(x, y) > 0$ and therefore $\lambda > 0$.

To prove the first condition, that $P_x(T_k < \infty) = 1$ for every $x \in K^C$, we can use the method of Lyapunov functions. A Lyapunov function has $g \geq 0$ and $E_x[g(X_1)] \leq g(x) - 1$ for every $x \in K^C$. What that means is that the expected value of the value at the next time step is always at least 1 smaller than the value at the current time step. We can think of $g(x)$ as the potential energy of being in state x, and at every time step, we lose energy until the point where we go into the compact set K. For example, if we start at 100, then within 100 time steps, we must get to the point where g starts to become negative, but $g \geq 0$, so therefore that must mean that g has reached the compact set K within the

first 100 time steps. In other words, $E_x[T_k] \leq g(x)$ for every $x \in K^C$. Therefore, if $g(x)$ is a finite value, then $P_x(T_K < \infty) = 1$.

A summary of what we have seen is as follows. If there exists a compact set $K \subseteq S$ and a Lyanpunov function $g : S \to \mathbb{R}_+$ such that

- $E_x[g(X_1)] \leq g(x) - 1$, for every $x \in K^C$
- $\int_S inf_{x \in K} p(x, y)dy > 0$

then $X = (X_n : n \geq 0)$ has infinitely many regenerations. If, in addition, we have

- $sup_{x \in K} E_x[g(X_1)] < \infty$

then X is positive recurrent. In that case, there exists a probability solution of $\pi(y) = \int_S \pi(x)p(x, y)dx$ and $\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) \overset{a.s.}{\to} \int_S r(y)\pi(y)dy$ as $n \to \infty$, provided $r(\cdot)$ is non-negative.

To show why this positive recurrent condition makes sense, we recall that $g(x) \geq E_x[T_K]$ for $x \in K^C$. Then using first transition analysis, we have

$E_x[T_K] = 1 + \int_{K^C} p(x, y)E_y[T_K]dy$
$\leq 1 + \int_{K^C} p(x, y)g(y)dy$
$\leq 1 + \int_S p(x, y)g(y)dy$
$= 1 + E_x[g(X_1)]$

Therefore, if $sup_{x \in K} E_x[g(X_1)] < \infty$, then we have an upper bound on $E_x[T_K]$, and thus it is positive recurrent.

One thing we note for Lyapunov functions is that our definition has that $E_x[g(X_1)] \leq g(x) - 1$ for every $x \in K^C$. One potential choice is to choose the case when they are equal. We can then rewrite $g(x) = E_x[g(X_1)] + 1$. Therefore, one choice is to use $E_x[T_K]$. Typically, we won't be able to compute this in closed form, but what this suggests is that one possible way of guessing Lyapunov functions is to use your model intuition to approximate this expected value and then plug the approximation into these Lyapunov inequalities and then hope that the approximation satisfies the inequalities and then you have your Lyapunov function.

26.1.1 Example of Checking for Equilibrium in Non-Compact State Space

The first example comes from queueing theory where we write $W_{n+1} = [W_n + Z_{n+1}]^+$, where W_{n+1} represents the waiting time for customer n. We assume the Z_i's are iid and that $E[Z_1] < 0$, meaning that customers are being served at a faster rate than they are lining up. The random variable Z also has a positive continuous density. Therefore, any interval, $K = [0, c]$ satisfies the second condition, that $\int_S inf_{x \in K} p(x, y)dy > 0$.

We see that when W_n is large, it is approximately equal to x, and taking expectation of both sides, we get that $E[W_{n+1}] = x + E[Z_1]$. Therefore, we see that the system is decreasing linearly in x, and so we guess a Lyapunov function $g(x) = ax$. Therefore, we have

$E_x[g(X_1)] - g(x) = aE_x[X_1] - ax$
$= aE[x + Z_1]^+ - ax$

We see that $[x + Z_1]^+ - x \to Z_1$ as $x \to \infty$, so therefore $E[x + Z_1]^+ - x \to E[Z_1]$ and scaling by a, we get that $aE[x + Z_1]^+ - ax \to aE[Z_1]$. We choose a so that $aE[Z_1] = -2$. There exists an $x_0 < \infty$ such that $E[x + Z_1]^+ - ax \leq -1$ for every $x \geq x_0$. We choose $K = [0, x_0]$. Therefore, we have satisfied the first condition, that $E_x[g(X_1)] \leq g(x) - 1$, for every $x \in K^C$.

We then have that $sup_{x \in K} E_x[g(X_1)] = asup_{x \in [0,x_0]} E[x + Z_1]^+ \leq a[x_0 + E[Z_1]] < \infty$. Therefore, we have satisfied the third condition, that $sup_{x \in K} E_x[g(X_1)] < \infty$. Therefore, we have a positive recurrent Markov chain!

The second example is an autoregressive process of order 1. We have $X_{n+1} = \rho X_n + Z_{n+1}$ where the Z_i's are iid random variables with a continuous positive density. Therefore, any compact subset $K \subseteq \mathbb{R}$ will satisfy condition 2. We can see this when we write out $p(x, y) = f_Z(y - \rho x)$. We assume that $E[log(1 + |Z_1|)] < \infty$, and we claim that under this assumption, that we have a positive recurrent chain. When X is large, we have that $X_{n+1} \approx \rho X_n$ and we estimate that it will take about $\rho^n x$ jumps to get to the compact set since we jump by a factor of ρ each time. Therefore, if our compact set is $K = [-c, c]$, then we have that $\rho^n x \leq |c|$.

We rewrite this as $nlog(p) + log(x) \leq log(|c|)$. Solving this, we get that $n \approx log(x)$. Therefore, we use a Lyapunov function of $g(x) = alog(1 + |x|)$. The reason we add 1 to the value inside the log is because the log function is not positive until after 1 and we need $g \geq 0$. Therefore we have that

$E_x[g(X_1)] - g(x) = aE[1 + log(1 + |\rho x + Z_1|)] - alog(1 + |x|)$
$= Elog(\frac{1+|\rho x+Z_1|}{1+|x|}) \to log(|\rho|)$ as $n \to \infty$, since $|\rho| < 1$

Therefore, $log(|\rho|)$ is negative and it can be easily proved from here that it satisfies condition 1 and 3. Therefore we have a positive recurrent Markov chain.

