Conditional expectation

Jason Swanson

April 23, 2014

1 Conditioning on σ-algebras

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A \in \mathcal{F}$ with $P(A) > 0$. Define

$$Q(B) = P(B \mid A) = \frac{P(B \cap A)}{P(A)}, \qquad \text{for all } B \in \mathcal{F}.$$

It is easy to to check that Q is a probability measure on $(\Omega, \mathcal{F})$. If X is a random variable, we define the **conditional expectation of X given A** as

$$E[X \mid A] = \int X \, dQ, \tag{1.1}$$

whenever this integral is well-defined. Note that $E[1_B \mid A] = P(B \mid A)$.

Theorem 1.1. *If $E[|X|1_A] < \infty$, then X is Q-integrable. If $X \geqslant 0$ or $E[|X|1_A] < \infty$, then*

$$E[X \mid A] = \frac{E[X1_A]}{P(A)}. \tag{1.2}$$

Remark 1.2. Note that (1.2) may be written as

$$E[X \mid A] = \frac{\alpha(A)}{P(A)}, \tag{1.3}$$

where $d\alpha = X \, dP$. Also note that (1.2) gives us the formula $E[X1_A] = P(A)E[X \mid A]$. If $X = 1_B$, then this reduces to the familiar multiplication rule, $P(A \cap B) = P(A)P(B \mid A)$.

Proof of Theorem 1.1. Note that if $P(B) = 0$, then $Q(B) = 0$. Hence $Q \ll P$. Also note that

$$Q(B) = \int_B \frac{1_A}{P(A)} \, dP, \qquad \text{for all } B \in \mathcal{F}.$$

Thus, $dQ/dP = 1_A/P(A)$. It follows that if $X \geqslant 0$, then

$$E[X \mid A] = \int X \, dQ = \int X \frac{dQ}{dP} \, dP = E\left[X \frac{1_A}{P(A)}\right] = \frac{E[X1_A]}{P(A)}.$$

Therefore, if $E[|X|1_A] < \infty$, then X is Q-integrable, and the same formula holds. $\qquad \square$

1

Lemma 1.3. *Any finite σ-algebra $\mathcal{F}$ on Ω can be written as $\mathcal{F} = \sigma(\{A_j\}_{j=1}^n)$, where $\{A_j\}_{j=1}^n$ is a partition of Ω, that is, $\Omega = \biguplus_{j=1}^n A_j$. Moreover, the partition $\{A_j\}_{j=1}^n$ is unique.*

Proof. For each $\omega \in \Omega$, let A_ω be the smallest measurable set containing ω. That is, $A_\omega = \bigcap \mathcal{F}_\omega$, where $\mathcal{F}_\omega = \{A \in \mathcal{F} : \omega \in A\}$. Since this is a finite intersection, $A_\omega \in \mathcal{F}$. In particular, $\mathcal{E} = \{A_\omega : \omega \in \Omega\}$ is a finite set. We claim that $\mathcal{E}$ is a partition of Ω and that $\mathcal{F} = \sigma(\mathcal{E})$.

Clearly, $\Omega = \bigcup_{\omega \in \Omega} A_\omega$, so to show that $\mathcal{E}$ is a partition, it suffices to show that this is a disjoint union. More specifically, we wish to show that if $\omega, \omega' \in \Omega$, then either $A_\omega = A_{\omega'}$ or $A_\omega \cap A_{\omega'} = \varnothing$. Let $\omega, \omega' \in \Omega$. Note that for any $A \in \mathcal{F}$, if $\omega \in A$, then $A \in \mathcal{F}_\omega$, which implies $A_\omega \subset A$. Hence, if $\omega \in A_{\omega'}$, then $A_\omega \subset A_{\omega'}$; and if $\omega \in A_{\omega'}^c$, then $A_\omega \subset A_{\omega'}^c$. That is, either $A_\omega \subset A_{\omega'}$ or $A_\omega \subset A_{\omega'}^c$. By symmetry, either $A_{\omega'} \subset A_\omega$ or $A_{\omega'} \subset A_\omega^c$. Taken together, this shows that either $A_\omega = A_{\omega'}$ or $A_\omega \cap A_{\omega'} = \varnothing$.

To see that $\mathcal{F} = \sigma(\mathcal{E})$, simply note that any $A \in \mathcal{F}$ can be written as $A = \bigcup_{\omega \in A} A_\omega$, and that this is a finite union.

For uniqueness, suppose that $\mathcal{F} = \sigma(\{B_j\}_{j=1}^n)$, where $\Omega = \biguplus_{j=1}^n B_j$. If $\omega \in B_j$, then $A_\omega = B_j$. Therefore, $\mathcal{E} = \{B_j\}_{j=1}^n$. $\qquad\square$

Exercise 1.4. Show that every infinite σ-algebra is uncountable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable. Let $\mathcal{G} \subset \mathcal{F}$ be a finite σ-algebra. Write $\mathcal{G} = \sigma(\{A_j\}_{j=1}^n)$, where $\{A_j\}_{j=1}^n$ is a partition of Ω. The **conditional expectation of X given $\mathcal{G}$**, written $E[X \mid \mathcal{G}]$, is a random variable defined by

$$
E[X \mid \mathcal{G}](\omega) = \begin{cases} E[X \mid A_1] & \text{if } \omega \in A_1, \\ E[X \mid A_2] & \text{if } \omega \in A_2, \\ \quad\vdots \\ E[X \mid A_n] & \text{if } \omega \in A_n. \end{cases}
$$

Note that we may write

$$
E[X \mid \mathcal{G}] = \sum_{j=1}^n E[X \mid A_j] 1_{A_j}. \tag{1.4}
$$

We also define the **conditional probability of A given $\mathcal{G}$** as $P(A \mid \mathcal{G}) = E[1_A \mid \mathcal{G}]$. Note that $P(A \mid \mathcal{G})(\omega) = P(A \mid A_j)$ if $\omega \in A_j$.

To illustrate this concept, consider the following heuristic example. Imagine my friend is at the local bar, and is about to throw a dart at a dartboard. If I model the dart board by a unit circle, which I call Ω, then his dart will land at some point $\omega \in \Omega$.

Unfortunately, I am not there with him and will not be able to observe the exact location of ω. But after he throws the dart, he is going to call me on the phone and tell me what his score for that shot was. This information will not be enough for me to determine ω. It will, however, narrow it down. Before I receive his call, I can partition the dartboard Ω into several pieces, $A_1, \ldots, A_n$, with each piece corresponding to a unique score. Once he calls me, I will know which piece contains his dart.

Let X be the distance from his dart to the bullseye. Suppose he calls me and I determine that his dart is somewhere inside A_j. I can then compute $E[X \mid A_j]$. However, before

he calls, I can get prepared by computing $E[X \mid A_j]$ for all j, and then encoding all this information into the single random variable $E[X \mid \mathcal{G}]$.

