Stanford CME 308 Notes

Yikun Chi

September 26, 2022

Most of the contents in this notes are from Stanford CME 308.

Contents

Chapter 1

Primer

1.1 Random Variables Convergence

A random variable is a **function** that maps a point in the sample space Ω to measurable space, such as $\mathbb{R}$. This means that when we say a r.v. converges, it means a sequence of function converges. E.g.: let $S_n(w)$ be the number of heads in the n toss. If we want to say as $n \to \infty$, $\frac{1}{n}S_n \to \frac{1}{2}$, there are actually many different way of converges, such as

- point-wise convergence: $f_n(x) \to f_\infty(x) \forall x$
- Convergence almost for every point: $f_n(x) \to f_\infty(x)$ for $a.l.x.$ i.e.: it only not convergence for a set that measures 0. For example rational vs. real.
- Uniform convergence: $\sup_x |f_n(x) - f_\infty(x)| \to 0$

1.2 Problem of Points

Problem of points motivates the intuition of expectation, law of large number, and conditional probability. Player A and B play for \$1 and the first one wins 5 rounds gets the price. Current score is (4, 3). How to split the pot?

$$\begin{aligned} u(4,3) &= P(\textbf{A wins}|(4,3)) \\ &= \frac{1}{2}u(5,3) + \frac{1}{2}u(4,4) \\ &= \frac{1}{2} + \frac{1}{2}(u(5,4) + u(4,5)) \\ &= \frac{3}{4} \end{aligned}$$

Chapter 2

Probability

2.1 Measure Space

A measure space is a triple $(\mathcal{X}, \mathcal{A}, \mu)$ where

- $\mathcal{X}$ is a set
- $\mathcal{A}$ is a $\sigma-$algebra on the set $\mathcal{X}$
- μ is a measure on $(\mathcal{X}, A)$

2.1.1 Sigma algebra

A $\sigma-$algebra on a set $\mathcal{X}$ is a collection Σ of subsets of $\mathcal{X}$ that are:

- Closed under complement
- Closed under countable unions and countable intersections

2.1.2 Measure

Let $\mathcal{X}$ be a set and Σ a $\sigma-$ algebra over $\mathcal{X}$. A function μ that maps Σ to extended real number line is called a measure if the following are true:

- Non-negativity: $\mu(E) \geq 0 \forall E \in \pm$
- Null empty set: $\mu(\emptyset) = 0$
- Countable additivity: For all countable collections $\{E_k\}_{k=1}^{\infty}$ of pairwise disjoint sets in Σ, we have $\mu(\bigcup E_k) = \sum \mu(E_k)$

2.2 Consistency Property

For $n \geq 0$. Let P_n be a probability of $\mathbb{R}^{n+1}$. The sequence $P_n : n \geq 1$ is said to have the consistency property if for $n \geq m$:

$$P_n((X_0, ..., X_m) \in \cdot) = P_(m_0, ..., X_m) \in \cdot)$$

2.3 Kolmogorov's Extension Theorem

Let $(P_n : n \geq 1)$ be a sequence of probabilities having the consistency property, then there exists a probability P on Ω for which

$$P((X_0, ..., X_m) \in \cdot) = P_m((X_0, ..., X_m) \in \cdot)$$

for $m \geq 0$ and P is unique.

2.4 Convergence

2.4.1 Convergence in Probability

We say $Z_n \xrightarrow{P} Z_\infty$ iff $\forall \epsilon > 0$:

$$P(|Z_n - Z_\infty| > \epsilon) \to 0 \textbf{ as } n \to \infty$$

2.4.2 Bounded Convergence Theorem

Suppose $Y_n \xrightarrow{P} Y_\infty$ as $n \to \infty$. If there exists $c < \infty$ such that $P(|Y_n| \leq c) = 1$ for $n \geq 1$, then $E[Y_n] \to E[Y_\infty]$ as $n \to \infty$. So it establishes the relationship between convergence in probability and convergence of expectation.

Proof

$$\begin{aligned}
|E[Y_n] - E[Y_\infty]| &\leq E[|Y_n - Y_\infty|] \\
&\leq E\left[|Y_n - Y_\infty| * \mathbb{I}\left\{|Y_n - Y_\infty| \leq \epsilon\right\}\right] + E\left[|Y_n - Y_\infty| * \mathbb{I}\left\{|Y_n - Y_\infty| > \epsilon\right\}\right] \\
&\leq \epsilon * P(|Y_n - Y_\infty| \leq \epsilon) + 2cP(|Y_n - Y_\infty| > \epsilon) \\
&\leq \epsilon + 2cP(|Y_n - Y_\infty| > \epsilon)
\end{aligned}$$

As $n \to \infty$, we have $\lim_{n\to\infty} sup\, |E[Y_n] - E[Y_\infty]| \leq \epsilon$

2.4.3 Dominated Convergence Theorem

Suppose $Y_n \xrightarrow{P} Y_\infty$ as $n \to \infty$ and that there exists a r.v. $W \geq 0$ for which $E[W] < \infty$ and

$$|Y_n(\omega)| \leq W(\omega)$$

for $\omega \in \Omega$ and $n \geq 1$, then

$$E[Y_n] \to E[Y_\infty]$$

as $n \to \infty$

2.4.4 Monotone Convergence Theorem

Suppose $Y_n : n \geq 0$ is a sequence of non-negative r.v., and

$$Y_n(\omega) \leq Y_{n+1}(\omega)$$

as $n \to \infty$ for $n \geq 0$ and $Y_\infty(\omega) \triangleq \lim_{n\to\infty} Y_n(\omega)$, then we have

$$E[Y_n(\omega)] \leq E[Y_{n+1}(\omega)]$$

2.4.5 Fatou's Lemma

Suppose Y_n is a sequence of non-negative r.v., then

$$E[\lim_{n\to\infty} Y_n] \leq \lim_{n\to\infty} E[Y_n]$$

2.4.6 Small O notation

$f(\theta) = o(g(\theta))$ as $\theta \to 0$ if $\frac{f(\theta)}{g(\theta)} \to 0$ as $\theta \to 0$

2.4.7 big O notation

If $|f(\theta)| \leq cg(\theta)$ then $f(\theta) = O(g(\theta))$

2.4.8 Convergence in Distribution

This is also called convergence weakly or converge in law. Let $Z_n \in \mathbb{R}$ be a sequence of r.v.s. Then We say $Z_n \xrightarrow{D} Z_\infty$ or $Z_n \Rightarrow Z_\infty$ iff

$$\lim_{n\to\infty} F_n(x) = F_\infty(x)$$

for every number $x \in \mathbb{R}$ at which $F_\infty(x)$ is continuous. Notice that we do not need this to hold at all n.
For example, let X_n to be a uniform distribution on the interval of $(0, \frac{1}{n})$. It's cdf is

$$F_{X_n}(x) = \begin{cases} 0 & x \leq 0 \\ xn & x \in (0, \frac{1}{n}) \\ 1 & x \geq \frac{1}{n} \end{cases}$$

Now consider the degenerate random variable $X = 0$, we know that $F_X(0) = 1$, But $F_{X_n}(0) = 0$, so those two cdf failed to converge at point $x = 0$ where F is discontinuous. But they $X_n \xrightarrow{D} X$

2.4.8.1 Extention to Skorohod Representation

If $Z_n \Rightarrow Z_\infty$, then there exists a probability space supporting a sequence $Z'_n : 1 \leq n \leq \infty$ for which

- $Z'_n \stackrel{D}{=} Z_n$
- $Z'_n \to Z'_\infty$ a.s. as $n \to \infty$

2.4.8.2 Expansion

$Z_n \stackrel{D}{\to}$ iff any of the three

- $F_{Z_n} \to F_{Z_\infty}$ at all continuity points of F_{Z_∞}
- $F^{-1}_{Z_n}(z) \to F^{-1}_{Z_\infty}(z)$ holds at all except countably many z
- $P(F^{-1}_{Z_n}(z) \to F^{-1}_{Z_\infty}(z)) = 1$ as $n \to \infty$

2.4.8.3 Sampling from Uniform Distribution

Given $u_1, ...$ iid from uniform $[0, 1]$, we can obtain x_i iid with cdf F_x through inversion

$$F_X(F_X^{-1}(x)) = X = F^{-1}(F(X))$$

where $F_x^{-1}(x) = inf\{y : F_X(y) \geq x\}$. This means that $F_X^{-1}(u)$ has cdf F_X.

Example Suppose we want to generate X_i from $Exp(\lambda)$. We know that $f(x) = \lambda e^{-\lambda x}$ and $F(x) = 1 - e^{-\lambda x}$. In order to generate this from uniform $[0, 1]$, we need to

1. $y = F^{-1}(x)$. So $F(y) = x = 1 - e^{-\lambda y} \Longrightarrow y = -\frac{1}{\lambda}(1 - x)$
2. $X_i = -\frac{1}{\lambda}(1 - U_i)$

2.4.9 Almost Sure Convergence

This is also called w/probability 1 or almost certain convergence.

$$P(\lim_{n\to\infty} X_n = X) = 1$$

Alternative definition:

$$P(A) = 1$$
$$A = \{\omega : Z_n(\omega) \to Z_\infty(\omega) \textbf{ as } n \to \infty\}$$

2.4.10 Hierarchy of Convergence

Almost sure $Y_n \to Y_\infty$ implies convergence in probability $Y_n \stackrel{P}{\to}$, which implies convergence in distribution $Y_n \Rightarrow Y_\infty$

2.4.11 Continuous Mapping Theorem

Suppose h is continuous, then if $X_n \to X$ (in any of the three convergence definition), we have $h(X_n) \to h(X)$ (in the corresponding convergence definition)

2.4.12 TFAE

TFAE:

- $Z_n \Rightarrow Z_\infty$
- $\exists$ a probability space such that $Z'_n \to Z'_\infty$ a.s.with $Z'_n \Rightarrow Z_n$
- $E[h(Z_n)] \to E[h(Z_\infty)] \forall h \in$ bounded continuous space
- $c_{Z_n}(\theta) \to c_{Z_\infty}(\theta) \forall \theta$

2.5 Markov Inequality

Let W be a random variable. $W \geq 0$, $E[W] < \infty$, $x \geq 0$ then we have

$$P(W \geq x) \leq \frac{E[W]}{x}$$

Proof:

$$\begin{aligned} P(W \geq x) &= E[\mathbb{I}(W > x)] \\ &\leq E[\frac{W}{x} * \mathbb{I}(W > x)] && (\tfrac{W}{x} > 1) \\ &\leq E[\frac{W}{x}] && \text{(Indicator variable can only be 0, 1)} \\ &= \frac{E[W]}{x} \end{aligned}$$

2.6 Chebyshev's inequality

Let $E[W^2] < \infty$ (finite variance), then

$$P(|W - E[W]| \geq \epsilon) \leq \frac{Var(W)}{\epsilon^2}$$

Proof : Let $|W - E[W]|^2$ be the W' and apply Markov Inequality

$$\begin{aligned} P(|W - E[W]| \geq \epsilon) &= P(|W - E[W]|^2 \geq \epsilon^2) \\ &\leq \frac{E[|W - E[W]|^2]}{\epsilon^2} \\ &= \frac{Var(W)}{\epsilon^2} \end{aligned}$$

2.7 Jensen's Inequality

Let X be a random variable, and ϕ is a convex function, then

$$\phi(E[X]) \leq E[\phi(X)]$$

Proof Sketch : Given convex function, we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) \geq \phi(\frac{1}{n}\sum_{i=1}^{n}X_i)$$

The left hand side converges in probability to $E[\phi(X)]$ due to LLN, the right hand side converges in probability to $\phi(E[X])$ with a lemma that says if $Z_n \xrightarrow{P} Z_\infty$, and h is continuous, then $h(Z_n) \xrightarrow{P} h(Z_\infty)$

2.7.1 Cauchy-Schwarz Inequality

Cauchy-Schwarz inequality is a special case of Jensen's inequality when $\phi(x) = x^2$

2.7.2 Lyapunov's Inequality

Cauchy-Schwarz inequality can also be seen as a special case of Lyapunov's inequality. It states that if $0 < s < t$, then

$$E[|Z|^s]^{\frac{1}{s}} \leq E[|Z|^t]^{\frac{1}{t}}$$

2.8 Characteristic function

Given a real-valued r.v. X, the characteristic function of X is defined as

$$c_X(\theta) = E[e^{i\theta X}]$$

2.8.1 Characteristic function and pdf

If X has a pdf f_X, following the definition of expected value, the characteristic function is given by

$$c_X(\theta) = \int_{-\infty}^{\infty} e^{i\theta x} f_X(x)\, dx$$

Characteristic function can be seen as a Fourier transform of the density.

2.8.2 Existence of characteristic function

Since $\left|e^{i\theta X}\right| \leq 1$ (Euler's Formula), the characteristic function c_X exists for any random variable X.

2.8.3 Characteristic function and density

Characteristic function specifies the distribution function. If $a < b$, then

$$P\left(X \in (a,b)\right) + \frac{1}{2}\left(P(X=a) + P(X=b)\right) = \lim_{t\to\infty} \frac{1}{2\pi} \int_{-t}^{t} \frac{e^{-i\theta a} - e^{-i\theta b}}{i\theta} c_X(\theta)\, d\theta$$

2.8.4 Characteristic function and moment

Characteristic function can recover moments. Suppose $E[|X|^k] < \infty$ for some $k \geq 1$, then c_X is k times continuously differentiable, and its kth derivatives can be given by

$$C_X^{(k)}(\theta) = i^k E[X^k e^{i\theta X}]$$

Example of when K=1

$$\frac{c_X(\theta+h) - c_X(\theta)}{h} = \frac{E[e^{i(\theta+h)X}] - E[e^{i\theta X}]}{h}$$

According to Euler's formula, the real part of this is

$$\frac{\cos((\theta+h)X) - \cos(\theta X)}{h}$$

We can think of this as the average rate of change of $\cos(\theta X)$ w.r.t θ from θ to $\theta + h$, so by mean value theory we know that this average rate of change will equal to $-X\sin(\xi X)$ where $\xi \in [\theta, \theta+h]$. So we know

$$\left|\frac{\cos((\theta+h)X) - \cos(\theta X)}{h}\right| \leq |X|$$

Since $E[X] < \infty$, by Dominated Convergence Theorem we have

$$E\left[\frac{\cos((\theta+h)X) - \cos(\theta X)}{h}\right] \to -E[X\sin(\theta X)]$$

as $h \to 0$, similarly we have

$$E\left[\frac{\sin((\theta+h)X) - \sin(\theta X)}{h}\right] \to E[X\cos(\theta X)]$$

So we have

$$\begin{aligned}
\frac{c_X(\theta+h) - c_X(\theta)}{h} &= \frac{E\left[e^{i(\theta+h)X} - e^{i\theta X}\right]}{h} \\
&= \frac{E\left[\cos((\theta+h)X) - \cos(\theta X)\right]}{h} + \frac{iE\left[\sin((\theta+h)X) - \sin(\theta X)\right]}{h} \\
&\to -E[X\sin(\theta X)] + iE[X\cos(\theta X)] \\
&= i(E[X\cos(\theta X) + iX\sin(\theta X)]) \\
&= iE[Xe^{i\theta X}]
\end{aligned}$$

2.8.5 Characteristic function of sums of independent random variables

$$\begin{aligned} c_{S_n}(\theta) &= E[e^{i\theta S_n}] \\ &= E[\prod_{j=1}^{n} e^{i\theta X_j}] \\ &= \prod_{j=1}^{n} E[e^{i\theta X_j}] && \text{(By independence)} \\ &= \prod_{j=1}^{n} c_{X_j}(\theta) \end{aligned}$$

2.8.6 Characteristic function approximation

Using Taylor Series expansion, if $E[X] = 0$ and $Var(x) = \sigma^2 < \infty$, then $c_X(\theta) = 1 - \frac{\theta^2}{2}\sigma^2 + o(\theta^2)$

2.8.7 Characteristic function convergence

Let X_n be a sequence of r.v. then

$$X_n \Rightarrow X_\infty \qquad \text{(Convergence in distribution)}$$

as $n \to \infty$ if and only if

$$c_{X_n}(\theta) \to C_{X_\infty}(\theta) \qquad \text{(Convergence that is not related to randomness)}$$

as $n \to \infty$ at each $\theta \in \mathbb{R}$. Similarly, if there exists a function f_∞ that is continuous in a neighborhood of 0 for which

$$c_{X_n}(\theta) \to f_\infty(\theta)$$

as $n \to \infty$ at each $\theta in \mathbb{R}$, then f_∞ is the characteristic function of a finite-valued r.v. X_∞ and

$$X_n \Rightarrow X_\infty$$

2.9 Law of Large Number

2.9.1 Weak Law of Large Number

Let $x_1, ..., x_n$ be i.i.d samples and $E[x_i] < \infty$, then

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{P} E[x_i]$$

Proof for a weaker version by assuming $E[x_i^2] < \infty$ and iid:

$$\begin{aligned} P(\left|\bar{X}_n - E[X]\right| > \epsilon) &\leq \frac{Var(\bar{X}_n)}{\epsilon^2} \\ &= \frac{Var(X_1)}{n\epsilon^2} \\ &\to 0 \textbf{ as } n \to \infty \end{aligned}$$

