Chapter I
Stochastic Dependence

In almost all interesting stochastic processes that arise in practice, what will happen in period t *depends* in some probabilistic manner to what has happened in periods before t. So, to study such processes, one needs a method of understanding how certain random events affect the occurrence of other random events. The systematic development of this method is the subject matter of the first half of this chapter. We begin with discussing at length the idea behind conditioning on a σ-algebra and that behind defining the notion of conditional expectation of a random variable itself as a random variable. These concepts may appear rather strange to the novice, so our exposition is particularly leisurely at the outset. Once the formal definitions are introduced and some examples are worked out, however, it becomes more streamlined. At this point we derive several useful properties of the conditional expectation operator, and establish the existence of this operator. Our existence proof is based on Hilbert space methods (which were reviewed in Appendix B).[1]

1 Conditional Expectation

1.1 Conditioning on a Countable Decomposition

Let x be an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$. (That is, $x \in \mathcal{L}^0(X, \Sigma)$ and $\mathbb{E}(|x|) < \infty$, or put differently, $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$.) For any event $C \in \Sigma$ with $\mathbf{p}(C) > 0$, the **conditional expectation** *of x given C* is defined as

$$\mathbb{E}(x \,|\, C) := \frac{1}{\mathbf{p}(C)} \int_C x d\mathbf{p}. \tag{1}$$

The idea is simple. When it is known that the event C has occurred, one is less uncertain about the outcome of the underlying experiment. The relevant probability space in this case can be thought of as $(C, \Sigma \cap C, \mathbf{q})$, where $\Sigma \cap C := \{A \cap C : A \in \Sigma\}$

---

[1] A formal development of the conditional expectation operator can be found in any of the (measure-theoretic) probability texts I have mentioned in the earlier chapters. However, it is my experience that economics students often find these expositions not so easy to follow. This is why the pace of analysis is rather slow earlier in this chapter.

and

$$\mathbf{q} := \frac{1}{\mathbf{p}(C)}\mathbf{p}|_{\Sigma\cap C}.$$

Consequently, we compute the conditional expectation of x given C simply as the expectation of $x|_C$ on this space.

Precisely the same intuition tells us that we should define the **conditional probability** *of an event* $A \in \Sigma$ *given* C as $\mathbf{p}(A \,|\, C) := \mathbf{p}(A \cap C)/\mathbf{p}(C)$, a definition which you have undoubtedly come across before. This concept is closely linked to that of the conditional expectation. Just as the probability of an event is the expectation of the indicator function of that event, the conditional probability of an event is the conditional expectation of the indicator function of that event:

$$\mathbf{p}(A \,|\, C) = \frac{\mathbf{p}(A \cap C)}{\mathbf{p}(C)} = \mathbb{E}(\mathbf{1}_A \,|\, C).$$

These notions are only partly satisfactory, however. They apply only when we know that an event has already occurred. In most cases, we need to talk about the conditional expectation of a random variable before the ambient experiment is performed, but given that we *will* get some information after the experiment is run. A gambler, for instance, would like to have a strategy which is contingent on what *will* happen through the sequence of games she will repeatedly participate in.

To be more precise, let us first agree on some terminology. In what follows, by a **countable (finite) Σ-decomposition** of X we mean a countable (finite) partition $\mathcal{C}$ of X such that $\mathcal{C} \subseteq \Sigma$ and $\mathbf{p}(C) > 0$ for all $C \in \mathcal{C}$. Now take any countable Σ-decomposition of X, and suppose that once the experiment underlying the probability space $(X, \Sigma, \mathbf{p})$ is performed, we will be told which of the events in $\mathcal{C}$ has actually occurred. For instance, if $\mathcal{C} = \{X\}$, then we will get no information whatsoever under this scenario. At the opposite extreme, if $\Sigma = \sigma\{\{\omega_i\} : i = 1, ..., k\}$ for some positive integer k, and $\mathcal{C} = \{\{\omega_1\}, ..., \{\omega_k\}\}$, then we will be given all the information that can be deduced from the experiment.[2] This prompts us defining the **conditional expectation** *of* x *given* $\mathcal{C}$ (that is, given that we will observe which member of $\mathcal{C}$ has occurred *once the experiment is performed*) as the discrete random variable $\mathbb{E}(x \,|\, \mathcal{C}) : X \to \mathbb{R}$ with

$$\mathbb{E}(x \,|\, \mathcal{C})(\omega) := \mathbb{E}(x \,|\, C_\omega), \tag{2}$$

---

[2] Recall that what one may observe in the experiment is given through the σ-algebra Σ. That is, an "event" that can be deduced to occur in the ambient experiment is a member of Σ which may or may not contain all subsets of X.

where C_ω is the member of $\mathcal{C}$ that contains the outcome ω. (This well-defines $\mathbb{E}(x \,|\, \mathcal{C})$ precisely because $\mathcal{C}$ is a partition of X and $\mathbf{p}(C) > 0$ for each $C \in \mathcal{C}$.) More compactly, we may write

$$\mathbb{E}(x \,|\, \mathcal{C}) := \sum_{C \in \mathcal{C}} \mathbb{E}(x \,|\, C)\mathbf{1}_C.$$

The **conditional probability** of an event A in Σ given $\mathcal{C}$ is similarly defined as the discrete random variable $\mathbf{p}(A \,|\, \mathcal{C}) : X \to \mathbb{R}$ with

$$\mathbf{p}(A \,|\, \mathcal{C})(\omega) := \frac{\mathbf{p}(A \cap C_\omega)}{\mathbf{p}(C_\omega)},$$

or equivalently,

$$\mathbf{p}(A \,|\, \mathcal{C}) := \sum_{C \in \mathcal{C}} \mathbf{p}(A \,|\, C)\mathbf{1}_C.$$

Of course, we have $\mathbf{p}(A \,|\, \mathcal{C}) = \mathbb{E}(\mathbf{1}_A \,|\, \mathcal{C})$.

The conceptual point here is that we are modeling conditional expectation not as a number, but as a random variable. On a closer inspection this makes good sense. For instance, $\mathbf{p}(A \,|\, \mathcal{C})$ is a random variable, because we do not know *a priori* (that is, before the experiment is performed) which of the events in $\mathcal{C}$ will occur. (By contrast, $\mathbf{p}(A \,|\, C)$ is a number that tells us the probability of A given the information that C has actually occurred.) If the outcome of the experiment is ω, then we are pointed to "the" event in $\mathcal{C}$ that contains ω (that is, C_ω) – we learn that "something in C_ω" has occurred. Since the event that we will be informed to have occurred depends on ω – remember, the experiment is not run yet – we should indeed think of $\mathbf{p}(A \,|\, \mathcal{C})$ as a random variable, and set $\mathbf{p}(A \,|\, \mathcal{C})(\omega)$ equal to $\mathbf{p}(A \,|\, C)$ whenever the outcome of the experiment ω belongs to C. (The "randomness" of $\mathbf{p}(A \,|\, C)$ thus arises from the fact that we will never observe an outcome ω once the experiment is run (unless, of course, $\{\omega\} \in \mathcal{C}$), but we will observe the event in $\mathcal{C}$ that contains ω.)

Example 1.1. Let $(X, \Sigma, \mathbf{p})$ be a probability space, and take any $A, C \in \Sigma$ with $\mathbf{p}(C) > 0$. If $\mathcal{C} := \{C, X \backslash C\}$, then

$$\mathbf{p}(A \,|\, \mathcal{C})(\omega) := \begin{cases} \mathbf{p}(A \,|\, C), & \text{if } \omega \in C \\ \mathbf{p}(A \,|\, X \backslash C), & \text{if } \omega \notin C \end{cases}.$$

In particular, $\mathbf{p}(C \,|\, \mathcal{C}) = \mathbf{1}_C = \mathbb{E}(\mathbf{1}_C \,|\, \mathcal{C})$. (Interpretation?) □

Example 1.2. An urn contains four balls labeled $1, 2, 3, 4$, respectively. Consider the experiment of simultaneously drawing two balls from this urn at random. The probability space of the experiment is $(X, 2^X, \mathbf{p})$, where $X := \{\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\},$

$\{2,4\},\{3,4\}\}$ and $\mathbf{p}(\omega) := 1/6$ for each $\omega \in X$. Suppose, once the two balls are drawn, we will be informed if "one of these balls is 1," and if "none is 1 but one is 2," and if "none is 1 or 2." So letting $C_1 := \{\{1,2\},\{1,3\},\{1,4\}\}$, $C_2 := \{\{2,3\},\{2,4\}\}$ and $C_3 := \{\{3,4\}\}$, we understand that the information that we will be given (once the experiment is run) will let us find out which of the events in the finite 2^X-decomposition $\mathcal{C} := \{C_1, C_2, C_3\}$ of X has occurred. Therefore, if A is the event that the product of the numbers on the balls drawn is a prime number, we have

$$\mathbf{p}(A \,|\, \mathcal{C})(\omega) = \begin{cases} 2/3, & \text{if } \omega \in C_1 \\ 0, & \text{if } \omega \in C_2 \cup C_3. \end{cases}$$

If the real map x on X is defined as $x(\{i,j\}) := ij$, and if $\omega \in C_1$, then

$$\mathbb{E}(x \,|\, \mathcal{C})(\omega) = \frac{1}{\mathbf{p}(C_1)} \int_{C_1} x d\mathbf{p} = 2(\tfrac{2}{6} + \tfrac{3}{6} + \tfrac{4}{6}) = 3.$$

Similar computations yield $\mathbb{E}(x \,|\, \mathcal{C})(\omega) = 7$ for $\omega \in C_2$, and $\mathbb{E}(x \,|\, \mathcal{C})(\omega) = 12$ for $\omega \in C_3$. □