In probability theory, we model information by σ-algebras. In this example, the σ-algebra $\mathcal{G}$ generated by the partition $\{A_j\}$ models the information I will receive from my friend's phone call. Imagine that while I am waiting for my friend's phone call, an interviewer starts asking me questions. For various events A, the interviewer asks me, "After your friend calls, will you know with certainty whether or not A has occurred?" Depending on the event A, I will have to answer either yes, no, or maybe. The events $A \in \mathcal{G}$ are precisely those events for which I can answer yes.

Theorem 1.5. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable. Let $\mathcal{G} \subset \mathcal{F}$ be a finite σ-algebra, and let $d\alpha = X \, dP$. Then*

$$E[X \mid \mathcal{G}] = \frac{d(\alpha|_{\mathcal{G}})}{d(P|_{\mathcal{G}})}. \tag{1.5}$$

Or equivalently, $Z = E[X \mid \mathcal{G}]$ is the unique random variable such that

(i) Z is $\mathcal{G}$-measurable, and

(ii) $E[X 1_A] = E[Z 1_A]$, for all $A \in \mathcal{G}$.

Remark 1.6. First, note the similarity between (1.5) and (1.3). Second, it should be mentioned that $Z = E[X \mid \mathcal{G}]$ is unique in the following sense. If $\widetilde{Z}$ is another $\mathcal{G}$-measurable random variable such that $E[X 1_A] = E[\widetilde{Z} 1_A]$ for all $A \in \mathcal{G}$, then $Z = \widetilde{Z}$ a.s.

Proof of Theorem 1.5. Let $Z = E[X \mid \mathcal{G}]$. To prove (1.5), we must show that Z is $\mathcal{G}$-measurable and that

$$(\alpha|_{\mathcal{G}})(A) = \int_A Z \, d(P|_{\mathcal{G}}), \quad \text{for all } A \in \mathcal{G}.$$

If $A \in \mathcal{G}$, then $(\alpha|_{\mathcal{G}})(A) = \alpha(A) = E[X 1_A]$; and if Z is $\mathcal{G}$-measurable, then $\int_A Z \, d(P|_{\mathcal{G}}) = \int_A Z \, dP = E[Z 1_A]$. It therefore follows that (1.5) is equivalent to (i) and (ii). Uniqueness follows from the uniqueness for the Radon-Nikodym derivative.

Write $\mathcal{G} = \sigma(\{A_j\}_{j=1}^n)$, where $\{A_j\}_{j=1}^n$ is a partition of Ω. By (1.4), it is clear that Z is $\mathcal{G}$-measurable, so we need only verify (ii). By the linearity of the expected value, it will suffice to show that $E[X 1_{A_j}] = E[Z 1_{A_j}]$ for all j. Using (1.4), we have

$$E[Z 1_{A_j}] = E\left[\left(\sum_{i=1}^n E[X \mid A_i] 1_{A_i}\right) 1_{A_j}\right] = E\left[\sum_{i=1}^n E[X \mid A_i] 1_{A_i \cap A_j}\right]$$

$$= E[E[X \mid A_j] 1_{A_j}] = E[X \mid A_j] P(A_j) = E[X 1_{A_j}],$$

where the last equality follows from (1.2). $\qquad\square$

We now extend the definition of $E[X \mid \mathcal{G}]$ to infinite σ-algebras. Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. The **conditional expectation of X given $\mathcal{G}$** is the unique random variable $Z = E[X \mid \mathcal{G}]$ satisfying (i) and (ii) of Theorem 1.5. Equivalently, $E[X \mid \mathcal{G}]$ can be defined by (1.5), where $d\alpha = X \, dP$. We also define the **conditional probability of A given $\mathcal{G}$** as $P(A \mid \mathcal{G}) = E[1_A \mid \mathcal{G}]$.

2 Conditioning on random variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable. Let Y be a discrete random variable taking values in a finite set $S = \{k_1, \ldots, k_n\}$, and assume that $P(Y = k_j) > 0$ for all j. Define $f(k) = E[X \mid Y = k]$ for all $k \in S$. Since $\{Y = k\}$ is an event with positive probability, this is the simplest kind of conditional expectation we have used so far, namely, the one defined in (1.1).

We now define the **conditional expectation of X given Y** as $E[X \mid Y] = f(Y)$. In other words, $E[X \mid Y]$ is a random variable whose outcome depends on Y. Specifically, for all $\omega \in \{Y = k\}$, we have $E[X \mid Y](\omega) = f(k)$.

Theorem 2.1. *Under the above definitions, $E[X \mid Y] = E[X \mid \sigma(Y)]$.*

Proof. Under the above definitions, $E[X \mid Y] = f(Y)$. In order to show that $f(Y) = E[X \mid \sigma(Y)]$, we must show that (i) $f(Y)$ is $\sigma(Y)$-measurable, and (ii) $E[X1_A] = E[f(Y)1_A]$ for all $A \in \sigma(Y)$.

Recall that a random variable Z is $\sigma(Y)$-measurable if and only if $Z = g(Y)$ for some measurable function g. Hence, (i) clearly holds. For (ii), let $A \in \sigma(Y)$ be arbitrary. Then $A = \{Y \in B\}$ for some $B \in \mathcal{R}$. This implies there is some subset $C \subset S$ such that $A = \bigcup_{k \in C} \{Y = k\}$. By the linearity of the expected value, we may assume without loss of generality that $A = \{Y = k\}$. In this case,

$$
\begin{aligned}
E[f(Y)1_A] = E[f(Y)1_{\{Y=k\}}] = E[f(k)1_{\{Y=k\}}] &= f(k)P(Y = k) \\
&= E[X|Y = k]P(Y = k) = E[X1_{\{Y=k\}}] = E[X1_A].
\end{aligned}
$$

Note that we have again used (1.2). $\qquad \square$

We now extend the definition of $E[X \mid Y]$ to arbitrary random variables. Let $(\Omega, \mathcal{F}, P)$ be a probability space and X an integrable random variable. Let Y be an arbitrary random variable. The **conditional expectation of X given Y** is $E[X \mid Y] = E[X \mid \sigma(Y)]$. We also define the **conditional probability of A given Y** as $P(A \mid Y) = E[1_A \mid Y]$. Note that since $E[X \mid Y]$ is $\sigma(Y)$-measurable, there exists a measurable function g (which depends on X) such that $E[X \mid Y] = g(Y)$.

Example 2.2. Suppose X and Y are random variables with a joint density function $f(x, y)$. Note that $P(X \in A \mid Y = k)$ is undefined, since $P(Y = k) = 0$. Nonetheless, it should be intuitively clear that the following equation ought to hold:

$$
P(X \in A \mid Y = k) = \frac{\displaystyle\int_A f(x, k) \, dx}{\displaystyle\int_{\mathbb{R}} f(x, k) \, dx}. \tag{2.1}
$$

The integral in the denominator is necessary in order to make the function $x \mapsto f(x, k)$ a probability density function. In this example, we will explore the sense in which this formula is rigorously valid. (In an undergraduate class, it may be rigorously valid *by definition*. But for us, as usual, it is a special case of something more general.)

