Lecture 3: Conditional Expectation

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

March 16,18,23,25, 2020

Outline

Motivation

- 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: $E(Y|X)$

Outline

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

$$E(Y|A) = \sum_y yP(Y = y|A),$$

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

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

where the conditional PDF $f(y|A)$ is defined as the derivative of the conditional CDF $F(y|A) = P(Y \leq y|A)$, and can also be computed by a hybrid version of Bayes' rule:

$$f(y|A) = \frac{P(A|Y = y)f(y)}{P(A)}.$$

Direct Computing of $E(Y|A)$

① Y is continuous r.v.

$$P(Y\in(y-\varepsilon, y+\varepsilon)|A) = \frac{P(A|Y\in(y-\varepsilon,y+\varepsilon))P(Y\in(y-\varepsilon,y+\varepsilon))}{P(A)}$$

$$\approx 2\varepsilon f(y|A) = \frac{P(A|Y\in(y-\varepsilon,y+\varepsilon))\cdot 2\varepsilon f(y)}{P(A)}$$

$\lim_{\varepsilon\to 0}$

$$f(y|A) = \frac{P(A|Y=y)f(y)}{P(A)}$$

② $$E(Y|A) = \int_{-\infty}^{\infty} y f(y|A)\,dy, \quad y\in(-\infty,\infty)$$

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

① T : life span.

$E(T) = 70 = E(T \mid T \geq 0)$

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

- Suppose you are 20 years old now.
- Average life time of your country is 70 years old.
- So you have 50 years to live on average. True or not?

② In general, $E(T) \neq E(T \mid T \geq 20)$

③ $T \sim \text{expo}(\lambda)$. memoryless.

Law of Total Expectation

LOTE $\longrightarrow$ LOTP.

$Y = I_B$ (indicator of event B occurring).

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

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

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

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

Outline

Conditional Expectation Given An R.V.

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

① $M=\max(X,Y)$, $L=\min(X,Y)$

By memoryless property, $M-L$ is independent of L

$M-L \sim \text{expo}(\lambda)$

② $E(M|L)$

For $X, Y \overset{\text{i.i.d.}}{\sim} \text{Expo}(\lambda)$, find $E(\max(X,Y)|\min(X,Y))$.

$$E(M|L=l) = E(M-L+L|L=l)$$

$$= E(M-L|L=l) + E(L|L=l)$$

$$= E(M-L) + l$$

$$= \frac{1}{\lambda} + l = g(l)$$

③ $E(M|L) = \frac{1}{\lambda} + L$

$g(L)$

Solution

Outline

Dropping What's Independent

Theorem

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

Taking Out What's Known

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

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

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

Theorem

For any function h,

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

Linearity

$E(Y|X_1+X_2) \neq E(Y|X_1) + E(Y|X_2)$

Theorem

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

Example

By symmetry property, $E[X_1|S_n] = E[X_2|S_n] = \cdots = E(X_n|S_n)$

By Linearity, $E[X_1|S_n] + \cdots + E[X_n|S_n]$

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

$= E[S_n|S_n] = S_n$

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

$\Rightarrow \quad E[X_1|S_n] = \frac{1}{n} S_n$

Adam's Law

W.L.O.G. Focus on the scenario that X and Y are both discrete.

1°. $g(X) = E(Y|X)$

$g(x) = E(Y|X=x) = \sum_y y P(Y=y|X=x)$

Theorem

For any r.v.s X and Y,

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

2°. $E[g(X)] = \sum_x g(x) P(X=x)$

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

$= \sum_y y \sum_x P(Y=y|X=x) \cdot P(X=x) = \sum_y y \cdot \sum_x P(Y=y, X=x)$

$= \sum_y y \cdot P(Y=y) = E(Y)$

Proof

Adam's Law and LOTE

LOTE: $E(Y) = \sum_x E[Y|X=x]\,P(X=x)$

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

via $g(x) = E(Y|X=x)$, $g(X) = E(Y|X)$

since $E[E(Y|X)] = E[g(X)] = \sum_x g(x)\,P(X=x)$

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

$= E(Y)$

Adam's Law with Extra Conditioning

$\hat{E}(\cdot) = E(\cdot|Z)$ Conditional expectation is expectation.

Adam's Law. $\hat{E}(\hat{E}(Y|X)) = \hat{E}(Y)$

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

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

Theorem

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

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

important! Bandit / RL. Bellman equation.

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

Conditional Variance

Definition

The *conditional variance of Y given X* is

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

This is equivalent to

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

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

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

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

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

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

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

Example: Random Sum

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

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

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

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

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

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

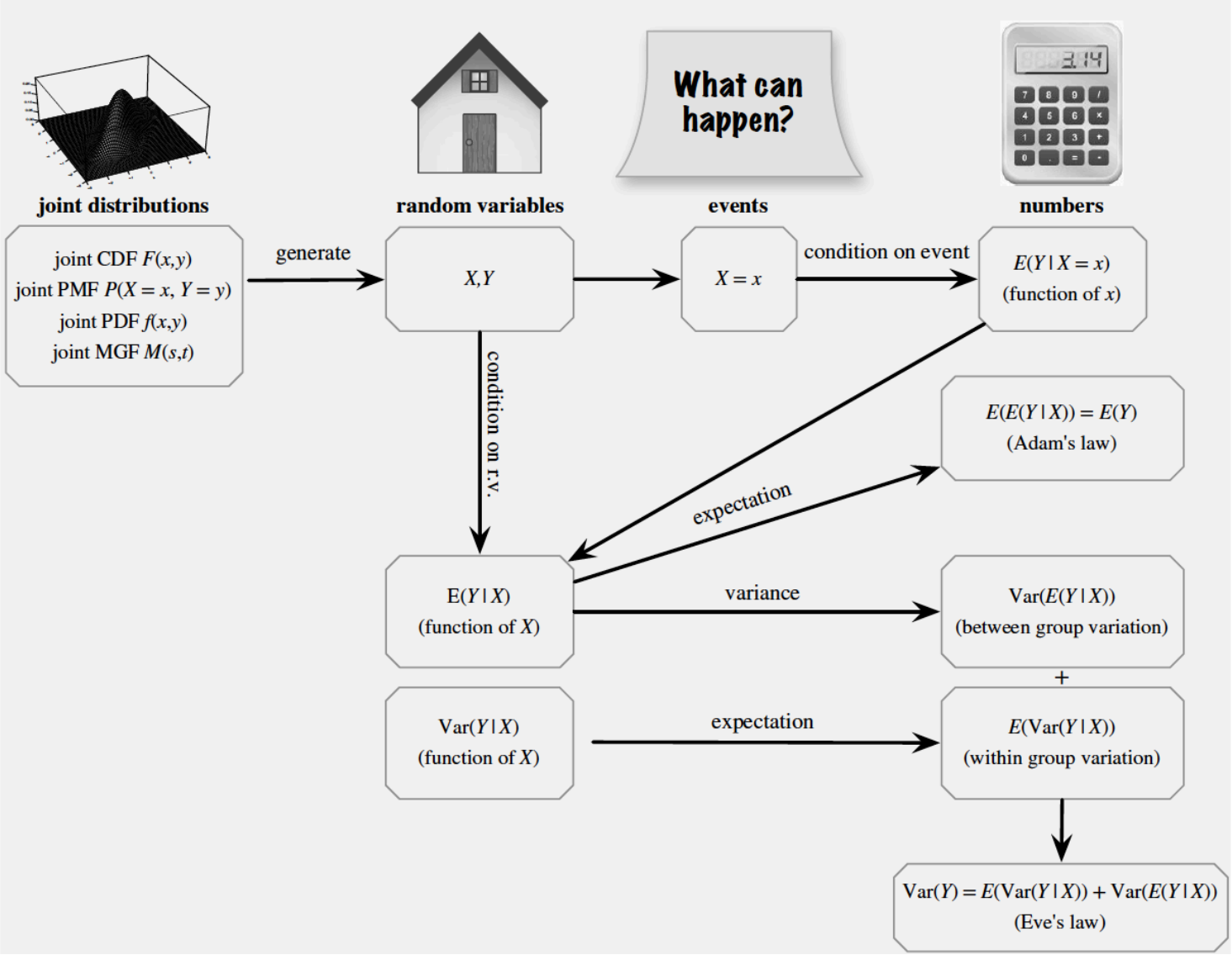
Solution

③ By Eve's Law, $\mathrm{Var}(X) = E[\mathrm{Var}(X|N)] + \mathrm{Var}[E(X|N)]$

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

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

Summary


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

Outline

Prediction Perspective

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

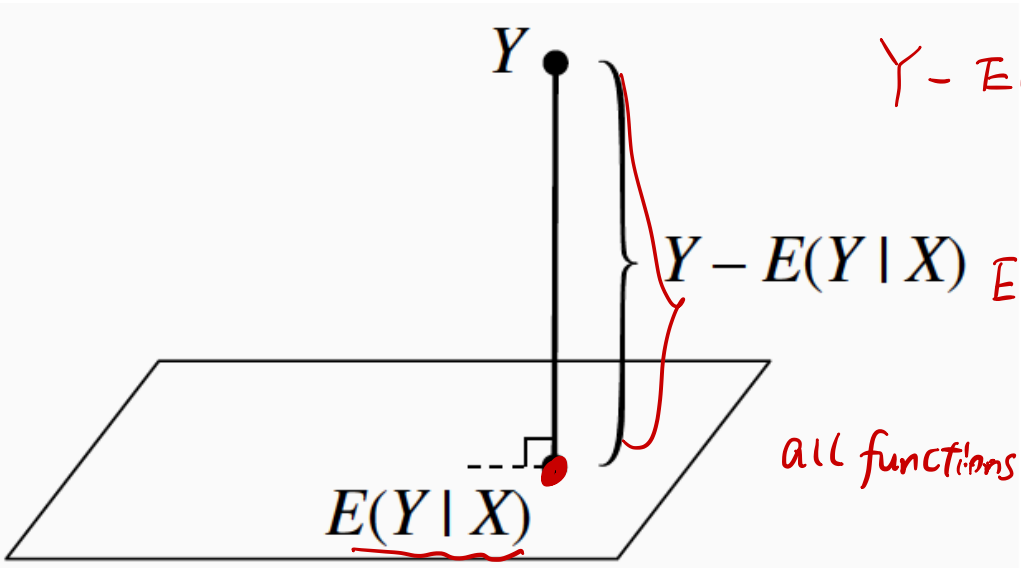
Geometric Perspective

1°. Projection of Y onto the plane of all functions of X.

2°. $Y - E(Y|X)$ is orthogonal to the plane.

$a \perp b$

$E(a \cdot b) = 0$



$Y - E(Y|X) \perp \{h(X)\}$

$\Downarrow$

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

all functions of X. $\{h(X)\}$

Projection Interpretation

Theorem

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

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

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

Proof

1^0. $Z = Y - E(Y|X)$, $E(Z) = E[Y - E(Y|X)] = E(Y) - E[E(Y|X)]$
$= E(Y) - E(Y) = 0$

$$Cov(Y - E(Y|X), h(X)) = Cov(Z, h(X)).$$
$$= E[(Z - EZ)(h(X) - E(h(X)))]$$
$$= E[Z(h(X) - E(h(X)))] = E[Zh(X)] - E[h(X)] \cdot E(Z)$$
$$= E[Zh(X)]$$

Thus $Cov(Y - E(Y|X), h(X)) = 0 \iff E(Zh(X)) = 0$

2^0. $E[Zh(X)] = E[(Y - E(Y|X))h(X)]$
$= E(Yh(X)) - E[h(X)E(Y|X)]$
$= E(Yh(X)) - E[E[Yh(X)|X]]$ taking back what's known
$= E(Yh(X)) - E(Yh(X))$ Adam's law
$= 0$

Proof

Prediction Perspective

$\min_{\hat{Y}} E[(Y-\hat{Y})^2]$

$\hat{Y}^* = \arg\min_{\hat{Y}} E[(Y-\hat{Y})^2]$

$= E(Y|X)$

MMSE

Minimum Mean Squared Error.

- Predict or estimate the future observations or unknown parameters based on data
- $E(Y|X)$ is our **best predictor** of Y based on data 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^o. $\hat{Y}$: predictor of Y, a function of X. (info/data)

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

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

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

2^o. Let $h(X) = E(Y|X) - g(X)$ is a function of X.

By projecton theorem, we know

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

Proof

3^0. Thus

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

$$= E[(Y-E(Y|X))^2] + E[(E(Y|X)-g(X))^2] \geq 0 \quad \forall g(X)$$

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

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

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

Proof

Perspective of Statistical Learning

In general, $E(Y|X)$ is a nonlinear function of X, which is hard to compute. At least half contents of statistical learning courses tell you how to approximately compute $E(Y|X)$.