27 Markov Chain Monte Carlo

Markov Chain Monte Carlo is a method that can be used in Bayesian calculations to find the posterior distribution even when the calculation is analytically untractable. More generally, our goal is to generate a random variable Z taht follows a density/pmf of the form $f_z(\cdot) = \frac{h(\cdot)}{\int_S h(y)dy}$. While the numerator is known in closed form, the denominator, the normalization constant, is hard to get analytically.

In Bayesian statistics, where have have the $f(\theta|data) \propto L(data|\theta)p(\theta) = \frac{L(data|\theta)p(\theta)}{\int_S L(data|\theta')p(\theta')d\theta'}$, again, the denominator is often very hard to calculate. Thus we turn to Markov Chain Monte Carlo. It is an algorithm that gets set up in such a way that we can sample from the posterior without knowing this normalization constant. We will run a Markov chain that will allow us to sample from this posterior. More precisely, we will sample from a Markov chain with an equilibrium density $\pi(y) = \frac{h(y)}{\int_S h(z)dz}$.

From LLN, we know that the sample $X_0, X_1, X_2, ..., X_n$ from this Markov chain will converge: $\frac{1}{n}\sum_{j=0}^{n-1} \mathbb{1}(X_j \in B) \overset{a.s.}{\to} \int_B \pi(y)dy$. In practice, we would run the Markov chain until a large number k where there is equilibrium, and then we sample at $X_k, X_{k+m}, X_{k+2m}..., X_{k+nm}$, where m is chosen so that the time in between these samples is long enough to lose memory. That way, we basically have independence between the samples. More research is currently being done on how to reliably choose a good k and m.

To get the $pi(y)$, we need to construct a transition kernel P such that $pi = piP$. If we have d dimensions, we have d equations and $d(d-1)$ unknowns, so we have far more unknowns than equations. Therefore, we need to find a way to add more equations.

27.1 Balance

We can write our equilibrium formula $\pi = \pi P$ as $\pi(y) = \int_S \pi(x)P(x,y)dx$. The left hand side is the rate at which we transition out of y in equilibrium, and by the definition of equilibrium, the right hand side must be the rate at which we transition into y in equilibrium. Thus, we have "global balance."

We have "detailed balance" when the rate at which we transition in and out of any two states x, y is equal. Mathematically, this is seen as $\pi(x)p(x,y) = \pi(y)p(y,x)$ for all $x, y \in S$. Detailed balance implies global balance, but global balance does not imply detailed balance. An example of this is if we have a Markov chain where $p(x,y) > 0$ and $p(y,x) = 0$.

We claim that birth-death chains satisfy detailed balance. In a birth-death chain, we have states $S = \{0, 1, 2, ...\}$. Recall the equilibrium equation is $\pi(x) = \pi(0)\frac{p_0 p_1 \cdots p_{x-1}}{q_1 q_2 \cdots q_x} = \pi(x-1)\frac{p_{x-1}}{q_x}$. Therefore, we can write this as $q_x\pi(x) = \pi(x-1)p_{x-1}$. But q_x is just the transition probability from x to $x-1$, or in other words, $P(x, x-1)$. p_{x-1} is the transition probability from $x-1$ to x, or in other words $P(x-1, x)$. Therefore, we have $pi(x)P(x,y) = \pi(y)P(y,x)$ for every x, y and we have detailed balance.

27.2 Time-Reversed Markov Chains

One important characterization of Markov chains is that if we look at the past, conditional on the current state, and the future, conditional on the current state, that the past and the future are conditionally independent. That suggests that if we reverse the role of the past and the future, the independence will be preserved.

Let us assume that we have a Markov chain with an equilibrium distribution pi. We also assume that $X_0 \overset{D}{=} \pi$, therefore it is a stationary process. However, we can start the process at $X_{-j} \overset{D}{=} \pi$ and let $j \to \infty$. Therefore, we have a two-sided stationary process $(X_n : -\infty < n < \infty)$. Let $Y_n = X_{-n}$. This is what we call the "time-reversed Markov chain." Let us denote $Q(x,y)$ the transition matrix that goes with the Y-chain.

$$
\begin{aligned}
&Q(x,y) = P(Y_{n+1} = y | Y_n = x) \\
&= \frac{P(Y_n = x, Y_{n+1} = y)}{P(Y_n = x)} \\
&= \frac{P(X_{-n-1} = y, X_{-n} = x)}{P(X_{-n} = x)} \\
&= P(X_{-n} = x | X_{-n-1} = y)\frac{P(X_{-n-1} = y)}{P(X_{-n} = x)} \\
&= P(y,x)\frac{\pi(y)}{\pi(x)}
\end{aligned}
$$

We want to figure out when $Q = P$, which would mean that $X \overset{D}{=} Y$, and that we have a time-reversible Markov chain. If $Q = P$, then $Q(x,y) = P(x,y)$ so we write $P(x,y) = \frac{\pi(y)}{\pi(x)}P(y,x)$, which can be rewritten as $\pi(x)P(x,y) = \pi(y)P(y,x)$ for all $x, y \in S$, which is exactly detailed balance!

Therefore, detailed balance holds iff X is time-reversible. In matrix form, this means that our transition matrix is symmetric, and specifically when $\pi(x)$ is uniform, this means that $P(x,y) = P(y,x)$. The benefits of a symmetric matrix are that they have real eigenvalues and are diagonalizable, therefore we have a nice representation of P^n.

27.3 Metropolis Algorithm

Suppose we have a transition matrix Q that is irreducible and aperiodic, so all entries are positive. Then we define a density P, such that for $x \neq y$, we have $p(x,y) = q(x,y)min(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1)$. By multiplying $\pi(x)$ to both sides of the equation, we can see that $\pi(x)p(x,y) = \pi(y)q(y,x)$. Therefore, we have detailed balance holding for all $x \neq y$. Detailed balance always holds when $x = y$, so we have detailed balance for all possible pairs of states.

