Lecture 9: Statistical Inference

Ziyu Shao

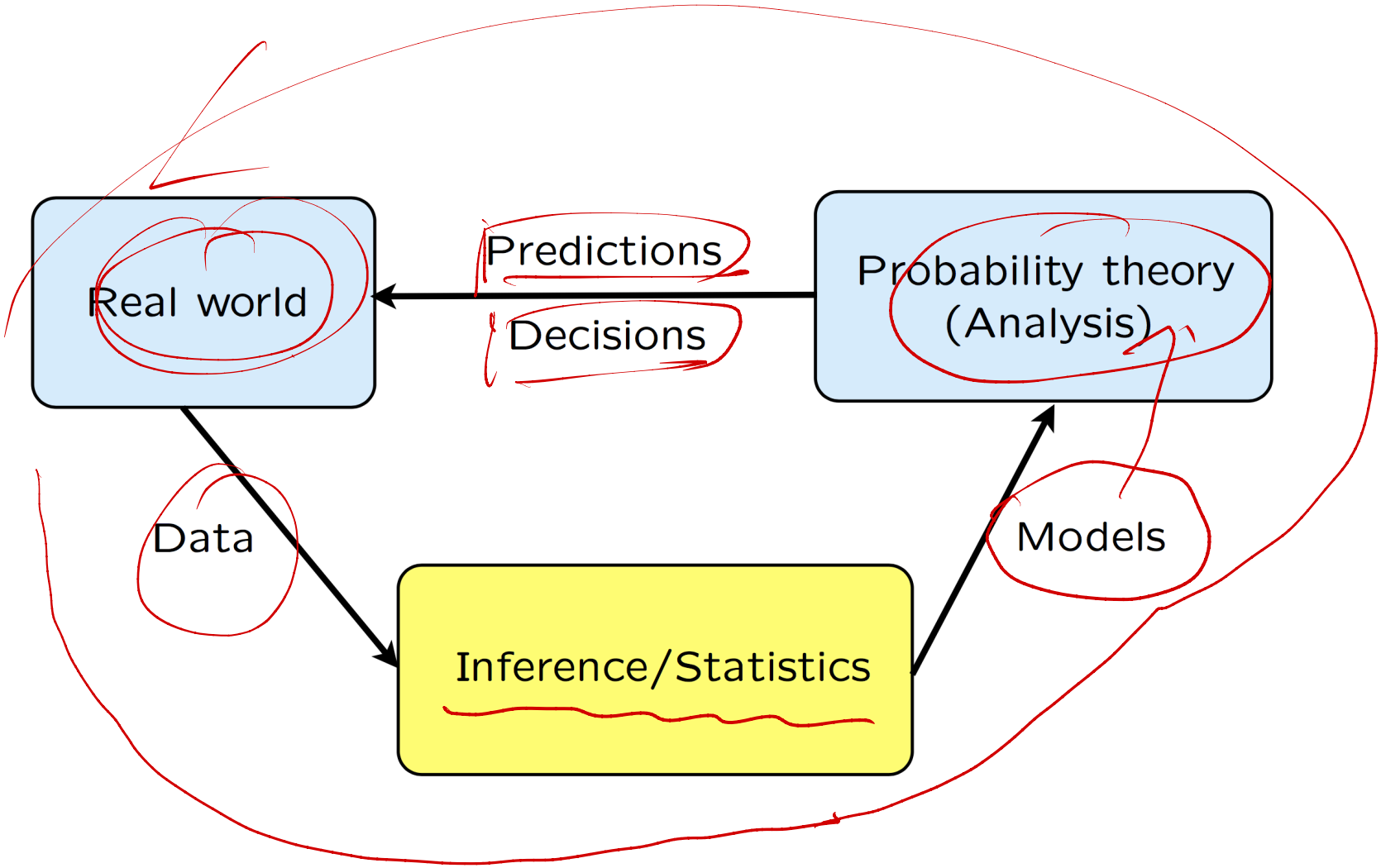
School of Information Science and Technology
ShanghaiTech University

May 9, 2023

Outline

Outline

Inference


Real world
Predictions
Decisions
Probability theory
(Analysis)
Data
Models
Inference/Statistics

Statistical Inference

- The process of extracting information about an unknown variable or an unknown model from available data
- Called "Learning" in CS
- Called "Signal Processing" in EE
- A typical question is: given a sample $X_1, \ldots, X_n \sim F$, how do we infer F or some features of F?

Core Parts of Statistical Inference

- Point Estimation
- Interval Estimation (Confidence Interval)
- Hypothesis Testing

Point Estimation

Point Estimation refers to providing a single "best guess" of some quantity of interest such as

- a parameter θ (possibly multi-dimensional) in a parametric model
- a CDF F
- a probability density function f
- a prediction for a future value Y of some random variable

Interval Estimation

Definition

An interval estimate of a real-valued parameter θ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. When $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made. The random interval $[L(\mathbf{X}), U(\mathbf{X})]$ is called an interval estimator.

Confidence Interval

Definition

A $1-\alpha$ confidence interval for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \forall \theta \in \Theta.$$

In words, (a, b) traps θ with probability $1-\alpha$. We call $1-\alpha$ the coverage of the confidence interval. If θ is a vector, then we use a confidence set instead of an interval.

Hypothesis Testing

Start with a finite number of competing hypotheses and use the available data to decide which of them is true.

- Example 1: given a noisy picture, decide whether there is a person in the picture or not
- Example 2: given a noisy received signal, decide whether symbol 1 or 0 was sent by the transmitter
- Example 3: given a set of trials with three alternative medical treatments, decide which treatment is the most effective

Hypothesis Testing

- A hypothesis is a statement about the data
- The two complementary hypothesis in a hypothesis testing problem are called the Null Hypothesis (denoted by H_0) and the Alternative Hypothesis (denoted by H_1).

Example: Testing if a coin is Fair

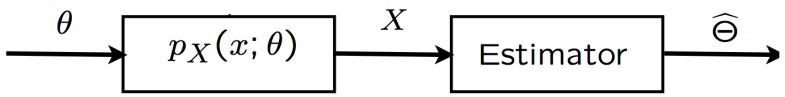
- $X_1, \ldots, X_n \sim Bernoulli(p)$ are n independent coin flips.
- H_0: the coin is fair (Null Hypothesis)
- H_1: the coin is not fair (Alternative Hypothesis)
- $H_0 : p = \frac{1}{2}$ versus $H_1 : p \neq \frac{1}{2}$
- It seems reasonable to reject H_0 if $T = |\hat{p}_n - \frac{1}{2}|$ is large

Bayesian versus Classical Statistical Inference

- Difference relates to the nature of the unknown models or variables
- Treated as random variables with prior (known) distributions: **Bayesian approach**
- Treated as an unknown constants: **classical/frequentist approach**

Bayesian vs. Classical

- Inference using the Bayes rule:
  unknown Θ and observation X are both random variables
  - Find $p_{\Theta|X}$
- Classical statistics: unknown constant θ

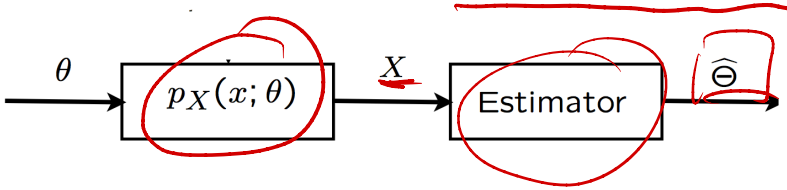


  - also for vectors X and θ: $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$
  - $p_X(x;\theta)$ are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ

Outline

Classical Statistical Inference

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
- Estimation: design an **estimator** $\widehat{\Theta}$, to "keep estimation **error** $\widehat{\Theta} - \theta$ small"

Inference Rule: Maximum Likelihood Estimation (MLE)

PDF: $f(x;\theta)$ = PDF: a function of x with θ fixed; likelihood: a function of θ with x fixed;

- Joint distribution of the vector of observations $X = (X_1, \ldots, X_n)$: PMF $P_X(x;\theta)$ (or PDF $f_X(x;\theta)$)
- θ: unknown (scalar or vector) parameter θ.
- We observe a particular value $x = (x_1, \ldots, x_n)$ of X, then a **maximum likelihood estimate** (MLE) is a value of the parameter that maximizes the numerical function $P_X(x_1, \ldots, x_n;\theta)$ (or $f_X(x_1, \ldots, x_n;\theta)$) over all θ:

$$\hat{\theta}_n = \arg\max_{\theta} P_X(x_1, \ldots, x_n;\theta)$$

$$\hat{\theta}_n = \arg\max_{\theta} f_X(x_1, \ldots, x_n;\theta)$$

MLE under Independent Case

- Observations X_i are independent, and we observe a particular value $x = (x_1, \ldots, x_n)$ of X.
- We define the log-likelihood function as follows:

$$\log[P_X(x_1, \ldots, x_n; \theta)] = \log \prod_{i=1}^{n} P_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log[P_{X_i}(x_i; \theta)]$$

$$\log[f_X(x_1, \ldots, x_n; \theta)] = \log \prod_{i=1}^{n} f_{X_i}(x_i; \theta) = \sum_{i=1}^{n} \log[f_{X_i}(x_i; \theta)]$$

MLE under Independent Case

- Thus a **maximum likelihood estimate** (MLE) under independent case is a value of the parameter that maximizes the numerical function $P_X(x_1, \ldots, x_n; \theta)$ (or $f_X(x_1, \ldots, x_n; \theta)$) over all θ:

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \log[P_{X_i}(x_i; \theta)]$$

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \log[f_{X_i}(x_i; \theta)]$$

Example: Biased Coin Problem

Coin : (P)

1°. n coin tosses; n independent Bernoulli trials.

