

# Lecture 8: Conditional Expectation

Ziyu Shao

School of Information Science and Technology  
ShanghaiTech University

May 11, 2024

# Outline

- 1 Conditional Expectation: Given An Event
- 2 Conditional Expectation: Given A Random Variable
- 3 Prediction & Estimation
- 4 Application Case: Kalman Filter

# Outline

- 1 Conditional Expectation: Given An Event
- 2 Conditional Expectation: Given A Random Variable
- 3 Prediction & Estimation
- 4 Application Case: Kalman Filter

## 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  $A_1, \dots, A_n$ , each  $A_i$  with a positive probability  $P(A_i) > 0$ ,  $i = 1, 2, \dots, n$ :

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

# 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

$$f_{X|A}(x) = \underbrace{(P(X \leq x|A))'}_{\text{red underline}}$$

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

$$f_X(x) = \underbrace{\sum_{i=1}^n P(A_i) f_{X|A_i}(x)}_{\text{red underline}}$$

# 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

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

where the sum is over the support of  $Y$ .

If  $Y$  is a continuous r.v. with PDF  $f$ , then

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

## Example

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

Method 1 :

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

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

$$E[X^2|X > 1]$$

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

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

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

$$E(X) = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}$$

$$= 1 + E[X]$$

$$\text{Var}(X).$$

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

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

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

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

$$\text{Var}(X|X > 1) = \text{Var}(X-1|X > 1)$$

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

$$= E[X] + 1$$

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

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

$$\text{Var}(X-1|X > 1) = \text{Var}(X)$$

# 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.

$$E(T) = 70 \text{ ,}$$

$$\underline{E[T | T \geq 20]}$$

$$\neq \underline{E[T]}$$

if  $T \sim \text{Expo}(\lambda)$  ;  $E[T | T \geq 20]$

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

# Law of Total Expectation LOTE

$$\begin{aligned} \text{LOTE} \rightarrow \text{LoTP} : Y = I_B \Rightarrow \underline{P(B)} &= E(I_B) = E(Y) \stackrel{\text{LoTE}}{=} \\ &= \sum_{i=1}^n E(Y|A_i) \cdot P(A_i) = \sum_{i=1}^n \underline{E(I_B|A_i) \cdot P(A_i)} \\ &= \sum_{i=1}^n \underline{P(B|A_i) \cdot P(A_i)} \end{aligned}$$

## Theorem

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

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

$X \sim \text{Exp}(\lambda)$ . Find  $E[X | X \leq 1]$  ?

$$\begin{aligned} \frac{E(X)}{\lambda} &= \frac{E(X | X > 1) \cdot P(X > 1)}{\lambda} + \frac{E(X | X \leq 1) \cdot P(X \leq 1)}{\lambda} \\ &= [1 + \frac{1}{\lambda}] \cdot e^{-\lambda} + \underline{(E(X | X \leq 1) \cdot (1 - e^{-\lambda}))} \end{aligned}$$

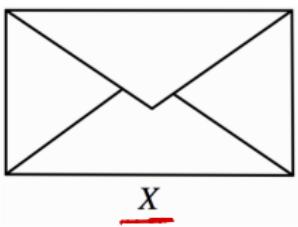
## Two-envelope Paradox

$$\textcircled{1} \quad Y = 2X \quad \text{or} \quad Y = \frac{X}{2} \quad \text{w.p. 0.5}$$

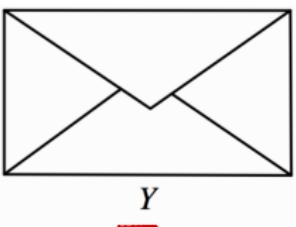
$$E(Y) = E(2X) \cdot \frac{1}{2} + E\left(\frac{X}{2}\right) \cdot \frac{1}{2} = \frac{5}{4}E(X)$$

A stranger presents you with two identical-looking, sealed envelopes, each of which contains a check for some positive amount of money.  $> E(X)$

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!)  $\textcircled{2} \quad X = \frac{Y}{2} \quad \text{or} \quad X = 2Y$



X



Y

$$E(X) = \frac{5}{4}E(Y)$$

$> E(Y)$

**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

NOTE :

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

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

$$= \underbrace{E[2X | Y=2X]}_{= E[2X]} \cdot 0.5 + \underbrace{E[\frac{X}{2} | Y=\frac{X}{2}]}_{\neq E[\frac{X}{2}]} \cdot 0.5$$

$$\neq E[2X]$$

$$\neq E[\frac{X}{2}]$$

$$\overline{E[Y | Y=2X]} = 2$$

$$E[Y | Y=2X]$$

$$\neq E[2X]$$

# Geometric Expectation Redux

$X \sim \text{Geom}(p)$ . Find  $E(X)$ .

