Supplemental Notes

EE503 Week 13 Dr. Franzke

HW: (none)

① Read: Gubner, Ch. 12

Leon Garcia, Ch. 11

② Daily Derivations (Final Exam prep)

Topics: Markov Chains and Estimation.

① M/M/1 Queue "Birth Death" Process

"Global Balance Equations"

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda+\mu) p_n = \lambda p_{n-1} + \mu p_{n+1} \quad \text{if } n \geq 1 \end{cases}$$

("birth" λp_{n-1}, "death" μp_{n+1})

$\therefore p_n = (1-\rho)\rho^n$ Geometric pdf

② (DTFS) Markov Chain: $\pi_e P = \pi_e$ (π_e: equilibrium pdf.)

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \Rightarrow \pi_e = \left(\frac{b}{a+b}, \frac{a}{a+b} \right)$$

(π_e: the "google eigenvector")

M/M/1 Queue

- Queue with 1 server & infinite buffer

  ∴ State space = $\{0, 1, 2, 3, \ldots\}$

- First "M": Poisson arrivals (customers)

  with rate $\lambda > 0$ (Poisson process)

  - i.i.d. exponential inter-arrival times

    (renewal process)

- Second "M": Exponential service times

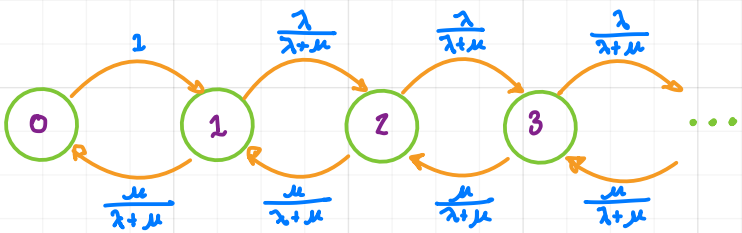
  with rate $\mu > 0$

  ∴ pdf $f(t) = \mu e^{-\mu t}$

∴ Continuous Markov Chain: with "jumps"

(Birth-Death process)

'jumps' ⟶ EMBEDDED CHAIN.



∴ λ = average "birth" rate

μ = average "death" rate

$P_n(t)$ = Prob (System is in state (n) at time t)

∴ at equilibrium : p_n = proportion of time system is in state (n)

$p_n = P_n(\infty) = \lim_{t \to \infty} p_n(t)$ (if exists)

Assume: $\lambda < \mu$ (else unstable)

Define: $\rho \triangleq \frac{\lambda}{\mu}$ ∴ $0 < \rho < 1$.

✡ Global Balance Equations (memorize)

$$\begin{cases} \lambda p_0 = \mu p_1 \\ (\lambda + \mu) p_n = \lambda p_{n-1} + \mu p_{n+1} \quad \text{if } n \geq 1 \end{cases}$$

("birth" → λp_{n-1}, "death" → μp_{n+1}) At equilibrium

Will show: $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 = \rho^n p_0$ (pdf: $\sum_{n=0}^{\infty} p_n = 1$)

$\therefore \sum_{n=0}^{\infty} p_0 \cdot \rho^n = \frac{p_0}{1-\rho}$

$\therefore p_0 = 1 - \rho.$

$$\therefore \boxed{p_n = (1-\rho)\rho^n}$$

Geometric pdf

$\therefore P(N \geq n) = \rho^n$

$\therefore E[N] = \boxed{\frac{\rho}{1-\rho}} = \frac{\lambda}{\mu - \lambda}$ ← expected number of customers in the system

$V[N] = \boxed{\frac{\rho}{(1-\rho)^2}}$

Derivation

Assume: - $P(1 \overset{B}{\text{BIRTH}} \text{ in } (t, t+\Delta t)) \approx \lambda \cdot \Delta t + O(\Delta t)$ (B: arrival)

- $P(\geq 2 \text{ BIRTHS in } (t+\Delta t)) \approx O(\Delta t)$

- $P(1 \overset{D}{\text{DEATH}} \text{ in } (t+\Delta t)) \approx \mu \cdot \Delta t + O(\Delta t)$ (D: service)

for $\lim_{\Delta t \to 0} \frac{O(\Delta t)}{\Delta t} = 0$ $\quad \therefore c\,O(\Delta t) = O(\Delta t)$, $O(\Delta t) + O(\Delta t) = O(\Delta t)$

Put $p_n(t) = P(\overset{N(t)}{\text{population}} = n \text{ at time } t)$ for $n \geq 0$

($\lambda_n = \lambda$ $\forall n$, but $\mu_0 = 0$ & $\mu_n = \mu$ for $n \geq 1$)

Case 1: $n \geq 1$

4 birth-death (D-B) cases plus $O(\Delta t)$ → B: 0 or 1 and D: 0 or 1

$$\therefore p_n(t+\Delta t) = p_n(t)\underbrace{(1-\lambda\Delta t)}_{B=0}\underbrace{(1-\mu\Delta t)}_{\&\ D=0} + p_n(t)\cdot\underbrace{(\lambda\Delta t)}_{B=1}\underbrace{(\mu\Delta t)}_{\&\ D=1}$$

$$+ p_{n-1}(t)\underbrace{(\lambda\Delta t)}_{B=1}(1-\mu\Delta t) + p_{n+1}(t)\cdot(1-\lambda\Delta t)\underbrace{(\mu\Delta t)}_{B=0\ \&\ D=1} + O(\Delta t)$$

$$= p_n(t)\left[1-\lambda\Delta t-\mu\Delta t+\lambda\mu(\Delta t)^2\right] + p_n(t)\lambda\mu(\Delta t)^2$$

$$+ p_{n-1}(t)(\lambda\Delta t-\lambda\mu(\Delta t)^2) + p_{n+1}(t)\left(\mu\Delta t-\lambda\mu(\Delta t)^2\right) + O(\Delta t)$$

$$= p_n(t)(1-\lambda\Delta t-\mu\Delta t) + p_{n-1}(t)\lambda\Delta t + p_{n+1}(t)\mu\Delta t$$

$$+ \underbrace{\left[2p_n(t)\lambda\mu - p_{n-1}(t)\lambda\mu - p_{n+1}(t)\lambda\mu\right](\Delta t)^2}_{= O(\Delta t) \text{ since } \lim_{\Delta t\to 0}\frac{(\Delta t)^2}{(\Delta t)} = \lim_{\Delta t\to 0}\Delta t = 0} + O(\Delta t)$$

$$= p_n(t)\left(1-\lambda\Delta t-\mu\Delta t\right) + p_{n-1}(t)\,\lambda\Delta t + p_{n+1}(t)\,\mu\Delta t + O(\Delta t)$$

since $O(\Delta t)+O(\Delta t)=O(\Delta t)$

$$= p_n(t) - p_n(t)(\lambda+\mu)\Delta t + p_{n-1}(t)\lambda\Delta t + p_{n+1}(t)\mu\Delta t + O(\Delta t)$$

$$\therefore \dot{p}_n(t) = \frac{dp_n(t)}{dt} = \lim_{\Delta t\to 0}\frac{p_n(t+\Delta t)-p_n(t)}{\Delta t}$$

$$= \lim_{\Delta t\to 0}\frac{-p_n(t)(\lambda+\mu)\Delta t + p_{n-1}(t)\lambda\Delta t + p_{n+1}(t)\mu\Delta t}{\Delta t} + \lim_{\Delta t\to 0}\frac{O(\Delta t)}{\Delta t}$$

$$= -p_n(t)(\lambda+\mu) + p_{n-1}(t)\lambda + p_{n+1}(t)\mu.$$

$\therefore$ At equilibrium: $\dot{p}_n(t) = 0$

$$\longleftrightarrow \quad (\lambda+\mu)p_n = \lambda p_{n-1} + \mu p_{n+1}$$

Case 1: $n = 0$

(i.e. $p_{-1}(t)=0$)