Proof 2 for a weaker version by assuming $E[x_i^2] < \infty$ and X_is are stationary sequence ($Cov(X_m, X_n) = Cov(X_0, X_{n-m})$ distribution are time invariant) instead of iid.

$$\begin{aligned} P(\left|\frac{1}{n}S_n - E[X_i]\right| > \epsilon) &\leq \frac{Var(\bar{X})}{\epsilon^2} \\ &= \frac{Var(S_n)}{n^2\epsilon^2} \\ &= \frac{1}{n^2\epsilon^2}E[(S_n - n * \bar{X})^2] \\ &= \frac{1}{n^2\epsilon^2}E[\sum_{i=1}^{n}\tilde{X}_i^2] \qquad (\tilde{X}_i = X_i - E[X_i]) \\ &= \frac{1}{n^2\epsilon^2}E[\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{X}_i\tilde{X}_j] \\ &= \frac{1}{n^2\epsilon^2}E[\sum_{i=1}^{n}\sum_{j=1}^{n}Cov(X_i, X_j) \\ &= \frac{1}{n^2\epsilon^2}E[\sum_{i=1}^{n}\sum_{j=1}^{n}C(i-j) \quad \text{(Covariate time invariant)} \\ &= \frac{1}{n^2\epsilon^2}\left(n * C(0) + 2 * \sum_{i=1}^{n-1}(n-i)C(i))\right) \\ &= \frac{C(0)}{n\epsilon^2} + \frac{2}{n}\frac{\sum(1-\frac{i}{n})C(i)}{\epsilon^2} \\ &\leq \frac{C(0)}{n\epsilon^2} + \frac{2}{n\epsilon^2}\sum C(i) \end{aligned}$$

We know the first part goes to 0 as $n \to \infty$. Now for the second part, we know that $C(n) \to 0$ as $n \to \infty$. $\forall \epsilon > 0$, we have $\exists \tilde{n}$ such that $|C(j)| \leq \epsilon_j \forall j \geq \tilde{n}$. So

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \left|\sum C(j)\right| &\leq \frac{1}{n} \sum_{i=1}^{\tilde{n}} |C(j)| + \frac{1}{n} \sum_{i=\tilde{n}+1}^{\infty} |C(j)| \\ &\leq \frac{1}{n} \sum_{i=1}^{\tilde{n}} |C(j)| + \frac{1}{n} n\epsilon_j \\ &\leq \frac{1}{n} \sum_{i=1}^{\tilde{n}} |C(j)|\, \epsilon_j\end{aligned}$$

But the choice of ϵ_j is arbitrary. So the whole thing converges to 0 too. Hence complete the proof.

2.9.2 Strong Law of Large Numbers

Let X_i be an iid sequence of r.v. for which $E[X_i] < \infty$, then

$$\frac{S_n}{n} \overset{a.s}{\to} E[X_i]$$

Proof of bounded iid case

Lemma 1:
Let $C_k : k \geq 1$ to be a sequence of events for which $P(C_k) = 1$ for $k \geq 1$, then

$$P(\bigcap_{k=1}^{\infty} C_k) = \lim_{n\to\infty} P(\bigcap_{k=1}^{\infty} C_k) = 1$$

Proof Part 1:
Notice that in the finite case $P(\bigcap_{k=1}^{n} C_k) = 1$ can be shown by looking at base case $P(C_1 \cap C_2) = P(C_1) + P(C_2) - P(C_1 \cup C_2) \geq 1 + 1 - 1$, and do induction. So we have $\lim_{n\to\infty} P(\bigcap_{k=1}^{\infty} C_k) = 1$

Proof Part 2:
First we can connect probability to expected value of indicator variable:

$$P(\bigcap_{k=1}^{\infty} C_k) = E\left[\mathbb{I}\left\{\bigcap_{k=1}^{\infty} C_k\right\})\right] \qquad P(\bigcap_{k=1}^{n} C_k) = E\left[\mathbb{I}\left\{\bigcap_{k=1}^{n} C_k\right\})\right]$$

Second we can transform indicator of intersection of events to product of individual indicator

$$\mathbb{I}\left\{\bigcap_{k=1}^{\infty} C_k\right\}) = \prod_{k=1}^{\infty} \mathbb{I}\{C_k\} \qquad \mathbb{I}\left\{\bigcap_{k=1}^{n} C_k\right\}) = \prod_{k=1}^{n} \mathbb{I}\{C_k\}$$

We can define the distribution of the product of indicator varibales

$$Y_n \triangleq \prod_{k=1}^{n} \mathbb{I}\{C_k\} \qquad\qquad Y_\infty \triangleq \prod_{k=1}^{\infty} \mathbb{I}\{C_k\}$$

Since indicator variables can only take on value 0 or 1, we have a trivial monotone sequence

$$Y_n \searrow Y_\infty$$

By BCT we have $1 = E[Y_n] \to E[Y_\infty]$. Hence

$$1 = E[Y_n] = P(\bigcap_{k=1}^{n} C_k) \to P(\bigcap_{k=1}^{\infty} C_k)$$

Proof Statement: Let X_i be an iid sequence for which there exists $c < \infty$ such that $P(|X_1| \leq c) = 1$, then $P(A) = 1$ where $A = \{\omega : \frac{1}{n}(X_1(\omega) + ... + X_n(\omega)) \to E[X_1] \text{ as } n \to \infty\}$

Proof:
$\bar{X}_n \to E[X_1]$ implies that $\forall \epsilon > 0 \, \exists N \, s.t. \, n \geq N(\epsilon) \implies \left|\bar{X}_n - X_\infty\right| < \epsilon$. In another word, there are only finite number of times where $\left|\bar{X}_n - X_\infty\right| > \epsilon$. Further notice that we do not have to check every ϵ, we can just check ϵ in a countable set such as $\epsilon = \frac{1}{n}$ So we can rewrite A as

$$A = \bigcap m \geq 1 A_m$$
$$A_m = \{\omega : \left|\frac{S_n(\omega)}{n} - E[X_1]\right| > \frac{1}{m} \text{ only finite many times}\}$$

Notice that $\omega \in A_m$ iff $\sum_{i=1}^{\infty} \mathbb{I}\left\{\left|\frac{S_n(\omega)}{n} - E[X_1]\right| > \frac{1}{m}\right\} < \infty$, i.e.: finite number of times violating the inequality.

$$\begin{aligned}
& E\left[\sum_{i=1}^{\infty} \mathbb{I}\left\{\left|\frac{S_n(\omega)}{n} - E[X_1]\right| > \frac{1}{m}\right\}\right] \\
&= \sum_{i=1}^{\infty} E\left[\mathbb{I}\left\{\left|\frac{S_n(\omega)}{n} - E[X_1\right| > \frac{1}{m}\right\}\right] && \text{(Fibini's Theorem)} \\
&\leq \sum_{i=1}^{\infty} \frac{Var(X_1)}{i\epsilon^2} && \text{(Chebyshev's Inequality)}
\end{aligned}$$

Notice that this series does not converge, but if we replace i with i^2, it converges. If the expected value is not infinite, then we know that the actual value is also not finite, which means $P(A_m) = 1$. Then by Lemma 1 we have $P(A) = 1$.

This final bits shows why we can replace i with i^2:

$$\begin{aligned}\frac{S_{(n+1)^2}}{(n+1)^2} - \frac{S_{n^2}}{n^2} &\leq \sum_{n^2+1}^{(n+1)^2} |X_j| / n^2 \\ &\leq \sum_{n^2+1}^{(n+1)^2} \frac{c}{n^2} \\ &= \frac{c(2n+1)}{n^2} \to 0 \text{ as } n \to \infty\end{aligned}$$

2.9.3 Generalization of the LLN using ergodic theorem

An $S-$valued sequence $(X_j : j \geq 0)$ is stationary if

$$(X_{n+j}) \stackrel{D}{=} (X_j)$$

Birkhoff's version of the Ergodic Theorem says that if X_j is a real-valued stationary sequence with $E[X_0] < \infty$, then there exists a random variable W for which

$$\frac{1}{n} \sum_{j=0}^{n-1} X_j \stackrel{a.s.}{\to} W$$

as $n \to \infty$ and $E[W] = E[X_0]$

2.9.4 SLLN and Infinite Sequence Probability Example

Setup: Consider X_i to be iid Bernulli with parameter $\frac{1}{2}$. So the natural sample space is $\Omega = \{0,1\}^{\mathbb{Z}_+}$. Let $X_i(\omega) = \omega_i$ where $\omega = (\omega_0, \omega_1, ...)$.

Now let's consider event A in the almost surely convergence definition. It is the infinite-dimensional event consisting of the sequence $\omega \in \Omega$ such that:

$$A = \{\omega : \frac{1}{n} \sum_{i=0}^{n-1} X_i(w) \to \frac{1}{2} \textbf{ as } n \to \infty\}$$

In finite-dimensional computation, we know that the probability of a particular sequence is $P(X_0 = i_0, ..., X_{n-1} = i_{n-1}) = 2^{-n}$. So we can get the probability of any finite-dimensional event A_n through $2^{-n} |A_n|$ where $|A_n|$ is the number of sequences lying in A_n. This approaches fails because if the sequence is infinite dimension, its probability equals to 0.
In order to think about the infinite sequence probability $P((X_0, X_1, ...) \in \cdot)$, we approach it from the binary representation of a $[0,1]$ uniform distribution.
Let $\tilde{\Omega} = [0,1]$ and let $\tilde{P}$ be the uniform distribution on $\tilde{\Omega}$ and let $Y(\tilde{\omega}) = \tilde{\omega}$, then

$$\tilde{P}(Y \in B) = \int_B dy$$

For $\tilde{\omega} \in [0,1]$, there exists a dyadic expansion (binary representation):

$$\tilde{\omega} = \sum_{j=1}^{\infty} I_j(\tilde{\omega}) 2^{-j}$$

where $I_j(\tilde{\omega}) \in \{0,1\}$. We can then construct the probability P as follows:

$$P((X_0, ...) \in \cdot) \stackrel{\triangle}{=} \tilde{P}\{(I_1, ...) \in \cdot\}$$

So we transformed the probability of an infinite sequence into the probability of the binary representation of a r.v. b/t $[0,1]$.

2.9.5 Application of LLN: Newstand Model

Let D be the demand, x be the order amount, profit be

$$W(x,D) = r * min(x,D) - cx + s * (x - min(x,D))$$

where r is the revenue per newspaper, c is the cost and s is the rebate/trash recovery rate. We have

$$\begin{aligned} E[W(x,D)] &= E[r * min(x,D) - cx + s * (x - min(x,D))] \\ &= rx * P(D > x) + r * \int_0^x y f_D(y)\, dy - E[cx + s * (x - min(x,D))] \\ &= rx * P(D > x) + r * \int_0^x y f_D(y)\, dy - cx + E[s * (x - min(x,D))] \\ &= rx * P(D > x) + r * \int_0^x y f_D(y)\, dy - cx + s * \int_0^x ((x-y) f_D(y)\, dy \end{aligned}$$

Taking derivative w.r.t x and set it to 0 we get

$$0 = rP(D > x) - rx f_D(x) + rx f_D(x) - c + s * \int_0^x f_D(y)\, dy + sx f_D(x) - sx f_D(x)$$

$$\implies P(D \leq x^*) = \frac{r-c}{r-s}$$

2.9.6 Application of LLN: Investment

Setup: Riskless investment return rate is $1+\delta$, and risky investment return rate is W_i at i-th time where $E[W_i] > 1+\delta$. Let f be the fraction of money allocated to riskless investment. So at any time, the return of investment is

$$V_n = V_{n-1}(f(1+\delta) + (1-f)W_n)$$

Strategy 1: Maximizing expected return

$$
\begin{aligned}
R_i &\equiv f(1+\delta) + (1-f)W_i \\
V_n &= V_0 \prod_{i=1}^{n} R_i \\
E[V_n] &= V_0 * \prod_{i=1}^{n} E[R_i] && \text{(Independent assumption)} \\
&= V_0 * E[R_1]^n
\end{aligned}
$$

So maxing $E[V_n]$ is equivalent to maximizing R_1, which means $f = 0$. However, this is actually very risky. If $P(W_i = 0) > 0$, then after n rounds, the chance of encountering at least 1 wipe out time is $1 - P(W_i \neq 0)^n$. So as n increases, this probability goes to 1. Once you are wiped out, there is no way to recover. So we need a different maximizing objective.

Strategy 2: Maximizing log expected return

$$
\begin{aligned}
\frac{1}{n} E[\log V_n] &= \frac{1}{n} E[\log V_0 + \sum_{i=1}^{n} \log R_i] \\
&= \frac{\log V_0}{n} + \frac{1}{n} \sum_{i=1}^{n} E[\log R_i] \\
&\xrightarrow{P} E[\log R_1]
\end{aligned}
$$

So the objective is to maximizes $E[\log R_1]$, which is smaller than $\log E[R_i]$ due to Jensen's Inequality.

2.10 Central Limit Theorem

2.10.1 Motivation

From LLN we know that $S_n \approx E[S_n]$, however, if we want to answer $P(S_n \in A)$, this result can only gives us a binary answer: 1 if $E[S_n] \in A$ and 0 otherwise. So this is not very useful. We want to approximate S_n by a random variable instead of just a number.

2.10.2 Definition

$$
\frac{S_n - nE[X_1]}{\sqrt{n}\sigma} \Rightarrow N(0,1)
$$

2.10.3 Proof

Statement: Suppose that X_n is an iid sequence of r.v. for which $\sigma^2 = var(X_1) < \infty$, then as $n \to \infty$

$$\frac{S_n - nE[X_1]}{\sqrt{n}} \Rightarrow \sigma N(0,1)$$

Notice the result is trivial if $\sigma = 0$, so we can add the assumption that $\sigma^2 > 0$. Let $\tilde{X}_i = \frac{X_i - E[x_1]}{\sigma}$, so we have $E[\tilde{X}_i] = 0$ and $Var(\tilde{X}_i) = 1$. Then we have

$$\begin{aligned} E[e^{i\theta \frac{\tilde{S}_n}{\sqrt{n}}}] &= E\left[e^{i\theta \frac{\tilde{X}_i}{\sqrt{n}}}\right]^n \\ &= c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}})^n \end{aligned}$$

Based on the differentiability of characteristic function and Taylor Series expansion, we have

$$\begin{aligned} c_{\tilde{X}_1}(\frac{\theta}{\sqrt{n}}) &= 1 + i\frac{\theta}{\sqrt{n}}E[\tilde{X}_1] - \frac{\theta^2}{2n}E[\tilde{X}_1^2] + o(\frac{1}{n}) \\ &= 1 - \frac{\theta^2}{2n} + o(\frac{1}{n}) \end{aligned}$$

So we have

$$\begin{aligned} E[e^{i\theta \frac{\tilde{S}_n}{\sqrt{n}}}] &= (1 - \frac{\theta^2}{2n} + o(\frac{1}{n}))^n \\ &\to e^{-\frac{\theta^2}{2}} \end{aligned}$$

Since $f_\theta(\theta) = e^{-\frac{\theta^2}{2}}$ is continuous in a neighborhood of the origin, then we know that must exsts a r.v. W having characteristic $f_\infty(\cdot)$ for which $\frac{\tilde{S}_n}{\sqrt{n}} \Rightarrow W$. We can then show W has to be normal random variable by showing that the normal random variable's character function is $e^{-\theta^2/2}$

2.11 Generalization of the CLT

2.11.1 To multiple dimension

Suppose that $X_1, ...$ is a sequence of iid $\mathbb{R}^d$ -valued random vectors for which $E[\|X_1\|^2] < \infty$, then as $n \to \infty$

$$\frac{S_n - nE[X_1]}{\sqrt{n}} \Rightarrow N(0, C)$$

Where $C = E[X_1 X_1^T] - E[X_1]E[X_1]^T$

2.11.2 Delta Method

Let $\bar{X}_n \xrightarrow{P} E[x_1]$, X_i are iid, $E[\|X_i\|_2^2] < \infty$, g is a continuous function that maps $\mathbb{R}^d \to \mathbb{R}$, then we have $g(\bar{X}_n) \xrightarrow{P} g(E[X])$. If g is differentiable in a neighborhood of $E[X]$, we have

$$\sqrt{n}(g(\bar{X}_n) - g(E[X_1])) \Rightarrow \nabla g(E[X])N(0, C)$$

In univariate case, we have:
If $\sqrt{n}(X_n - \theta) \Rightarrow N(0, \sigma^2)$, θ, and σ^2 are finite valued constant, then

$$\sqrt{n}(g(X_n) - g(\theta)) \Rightarrow N(0, \sigma^2 * g'(\theta)^2)$$

2.12 Tail

A real-valued r.v. X is light right tail if $E[e^{\theta X}] < \infty$ for some $\theta < 0$. It has a light left tail if $E[e^{\theta X}] < \infty$ for some $\theta < 0$. It has light tail if statement is true in some neighborhood of the origin. If X does not have light tails, it is heavy tailed.