- Linear regression: using linear function to approximate
- Logistic regression: using log-linear function to approximate
- Polynomial regression: using polynomial function to approximate
- Regression splines: using spline function to approximate
- Deep learning for prediction: using neural networks to approximate

Outline

Basic Estimation Problem

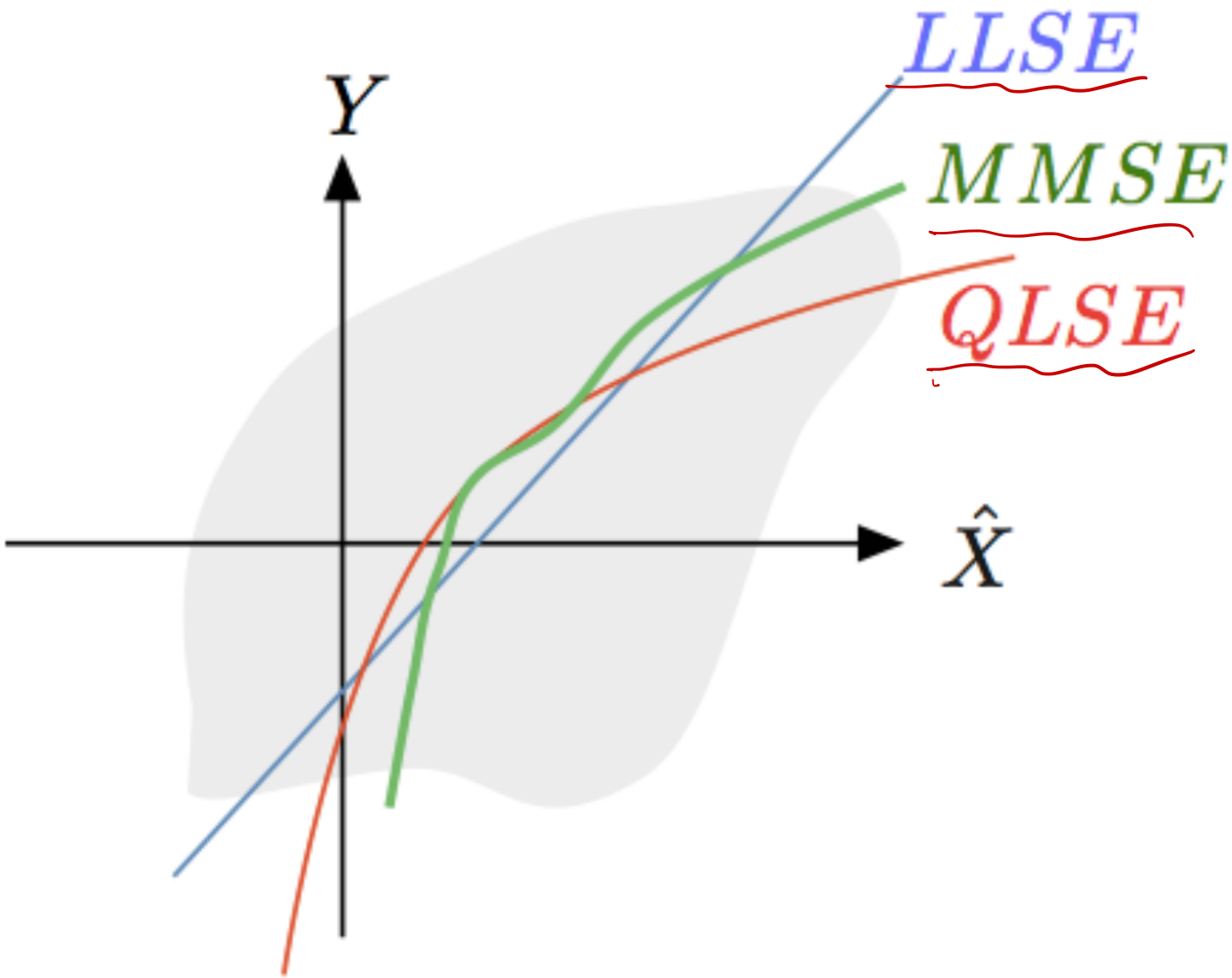
- Estimate X from the observed value Y
- Choose the estimator (inference function) $g(\cdot)$ to minimize the expected error $E(c(X, g(Y)))$
- $c(X, \hat{X})$ is the cost of guessing $\hat{X}$ when the actually value is X.

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

Basic Estimation Problem

- When $c(X, \hat{X}) = ||X - \hat{X}||^2$, the problem is called "the least square estimate (LSE)" problem.
- Further, if the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a + bY$, it is called "the Linear Least Square Estimate (LLSE)" estimate of X given Y.
- Further, if the function $g(\cdot)$ is restricted to be quadratic, i.e., of the form $a + bY^2$, it is called "the Quadratic Least Square Estimate (QLSE)" estimate of X given Y.
- Further, if the function $g(\cdot)$ can be arbitrary, it is called "the Minimum Mean Square Estimate (MMSE)" estimate of X given Y.

Illustration



Linear Least Square Estimate

Theorem

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

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

Proof

1°. cost function $g(a,b) = E[(X-a-bY)^2]$

$= a^2 - 2aE(X) + 2abE(Y) + b^2E(Y^2) - 2bE(XY) + E(X^2)$

To minimize $g(a,b)$. $\frac{\partial g(a,b)}{\partial a} = 0$, $\frac{\partial g(a,b)}{\partial b} = 0$

2°

$$\left.\frac{\partial g(a,b)}{\partial a}\right|_{a^*} = 2a^* - 2E(X) + 2b^*E(Y) = 0$$

$$\left.\frac{\partial g(a,b)}{\partial b}\right|_{b^*} = 2a^*E(Y) + 2b^*E(Y^2) - 2E(XY) = 0$$

$\Rightarrow$ $b^* = \frac{Cov(X,Y)}{Var(Y)}$, $a^* = E(X) - \frac{Cov(X,Y)}{Var(Y)} \cdot E(Y)$

3°. $L[X|Y] = a^* + b^*Y = E(X) - \frac{Cov(X,Y)}{Var(Y)} \cdot E(Y) + \frac{Cov(X,Y)}{Var(Y)} \cdot Y$

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

Proof

4°. $\min_{a,b} g(a,b)$. $\frac{\partial g(a,b)}{\partial a}=0$, $\frac{\partial g(a,b)}{\partial b}=0$

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

positive semidefinite.

$\forall z^T=[c,d]$, $z^T H z \geq 0$

$$H=\begin{bmatrix} 2 & 2E[Y] \\ 2E[Y] & 2E[Y^2] \end{bmatrix}=2\begin{bmatrix} 1 & E[Y] \\ E[Y] & E[Y^2] \end{bmatrix}$$

$$\Rightarrow z^T H z = 2[c,d]\begin{bmatrix} 1 & E[Y] \\ E[Y] & E[Y^2] \end{bmatrix}\begin{bmatrix} c \\ d \end{bmatrix}$$

$$= 2\left[(c+dE(Y))^2 + d^2 \mathrm{Var}(Y)\right]$$

$$\geq 0$$

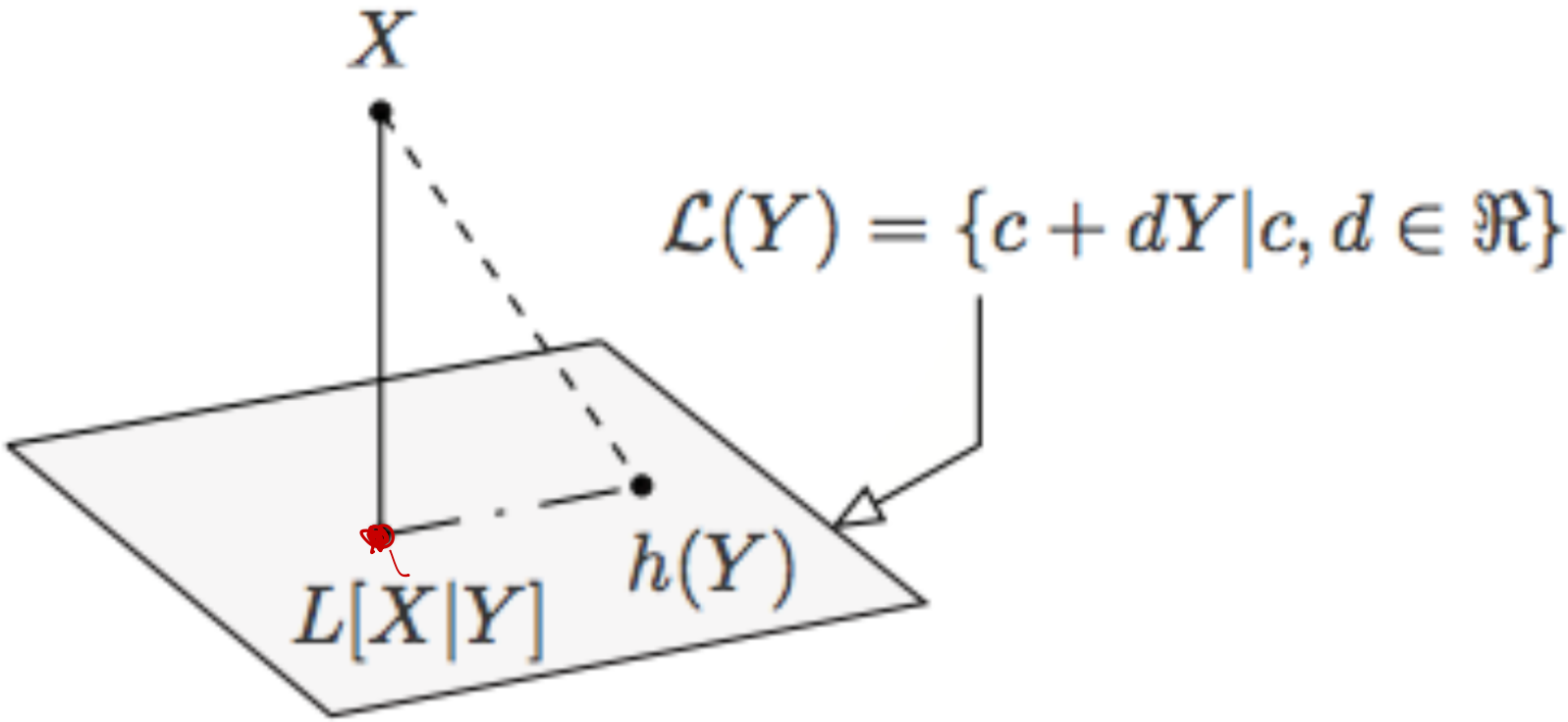
Proof

Projection Perspective

- Random variables X and Y are called "orthogonal" if $E(XY) = 0$, denoted by $X \perp Y$.
- $E(X - E(X)) = 0$: $X - E(X) \perp c$ for any constant c.
- $E[(X - L(X|Y))g(Y)] = 0$: $X - L(X|Y) \perp g(Y)$ for any linear functions of Y.

Projection Perspective

$L[X|Y]$ is the projection of X onto the set $\mathcal{L}(Y)$ of linear functions of Y: $X - L[X|Y]$ is orthogonal to every linear function of Y.



Linear Regression

Suppose we observe K i.i.d. samples $(X_1, Y_1), \ldots, (X_K, Y_K)$ of X, Y. Find a function $g(Y) = a + bY$ such that $E[(X - a - bY)^2]$ is minimized.