① First step Analysis... conditioning on the outcome of the first toss.

$$O_1 = H \text{ or } T.$$

$$\begin{aligned} \textcircled{2} \quad E(X) &\stackrel{\text{LoTE}}{=} \frac{E(X|O_1=H) \underbrace{p(O_1=H)}_{\text{memoryless}} + E(X|O_1=T) \underbrace{p(O_1=T)}_{\text{memoryless}}}{p(O_1=H) + p(O_1=T)} \\ &= 0 \cdot p + \frac{(1 + E(X)) \cdot (1-p)}{p} \\ \Rightarrow E[X] &= \frac{1-p}{p}, \end{aligned}$$

Coin Tosses.

# of tosses  $\times$   
before the  
appearance of  
the first Head?

H = Head

T = Tail.

# Time until $HH$ vs. $HT$

①  $HT$ : partial progress.



PGF

Condition Expectation (LOTE)

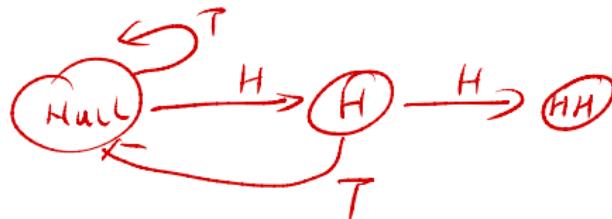
Markov chain.

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".

  
 $W_1$        $W_2$

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

$$W_2 \sim FS(1/2) \text{ (memoryless)}$$

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

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

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

Solution  $\text{W}_{\text{HH}}$ : First step Analysis.

$O_1$ : result of the first toss.

1<sup>o</sup>.  $O_1 = H$  or  $T$

$O_2$ : ... second ...

$$E(W_{\text{HH}}) \stackrel{\text{LOTE}}{=} \underbrace{E(W_{\text{HH}} | O_1=H)}_{\frac{1}{2}} \cdot P(O_1=H) + \underbrace{E(W_{\text{HH}} | O_1=T)}_{[1+E(W_{\text{HH}})]} \cdot \underbrace{P(O_1=T)}_{\frac{1}{2}}$$

$THTTTT\overbrace{HHHH\dots}$

↑  
start over

2<sup>o</sup>.  $O_2 = H$  or  $T$

$E(W_{\text{HH}} | O_1=H)$   
LOTE with extra  
conditioning.

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

$$= \frac{E(W_{\text{HH}} | O_1=H, O_2=H) \cdot P(O_2=H | O_1=H)}{2} + \frac{E(W_{\text{HH}} | O_1=H, O_2=T) \cdot P(O_2=T | O_1=H)}{2 + E(W_{\text{HH}})}$$

Solution 3°.  $E(W_{HH}|O_1=H) = 2 \cdot \frac{1}{2} + [2 + E(W_{HH})] \cdot \frac{1}{2}$

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

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

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

$HHTHHTTHHHHHHTHTHTHTTHTT$

$HHTHHTTHHHHHHTHTHTHTTHTT$

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

# Outline

- 1 Conditional Expectation: Given An Event
- 2 Conditional Expectation: Given A Random Variable
- 3 Prediction & Estimation
- 4 Application Case: Kalman Filter

# Conditional Expectation Given An R.V.

$$\underline{g(x) = E[Y|X=x]} \quad \text{Real Number.} \quad \text{estimate.}$$

$$\underline{g(X) = E[Y|X]} \quad \text{R.V.} \quad \text{estimator.}$$

## Definition

Let  $\underline{g(x) = E(Y|X=x)}$ . Then the *conditional expectation of Y given X*, denoted  $\underline{E(Y|X)}$ , is defined to be the random variable  $\underline{g(X)}$ . 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  $\text{Var}(E(Y|X))$ .

## Example: Stick Length



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

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

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

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.

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

---

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

②  $g(x) \rightarrow g(X) = E[Y|X]$

$$x \rightarrow X$$

# Solution

# Dropping What's Independent $Y_1, Y_2, X$ r.v.s.

Linearity :  $E[Y_1 + Y_2 | X] = \underline{E[Y_1 | X]} + E[Y_2 | X]$

## Theorem

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

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

$\Rightarrow \underline{g(x)} = \underline{E[Y]}$

# Taking Out What's Known

$$\underline{g(x)} = E[h(x)Y | X=x] = E[\underline{h(x)} \cdot Y | X=x]$$
$$= \underline{h(x)} \cdot \underline{E[Y | X=x]}$$

Theorem

$$\Rightarrow \underline{g(x)} = h(x) \cdot E[Y|X]$$

For any function  $h$ ,

$$\underline{E[h(x)Y|X]}$$

$$\underline{E(h(X)Y|X)} = \underline{h(X)} \underline{E(Y|X)}$$

# Linearity

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

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

## Theorem

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

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

2°. By Linearity . . .

