Lecture 8: Conditional Expectation

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

May 11, 2024

Outline

Outline

Conditional PMF

- Let A be an event with positive probability. If X is a discrete r.v., then the *conditional PMF of X given A* is

$$P_{X|A}(x) = P(X = x|A) = \frac{P(\{X = x\} \cap A)}{P(A)}.$$

- Bayes' Rule:

$$P_{X|A}(x) = P(X = x|A) = \frac{P(A|X = x)P(X = x)}{P(A)}.$$

- LOTP: with a partition $A_1, \ldots, A_n$, each A_i with a positive probability $P(A_i) > 0$, $i = 1, 2, \ldots, n$:

$$P(X = x) = \sum_{i=1}^{n} P_{X|A_i}(x)P(A_i).$$

Conditional PDF

- Let A be an event with positive probability. If X is a continuous r.v., then the *conditional PDF of X given A* is

$$f_{X|A}(x) = (P(X \leq x|A))'.$$

- LOTP: with a partition $A_1, \ldots, A_n$, each A_i with a positive probability $P(A_i) > 0$, $i = 1, 2, \ldots, n$:

$$f_X(x) = \sum_{i=1}^{n} P(A_i) f_{X|A_i}(x).$$

Conditional PDF

- Bayes' Rule: given an event A with $P(A) > 0$, then

$$f_{X|A}(x) = \frac{P(A|X = x)}{P(A)} \cdot f_X(x).$$

- Bayes' Rule: given event $A =$ "$a \leq X \leq b$" and $P(A) > 0$, then

$$\begin{aligned} f_{X|A}(x) &= \frac{\mathbf{1}_{x\in[a,b]}}{P(A)} \cdot f_X(x) \\ &= \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Conditional Expectation Given An Event

Definition

Let A be an event with positive probability. If Y is a discrete r.v., then the *conditional expectation of Y given A* is

$$E(Y|A) = \sum_y y \cdot P(Y = y|A) = \sum_y y \cdot P_{Y|A}(y),$$

where the sum is over the support of Y. If Y is a continuous r.v. with PDF f, then

$$E(Y|A) = \int_{-\infty}^{\infty} y \cdot f_{Y|A}(y)\, dy.$$

LOTUS Given An Event

Definition

Let A be an event with positive probability and g is a function from $\mathbf{R}$ to $\mathbf{R}$. If Y is a discrete r.v., then the *conditional expectation of* $g(Y)$ *given* A is

$$E(g(Y)|A) = \sum_y g(y) \cdot P_{Y|A}(y),$$

where the sum is over the support of Y.
If Y is a continuous r.v. with PDF f, then

$$E(g(Y)|A) = \int_{-\infty}^{\infty} g(y) \cdot f_{Y|A}(y)\, dy.$$

Example

Let $X \sim Expo(\lambda)$, find $E(X|X>1)$ and $Var(X|X>1)$.

1°. event $A = "X>1"$; $P(A) = P(X>1) = \int_1^\infty \lambda e^{-\lambda x} dx = e^{-\lambda} > 0$.

Method 1:

$$f_{X|A}(x) = \frac{f(x)}{P(A)} = \lambda e^{-\lambda(x-1)}, \quad x>1.$$

2°. $E[X|X>1] = \int_1^\infty x \cdot f_{X|A}(x)dx = \int_1^\infty x \cdot \lambda e^{-\lambda(x-1)} dx = 1+\frac{1}{\lambda}$

$E[X|X>2024]$

$E[X^2|X>1] = \int_1^\infty x^2 \cdot f_{X|A}(x)dx = \int_1^\infty x^2 \cdot \lambda e^{-\lambda(x-1)} dx = \frac{\lambda^2+2\lambda+2}{\lambda^2}$

$\Rightarrow Var(X|X>1) = E[X^2|X>1] - E[X|X>1]^2 = \frac{1}{\lambda^2}$

$E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$ $\quad = 1+E[X]$ $\quad Var(X)$.

Method 2: $\forall s,t>0$, $P(X>s+t|X>t) = P(X>s)$; $t=1$

$\Rightarrow P(X>1+s|X>1) = P(X>s) \Rightarrow P(X-1>s|X>1) = P(X>s)$

$X-1|X>1 \sim X$.

$E(X-1|X>1) = E(X)$

$Var(X-1|X>1) = Var(X)$

$E[X|X>1] = E[X-1+1|X>1] = E[X-1|X>1] + 1$

$= E[X] + 1$

$= \frac{1}{\lambda} + 1$.

$Var(X|X>1) = Var(X-1|X>1)$

$= Var(X) = \frac{1}{\lambda^2}$.

Solution

Motivation of Conditional Expectation

- Conditional expectation is a powerful tool for calculating expectations: first-step analysis
- Conditional expectation allows us to predict or estimate unknowns based on whatever evidence is currently available.
- Conditional Expectation given an event: $E(Y|A)$
- Conditional Expectation given a random variable: $E(Y|X)$

Intuition for $E(Y|A)$

① Y r.v. (n samples. $y_1, \dots y_n$

$$E(Y) \approx \frac{1}{n} \sum_{j=1}^{n} y_j$$

② $E(Y|A)$ n samples.

I_j : in the j^{th} sampling process event A's indicator

$$E(Y|A) \approx \frac{\sum_{j=1}^{n} I_j y_j}{\sum_{j=1}^{n} I_j}$$

Intuition for $E(Y|A)$

Principle

$E(Y|A)$ is approximately the average of Y in a large number of simulation runs in which A occurred.

Life Expectancy

T : Life span.

$E(T) = 70$;

$E[T \mid T \geq 20] \neq E[T]$

If $T \sim Expo(\lambda)$; $E[T \mid T \geq 20]$

$= 20 + E[T] = 90 > 70.$

Law of Total Expectation LOTE

LOTE $\Rightarrow$ LOTP : $Y = I_B \Rightarrow P(B) = E(I_B) = E(Y) \overset{\text{LOTE}}{=}$

$$= \sum_{i=1}^{n} E(Y|A_i) \cdot P(A_i) = \sum_{i=1}^{n} E(I_B|A_i) \cdot P(A_i)$$

$$= \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

Theorem

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

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

$X \sim \text{Expo}(\lambda)$ Find $E[X|X \le 1]$?

$$E(X) = E[X|X>1] \cdot P(X>1) + E[X|X \le 1] \cdot P(X \le 1)$$

$$\frac{1}{\lambda} = \left[1 + \frac{1}{\lambda}\right] \cdot e^{-\lambda} + E[X|X \le 1] \cdot (1 - e^{-\lambda})$$

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!)