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

Monte Carlo

Sample average $\longrightarrow$ Expectation

Linear Regression

$$L^k(X|Y) = E_k(X) + \frac{Cov_k(X,Y)}{var_k(Y)}(Y - E_k(Y))$$

1^0. $E_k(X) = \frac{1}{k}\sum_{j=1}^{k} x_j$

2^0. $E_k(y) = \frac{1}{k}\sum_{j=1}^{k} y_j$

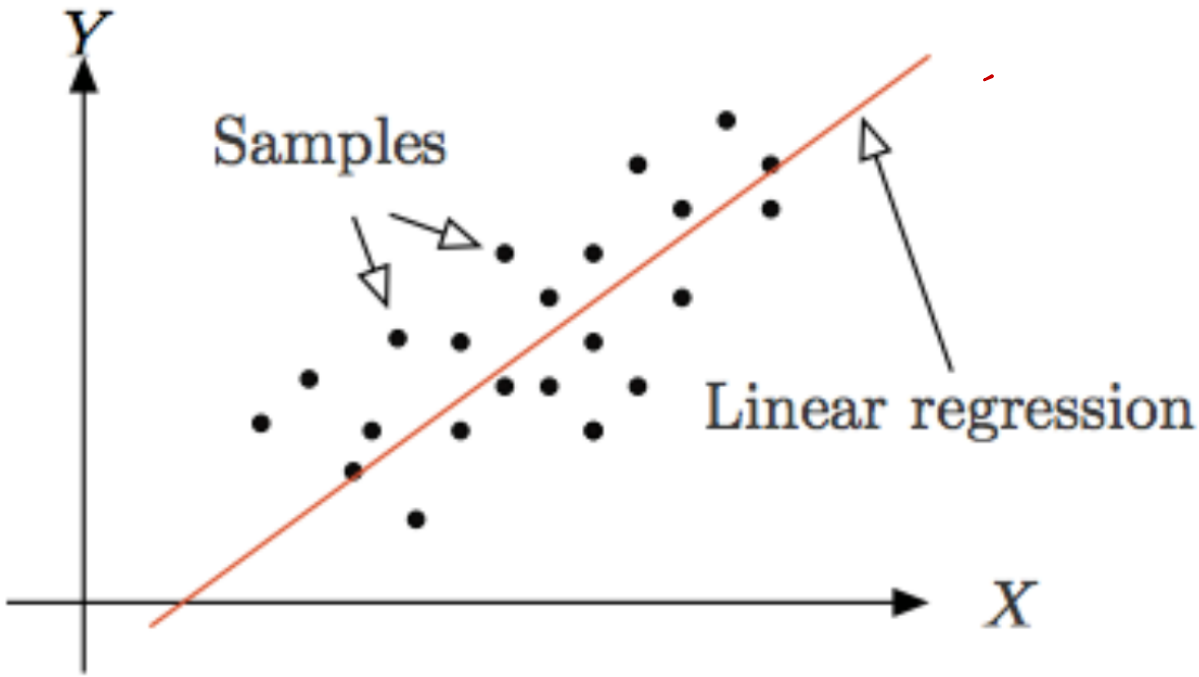
3^0. $Cov_k(X,Y) = E_k(XY) - E_k(X)\cdot E_k(Y)$

$$= \frac{1}{k}\sum_{j=1}^{k} x_j y_j - E_k(X)E_k(Y)$$

4^0. $Var_k(Y) = E_k(Y^2) - (E_k(Y))^2$

$$= \frac{1}{k}\sum_{j=1}^{k} y_j^2 - (E_k(Y))^2$$

Linear Regression



Linear Regression

As $k \to \infty$, by SLLN.

$E_k(X) \to E(X)$

$E_k(Y) \to E(Y)$

$\mathrm{Cov}_k(X,Y) \to \mathrm{Cov}(X,Y)$

$\mathrm{Var}_k(Y) \to \mathrm{Var}(Y)$ w.p.1

$\Rightarrow L^k(X|Y) \to L(X|Y)$

Theorem

As the number of samples increases, the linear regression approaches the LLSE.

New Perspectives of Covariance

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

1°. if $Cov(X,Y) = 0$, $L[X|Y]$ does not depend on Y

2°. if $Cov(X,Y) > 0$ $L[X|Y]$ ↑ with Y $(Y - E(Y))$

< 0 ↓ Y. $(Y - E(Y))$

$Cov(X,Y)$ measures a form of dependency in terms of LR.

Minimum Mean Squared Error Estimator

Theorem

The MMSE of X given Y is given by

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

Orthogonality Property of MMSE

Theorem

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

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

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

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

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

Proof

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

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

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

$= E[X\phi(Y)] - E[E[X\phi(Y)|Y]]$ Adam's Law

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

$= 0$

$\forall \phi(\cdot)$

Proof

(b) $E\left[\{(g(Y)-E(X|Y)\}^2\right]$

$= E\left[\left(g(Y)-E(X|Y)\right)\{(g(Y)-X)+(X-E(X|Y))\}\right]$

$= \underbrace{E\left[\left(g(Y)-E(X|Y)\right)\cdot(g(Y)-X)\right]}_{0} + \underbrace{E\left[(g(Y)-E(X|Y))\cdot(X-E(X|Y))\right]}_{0}$

$\phi(Y) = g(Y)-E(X|Y)$ a function of Y.

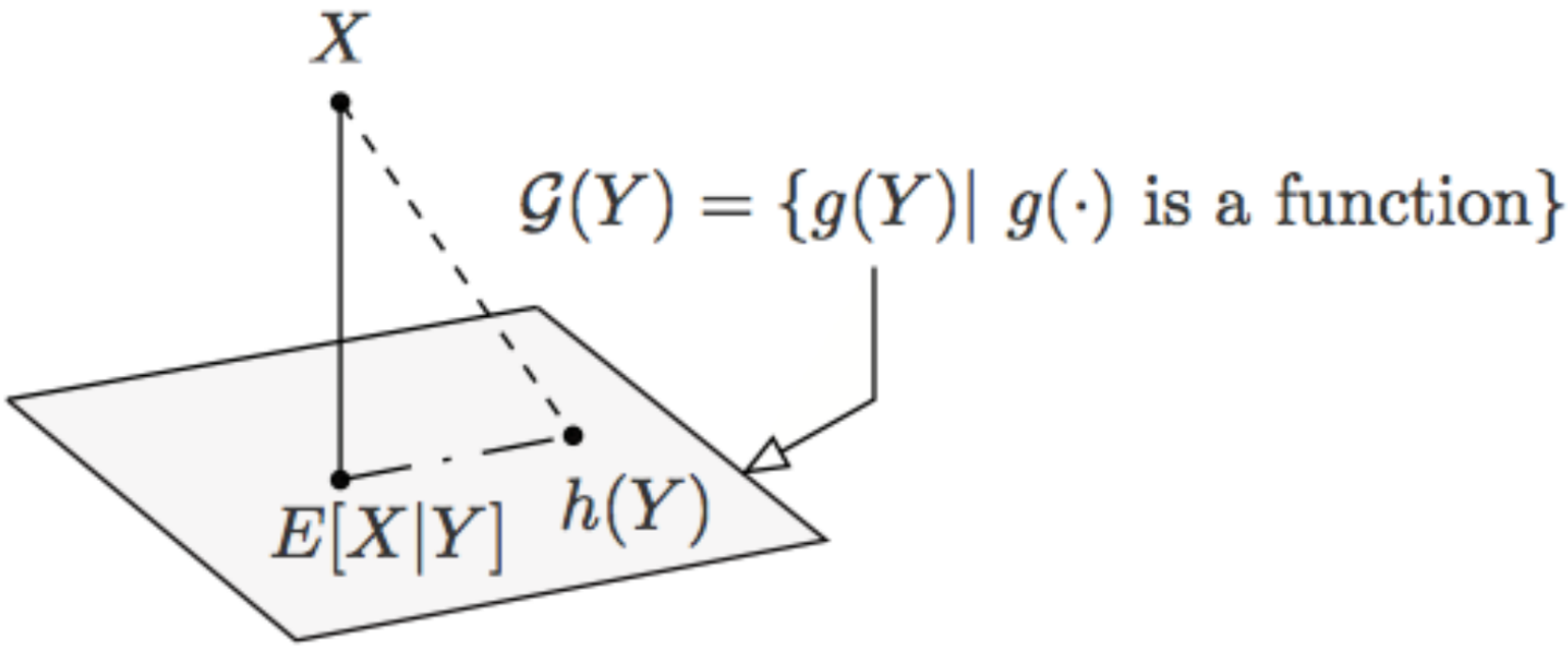
$g(Y)-X \perp \phi(Y)$ by (b).

$X-E(X|Y) \perp \phi(Y)$ by (a).

$= 0$

$\forall g(\cdot) \Rightarrow \hat{g}(Y) = E(X|Y)$

Projection Perspective of MMSE



MMSE and LLSE

- In general, the MMSE estimator is different from the the LLSE: $E[X|Y] \neq L[X|Y]$.
- The ordering of mean squared errors is

$$E[(X - E[X|Y])^2] \leq E[(X - L[X|Y])^2] \leq Var(X)$$

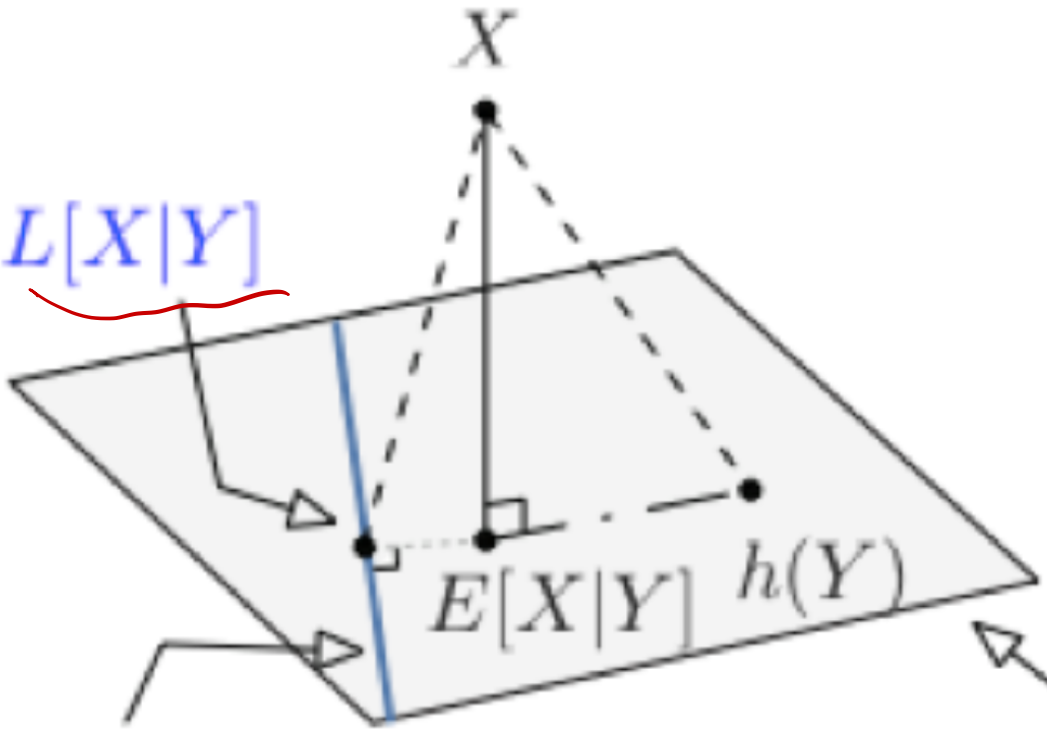
$E[(X-E(X))^2]$

MMSE.

$E(X)$ as estimator.

a special case of linear function of X.

MMSE and LLSE



$$\mathcal{L}(Y) = \{a + bY | a, b \in \Re\}$$

$$\mathcal{G}(Y) = \{g(Y) |\ g(\cdot) \text{ is a function}\}$$

MMSE for Jointly Gaussian Random Variables

Theorem

Let X, Y be jointly Gaussian random variables. Then

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

Example I

Let X, Y be jointly continuous random variables with the pdf

$$f_{X,Y}(x,y) = \begin{cases} x+y & \text{if } 0 \le x, y \le 1 \\ 0 & \textit{otherwise} \end{cases}$$

Find the MMSE $E(X|Y)$ and LLSE $L(X|Y)$.

Solution

1^0. Find MMSE $E[X|Y]$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \begin{cases} \frac{1}{2}+y & 0\le y\le 1 \\ 0 & \text{else} \end{cases}$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{x+y}{\frac{1}{2}+y} & 0\le x\le 1 \\ 0 & \text{else} \end{cases}$$

Thus for $0\le y\le 1$, $E[X|Y=y] = \int_0^1 x f_{X|Y}(x|y)dx$

$$= \int_0^1 \frac{x^2+xy}{\frac{1}{2}+y}dx = \frac{2+3y}{3+6y}$$

$$\Rightarrow E[X|Y] = \frac{2+3Y}{3+6Y}$$

Solution

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

$E(X) = E(Y) = \frac{7}{12}$

$Var(Y) = \frac{11}{144}$, $Cov(X,Y) = -\frac{1}{144}$

$\Rightarrow L[X|Y] = \frac{7}{12} - \frac{1}{11}(Y - \frac{7}{12})$

Example II: Revisit Biased Coin Problem

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 $L(\Theta|X)$.