$X_1, \dots X_n \sim \text{Bern}(p)$.

2°. $x_1, \dots x_n \in \mathbb{R}$ (observation value)

$x_i = 0$ or 1.

$P(X_1 = 1) = p$

$P(X_1 = 0) = 1-p$

$\Rightarrow P(X_1 = x) = p^x (1-p)^{1-x}$, $x = 0$ or 1

$$P_{\mathbf{X}}(\mathbf{x}; p) = \prod_{i=1}^{n} P_{X_i}(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n-\sum_{i=1}^n x_i} = p^{S_n}(1-p)^{n-S_n}$$

$S_n = X_1 + \dots + X_n$: # of heads in n coin tosses.

3°. $\log P_X(x; p) = S_n \log p + (n - S_n)\log(1-p) = f(p)$.

$$\hat{p}_{MLE} = \arg\max_p f(p)$$

[$f'(p) = 0$; $f''(p) \le 0$]

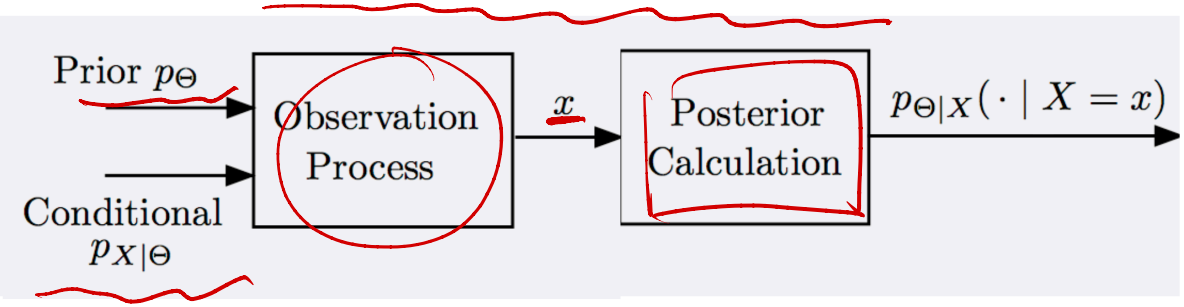
$$\Rightarrow \hat{p}_{MLE} = \frac{1}{n} S_n = \frac{1}{n}(X_1 + \dots + X_n)$$

Example: Biased Coin Problem

Outline

The Bayesian Statistical Inference Framework

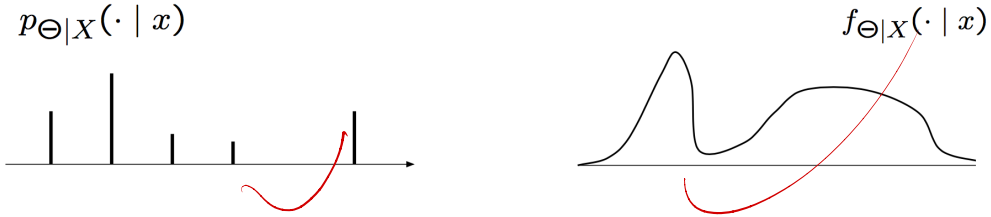
- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

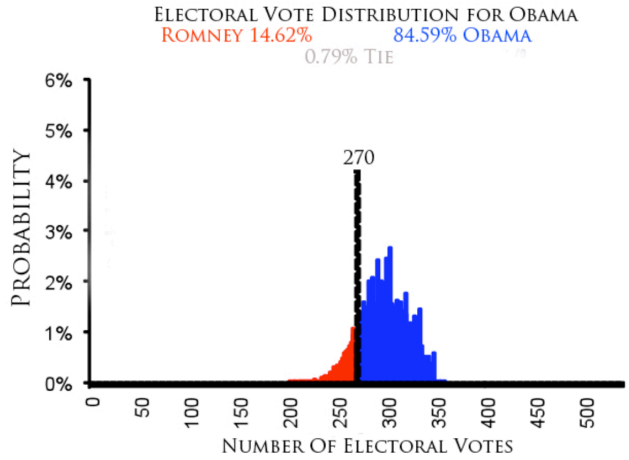


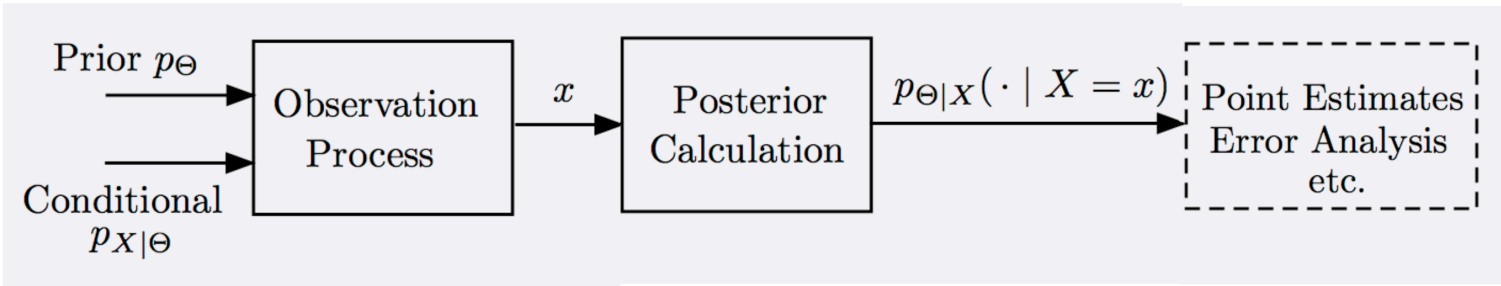
- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

The Output of Bayesian Statistical Inference

The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$







General LOTP

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

General Bayes' Rule

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

General Bayes' Rule: Another Perspective

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')}.$$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \,|\, \theta')}.$$

- Θ continuous, X discrete:

$$f_{\Theta|X}(\theta \,|\, x) = \frac{f_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\int f_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')\, d\theta'}.$$

- Θ continuous, X continuous:

$$f_{\Theta|X}(\theta \,|\, x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta)}{\int f_\Theta(\theta') f_{X|\Theta}(x \,|\, \theta')\, d\theta'}.$$

Example of Inference Rule: The Maximum A Posteriori Probability (MAP)

- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes over θ the posterior distribution $p_{\Theta|X}(\theta|x)$ (if Θ is discrete) or $f_{\Theta|X}(\theta|x)$ (if Θ is continuous).
- Equivalently, it selects $\hat{\theta}$ that maximizes over θ:
  - $p_\Theta(\theta)p_{X|\Theta}(x|\theta)$: if both Θ and X are discrete
  - $p_\Theta(\theta)f_{X|\Theta}(x|\theta)$: if Θ is discrete and X is continuous
  - $f_\Theta(\theta)p_{X|\Theta}(x|\theta)$: if Θ is continuous and X is discrete
  - $f_\Theta(\theta)f_{X|\Theta}(x|\theta)$: if both Θ and X are continuous

The MAP Rule for Point Estimates

- An estimator is a random variable of the form $\hat{\Theta} = g(X)$, for some function g. Different choices of g correspond to different estimators.
- An estimate is the value $\hat{\theta}$ of an estimator, as determined by the realized value x of the observation X.
- Once the value x of X is observed, the MAP estimator, sets the estimate $\hat{\theta}$ to a value that maximizes the posterior distribution over all possibly values of θ.

Example: Inference of A Biased Coin

1° $\Theta \sim Unif(0,1) = Beta(1,1)$; # of heads $X|_{\Theta=\theta} \sim Bin(n,\theta)$.

By Beta-Binomial conjugacy, $\Theta|_{X=k} \sim Beta(1+k, 1+n-k)$

Another Inference Rule

$\hat{\Theta} = E[\Theta|_{X=k}] = \frac{k+1}{k+1+n-k+1} = \frac{k+1}{n+2}$ $\therefore \hat{\Theta} = \frac{X+1}{n+2}$.

