通知：
1、会议中请使用真实姓名
2、会议开始后将锁定会议
3、会议slides已发布至piazza最新note，可自行下载

Conditional Expectation

SI252 Reinforcement Learning

School of Information Science and Technology
ShanghaiTech University

March 18, 2020

Law of Total Expectation

Theorem (Law of Total Expectation)

Let $A_1, \ldots, A_n$ be a partition of a sample space, with $\Pr(A_i) > 0$ for all i, and let Y be a random variable on this sample space. Then

$$\mathbb{E}(Y) = \sum_{i=1}^{n} \mathbb{E}(Y|A_i) \Pr(A_i).$$

Example: Two-envelope Paradox

Example (Two-envelope Paradox)

A stranger presents you with two identical-looking, sealed envelopes, each of which contains a check for some positive amount of money. You are informed that one of the envelopes contains exactly twice as much money as the other. You can choose either envelope. Which do you prefer: the one on the left or the one on the right? (Assume that the expected amount of money in each envelope is finite – certainly a good assumption in the real world!)

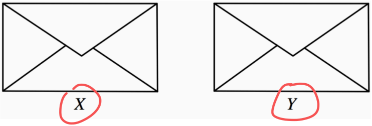


FIGURE 9.1
Two envelopes, where one contains twice as much money as the other. Either $Y = 2X$ or $Y = X/2$, with equal probabilities. Which would you prefer?

Example: Two-envelope Paradox (Solution)

$$E(Y) = E(Y \mid Y=2X)\cdot P_r(Y=2X) + E(Y \mid Y=\tfrac{1}{2}X)\, P_r(Y=\tfrac{1}{2}X)$$

$$= \underbrace{\tfrac{1}{2}E(Y \mid Y=2X)}_{\times} + \tfrac{1}{2}\underbrace{E(Y \mid Y=\tfrac{1}{2}X)}_{\times}$$

$$= \tfrac{1}{2}\underline{E(2X)} + \tfrac{1}{2}E(\underline{\tfrac{1}{2}X}) = 1.25E(X) > E(X)$$

$$E(X) \stackrel{?}{=} E(Y)$$

$\underline{X, Y \le 100}$

$Y \qquad Y=2X \qquad Y=\tfrac{1}{2}X$

$Y = 100 \qquad \underline{Y=2X}$

$X \le 50$

Example: Geometric Expectation Redux

Consider a sequence of iid coin flips Head w.p. p

X : number of Tails before the first Heads

$Pr(X=k) = (1-p)^k \cdot p$

$E(X) = \sum_{k=0}^{\infty} (1-p)^k p \cdot k = \frac{q}{p}$

Example (Geometric Expectation Redux)

Let $X \sim \mathrm{Geom}(p)$. Derive $\mathbb{E}(X)$ using first-step analysis.

First-step analysis ; $q \triangleq 1-p$

$E(X) = E(X \mid \text{first toss Head}) \cdot p + E(X \mid \text{first toss Tail}) \cdot q$

$= 0 \cdot p + (E(X)+1) \cdot q$

$E(X) = \frac{q}{p}$

Example: Time until HH vs. HT

Example (Time until HH vs. HT)

You toss a fair coin repeatedly. What is the expected number of tosses until the pattern HT appears for the first time? What about the expected number of tosses until HH appears for the first time?

Example: Time until HH vs. HT (Solution)

(1) HT

$W_{HT} \triangleq$ number of tosses until HT appears

$W_1 \triangleq$ waiting time for the first H $\qquad W_1 \sim FS(\frac{1}{2})$

$W_2 \triangleq$ additional waiting time for the first T $\qquad W_2 \sim FS(\frac{1}{2})$

$$\underbrace{TTTH}_{W_1}\underbrace{HHHT}_{W_2}HHT\ldots$$

$$E(W_1) = E(W_2) = \frac{1}{\frac{1}{2}} = 2$$

$$E[W_{HT}] = E[W_1 + W_2] = E(W_1) + E(W_2) = 4$$

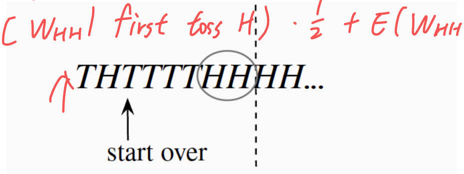
Example: Time until HH vs. HT (Solution)