Solution

MMSE

1°. $\Theta \sim \text{unif}(0,1) = \text{Beta}(1,1)$, prior PDF

# of heads $X|\Theta=\theta \sim \text{Bin}(n,\theta)$

By Beta-binomial conjugacy,

$$\Theta|X=k \sim \text{Beta}(k+1, n-k+1)$$

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

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

MMSE

Linear of X.

$$E[\Theta|X] = L[\Theta|X]$$

Solution

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

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

$X|\theta = \theta \sim Bin(n,\theta)$

$E[X] = E[E(X|\theta)] = E[n\theta] = n \cdot E(\theta) = \frac{n}{2}$

$Var(X) \overset{\text{Eve's Law}}{=} E[Var(X|\theta)] + Var(E(X|\theta))$

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

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

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

$= \frac{n}{12}(n+2)$

Solution

3°. $Cov(X,\Theta) = E[\Theta X] - E(\Theta)\cdot E(X)$

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

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

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

$= n E(\Theta^2) - \frac{n}{4}$

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

4° $L[\Theta|X] = E(\Theta) + \frac{Cov(\Theta,X)}{Var(X)}[X - E(X)]$

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

$= E[\Theta|X]$

Application Case: Kalman Filter

- 1960: Rudoph Emil Kalman (1930-2016) introduced what is known as Kalman filter.
- Milestones in Statistics & Signal Processing

Reasons for Popularity of Kalman Filter

- Good results in practice due to optimality and structure: LLSE estimation in general and 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"

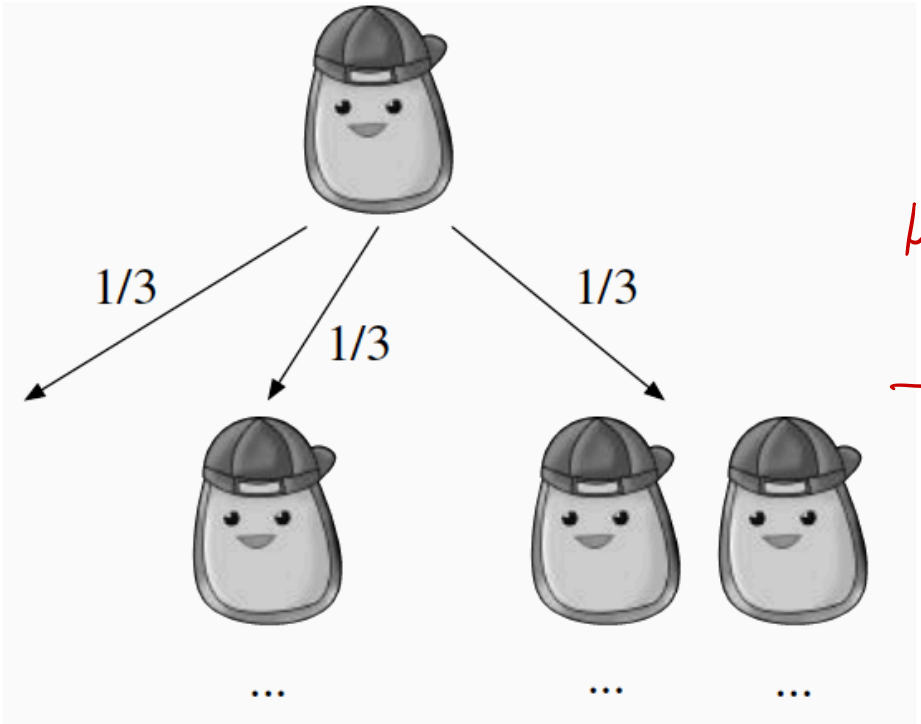
- The process of finding the "best estimate" from noisy data amounts to "filtering out" the noise.
- Estimation (statistical perspective) vs. Filtering (signal processing perspective)
- A Kalman filter not only clean up the data measurements
- A Kalman filter also projects these measurements onto the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

Outline

Revisit the Story of Bobo

A single amoeba, Bobo, lives in a pond. After one minute Bobo will either die, split into two amoebas, or stay the same, with equal probability, and in subsequent minutes all living amoebas will behave the same way, independently. What is the probability that the amoeba population will eventually die out?

First step Analysis.



$e = 1$

$\mu = \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 2$

$= 1$

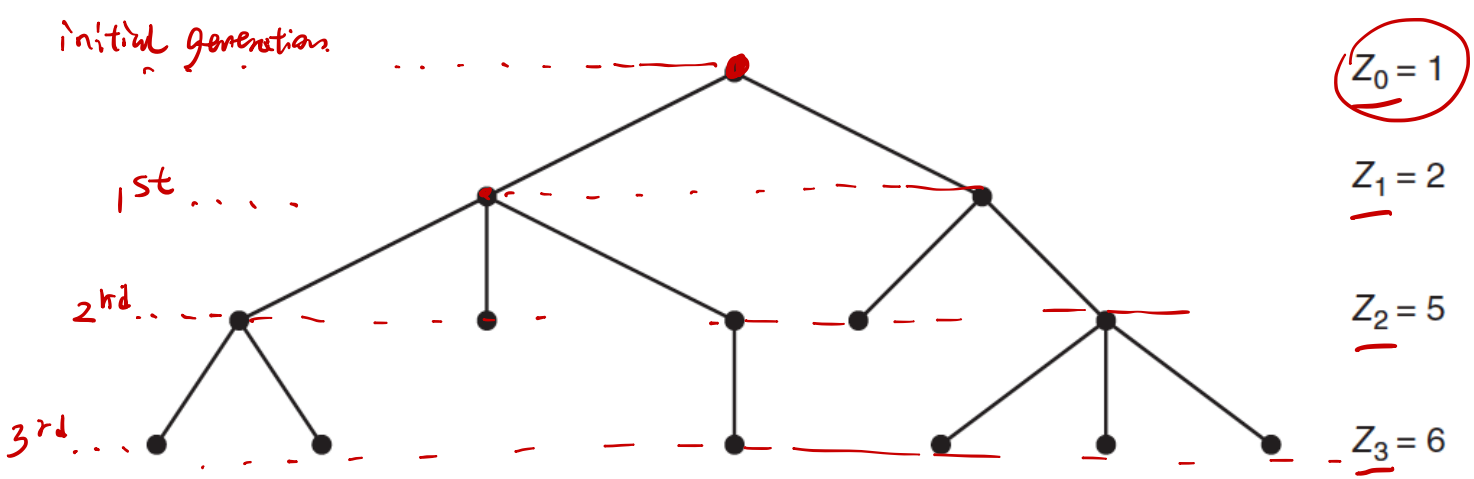
$\mu < 1$

$\mu > 1$

Branching Process

- Very useful model for the growth of populations
- Galton-Watson process: Sir Francois Galton & Reverend Henry William Watson made important contributions.
- Various Applications
  - Extinction of family surnames (original motivation)
  - The spread of infectious diseases and epidemics (biology and epidemiology)
  - The spread of computer software viruses
  - Rumor(information) spread in social networks

Illustration of A Branching Process



$Z_0, Z_1, \ldots$ branching process.

offspring distribution $a = (a_0, a_1, a_2, \ldots)$

a_k: an individual (prob. of) gives birth to k children.

$a_0 = 1 \Rightarrow Z_n = 0,\ n \geq 1.$

usually $0 < a_0 < 1$, $0 < a_0 + a_1 < 1$, ≥ 2 children w.p. > 0.

Setting

- Offspring distribution $(a) = (a_0, a_1, a_2, \ldots)$: distribution of each individual independently produces a random number of children
- An individual gives birth to k children with probability $a_k, k \geq 0$, independently of other individuals.
- Z_n: the size (number of individuals) of the n^{th} generation, $n \geq 0$.
- $Z_0, Z_1, \ldots$ is a branching process.

Mean Generation Size

① X_i : the i^{th} person's children number in the $(n-1)^{th}$ generation

$X_1, \cdots$ i.i.d. a.

$E(X_i) = \mu$, X_i and Z_{n-1} are independent

Theorem

Let $\mu = \sum_{k=0}^{\infty} k \cdot a_k$. *Suppose* $Z_0 = 1$, *then the mean of* n^{th} *generation size is*

$$E(Z_n) = \mu^n, n \geq 0.$$

② $Z_n = \sum_{i=1}^{Z_{n-1}} X_i \Rightarrow E[Z_n | Z_{n-1}=k] = E[\sum_{i=1}^{Z_{n-1}} X_i | Z_{n-1}=k]$

$= E[\sum_{i=1}^{k} X_i | Z_{n-1}=k] = E[\sum_{i=1}^{k} X_i] = k E(X_i) = k.\mu \Rightarrow E(Z_n|Z_{n-1}) = k Z_{n-1}$

$\Rightarrow E[Z_n] = E[E[Z_n|Z_{n-1}]] = E[Z_{n-1}\mu] = \mu \cdot E[Z_n] \Rightarrow E(Z_n) = \mu^n, n \geq 0$

Proof

Three Cases

$$E(Z_n) = \mu^n$$

$$\lim_{n\to\infty} E(Z_n) = \lim_{n\to\infty} \mu^n = \begin{cases} 0 & \text{if } \mu<1 \quad \text{Subcritical.} \\ 1 & \text{if } \mu=1 \quad \text{Critical} \\ \infty & \text{if } \mu>1 \quad \text{Supercritical.} \end{cases}$$

Three Cases

TABLE 4.1 Simulations of a Branching Process for Three Choices of μ

μ	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
0.75	1	2	3	3	2	1	1	0	0	0	0
0.75	1	0	0	0	0	0	0	0	0	0	0
0.75	1	2	0	0	0	0	0	0	0	0	0
0.75	1	0	0	0	0	0	0	0	0	0	0
0.75	1	3	3	1	3	1	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0
1	1	2	0	0	0	0	0	0	0	0	0
1	3	6	6	5	6	7	8	8	8	6	5
1	1	3	4	1	2	1	0	0	0	0	0
1	1	2	1	1	2	1	2	1	0	0	0
1.5	1	2	3	10	22	41	93	173	375	763	1,597
1.5	1	1	1	1	2	4	7	9	11	19	29
1.5	1	4	5	18	34	68	127	246	521	1,011	2,065
1.5	1	1	2	0	0	0	0	0	0	0	0
1.5	1	2	5	3	2	6	9	17	18	13	19

Extinction in the Subcritical Case

Theorem

With probability 1, a subcritical branching process eventually goes extinct.

Variance of Generation Size

Theorem

Let σ^2 denote the variance of the offspring distribution. Then we have

$$Var(Z_n) = \begin{cases} n\sigma^2 & \text{if } \mu = 1 \\ \sigma^2 \mu^{n-1} \left(\frac{\mu^n - 1}{\mu - 1} \right) & \text{if } \mu \neq 1 \end{cases}$$

Proof

$Z_n = \sum_{i=1}^{Z_{n-1}} X_i$, $\mathrm{Var}(Z_n) = E[\mathrm{Var}(Z_n|Z_{n-1})] + \mathrm{Var}(E(Z_n|Z_{n-1}))$ $n \geq 1$

1°. $E[Z_n | Z_{n-1} = k] = E(\sum_{i=1}^{k} X_i) = k\mu \Rightarrow E[Z_n | Z_{n-1}) = \mu \cdot Z_{n-1}$

2°. $\mathrm{var}(Z_n | Z_{n-1} = k) = \mathrm{var}(\sum_{i=1}^{k} X_i) = k\sigma^2 \Rightarrow \mathrm{var}(Z_n | Z_{n-1}) = \sigma^2 Z_{n-1}$

$\Rightarrow \mathrm{Var}(Z_n) = E(\mathrm{Var}(Z_n|Z_{n-1})) + \mathrm{var}(E(Z_n|Z_{n-1}))$

$= E(\sigma^2 Z_{n-1}) + \mathrm{Var}(\mu Z_{n-1})$

$= \sigma^2 E(Z_{n-1}) + \mu^2 \mathrm{var}(Z_{n-1}) = \sigma^2 \cdot \mu^{n-1} + \mu^2 \mathrm{var}(Z_{n-1})$ $n \geq 1$

(a) $\mu = 1 \Rightarrow \mathrm{var}(Z_n) = \mathrm{var}(Z_{n-1}) + \sigma^2$ $n \geq 1$ $\Rightarrow \mathrm{var}(Z_n) = n\sigma^2$, $n \geq 1$

(b) $\mu \neq 1$ $\Rightarrow \mathrm{Var}(Z_1) = \sigma^2$, $\mathrm{var}(Z_2) = \sigma^2 \mu(\mu+1)$

$\mathrm{Var}(Z_3) = \sigma^2\mu^2(1+\mu+\mu^2)$ $\cdots$ $\mathrm{var}(Z_n) = \sigma^2 \mu^{n-1} (\sum_{k=0}^{n-1} \mu^k)$

$= \sigma^2 \cdot \mu^{n-1} \cdot \frac{\mu^n - 1}{\mu - 1}$ $n \geq 1$

Proof

$$E(Z_n) = \mu^n$$

$$\mathrm{Var}(Z_n) = \begin{cases} n\sigma^2 & \mu = 1 \\ \sigma^2 \mu^{n-1} \dfrac{\mu^n - 1}{\mu - 1} & \mu \neq 1 \end{cases}$$

$\mu < 1$ Subcritical. $E(Z_n) \to 0$ $\mathrm{Var}(Z_n) \to 0$

$\mu = 1$ Critical. $E(Z_n) = 1$, $\mathrm{Var}(Z_n) = n\sigma^2 \to \infty$

$\mu > 1$ Supercritical $E(Z_n) > 1$, $\mathrm{Var}(Z_n) \to \infty$

Z

Probability Generation Function (PGF)

Theorem

The PGF for Z_n is the n-fold composition of the PGF for offspring distribution.