Notice that $\pi(y) = \frac{h(y)}{\int_S h(z)dz}$ and that $\pi(x) = \frac{h(x)}{\int_S h(z)dz}$, so that when we divide them, the normalization constant cancels. $\frac{\pi(y)}{\pi(x)} = \frac{h(y)}{h(x)}$. Therefore, we can simplify our expression to be $p(x,y) = q(x,y)min(\frac{h(y)q(y,x)}{h(x)q(x,y)}, 1)$, which now has no normalization constant!

We create a q that is easily simulatable. We propose a move from x to y using the transition density of q. We then accept the move to state y with probability $min(\frac{h(y)q(y,x)}{h(x)q(x,y)}, 1)$. If we reject the move, we stay in state x for 1 additional unit of time.

28 Markov Jump Process

A Markov jump process is a Markov process in continuous time, but with a discrete state space. Since it is a continuous time Markov process, we write $P(X(t+s) \in B | X(u) : 0 \leq u \leq t) = P(X(t+s) \in B | X(t))$ for all $s, t \geq 0$ and for all $B \subseteq S$. Since it is in a discrete space, we write $P(X(t+s) = y | X(u) : 0 \leq u \leq t) = P(t, t+s, x, y)$. To simplify assumptions, we specifically look at the case where we have stationary transition probabilities. In this scenario, the transition probabilities depend only on the length of time, and we can therefore write $P(s) = (P(s, x, y) : x, y \in S)$.

In the stationary discrete time setting, we can write $P_n = P^n$, and more generally, $P_{n+m} = P_n * P_m$. Analogously, in the continuous time setting that we are currently looking at, we can write $P(t+s, x, y) = \sum_{z \in S} P(t, x, z) P(s, z, y) = (P(t)P(s))(x, y)$. Therefore, $P(t+s) = P(t)P(s)$. Now, if we were looking at scalars, rather than Markov processes, we would notice that $p(t+s) = p(t)p(s)$ is an exponential distribution. In this exponential distribution, we would have $p(t) = e^{ct}$, and we could solve for $c = p'(0)$. What do we see when we take the derivative in this Markov process setting?

We define matrix $Q = P'(0)$, where the derivative means the derivatives entry-by-entry. Then, $Q = P'(0) = lim_{n \to \infty} \frac{P(h) - P(0)}{h}$. We notice that $P(0) = P(X(0) = y | X(0) = x) = I$, so we have $Q = P'(0) = lim_{n \to \infty} \frac{P(h) - I}{h}$. Let us look at the value of Q for diagonal entries and non-diagonal entries separately. For non-diagonal entries, $Q(x, y) = lim_{h \to 0} \frac{P(h, x, y)}{h}$. We can rewrite as $P(h, x, y) = Q(x, h)h + o(h)$. Intuitively, $Q(x, y)$ is the rate at which X jumps from state x to state y. Now looking at the diagonal entries, we have that $Q(x, x) = lim_{h \to 0} \frac{P(h, x, x) - 1}{h} = -lim_{h \to 0} \frac{1 - P(h, x, x)}{h} = -lim_{h \to 0} \frac{\sum_{y \neq x} P(h, x, y)}{h} = -\sum_{y \neq x} Q(x, y)$. Intuitively, this value is the rate at which X jumps out of state x.

We notice that $Qe = 0$. This implies that Q is always a singular matrix. Q is called the "rate matrix" of X. By combining our equations for the diagonal and non-diagonal entires, we can write $P(h) = P(0) + Qh + o(h) = I + Qh = o(h)$. If we multiply a column vector g to the right, we get $P(h)g = g + Qgh + o(h)$. If we look at the x entry of the product, we see that $((Ph)g)(x) = \sum_y g(y) P_x(X(h) = y) = E_x[g(X(h)]$. Therefore, we have that $E_x[g(X(h)] = g(x) + h(Qg)(x) + o(h)$. This is the "short-time" expansion for how the expected values evolve.

Returning back to the property that $P(t+s) = P(t)P(s)$, similar to the exponential distribution, we have that $P(t) = exp(Qt)$. Since these are matrices, we can use the exponential expansion where $exp(Qt) = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!} = \sum_{n=0}^{\infty} \frac{Q^n t^n}{n!}$. This expansion will always work in the finite discrete space case, and will work mostly in the countably infinite space case, except for a few scenarios when the sum is "explosive."

Returning back to the property that $P(t+s) = P(t)P(s)$, we can take the derivative of both sides with respect to t, so $\frac{d}{dt} P(t+s) = \frac{d}{dt} P(t)P(s)$. Rewriting this as $P'(t+s) = P'(t)P(s)$, setting $t = 0$, we get a system of equations $P'(s) = QP(s)$ subject to $P(0) = I$. We can solve this system using a time-stepping method. The advantage of this method as opposed to just expanding the exponential is that we can now compute the whole range of values, not just at one value of t. Similarly, if we take the derivative of both sides with respect to s instead of t, we get that $P'(t+s) = P(t)P'(s)$. Setting $s = 0$, we have the system of equations that $P'(t) = P(t)Q$ subject to $P(0) = I$.

If we right multiply by a column vector g, we have $P'(s)g = QP(s)g$. Let us denote the column vector $P(s)g$ as $u(s)$. Therefore, our system of equations is now $u'(s) = Qu(s)$ subject to $u(0) = g$. These are the backwards differential equations and $u(s, x) = (P(s)g)(x) = E_x[g(X(s))]$. Intuitively, this is a vector of the expectations. If we instead left multiply by a vector μ, we have $\mu P'(t) = P(t)Q$. Let us denote the row vector $\mu P(t)$ as $\mu(t)$. Therefore, our system of equations is now $\mu'(t) = \mu(t)Q$ subject to $\mu(0) = \mu$. These are the forwards differential equations and $\mu(t, y) = (\mu P(t))(y) = \sum_x \mu(x) P_x(X(t) = y)$. Intuitively, this row vector represents the probability of going from state x to state y.

