

Introduction to Conditional Probabilities and Expectations

Steve Cheng

March 2, 2008

Contents

1 Basic definition	2
2 Intuitive explanations	2
3 Conditional expectation given a random variable	3
4 Basic properties of the conditional expectation	4
5 Conditional probabilities	6
6 Convergence properties of the conditional expectation	8
7 Conditional probability measures	9
8 Calculating conditional expectations	13
9 Change of variable	14
10 What is being conditioned on can be set constant	15
11 Hilbert space theory of conditional expectations	16
12 Bibliography	18

Purpose

In this note, we give a rigorous definition of conditional probabilities and expectations, and some fundamental results about them. We assume that the reader already is familiar with the intuitive notion of conditional probabilities ($P(A | B)$ for $P(B) > 0$). For our exposition, we will also depend on, of course, some measure theory, including the Lebesgue-Radon-Nikodym theorem.

The author has written this note because he still does not readily encounter introductions to conditional probability that are theoretically rigorous and yet not afraid to delve into, explain and justify the intuition behind the concepts. (Though J. Michael Steele's

book referenced in the bibliography comes close, even as that author remarks that the abstract definition of conditional probability is “not easy to love; fortunately, love is not required.”)

Copyright matters

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, with no Front-Cover Texts, and with no Back-Cover Texts.

1 Basic definition

Let (Ω, Σ, P) be a probability space. Suppose we are given $\mathcal{G} \subseteq \Sigma$, a σ -algebra of Ω , and $Y: \Omega \rightarrow \mathbb{R}$, a random variable with $\mathbb{E}|Y| < \infty$. We define the *conditional expectation of Y given \mathcal{G}* by the following procedure.

Let ν be a signed measure given by

$$\nu(A) = \mathbb{E}[Y 1_A], \quad \text{for } A \in \mathcal{G}.$$

Clearly $\nu(\emptyset) = 0$, while if A_1, A_2, \dots are disjoint events, then

$$\begin{aligned} \nu(A_1 \cup A_2 \cup \dots) &= \mathbb{E}[Y 1_{A_1 \cup A_2 \cup \dots}] \\ &= \mathbb{E}[Y (1_{A_1} + 1_{A_2} + \dots)] \\ &= \mathbb{E}[Y 1_{A_1}] + \mathbb{E}[Y 1_{A_2}] + \dots \\ &= \nu(A_1) + \nu(A_2) + \dots \end{aligned}$$

Let $P|_{\mathcal{G}}$ denote the restriction of P to \mathcal{G} . If $P|_{\mathcal{G}}(A) = 0$ then $\mathbb{E}[Y 1_A] = 0$, so the measure ν is absolutely continuous with respect to $P|_{\mathcal{G}}$. By the Lebesgue-Radon-Nikodym theorem, there exists a \mathcal{G} -measurable function $g: \Omega \rightarrow \mathbb{R}$, such that

$$\mathbb{E}[Y 1_A] = \nu(A) = \int_A g dP|_{\mathcal{G}} = \mathbb{E}[g 1_A], \quad \text{for all } A \in \mathcal{G}. \quad (1)$$

Moreover, the theorem says that the function g is actually unique up to $P|_{\mathcal{G}}$ -null sets.

Definition 1.1. The *conditional expectation of Y given \mathcal{G}* , denoted by $\mathbb{E}[Y | \mathcal{G}]$, is defined to be any of the \mathcal{G} -measurable functions $g: \Omega \rightarrow \mathbb{R}$ that satisfy equation (1).

Though there are many functions g that are candidates, each pair of them only differ on a set of $P|_{\mathcal{G}}$ -measure zero.

As the conditional expectation can only be defined when Y is integrable ($\mathbb{E}|Y| < \infty$), we will tacitly make such assumptions in our work unless stated otherwise.

2 Intuitive explanations

We now want to give some intuitions about the strange definition of conditional expectation just given.

Firstly, a σ -algebra \mathcal{G} on the sample space Ω , roughly speaking, is a lens or filter through which we observe what is happening in the probability space. If \mathcal{G} is coarse and contains few events — the extreme example being $\mathcal{G} = \{\emptyset, \Omega\}$ — then we can calculate probabilities for those events only. To say that a function $g: \Omega \rightarrow \mathbb{R}$ is \mathcal{G} -measurable for a smaller σ -algebra \mathcal{G} , means that it is blurred by looking at it through \mathcal{G} . This vague statement is made precise by this simple fact:

Proposition 2.1. *Let (Ω, \mathcal{G}) be a measurable space, and $g: \Omega \rightarrow \mathbb{R}$ be a \mathcal{G} -measurable function. If $\omega_1, \omega_2 \in \Omega$ are never separated in \mathcal{G} , i.e. there is no set in \mathcal{G} that contains one of these points but not the other, then $g(\omega_1) = g(\omega_2)$.*

Proof. Let $B = \{g(\omega_1)\}$. As $g^{-1}(B)$ is \mathcal{G} -measurable and contains ω_1 , it must contain ω_2 . ■

We can now say that, $\mathbb{E}[Y | \mathcal{G}]$ is an average value of the random variable Y when it is viewed through the lens of a more constricting σ -algebra \mathcal{G} . It is an average value, because, when its expected value is taken in equation (1), it is the same as the expectation of Y , across *all subsets* of \mathcal{G} . The function $\mathbb{E}[Y | \mathcal{G}]$ is a “blurred” version of Y where we cannot tell the difference in the value of Y , at sample points that are not separated in \mathcal{G} .

To further illustrate the intuitive explanations, consider conditioning Y on the trivial σ -algebra $\mathcal{G} = \{\emptyset, \Omega\}$, which separates no points. The \mathcal{G} -measurable function $\mathbb{E}[Y | \mathcal{G}]$ must clearly be a constant. By equation (1), the constant must be $\mathbb{E}Y$, i.e. the average value of Y on the whole sample space.

Another important case to consider is when Y is already \mathcal{G} -measurable — that is, when

$$\mathcal{G} \supset \sigma(Y) = \{Y^{-1}(B) : B \text{ is a Borel set in } \mathbb{R}\},$$

or that \mathcal{G} separates points finely enough for Y . Then $\mathbb{E}[Y | \mathcal{G}] = Y$ immediately from the definition. So constricting Y to \mathcal{G} just amounts to doing nothing at all.

3 Conditional expectation given a random variable

We commonly condition random variables given the σ -algebra

$$\mathcal{G} = \sigma(X) = \{X^{-1}(B) : B \text{ is measurable}\}$$

that is generated by another random variable X .

If $X: \Omega \rightarrow \mathbb{R}$ takes the same value at two sample points $\omega_1, \omega_2 \in \Omega$, then $\mathcal{G} = \sigma(X)$ does not separate ω_1 and ω_2 . Therefore, by Proposition 2.1, $\mathbb{E}[Y | \mathcal{G}](\omega)$ takes on the same values wherever $X(\omega)$ are at the same values. In other words, $\mathbb{E}[Y | \mathcal{G}]$ is really a function of X ; that is, $\mathbb{E}[Y | \mathcal{G}] = h \circ X$ for some $h: \mathbb{R} \rightarrow \mathbb{R}$.

We will want to ascertain some measurability properties of h , and to do this it is best to go modify the original construction of $\mathbb{E}[Y | \mathcal{G}]$ to produce the function h directly.