(2) HH

$W_{HH} \triangleq$ number of tosses until HH appears

First-step analysis:

$$E[W_{HH}] = E[W_{HH} \mid \text{first toss } H) \cdot \frac{1}{2} + E(W_{HH} \mid \text{first toss } T) \cdot \frac{1}{2}$$



$$E[W_{HH} \mid \text{first toss } T] = 1 + E[W_{HH}]$$

$$E(W_{HH} \mid \text{first toss } H] = E[W_{HH} \mid \text{first toss } H, \text{ second toss } H) \cdot \frac{1}{2} + E[W_{HH} \mid \text{first toss } H, \text{ second toss } T) \cdot \frac{1}{2}$$

$$= 2 \times \frac{1}{2} + (E[W_{HH}] + 2) \times \frac{1}{2}$$

$$E[W_{HH}] = 6$$

Example: Mystery Prize

Example (Mystery Prize)

You have an opportunity to bid on a mystery box containing a mystery prize! The value of the prize is completely unknown. The true value V of the prize is considered to be Uniform on $[0, 1]$ (measured in millions of dollars). Specifically, if $b < 2V/3$, then the bid is rejected and nothing is gained or lost. If $b \geq 2V/3$, then the bid is accepted and your net payoff is $V - b$ (since you pay b to get a prize worth V). What is your optimal bid b, to maximize the expected payoff?

$V \sim \text{Unif}(0, 1)$

Example: Mystery Prize (Solution)

W : payoff

$$E(W) = E(W \mid b \geq \tfrac{2}{3}V)\, Pr(b \geq \tfrac{2}{3}V) + E(W \mid b < \tfrac{2}{3}V)\, Pr(b < \tfrac{2}{3}V)$$

$$= E(V - b \mid b \geq \tfrac{2}{3}V)\, Pr(b \geq \tfrac{2}{3}V) + 0$$

$$= [E(V \mid V \leq \tfrac{3}{2}b) - b]\, Pr(V \leq \tfrac{3}{2}b)$$

(i) $b \geq \frac{2}{3} \Rightarrow \frac{3}{2}b \geq 1 \qquad Pr(V \leq \frac{3}{2}b) = 1$

$$E(W) = (E(V) - b) \cdot 1 = (\tfrac{1}{2} - b) < 0$$

(ii) $0 \leq b < \frac{2}{3} \Rightarrow 0 \leq \frac{3}{2}b < 1 \qquad Pr(V \leq \frac{3}{2}b) = \frac{3}{2}b$

$$V \mid V \leq \tfrac{3}{2}b \sim Unif(0, \tfrac{3}{2}b)$$

$$E(W) = (\tfrac{3}{4}b - b) \times (\tfrac{3}{2}b) = -\tfrac{3}{8}b^2 \leq 0$$

$b = 0$

Conditional Variance

$Var(Y) = E[(Y - E(Y))^2]$

Definition (Conditional Variance)

The **conditional variance of** Y **given** X is

$$\mathrm{Var}(Y|X) = E\left((Y - E(Y|X))^2 \middle| X\right).$$

This is equivalent to

$$\mathrm{Var}(Y|X) = E(Y^2|X) - (E(Y|X))^2.$$

Example

① $Var(Y|Z) = Var(Z^2|Z) = 0$

② $Var(Z|Y) = E(Z^2|Y) - (E(Z|Y))^2 = E(Z^2|Z^2) - (E(Z|Z^2))^2$
$= Z^2 - 0 = Z^2 = Y$

Example

Let $Z \sim \mathcal{N}(0,1)$ and $Y = Z^2$. Find $\mathrm{Var}(Y|Z)$ and $\mathrm{Var}(Z|Y)$.