Theorem 2.3. *Let X and Y have joint density $f(x, y)$. Let g be a measurable function such that $E|g(X)| < \infty$. Define*

$$h(k) = \frac{\displaystyle\int_{\mathbb{R}} g(x)f(x, k)\, dx}{\displaystyle\int_{\mathbb{R}} f(x, k)\, dx},$$

whenever $\int_{\mathbb{R}} f(x, k)\, dx > 0$, and $h(k) = 0$ otherwise. Then $E[g(X) \mid Y] = h(Y)$.

Remark 2.4. If $g(x) = 1_A(x)$, then $h(k)$ agrees with the right-hand side of (2.1). Also, as can be seen from the proof below, we could have defined $h(k)$ arbitrarily when $\int_{\mathbb{R}} f(x, k)\, dx = 0$.

Proof of Theorem 2.3. Since $h(Y)$ is $\sigma(Y)$-measurable, it will suffice for us to show that $E[h(Y)1_A] = E[g(X)1_A]$ for all $A \in \sigma(Y)$. Let $A \in \sigma(Y)$. Then $A = \{Y \in B\}$ for some $B \in \mathcal{R}$. We now have

$$E[h(Y)1_A] = E[h(Y)1_B(Y)] = \int_B \int_{\mathbb{R}} h(y)f(x, y)\, dx\, dy$$

$$= \int_B \left(h(y) \int_{\mathbb{R}} f(x, y)\, dx \right) dy = \int_{B \cap C} \left(h(y) \int_{\mathbb{R}} f(x, y)\, dx \right) dy,$$

where $C = \{y : \int_{\mathbb{R}} f(x, y)\, dx > 0\}$. Note that for all $y \in C$, we have $h(y) \int_{\mathbb{R}} f(x, y)\, dx = \int_{\mathbb{R}} g(x)f(x, y)\, dx$. Also, for all $y \in C^c$, we have $f(x, y) = 0$ for Lebesgue almost every x. Thus, $y \in C^c$ implies $\int_{\mathbb{R}} g(x)f(x, y)\, dx = 0$. It therefore follows that

$$E[h(Y)1_A] = \int_{B \cap C} \int_{\mathbb{R}} g(x)f(x, y)\, dx\, dy = \int_B \int_{\mathbb{R}} g(x)f(x, y)\, dx\, dy$$

$$= E[g(X)1_B(Y)] = E[g(X)1_A],$$

which was what we needed to prove. $\qquad\square$

In general, we interpret $E[X \mid Y = y]$ to mean $g(y)$, where g is a measurable function such that $E[X \mid Y] = g(Y)$. Some caution is needed in these cases, though, since such a function g is only defined μ_Y-a.s., where μ_Y is the distribution of Y.

3   Properties of conditional expectation

3.1   Basic properties

Let $(\Omega, \mathcal{F}, P)$ be a probability space, X an integrable random variable, and $\mathcal{G} \subset \mathcal{F}$.

Proposition 3.1. $E[E[X \mid \mathcal{G}]] = EX$.

Proof. By the definition of conditional expectation, we have $E[E[X \mid \mathcal{G}]1_A] = E[X1_A]$ for all $A \in \mathcal{G}$. Take $A = \Omega$. $\qquad\square$

Proposition 3.2. *If X is $\mathcal{G}$-measurable, then $E[X \mid \mathcal{G}] = X$.*

Proof. It follows trivially, since X is $\mathcal{G}$-measurable and $E[X1_A] = E[X1_A]$ for all $A \in \mathcal{G}$. $\square$

Recall that X and $\mathcal{G}$ are independent if $\sigma(X)$ and $\mathcal{G}$ are independent, which in turn means that $P(A \cap B) = P(A)P(B)$ whenever $A \in \sigma(X)$ and $B \in \mathcal{G}$. Hence, X and $\mathcal{G}$ are independent if and only if $P(\{X \in C\} \cap B) = P(X \in C)P(B)$ for all $C \in \mathcal{R}$ and $B \in \mathcal{G}$.

Proposition 3.3. *If X and $\mathcal{G}$ are independent, then $E[X \mid \mathcal{G}] = E[X]$. In particular, $E[X \mid \{\varnothing, \Omega\}] = E[X]$.*

Proof. A constant random variable is measurable with respect to every σ-algebra, so $E[X]$ is trivially $\mathcal{G}$-measurable. Also, for all $A \in \mathcal{G}$, we have $E[X1_A] = E[X]E[1_A] = E[E[X]1_A]$. The final claim holds since every random variable is independent of the trivial σ-algebra. $\square$

Theorem 3.4. *If $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{F}$, then $E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = E[E[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = E[X \mid \mathcal{G}_1]$.*

Remark 3.5. In words, this says that in a battle between nested σ-algebras, the smallest σ-algebra always wins.

Proof of Theorem 3.4. Since $\mathcal{G}_1 \subset \mathcal{G}_2$ and $E[X \mid \mathcal{G}_1]$ is $\mathcal{G}_1$-measurable, it is also $\mathcal{G}_2$-measurable. Hence, by Proposition 3.2, $E[E[X \mid \mathcal{G}_1] \mid \mathcal{G}_2] = E[X \mid \mathcal{G}_1]$. The other equality holds since $E[X \mid \mathcal{G}_1]$ is $\mathcal{G}_1$-measurable and, for all $A \in \mathcal{G}_1 \subset \mathcal{G}_2$, we have $E[E[X \mid \mathcal{G}_1]1_A] = E[X1_A] = E[E[X \mid \mathcal{G}_2]1_A]$. $\square$

Theorem 3.6. *If Y and XY are integrable, and X is $\mathcal{G}$-measurable, then*

$$E[XY \mid \mathcal{G}] = XE[Y \mid \mathcal{G}] \text{ a.s.}$$

Proof. Let $d\alpha = Y\,dP$, so that $E[Y \mid \mathcal{G}] = d(\alpha|_\mathcal{G})/d(P|_\mathcal{G})$. Let $d\beta_1 = X\,d\alpha_+$, $d\beta_2 = X\,d\alpha_-$, and $\beta = \beta_1 - \beta_2$, so that $d\beta = XY\,dP$ and $E[XY \mid \mathcal{G}] = d(\beta|_\mathcal{G})/d(P|_\mathcal{G})$. Since X is $\mathcal{G}$-measurable, we have $d(\beta_1|_\mathcal{G})/d(\alpha_+|_\mathcal{G}) = d(\beta_2|_\mathcal{G})/d(\alpha_-|_\mathcal{G}) = X$. Hence,

$$
\begin{aligned}
E[XY \mid \mathcal{G}] &= \frac{d(\beta|_\mathcal{G})}{d(P|_\mathcal{G})} = \frac{d(\beta_1|_\mathcal{G})}{d(P|_\mathcal{G})} - \frac{d(\beta_2|_\mathcal{G})}{d(P|_\mathcal{G})} \\
&= \frac{d(\beta_1|_\mathcal{G})}{d(\alpha_+|_\mathcal{G})} \cdot \frac{d(\alpha_+|_\mathcal{G})}{d(P|_\mathcal{G})} - \frac{d(\beta_2|_\mathcal{G})}{d(\alpha_-|_\mathcal{G})} \cdot \frac{d(\alpha_-|_\mathcal{G})}{d(P|_\mathcal{G})} \\
&= X\left(\frac{d(\alpha_+|_\mathcal{G})}{d(P|_\mathcal{G})} - \frac{d(\alpha_-|_\mathcal{G})}{d(P|_\mathcal{G})}\right) = X\frac{d(\alpha|_\mathcal{G})}{d(P|_\mathcal{G})} = XE[Y \mid \mathcal{G}], \ P\text{-a.s.,}
\end{aligned}
$$

and we are done. $\square$

Theorem 3.7. (linearity) $E[aX + Y \mid \mathcal{G}] = aE[X \mid \mathcal{G}] + E[Y \mid \mathcal{G}]$.

