CS70: Jean Walrand: Lecture 23.

Conditional Expectation

1. Review: LR and LLSE
2. Conditional expectation
3. Applications: Diluting, Mixing, Rumors
4. CE = MMSE

Review: LLSE and LR

Definitions Let X and Y be RVs on Ω.

- Covariance: $cov(X,Y) := E[XY] - E[X]E[Y]$
- LLSE: $L[Y|X] = a + bX$ where a, b minimize $E[(Y - a - bX)^2]$.

We saw that

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

Then,

$$E[(Y - L[Y|X])^2] = var(Y) - cov(X,Y)^2/var(X).$$

Non-Bayesian (LR): We are given samples $(X_1, Y_1), \ldots, (X_K, Y_K)$, no distribution.

We define the RVs (X, Y) so that

$$Pr[(X,Y) = (X_k, Y_k)] = 1/K, k = 1, \ldots, K.$$

Then, as before.

Review: LLSE and LR

Consider the non-Bayesian case: sample $(X_1, Y_1), \ldots, (X_K, Y_K)$.

Then,

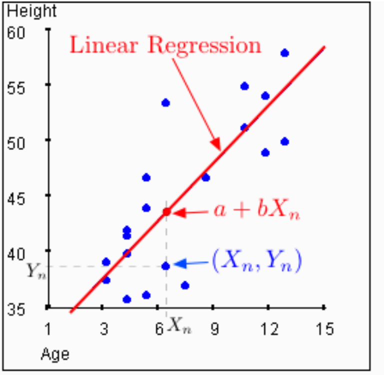
$$L[Y|X] = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X]).$$

Here,

$$\begin{aligned}
E[X] &= \frac{1}{K}\sum_{k=1}^{K} X_k \\
E[Y] &= \frac{1}{K}\sum_{k=1}^{K} Y_k \\
E[X^2] &= \frac{1}{K}\sum_{k=1}^{K} X_k^2 \\
E[XY] &= \frac{1}{K}\sum_{k=1}^{K} X_k Y_k \\
cov(X,Y) &= E[XY] - E[X]E[Y] \\
var(X) &= E[X^2] - E[X]^2.
\end{aligned}$$

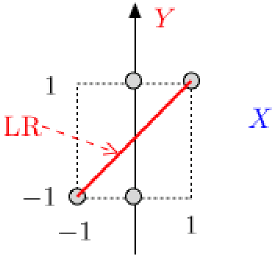
Linear Regression Examples

Example 1:



Linear Regression Examples

Example 2: Four equally likely values of (X, Y), or four samples.

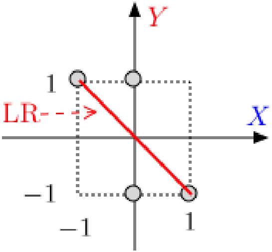


We find:

$$E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = 1/2;$$
$$var[X] = E[X^2] - E[X]^2 = 1/2; cov(X, Y) = E[XY] - E[X]E[Y] = 1/2;$$
$$\text{LR: } \hat{Y} = E[Y] + \frac{cov(X, Y)}{var[X]}(X - E[X]) = X.$$

Linear Regression Examples

Example 3: Four equally likely values of (X, Y), or four samples.

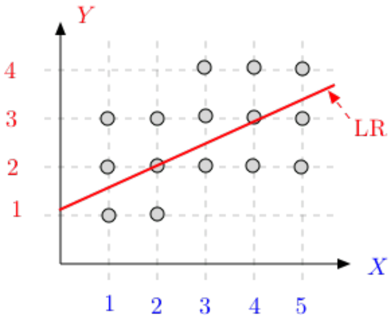


We find:

$E[X] = 0; E[Y] = 0; E[X^2] = 1/2; E[XY] = -1/2;$
$var[X] = E[X^2] - E[X]^2 = 1/2; cov(X, Y) = E[XY] - E[X]E[Y] = -1/2;$
LR: $\hat{Y} = E[Y] + \dfrac{cov(X, Y)}{var[X]}(X - E[X]) = -X.$

Linear Regression Examples

Example 4: Equally likely values of (X, Y), or samples.