Example 1.3. Let x be an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$, and y a discrete random variable on $(X, \Sigma, \mathbf{p})$ such that $\mathbf{p}_y\{a\} > 0$ for every $a \in y(X)$. Suppose that once the experiment is performed, we will learn the value of y. This means that if the outcome $\omega \in X$ occurs in the experiment and $y(\omega) = a$, then we will know that the event $\{y = a\}$ has occurred (but we won't know that the outcome is ω), and compute the expectation of x accordingly. Since $\mathcal{C} := \{\{y = a\} : a \in y(X)\}$ is a countable Σ-decomposition of X, it is then in the nature of things to define the **conditional expectation** *of* x *given* y as the discrete random variable $\mathbb{E}(x \,|\, y)$ on (X, Σ) with $\mathbb{E}(x \,|\, y) := \mathbb{E}(x \,|\, \mathcal{C})$. That is, for any given $\omega \in X$, we have $\mathbb{E}(x \,|\, y)(\omega) := \mathbb{E}(x \,|\, \{y = a_\omega\})$, or put more explicitly,

$$\mathbb{E}(x \,|\, y)(\omega) = \frac{1}{\mathbf{p}_y\{a_\omega\}} \int_{\{y = a_\omega\}} x d\mathbf{p}, \tag{3}$$

where $a_\omega := y(\omega)$. □

Example 1.4. Let y be as in the previous example. What is $\mathbb{E}(y \,|\, y)$? Think about it intuitively first. If $\omega \in X$ comes up in the experiment, we will be told that the value of y is $a := y(\omega)$. So what would we "expect" the value of y to be in light of this information? The answer is, of course, a. If we are told what value of y has come up, the uncertainty about y resolves completely – that value is our expectation of y. That is, $\mathbb{E}(y\,|\,y) = y$. (Does (3) agree with this intuition?)

More generally, we have $\mathbb{E}(f(y)\,|\,y) = f(y)$ for any real map f on $y(X)$. And conversely, if x is a discrete random variable on (X, Σ) and $\mathbb{E}(x\,|\,y) = x$, then we understand that the occurrence of y gives us full information about the occurrence of x.[3] $\square$

Example 1.5. Consider the probability space $([0,1), \mathcal{B}[0,1), \ell)$, and define the random variables x and y_m on this space by $x(\omega) := 2\omega$ and

$$y_m(\omega) := i, \qquad \tfrac{i}{m} \leq \omega < \tfrac{i+1}{m},\ i = 0, ..., m-1,$$

where m is a positive integer. Let us compute $\mathbb{E}(x\,|\,y_m)$ for an arbitrarily fixed m. Since y_m may assume the values $0, ..., m-1$ with positive probability, the relevant events here are $\{y_m = 0\}, ..., \{y_m = m-1\}$, and clearly, we have $\ell\{y_m = i\} = 1/m$ for each $i = 0, ..., m-1$. Thus, $\mathbb{E}(x\,|\,y_m)$ is the simple random variable on $([0,1), \mathcal{B}[0,1), \ell)$ given by

$$\begin{aligned}\mathbb{E}(x\,|\,y_m)(\omega) &= \frac{1}{1/m}\int_{[i/m,(i+1)/m)} x d\ell \\ &= m\int_{i/m}^{(i+1)/m} 2\omega d\omega \\ &= \frac{2i+1}{m},\end{aligned}$$

for any ω in the interval $[i/m, (i+1)/m)$ and any nonnegative integer i less than $m-1$. Notice that y_1 is a constant random variable, so learning its value cannot possibly give us any information about the outcome of the experiment. In accordance with this intuition, we find here that $\mathbb{E}(x\,|\,y_1) = 1 = \mathbb{E}(x)$. On the other hand, for large m, learning the value of y_m gives us a lot of information in the sense that with this information we understand that the actual outcome of the experiment lies in a particular interval of length $1/m$. Clearly, the higher m, the more refined the conditional expectation $\mathbb{E}(x\,|\,y_m)$ becomes. $\square$

---

[3] *Quiz.* Show that, for any discrete random variable x on $(X, \Sigma, \mathbf{p})$, $\mathbb{E}(x\,|\,y) = x$ implies that $x = f(y)$ for some real map f on $y(X)$.

Exercise 1.1. Consider the probability space $([0,1], \mathcal{B}[0,1], \ell)$ and consider the following functions defined on $[0,1]$:

$$x(\omega) := \begin{cases} 1, & \text{if } 0 \leq \omega < 1/4 \\ 2, & \text{if } 1/4 \leq \omega < 1/2 \\ 3, & \text{if } 1/2 \leq \omega < 3/4 \\ 4, & \text{if } 3/4 \leq \omega \leq 1, \end{cases}$$

$y := \mathbf{1}_{[0,1/2)} + 2\mathbf{1}_{[1/2,1]}$, and $z := 3\mathbf{1}_{[0,1/2)} + 4\mathbf{1}_{[1/2,1]}$.
(a) What is $\mathbb{E}(y \,|\, x)$? (It should take you at most 30 seconds to answer this.)
(b) Compute $\mathbb{E}(x \,|\, y)$ and $\mathbb{E}(x \,|\, z)$.

Exercise 1.2.$^{\mathrm{H}}$ Let F be the distribution function such that $F(t) = t^2$ for every real number t in $[0,1]$. Consider the probability space $([0,1], \mathcal{B}[0,1], \mathbf{p}_F)$, where $\mathbf{p}_F$ is the Lebesgue-Stieltjes measure induced by F, and define

$$x(\omega) := \begin{cases} 1, & \text{if } 0 \leq \omega < 1/4 \\ 0, & \text{if } 1/4 \leq \omega < 1/2 \\ -1, & \text{if } 1/2 \leq \omega < 3/4 \\ 0, & \text{if } 3/4 \leq \omega \leq 1. \end{cases}$$

Compute $\mathbb{E}(x \,|\, \mathbf{1}_{[0,1/2)})$ and $\mathbb{E}(x^2 \,|\, \mathbf{1}_{[0,1/2)})$.

Exercise 1.3.$^{\mathrm{H}}$ In the setup of Example 1.3, show that $\mathbb{E}(x \,|\, y)$ is $\sigma(y)$-measurable and

$$\int_B \mathbb{E}(x \,|\, y) d\mathbf{p} = \int_B x d\mathbf{p} \qquad \text{for every } B \in \sigma(y).$$

1.2 Conditioning on an Arbitrary σ-algebra

So far so good, but the development of the conditional expectation above allows us to model the information that we will be given (once the experiment is performed) only through a countable Σ-decomposition of the sample space. This is too restrictive. For instance, consider the probability space $([0,1], \mathcal{B}[0,1], \ell)$. Since $\mathcal{B}[0,1]$ is not countable – it is not even generated by a countable partition of $[0,1]$ – presently we cannot talk about the probability of the occurrence of $A \in \mathcal{B}[0,1]$, given that we will be told in the aftermath of the experiment which members of $\mathcal{B}[0,1]$ have occurred. This is pretty silly, for it is obvious what probability to assign to A conditional on knowing $\mathcal{B}[0,1]$. After all, $\mathcal{B}[0,1]$ contains all the information that can be deduced through the experiment, that is, we will learn all there is to know once the experiment is run. Formally, if $\omega \in A$ (we will learn this), we should naturally say $\mathbf{p}(A \,|\, \mathcal{B}[0,1])(\omega) = 1$, and let $\mathbf{p}(A \,|\, \mathcal{B}[0,1])(\omega) = 0$ otherwise.

Let us elaborate on this point a little further, this time using the random variable terminology. You probably know already that the most common way of receiving information about the outcome of an experiment with the probability space $(X, \Sigma, \mathbf{p})$ is through learning the value of a random variable $y \in \mathcal{L}^0(X, \Sigma)$. This means that, once the experiment is performed, we will learn the value of y, that is, prior to the experiment, we know that we will be told which one of the events in $\{\{y = a\} : a \in y(X)\}$ has occurred. But what if $\mathbf{p}\{y = a\} > 0$ does not hold for some $a \in y(X)$, as it does not in many interesting cases? For instance, in Example 1.5, what would we understand from $\mathbb{E}(x \mid y)$ if y was the identity function on $[0, 1]$? Surely, learning y would give us a lot of information – this would tell us the exact outcome of the experiment – but we still cannot talk about $\mathbb{E}(x \mid y)$ by means of a definition like (3) since $\{y = a\}$, that is, $\{a\}$, is a zero probability event for any $a \in [0, 1]$. Yet, it is intuitively clear that $\mathbb{E}(x \mid y)(\omega)$ must equal $x(\omega)$ for every $\omega \in [0, 1]$ – if we are told that "ω comes up in the experiment," we would surely "expect" that the value of x is $x(\omega)$!

To give another example, consider the following experiment that is to take place in two stages. In the first stage, a real number α between 0 and 1 will be randomly selected, and in the second stage a lottery that pays 1 with probability α and -1 with probability $1 - \alpha$ will be played. Suppose y corresponds to the outcome of the first stage of the experiment – it is uniformly distributed on $[0, 1]$ – and x that of the second. What would be your expectation of x conditional on learning that $y = \frac{1}{2}$? The answer seems straightforward. Given $y = \frac{1}{2}$, the second stage lottery is one that pays 1 with probability $1/2$ and -1 with probability $\frac{1}{2}$ – obviously, given $y = 1/2$, we expect the value of x to be 0. But, formally speaking, we have no way of concluding this from our definitions so far, since y is not discrete. To make matters worse, we cannot even write a formula for $\mathbb{E}(x \mid \{y = \frac{1}{2}\})$ in a natural way, for $\mathbf{p}\{y = \frac{1}{2}\} = 0$.