Proof. The right-hand side is clearly $\mathcal{G}$-measurable. Let $A \in \mathcal{G}$. Then

$$
\begin{aligned}
E[(aE[X \mid \mathcal{G}] + E[Y \mid \mathcal{G}])1_A] &= aE[E[X \mid \mathcal{G}]1_A] + E[E[Y \mid \mathcal{G}]1_A] \\
&= aE[X1_A] + E[Y1_A] = E[(aX + Y)1_A],
\end{aligned}
$$

and we are done. $\square$

Lemma 3.8. *Suppose U and V are $\mathcal{H}$-measurable random variables. If $E[U1_A] \leqslant E[V1_A]$ for all $A \in \mathcal{H}$, then $U \leqslant V$ a.s. If $E[U1_A] = E[V1_A]$ for all $A \in \mathcal{H}$, then $U = V$ a.s.*

Proof. By reversing the roles of U and V, the second claim follows from the first. To prove the first, suppose $E[U1_A] \leqslant E[V1_A]$ for all $A \in \mathcal{H}$. Let $A = \{U > V\} \in \mathcal{H}$ and define $Z = (U - V)1_A$, so that $Z \geqslant 0$. Note that $EZ = E[U1_A] - E[V1_A] \leqslant 0$. Hence, $EZ = 0$, so $Z = 0$ a.s., which implies $P(A) = 0$. $\qquad\square$

Theorem 3.9. (monotonicity) *If $X \leqslant Y$ a.s., then $E[X \mid \mathcal{G}] \leqslant E[Y \mid \mathcal{G}]$ a.s.*

Proof. For all $A \in \mathcal{G}$, we have $E[E[X \mid \mathcal{G}]1_A] = E[X1_A] \leqslant E[Y1_A] = E[E[Y \mid \mathcal{G}]1_A]$. Hence, by Lemma 3.8, $E[X \mid \mathcal{G}] \leqslant E[Y \mid \mathcal{G}]$ a.s. $\qquad\square$

Theorem 3.10. *Suppose X and Y are independent and φ is a measurable function such that $E|\varphi(X,Y)| < \infty$, then $E[\varphi(X,Y) \mid X] = g(X)$, where $g(x) = E[\varphi(x,Y)]$.*

Remark 3.11. It is important here that X and Y are independent. This result is not true when X and Y are dependent.

Proof of Theorem 3.10. Clearly, $g(X)$ is $\sigma(X)$-measurable. Let $A \in \mathcal{R}$. Then

$$
E[\varphi(X,Y)1_{\{X \in A\}}] = \int \int \varphi(x,y) 1_A(x) \mu_Y(dy) \mu_X(dx)
$$
$$
= \int 1_A(x) \left(\int \varphi(x,y) \mu_Y(dy) \right) \mu_X(dx) = \int 1_A(x) g(x) \mu_X(dx) = E[g(X)1_{\{X \in A\}}],
$$

and we are done. $\qquad\square$

Example 3.12. Let X, Y, Z be iid, uniformly distributed on $(0,1)$. We shall compute the distribution of $(XY)^Z$. We begin by computing the distribution of $W = XY$. Let $w \in (0,1)$. Then
$$
P(W \leqslant w) = P(XY \leqslant w) = E[1_{\{XY \leqslant w\}}] = E[E[1_{\{XY \leqslant w\}} \mid X]].
$$
By Theorem 3.10, $E[1_{\{XY \leqslant w\}} \mid X] = f(X)$, where
$$
f(x) = E[1_{\{xY \leqslant w\}}] = P(xY \leqslant w) = P\left(Y \leqslant \frac{w}{x}\right) = 1_{\{x < w\}} + \frac{w}{x} 1_{\{x \geqslant w\}}.
$$
Thus,
$$
P(W \leqslant w) = E[f(X)] = E\left[1_{\{X < w\}} + \frac{w}{X} 1_{\{X \geqslant w\}}\right] = w + \int_w^1 \frac{w}{x}\,dx = w - w \log w.
$$

Differentiating, we find that W has density $f_W(w) = (-\log w)1_{(0,1)}(w)$.

Similarly, for $x \in (0,1)$, we now compute
$$
P((XY)^Z \leqslant x) = E[P(W^Z \leqslant x \mid W)] = E[g(W)],
$$
where
$$
g(w) = P(w^Z \leqslant x) = P\left(Z \geqslant \frac{\log x}{\log w}\right) = \left(1 - \frac{\log x}{\log w}\right)1_{\{w \leqslant x\}}.
$$

Thus,

$$P((XY)^Z \leqslant x) = E\left[\left(1 - \frac{\log x}{\log W}\right)1_{\{W \leqslant x\}}\right] = \int_0^x \left(1 - \frac{\log x}{\log w}\right)(-\log w)\,dw$$

$$= -\int_0^x \log w\,dw + x\log x = x.$$

In other words, $(XY)^Z$ is uniformly distributed on $(0,1)$.

3.2   Limit theorems and inequalities

Theorem 3.13. (monotone convergence) *If $0 \leqslant X_n \uparrow X$ a.s. and X is integrable, then $E[X_n \mid \mathcal{G}] \uparrow E[X \mid \mathcal{G}]$ a.s.*

Proof. By monotonicity, there exists a $\mathcal{G}$-measurable random variable Z such that $E[X_n \mid \mathcal{G}] \uparrow Z$ a.s. Let $A \in \mathcal{G}$. Using monotone convergence,

$$E[Z1_A] = \lim_{n\to\infty} E[E[X_n \mid \mathcal{G}]1_A] = \lim_{n\to\infty} E[X_n 1_A] = E[X1_A],$$

which shows $Z = E[X \mid \mathcal{G}]$. $\qquad\square$

Theorem 3.14. (Fatou's lemma) *If $X_n \geqslant 0$ a.s., each X_n is integrable, and $\liminf_{n\to\infty} X_n$ is integrable, then*

$$E[\liminf_{n\to\infty} X_n \mid \mathcal{G}] \leqslant \liminf_{n\to\infty} E[X_n \mid \mathcal{G}] \text{ a.s.}$$