$Z | Z^2 = a = \begin{cases} \sqrt{a} & w.p.\ \frac{1}{2} \\ -\sqrt{a} & w.p.\ \frac{1}{2} \end{cases}$ $\quad E(Z|Z^2) = \sqrt{a} \cdot \frac{1}{2} + (-\sqrt{a})\frac{1}{2} = 0$

Linear Least Square Estimate

Theorem (Linear Least Square Estimate)

The Linear Least Square Estimate (LLSE) of X given Y, denoted by $L[X|Y]$, is the linear function $a + bY$ that minimizes $E[(X - a - bY)^2]$. In fact

$$L[X|Y] = E(X) + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}(Y - E(Y))$$

Example: Linear Signal Detection Over Noise

Example (Linear Signal Detection Over Noise)

X is a signal transmitted over a noise channel. The observation $Y = \alpha X + Z$, where Z is the noise, α is the scale coefficient. Both X and Z are zero-mean and independent. Find $L[X|Y]$.

Example: Linear Signal Detection Over Noise (Solution)

$$L[X|Y] = E(X) + \frac{Cov(X,Y)}{Var(Y)}(Y - E(Y))$$

$E(X) = E(Z) = 0$ X, Z independent

$E(Y) = E(\alpha X + Z) = \alpha E(X) + E(Z) = 0$

$$Cov(X,Y) = E(XY) - \underbrace{E(X)}_{=0} E(Y) = E(XY) = E(X(\alpha X + Z))$$

$$= \alpha E(X^2) + E(XZ) = \alpha E(X^2) + E(X)E(Z) = \alpha E(X^2)$$

$$Var(Y) = Var(\alpha X + Z) = \alpha^2 Var(X) + Var(Z) = \alpha^2 E(X^2) + E(Z^2)$$

$$L[X|Y] = \frac{\alpha E(X^2)}{\alpha^2 E(X^2) + E(Z^2)} \cdot Y = \frac{1}{\alpha} Y \cdot \frac{1}{1 + \frac{1}{\frac{\alpha^2 E(X^2)}{E(Z^2)}}}$$

$\frac{\alpha^2 E(X^2)}{E(Z^2)} \triangleq SNR$

signal-to-noise ratio

Example: Linear Signal Detection Over Noise (Solution)

$$L[X|Y] = \frac{\alpha^{-1} Y}{1 + SNR^{-1}}$$

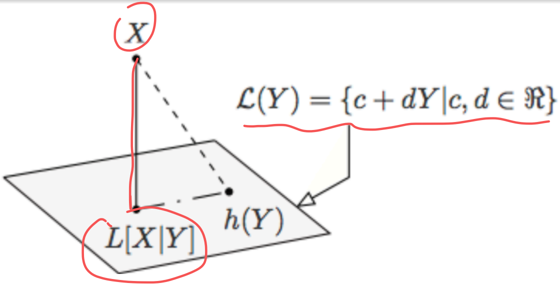
$SNR \ll 1 \qquad L[X|Y] \approx 0$

$SNR \gg 1 \qquad L[X|Y] \approx \alpha^{-1} Y$

Projection Perspective

Theorem

The Linear Least Square Estimate (LLSE) of X given Y is the projection of X onto the set $\mathcal{L}(Y)$ of linear functions of Y.



Projection Perspective (Proof)

$f(Y) \triangleq$ Projection of X onto set $\mathcal{L}(Y)$

$f(Y) = L(X|Y)$

If $f(Y)$ satisfies $E[(X - f(Y))^2] \leq E[(X - h(Y))^2]$ $\forall\, h(Y) \in \mathcal{L}(Y)$

then $f(Y) = L(X|Y)$.