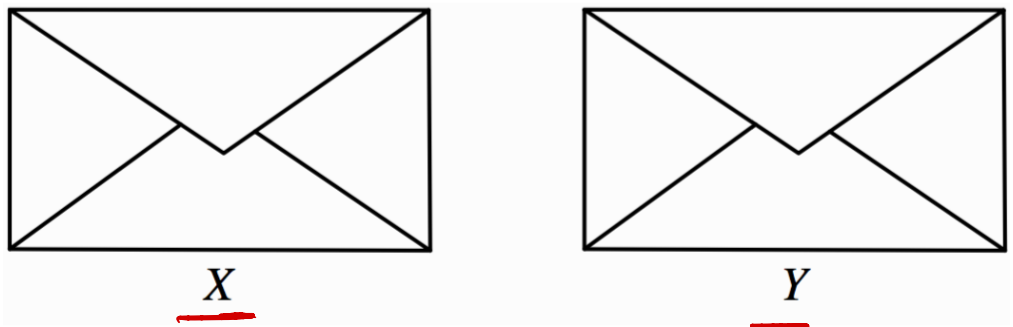


FIGURE 9.1
Two envelopes, where one contains twice as much money as the other. Either $Y = 2X$ or $Y = X/2$, with equal probabilities. Which would you prefer?

Solution

$$Y = \begin{cases} 2X & \text{w.p. } 0.5 \\ \frac{X}{2} & \text{w.p. } 0.5 \end{cases}$$

LOTE:

$$E(Y) = E[Y \mid Y=2X] \cdot P(Y=2X) + E[Y \mid Y=\tfrac{X}{2}] \cdot P(Y=\tfrac{X}{2})$$

$$= \underline{E[2X \mid Y=2X]} \cdot 0.5 + \underline{E[\tfrac{X}{2} \mid Y=\tfrac{X}{2}]} \cdot 0.5$$

$$\neq E[2X] \qquad \neq E[\tfrac{X}{2}]$$

$$E[Y \mid Y=2] = 2$$

$$E[Y \mid Y=2X] \neq E[2X]$$

Geometric Expectation Redux

$X \sim Geom(p)$. Find $E(X)$.

Coin Tosses.

# of tosses (X) before the appearance of the first Head?

H: Head

T: Tail.

① First Step Analysis... Conditioning on the outcome of the first toss:

$O_1 = H$ or T.

② $E(X) \overset{LOTE}{=} E(X|O_1=H)\,P(O_1=H) + E(X|O_1=T)\,P(O_1=T)$

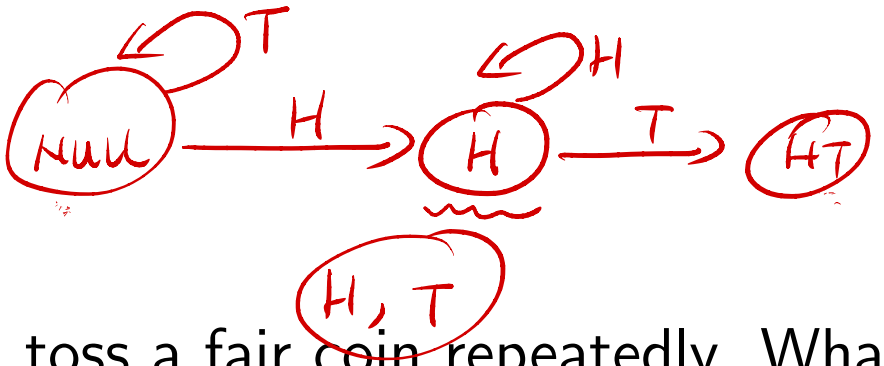
(memoryless)

$= 0 \cdot P + (1 + E[X]) \cdot (1-P)$

$\Rightarrow E[X] = \frac{1-P}{P}$.

Time until *HH* vs. *HT*

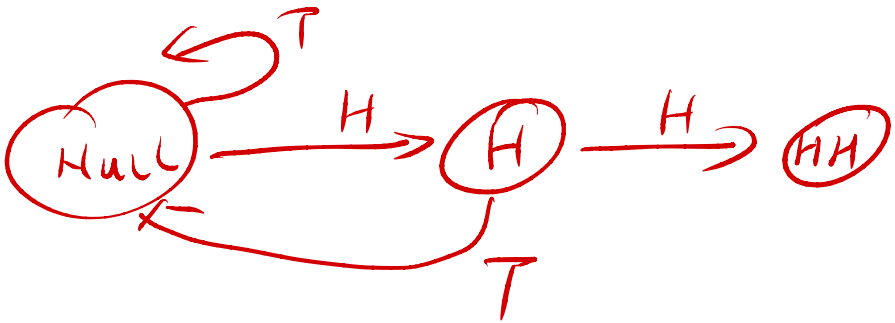
① HT : partial progress.



PGF
Condition Expectation (LOTE)
Markov chain.
Renewal Theorem
Martingale.

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?

② HH :



Solution

W_{HT} : # of tosses until "HT" for the first time.

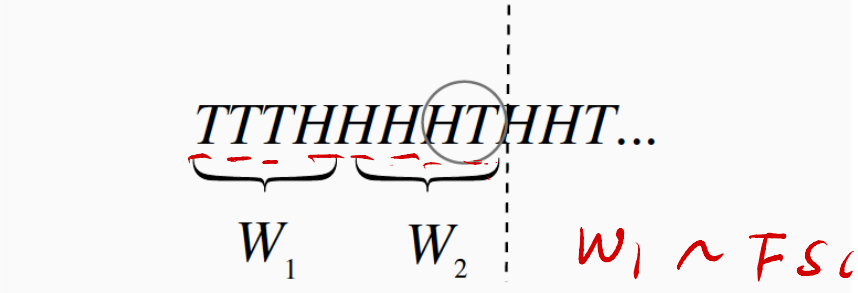
$W_{HT} = W_1 + W_2$.

W_1 : # of tosses for the first "H".

W_2 : after the first "H", # of additional tosses for the first "T".

$X \sim Fs(p)$.

$E(X) = \frac{1}{p}$



$W_1 \sim FS(1/2)$.

$W_2 \sim FS(1/2)$ (memoryless)

$\Rightarrow E(W_1) = 2$; $E(W_2) = 2$;

$\Rightarrow E(W_{HT}) = E(W_1 + W_2)$

$= E(W_1) + E(W_2) = 4$.

Solution

W_{HH}: First step Analysis.

O_1: result of the first toss.

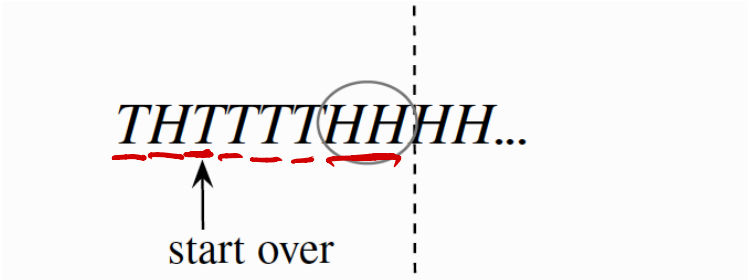
O_2: ..... second ...

1^0. $O_1 = H$ or T

$$E(W_{HH}) \overset{\text{LOTE}}{=} E(W_{HH} \mid O_1 = H)\, P(O_1 = H) + E(W_{HH} \mid O_1 = T)\, P(O_1 = T)$$

$P(O_1 = H) = \frac{1}{2}$, $P(O_1 = T) = \frac{1}{2}$

$E(W_{HH} \mid O_1 = T) = [1 + E(W_{HH})]$.



