Conditional Expectation

- Random vectors
- Conditional Expectation - basic defn
- Conditional Expectation as an MMSE estimator
- Conditioning on a σ-field

Random Vectors

- Random vector $\underline{X}$ of dimension n is a collection of n random variables $\underline{X} = (X_1, X_2, \dots, X_n)$

- CDF $\quad F_{\underline{X}}(x_1, x_2, \dots, x_n) = \mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n]$
  intersection of events

- If $\underline{X}$ is discrete, then $\underline{X}$ has a pmf
  $$P_{\underline{X}}(x_1, x_2, \dots, x_n) = \mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n]$$

  If $\underline{X}$ is absolutely continuous, then it has a pdf f_X
  s.t. $$F_{\underline{X}}(x_1, x_2, \dots, x_n) = \int_{-\infty}^{x_n}\int_{-\infty}^{x_{n-1}}\cdots\int_{-\infty}^{x_1} f(z_1, z_2, \dots, z_n)\, dz_1\, dz_2 \dots dz_n$$

- For fn $g : \mathbb{R}^n \to \mathbb{R}$, its expectation is
  $$\mathbb{E}[g(\underline{X})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n)\, dx_1\, dx_2 \dots dx_n$$
  This inherits all the properties of $\mathbb{E}[\cdot]$ in one-dimension

- If $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n$ (mutually independent)
  then $F_{\underline{X}}(x_1, x_2, \dots, x_n) = F_{X_1}(x_1) F_{X_2}(x_2) \dots F_{X_n}(x_n)$

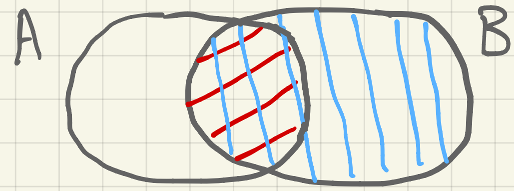
Basic Conditional Probability

(Revision of what you should have seen before)

- For $A, B \in \mathcal{F}$ s.t. $\mathbb{P}[B] > 0$,

$$\mathbb{P}[A \mid B] = \mathbb{P}[A \cap B] / \mathbb{P}[B]$$

- Pictorially



- Similarly we can extend this to r.v.s conditioned on events
  - For any r.v. X and event A
    - Conditional CDF $\quad F_{X|A}(t) = \mathbb{P}[X \leq t \mid A]$
    - natural event - $A = \{Y = y\}$ for some r.v. Y
  - For discrete r.v. X

$$P_{X|A}(t) = \mathbb{P}[X = t \mid A], \quad \mathbb{E}[X \mid A] = \sum_x x \, P_{X|A}(x)$$

  - For continuous r.v. X, Y, and y s.t. $f_Y(y) > 0$

$$f_{X|Y=y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}, \quad \mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \, f_{X|Y=y}(x) \, dx$$

- Useful fact - for any r.v. X and event A

$$\mathbb{E}[X \, \mathbb{1}_A] = \mathbb{E}[X \mid A] \, \mathbb{P}[A]$$

Conditional Expectation (via conditional probabilities)

- Now consider r.v. X and Y and let $\phi(y) \triangleq \mathbb{E}[X \mid Y = y]$.

Then the conditional expectation of X given Y is defined as

$$\mathbb{E}[X \mid Y] = \phi(Y)$$

- <u>Notes</u>

i) $\mathbb{E}[X \mid Y]$ is a random variable!

ii) Sometimes denoted as $\mathbb{E}^{Y}[X]$ (see Brémaud)

Properties of $\mathbb{E}[X|Y]$

We first look at some props of $\mathbb{E}[X|Y]$, before trying to understand it in more detail.