2.12.1 Tail relation to decay

If $E[e^{\theta X_1}] < \infty$ for some $\theta > 0$, Markov's inequality ensures that for $x > 0$

$$P(X_1 > x) \leq \frac{E[e^{\theta X_1}]}{e^{\theta x}}$$

So $P(X_1 > x)$ decays to 0 at least exponentially quickly. So the presence of light tails is equivalent to assuming that the tails decay to zero at least exponentially.

2.13 Large Deviations

Motivation: If X_i are iid r.v. finite variance σ^2, CLT gives us

$$P(\frac{\tilde{S}_n}{\sqrt{n}} > x) \Rightarrow 1 - \Phi(x)$$

However, for a super large number, CLT tells us nothing except that both the left side and right side tends to 0. But in reality, they could converges to 0 at different rates.

2.13.1 Simple Upper Bound Via Exponential Inequality

Recall $\tilde{X}_i = \frac{X_i - E[X_i]}{\sigma}$. We wish to explore the behavior of

$$P(\frac{\tilde{S}_n}{n} > \gamma\sqrt{n}) = P(\frac{S_n}{n} > E[X_1] + \gamma\sigma) = P(S_n > ny)$$

Using Exponential inequality and Markov Inequality, we have

$$\begin{aligned} P(S_n > ny) &= P(e^{\theta S_n} > e^{\theta ny}) \\ &\leq \frac{E[e^{\theta S_n}]}{e^{\theta ny}} \end{aligned}$$

Based on iid, we have

$$E[e^{\theta S_n}] = E[e^{\theta X_1}]^n = e^{n\psi(\theta)} \qquad \psi(\theta) = \log E[e^{\theta X}]$$
(log moment generating function)

So

$$\begin{aligned} P(S_n > ny) &\leq \frac{E[e^{\theta S_n}]}{e^{\theta ny}} \\ &= \frac{e^{n\psi(\theta)}}{e^{\theta ny}} \\ &= e^{-n(\theta y - \psi(\theta)} \end{aligned}$$

If we want to find the tighest upper bound, we want to minimize $e^{-n(\theta y-\psi(\theta)}$, which means choosing θ that maximizes $\theta y - \psi(\theta)$, so the optimal solution $\theta^* = \theta(y)$ (simply means the optimal solution is a function of y) satisfies

$$\psi'(\theta^*) = y$$

To summarize, our upper bound is

$$P(S_n > ny) \leq e^{-n(y\theta(y)-\psi(\theta(y)))}$$

Or alternatively

$$\begin{aligned} &P(S_n > ny) \leq \exp(-nI(y)) \\ &I(y) = y\theta(y) - \psi(\theta(y)) \end{aligned} \qquad \text{(Rate function)}$$

2.13.2 Change of Measure

For $\theta > 0$, define the probability $P_\theta(\cdot)$ such that X_i are iid with

$$P_\theta(X_1 \in dx) \propto e^{\theta x} P(X_1 \in dx)$$

for some $x \in \mathbb{R}$. Intuitively, this means that for probability P_θ, the probability of X_1 in a infinitesimal area around x (analogous to $p(x)dx$) is proportional to a weighted original probability. Because $\theta > 0$, the probability puts higher likelihood on large values of x for X_1 then original P. In order to formalize it as a probability, we have

$$P_\theta(X_1 \in dx) = e^{\theta x - \psi(\theta)} P(X_1 \in dx)$$

If a function $f : \mathbb{R}^n \to \mathbb{R}_+$ is a non-negative function, it follows that

$$E[f(X_1, ..., X_n)] = E_\theta[e^{-\theta S_n + n\psi(\theta)} f(X_1, ..., X_n)]$$

This formula shows how the expectation in the original setting can be computed as the expectation in terms of the new probability distribution.

2.13.3 Lower Bound

From previous change of measure, we have

$$P(S_n > ny) = E_\theta[e^{-\theta S_n + n\psi(\theta)} \mathbb{I}\{S_n > ny\}]$$

Conceptually, given $S_n > ny$ is a rare event, S_n is more likley to be close to ny when this happens.

$$\begin{aligned} E[X_1|S_n = ny] &= \frac{1}{n}\sum_{j=1}^{n} E[X_j|S_n = ny] \\ &= E[\frac{S_n}{n}|S_n = ny] \\ &= y \end{aligned}$$

Notice the above argument is valid for all X_i due to iid, so we expect the conditional mean for any X_i to be close to y. So we want to choose θ such that $E_\theta[X_1] = y$, so we have

$$\begin{aligned} y = E_\theta[X_1] &= \int_{\mathbb{R}} x e^{-\theta x - \psi(\theta)} P(X_1 \in dx) \\ &= e^{-\psi(\theta)} \frac{\partial}{\partial\theta} \int_{\mathbb{R}} e^{\theta x} P(X_1 \in dx) \\ &= \frac{\frac{\partial}{\partial\theta} E[e^{\theta X_1}]}{E[e^{\theta X_1}]} \qquad \text{(By DCT)} \\ &= \psi'(\theta) \end{aligned}$$

Note that the idea θ actually is the same $\theta(y)$ as the upper bound. From LLN we have

$$P_{\theta(y)}(\left|\frac{S_n}{n} - y\right| \leq \epsilon) \to 1$$

So we have

$$\begin{aligned} P(S_n > ny) &= E_\theta[e^{-\theta S_n + n\psi(\theta)} \mathbb{I}\{S_n > ny\}] \\ &\approx E_\theta[e^{-\theta ny + n\psi(\theta)} \mathbb{I}\{S_n > ny\}] \\ &= e^{-\theta ny + n\psi(\theta)} P_{\theta(y)}(S_n > ny) \end{aligned}$$

Suppose that X_1 is such that

$$\begin{aligned} &\exists \theta(y)\, s.t.\, \psi'(\theta(y)) = y \\ &\psi(\theta) < \infty \text{ for } \theta \text{ in a neighborhood of } \theta(y) \end{aligned}$$

then

$$\frac{1}{n}\log P(S_n > ny) \to -I(y) \qquad\qquad I(y) = y\theta(y) - \psi((\theta(y))$$

A better result is that

$$P(S_n > ny) \approx \frac{1}{\sqrt{2\pi n \psi''(\theta(y))}} \exp(-nI(y))$$

2.14 Monte Carlo

2.14.1 Set up

Let X be a function of other random variables, such as $X = g(Y_1, ..., Y_n)$. Often we can simulate Y_i easily, but direct computation of X is hard due to g being a complicated / algorithmically implemented function. So we can use Monte Carlo Method:

1. Sample $Y = (Y_1, ..., Y_d)^T$
2. Compute $X = g(Y)$
3. Repeat step 1 and 2 n times to get the empirical expected value and variance.

then we know

$$\bar{X}_n \xrightarrow{P} E[X] \quad \text{(LLN)}$$

$$\sqrt{n}(\bar{X}_n - E[X]) \Rightarrow \sigma N(0,1) \quad \text{(CLT)}$$

$$\bar{X}_n \overset{D}{\approx} E[X] + \frac{\sigma}{\sqrt{n}} N(0,1) \quad \text{(Informal CLT)}$$

2.14.2 Convergence

From CLT, we can get the confidence interval

$$P(-Z \leq N(0,1) \leq Z) = 1 - \delta \Rightarrow P(E[X] \in \bar{X}_n \pm \frac{Z\sigma}{\sqrt{n}}) \approx 1 - \delta$$

But we don't know σ, but it can be estimated from sample standard deviation S_n. This is because

$$\frac{S_n}{\sigma} \xrightarrow{P} 1 \quad \text{(LLN)}$$

$$\sqrt{n}\frac{\bar{X}_n - E[X]}{\sigma} \Rightarrow N(0,1) \quad \text{(CLT)}$$

Combine both result with Slutsky theorem to have $\sqrt{n}\frac{\bar{X}_n - E[X]}{S_n} \Rightarrow N(0,1)$. So the key takeaway is that the performance of Monte Carlo method relies on the variance

2.14.3 Variance Reduction Method: Importance Sampling

Motivation

Let f be the pdf of X and let h be the pdf of another distribution Z. We have

$$E_X[X] = \int_{\mathbb{R}} x f(x)\, dx = \int_{\mathbb{R}} x \frac{f(x)}{h(x)} h(x)\, dx = E_Z[z * \frac{f(z)}{h(z)}]$$

Now we transformed from sampling from X to sampling from Z. We can carefuly choose h to minimize the variance. Conceptually, we want to choose the sample from part of the domain where $xf(x)$ is large.

Formalization
Let Y be a random variable (so it maps $\omega \in \Omega$ to $\mathbb{R}$. P is the underlying distribution of Y.

$$\begin{aligned} E_P[Y] &= \int_\Omega y(\omega)p(\omega)\,d\omega \\ &= \int_\Omega y(\omega)P(d\omega) && \text{(Notation change)} \\ &= \int_\Omega y(\omega)\frac{P(d\omega)}{Q(d\omega)}Q(d\omega) \\ &= E_Q[YL] && \text{(L is likelihood ratio)} \end{aligned}$$

Prerequisite
Notice that the if we can find an x such that $Q(x) = 0$ but $P(x) \neq 0$, then we are in trouble. So the support of P has to be the same or contained within the support of Q. This is called $P << Q$. If this condition is satisfied, then the likelihood ratio always exists by Radon–Nikodym theorem.

Picking Optimal Q given P
Let's say we want to get $a = E_P[Y]$, we can define an optimal Q^* as

$$\begin{aligned} Q^*(d\omega) &= \frac{|Y(\omega)|\,P(d\omega)}{\int_\Omega |Y(\omega)|\,P(d\omega)} \\ &= \frac{Y(\omega)P(d\omega)}{a} && \text{(Assume } Y > 0\text{)} \end{aligned}$$

Notice that when we sample from Q^*, the r.v. returned by the algorithm is

$$(YL)(\omega) = (Y(\omega)\frac{P(d\omega)a}{Y(\omega)P(d\omega)} = a$$

This would mean that it can get the exact quantity a with 0 variance. But notice that Q^* has a in it, so it is impossible to get Q^*.

Chapter 3

Statistics

3.1 Frequentist Parametric Estimators

3.1.1 Maximum Likelihood

Estimate the parameter $\hat{\theta}$ via the value that maximize the likelihood of observing the dataset. Generally we ended up maximizing the log likelihood.

Example: MLE for Uniform Distribution Assume the data are drawn from uniform distribution in $[0, \theta]$

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} \frac{\mathbb{I}\{0 \leq x_i \leq \theta\}}{\theta} \\ &= \mathbb{I}\{\theta \geq \max X_i\} * \frac{1}{\theta^n} \end{aligned}$$

So the θ that minimze the likelihood is the minimum θ which satisfies the indicator variable. Hence $\hat{\theta} = \max X_i$. Notice that this particular estimator actually converges in n to exponential distribution instead of the slower $\sqrt{n}$ with CLT.

3.1.2 Method of Moments

The general method of moments is pick $k(x)$ such that $E_{\theta_1}[k(x)] = E_{\theta_2}[k(x)] \iff \theta_1 = \theta_2$. Then we have

$$\frac{1}{n} \sum k(x_i) \approx E_\theta[k(x)]$$

So we use the theoretical result of the first t moments and equalizing them to the empirical sample moments in order to solve for θ.

3.1.3 Estimating Equations

Estimating equation generalizes both Method of Moments and MLE.

$$E_{\theta_1}[g(\theta_2, X)] = 0 \text{ iff } \theta_1 = \theta 2$$

So then we can estimate θ_1 with the root $\hat{\theta}$ of the equation:

$$\frac{1}{n}\sum_{i=1}^{n} g(\hat{\theta}, X_i) = 0$$

In MLE, we have

$$g(\theta, x) = \frac{\nabla_\theta f(\theta, x)}{f(\theta, x}$$

where as in MoM, we have

$$g(\theta, x) = E_\theta[k(X)] - k(x)$$

3.1.3.1 Large Sample Behavior of Estimator

In estimating equation framework, assuming consistency, i.e.: $\hat{\theta}_n \xrightarrow{P} \theta^*$ the estimator is a root of:

$$\frac{1}{n}\sum g(\hat{\theta}_n, x_i) = 0$$

So now we can expand to look at the convergence

$$\frac{1}{n}\sum \Big(g(\hat{\theta}_n, x_i) - g(\theta^*, x_i)\Big) = -\frac{1}{n}\sum g(\theta^*, x_i)$$

$$\frac{1}{n}\sum g'(\xi_n, x_i)(\hat{\theta}_n - \theta^*) = -\frac{1}{n}\sum g(\theta^*, x_i) \qquad \text{(By mean value theorem)}$$

$$\frac{1}{n}\sum g'(\xi_n, x_i)\sqrt{n}(\hat{\theta}_n - \theta^*) = -\frac{1}{\sqrt{n}}\sum g(\theta^*, x_i)$$

$$\frac{1}{n}\sum g'(\xi_n, x_i)\sqrt{n}(\hat{\theta}_n - \theta^*) = -\frac{1}{\sqrt{n}}\Big(\sum g(\theta^*, x_i) - n * E[g(\theta^*, x_i)]\Big)$$

(Expectation = 0 due to estimating equation setup)

$$\frac{1}{n}\sum g'(\xi_n, x_i)\sqrt{n}(\hat{\theta}_n - \theta^*) = N(0, Var_{\theta^*}(g(\theta^*, x_i)) \qquad \text{(By CLT)}$$

$$\frac{1}{n}\sum g'(\xi_n, x_i)\sqrt{n}(\hat{\theta}_n - \theta^*) = N(0, E_{\theta^*}[g(\theta^*, X_i)^2]$$

Now couple observations:

1. $\frac{1}{n}\sum g'(\xi_n, x_i) \xrightarrow{P} E_{\theta^*}[g'(\theta^*, X_1)]$ (shown later)
2. $\hat{\theta}_n - \theta^* \xrightarrow{P} 0$ by consistency

3. The RHS converged to a normal random variable

So we can use Slutsky's algorithm to actually establish the convergence

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta^*) &\Rightarrow \frac{N(0, E_{\theta^*}[g(\theta^*, X_i)^2]}{E_{\theta^*}[g'(\theta^*, X_i)]} \\ &\stackrel{D}{=} N\left(0, \frac{E_{\theta^*}[g(\theta^*, X_i)^2]}{E_{\theta^*}[g'(\theta^*, X_i)]^2}\right)\end{aligned}$$

Therefore, we can establish the confidence interval for estimating equation estimator $[\hat{\theta}_n - 2\frac{\hat{S}_n}{\sqrt{n}}, \hat{\theta}_n + 2\frac{\hat{S}_n}{\sqrt{n}}]$ where

$$\begin{aligned}\hat{S}_n &= \sqrt{\frac{E_{\theta^*}[g(\theta^*, X_i)^2]}{E_{\theta^*}[g'(\theta^*, X_i)]^2}} \\ &\approx \sqrt{\frac{\frac{1}{n}\sum g(\hat{\theta}_n, x_i)^2}{(\frac{1}{n}\sum g'(\hat{\theta}_n, x_i))^2}}\end{aligned}$$

Now just the last bits of clean up: Proving $\frac{1}{n}\sum g'(\xi_n, x_i) \xrightarrow{P} E_{\theta^*}[g'(\theta^*, X_1)]$. Notice that $g'(\xi_n, x_i)$is not iid becuase ξ_n is constructed based on $\hat{\theta}_n$, which is dependent on all x_i. We can do something similar by subtracting certain terms from both side. i.e.: We want to show

$$\begin{aligned}&\frac{1}{n}\sum g'(\xi_n, x_i) - \frac{1}{n}\sum g'(\theta^*, x_i) \xrightarrow{P} E_{\theta^*}[g'(\theta^*, X_1)] - \frac{1}{n}\sum g'(\theta^*, x_i) \\ &\frac{1}{n}\sum g''(\beta_n, x_i)(\xi_n - \theta^*) \xrightarrow{P} E_{\theta^*}[g'(\theta^*, X_1)] - \frac{1}{n}\sum g'(\theta^*, x_i) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By mean value theorem)}\end{aligned}$$

Notice that the right hand side converges to 0 by SLLN. So now we just need to show that the left hand side also converges to 0. Notice that β_n is between $\hat{\theta}_n$ and ξ_n. If we assume that the second derivative is global bounded, hence $\sup_\theta |g''(\theta, x)| \leq s(x)$, then we have

$$\left|\frac{1}{n}\sum g''(\beta_n, x_i)(\xi_n - \theta^*)\right| \leq \frac{1}{n}\sum s(x_i)\,|\xi_n - \theta^*|$$

Notice on the right hand side, the first part converges to $E_{\hat{\theta}_n}[s(X_i)]$ and the second part converges to 0. So by Slutsky the product converge to 0.