2^0. $O_2 = H$ or T

$$E(W_{HH} \mid O_1 = H) = E(W_{HH} \mid O_1 = H, O_2 = H) \cdot P(O_2 = H \mid O_1 = H)$$

LOTE with extra conditioning.

$P(O_2 = H) = \frac{1}{2}$

$$+ E(W_{HH} \mid O_1 = H, O_2 = T) \cdot P(O_2 = T \mid O_1 = H)$$

$$\frac{2}{2} \qquad\qquad \frac{2 + E(W_{HH})}{\frac{1}{2}}.$$

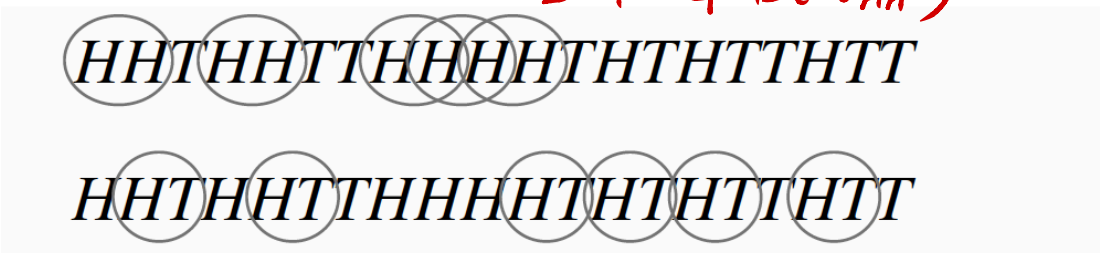
Solution

3°. $E(W_{HH} \mid \theta_1 = H) = 2 \cdot \frac{1}{2} + [2 + E(W_{HH})] \cdot \frac{1}{2}$

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

$\Rightarrow$ $E(W_{HH}) = [2 + \frac{1}{2} E(W_{HH})] \cdot \frac{1}{2} + [4 + E(W_{HH})] \cdot \frac{1}{2}$

$= \frac{3}{2} + \frac{3}{4} E(W_{HH})$



$\Rightarrow$ $E(W_{HH}) = 6$ vs. $E(W_{HT}) = 4$

Outline

Conditional Expectation Given An R.V.

$g(x) = E[Y|X=x]$ Real Number. estimate.

$g(X) = E[Y|X]$ R.V. estimator.

Definition

Let $g(x) = E(Y|X=x)$. Then the *conditional expectation of Y given X*, denoted $E(Y|X)$, is defined to be the random variable $g(X)$. In other words, if after doing the experiment X crystallizes into x, then $E(Y|X)$ crystallizes into $g(x)$.

Remark

- $E(Y|X)$ is a function of X, and it is a random variable.
- It makes sense to computer $E(E(Y|X))$ and $Var(E(Y|X))$.

Example: Stick Length

① $X \sim \text{Unif}(0,1)$; $Y|X=x \sim \text{Unif}(0,x)$

$E[Y|X=x] = \frac{x}{2} = g(x)$; $\Rightarrow E[Y|X] = g(X) = \frac{X}{2}$

② $E[E[Y|X]] = E[\frac{X}{2}] = \frac{1}{2}E[X] = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$

Suppose we have a stick of length 1 and break the stick at a point X chosen uniformly at random. Given that $X = x$, we then choose another breakpoint Y uniformly on the interval $[0, x]$. Find $E(Y|X)$, and its mean and variance.

$\text{Var}[E[Y|X]] = \text{Var}(\frac{X}{2}) = \frac{1}{4}\text{Var}(X) = \frac{1}{4}\cdot\frac{1}{12} = \frac{1}{48}$

Find $E[Y|X]$ $\Rightarrow$ ① $g(x) = E[Y|X=x]$

② $g(x) \to g(X) = E[Y|X]$

$x \to X$

Solution

Dropping What's Independent

Y_1, Y_2, X r.v.s.

Linearity: $E[Y_1 + Y_2 | X] = E[Y_1 | X] + E[Y_2 | X]$

Theorem

If X and Y are independent, then $E(Y|X) = E(Y)$.

$\forall x$, $E[Y | X = x] = E[Y] = g(x)$.

$\Rightarrow$ $g(X) = E[Y]$

Taking Out What's Known

$$g(x) = E[h(x)Y \mid X=x] = E[h(x)\cdot Y \mid X=x]$$

$$= h(x)\cdot E[Y \mid X=x]$$

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

Theorem

For any function h,

$$E[h(X)\cdot Y \mid X]$$

$$E(h(X)\,Y|X) = h(X)\,E(Y|X)$$

Linearity

$$g(x) = E[Y_1+Y_2 \mid X=x] = E[Y_1 \mid X=x] + E[Y_2 \mid X=x]$$

$$\Rightarrow g(X) = E[Y_1+Y_2 \mid X] = E[Y_1 \mid X] + E[Y_2 \mid X]$$

Theorem

$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$.

Example

1°. By symmetry, $E(X_1|S_n) = E(X_2|S_n) = \cdots = E(X_n|S_n)$

2°. By Linearity, $E(X_1|S_n) + E(X_2|S_n) + \cdots + E(X_n|S_n)$

$= E(X_1 + \cdots + X_n|S_n)$

$= E(S_n|S_n) = S_n$

$E[S_n|S_n=s] = E[s|S_n=s] = s = g(s)$

$\Rightarrow E[S_n|S_n] = g(S_n) = S_n$

Let $X_1, \ldots, X_n$ be i.i.d., and $S_n = X_1 + \cdots + X_n$. Find $E(X_1|S_n)$.

$3^\circ \Rightarrow \quad n\, E(X_1|S_n) = S_n$

$E(X_1|S_n) = \frac{1}{n} S_n$

Adam's Law

The Law of Iterated Expectation

The Tower Rule.

The Smoothing Theorem

Theorem

For any r.v.s X and Y,

$$E(E(Y|X)) = E(Y).$$

W.l.o.G., X and Y are both discrete r.v.s.

1° $g(X) = E[Y|X]$; $g(x) = E[Y|X=x] = \sum_y y \cdot P(Y=y|X=x)$

2° LHS $E[E[Y|X]] = E[g(X)] = \sum_x g(x) \cdot P(X=x)$

$$= \sum_x \left[\sum_y y \cdot P(Y=y|X=x)\right] \cdot P(X=x)$$

$$= \sum_y y \cdot \left[\sum_x P(Y=y|X=x) \cdot P(X=x)\right]$$

$$= \sum_y y \cdot \left[\sum_x P(Y=y, X=x)\right] = \sum_y y \cdot P(Y=y) = E[Y]$$ RHS.