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

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

Let  $X_1, \dots, X_n$  be i.i.d., and  $S_n = \underline{X_1 + \dots + X_n}$ . Find  $\underline{E(X_1|S_n)}$ .

$$\begin{aligned} & E(X_1|S_n) + E(X_2|S_n) + \dots + E(X_n|S_n) \\ &= E(\underline{X_1 + \dots + X_n}|S_n) \\ &= E(\underline{S_n}|S_n) = \underline{S_n} \end{aligned}$$

$$3°. \Rightarrow n E(X_1|S_n) = S_n$$

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

## Adam's Law

## The Law of Iterated Expectation

The Tower Rule.

The Smoothing Theorem

Theorem

For any r.v.s  $X$  and  $Y$ ,

N.L.O.G...  $X$  and  $Y$  are both discrete r.v.s.

$$1^{\circ} \quad g(X) = E[Y|X]; \quad g(x) = E[Y|X=x]$$

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

$$\left( \sum y \cdot P(Y=y|X=x) \right)$$

$$2^{\circ}. \quad \text{LHS} \quad E[E(Y|X)] = E[g(X)] = \sum_x g(x) \cdot P(X=x)$$

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

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

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

RHS



# Proof

# Adam's Law with Extra Conditioning

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

Adam's Law:  $E[\hat{E}[Y|X]] = E[Y]$

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

## Theorem

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

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

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

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

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

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

## Conditional Variance

$$1^{\circ} \quad \text{Var}(Y) = E[(Y - E(Y))^2]$$

$$\widehat{E}(\cdot) = \underline{E}(\cdot | X)$$

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

### Definition

The *conditional variance of Y given X* is

$$\underline{\text{Var}}(Y|X) = E \left[ (Y - E(Y|X))^2 | X \right].$$

This is equivalent to

$$2^{\circ} \quad \text{Var}(Y) = \underline{E}(Y^2) - (\underline{E}(Y))^2$$

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

$$\text{Var}(Y|X) = \widehat{E}(Y^2) - (\widehat{E}(Y))^2$$

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

# Eve's law

## Theorem

For any r.v.s  $X$  and  $Y$ ,

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

The ordering of  $E$ 's and  $\text{Var}$ '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

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

①  $\underline{g(x)} = \underline{E[Y|X]}$ ; By Adam's Law,  $E[g(x)] = E[E(Y|X)] = \underline{E[Y]}$

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

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

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

# Proof

## Example: Random Sum

$N, X_j$   
independent

$$\textcircled{1} \quad E[X] : E[X|N=n] = E\left(\sum_{j=1}^N X_j | N=n\right) = E\left(\sum_{j=1}^n X_j | N=n\right)$$
$$\Rightarrow E[X|N] = N \cdot \mu.$$
$$\Rightarrow E[X] = E[E[X|N]] = E[N \cdot \mu] = \mu \cdot E(N).$$

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  $j$ th customer at the store. Assume that each  $X_j$  has mean  $\mu$  and variance  $\sigma^2$ , 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  $\mu, \sigma^2, E(N),$  and  $\text{Var}(N)$ .

$$\textcircled{2} \quad \text{Var}(X|N=n) = \text{Var}\left(\sum_{j=1}^N X_j | N=n\right) = \text{Var}\left(\sum_{j=1}^n X_j | N=n\right)$$
$$= \text{Var}\left(\sum_{j=1}^n X_j\right) = \sum_{j=1}^n \text{Var}(X_j) = n\sigma^2.$$
$$\Rightarrow \text{Var}(X|N) = N \cdot \sigma^2$$

# Solution

③ By Euds Law ,

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

# Solution

# Outline

- 1 Conditional Expectation: Given An Event
- 2 Conditional Expectation: Given A Random Variable
- 3 Prediction & Estimation
- 4 Application Case: Kalman Filter

