Lecture 19 : Conditional Probability & Expectation

STAT205 *Lecturer: Jim Pitman* Scribe: *Saurabh Amin* <amins@berkeley.edu>

This set of notes is a revision of the work of Charles C. Fowlkes. Reference: [1], section 4.1.

19.1 Definition of Conditional Expectation

We present the definition of conditional expectation due to Kolmogorov (1933).

Definition 19.1 *Given the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, some sub-sigma field $\mathcal{G} \subset \mathcal{F}$, and a random variable $X \in \mathbf{L}^1(\mathcal{F})$ (meaning that X is $\mathcal{F}$-measurable and $\mathbb{E}|X| < \infty$), the conditional expectation of X given $\mathcal{G}$ is the (almost surely unique) random variable $\hat{X}$ such that*

i. $\hat{X} \in L^1(\mathcal{G})$ *that is,* $\hat{X}$ *is* $\mathcal{G}$*-measurable; and*

ii. $\mathbb{E}(\hat{X}\mathbf{1}_G) = \mathbb{E}(X\mathbf{1}_G)$ *for all* $G \in \mathcal{G}$*: that is,* $\hat{X}$ *integrates like* X *over all* $\mathcal{G}$*-sets.*

Recall that $\mathbb{E}(\hat{X}\mathbf{1}_G) = \int_G X d\mathbb{P}$. The random variable $\hat{X}$ is denoted by $\mathbb{E}(X|\mathcal{G})$. We now discuss the motivation behind this definition by elementary considerations. Recall the 'undergraduate' definition of conditional probability given by Bayes' Rule

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

for $\mathbb{P}(B) > 0$. Now if $G_1, G_2, \ldots$ is a partition of Ω into measurable sets, then

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap G_i) = \sum_i \mathbb{P}(A|G_i)\mathbb{P}(G_i) \tag{19.1}$$

Recall that this is a form of the law of total probability. $\mathbb{P}(\cdot|B)$ is a new probability measure on Ω which concentrates on B. We can then generalize naturally to conditional expectations because $\mathbb{P}(d\omega|B)$ can be used to integrate as any $\mathbb{P}(d\omega)$. Using this fact, Bayes' rule, and the law of total probability, we have

$$\begin{aligned}
\mathbb{E}(X|B) &= \int_\Omega X(\omega)\mathbb{P}(dw|B) \\
&= \frac{\int X(\omega)\mathbf{1}(\omega \in B)\mathbb{P}(dw)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{E}(X\mathbf{1}_B)}{\mathbb{P}(B)}
\end{aligned}$$

Note that equation (19.1) is obtained by multiplying the identity $\mathbf{1} = \sum_i \mathbf{1}_{G_i}$ on both sides by $\mathbf{1}_A$ and taking expectations. This can easily be generalized to the following by multiplying the identity on both sides by X and taking expectations.

$$\mathbb{E}(X) = \sum_i \mathbb{E}(X\mathbf{1}_{G_i}) = \sum_i \mathbb{E}(X|G_i)\mathbb{P}(G_i) \tag{19.2}$$

This will be true provided $\mathbb{E}|X| < \infty$. A variation of equation (19.2) can be obtained as follows. Let G be any union of the G_i, i.e., $G \in \mathcal{G}$ where $\mathcal{G} = \sigma(G_i, i = 1, 2, , \ldots)$. Again, multiplying the identity $\mathbf{1}_G = \sum_{i:G_i \subset G} \mathbf{1}_{G_i}$ on both sides by X and taking expectations we obtain the following:

$$\begin{aligned}
\mathbb{E}(X\mathbf{1}_G) &= \sum_{i:G_i \subset G} \mathbb{E}(X\mathbf{1}_{G_i}) \\
&= \sum_{i:G_i \subset G} \mathbb{E}(X|G_i)\mathbb{P}(G_i) \qquad (19.3)
\end{aligned}$$

The R.H.S. of equation (19.3) can interpreted as the expectation of a random variable $\hat{X} = \sum_{i:G_i \subset G} \mathbb{E}(X|G_i)\mathbf{1}_{G_i}$, that is, $\hat{X}$ takes the value $\mathbb{E}(X|G_i)$ if G_i occurs. We have just shown that $\mathbb{E}(X\mathbf{1}_G) = \mathbb{E}(\hat{X}\mathbf{1}_G)$ for every G which is a union of the G_is' (refer to condition *(ii)* of definition 19.1). Obviously, since $\hat{X}$ is measurable w.r.t. $\mathcal{G}$, condition *(i)* of definition 19.1 is satisfied. So we have just constructed $\hat{X} = \mathbb{E}(X|\mathcal{G})$ explicitly in the case when $\mathcal{G} = \sigma(G_1, G_2, \ldots)$.