Proof

Adam's Law with Extra Conditioning

① $\hat{P}(\cdot) = P(\cdot|Z)$; $\hat{E} = E(\cdot|Z)$

Adam's Law. $E[E[Y|X]] = E[Y]$

$\hat{E}[\hat{E}[Y|X]] = \hat{E}(Y)$

Theorem

For any r.v.s X, Y, Z, we have

$$E(E(Y|X,Z)|Z) = E(Y|Z)$$
$$E(E(X|Z,Y)|Y) = E(X|Y)$$

(a) $E(Y|X) = E(Y|X,Z)$

$E(Y) = E(Y|Z)$

(b) $\hat{E}[\hat{E}[Y|X]]$

$= E[\hat{E}[Y|X]|Z]$

$= E[E[Y|X,Z]|Z]$

Conditional Variance

1°. $\mathrm{Var}(Y) = E[(Y - E(Y))^2]$

$\hat{E}(\cdot) = E(\cdot|X)$

$\Rightarrow \mathrm{Var}(Y|X) = \hat{E}[(Y - \hat{E}(Y))^2]$

Definition

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

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

This is equivalent to

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

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

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

$= E(Y^2|X) - (E(Y|X))^2$

Eve's law

Theorem

For any r.v.s X and Y,

$$\operatorname{Var}(Y) = E(\operatorname{Var}(Y|X)) + \operatorname{Var}(E(Y|X)).$$

The ordering of E's and Var's on the right-hand side spells EVVE, whence the name Eve's law. Eve's law is also known as the law of total variance or the variance decomposition formula.

Proof

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|X)] + \mathrm{Var}(E[Y|X])$$

① $g(X) = E[Y|X]$; By Adam's Law, $E[g(X)] = E[E[Y|X]] = E[Y]$

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

$= E[Y^2] - E[g^2(X)]$

③ $\mathrm{Var}[E(Y|X)] = \mathrm{Var}(g(X)) = E[g^2(X)] - (E[g(X)])^2$

$= E[g^2(X)] - E^2(Y)$

② + ③ $\Rightarrow E[\mathrm{Var}(Y|X)] + \mathrm{Var}[E(Y|X)] = E(Y^2) - E^2(Y)$

$= \mathrm{Var}(Y)$

Proof

Example: Random Sum

N, X_j independent

① $E[X]$: $E[X|N=n] = E\left(\sum_{j=1}^{N} X_j \,|\, N=n\right) = E\left(\sum_{j=1}^{n} X_j \,|\, N=n\right)$

$= E\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} E(X_j) = n\cdot\mu.$ $(=g(n))$

$\Rightarrow E[X|N] = N\cdot\mu.$

$\Rightarrow E[X] = E[E[X|N]] = E[N\cdot\mu] = \mu\cdot E[N].$

A store receives N customers in a day, where N is an r.v. with finite mean and variance. Let X_j be the amount spent by the jth customer at the store. Assume that each X_j has mean μ and variance σ^2, and that N and all the X_j are independent of one another. Find the mean and variance of the random sum $X = \sum_{j=1}^{N} X_j$, which is the store's total revenue in a day, in terms of μ, σ^2, $E(N)$, and $\mathrm{Var}(N)$.

② $\mathrm{Var}(X|N=n) = \mathrm{Var}\left(\sum_{j=1}^{N} X_j \,|\, N=n\right) = \mathrm{Var}\left(\sum_{j=1}^{n} X_j \,|\, N=n\right)$

$= \mathrm{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathrm{Var}(X_j) = n\sigma^2.$

$\Rightarrow \mathrm{Var}(X|N) = N\cdot\sigma^2$

Solution

③ By Eve's Law,

$$\begin{aligned}
\mathrm{Var}(X) &= E[\mathrm{Var}(X|N)] + \mathrm{Var}[E[X|N]] \\
&= E[N\cdot\sigma^2] + \mathrm{Var}(N\cdot\mu) \\
&= \sigma^2 E[N] + \mu^2 \mathrm{Var}(N)
\end{aligned}$$

Solution

Outline

Basic Problem

- Estimate Y from the observed value X
- Choose the estimator (inference function) $g(\cdot)$ to minimize the expected error $E(c(Y, g(X)))$
- $c(Y, \hat{Y})$ is the cost of guessing $\hat{Y}$ when the actually value is Y.
- When $c(Y, \hat{Y}) = ||Y - \hat{Y}||^2$, the best guess is called "the least square estimate (LSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a + bX$, it is called "the Linear Least Square Estimate (LLSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ can be arbitrary, it is called "the Minimum Mean Square Estimate (MMSE)" estimate of Y given X.

Linear Least Square Estimate

① $f(a,b) = E[(Y-a-bX)^2] = E[Y^2+a^2+b^2X^2-2Ya-2bXY+2abX]$

$= a^2 + E[Y^2] + b^2E[X^2] - 2aE[Y] + 2abE[X] - 2bE[XY]$

1°. $\frac{\partial f(a,b)}{\partial a} = 2a - 2E[Y] + 2bE[X] = 0 \Rightarrow a + bE[X] = E[Y]$

Theorem

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

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

2°. $\frac{\partial f(a,b)}{\partial b} = 2bE[X^2] + 2aE[X] - 2E[XY] = 0 \Rightarrow aE[X] + bE[X^2] = E[XY]$.

Proof

$b = \frac{Cov(X,Y)}{Var(X)}$, $a = E[Y] - \frac{Cov(X,Y)}{Var(X)} \cdot E[X]$.

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

Verify : Hessian Matrix.

$H \succeq 0$ $\quad z^T H z \geq 0$.

$$H = \begin{bmatrix} \frac{\partial^2 f(a,b)}{\partial a^2} & \frac{\partial^2 f(a,b)}{\partial a \partial b} \\ \frac{\partial^2 f(a,b)}{\partial b \partial a} & \frac{\partial^2 f(a,b)}{\partial b^2} \end{bmatrix} \succeq 0.$$

$$H = \begin{bmatrix} 2 & 2E[X] \\ 2E[X] & 2E[X^2] \end{bmatrix} = 2 \begin{bmatrix} 1 & E[X] \\ E[X] & E[X^2] \end{bmatrix} \; ; \; z = (z_1, z_2)^T \; ;$$

$$z^T H z = (z_1, z_2) H \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = z_1^2 + 2 z_1 z_2 E[X] + z_2^2 E[X^2]$$

$$= (z_1 + z_2 E[X])^2 + z_2^2 (E[X^2] - E[X]^2) \geq 0.$$

$Var(X)$

Proof

Linear Regression

k data points $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$

$y_j \approx a + b x_j$

$$\min_{a,b} \sum_{j=1}^{k} (y_j - a - b x_j)^2 \Rightarrow L_k[Y|X]$$

Data-Driven LLSE: Linear Regression

- $L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$
- Now we only have k i.i.d samples: $(X_1, Y_1), \ldots, (X_k, Y_k)$
- Use sample mean to replace expectation