# Basic Problem

- Estimate  $\hat{Y}$  from the observed value  $X$   $\hat{Y} = g(X)$  estimator of  $Y$ .
- 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

$$\textcircled{1} \quad f(a, b) = E[(Y - a - bX)^2] = E[Y^2 + a^2 + b^2 X^2 - 2Ya - 2bXY + 2abX] \\ = a^2 + E[Y^2] + b^2 E[X^2] - 2a E[Y] + 2ab E[X] - 2b E[XY] \\ \textcircled{2} \cdot \frac{\partial f(a, b)}{\partial a} = \underline{2a - 2 E[Y]} + \underline{2b E[X]} = 0 \Rightarrow \boxed{a + b E[X] \\ = E[Y]}$$

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,

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

$$\textcircled{2} \cdot \frac{\partial f(a, b)}{\partial b} = \underline{2b E[X^2]} + \underline{2a E[X]} - \underline{2E[XY]} = 0 \Rightarrow \boxed{a E[X] \\ + b E[X^2] \\ = E[XY].}$$

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

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

Verify : Hessian Matrix.

$$H > 0$$

$$\mathbf{z}^T H \mathbf{z} \geq 0.$$

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

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

$$\begin{aligned} \mathbf{z}^T H \mathbf{z} &= (z_1, z_2) H \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = z_1^2 + 2z_1 z_2 E[X] + z_2^2 E[X^2] \\ &= \underbrace{(z_1 + z_2 E[X])^2}_{\geq 0} + \underbrace{z_2^2 (E[X^2] - E[X]^2)}_{\text{Var}(X)} \geq 0. \end{aligned}$$

# Proof

## Linear Regression

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

$$\underline{y_j} \approx \underline{a + b x_j}$$

$$\min_{a,b} \sum_{j=1}^k (y_j - a - b x_j)^2$$

$\Rightarrow$

$$L_k(\gamma | x)$$

## Data-Driven LLSE: Linear Regression

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

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

$E \leftarrow$  Sample mean.