Given a measurable function $X: \Omega \rightarrow \Omega'$, where (Ω', Σ') is another measurable space¹, define the signed measure ν' on (Ω', Σ') by

$$\nu'(B) = \mathbb{E}[Y 1_{X \in B}] = \mathbb{E}[Y 1_{X^{-1}(B)}], \quad \text{for all } B \in \Sigma'.$$

¹The most simplest case is $\Omega' = \mathbb{R}$ and $\Sigma' = \mathcal{B}_{\mathbb{R}}$, but Ω' could also be \mathbb{R}^n . Usually we will assume that Σ' contains all singleton sets; otherwise Proposition 2.1 might fail to hold with \mathbb{R} replaced by Ω' . Also we want to be allowed to talk about seemingly simple sets like $\{X = x\}$.

Let P_X be the pull-back measure $B \mapsto P(X^{-1}(B))$. If $P_X(B) = 0$, then $\nu'(B) = 0$, so ν' is absolutely continuous with respect to P_X . By the Lebesgue-Radon-Nikodym theorem, we obtain a measurable function $h: \Omega' \rightarrow \mathbb{R}$, unique up to P_X -null sets, such that

$$\mathbb{E}[Y 1_{X \in B}] = \nu'(B) = \int_B h dP_X, \quad \text{for all } B \in \Sigma'. \quad (2)$$

By a change of variables with $A = X^{-1}(B) \in \mathcal{G} = \sigma(X)$, we have

$$\begin{aligned} \int_A h \circ X dP|_{\mathcal{G}} &= \int_B h dP_X \\ &= \mathbb{E}[Y 1_{X \in B}] = \mathbb{E}[Y 1_E] = \int_A g dP|_{\mathcal{G}}, \quad g = \mathbb{E}[Y | \mathcal{G}]. \end{aligned}$$

Since the integrals on the extreme left and right are equal on all $A \in \mathcal{G}$, the integrands $h \circ X$ and g must be equal $P|_{\mathcal{G}}$ -almost surely.

Definition 3.1. The symbol $\mathbb{E}[Y | X]$ is to mean the same thing as $\mathbb{E}[Y | \sigma(X)]$. It is called the *conditional expectation of Y given the random variable X* .

There always exists a measurable function $h: \Omega' \rightarrow \mathbb{R}$ such that $h(X(\omega)) = \mathbb{E}[Y | X](\omega)$ for $P|_{\mathcal{G}}$ -almost all $\omega \in \Omega$. The value $h(x)$ is often denoted by $\mathbb{E}[Y | X = x]$, even though it may not always be well-defined as a single number. When we use this notation we will usually have a specific version of the function h in mind.

Equation (2) stated in the new notation becomes:

$$\mathbb{E}[Y 1_{X \in B}] = \mathbb{E}[\mathbb{E}[Y | X] 1_{X \in B}] = \int_{x \in B} \mathbb{E}[Y | X = x] dP_X, \quad \text{for all } B \in \Sigma'. \quad (3)$$

To summarize, the construction just given is basically the same as the first construction used for $\mathbb{E}[Y | \mathcal{G}]$, only that (Ω', Σ', P_X) replaces $(\Omega, \Sigma, P|_{\mathcal{G}})$ in the original. Since they are so similar, when we discuss properties about $\mathbb{E}[Y | \mathcal{G}]$ and $\mathbb{E}[Y | X = x]$ hereafter, we will usually state and prove the properties for only $\mathbb{E}[Y | \mathcal{G}]$, as the modifications for $\mathbb{E}[Y | X = x]$ are trivial.

Intuitively, the function $\mathbb{E}[Y | X]$ answers the question: “What is the average value of Y given the values that the random variable X takes on?”

Or, the operation $\mathbb{E}[\cdot | X]$ can be thought of as extracting the part of a random variable that can be predicted as a function of X (a $\sigma(X)$ -measurable random variable).

Note the arguments in this section also give an easy-to-digest characterization of real-valued integrable functions² that are $\sigma(X)$ -measurable for some measurable X : they are exactly those functions that can be expressed as an integrable function of X almost everywhere.

4 Basic properties of the conditional expectation

The following property is trivial but well-known in naïve probability theory (when $\mathcal{G} = \sigma(X)$):

²This characterization is also valid in general measure theory, provided the measure involved is σ -finite.

Proposition 4.1 (The law of total probability). *For any real-valued random variable Y ,*

$$\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}]] = \mathbb{E}Y.$$

Proof. Take A to be the entire probability space $\Omega \in \mathcal{G}$ in equation (1). ■

In a sense, the conditional expectation has been defined in such a way that functions that are conditional expectations must satisfy a generalization of the law of total probability — that it must hold on all \mathcal{G} -measurable subsets and not just on Ω . The generalized law of total probability makes the satisfying functions unique up to sets of measure zero, whereas lots of other \mathcal{G} -measurable functions Z can satisfy only $\mathbb{E}[Z1_\Omega] = \mathbb{E}[Y1_\Omega]$, including of course the constant $Z = \mathbb{E}Y$.

It is easy to check the linearity property of conditional expectations, as suggested by the notation:

Proposition 4.2. *For constants $a, b, c \in \mathbb{R}$, almost surely*

$$\mathbb{E}[aX + bY + c \mid \mathcal{G}] = a\mathbb{E}[X \mid \mathcal{G}] + b\mathbb{E}[Y \mid \mathcal{G}] + c.$$

Proof. The Radon-Nikodym derivative operates linearly on signed measures, and obviously $\mathbb{E}[c \mid \mathcal{G}] = c$. ■

And the comparison property:

Proposition 4.3. *If $Y \geq 0$, then almost surely*

$$\mathbb{E}[Y \mid \mathcal{G}] \geq 0,$$

which in combination of Proposition 4.2 leads to the generalized triangle inequality:

$$\left| \mathbb{E}[Y \mid \mathcal{G}] \right| \leq \mathbb{E}[|Y| \mid \mathcal{G}].$$

Proof. The Radon-Nikodym derivative of a positive measure is almost everywhere non-negative. ■

The same properties hold when \mathcal{G} is replaced with a random variable X .

We end this section with the tower property of conditional expectation, that is so frequently used that it would not do justice to omit:

Proposition 4.4. *If \mathcal{G}_1 and \mathcal{G}_2 are two σ -algebras with $\mathcal{G}_1 \subseteq \mathcal{G}_2$, then for any random variable Y ,*

$$\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}_1] \mid \mathcal{G}_2] = \mathbb{E}[Y \mid \mathcal{G}_1] = \mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1].$$

Explanation: Filtering a random variable on a coarser σ -algebra, then a sharper/finer one, or vice versa, is the same as filtering just on the coarser one. Or, conditioning can only subtract from, and not add to, the randomness of a random variable.

Proof. The first equality is immediate: as $\mathbb{E}[Y \mid \mathcal{G}_1]$ is \mathcal{G}_1 -measurable by definition, and hence is already \mathcal{G}_2 measurable, so re-conditioning on \mathcal{G}_2 changes nothing.