Proof. Let $\overline{X}_n = \inf_{j \geqslant n} X_j$ and $X = \liminf_{n\to\infty} X_n$. Note that $0 \leqslant \overline{X}_n \uparrow X$. In particular, $\overline{X}_n$ is integrable. For each $j \geqslant n$, we have $\overline{X}_n \leqslant X_j$ a.s. Hence, by monotonicity, $E[\overline{X}_n \mid \mathcal{G}] \leqslant E[X_j \mid \mathcal{G}]$ a.s. It follows that

$$E[\overline{X}_n \mid \mathcal{G}] \leqslant \inf_{j \geqslant n} E[X_j \mid \mathcal{G}] \text{ a.s.}$$

Monotone convergence implies

$$E[X \mid \mathcal{G}] = \lim_{n\to\infty} E[\overline{X}_n \mid \mathcal{G}] \leqslant \liminf_{n\to\infty} E[X_n \mid \mathcal{G}] \text{ a.s.,}$$

and we are done. $\qquad\square$

Theorem 3.15. (dominated convergence) *Let X_n be random variables with $X_n \to X$ a.s. Suppose there exists an integrable random variable Y such that $|X_n| \leqslant Y$ a.s. for all n. Then*

$$\lim_{n\to\infty} E[X_n \mid \mathcal{G}] = E[X \mid \mathcal{G}] \text{ a.s.}$$

Exercise 3.16. Prove Theorem 3.15 by mimicking the proof of the ordinary dominated convergence theorem.

Exercise 3.17. Show that if $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then the left-hand derivative,

$$\varphi'_-(c) = \lim_{h \downarrow 0} \frac{\varphi(c) - \varphi(c - h)}{h}$$

exists for all c. Moreover,

$$\varphi(x) - \varphi(c) - (x - c)\varphi'_-(c) \geqslant 0, \tag{3.1}$$

for all x and c.

Theorem 3.18. (Jensen's inequality) *If φ is convex and X and $\varphi(X)$ are integrable, then $\varphi(E[X \mid \mathcal{G}]) \leqslant E[\varphi(X) \mid \mathcal{G}]$.*

Proof. Let $Z = (X - E[X \mid \mathcal{G}])\varphi'_-(E[X \mid \mathcal{G}])$, so that by (3.1), $\varphi(X) - \varphi(E[X \mid \mathcal{G}]) - Z \geqslant 0$, which implies

$$0 \leqslant E[\varphi(X) - \varphi(E[X \mid \mathcal{G}]) - Z \mid \mathcal{G}] = E[\varphi(X) \mid \mathcal{G}] - \varphi(E[X \mid \mathcal{G}]) - E[Z \mid \mathcal{G}].$$

It therefore suffices to show that $E[Z \mid \mathcal{G}] = 0$. To see this, we calculate

$$\begin{aligned}
E[Z \mid \mathcal{G}] &= E[(X - E[X \mid \mathcal{G}])\varphi'_-(E[X \mid \mathcal{G}]) \mid \mathcal{G}] \\
&= \varphi'_-(E[X \mid \mathcal{G}])E[X - E[X \mid \mathcal{G}] \mid \mathcal{G}] \\
&= \varphi'_-(E[X \mid \mathcal{G}])(E[X \mid \mathcal{G}] - E[E[X \mid \mathcal{G}] \mid \mathcal{G}]) \\
&= \varphi'_-(E[X \mid \mathcal{G}])(E[X \mid \mathcal{G}] - E[X \mid \mathcal{G}]) = 0,
\end{aligned}$$

and we are done. $\square$

Theorem 3.19. (Hölder's inequality) *Let $p, q \in (1, \infty)$ be conjugate exponents, so that $1/p + 1/q = 1$. Suppose that $|X|^p$ and $|Y|^q$ are integrable. Then*

$$E[|XY| \mid \mathcal{G}] \leqslant (E[|X|^p \mid \mathcal{G}])^{1/p}(E[|Y|^q \mid \mathcal{G}])^{1/q} \ a.s.$$

Proof. Note that by the ordinary Hölder's inequality, XY is integrable, so that $E[|XY| \mid \mathcal{G}]$ is well-defined. Let $U = (E[|X|^p \mid \mathcal{G}])^{1/p}$ and $V = (E[|Y|^q \mid \mathcal{G}])^{1/q}$. Note that both U and V are $\mathcal{G}$-measurable. Observe that

$$E[|X|^p 1_{\{U=0\}}] = E[E[|X|^p 1_{\{U=0\}} \mid \mathcal{G}]] = E[1_{\{U=0\}}E[|X|^p \mid \mathcal{G}]] = E[1_{\{U=0\}}U^p] = 0.$$

Hence, $|X|1_{\{U=0\}} = 0$ a.s., which implies

$$E[|XY| \mid \mathcal{G}]1_{\{U=0\}} = E[|XY|1_{\{U=0\}} \mid \mathcal{G}] = 0.$$

Similarly, $E[|XY| \mid \mathcal{G}]1_{\{V=0\}} = 0$. It therefore suffices to show that $E[|XY| \mid \mathcal{G}]1_H \leqslant UV$, where $H = \{U > 0, V > 0\}$. For this, we will use Lemma 3.8 to prove that

$$\frac{E[|XY| \mid \mathcal{G}]}{UV}1_H \leqslant 1 \text{ a.s.}.$$

Note that the left-hand side is defined to be zero on H^c.

9

Let $A \in \mathcal{G}$ be arbitrary and define $G = H \cap A$. Then

$$E\left[\frac{E[|XY| \mid \mathcal{G}]}{UV}1_H 1_A\right] = E\left[E\left[\frac{|XY|}{UV}1_G \,\Big|\, \mathcal{G}\right]\right]$$

$$= E\left[\frac{|X|}{U}1_G \cdot \frac{|Y|}{V}1_G\right]$$

$$\leqslant \left(E\left[\frac{|X|^p}{U^p}1_G\right]\right)^{1/p}\left(E\left[\frac{|Y|^q}{V^q}1_G\right]\right)^{1/q}$$

$$= \left(E\left[\frac{E[|X|^p \mid \mathcal{G}]}{U^p}1_G\right]\right)^{1/p}\left(E\left[\frac{E[|Y|^q \mid \mathcal{G}]}{V^q}1_G\right]\right)^{1/q}$$

$$= (E[1_G])^{1/p}(E[1_G])^{1/q} = E[1_G] \leqslant E[1_A].$$

Applying Lemma 3.8 finishes the proof. $\qquad\square$

3.3 Minimizing the mean square error

We say that a random variable X is **square integrable** if $E|X|^2 < \infty$. Let X be square integrable and consider the function $f(a) = E|X - a|^2 = a^2 - 2(EX)a + E|X|^2$. This function has a minimum at $a = EX$. In other words, if we wish to approximate X by a constant, then the constant EX is the one which minimizes our mean square error.

The conditional expectation has a similar property. If we wish to approximate X by a square integrable, $\mathcal{G}$-measurable random variable, then $E[X \mid \mathcal{G}]$ is the random variable which minimizes our mean square error. This is made precise in the following theorem.