3.1.3.2 Asymptotic Variance of MLE in Estimating Equation Format

In MLE formulation of Estimating equation, we already have

$$
\begin{aligned}
&E_{\theta^*}[\frac{f'(\hat{\theta}_n, x_1)}{f(\hat{\theta}_n, x_1)}] = 0 \\
&g(\theta, x) = \frac{f'(\theta, x_1)}{f(\theta, x_1)} \\
&g'(\theta, x) = \frac{f''(\theta, x)}{f(\theta, x)} - (\frac{f'(\theta, x)}{f(\theta, x}))^2 \\
&E_{\theta^*}[g(\theta^*, X_1)^2] = E[\left(\frac{f'(\theta, x_1)}{f(\theta, x_1)}\right)^2]
\end{aligned}
$$

First we simplify $E[g'(\theta, x)]$:

$$
\begin{aligned}
E_{\theta^*}[g'(\theta, x)] &= E_{\theta^*}[\frac{f''(\theta, x)}{f(\theta, x)} - (\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= E_{\theta^*}[\frac{f''(\theta, x)}{f(\theta, x)}] - E[(\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= \int_{\mathbb{R}} \frac{f''(\theta, x)}{f(\theta, x)} f(\theta, x)\, dx - E[(\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} f(\theta, x)\, dx - E[(\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} \int_{\mathbb{R}} f(\theta, x)\, dx - E[(\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= \frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta} 1 - E[(\frac{f'(\theta, x)}{f(\theta, x)})^2] \\
&= -E[(\frac{f'(\theta, x)}{f(\theta, x)})^2]
\end{aligned}
$$

So now we can start to simplify the previous variance $\frac{E_{\theta^*}[g(\theta^*, X_i)^2]}{E_{\theta^*}[g'(\theta^*, X_i)]^2}$:

$$
\begin{aligned}
\frac{E_{\theta^*}[g(\theta^*, X_i)^2]}{E_{\theta^*}[g'(\theta^*, X_i)]^2} &= E[\left(\frac{f'(\theta, x_1)}{f(\theta, x_1)}\right)^2] * \frac{1}{E[(\frac{f'(\theta,x)}{f(\theta,x)})^2]^2} \\
&= \frac{1}{E[(\frac{f'(\theta,x)}{f(\theta,x)})^2]}
\end{aligned}
$$

Hence the asymptotic variance $\hat{\theta}_n$ for MLE is

$$
\frac{1}{n}\frac{1}{E_{\theta^*}\left[\left(\frac{f'(\theta,x)}{f(\theta,x)}\right)^2\right]}
$$

3.1.4 Cramer-Rao Bound

Cramer-Rao bound establishes that MLE is the best estimator for unbiased estimator. Let $\hat{\theta}_n = w_n(X_1, ..., X_n)$ and assume that the estimator is unbiased at every point $\theta \in \Lambda$. We have

$$\theta = E_\theta[w_n(X_1, ..., X_n)] = \int_{\mathbb{R}^n} w_n(x_1, ..., x_n) \prod_{i=1}^{n} f(\theta, x_i)\, dx_1, ...dx_n$$

$$\implies \int_{\mathbb{R}^n} w_n(x_1, ..., x_n) \sum_{i=1}^{n} f'(\theta, x_i) \prod_{j \neq i} f(\theta, x_j)\, dx_1, ...dx_n = 1$$

(Chain rule and derivative both side)

$$\implies \int_{\mathbb{R}^n} w_n(x_1, ..., x_n) \sum_{i=1}^{n} \frac{f'(\theta, x_i)}{f(\theta, x_i)} * \prod_{j=1}^{n} f(\theta, x_j)\, dx_1, ...dx_n = 1$$

$$\implies E_\theta \left[w_n(X_1, ..., X_n) \sum_{i=1}^{n} \frac{f'(\theta, X_i)}{f(\theta, X_i)} \right] = 1$$

$$A \equiv w_n(X_1, ..., X_n)$$

$$B \equiv \sum_{i=1}^{n} \frac{f'(\theta, X_i)}{f(\theta, X_i)}$$

Notice that we care about the variance of our estimator, hence $E[A^2] = E_\theta[w_n(X_1, ..., X_n)^2]$. By Cauchy Schwartz we have

$$\sqrt{E[A^2]E[B^2]} \geq |E[AB]| = 1$$

Notice that

$$\begin{aligned} E_[B^2] &= E\left[\left(\sum_{i=1}^{n} \frac{f'(\theta, X_i)}{f(\theta, X_i)} \right)^2 \right] \\ &= Var\left(\sum_{i=1}^{n} \frac{f'(\theta, X_i)}{f(\theta, X_i)} \right) \qquad \text{(Mean 0)} \\ &= n * Var\left(\frac{f'(\theta, X_1)}{f(\theta, X_1)} \right) \\ &= n * E\left[\left(\frac{f'(\theta, X_1)}{f(\theta, X_1)} \right)^2 \right] \end{aligned}$$

So we have

$$E_\theta[\hat{\theta}^2] = E[A^2] \geq \frac{1}{n * E\left[\left(\frac{f'(\theta, X_1)}{f(\theta, X_1)} \right)^2 \right]}$$

We can look at the connection to fisher information. Individual variable in B can be seen as the score function

$$\begin{aligned} E[\partial\theta \log f(X,\theta)] &= \int_{\mathbb{R}} \frac{f'(x,\theta}{f(x,\theta)} f(x,\theta)\, dx \\ &= \int_{\mathbb{R}} \frac{\partial}{\partial\theta} f(x,\theta)\, dx \\ &= 0 \end{aligned}$$

and the variance is the of the score, which is $E\left[\left(\frac{f'(\theta,X_1)}{f(\theta,X_1)}\right)^2\right]$.

3.2 Frequentist Non-parametric Perspective

3.2.1 Setup and Glivenko–Cantelli theorem

In frequentist non-parametric perspective, we have X_i iid from an unknown distribution F. We attach mass $\frac{1}{n}$ to each point X_i. So the empirical CDF is defined as

$$F_n(x) = \frac{1}{n} \sum \mathbb{I}\{x_i \leq x\}$$

From SLLN, we know that as $n \to \infty$, $\forall x \in \mathbb{R}$ (note that below theorem is for a given x), we have

$$F_n(x) \overset{a.s.}{\to} F(x)$$

But Glivenko-Cantelli Theorem provides a stronger point that says

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{a.s.}{\to} 0$$

The proof can be found here.

Another restatement of GC theorem is that

$$\begin{aligned} &\sup_{h\in S} \left| \int h(x) F_n(dx) - \int h(x) F(dx) \right| \overset{a.s.}{\to} 0 \\ &S = \{h : h(y) = \mathbb{I}\{y \leq x\} \text{ for some } x \in \mathbb{R}\} \end{aligned}$$

Notice that this result also applies to a broader function space (GC class), but not to every function.

3.2.2 Plug-in Principle

For any fixed function h, we have that

$$\int h(x) F_n(dx) = \frac{1}{n} \sum h(x_i) \to \int h(x) F(dx)$$

So if we want to estimate a function of the distribution $T(F)$, we can just use the empirical distribution $T(F_n)$. So if we want to calculate the median of F, we can esimate it with sample median.

$$
\begin{aligned}
&T(F) = \int_{\mathbb{R}} h(x)F(dx) = E[h(x)] \\
&T(F_n) = \frac{1}{n}\sum h(x_i) \\
&\sqrt{n}(T(F_n) - T(F)) \Rightarrow \sqrt{Var(h(x))}N(0,1) \qquad \text{(CLT)} \\
&\Rightarrow CI : [T(F_n) \pm Z\frac{\sigma(F_n)}{\sqrt{n}}]
\end{aligned}
$$

3.2.3 Skorokhod space

A Skorokhod space e.g.: $D(-\infty, \infty)$ is a space of all functions Z such that

- Z is right continuous: $\lim_{t\downarrow s} z(t) = z(s)$
- Z has left limit wverywhere: $\lim_{t\uparrow s} z(t) = z(s-)$ exists

Z is defined on the interval of the real line (e.g.: $(-\infty, \infty)$), and takes values on the real line or some metric space.

3.2.4 Alternative Definition of Weak Convergence

An alternative definition of weak convergence is $Y_n \Rightarrow Y_\infty$ in $\mathbb{R}$ iff:

$$E[g(Y_n)] \to E[g(Y_\infty]$$

for all g that are bounded converges. This allows us to think about convergence of random variable when they are in a more complicated space such as the Skorokhod space.

3.2.5 Convergence in abstract metric space

Given a metric space S and a distance measure $d : S \times S \to \mathbb{R}$. The distance d has to satisfies:

- $d(x, x) = 0$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(z, y)$

then we have

$$y_n \to y_\infty \text{ iff } d(y_n, y_\infty) \to 0 \text{ as } n \to \infty$$

3.2.6 Completeness

Cauchy Sequence is defined as :

$$\forall \epsilon > 0\,, \exists n = n(\epsilon)\, s.t.\, m_1, m_2 \geq n(\epsilon) \Longrightarrow |X_{m_1} - X_{m_2}| < \epsilon$$

The completeness of real line means that Cauchy sequence on real line always converges to an element in real. Equivalently convergence sequence in real line has to be a Cauchy Sequence. This definition of completeness can be extended to abstract metric space S by replacing the definition of $-$ in Cauchy sequence with general distance measurement.

3.2.7 Separability

Metric space S is separable iff a countable set of element in S can be used to approximate all element in S (have a dense countable subset).

3.2.8 Convergence in Polish Space

Let metric space S with measurement d be a complete separable metric space (Polish Space), then

$$Y_n \Rightarrow Y_\infty \text{ in } (s, d)$$

iff

$$E[g(Y_n)] \to E[g(Y_\infty)] \forall g \in b.c.(\text{bounded converges})$$

iff

$$\begin{aligned} &\exists \text{ a probability space supporting sequence } Y_n\text{' s.t. } Y_n' \overset{D}{=} Y_n \\ &\text{and } d(Y_n', Y_\infty) \to 0\, a.s. \text{ as } n \to \infty \end{aligned}$$

3.2.9 Skorokhod's Representation

Let μ_n be a sequence of probability measures on a metric space S such that μ_n converges weakly to some probability measure μ_∞ on S as $n \to \infty$. Suppose that the support of μ_∞ is separable. Then there exists S-valued random variables X_n defined a a common probability space such that the law of X_n is μ_n for all n and $X_n \to X_\infty$ almost surely.

3.2.10 Brownian Motion

A standard Wiener process(Brownian Motion) is a stochastic process $\{W_t\}$ indexed by nonnegative real numbers t with the following propertyies:

- $W_0 = 0$
- With probability 1, the function $t \to W_t$ is continuous in t

- Stationary independent increment
- $W_{t+s} - W_s$ has $N(0,t)$ distribution

Some result include that if $X_n \Rightarrow X_\infty$, then it must be that X_∞ is Gaussian, $E[X_\infty] = 0$ with finite variance $\sigma^2 t$.

Another result is that

3.2.11 Donsker's Theorem

Donsker's theorem is a function extension of the CLT. Let X_i be iid r.v. with mean 0 and variance 1. Define the diffusively rescaled random walk as

$$W^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \quad t \in [0,1]$$
$$S_n := \sum_{i=1}^{n} X_i$$

CLT: $W^{(n)}(1)$ converges in distribution to standard Gaussian r.v. $W(1)$ as $n \to \infty$. Donsker's invariance principle extends this converges to the whole function.

First notice that S_n and $W^{(n)}(t)$ both have independent increment property (i.e.: $S_{m+n} - S_m$ is independent from $S_m - S_0$). $W^{(n)}(t) \in D[0,\infty)$ due to the floor function.

Donsker's provides convergence in the Skorokhod space.

$$W^{(n)}(t) \Rightarrow W_\infty \text{ in } (D[(0,\infty), d)$$
$$W^{(n)}(t) \Rightarrow \sigma B \qquad \text{(Standard Brownian Motion)}$$

The Skorokhod representation further says that there exists

$$W'^{(n)} \overset{a.s.}{\to} \sigma B'$$
$$d(W'^{(n)}, \sigma B') \overset{a.s.}{\to} 0$$

For $y \in D[0,\infty)$, let $h(y) = \max_{0 \le t \le 1} y(y)$, then we have

$$y(W^{(n)}(t)) \Rightarrow y(\sigma B)$$

3.2.12 Application: Quantile Estimation

Estimate q such that $F(q) = p$ with empirical estimator $F_n(Q_n) = p$. Note that $\sqrt{n}(F_n(x) - F(x)) = X_n \Rightarrow Z$ if x is in finite dimension, in infinite dimension, we know that $X_n \Rightarrow Z$ in D space (because Q_n) is discontinuous.

$$\sqrt{n}\left((F_n(Q_n) - F(Q_n)) - (F_n(q) - F(q)) + (F(Q_n) - F(q)) + (F_n(q) - F(q))\right) = \sqrt{n}(F_n(Q_n) - F(q))$$

Observation:

$F_n(Q_n) - F(Q_n) = X_n(Q_n)$
$(F_n(q) - F(q)) = X_n(q)$
$(F(Q_n) - F(q)) = (Q_n - q)f(q) + O(f''(q)(Q_n - q)^2)$ (Taylor Expansion)
$F_n(q) - F(q)$can use CLT
$F_n(Q_n)$ is $O(\frac{1}{n})p$

So we have

$$\sqrt{n}(X_n(Q_n) - X_n(q)) + \sqrt{n}(Q_n - q)f(q) + \sqrt{F(q)(1 - F(q))}N(0,1) = 0$$

So now if we can show $X_n(Q_n) - X_n(q) \xrightarrow{P} 0$, then we can apply Slutsky Theorem and have

$$Q_n - q \Rightarrow \frac{\sqrt{p(1-p)}}{f(q)}N(0,1)$$

We can show $X_n(Q_n) - X_n(q) \xrightarrow{P} 0$ by the following argument: Assume F is continuous everywhere, from Densker we know that $X_n \Rightarrow Z$ in D space, and subsequently have uniform convergence in a compact set. From Skorhod Representation we know that there exists

$$d(X'_n, Z') \to 0 \text{ a.s.}$$

So we have

$(X_n, Q_n) \Rightarrow (Z, q)$ (From Slutksy in multiple dimension)
$(X'_n, Q'_n) \to (Z', q)$ a.s.
$X'_n(Q'_n) \to Z'(q)$
$X'_n(q) \to Z'(q)$
$\Rightarrow X'_n(Q'_n) - X'_n(q) \to 0 a.s.$
$\Rightarrow X_n(Q_n) - X_n(q) \xrightarrow{P} 0$

3.2.13 Kolmogorov-Smirnov Test

K.S. test whether $X_1, ..., X_n$ is iid from F.

Intuition How big is $\sup_{x\in\mathbb{R}} \sqrt{n}\,|F_n(x) - F(x)| = \sup_{x\in\mathbb{R}} |X_n(x)| \Rightarrow \sup_{x\in\mathbb{R}} |Z(x)|$. Then we can set up confidence interval through $P(\sup_x |Z(x)| > w) = 0.01$. All we need to know is the limit distribution Z (Densker only gives us its existence).

Equivalent Turns out that if Z is the limit associate with F, let $\tilde{Z}$ be the limit associate with uniform $[0,1]$, then

$$Z(x) = \tilde{Z}(F(x)) \quad \forall x$$

So we have

$$\sup_x |Z(x)| \stackrel{D}{=} \sup_x \left|\tilde{Z}(F(x))\right| = \sup_{y\in[0,1]} \left|\tilde{Z}(y)\right|$$

To summarize, as long as we can find the limit distribution $\tilde{Z}$ for uniform iid, we can apply KS test for every distribution. Turns out that such $\tilde{Z}$ has a closed form and can be easily computed.