All this is to say that we need to improve upon our current definition of the conditional expectation. Okay, then let's recall that, in probability theory, we think of "information" about an experiment as a particular σ-algebra that tells us which events can be understood to have occurred once the experiment is performed. Therefore, what we should be after is a notion of conditional expectation that will correspond to the idea that "getting some information about the experiment" is to learn which events in a σ-algebra Σ_0 (contained within the original σ-algebra of the experiment) have occurred after the experiment is run. That is, we need to find a way to condition on an arbitrary σ-algebra.[4] Unfortunately, this is not at all a trivial matter. After

---

[4] The notion of "learning the values of y" is equivalent to "learning $\sigma(y)$" for a discrete random variable y, and the notion "learning $\sigma(y)$" is actually meaningful for *any* random variable y. So

all, since numerous events in a σ-algebra may have zero probability of occurring, we cannot use a formula like (2) for this purpose.

So, what do we do? Well, let us ask ourselves what we would require of this object, call it, $\mathbb{E}(x \mid \Sigma_0)$. We certainly want it to be a random variable on (X, Σ), that much follows from the discussion of the previous subsection. But in fact this is not good enough; we actually need the statements like "$\mathbb{E}(x \mid \Sigma_0) = 1$" or "$1 < \mathbb{E}(x \mid \Sigma_0) \leq 2$" correspond to events in Σ_0. After all, we would like to be able to compute the values of $\mathbb{E}(x \mid \Sigma_0)$ using only the information in Σ_0, or put differently, we would like to "know" $\mathbb{E}(x \mid \Sigma_0)(\omega)$ once we are told which events in Σ_0 contains ω. This means that

$$\mathbb{E}(x \mid \Sigma_0) \text{ is } \Sigma_0\text{-measurable.} \tag{4}$$

Intuitively, this means that $\mathbb{E}(x \mid \Sigma_0)$ does not contain more information than that contained in Σ_0. For instance, if $\{\mathbb{E}(x \mid \Sigma_0) = 1\} \notin \Sigma_0$ was the case, then we could not decide whether or not $\mathbb{E}(x \mid \Sigma_0) = 1$ even after we observe Σ_0 (that is, after we learn which members of Σ_0 have occurred) – this would mean that somehow $\mathbb{E}(x \mid \Sigma_0)$ uses more information than what is contained in Σ_0. It is precisely this sort of an oddity that we avoid by positing that $\mathbb{E}(x \mid \Sigma_0)$ is Σ_0-measurable.[5]

Okay, (4) is in the pocket, what else? Well, we wish to define $\mathbb{E}(x \mid \Sigma_0)$ via something like (2), but we cannot in general do this directly because of the division-by-zero problem (unless, of course, Σ_0 is generated by a countable Σ-decomposition of X). Indirectly, however, we may bypass this problem by observing first that (2) can be written equivalently as

$$\mathbf{p}(C_\omega)\mathbb{E}(x \mid \mathcal{C})(\omega) = \int_{C_\omega} x d\mathbf{p} \quad \text{for every } \omega \in X,$$

and hence

$$\mathbf{p}(C)\mathbb{E}(x \mid \mathcal{C})(\omega) = \int_{C} x d\mathbf{p} \quad \text{for every } C \in \mathcal{C}.$$

But since $\mathcal{C}$ is a countable Σ-decomposition of X here, $\mathbb{E}(x \mid \mathcal{C})$ is a discrete random variable which is constant on each $C \in \mathcal{C}$. Therefore, $\mathbf{p}(C_\omega)\mathbb{E}(x \mid \mathcal{C})(\omega)$ equals

---

if we find out how to condition on an arbitrary σ-algebra, it would be quite natural to think of "conditioning on y" as "conditioning on $\sigma(y)$."

[5] To see this point a bit more clearly, consider any $y \in \mathcal{L}^0(X, \Sigma)$, and suppose that $\mathbb{E}(x \mid y)$ is not $\sigma(y)$-measurable. Then it may well be the case that $\mathbb{E}(x \mid y)$ (which is shortly to be defined as $\mathbb{E}(x \mid \sigma(y))$) is not constant on $y^{-1}(a)$ for some $a \in y(X)$. This would clearly mean that $\mathbb{E}(x \mid y)$ is somehow using more information than contained in $\sigma(y)$, which is absurd. Thanks to (4), no such absurdity is allowed; we are guaranteed to have a unique value for the expectation of x conditional on the event $\{y = a\}$. (See Remark 1.1.)

$\int_{C_\omega} \mathbb{E}(x \,|\, \mathcal{C}) d\mathbf{p}$ for every $\omega \in X$. It follows that another way of writing (2) is:

$$\int_{C_\omega} \mathbb{E}(x \,|\, \mathcal{C}) d\mathbf{p} = \int_{C_\omega} x d\mathbf{p} \quad \text{for every } \omega \in X.$$

The upshot is that this property of $\mathbb{E}(x \,|\, \mathcal{C})$ generalizes to the case where we condition the expectation of x on Σ_0. So, it seems duly reasonable to demand $\mathbb{E}(x \,|\, \Sigma_0)$ satisfy the following:

$$\int_B \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} = \int_B x d\mathbf{p} \quad \text{for every } B \in \Sigma_0. \tag{5}$$

Good, what else? Well, as a matter of fact, we're done at this point, because (4) and (5) can be satisfied by essentially only one real function on X. Indeed, if z_i is a Σ_0-measurable real function on X and $\int_B z_i d\mathbf{p} = \int_B x d\mathbf{p}$ for every $B \in \Sigma_0$, $i = 1, 2$, then we must have $z_1 =_{\text{a.s.}} z_2$. For, in this case we have

$$y \in \mathcal{L}^0(X, \Sigma_0) \quad \text{and} \quad \int_B y d\mathbf{p} = 0 \text{ for every } B \in \Sigma_0,$$

where $y := z_1 - z_2$. Choosing $B_m := \{y \geq \frac{1}{m}\} \in \Sigma_0$, we thus get

$$0 = \int_{B_m} y d\mathbf{p} \geq \tfrac{1}{m}\mathbf{p}(B_m)$$

so that $\mathbf{p}(B_m) = 0$ for every positive integer m. Applying this argument to $-y$ as well, we find that $\mathbf{p}\{|y| \geq \frac{1}{m}\} = 0$ for every $m \in \mathbb{N}$, so by Proposition B.2.2, $\mathbf{p}\{y = 0\} = 1$, that is, $z_1 =_{\text{a.s.}} z_2$.

This observation gives us, therefore, an indirect way of defining the conditional expectation of a random variable given an arbitrary σ-algebra (or an arbitrary random variable) which reduces to the definitions given earlier in the special case where this σ-algebra is generated by a countable decomposition of the sample space (or by a discrete random variable). Here comes the formal definition.[6]

Definition. Let $(X, \Sigma, \mathbf{p})$ be a probability space. By a **sub-σ-algebra** of Σ, we mean a subset of Σ which is itself a σ-algebra. For any sub-σ-algebra Σ_0 of Σ, and any $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$, the **conditional expectation of x given** Σ_0 is defined as an extended real function $\mathbb{E}(x \,|\, \Sigma_0)$ on X which satisfies (4) and (5), that is,

$$\mathbb{E}(x \,|\, \Sigma_0) \text{ is } \Sigma_0\text{-measurable}$$

---

[6] I will give the definition of conditional expectation only for integrable random variables here, as I will work with only such random variables in what follows. At any rate, it is obvious how to extend this definition to the case of all random variables whose expectations exist.

and

$$\int_B \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} = \int_B x d\mathbf{p} \text{ for every } B \in \Sigma_0.$$

For any $y \in \mathcal{L}^0(X, \Sigma)$, we define the **conditional expectation of** x **given** y, denoted $\mathbb{E}(x \,|\, y)$, as $\mathbb{E}(x \,|\, y) := \mathbb{E}(x \,|\, \sigma(y))$. Finally, for any $A \in \Sigma$, we define the **conditional probability of** A **given** Σ_0, denoted $\mathbf{p}(A \,|\, \Sigma_0)$, as an extended real function on X that equals $\mathbb{E}(\mathbf{1}_A \,|\, \Sigma_0)$, or equivalently, that satisfies

$$\mathbf{p}(A \,|\, \Sigma_0) \text{ is } \Sigma_0\text{-measurable}$$

and

$$\int_B \mathbf{p}(A \,|\, \Sigma_0) d\mathbf{p} = \mathbf{p}(A \cap B) \text{ for every } B \in \Sigma_0.$$

There are two issues that we should discuss before moving on to examples that will illustrate the content of these definitions. Let us first think about the uniqueness matters. What does it mean to define "the" conditional expectation of x given Σ_0 as "a" function that satisfies (4) and (5)? The answer is that the definition given above identifies $\mathbb{E}(x \,|\, \Sigma_0)$ only almost uniquely – a nuisance we will have to live with. Formally speaking, one should really think of $\mathbb{E}(x \,|\, \Sigma_0)$ as an *equivalence class* of functions in $\mathcal{L}^0(X, \Sigma_0)$, where the equivalence relation is "$=_{\text{a.s.}}$". (This situation is reminiscent of how one may view $\mathcal{L}^p$-spaces as normed linear spaces. Recall Section C.6.) Any member of this equivalence class is called a **version** of $\mathbb{E}(x \,|\, \Sigma_0)$. In practice, one often speaks of $\mathbb{E}(x \,|\, \Sigma_0)$ in terms of its versions (which are of course identical to each other with probability one).