For the second equality, we must show that $\mathbb{E}[\mathbb{E}[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1]$ is just another version of $\mathbb{E}[Y \mid \mathcal{G}_1]$, satisfying its defining equation (1). All we need to do is to let $B \in \mathcal{G}_1 \subseteq \mathcal{G}_2$ and push some symbols around:

$$\mathbb{E}[\mathbb{E}[(\mathbb{E}[Y \mid \mathcal{G}_2]) \mid \mathcal{G}_1] 1_B] = \mathbb{E}[(\mathbb{E}[Y \mid \mathcal{G}_2]) 1_B] = \mathbb{E}[Y 1_B]. \quad \blacksquare$$

Proposition 4.5. *If X is \mathcal{G} -measurable, then X can be factored out of a conditional expectation with respect to \mathcal{G} :*

$$\mathbb{E}[XY \mid \mathcal{G}] = X \mathbb{E}[Y \mid \mathcal{G}].$$

Proof. We have to prove that, for all $B \in \mathcal{G}$,

$$\mathbb{E}[X \mathbb{E}[Y \mid \mathcal{G}] 1_B] = \mathbb{E}[XY 1_B].$$

If $X = 1_A$ for some $A \in \mathcal{G}$, this is trivial. By linearity, if X is a simple random variable of the form

$$X = \sum_{i=1}^n x_i 1_{A_i}, \quad x_i \in \mathbb{R}, \quad A_i \in \mathcal{G},$$

the conclusion also holds.

Now assume, for the moment, that X and Y are non-negative. Take a sequence of random variables X_n that increase to X , and apply the monotone convergence theorem twice:

$$\mathbb{E}[(XY) 1_B] = \lim_{n \rightarrow \infty} \mathbb{E}[(X_n Y) 1_B] = \lim_{n \rightarrow \infty} \mathbb{E}[X_n \mathbb{E}[Y \mid \mathcal{G}] 1_B] = \mathbb{E}[X \mathbb{E}[Y \mid \mathcal{G}] 1_B].$$

Finally, if X and Y are not necessarily non-negative, write them in terms of their positive and negative parts, and apply linearity to obtain the same conclusion. \blacksquare

5 Conditional probabilities

We have delayed the discussion of conditional probabilities, because they are defined by a similar process as conditional expectations, and we want to avoid repeating proofs of what are mostly the same thing. But with conditional expectations available, they are easy to define:

Definition 5.1. The conditional probability of an event $A \in \Sigma$ given \mathcal{G} is the function

$$P(A \mid \mathcal{G}) = \mathbb{E}[1_A \mid \mathcal{G}].$$

Definition 5.2. The conditional probability of an event $A \in \Sigma$ given a random variable $X: \Omega \rightarrow \Omega'$ is the function

$$P(A \mid X) = \mathbb{E}[1_A \mid X].$$

Intuitively, the right-hand sides are “the average value of 1_A given \mathcal{G} (or X)”. This average value of the indicator function 1_A is the probability of the event A occurring.

Also, if given a value $X(\omega) = x \in \Omega'$, evaluating $P(A \mid X = x) = \mathbb{E}[1_A \mid X = x] = \mathbb{E}[1_A \mid X](\omega)$ gives the conditional probability of A occurring when we know that $X = x$. However, we cannot take this too literally, because if the event $[X = x]$ occurs with probability zero, $P(A \mid X = x)$ might not have a single value across all versions of the function $P(A \mid X)$.

Reinterpreting equations (1) and (3), we arrive at the following defining relations:

$$P(A \cap E) = \int_E P(A \mid E) dP_{|\mathcal{G}}, \quad E \in \mathcal{G}, \quad (4)$$

$$P(A \cap [X \in B]) = \int_{x \in B} P(A \mid X = x) dP_X, \quad B \in \Sigma'. \quad (5)$$

But these are just the *Bayes' rules* for conditional probabilities!

Definition 5.3. For any event $B \in \Sigma$ with positive probability of occurrence, we also define

$$\begin{aligned} P(A | B) &= P(A | 1_B = 1) \\ \mathbb{E}(Y | B) &= \mathbb{E}(Y | 1_B = 1). \end{aligned}$$

Applying equation (3) shows that $P(A | B)$ just defined agrees with the usual definition:

$$\begin{aligned} P(A \cap B) &= \mathbb{E}[1_A 1_B] = \int_{\{1\}} P(A | 1_B) dP_{1_B} \\ &= P(A | 1_B = 1) \cdot P(1_B = 1) = P(A | B) \cdot P(B), \end{aligned}$$

so that

$$P(A | B) = \frac{P(A \cap B)}{P(B)}. \quad (6)$$

If B is a null set, $P(A | B)$ is not well-defined, and this is of course where the traditional definition of $P(A | B)$ fails.

If B is the event $[X = x]$, the definition just made is (fortunately!) consistent with that of $P(A | X = x)$.

For the sake of completeness, we mention the following intuitive fact, obviously true for the naïve definition of conditional probability:

Proposition 5.1. *Let $A \in \Sigma$ be an event. Then $P(A | \mathcal{G}) = P(A)$ almost surely if and only if A is independent from all events in \mathcal{G} ,*

Proof. For the “if” direction: for all $E \in \mathcal{G}$,

$$\mathbb{E}[P(A) 1_E] = P(A) \cdot P(E) = P(A \cap E) = \mathbb{E}[1_A 1_E] = \mathbb{E}[P(A | \mathcal{G}) 1_E],$$

so $P(A)$ satisfies the condition for being $P(A | \mathcal{G})$. The converse follows from the same equation. ■

The proof of the following theorem requires a result from the next section. But rest assured, the next section does not depend on the results here, nor is the present result crucial to understand immediately. The material is put into this order only to make it easier to digest, and to build intuition on how independence and conditional expectations are related.

Corollary 5.2. $\mathbb{E}[f(Y) | \mathcal{G}] = \mathbb{E}[f(Y)]$ for all P_Y -integrable functions f , if and only if Y is a random variable independent of all the events in \mathcal{G} .

(The random variable Y can have a codomain Ω' that is not necessarily \mathbb{R} .)

Proof. Note that Y being independent of all events in \mathcal{G} means exactly that the events $[Y \in B]$ are independent of all events in \mathcal{G} .

For the “if” direction, Proposition 5.1 already proves the case for $f = 1_B$ for measurable sets $B \subseteq \Omega'$, setting $A = [Y \in B]$. For arbitrary integrable f , approximate it with a sequence of linear combinations f_n of indicator functions converging to it pointwise everywhere and dominated by $|f|$ itself. Employing dominated convergence (Proposition 6.4) to take limits, we find $\mathbb{E}[f(Y) | \mathcal{G}] = \mathbb{E}[f(Y)]$.

For the “only if” direction, simply select the particular cases $f = 1_B$ for measurable sets $B \subseteq \Omega'$, and apply the “only if” direction from Proposition 5.1. ■

For a little intuition of Corollary 5.2, take the example of $\mathcal{G} = \sigma(X)$ and $h = \text{identity}$. If Y is independent from X , then Y cannot be expressed as a function of X at all, unless it is constant.³ Thus knowing the value of X cannot possibly give additional information on Y — this is the content of the equation $\mathbb{E}[Y | X] = \mathbb{E}[Y]$, the constant average value.

6 Convergence properties of the conditional expectation

This section is devoted to developing the analogues of the convergence theorems for normal expectations or integrals.