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim Unif(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MAP estimator of Θ.

2°. MAP: $f_{\Theta|X=k}(\theta) \propto \theta^k(1-\theta)^{n-k}$. $\theta \in (0,1)$

$\hat{\Theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} f_{\Theta|X=k}(\theta) = \underset{\theta}{\operatorname{argmax}} \theta^k(1-\theta)^{n-k}$

$= \underset{\theta}{\operatorname{argmax}} [\log(\theta^k(1-\theta)^{n-k})]$

Solution

$g(\theta) = \log(\theta^k (1-\theta)^{n-k}) = k\log\theta + (n-k)\log(1-\theta)$

$g'(\theta) = 0; \quad g''(\theta) \le 0.$

$\Rightarrow \theta^* = \frac{k}{n}$;

$\Rightarrow \hat{\theta}_{MAP|X=k} = \frac{k}{n}$. estimate . (real number)

$\hat{\theta}_{MAP|X} = \frac{X}{n}$; estimator

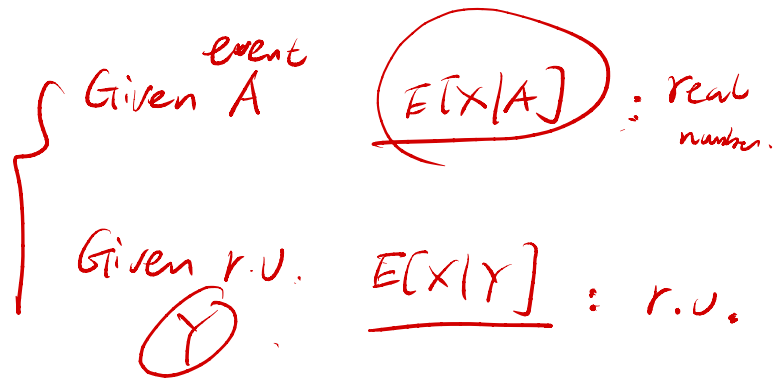
(r.v.)

MAP + Unif prior distribution = MLE.

Bayesian. classical.

Solution

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 $P(A_i) > 0$, $i = 1, 2, \ldots, n$:

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

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

Conditional PDF $f_{X|A}(x) = \lim_{\delta \to 0} \frac{P(x \le X \le x+\delta \,|\, A)}{\delta}$.

- 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 $A =$ "$a \le X \le b$" and $P(A) > 0$, then

$$f_{X|A}(x) = \frac{\mathbf{1}_{x\in[a,b]}}{P(A)} \cdot f_X(x)$$

$$= \begin{cases} \frac{f_X(x)}{P(A)} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$= \lim_{\delta \to 0} \frac{P(A \,|\, x \le X \le x+\delta) \cdot P(x \le X \le x+\delta)}{\delta \cdot P(A)}$$

$$= \lim_{\delta \to 0} \frac{1}{P(A)} \left[P(A \,|\, x \le X \le x+\delta) \right] \cdot \frac{P(x \le X \le x+\delta)}{\delta}$$

$$= \frac{1}{P(A)} \cdot P(A \,|\, X = x) \cdot f_X(x)$$

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

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

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

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

Method 1:

$$f_{X|A}(x) = \begin{cases} \frac{f(x)}{P(A)} = \lambda e^{-\lambda(x-1)} & \text{if } x > 1 \\ 0 & \text{otherwise} \end{cases}$$

PDF. $f(x) = \lambda e^{-\lambda x}$, $x > 0$.

$X \sim Expo(\lambda)$.

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

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

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

2^0. $E(X|X>1) = \int_1^\infty x \cdot f_{X|A}(x)\,dx = \int_1^\infty x \cdot \lambda e^{-\lambda(x-1)}\,dx = 1 + \frac{1}{\lambda}$

$E(X^2|X>1) = \int_1^\infty x^2 f_{X|A}(x)\,dx = \int_1^\infty x^2 \lambda e^{-\lambda(x-1)}\,dx$

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

$= \frac{1}{\lambda^2}$.

Solution

Memoryless. $\forall s,t>0$

Method 2 : ① $P(X>s+t \mid X>t) = P(X>s)$

? $E[X \mid X<1]$

Let $t=1 \Rightarrow P(X>s+1 \mid X>1) = P(X>s)$.

$\Rightarrow P(X-1>s \mid X>1) = P(X>s)$.

$E[X-1 \mid X>1] = E[X]$ ✓

$E[X \mid X>m] = m+\frac{1}{\lambda}$.

② $E[X \mid X>1] = E[(X-1)+1 \mid X>1] = E[X-1 \mid X>1] + 1$

$= E[X] + 1$

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

③ $Var(X \mid X>1)$

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

if $X \sim Geom(p)$; $E[X \mid X>1]$? $Var[X \mid X>1]$?

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: $E(Y|A)$ Real Number
- Conditional Expectation given a random variable: $E(Y|X)$

Intuition for $E(Y|A)$

① Y r.v.

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

$y_1, \ldots, y_n$

② $E[Y|A]$

n measurements.

where event A occurs.

I_j : in the jth measurements.

$$\approx \frac{\sum_{j=1}^{n} I_j \cdot y_j}{\sum_{j=1}^{n} I_j}$$

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

$E(T) = 70$;

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

---

if $T \sim Expo(\lambda)$;

$E[T \mid T \geq 20] = 20 + E[T] = 90$

Law of Total Expectation

LoTE

$E[X|X\le 1]$?

$E[X|X>1]$

$X \sim Exp(\lambda)$

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

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

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

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

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

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!)

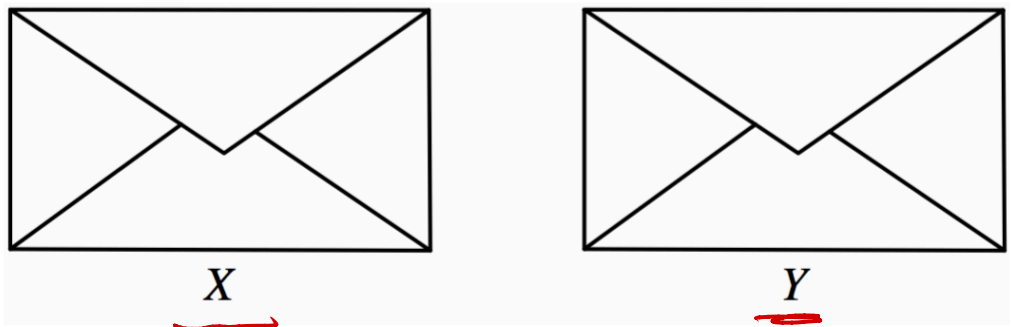


FIGURE 9.1
Two envelopes, where one contains twice as much money as the other. Either $Y = 2X$ or $Y = X/2$, with equal probabilities. Which would you prefer?

Solution

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

LOTE :

$$E(Y) = \underline{E[Y \mid Y=2X]}\, P(Y=2X) + E\left[Y \mid Y=\frac{X}{2}\right] P\left(Y=\frac{X}{2}\right)$$

$$\neq E[2X]$$

$$E[Y \mid Y=2X] = \underline{E[2X \mid Y=2X]} \neq E[2X]$$

Geometric Expectation Redux

$X \sim Geom(p)$. Find $E[X]$?

Coin Tosses.

# of tosses before 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 $E[X] \overset{LOTE}{=} E[X|O_1=H)P(O_1=H) + E[X|O_1=T) \cdot P(O_1=T)$

$0 \cdot p$

$+ [1+E[X]) \cdot (1-p)$

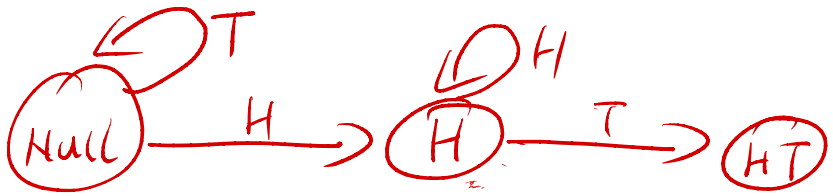
$\Rightarrow E[X] = \frac{1-p}{p}$

Time until *HH* vs. *HT*

H: Head ; T: Tail.