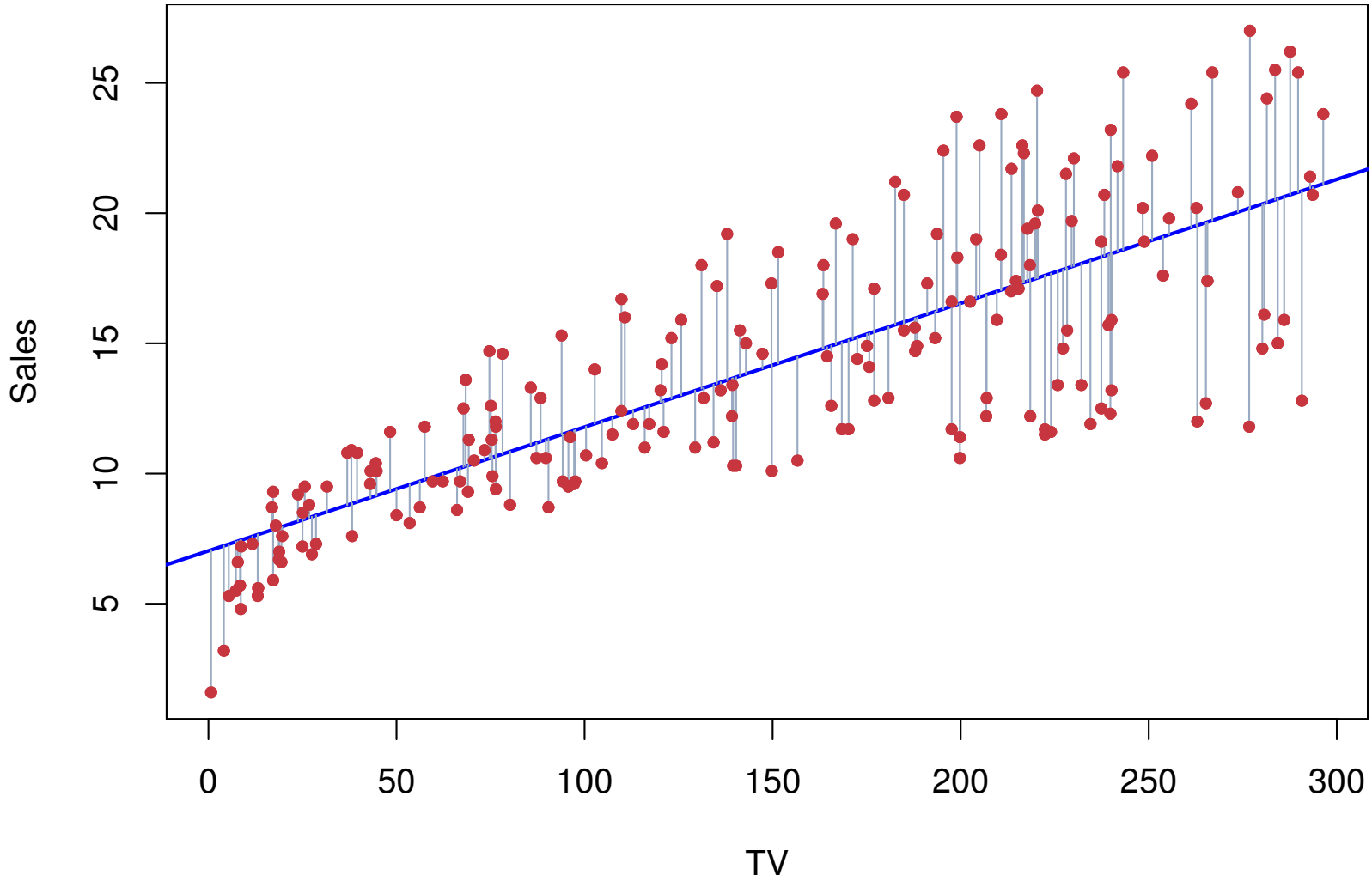
$$E(X) \leftarrow E_k(X) = \frac{1}{k}\sum_{j=1}^{k} X_k$$

$$E(Y) \leftarrow E_k(Y) = \frac{1}{k}\sum_{j=1}^{k} Y_k$$

$$Cov(X, Y) = E(XY) - E(X)E(Y) \leftarrow \frac{1}{k}\sum_{j=1}^{k} X_k Y_k - E_k(X)E_k(Y)$$

$$Var(X) = E(X^2) - (E(X))^2 \leftarrow \frac{1}{k}\sum_{j=1}^{k} X_k^2 - (E_k(X))^2$$

Data-Driven LLSE: Linear Regression



Minimum Mean Square Error Estimator

MMSE

LLSE : $\min_{\hat{Y}} E[(Y-\hat{Y})^2]$ $\hat{Y} = a+bX$ $\Rightarrow \hat{Y}^* = L[Y|X]$

MMSE : $\min_{\hat{Y}} E[(Y-\hat{Y})^2]$ $\hat{Y} = g(X)$ $\Rightarrow \hat{Y}^* = E[Y|X]$

Theorem

The MMSE of Y given X is given by

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

Geometric Perspective of Conditional Expectation

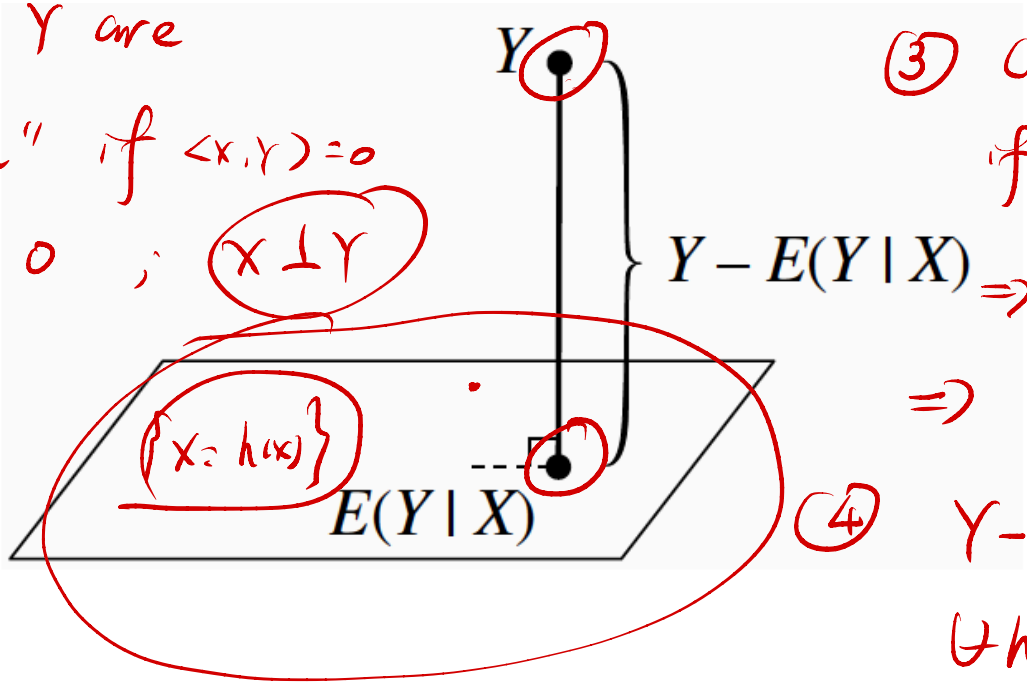
① inner product $\langle X, Y \rangle = E[X \cdot Y]$; $dist(X,Y) = \sqrt{\langle X-Y, X-Y \rangle}$

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

$\cos\theta = \frac{\langle X,Y \rangle}{\|X\| \cdot \|Y\|}$

② X and Y are "orthogonal" if $\langle X,Y \rangle = 0$

$E[XY] = 0$; $X \perp Y$



③ $Cov(X,Y) = E(XY) - E(X)E(Y)$

if $E(X) = 0$ or $E(Y) = 0$ or both ;

$\Rightarrow Cov(X,Y) = E(XY)$

$\Rightarrow$ uncorrelated $\Leftrightarrow$ orthogonal.