Theorem 6.1. *Taking conditional expectations is a \mathbf{L}^1 contraction:*

$$\mathbb{E}\left[\left|\mathbb{E}[Y | \mathcal{G}]\right|\right] \leq \mathbb{E}|Y| \quad \text{or equivalently,} \quad \left\|\mathbb{E}[Y | \mathcal{G}]\right\|_{\mathbf{L}^1} \leq \|Y\|_{\mathbf{L}^1}.$$

Proof. $\mathbb{E}\left[\left|\mathbb{E}[Y | \mathcal{G}]\right|\right] \leq \mathbb{E}\left[\mathbb{E}[|Y| | \mathcal{G}]\right] = \mathbb{E}|Y|$ using the generalized triangle inequality (Proposition 4.3). ■

Corollary 6.2. *If the real-valued random variables X_1, X_2, \dots converge to X in \mathbf{L}^1 , then $\mathbb{E}[X_n | \mathcal{G}]$ converge to $\mathbb{E}[X | \mathcal{G}]$ in \mathbf{L}^1 .*

Proof. Apply the previous theorem to $Y = Y_n = X_n - X$ and take $n \rightarrow \infty$. ■

Theorem 6.3 (Monotone convergence). *If the non-negative random variables X_n are increasing almost surely to a random variable X , and then the conditional expectations $\mathbb{E}[X_n | \mathcal{G}]$ increase almost surely to $\mathbb{E}[X | \mathcal{G}]$.*

Direct proof. First of all, we know that almost surely $\mathbb{E}[X_n | \mathcal{G}]$ are increasing and bounded by $\mathbb{E}[X | \mathcal{G}]$ (justified by Proposition 4.3), but whether they increase to the bound $\mathbb{E}[X | \mathcal{G}]$ is still in question. To settle this, let B be the event where $\sup_n \mathbb{E}[X_n | \mathcal{G}]$ is less than $\mathbb{E}[X | \mathcal{G}] - \epsilon$, for some fixed $\epsilon > 0$.

Notice that $B \in \mathcal{G}$ because both $\mathbb{E}[X_n | \mathcal{G}]$ and $\mathbb{E}[X | \mathcal{G}]$ are, by definition, \mathcal{G} -measurable. Then:

$$\begin{aligned} \mathbb{E}[X 1_B] &= \lim_{n \rightarrow \infty} \mathbb{E}[X_n 1_B] && \text{normal monotone convergence} \\ &= \lim_{n \rightarrow \infty} \mathbb{E}[\mathbb{E}[X_n | \mathcal{G}] 1_B] && \text{definition of conditional expectation} \\ &= \mathbb{E}\left(\lim_{n \rightarrow \infty} \mathbb{E}[X_n | \mathcal{G}] 1_B\right) && \text{normal monotone convergence} \\ &\leq \mathbb{E}\left(\lim_{n \rightarrow \infty} (\mathbb{E}[X | \mathcal{G}] - \epsilon) 1_B\right) \\ &= \mathbb{E}[X 1_B] - \epsilon P(B), \end{aligned}$$

and this is impossible unless $P(B) = 0$. Letting $\epsilon \searrow 0$, we see that the event where $\mathbb{E}[X_n | \mathcal{G}]$ does not increase to $\mathbb{E}[X | \mathcal{G}]$ must have probability zero. ■

³This conclusion also follows if Y is merely uncorrelated with all functions of X . For a proof, refer to the last remarks in Section 2.

Simpler proof. Apply the next result on dominated convergence, as X_n are dominated by X . Of course, normally we cannot reduce the Dominated Convergence Theorem to the Monotone Convergence Theorem this way, because $\mathbb{E}|X|$ in general might be infinite, but in that case $\mathbb{E}[X | \mathcal{G}]$ cannot be defined anyway. ■

Another formulation of the monotone convergence is possible that perhaps does not so trivially reduce to dominated convergence. Suppose we do not assume that X has finite mean, but that $Z = \lim_n \mathbb{E}[X_n | \mathcal{G}]$ has finite mean. Then we can conclude that X has finite mean and $Z = \mathbb{E}[X | \mathcal{G}]$.

Theorem 6.4 (Dominated convergence). *If the random variables X_n converge to a random variable X almost surely, and $|X_n|$ are dominated by another random variable Z with finite expectation, then almost surely,*

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X| | \mathcal{G}] = 0 \quad \text{and hence} \quad \lim_{n \rightarrow \infty} \mathbb{E}[X_n | \mathcal{G}] = \mathbb{E}[X | \mathcal{G}].$$

Proof. Set $Y_n = |X_n - X|$. Then almost surely, Y_n are dominated by $2Z$, while $\mathbb{E}[Y_n | \mathcal{G}]$ are dominated by $2\mathbb{E}[Z | \mathcal{G}]$. The last quantity has a finite expectation of $2\mathbb{E}Z$. For any $B \in \mathcal{G}$, we have:

$$\begin{aligned} & \mathbb{E}\left[\left(\lim_{n \rightarrow \infty} \mathbb{E}[Y_n | \mathcal{G}]\right) 1_B\right] \\ &= \lim_{n \rightarrow \infty} \mathbb{E}[\mathbb{E}[Y_n | \mathcal{G}] 1_B] && \text{dominated convergence on } \mathbb{E}[Y_n | \mathcal{G}] 1_B \\ &= \lim_{n \rightarrow \infty} \mathbb{E}[Y_n 1_B] \\ &= 0 && \text{dominated convergence on } Y_n. \end{aligned}$$

Thus the \mathcal{G} -measurable random variable $\lim_n \mathbb{E}[Y_n | \mathcal{G}]$ is equal to 0 almost surely. ■

7 Conditional probability measures

Suppose X is a discrete random variable. If we look at $P(A | X = x) = P(A \cap [X = x]) / P([X = x])$, as a function of A , it describes a probability measure with each fixed x . So we are led to ask: is $A \mapsto P(A | \mathcal{G})(\omega)$ a probability measure for each fixed $\omega \in \Omega$?

The answer is no, generally. For one thing, $P(A | \mathcal{G})(\omega)$ is not always well-defined for pointwise ω . We can see in more detail what goes wrong when we attempt to prove that it is a probability measure.

Proposition 7.1. *Let (Ω, Σ, P) be a probability space, and $\mathcal{G} \subseteq \Sigma$ be a σ -algebra.*

- i. If $A \in \Sigma$ is any event, then $0 \leq P(A | \mathcal{G}) \leq 1$ almost surely (with respect to $P|_{\mathcal{G}}$).*
- ii. $P(\emptyset | \mathcal{G}) = 0$ and $P(\Omega | \mathcal{G}) = 1$ almost surely.*
- iii. Given disjoint events $A_1, A_2, \dots \in \Sigma$, we have almost surely*

$$P(A_1 \cup A_2 \cup \dots | \mathcal{G}) = P(A_1 | \mathcal{G}) + P(A_2 | \mathcal{G}) + \dots$$

Proof.

i. Let $E = [P(A | \mathcal{G}) \leq 0]$. Then

$$0 \leq P(A \cap E) = \int_E P(A | \mathcal{G}) dP_{|\mathcal{G}} \leq 0,$$

so $P(A | \mathcal{G}) = 0$ almost surely on E . i.e. $P(A | \mathcal{G})$ almost never takes on negative values. Similarly, if $E = [P(A | \mathcal{G}) \geq 1]$,

$$P(E) \geq P(A \cap E) = \int_E P(A | \mathcal{G}) dP_{|\mathcal{G}} \geq P(E), \quad \int_E (P(A | \mathcal{G}) - 1) dP_{|\mathcal{G}} = 0.$$

i.e. $P(A | \mathcal{G})$ is almost never greater than one.

ii.

$$1 = P(\Omega) = \int_{\Omega} P(\Omega | \mathcal{G}) dP_{|\mathcal{G}}, \quad \int_{\Omega} (1 - P(\Omega | \mathcal{G})) dP_{|\mathcal{G}} = 0.$$

