Advanced Probability Theory **Fall 2023**

Lecture 2: Conditional Expectation

Lecturer: ***Ioannis Karatzas*** *Scribes: Heyuan Yao*

2.1 Background, Definition and Connection to Hilbert Space

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider random variables X taking values in $\{x_1, ..., x_m\}$ and Z taking values in $\{z_1, ..., z_m\}$. Elementary conditional probabilities and expectations are defined by:

$$\mathbb{P}(X = x_i | Z = z_j) = \frac{\mathbb{P}(X = x_i, Z = z_j)}{\mathbb{P}(Z = z_j)} \tag{2.1}$$

$$\mathbb{E}(X | Z = z_j) = \sum_{i=1}^{m} x_i \mathbb{P}(X = x_i | Z = z_j) =: h_j. \tag{2.2}$$

The random variable $H = \mathbb{E}(X|Z)$, the conditional expectation of X given Z, is defined as: $H(\omega) = h_j$, on $\omega \in \{Z = z_j\}$. Formally,

$$H = \sum_{j=1}^{n} h_j \mathbb{I}_{\{Z = z_j\}}.$$

Yes, conditional expectation is a random variable. Life expectancy at birth is an expectation; but life expectancy at birth conditional on sex is a random variable, taking one value for males and another for females.

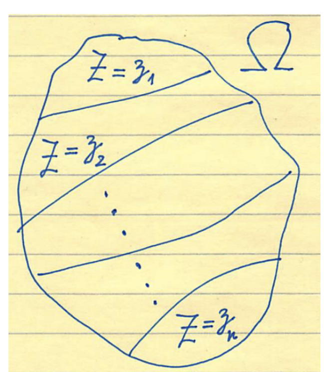

Figure 2.1: The Partition of Ω

Now look at this definition as follows (see Figure 2.1): "Reporting the value of $Z(\omega)$" amounts to partitioning

ω into "Z-atoms"; and

$$\mathcal{G} = \sigma(Z),$$

the $\sigma-$algebra generated by Z, consisted of the 2^n possible unions of there $Z-$atoms $\{Z = z_1\}, ..., \{Z = z_n\}$. Clearly, $H = \sum_{j=1}^{n} h_j \mathbb{I}_{\{Z=z_j\}}$ is constant on each of these atoms; to wit, H is $\mathcal{G}-$measurable. Furthermore,

$$\begin{aligned}\int_{Z=z-j} H d\mathbb{P} = h_j \mathbb{P}(Z = z_j) &= \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i | Z = z_j)\mathbb{P}(Z = z_j) \\ &= \sum_{i=1}^{n} x_i \mathbb{P}(X_i = x_i, Z = z_j) = \int_{Z=z_j} X d\mathbb{P}\end{aligned}$$

with $G_j := \{Z = z_j\}$, this reads: $\mathbb{E}(H\mathbb{I}_{G_j}) = \mathbb{E}(X)\mathbb{I}_{G_j}$.

Now, for every event $G \in \mathcal{G}$, the indicator $\mathbb{I}_G$ is a sum of $\mathbb{I}_{G_j}$'s, so the linearity of the expectation gives

$$\mathbb{E}(H\mathbb{I}_{G_j}) = \mathbb{E}(X)\mathbb{I}_{G_j}, \text{ or } \int_G H d\mathbb{P} = \int_G X d\mathbb{P}. \tag{2.3}$$

Okay, so what?

Well, it turns out that for many years up until 1930, centuries really, nobody had any idea how to define conditional probabilities or expectations in any generality. This was done finally by A.N. Kolmogorov in 1933: the biggest accomplishment in his *Grundbegriffe der Wahrscheinlichkeitsrechnung*.

Theorem 2.1 (Conditional Expectation) *On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let X be an integrable random variable and $\mathcal{G}$ a $\sigma-$algebra $\mathcal{G} \subseteq \mathcal{F}$.*

There exists then a $\mathbb{P}-$a.e. unique r.v. $H : \Omega \to \mathbb{R}$, which is $\mathcal{G}-$measurable and satisfies

$$\int_G H d\mathbb{P} = \int_G X d\mathbb{P}, \qquad \forall G \in \mathcal{G}. \tag{2.4}$$

This random variable is called (the) ***conditional expectation of** X **given** $\mathcal{G}$*, and denoted*

$$H = \mathbb{E}(X|\mathcal{G}).$$

Proof: Suppose $X \geq 0$. Then the set function $\mathcal{G} \ni G \mapsto \nu(G) := \int_G X d\mathbb{P} \in [0, \infty)$ is $\sigma-$additive, thus a measure, and finite: $\nu(\Omega) = \mathbb{E}(X) < \infty$. The Radon-Nikodým theorem gives then the existence of $\mathcal{G}-$measurable $H : \Omega \to [0, \infty)$ with $\nu(G) = \int_G H d\mathbb{P}$, $\forall G \in \mathcal{G}$, proving the claim.