$E[(X - h(Y))^2] = E[(\underbrace{X - f(Y)}_{A} + \underbrace{f(Y) - h(Y)}_{B})^2]$

$= \underbrace{E[(X - f(Y))^2]}_{A^2} + \underbrace{E[(f(Y) - h(Y))^2]}_{B^2}$

$+2\underbrace{E[(X - f(Y)) \cdot (f(Y) - h(Y))]}_{A \quad B}$

Projection Perspective (Proof)

$$X - f(Y) \perp f(Y) - h(Y) \Rightarrow E[(X - f(Y))(f(Y) - h(Y))] = 0$$

$$E[(X - h(Y))^2] = E[(X - f(Y))^2] + \underbrace{E[(f(Y) - h(Y))^2]}_{\geq 0}$$

$$E[(X - h(Y))^2] \geq E[(X - f(Y))^2] \quad \forall h(Y) \in \mathcal{L}(Y)$$

Example

Example

Here $X \sim \mathcal{N}(0,1)$, $Z \sim \mathcal{N}(0,\sigma^2)$, $X \perp Z$ and $Y = X + Z$. Find $\hat{X} = L[X|Y]$.

Example (Solution)

$X \sim N(0,1) \quad Z \sim N(0, \sigma^2)$

$X \perp Z$

Method 1:

$$L(X|Y) = E(X) + \frac{Cov(X,Y)}{Var(Y)} (Y - E(Y))$$

$E(X) = 0 \qquad E(Y) = E(X+Z) = 0$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E(XY) = E(X(X+Z))$$

$$= E(X^2) + E(XZ) = E(X^2) = Var(X) = 1$$

$$Var(Y) = Var(X+Z) = Var(X) + Var(Z) + \underline{Cov(X,Z)}$$

$$= 1 + \sigma^2$$

$$L(X|Y) = \frac{Y}{1+\sigma^2}$$

Example (Solution)

Method 2: $X \sim N(0,1)$ $\quad Z \sim N(0, \sigma^2)$ $\quad X \perp Z$ $\quad Y = X + Z$

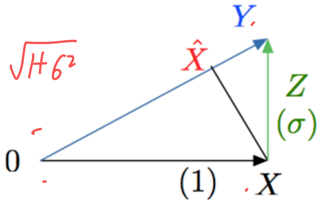
$\|X\| = \sqrt{E(X^2)} = 1$

$\|Z\| = \sqrt{E(Z^2)} = \sigma$

$\|Y\| = \sqrt{\|X\|^2 + \|Z\|^2} = \sqrt{1+\sigma^2}$

$\hat{X} = L[X|Y]$

$\hat{X} \perp Y$



$0\hat{X}X \quad 0XY$ similar

$$\frac{\|\hat{X}\|}{\|X\|} = \frac{\|X\|}{\|Y\|} \Rightarrow \|\hat{X}\| = \frac{1}{\sqrt{1+\sigma^2}} = \frac{\sqrt{1+\sigma^2}}{1+\sigma^2} = \frac{\|Y\|}{1+\sigma^2}$$

$$L[X|Y] = \hat{X} = \frac{Y}{1+\sigma^2}$$

Orthogonality Property of $\mathbb{E}[X|Y]$

Theorem

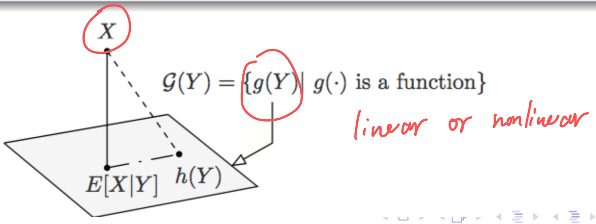
1. *For any function $\phi(\cdot)$, one has*

$$E[(X - E[X|Y])\phi(Y)] = 0.$$

2. *Moreover, if the function $g(Y)$ is such that*

$$E[(X - g(Y))\phi(Y)] = 0, \quad \forall\phi(\cdot),$$

then $g(Y) = E[X|Y]$.