By (i), the integrand of the last integral is non-negative, and therefore $P(\Omega | \mathcal{G}) = 1$ almost surely. Similarly,

$$0 = P(\emptyset) = P(\emptyset \cap \Omega) = \int_{\Omega} P(\emptyset | \mathcal{G}) dP_{|\mathcal{G}}$$

implies that $P(\emptyset | \mathcal{G})$ must be zero almost surely.

iii. Since for all $E \in \mathcal{G}$,

$$\begin{aligned} \int_E P\left(\bigcup_n A_n | \mathcal{G}\right) dP_{|\mathcal{G}} &= P\left(\bigcup_n A_n \cup E\right) = \sum_n P(A_n \cup E) \\ &= \sum_n \int_E P(A_n | \mathcal{G}) dP_{|\mathcal{G}} = \int_E \left(\sum_n P(A_n | \mathcal{G})\right) dP_{|\mathcal{G}}, \end{aligned}$$

the first and final integrands are equal almost surely. (The exchange of summation and integration is allowed since the integrands are non-negative.) ■

The key phrase in Proposition 7.1 is *almost surely*. There is a $P_{|\mathcal{G}}$ -null set $N \subset \Omega$ where the (in)equalities may fail. The set N depends on the events A , because each conditional probability is separately constructed for each A . There may not necessarily exist a null set N for which the (in)equalities hold everywhere else *for every event* $A \in \Sigma$.

On the other hand, if we were to identify a *countable* set of events A that we are interested in, then we avoid this problem. For each A there is an exceptional null set, and the countable union of all of these is again a null set; everywhere else on Ω all the relevant relations hold.

Given a random variable $Y: \Omega \rightarrow \mathbb{R}$, a category of events that we can look at are those of the form $Y^{-1}((-\infty, y])$. If we restrict $y \in \mathbb{Q} \cup \{-\infty, +\infty\}$. There are at most countably many of these, and yet knowing only their probabilities already determines the probability distribution of Y . So the idea is to construct a version of $B \mapsto P([Y \in B] | \mathcal{G})$, where almost at all sample points in Ω it becomes a probability distribution.

In our formal construction, we will also generalize to random variables that are \mathbb{R}^n -valued.

Theorem 7.2. *Let (Ω, Σ, P) be a probability space, $\mathcal{G} \subseteq \Sigma$ be a σ -algebra, and $Y: \Omega \rightarrow \mathbb{R}^n$ a random variable. For each $\omega \in \Omega$, there exists a probability measure μ_ω on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, such that for $P|_{\mathcal{G}}$ -almost all $\omega \in \Omega$,*

$$\mu_\omega(B) = P([Y \in B] | \mathcal{G})(\omega).$$

Proof.

Construction of the cumulative distribution function. Let \mathcal{D} be the countable collection of all events $D_y = (-\infty, y_1] \times \cdots \times (-\infty, y_n]$ for $y \in \mathbb{Q} \cup \{-\infty, +\infty\}$.

By taking countable unions of the exceptional null sets, we can obtain a null set N for which properties (i) and (ii) of Proposition 7.1 hold everywhere else for events in $Y^{-1}(\mathcal{D})$.

If $A, B \in \mathcal{D}$ and $A \subseteq B$, then we have

$$P([Y \in A] | \mathcal{G}) \leq P([Y \in B] | \mathcal{G}), \quad (7)$$

except on a null set. For each of the countably many pairs $(A, B) \in \mathcal{D} \times \mathcal{D}$ there is such a null set. Taking their union, we obtain a null set M for which relation (7) holds everywhere except on M .

For each $y \in \mathbb{Q}^n$, define

$$F_\omega(y) = P(Y \in D_y | \mathcal{G})(\omega), \quad \omega \in \Omega \setminus (N \cup M) \quad (8)$$

for one of the versions of $P(\cdots | \mathcal{G})$. And for $z \in \mathbb{R}^n \setminus \mathbb{Q}^n$, set

$$F_\omega(z) = \inf_{y \in \mathbb{Q}^n: z_j \leq y_j} F_\omega(y), \quad \omega \in \Omega \setminus (N \cup M). \quad (9)$$

We claim that F_ω is a multi-dimensional cumulative distribution function, for each $\omega \in \Omega \setminus (N \cup M)$. Clearly $F_\omega(z) \geq 0$ for all $z \in \mathbb{R}^n$. Also, by relation (7), F_ω is an increasing function of each argument if they are all rational. For non-rational arguments it is also seen that F_ω is increasing by virtue of equation (9).

These facts also mean that relation (9) can be broadened to:

$$F_\omega(z) = \inf_{y \in \mathbb{R}^n: z \neq y, z_j \leq y_j} F_\omega(y), \quad (10)$$

for $z \in \mathbb{R}^n \setminus \mathbb{Q}^n$.

Actually equation (10) holds for $z \in \mathbb{Q}^n$ as well. Firstly, because F_ω is increasing in each variable, the infimum can be taken over only points of the form $y_n = z + (1/n, \dots, 1/n)$, for $n \in \mathbb{N}$, with no change. And secondly, by Fatou's Lemma,

$$\begin{aligned} 0 &\leq \int_{\Omega} \liminf_{n \rightarrow \infty} [F_\omega(y_n) - F_\omega(z)] dP|_{\mathcal{G}} \\ &\leq \liminf_{n \rightarrow \infty} \int_{\Omega} [F_\omega(y_n) - F_\omega(z)] dP|_{\mathcal{G}} \\ &= \liminf_{n \rightarrow \infty} P(Y \in D_{y_n}) - P(Y \in D_z) = 0. \end{aligned}$$

Thus, $\liminf_n F_\omega(y_n) - F_\omega(z) = \inf_n F_\omega(y_n) - F_\omega(z) = 0$ for all $\omega \in \Omega$ except on a null set N_z . Provided we strip away these null sets too (for $z \in \mathbb{Q}^n$) at the beginning, equation (10), equivalent to right-continuity of F_ω , holds in general.

With the same sort of argument using Fatou's Lemma, we can prove that, save for a null set, $F_\omega \rightarrow 0$ if one of the variables tends to $-\infty$, and $F_\omega \rightarrow 1$ if all the variables tend to $+\infty$.

For those $\omega \in \Omega$ which are on the exceptional null sets, we can define $F_\omega(z)$ arbitrarily as $P([Y \in D_z])$.

Thus we now know F_ω is a cumulative distribution function, which then has a corresponding probability measure μ_ω .

$\mu(B)$ is the conditional probability. All that remains is to show that $\mu_\omega(B) = P([Y \in B] | \mathcal{G})(\omega)$.

The first point is that $\omega \mapsto \mu_\omega(B)$ should be \mathcal{G} -measurable. This is unfortunately somewhat technical: it involves the monotone class theorem, the same sort of argument used to prove measurability in Fubini's Theorem.

Let $\Sigma' = \{B \in \mathcal{B}_{\mathbb{R}^n} : \omega \mapsto \mu_\omega(B) \text{ is } \mathcal{G}\text{-measurable}\}$. By equation (8), $\mu(D)$ is measurable for all the sets in $D \in \mathcal{D}$. For finite disjoint unions and complements B of sets from \mathcal{D} , $\mu(B)$ is measurable too, because it can be obtained by addition and subtraction of various functions $\mu(D)$ for $D \in \mathcal{D}$. And if B_n are sets in Σ' increasing or decreasing to B , then $\mu(B) = \lim_{n \rightarrow \infty} \mu(B_n)$ is a limit of \mathcal{G} -measurable functions and hence is measurable. This shows Σ' is a monotone class, containing the algebra generated \mathcal{D} ; by the monotone class theorem, Σ' equals the σ -algebra generated by \mathcal{D} , that is, $\mathcal{B}_{\mathbb{R}^n}$.

