CONDITIONAL EXPECTATION

STEVEN P. LALLEY

1. CONDITIONAL EXPECTATION: L^2–THEORY

Definition 1. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\mathscr{G}$ be a σ–algebra contained in $\mathscr{F}$. For any real random variable $X \in L^2(\Omega, \mathscr{F}, P)$, define $E(X|\mathscr{G})$ to be the *orthogonal projection* of X onto the closed subspace $L^2(\Omega, \mathscr{G}, P)$.

This definition may seem a bit strange at first, as it seems not to have any connection with the naive definition of conditional probability that you may have learned in elementary probability. However, there is a compelling rationale for Definition 1: the orthogonal projection $E(X|\mathscr{G})$ minimizes the expected squared difference $E(X-Y)^2$ among all random variables $Y \in L^2(\Omega, \mathscr{G}, P)$, so in a sense it is the *best predictor* of X based on the information in $\mathscr{G}$. It may be helpful to consider the special case where the σ–algebra $\mathscr{G}$ is generated by a single random variable Y, i.e., $\mathscr{G} = \sigma(Y)$. In this case, every $\mathscr{G}$–measurable random variable is a Borel function of Y (exercise!), so $E(X|\mathscr{G})$ is the unique Borel function $h(Y)$ (up to sets of probability zero) that minimizes $E(X-h(Y))^2$. The following exercise indicates that the special case where $\mathscr{G} = \sigma(Y)$ for some real-valued random variable Y is in fact very general.

Exercise 1. Show that if $\mathscr{G}$ is *countably generated* (that is, there is some countable collection of set $B_j \in \mathscr{G}$ such that $\mathscr{G}$ is the *smallest* σ–algebra containing all of the sets B_j) then there is a $\mathscr{G}$–measurable real random variable Y such that $\mathscr{G} = \sigma(Y)$.

The following exercise shows that in the special case where the σ–algebra $\mathscr{G}$ is finite, Definition 1 is equivalent to the naive definition of conditional expectation.

Exercise 2. Suppose that the σ–algebra $\mathscr{G}$ is finite, that is, suppose that there is a finite measurable partition $B_1, B_2, \ldots, B_n$ of Ω such that $\mathscr{G}$ is the σ–algebra generated by the sets B_i. Show that for any $X \in L^2(\Omega, \mathscr{F}, P)$,

$$E(X|\mathscr{G}) = \sum_{i=1}^{n} \frac{E(X\mathbf{1}_{B_i})}{P(B_i)} \mathbf{1}_{B_i} \quad a.s.$$

Because conditional expectation is defined by orthogonal projection, all of the elementary properties of orthogonal projection operators translate to corresponding properties of conditional expectation. Following is a list of some of the more important of these.

Properties of Conditional Expectation: Let $X \in L^2(\Omega, \mathscr{F}, P)$ and let $\mathscr{G}$ be a σ–algebra contained in $\mathscr{F}$. Then

(0) **Linearity:** $E(aX_1 + bX_2|\mathscr{G}) = aE(X_1|\mathscr{G}) + bE(X_2|\mathscr{G})$.
(1) **Orthogonality:** $X - E(X|\mathscr{G}) \perp L^2(\Omega, \mathscr{G}, P)$.

(2) **Best Prediction:** $E(X\,|\,\mathscr{G})$ minimizes $E(X-Y)^2$ among all $Y \in L^2(\Omega,\mathscr{G},P)$.

(3) **Tower Property:** If $\mathscr{H}$ is a $\sigma-$algebra contained in $\mathscr{G}$, so that $\mathscr{H} \subseteq \mathscr{G} \subseteq \mathscr{F}$, then

$$E(X\,|\,\mathscr{H}) = E(E(X\,|\,\mathscr{G})\,|\,\mathscr{H}).$$

(4) **Covariance Matching:** $E(X\,|\,\mathscr{G})$ is the unique random variable $Z \in L^2(\Omega,\mathscr{G},P)$ such that for every $Y \in L^2(\Omega,\mathscr{G},P)$,

$$E(XY) = E(ZY).$$

Property (4) is just a re-statement of the orthogonality law. It is usually taken to be the *definition* of conditional expectation (as in Billingsley). Observe that for the equation in (4) to hold for *all* $Y \in L^2(\Omega,\mathscr{G},P)$ it is necessary that Z be square-integrable (because the equation must hold for $Y = Z$). That there is only one such random variable Z (up to change on events of probability 0) follows by an easy argument using indicators of events $B \in \mathscr{G}$: if there were two $\mathscr{G}-$measurable random variables Z_1, Z_2 such that the equation in (4) were valid for both $Z = Z_1$ and $Z = Z_2$, and all Y, then for every $B \in \mathscr{G}$,

$$E(Z_1 - Z_2)\mathbb{1}_B = 0.$$

But any $\mathscr{G}-$measurable random variable that integrates to 0 on *every* event $B \in \mathscr{G}$ must equal 0 a.s. (why?).