④ $Y - E[Y|X] \perp h(X)$ $\forall h(\cdot)$.

⑤ $E[Y|X]$: a projection of Y onto the space of (arbitrary function of X)

⑥ $L[Y|X]$: a projection of Y onto $\mathcal{L}(X) = \{aX + b : a, b \in \mathbb{R}\}$ the space of (linear function of X)

LLSE.

Projection Interpretation

1°. $E(Y - E[Y|X]) = E(Y) - E[E[Y|X]]$

$= E(Y) - E[Y] = 0$

Uncorrelated $\Leftrightarrow$ orthogonal

Theorem

For any function h, the r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$. Equivalently,

$$E((Y - E(Y|X))h(X)) = 0.$$

$Y - E[Y|X] \perp h(X)$

(This is equivalent since $E(Y - E(Y|X)) = 0$, by linearity and Adam's law.)

2°. show $Y - E[Y|X] \perp h(X)$

Proof

$Y - E[Y|X] \perp h(X)$

$\Leftrightarrow E[(Y - E[Y|X]) \cdot h(X)] = 0$

$= E[Yh(X) - h(X) \cdot E[Y|X]]$

$= E[Yh(X)] - E[h(X)E[Y|X]]$

$= E[Yh(X)] - E[E[h(X) \cdot Y|X]]$

$= E[Yh(X)] - E[h(X) \cdot Y]$ Adam's Law.

$= 0$

Proof

Prediction Perspective

$E[(Y-E[Y|X])^2] \le E[(Y-g(X))^2]$

$\forall\, g(\cdot)$

$\min_{g(x)} E[(Y-g(X))^2] \Rightarrow g^*(x) = E[Y|X]$

- Predict or estimate the future observations or unknown parameters based on data
- $E(Y|X)$ is our **best predictor** of Y based on X.
- Best means it is the function of X with the lowest **mean squared error** (expected squared difference between Y and prediction of Y).
- It is called the Minimum Mean Square Estimate (MMSE)

Proof

1°. $\hat{Y}$ = estimator of Y. ($\hat{Y} = g(X)$)

$E[(Y-\hat{Y})^2] = E[(Y-g(X))^2]$ ∵ $Y-g(X) = Y-E[Y|X] + (E[Y|X]-g(X))$

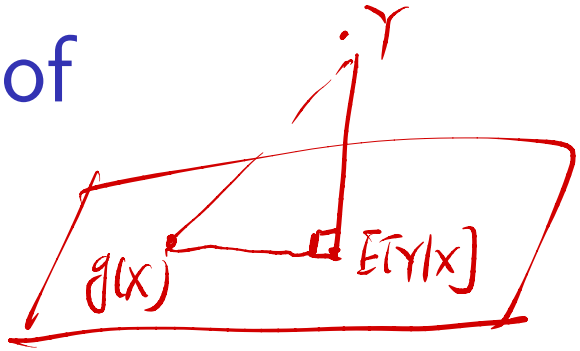
2°. $E[(Y-g(X))^2] = E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2] + 2E[(Y-E[Y|X])(E[Y|X]-g(X))]$

$Y-E[Y|X] \perp h(X)$

3°. $h(X) \triangleq E[Y|X]-g(X)$: a function of X $\Rightarrow E[(Y-E[Y|X])\cdot h(X)] = 0$

4°. $E[(Y-g(X))^2] = E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2]$

Proof



$dist^2(Y, E[Y|X]) + dist^2(g(X), E[Y|X])$

$= dist^2(Y, g(X))$

5° $E[(Y-\hat{Y})^2] = E[(Y-g(X))^2]$

$= E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2]$

$\geq 0. \quad \forall g(\cdot)$

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

$\hat{Y}^* = g^*(X) = E[Y|X]$

MMSE

Proof

MMSE for Jointly Normal Random Variables

Theorem

Let X, Y be jointly Normal random variables. Then

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

Remark: Statistical Learning Perspective

- In general, MMSE is a highly nonlinear function.
- Adoption of various approximation methods leads to various learning methods
  - Linear regression
  - Logistic regression
  - Polynomial regression
  - Regression with Spline functions
  - Neural network