- $\mathbb{E}[\lambda_1 X_1 + \lambda_2 X_2 | Y] = \lambda_1 \mathbb{E}[X_1|Y] + \lambda_2 \mathbb{E}[X_2|Y]$
  (linearity)
- $g_1(x) > g_2(x)\ \forall x \Rightarrow \mathbb{E}[g_1(X)|Y] \geq \mathbb{E}[g_2(X)|Y]$
  (monotonicity)

These follow from properties of $\mathbb{E}[\cdot]$

Thm - $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ (assuming $\mathbb{E}[|X|] < \infty$)

(tower rule)

Pf. $\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} f_Y(y)\, \mathbb{E}[X|Y=y]\, dy$

$$= \int_{-\infty}^{\infty} f_Y(y) \left(\int_{-\infty}^{\infty} \frac{f_{XY}(x,y)}{f_Y(y)}\, x\, dx \right) dy$$

Fubini
(Assuming $\mathbb{E}[|X|] < \infty$)

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x f_{XY}(x,y)\, dy \right) dx$$

$$= \int_{-\infty}^{\infty} x f_X(x)\, dx = \mathbb{E}[X]$$

Thm - $\mathbb{E}[g(Y) \mid Y] = g(Y)$, and more generally $\mathbb{E}[g(Y)h(X,Y) \mid Y] = g(Y)\,\mathbb{E}[h(X,Y) \mid Y]$

(pull-out property)

Pf - Again we will assume X, Y have a joint pdf f_{XY}. Now for any y s.t. $f_Y(y) > 0$

$$\mathbb{E}[g(Y)h(X,Y) \mid Y = y] = \int_{-\infty}^{\infty} g(y)h(x,y) \frac{f_{XY}(x,y)}{f_Y(y)}\,dx$$

$$= g(y) \int_{-\infty}^{\infty} h(x,y) \frac{f_{XY}(x,y)}{f_Y(y)}\,dx$$

$$= g(y)\,\mathbb{E}[h(X,y) \mid Y = y]$$

Thus $\mathbb{E}[g(Y)h(X,Y) \mid Y] = g(Y)\,\mathbb{E}[h(X,Y) \mid Y]$

Thm - If $X \perp\!\!\!\perp Y$, $\mathbb{E}[g(X) \mid Y] = \mathbb{E}[g(X)]$

(independence & conditioning)

Pf - $\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) \frac{f_{XY}(x,y)}{f_Y(y)}\,dx$ $\quad \because X \perp\!\!\!\perp Y$

$$= \int_{-\infty}^{\infty} g(x) \frac{f_X(x)\cancel{f_Y(y)}}{\cancel{f_Y(y)}}\,dx = \mathbb{E}[g(X)]$$

Conditional Expectation $\equiv$ Estimation

- The best way to understand $\mathbb{E}[X|Y]$ is in terms of estimation - In particular, suppose we have access to a random variable Y, and want to use it to approximate some other r.v. X as $\hat{X} = g(Y)$ for some fn g.

$$\boxed{\begin{array}{l}\underline{\text{Claim}} - \ g^*(Y) = \mathbb{E}[X|Y] \text{ is the MMSE (minimum mean-squared error) approximation} \\ \text{of } X \text{, i.e, it minimizes } \mathbb{E}[(X-g(Y))^2] \text{ over} \\ \text{all } g \text{ s.t } \mathbb{E}[(g(Y))^2] < \infty .\end{array}}$$

- This can actually be used to <u>define</u> $\mathbb{E}[X|Y]$!
- You will see this in more detail in 6700; however we will now see a brief proof of this.