3.2.14 Estimating Density Through Kernel Density Estimation

3.2.14.1 Formulation

Let X_i be iid from unknown distribution. We can estimate the density by put a continuous distribution (e.g.: $N(x_i, h_n^2)$) around each point x_i.

$$\begin{aligned} P(N(x_i,h^2) \le y) &= P(x_i + h * N(0,1) \le y) \\ &= P(N(0,1) \le \frac{y-x_i}{h}) \\ &= \Phi(\frac{y-x_i}{h}) \\ \frac{\partial}{\partial y}\Phi(\frac{y-x_i}{h}) &= \frac{1}{h}\phi(\frac{y-x_i}{h}) \end{aligned}$$

The final density is the weighted sum of the density:

$$f_n(y) = \frac{1}{h}\frac{1}{n}\sum_{i=1}^{n}\phi(\frac{y-x_i}{h})$$

3.2.14.2 Convergence

In general, we can decompose MSE:

$$\begin{aligned} E[f_n(y) - f(y)]^2 &= E\left[f_n(y) - E[f_n(y)] + E[f_n(y)] - f(y)\right]^2 \\ &= E[f_n(y) - E[f_n(y)]]^2 + E[E[f_n(y)] - f(y)]^2 + 2 * E[f_n(y) - E[f_n(y)]]E[E[f_n(y)] - f(y)] \\ &= E[f_n(y) - E[f_n(y)]]^2 + E[E[f_n(y)] - f(y)]^2 \\ &= Var(f_n(y)) + Bias^2 \end{aligned}$$

The expected value of the estimator is :

$$\begin{aligned}
E[f_n(y)] &= E[\phi(\frac{y-x_1}{h})\frac{1}{h}] \\
&= \int_{-\infty}^{\infty} \phi(\frac{y-x}{h})\frac{1}{h} f(x)\,dx \\
&= \int_{-\infty}^{\infty} \phi(z) f(y-zh)\,dz && (z = \tfrac{y-x}{h}) \\
&= \int_{-\infty}^{\infty} \phi(z)\left(f(y) - zhf'(y) + \frac{z^2h^2}{2}f''(y) + o(h^2)\right) dz \\
& && \text{(Taylor Expansion)} \\
&= f(y)\int_{-\infty}^{\infty} \phi(z)\,dz - f'(y)h\int_{-\infty}^{\infty} z\phi(z)\,dz + \frac{f''(y)h^2}{2}\int_{-\infty}^{\infty} z^2\phi(z)\,dz + o(h^2) \\
&= f(y) - f'(y)h\int_{-\infty}^{\infty} z\phi(z)\,dz + \frac{f''(y)h^2}{2}\int_{-\infty}^{\infty} z^2\phi(z)\,dz + o(h^2) \\
& && \text{(Integral of a pdf)} \\
&= f(y) + \frac{f''(y)h^2}{2}\int_{-\infty}^{\infty} z^2\phi(z)\,dz + o(h^2) && \text{(Property of } \phi) \\
&= f(y) + \frac{h^2}{2}f''(y) + o(h^2) && \text{(Property of } \phi)
\end{aligned}$$

So the intuition is that we want to have small h, as little smoothing as possible. The variance of the estimator is:

$$\begin{aligned}
\frac{1}{n}Var\left(\frac{1}{h}\phi(\frac{y-x_1}{h})\right) &= \frac{1}{nh^2}Var\left(\phi(\frac{y-x_1}{h})\right) \\
&= \frac{1}{nh^2}\left(E[\phi^2(\frac{y-x_1}{h})] - E[\phi(\frac{y-x_1}{h})]^2\right) \\
&= \frac{1}{nh^2}\left(\int_{-\infty}^{\infty}\phi^2(\frac{y-x}{h})f(x)\,dx - E[\phi(\frac{y-x_1}{h})]^2\right) \\
&= \frac{1}{n}\left(\frac{1}{h}\int_{-\infty}^{\infty}\phi^2(z)f(y-zh)\,dz - \frac{1}{h}E[\phi(\frac{y-x_1}{h})]^2\right) \\
&= \frac{1}{n}(\frac{1}{h}O(1) + O(h)) \\
&\approx \frac{1}{nh} \text{ as } h \to 0
\end{aligned}$$

So the variance is $O(\frac{1}{nh})$, which means we want to have h big, as much smoothing as possible.

3.2.14.3 Bias Variance Tradeoff

Given above result, we can set the bias squared equal to variance to get the optimal h

$$\frac{1}{nh} \approx h^4 \Longrightarrow h \approx n^{-1/5}$$

The optimal RMSE under this condition is

$$\frac{1}{nh} + h^4 = \sqrt{n^{-4/5}} = n^{-2/5}$$

The optimal h will give us CLT $n^{2/5}(f_n(y) - f(y)) \Rightarrow N(a, b^2)$. In reality, bias is hard to estimate and the variance is easier to estimate. So we might want to set $h > n^{-1/5}$ to have a low bias high variance estimator.

3.2.15 Estimating Mode

Mode is defined as

$$m = \underset{x}{\operatorname{argmax}}\, pdf(x)$$

The intuition is to estimate $f'(x)$ through:

$$\begin{aligned} &\frac{f_n(y + r_n) - f_n(y)}{r_n} \\ &h_n \to 0 && \text{(Used for estimate } f_n\text{)} \\ &r_n \to 0 \end{aligned}$$

But notice tha t as $r_n \to 0$, the variance of this estimator gets big. So we need to let $h_n \to 0$ slower compare to the previous optimal h. In general, as we estimate high order derivative of pdf, we depend more on local structure and requires smoother curve, i.e.: bigger h.

3.2.16 CDF Estimation: Integrating PDF vs. Plug In

So far we have two CDF estimator:

$$\begin{aligned} F_n(y) &= \frac{1}{n}\sum \mathbb{I}\{x_i \le y\} && \text{(Plug In Estimator } h = 0\text{)} \\ \hat{F}_n(y) &= \int_{-\infty}^{y} f_n(z)\, dz && \text{(pdf Estimator } h \propto n^{-1/5}\text{)} \end{aligned}$$

Here we demonstrate why the plug in estimator is better:

$$\begin{aligned} \hat{F}_n(y) &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{y} \phi(\frac{z - x_i}{h})\frac{1}{h}\, dz \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{y - x_i}{h}) \end{aligned}$$

Notice that as $h \to 0$, $\Phi(\cdot)$ will either goes to 1 (if $x_i < y$) or to 0 otherwise. So the function is bounded and we can apply bounded convergence theorem:

$$E[\Phi(\frac{y-x_i}{h})] \to E[\mathbb{I}\{x_i \le y\}] = P(X_i \le y)$$
$$E[\Phi^2(\frac{y-x_i}{h})] \to P(X_i \le y) \qquad \text{(By the same 0, 1 argument)}$$

We can then look at variance of the estimator:

$$\begin{aligned} Var(F_n(y)) &= \frac{1}{n} Var(\mathbb{I}\{X_i \le y\}) \\ Var(\hat{F}_n(y)) &= \frac{1}{n} Var(\Phi(\frac{y-x_i}{h})) \\ &\to \frac{1}{n}(P(X_i \le y) - P^2(X_i \le y)) \end{aligned}$$

Now for bias:

$$\begin{aligned} E[\hat{F}_n(y) - F(y)] &= E[\Phi(\frac{y-x}{h})] - F(y) \\ &= \int_{-\infty}^{\infty} \Phi(\Phi(\frac{y-x}{h}) f(x)\, dx - F(y) \\ &= h \int_{-\infty}^{\infty} \Phi(z) f(y - zh)\, dz - F(y) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{0}^{\infty} \Phi(z) f(y - zh)\, dz \right) - F(y) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{0}^{\infty} (1 - \bar{\Phi}(z)) f(y - zh)\, dz \right) - F(y) \\ &\qquad (\bar{\Phi}(x) \equiv 1 - \Phi(x)) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz - \int_{y}^{-\infty} (1 - \bar{\Phi}(\frac{y-\omega}{h})) f(\omega) \frac{1}{h}\, d\omega \right) - F(y) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{y}^{-\infty} \bar{\Phi}(\frac{y-\omega}{h}) f(\omega) \frac{1}{h}\, d\omega \right) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{0}^{\infty} \bar{\Phi}(z) f(y - zh)\, dz \right) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{0}^{-\infty} \bar{\Phi}(-t) f(y + th) * -1\, dt \right) \\ &= h \left(\int_{-\infty}^{0} \Phi(z) f(y - zh)\, dz + \int_{0}^{-\infty} \Phi(t) f(y + th)\, dt \right) \\ &\qquad (\Phi(\bar{-}x) = \Phi(x)) \\ &= h \left(\int_{-\infty}^{0} \Phi(z)(f(y - zh) - f(y + zh))\, dz \right) \end{aligned}$$

So bias squared is in $h^2 = n^{-2/5}$ term, and variance is in $\frac{1}{n}$ term.

3.2.17 Multidimensional density estimation with Multivariate Normal Smoother

Result: bias $O(h^2)$ and variance $O(\frac{c}{nh^d})$ where d is the dimension. RMSE is $O(n^{-2/(d+4)})$. So as d increases, we will have less and less convergence.

3.3 Bayesian Statistics

3.3.1 General Formulation

$$f(p|data) = \frac{p(data|p) * f(p)}{p(data)}$$

On the top is the data's likelihood times the prior. The bottom is the likelihood of the data times prior integrated over entire support.

3.3.2 Example

Setup: Bernoulli trial with result 1, 1, 1, 1. Let $f(p) = unif[0, 1]$ to be the prior (uninformative prior). The posterior distribution is :

$$\begin{aligned} f(p|data) &= \frac{p(data|p) * p(p)}{p(data)} \\ &= \frac{p^4 * 1}{\int_0^1 1 * p^4 \, dp} \\ &= 5p^4 \end{aligned}$$

So the posterior mode is

$$\underset{p}{\operatorname{argmax}} \, 5p^4 = 1$$

The posterior mean is

$$\int_0^1 p * f(p|data) \, dp = \int_0^1 5p^5 \, dp = \frac{5}{6}$$

3.3.3 Beta Distribution as Conjugate Prior for Binomial Distribution

Beta distribution with parameter α, β is defined as:

$$f(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Beta(\alpha, \beta)}$$

If an experiment has k success and $n - k$ failure. The likelihood of the data is:

$$p^k(1-p)^{n-k}$$

The posterior can always be looked at from proportion perspective (since $p(data)$ is a constant):

$$\begin{aligned} f(p|data) &\propto p(data|p) * f(p) \\ &= p^k(1-p)^{n-k}p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha+k-1}(1-p)^{\beta+n-k-1} \end{aligned}$$

So the posterior distribution is also a new beta density with $\alpha' = \alpha + k$ and $\beta' = \beta + n - k$

3.3.4 Gaussian Distribution as Conjugate Prior for Gaussian

Let X_i be iid from $N(\mu^*, 1)$, and assume the prior distribution for μ^* to be $N(0, 1)$. The posterior is

$$\begin{aligned} f(\mu|data) &\propto f(data|\mu) * f(\mu) \\ &= (\frac{1}{\sqrt{2\pi}})^n \exp(-\sum \frac{(x_i-\mu)^2}{2}) \frac{1}{\sqrt{2\pi}} \exp(-\mu^2/2) \\ &\propto \exp(-\sum \frac{(x_i-\mu)^2}{2}) \exp(-\mu^2/2) \\ &= \exp(-\frac{1}{2}(\mu^2 + \sum x_i^2 + n*\mu^2 - 2\mu \sum x_i)) \\ &\propto \exp(-\frac{1}{2}((n+1)\mu^2 - 2\mu \sum x_i)) \\ &= \exp(-\frac{n+1}{2}(\mu^2 - \frac{2\mu \sum x_i)}{n+1}) \\ &\propto \exp(-\frac{n+1}{2}(\mu - \frac{\sum x_i}{n+1})^2) \end{aligned}$$

So this is a new Gaussian distribution with $\mu' = \frac{\sum x_i}{n+1}$ and $\sigma' = \frac{1}{n+1}$

3.3.5 Gamma Distribution as Conjugate Prior

Overall result: If X_i are iid from $Gamma(\lambda, \alpha)$, then there exists a conjugate prior on λ^* which is Gamma distribution.

3.3.6 Bayesian Regression

3.3.6.1 Non-Bayesian Regression Formulation

Regression normally has the form of

$$Y_i = \alpha_0 x_0 + \alpha_1 x_1 + ... + \alpha_d x_d + \epsilon_i$$

where $x_0 = 1$ and ϵ_i is iid $N(0, \sigma^2)$. The goal is to estimate α_i and σ. By setting the equation to be the form of $\epsilon_i = \cdot$, we can get Gaussian likelihood

$$L(\alpha_0, ..., \alpha_d, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} (\frac{y_i - \sum_{j=0}^{d} \alpha_j x_{ji}}{\sigma})^2 \right) \right)$$

$$l(\alpha_0, ..., \alpha_d, \sigma^2) = \log(\frac{1}{\sqrt{2\pi}\sigma}) - \sum_{i=0}^{n} \frac{1}{2} \left(\frac{y_i - \sum_{j=0}^{d} \alpha_j x_{ji}}{\sigma} \right)^2$$

Notice that the α_j that maximizes the log likelihood is equivalent to minimize $\sum_{i=0}^{n} (y_i - \sum_{j=0}^{d} \alpha_j x_{ji})^2$.
In matrix vector formulation, we let Y and a be a $n \times 1$ matrix, X is $n \times d$ matrix. So we have

$$Y = Xa + \Sigma \qquad (\Sigma \sim N(0, \sigma^2 C))$$

C in general is a known typically diagonal non-identity matrix

Then the MLE estimator is equivalent to minimize

$$(Y - Xa)^T C^{-1} (Y - Xa)$$

3.3.6.2 Bayesian Regression Formulation

In Bayesian formulation, we have

$$\begin{aligned} &Y = Xa + \Sigma \\ &\Sigma \sim N(0, \sigma^2 C) && (C \text{ is known, but } \sigma^2 \text{ is unknown}) \\ &a \sim N(O, I) && (\text{Gaussian Prior}) \end{aligned}$$

From this formulation, we know that the posterior is

$$f(\alpha | data) \propto \exp\left(-\frac{\sum_{j=0}^{d} \alpha_i^2}{2} - \frac{(Y - X\alpha)^T C^{-1} (Y - X\alpha)}{2\sigma^2} \right)$$

Notice that the first part is from $f(a)$, and the second part is from $p(data|a)$. So we can derive the mode of the posterior (the a that maximizes the probability) to be

$$\underset{\alpha}{\operatorname{argmin}} (Y - X\alpha)^T C^{-1} (Y - X\alpha) + \frac{1}{2} \sum_{j=0}^{d} \alpha_i^2$$

Hence Regression with Gaussian prior is equivalent to regression with L2 regularization.

If we use Laplace distribution as prior, such as $\frac{1}{2}e^{-|\alpha|}$, then the final mode is equivalent to L1 regularization.

$$\underset{\alpha}{\operatorname{argmin}}(Y - X\alpha)^T C^{-1}(Y - X\alpha) + \frac{1}{2}\sum_{j=0}^{d} |\alpha_i|$$

Chapter 4

Gaussian Modeling

4.1 Primer: First Order AR Model

Assume a first order AR model, we have

$$\begin{aligned} x_{n+1} &= \alpha_0 x_n + c + \epsilon_{n+1} \\ \epsilon &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

So given data point $X_0, ..., X_n$, the likelihood is

$$\prod_{j=0}^{n-1} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2}\left(\frac{x_{j+1} - \alpha_0 x_j - c}{\sigma}\right)^2\right)$$

The covariance is

$$\begin{aligned} Cov(X_{n+1}, X_n) &= Cov(\alpha_0 X_n + c + \epsilon_{n+1}, x_n) \\ &= Cov(\alpha_0 x_n, x_n) \\ &= \alpha_0 Var(X_n) \\ &\implies \alpha_0 = \frac{Cov(x_{n+1}, x_n)}{Var(x_n)} \\ &\implies \hat{\alpha}_0 = \frac{\frac{1}{n}\sum_{j=0}^{n-1}(x_j x_{j+1} - \bar{x}^2)}{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2} \end{aligned}$$

If we know α_i and c, then generating the next value is easy:

$$\begin{aligned} E[X_{n+1}|x_j : j \leq n] &= E[\alpha_0 x_n + ... + \alpha_p x_{n-p} + c + \epsilon_{n+1}|x_j : j \leq n] \\ &= \alpha_0 x_n + ... + \alpha_p x_{n-p} + c \end{aligned}$$

4.2 Gaussian Distribution

Multivariate Gaussian distribution pdf is:

$$\frac{1}{\sqrt{2\pi}^d \sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)\right)$$

4.3 Generating Multivariate Gaussian from Standard Multivariate Gaussian

One important property of multivariate is that it is closed under affine transformation:

$$\begin{aligned} &X \sim N(\mu, \Sigma) \\ &Y = AX + b \\ &\implies Y \overset{D}{=} N(A\mu + b, A\Sigma A^T) \end{aligned}$$

This means that if we want to generate any multivariate Gaussian r.v. $Y \sim N(\mu, C)$, we can:

1. Generate $X \sim N(0, I)$
2. Set $Y = \mu + AX$ where $AA^T = C$. (We can easily find such A through Cholesky factorization

Example of Cholesky Factorization

$$\begin{aligned} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} &\overset{D}{=} N(0, C) \\ C &\equiv \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\ \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} &\overset{D}{=} N(0, I) \end{aligned}$$

So we begin the factorization process by

$$\begin{aligned} X_1 &= \sigma_1 Z_1 \\ X_2 &= \alpha X_1 + bZ_2 \end{aligned}$$

So we have:

$$\begin{aligned} &\alpha\sigma_1^2 = Cov(X_1, \alpha X_1 + bZ_2) = Cov(X_1, X_2) = \rho\sigma_1\sigma_2 \\ &\implies \alpha = \frac{\rho\sigma_2}{\sigma_1} \end{aligned}$$

Then we can look at $Var(X_2)$:

$$\alpha^2\sigma_1^2 + b^2 = Cov(X_2, X_2) = Var(X_2) = \sigma_2^2$$
$$\implies b = \sqrt{1-\rho^2}\sigma_2$$

4.4 Generating Standard Multivariate Gaussian Random Variable

We have shown that we can generate any $Y = N(\mu, C)$ by generating a standard multivariate normal r.v. X and transform it via $Y = \mu + AX$ such that $AA^T = C$. So we need to a good algorithm to generate $X \sim N(0, I)$, or essentially generating scalar $N(0, 1)$. Turns out, we can generate $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N(0, I)$ from polar coordinates.