2. Conditional Expectation: L^1-Theory

The major drawback of Definition 1 is that it applies only to square-integrable random variables. Thus, our next objective will be to extend the definition and basic properties of conditional expectation to all *integrable* random variables. Since an integrable random variable X need not be square-integrable, its conditional expectation $E(X|\mathscr{G})$ on a $\sigma-$algebra $\mathscr{G}$ cannot be defined by orthogonal projection. Instead, we will use the covariance property (4) as a basis for a general definition.

Definition 2. Let $(\Omega,\mathscr{F},P)$ be a probability space and let $\mathscr{G}$ be a $\sigma-$algebra contained in $\mathscr{F}$. For any real random variable $X \in L^1(\Omega,\mathscr{F},P)$, define $E(X\,|\,\mathscr{G})$ to be the unique random variable $Z \in L^1(\Omega,\mathscr{G},P)$ such that for every *bounded*, $\mathscr{G}-$measurable random variable Y,

$$E(XY) = E(ZY). \tag{5}$$

In particular, equation (5) must hold for every indicator $Y = \mathbf{1}_G$, where $G \in \mathscr{G}$. We have already seen that there can be at most one such random variable $Z \in L^1(\Omega,\mathscr{G},P)$; thus, to verify that Definition 2 is a valid definition, we must prove that there is *at least one* random variable $Z \in L^1(\Omega,\mathscr{G},P)$ satisfying equation (5).

Proposition 1. *For every $X \in L^1(\Omega,\mathscr{F},P)$ there exists $Z \in L^1(\Omega,\mathscr{G},P)$ such that equation* (5) *holds for all bounded , $\mathscr{G}-$measurable random variables Y.*

First Proof. There is a short, easy proof using the Radon-Nikodym theorem. It is enough to consider the case where X is nonnegative, because in general an integrable random variable can be split into positive and negative parts. Assume, then, that $X \geq 0$. Define a finite, positive measure ν on $(\Omega,\mathscr{G})$ by

$$\nu(G) = EX\mathbf{1}_G.$$

It is easily checked that $\nu \ll P$ on $\mathscr{G}$. Therefore, the Radon-Nikodym theorem implies that there exists a nonnegative, $\mathscr{G}-$measurable random variable Z such that for every bounded, $\mathscr{G}-$measurable Y,

$$\int Y\, d\nu = E(ZY).$$

□

Although it is short and elegant, the preceding proof relies on a deep theorem, the Radon-Nikodym theorem. In fact, the use of the Radon-Nikodym theorem is superfluous; the fact that every L^1 random variable can be arbitrarily approximated by L^2 random variables makes it possible to construct a solution to (5) by approximation. For this, we need several more properties of the conditional expectation operator on L^2.

(6) **Normalization:** $E(1\,|\,\mathscr{G}) = 1$ almost surely.

(7) **Positivity:** For any nonnegative, bounded random variable X,

$$E(X\,|\,\mathscr{G}) \geq 0 \quad \text{almost surely.}$$

(8) **Monotonicity:** If X, Y are bounded random variables such that $X \leq Y$ a.s., then

$$E(X\,|\,\mathscr{G}) \leq E(Y\,\mathscr{G}) \quad \text{almost surely.}$$

The normalization property (6) is almost trivial: it holds because any constant random variable c is measurable with respect to *any* $\sigma-$algebra, and in particular $\mathscr{G}$; and any random variable $Y \in L^2(\Omega, \mathscr{G}, P)$ is its own projection. The positivity property (7) can be easily deduced from the covariance matching property (4): since $X \geq 0$ a.s., for any event $G \in \mathscr{G}$,

$$E(X\mathbf{1}_G) = E(E(X\,|\,\mathscr{G})\mathbf{1}_G) \geq 0;$$

consequently, $E(X|\mathscr{G})$ must be nonnegative almost surely, because its integral on any event is nonnegative. The monotonicity property (8) follows directly from *linearity* and *positivity*.