The second issue that we need to worry about is existence. How do we know that $\mathbb{E}(x \,|\, \Sigma_0)$ exists for any $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$ and any sub-σ-algebra Σ_0 of Σ? This is a highly nontrivial matter which is commonly settled by an appeal to a fundamental theorem of real analysis, the celebrated Radon-Nikodym Theorem. For the time being, however, we shall accept the following theorem on faith. This situation will be remedied later in the chapter.

Therorem 1.1. *If* x *is an integrable random variable on the probability space* $(X, \Sigma, \mathbf{p})$, *and* Σ_0 *is a sub-*σ*-algebra of* Σ, *then* $\mathbb{E}(x \,|\, \Sigma_0)$ *exists.*

Note that we do not need to invoke this theorem in the following examples, for we can settle the issue of existence in these particular cases by means of direct computation.

Example 1.6. Let $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$, and let Σ_0 be a sub-σ-algebra of Σ.

[1] If $\Sigma_0 := \{\emptyset, X\}$, then $\mathbb{E}(x \,|\, \Sigma_0) =_{\text{a.s.}} \mathbb{E}(x)$, that is, $\mathbb{E}(x)$ is a version of $\mathbb{E}(x \,|\, \Sigma_0)$. This case corresponds to the scenario in which we will be given no information after the experiment is performed (or more precisely, we will learn only that the outcome of the experiment belongs to X, which we already knew).

[2] If x is Σ_0-measurable, then we have $\mathbb{E}(x \,|\, \Sigma_0) =_{\text{a.s.}} x$, that is, x is a version of $\mathbb{E}(x \,|\, \Sigma_0)$. This fact, which generalizes that noted in Example 1.4, can be interpreted as follows: Since x is Σ_0-measurable, its informational content $\sigma(x)$ is less than Σ_0 (i.e. $\sigma(x) \subseteq \Sigma_0$), so learning Σ_0 would reveal $\sigma(x)$, and hence x. Put differently, since we will be told which members of Σ_0 have occurred in the experiment, and $\sigma(x) \subseteq \Sigma_0$, we will know, for each $\omega \in X$, whether or not $\{\nu \in X : x(\nu) = x(\omega)\}$ has occurred. Thus in this case, we will be able to pinpoint the value of x almost surely.[7] In particular,

$$\mathbb{E}(x \,|\, x) \underset{\text{a.s.}}{=} x.$$

[3] Let $\mathcal{C}$ be a countable Σ-decomposition of X. Then $\mathbb{E}(x \,|\, \sigma(\mathcal{C})) =_{\text{a.s.}} \mathbb{E}(x \,|\, \mathcal{C})$, where $\mathbb{E}(x \,|\, \mathcal{C})$ is defined in (2). More precisely, we have

$$\mathbb{E}(x \,|\, \sigma(\mathcal{C})) \underset{\text{a.s.}}{=} \sum_{C \in \mathcal{C}} \mathbb{E}(x \,|\, C) \mathbf{1}_C.$$

We leave proving this as an exercise. (*Hint.* Any element of $\sigma(\mathcal{C})$ is a countable union of the elements of $\mathcal{C}$.) *Insight*: The general definition of $\mathbb{E}(x \,|\, \Sigma_0)$ given above reduces to the one considered in (2) in the special case where Σ_0 is generated by a countable Σ-decomposition of the sample space. $\square$

Example 1.7. Let $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. If $C \in \Sigma_0$ and $\mathbf{p}(C) > 0$, then

$$\mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0) \,|\, C) = \mathbb{E}(x \,|\, C),$$

---

[7] This is an important point, make sure you grasp the underlying intuition. In general, the two requirements that determine a conditional expectation work against each other. In particular, the constant random variable $\mathbb{E}(x)$ need not be a version of $\mathbb{E}(x \,|\, \Sigma_0)$ due to the requirement (5), while the only reason why x need not be a version of $\mathbb{E}(x \,|\, \Sigma_0)$ is because it need not be Σ_0-measurable. Interpretation?

because

$$\begin{aligned}\mathbb{E}(\mathbb{E}(x\,|\,\Sigma_0)\,|\,C) &= \frac{1}{\mathbf{p}(C)}\int_C \mathbb{E}(x\,|\,\Sigma_0)d\mathbf{p} \\ &= \frac{1}{\mathbf{p}(C)}\int_C x d\mathbf{p} \\ &= \mathbb{E}(x\,|\,C).\end{aligned}$$

Interpretation should be clear. $\mathbb{E}(x\,|\,\Sigma_0)$ is the random variable that summarizes one's expectation of x today given that tomorrow (that is, after the experiment is run), she will learn which members of Σ_0 have occurred. So, the expectation of this random variable given that $C \in \Sigma_0$ has occurred is none other than what one would conclude today about the conditional expectation of x given that she is certain that C will occur. □

Remark 1.1. One of the important consequences of the condition (4) is the following: If $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$ and $y \in \mathcal{L}^0(X, \Sigma)$, then

$$\mathbb{E}(x\,|\,y)(\omega) = \mathbb{E}(x\,|\,y)(\nu) \quad \text{whenever} \quad y(\omega) = y(\nu).$$

To see this, take any $a \in y(X)$ and consider any two outcomes ω and ν in $y^{-1}(a)$. If $\mathbb{E}(x\,|\,y)(\omega) = \alpha$, then by (4), we have $A := \mathbb{E}(x\,|\,y)^{-1}(\alpha) \in \sigma(y)$, that is, $A = y^{-1}(S)$ for some Borel subset S of $\mathbb{R}$. But, since $\omega \in y^{-1}(S)$ and $y(\nu) = y(\omega)$, we clearly have $y(\nu) \in S$, that is, $\nu \in y^{-1}(S) = A$, and this means that $\mathbb{E}(x\,|\,y)(\nu) = \alpha$, as we sought.

For any $a \in y(X)$, this observation allows us to introduce unambiguously the *conditional expectation of* x *given that* $y = a$ as the number

$$\mathbb{E}(x\,|\,y = a) := \mathbb{E}(x\,|\,y)(\omega)$$

where ω is *any* member of $y^{-1}(a)$. Observe that $\mathbb{E}(x\,|\,y = a)$ is well-defined for any number $a \in y(X)$, even though $\{y = a\}$ may be a probability zero event. □

Example 1.8. Consider the probability space $([0,1], \mathcal{B}[0,1], \ell)$ and define the random variable x on this space by $x(\omega) := 2\omega$.

[**1**] Let y be a strictly increasing (or strictly decreasing) real function on $[0,1]$. (Intuitively speaking, the knowledge of y is fully informative about the underlying probabilistic model – if we know $y(\omega)$, then we know ω, and hence $x(\omega)$.) Clearly,

$\sigma(y)$ equals $\sigma\{\{y \leq a\} : a \in \mathbb{R}\}$ – yes? – whereas every $\{y \leq a\}$ is a closed interval in $[0,1]$. It follows that x is $\sigma(y)$-measurable, and hence, by the second part of Example 1.6, we have $\mathbb{E}(x \,|\, y) =_{\text{a.s.}} x$.

[2] Let $y_1(\omega) := 0$ for each $\omega \in [0,1]$, and define the real map y_m on $[0,1]$ as

$$y_m(\omega) := \begin{cases} 0, & \text{if } 0 \leq \omega \leq \frac{1}{m} \\ \omega, & \text{if } \frac{1}{m} < \omega \leq 1 \end{cases}$$

for any integer $m \geq 2$. (Would you agree that the larger m, the more informative is y_m?) Let us compute (a version of) $\mathbb{E}(x \,|\, y_m)$ for each m. To this end, we first need to find $\sigma(y_m)$ for each m. Obviously, $\sigma(y_1) = \{\emptyset, [0,1]\}$. To settle the general case, fix an arbitrary $m \geq 2$, and observe that, for any Borel subset S of $\mathbb{R}$, we have

$$y_m^{-1}(S) = S \cap (\tfrac{1}{m}, 1]$$

if 0 does not belong to S, and

$$y_m^{-1}(S) = [0, \tfrac{1}{m}] \cup (S \cap (\tfrac{1}{m}, 1])$$

if 0 belongs to S. It follows that $\sigma(y_m)$ is equal to

$$\left\{S \cap \left(\tfrac{1}{m}, 1\right] : S \in \mathcal{B}(\mathbb{R})\right\} \bigcup \left\{\left[0, \tfrac{1}{m}\right] \cup \left(S \cap \left(\tfrac{1}{m}, 1\right]\right) : S \in \mathcal{B}(\mathbb{R})\right\}.$$

Now, since y_m is constant on $[0, 1/m]$, so must $\mathbb{E}(x \,|\, y_m)$. (This was the content of Remark 1.1.) Therefore, for any ω in $[0, 1/m]$, we have

$$\begin{aligned} \mathbb{E}(x \,|\, y_m)(\omega) &= \frac{1}{\ell([0,1/m])} \int_{[0,1/m]} x d\ell \\ &= m \int_0^{1/m} 2\omega d\omega \\ &= \tfrac{1}{m}. \end{aligned}$$

(Right?) On the other hand, if we let $\mathbb{E}(x \,|\, y_m)(\omega) = x(\omega)$ for each $\omega \in (1/m, 1]$, we obviously have

$$\int_C \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} = \int_C x d\mathbf{p}$$

for every Borel subset C of $(1/m, 1]$ (that is, for every C that can be expressed as $S \cap (1/m, 1]$ for some Borel subset S of $\mathbb{R}$). We conclude that $z_m \in \mathcal{L}^0([0,1], \mathcal{B}[0,1])$ is a version of $\mathbb{E}(x \,|\, y_m)$, where

$$z_1 := 1 \quad \text{and} \quad z_m(\omega) := \begin{cases} 1/m, & \text{if } 0 \leq \omega \leq \frac{1}{m} \\ 2\omega, & \text{if } \frac{1}{m} < \omega \leq 1 \end{cases}$$

for every integer $m \geq 2$. (Check this by using the definition of $\mathbb{E}(x \,|\, y_m)$!) As one would expect, the knowledge of y_1 does not at all improve our ability to come up with a "guess" for the value of x, but for large m, the knowledge of y_m refines our expectation of x considerably. □