$$\text{SLLN} \Rightarrow k \rightarrow \infty$$

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

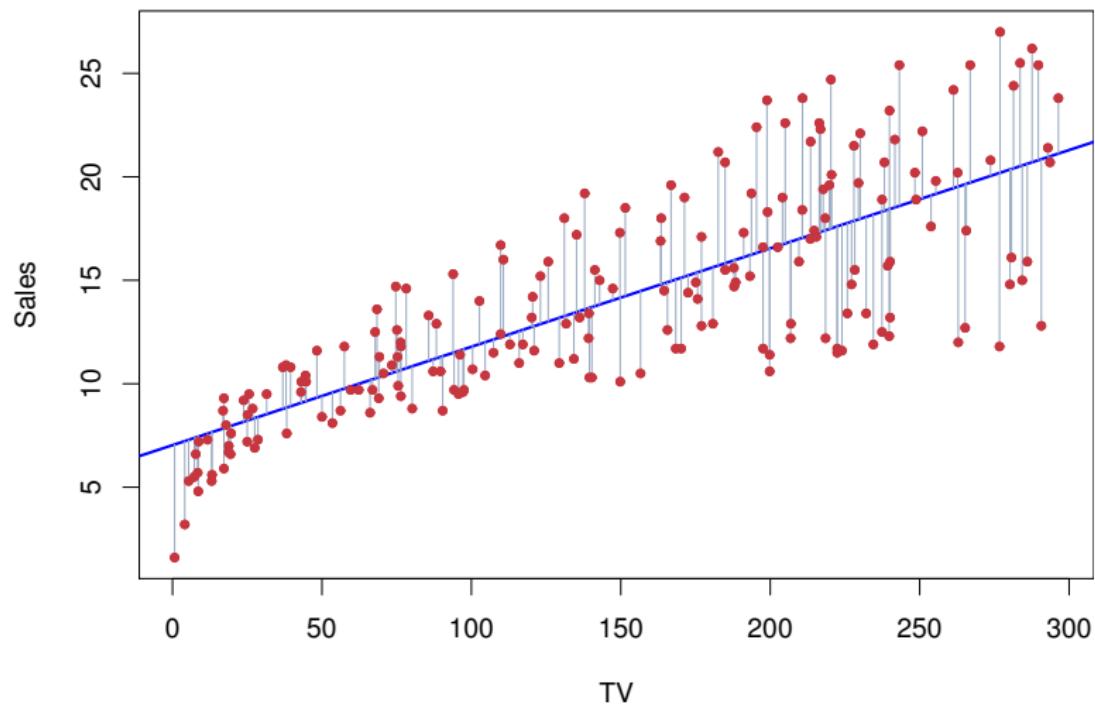
$\angle R \rightarrow \text{LLSE}$

$$\underline{L_k[Y|X]}$$

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

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

# Data-Driven LLSE: Linear Regression



# Minimum Mean Square Error Estimator

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

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

Theorem

The MMSE of  $Y$  given  $X$  is given by

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

# Geometric Perspective of Conditional Expectation

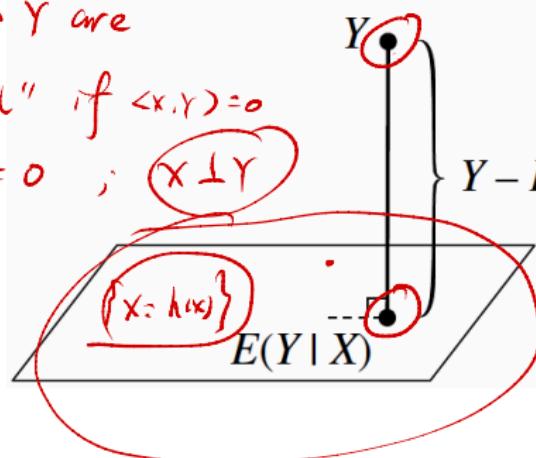
(1) inner product  $\langle X, Y \rangle = E[X \cdot Y]$ ;  $\text{dist}(x, y) = \sqrt{\langle X - Y, X - Y \rangle}$

$$\cos \theta = \frac{\langle X, Y \rangle}{\|X\| \cdot \|Y\|} = \sqrt{E[(XY)^2]}$$

(2)  $X$  and  $Y$  are

"Orthogonal" if  $\langle X, Y \rangle = 0$

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



(3)  $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$

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

$$\Rightarrow \text{cov}(X, Y) = E(XY)$$

$\Rightarrow$  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

(6)  $L(Y|X)$ : a projection of  $Y$  onto  $\{L(x) = f + bx : a, b \in \mathbb{R}\}$   
the space of (Linear function of  $x$ )

# Projection Interpretation

$$\begin{aligned} 1^{\circ}. \quad E(Y - E(Y|X)) &= E(Y) - \underline{E[E(Y|X)]} \\ &= E(Y) - E[Y] = 0 \end{aligned}$$

## Theorem

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

$$E((Y - E(Y|X))h(X)) = 0. \quad \underbrace{Y - E(Y|X)}_{\perp h(X)}$$

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

$$2^{\circ}. \quad \text{show} \quad Y - E(Y|X) \perp h(X)$$

# Proof

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

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

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

$$= E[Yh(X)] - \underbrace{E[h(X)E[Y|X]]}$$

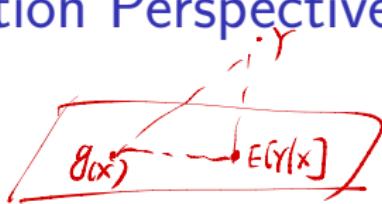
$$= E[Yh(X)] - \underbrace{E[E[h(X)Y|X]]}$$

$$= \underbrace{E[Yh(X)]} - \underbrace{E[h(X)Y]}_{\text{Adam's Law}}.$$

$$= 0$$

# Proof

## Prediction Perspective



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

$\forall g(\cdot)$

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

- 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<sup>o</sup>.  $\hat{Y}$  = estimator of  $Y$ . ( $\hat{Y} = g(x)$ )

$$E[(Y - \hat{Y})^2] = E[(Y - g(x))^2] \quad , \quad Y - g(x) = \frac{Y - E[Y|x]}{\underline{+ (E[Y|x] - g(x))}}$$

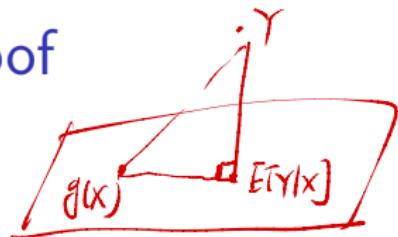
2<sup>o</sup>.  $E[(Y - g(x))^2] = E[(Y - E[Y|x])^2] + E[(E[Y|x] - g(x))^2]$

$$+ \underbrace{2E[(Y - E[Y|x])(E[Y|x] - g(x))]}_{\begin{matrix} 0 \\ o \\ h(x) \end{matrix}} \quad \begin{matrix} Y - E[Y|x] \\ + h(x) \end{matrix}$$

3<sup>o</sup>.  $h(x) \triangleq \underline{E[Y|x] - g(x)}$ : a function of  $x$ .  $\Rightarrow E[(Y - E[Y|x]), h(x)] = 0$

4<sup>o</sup>.  $\underline{\underline{E[(Y - g(x))^2]}} = E[(Y - E[Y|x])^2] + E[(E[Y|x] - g(x))^2]$