Second Proof of Proposition 1. First, observe again that it suffices to consider the case where X is nonnegative. Next, recall that $X = \lim \uparrow X \wedge n$. Each of the random variables $X \wedge n$ is bounded, hence in L^2, and so its conditional expectation is well-defined by orthogonal projection. Moreover, by the positivity and monotonicity laws (7) – (8), the conditional expectations $E(X \wedge n\,|\,\mathscr{G})$ are nonnegative and non-decreasing with n. Consequently,

$$Z := \lim \uparrow E(X \wedge n\,|\,\mathscr{G})$$

exists and is $\mathscr{G}-$measurable. Now by the Monotone Convergence Theorem, for any bounded, nonnegative, $\mathscr{G}-$measurable random variable Y,

$$\begin{aligned} E(XY) &= \lim \uparrow E((X \wedge n)Y) \\ &= \lim \uparrow E(E(X \wedge n\,|\,\mathscr{G})Y) \\ &= E(ZY). \end{aligned}$$

This proves equation (5) for *nonnegative* Y; it then follows for arbitrary bounded Y by linearity, using $Y = Y_+ - Y_-$. □

2.1. **Properties of Conditional Expectation.** Henceforth we shall take Definition 2 to be the definition of conditional expectation. By the covariance matching property (4), this definition agrees with Definition 1 for $X \in L^2$. Given Definition 2, the following properties are all easily established. (Exercise: Check any that you do not find immediately obvious.)

(1) **Definition:** $EXY = EE(X|\mathcal{G})Y$ for all bounded, $\mathcal{G}-$measurable random variables Y.
(2) **Linearity:** $E(aU + bV|\mathcal{G}) = aE(U|\mathcal{G}) + bE(V|\mathcal{G})$ for all scalars $a, b \in \mathbb{R}$.
(3) **Positivity:** If $X \geq 0$ then $E(X|\mathcal{G}) \geq 0$.
(4) **Stability:** If X is $\mathcal{G}-$measurable, then $E(XZ|Y) = XE(Z|Y)$.
(5) **Independence Law:** If X is independent of $\mathcal{G}$ then $E(X|\mathcal{G}) = EX$ is constant a.s.
(6) **Tower Property:** If $\mathcal{H} \subseteq \mathcal{G}$ then $E(E(X|\mathcal{G})|\mathcal{H}) = E(X|\mathcal{H}$.
(7) **Expectation Law:** $E(E(X|\mathcal{G})) = EX$.
(8) **Constants:** For any scalar a, $E(a|\mathcal{G}) = a$.
(9) **Jensen Inequalities:** If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $E|X| < \infty$ then

$$E(\varphi(X)) \geq \varphi(EX) \text{ and}$$
$$E(\varphi(X)|Y) \geq \varphi(E(X|Y).$$

In all of these statements, the relations = and ≤ are meant to hold *almost surely*. Properties (3)–(7) extend easily to nonnegative random variables X with infinite expectation. Observe that, with the exceptions of the Stability, Tower, and Independence Properties, all of these correspond to basic properties of *ordinary* expectation. Later, we will see a deeper reason for this. Following is another property of conditional expectation that generalizes a corresponding property of ordinary expectation.

Proposition 2. *(Jensen Inequality) If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex and $E|X| < \infty$ then*

$$E(\varphi(X)|\mathcal{G}) \geq \varphi(E(X|\mathcal{G}). \tag{9}$$

REMARK. Since (9) holds for the trivial $\sigma-$algebra $\{\emptyset, \Omega\}$, the usual Jensen inequality $E\varphi(X) \geq \varphi(EX)$ is a consequence.

Proof of the Jensen Inequalities. One of the basic properties of convex functions is that every point on the graph of a convex function φ has a *support line*: that is, for every argument $x_* \in \mathbb{R}$ there is a linear function $y_{x_*}(x) = ax + b$ such that

$$\varphi(x_*) = y_{x_*}(x_*) \quad \text{and}$$
$$\varphi(x) \geq y_{x_*}(x) \quad \text{for all } x \in \mathbb{R}.$$

Let X be a random variable such that $E|X| < \infty$, so that the expectation EX is well-defined and finite. Let $y_{EX}(x) = ax + b$ be a support line to the convex function at the point $(EX, \varphi(EX))$. Then by definition of a support line, $y_{EX}(EX) = \varphi(EX)$; also, $y_{EX}(X) \leq \varphi(X)$, and so

$$Ey_{EX}(X) \leq E\varphi(X).$$

But because $y_{EX}(x) = ax + b$ is a linear function of x,

$$Ey_{EX}(X) = y_{EX}(EX) = \varphi(EX).$$

This proves the Jensen inequality for ordinary expectation. The proof for conditional expectation is similar. Let $y_{E(X|\mathcal{G})}(x)$ be the support line at the point $(E(X|\mathcal{G}), \varphi(E(X|\mathcal{G})))$. Then

$y_{E(X|\mathscr{G})}(E(X|\mathscr{G})) = \varphi(E(X|\mathscr{G}))$, and for every value of X, $y_{E(X|\mathscr{G})}(X) \le \varphi(X)$. Consequently, by the linearity and positivity properties of conditional expectation,

$$\begin{aligned}\varphi(E(X|\mathscr{G})) &= y_{E(X|\mathscr{G})}(E(X|\mathscr{G})) \\ &= E(y_{E(X|\mathscr{G})}(X)|\mathscr{G}) \\ &\le E(\varphi(X)|\mathscr{G}).\end{aligned}$$