First, any pair of coordinate (x_1, x_2) from polar coodinates can be represented as $(R\cos\theta, R\sin\theta)$ and $\theta = unif[0, 2\pi]$, independent of R.
Second, notice that $R^2 = x_1^2 + x_2^2 = X^2$ r.v. with 2 degree of freedom, which is equivalent to $2 * Exp(1)$, which can be generated through inversion. Specifically, we can generate $U \sim uniform[0, 1]$ and then take $-2\log(1-u)$, which is equivalent to $-2\log(u)$.
So in summary, we can generate $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N(0, I)$ by:

1. Generuate u_1, u_2 from $uniform[0, 1]$ (1 for generate θ, and another for generate R.
2. Set $x_1 = \sqrt{-2\log(u_1)}\cos(2\pi u_2)$ and $x_2 = \sqrt{-2\log(u_1)}\sin(2\pi u_2)$

4.5 Structure of Conditional Distribution

In block matrix from, we have:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = N(\mu, C)$$
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$
$$C = \begin{bmatrix} E[\tilde{X}_1\tilde{X}_1^T] & E[\tilde{X}_1\tilde{X}_2^T] \\ E[\tilde{X}_2\tilde{X}_1^T & E[\tilde{X}_2\tilde{X}_2^T]] \end{bmatrix}$$
$$\tilde{X}_i = X_i - \mu_i$$

The conditional mean of $X_1|X_2 = x_2$ is :

$$\mu_1 + E[\tilde{X}_1\tilde{X}_2^T](E[\tilde{X}_2\tilde{X}_2^T])^{-1}(x_2 - \mu_2)$$

The conditional variance actually does not depend on x_2. This means that it can be pre-computed without actually seeing the data.

4.6 Gaussian Stochastic Process and Gaussian Random Field

Given Gaussian Stochastic Process $z(t), t \geq 0$ or Random Field $z(x), x \in \mathbb{R}^d$, we have below notation:

$$m(x) = E[z(x)]$$
$$c(x, y) = Cov(z(x), z(y))$$

4.6.1 Fixed Mean Stationary Isotropic Case

Add the following simplifying assumptions:

$m(x) = m$ (Fixed mean)
$c(x, y) = c'(x - y)$ (Stationary: cov only dependent on difference)
$c(z) = c'(\|z\|)$ (Isotropic: cov only dependent on Euclidean distance)
$\implies c(x, y) \equiv Cov(z(x), z(y)) = c(\|x - y\|)$

One good example of such case is Brownian model (note it is technically not stationary, but have stationary increment) :

$$B(t + h) - B(t) \stackrel{D}{=} B(h) - B(0) \stackrel{D}{=} N(0, h) \stackrel{D}{=} \sqrt{h}N(0, 1)$$

However Brownian motion is not differentiable. We can see this intuitively by looking at as $h \to 0$, the approximation doesn't converges to anything good.

$$\frac{B(t + h) - B(t)}{h} \stackrel{D}{=} \sqrt{h}N(0, 1)$$

But we could look at what it looks like for a general process that is smooth. So we want

$$\frac{z(x + h) - z(x)}{h} \Rightarrow z'(x)$$

One intuition to make such thing happen is to make the variance of the LHS and RHS equivalent.

$$\begin{aligned} Var(\frac{z(x + h) - z(x)}{h}) &= (var(z(x + h)) + var(z(h)) - 2cov(z(x), z(x + h))) * \frac{1}{h^2} \\ &= \frac{2}{h^2}(var(z(0)) - cov(z(0), z(h))) \\ &\qquad \text{(stationary property)} \\ &= \frac{2}{h^2}Var(z(0))(1 - \rho(h)) * \frac{1}{h^2} \end{aligned}$$

Notice that $\rho(h) = 1 + h\rho'(0) + \frac{h^2}{2}\rho''(0) + o(h^2)$. So if we want to converge, we need to have $\rho'(0) = 0$ since the outside is $\frac{1}{h^2}$ term. In addition, ρ has to have positive definite property in order to be a proper correlation function.

4.6.2 Principal Component of a Multivariate Gaussian

$$X \overset{D}{=} N(0, C) \qquad (C \text{ is positive definite symmetric})$$

We can show that there exists a basis of orthogonal eigenvectors for C:

$$C = U^T D U$$
$$U^T U = I$$

Let $Z = UX$, then

$$Z \overset{D}{=} N(0, UCU^T) \overset{D}{=} N(0, UU^T DUU^T) \overset{D}{=} N(0, D)$$

So we have

$$Cov(z_i, z_j) = \begin{cases} \lambda_i & i = j \\ 0 & i \neq j \end{cases}$$

and $X = U^T Z$ where Z is the principal component of X

4.6.3 Generalization of Principal Component to Gaussian Process

We have

$$X = (X(t) : 0 \leq t \leq 1)$$
$$E[X(t)] = 0$$
$$\int_0^1 E[X^2(t)]\, dt < \infty$$
$$c(s,t) = Cov(X(s), X(t))$$

Mercer's Theorem says that there exists eigen function $e_k(t)$ such that:

$$\int_0^1 c(s,t) e_k(t)\, dt = \lambda_k e_k(s), k \geq 1$$
$$\int_0^1 e_k(s) e_g(s)\, ds = \begin{cases} 1 & k = g \\ 0 & \text{otherwise} \end{cases}$$

Some quick notation

$$< f, g > \triangleq \int_0^1 f(t)g(t)\,dt$$
$$\|f\| = \sqrt{< f, f >}$$
$$L^2[0,1] = \{f : \int_0^1 f^2(t)\,dt < \infty\} \qquad \text{(Square-integrable function)}$$

Eigen functions complete in L^2

Define Z_i as the following

$$Z_i \triangleq \int_0^1 x(t)e_i(t)\,dt$$

We know Z_i is a linear process (think of any Ramon sum approximation. We can look at the covariance of Z_i (showing independence):

$$\begin{aligned} E[Z_i Z_j] &= \int_0^1 \int_0^1 E[x(s)x(t)e_i(s)e_j(t)]\,ds\,dt \\ &= \int_0^1 \int_0^1 c(s,t)e_j(t)e_i(s)\,ds\,dt \\ &= \int_0^1 \lambda_j e_j(s)e_i(s)\,ds \\ &= \int \lambda_i\,ds \qquad \text{(If } i = j\text{, because } \textstyle\int_0^1 e_i e_j = 0 \text{ if } i \neq j\text{)} \end{aligned}$$

Overall we have the Kosambi–Karhunen–Loève theorem, i.e.: stochastic process can be represented as an infinite linear combination of orthogonal functions

$$x(t) = \sum_{k=0}^{\infty} z_k e_k(t)$$

Z_k are pairwise uncorrelated random variables and the function e_k are continuous real valued functions that are pairwise orthogonal in L^2

4.6.4 Example: Kosambi–Karhunen–Loève Expansion in Weiner Process

We have

$$B(t) : 0 \leq t \leq 1$$
$$E[B(t) = 0] \quad \forall t$$
$$B(t) \overset{D}{=} N(0,t)$$

We can look at $c(s,t)$:

$$\begin{aligned} C(s,t) &= Cov(B(s), B(t)) \\ &= Cov(B(s), B(s) + (B(t) - B(s))) \\ &\qquad \text{(Notice } B(t) - B(s) \text{ is independent from } B(s)) \\ &= Cov(B(s), B(s)) \qquad\qquad (s = min(s,t)) \\ &= s \end{aligned}$$

So then what happens to eigenfunction e_k

$$\begin{aligned} \int_0^1 c(s,t)e_k(t)\,dt &= \int_0^1 min(s,t)e_k(t)\,dt \\ &= \int_0^s min(s,t)e_k(t)\,dt + \int_s^1 min(s,t)e_k(t)\,dt \\ &= \int_0^s te_k(t)\,dt + \int_s^1 se_k(t)\,dt \\ &= \lambda_k e_k(s) \qquad\qquad \text{(From Mercer's Theorem)} \end{aligned}$$

Take the derivative w.r.t s on both side of the last two lines we have

$$se_k(s) + s(-e_k(s)) + \int_s^1 e_k(t)\,dt = \lambda_k e_k'(s)$$

take the derivative again we have

$$\begin{aligned} &- e_k(s) = \lambda_k e_k''(s) \\ &\implies e_k''(s) + \frac{1}{\lambda_k} e_k(s) = 0 \\ &\implies \lambda_k = \frac{1}{(k - \frac{1}{2})^2 \pi^2} \\ &e_k(t) = \sqrt{2}\sin((k - \frac{1}{2})\pi t) \end{aligned}$$

So we have the expansion of Brownian process:

$$B(t) = \sum_{k=0}^{\infty} Z_k e_k(t)$$

Z_k is independent Gaussian with variance λ_k

e_k is defined above

Chapter 5

Markov Chains

5.1 Primer: Dynamical System

In discrete time, a deterministic dynamical system can be modeled as

$$X_{n+1} = f(X_n) \qquad \text{(Time invariant version)}$$
$$X_{n+1} = f_n(X_n) \qquad \text{(Time variant version)}$$

We can add a sequence of iid random variables Z_i to obtain stochastic model

$$X_{n+1} = f(X_n, Z_{n+1}) \qquad \text{(Time invariant version)}$$
$$X_{n+1} = f_n(X_n, Z_{n+1}) \qquad \text{(Time variant version)}$$

In continuous time, we can model a deterministic dynamical system via differential equation

$$\frac{\partial}{\partial t}x(t) = f(x(t)) \qquad \text{(Time invariant version)}$$
$$\frac{\partial}{\partial t}x(t) = f(t, x(t)) \qquad \text{(Time variant version)}$$

We can add white noise, which is continuous analog to a sequence of iid r.v. to obtain stochastic version

$$\frac{\partial}{\partial t}x(t) = f(x(t)) + \epsilon(t) \qquad \text{(Time invariant version)}$$
$$\frac{\partial}{\partial t}x(t) = f(t, x(t)) + \epsilon(t) \qquad \text{(Time variant version)}$$

5.2 Markov Property

Markov property essentially says that the future state only depend on the most recent past state (asymptotically loss of memory / asymptotic independence).

In discrete time:

$$P(X_{n+1} \in \cdot | X_0, ..., X_n) = P(X_{n+1} \in \cdot | X_n)$$

In continuous time:

$$P(X(t+s) \in \cdot | X(u), 0 \leq u \leq t) = P(X(t+s) \in \cdot | X(t))$$

5.3 Discrete One-step transition Kernel

The one step transition kernel denotes the probability of next state given the current state. Generally it is defined as

$$P(n+1, x, B) = P(X_{n+1} \in B | X_n = x)$$

In stationary transition setting, the time doesn't matter, hence regardless what n is, we have the transition kernel:

$$P(x, B) = P(X_{n+1} \in B | X_n = x)$$

So the matrix $[P(n+1)]$ is defined as

$$[P(n+1, x, y) : x, y \in S]$$

The stationary transition probability matrix is

$$[P(x, y) : x, y \in S]$$

5.4 Forward and Backward Equation

Motivation Suppose we want to compute distribution of

$$X_2 | X_0 = x$$

We can obtain it via the product of one-step transition matrix:

$$\begin{aligned} P(X_2 = y | X_0 = x) &= \sum_{z \in S} P(1, x, z) * P(2, z, y) \\ &= P(1) * P(2) \qquad (P^2 \text{ if process is stationary}) \end{aligned}$$

We define the n step transition probability matrix as

$$P_n = P(1) * P(2) ... * P(n)$$

But notice this is $O(nd^3)$ operation. We want to compute $P(1)...P(n)$ recursively via matrix vector instead of matrix matrix computation.

Forward Equation Let μ_0 be a row vector on S represent as the initial probability distribution, then we have

$$\begin{aligned} &\mu_0 P(1)...P(n) = \mu_n \\ &\implies \mu_n = \mu_{n-1} P(n) \qquad \text{(Forward Equation)} \\ &\mu_0(x) = P(X_0 = x) \\ &\mu_n(x) = P(X_n = x) \end{aligned}$$

So for a specific x, we have:

$$\mu_i(x) = (\mu_0 P(1)...P(i))(x)$$

The alternative form is

$$\mu_{i+1} - \mu_i = \mu_i(P(i+1) - I)$$

In stationary transition probability setting $P(i) = P$, the simplifed form is

$$\mu_{i+1} - \mu_i = \mu_i(P - I)$$

Backward Equation Let $r(x) : x \in S$ be a column vector on S represent reward function of state $x \in S$. then we can calculate the utility u_i via

$$\begin{aligned} &u_n = P(1)...P(n)r \\ &u_0 = r \\ &\implies \mu_i = P(n-i+1)...P(n-1)P(n)r \qquad \text{(Backward Equation)} \end{aligned}$$

The interpretation of $u_i(x)$ is the expected reward at step X_n given $n - i$ step ago $X_{n-i} = x$

$$\begin{aligned} u_1(x) &= (P(n)r)(x) \\ &= \sum_{y \in S} p(n, x, y) r(y) \\ &= E[r(X_n) | X_{n-1} = x] \\ u_n(x) &= E[r(X_n) | X_0 = x] \end{aligned}$$

The alternative form is

$$u_i - u_{i-1} = (P(n-i+1) - I)u_{i-1}$$

In stationary transition probability setting $P(i) = P$, the simplifed form is

$$u_i - u_{i-1} = (P - I)u_{i-1}$$

5.5 Intuition for Markov Convergence

In deterministic system $X_n = f(X_{n-1})$, converges means that if $X_n \to X_\infty$, then $x_\infty = f(x_\infty)$. In Markov Chain setting $X_{n+1} = f(X_n, Z_{n+1})$, since at each time $n+1$ we introduce a noise Z_{n+1}, strictly as surely convergence won't be possible. So we have that

$$X_n \Rightarrow X_\infty$$
$$\text{then } X_\infty \overset{D}{=} f(X_\infty, Z) \qquad \text{(a fresh } Z\text{)}$$

Example: 2 State Markov Chains Suppose we have only two state, with transition probability matrix

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

First notice that P will always be a stochastic matrix (cell value ≥ 0 and row sums to 1). For any stochastic matrix, we have

$$Pe = e \qquad (e = \text{col vector with all 1})$$

So we know 1 is an eigenvalue of P and e is the associated eigenvector. We can easily get the decomposition $\lambda_2 = 1 - \alpha - \beta$

$$R = \begin{bmatrix} 1 & \frac{-\alpha}{\alpha+\beta} \\ 1 & \frac{\beta}{\alpha+\beta} \end{bmatrix}$$

So after n steps, we have

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

So if $|1-\alpha-\beta| < 1$, then:

$$P^n(x,y) = P(X_n = y | X_0 = x) = \begin{cases} \frac{\beta}{\alpha+\beta} & y = 1 \\ \frac{\alpha}{\alpha+\beta} & y = 2 \end{cases}$$

So the final state after n steps is independent of the initial state X_0. The other two conditions are

- $1-\alpha-\beta = 1$ which means $\alpha = \beta = 0$. So the 2 state markov chains can be reduced to two 1-state chain that always transition to self state.
- $1-\alpha-\beta = -1$ which means $\alpha = \beta = 1$. So it is always transit to the other state. Which means in even n, we have $X_n = X_0$

5.6 Example of Markov Chain

Notice the setup itself gives us stationary transition probability:

5.6.1 Autoregressive Process

$$X_{n+1} = \rho x_n + Z_{n+1}$$

We can get its distribution (transition density) by differentiating the CDF of Z

$$\begin{aligned} p(x,y) &= \frac{\partial}{\partial y} P_x(X_1 \le y) && (P(X_1 \le y | X_0 = x) \\ &= \frac{\partial}{\partial y} P(\rho x + z_1 \le y) \\ &= \frac{\partial}{\partial y} P(z_1 \le y - \rho x) \\ &= f_z(y - \rho x) \\ &\implies p(x,B) = \int_B f_z(y - \rho x)\, dy \end{aligned}$$