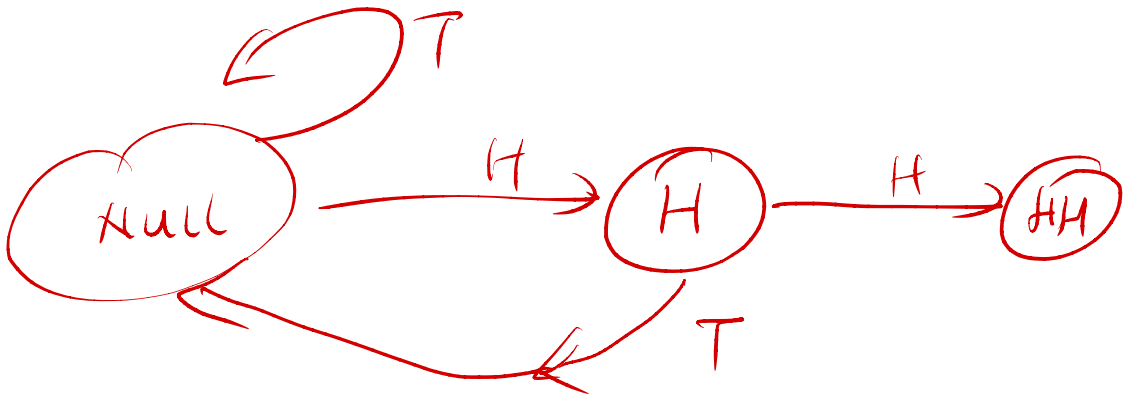
PGF.
LoTE.
Markov chain
Renewal process.
Martingale.

① HT : partial progress



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.

$W_{HT} = W_1 + W_2$

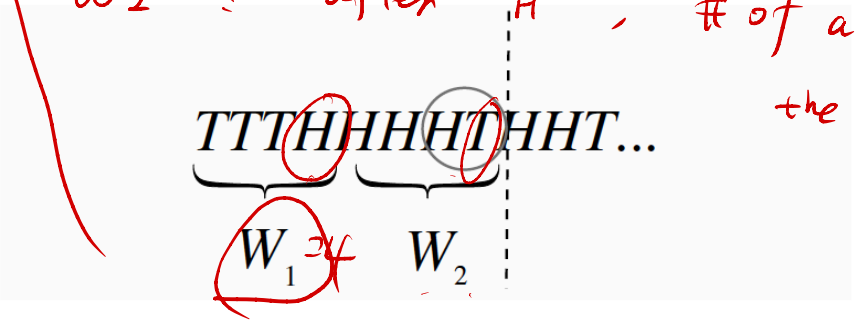
W_1 : # of tosses until the "H" appears for the first time.

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

$X \sim FS(p)$.

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

FS : Memoryless.



$W_1 \sim FS(1/2)$; $W_2 \sim FS(1/2)$

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

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

$= 2 + 2 = 4$.

Solution

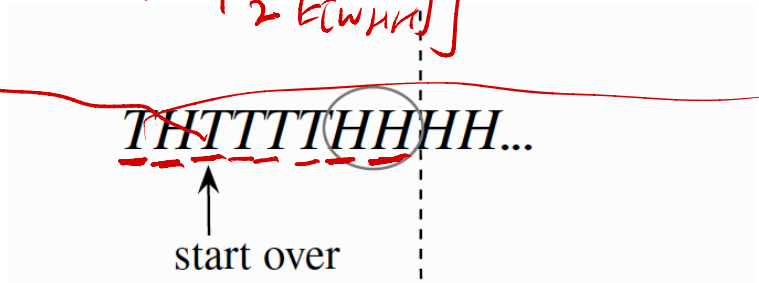
$E(W_{HH})$: First-step Analysis. O_1: outcome of 1st toss; O_2: ... 2nd toss.

1^0 $E(W_{HH}) \overset{\text{LoTE}}{=} E(W_{HH} \mid O_1=H)\,P(O_1=H) + E(W_{HH} \mid O_1=T)\cdot P(O_1=T)$

$\frac{1}{2}$ $\qquad$ $+\,[1+E(W_{HH})]\cdot\frac{1}{2}$.

$\downarrow$?

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



LoTE with extra conditioning.

2^0. $E(W_{HH} \mid O_1=H) = E[W_{HH} \mid O_1=H, O_2=H]\cdot P(O_2=H \mid O_1=H)$ $\qquad$ 2 $\qquad$ $P(O_2=H)=\frac{1}{2}$ $\qquad$ O_1, O_2 independent.

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

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

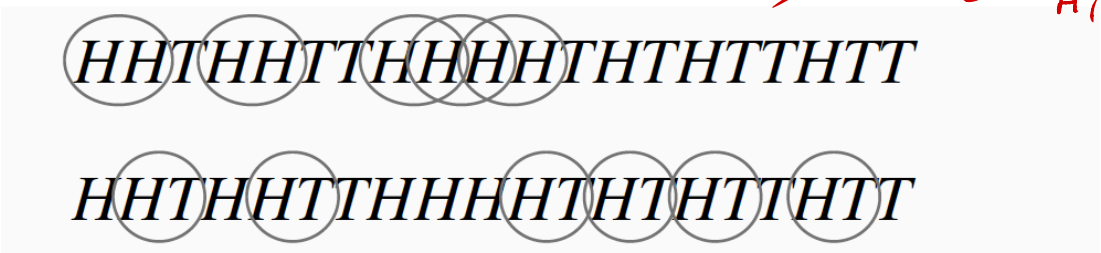
$= 2+\frac{1}{2}E(W_{HH})$.

Solution

3°. $E[W_{HH}] = [2 + \frac{1}{2}E[W_{HH}]]\cdot\frac{1}{2} + [1 + E[W_{HH}]]\cdot\frac{1}{2}$

$= \frac{3}{2} + \frac{3}{4}E[W_{HH}]$

$\Rightarrow E[W_{HH}] = 6 > E[W_{HT}] = 4$



Conditional Expectation Given An R.V.

$g(x) = E[Y \mid X=x]$: Real Number. ; estimate.

$g(X) = E[Y \mid X]$: Random Variable. ; 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 $g(x)$.

$A = "X=x"$

$X \longrightarrow \boxed{g(\ )} \longrightarrow \hat{Y} = g(X)$: estimator.

$X = x$.

$g(x)$ estimate.

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 $X \sim \text{unif}(0,1)$.

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

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

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.

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

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

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

② $g(x) \to g(X)$ $x \to X$

Solution

Dropping What's Independent

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

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

Theorem

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

Taking Out What's Known

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

① $g(x) = E[h(x)\cdot Y | X=x]$

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

② $g(X)$

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

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

Theorem

For any function h,

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

Linearity

① $g(x) = E[Y_1+Y_2 / X=x]$

$= E[Y_1 / X=x] + E[Y_2/X=x]$

② $g(X) = E[Y_1|X] + E[Y_2/X]$

Theorem

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

$E[Y|X_1+X_2] \neq E[Y|X_1] + E[Y|X_2]$

Example

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

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

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

$= E[X_1 + \cdots + X_n|S_n]$

$= E[S_n|S_n]$

$= S_n$.

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

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

$= S_n$

3°. $E[X_1|S_n] = \frac{1}{n} S_n$.

Adam's Law

The Law of Iterated Expectation

Theorem

For any r.v.s X and Y,

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

Proof

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

LHS ✓ RHS.

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

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

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

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

$$= E[Y] \quad \text{RHS.}$$

Adam's Law and LOTE

discrete X.

1°. LOTE. $E[Y] = \sum_x E[Y|X=x] \cdot P(X=x)$

2°. Adam's Law. $g(x) = E[Y|X=x]$; $g(X) = E[Y|X]$

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

$= \sum_x g(x) \cdot P(X=x)$

$= \sum_x E[Y|X=x] \cdot P(X=x)$

LOTE $= E[Y]$

Adam's Law with Extra Conditioning

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

② Adam's Law. $E[E[Y|X]] = E[Y]$

Theorem

③ $\hat{E}[\hat{E}[Y|X]] = \hat{E}[Y]$

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

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

1° $\hat{E}(Y) = E(Y|Z)$.

2° $\hat{E}(Y|X) = E(Y|X,Z)$

3° $\hat{E}[\hat{E}[Y|X]] = E(\hat{E}(Y|X)|Z)$

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

Conditional Variance

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

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

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

Definition

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

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

The *conditional variance of Y given X* is

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

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

$= E[E[Y^2|X]] - E[g^2(X)]$ $\overset{\text{Adam's Law}}{=} E[Y^2] - E[g^2(X)]$

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

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

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

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 $\mathrm{Var}(N)$.

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

Adam's Law

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

Solution

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

$= \mathrm{Var}\left(\sum_{j=1}^{n} X_j \mid N=n\right) = \mathrm{Var}\left(\sum_{j=1}^{n} X_j\right)$

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

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

Eve's Law.