The rest is easy. Consider

$$B \mapsto \int_{\omega \in E} \mu_\omega(B) dP|_{\mathcal{G}},$$

which defines a positive measure, and, by definition, agrees with the measure $B \mapsto P([Y \in B] \cap E)$ for $B \in \mathcal{D}$. Since \mathcal{D} generates $\mathcal{B}_{\mathbb{R}^n}$, the two measures are ultimately equal. As this is true for all $E \in \mathcal{G}$, we have $\mu(B) = P([Y \in B] | \mathcal{G})$ as desired. ■

Definition 7.1. Let $Y: \Omega \rightarrow \Omega'$ be measurable, for a measurable space (Ω', Σ') . Any function $\mu: \Omega \times \Sigma' \rightarrow [0, 1]$ such that

- (i) for each $\omega \in \Omega$, $\mu_\omega: \Sigma' \rightarrow [0, 1]$ is a probability measure, and
- (ii) for each $B \in \Sigma'$, $\mu(B) = P([Y \in B] | \mathcal{G})$ $P|_{\mathcal{G}}$ -almost surely

is called a *conditional probability measure for Y given \mathcal{G}* . In general, we denote these functions μ by $P_{Y|\mathcal{G}}$.

A *conditional probability measure for Y given a random variable X* is similarly defined, and denoted $P_{Y|X}$.

Theorem 7.2 says that $P_{Y|\mathcal{G}}$ (or $P_{Y|X}$) exists at least if $\Omega' = \mathbb{R}^n$.

We make a brief note, that it exists also if $\Omega' = \mathbb{R}^{\mathbb{N}}$. That is, given a countable number of random variables $Y_n: \Omega \rightarrow \mathbb{R}$, we can still construct $P_{Y|\mathcal{G}}$ for $Y = (Y_1, Y_2, \dots)$. This is done by the same kind of procedures used to construct sample spaces for an infinite

number of random variables — namely, by the Kolmogorov Existence Theorem. For it to work, we only need to verify the consistency conditions:

$$P_\alpha(E) = P_\beta(E \times \mathbb{R}^{|\beta|-|\alpha|}), \quad (11)$$

where α, β are ordered finite subsets of \mathbb{N} (with no repetition of members), and P_α stands for the finite-dimensional conditional probability measures for $(Y_{\alpha(1)}, Y_{\alpha(2)}, \dots, Y_{\alpha(|\alpha|)})$ constructed in Theorem 7.2. For each α, β , equation (11) is found to hold for every measurable $E \in \mathcal{B}_{\mathbb{R}^{|\alpha|}}$ except for a null set. But there are only countably many possible pairs of α, β , so we can obtain a single null set where equation (11) holds everywhere else. Then the Kolmogorov Existence Theorem allows us to construct

$$P(a_1 < Y_1 \leq b_1, a_2 < Y_2 \leq b_2, \dots | \mathcal{G})(\omega)$$

as a probability measure for each $\omega \in \Omega$.

(We cannot go much further than this, to construct conditional probability measures for an uncountable number of variables, because by taking the variables 1_E for every event E , we would be able to construct $P(E | \mathcal{G})(\omega)$.)

8 Calculating conditional expectations

In elementary courses of probability theory, one learns a definition of the conditional probability measure $\mu_x(B) = P(Y \in B | X = x)$, and defines $\mathbb{E}[Y | X = x]$ as the expectation of a random variable whose distribution is given by μ_x .

Since our definition of conditional expectation does not invoke conditional probability measures, the above “definition” — though quite convenient for practical computations — has to be proven. With well-defined conditional probability measures at our disposal, we can do this.

Theorem 8.1. *Let (Ω, Σ, P) be a probability space, $\mathcal{G} \subseteq \Sigma$ be a σ -algebra, and $Y: \Omega \rightarrow \mathbb{R}$ be a random variable. Then*

$$\mathbb{E}[Y | \mathcal{G}] = \int_{y \in \mathbb{R}} y dP_{Y|\mathcal{G}} \quad \text{i.e.,} \quad \mathbb{E}[Y | \mathcal{G}](\omega) = \int_{y \in \mathbb{R}} y dP_{Y|\mathcal{G}(\omega)} \quad \text{for } \omega \in \Omega.$$

Proof. The approximation theorem for measurable functions furnishes a sequence of random variables Y_1^+, Y_2^+, \dots , such that $Y_n^+ \geq 0$ and $Y_n^+ \nearrow Y^+ = \max(0, Y)$. In fact they have the explicit expression:

$$Y_n^+ = \sum_{k=1}^{n2^n-1} \frac{k}{2^n} 1_{[Y \in D_{n,k}]} + n 1_{[Y \in D_{n,\infty}]}, \quad D_{n,k} = \left[\frac{k}{2^n}, \frac{k+1}{2^n} \right), \quad D_{n,\infty} = [n, \infty).$$

Then we have

$$\begin{aligned} \mathbb{E}[Y_n^+ | \mathcal{G}] &= \sum_{k=1}^{n2^n-1} \frac{k}{2^n} P(Y \in D_{n,k} | \mathcal{G}) + n P(Y \in D_{n,\infty} | \mathcal{G}) \quad (\text{linearity of } \mathbb{E}[\cdot | \mathcal{G}]) \\ &= \sum_{k=1}^{n2^n-1} \frac{k}{2^n} P_{Y|\mathcal{G}}(D_{n,k}) + n P_{Y|\mathcal{G}}(D_{n,\infty}) \\ &= \int_{\mathbb{R}} \left(\sum_{k=1}^{n2^n-1} \frac{k}{2^n} 1_{D_{n,k}} + n 1_{D_{n,\infty}} \right) dP_{Y|\mathcal{G}}. \end{aligned}$$

The last integrand is the n th approximation for the positive part of the identity function $y \mapsto y$.

By monotone convergence (Proposition 6.3), $\mathbb{E}[Y^+ | \mathcal{G}]$ may be obtained as a limit of $\mathbb{E}[Y_n^+ | \mathcal{G}]$. Therefore,

$$\mathbb{E}[Y^+ | \mathcal{G}] = \lim_{n \rightarrow \infty} \mathbb{E}[Y_n^+ | \mathcal{G}] = \int_{y \in (0, \infty)} y dP_{Y|\mathcal{G}}.$$

We apply a similar argument for $Y^- = \max(0, -Y)$: by taking limits through a sequence of functions Y_1^-, Y_2^-, \dots such that $Y_n^- \geq 0$ and $Y_n^- \nearrow Y^-$, we have

$$-\mathbb{E}[Y^- | \mathcal{G}] = \lim_{n \rightarrow \infty} \mathbb{E}[-Y_n^- | \mathcal{G}] = \int_{y \in (-\infty, 0)} y dP_{Y|\mathcal{G}}.$$

Hence

$$\mathbb{E}[Y | \mathcal{G}] = \mathbb{E}[Y^+ | \mathcal{G}] - \mathbb{E}[Y^- | \mathcal{G}] = \int_{y \in (-\infty, \infty)} y dP_{Y|\mathcal{G}}. \quad \blacksquare$$

Example 8.1. If X, Y are real-valued random variables, with a joint probability density $f_{X,Y}$, then we calculate that

$$\mathbb{E}[Y | X = x] = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x, y)}{f_X(x)} dy, \quad f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy.$$

Not suprisingly, the conditional probability density — that is, the Radon-Nikodym derivative — appears to be the infinitesimal version of the elementary equation (6) for the conditional probability.