3 B-D cases to (Δt) since can't have B=0 & D=1 $\therefore \mu_0 = 0$

$$\therefore p_0(t+\Delta t) = p_0(t)\underbrace{(1-\lambda\Delta t)}_{B=0}\underbrace{(1-\mu_0\Delta t)}_{D=0} + p_0(t)\underbrace{(\lambda\Delta t)}_{B=1}\underbrace{(\mu\Delta t)}_{D=1}$$

$$+ p_1(t)\underbrace{(1-\lambda\Delta t)}_{B=0}\underbrace{(\mu\Delta t)}_{D=1} + O(\Delta t)$$

$$= p_0(t)\left(1-\lambda\Delta t-\mu_0\Delta t+\lambda\mu_0(\Delta t)^2\right) + p_0(t)\lambda\mu(\Delta t)^2$$

$$+ p_1(t)\mu\Delta t - p_1(t)\lambda\mu(\Delta t)^2 + O(\Delta t)$$

$$= p_0(t) - p_0(t)(\lambda+\mu_0)\Delta t + p_1(t)\mu\Delta t + \underbrace{\lambda\mu(p_0(t)-p_1(t))(\Delta t)^2}_{O(\Delta t)} + O(\Delta t)$$

$$= p_0(t) - (\lambda+\underbrace{\mu_0}_{=0})p_0(t)\Delta t + \mu_1 p_1(t)\Delta t + O(\Delta t)$$

$$= p_0(t) - \lambda p_0(t)\Delta t + \mu_1 p_1(t)\Delta t + O(\Delta t)$$

$$\therefore 0 = \dot{p}_0(t) := \lim_{\Delta t\to 0} \frac{p_0(t+\Delta t) - p_0(t)}{\Delta t}$$

$$= \lim_{n\to\infty} \frac{-p_0(t)\lambda\Delta t + p_1(t)\mu\Delta t}{\Delta t} + \lim_{n\to\infty} \frac{O(\Delta t)}{\Delta t}$$

$$= -\lambda p_0(t) + \mu p_1(t)$$

$$\therefore \dot{p}_0(t) = 0 \longleftrightarrow \lambda p_0 = \mu p_1$$

QED - Global Balance Equations

Thm: (for Global Balance Equations) $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0$

Prf: (by induction on n)

Basis step: $n=1$ (& $n=2$)

For $n=1$: $p_0\lambda = p_1\mu \longleftrightarrow p_1 = \left(\frac{\lambda}{\mu}\right)^1 p_0$

For $n=2$: $p_1(\lambda+\mu) = p_0\lambda + p_2\mu$

$$\therefore p_2 = \frac{1}{\mu}\left[p_1(\lambda+\mu) - p_0\lambda\right] \overset{n=1}{=} \frac{1}{\mu}\cdot\left[\left(\frac{\lambda}{\mu}\right)p_0(\lambda+\mu) - p_0\lambda\right)$$

$$= \frac{1}{\mu}\left(\frac{\lambda^2}{\mu}p_0 + \cancel{\lambda p_0} - \cancel{\lambda p_0}\right) = \left(\frac{\lambda}{\mu}\right)^2 \cdot p_0$$

QED: Basis

Induction step:

Assume Induction Hypothesis (IH): $p_n = \left(\frac{\lambda}{\mu}\right)^n p_0$

Show: $p_{n+1} = \left(\frac{\lambda}{\mu}\right)^{n+1} p_0$

$p_{n+1}\cdot\mu = p_n(\lambda+\mu) - p_{n-1}\lambda$ from G.B. $(\lambda+\mu)p_n = \lambda p_{n-1} + \mu p_{n+1}$

$$\therefore p_{n+1} = \left(\frac{\lambda+\mu}{\mu}\right)p_n - \left(\frac{\lambda}{\mu}\right)p_{n-1}$$

$$\overset{\text{I.H.}}{=} \left(\frac{\lambda+\mu}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n p_0 - \left(\frac{\lambda}{\mu}\right)p_{n-1}$$

$$\overset{\text{I.H.}}{=} \left(\frac{\lambda+\mu}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n p_0 - \left(\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^{n-1} p_0$$

$$= \left(\frac{\lambda+\mu}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n p_0 - \left(\frac{\lambda}{\mu}\right)^n p_0$$

$$= \left[\frac{\lambda+\mu}{\mu} - 1\right]\left(\frac{\lambda}{\mu}\right)^n p_0$$

$$= \left[\frac{\lambda}{\mu} + \cancel{1} - \cancel{1}\right]\left(\frac{\lambda}{\mu}\right)^n p_0 = \left(\frac{\lambda}{\mu}\right)^{n+1} p_0$$

QED

$\therefore p_n = \rho^n p_0$ if $\rho = \frac{\lambda}{\mu}$

$$\therefore\ 1 \overset{\text{pdt}}{=} \sum_{n=0}^{\infty} p_n \overset{\text{thm}}{=} \sum_{n=0}^{\infty} \rho^n p_0 = p_0 \cdot \sum_{n=0}^{\infty} \rho^n$$

$$\overset{\text{G.S.}}{=} p_0 \cdot \frac{1}{1-\rho}$$

$$\therefore\ p_0 = 1-\rho$$

geometric structure

$$\therefore\ \boxed{p_n = \rho^n(1-\rho)} \quad \star$$

$$\therefore\ P(N \geq n) = \sum_{k=n}^{\infty} p_k \overset{\text{Thm}}{=} \sum_{k=n}^{\infty} (1-\rho)\rho^k = (1-\rho)\sum_{k=n}^{\infty} \rho^k$$

$$\overset{\text{G.S.}}{=} (1-\rho)\frac{\rho^n}{1-\rho} = \rho^n$$

$$\therefore\ P(N \geq n) = \rho^n.$$

Ex: In-n-Out Burger drivethru $(M|M|1)$

Say 3 customers arrive per minute (late night): $\lambda = 3$

Find service time μ so that 95% of time queue has fewer than 10 customers

$$\therefore\ P(N \geq 10) = \rho^{10} = \left(\frac{\lambda}{\mu}\right)^{10} = \left(\frac{3}{\mu}\right)^{10}$$

$$\therefore\ 0.95 = P(N < 10) = 1 - P(N \geq 10) = 1 - \left(\frac{3}{\mu}\right)^{10} = \frac{19}{20}$$

$$\therefore\ \left(\frac{3}{\mu}\right)^{10} = \frac{1}{20} \longleftrightarrow \mu = 3\sqrt[10]{20} \approx 4.048 \text{ customers per minute.}$$

Recall: $|a| < 1 \longrightarrow \sum_{k=n}^{\infty} a^k \overset{\text{G.S.}}{=} \frac{a^n}{1-a}$

$\therefore \sum_{n=0}^{\infty} n a^n = \frac{a}{(1-a)^2} \quad \left(= \sum_{n=1}^{\infty} n a^n \right)$