Theorem 3.20. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be square integrable. Let $\mathcal{G} \subset \mathcal{F}$ and define $Z = E[X \mid \mathcal{G}]$. If Y is any square integrable, $\mathcal{G}$-measurable random variable, then $E|X - Z|^2 \leqslant E|X - Y|^2$.*

Proof. First note that by Jensen's inequality,

$$|Z|^2 = |E[X \mid \mathcal{G}]|^2 \leqslant E[|X|^2 \mid \mathcal{G}] \text{ a.s.}$$

Hence, $E|Z|^2 \leqslant E[E[|X|^2 \mid \mathcal{G}]] = E|X|^2 < \infty$ and Z is square integrable. Let $W = Z - Y$. Since W is $\mathcal{G}$-measurable,

$$E[WZ] = E[WE[X \mid \mathcal{G}]] = E[E[WX \mid \mathcal{G}]] = E[WX].$$

Hence, $E[W(X - Z)] = 0$, which implies

$$E|X - Y|^2 = E|X - Z + W|^2 = E|X - Z|^2 + 2E[W(X - Z)] + E|W|^2$$
$$= E|X - Z|^2 + E|W|^2 \geqslant E|X - Z|^2,$$

and we are done. $\qquad\square$

Remark 3.21. In the language of Hilbert spaces and L^p spaces, this theorem says the following: X is an element of the Hilbert space $L^2(\Omega, \mathcal{F}, P)$, and $E[X \mid \mathcal{G}]$ is the orthogonal projection of X onto the subspace $L^2(\Omega, \mathcal{G}, P)$.

4 Regular conditional distributions

4.1 Introduction

If X is a real-valued random variable, then $\mu_X(A) = P(X \in A)$ defines a measure μ_X on the real line which we call the distribution (or law) of X. One feature of the distribution is that it provides us with a way to calculate expectations:

$$E[f(X)] = \int_{\mathbb{R}} f(x)\, \mu_X(dx).$$

Likewise, if B is an event, then $\mu_{X,B}(A) = P(X \in A \mid B)$ defines a measure $\mu_{X,B}$ on $\mathbb{R}$ which is the conditional distribution of X given B, and we have

$$E[f(X) \mid B] = \int_{\mathbb{R}} f(x)\, \mu_{X,B}(dx).$$

If $\mathcal{G}$ is a finite σ-algebra, so that $\mathcal{G} = \sigma(\{A_j\}_{j=1}^n)$, where $\{A_j\}_{j=1}^n$ is a partition of Ω, then

$$P(X \in A \mid \mathcal{G})(\omega) = \begin{cases} P(X \in A \mid A_1) & \text{if } \omega \in A_1, \\ P(X \in A \mid A_2) & \text{if } \omega \in A_2, \\ \quad\vdots \\ P(X \in A \mid A_n) & \text{if } \omega \in A_n. \end{cases}$$

In other words, $P(X \in \cdot \mid \mathcal{G})$ is just a conditional probability distribution that happens to depend on ω. Another way of saying it is that $P(X \in \cdot \mid \mathcal{G})$ is a random probability measure on the real line.

Conditional expectations can be computed by integrating against this random measure. That is, if we define $\mu_{X,\mathcal{G}}(\omega, A) = \mu_{X,A_j}(A)$ for $\omega \in A_j$, then

$$E[f(X) \mid \mathcal{G}](\omega) = \int_{\mathbb{R}} f(x)\, \mu_{X,\mathcal{G}}(\omega, dx).$$

With a structure such as this, expectations conditioned on σ-algebras behave very much like ordinary expectations. When this happens, we are able to make valuable intuitive connections to mathematical ideas that we are already familiar with. It would be nice if $P(X \in \cdot \mid \mathcal{G})$ was always a random measure, even when $\mathcal{G}$ is infinite. The following theorem is a step in this direction.

Theorem 4.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ a measurable space. Let X be an S-valued random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Then*

  (i) $P(X \in A \mid \mathcal{G}) \in [0,1]$ a.s., for all $A \in \mathcal{S}$.

  (ii) $P(X \in \varnothing \mid \mathcal{G}) = 0$ a.s., for all $A \in \mathcal{S}$.

  (iii) $P(X \in S \mid \mathcal{G}) = 1$ a.s., for all $A \in \mathcal{S}$.

(iv) $P(X \in \biguplus_{n=1}^{\infty} A_n \mid \mathcal{G}) = \sum_{n=1}^{\infty} P(X \in A_n \mid \mathcal{G})$ *a.s., for all disjoint collections* $\{A_n\} \subset \mathcal{S}$.

Exercise 4.2. Prove Theorem 4.1.

Unfortunately, Theorem 4.1 does not show that $A \mapsto P(X \in A \mid \mathcal{G})(\omega)$ is a measure for P-a.e. $\omega \in \Omega$. This is because the null set in (iv) can depend on the collection $\{A_n\}$. So there may not exist a single event of probability one on which (iv) holds simultaneously for all disjoint collections.

However, when X takes values in a space that is "nice", such as the real line, it is possible to express $P(X \in \cdot \mid \mathcal{G})$ as a genuine random measure. The remainder of this sections elaborates on this topic.

4.2   Random measures

Let $(S, \mathcal{S})$ be a measurable space and let $M(S)$ be the space of all σ-finite measures on $(S, \mathcal{S})$. Let $\mathcal{M}(S)$ be the σ-algebra on $M(S)$ generated by sets of the form $\{\nu : \nu(A) \in B\}$, where $A \in \mathcal{S}$ and $B \in \mathcal{R}$. Note that $\mathcal{M}(S)$ is the smallest σ-algebra such that the projection functions $\pi_A : M(S) \to \mathbb{R}$, defined by $\pi_A(\nu) = \nu(A)$, are Borel measurable for all $A \in \mathcal{S}$. Taking $A = S$ and $B = \{1\}$ shows that $M_1(S)$, the space of all probability measures on $(S, \mathcal{S})$ is measurable. Let $\mathcal{M}_1(S)$ denote $\mathcal{M}(S)$ restricted to $M_1(S)$.

Let $(T, \mathcal{T})$ be another measurable space. If $\mu : T \to M(S)$, we will write $\mu(t, A) = (\mu(t))(A)$. Note that μ is $(\mathcal{T}, \mathcal{M}(S))$-measurable if and only if $\pi_A \circ \mu = \mu(\cdot, A)$ is $(\mathcal{T}, \mathcal{R})$-measurable for all $A \in \mathcal{S}$. Any such measurable function is called a **kernel** from T to S. If μ takes values in $M_1(S)$, then μ is a **probability kernel**.

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A **random measure** on S is an $M(S)$-valued random variable. In other words, it is a kernel from Ω to S. If a random measure takes values in $M_1(S)$, then it is a **random probability measure** on S. Note that if μ is a kernel from T to S and Y is a T-valued random variables, then $\mu(T)$ is a random measure.