5.6.2 Storage Model

Let S_n be the amount stored at time n, Z_n is the inflow at time n, generally modeled as iid, and O_n is the out flow, generally modeled as aS_n^b

$$\begin{aligned} & S_{n+1} = S_n + Z_{n+1} - O_{n-1} && (\text{New} = \text{Old} + \text{In - Out}) \\ & \iff \\ & S_{n+1} + aS_{n+1}^b = S_n + Z_{n+1} \end{aligned}$$

We can again start with the cdf. Let $g(x) = x + ax^b$

$$\begin{aligned} p(x,y) &= \frac{\partial}{\partial y} P_x(S_1 \le y) \\ &= \frac{\partial}{\partial y} P_x(g(S_1) \le g(y)) \\ &= \frac{\partial}{\partial y} P_x(S_1 + aS_1^b \le y + ay^b) \\ &= \frac{\partial}{\partial y} P(x + Z_1 \le y + ay^b) \\ &= \frac{\partial}{\partial y} F_z(y + ay^b - x) \\ &= f_z(y + ay^b - x)(1 + aby^{b-1}) \end{aligned}$$

So the transition probability is : (notice we have to do $P(x,B)$ because in continuous setting, $P(x,y) = 0$ just like in continuous distribution $P(X = x) = 0$ technically.

$$P(x,B) = \int_B f_z(y + ay^b - x)(1 + aby^{b-1})\, dy$$

5.6.3 Congestion Modeling

This congestion model will have a mixed distribution as result instead of one $P(x, B)$. The setting consists of a waiting room with infinite capacity, and a single server using FIFO order:

W_n	(Waiting time for customer n)
A_n	(Arrival time for customer n)
D_n	(Departure time for customer n)
$X_n \equiv A_n - A_{n-1}$	(iid Inter arrival time between customer n and $n-1$)
V_n	(iid service time for customer n)

In general, we have 1) departure time = arrival + wait + service and 2): If customer n departs before $n+1$ arrivals, then wait time is 0. Otherwise, the wait time is the difference between n departure and $n+1$ arrival:

$$\begin{aligned} D_{n+1} &= A_{n+1} + W_{n+1} + V_{n+1} \\ W_{n+1} &= max(0, D_n - A_{n+1}) = [D_n - A_{n+1}]^+ \end{aligned}$$

Combine together we have

$$\begin{aligned} W_{n+1} &= [A_n + W_n + V_n - A_{n+1}]^+ \\ &= [W_n + V_n - X_{n+1}]^+ \\ &= [W_n + Z_{n+1}]^+ \qquad (Z_{n+1} \triangleq V_n - X_{n+1}) \end{aligned}$$

Assume $y \geq 0$

$$\begin{aligned} P_x(W_1 \leq y) &= P(x + Z_1 \leq y) \\ &= F_z(y - x) \\ P_x(W_1 = 0) &= P_X(W_1 \leq 0) \\ &= F_z(-x) \end{aligned}$$

Taking partial derivative:

$$\frac{\partial}{\partial y} P_x(W_1 \leq y) = f_z(y - x)$$

So we have:

$$\begin{aligned} P(x, dy) &= F_z(-x)\zeta_0(dy) + f_z(y - x)dy \\ \zeta_0(A) &= \begin{cases} 1 & 0 \in A \\ 0 & \text{else} \end{cases} \\ P(x, B) &= F_z(x)\zeta_0(B) + \int_B f_z(y - x)dy \end{aligned}$$

5.7 Markov Chain Convergence Proof

Statement : Let Λ be a stochastic matrix with identical rows, and $\delta > 0$, then as $n \to \infty$:

$$P \geq \delta\Lambda \Longrightarrow P^n \to \Pi$$

where Π is a stochastic matrix with identical rows (For each row, each column has the same value). This means that the chain converge to state that is independent of initial state.

Proof We can represent P as:

$$P = \delta\Lambda + (1-\delta)Q$$
$$Q \triangleq \frac{(P - \delta\Lambda)}{1-\delta}$$

Lemma 1: Note that if J is a stochastic rank 1 matrix with identical rows, then $RJ = J$ for any stochastic matrix R.

$$\begin{aligned} P^2 &= P\left(\delta\Lambda + (1-\delta)Q\right) \\ &= \delta\Lambda + (1-\delta)PQ \qquad (P\Lambda = \Lambda \text{ due to lemma 1}) \\ &= \delta\Lambda + (1-\delta)(\delta\Lambda + (1-\delta)Q)Q \\ &= \delta\Lambda + \delta(1-\delta)\Lambda Q + (1-\delta)^2 Q^2 \end{aligned}$$

By induction, we can derive that

$$P^n = \delta\Lambda + \delta(1-\delta)\Lambda Q + ... + \delta(1-\delta)^{n-1}\Lambda Q^{n-1} + (1-\delta)^n Q^n$$

Define a new norm of max matrix row sum

$$|||A||| \triangleq \max_{x \in S} \sum_y |A(x,y)|$$

Given the nature of Stochastic matrix, we know that

$$|||\Lambda Q^n||| = 1$$

Given the boundness, we have

$$\delta\Lambda + \delta(1-\delta)\Lambda Q + ... + \delta(1-\delta)^{n-1}\Lambda Q^{n-1} + (1-\delta)^n Q^n \to \sum_{j=0}^{\infty} \delta(1-\delta)^j \Lambda Q^j \triangleq \Pi$$

Hence $P^n \to \Pi$ and $|||P^n - \Pi||| = O(|1-\delta|^n)$.

Extension 1: Suppose that $\exists m \geq 1$ s.t. $P^m \geq \delta\Lambda$, then we also have $P^n \to \Pi$ as $n \to \infty$.

Extension 2: $\exists\, y$ s.t. $P_x(X_m = y) \geq \delta \quad \forall x$ iff $P^m \geq \delta\Lambda$. This former statement is also equivalent to $\min_{x\in S} P_x(X_n = y) > 0$.

5.8 Markov Chain Convergence in Infinite State Space

Let $|S| = \infty$ and S is discrete, then the condition is $\exists\, y \in S\, m \geq 1$ such that

$$\inf_{x\in S} P_x(X_m = y) > 0$$

i.e.: Uniformly high probability that the chain will hit y in m steps. Then we know that $P^n \to \Pi$, the difference is still in order $O(|1-\delta|^n)$

5.9 Markov Chain Convergence in General State Space

More formally, the distribution Π is a stationary distribution of P if

$$\Pi(A) = \int_{x\in S} \Pi(dx)P(x, A)$$

We first define total variation metric as

$$\sup_{A\subseteq S, x\in S} |P_x(X_n, A) - \Pi(A)|$$

We can then define Doeblin minorization condition as $\exists\epsilon > 0$ and a probability measure $\mu(\cdot)$ such that for all $x \in S$ and measurable subsets $A \subseteq S$, we have

$$P(x, A) \geq \epsilon\mu(A)$$

The convergence theorem says that if above condition is met, then for any $x \in S$ and given the stationary distribution $\Pi(\cdot)$ we have

$$\sup_{A\subseteq S, x\in S} |P_x(X_n, A) - \Pi(A)| \leq (1-\epsilon)^m$$

We can apply Doeblin minorization condition on the Markov chain, then the rate of convergence can be bounded by a function that goes to 0.

5.10 Martingale Sequence

5.10.1 Definition

Sequence $(M_n : n \geq 0)$ is a martingale adapted to sequence $(Z_n : n \geq 0)$ if

- Adaptedness: For each $n \geq 0$, there exists a deterministic function f_n such that $M_n = f_n(Z_0, ..., Z_n)$.
- Finite expectation: $E[M_n] < \infty, n > 0$
- Stable expectation: $E[M_{n+1}|Z_0, ..., Z_n] = M_n$.
  - Sub-Martingale: $E[M_{n+1}|Z_0, ..., Z_n] \geq M_n$.
  - Super-Martingale: $E[M_{n+1}|Z_0, ..., Z_n] \leq M_n$.

5.10.2 Examples

5.10.2.1 Random Walk

$$\begin{aligned} &S_n = Z_1 + ... + Z_n \\ &Z_i \sim iidZ \\ &E[Z] = 0 \end{aligned}$$

Adaptedness (sum function) and finite expectation (0) is trivial. For stable expectation:

$$\begin{aligned} E[S_{n+1}|Z_0, ..., Z_n] &= E[S_n + Z_{n+1}|Z_0, ..., Z_n] \\ &= S_n + E[Z_{n+1}|Z_0, ..., Z_n] \\ &= S_n \end{aligned}$$

5.10.2.2 Random Walk 2

$$\begin{aligned} &S_n = Z_1 + ... + Z_n \\ &Z_i \sim iidZ \\ &E[Z] = 0 \\ &E[Z^2] < \infty \\ &M_n = S_n^2 - n\sigma^2 \end{aligned}$$

Adaptedness is easy. For expectation:

$$\begin{aligned} E[M_n] &= E[S_n^2] - n * Var(Z) \\ &= E[S_n]^2 + Var(S_n) - n * Var(Z) \\ &= n^2 * Var(Z) - nVar(Z) < \infty \end{aligned}$$

For conditional expectation

$$\begin{aligned}
E[M_{n+1}|Z_0,...,Z_n] &= E[S_{n+1}^2 - (n+1)\sigma^2|Z_0,...,Z_n] \\
&= E[(S_n + Z_{n+1})^2 - (n+1)\sigma^2|Z_0,...,Z_n] \\
&= E[S_n^2 + Z_{n+1}^2 - 2S_nZ_{n+1} - (n+1)\sigma^2|Z_0,...,Z_n] \\
&= S_n^2 + E[Z_{n+1}^2|Z_0,...,Z_n] - 2S_nE[Z_{n+1}|Z_0,...,Z_n] - (n+1)\sigma^2 \\
&= S_n^2 + \sigma^2 - (n+1)\sigma^2 \\
&= S_n^2 - n\sigma^2 = M_n
\end{aligned}$$

5.10.2.3 Markov Chain

Let $f : S \to \mathbb{R}$ be a bounded function that $Pf = f$. Notice adaptedness is trivial, finite expectation is given by boundness. For conditional expectation:

$$\begin{aligned}
E[f(X_{n+1})|X_1,...,X_n] &= \sum_{y\in S} f(y)P(X_{n+1} = y|X_1,...,X_n) \\
&\quad \text{(Notice the probability is the } (X_n, y) \text{ entry in } P) \\
&= (Pf)(X_n) \ \text{ (Entry associated with } X_n \text{ in vector } Pf) \\
&= f(X_n)
\end{aligned}$$

5.10.3 Represent Martingale as Sum of Differences

Martingale sequence can be re-written as

$$\begin{aligned}
M_n &= M_0 + D_1 + ... + D_n \\
D_n &= M_n - M_{n-1}
\end{aligned}$$

Notice that $E[D_n] = 0$ given how Martingale sequence is defined. Also, $Cov(D_i, D_j) = 0$ given the history. Proof: Assume $i < j$ and that MG is square integrable $(E[M_n^2] < \infty)$:

$$\begin{aligned}
E[D_iD_j|Z_0,...,Z_{j-1}] &= D_i * E[D_j|Z_0,...,Z_{j-1}] \\
&= D_i * E[M_j - M_{j-1}|Z_0,...,Z_{j-1}] \\
&= D_i * (E[M_j|Z_0,...,Z_{j-1}] - M_{j-1}) \\
&= D_i * 0 = 0
\end{aligned}$$

Another point is $Cov(M_0, D_i) = 0, i \geq 1$. So overall we have

$$Var(M_n) = Var(M_0) + \sum_{i=1}^{n} Var(D_i)$$

5.10.4 WLLN For Martingale Sequence

Assume $Var(D_i) \leq c < \infty$:

$$\begin{aligned} P(\left|\frac{M_n}{n}\right| > \epsilon) &\leq \frac{Var(M_n)}{\epsilon^2 n^2} && \text{(Chebyshev's Inequality)} \\ &= \frac{Var(M_0) + \sum_{i=1}^{n} E[D_i^2]}{n^2\epsilon^2} \\ &= 0 \end{aligned}$$

Hence $\frac{M_n}{n} \xrightarrow{P} 0$

5.10.5 SLLN For Martingale Sequence

Martingale Converge Theorem: Let $M_n : n \geq 0$ be a generic Martingale sequence adopted to $Z_n, n \geq 0$ and $\sup_{n\geq 0} E[|M_n|] < \infty$, then there eixtis a finite-valued r.v. M_∞ such that

$$M_n \xrightarrow{a.s.} M_\infty$$

Some Intuition a value higher then interval $[a, b]$ is called an upcrossing of $[a, b]$. Martingale have finite number of upcrossing (upcrossing inequality). A sequence doesn't converge iff it osscilates or go up / down infinitey. The former is prevented by the upcrossing inequality of MG, and the later is prevented by the finite upper bound $\sup_{n\geq 0} |E[M_n]| < \infty$

Proof Intuition : Notice that $\frac{1}{n}M_n$ is not Martingale, so we can't directly use Martingale Convergence Theorem. We need to construct $\alpha_j, j \geq 0$ deterministicly such that

$$\hat{M}_n = \alpha_0 M_0 + \sum_{i=1}^{n} \alpha_j D_j$$

is always a MG sequence.

Formal Proof Let $\hat{M}_n = M_0 + \sum_{j=1}^{n} \frac{1}{j} D_j$.

$$\begin{aligned} Var(\hat{M}_n) &= Var(M_0) + \sum_{j=1}^{n} \frac{1}{j^2} E[D_j^2] && \text{(Assume bounded)} \\ &\leq var(M_0) + \sum_{j=1}^{\infty} \frac{1}{j^2} c \\ &\implies \sup_{n\geq 1} E[\hat{M}_n^2] < \infty \\ &\implies \sup_{n\geq 1} E[\left|\hat{M}_n\right|] < \infty && \text{(By Cauchy-Schwartz Inequality)} \end{aligned}$$

So by Martingale Convergence Theorem, we know that there exists finite valued r.v. $\hat{M}_\infty$ such that

$$\hat{M}_n \overset{a.s.}{\to} \hat{M}_\infty$$
$$\implies M_0 + \sum_{j=1}^{n} \frac{1}{j} D_j \overset{a.s.}{\to} \hat{M}_\infty$$

Apply Kronecker's lemma path by path ($\alpha_j = j, x_j = D_j$), we have

$$M_0 + \frac{1}{n}(D_1 + ... + D_n) \overset{a.s.}{\to} 0$$

Kronecker's Lemma Let $\alpha_n, n \geq 0$, $\alpha_n \nearrow \infty$, and

$$\sum_{j=1}^{n} \frac{X_j}{\alpha_j} \to c < \infty$$

then

$$\frac{1}{\alpha_n} * \sum_{i=1}^{n} x_i \to 0$$

5.10.6 CLT For Martingale

Let $M_n : n \geq 0$ adopted to $Z_n : n \geq 0$, and M_n is square-integrable.D_i form a stationary sequence, then

$$\frac{1}{\sqrt{n}} M_n \Rightarrow \sigma N(0,1)$$
$$\sigma^2 = Var(D_1) = E[D_1^2]$$

5.11 Connect Martingale to Markov Chains

5.11.1 Reward function difference as MG difference

Let $g : S \to \mathbb{R}$ and bounded. $D_i = g(X_i) - E[g(X_i)|X_{i=1}]$, then D_i is a MG differences adopted to Markov Chain $X_n : n \geq 0$

Proof The adaptedness and finite expectation are obvious. Now we want to show stable expectation, hence

$$\begin{aligned}
0 &= E[D_i|X_0, ..., X_{i-1}] \\
&\qquad \text{(Notice } D_i \text{ is the MG difference)} \\
&= E[g(X_i)|X_0, ..., X_{i-1}] - E[g(X_i)|X_{i-1}] \\
&= E[g(X_i)|X_{i-1}] - E[g(X_i)|X_{i-1}] \qquad \text{(Markov Property)} \\
&= 0
\end{aligned}$$

Example if S is discrete, we have

$$\begin{aligned} E[g(X_i)|X_{i-1}] &= \sum_{y\in S} P(X_{i-1}, y)g(y) \\ &= (Pg)(X_{i-1}) \end{aligned}$$

(The X_{i-1} corresponding entry to Matrix Vector product)

5.11.2 Construct Markov Chain

We have $D_i = g(X_i) - E[g(X_i)|X_{i-1}]$ is a MG difference, we can get a MG by

$$\begin{aligned} M_n &= \sum_{i=1}^{n} D_i \\ &= \sum_{i=1}^{n} [g(X_i) - (Pg)(X_{i-1})] \\ &= \sum_{i=0}^{n-1} [g(X_i) - (Pg)(X_{i-1})] - g(X_0) + g(X_n) \end{aligned}$$

5.12 Markov Chains Reward Function Converges

We know that if $p^m \geq \delta\Lambda$, then $p^n \to \Pi$, hence $P_x(X_n = y) \to \pi(y)$ as $n \to \infty$. This is the convergence of probability distribution function. If we look at the backward function, the intuition is follows:

Let $r(x)$ be a function $S \to \mathbb{R}$, let $r(X_n)$ denotes the entry in vector r associated with state X_n. (Notice $X_n \in \S$). Let π be a row in stable distribution Π. Then LLN of Markov Chains states:

$$\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) \xrightarrow{P} \sum_{x\in S} \pi(x)r(x) = \pi r$$

In another word, the long term cumulative reward of each state can be decomposed into the probability at that state times the reward at that state.