We find:

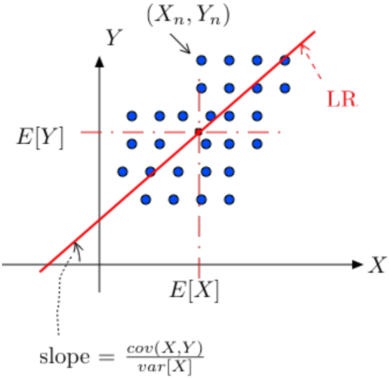
$$E[X] = 3; E[Y] = 2.5; E[X^2] = (3/15)(1 + 2^2 + 3^2 + 4^2 + 5^2) = 11;$$

$$E[XY] = (1/15)(1 \times 1 + 1 \times 2 + \cdots + 5 \times 4) = 8.4;$$

$$var[X] = 11 - 9 = 2; cov(X, Y) = 8.4 - 3 \times 2.5 = 0.9;$$

$$\text{LR: } \hat{Y} = 2.5 + \frac{0.9}{2}(X - 3) = 1.15 + 0.45X.$$

LR: Another Figure



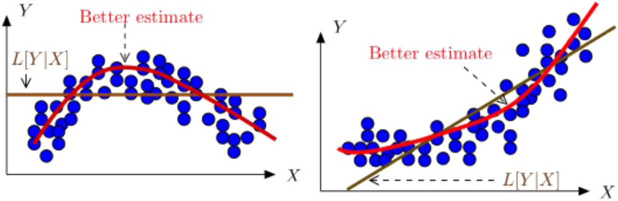
Note that

- the LR line goes through $(E[X], E[Y])$
- its slope is $\frac{cov(X,Y)}{var(X)}$.

Conditional Expectation: Motivation

There are many situations where a good guess about Y given X is not linear.

E.g., (diameter of object, weight), (school years, income), (PSA level, cancer risk).

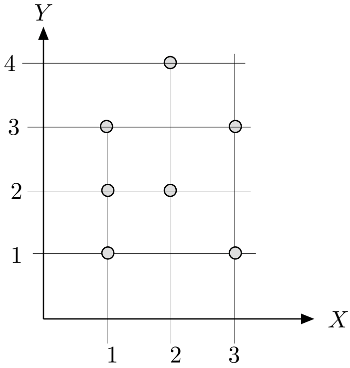


Our goal: Derive the best estimate of Y given X!

That is, find the function $g(\cdot)$ so that $g(X)$ is the best guess about Y given X.

Ambitious! Can it be done? Amazingly, yes!

Conditional Expectation: Intuition



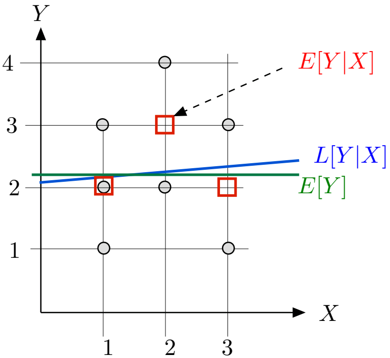
Without any observation, our guess for Y is $E[Y] = 2.3$.

Assume now we observe X. We can calculate
$L[Y|X] = a + bX \approx 2.1 + 0.1x$. .

A better guess when $X = 1$ is 2; when $X = 2$: 3; when $X = 3$: 2.

.

Conditional Expectation: Intuition



Here, $E[Y|X=1]$ is the mean value of Y given that $X=1$. Also, $E[Y|X=2]$ is the mean value of Y given that $X=2$ and $E[Y|X=3]$ is the mean value of Y given that $X=3$.

When we know that $X=1$, Y has a new distribution: Y is uniform in $\{1,2,3\}$.
Thus, our guess is $E[Y|X=1] = 1(1/3)+2(1/3)+3(1/3) = 2$.

Conditional Expectation

Definition Let X and Y be RVs on Ω. The conditional expectation of Y given X is defined as

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

where

$$g(x) := E[Y|X = x] := \sum_y yPr[Y = y|X = x],$$

with $Pr[Y = y|X = x] := \frac{Pr[X=x, Y=y]}{Pr[X=x]}$.

