5. Conditional Expectation

A. Definition of conditional expectation

Suppose that we have partial information about the outcome ω, drawn from Ω according to probability measure $I\!P$; the partial information might be in the form of the value of a random vector $Y(\omega)$ or of an event B in which ω is known to lie. The concept of conditional expectation tells us how to calculate the expected values and probabilities using this information.

The general definition of conditional expectation is fairly abstract and takes a bit of getting used to. We shall first build intuition by recalling the definitions of conditional expectaion in elementary elementary probability theory, and showing how they can be used to compute certain Radon-Nikodym derivatives. With this as motivation, we then develop the fully general definition. Before doing any probability, though, we pause to review the Radon-Nikodym theorem, which plays an essential role in the theory.

The Radon-Nikodym theorem. Let ν be a positive measure on a measurable space $(S, \mathcal{S})$. If μ is a signed measure on $(S, \mathcal{S})$, μ is said to be absolutely continuous with respect to ν, written $\mu \ll \nu$, if

$$\mu(A) = 0 \quad \text{whenever} \quad \nu(A) = 0.$$

Suppose that f is a ν-integrable function on $(S, \mathcal{S})$ and define the new measure $\mu(A) = \int_A f(s)\nu(ds)$, $A \in \mathcal{S}$. Then, clearly, $\mu \ll \nu$. The Radon-Nikodym theorem says that all measures which are absolutely continuous to ν arise in this way.

Theorem (Radon-Nikodym). Let ν be a σ-finite, positive measure on $(S\, sS)$. (This means that there is a countable covering $A_1, A_2, \ldots$ of S by subsets of $\mathcal{S}$ such that $\nu(A_i) < \infty$ for all i; in particular, bounded measures, and hence probability measures, are σ-finite.) Let μ be absolutely continuous w.r.t ν. Then there is a measurable function g on $(S, \mathcal{S})$ such that

$$\mu(A) = \int_A g(s)\nu(ds), \quad A \in \mathcal{S}.$$

The function g is unique up to ν-a.e. equivalence. That is, if μ can also be defined as above using f instead of g, then $f = g$, ν-a.e. The function g is called the Radon-Nikodym derivative of μ w.r.t. ν and is denoted by $\dfrac{d\mu}{d\nu}$.

A discussion of the Radon-Nikodym derivative may be found in any good text on measure and integration theory.

Conditional expectation given an event; elementary definition and characterization as a Radon-Nikodym derivative. In elementary probability theory one defines the *conditional probability of A given B* by

$$I\!P(A/B) = \frac{I\!P(A \cap B)}{I\!P(B)},$$

if B is an event of positive probability. One thinks of $\mathbb{P}(A/B)$ as the probability that A occurs given that B is known to have occured. This is certainly correct for the frequentist interpretation of probability; in independent repeated trials, the law of large numbers says that $\mathbb{P}(A \cap B)/\mathbb{P}(B)$ is the asymptotic frequency of occurence of A in those trials in which B occurs, since $\mathbb{P}(A \cap B)$ and $\mathbb{P}(B)$ are, respectively, the limiting frequencies of occurrence of $A \cap B$ and B. Continuing in this vein, the conditional expectation of a random variable X given the event B has occured should be the long-time average value of X in those trials in which B occurs. This leads to the definition

$$E[X/B] := \frac{E[\mathbf{1}_B X]}{\mathbb{P}(B)}. \tag{1}$$

That is, $E[X/B]$ is the $\mathbb{P}$-weighted average of X over B.

So far, we have just followed the definitions of elementary probability theory. To understand them better, we will take a deeper look. Consider a partition of Ω into disjoint events $B_1, \ldots, B_n$. Let $\mathcal{G}$ be the (finite) σ-algebra consisting of all possible unions of these events. Let X be an integrable random variable. We shall define a new random variable,

$$E[X/\mathcal{G}](\omega) = \sum_{i=1}^{n} E[X/B_i]\mathbf{1}_{B_i}(\omega) \tag{2}$$

Think of this random variable as follows: a sample point ω is drawn at random from the probability space Ω according to probability measure $\mathbb{P}$. The experimenter does not learn the exact value of ω, only the set B_i into which it falls. She then computes the expected value of X given this information, which, according to (1), is $E[X/B_i]$.