9 Change of variable

Suppose X is a random variable on a probability space (Ω, Σ, P) , and $f: \mathbb{R} \rightarrow \mathbb{R}$ is a measurable function. Let $Y = f(X)$. Then $\mathbb{E}Y$ can be computed in any of these three ways:

$$\int_{\omega \in \Omega} f(X(\omega)) dP, \quad \int_{x \in \mathbb{R}} f(x) dP_X, \quad \int_{y \in \mathbb{R}} y dP_Y. \quad (12)$$

So what about $\mathbb{E}[Y | \mathcal{G}]$?

Theorem 8.1 says that it can be calculated as

$$\int_{y \in \mathbb{R}} y dP_{Y|\mathcal{G}},$$

which is the counterpart to the last integral in (12). The first integral in (12) has no counterpart for conditional expectations, since we do not have available a conditional measure $B \mapsto P(B | \mathcal{G})(\omega)$ that is defined for all events B . But the second integral in (12) ought to have an analogy for conditional expectations, namely:

$$\int_{x \in \mathbb{R}} f(x) dP_{X|\mathcal{G}}.$$

We shall prove this.

Theorem 9.1. *Let (Ω, Σ, P) be a probability space, and $\mathcal{G} \subseteq \Sigma$ be a σ -algebra. Let $X: \Omega \rightarrow \Omega'$ be a measurable function to another measurable space Ω' , and $f: \Omega' \rightarrow \mathbb{R}$ be measurable also. Then*

$$\mathbb{E}[f(X) | \mathcal{G}] = \int_{x \in \Omega'} f(x) dP_{X|\mathcal{G}}.$$

Proof. Let $Y = f(X)$. We show $P_{Y|\mathcal{G}} = P_{X|\mathcal{G}} \circ f^{-1}$ almost surely. For each $B \in \mathcal{B}_{\mathbb{R}}$, almost surely we have

$$\begin{aligned} P_{Y|\mathcal{G}}(B) &= P(Y^{-1}(B) | \mathcal{G}) = P((f \circ X)^{-1}(B) | \mathcal{G}) \\ &= P(X^{-1}(f^{-1}(B)) | \mathcal{G}) = P_{X|\mathcal{G}}(f^{-1}(B)). \end{aligned}$$

There is a single $P_{|\mathcal{G}}$ -null set on which the above equations are true for every $B = (-\infty, y]$, with $y \in \mathbb{Q} \cup \{-\infty, +\infty\}$. Since the left- and right-side expressions define measures, the equations must then hold for every $B \in \mathcal{B}_{\mathbb{R}}$, on the same $P_{|\mathcal{G}}$ -null set. The result now follows from the change-of-variables theorem for ordinary integrals. \blacksquare

10 What is being conditioned on can be set constant

The aim of this section is to rigorously generalize two facts well known from the naïve definition of conditional probability:

1. For any events A, B (with $P(B) > 0$), $P(A \cap B | B) = P(A | B)$. i.e. if we are given B , and asked to calculate conditional probabilities, then of course B happens “for certain”. For the same reason, $P(A \cap B^c | B) = 0$.
2. This is related to the first fact. Suppose $f(X, Y)$ is a measurable function of two random variables X and Y , and we want to compute $\mathbb{E}[f(X, Y) | X]$. Since X is a given in the conditional probability, in the integral calculations X may be assumed to be constant. So, for instance (Proposition 4.5), $\mathbb{E}[XY | X] = X \mathbb{E}[Y | X]$.

In what follows, (Ω, Σ, P) is a probability space, $\mathcal{G} \subseteq \Sigma$ is a σ -algebra, and Ω_1, Ω_2 are two other measurable spaces. Also $X: \Omega \rightarrow \Omega_1$ will be \mathcal{G} -measurable, while $Y: \Omega \rightarrow \Omega_2$ will be Σ -measurable.

Theorem 10.1. *Let ν be a version of the conditional probability measure $P_{Y|\mathcal{G}}$. Then the product measure $\mu = \delta_X \otimes \nu$ gives a version of the conditional probability measure $P_{X,Y|\mathcal{G}}$. (Here δ_x denotes the point-mass measure at $x \in \Omega_1$.)*

Proof. Clearly, μ is a probability measure everywhere on Ω .

We prove that $\mu(S)$ is \mathcal{G} -measurable for every measurable $S \subseteq \Omega_1 \times \Omega_2$ by appealing to the monotone class theorem (again). Taking S of the form $A \times B$, where $A \subseteq \Omega_1$, $B \subseteq \Omega_2$ are measurable, the function $\mu(S) = \delta_X(A) \nu(B) = 1_{[X \in A]} \nu(B)$ is \mathcal{G} -measurable because X and $\nu(B)$ are. And \mathcal{G} -measurability is preserved under finite disjoint unions of sets $A \times B$, and under increasing and decreasing limits. So it follows that $\mu(S)$ is \mathcal{G} -measurable for every S in the product σ -algebra of $\Omega_1 \times \Omega_2$.

Also, for each $E \in \mathcal{G}$, consider the measure $S \mapsto \mathbb{E}[\mu(S) 1_E]$. It agrees with the measure $S \mapsto \mathbb{P}([(X, Y) \in S] \cap E)$ on sets S of the form $A \times B$:

$$\begin{aligned} \mathbb{E}[\mu(A \times B) 1_E] &= \mathbb{E}[1_{[X \in A]} \mathbb{P}([Y \in B] | \mathcal{G}) 1_E] \\ &= \mathbb{E}[\mathbb{P}([Y \in B] | \mathcal{G}) 1_{[X \in A] \cap E}] \\ &= \mathbb{P}([X \in A] \cap [Y \in B] \cap E) \quad (\text{note } [X \in A] \cap E \in \mathcal{G}). \end{aligned}$$

And hence the two measures agree on all measurable $S \subseteq \Omega_1 \times \Omega_2$. Since $E \in \mathcal{G}$ is arbitrary, it follows that $\mu(S)$ is a version of $\mathbb{P}([(X, Y) \in S] | \mathcal{G})$. ■

Theorem 10.2. *Let $f: \Omega_1 \times \Omega_2 \rightarrow \mathbb{R}$ be measurable in the product measure space. Then*

$$\mathbb{E}[f(X, Y) | \mathcal{G}] = \int_{y \in \Omega_2} f(X, y) d\mathbb{P}_{Y|\mathcal{G}}.$$

Proof. Given a version of $\mathbb{P}_{Y|\mathcal{G}}$, we calculate using the version of $\mathbb{P}_{X, Y|\mathcal{G}}$ that Theorem 10.1 constructs.

$$\begin{aligned} \mathbb{E}[f(X, Y) | \mathcal{G}] &= \int_{\Omega_1 \times \Omega_2} f d\mathbb{P}_{X, Y|\mathcal{G}} && (\text{Theorem 9.1}) \\ &= \int_{\Omega_1 \times \Omega_2} f d(\delta_X \otimes \mathbb{P}_{Y|\mathcal{G}}) && (\text{Theorem 10.1}) \\ &= \int_{y \in \Omega_2} f(X, y) d\mathbb{P}_{Y|\mathcal{G}}. && \blacksquare \end{aligned}$$

Corollary 10.3. *Under the same hypotheses as Theorem 10.2,*

$$\mathbb{E}[f(X, Y) | X = x] = \int_{y \in \Omega_2} f(x, y) d\mathbb{P}_{Y|X=x} = \mathbb{E}[f(x, Y) | X = x].$$

11 Hilbert space theory of conditional expectations

In section 2, we observed that, in general, the conditional expectation $\mathbb{E}[Y | \mathcal{G}]$ extracts “the part of Y that is \mathcal{G} -measurable”.