19.2 Existence and Uniqueness of Conditional Expectation

Proposition 19.2 $\mathbb{E}(X|\mathcal{G})$ *is unique up to almost sure equivalence.*

Proof Sketch: Suppose that two random variables $\hat{X}_1$ and $\hat{X}_2$ are candidates for the conditional expectation $\mathbb{E}(X|\mathcal{G})$. Let $Y := \hat{X}_1 - \hat{X}_2$. So we have $Y \in L^1(\mathcal{G})$ and $\mathbb{E}(Y\mathbf{1}_G) = 0 \ \forall G \in \mathcal{G}$. In particular, choose $G = \{Y > \epsilon\}$ and so we have

$\mathbb{E}(Y\mathbf{1}(Y > \epsilon)) = 0$. By Markov's inequality, $\mathbb{P}(Y > \epsilon) \leq \mathbb{E}(Y\mathbf{1}(Y > \epsilon))/\epsilon = 0$. Interchanging the roles of X_1 and X_2, we have $\mathbb{P}(Y < -\epsilon) = 0$. And since ϵ is arbitrary, $\mathbb{P}(Y = 0) = 1$. ∎

Proposition 19.3 $\mathbb{E}(X|\mathcal{G})$ *exists.*

We give three different approaches for dealing with the general case.

19.2.1 Measure Theory Proof

Here we pull out some power tools from measure theory.

Theorem 19.4 (Lebesgue-Radon-Nikodym) *(see [1], p.$\tilde{4}$77) 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 $G \in \mathcal{G}$. If $\hat{X}$ is another such function then $\hat{X} = \hat{Y}$ μ a.e.

Proof Sketch: (existence via Lebesgue-Radon-Nikodym) Assume $Y \geq 0$ and define the measure

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

which is non-negative and finite because $\mathbb{E}|Y| < \infty$. Note that 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 $\mathbb{E}(Y^+|\mathcal{G}) - \mathbb{E}(Y^-|\mathcal{G})$. ∎

19.2.2 Hilbert Space Method

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

Lemma 19.5 *Every nonempty, closed, convex set E in a Hilbert space H contains a unique element of smallest norm.*

Lemma 19.6 (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 Rudin 87 (p.$\tilde{7}$9) for a full discussion of Lemma 19.5.

Proof Sketch: (Existence via Hilbert Space Projection) Suppose $Y \in \mathbf{L}^2(\mathcal{F})$ and $X \in \mathbf{L}^2(\mathcal{G})$. Requirement *(ii)* 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 *(i)* 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. ■

19.2.3 "Hands On" Proof

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

Lemma 19.7 (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 $G \in \mathcal{G} = \sigma(\{G_i\})$ we have shown that

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

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 $\{B_i = (X \le r_i) : r \in \mathbb{Q}\}$. 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 serve as a version of $\mathbb{E}(Y|\mathcal{G})$. We will assume that $\mathbb{E}|Y|^2 < \infty$

Write $Y_n = \mathbb{E}(Y|\mathcal{G}_n) = Y_1 + (Y_2 - Y_1) + (Y_3 - Y_2) + \ldots + (Y_n - Y_{n-1})$. 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 $s^2 = \mathbb{E}(Y^2) = \mathbb{E}(Y_n + (Y - Y_n)) < \infty$. Then $s_n^2 \uparrow s_\infty^2 \leq s^2 < \infty$. For $n > m$ we know again by orthogonality that $\mathbb{E}((Y_n - Y_m)^2) = s_n^2 - s_m^2 \to 0$ 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 *(i)* since as a limit of $\mathcal{G}$-measurable variables it is $\mathcal{G}$-measurable. To check *(ii)* we need to show that $\mathbb{E}(YG) = \mathbb{E}(Y_\infty G)$ 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 $\mathcal{A} = \cup_m \mathcal{G}_m$.

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

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 the 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 19.8 *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 19.9 *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.)*

References

[1] Richard Durrett. *Probability: theory and examples, 3rd edition.* Thomson Brooks/Cole, 2005.