$\therefore \sum_{n=0}^{\infty} n^2 a^n = \frac{a+a^2}{(1-a)^3} \quad \left(= \sum_{n=1}^{\infty} n^2 a^n \right)$

$$\therefore E_N[N] = \sum_{n=0}^{\infty} n \cdot p_n = \sum_{n=0}^{\infty} n (1-\rho)\rho^n = (1-\rho) \sum_{n=0}^{\infty} n\rho^n$$

$$\overset{\text{G.S.}}{=} (1-\rho) \cdot \frac{\rho}{(1-\rho)^2} = \boxed{\frac{\rho}{1-\rho}} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda}.$$

$$E_N[N^2] = \sum_{n=0}^{\infty} n^2 \cdot p_n = \sum_{n=0}^{\infty} n^2 \cdot \rho^n (1-\rho) = (1-\rho) \sum_{n=0}^{\infty} n^2 \rho^2$$

$$\overset{\text{G.S.}}{=} (1-\rho) \cdot \frac{\rho+\rho^2}{(1-\rho)^3} = \frac{\rho+\rho^2}{(1-\rho)^2}$$

$$\therefore V_N[N] = E_N[N^2] - E_N^2[N] = \frac{\rho+\rho^2}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} = \boxed{\frac{\rho}{(1-\rho)^2}}$$

$$\therefore E_N[N] = \frac{\rho}{1-\rho} \quad \text{and} \quad V_N[N] = \frac{\rho}{(1-\rho)^2}$$

Ex: from above: $\lambda = 3$ and $\mu \approx 4$ $\quad \therefore \rho = 3/4$

$$\therefore E_N[N] = \frac{\rho}{1-\rho} = \frac{3/4}{1-3/4} = \frac{3/4}{1/4} = 3 \text{ customers}$$

$$V_N[N] = \frac{\rho}{(1-\rho)^2} = \frac{3/4}{(1-3/4)^2} = \frac{3/4}{1/16} = \frac{3 \cdot 16}{4} = 12$$

$$\therefore \sigma_N = \sqrt{V_N[N]} = \sqrt{12} = 2\sqrt{3} \text{ customers.}$$

Markov chains

Markov idea: Future = f(Present)

↑ does not depend on the Past.

$$f(x_{n+1} \mid x_1, \ldots, x_n) = f(x_{n+1} \mid x_n)$$

Multiplication Theorem: (always holds)

$$P\left(\bigcap_{k=1}^{n} A_k\right) = P(A_1) \cdot P(A_2 \mid A_1) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1})$$

Independence: $P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} P(A_k)$

Markovity lies in between

$$P\left(\bigcap_{k=1}^{n} A_k\right) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_2) \cdots P(A_n \mid A_{n-1})$$

DTFS: Discrete Time and Finite State

State: $S_1, S_2, \ldots, S_m$

Ex:
- Frog jumping from lily pad to lily pad ("random walk")
- Chess-board configuration
- Google "Page rank" algorithm (Larry page)
  - ↳ US Patent # 6,285,999 B1.

page rank ⟷ main eigenvector from Markov-chain matrix.

Defn: $X_1, X_2, \ldots$ is a Markov Chain iff

$$P(X_{k+1} = x_{k+1} \mid X_1 = x_1, \ldots, X_k = x_k) = P(X_{k+1} = x_{k+1} \mid X_k = x_k) \quad \forall k$$

Ex: iid sum $S_n = \sum_{k=1}^{n} X_k$ is a Markov Chain (infinite state space)

Prf: (SIT technique)

$$P(S_{n+1} = s_{n+1} \mid S_1 = s_1, \ldots, S_n = s_n)$$

$$= P(X_{n+1} + S_n = s_{n+1} \mid S_1 = s_1, \ldots, S_n = s_n)$$

$$\overset{\text{Sub}}{=} P(X_{n+1} + s_n = s_{n+1} \mid S_1 = s_1, \ldots, S_n = s_n)$$

$$= P(X_{n+1} = s_{n+1} - s_n \mid S_1 = s_1, \ldots, S_n = s_n)$$

$$\overset{\text{ind}}{=} P(X_{n+1} = s_{n+1} - s_n) \quad \text{by defn of } S_n$$

$$= P(X_{n+1} + s_n = s_{n+1})$$

$$\overset{\text{Sub}}{=} P(X_{n+1} + S_n = s_{n+1} \mid S_n = s_n)$$

$$= P(S_{n+1} = s_{n+1} \mid S_n = s_n)$$

$\therefore \{S_n\}$ is a Markov Chain

QED.

History: Andrei Markov – 1906 (St. Petersburg)

- Markov wanted to show that independence was not needed for the LLN and CLT
  He succeeded: there are Markov LLN and CLT.

- 1913: Analyzed 20,000 letters of Pushkin's poem
  Eugeny Onegin

found: $p(\text{vowel}) = 0.432$ ← stationary value

$p(\text{vowel} \mid \text{vowel just occured}) = 0.128 = P(v \to v)$

$p(\text{vowel} \mid \text{consonant just occured}) = 0.663 = P(c \to v)$

Simplest case: DTFS

n-states: $①, ②, \dots, ⓝ$ (lily pads or web pages)

$p_{ij} = P(ⓘ \to ⓙ) = P(X_{k+1} = j \mid X_k = i)$

↑ "transition probability"

Time-Homogeneous: $P(X_{k+1} = j \mid X_k = i) = p_{ij}$ for $\forall k$.

↑ $p_{ij} \geq 0$

Transition Matrix P:

$$P_{n\times n} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

← each row is a pdf ∴ stochastic matrix.

infinite matrix for infinite state Markov chain

So that each row sums to 1:

$$\sum_{k=1}^{n} p_{jk} = 1 \quad \text{for } \forall j.$$

∴ $P: S^n \longrightarrow S^n$ maps probability vector π to probability vector

(S^n: n-simplex)

∴ P has a fixed point pdf π:

$$\boxed{\pi P = \pi}$$

(from Brouwer's fixed point theorem since P is linear and continuous and since S^n is convex and compact (closed and bounded))

∴ $\underset{1\times n}{\pi}$ is an eigenvector with eigenvalue $\lambda = 1$

Fact: All other eigenvalues λ': $|\lambda'| < 1$.
↳ from Perron-Frobenius Theorem for non-negative matrices.

For any pdf u: $u(k) = u(0)P^k$
for time-evolution $u(0), u(1), \ldots, u(k)$
↳ from Chapman-Kolmogorov equations)

Defn: $P_{n\times n}$ is (regular) irreducible if $\exists k > 0$: P^k is a positive matrix: $p_{ij}^{(k)} > 0 \quad \forall i \forall j$

↑ cannot happen if P has a 1 on a main diagonal unless $\sum_{i\neq 1} p = 1$. $\therefore p_{kk} = 1 \longrightarrow$ ~irreducible

Defn: P is aperiodic iff the number of steps between any two states (i) and (j) is an integer multiple (not forced into fixed cycle).