# Proof



$$\text{dist}^2(Y, E[Y|x]) + \text{dist}^2(g(x), E[Y|x]) \\ = \text{dist}^2(Y, g(x))$$

$$5^\circ. \quad \underline{E[(Y-\hat{Y})^2]} = \underline{E[(Y-g(x))^2]}$$

$$= \underline{E[(Y-E[Y|x])^2]} + \underline{E[(E[Y|x]-g(x))^2]}$$

$$\geq \underline{E[(Y-E[Y|x])^2]}$$

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

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

MMSE

# Proof

# MMSE for Jointly Normal Random Variables

LLSE

$E[Y|X]$

Methods : 1<sup>o</sup>. Joint PDF

2<sup>o</sup>. Projection

MMSE  
LLSE

## Theorem

Let  $X, Y$  be jointly Normal random variables. Then

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

# Remark: Statistical Learning Perspective

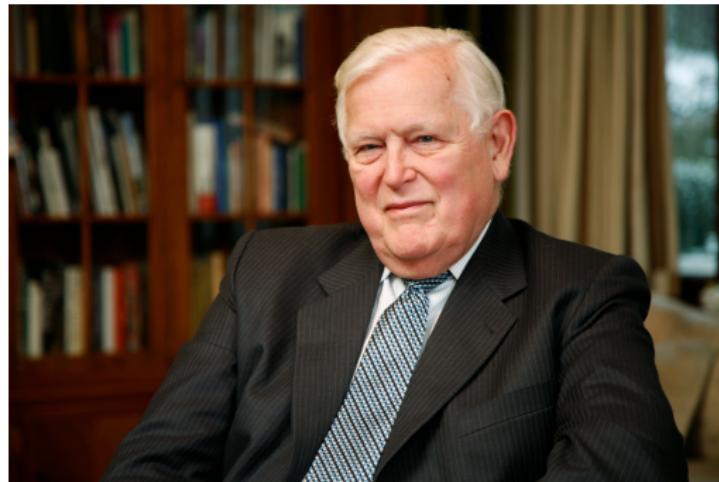
- In general, MMSE is a highly nonlinear function.
- Adoption of various approximation methods leads to various learning methods
  - ▶ Linear regression  $g(\cdot) = a + bX$
  - ▶ Logistic regression  $g(\cdot) =$
  - ▶ Polynomial regression  $g(\cdot) =$
  - ▶ Regression with Spline functions
  - ▶ Neural network  $g(\cdot)$   $N\pi k$

# Outline

- 1 Conditional Expectation: Given An Event
- 2 Conditional Expectation: Given A Random Variable
- 3 Prediction & Estimation
- 4 Application Case: Kalman Filter

# Milestones in Statistics & Signal Processing

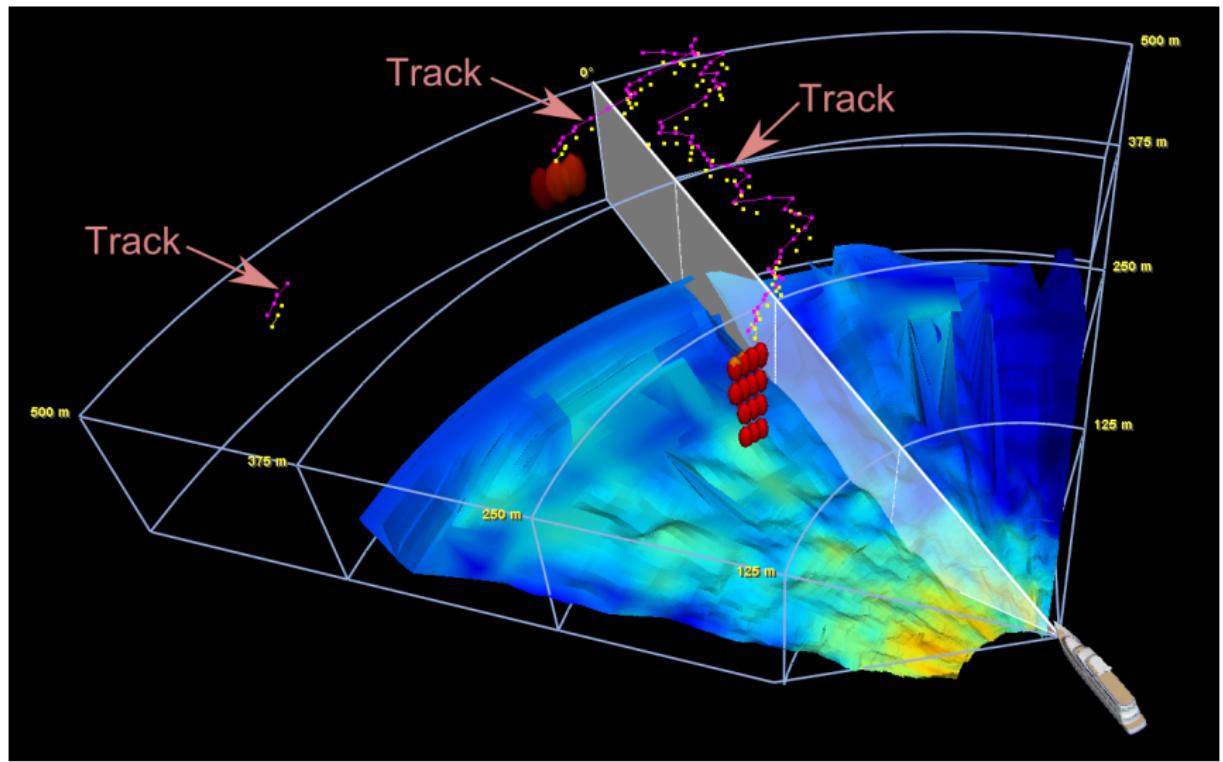
- 1960: Rudolph Emil Kalman (1930-2016) introduced what is known as Kalman filter.