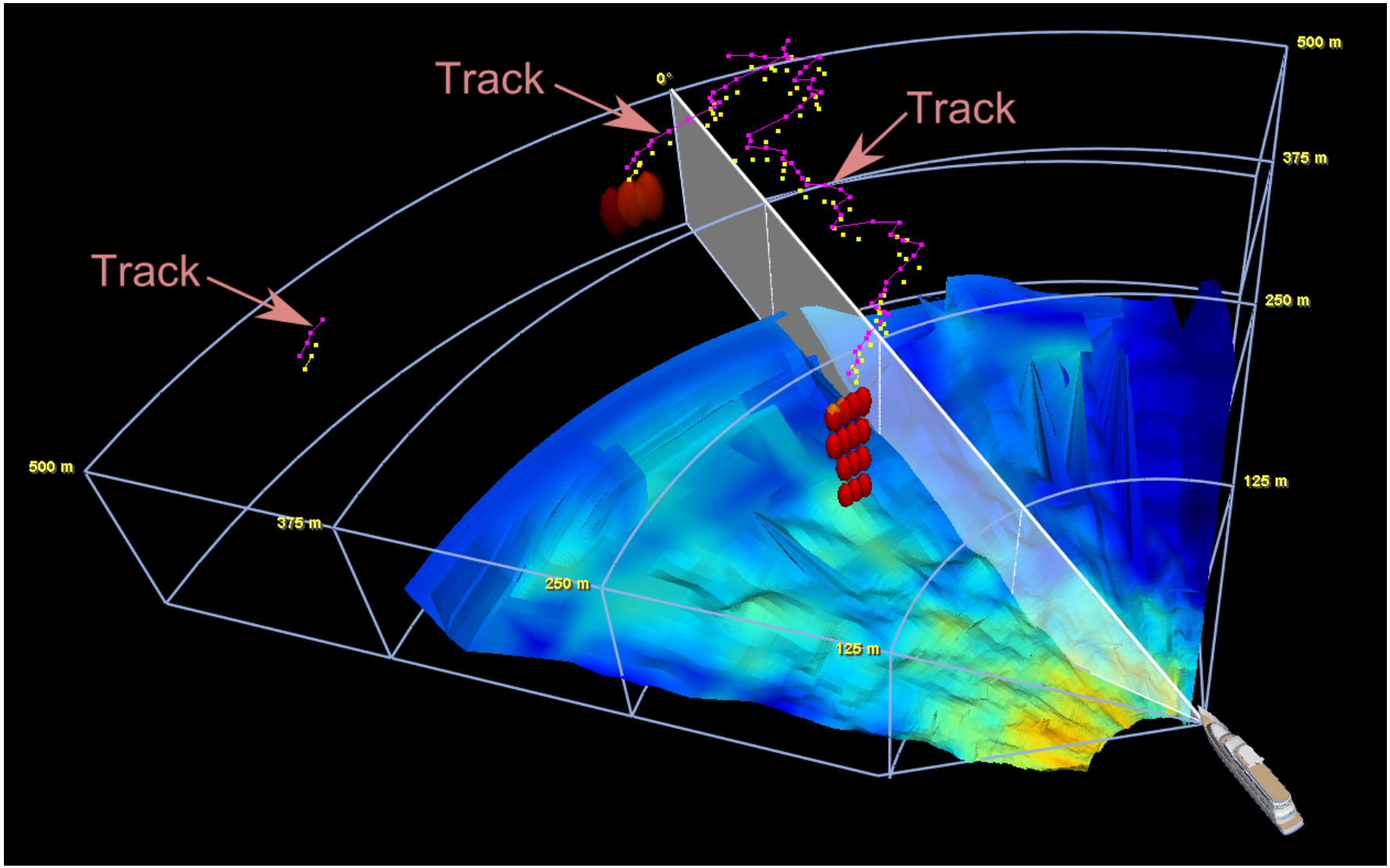
Outline

Milestones in Statistics & Signal Processing

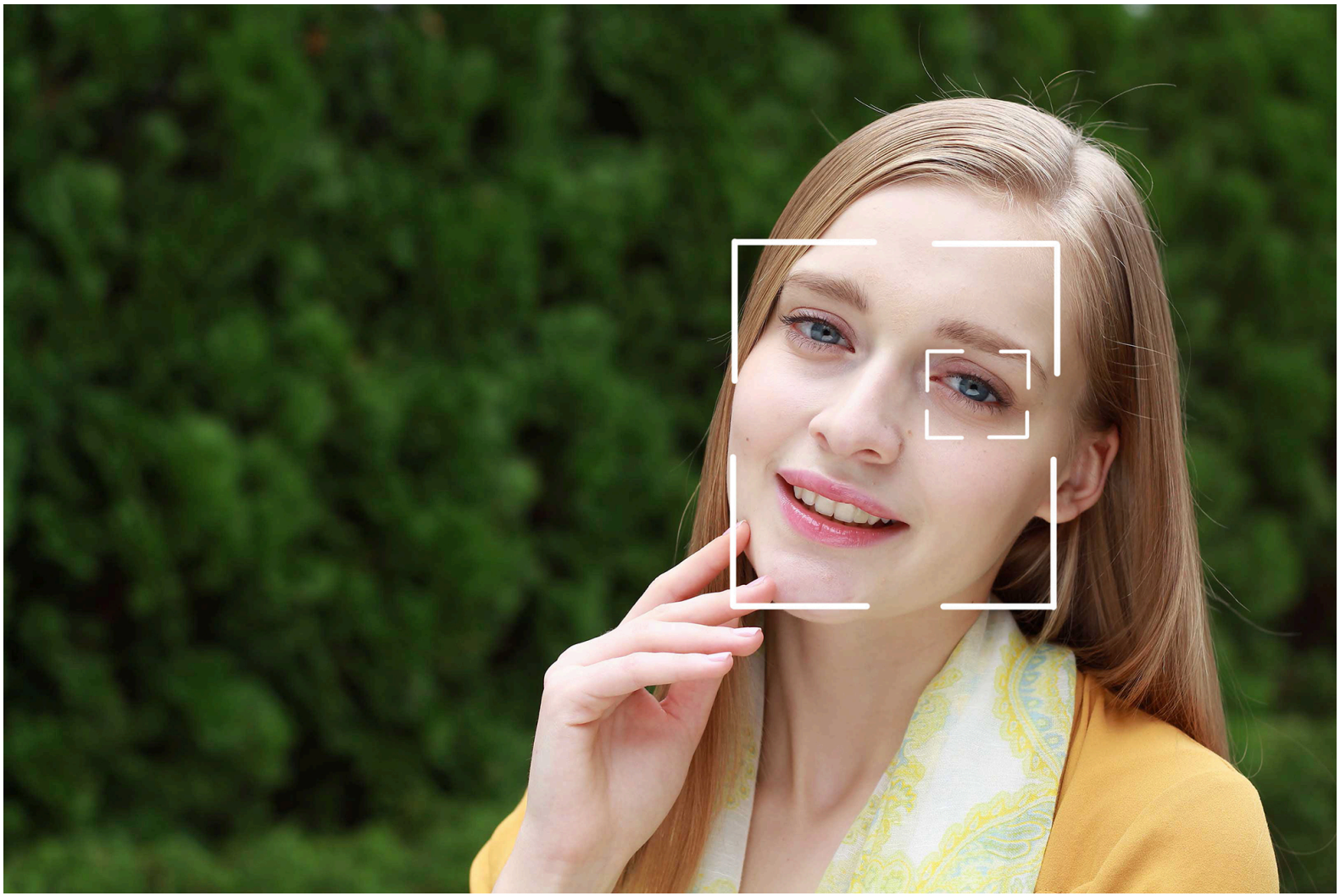
- 1960: Rudoph Emil Kalman (1930-2016) introduced what is known as Kalman filter.