Note: P is a periodic if no eigenvalue λ' : $\lambda' = -1$

★

Thm: ("AI") If the transition matrix $P_{n\times n}$ is (1) aperiodic and (2) irreducible then P has a stationary pdf π_e: (unique since finite n)

(1) $\pi_e P = \pi_e$

(2) $\lim_{k\to\infty} P^k = \begin{pmatrix} -\pi_e- \\ -\pi_e- \\ \vdots \\ -\pi_e- \end{pmatrix}$ "in the long run"

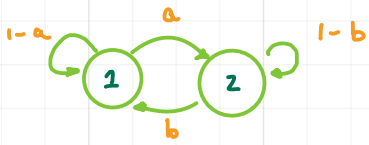
(3) $\lim_{k\to\infty} uP^k = \pi_e$ for $\forall u \in S^n$.

Basis of MCMC approximation (Gibbs sampler)

↳ identity problem with solution = π_e.

Ex: Special case: 2×2 P

$$P_{2\times 2} = \begin{array}{c} \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix} \end{array}$$



for $a, b \in [0,1]$.

$\therefore$ **Proposition:** $\pi_e = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)$ if P is 2×2

Prf:

$$\pi_e P = \left(\frac{b}{a+b}, \frac{a}{a+b}\right)\begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

$$= \left(\frac{b - ab + ab}{a+b}, \frac{ab + a - ab}{a+b}\right)$$

$$= \left(\frac{b}{a+b}, \frac{a}{a+b}\right) = \pi_e$$

$\therefore$ fixed point.

QED

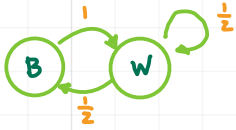
Ex: Suppose each day John walks (W) or bikes (B) to class. If John bikes one day then he never bikes the next day. If John walks one day then the next day he is equally likely to bike or walk.

Q: In the long run... how often does John bike to class?

A:

$$P = \begin{array}{c} \\ B \\ W \end{array} \begin{array}{c} \begin{array}{cc} B & W \end{array} \\ \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \end{array}$$ from above facts

$$= \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$



Note: $\lambda_1 = 1$ and $\lambda_2 = 1 - a - b = -\frac{1}{2} \neq -1$ $\therefore$ P is aperiodic (A)

$$P^2 = \begin{pmatrix} 1/2 & 1/2 \\ 1/4 & 3/4 \end{pmatrix}$$ ∴ all positive entries

∴ P is irreducible (I)

∴ Theorem applies for unique equilibrium pdf: π_e.

$$\pi_e = \left(\frac{\overset{B}{b}}{a+b}, \frac{\overset{W}{a}}{a+b}\right) = \left(\frac{1/2}{3/2}, \frac{1}{3/2}\right) = \left(\overset{B}{1/3}, \overset{W}{2/3}\right)$$

∴ Bike 1/3 the time.

Check: $\pi_e P = \left(\frac{1}{3}, \frac{2}{3}\right)\begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} = \left(\frac{2}{3}\cdot\frac{1}{2}, \frac{1}{3} + \frac{2}{3}\cdot\frac{1}{2}\right)$

$= \left(\frac{1}{3}, \frac{2}{3}\right)$

$= \pi_e$

Extension: Hidden Markov Model (HMM)
- train with E-M algorithm.

Ex: Google Page Rank Algorithm "random surfer on N web pages"

$r(A)$ = rank of page A

$$= \frac{\alpha}{N} + (1-\alpha)\cdot\left(\frac{r(B_1)}{|B_1|} + \cdots + \frac{r(B_n)}{|B_n|}\right)$$ ∴ $\Sigma r(A) = 1$

where $B_1, \ldots, B_n$ are the "backlink" pages of A with ranks $r(B_1), \ldots, r(B_n)$

- $\alpha \in [0,1]$
- $|B_k|$ = # "forward links" of B_k
- N = total # of pages on the web

$\alpha \approx P(\text{surfer jumps randomly to any web page})$

$\alpha =$ "damping factor" ← keeps algorithm from getting stuck in loops.

≈ 0.15

Put $\underset{N\times N}{P} = \frac{\alpha}{N} \underset{N\times N}{\mathbb{1}} + (1-\alpha)\underset{N\times N}{B}$ where ($\mathbb{1}$: matrix of all 1's)

- $b_{ij} = \begin{cases} \frac{1}{n_i} & \text{if node } i \text{ points to node } j \\ 0 & \text{else} \end{cases}$
- $n_i =$ total # of forward links from i

$\therefore \exists$ unique $\underset{1\times N}{\pi_e}$: $\pi_e P = \pi_e$ since P is aperiodic and irreducible.

$\therefore \pi_e = p_\infty$ is Google ranking of all N webpages

---

Markov chains $\{X_k\}$ obey forms of LLNs and CLTs if they are suitably "ergodic" (aperiodic and irreducible and positive recurrent) - even though the chain $X_1, X_2, \ldots$ is NOT independent.

Markov Law of Large Numbers (MLLLN)

If
- π is the stationary pdf
- Markov chain $X_1, X_2, \ldots$ is (Harris) ergodic
- $f: X \longrightarrow \mathbb{R}$ is measureable
- $E_\pi[|f|] < \infty$

Then with probability one ("a"):

$$\bar{f}_n(x) = \frac{1}{n} \cdot \sum_{k=1}^{n} f(x_k) \xrightarrow{a} E_\pi[f(x)]$$

With these and further technical conditions:

MC CLT:

$$\sqrt{n}\left(\bar{f}_n(x) - E_\pi[f]\right) \xrightarrow{d} W \sim N(0, \sigma_f^2)$$

if $\sigma_f^2 < \infty$ but now σ_f^2 has the form:

$$\sigma_f^2 = V_\pi[f(X_1)] + Cov_\pi[f(X_1), f(X_k)].$$

∴ Markov was right: independence is not necessary for LLN or CLT.

Optimal Estimators

Recall: ① Maximum Likelihood (ML)

$$\hat{\theta}_n^{ML} = \arg\max_\theta g(x|\theta)$$

- need closed form pdf g
- Consistency: $\hat{\theta}_n^{ML} \xrightarrow{P} \theta$ (smoothness condition)
- Asymptotic normality $\sqrt{n}(\hat{\theta}_n^{ML} - \theta) \xrightarrow{d} W \sim N(0, \frac{1}{I(\theta)})$ for Fisher Information I (Cramer-Rao)
- Invariance principle $\widehat{g(\theta)}^{ML} = g(\hat{\theta}^{ML})$

- Applications: - EM algorithm for missing data
  - Viterbi algorithm (dynamic programming on "hidden" Markov chain)

② Maximum A Posteriori (MAP)

$$\hat{\theta}_n^{MAP} = \arg\max_\theta f(\theta|x) \quad \text{if } \theta \sim h(\theta) \text{ and } f(\theta|x) \propto h(\theta) \cdot g(x|\theta)$$

($h(\theta)$: prior; $g(x|\theta)$: likelihood)

- MAP = MLE if $h(\theta) \sim$ uniform.
- Bayes minimization of squared error loss $\longrightarrow E[\theta|X]$ is optimal estimator (MSE result)
- Gibbs Sampler (MCMC)
- Subjectivity: who gives prior $h(\theta)$?

③ Mean square estimation (MSE)

$$\hat{\theta}_n^{ms} = E[\theta|X] \text{ for random } \theta \text{ (in general)}$$

focus of these notes.

Data set $\mathcal{X}_n = \{X_1, X_2, \ldots, X_n\}$ for random vectors X_k

Estimator $\underset{k\times 1}{\hat{\theta}_n} = g(\mathcal{X}_n)$

$= g(X_1, \ldots, X_n)$ $\therefore \hat{\theta}_n$ is a statistic.