Eg- For any r.v. X, suppose we want to approximate it by some constant $b \in \mathbb{R}$, such that we minimize the mean-squared error $\mathbb{E}[(X-b)^2]$

Then we have

$$\mathbb{E}[(X-b)^2] = \mathbb{E}[((X-\mathbb{E}X)+(\mathbb{E}X-b))^2]$$

$$= \underbrace{\mathbb{E}[(X-\mathbb{E}X)^2]}_{Var(X)} + \mathbb{E}[(\mathbb{E}X-b)^2] \quad \left(\begin{array}{c}\text{linearity of} \\ \text{Expectation}\end{array}\right)$$

$$= 2(\mathbb{E}X-b)\underbrace{\mathbb{E}[X-\mathbb{E}[X]]}_{=0} \longrightarrow + 2\mathbb{E}[(X-\mathbb{E}X)(\mathbb{E}X-b)]$$

$$= Var(X) + \mathbb{E}[(b-\mathbb{E}X)^2]$$

$\therefore$ to minimize $\mathbb{E}[(X-b)^2]$, we choose $b^* = \mathbb{E}[X]$

- Now we can extend this to estimating X by $g(Y)$.

We have $\mathbb{E}[(X-g(Y))^2] = \int_{-\infty}^{\infty} \mathbb{E}[(X-g(Y))^2]f_Y(y)\,dy$

where $\mathbb{E}[(X-g(Y))^2 | Y=y] = \int_{-\infty}^{\infty} (x-g(y))\underbrace{f_{X|Y=y}(x)}_{= f_{XY}(x,y)/f_Y(y)}\,dx$

From above, we know $\mathbb{E}[(X-g(y))^2 | Y=y]$ is minimized by setting $g^*(y) = \mathbb{E}[X|Y=y]$

$\Rightarrow$ The MMSE estimator $g^*(Y) = \mathbb{E}[X|Y]$

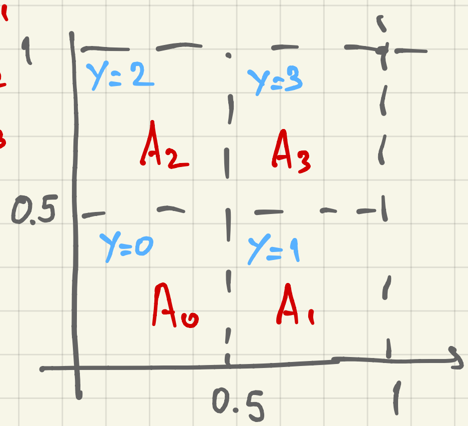
Eg - 'Visualizing' $\mathbb{E}[X|Y]$

'distributed as'

Suppose $\Omega = [0,1]^2$, $X = (X_1, X_2)$, $X_1, X_2 \sim \text{Unif}[0,1]$, $X_1 \perp\!\!\!\perp X_2$

and $Y = \begin{cases} 0 \; ; \; \omega \in [0,0.5]\times[0,0.5] \equiv A_0 \\ 1 \; ; \; \omega \in (0.5,1]\times[0,0.5] \equiv A_1 \\ 2 \; ; \; \omega \in [0,0.5]\times(0.5,1] \equiv A_2 \\ 3 \; ; \; \omega \in (0.5,1]\times(0.5,1] \equiv A_3 \end{cases}$

(see fig on right)



- Recall we defined $\sigma(\mathcal{C})$ to be the smallest σ-field containing a given collection of sets $\mathcal{C}$

The σ-field $\sigma(\{A_0, A_1, A_2, A_3\})$ is referred to as the **σ-field generated by Y**, and denoted $\mathcal{F}_Y$, and the conditional expectation $\mathbb{E}[X|Y]$ can now be alternately written as $\mathbb{E}[X|\mathcal{F}_Y]$ - this makes it clear that $\mathbb{E}[X|Y]$ is a <u>function</u> that associates a number with each set $A \in \mathcal{F}_Y$ (and such that the numbers obey Kolmogorov's axioms)

$f_{X_1|Y}(x) = \frac{1}{\mathbb{P}[Y=0]} \;\; \forall x \in [0,0.5]$

- In this case $\mathbb{E}[X|Y=0] = \left(\int_0^{0.5} x_1\, 2\, dx_1, \int_0^{0.5} x_2\, 2\, dx_2\right)$
$= (0.25, 0.25)$

and $\mathbb{E}[X|Y] = \begin{cases} (0.25, 0.25) \; ; \; Y=0 \\ (0.75, 0.25) \; ; \; Y=1 \\ (0.25, 0.75) \; ; \; Y=2 \\ (0.75, 0.75) \; ; \; Y=3 \end{cases}$

