Lecture 10 : Conditional Expectation

STAT205 *Lecturer: Jim Pitman* Scribe: *Charless C. Fowlkes* <fowlkes@cs.berkeley.edu>

10.1 Definition of Conditional Expectation

Recall the "undergraduate" definition of conditional probability associated with Bayes' Rule

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$$

For a discrete random variable X we have

$$\mathbb{P}(A) = \sum_x \mathbb{P}(A, X = x) = \sum_x \mathbb{P}(A|X = x)\mathbb{P}(X = x)$$

and the resulting formula for conditional expectation

$$\begin{aligned}
\mathbb{E}(Y|X = x) &= \int_\Omega Y(\omega)\mathbb{P}(dw|X = x) \\
&= \frac{\int_{X=x} Y(\omega)\mathbb{P}(dw)}{\mathbb{P}(X = x)} \\
&= \frac{\mathbb{E}(Y\mathbf{1}_{(X=x)})}{\mathbb{P}(X = x)}
\end{aligned}$$

We would like to extend this to handle more general situations where densities don't exist or we want to condition on very "complicated" sets.

Definition 10.1 *Given a random variable Y with $\mathbb{E}|Y| < \infty$ on the space $(\Omega, \mathcal{F}, \mathbb{P})$ and some sub-σ-field $\mathcal{G} \subset \mathcal{F}$ we will define the* **conditional expectation** *as the almost surely unique random variable $\mathbb{E}(Y|\mathcal{G})$ which satisfies the following two conditions*

1. *$\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable*
2. *$\mathbb{E}(YZ) = \mathbb{E}(\mathbb{E}(Y|\mathcal{G})Z)$ for all Z which are bounded and $\mathcal{G}$-measurable*

For $\mathcal{G} = \sigma(X)$ when X is a discrete variable, the space Ω is simply partitioned into disjoint sets $\Omega = \sqcup G_n$. Our definition for the discrete case gives

$$\begin{aligned}
\mathbb{E}(Y|\sigma(X)) &= \mathbb{E}(Y|X) \\
&= \sum_n \frac{\mathbb{E}(Y\mathbf{1}_{X=x_n})}{\mathbb{P}(X = x_n)}\mathbf{1}_{X=x_n} \\
&= \sum_n \frac{\mathbb{E}(Y\mathbf{1}_{G_n})}{\mathbb{P}(G_n)}\mathbf{1}_{G_n}
\end{aligned}$$

which is clearly $\mathcal{G}$-measurable.

Exercise 10.2 *Show that the discrete formula satisfies condition 2 of Definition 10.1. (**Hint:** show that the condition is satisfied for random variables of the form $Z = \mathbf{1}_G$ where $G \in \mathcal{C}$ is a collection closed under intersection and $\mathcal{G} = \sigma(\mathcal{C})$ then invoke Dynkin's $\pi - \lambda$)*

10.2 Conditional Expectation is Well Defined

Proposition 10.3 $E(X|\mathcal{G})$ *is unique up to almost sure equivalence.*

Proof Sketch: Suppose that both random variables $\hat{Y}$ and $\hat{\hat{Y}}$ satisfy our conditions for being the conditional expectation $E(Y|X)$. Let $W = \hat{Y} - \hat{\hat{Y}}$. Then W is $\mathcal{G}$-measurable and $E(WZ) = 0$ for all Z which are $\mathcal{G}$-measurable and bounded. If we let $Z = \mathbf{1}_{W>\epsilon}$ (which is bounded and measurable) then

$$\epsilon P(W > \epsilon) \leq E(W\mathbf{1}_{W>\epsilon}) = 0$$

for all $\epsilon > 0$. A similar argument applied to $P(W < -\epsilon)$ allows us to conclude that $P(|W| > \epsilon) = 0$ holds for all ϵ and hence $W = 0$ almost surely making $E(Y|X)$ almost surely unique. ■

Proposition 10.4 $\mathbb{E}(X|\mathcal{G})$ *exists*

We've shown that $\mathbb{E}(Y|\mathcal{G})$ exists in the discrete case by writing out an explicit formula so that "$\mathbb{E}(Y|X)$ to integrates like Y over $\mathcal{G}$-measurable sets." We give three different approaches for attacking the general case.