Random Parameter $\underset{k\times 1}{\theta} \sim f(\theta_1, \ldots, \theta_k)$

generalizes deterministic case

Assume: Data $\mathcal{X}_n$ depends on θ

$\therefore$ there exists a joint pdf: $f(\theta, \mathcal{X}_n) = f(\theta, X_1, \ldots, X_n)$

★ **Thrm:** $\boxed{\underset{k\times 1}{\hat{\theta}_n^{ms}} = E[\underset{k\times 1}{\theta} | \mathcal{X}_n]}$ if $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and θ obey joint pdf $f(\theta, \mathcal{X}_n)$

Prf: $0 \le \underset{1\times 1}{MSE}[\hat{\theta}_n^{ms}]$

$= E_{\theta\mathcal{X}}[\underset{1\times k}{(\theta - \hat{\theta}_n^{ms})^T} \underset{k\times 1}{(\theta - \hat{\theta}_n^{ms})}]$

where $\theta\mathcal{X} \longleftrightarrow f(\theta, \mathcal{X}_n) = f(\theta_1, \ldots, \theta_n, X_1, \ldots X_n)$

total exp. $= E_{\mathcal{X}}[E[(\theta - \hat{\theta}_n^{ms})^T(\theta - \hat{\theta}_n^{ms}) | \mathcal{X}_n]]$.

linear $= E_{\mathcal{X}}[E[\theta^T\theta | \mathcal{X}_n] + E[\hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} | \mathcal{X}_n]$

$- E[\theta^T\hat{\theta}_n^{ms} | \mathcal{X}_n] - E[\hat{\theta}_n^{ms}\theta | \mathcal{X}_n]]$.

$$= E_x\Big[E[\theta^T\theta \mid \chi_n] + \hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} - E[\theta^T \mid \chi_n]\cdot\hat{\theta}_n^{ms} - \hat{\theta}_n^{ms\,T} E[\theta \mid \chi_n]\Big].$$

since $\hat{\theta}_n^{ms} = g(\chi_n)$ and $E[g(x)z \mid x] = g(x)\cdot E[z \mid x]$

comp. the square $= E_x\Big[E[\theta^T\theta \mid \chi_n] - E[\theta^T \mid \chi_n]\cdot E[\theta \mid \chi_n] + \big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

since $\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)$

$= \hat{\theta}_n^{ms\,T}\hat{\theta}_n^{ms} - \hat{\theta}_n^{ms\,T}E[\theta \mid \chi_n] - E[\theta^T \mid \chi_n]\,\hat{\theta}_n^{ms} + E[\theta^T \mid \chi_n]\cdot E[\theta \mid \chi_n]$

linear $= E_x\Big[E[\theta^T\theta \mid \chi_n] - E[\theta^T \mid \chi_n] E[\theta \mid \chi_n]\Big] + E_x\Big[\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

↑ only this term depends on $\hat{\theta}_n^{ms}$

$\therefore\ \hat{\theta}_n^{ms}$ minimizes $MSE[\hat{\theta}_n^{ms}] \geq 0$ if it minimizes $E_x\Big[\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)^T\big(\hat{\theta}_n^{ms} - E[\theta \mid \chi_n]\big)\Big]$

≥ 0.

Latter occurs iff $\hat{\theta}_n^{ms} = E[\theta \mid \chi_n]$.

QED

Same result if minimize $E_{\theta x}\big[(\theta - \hat{\theta}_n^{ms})^T W (\theta - \hat{\theta}_n^{ms})\big]$.

weight matrix $W = W^T$ is positive semi-definite

Unbiased: $E_x[\hat{\theta}_n^{ms}] \overset{\text{thm}}{=} E_x[E[\theta|X_n]] \overset{\text{total exp.}}{=} E_\theta[\theta] \quad \forall n.$

★ **Thrm:** (Orthogonality Principle) MSE error is orthogonal to data (functions of data): $\mathcal{E}^{ms} \perp f(X)$.

$$E_{\theta x}[(\theta - \hat{\theta}_n^{ms}) g^T(X_n)] = 0$$

for ANY function $g(X_n) = g(X_1, \ldots, X_n)$

- applications of orthogonal projections.
  - Wiener-filter (Wiener-Hopf equation) EE562
  - Yule-Walker equations EE583
  - Kalman-filter (linear Gauss-Markov model)

Prf:

$$E_{\theta x}[(\theta - \hat{\theta}_n^{ms}) g^T(X_n)]$$

$$\overset{\text{total exp.}}{=} E_x[E[(\theta - \hat{\theta}_n^{ms}) g^T(X_n) | X_n]]$$

$$= E_x[E[\theta - \hat{\theta}_n^{ms} | X_n] \cdot g(X_n)].$$

since $E[g(X)Y|X] = g(X) \cdot E[Y|X]$

$$\overset{\text{MSE thm.}}{=} E_x[E[\theta - E[\theta|X_n] | X_n] g^T(X_n)]$$

$$\overset{\text{linear}}{=} E_x[(E[\theta|X_n] - E[\theta|X_n] \cdot E[1|X_n]) g^T(X_n)].$$

since $E[\theta|X_n] = \phi(X_n)$ and $E[\phi(X_n)|X_n] = \phi(X_n) \cdot E[1|X_n]$

$$= E_x\left[\left(\underbrace{E[\theta|X_n] - E[\theta|X_n]}_{=0} \right) g^T(X_n) \right].$$

$$= E_x[\, 0 \; g^T(X_n)]$$

$$= 0$$

QED

- Hard to find closed form $E[\theta|X_n]$ in general
- $\hat{\theta}_n^{MS}$ is <u>not robust</u> to outliers or impulsive data (unlike conditional median)

Special (but common) case: Jointly-Gaussian $\underset{n\times 1}{X}$ and $\underset{p\times 1}{Y}$

$$E[\underset{n\times 1}{X}|Y] = \underset{n\times 1}{\mu_X} + \underset{n\times p}{K_{XY}} \underset{p\times p}{K_{YY}^{-1}} (\underset{p\times 1}{Y} - \mu_Y)$$

$$\underset{n\times n}{K_{X|Y}} = \underset{n\times n}{K_{XX}} - \underset{n\times p}{K_{XY}} \underset{p\times p}{K_{YY}^{-1}} \underset{p\times n}{K_{YX}}$$

<u>Warning</u>: Must <u>TEST</u> for normality (EE517)

<u>Ex</u>: Bayesian conjugacy for finding $E[\theta|X]$.
(proved in Week 5: "beta is conjugate to binomial")

<u>Thm</u>: If $\theta \sim h(\theta) = \text{Beta}(\alpha, \beta)$
$g(x|\theta) = \text{binomial}(n, x, \theta)$ ← Bernoulli trials

then $f(\theta|x) = \text{Beta}(\alpha + x, \beta + n - x)$

where θ estimates "success probability" p after n-trials and n successes (∴ $n-x$ failures)

Recall: Also proved $E[\theta] = \frac{\alpha}{\alpha+\beta}$ if $\theta \sim \text{Beta}(\alpha, \beta)$