- Thus $\mathbb{E}[X|Y] = \mathbb{E}[X|\mathcal{F}_Y]$ essentially takes every set in $\mathcal{F}_Y$ and associates the 'most likely' (in a mean-squared sense) number for X in that set.
- You can think of this as a form of 'data compression' $\equiv$ given some σ-field $\mathcal{F}_Y$ (generated by Y), and a r.v. X, we 'smear' the information of X over $\mathcal{F}_Y$

- We next use this idea to give a more general definition of $\mathbb{E}[X|Y]$, which covers the two definitions we have seen -

i) $\mathbb{E}[X|Y] = g(Y)$, where $g(y) = \mathbb{E}[X|Y=y]$

ii) $\mathbb{E}[X|Y]$ is the (unique) fn $g(Y)$ with $\mathbb{E}[|g(Y)|^2] < \infty$ which minimizes the mean-squared error $\mathbb{E}[(X-g(Y))^2]$

Conditioning on a σ-field

We now see a more abstract defn of $\mathbb{E}[X|Y]$ that generalizes the previous defns, and also the previous discussion.

- It is more general as it makes less assumptions (Note: for defn (i), we assumed X and Y have a pdf, for (ii), we needed $\mathbb{E}[g(Y)]^2]$ $< \infty$; in contrast we will now only need $\mathbb{E}[|g(Y)|] < \infty$, which is weaker).
- It is more intuitive (even though more abstract!) once you get comfortable with the use of σ-fields
- It captures the idea of $\mathbb{E}[X|Y]$ as a means of 'compressing information'.
- It will be important later when we talk about Markov chains & Martingales

We first need some defns. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space.

i) A collection $\mathcal{D}$ is a sub σ-field of $\mathcal{F}$ if $\mathcal{D}$ is a σ-field and $\mathcal{D} \subseteq \mathcal{F}$

ii) A r.v X is said to be $\mathcal{D}$-measurable or adapted to $\mathcal{D}$ if $\{X \leq t\} \in \mathcal{D}$

iii) For any collection of r.v $\underline{Y} = \{Y_i ; i \in I\}$, the σ-field generated by $\underline{Y}$ (denoted as $\sigma(\underline{Y})$ or $\mathcal{F}_Y$) is defined as the smallest sub-σ-field of $\mathcal{F}$ containing all sets of the form $\{Y_i \leq t\}$, $i \in I$, $t \in \mathbb{R}$.

iv) The σ-field $\mathcal{D} \triangleq \{\Omega, \phi\}$ is referred to as the trivial σ-field. The only r.v. which are measurable w.r.t $\mathcal{D}$ are constants, i.e., $X(\omega) = c \ \forall \omega \in \Omega$ (for some $c \in \mathbb{R}$)

<u>Defn</u> - Given prob space $(\Omega, \mathcal{F}, \mathbb{P})$, r.v X with $\mathbb{E}[X] < \infty$, and sub σ-field $\mathcal{D} \subseteq \mathcal{F}$, the conditional expectation of X given $\mathcal{D}$, denoted $\mathbb{E}[X|\mathcal{D}]$ is the (almost-sure) unique r.v on $(\Omega, \mathcal{F}, \mathbb{P})$ s.t.

i) $\mathbb{E}[X|\mathcal{D}]$ is $\mathcal{D}$-measurable

ii) $\mathbb{E}\left[(X - \mathbb{E}[X|\mathcal{D}])\mathbb{1}_A\right] = 0 \;\; \forall A \in \mathcal{D}$

- <u>Fact</u> - $\exists$ unique r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$ satisfying the above (See for example, Hajek Ch 10.1)