10.2.1 "Hands On" Proof

The first is a hands on approach by extending the discrete case via limits. We will make use of

Lemma 10.5 William's Tower Property *Suppose $\mathcal{G} \subset H \subset F$ are nested σ-fields and $\mathbb{E}(\cdot|\mathcal{G})$ and $\mathbb{E}(\cdot|\mathcal{H})$ are both well defined then $\mathbb{E}(\mathbb{E}(Y|\mathcal{H})|\mathcal{G}) = \mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(\mathbb{E}(Y|\mathcal{G})|\mathcal{H})$*

A special case is when $\mathcal{G} = \{\emptyset, \Omega\}$ then $\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}Y$ is a constant so it's easy to see $\mathbb{E}(\mathbb{E}(Y|\mathcal{H})|\mathcal{G}) = \mathbb{E}(\mathbb{E}(Y)|\mathcal{H}) = \mathbb{E}(Y)$ and $\mathbb{E}(\mathbb{E}(Y|\mathcal{G})|\mathcal{H}) = \mathbb{E}(\mathbb{E}(Y)|\mathcal{H}) = \mathbb{E}(Y)$

Proof Sketch: Existence via Limits For a disjoint partition $\sqcup G_i = \Omega$ and $G \in \mathcal{G} = \sigma(\{G_i\})$ define

$$E(Y|\mathcal{G}) = \sum_i \frac{E(Y\mathbf{1}_{G_i})}{P(G_i)} 1_{G_i}$$

where we deal appropriately with the niggling possibility of $\mathbb{P}(G_i) = 0$ by either throwing out the offending sets or defining $\frac{0}{0} = 0$.

We now consider an arbitrary but countably generated σ-field $\mathcal{G}$. This situation is not too restrictive, for example the σ-field associated with an $\mathbb{R}$-valued random variable X is generated by the countable collection $\{B_i = (X \leq r_i) : r \in \mathbb{Q}\}$. If we set $\mathcal{G}_n = \sigma(B_1, B_2, \ldots, B_n)$ then $\mathcal{G}_n$ is increasing to the limit $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots \subset \mathcal{G} = \sigma(\cup \mathcal{G}_n)$. For a given n the random variable $Y_n = \mathbb{E}(Y|\mathcal{G}_n)$ exists by our explicit definition above since we can decompose the generating set into a disjoint partition of the space.

Now we show that Y_n converges in some appropriate manner to a Y_∞ which will then function as a version of $E(Y|\mathcal{G})$. We will assume that $\mathbb{E}|Y|^2 < \infty$

Write $Y_n = \mathbb{E}(Y|G_n) = Y_1 + (Y_2 - Y_1) + (Y_3 - Y_2) + \ldots + (Y_n - Y_{n-1})$. The terms in this summation are orthogonal in $\mathbf{L}^2$ so we can compute the variance as

$$s_n^2 = \mathbb{E}(Y_n^2) = \mathbb{E}(Y_1^2) + \mathbb{E}((Y_2 - Y_1)^2) \ldots + \mathbb{E}((Y_n - Y_{n-1})^2)$$

where the cross terms are zero. Let $s^2 = E(Y^2) = E(Y_n + (Y - Y_n)) < \infty$. Then $s_n^2 \uparrow s_\infty^2 \leq s^2 < \infty$. For $n > m$ we know again by orthogonality that $E((Y_n - Y_m)^2) = s_n^2 - s_m^2 \to 0$ as $m \to \infty$ since s_n^2 is just a bounded real sequence. This means that the sequence Y_n is Cauchy in $\mathbf{L}^2$ and invoking the completeness of $\mathbf{L}^2$ we conclude that $Y_n \to Y_\infty$.

All that remains is to check that Y_∞ is a conditional expectation. It satisfies requirement (1) since as a limit of $\mathcal{G}$-measurable variables it is $\mathcal{G}$-measurable. To check (2) we need to show that $E(YG) = E(Y_\infty G)$ for all G which are bounded and $\mathcal{G}$-measurable. As usual, it suffices to check for a much smaller set $\{\mathbf{1}_{A_i} : A_i \in \mathcal{A}\}$ where $\mathcal{A}$ is an intersection closed collection and $\sigma(\mathcal{A}) = \mathcal{G}$. Take this collection to be $\mathcal{A} = \cup_m \mathcal{G}_m$.

$$\mathbb{E}(Y G_m) = \mathbb{E}(Y_m G_m) = \mathbb{E}(Y_n G_m)$$

holds by the tower property for any $n > m$. Noting that $\mathbb{E}(Y_n Z) \to \mathbb{E}(Y_\infty Z)$ is true for all $Z \in \mathbf{L}^2$ by the continuity of inner product this sequence must go to the desired limit which gives $\mathbb{E}(Y\mathcal{G}_m) = \mathbb{E}(Y_\infty \mathcal{G}_m)$ ■