$\therefore$ For squared-error loss and $f(\theta|x)$

$$\hat{\theta}_n^{ms} = E[\theta \mid X=x]$$

$$\overset{\text{conjugacy}}{=} \frac{\alpha'}{\alpha'+\beta'} \quad \text{since } f(\theta|x) = \text{Beta}(\alpha', \beta')$$

$$= \frac{\alpha+x}{\alpha+x+\beta+n-x} \quad \text{since } f(\theta|x) = \text{Beta}(\alpha+x, \beta+n-x)$$

$$= \boxed{\frac{\alpha+x}{\alpha+\beta+n}} \leftarrow \text{closed form answer.}$$

$$= \frac{\alpha}{\alpha+\beta+n} + \frac{x}{\alpha+\beta+n}$$

$$= \frac{\alpha}{\alpha+\beta+n} \cdot \frac{\alpha+\beta}{\alpha+\beta} + \frac{x}{\alpha+\beta+n} \cdot \frac{n}{n}$$

$$= \underbrace{\frac{\alpha+\beta}{\alpha+\beta+n}}_{\lambda_n} \cdot \frac{\alpha}{\alpha+\beta} + \underbrace{\frac{n}{\alpha+\beta+n}}_{1-\lambda_n} \cdot \frac{x}{n}$$

$$= \lambda_n \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{E[\theta]} + (1-\lambda_n) \underbrace{\frac{x}{n}}_{\bar{X}_n = \hat{\theta}_n^{ml}} \quad \text{for convex coefficients } 0 \le \lambda_n \le 1$$

$$= \lambda_n \cdot E[\theta] + (1-\lambda_n)\bar{X}_n$$

$\longrightarrow \bar{X}_n$ as $n \longrightarrow \infty$ because $\lambda_n = \frac{\alpha+\beta}{\alpha+\beta+n} \downarrow 0$

$\therefore 1-\lambda_n \longrightarrow 1.$

$\therefore$ Effect $E[\theta]$ of subjective prior "washes out" for large n.

$\therefore \hat{\theta}_n^{ms}$ is asymptotically unbiased for θ. since

$$E[\bar{X}_n] = \frac{E[X]}{n} = \frac{np}{n} = p \quad \text{for binomial } n$$

$$- V[E[\theta|X]] = V\left[\frac{\alpha+X}{\alpha+\beta+n}\right]$$

$$= \frac{1}{(\alpha+\beta+n)^2} V[X]$$

$$= \frac{npq}{(\alpha+\beta+n)^2} \quad \text{since } X \sim \text{binomial}$$

$$\downarrow 0 \quad \text{as} \quad n \uparrow \infty$$

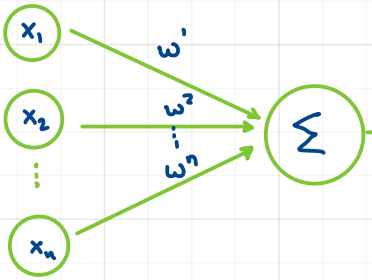
$$\therefore E[\theta|X] \xrightarrow{m} p$$

$\therefore$ Consistent: $E[\theta|X] \xrightarrow{\text{plim}} p$ by $m \longrightarrow p$

Adaptive MS Estimation (LMS) EE583

"Linear combiner" → adaptive noise suppression
→ adaptive equalizer and modems

Weights: $w \in \mathbb{R}^n$



desired output $d_j \in \mathbb{R}$

Learn MS optimal w^* from $(x_1, d_1), \ldots, (x_m, d_m)$

r.v. $\varepsilon_k = d_k - y_k = d_k - x_k^T w$

$\therefore$ MSE $= J_k(w) = E[\varepsilon_k^2]$

← unknown pdf $p(x)$

$\therefore \nabla_w J = 0 : \; w^* = R^{-1} P$

$= (E[X_k X_k^T])^{-1} E[d_k x_k^T]$

unknown pdfs

★

LMS idea: $E[\varepsilon_k^2] \approx \varepsilon_k^2$ (Widrow-Hopf, late 1950's)

single realization of r.v.

∴ **LMS algorithm**

LMS estimate

$$w_{k+1} = w_k - \mu \hat{\nabla}_w J \leftarrow -2\varepsilon_k x_k \text{ (derivative below)}$$

constant learning rate

general GRADIENT DESCENT algorithm

$$= \boxed{w_k + 2\mu \varepsilon_k x_k} \quad \star$$

↑ arguably most widely used EE algorithm.

if WSS process (wide-sense stationary) EE562

Not robust to outliers or impulsive ($S\alpha S$-type) noise

Robust LMS: $\varepsilon_k \longmapsto \text{sign}(\varepsilon_k)$ "signed LMS"

$S\alpha S$ noise: $\alpha = 2$ Gaussian; $\alpha = 1$ Cauchy

$\alpha \downarrow \Longrightarrow$ tail-thickness ↑ ($\alpha < 2 \longrightarrow$ infinite variance)

∴ LMAD: "least mean absolute deviation"

- LMS similar to CENTROID learning in clustering models (pattern recognition)
- LMS: constant learning rate μ (unlike most neural network algorithms) ↔ still converges if WSS signal

Fact: $w_k \xrightarrow{\text{LMS}} w^*$ if $\boxed{0 < \mu < \frac{1}{\lambda_{max}}}$

eigenvalues of R: $0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_{max}$

$$\hat{\nabla}_w J = \nabla_w \varepsilon_k^2 \qquad \text{since LMS idea } E[\varepsilon_k^2] = \varepsilon_k^2$$

$$= \left(\frac{\partial \varepsilon_k^2}{\partial w^1}, \cdots, \frac{\partial \varepsilon_k^2}{\partial w^n} \right)$$

$$= \left(\frac{\partial}{\partial w^1}(d_k - y_k)^2, \cdots, \frac{\partial}{\partial w^n}(d_k - y_k)^2 \right)$$

$$= \left(-2\underbrace{(d_k - y_k)}_{\varepsilon_k} \frac{\partial y_k}{\partial w^1}, \cdots, -2\underbrace{(d_k - y_k)}_{\varepsilon_k} \frac{\partial y_k}{\partial w^n} \right)$$

$$= -2\varepsilon_k \left(\frac{\partial y_k}{\partial w^1}, \cdots, \frac{\partial y_k}{\partial w^n} \right)$$

$$= -2\varepsilon_k (x_k^1, \cdots, x_k^n) \qquad \text{since } y_k = x_k^T w = \sum_{i=1}^{n} x_k^i w^i$$

$$= -2\varepsilon_k x_k$$

QED.

$$\therefore \; w_{k+1} = w_k + 2\mu\varepsilon_k x_k \qquad \text{(LMS)}$$

Simple Linear Regression (OLS)

$$y_k = f(x_k)$$

$$= \underbrace{a}_{\text{constant}} + b x_k + \underbrace{\varepsilon_k}_{\text{noise}}$$

ε_k: modelling error

$$\triangleq \beta_0 + \beta_1 x_k + \varepsilon_k .$$

y_k is a (normal) r.v. because iid $\varepsilon_k \sim N(0, \sigma^2)$

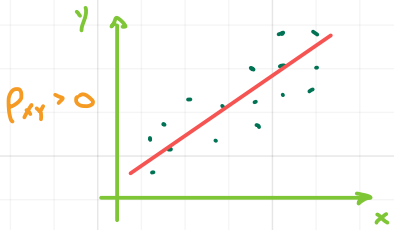
"TRUE" model: $y = a + bx$

Result: ★ $a_n^{LS} = \bar{y}_n - \hat{b}^{LS} \bar{x}_n$

$= \hat{\beta}_0$ unbiased and consistent.