Example 1.9. At the beginning of this section, we considered the two-stage experiment in which, in the first stage, a real number α is randomly selected from $[0, 1]$, and in the second stage, a lottery that pays 1 with probability α and -1 with probability $1 - \alpha$ is played. A probabilistic model for the experiment is constructed as follows. Clearly, the associated outcome and event spaces are $X := [0, 1] \times \{-1, 1\}$ and $\Sigma := \mathcal{B}(X) = \mathcal{B}[0, 1] \times 2^{\{-1,1\}}$, respectively. (We denote a generic element (ω_1, ω_2) of X by ω in what follows.) A moment's reflection will show that our (product) probability measure $\mathbf{p}$ should be defined on Σ as:

$$\mathbf{p}(A \times \{1\}) := \int_A \omega_1 \ell(d\omega_1)$$

and

$$\mathbf{p}(A \times \{-1\}) := \int_A (1 - \omega_1)\ell(d\omega_1)$$

for every $A \in \mathcal{B}[0, 1]$. (Why?) In particular, we have

$$\mathbf{p}(A \times \{-1\}) = \ell(A) - \mathbf{p}(A \times \{1\})$$

for every Borel subset A of $[0, 1]$.)

Now consider the integrable random variables x and y defined on $(X, \Sigma, \mathbf{p})$ by $x(\omega) := \omega_2$ and $y(\omega) := \omega_1$. The question is: What is $\mathbb{E}(x \,|\, y)$? Clearly, the intuition suggests that

$$\mathbb{E}(x \,|\, y)(\omega) = \omega_1 + (1 - \omega_1)(-1) = 2\omega_1 - 1 \quad \text{for every } \omega \in X. \tag{6}$$

After all, if ω came up in the experiment, then the value of x is determined by a lottery that pays 1 with probability ω_1 and -1 with probability $1 - \omega_1$. In particular, this would imply that $\mathbb{E}(x \,|\, y = \frac{1}{2}) = 0$, as we surely wish to find.

Let us now show that our intuition is well supported by the formal definition of the conditional expectation by proving (6) – there is no escape, we have to get our hands dirty for this. Clearly, $\sigma(y) = \mathcal{B}[0, 1] \times \{-1, 1\}$, so the map $\omega \mapsto 2\omega_1 - 1$ (defined on X) is obviously $\sigma(y)$-measurable. Thus our main claim is that

$$\int_{A \times \{-1,1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) = \int_{A \times \{-1,1\}} \omega_2 \mathbf{p}(d\omega)$$

for every $A \in \mathcal{B}[0,1]$. Since $\mathcal{B}[0,1]$ equals $\sigma(\{[a,b] : 0 \leq a \leq b \leq 1\})$, it is enough to show that

$$\int_{[a,b]\times\{-1,1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) = \int_{[a,b]\times\{-1,1\}} \omega_2 \mathbf{p}(d\omega)$$

for any two real numbers a and b in $[0,1]$ with $a \leq b$. (See Exercise 1.7 below.) But we have

$$\begin{aligned}\int_{[a,b]\times\{-1,1\}} \omega_2 \mathbf{p}(d\omega) &= \int_{[a,b]\times\{-1\}} \omega_2 \mathbf{p}(d\omega) + \int_{[a,b]\times\{1\}} \omega_2 \mathbf{p}(d\omega) \\ &= (-1)\mathbf{p}([a,b]\times\{-1\}) + \mathbf{p}([a,b]\times\{1\}) \\ &= 2\mathbf{p}([a,b]\times\{1\}) - \ell([a,b]) \\ &= (b^2 - a^2) - (b - a),\end{aligned}$$

so our problem becomes proving that

$$\int_{[a,b]\times\{-1,1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) = (b^2 - a^2) - (b - a)$$

for any two real numbers a and b in $[0,1]$ with $a \leq b$. Let us take any two such numbers a and b. Define the finite measure $\mathbf{p}_1$ on $\mathcal{B}[0,1]$ by $\mathbf{p}_1(A) := \mathbf{p}(A \times \{1\})$, and notice that

$$\int_{[a,b]\times\{1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) = \int_{[a,b]} (2\omega_1 - 1)\mathbf{p}_1(d\omega).$$

But $\mathbf{p}_1$ is none other than the Lebesgue-Stieltjes measure induced by the real map F defined on $[0,1]$ by $F(t) := \int_0^t \omega_1 d\omega_1 = t^2/2$. Therefore,

$$\begin{aligned}\int_{[a,b]\times\{1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) &= \int_a^b (2t - 1)dF(t) \\ &= \int_a^b (2t - 1)t dt.\end{aligned}$$

One can similarly show that

$$\int_{[a,b]\times\{-1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) = \int_a^b (2t - 1)(1 - t)dt,$$

and hence

$$\begin{aligned}\int_{[a,b]\times\{-1,1\}} (2\omega_1 - 1)\mathbf{p}(d\omega) &= \int_a^b (2t - 1)t dt + \int_a^b (2t - 1)(1 - t)dt \\ &= (b^2 - a^2) - (b - a),\end{aligned}$$

as we sought. □

Exercise 1.4. Consider the probability space $([0,1),\mathcal{B}[0,1),\ell)$, and define the random variables x and y on this space as follows:

$$x(\omega) := 2\omega, \qquad y(\omega) := \begin{cases} 2\omega, & \text{if } 0 \leq \omega < 1/2 \\ 2\omega - 1, & \text{if } 1/2 \leq \omega < 1 \end{cases}.$$

(a) Compute $\sigma(y)$.
(b) Compute (a version of) $\mathbb{E}(x \,|\, y)$.

Exercise 1.5. Compute $\mathbb{E}(y \,|\, x)$ in the previous exercise.

Exercise 1.6. Consider the probability space $([0,1),\mathcal{B}[0,1),\ell)$, and define the random variables x and y on this space as follows:

$$x(\omega) := 2\omega, \qquad y(\omega) := 1 - |2\omega - 1|\,.$$

Compute (a version of) $\mathbb{E}(x \,|\, y)$.

Exercise 1.7.$^{\mathrm{H}}$ Let $(X,\Sigma,\mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. Take any nonempty collection $\mathcal{A}$ of subsets of X that includes X and that is closed under taking finite intersections. Prove that if $\sigma(\mathcal{A}) = \Sigma_0$ and if $y \in \mathcal{L}^0(X,\Sigma_0)$ satisfies $\int_A y d\mathbf{p} = \int_A x d\mathbf{p}$ for every $A \in \mathcal{A}$, then $y =_{\text{a.s.}} \mathbb{E}(x \,|\, \Sigma_0)$.

Exercise 1.8. (*Bayes' Theorem*) Let $(X,\Sigma,\mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. If $(A,C) \in \Sigma \times \Sigma_0$, then

$$\mathbf{p}(C \,|\, A) = \frac{\int_C \mathbf{p}(A \,|\, \Sigma_0) d\mathbf{p}}{\int_X \mathbf{p}(A \,|\, \Sigma_0) d\mathbf{p}}.$$

How would you write this formula if, for some finite Σ-decomposition $\mathcal{C}$ of X, we had $\Sigma_0 = \sigma(\mathcal{C})$?

Exercise 1.9.$^{\mathrm{H}}$ Let x be an integrable random variable on a probability space $(X,\Sigma,\mathbf{p})$, and $y \in \mathcal{L}^0(X,\Sigma)$. Prove or disprove: $\mathbb{E}(x \,|\, y) =_{\text{a.s.}} x$ iff $x =_{\text{a.s.}} f(y)$ for some $f \in \mathcal{L}^0(y(X),\mathcal{B}(y(X)))$.

Exercise 1.10.$^{\mathrm{H}}$ Let x and y be two integrable random variables on a probability space $(X,\Sigma,\mathbf{p})$ such that $\mathbb{E}(x \,|\, y) =_{\text{a.s.}} y$ and $\mathbb{E}(y \,|\, x) =_{\text{a.s.}} x$. Prove that $x =_{\text{a.s.}} y$.

Exercise 1.11. Let $(X,\Sigma,\mathbf{p})$ be a probability space and $x, y \in \mathcal{L}^2(X,\Sigma,\mathbf{p})$. Suppose $\mathbb{E}(x \,|\, y) =_{\text{a.s.}} f(y)$ for some increasing Borel measurable real function f on $y(X)$. Prove that $\mathrm{Cov}(x,y) \geq 0$.

Exercise 1.12. (*Conditional Markov Inequality*) Let x be an integrable random variable on a probability space $(X,\Sigma,\mathbf{p})$. Prove that

$$\mathbf{p}(|x| \geq a \,|\, \Sigma_0) \leq_{\text{a.s.}} \frac{\mathbb{E}(|x| \,|\, \Sigma_0)}{a}$$

for any sub-σ-algebra Σ_0 of Σ and real number $a > 0$.