More generally, we write $X = X^+ + X^-$ and proceed as above and by additivity. ■

It is instructive to write the requirement (2.4) as

$$\mathbb{E}[(H - X)\mathbb{I}_G] = 0, \qquad , \forall G \in \mathcal{G};$$

then building up from indicators to simple, and then to general measurable functions, as

$$\mathbb{E}[(H - X)Y] = 0 \tag{2.5}$$

for every $\mathcal{G}$-measurable $Y : \Omega \to \mathbb{R}$ for with the product $(H - X)Y$ in (2.5) is integrable.

In particular, suppose $X \in \mathbb{L}^2(\mathcal{F})$. Then, as we will see precisely,

$$H \in \mathbb{L}^2(\mathcal{G}),$$

and by Cauchy-Schwarz, the product in (2.5) is integrable for every $Y \in \mathbb{L}^2(\mathcal{G})$.

But then 2.5) becomes the orthogonality relation (and you can see a visualization below)

$$\langle H - X, Y\rangle_{\mathbb{L}^2(\mathcal{F})} = \mathbb{E}[(H - X)Y] = 0, \quad \forall Y \in \mathbb{L}^2(\mathcal{G}).$$

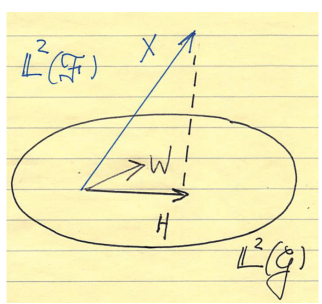

Figure 2.2: The difference X_H is orthogonal to every $Y \in \mathbb{L}^2(\mathcal{G})$

Then, conditional expectation becomes projection in the subspace $\mathbb{L}^2(\mathcal{G})$ of the Hilbert space $\mathbb{L}^2(\mathcal{F})$:

$$\mathbb{E}(H - X)^2 = \inf_{W \in \mathbb{L}^2(\mathcal{G})} \mathbb{E}(W - X)^2.$$

This interpretation, of conditional expectation as projection, makes absolutely clear properties that are valid in complete generality.

2.2 Properties of Conditional Expectation

Tower Property:

If $\mathcal{G}_1 \subseteq \mathcal{G}_2$ are sub$-\sigma-$algebras pf $\mathcal{F}$, then

$$\mathbb{E}[\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1] = \mathbb{E}(X|\mathcal{G}_1). \tag{2.6}$$

Taking Out What is Known:

If $Y : \Omega \to \mathbb{R}$ is $\mathcal{G}-$measurable and $\mathbb{E}[|X|(1 + |Y|)] < \infty$, then

$$\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}). \tag{2.7}$$

Independence Makes Conditioning Irrelevant:

If the $\sigma-$algebra $\mathcal{G}$ and $\sigma(X)$ are independent,

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X). \tag{2.8}$$

A bit more generally, if $\mathcal{H}$ is another sub-σ-algebra pf $\mathcal{F}$ such that $\mathcal{G}$ and $\sigma(\sigma(X), (H))$ are independent, then

$$\mathbb{E}(X|\mathcal{G} \vee \mathcal{H}) = \mathbb{E}(X|\mathcal{H}). \tag{2.9}$$

Here, $\mathcal{G} \vee \mathcal{H}$ is the $\sigma-$algebra generated by

$$\mathcal{J} := \{A \in \mathcal{F} : A \in G \cap H, \;\; with \;\; G \in \mathcal{G}, \;\;\; H \in \mathcal{H}\}. \tag{2.10}$$

Linearity:

For real α, β, we have

$$\mathbb{E}(\alpha X_1 + \beta X_2|\mathcal{G}) = \alpha\mathbb{E}(X_1\mathcal{G}) + \beta\mathbb{E}(X_2|\mathcal{G}). \tag{2.11}$$

Monotonicity:

If $X_1 \leq X_2$, then

$$\mathbb{E}(X_1|\mathcal{G}) \leq \mathbb{E}(X_2|\mathcal{G}). \tag{2.12}$$

MCT:

If $) \leq X_1 \leq X_2 \leq ...$, and $X = \lim_n \uparrow X_n$, then

$$\lim_n \uparrow \mathbb{E}(X_n|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}). \tag{2.13}$$

DCT:

If $X_n \leq Y$, $\forall n \in \mathbb{N}$, for $\mathbb{E}Y < \infty$, and $\lim_n X_n = X$, then

$$\lim_n \mathbb{E}(X_n|\mathcal{G}) = \mathbb{E}(X|\mathcal{G}). \tag{2.14}$$

Conditional Jensen:

If $f : \mathbb{R} \to \mathbb{R}$ is convex, then

$$f(\mathbb{E}(X|\mathcal{G})) \leq \mathbb{E}[f(X)|\mathcal{G}]. \tag{2.15}$$

Conditioning Decreases $\mathbb{L}^p$ norms: For any $q \leq p \leq \infty$, we have

$$\|\mathbb{E}(X|\mathcal{G})\|_p \leq \|X\|_p \tag{2.16}$$

Conditioning Decreases Relative Entropy: Let $\mathbb{P}, \mathbb{Q}$ be probability measures on $(\Sigma, \mathcal{F})$ with $\mathbb{Q} < \mathbb{P}$, and let

$$X := \frac{d\mathbb{Q}}{d\mathbb{P}}|_{\mathcal{F}}, \qquad Z := \frac{d\mathbb{Q}}{d\mathbb{P}}|_{\mathcal{G}},$$

$$\mathcal{D}_{\mathcal{F}}(\mathbb{Q}|\mathbb{P}) = \mathbb{E}^{\mathbb{Q}}(log(X)) = \mathbb{E}^{\mathbb{P}}(Xlog(X))$$

$$\mathcal{D}_{\mathcal{G}}(\mathbb{Q}|\mathbb{P}) = \mathbb{E}^{\mathbb{Q}}(log(Z)) = \mathbb{E}^{\mathbb{P}}(Zlog(Z))$$

Then

$$Z = \mathbb{E}^{\mathbb{P}}(X|\mathcal{G}), \qquad \mathcal{D}_{\mathcal{F}}(\mathbb{Q}|\mathbb{P}) \geq \mathcal{D}_{\mathcal{G}}(\mathbb{Q}|\mathbb{P}) \tag{2.17}$$

Conditional probability as Conditional Expectation:

For $A \in \mathcal{F}$, we set

$$\mathbb{P}(A|\mathcal{G}) := \mathbb{E}(\mathbb{I}_A|\mathcal{G}). \tag{2.18}$$

Proof: [Independence Makes Conditioning Irrelevant]

For equation (2.8), with $G\mathcal{G}$, the r.v.'s X, $\mathbb{I}_G$ are independent. Thus

$$\mathbb{E}(X\mathbb{I}_G) = \mathbb{E}(X)\mathbb{E}(\mathbb{I}_G) = \mathbb{E}[\mathbb{E}(X)\mathbb{I}_G].$$

For equation (2.9), assume $X \geq 0$ for simplicity; denote $Y := \mathbb{E}(X|\mathcal{H})$; take arbitrary $G \in \mathcal{G}$, $H \in \mathcal{H}$ and set $A := G \cap \mathcal{H} \in \mathcal{J}$. We have

$$\mathbb{E}(X\mathbb{E}(\mathbb{I}_A) = \mathbb{E}(X\mathbb{E}(\mathbb{I}_H\mathbb{E}(\mathbb{I}_G) = \mathbb{E}(X\mathbb{E}(\mathbb{I}_H)\mathbb{P}(G),$$
$$\mathbb{E}(Y\mathbb{E}(\mathbb{I}_A) = \mathbb{E}(Y\mathbb{E}(\mathbb{I}_H\mathbb{E}(\mathbb{I}_G) = \mathbb{E}(Y\mathbb{E}(\mathbb{I}_H)\mathbb{P}(G).$$

Now, the last terms in these two expressions are the same; so the first are the same as well.

This means that the measures

$$\mu(A) := \mathbb{E}(X\mathbb{I}_A), \quad \nu(A) := \mathbb{E}(Y\mathbb{I}_A)$$

agree on $\mathcal{J}$. But $\mu(\Omega) = \mathbb{E}(X) = \mathbb{E}(Y) = \nu(\Omega) \leq \infty$, and $\mathcal{J}$ is closed under finite interesction; so, by Monote-Class type results, the two measures μ, ν agree also on $\sigma(\mathcal{J}) =: \mathcal{G} \vee \mathcal{H}$. This means

$$\mathbb{E}(X\mathbb{I}_A) = \mathbb{E}(Y\mathbb{I}_A), \quad \forall A \in \mathcal{G} \vee \mathcal{H}$$

or equivalently

$$\mathbb{E}(X|\mathcal{G} \vee \mathcal{H}) = Y \equiv \mathbb{E}(X|\mathcal{H}).$$

■

Proof: [Conditional Jensen]

Every convex function $\phi : \mathbb{R} \to \mathbb{R}$ can be expressed as the upper envelope of a countable collection of affine (tangent) functions: $\phi(x) = \sup_n(a_n x + b)$.