Proof

1°. Z_n. PGF of Z_n : $G_n(s) = E[s^{Z_n}] = \sum_{k=0}^{\infty} s^k P(Z_n = k)$

On the other hand, Let $G(s) = \sum_{k=0}^{\infty} s^k a_k = E(s^X)$ be the PGF of the offspring distribution

$$\Rightarrow G_n(s) = E(s^{Z_n}) = E\left(s^{\sum_{j=1}^{Z_{n-1}} X_j}\right) = E\left[E\left[s^{\sum_{j=1}^{Z_{n-1}} X_j} \middle| Z_{n-1}\right]\right]$$

Independence of X_j and Z_{n-1}

$$E\left(s^{\sum_{j=1}^{Z_{n-1}} X_j} \middle| Z_{n-1} = z\right) = E\left(s^{\sum_{j=1}^{z} X_j} \middle| Z_{n-1} = z\right) = E\left(s^{\sum_{j=1}^{z} X_j}\right)$$

$$= E\left(\prod_{j=1}^{z} s^{X_j}\right) = \prod_{j=1}^{z} E(s^{X_j}) = [G(s)]^z \quad \forall z, (X_1, \ldots X_j \sim \text{i.i.d.})$$

$$\Rightarrow E\left(s^{\sum_{j=1}^{Z_{n-1}} X_j} \middle| Z_{n-1}\right) = [G(s)]^{Z_{n-1}}$$

Tahly expectation. $G_n(s) = E[s^{Z_n}] = E[(G(s))^{Z_{n-1}}] = G_{n-1}(G(s))$

$$\Rightarrow G_n(s) = G_{n-1}(G(s)) \quad n \geq 1$$

$G_0(s) = E[s^{Z_0}] = s$, $G_1(s) = G_0(G(s)) = G(s)$

$G_2(s) = G_1(G(s)) = G[G(s)] = G(G_1(s))$

Proof

$$G_n(s) = \underbrace{G[\cdots G[G(s)]\cdots]}_{n\text{-fold}} = G[G_{n-1}(s)]$$

Extinction Probability

$e = \lim_{n\to\infty} e_n = \lim_{n\to\infty} P(Z_n = 0)$

$e_n = P(Z_n = 0) = G_n(0) = G(G_{n-1}(0)) = G(e_{n-1})$

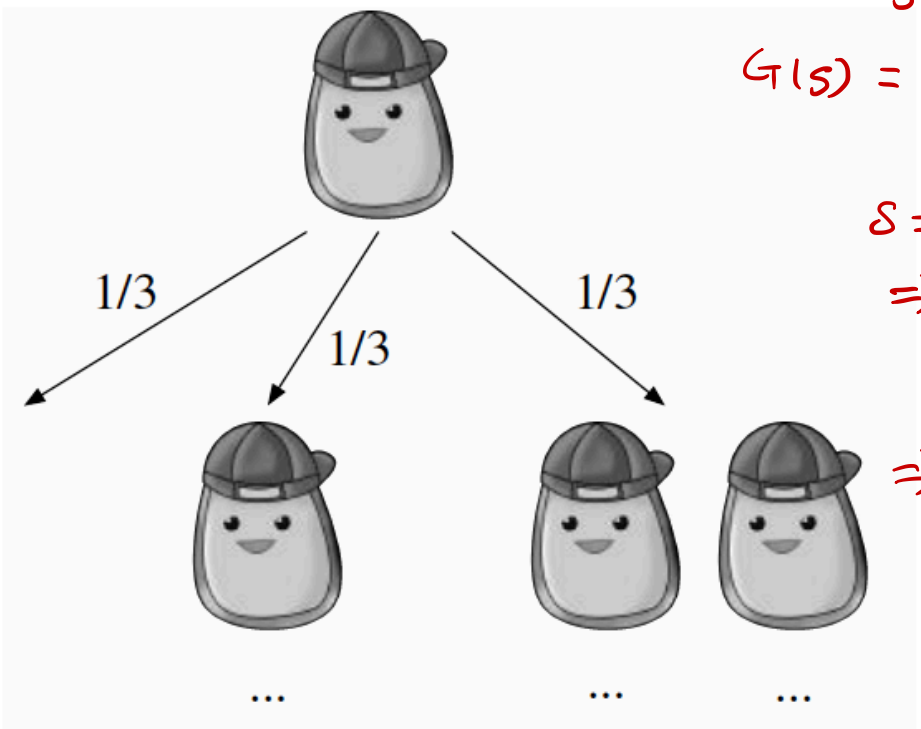
$\lim_{n\to\infty}$, $\Rightarrow e = G(e)$ fixed point.

Theorem

Given a branching process, let G be the probability generating function of the offspring distribution. Then, the probability of eventual extinction is the smallest positive root of the equation $s = G(s)$. Further, If $\mu \leq 1$, that is, in the subcritical and critical cases, the extinction probability is equal to 1.

Example: Story of Bobo

A single amoeba, Bobo, lives in a pond. After one minute Bobo will either die, split into two amoebas, or stay the same, with equal probability, and in subsequent minutes all living amoebas will behave the same way, independently. What is the probability that the amoeba population will eventually die out?



$a = (0, 1, 2)$

$\frac{1}{3}\ \frac{1}{3}\ \frac{1}{3}$

$G(s) = \sum_{k=0}^{\infty} s^k a_k = \frac{1}{3} + \frac{1}{3}s + \frac{1}{3}s^2$

$s = G(s)$

$\Rightarrow s^2 - 2s + 1 = 0$

$\Rightarrow s = 1$

$\Rightarrow e = 1$

Example: First-Success Distribution of Offspring

1^0. $G(s) = \sum_{k=0}^{\infty} s^k a_k = \sum_{k=0}^{\infty} s^k (1-p)^k p = \frac{p}{1-s(1-p)}$ $|s(1-p)|<1$.

$\mu = G'(1) = \frac{1-p}{p}$. supercritical $\Rightarrow \mu > 1 \Rightarrow p < \frac{1}{2}$.

$G(s) = a_0 + \sum_{k=1}^{\infty} s^k a_k$

$G'(s) = \sum_{k=1}^{\infty} k s^{k-1} a_k \Rightarrow G'(1) = \sum_{k=1}^{\infty} k a_k = \sum_{k=0}^{\infty} k a_k = \mu$

A branching process has offspring distribution $a_k = (1-p)^k p$, for $k = 0, 1, \ldots$. Find the extinction probability in the supercritical case.

2^0. solve the equation $s = G(s) = \frac{p}{1-s(1-p)}$

$\Rightarrow (1-p)s^2 - s + p = 0$

$\Rightarrow s_1 = 1, s_2 = \frac{p}{1-p}$ since $0<p<1$

$\Rightarrow s_2 < s_1$

$\Rightarrow e = s_2 = \frac{p}{1-p}$

Simulation Results

$p=\frac{1}{4}$, $\mu=\frac{1-p}{p}=3$, $e=\frac{p}{1-p}=\frac{1}{3}$.

$a_k=(\frac{3}{4})^k\cdot\frac{1}{4}$.

TABLE 4.2 Simulation of a Supercritical Branching Process, with $\mu = 3$. Four of the 12 runs go extinct by the 10th generation.

Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
1	4	25	97	394	1160	3475	10685	31885	95757	287130
1	1	1	3	7	11	47	165	515	1525	4689
1	10	37	115	350	1124	3455	10073	29896	88863	267386
1	1	1	2	2	3	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
1	8	31	71	248	779	2282	6864	19895	59196	178171
1	0	0	0	0	0	0	0	0	0	0
1	4	7	13	34	106	380	1123	3385	10200	30090
1	16	49	163	447	1284	3794	11592	34626	104390	312704
1	1	1	5	16	51	155	559	1730	5378	15647
1	1	3	31	79	267	883	2637	8043	23970	71841
1	0	0	0	0	0	0	0	0	0	0

Outline

Poisson Process

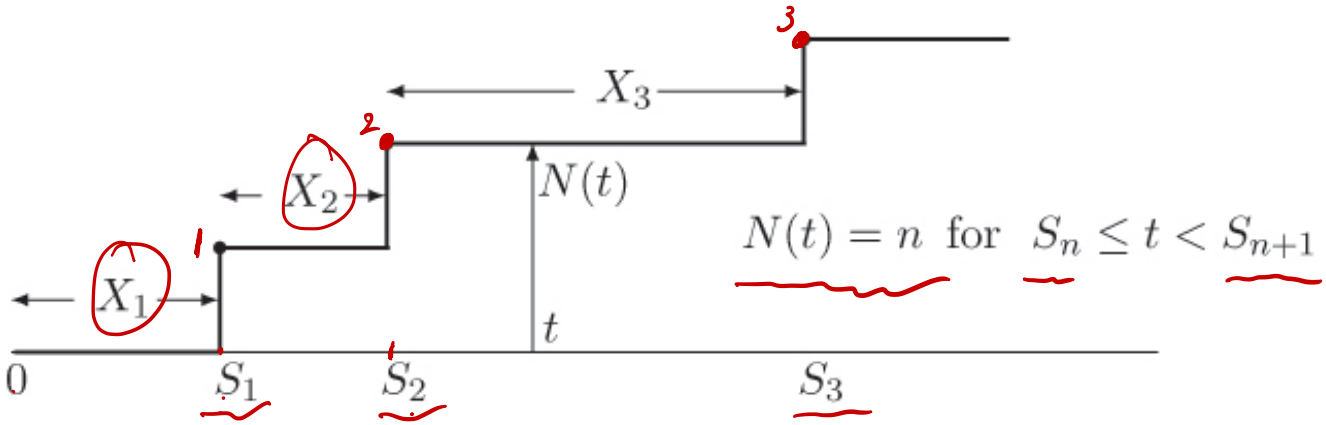


Figure : Illustration of a Poisson process and its arrival times $\{S_1, S_2, \ldots, \}$, its interarrival times $\{X_1, X_2, \ldots, \}$, and its counting process $N(t) : t \geq 0$.

Various Perspectives

The Poisson process can be described in three ways:

- The interarrival times $X_1, X_2, X_3, \cdots$
- The arrival times $S_0 = 0, S_1, S_2, S_3, \cdots$
- The arrival counts $N(t)$ in $[0, t], t \geq 0$

Important Relationships

$$S_n = \sum_{j=1}^{n} X_j, n \geq 1$$

$$X_n = S_n - S_{n-1}, n \geq 1$$

$$N(t) = \max\{n : S_n \leq t\}$$

$$N(t) \geq n \Leftrightarrow S_n \leq t$$

Equivalent Definitions of Poisson Process

Poisson Process - Definition 1

Definition

A *Poisson process* with parameter λ is a counting process $(N_t)_{t\geq 0}$ with the following properties:

1. $N_0 = 0$.
2. For all $t > 0$, N_t has a Poisson distribution with parameter λt.
3. (Stationary increments) For all $s, t > 0$, $N_{t+s} - N_s$ has the same distribution as N_t. That is,

$$P(N_{t+s} - N_s = k) = P(N_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \text{ for } k = 0, 1, \ldots$$

4. (Independent increments) For $0 \leq q < r \leq s < t$, $N_t - N_s$ and $N_r - N_q$ are independent random variables.