★ $b_n^{LS} = \frac{S_{xy}}{S_x^2} = R_{xy} \cdot \frac{S_y}{S_x}$

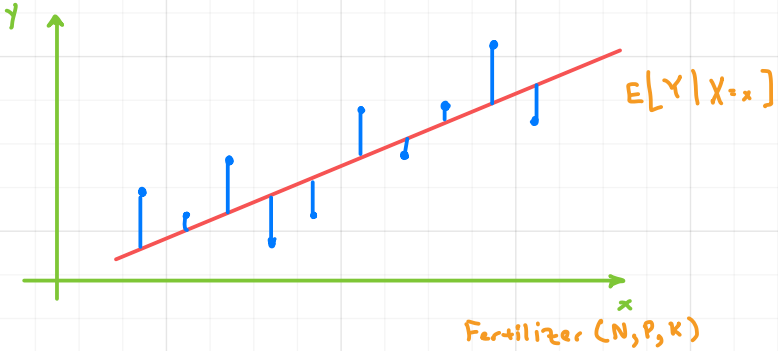
$= \hat{\beta}_1$ unbiased and consistent.



Least Squares

Model: $y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$

Data: $(x_1, y_1), \ldots, (x_n, y_n)$; iid $\varepsilon_k \sim N(0, \sigma^2)$



Check: iid $\varepsilon_k \sim N(0, \sigma^2)$

Independence All ε_k independent

- "serial correlation" → time series (AR(1))
  Durbin-Watson test.

Homoscedasticity Constant σ_k^2

- pattern in residuals?

Normality $y_k - \hat{y}_k \sim N$

$K_{\varepsilon\varepsilon} = \sigma^2 I = \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

else look at WLS.

Find: $\hat{\theta}_n^{ms} = \underset{2\times 1}{\hat{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

Minimize J:

$$J = SSE = \sum_{k=1}^{n} \varepsilon_k^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

$$\overset{LS}{=} \sum_{k=1}^{n} (y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))^2$$

$$\therefore \quad \odot = \nabla_\beta J = \begin{pmatrix} \partial J/\partial \beta_0 \\ \partial J/\partial \beta_1 \end{pmatrix}$$

$$= \begin{cases} \sum_{k=1}^{n} 2(y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))(-1) \\ \sum_{k=1}^{n} 2(y_k - (\hat{\beta}_0 + \hat{\beta}_1 x_k))(-x_k) \end{cases}$$

Solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ (positive definite Hessian matrix)

OLS Normal Equations (Ordinary Least Squares)

$$\underset{(k+1)\times 1}{\hat{\beta}^{LS}} = (X^T X)^{-1} X^T Y \quad \bigstar$$

Observation matrix

$$\underset{n\times 1}{Y} = \underbrace{\underset{n\times(k+1)}{X}\ \underset{(k+1)\times 1}{\beta}}_{n\times 1} + \underset{n\times 1}{\varepsilon}$$

$$y_k = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_k$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$Y = X \beta + \varepsilon$$

- $\varepsilon \sim N(0, \sigma^2 I)$

∴ White and normal and homoscedastic (additive noise)

Recall:

(1) $\frac{d}{dm}\underbrace{(\underset{1\times n}{b^T}\ \underset{n\times 1}{m})}_{1\times 1} = \underset{n\times 1}{b}$ (linear form) $\quad \left(\frac{d}{dx} cx = c\right)$

(2) If $A = A^T$: $\frac{d}{dm}\underbrace{(\underset{1\times n}{m^T}\ \underset{n\times n}{A}\ \underset{n\times 1}{m})}_{1\times 1} = \underbrace{2\ \underset{n\times n}{A}\ \underset{n\times 1}{m}}_{n\times 1}$ (quadratic form) $\quad \left(\frac{d}{dx} x^2 = 2x\right)$

Minimize $(Y-\hat{Y})^T(Y-\hat{Y})$

$$J \triangleq SSE \quad \hat{Y} = \underset{1\times n}{(Y - X\beta)^T} \underset{n\times 1}{(Y - X\beta)} \quad \text{scalar.}$$

$$= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta.$$

$$\therefore \nabla_\beta J = 0 = -2X^TY + 2X^TX\hat{\beta}.$$

$\longleftrightarrow$ "Normal Equations"

$$\boxed{(X^TX)\,\hat{\beta} = X^TY} \quad \star$$

$$\therefore \hat{\beta}^{LS} = (X^TX)^{-1}X^TY = C \cdot Y \quad \leftarrow \text{linear in } Y.$$

if $(X^TX)^{-1}$ exists, iff $\det(X^TX) \neq 0$

iff $\forall$ eigenvalues $\lambda > 0$

$$X^TX = \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum x_{i2}x_{i1} & & \vdots \\ \vdots & \vdots & \ddots & \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

$\therefore X^TX$ is real and symmetric. $(X^TX)^T = X^T(X^T)^T = X^TX.$

$\therefore$ for $\forall w \neq 0$

$$\underbrace{w^T (X^T X) w}_{1\times 1} \overset{\text{assoc.}}{=} (w^T X^T)(Xw)$$

$$= (Xw)^T (Xw)$$

$$= z^T \cdot z \qquad \text{if } z = Xw$$

$$\geq 0 \qquad \text{since } z \text{ real vector}$$

$\therefore X^T X \geq 0$ Positive Semi-definite

$\Longleftrightarrow$ all eigenvalues $\lambda \in \mathbb{R}^+ : \lambda_i \geq 0$

$\therefore \det(X^T X) = \prod_{i=1}^{k+1} \lambda_i \geq 0.$

- Problems with numerical stability if λ_i "small" (multicolinearity)
- Pseudo-inverse always exists and is unique

Thm: $Y = X\beta + \varepsilon$ and $\varepsilon \sim N(0, \sigma^2 I)$ and

$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y$ then

$$\boxed{\hat{\beta}^{LS} \sim N(\beta, \sigma^2 (X^T X)^{-1})} \quad \star$$

Prf:

(1) **Normality**

$$\hat{\beta} = (X^TX)^{-1}X^TY = C \cdot Y$$ - linear function of Y

But $Y = X\beta + \varepsilon \sim N$ because $\varepsilon \sim N(0, \sigma^2 I)$

$\therefore \hat{\beta} \sim N$ QED (1)

(2) **Unbiasedness**

$$E[\hat{\beta}] \overset{LS}{=} E[(X^TX)^{-1}X^TY]$$

$$\overset{\text{def}Y}{=} E[(X^TX)^{-1}X^T(X\beta+\varepsilon)].$$

$$= E[(X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon]$$

$$\overset{\text{linear}}{=} E[\underbrace{(X^TX)^{-1}(X^TX)}_{=I}\beta] + E[(X^TX)^{-1}X^T\varepsilon].$$

$$= E[\beta] + E[(X^TX)^{-1}X^T\varepsilon].$$

$$= \beta + E[(X^TX)^{-1}X^T\varepsilon]$$ since β not random

$$= \beta + (X^TX)^{-1}X^T E[\varepsilon]$$ since X not random

$$= \beta + \underline{0}$$ since $\varepsilon \sim N(\underline{0}, \sigma^2 I)$

$$= \beta$$

$\therefore \hat{\beta}$ is an unbiased estimator for β.