$\Rightarrow \mathrm{Var}(X) = E[\mathrm{Var}(X \mid N)] + \mathrm{Var}[E(X \mid N)]$

$= E[N\sigma^2] + \mathrm{Var}(N \cdot \mu)$

$= \sigma^2 E[N] + \mu^2 \cdot \mathrm{Var}(N)$

Solution

Outline

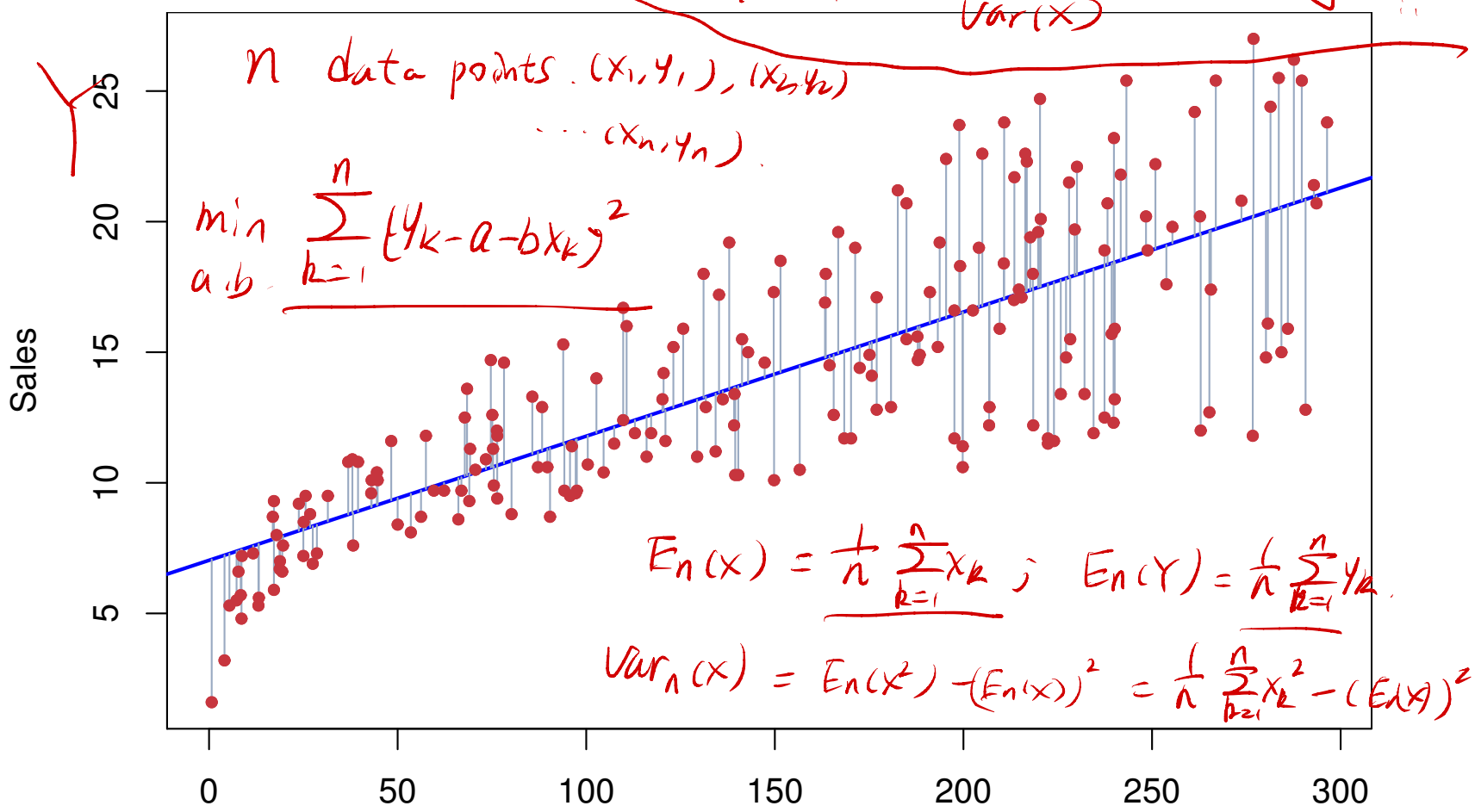
Basic Problem

- 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 Regression

$\hat{Y} = E[Y|X] = a + bX$

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



n data points $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$.

$\min_{a,b} \sum_{k=1}^{n} (y_k - a - bx_k)^2$

$E_n(X) = \frac{1}{n}\sum_{k=1}^{n} x_k$; $E_n(Y) = \frac{1}{n}\sum_{k=1}^{n} y_k$.

$Var_n(X) = E_n(X^2) - (E_n(X))^2 = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - (E_n(X))^2$

$Cov_n(X,Y) = E_n(XY) - E_n(X) \cdot E_n(Y)$

$= \frac{1}{n}\sum_{k=1}^{n} x_n y_n - E_n(X) \cdot E_n(Y)$.

$n \to \infty$; SLLN

$E_n \to E$.

Linear Regression

An extremely widely used method for data analysis in statistics is *linear regression*. In its most basic form, the linear regression model uses a single explanatory variable X to predict a response variable Y, and it assumes that the conditional expectation of Y is *linear* in X:

$$E(Y|X) = a + bX.$$

$\hat{Y} = a + bX$.

$Y = \hat{Y} + \varepsilon$ → error.

(a) Show that an equivalent way to express this is to write

$$Y = a + bX + \epsilon,$$

where ϵ is an r.v. (called the *error*) with $E(\epsilon|X) = 0$.

(b) Solve for the constants a and b in terms of $E(X)$, $E(Y)$, $\mathrm{Cov}(X, Y)$, and $\mathrm{Var}(X)$.

Solution

(a) 1°. if $Y = a + bX + \varepsilon$, $E[\varepsilon|X] = 0$ $\overset{\text{def}}{\Rightarrow}$ $E[Y|X] = a + bX$.

$$E[Y|X] = E[a + bX + \varepsilon|X] = a + b\,E[X|X] + E[\varepsilon|X]$$
$$= a + b \cdot X + 0 = a + bX$$

2°. if $E[Y|X] = a + bX$ $\overset{\text{def}}{\Rightarrow}$ $Y = a + bX + \varepsilon$, $E[\varepsilon|X] = 0$

Let $\varepsilon \triangleq Y - (a + bX)$. $\Rightarrow Y = a + bX + \varepsilon$.

$$E[\varepsilon|X] = E[Y - (a + bX)|X] = E[Y|X] - (a + bX)$$
$$= a + bX - (a + bX) = 0.$$

3°. $E[\varepsilon] = E[E[\varepsilon|X]] = E[0] = 0$.

Solution

(b) $\because Y = a + bX + \varepsilon$, $E[Y] = a + bE[X] + E[\varepsilon]$

$= a + bE[X]$

$\Rightarrow a = E[Y] - b \cdot E[X]$;

1° $Cov(X, \varepsilon) = \underbrace{E[\varepsilon X]}_{0} - E[X] \cdot \underbrace{E[\varepsilon]}_{0} = 0$

$E[\varepsilon X] \overset{\text{Adam's law}}{=\!=} E[\underbrace{E[\varepsilon X | X]}] = E[X \cdot \underbrace{E[\varepsilon | X]}_{0}] = 0$

3° $Cov(X, Y) = Cov(X, a + bX + \varepsilon) = \underbrace{Cov(X, a)}_{0} + \underbrace{Cov(X, bX)}$

$+ \underbrace{Cov(X, \varepsilon)}_{0} = b\, Cov(X, X)$

$= b \cdot Var(X)$

$\Rightarrow b = \frac{Cov(X, Y)}{Var(X)}$

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

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

Linear Least Square Estimate

$f(a,b) = E[(Y-a-bX)^2]$

$\min_{a,b} E[(Y-a-bX)^2] = \min_{a,b} f(a,b)$

$(a^*, b^*) \in \operatorname{argmin}_{(a,b)} f(a,b) \Leftarrow ① \frac{\partial f(a,b)}{\partial a}\Big|_{a=a^*,b=b^*} = 0$

② $\frac{\partial f(a,b)}{\partial b}\Big|_{a=a^*,b=b^*} = 0$.

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,

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

③ Hessian Matrix.

$$\begin{vmatrix} \frac{\partial f^2(a,b)}{\partial a^2} & \frac{\partial^2 f(a,b)}{\partial a \partial b} \\ \frac{\partial^2 f(a,b)}{\partial b \partial a} & \frac{\partial^2 f(a,b)}{\partial b^2} \end{vmatrix} \geq 0$$

positive semidefinite.

Proof

Proof

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

X, Y r.v.