Minimum Mean Square Error Estimator

Theorem (Minimum Mean Square Error Estimator)

The MMSE of X given Y is given by

$$g(Y) = \mathbb{E}[X|Y].$$

Minimum Mean Square Error Estimator (Proof)

To prove MMSE of X given Y is $E[X|Y]$

$$E[(X-\phi(Y))^2] \geq E[(X-E[X|Y])^2] \quad \forall \phi(\cdot)$$

$$E[(X-\phi(Y))^2]$$

$$= E[(\underbrace{X-E[X|Y]}_{A} + \underbrace{E[X|Y]-\phi(Y)}_{B \geq 0})^2]$$

$$= E[A^2] + E[B^2] + 2E[AB]$$

$$E(AB) = E[(X-E[X|Y]) \cdot (E(X|Y)-\phi(Y))] = 0$$

$$E[(X-\phi(Y))^2] \geq E[(X-E(X|Y))^2] \quad \forall \phi(\cdot)$$

A

$E(X|Y)$ is the MMSE of X given Y.

MMSE for Jointly Gaussian Random Variables

Theorem (MMSE for Jointly Gaussian Random Variables)

Let X, Y be jointly Gaussian random variables. Then

$$E[X|Y] = L[X|Y] = E(X) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}(Y - E(Y)).$$

MMSE for Jointly Gaussian Random Variables (Proof)

Definition (Multivariate Normal distribution)

A random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ is said to have a *Multivariate Normal* (MVN) distribution if every linear combination of the X_j has a Normal distribution. That is, we require

$$t_1 X_1 + \cdots + t_k X_k$$

to have a Normal distribution for any choice of constants $t_1, \ldots, t_k$. Equivalently, we say $X_1, \ldots, X_K$ are jointly Gaussian or $\boldsymbol{X}$ is a Gaussian vector.

MMSE for Jointly Gaussian Random Variables (Proof)

Properties of Multivariate Normal (MVN) distribution:

Property A: Linear combinations of MVN (JG) are still MVN (JG).

Property B: MVN (JG) rvs are independent iff uncorrelated.

$$X - L[X|Y] \perp \mathcal{L}(Y) = \{a + bY \mid a, b \in \mathbb{R}\}$$

$$\underline{X - L[X|Y] \perp Y \Rightarrow E[(X - L(X|Y))Y] = 0}$$

$$E[\underline{X - L[X|Y]}] = E(X) - E(\underline{L(X|Y)})$$

$$= E(X) - E\left(E(X) + \frac{Cov(X,Y)}{Var(Y)}(Y - E(Y))\right)$$

$$= E(X) - \left(E(X) + \frac{Cov(X,Y)}{Var(Y)}\underbrace{(E(Y) - E(Y))}_{=0}\right)$$

$$= 0$$

MMSE for Jointly Gaussian Random Variables (Proof)

$X - L[X|Y] \perp Y$

$\Leftrightarrow$ $X - L[X|Y]$ and Y are uncorrelated since $E[X - L(X|Y)] = 0$

$\mathrm{Cov}(X - L(X|Y), Y) = E[(X - L(X|Y))Y] - \underbrace{E(X - L(X|Y))}_{=0} E(Y) = 0$

$=0$

$\Leftrightarrow$ $X - L[X|Y]$ is independent of Y since $X - L(X|Y)$ and Y are JG

(Property A)

$\Leftrightarrow$ $X - L[X|Y]$ is independent of $\phi(Y)$ $\forall \phi(Y)$

$\Rightarrow$ $X - L[X|Y]$ and $\phi(Y)$ are uncorrelated (Property B)

$f(Y)$

$\Leftrightarrow$ $X - L[X|Y] \perp \phi(Y) \quad \forall \phi(Y)$ $\quad E[X - L(X|Y)] = 0$

MMSE = LLSE $\quad E[X|Y] = L[X|Y]$