This is literally true if Y is a \mathbf{L}^2 random variable: then Y can be decomposed uniquely into random variables U and V such that

$$\begin{aligned} Y &= U + V \text{ (almost surely),} \\ U &\in \mathcal{M} = \{\text{all } \mathcal{G}\text{-measurable, } \mathbf{L}^2 \text{ random variables}\}, \quad V \in \mathcal{M}^\perp. \end{aligned}$$

The U part of Y can be obtained by projecting Y orthogonally onto \mathcal{M} , a closed subspace of the Hilbert space \mathbf{L}^2 , with inner product $\langle X, Y \rangle = \mathbb{E}[XY]$. The part left over, the random variable V , will be in \mathcal{M}^\perp , orthogonal to \mathcal{M} .

The projection operator is exactly $\mathbb{E}[\cdot | \mathcal{G}]$ because,

$$\mathbb{E}[Y 1_B] = \mathbb{E}[U 1_B] + \mathbb{E}[V 1_B], \quad \text{for every } B \in \mathcal{G},$$

and $\mathbb{E}[V 1_B] = 0$ by the definition of the orthogonal projection. Since the conditional expectation is unique, we must have $U = \mathbb{E}[Y | \mathcal{G}]$.

It is also not hard to give an intuitive description of \mathcal{M}^\perp : it consists of all \mathbf{L}^2 random variables, with zero mean, that are uncorrelated with *every* \mathcal{G} -measurable random variable. Indeed, since $1 \in \mathcal{M}$, we must have $\mathbb{E}(V \cdot 1) = \mathbb{E}V = 0$ for every $V \in \mathcal{M}^\perp$. Then $\mathbb{E}[UV] = \mathbb{E}[(U - \mathbb{E}U)(V - \mathbb{E}V)]$, so $V \in \mathcal{M}^\perp$ is orthogonal to U if and only if V is uncorrelated to U .

A slicker way of recognizing that $U = \mathbb{E}[Y \mid \mathcal{G}]$ is to recall that the image of the orthogonal projection onto \mathcal{M} can be characterized as the unique vector in \mathcal{M} closest in norm to the pre-image:

Proposition 11.1. *If Y is a \mathbf{L}^2 real-valued random variable, then the best estimate of Y , in the least-squares sense, using only \mathcal{G} -measurable functions, is $\mathbb{E}[Y \mid \mathcal{G}]$. That is,*

$$\mathbb{E}[(Y - X)^2] \geq \mathbb{E}[(Y - \mathbb{E}[Y \mid \mathcal{G}])^2]$$

for all \mathcal{G} -measurable random variables X , with equality if and only if $X = \mathbb{E}[Y \mid \mathcal{G}]$.

(The lower bound can also be written as: $\mathbb{E}[\text{Var}(Y \mid \mathcal{G})] = \text{Var}(Y) - \text{Var}(\mathbb{E}[Y \mid \mathcal{G}])$.)

Proof. The proof is completely analogous to the proof of the well-known case when X are restricted to constants. We have:

$$\begin{aligned} \mathbb{E}[(Y - X)^2] &= \mathbb{E}(X^2) - 2\mathbb{E}(XY) + \mathbb{E}(Y^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}[X\mathbb{E}(Y \mid \mathcal{G})] + \mathbb{E}[\mathbb{E}(Y^2 \mid \mathcal{G})] \\ &= \mathbb{E}[X^2 - 2\mathbb{E}(Y \mid \mathcal{G})X + \mathbb{E}(Y^2 \mid \mathcal{G})]. \end{aligned}$$

The outermost integrand is a quadratic in X , which is minimized when X equals the \mathcal{G} -measurable function $\mathbb{E}[Y \mid \mathcal{G}]$. ■

In fact, this Hilbert-space argument can be turned around, to prove the existence of $\mathbb{E}[\cdot \mid \mathcal{G}]$ without recourse to the Lebesgue-Radon-Nikodym theorem!⁴

Example 11.1 (Orthonormal basis expansion of conditional expectation). Let X and Y be two \mathbf{L}^2 random variables, with some joint distribution that is known, and we want to compute the conditional expectation $\mathbb{E}[Y \mid X] = \mathbb{E}[Y \mid \sigma(X)]$.

Recall, from linear algebra, that an orthogonal projection can be evaluated from its known actions on an orthonormal basis $\{Z_n\}$ that spans $\sigma(X)$.

$$\mathbb{E}[Y \mid \sigma(X)] = \sum_n \langle Y, Z_n \rangle Z_n.$$

Let F be the cumulative distribution function of X . Then $U = F(X)$ takes on values in $[0, 1]$ and has the uniform distribution. Since every element of $\sigma(X) = \sigma(U)$ can be represented as a function of U , we can set $Z_n = \phi_n(U)$ for some $\phi_n \in \mathbf{L}^2[0, 1]$ with Lebesgue measure. This latter Hilbert space $\mathbf{L}^2[0, 1]$ is separable, so the orthonormal basis is countable.

Thus, for any orthonormal basis $\{\phi_n\}$ of $\mathbf{L}^2[0, 1]$, we can expand:

$$\mathbb{E}[Y \mid X] = \sum_n \mathbb{E}[Y \overline{\phi_n(U)}] \phi_n(U), \quad U = F(X),$$

the series of random variables being convergent in \mathbf{L}^2 .

⁴ Incidentally, the Lebesgue-Radon-Nikodym theorem has a nice proof using Hilbert-space methods also.

Example 11.2 (Fourier expansion of conditional expectation). One popular orthonormal basis is the complex Fourier basis $\phi_n(u) = e^{2\pi i n u}$. So:

$$\mathbb{E}[Y | X] = \sum_{n=-\infty}^{\infty} \mathbb{E}[Y e^{-2\pi i n U}] e^{2\pi i n U}, \quad U = F(X).$$

Example 11.3 (Conditional expectation for discrete random variables). The only time that the orthonormal basis in Example 11.1 can be taken to be a finite set is when X has finite range $\{x_1, \dots, x_n\}$. In this case, the obvious orthonormal basis to use is $Z_n = 1(X = x_n)/\sqrt{P(X = x_n)}$. Then we arrive at the familiar expression:

$$\mathbb{E}[Y | X] = \sum_{n=1}^N \frac{\mathbb{E}[Y 1(X = x_n)]}{P(X = x_n)} 1(X = x_n).$$

12 Bibliography

References

- [Bouveau] Nicolas Bouveau, Dominique Lépingle. *Numerical Methods for Stochastic Processes*. Wiley-Interscience, 1994.
- [Folland] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, second ed. Wiley-Interscience, 1999.
- [Rosenthal] Jeffrey S. Rosenthal, *A First Look at Rigorous Probability Theory*. World Scientific, 2000.
- [Schmetterer] Leopold Schmetterer, *Introduction to Mathematical Statistics*. Trans. Kenneth Wickwise. Springer-Verlag, 1974.
- [Steele] J. Michael Steele, *Stochastic Calculus and Financial Applications*. Springer-Verlag, 2001.