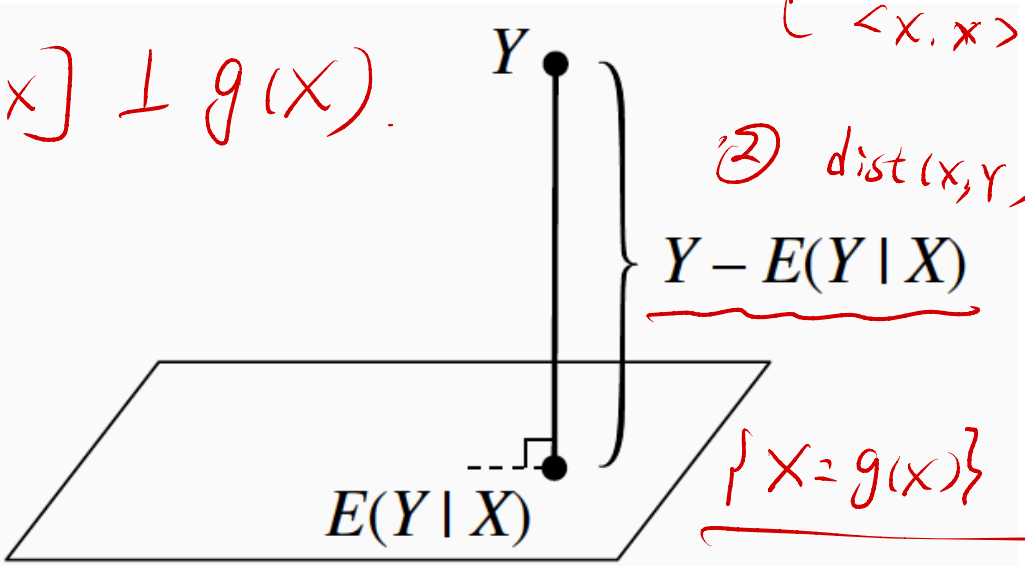
① Inner product $\langle X, Y \rangle = E[XY]$

$\begin{cases} \langle X, Y \rangle = \langle Y, X \rangle \\ \langle \alpha X + \beta Y, Z \rangle = \alpha \langle X, Z \rangle + \beta \langle Y, Z \rangle \\ \langle X, X \rangle \geq 0 \text{ and } \langle X, X \rangle = 0 \Rightarrow X = 0 \end{cases}$

② $\text{dist}(X, Y) = \langle X - Y, X - Y \rangle^{\frac{1}{2}} = \sqrt{E[(X-Y)^2]}$

③ X and Y are "orthogonal" if $E[XY] = 0$, $X \perp Y$

④ $Y - E[Y|X] \perp g(X)$.



$\{X = g(X)\}$

$\{E[X^2] < \infty\}$

⑤ $E[Y|X]$: a projection of Y onto the space of arbitrary functions of X.

Projection Interpretation

$Cov(X,Y) = E(XY) - E(X)\cdot E(Y)$

if $E[X]$ or $E[Y] = 0$ or both

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

$\Rightarrow$ 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.)

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

Proof

$Y - E[Y|X] \perp h(X)$ ✓

$E[(Y - E[Y|X])\, h(X)] \overset{?}{=} 0$ ✓

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

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

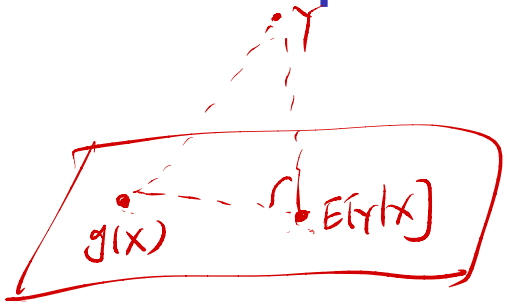
$= E[Y h(X)] - E[E[h(X) \cdot Y(X)]]$

Adam's Law

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

Proof

Prediction Perspective



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

$\leq E[(Y-g(X))^2]$ $\forall g(\cdot)$

- 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).

Proof

1^0. $\hat{Y}$: estimator of Y. $(\hat{Y}=g(X))$.

$$E[(Y-\hat{Y})^2] = E[(Y-g(X))^2] \quad ; \quad Y-g(X) = \underbrace{(Y-E[Y|X])}_{A} + \underbrace{(E[Y|X]-g(X))}_{B}$$

2^0. $$E[(Y-g(X))^2] = E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2] + 2E[(Y-E[Y|X])(E[Y|X]-g(X))]$$

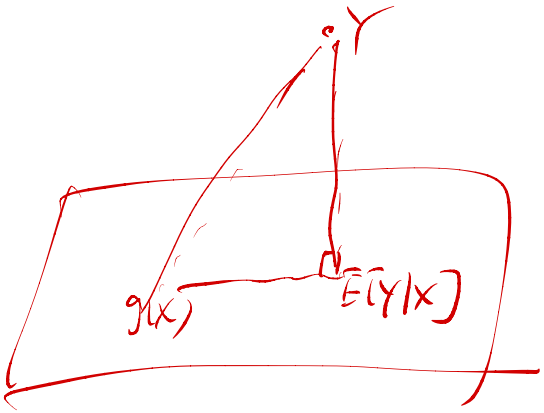
3^0. $E[Y|X]-g(X)$ is still a function of X.

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

$$\Rightarrow E[(Y-E[Y|X])\cdot(E[Y|X]-g(X))]=0$$

Proof

4^0 $E[(Y-g(X))^2] = E[(Y-E(Y|X))^2] + E[(E(Y|X)-g(X))^2]$



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

≥ 0

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

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

$\hat{Y}^* = g^*(X) = E(Y|X)$ MMSE.

Proof

Orthogonality Property of MMSE

Theorem

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

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

(b) Moreover, if the function $g(X)$ *is such that*

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

then $g(X) = E(Y|X)$

Proof

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

$E[(g(X)-E[Y|X])^2] \overset{!}{=} 0 \Rightarrow g(X)=E[Y|X]$

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

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

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

$= 0$

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

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

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

$= 0$

Proof

MMSE for Jointly Gaussian Random Variables

Normal

LLSE : Suboptimal ; low complexity.

MMSE : optimal ; high complexity.

Theorem

Let X, Y be jointly Gaussian random variables. Then

Normal

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

Optimal / low complexity

LLSE : a projection of Y onto the space of arbitrary linear function of X

$(aX+b)$

Proof

Proof sketch:

1°. $Y - L[Y|X] \perp X$.

2°. $E[Y - L[Y|X]] = 0 \Leftrightarrow$ Uncorrelation $\Leftrightarrow$ orthogonal

3°. $Y - L[Y|X]$ is uncorrelated with X.

4°. $Y - L[Y|X]$ and X joint Normal.

$t_1(Y - L[Y|X]) + t_2 X \sim$ Normal.

$t_1 Y + t_2' X \sim$ Normal

5°. $Y - L[Y|X]$ is independent of X. $\Rightarrow$ independent of $\phi(X)$.

6°. $Y - L[Y|X]$ is uncorrelated of $\phi(X)$.

7°. $Y - L[Y|X] \perp \phi(X)$.

8°. $L[Y|X] = E[Y|X]$

Proof

Example: Revisit Biased Coin Problem

1^0. MMSE: $E[\Theta|X=k]$; $\Theta \sim unif(0,1) = Beta(1,1)$.

$E[\Theta|X]$ # of heads $X|\Theta=\theta \sim Bin(n,\theta)$.

By Beta-Binomial conjugacy, $\Theta|X=k \sim Beta(1+k, 1+n-k)$

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim Unif(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta|X)$ and LLSE $L(\Theta|X)$.

$$\Rightarrow E[\Theta|X=k] = \frac{k+1}{1+k+1+n-k} = \frac{k+1}{n+2}$$

$$\Rightarrow E[\Theta|X] = \frac{X+1}{n+2}, \quad \text{MMSE}.$$

Solution

2°. LLSE. $L[\theta|X] = E[\theta] + \frac{Cov(\theta, X)}{Var(X)}(X - E(X))$.

$\theta \sim unif(0,1)$. $\Rightarrow E(\theta) = \frac{1}{2}$; $Var(\theta) = \frac{1}{12}$, $E(\theta^2) = \frac{1}{3}$.

$X|\theta=\theta \sim Bin(n,\theta)$ $\Rightarrow E[X|\theta=\theta] = n\theta$ $\Rightarrow E[X|\theta] = n\theta$.