Now, the main point is that $E[X/\mathcal{G}]$ has a natural, measure theoretic interpretation. We claima that

$$E\left[\mathbf{1}_A X\right] = E\left[\mathbf{1}_A E[X/\mathcal{G}]\right], \quad \text{for every } A \in \mathcal{G}, \tag{3}$$

and $E[X/\mathcal{G}]$ is the unique $\mathcal{G}$-measurable function with this property. Let us check that (3) is true. Set $Z = \sum_{i=1}^{n} E[X/B_i]\mathbf{1}_{B_i}$ for convenience of notation . Certainly, Z is $\mathcal{G}$-measurable. If A is an event in $\mathcal{G}$, there is a subset $I \subset \{1, \ldots, n\}$ such that $A = \cup_{i \in I} B_i$. Then, since the B_i, $1 \leq i \leq n$ are disjoint,

$$\begin{aligned} E\left[\mathbf{1}_A Z\right] &= E\left[\sum_{i \in I} \frac{E[\mathbf{1}_{B_i} X]}{\mathbb{P}(B_i)} \mathbf{1}_{B_i}\right] \\ &= \sum_{i \in I} E[\mathbf{1}_{B_i} X] \\ &= E\left[\mathbf{1}_A Z\right]. \end{aligned}$$

Conversely, suppose Z' is a $\mathcal{G}$ measurable random variable satisfying

$$E\left[\mathbf{1}_A X\right] = E\left[\mathbf{1}_A E[X/\mathcal{G}]\right], \quad \text{for every } A \in \mathcal{G}. \tag{4}$$

Since Z' is $\mathcal{G}$ measurable, it can be written in the form $Z' = \sum_i c_i \mathbf{1}_{B_i}$. By letting $A = B_j$ in (4), we get

$$E[\mathbf{1}_{A_j} X] = c_j E[\mathbf{1}_{A_j}],$$

and hence $Z' = Z$. Thus (3) characterizes $E[X/\mathcal{G}]$ as claimed.

But what is the meaning of (3) ? Let $I\!P_{\mathcal{G}}$ be the measure $I\!P$ restricted to $\mathcal{G}$. Define a second, possibly signed measure on $\mathcal{G}$ by

$$I\!M(A) = E[\mathbf{1}_A X], \quad A \in \mathcal{G}.$$

Then (3) says precisely that

$$I\!M(A) = \int_A E[X/\mathcal{G}] dI\!P, \qquad \text{for every } A \in \mathcal{G}.$$

In other words, $E[X/\mathcal{G}]$ is precisely the Radon-Nikodym derivative of $I\!M$ with respect to $I\!P_{\mathcal{G}}$. This insight tells us how to define the conditional expectation quite generally.

Conditional expectation: abstract definition. Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, I\!P)$. Suppose $\mathcal{G} \subset \mathcal{F}$ is a sub–σ-algebra of $\mathcal{F}$. Let $I\!P_{\mathcal{G}}$ denote $I\!P$ restricted to $\mathcal{G}$. Now define a bounded, signed measure $I\!M_X$ on $(\Omega, \mathcal{G})$ by

$$I\!M_X(A) = E[\mathbf{1}_A X] = \int X \mathbf{1}_A \, dI\!P, \qquad A \in \mathcal{G}.$$