Let us now look at the backwards and forwards equations in the context of non-stationary probabilities. In this scenario, $P(t, t+s+u) = P(t, t+s)P(t+s, t+s+u)$. We now have $u(s, x) = E[g(X(t)) | X(t-s) = x]$ and $u(0, x) = g(x)$. Therefore, our backwards equations are now $u'(s) = Q(s)u(s)$ subject to $u(0) = s$, where $Q(s) = lim_{h \to 0} \frac{P(s, s+h) - I}{h}$. We now also have $P(X(t) = y) = \mu(t, y)$ and $P(X(0) = x = \mu(x))$, so now our forwards equations are $\mu'(t) = \mu(t)Q(t)$ subject to $\mu(0) = \mu$.

28.1 Path Structure

Let us take a closer look at the system of equations $P'(t) = QP(t)$ subject to $P(0) = I$. Let us denote the jump function from state x as $\lambda(x) = \sum_{y \neq x} Q(x, y)$. Therefore, we have

$P'(t, x, y) = \sum_z Q(x, z) P(t, z, y)$
$= -\lambda(x) P(t, x, y) + \sum_{z \neq x} Q(x, z) P(t, z, y)$

Moving the term to the left side, we have

$P'(t,x,y) + \lambda(x)P(t,x,y) = \sum_{z \neq x} Q(x,z)P(t,z,y)$
$e^{\lambda(x)t}P'(t,x,y) + \lambda(x)P(t,x,y) = e^{\lambda(x)t}\sum_{z \neq x} Q(x,z)P(t,z,y)$
$\frac{d}{dt}[e^{\lambda(x)}P(t,x,y)] = e^{\lambda(x)t}\sum_{z \neq x} Q(x,z)P(t,z,y)$
$e^{\lambda(x)t}P(t,x,y) - \delta_{xy} = \sum_{z \neq x}\int_0^t Q(x,z)e^{\lambda(x)s}P(t,z,y)ds$
$P(t,x,y) = e^{\lambda(x)t}\delta_{xy} + \sum_{z \neq x} Q(x,z)\int_0^t e^{\lambda(x)(s-t)}P(t,z,y)ds$

Let $r = t - s$. Then we have

$P(t,x,y) = e^{\lambda(x)t}\delta_{xy} + \sum_{z \neq x} Q(x,z)\int_0^t e^{-\lambda(x)r}P(t-r,z,y)dr$

Let $R(x,z) = \frac{Q(x,z)}{\lambda(x)}$. Note that $\sum_{y \neq x} R(x,y) = 1$.

$P(t,x,y) = e^{\lambda(x)t}\delta_{xy} + \sum_{z \neq x} R(x,z)\int_0^t \lambda(x)e^{-\lambda(x)r}P(t-r,z,y)dr$

In the case where $x \neq y$, then $\delta_{xy} = 0$. We have

$P(t,x,y) = \int_0^t \lambda(x)e^{-\lambda(x)r}\sum_{z \neq x} R(x,z)P(t-r,z,y)dr$
$P(t,x,y) = \int_0^t P(exp(\lambda(x) \in dr))\sum_{z \neq x} R(x,z)P(t-r,z,y)dr$

Therefore, from this derivation, we can see that the holding time, or the time that the system stays in the same state is exponential with parameter $\lambda(x)$. The choice of the next state is given by $R(x,z)$.

If we wanted to simulate X, this is how we could do it. We could simulate the amount of time spent in a state as an exponential distribution with parameter $\lambda(x)$. The sequence of states visited is determined by a discrete time Markov chain $Y = (Y_n : n \geq 0)$ having transition matrix R. We could then run Monte Carlo in order to obtain results from our simulation.

Given the exponential holding time, one thing to note is that if we wanted to find the probability that starting at state x, we jump out of within a time period h, we can calculate $\int_0^h \lambda(x)e^{-\lambda(x)s}ds$, which equals $\lambda(x)h + o(h)$. If we wanted to find the probability that starting at state x, we perform 2 or more jumps in within a time period h, it is $O(h^2)$, which is quite unlikely. One other thing to note is that because we are now in a continuous time setting rather than a discrete time setting, we can spend different amounts of time for each time period. As a result, it is possible to jump infinitely often in a finite amount of time, which is what is called an "explosion." This is a very unlikely scenario, and practically people do not normally model this explosion scenario, but rather assume that the process is non-explosive.

28.2 Examples

We start with a simple example of customers arriving in a queue, such as the security line at an airport. We have a discrete state space of $\{0,1,2...\}$, where $N(t)$ is the number of arrivals in the time $[0,1]$. We jump from one state to another exponentially at a rate of λ. Therefore, we can write our rate matrix as $Q(i,i+1) = \lambda$ and $Q(i,i) = -\lambda$, for $i \geq 0$. This means that our P matrix is $P(t,i,i+j) = e^{-\lambda t}\frac{(\lambda t)^j}{j!}$, which is a Poisson random variable with mean λt. This is called a Poisson process.

Another example is a single server queueing model, where we assume we jump forwards exponentially with parameter λ, but backwards with parameter μ. This is an M/M/1 queue. Therefore, we can write our rate matrix as $Q(i,i+1) = \lambda$ and $Q(i,i-1) = \mu$ and $Q(i,i) = -\lambda - \mu$ for $i \geq 1$. We also have $Q(0,0) = -\lambda$ and $Q(x,y) = 0$.

If we take the single serve queueing model and increase it to two servers, what we get the same jumping forwards, but we jump backwards with parameter 2μ for all jumps except from 1 to 0, which is with parameter μ. Therefore, our rate matrix is $Q(i,i+1) = \lambda$ and $Q(i,i-1) = 2\mu$ and $Q(i,i) = -\lambda - 2\mu$ for $igeq2$. We also have $Q(1,0) = \mu$, $Q(x,y) \neq 0$, $Q(1,1) = -\lambda - \mu$, and $Q(0,0) = -\lambda$.

For an infinite server model, such as with a call center where there are so many customer service agents that it is approximately infinite, we again jump forwards exponentially with parameter λ, but jump backwards from $Q(i,i-1) = i\mu$.