Translated Poisson Process

Let $(N_t)_{t\geq 0}$ be a Poisson process with parameter λ. For $s > 0$, let $\tilde{N}_t = N_{t+s} - N_s$, for $t \geq 0$. Then we have

- $(\tilde{N}_t)_{t\geq 0}$ is called "Translated Poisson Process".
- $(\tilde{N}_t)_{t\geq 0}$ is a Poisson process with parameter λ

Poisson Process - Definition 2

Definition

Let $X_1, X_2, \ldots$ be a sequence of *i.i.d.* exponential random variables with parameter λ. For $t > 0$, let

$$N_t = \max\{n : X_1 + \cdots + X_n \leq t\},$$

with $N_0 = 0$. Then, $(N_t)_{t \geq 0}$ defines a Poisson process with parameter λ.

Arrival Times & Gamma Distribution

$S_n = X_1 + \cdots + X_n$ $X_1, \ldots X_n$ i.i.d. expo(λ).

For $n = 1, 2, \ldots$, let S_n be the time of the n-th arrival in a Poisson process with parameter λ. Then, S_n has a gamma distribution with parameters n and λ. The density function of S_n is

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}, \text{ for } t > 0.$$

Mean and variance are

$$E(S_n) = \frac{n}{\lambda} \text{ and } Var(S_n) = \frac{n}{\lambda^2}$$

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

$= \frac{1}{\lambda} \cdot n = \frac{n}{\lambda}$.

One New Perspective

CDF $\xrightarrow{\prime}$ PDF

Find $f_{S_n}(t)$

1° $N(t) \geq n \Leftrightarrow S_n \leq t$

$\Rightarrow$ CDF of S_n $\quad F_{S_n}(t) = P(S_n \leq t) = P(N(t) \geq n)$

$$= \sum_{j=n}^{\infty} P(N(t)=j) = \sum_{j=n}^{\infty} \frac{e^{-\lambda t}\cdot(\lambda t)^j}{j!}$$

PDF $f_{S_n}(t) = F'_{S_n}(t) = \left(\sum_{j=n}^{\infty} \frac{e^{-\lambda t}\cdot(\lambda t)^j}{j!}\right)'$

$= \cdots$

$$= \frac{\lambda e^{-\lambda t}\cdot(\lambda t)^{n-1}}{(n-1)!}$$

Another New Perspective

$$f_{S_n}(t)dt = P(t<S_n<t+dt) = P(N(t)=n-1,\ 1 \text{ arrival in } (t,t+dt))$$

Independent increment:
$$= P(N(t)=n-1)\cdot P\{\text{One arrival in } (t,t+dt)\}$$

Stationary increment:
$$= \frac{e^{-\lambda t}\cdot(\lambda t)^{n-1}}{(n-1)!}\cdot P\{N(dt)=1\}$$

$$= \frac{e^{-\lambda t}\cdot(\lambda t)^{n-1}}{(n-1)!}\cdot e^{-\lambda dt}\cdot\lambda dt$$

$$\Rightarrow f_{S_n}(t) = \lim_{dt\to 0}\frac{P(t<S_n<t+dt)}{dt} = \lim_{dt\to 0}\frac{e^{-\lambda t}\cdot(\lambda t)^{n-1}}{(n-1)!}\cdot\lambda e^{-\lambda dt}$$

$$= \frac{\lambda e^{-\lambda t}\cdot(\lambda t)^{n-1}}{(n-1)!}$$

Arrival Times & Uniform Distribution

Let S_1, S_2, ..., be the arrival times of a Poisson process with parameter λ. Conditional on $N_t = n$, the joint distribution of $(S_1, \ldots, S_n)$ is the distribution of the order statistics of n *i.i.d.* uniform random variables on $[0, t]$. That is, the joint density function of $S_1, \ldots, S_n$ is

$$f(s_1, \ldots, s_n | N_t = n) = \frac{n!}{t^n}, \text{ for } 0 < s_1 < \cdots < s_n < t.$$

Equivalently, let $U_1, \ldots, U_n$ be an *i.i.d.* sequence of random variables uniformly distributed on $[0, t]$. Then, conditional on $N_t = n$,

$$(S_1, \ldots, S_n) \text{ and } (U_{(1)}, \ldots, U_{(n)})$$

have the same distribution.

Example

1°. $N(t)$: # of people arriving by the time t.

$N(t) \sim pois(\lambda t)$

2°. total waiting time of people who arrive before the band starts.

$\sum_{k=1}^{N(t)} (t - S_k)$.

Concert-goers arrive at a show according to a Poisson process with parameter λ. The band starts playing at time t. The k-th person to arrive in $[0, t]$ waits $t - S_k$ time units for the start of the concert, where S_k is the k-th arrival time. Find the expected total waiting time of concert-goers who arrive before the band starts.

3°. We focus. $E\left[\sum_{k=1}^{N(t)} (t - S_k)\right]$

Solution 4.0

$$E\left(\sum_{k=1}^{N(t)}(t-S_k)\right) = E\left[E\left[\sum_{k=1}^{N(t)}(t-S_k)\right]\middle|N(t)\right]$$

Consider $E\left[\sum_{k=1}^{N(t)}(t-S_k)\middle|N(t)=n\right] = E\left[\sum_{k=1}^{n}(t-S_k)\middle|N(t)=n\right]$

Given $N(t)=n$, $(S_1,\dots S_n) \sim (U_{(1)},\dots,U_{(n)})$

$$\Rightarrow E\left(\sum_{k=1}^{n}(t-S_k)\middle|N(t)=n\right) = E\left(\sum_{k=1}^{n}(t-U_{(k)})\middle|N(t)=n\right)$$

$$= E\left[\left(tn-\sum_{k=1}^{n}U_{(k)}\right)\middle|N(t)=n\right] \qquad \left(\sum_{k=1}^{n}U_{(k)}=\sum_{k=1}^{n}U_k\right)$$

$U_i,\dots$ i.i.d. $\sim f(0,t)$

$$= E\left[\left(tn-\sum_{k=1}^{n}U_k\right)\middle|N(t)=n\right]$$

$$= E\left[tn-\sum_{k=1}^{n}U_k\right] = tn-\sum_{k=1}^{n}E(U_k) = tn-n\cdot\frac{t}{2} = \frac{1}{2}tn$$

Solution

We have $E\left(\sum_{k=1}^{N(t)}(t-S_k)\,\middle|\,N(t)=n\right)=\frac{1}{2}t\cdot n \qquad \forall n$

$$\Rightarrow E\left(\sum_{k=1}^{N(t)}(t-S_k)\,\middle|\,N(t)\right)=\frac{1}{2}tN(t).$$

$$\Rightarrow E\left(\sum_{k=1}^{N(t)}(t-S_k)\right)=E\left[E\left[\sum_{k=1}^{N(t)}(t-S_k)\,\middle|\,N(t)\right]\right]$$

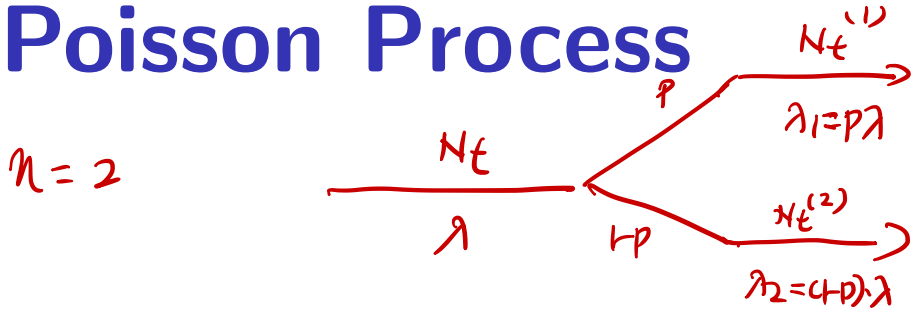
$$=E\left[\frac{1}{2}tN(t)\right]$$

$$=\frac{1}{2}tE[N(t)]$$

$$=\frac{1}{2}t\cdot\lambda t$$

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

Thinned Poisson Process

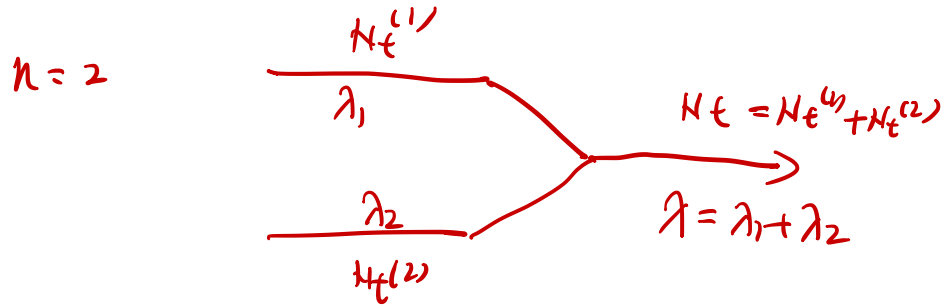


Let $(N_t)_{t\geq 0}$ be a Poisson process with parameter λ. Assume that each arrival, independent of other arrivals, is marked as a type k event with probability p_k, for $k = 1, \ldots, n$, where $p_1 + \cdots + p_n = 1$. Let $N_t^{(k)}$ be the number of type k events in $[0, t]$. Then, $\left(N_t^{(k)}\right)_{t\geq 0}$ is a Poisson process with parameter with λp_k. Furthermore, the processes

$$\left(N_t^{(1)}\right)_{t\geq 0}, \ldots, \left(N_t^{(n)}\right)_{t\geq 0}$$

are independent. Each process is called a *thinned Poisson process*.

Superposition Process



Assume that $\left(N_t^{(1)}\right)_{t\geq 0}, \ldots, \left(N_t^{(n)}\right)_{t\geq 0}$ are n independent Poisson processes with respective parameters $\lambda_1, \ldots, \lambda_n$. Let $N_t = N_t^{(1)} + \cdots + N_t^{(n)}$, for $t \geq 0$. Then, $(N_t)_{t\geq 0}$ is a Poisson process with parameter $\lambda = \lambda_1 + \cdots + \lambda_n$.

Revisited Coupon Collector Problem

Embedding Trick. (Poisson process $N(t)$, $\lambda = 1$)

Discrete — Continuous.

- Previously we introduce the coupon collector problem: n toy types, collected one by one, sampling with replacement from the set of toy types each time, and all toy types were equally likely to be collected.
- Now suppose that at each stage, the j^{th} toy type is collected with probability p_j, where the $p_j, j = 1, \ldots, n$ are not necessarily equal.
- Let N be the number of toys needed until we have a full set.
- Find $E(N)$.

Solution

(a). Suppose that the toy arrive accordy to a poisson process. with rate $\lambda=1$.

interarrival times between toys are i.i.d. $X_j \sim \text{expo}(1)$, $j=1,2,\dots,n$

Let Y_j : the waitty time until the first toy of type j

[By thinning the poisson process $\Longrightarrow$ n independent poisson process.

$[0,t]$ # of arrives type-j toys. $N_j(t) \sim \text{Pois}(\lambda p_j) = \text{Pois}(p_j)$]

$\Rightarrow Y_j \sim \text{expo}(p_j)$

(b). The waitty time T until all toy types are collected as

$$T = \max(Y_1, \dots, Y_n)$$

N : # of toys collected, $T = X_1 + X_2 + \dots + X_N$ ($\text{Pois}(1)$)

(arrival time of N^{th} toys)

$$\Rightarrow E(T) = E(X_1+\dots+X_N) = E[E(X_1+\dots+X_N|N)]$$

$$E[X_1+\dots+X_N|N=n] = E(X_1+\dots+X_n|N=n] = nE[X_1|N=n] = nE[X_1] = n\cdot 1 = n$$

$$\Rightarrow E[X_1+\dots+X_N|N] = N$$

Solution

$\Rightarrow E(T) = E[E(X_1+\cdots+X_N \mid N)] = E(N)$

Continuous $\Leftarrow$ bridge $\begin{matrix}\text{pois}(\lambda)\\ \text{expo}(\lambda)\end{matrix} \longrightarrow$ discrete.

(c). $E(N) = E(T) \overset{T\geq 0}{=} \int_0^\infty P(T>t)\,dt$

$P(T>t) = 1 - P(T\leq t) = 1 - P(\max(Y_1,\cdots Y_n)\leq t)$

$= 1 - P(Y_1\leq t, \cdots Y_n \leq t) = 1 - P(Y_1\leq t)\cdots P(Y_n\leq t)$ $\quad Y_j \sim \exp(p_j)$

$= 1 - \prod_{j=1}^{n}(1-e^{-p_j t})$

$\Rightarrow E(T) = \int_0^\infty \left[1 - \prod_{j=1}^{n}(1-e^{-p_j t})\right] dt$

$\Rightarrow E(N) = \int_0^\infty \left[1 - \prod_{j=1}^{n}(1-e^{-p_j t})\right] dt$

Solution

Birthday Match Problem: Online Version

① People enter the room according to a Poisson process N_t $(t \geq 0)$ with rate $\lambda = 1$

each person is independently marked with one of 365 days. w.p. $\frac{1}{365}$.