Exercise 1.13.$^{\mathrm{H}}$ Let (x_m) be a sequence of integrable and i.i.d. random variables on a probability space. For every positive integer m, define $y_m := x_1 + \cdots + x_m$. Show that

$$\mathbb{E}(x_1 \,|\, y_m) \underset{\text{a.s.}}{=} \cdots \underset{\text{a.s.}}{=} \mathbb{E}(x_m \,|\, y_m) \underset{\text{a.s.}}{=} \frac{y_m}{m}.$$

1.3 Conditional Density Functions

Let x and y be two integrable random variables on a probability space $(X, \Sigma, \mathbf{p})$, and let f be a joint density for x and y. The **conditional density of** x **given** y is defined as

$$f_{x|y}(s,t) := \begin{cases} f(s,t)/f_y(t), & \text{if } f_y(t) > 0, \\ 0, & \text{otherwise} \end{cases},$$

where f_y is the marginal density of y. We next show that

$$\mathbb{E}(x \,|\, y) \underset{\text{a.s.}}{=} \int_{-\infty}^{\infty} s f_{x|y}(s, y(\cdot))ds \tag{7}$$

provided that $f_{x|y}$ is nice.[8] We leave it to you to prove that $\omega \mapsto \int_{-\infty}^{\infty} s f_{x|y}(s, y(\omega))ds$ is $\sigma(y)$-measurable. To establish the rest of the claim, take any real numbers a and b with $a \leq b$, and check that the niceness of $f_{x|y}$ implies the piecewise continuity of the self-map $\phi_{a,b}$ on $\mathbb{R}$, where

$$\phi_{a,b}(t) = \begin{cases} \frac{1}{f_y(t)} \int_{-\infty}^{\infty} s f(s,t)ds, & \text{if } f_y(t) > 0 \text{ and } t \in (a,b) \\ 0, & \text{otherwise} \end{cases}.$$

Applying Corollary D.2.3 and F.4.6, and Exercise F.4.8, we have

$$\begin{aligned} \mathbb{E}(\phi_{a,b} \circ y) &= \int_a^b \phi_{a,b}(t) f_y(t) dt \\ &= \int_a^b \int_{-\infty}^{\infty} s f(s,t) ds dt \\ &= \mathbb{E}(x \mathbf{1}_{y^{-1}((a,b))}). \end{aligned}$$

But, obviously, $\mathbb{E}(\phi_{a,b} \circ y) = \int_{y^{-1}((a,b))} \phi_{a,b}(y) d\mathbf{p}$, so we conclude:

$$\int_{y^{-1}((a,b))} \phi_{a,b}(y) d\mathbf{p} = \int_{y^{-1}((a,b))} x d\mathbf{p}$$

for every real numbers a and b with $a \leq b$. Since $\sigma(y)$ equals $\sigma(y^{-1}((a,b)) : -\infty \leq a \leq b \leq \infty\}$, therefore, (7) follows from Exercise 1.7.

Exercise 1.13. Let x be a uniformly distributed random variable on $[0,1]$. Compute $\mathbb{E}(x \,|\, x(1-x))$.

---

[8] Since $\mathbb{E}(x \,|\, y)$ is determined up to probability one, how one may define $f_{x|y}$ at a (u,v) with $f_y(v) = 0$ is irrelevant. After all, $\mathbf{p}\{f_y \circ y = 0\} = 0$, right?

Exercise 1.14. Let x and y be two random variables with the joint density $f_{x,y}$ defined by

$$f_{x,y}(s,t) := \begin{cases} 1, & \text{if } 0 \leq s,t \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Compute $\mathbb{E}(x \,|\, y)$ and $\mathbb{E}(x \,|\, x+y)$.

Exercise 1.15. The same question as the previous one, but with

$$f_{x,y}(s,t) := \begin{cases} s+t, & \text{if } 0 \leq s,t \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

Exercise 1.16. Suppose that x and y are two random variables whose distribution functions satisfy $F_x(t) = t^2$ and $F_y(t) = 2(t - t^2/2)$ for any t in $[0,1]$. Compute $\mathbb{E}(x \,|\, y)$ and $\mathbb{E}(x^2 \,|\, y)$.

2 Properties of Conditional Expectation

2.1 Elementary Properties

The following two propositions summarize some of the most basic properties of conditional expectations. We will use these properties frequently in the sequel.

Proposition 2.1. *Let x and y be integrable random variables on a probability space $(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. Then, for every real number α,*

$$\mathbb{E}(\alpha x \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \alpha \mathbb{E}(x \,|\, \Sigma_0)$$

and

$$\mathbb{E}(x + y \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \mathbb{E}(x \,|\, \Sigma_0) + \mathbb{E}(y \,|\, \Sigma_0).$$

Moreover,

$$x \underset{\text{a.s.}}{=} y \quad \textit{implies} \quad \mathbb{E}(x \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \mathbb{E}(y| \, \Sigma_0).$$

Exercise 2.1.$^{\mathrm{H}}$ Prove Proposition 2.1.

Proposition 2.2. *Let x be an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. Then,*

$$\mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0)) = \mathbb{E}(x), \tag{8}$$

and if Σ_1 *is another sub-*σ*-algebra of* Σ*, then*

$$\Sigma_0 \subseteq \Sigma_1 \quad \textit{implies} \quad \mathbb{E}(\mathbb{E}(x \,|\, \Sigma_1) \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \mathbb{E}(x \,|\, \Sigma_0).$$

Exercise 2.2.$^{\text{H}}$ Prove Proposition 2.2.

Exercise 2.3. True or False: Under the conditions of Proposition 2.2, $\Sigma_0 \subseteq \Sigma_1$ implies $\mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0) \,|\, \Sigma_1) =_{\text{a.s.}} \mathbb{E}(x \,|\, \Sigma_0)$.

While their statements may seem strange at first, the formulas in these results are in fact quite intuitive. Let us consider (8), for instance. Since one can intuitively think of $\mathbb{E}(x \,|\, \Sigma_0)$ as a random variable which is formed by an averaging process, taking the average of the values of $\mathbb{E}(x \,|\, \Sigma_0)$ should give us the average we would have found if we computed the *un*conditional expectation directly. After all, if A is a nonempty finite set of numbers, and $\{A_1, ..., A_k\}$ is a partition of A, then the average of the averages of each A_i must be the average of the numbers in A. This is the idea behind the fact that $\mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0)) = \mathbb{E}(x)$.[9] (Of course, we could not prove the formula this way, because the averaging idea is only an intuitive one – it is only implicitly contained in the formal definition of $\mathbb{E}(x \,|\, \Sigma_0)$.) The interpretation of the second claim of Proposition 2.2 – this is sometimes called the **filtering property** of conditional expectations – is analogous.

Proposition 2.3. *Let* x *be an integrable random variable on a probability space* $(X, \Sigma, \mathbf{p})$ *and* Σ_0 *a sub-*σ*-algebra of* Σ*. Then,*

$$x \underset{\text{a.s.}}{\geq} 0 \quad \textit{implies} \quad \mathbb{E}(x \,|\, \Sigma_0) \underset{\text{a.s.}}{\geq} 0,$$

and if x *is independent of* Σ_0 *(that is,* $\coprod\{\sigma(x), \Sigma_0\}$*), we have*

$$\mathbb{E}(x \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \mathbb{E}(x).$$

---

[9] *Example.* Consider the experiment of throwing a fair die, which induces the probability space $(X, 2^X, \mathbf{p})$ with $X := \{1, ..., 6\}$ and $\mathbf{p}\{\omega\} := 1/6$ for each $\omega \in X$. Define the real map x on X with $x(\omega) := \omega$, and let $\Sigma_0 := \{\emptyset, C, X \backslash C, X\}$ where $C := \{2, 4, 6\}$. Then $\mathbb{E}(x \,|\, \Sigma_0)$ is a random variable that takes two values (3 and 4), each with probability 1/2, and we have

$$\mathbb{E}(x) = \tfrac{1}{6} + \tfrac{2}{6} + \cdots + \tfrac{6}{6} = \tfrac{1}{2}\left(\tfrac{1}{1/2}\left(\tfrac{1}{6} + \tfrac{3}{6} + \tfrac{5}{6}\right)\right) + \tfrac{1}{2}\left(\tfrac{1}{1/2}\left(\tfrac{2}{6} + \tfrac{4}{6} + \tfrac{6}{6}\right)\right) = \mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0)).$$

Proof. For any positive integer m, let $B_m := \{\mathbb{E}(x \mid \Sigma_0) \leq -\frac{1}{m}\}$, and observe that $B_m \in \Sigma_0$. Then, $x \geq_{\text{a.s.}} 0$ implies

$$-\frac{\mathbf{p}(B_m)}{m} \geq \int_{B_m} \mathbb{E}(x \mid \Sigma_0) d\mathbf{p} = \int_{B_m} x d\mathbf{p} \geq 0$$

so that $\mathbf{p}(B_m) = 0$ for any $m \in \mathbb{N}$. But (B_m) is an increasing sequence with $\bigcup^{\infty} B_i = \{\mathbb{E}(x \mid \Sigma_0) < 0\}$, so the claim follows from Proposition B.2.2.

To prove the second claim, notice that the independence of x and Σ_0 entails that x and $\mathbf{1}_B$ are independent for any $B \in \Sigma_0$, so by Proposition G.2.1,

$$\int_B \mathbb{E}(x) d\mathbf{p} = \int_X \mathbb{E}(x) \mathbf{1}_B d\mathbf{p} = \mathbb{E}(x)\mathbb{E}(\mathbf{1}_B) = \mathbb{E}(x\mathbf{1}_B) = \int_B x d\mathbf{p}$$

for any $B \in \Sigma_0$. ■

An immediate corollary of Propositions 2.1 and 2.3 is that

$$x \underset{\text{a.s.}}{\geq} y \quad \text{implies} \quad \mathbb{E}(x \mid \Sigma_0) \underset{\text{a.s.}}{\geq} \mathbb{E}(y \mid \Sigma_0) \tag{9}$$

for any $x, y \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$. On the other hand, Proposition 2.3 implies that $\mathbb{E}(x \mid y) =_{\text{a.s.}} \mathbb{E}(x)$ for any two *independent* (and integrable) random variables x and y on some probability space. This is, of course, in keeping with our interpretation of conditional expectation. If x and y are independent, receiving information about the value of y does not change our expectation about x.