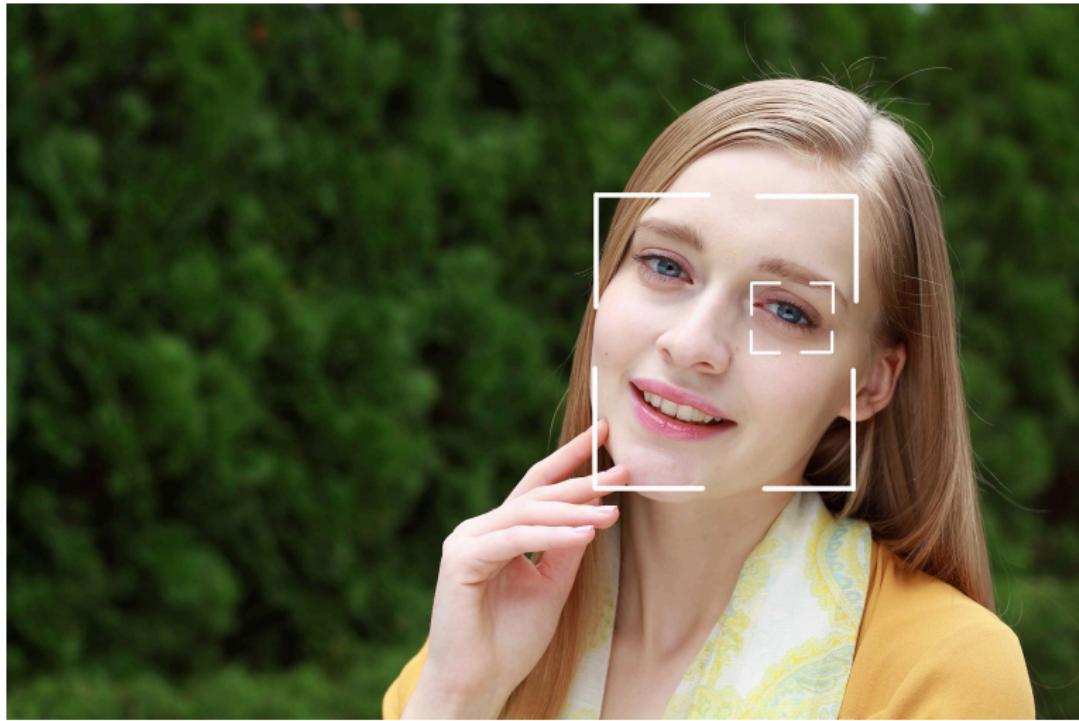
# Widely Applications: Location & Navigation & Map Building