$\Rightarrow$ 365 thinned Poisson processes, one for each birthday.

Assume a random person's birthday is uniformly distributed on the 365 days of the year. People enter the room one by one. How many people are in the room the first time that two people share the same birthday? Let K be the desired number. Find $E(K)$

$N_t^{(1)}$: # of arriving people with birthday at Jan. 1. with rate $\frac{1}{365}$. $(0,t)$

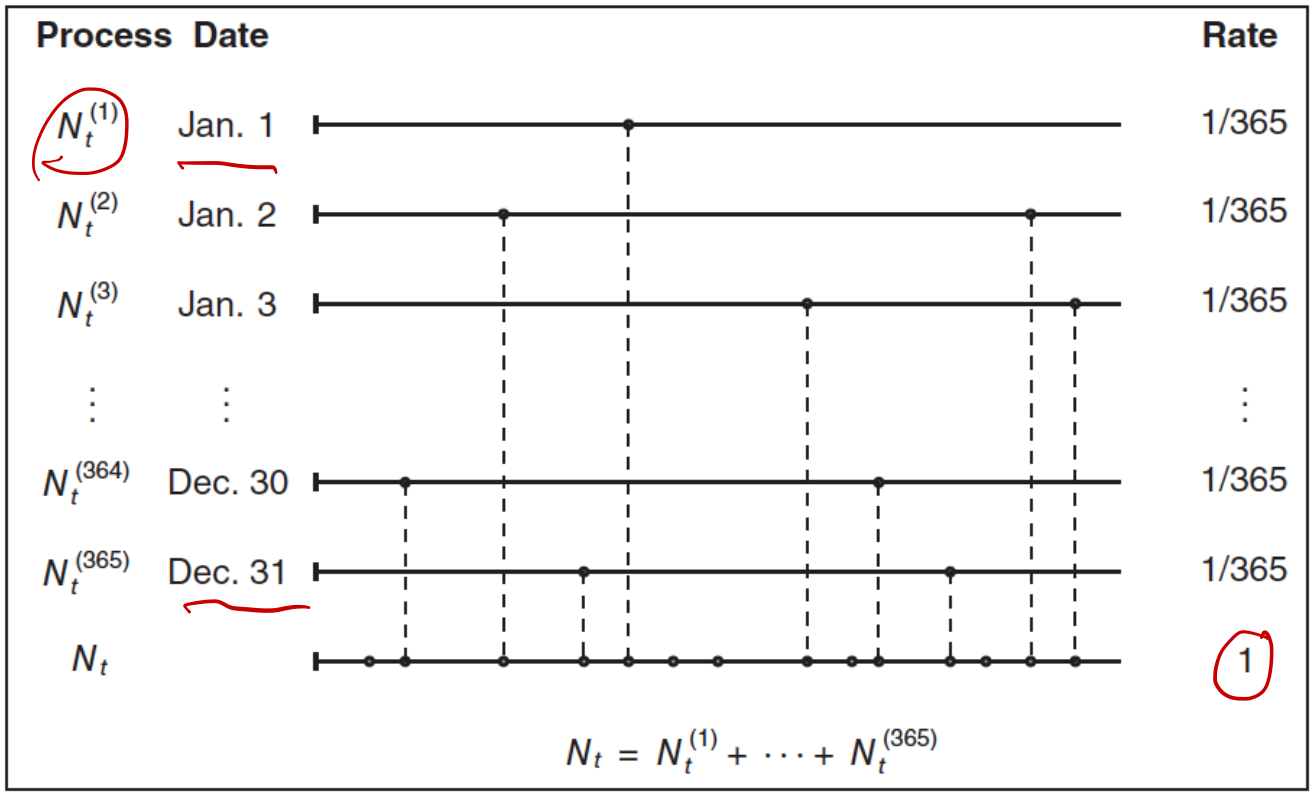
$N_t^{(365)}$: ..... Dec 31 ..... $\frac{1}{365}$.

Solution

② $X_1, \cdots X_n$: interarrival time of people entering the room.

$X_i \sim$ i.i.d. expo(1).

Let T be the first time that two people share the same birthday.



$$T = \sum_{i=1}^{k} X_i = S_k \quad (k \text{ is the desired number})$$

Solution

$T = \sum_{i=1}^{K} X_i = S_K$ $\quad X_i$ independent of K.

$\Rightarrow E(T) = E\left(\sum_{i=1}^{K} X_i\right) = E\left[E\left(\sum_{i=1}^{K} X_i \mid K\right)\right]$

$E\left(\sum_{i=1}^{K} X_i \mid K=n\right) = E\left(\sum_{i=1}^{n} X_i \mid K=n\right) = E\left(\sum_{i=1}^{n} X_i\right) = nE(X_i) = n.$

$\Rightarrow E\left(\sum_{i=1}^{K} X_i \mid K\right) = K.$

$\Rightarrow E(T) = E[K]$

③. $N_t(\lambda)$: $S_n \sim \text{Gamma}(n, \lambda)$.

for each thinned poisson process $N_t^{(j)}$, $j = 1, 2, \ldots 365$.

Let Z_j : the first time when the second person marked with birthday j enter the room.

$N_{Z_j}^{(j)} = 2$

$Z_j = S_2^{(j)} \sim \text{Gamma}(2, \frac{1}{365})$, $f_{Z_j}(t) = \frac{te^{-\frac{t}{365}}}{365^2}$ $t > 0$

PDF

Solution

CDF of Z_j

$$P(Z_j \le t) = \int_0^t \frac{s e^{-\frac{s}{365}}}{365^2} ds = 1 - e^{-\frac{t}{365}}\left(1+\frac{t}{365}\right)$$

④ $T = \min(Z_1, \dots, Z_{365})$

$$P(T>t) = P(\min(Z_1, \dots Z_{365}) > t) = P(Z_1 > t, \dots Z_{365} > t)$$

$$= P(Z_1 > t) \dots P(Z_{365} > t) = [P(Z_1 > t)]^{365} = \left(1+\frac{t}{365}\right)^{365} e^{-t}, \quad t>0$$

$$\Rightarrow E(T) = \int_0^\infty P(T>t)\,dt = \int_0^\infty \left(1+\frac{t}{365}\right)^{365} e^{-t} dt$$

⑤ $E(K) = E(T) = \int_0^\infty \left(1+\frac{t}{365}\right)^{365} e^{-t} dt = 24.617$

Thinned Poisson Process.

Discrete. $\xleftarrow[\text{bridge.}]{\text{Poisson (1)}}$ Continuous

X $\qquad$ Y.

$E(X) = E(Y) \cdot 1$

Solution

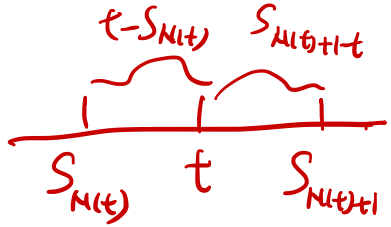
Feller's Classical Problem

E(inter-arrival time of two Buses) = 10 minutes.

E(..... include time t) ≈ 20 minutes.

Buses arrive at a bus stop according to a Poisson process. The time between buses, on average, is 10 minutes. Lisa gets to the bus stop at time t. How long can she expect to wait for a bus?

Length-biased Sampling



Theorem

Suppose buses arrive at a bus stop according to a Poisson process N_t with parameter λ. Given a fixed $t > 0$. The time of the last bus before t is S_{N_t}, and the time of the next bus after t is S_{N_t+1}. Then we have

$$E(S_{N_t+1} - S_{N_t}) = \frac{2 - e^{-\lambda t}}{\lambda}$$

$t \to \infty$.

$\approx \frac{2}{\lambda} = 2\left(\frac{1}{\lambda}\right)$.

Proof

$N_t \triangleq N(t)$.

1°. Compute $E(S_{N_t+1})$

$$E(S_{N(t)+1} \mid N(t)=k) = E(S_{k+1} \mid N(t)=k)$$

($(k+1)^{th}$ arrival time after time t independent of $N(t)$)

$$= E(S_{k+1}) = \frac{k+1}{\lambda}.$$

$$\Rightarrow E(S_{N(t)+1} \mid N(t)) = \frac{N(t)+1}{\lambda}.$$

Taking Expectation on both sides,

$$E(S_{N(t)+1}) = E\left(\frac{N(t)+1}{\lambda}\right) = \frac{E(N(t))+1}{\lambda} = \frac{\lambda t+1}{\lambda} = t+\frac{1}{\lambda}.$$

$$\Rightarrow E(S_{N(t)+1} - t) = \frac{1}{\lambda}.$$ (memoryless properties of exponential distribution)

Proof

2^0. Compute $E(S_{N(t)})$

$E(S_{N(t)} \mid N(t)=k) = E(S_k \mid N(t)=k)$

Conditioning on $N(t)=k$: the kth arrival time $S_k \wedge U_{(k)}$ ~ the maximum of k i.i.d. unif$(0,t)$ $U_1, \dots U_k$

$\Rightarrow$ $\forall 0 \le s \le t$, $P(U_{(k)} \le s) = P(\max(U_1, \dots U_k) \le s)$

$= P(U_1 \le s, \dots U_k \le s) = P(U_1 \le s) \dots P(U_k \le s)$

$= [P(U_1 \le s)]^k = (\frac{s}{t})^k$.

$\Rightarrow E(U_{(k)}) = \int_0^\infty P(U_{(k)} > s)\,ds = \int_0^t (1 - \frac{s^k}{t^k})\,ds = \frac{tk}{k+1}$

$\Rightarrow E(S_{N(t)} \mid N(t)=k) = E(S_k \mid N(t)=k) = E(U_{(k)}) = \frac{tk}{k+1}$

$\Rightarrow E(S_{N(t)} \mid N(t)) = \frac{t\,N(t)}{N(t)+1} = t - \frac{t}{N(t)+1}$

Taking expectation on both sides,

$E(S_{N(t)}) = E(t - \frac{t}{N(t)+1})$

Proof

$$\Rightarrow E(S_{N(t)}) = t - t\cdot E\left(\frac{1}{N(t)+1}\right)$$

$$E\left(\frac{1}{N(t)+1}\right) \overset{\text{LOTE}}{=} \sum_{k=0}^{\infty} \frac{1}{k+1} P(N(t)=k) = \sum_{k=0}^{\infty} \frac{1}{k+1} \cdot \frac{e^{-\lambda t}\cdot(\lambda t)^k}{k!}$$

$$= \frac{e^{-\lambda t}}{\lambda t}\left[\sum_{k=1}^{\infty} \frac{(\lambda t)^k}{k!}\right]$$

$$= \frac{e^{-\lambda t}}{\lambda t}\left[\sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} - 1\right]$$

$$= \frac{e^{-\lambda t}}{\lambda t}\left[e^{\lambda t} - 1\right] = \frac{1-e^{-\lambda t}}{\lambda t}$$

$$\Rightarrow \text{we have } E(S_{N(t)}) = t - t\cdot\frac{1-e^{-\lambda t}}{\lambda t} = t - \frac{1}{\lambda} + \frac{e^{-\lambda t}}{\lambda}$$