4.3   Regular conditional distributions

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ and $(T, \mathcal{T})$ measurable spaces. Let X and Y be S- and T-valued random variables, respectively. Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. If there exists a random measure $\mu = \mu_{X, \mathcal{G}}$ on S such that $P(X \in A \mid \mathcal{G}) = \mu(\cdot, A)$ a.s. for every $A \in \mathcal{S}$, then μ is a **regular conditional distribution** for X given $\mathcal{G}$, and we write $X \mid \mathcal{G} \sim \mu$. Similarly, if there exists a probability kernel $\mu = \mu_{X, Y}$ from T to S such that $P(X \in A \mid Y) = \mu(Y, A)$ a.s. for every $A \in \mathcal{S}$, then $\mu(Y)$ is a regular conditional distribution for X given $\sigma(Y)$, and we write $X \mid Y \sim \mu(Y)$.

A measurable space $(S, \mathcal{S})$ is a **Borel space** if there exists an $(\mathcal{S}, \mathcal{R})$-measurable bijection $\varphi : S \to \mathbb{R}$ such that φ^{-1} is $(\mathcal{R}, \mathcal{S})$-measurable. The following two theorems are Theorem 1.4.12 [1] and Theorem 5.3 in [2], respectively.

Theorem 4.3. *If S is a Borel subset of a complete separable metric space, and $\mathcal{S}$ is the collection of Borel subsets of S, then $(S, \mathcal{S})$ is a Borel space.*

Theorem 4.4. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ and $(T, \mathcal{T})$ measurable spaces. Let X and Y be S- and T-valued random variables, respectively. If S is a Borel space, then there exists a probability kernel $\mu = \mu_{X,Y}$ from T to S such that $X \mid Y \sim \mu(Y)$. If $\widetilde{\mu}$ is another such probability kernel, then $\mu = \widetilde{\mu}$, μ_Y-a.e.*

Corollary 4.5. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ a measurable space. Let X be an S-valued random variable and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. If S is a Borel space, then there exists a $\mathcal{G}$-measurable random probability measure $\mu = \mu_{X,\mathcal{G}}$ such that $X \mid \mathcal{G} \sim \mu$. If $\widetilde{\mu}$ is another such random probability measure, then $\mu = \widetilde{\mu}$ a.s.*

Proof. Apply Theorem 4.4 with $(T, \mathcal{T}) = (\Omega, \mathcal{G})$ and Y the identity function. $\square$

The first example of what we can do with regular conditional distributions is the following theorem, which can be regarded as a generalized version of Theorem 3.10.

Theorem 4.6. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ a measurable space. Let X be an S-valued random variable, $\mathcal{G} \subset \mathcal{F}$ a σ-algebra, and suppose $X \mid \mathcal{G} \sim \mu$. Let $(T, \mathcal{T})$ be a measurable space and Y a T-valued random variable. Let $f : S \times T \to \mathbb{R}$ be $(\mathcal{S} \times \mathcal{T}, \mathcal{R})$-measurable with $E|f(X, Y)| < \infty$. If $Y \in \mathcal{G}$, then*

$$E[f(X, Y) \mid \mathcal{G}] = \int_S f(x, Y) \, \mu(\cdot, dx) \quad a.s.$$

Proof. If $f = 1_{A \times B}$, where $A \in \mathcal{S}$ and $B \in \mathcal{T}$, then

$$E[1_{A \times B}(X, Y) \mid \mathcal{G}] = 1_B(Y) P(X \in A \mid \mathcal{G}) = 1_B(Y)\mu(\cdot, A) = \int_S 1_{A \times B}(x, Y) \, \mu(\cdot, dx) \quad \text{a.s.}$$

By the π-λ theorem, this proves the result for $f = 1_C$, where $C \in \mathcal{S} \times \mathcal{T}$. By linearity (Theorem 3.7), the result holds for all simple functions f. By monotone convergence (Theorem 3.13), the result holds for all nonnegative functions f satisfying $E|f(X, Y)| < \infty$. And finally, by considering the positive and negative parts, the result holds for all measurable functions f satisfying $E|f(X, Y)| < \infty$. $\square$

Corollary 4.7. *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(S, \mathcal{S})$ a measurable space. Let X be an S-valued random variable, $\mathcal{G} \subset \mathcal{F}$ a σ-algebra, and suppose $X \mid \mathcal{G} \sim \mu$. If $f : S \to \mathbb{R}$ is $(\mathcal{S}, \mathcal{R})$-measurable with $E|f(X)| < \infty$, then*

$$E[f(X) \mid \mathcal{G}] = \int_S f(x) \, \mu(\cdot, dx) \quad a.s.$$

Proof. Apply Theorem 4.6 with Y a constant random variable. $\square$

For our second example, we give a simple proof of Theorem 3.19 (Hölder's inequality).

Proof of Theorem 3.19. Since $(\mathbb{R}^2, \mathcal{R}^2)$ is a Borel space, there exists a random measure μ on $\mathcal{R}^2$ such that $(X, Y) \mid \mathcal{G} \sim \mu$. Since $|X|^p$ and $|Y|^q$ are integrable, the ordinary Hölder's inequality implies $|XY|$ is integrable. Thus, by Theorem 4.6,

$$E[|XY| \mid \mathcal{G}] = \int_{\mathbb{R}^2} |xy| \, \mu(\cdot, dx \, dy) \quad \text{a.s.}$$

For P-a.e. $\omega \in \Omega$, we may apply the ordinary Hölder's inequality to the measure $\mu(\omega, \cdot)$, yielding

$$E[|XY| \mid \mathcal{G}] \leqslant \left(\int_{\mathbb{R}^2} |x|^p \, \mu(\cdot, dx\, dy) \right)^{1/p} \left(\int_{\mathbb{R}^2} |y|^q \, \mu(\cdot, dx\, dy) \right)^{1/q} \quad \text{a.s.}$$

Applying Theorem 4.6 once again finishes the proof. $\qquad\square$

For our final example, let us first consider a property of unconditioned expectations. If X is a real-valued random variable, and $h : \mathbb{R} \to [0, \infty)$ is absolutely continuous with $h' \geqslant 0$ a.e. and $h(x) \downarrow 0$ as $x \to -\infty$, then

$$E[h(X)] = \int_{\mathbb{R}} h'(t) P(X > t)\, dt.$$

This is the content of Exercise 2.2.7 in [1], and one way to see it is to use Fubini's theorem:

$$E[h(X)] = \int_{\Omega} h(X)\, dP = \int_{\Omega} \int_{-\infty}^{X} h'(t)\, dt\, dP = \int_{\Omega} \int_{\mathbb{R}} 1_{\{X > t\}} h'(t)\, dt\, dP$$

$$= \int_{\mathbb{R}} h'(t) \int_{\Omega} 1_{\{X > t\}}\, dP\, dt = \int_{\mathbb{R}} h'(t) P(X > t)\, dt.$$

It is then natural to ask whether a similar thing is true for conditional expectations, and whether a similar proof can demonstrate it. We will answer both questions in the affirmative by using regular conditional probabilities.

