CS70: Lecture 23.

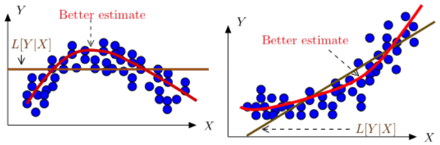
Conditional Expectation

1. Conditional Expectation (CE)
2. Applications: Diluting, Mixing, Wald's Identity
3. CE = MMSE (Minimum Mean Squares Estimate)

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).

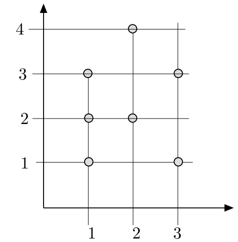


Our goal: Derive the best estimate of Y given X!

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



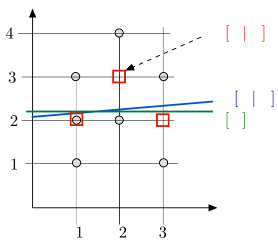
Without any observation, our guess for Y is $E[Y] = 2.3$.

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.

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:

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

□

Additonal Properties of Conditional Expectation

Theorem

(a) Linearity:

$$E[a_1 Y_1 + a_2 Y_2|X] = a_1 E[Y_1|X] + a_2 E[Y_2|X].$$

(b) Factoring Known Values:

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

(c) Smoothing:

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

(d) Independence: If Y and X are independent, then

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

Proof:
Follows easily from the definiton of CE. See Note 20 for a different proof using the projection property. □

Calculating $E[Y|X]$

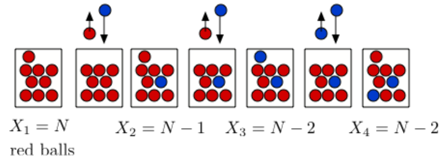
Let X, Y, Z be i.i.d. with mean 0 and variance 1. We want to calculate

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

We find

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

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,

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

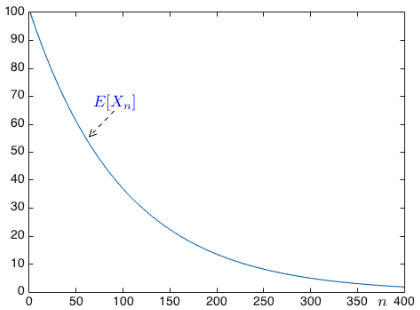
with $\rho := (N-1)/N$. Consequently,

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

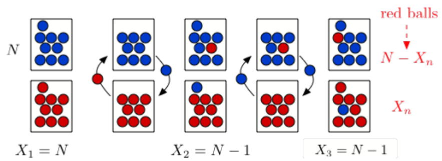
$$\implies E[X_n] = \rho^{n-1}E[X_1] = N(\frac{N-1}{N})^{n-1}, n \geq 1.$$

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,

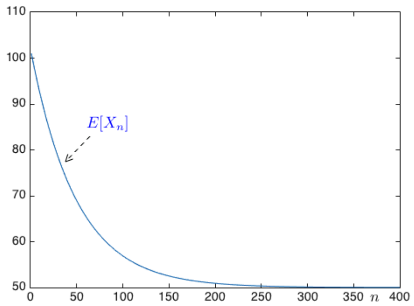
$$\begin{aligned} E[X_{n+1}] &= 1+\rho E[X_n] \\ E[X_2] &= 1+\rho N; E[X_3] = 1+\rho(1+\rho N) = 1+\rho+\rho^2 N \\ E[X_4] &= 1+\rho(1+\rho+\rho^2 N) = 1+\rho+\rho^2+\rho^3 N \\ E[X_n] &= 1+\rho+\cdots+\rho^{n-2}+\rho^{n-1}N. \end{aligned}$$

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: Wald's Identity

Theorem Wald's Identity

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

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

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

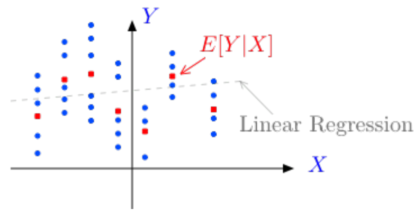
CE = MMSE

Theorem

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

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

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

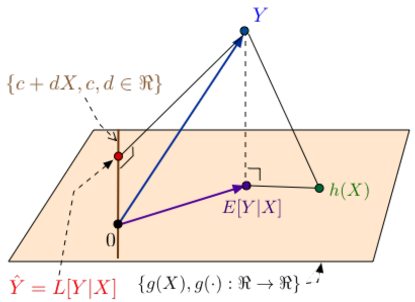
$$\begin{aligned} E[(Y - h(X))^2] &= E[(Y - g(X) + g(X) - h(X))^2] \\ &= E[(Y - g(X))^2] + E[(g(X) - h(X))^2] \\ &\quad + 2E[(Y - g(X))(g(X) - h(X))]. \end{aligned}$$

But,

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

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

$E[Y|X]$ 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
  - Wald
- MMSE: $E[Y|X]$ minimizes $E[(Y - g(X))^2]$ over all $g(\cdot)$