Widely Applications: Location & Navigation & Map Building

Widely Applications: Radar Tracking

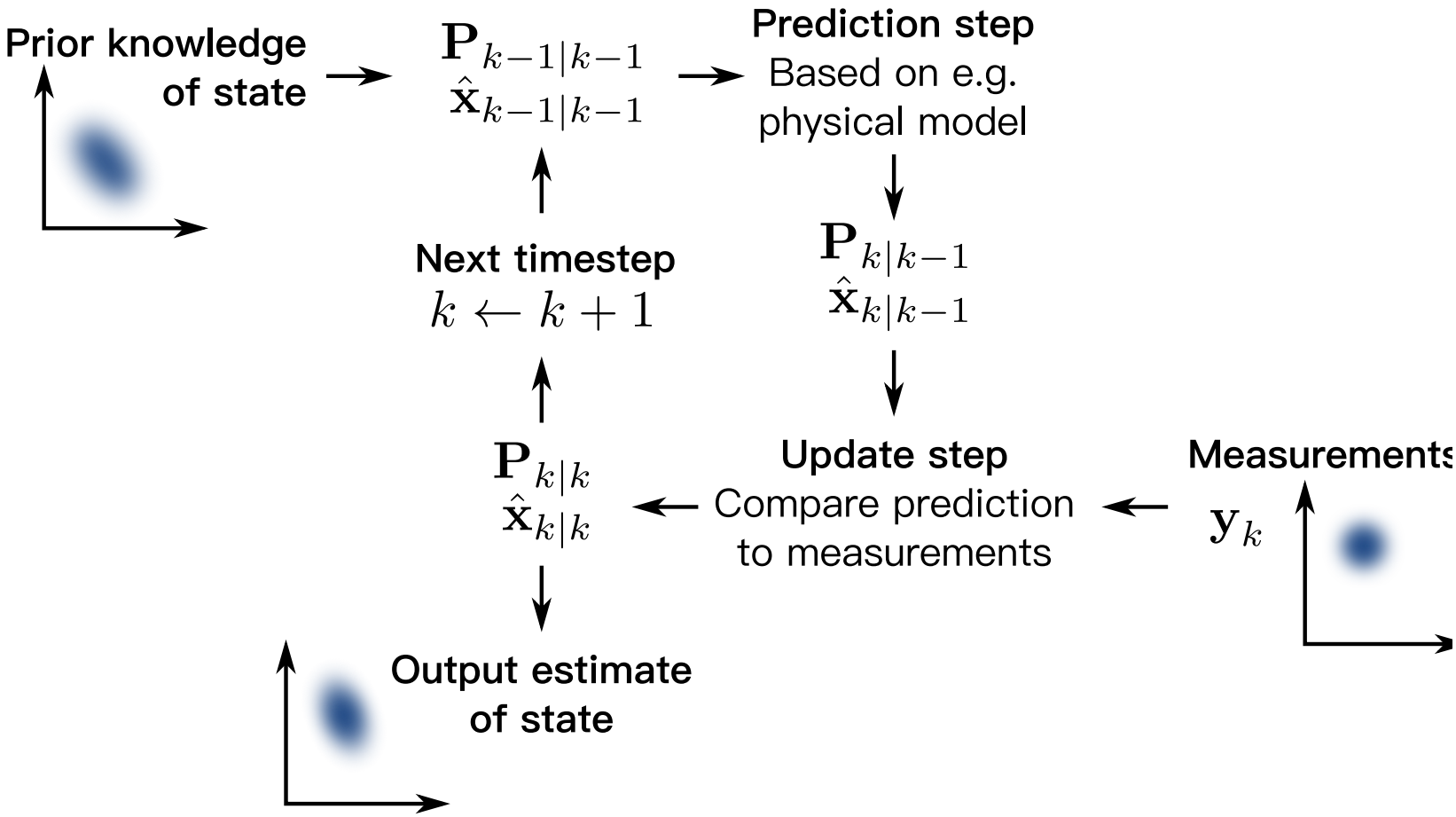


Widely Applications: Human Face & Eye Detection Autofocus



Widely Applications: Animal Eye Detection Autofocus

Essence of Kalman Filter



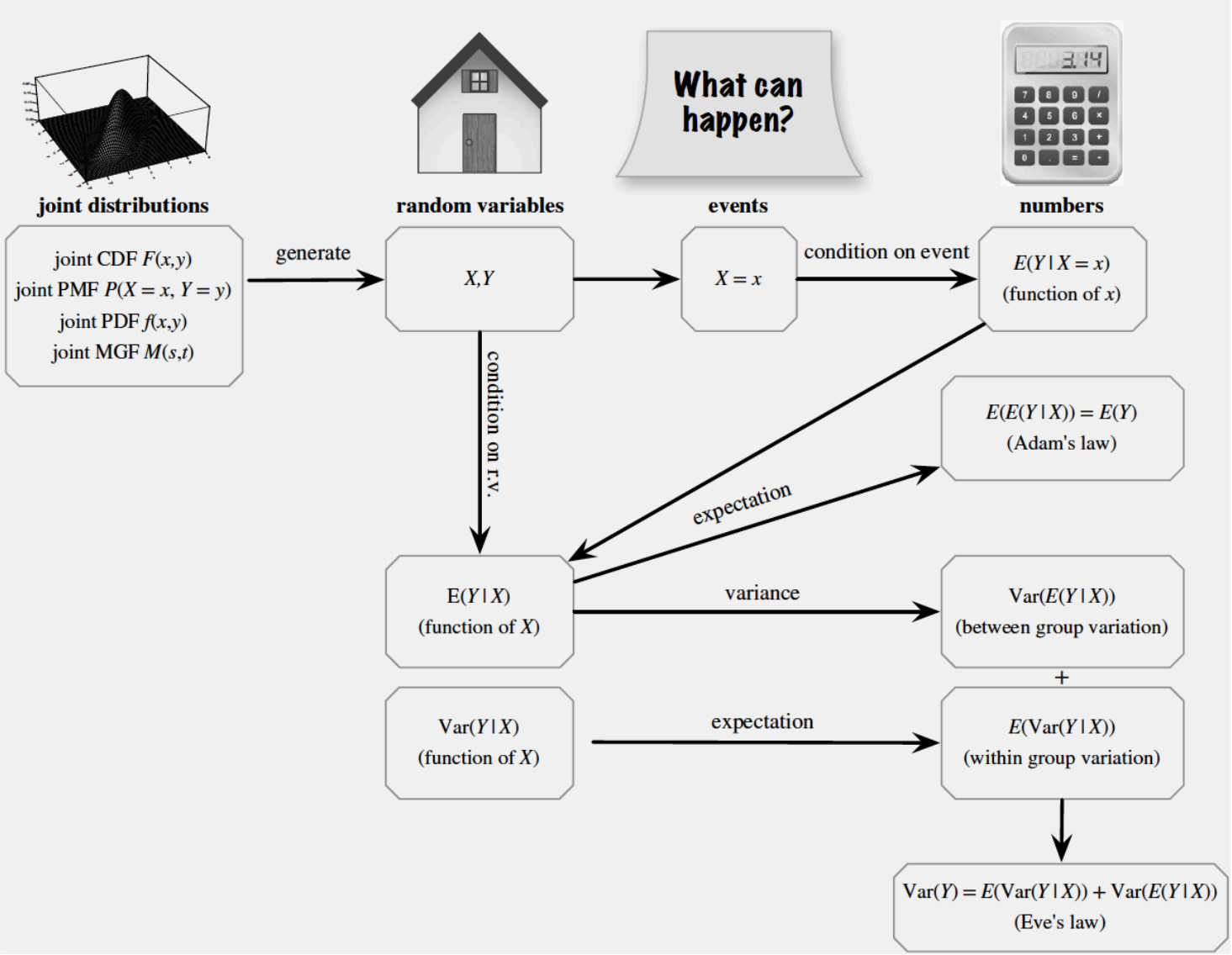
Reasons for Popularity of Kalman Filter

- Good results in practice due to optimality and structure: LLSE estimation in general, MMSE estimation under the setting of Gaussian noise.
- Convenient form for online real time processing: recursive equations.
- Easy to formulate and implement given a basic understanding.

Why Use The Word "Filter"

- The process of finding the "best estimate" from noisy data amounts to "filtering out" the noise.
- Estimation (statistical perspective) vs. Filtering (signal processing perspective)
- A Kalman filter not only cleans up the data measurements
- A Kalman filter also projects these measurements onto the state estimate

Summary 1


joint distributions
random variables
What can happen?
events
numbers
joint CDF $F(x,y)$
joint PMF $P(X=x, Y=y)$
joint PDF $f(x,y)$
joint MGF $M(s,t)$
generate
X,Y
$X=x$
condition on event
$E(Y|X=x)$
(function of x)
condition on r.v.
$E(E(Y|X)) = E(Y)$
(Adam's law)
expectation
$E(Y|X)$
(function of X)
variance
$\mathrm{Var}(E(Y|X))$
(between group variation)
+
$\mathrm{Var}(Y|X)$
(function of X)
expectation
$E(\mathrm{Var}(Y|X))$
(within group variation)
$\mathrm{Var}(Y) = E(\mathrm{Var}(Y|X)) + \mathrm{Var}(E(Y|X))$
(Eve's law)

References

- Chapter 9 of **BH**
- Chapters 4 & 6 of **BT**