Theorem 4.8. *Let X be a real-valued random variable on a probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{G} \subset \mathcal{F}$ a σ-algebra. Let $h : \mathbb{R} \to [0, \infty)$ be absolutely continuous with $h' \geqslant 0$ a.e. and $h(x) \downarrow 0$ as $x \to -\infty$. Suppose that $E|h(X)| < \infty$. Then it is possible to choose, for each $t \in \mathbb{R}$, a version of $P(X > t \mid \mathcal{G})$ so that the function $t \mapsto h'(t) P(X > t \mid \mathcal{G})$ is almost surely Lebesgue integrable on $\mathbb{R}$, and satisfies*

$$E[h(X) \mid \mathcal{G}] = \int_{\mathbb{R}} h'(t) P(X > t \mid \mathcal{G})\, dt \quad \text{a.s.}$$

Proof. Since $(\mathbb{R}, \mathcal{R})$ is a Borel space, there exists a random measure μ such that $X \mid \mathcal{G} \sim \mu$. For each $t \in \mathbb{R}$, let us choose the version of $P(X > t \mid \mathcal{G})$ determined by μ, that is, $P(X > t \mid \mathcal{G})(\omega) = \mu(\omega, (t, \infty))$. Then for P-a.e. $\omega \in \Omega$, we have

$$\int_{\mathbb{R}} \int_{\mathbb{R}} 1_{\{x > t\}} h'(t)\, dt\, \mu(\omega, dx) = \int_{\mathbb{R}} h(x)\, \mu(\omega, dx) = E[h(X) \mid \mathcal{G}](\omega) < \infty.$$

By Fubini's theorem, the function

$$t \mapsto \int_{\mathbb{R}} 1_{\{x > t\}} h'(t)\, \mu(\omega, dx) = h'(t) \mu(\omega, (t, \infty)) = h'(t) P(X > t \mid \mathcal{G})(\omega)$$

is Lebesgue integrable on $\mathbb{R}$ and

$$\int_{\mathbb{R}} h'(t) P(X > t \mid \mathcal{G})(\omega)\, dt = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_{\{x > t\}} h'(t)\, \mu(\omega, dx)\, dt$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} 1_{\{x > t\}} h'(t)\, dt\, \mu(\omega, dx) = E[h(X) \mid \mathcal{G}](\omega),$$

which proves the theorem. $\qquad\square$

5   A preview of stochastic processes

A **stochastic process** is a collection of random variable $\{X(t) : t \in T\}$ indexed by some set T. We usually think of T as time. A **discrete time stochastic process** is where $T = \mathbb{N}$, in which case the process is just a sequence of random variables.

Let $\{X_n : n \in \mathbb{N}\}$ be a discrete time stochastic process. Define $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The σ-algebra $\mathcal{F}_n$ represents all the information at time n that we would have from observing the values $X_1, \ldots, X_n$. Note that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$.

More generally, a **filtration** is a sequence of σ-algebras $\{\mathcal{F}_n\}_{n=1}^\infty$ such that $\mathcal{F}_n \subset \mathcal{F}_{n+1}$. A stochastic process $\{X_n : n \in \mathbb{N}\}$ is said to be **adapted** to the filtration $\{\mathcal{F}_n\}_{n=1}^\infty$ if X_n is $\mathcal{F}_n$-measurable for all n. The special case $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ is called the **filtration generated by** X, and is denoted by $\{\mathcal{F}_n^X\}_{n=1}^\infty$.

An important class of discrete time stochastic processes is the martingales. A stochastic process $\{X_n : n \in \mathbb{N}\}$ is a **martingale** with respect to the filtration $\{\mathcal{F}_n\}_{n=1}^\infty$ if

(i) X_n is integrable for all n,

(ii) $\{X_n : n \in \mathbb{N}\}$ is adapted to $\{\mathcal{F}_n\}_{n=1}^\infty$, and

(iii) $E[X_{n+1} \mid \mathcal{F}_n] = X_n$ for all n.

The critical item is (iii). Imagine that X_n models our cumulative wealth as we play a sequence of gambling games. Condition (iii) says that, given all the information up to time n, our expected wealth at time $n+1$ is the same as our wealth at time n. In other words, a martingale models a "fair" game.

Another important class of discrete time stochastic processes is the Markov chains. A stochastic process $\{X_n : n \in \mathbb{N}\}$ is a **Markov chain** with respect to the filtration $\{\mathcal{F}_n\}_{n=1}^\infty$ if

(i) $\{X_n : n \in \mathbb{N}\}$ is adapted to $\{\mathcal{F}_n\}_{n=1}^\infty$, and

(ii) $P(X_{n+1} \in B \mid \mathcal{F}_n) = P(X_{n+1} \in B \mid X_n)$ for all $B \in \mathcal{R}$.

Here, the critical item is (ii). It is called the **Markov property**. In words, it says that the conditional distribution of X_{n+1} given all the information up to time n is the same as if we were only given X_n. In other words, the future behavior of a Markov chain depends only on the present location of the chain, and not on how it got there.

The canonical example of a Markov chain is a random walk. If $\{X_j\}_{j=1}^\infty$ are iid and $S_n = X_1 + \cdots + X_n$, then $\{S_n : n \in \mathbb{N}\}$ is a **random walk**. The random walk is a Markov chain with respect to the filtration generated by S. Moreover, if each X_j is integrable with mean zero, then the random walk is also a martingale.

A continuous time stochastic process has the form $\{X(t) : t \in [0, \infty)\}$. Examples include the Poisson process and Brownian motion. Concepts such as filtrations, adaptedness, martingales, and the Markov property can all be extended to continuous time. Care is needed however, because (for one thing) the time domain is uncountable. Brownian motion is the continuous time analog of a random walk. It is the canonical example in continuous time of both a martingale and a Markov process. It can be realized as the limit of a sequence of random walks, where the step sizes are becoming smaller and the steps are occurring more frequently.

More specifically, let $\{S_n : n \in \mathbb{N}\}$ be a mean zero random walk. Let $S(t) = S_{\lfloor t \rfloor}$, where $\lfloor \cdot \rfloor$ denotes the greatest integer function. Then the sequence of processes

$$\left\{ \frac{S(nt)}{\sqrt{n}} : t \in [0, \infty) \right\}$$

converges (in a certain sense) as $n \to \infty$ to a continuous time stochastic process called Brownian motion. This is the conclusion of Donsker's theorem, which is a kind of central limit theorem for stochastic processes.

Differential equations that involve Brownian motion are referred to as stochastic differential equations (SDEs). SDEs are used to model dynamical systems that involve randomness, and are very common in scientific applications. In order to understand SDEs, one must first understand the stochastic integral (with respect to Brownian motion), which behaves quite differently from the ordinary Lebesgue-Stieltjes integral. In particular, the classical fundamental theorem of calculus no longer applies when one is working with stochastic integrals. It must be replaced by a new rule called Itô's rule. Itô's rule gives rise to a whole new calculus called stochastic calculus.

References

[1] Rick Durrett. *Probability: theory and examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[2] Olav Kallenberg. *Foundations of modern probability.* Probability and its Applications (New York). Springer-Verlag, New York, 1997.