In a network model where we have d stations, $x = (x_1, x_2, ..., x_d)$ where x_i is the number of customers at station i. Our rate matrix is $Q(x, x+e_i) = \lambda_i$, where e_i is the 0 vector with the value 1 in the ith component. We also have $Q(x, x-e_i, x+e_j) = \mu_i W_{ij}$, for $x_i \geq 1$, where μ_i is the service rate at station i and W_{ij} is the probability that a customer leaving i joins the queue at

station j. We also have $Q(x, x - e_i) = \mu_i(1 - \sum_{j=1}^{d} W_{ij})$ for $x_i \geq 1$. Lastly, we have $Q(x, z) = 0$ where $x \neq z$.

One more example is a model that is popular is epidemiology, the SIR (Susceptible, Infected, Recovery) model such as the one used for COVID. If we let x be the susceptibles, y be the infecteds, and z be the recovered, we have $Q(x, y, z), (x-1, y+1, z)) = \lambda xy$, for $x \geq 1$, which represents the rate at which the individuals become infected. We also have $Q((x, y, z), (x, y-1, z+1)) = \mu y$, for $y \geq 1$, which is the rate at which the infected individuals recover. All other transition probabilities are 0.

28.3 First Transition Analysis

For the following first transition analysis calculations, we will assume stationary transition probabilities. Recall from our backwards equation that $E_x[r(X(h))] = r(x) + h(Qr)(x) + o(h)$. Therefore, assume we want to find the expected value of the hitting time T. We want $u^*(x) = E_x[T]$, where $T = inf\{t \geq 0 : X(t) \in C^C\}$. We also have the boundary condition that $u^*(x) = 0$ if $x \in C^C$. For $x \in C$, in the discrete case we were able to step forward to time step 1, but in this continuous time case, we will step forward by h, and then send h to 0.

We have that $u(x) = E_x[min(T, h) + \mathbb{1}(T > h)u^*(X(h))]$

$= \int_0^h s\lambda(x)e^{-\lambda(x)s}ds \sum_{y \in C^C} R(x, y) + hP_x(T > h) + E_x[\mathbb{1}(T > h)u^*(X(h))]$
$= \int_0^h s\lambda(x)e^{-\lambda(x)s}ds \sum_{y \in C^C} R(x, y) + h[1 - \int_0^h \lambda(x)e^{-\lambda(x)s}ds] + E_x[\mathbb{1}(T > h)u^*(X(h))]$
$= \int_0^h s\lambda(x)e^{-\lambda(x)s}ds \sum_{y \in C^C} R(x, y) + [h + o(h)] + E_x[\mathbb{1}(T > h)u^*(X(h))]$
$= o(h) * \sum_{y \in C^C} R(x, y) + [h + o(h)] + E_x[\mathbb{1}(T > h)u^*(X(h))]$
$= o(h) + [h + o(h)] + E_x[u^*(X(h)) + o(h)]$
$= h + o(h) + E_x[u^*(X(h))]$
$= h + o(h) + [u^*(x) + (Qu^*)(x)h + o(h)]$
$= h + o(h) + u^*(x) + (Qu^*)(x)h$

Therefore, we have $u^*(x) = h + o(h) + u^*(x) + (Qu^*)(x)h$. Simplifying, we have $-1 = (Qu^*)(x)$, for $x \in C$, subject to $u^*(x) = 0$ on C^C. Note that this is completely analogous to the discrete case where we had $(P - I)u^* = -e$ for $x \in C$.

Let's look at the first transition analysis for when we want to calculate the expected investment return discounted by the time value of money. We want to find $u^*(x) = E_x[\int_0^\infty e^{-\alpha t} r(X(t))]dt$. Therefore, we have

$u(x) = E_x[\int_o^h e^{-\alpha t} r(X(s))ds + e^{-\alpha h}u^*(X(h))]$
$= hr(x) + o(h) + (1 - \alpha h + o(h)) * E_x[u^*(X(h))]$
$= hr(x) + o(h) + (1 - \alpha h + o(h)) * [u^*(x) + h(Qu^*)(x) + o(h)]$
$= hr(x) + u^*(x) - \alpha u^*(x)h + h(Qu^*)(x) + o(h)$

Subtracting $u^*(x)$ from both sides and dividing by h, we get that $0 = r(x) - \alpha u^*(x) + (Qu^*)(x)$, for $x \in S$.

One more example of first transition analysis is in stochastic control where we develop the Bellman's equation in a continuous time setting. Assume we want to maximize $v^*(x) = E_{(A(t):t \geq 0)}[\int_0^\infty e^{-\alpha x} r(X(s), A(s))ds | X(0) = x]$. To maximize this, we want to maximize at every time step h, so we have

$v(x) = max_{(A(t):t \geq 0)}[\int_0^h e^{-\alpha s} r(X(s), A(s))ds + e^{-\alpha h}v(X(h))]$
$= max_{a \in a(x)}[r(x, a)h + (1 - \alpha h + o(h))(v(x) + \sum_y Q(x, y)v(y)h) + o(h)]$

Subtracting $v(x)$ from both sides and dividing by h, we get that our Bellman's equation: $0 = max_{a \in a(x)}[r(x, a) + \sum_y Q_a(x, y)v(y) - \alpha v(x)]$.

28.4 Equilibrium Theory

Assume we have stationary transition probabilities and that our system is non-explosive. Looking back at our forward equations, we have $\mu(t, y) = P(X(t) = y) \to \pi(y)$ as $t \to \infty$. If we take the derivative of both sides with respect to t, we get $\frac{d}{dt}\mu(t, y) \to 0$ as $t \to \infty$. Therefore, we have $\mu'(t) = \mu(t)Q$, where $\mu'(t) \to 0$ and $\mu(t) \to \pi$. Therefore, our equilibrium equation is $0 = \pi Q$.

We have the following theorem. If X is irreducible and there exists a probability solution π of $\pi Q = 0$, then

- $\frac{1}{t}\int_0^t r(X(s))ds \overset{a.s.}{\to} \sum_y \pi(y)r(y)$ as $t \to \infty$

- $P_x(X(t) = y) \to \pi(y)$ as $t \to \infty$. Note that in the discrete case, we need aperiodicity for this to hold, but in our continuous case, we cannot have aperiodicity so this always holds.