Theorem: $E[Y|X]$ is the best guess about Y given X.

That is, for any function $h(\cdot)$, one has

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

Proof: Later.

Calculating $E[Y|X]$

Let X, Y, Z be i.i.d. with mean 0 and variance 1. We want to calculate

$$E[2 + 5X + 7XY + 11X^2 + 13X^3Z^2|X].$$

We find

$$\begin{aligned}
&E[2 + 5X + 7XY + 11X^2 + 13X^3Z^2|X] \\
&\quad = 2 + 5X + 7XE[Y|X] + 11X^2 + 13X^3E[Z^2|X] \\
&\quad = 2 + 5X + 7XE[Y] + 11X^2 + 13X^3E[Z^2] \\
&\quad = 2 + 5X + 11X^2 + 13X^3(var[Z] + E[Z]^2) \\
&\quad = 2 + 5X + 11X^2 + 13X^3.
\end{aligned}$$

Projection Property

The claim is that

$$E[(Y - E[Y|X])f(X)] = 0, \forall f(.).$$

That is,

$$E[Yf(X)] = E[E[Y|X]f(X)]$$

.

In particular, choosing $f(x) = 1$, we get

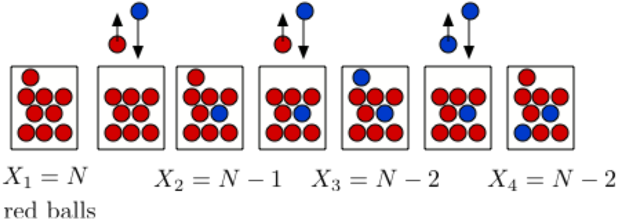
$$E[Y] = E[E[Y|X]].$$

Proof:

$$\begin{aligned}
E[E[Y|X]f(X)] &= \sum_x E[Y|X = x]f(x)Pr[X = x] \\
&= \sum_x [\sum_y yf(x)Pr[Y = y|X = x]]Pr[X = x] \\
&= \sum_x \sum_y yf(x)Pr[X = x, Y = y] \\
&= E[Yf(X)].
\end{aligned}$$

$\square$

Application: Diluting



At each step, pick a ball from a well-mixed urn. Replace it with a blue ball. Let X_n be the number of red balls in the urn at step n. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m-1$ w.p. m/N (if you pick a red ball) and $X_{n+1} = m$ otherwise. Hence,

$$E[X_{n+1}|X_n = m] = m - (m/N) = m(N-1)/N = X_n\rho,$$

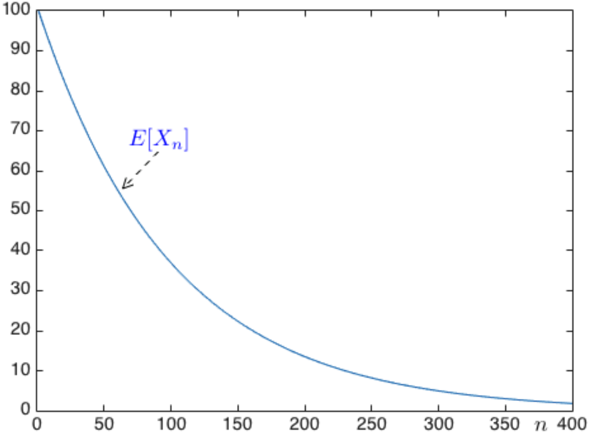
with $\rho := (N-1)/N$. Consequently,

$$E[X_{n+1}] = E[E[X_{n+1}|X_n]] = \rho E[X_n], n \geq 1.$$

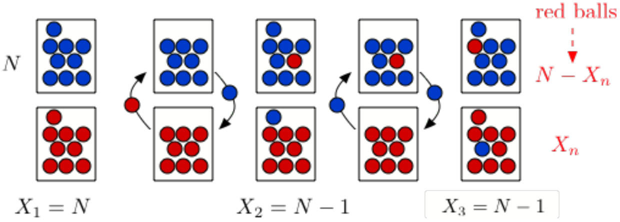
$$\Longrightarrow E[X_n] = \rho^{n-1}E[X_1] = N(\frac{N-1}{N})^{n-1}, n \geq 1.$$

Diluting

Here is a plot:



Application: Mixing



At each step, pick a ball from each well-mixed urn. We transfer them to the other urn. Let X_n be the number of red balls in the bottom urn at step n. What is $E[X_n]$?