Exercise 2.4. Prove: If x is an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$, then $|\mathbb{E}(x \mid \Sigma_0)| \leq_{\text{a.s.}} \mathbb{E}(|x| \mid \Sigma_0)$ for any sub-σ-algebra Σ_0 of Σ.

Exercise 2.5. Find two random variables x and y on a probability space such that $\mathbb{E}(x \mid y) =_{\text{a.s.}} \mathbb{E}(x)$ but x and y are not independent.

Remark 2.1. Let $(X, \Sigma, \mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. Let us denote $\mathcal{L}^1(X, \Sigma, \mathbf{p})$ simply by $\mathcal{L}^1$, and define $L : \mathcal{L}^1 \to \mathcal{L}^0(X, \Sigma)$ such that $L(x)$ is a real-valued version of $\mathbb{E}(x \mid \Sigma_0)$ for each $x \in \mathcal{L}^1$. By Exercise 2.4, and Proposition 2.2, we have

$$\int_X |L(x)| \, d\mathbf{p} = \mathbb{E}(|\mathbb{E}(x \mid \Sigma_0)|) \leq \mathbb{E}(\mathbb{E}(|x| \mid \Sigma_0)) = \mathbb{E}(|x|)$$

for any $x \in \mathcal{L}^1$. Conclusion: $L(\mathcal{L}^1) \subseteq \mathcal{L}^1$. Combining this observation with Proposition 2.1, therefore, we may conclude that L is a *linear operator* mapping the Banach space $\mathcal{L}^1$ into itself. Moreover, by Proposition 2.3, L is a *positive* operator on $\mathcal{L}^1$ in the sense that $L(x) \geq_{\text{a.s.}} 0$ for every $x \geq_{\text{a.s.}} 0$. This is what one means when she refers to $\mathbb{E}(\cdot \,|\, \Sigma_0)$ as a *positive linear operator.* $\square$

Under the conditions of Proposition 2.1, $\alpha \mathbf{1}_X \mathbb{E}(x \,|\, \Sigma_0)$ is a version of $\mathbb{E}(\alpha \mathbf{1}_X x \,|\, \Sigma_0)$ for any real number α. The following result gives a useful generalization of this fact showing that one can replace here the constant random variable $\alpha \mathbf{1}_X$ with any Σ_0-measurable random variable y such that xy is integrable (even if y itself is not integrable). Intuitively speaking, "we can take the "known" random variable out of the conditional expectation." (Why do we refer to y as "known" here?)

Proposition 2.4. *Let x be an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. If $y \in \mathcal{L}^0(X, \Sigma_0)$ and $E(|xy|) < \infty$, then*

$$\mathbb{E}(xy \,|\, \Sigma_0) \underset{\text{a.s.}}{=} y \mathbb{E}(x \,|\, \Sigma_0).$$

Proof. Note first that the Σ_0-measurability of $\mathbb{E}(x \,|\, \Sigma_0)$ and y imply that of $y\mathbb{E}(x \,|\, \Sigma_0)$. It is then enough to establish that

$$\int_B y \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} = \int_B yx d\mathbf{p} \quad \text{for every } B \in \Sigma_0.$$

To this end, we consider first the case where y is simple, that is, we assume there exist a positive integer m, real numbers $a_1, ..., a_m$, and sets $A_1, ..., A_m \in \Sigma_0$ such that $y = a_1 \mathbf{1}_{A_1} + \cdots + a_m \mathbf{1}_{A_m}$. In this case, we have

$$\begin{aligned} \int_B y \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} &= \sum_{i=1}^m a_i \int_{B \cap A_i} \mathbb{E}(x \,|\, \Sigma_0) d\mathbf{p} \\ &= \sum_{i=1}^m a_i \int_{B \cap A_i} x d\mathbf{p} \\ &= \int_B yx d\mathbf{p} \end{aligned}$$

for every $B \in \Sigma_0$. The case where y is not simple but nonnegative is then established by an appeal to Lemma C.1.2 and the Monotone Convergence Theorem 1, and finally, the general case is settled by splitting y into its positive and negative parts, and using the linearity of $\mathbb{E}(\cdot \,|\, \Sigma_0)$. $\blacksquare$

Exercise 2.6. Complete the proof of Proposition 2.4.

Exercise 2.7. Let x and y be two bounded random variables on a probability space $(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. Show that $\mathbb{E}(y\mathbb{E}(x \,|\, \Sigma_0)) =_{\text{a.s.}} \mathbb{E}(x\mathbb{E}(y \,|\, \Sigma_0))$.

Exercise 2.8. Let x, y and z be integrable random variables on a probability space $(X, \Sigma, \mathbf{p})$. Show that if both x and y are independent of z, then $\mathbb{E}(x \,|\, y, z) =_{\text{a.s.}} \mathbb{E}(x \,|\, y)$. (Here $\mathbb{E}(x \,|\, y, z)$ stands for $\mathbb{E}(x \,|\, \sigma\{y, z\})$, where $\sigma\{y, z\}$ is the smallest σ-algebra generated by y and z.)

Exercise 2.9. Let $(X, \Sigma, \mathbf{p})$ be a probability space and Σ_0 a sub-σ-algebra of Σ. Show that

$$\mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0)^2) = \mathbb{E}(x\mathbb{E}(x \,|\, \Sigma_0))$$

and

$$\mathbb{E}((x - \mathbb{E}(y \,|\, \Sigma_0))^2) = \mathbb{E}(x^2) - \mathbb{E}(\mathbb{E}(x \,|\, \Sigma_0)^2)$$

for any $x \in \mathcal{L}^2(X, \Sigma, \mathbf{p})$.

Exercise 2.10. Let $(X, \Sigma, \mathbf{p})$ be a probability space and Σ_0 a sub-σ-algebra of Σ. For any $x \in \mathcal{L}^2(X, \Sigma, \mathbf{p})$, the **conditional variance** of x **given** Σ_0, denoted as $\mathbb{V}(x \,|\, \Sigma_0)$, is defined as (a version of) the random variable $\mathbb{E}((x - \mathbb{E}(x \,|\, \Sigma_0))^2 \,|\, \Sigma_0)$.
(a) Show that

$$\mathbb{V}(x) = \mathbb{E}(\mathbb{V}(x \,|\, \Sigma_0)) + \mathbb{V}(\mathbb{E}(x \,|\, \Sigma_0)).$$

(b) (*Rao-Blackwell Theorem*) Prove: There is a $y \in \mathcal{L}^2(X, \Sigma_0, \mathbf{p})$ such that $\mathbb{E}(y) = \mathbb{E}(x)$ and $\mathbb{V}(y) \leq \mathbb{V}(x)$. (In particular, $\mathbb{E}(x \,|\, \Sigma_0)$ is up to the task.)

Exercise 2.11. Let $(X, \Sigma, \mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. For any two collections $\mathcal{A}$ and $\mathcal{B}$ of events in Σ, we say that $\mathcal{A}$ and $\mathcal{B}$ are **conditionally independent given** Σ_0, if

$$\mathbf{p}(A \cap B \,|\, \Sigma_0) \underset{\text{a.s.}}{=} \mathbf{p}(A \,|\, \Sigma_0)\mathbf{p}(B \,|\, \Sigma_0) \qquad \text{for every } (A, B) \in \mathcal{A} \times \mathcal{B}.$$

Prove that the following statements are equivalent:
(i) $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given Σ_0;
(ii) For every $A \in \mathcal{A}$, we have $\mathbf{p}(A \,|\, \sigma(\mathcal{B} \cup \Sigma_0)) =_{\text{a.s.}} \mathbf{p}(A \,|\, \Sigma_0)$.

2.2 Convergence Properties

We next extend some of the basic convergence theorems we have established in Chapter C for the expectation functional to the case of the conditional expectation operator. Since the expectation of a random variable on (X, Σ) is a version of its conditional expectation given the trivial sub-σ-algebra $\{\emptyset, X\}$, the following results are in fact generalizations of the corresponding theorems we obtained in that chapter.

The Conditional Monotone Convergence Theorem. *Let* $x, x_1, x_2, \ldots$ *be nonnegative integrable random variables on a probability space* $(X, \Sigma, \mathbf{p})$ *such that* $x_m \nearrow_{\text{a.s.}} x$.[10] *If* Σ_0 *is a sub-*σ*-algebra of* Σ*, then*

$$\mathbb{E}\left(x_m \,|\, \Sigma_0\right) \nearrow_{\text{a.s.}} \mathbb{E}(x \,|\, \Sigma_0).$$

Proof. For every positive integer m, let y_m be a version of $\mathbb{E}(x_m \,|\, \Sigma_0)$. We wish to show that the sequence (y_m) converges, almost surely, to a version of $\mathbb{E}(x \,|\, \Sigma_0)$.[11] Now, since each y_m is Σ_0-measurable, $S := \{0 \leq y_1 \leq y_2 \leq \cdots\}$ belongs to Σ_0. (Yes?) Moreover, as x_1 is nonnegative and $x_m \nearrow_{\text{a.s.}} x$, (9) guarantees that $\mathbf{p}(S) = 1$. (Yes?) Therefore, $z_m := y_m \mathbf{1}_S$ is also a version of $\mathbb{E}(x_m \,|\, \Sigma_0)$ for each m. As we have $0 \leq z_1 \leq z_2 \leq \cdots$, then, $z := \lim z_m$ is a Σ_0-measurable $\overline{\mathbb{R}}_+$-valued map on X (Example B.5.5). Furthermore, $y_m \nearrow_{\text{a.s.}} z$, so all we need to show here is that $z =_{\text{a.s.}} \mathbb{E}(x \,|\, \Sigma_0)$.