Proof Let

$$r_c(x) \triangleq r(x) - \sum_{y\in S} \pi(y)r(y)$$

Then we can represent $\sum_{j=0}^{n-1} r_c(X_j)$ in terms of MG sequence if we could find g such that $(I - P)g = r_c$ This is also called the poission equation. Then

$$\sum_{i=0}^{n-1} r_c(X_i) - g(X_1) + g(X_0)$$

would be MG sequence, then we have

$$\frac{1}{n}\sum_{i=0}^{n-1} r_c(X_i) \overset{a.s.}{\to} 0$$

So now we need to shown the existence of g. We know that $|S| > \infty$ and $p^m \geq \delta\Lambda$. Notice sicne P is a stochastic matrix, we have $Pe = e$, 1 is an eigenvalue of P. So $P - I$ is a singular matrix, so we cannot just take inverse to obtain the solution to $(I - p)g = r_c$. In another word, any solution of g is not unique, because $\hat{g} = g + ce$ is also a solution.

Also notice $(I - P)g = h$ will not be solvable for all h. Because $p^n \to \Pi$, we have $p^{n+1} = p^n p = pp^n \Rightarrow \Pi = \Pi p = p\Pi$. Since Π has identical rows, let π be a row of Π. Apply to $(I - P)g = h$ we get $0 = \pi h$. So h has to satisfies $\pi h = 0$, which r_c satisfies.
From $p^n \to \Pi$ we have

$$(P - \Pi)^2 = P^2 - p\Pi - \Pi p + \Pi^2 = p^2 - \Pi$$

by induction, we have

$$(P - \Pi)^n = p^n - \Pi$$

Notice due to convergence, $p^n - \Pi$ goes to 0 geometrically fast, so norm of $||P - \Pi||^r < 1$ for some $r \geq 1$. So $\sum_{n=0}^{\infty}(P - \Pi)^n$ converges and to $(I - (P - \Pi))^{-1}$ (inverse also exists).

Let $\tilde{g} = g + \Pi g$ (if g is a solution, $\tilde{g}$ is also a solution). We have

$$\begin{aligned}
&(I - P)(g + \Pi g) = r_c \\
&\Longrightarrow (I - P + \Pi)g = r_c \\
&\Longrightarrow g = (I - P + \Pi)^{-1} r_c
\end{aligned}$$

So overall with $g = (I - p + \Pi)^{-1} r_c$ we have the following MG sequence

$$\sum_{i=1}^{n-1} r_c(X_i) + g(X_n) - g(X_0)$$

and

$$\frac{1}{n}\sum_{i=0}^{n-1} r(X_i) \overset{a.s.}{\to} \sum_{x \in S} \pi(x) r(x)$$

5.13 CLT for Markov Chains

If $|s| < \infty$, $p^m \geq \delta\Lambda$, then we know that $g = (I - P + \Pi)^{-1} r_c$ exists.
Let $r(x)$ be a function $r \to \mathbb{R}$, let $r(X_n)$ denotes the entry in vector r associated

with state X_n. (Notice $X_n \in \S$). Let π be a row in stable distribution Π. Then CLT of Markov Chains states:

$$\sqrt{n}\left(\frac{1}{n}\sum_{j=0}^{n-1} r(X_j) - \pi r)\right) \Rightarrow \sigma N(0,1)$$
$$\sigma^2 = E_\pi[D_i^2]$$

5.14 First Transition Analysis

5.14.1 Recursion Equation

Let $u^*(x) = E_x[T]$ and $T = inf\{n \geq 0 : x_n \in C^c\}$. I.e.: u^* is the expected value of starting from x, first time step into C^c. Notice that given the setup, u^* has to satisfies the recursion

$$u(x) = 1 + \sum_{y \in C} P(x,y)u(y)$$

Let u be a vector represent $(u(x), x \in C)$, B be the block in transition matrix represent the transition probability from C to C, then we can also written the above recursion as

$$u = e + Bu$$

Alternatively, let u' be a vector represent $(u(x), x \in S)$ (notice here we are in space S instead of C), then we have

$$u' = e + Pu'$$
$$u'(x) = \begin{cases} 1 + \sum_y P(x,y)u'(y) & x \in C \\ 0 & x \in C^c \end{cases}$$

Note here we shown that $u^*(x)$ is a solution to the above recursion equation. Later we show that it is the minimal non-negative solution.

5.14.2 Potential Complication Example of FTA

5.14.2.1 Infinity

Let $|S| = 2$, and each state will only transition back to itself. Let $C^c = \{2\}$, $u^*(1) = E_1[T]$. Given the setup, it is obvious to see that $u^*(1) = \infty$. This showcases complication 1: ∞ can be a legit solution and always will be a solution to $u = e + Bu$ for non=trivial u.

5.14.2.2 Non-uniqueness

Let $|S|$ be a infinite birth-death chain. At any state m there is p go to state $m+1$ and q go to state $m-1$. At 0, there is Q go to sate 0 again. Let $C^c = \{0\}$. $u^*(x) = E_x[T]$. So we have

$$u(x) = \begin{cases} 1 + pu(x+1) + qu(x-1) & x \geq 1 \\ 0 & x = 0 \end{cases}$$

Notice that this equation has non-unique solutions since it only has 1 boundary condition.

5.14.3 Minimal Solution to Recursive Equation

$$\begin{aligned}
u^*(x) &= E_x[T] \\
&= E_x[\sum_{j=0}^{T-1} 1] \\
&= E_x[\sum_{j=0}^{\infty} \mathbb{I}\{T > j\}] \\
&= \sum_{j=0}^{\infty} p_x(T > j) \qquad \text{(Recall } T = inf\{n \geq 0 : x_n \in C^c\}) \\
&= \sum_{j=0}^{\infty} \sum_{y_1,\ldots,y_{j-1} \in C} p(x, y_1)p(y_1, y_2)...p(y_{j-2}, y_{j-1}) \\
&= \sum_{j=0}^{\infty} \sum_{y_1,\ldots,y_{j-1}} B(x, y_1)...B(y_{j-2}, y_{j-1}) \\
&= \sum_{j=0}^{\infty} B^j e
\end{aligned}$$

Notice above derivation will give us the same recursion equation

$$\begin{aligned}
u^* &= \sum_{j=0}^{\infty} B^j e \\
&= e + \sum_{j=1}^{\infty} B^j e \\
&= e + B \sum_{j=0}^{\infty} B^j e \\
&= e + Bu^*
\end{aligned}$$

Theorem : u^* is the minimal non-negative solution of $u = e + Bu$. Equivalently, if $u' \geq 0$ satisfies $u' = e + Bu'$, then $u'(x) \geq u^*(x), x \in C$

Proof : Let $u' \geq 0$ be a solution to $u = e + Bu$

$$\begin{aligned} u' &= e + Bu' \\ &= e + B(e + Bu') \\ &= E + Be + B^2u' \\ &... \qquad\qquad = e + Be + ... + B^n e + B^{n+1}u' \\ &\geq e + Be + ... + B^n e \qquad \text{(Both } B \text{ and } u' \text{ is non-negative)} \\ &= u^* \end{aligned}$$

5.14.4 Relation to Reward Function

If we now think about reward function.

$$u^*(x) = E_x[\sum_{j=0}^{T-1} r(x_j)]$$
$$u(x) = r(x) + \sum_{y \in C} P(x,y)u(y)$$

So the recursive equation becomes

$$u = r + Bu$$

Following the same set of derivation, we have

$$u^*(x) = \sum_{n=0}^{\infty} B^n r$$

5.14.5 Example: Gambler's ruin problem

Let state 0 be gambler 1 loose all money, and state N be gambler 2 loose all money. Then the game between two gambler can be representd as Markov Chain with state 0 to N. At each state $m \neq 1 \neq N$, there is p chance to mvoe to $m+1$, and q chance to move to $m-1$. State 0 and N are absorbing states.

Let $x_i \in S$ be the wealth of gambler 1 at time i. Let $T = inf\{n \geq 0 : x_n \in C^c\}$, $C^c = \{0, N\}$. Let $u^*(x) = P_x(X_T = 0)$, hence the probability of gambler 1 ruined. Notice this setup is a specific version of P(x enter C^c through A). So let $A = \{0\}$. Hence $u^*(x) = p_x(X_T \in A)$.

We know it has to satisfies recursion (step into A, or step int something that is not A)

$$u(x) = \sum_{y \in A} P(x,y) + \sum_{y \in C} P(x,y)u(y)$$

As a result, we will have linear system:

$$u = f + Bu$$
$$f = (f(x) : x \in C)$$
$$f(x) = \sum_{y \in A} P(x, y)$$
$$u^* = \sum_{n=0}^{\infty} B^n f$$

5.14.6 High Level FTA

At high level, we can summarize FTA as

$$u = f + Gu$$
$$G \geq 0, f \geq 0$$
$$u^* = \sum_{n=0}^{\infty} G^n f$$

If $G^n \to 0$, then $\sum_{n=0}^{\infty} G^n = (I - G)^{-1}$ and $(I - G)^{-1}$ is non-negative. Conversely, if $(I - G)^{-1}$ is singular or has negative entries, then G^n does not converges to 0,the infinite sum does not converge and $u^* \neq (I - G)^{-1} f$

Proof : $G^n \to 0$, so $||G^m|| < 1$ for some $m \geq 1$

$$\begin{aligned}\sum_{r=0}^{\infty} \sum_{e=0}^{m-1} ||G^{rm+e}|| &\leq \sum_{r=0}^{\infty} \sum_{e=0}^{m-1} ||G^e|| * ||G^m||^r \\ &\leq \sum_{r=0}^{\infty} \sum_{e=0}^{m-1} ||G^l|| * ||G^m||^r \\ &\leq \sum_{r=0}^{\infty} c * ||G^m||^r\end{aligned}$$

Converges, so the infinite sum is finite

Now we want to show that $\sum_{n=0}^{\infty} G^n = (I - G)^{-1}$. We have

$$(I - G)(I + G + ... + G^n) = I - G^{n+1}$$
$$\Rightarrow (I - G) \sum_{n=0}^{\infty} G_n = I$$

Proof Converse : Suppose $(I - G)^{-1}$ exists and is non-negative. We have

$$(I - G)(I + G + ... + G^n) = I - G^{n+1} \leq I$$
$$\Longrightarrow (I - G)^{-1}(I - G)(I + G + ... + G^n) \leq (I - G)^{-1}$$

5.14.7 Lyapunov Bound

So if we have $u^* = \sum_{n=0}^{\infty} G^n f$, how can I guarantee that u^* is finite-valued and can obtain an upper bound. It is not always easy to solve $u = f + Gu$ directly (for which u^* is the minimum non-negative solution). So we can relax the equality to inequality.

Suppose we have $h \geq 0$ such that $Gh \leq h - f$, h is finite-valued, then

$$u^* = \sum_{n=0}^{\infty} G^n f \leq h$$

Proof

$$\begin{aligned} & Gh \leq h - f \\ & \implies Gh \leq h && (f \text{ is non-negative}) \\ & \implies G^2 h \leq Gh \leq h \\ & \implies G^n h \leq h \end{aligned}$$

Since h is finite valued, so $G^n h$ is also finite valued.

$$\begin{aligned} & f \leq h - Gh && \text{(Assumption)} \\ & \implies Gf \leq Gh - G^2 h \\ & \ldots && \implies G^n f \leq G^n h - G^{n+1} h \end{aligned}$$

If we sum over all the inequalities (notice how $h - Gh + Gh - G^2 h$ cancels). we have

$$\sum_{i=0}^{n} G^i f \leq h - G^{n+1} h \leq h$$

Example Look at the infinite birth death chain again with transition probability p, q. Notice if $p > q$, the the process will go to infinity with probability 1, so we only look at the cases of $p < q$. Let $u^*(x) = E_x[T]$. So we have

$$\begin{aligned} & ph(x+1) + qh(x-1) \leq h(x) - 1 && x \geq 2 \\ & ph(2) \leq h(1) - 1 && x = 1 \end{aligned}$$

If we guess $h(x) = cx$, then we can solve it get c has to satisfies $c \geq \frac{-1}{p-q}$. So we have $E_x[T] \leq \frac{x}{q-p}$

5.15 General Reducible Markov Chain Decomposition

We know that in discrete state space:

$p^m \geq \delta\Lambda, \delta > 0, m \geq 1, \Lambda$rank one stochastic matrix with identical rows

$\iff$

$\exists y \in S s.t. \inf_{x \in S} p^m(x, y) > 0$

$\iff$

$|S| < \infty, X$ is irreducible and aperiodic

5.16 Principle of Regeneration

$|S| = \infty$ or S is continuous and non compact. How do we reason its convergence?

5.16.1 Motivation

Let $X_{n+1} = [X_n + Z_{n+1} - 1]^+$, $S = \{0, 1, ...\}$ We know that this chain is irreducible if Z_{n+1} has non-zero probability mass on 0 and 2.

Let's pick an arbitrary benchmark c and let $T(0)$ be the first time $X_n = c$, and $T(1)$ will be the second time and so on. Notice that the distribution of X_n from $T(0)$ to $T(1)$ is the same as the distribution of X_n from $T(1)$ to $T(1)$, and each distribution is independent from the other.

5.16.2 Recurrent

Let X be an irreducible MC, $x \in S$ is recurrent if $P_x(\tau(x) < \infty) = 1$ where $\tau(x) = inf\{n \geq 1 : x_n = x\}$. Notice this definition is a class property. If it is true for one state, it is true for all states.

5.16.3 LLN

Let $\tau_i = T(i) - T(i-1), i \geq 1$, let $r_i : S \to \mathbb{R}_+$, let $y_i = \sum_{j=T(i-1)}^{T(i)-1} r(x_j)$. If X is irreducible and recurrent, then (y_i, τ_i) are iid sequence. Notie tht y_i and τ_i are correlated, longer sequence generally means bigger reward. But each pair is indpendent from the other pair. The final convergence theorem is

$$\frac{1}{n} \sum_{i=0}^{n-1} r(x_i) \overset{a.s.}{\to} \frac{E_z\left[\sum_{j=0}^{\tau(z)-1} r(x_j)\right]}{E_z\left[\tau(z)\right]}$$

Extension: If X is irreducible and positive recurrent, then

$$\frac{1}{n}\sum_{i=0}^{n-1} r(x_i) \overset{a.s.}{\to} \sum_{x\in S} \pi(x)r(x)$$
$$\pi(x) = \frac{E_z\left[\sum_{j=0}^{\tau(z)-1} I(X_j = x)\right]}{E_z[\tau(z)]}$$

5.17 MC Chain Stability and Equilibrium Behavior

In discrete space, stability is equivalent to positive recurrence, and equilibrium is equivalent to $\pi = \pi P$.

Let $X = (X_n : n \geq 0)$ be s-valued. $|S| < \infty$, X is irreducible and aperiodic (equivalent to $p^m \geq \delta\Lambda$), then there exists a pmf Π such that as $n \to \infty$

$$p^n(x, y) \to \pi(y)$$
$$\frac{1}{n}\sum_{j=0}^{n-1} r(x_j) \overset{a.s.}{\to} \sum_{x\in S} \pi(x)r(x)$$

If we let $|s| \geq \infty$ (so include infinite), and X is still irreducible, then TFAE:

- MC positive recurrence: $E_z[\tau(z)] < \infty$
- Exists a pmf solution π to $\Pi = \Pi P$

then we have

$$\frac{1}{n}\sum_{i=0}^{n-1} r(x_i) \overset{a.s.}{\to} \sum_{w\in S} \pi(w)r(w)$$
$$\pi(x) = \frac{E_z\left[\sum_{j=0}^{\tau(z)-1} I(x_j = x)\right]}{E_z[\tau(z)]}$$

Extension: How to establish stability by showing positive recurrence through Lyapunov bound.
Recall: $u^* = E_x[T]$, $T = \inf\{n \geq 0 : x_n \in C^c\}$, $B = (B(x, y) : x, y \in C)$, $B(x, u) = P(x, y)$. If there exists a function g such that

$$Bg \leq g - e \iff \sum_{y\in C} B(x, y)g(y) \leq g(x) - 1 \qquad \forall x$$

then $E_x[T] \leq g(x)$.

1. Choose $C^c = \{Z\}$

2. Find a suitable Lyapunov function for g
3. $E_z[\tau(z)] \leq g(x)$
4. $E_z[\tau(z)] \leq 1 + \sum_{y \neq z} P(z, y)g(y)$

So if we can show the last part is bounded, then we have positive recurrence.