□

Proposition 3. *Let $\{X_\lambda\}_{\lambda\in\Lambda}$ be a (not necessarily countable) uniformly integrable family of real random variables on $(\Omega,\mathscr{F},P)$. Then the set of all conditional expectations $\{E(X_\lambda|\mathscr{G})\}_{\lambda\in\Lambda,\mathscr{G}\subset\mathscr{F}}$, where $\mathscr{G}$ ranges over all $\sigma-$algebras contained in $\mathscr{F}$, is uniformly integrable.*

Proof. We will use the equivalence of the following two characterizations of uniform integrability: a family $\{Y_\lambda\}_{\lambda\in\Lambda}$ of nonnegative random variables is uniformly integrable if either of the following holds.

(a) For each $\varepsilon > 0$ there exists $\alpha < \infty$ such that $EY_\lambda \mathbf{1}\{Y_\lambda \ge \alpha\} < \varepsilon$.
(b) The expectations EY_λ are bounded (that is, $\sup_{\lambda\in\Lambda} EY_\lambda < \infty$), and for each $\varepsilon > 0$ there exists $\delta > 0$ such that for every event A of probability less than δ,

$$\sup_{\lambda\in\Lambda} EY_\lambda \mathbf{1}_A < \varepsilon.$$

We may assume without loss of generality in proving the proposition that all of the random variables X_λ are nonnegative. Since the family $\{X_\lambda\}_{\lambda\in\Lambda}$ is uniformly integrable, for each $\varepsilon > 0$ there exists $\delta > 0$ such that for every event A satisfying $P(A) < \delta$ and every $\lambda \in \Lambda$,

$$EX_\lambda \mathbf{1}_A < \varepsilon. \tag{10}$$

Moreover, since the set $\{X_\lambda\}_{\lambda\in\Lambda}$ is uniformly integrable, the expectations are uniformly bounded, and so the set $\{E(X_\lambda|\mathscr{G})\}_{\lambda\in\Lambda,\mathscr{G}\subset\mathscr{F}}$ of all conditional expectations is bounded in L^1. Hence, by the Markov inequality,

$$\lim_{\alpha\to\infty} \sup_{\lambda\in\Lambda} \sup_{\mathscr{G}\subset\mathscr{F}} P\{E(X_\lambda\,|\,\mathscr{G}) \ge \alpha\} = 0.$$

To show that the set of all conditional expectations $\{E(X_\lambda|\mathscr{G})\}_{\lambda\in\Lambda,\mathscr{G}\subset\mathscr{F}}$ is uniformly integrable it suffices, by criterion (a), to prove that for every $\varepsilon > 0$ there exists a constant $0 < \alpha < \infty$ such that for each $\lambda \in \Lambda$ and each $\sigma-$algebra $\mathscr{G} \subset \mathscr{F}$,

$$EE(X_\lambda\,|\,\mathscr{G})\mathbf{1}_{\{E(X_\lambda\,|\,\mathscr{G})>\alpha\}} < \varepsilon. \tag{11}$$

But by the definition of conditional expectation,

$$EE(X_\lambda\,|\,\mathscr{G})\mathbf{1}_{\{E(X_\lambda\,|\,\mathscr{G})>\alpha\}} = EX_\lambda \mathbf{1}_{\{E(X_\lambda\,|\,\mathscr{G})>\alpha\}}.$$

If $\alpha < \infty$ is sufficiently large then by the preceding paragraph $P\{E(X_\lambda\,|\,\mathscr{G}) > \alpha\} < \delta$, where $\delta > 0$ is so small that the inequality (10) holds for all events A with probability less than δ. The desired inequality (11) now follows. □

3. Convergence Theorems for Conditional Expectation

Just as for ordinary expectations, there are versions of Fatou's lemma and the monotone and dominated convergence theorems.

Monotone Convergence Theorem . *Let X_n be a nondecreasing sequence of nonnegative random variables on a probability space $(\Omega, \mathscr{F}, P)$, and let $X = \lim_{n\to\infty} X_n$. Then for any $\sigma-$algebra $\mathscr{G} \subset \mathscr{F}$,*

$$E(X_n|\mathscr{G}) \uparrow E(X|\mathscr{G}). \tag{12}$$

Fatou's Lemma . *Let X_n be a sequence of nonnegative random variables on a probability space $(\Omega, \mathscr{F}, P)$, and let $X = \liminf_{n\to\infty} X_n$. Then for any $\sigma-$algebra $\mathscr{G} \subset \mathscr{F}$,*

$$E(X|\mathscr{G}) \le \liminf E(X_n|\mathscr{G}). \tag{13}$$

Dominated Convergence Theorem . *Let X_n be a sequence of real-valued random variables on a probability space $(\Omega, \mathscr{F}, P)$ such that for some integrable random variable Y and all $n \ge 1$,*

$$|X_n| \le Y. \tag{14}$$

Then for any $\sigma-$algebra $\mathscr{G} \subset \mathscr{F}$,

$$\lim_{n\to\infty} E(X_n|\mathscr{G}) = E(X|\mathscr{G}) \quad \textit{and} \quad \lim_{n\to\infty} E(|X_n - X||\mathscr{G}) = 0 \tag{15}$$