$Var[X|\theta=\theta] = n\theta(1-\theta)$. $\Rightarrow Var[X|\theta] = n\theta(1-\theta)$

$\Rightarrow E[X] = E[E[X|\theta]] = E[n\theta] = nE[\theta] = \frac{n}{2}$

$Var(X) = E[Var(X|\theta)] + Var[E[X|\theta]]$

$= E[n\theta(1-\theta)] + Var[n\theta]$

$= n(E(\theta) - E(\theta^2)) + n^2 Var(\theta)$

$= n(\frac{1}{2} - \frac{1}{3}) + n^2 \cdot \frac{1}{12} = \frac{n}{12}(n+2)$

Solution

$$\mathrm{Cov}(X,\theta) = E[\theta X] - E[\theta]\cdot E[X]$$

$$= E[E[\theta X|\theta]] - E[\theta]\cdot E[X]$$

$$= E[\theta E[X|\theta]] - E[\theta]\cdot E[X]$$

$$= E[\theta\cdot n\theta] - \frac{1}{2}\cdot\frac{n}{2}$$

$$= n\,E[\theta^2] - \frac{n}{4}$$

$$= n\cdot\frac{1}{3} - \frac{n}{4} = \frac{1}{12}n$$

$$\Rightarrow \text{LLSE}\quad L[\theta|X] = E[\theta] + \frac{\mathrm{Cov}(\theta,X)}{\mathrm{Var}(X)}[X-E(X)]$$

$$= \frac{1}{2} + \frac{\frac{1}{12}n}{\frac{n}{12}(n+2)}\left(X-\frac{n}{2}\right) = \frac{X+1}{n+2} = E[\theta|X]$$

Solution

Remark: Statistical Learning Perspective

$E[\theta|x]$

- 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

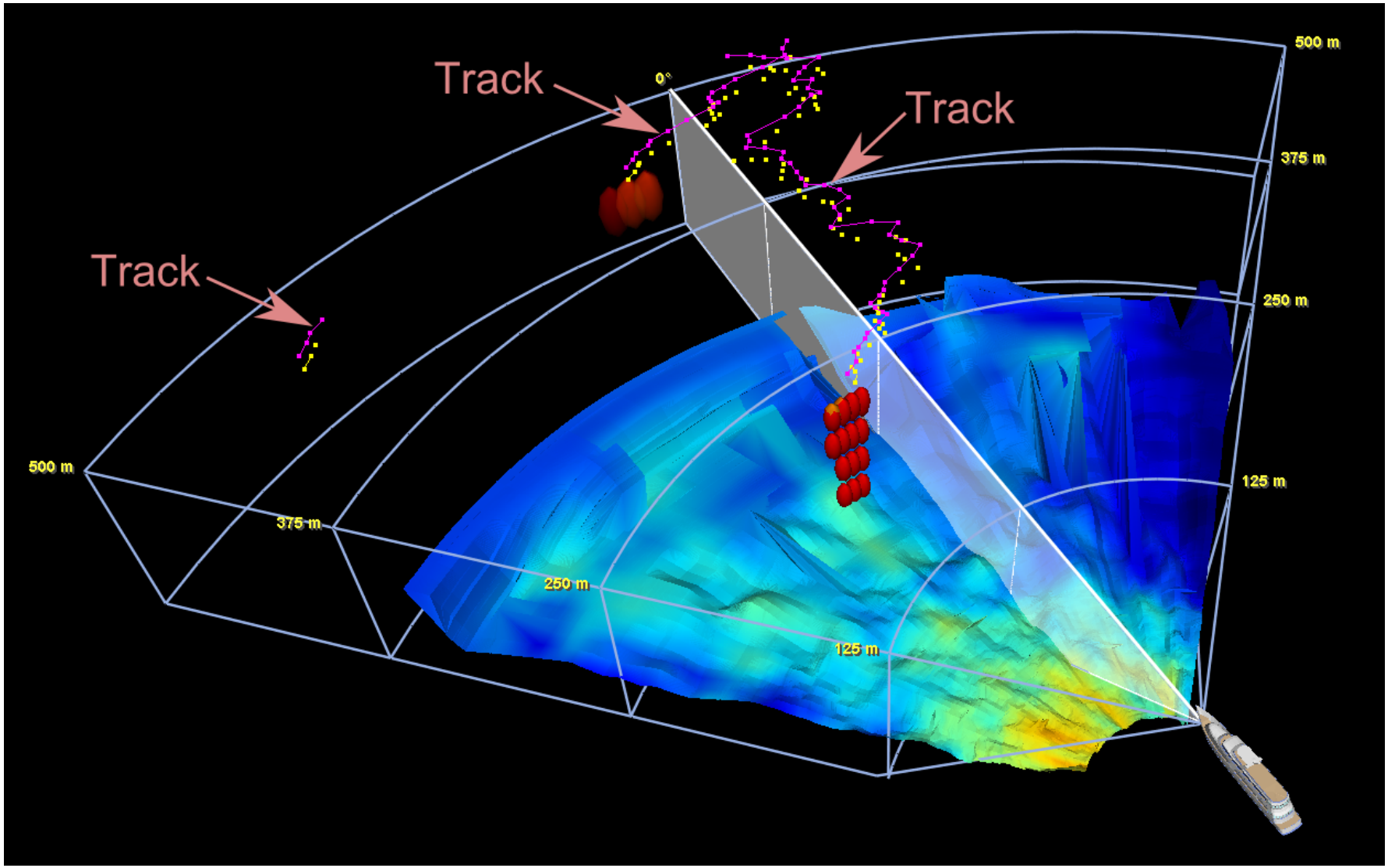
Outline

Milestones in Statistics & Signal Processing

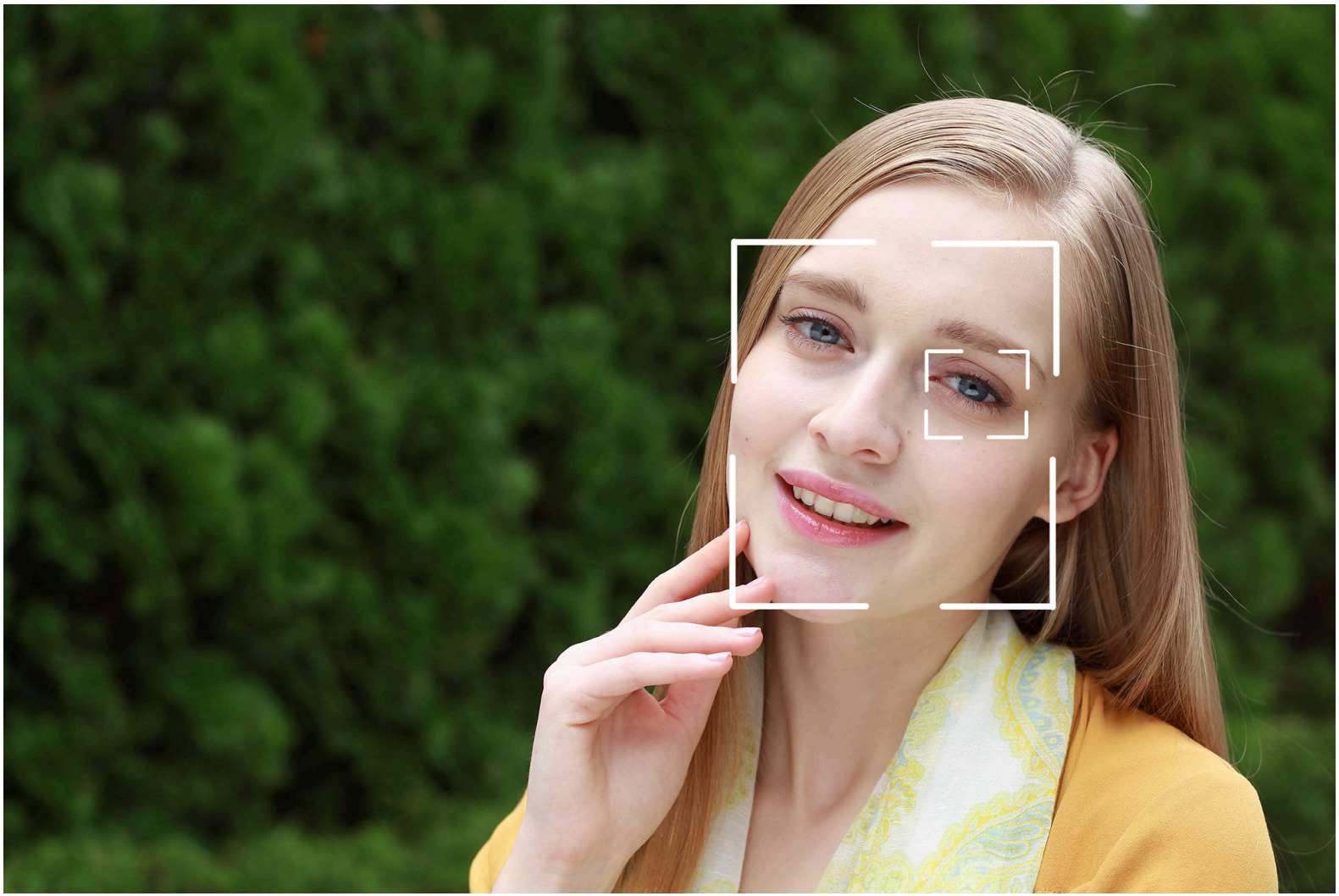
- 1960: Rudoph Emil Kalman (1930-2016) introduced what is known as Kalman filter.