Let us look at the equilibrium theory in the context of the server queues. First let us assume we are in the scenario with the single server queue. Recall that the forward jump in this case occurs exponentially with parameter λ and the backward jump with parameter μ. Therefore, since this is a birth-death chain, we have $\pi(x) = \pi(0)\frac{\lambda_0\lambda_1...\lambda_{x-1}}{\mu_1\mu_2...\mu_x} = \pi(0)(\frac{\lambda}{\mu})^x$. If we denote $\rho = \frac{\lambda}{\mu}$, then we have $\pi(x) = \pi(0)\rho^x$. If $\rho < 1$, we have an equilibrium where $\pi(x) = (1-\rho)\rho^x$ for $x \geq 0$. We can see that this distribution is geometric/exponential. If $\rho \geq 1$, the there is no equilibrium because the system is unstable. Intuitively, this occurs when $\lambda \geq \mu$, so the rate of jumping forward is faster than jumping backward and so there is no stability.

Let us now look at the infinite server case. In this case, the forward jump is always with parameter λ, just as in the single server case, but the backward jump is $\mu_i = i\mu$. Therefore, in this case, we have $\pi(x) = \pi(0)\frac{\lambda^x}{x!\mu^x}$. To have equilibrium, we choose $\pi(0) = e^{-\frac{\lambda}{\mu}}$, so then we have $\pi(x) = e^{-\frac{\lambda}{\mu}}(\frac{\lambda}{\mu})^x\frac{1}{x!}$, which is a Poisson distribution. Therefore, we can see that pi has a Poisson distribution with parameter $\frac{\lambda}{\mu}$. In the case where λ is large, which is often the case, a Poisson with a large mean is approximately Normal, so we can model this distribution as a $N(\frac{\lambda}{\mu}, \frac{\lambda}{\mu})$.

Notice that as λ increases relative to μ, then the standard deviation will shrink relative to the mean. For example, if $\lambda = 10000$ and $\mu = 1$, then we will have approximately a Normal with mean of 10000 and standard deviation of 100. Therefore, we won't fluctuate much from the mean.

One more example is when we look at a network model. Assume we have d stations and a vector state space now where $x = (x_1, x_2, ..., x_d)$. Therefore the equilibrium distribution is $\pi(x) = \pi(x_1, x_2, ..., x_d)$. Surprisingly, we can factor these out as $\pi(x_1, x_2, ..., x_d) = \pi_1(x_1)\pi_2(x_2)...\pi_d(x_d)$, which means that the number of customers at each station behave as if they are independent, which is surprising!

29 Stochastic Differential Equations/Diffusions

In this section, we are looking at Markov processes with continuous time as well as with continuous state space. These are known as Markov diffusion processes. We recall that in a stochastic process, we can write $X_{n+1} = f(X_n, Z_{n+1})$, where the Z_i's are iid. Rearranging the terms, we can write $X_{n+1} - X_n = \mu(X_n, Z_n)$. If we shrink the time between n and $n+1$ down to 0, we get that $\frac{dX}{dt} = \mu(X(t), \xi(t))$. $\xi(t)$, which is commonly referred to as white noise, has the independence property that $\xi(t)$ is independent of $\xi(s)$ for $s \neq t$, and that $\xi(t) \stackrel{D}{=} \xi(0)$. We realize that this mathematical object is very irregular and impossible to work with, because every infinitely small time period is independent, and so the values can jump all over the place. Therefore, we want to smooth it out using integration.

In a discretized setting, if we had $Z_1, Z_2, ...$ we could produce a sum $S_n = Z_1 + ... + Z_n$, a random walk. This random walk has two properties. The first is that $S_{n+m} = S_m + (S_{n+m} - S_m)$, where S_m is independent of $(S_{n+m} - S_m)$. This is called independent increments. The second is that $S_{n+m} - S_m \stackrel{D}{=} S_n - S_0$. This is called stationary increments. We want to have the analogous results in the continuous setting. Let us define $B = (B(t) : t \geq 0)$. Then the independent increments property says that for $t_1 < t_2 < ... < t_m$, that $B(t_1) - B(t_0)$ is independent of $B(t_2) - B(t_1)$, which is independent of $B(t_3) - B(t_2)$, etc. The stationary increments property states that $B(t+s) - B(s) = B(t) - B(0)$. One more property that we will enforce in this continuous setting is that $B(t) \stackrel{D}{=} N(0, t)$, that the mean is 0 and the variance increases linearly in t. These 3 conditions constitute the process known as standard Brownian motion, which is the integrated white noise. Note that there is no derivative of Brownian motion, since $\frac{B(t+h)-B(t)}{h} \stackrel{D}{=} \frac{N(0,h)}{h} \stackrel{D}{=} \frac{\sqrt{h}N(0,1)}{h}$, which has no limit as $h \to 0$. Therefore, Brownian motion is non-differentiable. However, Brownian motion has continuous sample paths! $B(\cdot) \in C[0, \infty)$.

Recall that our equation $\frac{dX}{dt} = \mu(X(t), \xi(t))$ was mathematically impossible to work with because we could not work with the white noise directly. Therefore, we rewrite as $\frac{dX}{dt} = \mu(X(t)) + \sigma(X(t))\xi(t)$. We then integrate both sides and get $X(t) - X(0) = \int_0^t \mu(X(s))ds + \int_0^t \sigma(X(t))\xi(s)ds$. Exchanging $\xi(s)$ for $dB(s)$, since we now know that Brownian motion is the integral of white noise, we have our equation

$X(t) - X(0) = \int_0^t \mu(X(s))ds + \int_0^t \sigma(X(t))dB(s)ds$, where $\mu(\cdot)$ is the "drift" and $\sigma(\cdot)$ is the "volatility."

If we assume linearity in both drift and volatility, where $\mu(x) = rx$ and $\sigma(x) = \sigma x$, then we have what is called geometric Brownian motion. The Black-Scholes option pricing model uses this type of model. A shorthand form of writing our general Brownian motion equation is to take the derivatives of both sides and write $dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t)$, which is called the stochastic differential equation. In the geometric Brownian motion example such as the Black-Scholes model, we would have $dX(t) = rX(t)dt + \sigma X(t)dB(t)$.