As in Properties 1–9 above, the limiting equalites and inequalities in these statements hold *almost surely*. The proofs are easy, given the corresponding theorems for *ordinary* expectations; I'll give the proof for the Monotone Convergence Theorem and leave the other two, which are easier, as exercises.

Proof of the Monotone Convergence Theorem. This is essentially the same argument as used in the second proof of Proposition 1 above. By the Positivity and Linearity properties of conditional expectation,

$$E(X_n|\mathscr{G}) \le E(X_{n+1}|\mathscr{G}) \le E(X|\mathscr{G})$$

for every n. Consequently, the limit $V := \lim_{n\to\infty} \uparrow E(X_n|\mathscr{G})$ exists with probability one, and $V \le E(X|\mathscr{G})$. Moreover, since each conditional expectation is $\mathscr{G}$-measurable, so is V. Set $B = \{V < E(X|\mathscr{G})\}$; we must show that $P(B) = 0$. Now $B \in \mathscr{G}$, so by definition of conditional expectation,

$$E(X\mathbf{1}_B) = E(E(X|\mathscr{G})\mathbf{1}_B) \quad \text{and} \quad E(X_n\mathbf{1}_B) = E(E(X_n|\mathscr{G})\mathbf{1}_B).$$

But the Monotone Convergence Theorem for ordinary expectation implies that

$$\begin{aligned} EX\mathbf{1}_B &= \lim_{n\to\infty} EX_n\mathbf{1}_B \quad \text{and} \\ EV\mathbf{1}_B &= \lim_{n\to\infty} EE(X_n|\mathscr{G})\mathbf{1}_B, \end{aligned}$$

so $EX\mathbf{1}_B = EV\mathbf{1}_B$. Since $V < X$ on B, this implies that $P(B) = 0$. □

4. Regular Conditional Distributions

Let X be a real random variable defined on a probability space $(\Omega, \mathscr{F}, P)$ and let $\mathscr{G}$ be a $\sigma-$algebra contained in $\mathscr{F}$. For each Borel set $B \subset \mathbb{R}$ define

$$P(X \in B \,|\, \mathscr{G}) = E(\mathbf{1}_{\{X \in B\}} \,|\, \mathscr{G}) \tag{16}$$

to be the *conditional probability* of the event $\{X \in B\}$ given $\mathscr{G}$. Observe that the conditional probability $P(X \in B \,|\, \mathscr{G})$ is a *random variable*, not a scalar. Nevertheless, conditional probability, like unconditional probability, is countably additive, in the sense that for any sequence B_n of pairwise disjoint Borel sets,

$$P(\cup_{n\geq 1} B_n \,|\, \mathscr{G}) = \sum_{n\geq 1} P(B_n \,|\, \mathscr{G}) \quad \text{a.s.} \tag{17}$$

(This follows by the linearity of conditional expectation and the monotone convergence theorem, as you should check.) Unfortunately, there are *uncountably* many sequences of Borel sets, so we cannot be sure *a priori* that the null sets on which (17) fails do not add up to an event of positive probability. The purpose of this section is to show that in fact they do not.

Definition 3. Let $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ and $(\mathscr{Y}, \mathscr{F}_{\mathscr{Y}})$ be measurable spaces. A *Markov kernel* (or *Markov transition kernel*) from $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ to $(\mathscr{Y}, \mathscr{F}_{\mathscr{Y}})$ is a family $\{\mu_x(dy)\}_{x\in\mathbb{X}}$ of probability measures on $(\mathscr{Y}, \mathscr{F}_{\mathscr{Y}})$ such that for each event $F \in \mathscr{F}_{\mathscr{Y}}$ the random variable $x \mapsto \mu_x(F)$ is $\mathscr{F}_{\mathscr{X}}-$measurable.