$$\Rightarrow t - E(S_{N(t)}) = \frac{1-e^{-\lambda t}}{\lambda}.$$

Proof

Thus $E(S_{N(t)+1} - S_{N(t)})$

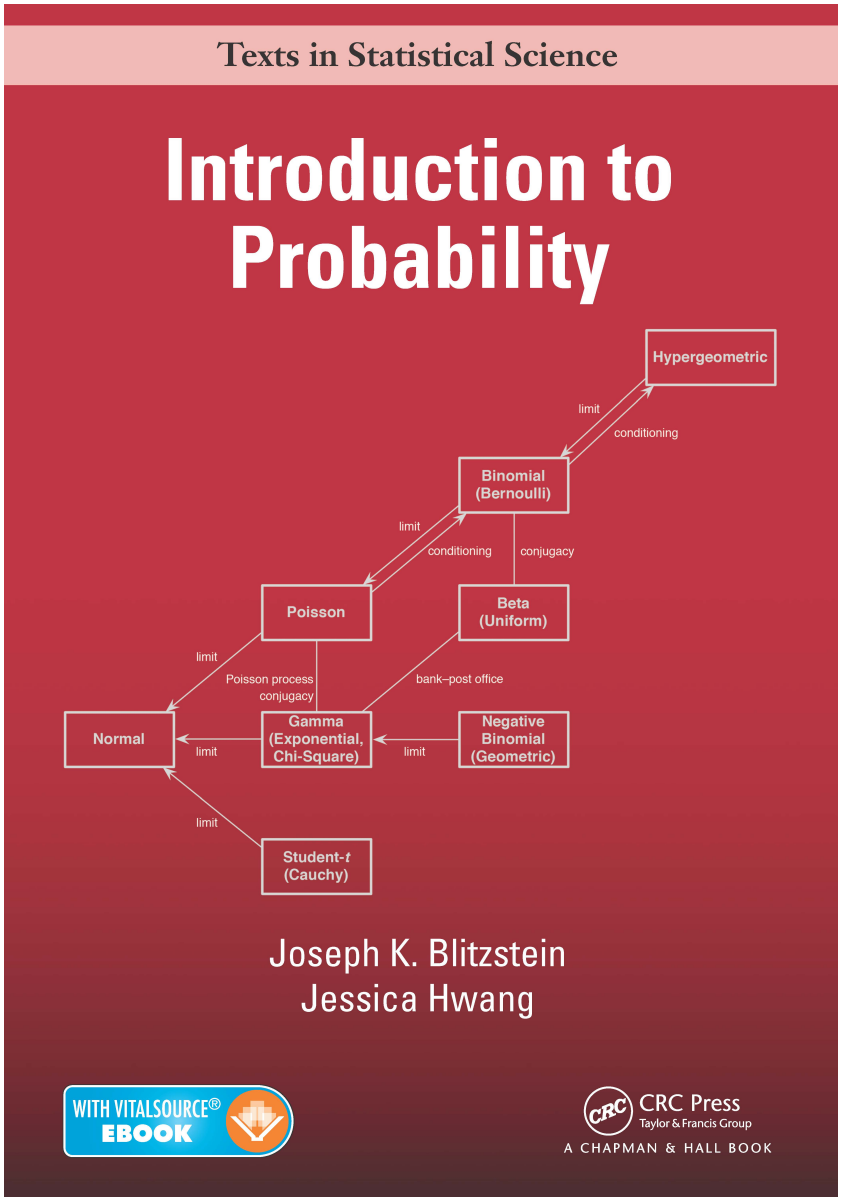
$= (t + \frac{1}{\lambda}) - [t - \frac{1}{\lambda} + \frac{e^{-\lambda t}}{\lambda}]$

$= \frac{2 - e^{-\lambda t}}{\lambda}$ $\lambda t >> 1$

$\approx \frac{2}{\lambda}$

Outline

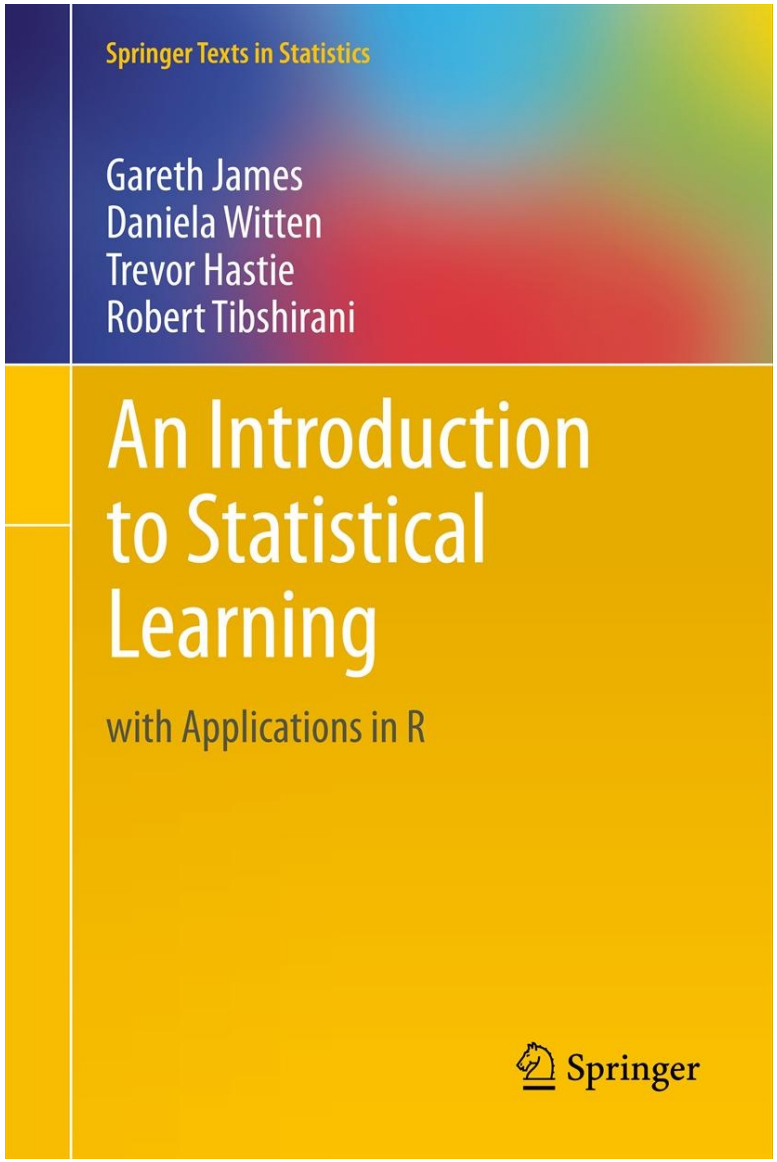
BH



BH

- Introduction to Probability
- Chapman & Hall/CRC, 2014.
- Chapman & Hall/CRC, 2019.
- Chapters 9 & 13

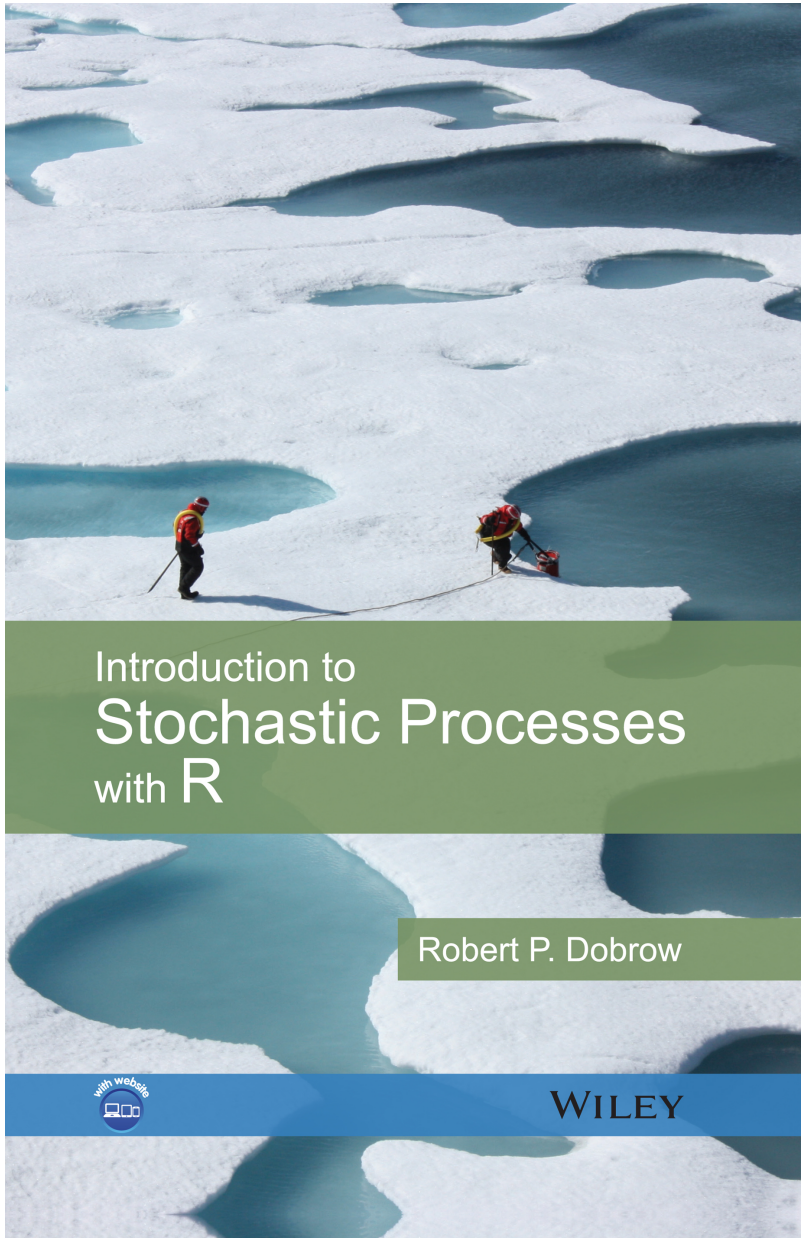
JWHT



JWHT

- An Introduction to Statistical Learning: with Applications in R
- Springer, 2013.

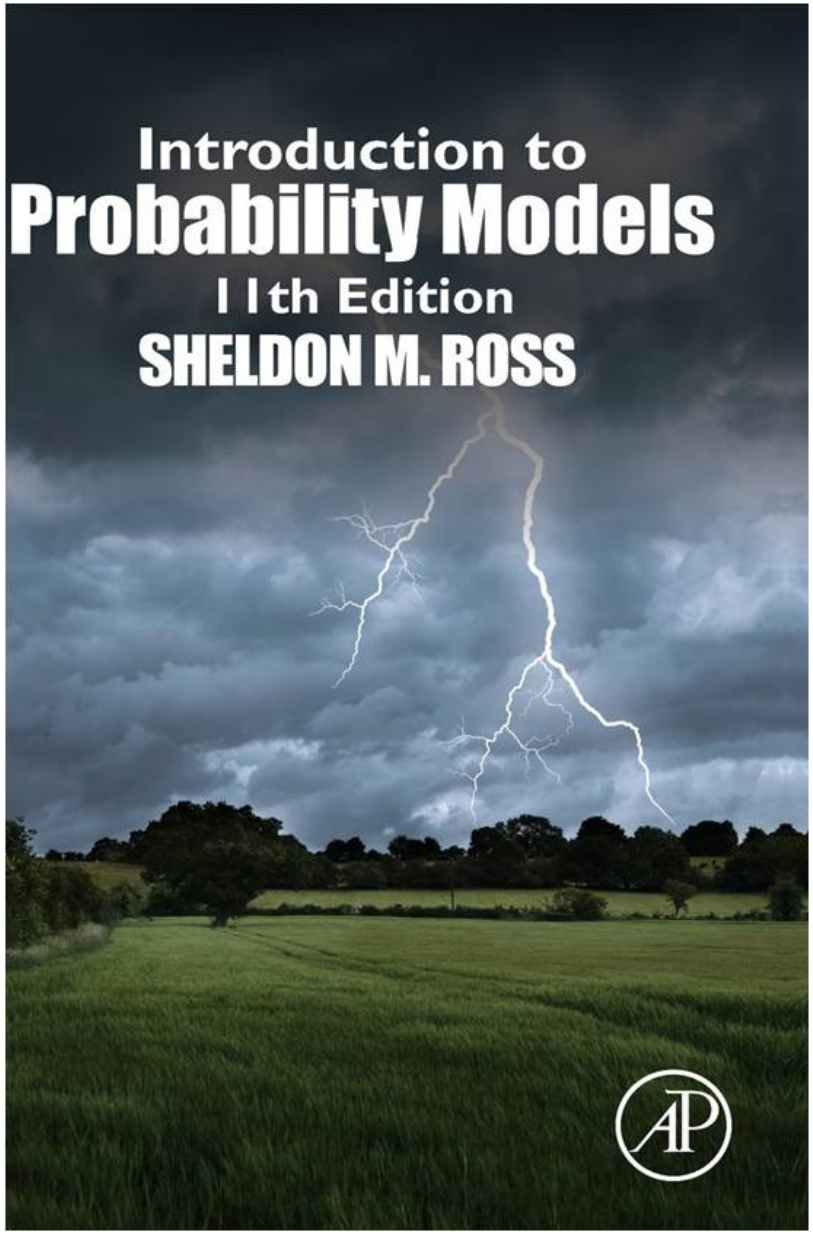
SRR



SPR

- Introduction to Stochastic Processes with R
- John Wiley & Son, 2016.
- Chapters 1 & 4 & 6

SMR



SMR

- Introduction to Probability Models
- Academic Press, 11 edition, 2014.
- Chapters 1 & 2 & 3 & 5