Exercise 10.6 *Remove the countably generated constraint on $\mathcal{G}$. (**Hint:** Be a bit more clever ... for $Y \in \mathbf{L}^2$ look at $\mathbb{E}(Y|\mathcal{G})$ for $\mathcal{G} \subset \mathcal{F}$ with $\mathcal{G}$ finite. Then as above $\sup_{\mathcal{G}} \mathbb{E}(\mathbb{E}(Y|\mathcal{G})^2) \leq \mathbb{E}Y^2$ so we can choose $\mathcal{G}_n$ with $\mathbb{E}(\mathbb{E}(Y|\mathcal{G}_n)^2)$ increasing to this supremum. The $\mathcal{G}_n$ may not be nested but argue that $\mathcal{C}_n = \sigma(\mathcal{G}_1 \cup \mathcal{G}_2 \cup \ldots \cup \mathcal{G}_n)$ are and let $\hat{Y} = \lim_n \mathbb{E}(Y|\mathcal{C}_n)$)).*

Exercise 10.7 *Remove the $\mathbf{L}^2$ constraint on Y. (**Hint:** Consider $Y \geq 0$ and show convergence of $\mathbb{E}(Y \wedge n \mid \mathcal{G})$ then turn crank on the standard machinery)*

10.2.2 Measure Theory Proof

Here we pull out some power tools from measure theory.

Theorem 10.8 Lebesgue-Radon-Nikodym *[2](p.121) If μ and λ are non-negative σ-finite measures on a collection $\mathcal{G}$ and $\mu(G) = 0 \implies \lambda(G) = 0$ (written $\lambda \ll \mu$, pronounced "λ is absolutely continuous with respect to μ") for all $G \in \mathcal{G}$ then there exists a non-negative $\mathcal{G}$ measurable function $\hat{Y}$ such that*

$$\lambda(G) = \int_G \hat{Y} d\mu$$

for all $G \in \mathcal{G}$.

Proof Sketch: Existence via Lebesgue-Radon-Nikodym Assume $Y \geq 0$ and define the probability measure

$$Q(C) = \int_C Y dP = \mathbb{E}Y\mathbf{1}_C$$

which is non-negative and finite because $\mathbb{E}|Y| < \infty$ and Q is absolutely continuous with respect to P. LRN implies the existence of $\hat{Y}$ which satisfies our requirements to be a version of the conditional expectation $\hat{Y} = \mathbb{E}(Y|\mathcal{G})$. For general Y we can employ $\mathbb{E}(Y^+|\mathcal{G}) - \mathbb{E}(Y^-|\mathcal{G})$. ■

10.2.3 Functional Analysis Proof

This gives a nice geometric picture for the case when $Y \in \mathbf{L}^2$

Lemma 10.9 *Every nonempty, closed, convex set E in a Hilbert space H contains a unique element of smallest norm*

Lemma 10.10 Existence of Projections in Hilbert Space *Given a closed subspace K of a Hilbert space H and element $x \in H$, there exists a decomposition $x = y + z$ where $y \in K$ and $z \in K^{\perp}$ (the orthogonal complement).*

The idea for the existence of projections is to let y be the element of smallest norm in $x + K$ and $z = x - y$. See [2](p.79) for a full discussion of Lemma 10.9.

Proof Sketch: Existence via Hilbert Space Projection Suppose $Y \in \mathbf{L}^2(\mathcal{F})$ and $X \in \mathbf{L}^2(\mathcal{G})$. Requirement (2) demands that for all X

$$\mathbb{E}((Y - \mathbb{E}(Y|\mathcal{G}))X) = 0$$

which has the geometric interpretation of requiring $Y - \mathbb{E}(Y|\mathcal{G})$ to be orthogonal to the subspace $\mathbf{L}^2(\mathcal{G})$. Requirement (1) says that $\mathbb{E}(Y|\mathcal{G}) \in \mathbf{L}^2(\mathcal{G})$ so $\mathbb{E}(Y|\mathcal{G})$ is just the orthogonal projection of Y onto the closed subspace $\mathbf{L}^2(\mathcal{G})$. The lemma above shows that such a projection is well defined. ■

10.3 Properties of Conditional Expectation

It's helpful to think of $\mathbb{E}(\cdot|\mathcal{G})$ as an operator on random variables that transforms $\mathcal{F}$-measurable variables into $\mathcal{G}$-measurable ones.