QED (2)

(3) **Covariance** (Consistency w/ Slutsky's Theorem)

$$K_{\hat\beta\hat\beta} = E\left[(\hat\beta - E[\hat\beta])(\hat\beta - E[\hat\beta])^T\right].$$

First, find K_{YY}:

$$K_{YY} = E\left[(Y - E[Y])(Y - E[Y])^T\right].$$

$$= E\left[(Y - X\beta)(Y - X\beta)^T\right] \quad \text{since } E[Y] = E[X\beta + \varepsilon] = X\beta + E[\varepsilon] = X\beta.$$

$$= E\left[\varepsilon\varepsilon^T\right] \quad \text{since } \varepsilon = Y - X\beta.$$

$$= E\left[(\varepsilon - \Theta)(\varepsilon - \Theta)^T\right]. \quad \text{by Hypo.}$$

$$= E\left[(\varepsilon - E[\varepsilon])(\varepsilon - E[\varepsilon])^T\right] \quad \text{since } \varepsilon \sim N(\Theta, \sigma^2 I)$$

$$\overset{\text{def}}{=} K_{\varepsilon\varepsilon}$$

$$\overset{\text{hypo.}}{=} \sigma^2 I \quad \text{since } \varepsilon \sim N(\Theta, \sigma^2 I)$$

$$\therefore K_{YY} = \sigma^2 I$$

$$K_{\hat\beta\hat\beta} = E\left[(\hat\beta - E[\hat\beta])(\hat\beta - E[\hat\beta])^T\right]$$

$$\overset{(3)}{=} E\left[(\hat\beta - \beta)(\hat\beta - \beta)^T\right] \quad \text{since } E[\hat\beta] = \beta. \text{ unbiased}$$

$$\overset{LS}{=} E\left[\left((X^TX)^{-1}X^TY - \beta\right)\left((X^TX)^{-1}X^TY - \beta\right)^T\right].$$

$$\overset{\text{"trick"}}{=} E\left[\left((X^TX)^{-1}X^TY - \underbrace{(X^TX)^{-1}X^TX}_{=I}\beta\right)\left((X^TX)^{-1}X^TY - \underbrace{(X^TX)^{-1}X^TX}_{=I}\beta\right)^T\right] \quad \text{since } \hat\beta = (X^TX)^{-1}X^TY.$$

dist

$$= E\left[\left((X^TX)^{-1}X^T(Y-\beta)\right)\left((X^TX)^{-1}X^T(Y-\beta)\right)^T\right]$$

$$= (X^TX)^{-1}X^T \cdot E\left[(Y-X\beta)(Y-X\beta)^T\right](X^TX)^{-1}X^T)^T$$

T

$$= (X^TX)X^T E\left[(Y-X\beta)(Y-X\beta)^T\right] \times (X^TX)^{-1}$$

since X is not random

since: $((X^TX)^{-1}X^T)^T = (X^T)^T((X^TX)^{-1})^T$

$= X \cdot ((X^TX)^T)^{-1}$

$= X \cdot (X^T(X^T)^T)^{-1}$

$= X \cdot (X^TX)^{-1}$

$$= (X^TX)^{-1}X^T E\left[(Y-E[Y])(Y-E[Y])^T\right] \cdot X(X^TX)^{-1}$$

since $E[Y] = E[X\beta+\varepsilon] = X\beta$.

$$= (X^TX)^{-1}X^T K_{YY} X (X^TX)^{-1}$$

lemma

$$= (X^TX)^{-1}X^T(\sigma^2 I) X (X^TX)^{-1}$$

$$= \sigma^2 \cdot (X^TX)^{-1} \underbrace{X^TX}_{=I} (X^TX)^{-1}$$

$$= \sigma^2 \cdot (X^TX)^{-1}$$

QED

$\therefore \hat{\beta} \xrightarrow{P} \beta$ consistent. by "m-p-d" (Slutsky)

Note: asymptotic normality (CLT) if only iid ε_k and $E[\varepsilon_k] = 0$

Thm: $\hat{\beta}^{LS} = \hat{\beta}^{ML}$ $\therefore$ invarance principle holds.

$$Y \sim N(X\beta, \sigma^2 I)$$

Prf:

$$L(\beta, \sigma^2) \overset{N}{=} \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\tfrac{1}{2}(Y-X\beta)^T K_{YY}^{-1} (Y-X\beta)\right]. \quad K_{YY} = \sigma^2 I.$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T I (Y-X\beta)\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(Y-X\beta)^T (Y-X\beta)\right].$$

$\therefore$ Maximize L $\longleftrightarrow$ minimize exponent.

$\longleftrightarrow$ minimize J

where $J = \varepsilon^T\varepsilon = (Y-X\beta)^T(Y-X\beta)$

$\therefore$ gives $\hat{\beta}^{LS}$

QED.

Similarly,

Thm: $\hat{\sigma}^{2\,ML} = \dfrac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{n}$

Weighted Least Squares (WLS)

Positive definite $W \gg 0$ often diagonal $W = \begin{pmatrix} w_{11} & & 0 \\ & w_{22} & \\ 0 & & \ddots \\ & & & w_{nn} \end{pmatrix}$

$\therefore$ W has a "square root" Q: $W = Q \cdot Q^T$
- find with Cholesky Decomposition, $O(n^3)$
- Q lower triangular

W can correct over-weighted errors if errors are uncorrelated (white) but Heteroscedastic

$$\hat{\beta}^{WLS} = (X^T W X)^{-1} X^T W Y \qquad (W = I)$$

$$= (X^T K_{\varepsilon\varepsilon}^{-1} X)^{-1} X^T K_{\varepsilon\varepsilon}^{-1} Y \qquad \text{if } W = K_{\varepsilon\varepsilon}^{-1}$$

where $K_{\varepsilon\varepsilon}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & 0 \\ & 1/\sigma_2^2 & \\ 0 & & \ddots \\ & & & 1/\sigma_n^2 \end{pmatrix}$ since $K_{\varepsilon\varepsilon}$ diagonal (uncorrelated errors)

Generalized Least Squares (GLS)

- Apply if errors heteroscedastic (non-constant variance)

  or correlated ($K_{\varepsilon\varepsilon} \neq$ diagonal)

Use WLS and positive-definite $K_{\varepsilon\varepsilon}^{-1}$ or estimate

$$\boxed{W = K_{\varepsilon\varepsilon}^{-1}} = (QQ^T)^{-1} \quad \text{since } K_{\varepsilon\varepsilon} = QQ^T$$

$\therefore$ Scale with Q^{-1} to <u>decouple</u>

Set $\varepsilon^* = Q^{-1}\varepsilon$

$\therefore K_{\varepsilon^*\varepsilon^*} = Q^{-1} K_{\varepsilon\varepsilon} Q = Q^{-1}(QQ^T)(Q^{-1})^T = I$

$\therefore \varepsilon^*$ white ($\therefore$ Gauss-Markov holds.) BLUE.

$\therefore J^{GLS} = (Y - X\beta)^T K_{\varepsilon\varepsilon}^{-1} (Y - X\beta)$ ← Malhalanobis distance

$= (\underbrace{Q^{-1}Y}_{Y^*} - (\underbrace{Q^{-1}X}_{X^*})\beta)^T (Q^{-1}Y - (Q^{-1}X)\beta)$

$= (Y^* - X^*\beta)^T (Y^* - X^*\beta)$

$\therefore$ OLS again with scaling $Y^* = Q^{-1}Y$ and $X^* = Q^{-1}X$.