# Widely Applications: Radar Tracking



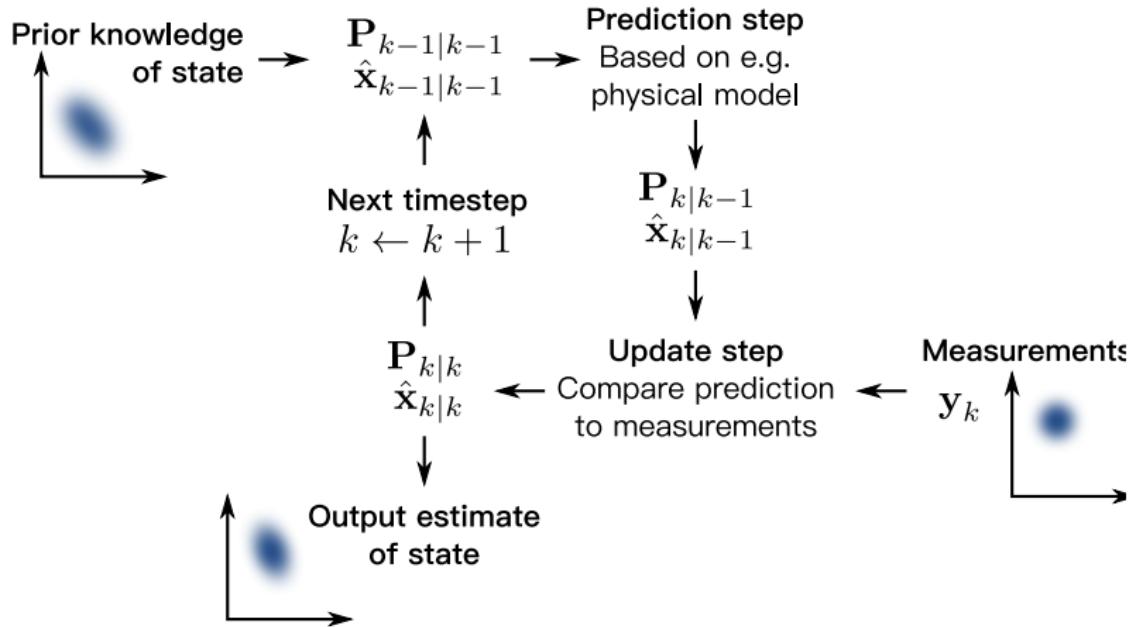
# Widely Applications: Human Face & Eye Detection Autofocus



# Widely Applications: Animal Eye Detection Autofocus



# Essence of Kalman Filter



# Reasons for Popularity of Kalman Filter

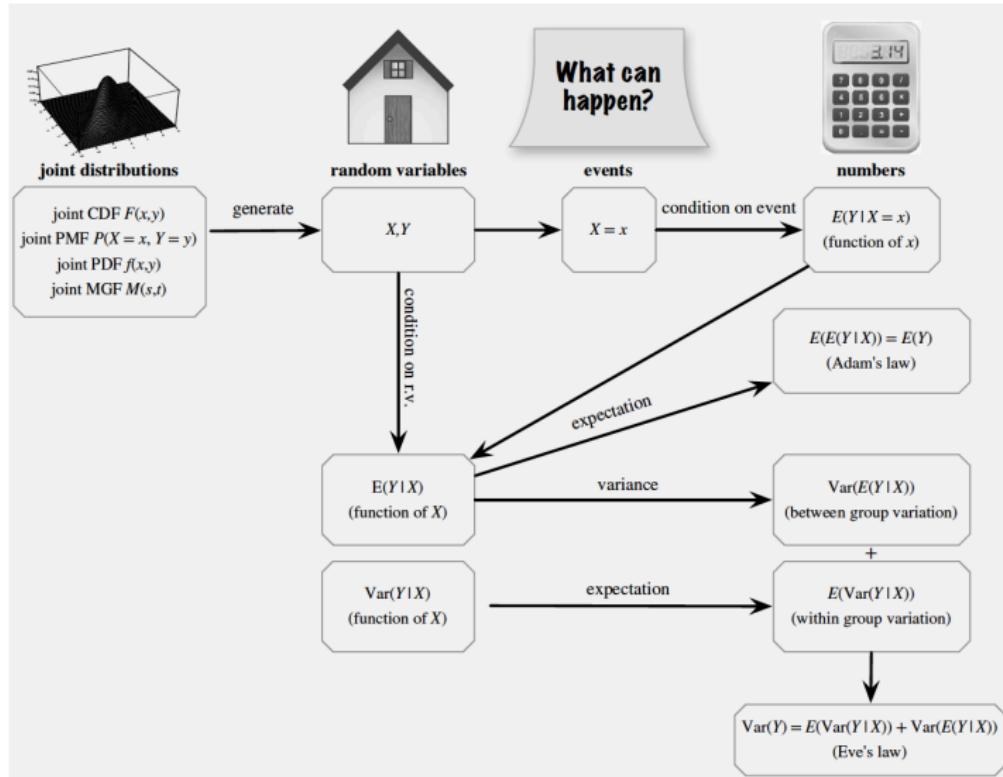
Online

- 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



# References

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