Given $X_n = m$, $X_{n+1} = m + 1$ w.p. p and $X_{n+1} = m - 1$ w.p. q

where $p = (1 - m/N)^2$ (B goes up, R down) and $q = (m/N)^2$ (R goes up, B down).

Thus,
$E[X_{n+1}|X_n] = X_n + p - q = X_n + 1 - 2X_n/N = 1 + \rho X_n,\ \rho := (1 - 2/N).$

Mixing

We saw that $E[X_{n+1}|X_n] = 1 + \rho X_n,\ \rho := (1 - 2/N)$. Hence,

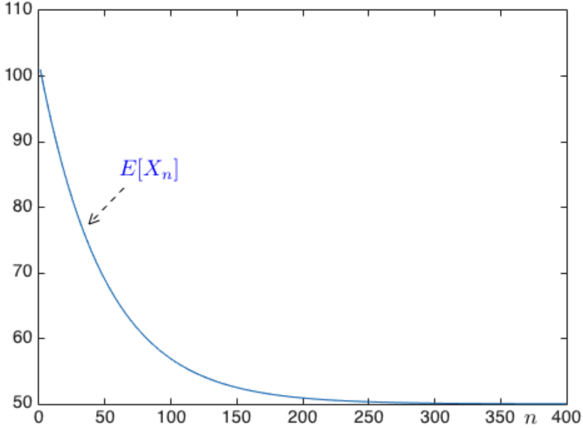
$$\begin{aligned}
&E[X_{n+1}] = 1 + \rho E[X_n] \\
&E[X_2] = 1 + \rho N; E[X_3] = 1 + \rho(1 + \rho N) = 1 + \rho + \rho^2 N \\
&E[X_4] = 1 + \rho(1 + \rho + \rho^2 N) = 1 + \rho + \rho^2 + \rho^3 N \\
&E[X_n] = 1 + \rho + \cdots + \rho^{n-2} + \rho^{n-1} N.
\end{aligned}$$

Hence,

$$E[X_n] = \frac{1 - \rho^{n-1}}{1 - \rho} + \rho^{n-1} N, n \geq 1.$$

Application: Mixing

Here is the plot.



Application: Going Viral

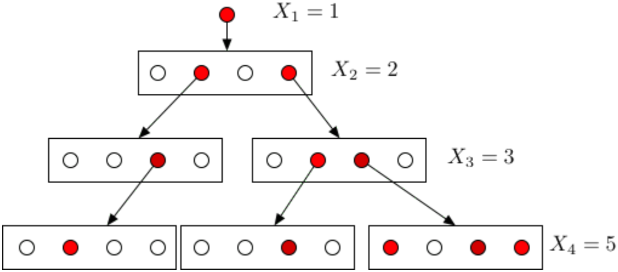
Consider a social network (e.g., Twitter).

You start a rumor (e.g., Walrand is really weird).

You have d friends. Each of your friend retweets w.p. p.

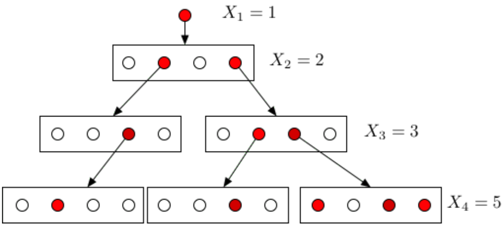
Each of your friends has d friends, etc.

Does the rumor spread? Does it die out (mercifully)?



In this example, $d = 4$.

Application: Going Viral



Fact: Let $X = \sum_{n=1}^{\infty} X_n$. Then, $E[X] < \infty$ iff $pd < 1$.

Proof:

Given $X_n = k$, $X_{n+1} = B(kd, p)$. Hence, $E[X_{n+1}|X_n = k] = kpd$.

Thus, $E[X_{n+1}|X_n] = pdX_n$. Consequently, $E[X_n] = (pd)^{n-1}, n \geq 1$.