Another example is the Ornstein-Uhlenbech process. In this case we have $dX(t) = -aX(t)dt - \sigma dB(t)$, where $a > 0$. Note that the $-aX(t)dt$ is mean reverting to 0. As it turns out, this process is very tractable and is a Gaussian process. One last model is a square root diffusion model, which would be written $dX(t) = -aX(t)dt - \sigma\sqrt{X(t)}dB(t)$. Again, it is mean reverting. This model is also called the CIR Model and Feller diffusion.

Note that for notational purposes, we have been assuming that the values in $dX(t) = \mu(X(t))dt + \sigma(X(t))dB(t)$ are scalar, but they can also be vector valued. $X(t)$ is a $d \times 1$ column vector. $\mu(x)$ is $d \times 1$. $\sigma(x)$ is a $d \times m$ matrix. $B(\cdot)$ is an $m \times 1$ vector-valued process, where each of the components are independent scalar standard Brownian motions. The solution to the stochastic differential equation $X(\cdot)$ is a Markov process and has continuous paths.

In a Markov jump process (discrete states, continuous time), we had an expression for $E_x[g(X(h))]$. Let us examine what this value would be in our diffusion (continuous states, continuous time). We have

$X(h) - X(0) = \int_0^h \mu(X(s))ds + \int_0^h \sigma(X(s))dB(s)$
$= h\mu(X(0)) + \sigma X(0)B(h) + o(h)$

If we assume a function g is smooth, then $g(X(h)) = g(X(0)) + g'(X(0))(X(h) - X(0)) + g''(\frac{X(0)}{2})(X(h) - X(0))^2 + g^{(3)}(\frac{X(0)}{6})(X(h) - X(0))^3 + ...$ Therefore combining these two equations, we have that

$E_x[g(X(h))] = g(x) + g'(x)E_x[X(h) - X(0)] + \frac{g''(x)}{2}E_x[(X(h) - X(0))^2] + \frac{g^{(3)}(x)}{6}E_x[(X(h) - X(0))^3] + ...$
$= g(x) + g'(x)[\mu(x)h + o(h)] + \frac{g''(x)}{2}[h\sigma^2(x) + o(h)] + o(h)$
$= g(x) = h[g'(x)\mu(x) + \frac{\sigma^2(x)}{2}g''(x) + o(h)]$

We can write this as $E_x[g(X(h))] = g(x) + h(Lg)(x) + o(h)$, where the operator $L = \mu(x)\frac{d}{dx} + \frac{\sigma^2(x)}{2}\frac{d^2}{dx^2}$. L is the stochastic differential equation analog of the rate matrix Q, and it called the "generator."

29.1 First Transition Analysis

Let us look at the first transition analysis for the hitting time for a diffusion. We would like to find $u^*(x) = E_x[T]$ where $T = inf\{t \geq 0 : X(t) \in C^C\}$. We can write $u^*(x) = E_x[T\mathbb{1}(T \leq h)] + E_x[\mathbb{1}(T > h)(h + u^*(X(h))]$. Note that $P_x(T \leq h) = O(e^{-\frac{b}{h}})$, which is a quantity that is exponentially small when h is small. Therefore, $\mathbb{1}(T > h) \approx 1$. Thus, we have

$u^*(x) = E_x[T\mathbb{1}(T \leq h)] + E_x[\mathbb{1}(T > h)(h + u^*(X(h))]$
$= E_x[h + u^*(X(h)) + o(h)]$
$= h + E_x[u^*(X(h))] + o(h)$
$= h + u^*(x)h(Lu^*)(x) + o(h)$

Subtracting $u^*(x)$ from both sides and dividing through by h we get

$-1 = (Lu^*)(x)$, when $x \in C$. For $x \in C^C$, $u^*(x) = 0$.

We can now look at the example for calculating discounted investment returns. We have $u^*(x) = E_x[\int_0^\infty e^{-\alpha t} r(X(t))dt]$. We can write this as

$u^*(x) = E_x[\int_0^h e^{-\alpha x} r(X(s))ds + \int_h^\infty e^{-\alpha s} r(X(s))ds]$
$= E_x[hr(x) + e^{-\alpha h} u^*(X(h))] + o(h)$
$= hr(x) + (1 - \alpha h)[u^*(x) + (Lu^*)(x)h] + o(h)$

Subtracting $u^*(x)$ from both sides and dividing through by h we get

$0 = r(x) + (Lu^*)(x) - \alpha u^*(x)$

If r is bounded, then u^* must be bounded, and then there is a unique bounded solution.

One more example is to look at the backwards equations. We have $u^*(t, x) = E_x[r(X(t))]$ and $u^*(0, x) = r(x)$ for all x. It is clear that $u^*(t, x) = E_x[u^*(t-h), X(h))] = u^*(t, x) - \frac{d}{dt}u^*(t, x)h + \frac{d}{dx}u^*(t, x)(X(h) - X(0)) + \frac{1}{2}\frac{d^2u^*(t,x)}{dx^2}(X(h) - X(0))^2 + o(h)$.

Subtracting $u^*(t, x)$ from both sides and dividing through by h we get

$\frac{du^*}{dt} = Lu^*$ subject to $u^*(0) = r$. Therefore, the PDE that we need to solve is $\frac{du}{dt} = \mu(x)\frac{du}{dx} + \frac{\sigma^2(x)}{2}\frac{d^2u}{dx^2}$ subject to $u(0, x) = r(x)$. When we have $\mu = 0$ and $\sigma^2 = 1$, then we have the "heat equation" or the diffusion equation.

Another example is if we want to calculate probability of exiting through B, where $B \in C^C$. Now we have $u^*(x) = E_x[u^*(X(h))] + o(h)$, since $P_x[X_T \in B, T < \infty] \approx 1$. Therefore, we have that $u^*(x) = u^*(x) + (Lu^*)(x)h + o(h)$. Subtracting both sides by $u^*(x)$ and dividing by h, we get that $0 = Lu^*$ for $x \in C$ subject to the boundary conditions that $u^* = 1$ for $x \in B$ and $u^* = 0$ for $x \in C^C \backslash B$.