Fix any event B in Σ_0. Indeed, by the Monotone Convergence Theorem 1 and since z_m is (a version of) $\mathbb{E}(x_m \,|\, \Sigma_0)$,

$$\int_B z d\mathbf{p} = \lim \int_B z_m d\mathbf{p} = \lim \int_B x_m d\mathbf{p}.$$

But by the Monotone Convergence Theorem 3,

$$\lim \int_B x_m d\mathbf{p} = \int_B x d\mathbf{p}.$$

It follows that $\int_B z d\mathbf{p} = \int_B x d\mathbf{p}$ for every $B \in \Sigma_0$. Since z is Σ_0-measurable, we may thus conclude that z is (a version of) $\mathbb{E}(x \,|\, \Sigma_0)$. ■

Exercise 2.12. Let $(X, \Sigma, \mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. For any events $S, S_1, S_2, \ldots$ in Σ, show that

$$S_m \nearrow S \quad \text{implies} \quad \mathbf{p}(S_m \,|\, \Sigma_0) \nearrow_{\text{a.s.}} \mathbf{p}(S \,|\, \Sigma_0).$$

Exercise 2.13. Let $(X, \Sigma, \mathbf{p})$ be a probability space, and Σ_0 a sub-σ-algebra of Σ. Prove: If $\mathcal{A} \subseteq \Sigma$ is a (nonempty) countable collection of disjoint events, then

$$\mathbf{p}(\bigcup \mathcal{A} \,|\, \Sigma_0) = \sum_{A \in \mathcal{A}} \mathbf{p}(A \,|\, \Sigma_0).$$

---

[10] *Reminder.* $x_m \nearrow_{\text{a.s.}} x$ means that $x_{m+1} \geq_{\text{a.s}} x_m$ for each positive integer m, and $x_m \to_{\text{a.s.}} x$.

[11] There is a room for a minor caution here: $\lim y_m$ need not be a version of $\mathbb{E}(x \,|\, \Sigma_0)$, for it may not exist on an event S, and even if it exists, it may not be Σ_0 measurable (unless $(X, \Sigma_0, \mathbf{p})$ is a complete probability space). But, of course, such a "bad" event S must have zero probability, so all should go well if we modify y_ms suitably on that event.

Other conditional convergence theorems are deduced from their unconditional versions by means of a similar method. For instance, we have:

The Conditional Dominated Convergence Theorem. *Let $x, y, x_1, x_2, \ldots$ integrable random variables on a probability space $(X, \Sigma, \mathbf{p})$. If $x_m \to_{\text{a.s.}} x$ and $|x_m| \leq_{\text{a.s.}} y$ for each m, then*

$$\mathbb{E}(x_m \,|\, \Sigma_0) \to_{\text{a.s.}} \mathbb{E}(x \,|\, \Sigma_0).$$

Exercise 2.14.$^{\text{H}}$ Prove the Conditional Dominated Convergence Theorem.

Exercise 2.15. State and prove a version of Fatou's Lemma for conditional expectations.

Finally, let us put on record the conditional version of Jensen's Inequality.

The Conditional Jensen's Inequality. *Let x be an integrable random variable on a probability space $(X, \Sigma, \mathbf{p})$, and Σ_0 a sub-σ-algebra of Σ. If ϕ is a concave self-map on $\mathbb{R}$ such that $\mathbb{E}(|\phi \circ x|) < \infty$, then*

$$\varphi(\mathbb{E}(x \,|\, \Sigma_0)) \underset{\text{a.s.}}{\geq} \mathbb{E}(\varphi \circ x \,|\, \Sigma_0).$$

Exercise 2.16.$^{\text{H}}$ Prove the Conditional Jensen's Inequality.

Exercise 2.17. Under the conditions of the Conditional Jensen's Inequality, show that

$$\int_X |\mathbb{E}(x \,|\, \Sigma_0)|^p \, d\mathbf{p} \leq \int_X |x|^p \, d\mathbf{p} \qquad \text{for every } p \geq 1.$$

Exercise 2.18. Let x and y be two nonnegative integrable random variables on a probability space. Assume that there exist integrable random variables z and w on some probability space such that $\mathbf{p}_x = \mathbf{p}_z$, $\mathbf{p}_y = \mathbf{p}_w$ and $z \geq_{\text{a.s.}} \mathbb{E}(w \,|\, z)$. Show that $x \succcurlyeq_{\text{SSD}} y$.[12]

2.3 The Existence of Conditional Expectation

As a final order of business in this section, we wish to prove Theorem 1.1. Our proof is geometric, which may come to you as a surprise. Here is the argument. Take any probability space $(X, \Sigma, \mathbf{p})$ and any $x \in \mathcal{L}^2(X, \Sigma, \mathbf{p})$. Let Σ_0 be a sub-σ-algebra of Σ. By the Riesz-Fischer Theorem (Section C.6), $Y := \mathcal{L}^2(X, \Sigma_0, \mathbf{p})$ is a closed linear subspace of the Hilbert space $\mathcal{L}^2(X, \Sigma, \mathbf{p})$. So, by the Orthogonal Projection Theorem – see Appendix B – there exists a $y \in Y$ such that

$$\int_X (x - y) z d\mathbf{p} = 0 \qquad \text{for every } z \in Y.$$

---

[12] Actually, the converse of this result is also true – this is sometimes called *Strassen's Theorem* – but it is much harder to prove.

(Of course, y is the closest point to x among all points in Y.) It is obvious that y is Σ_0-measurable, and by choosing $z = \mathbf{1}_B$, we get $\int_B x d\mathbf{p} = \int_B y d\mathbf{p}$ for every $B \in \Sigma_0$. Thus y is a version of $\mathbb{E}(x \,|\, \Sigma_0)$. We just proved:

Proposition 2.5. *For any probability space $(X, \Sigma, \mathbf{p})$ and sub-σ-algebra Σ_0 of Σ, the orthogonal projection of any $x \in \mathcal{L}^2(X, \Sigma, \mathbf{p})$ onto $\mathcal{L}^2(X, \Sigma_0, \mathbf{p})$ is a version of $\mathbb{E}(x \,|\, \Sigma_0)$.*

Thanks to the Orthogonal Projection Theorem, therefore, we now know that $\mathbb{E}(x \,|\, \Sigma_0)$ exists for any $x \in \mathcal{L}^2(X, \Sigma_0, \mathbf{p})$. This is not good enough though. For, to prove Theorem 1.1, we need to extend this fact to the case of $\mathcal{L}^1(X, \Sigma_0, \mathbf{p})$. We next show that this too can be accomplished by an easy approximation argument based on Proposition 2.5. (However, we should be careful not to invoke any of the results we proved earlier (such as the Conditional Monotone Convergence Theorem) which presumed the validity of Theorem 1.1.)

Proof of Theorem 1.1. Let $x \in \mathcal{L}^1(X, \Sigma, \mathbf{p})$. Assume first that $x \geq 0$. We know that there exists a sequence (x_m) of nonnegative simple random variables on $(X, \Sigma, \mathbf{p})$ such that $x_m \nearrow x$. Since each x_m is in $\mathcal{L}^2(X, \Sigma, \mathbf{p})$, it follows from Proposition 2.5 that $\mathbb{E}(x_m \,|\, \Sigma_0)$ exists for each m. So, by Proposition 2.3, for each positive integer m, there exists a version y_m of $\mathbb{E}(x_m \,|\, \Sigma_0)$ such that $y_m \geq 0$. (Why exactly?) Now define $y := \limsup y_m$ and note that y is Σ_0-measurable. But $y_m \nearrow_{\text{a.s.}} y$, so by the Monotone Convergence Theorem 3, we get

$$\int_B x d\mathbf{p} - \int_B y d\mathbf{p} = \lim \left(\int_B x_m d\mathbf{p} - \int_B y_m d\mathbf{p} \right) = 0$$

for every $B \in \Sigma_0$. Thus y is a version of $\mathbb{E}(x \,|\, \Sigma_0)$. This is, in fact, all we needed to complete the proof, for in the general case (where $x \geq 0$ need not hold but we are given the integrability of x), it is easy to verify that $y^+ - y^-$ is a version of $\mathbb{E}(x \,|\, \Sigma_0)$, where y^+ is a version of $\mathbb{E}(x^+ \,|\, \Sigma_0)$ and y^- that of $\mathbb{E}(x^- \,|\, \Sigma_0)$. ■

We conclude this section by noting that Proposition 2.5 provides one with a nontrivial interpretation of the notion of conditional expectation of a random variable with finite variance. Indeed, if $x \in \mathcal{L}^2(X, \Sigma, \mathbf{p})$ and Σ_0 is a sub-σ-algebra of Σ, then Proposition 2.5 entails that

$$\|x - \mathbb{E}(x \,|\, \Sigma_0)\|_2 = \min \left\{ \|x - z\|_2 : z \in \mathcal{L}^2(X, \Sigma_0, \mathbf{p}) \right\}.$$

In words, we may thus say that $\mathbb{E}(x \,|\, \Sigma_0)$ is the *best predictor* of x given Σ_0, in the sense that $\mathbb{E}(x \,|\, \Sigma_0)$ is "the" random variable in $\mathcal{L}^2(X, \Sigma, \mathbf{p})$ that is closest to x (with respect to the standard metric on $\mathcal{L}^2(X, \Sigma, \mathbf{p})$). We shall find a good occasion for adopting this interpretation in the next chapter.