For fixed $n \in \mathbb{N}$, the inequality $\phi(X) \geq a_n X + b_n$ gives

$$\mathbb{E}[\phi(X)|\mathcal{G}] \geq a_n\mathbb{E}(X|\mathcal{G}) + b_n, \quad a.e.$$

Here we need to be careful about the exceptional set E_n: its may depend on $n \in \mathbb{N}$. But even if it does, the union $E = \cup_{n\in\mathbb{N}} E_n$ of the exceptional sets is also a set of measure zero, so the inequality

$$\mathbb{E}[\phi(X)|\mathcal{G}] \geq \sup_{n\in\mathbb{N}} (a_n \mathbb{E}(X|\mathcal{G}) + b_n) = \phi(\mathbb{E}(X|\mathcal{G}))$$

holds on $\Omega n E$, i.e., a.e. ■

2.3 Convergence Result and Marginal Law

Conditional Expectations have excellent "continuity" properties relative to monotone sequences of $\sigma-$algebras.

We detail the most basic of these right below. The proofs have to wait for the development of the Theory of Martingales.

P. Lévy Convergence Results:

Suppose Z is an integrable variable.

(i) If $(\mathcal{G}_n)_{n\in\mathbb{N}}$ is an increasing sequence of $\sigma-$algebras (i.e., $\mathcal{G}_n \subseteq \mathcal{G}_{n+1}$, $\forall n \in \mathbb{N}$), and $\mathcal{G}_\infty := \sigma\left(\cup_{n\in\mathbb{N}} \mathcal{G}_n\right)$, then

$$\lim_{n\to\infty} \mathbb{E}(Z|\mathcal{G}_n) = \mathbb{E}(Z|\mathcal{G}_\infty), \quad a.e.$$

(ii) Same, if $(\mathcal{G}_n)_{n\in\mathbb{N}}$ is an decreasing sequence of $\sigma-$algebras, i.e., $\mathcal{G}_{n+1} \subseteq \mathcal{G}_n$, $\forall n \in \mathbb{N}$ and $\mathcal{G}_\infty := \sigma\left(\cap_{n\in\mathbb{N}} \mathcal{G}_n\right)$.

· Now suppose that $\mathcal{G} = \sigma(Z)$, for some random variable $Z : \Omega \to \mathbb{R}$, and that $X : \Omega \to [0, \infty)$ is integrable. Then

$$\mathbb{E}(X|Z) := \mathbb{E}(X|\mathcal{G}) = h(Z), \quad \mathbb{P} - a.e.$$

for some Borel-measurable $h : \mathbb{R} \to \mathbb{R}$.

· In fact, if we denote by $\mu(A) = \mathbb{P}(Z \in A)$ the distribution of Z, and introduce the finite measure

$$\nu(A) := \mathbb{E}(X\mathbb{I}_A(Z)) = \int_{Z^{-1}(A)} X d\mathbb{P}, \quad A \in \mathcal{B}(\mathbb{R})$$

then this function h is

$$h = \frac{d\nu}{d\mu}, \quad \mu - a.e.$$

Abusing notation a bit, we write

$$h(\xi) = \mathbb{E}(X|Z = \xi)$$

for $\xi \in \mathbb{R}$; but without ever forgetting that this function is only defined $\mu-$a.e., and that we should avoid making bold statements about it.

Special Structure:

Suppose now that the two random variables X, Z have joint probability distribution

$$\mathbb{P}[(X,Z) \in A] = \int\int_A f(x,\xi)dxd\xi, \quad A \in \mathcal{B}(\mathbb{R})$$

for some probability density function $f : \mathbb{R}^2 \to (0,\infty)$. thus also "marginal distribution"

$$\mathbb{P}[Z \in B] = \int\limits_B {}_Z(\xi)d\xi, \quad B \in \mathcal{B}(\mathbb{R})$$

with "marginal probability density function"

$$f_Z(\xi) = \int\limits_{\mathbb{R}} f(x,\xi)dx.$$

Then the function h from page 2-7 is

$$h(\xi) = \mathbb{E}(X|Z \in \xi) = \int_{\mathbb{R}} x\frac{f(x,\xi)}{f_Z(\xi)}dx, \quad \forall\xi \in \mathbb{R}. \tag{2.19}$$