Definition 4. Let X be a random variable defined on a probability space $(\Omega, \mathscr{F}, P)$ that takes values in a measurable space $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$, and let $\mathscr{G}$ be a $\sigma-$algebra contained in $\mathscr{F}$. A *regular conditional distribution* for X given $\mathscr{G}$ is a Markov kernel μ_ω from $(\Omega, \mathscr{G})$ to $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ such that for every set $F \in \mathscr{F}_{\mathscr{X}}$,

$$\mu_\omega(F) = P(X \in F \,|\, \mathscr{G})(\omega) \quad \text{a.s.} \tag{18}$$

Theorem 1. *If X is a real random variable defined on a probability space $(\Omega, \mathscr{F}, P)$ then for every $\sigma-$algebra $\mathscr{G} \subset \mathscr{F}$ there is a regular conditional distribution for X given $\mathscr{G}$.*

The proof will use the *quantile transform* method of constructing a random variable with a specified cumulative distribution function F from a uniform-[0,1] random variable. Given a c.d.f. F, define

$$F^-(t) = \inf\{x : F(x) \geq t\} \quad \text{for } t \in (0,1). \tag{19}$$

Lemma 1. *If U has the uniform distribution on $[0,1]$ then $F^-(U)$ has cumulative distribution function F, that is, for every $x \in \mathbb{R}$,*

$$P(F^-(U) \leq x) = F(x).$$

Proof. Exercise. □

Corollary 1. *For any cumulative distribution function F on $\mathbb{R}$ (that is, any right-continuous, nondecreasing function satisfying $F \to 0$ at $-\infty$ and $F \to 1$ at $+\infty$) there is a Borel probability measure μ_F on $\mathbb{R}$ such that for every $x \in \mathbb{R}$,*

$$\mu_F(-\infty, x] = F(x). \tag{20}$$

Proof. This follows directly from the following more general fact: if $(\Omega, \mathscr{F}, P)$ is any probability space and if $X : \Omega \to \mathscr{X}$ is a measurable transformation taking values in $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ then $\mu = P \circ X^{-1}$ is a probability measure on $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$. □

Thus, the quantile transform together with the existence of Lebesgue measure on $[0,1]$ implies the existence of a measure satisfying (20). In proving Theorem 1 we shall use this to reduce the problem of constructing probability measures $\mu_\omega(dx)$ to the simpler problem of constructing c.d.f.s $F_\omega(x)$ such that

$$F_\omega(x) = P(X \le x \,|\, \mathscr{G}) \quad \text{a.s.} \tag{21}$$

Proof of Theorem 1. For each *rational* x define $F_\omega(x) = P(X \le x \,|\, \mathscr{G})(\omega)$ to be some version of the conditional probability. Consider the following events:

$$\begin{aligned}
B_1 &= \{F_\omega \text{not monotone on } \mathbb{Q}\};\\
B_2 &= \{\exists\, x \in \mathbb{Q} : F_\omega(x) \neq \lim_{y \to x+} F_\omega(y)\};\\
B_3 &= \{\lim_{n\to\infty} F_\omega(n) \neq 1\};\\
B_4 &= \{\lim_{n\to\infty} F_\omega(-n) \neq 0\};\\
B &= B_1 \cup B_2 \cup B_3 \cup B_4.
\end{aligned}$$

(In B_2, the limit is through the set of rationals, and in B_3 and B_4 the limit is through the positive integers n.) These are all events of probability 0. (Exercise: why?)

On the event B *redefine* $F_\omega(x) = \Phi(x)$, where Φ is the standard normal c.d.f. On B^c, and on B^c extend F_ω to all real arguments by setting

$$F_\omega(x) = \inf_{y \in \mathbb{Q},\, y > x} F_\omega(x).$$

Note that for *rational* x this definition coincides with the original choice of $F_\omega(x)$ (because the event $B^c \subset B_2^c$). Then $F_\omega(x)$ is, for every $\omega \in \Omega$, a cumulative distribution function, and so there exists a Borel probability measure $\mu_\omega(dx)$ on $\mathbb{R}$ with c.d.f. F_ω. The proof will be completed by establishing the following two claims.

Claim 1: $\{\mu_\omega(dx)\}_{\omega\in\Omega}$ is a Markov kernel.
Claim 2: $\{\mu_\omega(dx)\}_{\omega\in\Omega}$ is a regular conditional distribution for X given $\mathscr{G}$.

Proof of Claim 1. Each μ_ω is by construction a Borel probability measure, so what must be proved is that for each Borel set B the mapping $\omega \mapsto \mu_\omega(B)$ is $\mathscr{G}-$measurable. This is certainly true for each interval $(-\infty, x]$ with $x \in \mathbb{Q}$, since $F_\omega(x)$ is a version of $P(X \le x \,|\, \mathscr{G})$, which by definition is $\mathscr{G}-$measurable. Moreover, by countable additivity, the set of Borel sets B such that $\omega \mapsto \mu_\omega(B)$ is $\mathscr{G}-$measurable is a $\lambda-$system (why?). Since the set of intervals $(-\infty, x]$ with $x \in \mathbb{Q}$ is a $\pi-$system, the claim follows from the $\pi - \lambda$ theorem. □

Proof of Claim 2. To show that $\{\mu_\omega(dx)\}_{\omega\in\Omega}$ is a regular conditional distribution for X, we must show that equation (18) holds for every Borel set F. By construction, it holds for every $F = (-\infty, x]$ with x rational, and it is easily checked that the collection of all F for which (18) holds is a $\lambda-$system (why?). Therefore, the claim follows from the $\pi - \lambda$ theorem. □

□

Regular conditional distributions are useful in part because they allow one to reduce many problems concerning conditional expectations to problems concerning only *ordinary* expectations. For such applications the following *disintegration formula* for conditional expectations is essential.