29.2 Stochastic Control

In stochastic control, our stochastic differential equation now is $dX(t) = \mu(X(t), A(t))dt + \sigma(X(t), A(t))dB(t)$, which is called a controlled diffusion. Again, we assume adaptedness. Let's look at the example of the expected infinite time horizon discounted cash flow. We want to find

$v^*(x) = E[\int_0^\infty e^{-\alpha s} r(X(s), A(s))|X_0 = x]$
$= max_a[E[\int_0^h e^{-\alpha s} r(X(S), a)ds] + E_a[\int_h^\infty e^{-\alpha s} r(X(s), A(s))ds|X(0) = x]]$
$= max_a[hr(x, a) + o(h) + e^{-\alpha h} E_a[v^*(X(h))]]$
$= max_a[hr(x, a) + o(h) + (1 - \alpha h + o(h))(v^*(x) + (L_a v^*)(x)h + o(h))]$
$= max_a[hr(x, a) + v^*(x) - \alpha v^*(x)h + (L_a v^*)(x)h + o(h)]$

Therefore, by subtracting $v^*(x)$ from both sides and dividing by h, we get the HJB equation

$0 = max_a[r(x,a) + (L_a v^*)(x) - \alpha v^*(x)]$, where $L_a = \mu(x,a)\frac{d}{dx} + \frac{\sigma^2(x,a)}{2}\frac{d^2}{dx^2}$

Let's look at a different example, this one about optimal stopping (such as an American option). Let T be the stopping time and we want to find $v^*(x) = sup_T E_x[r(X(T))]$. Therefore, we write $v^*(x) = max[r(x), E_x[v^*(X(h))]] + o(h)$. If X is in a continuation region, then we have $v^*(x) = E_x[v^*(X(h))]$ subject to $v^*(x) \geq r(x)$. Simplifying, we get $v^*(x) = v^*(x) + (Lv^*)(x)h + o(h)$, and therefore $Lv^*(x) = 0$. If X is in a stopping region, then we have $v^*(x) \geq E_x[v^*(X(h))]$, subject to $v^*(x) = r(x)$. Simplifying, we get $v^*(x) \geq v^*(x) + (Lv^*)(x)h + o(h)$, and therefore $Lv^*(x) \leq 0$. Putting these two pieces together, we get our HJB equation

$0 = max[r(x) - v^*(x), (Lv^*)(x)]$

29.3 Diffusion Approximations

At times, we may want to approximate discrete time modeling with continuum modeling (diffusion). To have this approximation, we need to rescale time and space. Let's take the example when we have $Z_1, Z_2, ..., Z_n$ iid Bernoulli where $P(Z_i = 1) = \frac{1}{2}$ and $P(Z_i = -1) = \frac{1}{2}$. The $E[Z_i] = 0$ and $Var[Z_i] = 1$. If we denote the sum $S_n = Z_1 + Z_2 + ... + Z_n$, then by the CLT, we have that $\frac{S_n}{\sqrt{n}} \Rightarrow N(0,1)$. We might want to model what happens when $n \to \infty$, so we use a diffusion to model this rather than a discrete model. We have a time scale of order n that goes with a spatial scale of order $\sqrt{n}$, as well as stationary independent increments. Then, we have

$X_n(t) = \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \to$ stationary independent increments

It can be proved that $(X_n(t_1), X_n(t_2), ..., X_n(t_m)) \Rightarrow (B(t_1), B(t_1), B(t_2), ..., B(t_m))$. X_n has finite dimensional distributions that converge to B. Since weak convergence extends to more abstract spaces, we have that $X_n \Rightarrow B$ in $D[0, \infty)$. By the continuous mapping theorem, any continuous function g preserves the weak convergence $g(X_n) \Rightarrow g(B)$. One particular function g that we want to look at is $g(x) = max_{0 \leq t \leq 1} x(t)$, the "maximum function." Therefore, we have $max_{0 \leq t \leq 1} X_n(t) \Rightarrow max_{0 \leq t \leq 1} B(t)$ as $n \to \infty$. Thus, $max_{0 \leq j \leq n} \frac{S_j}{\sqrt{n}} \Rightarrow max_{0 \leq t \leq 1} B(t)$ as $n \to \infty$.

Let us then try to calculate $P(max_{0 \leq t \leq 1} B(t) > x)$, or the probability that we exceed x before time 1. We have

$P(max_{0 \leq t \leq 1} B(t) > x) = P(T_x \leq 1)$
$= P(T_x \leq 1, B(1) > x) + P(T_x \leq 1, B(1) < x)$
$= 2P(T_x \geq 1, B(1) > x)$, since it is symmetric
$= 2P(B(1) > x)$
$= 2P(N(0,1) > x)$

Note that by using an continuous approximation, we were able to easily do this calculation due to path continuity. We could not have done this in a discrete model.

Our discrete variable $X_n(t)$ is a random walk with time unit 1 and space unit 1. As mentioned before, when doing a continuous approximation, we need to scale these units, and we can scale the time unit to $\frac{1}{n}$ and the space unit to $\frac{1}{\sqrt{n}}$. If we look at the discrete equation for expected hitting time, we have $u^*(x) = E_x[T]$, which satisfies the equation $u(x) = 1 + \frac{1}{2}u(x+1) + \frac{1}{2}u(x-1)$. With this new re-scaling, we can write

$u_n(x) = \frac{1}{n} + \frac{1}{2}u_n(x + \frac{1}{\sqrt{n}}) + \frac{1}{2}u_n(x - \frac{1}{\sqrt{n}})$
$u(x) = \frac{1}{n} + \frac{1}{2}[u(x) + u'(x)\frac{1}{\sqrt{n}} + \frac{u''(x)}{2} * \frac{1}{n} + ...] + \frac{1}{2}[u(x) - u'(x)\frac{1}{\sqrt{n}} + \frac{u''(x)}{2} * \frac{1}{n} + ...]$
$= \frac{1}{n} + u(x) + \frac{u''(x)}{2} * \frac{1}{n} + o(\frac{1}{n})$

Subtracting $u(x)$ from both sides and then multiplying by n, we get that

$-1 = \frac{1}{2}u''(x)$

We recognize this as the formula for the expected hitting time for Brownian motion and thus see a connection between discrete and continuous time modeling.