If $pd < 1$, then $E[X_1 + \cdots + X_n] \leq (1 - pd)^{-1} \implies E[X] \leq (1 - pd)^{-1}$.

If $pd \geq 1$, then for all C one can find n s.t.

$$E[X] \geq E[X_1 + \cdots + X_n] \geq C. \qquad \square$$

In fact, one can show that $pd \geq 1 \implies Pr[X = \infty] > 0$.

Application: Wald's Identity

Theorem Wald's Identity

Assume that $X_1, X_2, \dots$ and Z are independent, where

Z takes values in $\{0, 1, 2, \dots\}$

and $E[X_n] = \mu$ for all $n \geq 1$.

Then,

$$E[X_1 + \cdots + X_Z] = \mu E[Z].$$

Proof:

$E[X_1 + \cdots + X_Z | Z = k] = \mu k.$

Thus, $E[X_1 + \cdots + X_Z | Z] = \mu Z.$

Hence, $E[X_1 + \cdots + X_Z] = E[\mu Z] = \mu E[Z].$ $\square$

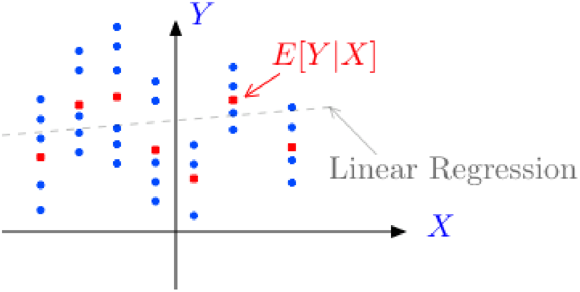
CE = MMSE

Theorem

$E[Y|X]$ is the 'best' guess about Y based on X.

Specifically, it is the function $g(X)$ of X that

$$\text{minimizes } E[(Y - g(X))^2].$$



CE = MMSE

Theorem CE = MMSE

$g(X) := E[Y|X]$ is the function of X that minimizes $E[(Y - g(X))^2]$.
Proof:
Let $h(X)$ be any function of X. Then

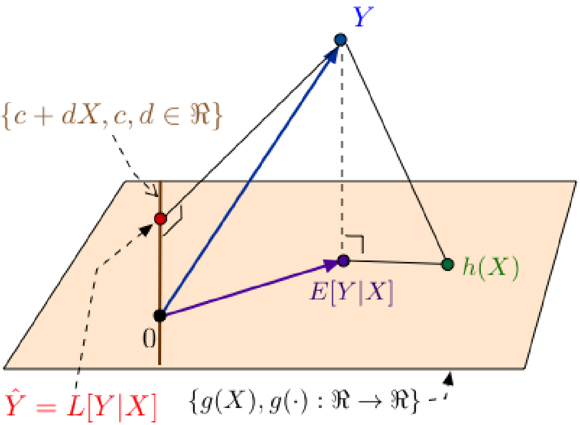
$$\begin{aligned} E[(Y - h(X))^2] &= E[(Y - g(X) + g(X) - h(X))^2] \\ &= E[(Y - g(X))^2] + E[(g(X) - h(X))^2] \\ &\qquad + 2E[(Y - g(X))(g(X) - h(X))]. \end{aligned}$$

But,

$E[(Y - g(X))(g(X) - h(X))] = 0$ by the projection property.

Thus, $E[(Y - h(X))^2] \geq E[(Y - g(X))^2]$. $\square$

$E[Y|X]$ and $L[Y|X]$ as projections



$L[Y|X]$ is the projection of Y on $\{a + bX, a, b \in \Re\}$: LLSE

$E[Y|X]$ is the projection of Y on $\{g(X), g(\cdot) : \Re \to \Re\}$: MMSE.

Summary

Conditional Expectation

- Definition: $E[Y|X] := \sum_y y Pr[Y = y|X = x]$
- Properties: Linearity,
  $Y - E[Y|X] \perp h(X)$; $E[E[Y|X]] = E[Y]$
- Some Applications:
  - Calculating $E[Y|X]$
  - Diluting
  - Mixing
  - Rumors
  - Wald
- MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$