Theorem 2. *Let $\mu_\omega(dx)$ be a regular conditional distribution for X given $\mathscr{G}$, let Y be $\mathscr{G}-$measurable, and let $f(x,y)$ be a jointly measurable real-valued function such that $E|f(X,Y)|<\infty$. Then*

$$E(f(X,Y)\,|\,\mathscr{G}) = \int f(x,Y(\omega))\,\mu_\omega(dx) \quad a.s. \tag{22}$$

Proof. By the usual arguments (linearity plus approximation by simple functions plus monotone convergence) it suffices to prove this in the special case where $f = \mathbf{1}_F$ is an indicator function. By the usual $\pi-\lambda$ argument, it suffices to consider sets F of the form $F = A\times B$, so $f(X,Y) = \mathbf{1}_A(X)\mathbf{1}_B(Y)$. Thus, our task is to show that

$$E(\mathbf{1}_A(X)\mathbf{1}_B(Y)\,|\,\mathscr{G}) = \int \mathbf{1}_A(x)\mathbf{1}_B(Y(\omega))\,\mu_\omega(dx).$$

Since the indicator $\mathbf{1}_B(Y)$ factors out on both sides (by the stability property of conditional expectation – see section 2.1 above) it is enough to show that for any measurable set A,

$$E(\mathbf{1}_A(X)\,|\,\mathscr{G})(\omega) = \int \mathbf{1}_A(x)\,\mu_\omega(dx).$$

But this follows directly from the definition of regular conditional distribution. □

Exercise 3. Let X and Y be real random variables defined on a probability space $(\Omega,\mathscr{F},P)$, and let $\mathscr{G} = \sigma(Y)$ be the $\sigma-$algebra generated by Y. Show that there is a Markov kernel $\mu_y(dx)$ from $(\mathbb{R},\mathscr{B})$ to $(\mathbb{R},\mathscr{B})$ such that a regular conditional distribution for X given $\mathscr{G}$ is given by

$$\mu_Y(B) = P(X\in B\,|\,\mathscr{G}) \quad \text{a.s.}$$

□

As an example of the use of regular conditional distributions, consider the Jensen inequality (9)

$$E(\varphi(X)|\mathscr{G}) \geq \varphi(E(X|\mathscr{G}).$$

The Jensen inequality for ordinary expectations is easy: see the proof of Proposition 2 above. Now let $\mu_\omega(dx)$ be a regular conditional distribution for X given $\mathscr{G}$. Then

$$\begin{aligned} E(\varphi(X)\,|\,\mathscr{G})(\omega) &= \int \varphi(x)\,\mu_\omega(dx) \quad \text{a.s.} \\ &\geq \varphi\left(\int x\,\mu_\omega(dx)\right) \\ &= \varphi(E(X\,|\,\mathscr{G})) \quad \text{a.s.} \end{aligned}$$

The inequality in the middle line follows from the Jensen inequality for ordinary expectations, since each μ_ω is a Borel probability.

5. Borel Spaces

It is not unusual – especially in the theory of Markov processes, and also in decision theory – that a random variable X of interest will take its values not in $\mathbb{R}$ but in some other space of objects. These might be *paths*, or *configurations*, or (in decision problems) *actions*. The argument given in section 4 to establish the existence of regular conditional distributions used in an essential way the hypothesis that the random variable X was real-valued. Are there regular conditional distributions for random variables that do not take real values? In this section we will show that if X takes values in a *Borel space* then rcds always exist.

Definition 5. A *Borel space*[1] is a measurable space $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ such that $\mathscr{F}_{\mathscr{X}}$ is *countably generated*, that is, such that for some countable collection $\{F_n\}_{n\geq 1}$,

$$\mathscr{F}_{\mathscr{X}} = \sigma(\{F_n\}_{n\geq 1}) \tag{23}$$

is the minimal $\sigma-$algebra containing each set F_n.

Example 1. The space $(\mathbb{R}^d, \mathscr{B}_d)$, where $\mathscr{B}_d$ is the $\sigma-$algebra of $d-$dimensional Borel sets, is a Borel space, because it is generated by the rectangles $\times_{i\leq d}(a_i, b_i]$.