We isolate some useful properties of conditional expectation which the reader will no doubt want to prove before believing

- $\mathbb{E}(\cdot|\mathcal{G})$ is positive:
$$Y \geq 0 \rightarrow \mathbb{E}(Y|\mathcal{G}) \geq 0)$$

- $\mathbb{E}(\cdot|\mathcal{G})$ is linear:
$$\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$$

- $\mathbb{E}(\cdot|\mathcal{G})$ is a projection:
$$\mathbb{E}(E(X|\mathcal{G})|\mathcal{G}) = E(X|\mathcal{G})$$

- More generally, the "tower property". If $\mathcal{H} \subset \mathcal{G}$ then
$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})\mathcal{G}) = \mathbb{E}(X|\mathcal{H})$$

- $\mathbb{E}(\cdot|\mathcal{G})$ commutes with multiplication by $\mathcal{G}$-measurable variables:
$$\mathbb{E}(XY|\mathcal{G}) = E(X|\mathcal{G})Y \text{ for } \mathbb{E}|XY| < \infty \text{ and } Y \in \mathcal{G}$$

- $\mathbb{E}(\cdot|\mathcal{G})$ respects monotone convergence:
$$0 \le X_n \uparrow X \implies \mathbb{E}(X_n|\mathcal{G}) \uparrow \mathbb{E}(X|\mathcal{G})$$

- If ϕ is convex and $\mathbb{E}|\phi(X)| < \infty$ then a conditional form of Jensen's inequality holds:
$$\phi(\mathbb{E}(X|\mathcal{G}) \le \mathbb{E}(\phi(X)|\mathcal{G})$$

- $\mathbb{E}(\cdot|\mathcal{G})$ is a continuous contraction of $\mathbf{L}^p$ for $p \ge 1$:
$$\|\mathbb{E}(X|\mathcal{G})\|_p \le \|X\|_p$$
and
$$X_n \xrightarrow{\mathbf{L}^2} X \text{ implies } \mathbb{E}(X_n|\mathcal{G}) \xrightarrow{\mathbf{L}^2} \mathbb{E}(X|\mathcal{G})$$

- Repeated Conditioning. For $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots$, $\mathcal{G}_\infty = \sigma(\cup \mathcal{G}_i)$, and $X \in \mathbf{L}^p$ with $p \ge 1$ then
$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{a.s.} \mathbb{E}(X|\mathcal{G}_\infty)$$
$$\mathbb{E}(X|\mathcal{G}_n) \xrightarrow{\mathbf{L}^p} \mathbb{E}(X|\mathcal{G}_\infty)$$

10.4 Regular Conditional Distributions

Definition 10.11 *Given random variable $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ and sub-σ-field $\mathcal{G} \subset \mathcal{F}$ we define the* **Markov kernel** *$Q(\omega, A) : \Omega \times \mathcal{S} \to [0,1]$ as a (carefully chosen) version of the conditional probability $\mathbb{P}(X \in A|\mathcal{G})$ which has the properties*

1. *$\omega \mapsto Q(\omega, A)$ is a ($\mathcal{G}$-measurable) version of $\mathbb{P}(X \in A|\mathcal{G})$ for fixed choice of A*
2. *$A \mapsto Q(\omega, A)$ is a probability measure on $(S, \mathcal{S})$*

When $S = \Omega$ and X is the identity map we call Q a **regular conditional probability**

For $G \in \mathcal{G}$ we have that
$$\mathbb{P}(X \in A, G) = \mathbb{E}(\mathbb{P}(X \in A|\mathcal{G})\mathbf{1}_G) = \int_G Q(\omega, A)P(d\omega)$$
and in the case when $\mathcal{G} = \sigma(Y)$ the kernel takes the form
$$Q(\omega, A) = \hat{Q}(Y(\omega), A)$$

for some $\hat{Q} : \mathbb{R} \times \mathcal{B}(\mathbb{R}) \to \mathbb{R}$ which we write as $P(X \in A|Y = y)$ and gives the slick formula

$$\mathbb{P}(X \in A, Y \in B) = \int_B P(X \in A|Y = y) P(Y \in dy)$$

reminiscent of Bayes' rule for discrete variables.