Widely Applications: Location & Navigation & Map Building

Widely Applications: Radar Tracking

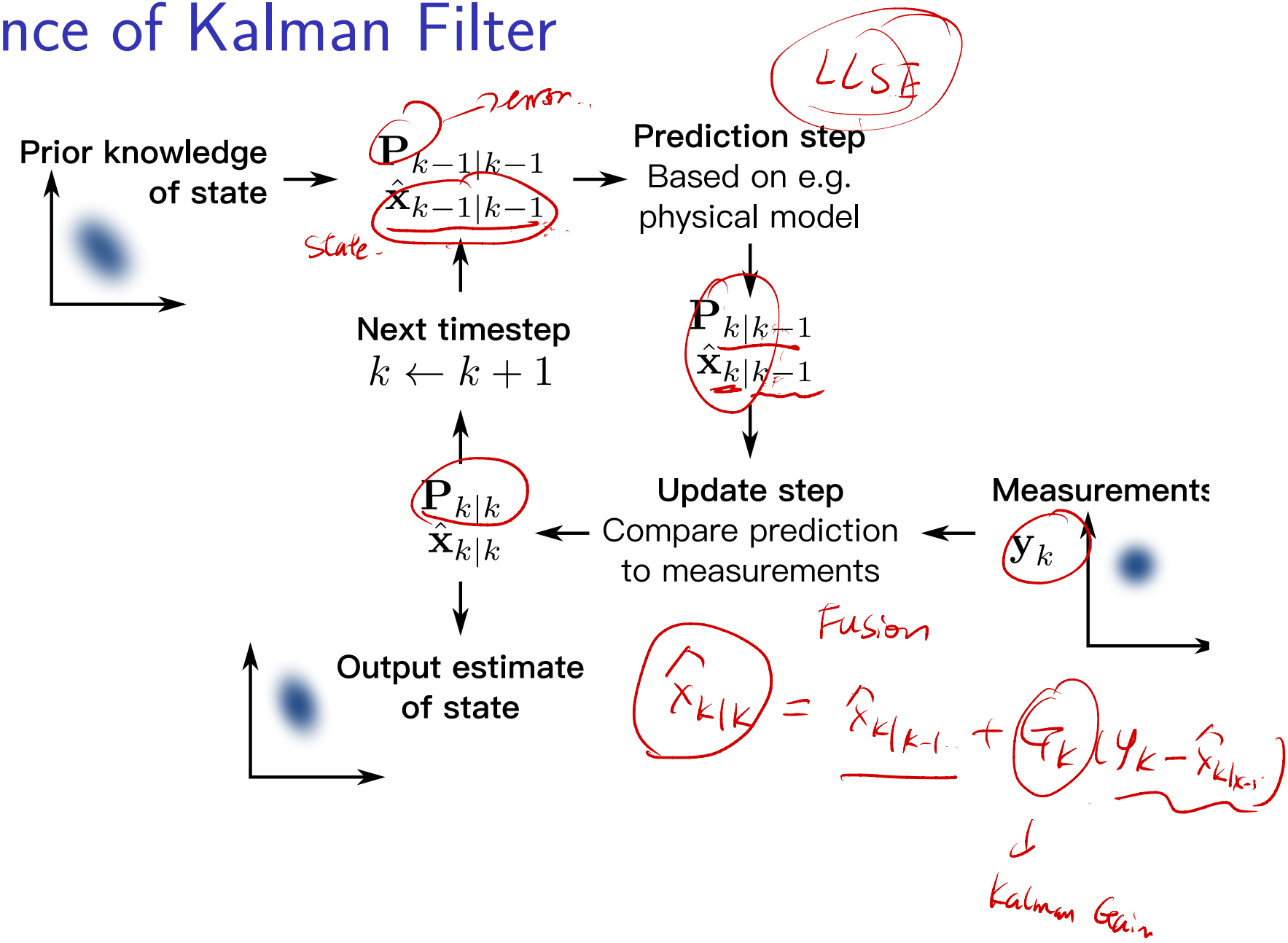


Widely Applications: Human Face & Eye Detection Autofocus



Widely Applications: Animal Eye Detection Autofocus

Essence of Kalman Filter



Essence of Kalman Filter

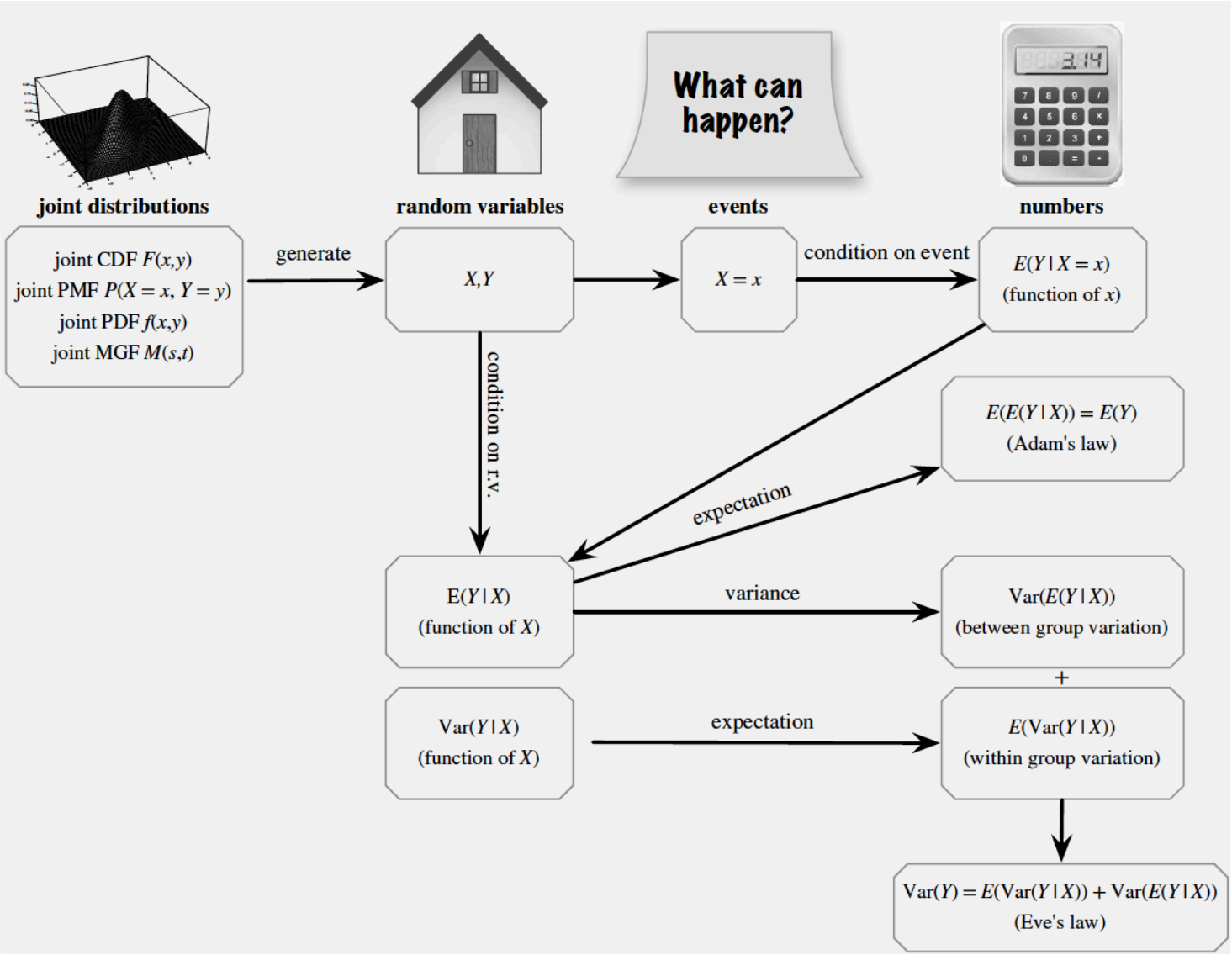
Reasons for Popularity of Kalman Filter

- 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
events
numbers
What can happen?
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
$E(Y \mid X = x)$
(function of x)
condition on r.v.
$E(E(Y \mid X)) = E(Y)$
(Adam's law)
expectation
$E(Y \mid X)$
(function of X)
variance
$\mathrm{Var}(E(Y \mid X))$
(between group variation)
+
$\mathrm{Var}(Y \mid X)$
(function of X)
expectation
$E(\mathrm{Var}(Y \mid X))$
(within group variation)
$\mathrm{Var}(Y) = E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X))$
(Eve's law)

References

- Chapters 9 of **BH**
- Chapters 4 & 6 & 8of **BT**