Example 2. If $(\mathscr{X}_n, \mathscr{F}_n)$ is a sequence of Borel spaces then the Cartesian product $(\times_{n=1}^{\infty}\mathscr{X}_n, \times_{n=1}^{\infty}\mathscr{F}_n)$ is a Borel space. Here $\times_{n=1}^{\infty}\mathscr{F}_n$ is the product $\sigma-$algebra, that is, the minimal $\sigma-$algebra containing all *finite-dimensional* cylinder sets. (A generating set is the collection of all cylinders $F_1 \times F_2 \times \cdots \times F_n$, where F_i is taken from the countable set of generators of $\mathscr{F}_n$.)

Thus, in particular, the sequence space $\mathbb{R}^{\infty}$ with the usual Borel field is a Borel space. This is perhaps the most important example in probability theory.

Example 3. A *Polish space* is a complete, separable metric space, together with its Borel field. (The Borel field of a metric space is the minimal $\sigma-$algebra containing all open balls of rational radii.) Every Polish space is a Borel space. (Exercise!)

A particular case of interest is the space $C[0,1]$ of continuous paths parametrized by time $t \in [0,1]$. The metric on $C[0,1]$ is induced by the supremum norm. That $C[0,1]$ is separable in the sup norm metric follows because the set of *polygonal paths* with rational endpoints is dense.

Proposition 4. *If $(\mathscr{X}, \mathscr{F})$ is a Borel space then there is a real-valued random variable $Y : \mathscr{X} \to \mathbb{R}$ such that $\mathscr{F} = \sigma(Y)$.*

Proof. In brief, the idea is *coding by indicators*. If $\mathscr{F}$ is countably generated then it is generated by a countable collection $\{F_n\}_{n\geq 1}$. Use the indicator functions of these sets F_n to construct a mapping to the Cantor set:

$$Y = \sum_{n=1}^{\infty} \frac{2\mathbf{1}_{F_n}}{3^n}.$$

This mapping is clearly measurable, and the value of Y uniquely determines the value of each indicator $\mathbf{1}_{F_n}$, because the ternary expansion provides a bijective mapping between the Cantor

[1] The terminology varies. In some places what I have defined to be a Borel space is called a *standard* Borel space, and in some places (see, especially, the fundamental paper of Mackey on the subject) it is also required that the $\sigma-$algebra *separates points*.

set and the set of all infinite sequences of 0s and 1s. Consider the $\sigma-$algebra consisting of all events of the form $\{Y \in B\}$, where B is a one-dimensional Borel set. This collection contains the events F_n, so it must also contain every member of $\mathscr{F}$, since $\mathscr{F}$ is generated by $\{F_n\}_{n\geq 1}$. Since Y is measurable, it follows that

$$\mathscr{F} = Y^{-1}(\mathscr{B}_1).$$

□

Theorem 3. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and X a random variable on Ω taking values in a Borel space $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$. Then for any $\sigma-$algebra $\mathscr{G}$ contained in $\mathscr{F}$ there exists a regular conditional distribution for X given $\mathscr{G}$.*

Proof. By Proposition 4 there is a measurable mapping $Y : \mathscr{X} \to \mathbb{R}$ such that $\mathscr{F}_{\mathscr{X}} = Y^{-1}(\mathscr{B}_1)$. By Theorem 1, there is a regular conditional distribution $\nu_\omega(dy)$ for the real random variable $Y \circ X$ given $\mathscr{G}$. Define a Markov kernel $\mu_\omega(dx)$ from $(\Omega, \mathscr{F})$ to $(\mathscr{X}, \mathscr{F}_{\mathscr{X}})$ as follows: for every Borel set $B \in \mathscr{B}_1$, set

$$\mu_\omega(Y^{-1}(B)) = \nu_\omega(B).$$

Since $\mathscr{F}_{\mathscr{X}} = Y^{-1}(\mathscr{B}_1)$, this defines $\mu_\omega(F)$ for every $F \in \mathscr{F}_{\mathscr{X}}$. That $\mu_\omega(dx)$ is a Markov kernel follows directly from the fact that $\nu_\omega(dy)$ is a Markov kernel. That $\mu_\omega(dx)$ is a regular conditional distribution for X given $\mathscr{G}$ can be seen as follows. For each $F \in \mathscr{F}_{\mathscr{X}}$ there is a Borel set B such that $F = Y^{-1}(B)$, and so

$$\begin{aligned} P(X \in F \,|\, \mathscr{G})(\omega) &= P(Y \circ X \in B \,|\, \mathscr{G})(\omega) \\ &= \nu_\omega(B) \\ &= \mu_\omega(F). \end{aligned}$$

□

Exercise 4. Use the existence of regular conditional distributions for random variables valued in $(\mathbb{R}^\infty, \mathscr{B}_\infty)$ to deduce the dominated convergence theorem for conditional expectations from the dominated convergence theorem for ordinary expectations. NOTE: You may also need the disintegration formula of Theorem 2.