- <u>Note</u> - $\because \Omega \in \mathcal{D}$ for any $\mathcal{D}$, we can set $A = \Omega$ in the above defn to get

$$\mathbb{E}\left[(X - \mathbb{E}[X|\mathcal{D}])\mathbb{1}_\Omega\right] = \mathbb{E}\left[(X - \mathbb{E}[X|\mathcal{D}])\right] = 0$$

$$\Rightarrow \mathbb{E}\left[\mathbb{E}[X|\mathcal{D}]\right] = \mathbb{E}[X] \quad \forall \mathcal{D}$$

Indeed $\mathbb{E}[X] = \mathbb{E}[X | \{\Omega, \phi\}]$, i.e., the trivial σ-field.

- We can now restate (and prove) properties of $\mathbb{E}[X|\mathcal{D}]$. Below, we assume $\mathbb{E}[X] < \infty$ $\forall$ r.v.

i) $\mathbb{E}[aX+bY|\mathcal{D}] = a\mathbb{E}[X|\mathcal{D}] + b\mathbb{E}[Y|\mathcal{D}]$ (linearity)

ii) If X is $\mathcal{D}$-measurable, then $\mathbb{E}[g(X)|\mathcal{D}] = g(X)$

(more generally, $\mathbb{E}[g(X)h(X,Y)|\mathcal{D}] = g(X)\,\mathbb{E}[h(X,Y)|\mathcal{D}]$)

(pull-out property)

iii) If $\mathcal{A} \subset \mathcal{D} \subset \mathcal{F}$

(tower rule)

$$\mathbb{E}[\mathbb{E}[X|\mathcal{A}]|\mathcal{D}] = \mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] = \mathbb{E}[X|\mathcal{A}]$$

- Note - The way to remember the tower rule is that if you condition A on multiple σ-fields, then this is same as conditioning on the smallest (or coarsest) σ-field. This corresponds to the notion of 'conditioning as compression' - if you compress X to a coarse σ-field, then you cannot recover information!

- <u>Pf of tower rule</u> - Let $\mathcal{A} \subset \mathcal{D} \subset \mathcal{F}$. We want to show that $\mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] = \mathbb{E}[X|\mathcal{A}]$. Note that both $\mathbb{E}[X|\mathcal{A}]$ and $\mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}]$ are $\mathcal{A}$-measurable (by defn, since they are r.v. of the form $\mathbb{E}[Y|\mathcal{A}]$). Also note that we can write

$$X - \mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] = (X - \mathbb{E}[X|\mathcal{D}]) - (\mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] - \mathbb{E}[X|\mathcal{D}])$$

Now for any $A \in \mathcal{A}$, we have $A \in \mathcal{D}$. By defn of $\mathbb{E}[\cdot|\mathcal{A}]$, we have

$$\mathbb{E}[(X - \mathbb{E}[X|\mathcal{D}])\mathbb{1}_A] = 0, \quad \mathbb{E}[(\mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] - \mathbb{E}[X|\mathcal{D}])\mathbb{1}_A] = 0$$

$$\Rightarrow \quad \mathbb{E}[(X - \mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}])\mathbb{1}_A] = 0 \quad \forall A \in \mathcal{A}$$

However, by the fact that $\mathbb{E}[X|\mathcal{A}]$ is a.s. unique, we must have $\mathbb{E}[\mathbb{E}[X|\mathcal{D}]|\mathcal{A}] = \mathbb{E}[X|\mathcal{A}]$

<u>Note</u> - Instead of defining $\mathbb{E}[X|Y]$ in terms of $\mathbb{E}[X|\mathcal{F}_Y]$ as above, we can directly define it as follows: given X, Y s.t $\mathbb{E}[|X|] < \infty$, then $\mathbb{E}[X|Y]$ is the (unique) fn $g(Y)$ s.t for every non-negative bounded fn ψ, we have

$$\mathbb{E}[(X - g(Y))\psi(Y)] = 0 \quad \text{a.s}$$

- See Brémaud Thm 2.3.15 for proof of existence & uniqueness