Regular conditional probabilities do not always exist. However, if we are dealing with a random variable whose range is a "nice" space (one for which there exists a measurable 1-1 map to $\mathbb{R}$ whose inverse is also measurable) the following sketch shows we are ok. ([1](p.230) gives full details)

Proof Sketch: Existence of "Regular" Conditional Probabilities First construct $\mathbb{P}(X \in A|\mathcal{G})$ for Borel sets so that it behaves as a probability with respect to A almost surely. Use intervals $\{(-\infty, q) : q \in \mathbb{Q}\}$. We can then choose $P(X \leq q|\mathcal{G})$ for $q \in \mathbb{Q}$ to be increasing and take on values of 0 and 1 at $-\infty$ and ∞ respectively. Uniquely extend this increasing function defined on $\mathbb{Q}$ to all of $\mathbb{R}$ in a right continuous manner by setting

$$P(X \leq r|\mathcal{G}) = \lim_{q \downarrow r} \mathbb{P}(x \leq q|\mathcal{G})$$

for any almost every ω. ■

Corollary 10.12 *For every joint distribution (X, Y) where Y's range is a nice space, say $(X, Y) \in \mathbb{R}^2$ then*

$$P(X \in dx, Y \in dy) = Q(x, dy) P(X \in dx)$$

for some Markov kernel Q.

It is important to note that while even when both Q_Y and Q_X exist so that

$$P(X \in dx, Y \in dy) = Q_X(y, dx) P(Y \in dy) = Q_Y(x, dy) P(X \in dx)$$

there is no general way to go from Q_X and $P(Y \in dy)$ to Q_Y unless we restrict ourselves to the case where X and Y have well defined densities.

10.5 A Word About $\mathbb{E}(Y|X = x)$

Suppose that $\mathbb{P}(X \in [a, b]) > 0$ then using the naive definition of conditional expectations we have

$$\mathbb{E}(Y|X \in [a, b]) = \frac{\mathbb{E}(Y \mathbf{1}_{(X \in [a,b])})}{\mathbb{P}(X \in [a, b])}$$

and we hope that this will give meaning to $\mathbb{E}(Y|X = x)$ in the context

$$\mathbb{E}(Y|X \in [a, b]) = \int_a^b \frac{\mathbb{E}(Y|X = x)}{\mathbb{P}(X \in [a, b])} dP(X \in dx)$$

Using our new definition of conditional expectation we have

$$\frac{\mathbb{E}(\mathbb{E}(X|Y) \mathbf{1}_{(X \in [a,b])}}{\mathbb{P}(X \in [a, b])} = \frac{\mathbb{E}(Y \mathbf{1}_{(X \in [a,b])})}{\mathbb{P}(X \in [a, b])}$$

which gives us

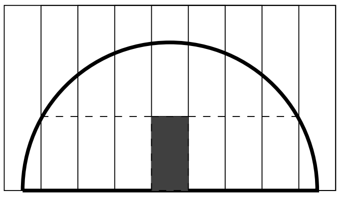
$$\mathbb{E}(Y \mathbf{1}_{(X \in [a,b])}) = \int_a^b \mathbb{E}(Y|X = x) P(X \in dx)$$

This is enough to define conditional expectations since the class of intervals $[a, b]$ is rich enough to extend the formula to each Borel set B so that

$$\mathbb{E}(Y\mathbf{1}_{(X\in B)}) = \int_B E(Y|X = x)P(X \in dx)$$

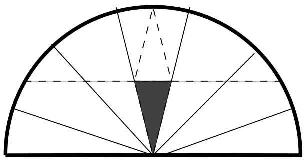
However, it is important not to attribute too much meaning to the notation $\mathbb{E}(A|X = x)$ since it is usually the case that $\mathbb{P}(X = x) = 0$ and so different versions of the conditional expectation may not agree.

This is highlighted by the following simple version of Borel's paradox:



Let (X, Y) be uniformly chosen on the half disc so that $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$ with $0 < R \leq 1$ and $\Theta \in [0, \pi]$. We should certainly believe the set equivalence

$$\{X = 0\} \iff \{\Theta = \frac{\pi}{2}\}$$



Now $P(Y > \frac{1}{2}|X = 0) = \frac{1}{2}$ has real meaning as there is a version of $\mathbb{P}(Y > \frac{1}{2}|X = x)$ which is continuous in X and it's value at 0 is $\frac{1}{2}$. On the other hand, there is a unique version of $P(Y > \frac{1}{2}|\Theta = \theta)$ whose value at $\theta = \frac{\pi}{2}$ is $\frac{3}{4}$. Slicing up a space in different ways can clearly give us surprisingly incommensurate[1] null sets!

References

[1] R. Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[2] W. Rudin. *Real and complex analysis.* McGraw-Hill Book Co., New York, third edition, 1987.

[1]From Webster's Revised Unabridged Dictionary (1913): Commensurate \ ke-'men(ts)-ret \, a. 1. Having a common measure; commensurable; reducible to a common measure; as, commensurate quantities. 2. Equal in measure or extent; proportionate.