Clearly, $I\!M_X$ is absolutely continuous with respect to $I\!P_{\mathcal{G}}$—that is, $I\!M_X(A) = 0$ whenever $A \in \mathcal{G}$ and $I\!P(A) = 0$. The Radon–Nikodym theorem then implies that there is a unique (up to $I\!P_{\mathcal{G}}$-a.s. equivalence) random variable called $dI\!M_X/dI\!P_{\mathcal{G}}$, which is the Radon–Nikodym derivative of $I\!M$ with respect to $I\!P_{\mathcal{G}}$, satisfying

a) $\dfrac{dI\!M_X}{dI\!P_{\mathcal{G}}}$ is $\mathcal{G}$–measurable; and,

b) For every $A \in \mathcal{G}$,

$$E\left[\mathbf{1}_A X\right] = I\!M_X(A) = \int_\Omega \mathbf{1}_A \frac{dI\!M_X}{dI\!P_{\mathcal{G}}} \, dI\!P_{\mathcal{G}} = E\left[\mathbf{1}_A \frac{dI\!M_X}{dI\!P_{\mathcal{G}}}\right]. \tag{5}$$

Definition: If X is an integrable random variable and $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, we use $E\left[X/\mathcal{G}\right]$ to denote $\dfrac{dI\!M_X}{dI\!P_{\mathcal{G}}}$. We call $E\left[X/\mathcal{G}\right]$ the conditional expectation of X given $\mathcal{G}$.

If $A \in \mathcal{F}$, the conditional probability of A given $\mathcal{G}$ is defined to be

$$I\!P(A/\mathcal{G}) = E\left[\mathbf{1}_A/\mathcal{G}\right].$$

Remarks 1. Our definition is slightly ambiguous because the Radon-Nikodym derivative is defined only up to $I\!P_{\mathcal{G}}$–a.s. equivalence. $E[X/\mathcal{G}]$ stands for any one of these $I\!P_{\mathcal{G}}$–a.s. equivalent random variables which serve as the Radon-Nikodym derivative. When working with $E[X/\mathcal{G}]$ in a calculation, we have in mind that it represents one fixed, but arbitrarily chosen random variable from the equivalence class. Sometimes, when we have explicitly constructed a candidate Z for $E[X/\mathcal{G}]$, we call Z a version of $E[X/\mathcal{G}]$. This is a little clumsy, but avoids having to treat random variables as equivalence classes; see, the relevant remarks on page 3 of chapter 2.

2. In probability theory, one generally suppresses the explicit dependence of random variables on $\omega \in \Omega$. However, clarity of exposition sometimes requires showing this dependence in the notation, and when we need to do this for conditional expectations we shall write $E[X/\mathcal{G}](\omega)$.

The following lemma characterizing conditional expectations is just a restatement of the definition and uniqueness of the Radon-Nikodym derivative. It is used often in identifying conditional expectations.

LEMMA A.1 Let X be an integrable r.v. and let $\mathcal{G} \subset \mathcal{F}$ be a sub–σ-algebra. If Z is a random variable satisfying

(i) Z is $\mathcal{G}$ measurable, and;
(ii) $E[\mathbf{1}_A X] = E[\mathbf{1}_A Z]$ for all $A \in \mathcal{G}$, then

$$Z = E[X/\mathcal{G}].$$

The definition of conditional expectation has the following simple consequence. Taking $A = \Omega$ in condition (ii) of Lemma A.1,

$$E[X] = E\left[E[X/\mathcal{G}]\right]. \tag{6}$$

For conditional probabilities, this becomes

$$I\!P(A) = E\left[I\!P(A/\mathcal{G})\right], \tag{7}$$

which is sometimes called the rule of total probabilities. These identities are often used, because it is often helpful to analyze a random variable by first conditioning on a σ-algebra.

Conditional expectation: further elementary examples. We have already shown that for a σ-algebra $\mathcal{G}$ generated by a finite disjoint partition $B_1, \ldots, B_n$ of Ω,

$$E[X/\mathcal{G}](\omega) = \sum_{i=1}^{n} E[X/B_i]\mathbf{1}_{B_i}(\omega)$$

For the particular case in which $X = \mathbf{1}_A$ for some measurable A not in $\mathcal{G}$, this gives

$$I\!P(A/\mathcal{G}) = \sum_1^n I\!P(A/B_i)\mathbf{1}_{B_i}$$

By applying (7) to this identity, one finds the rule of total probability encountered in elementary probability theory:

$$I\!P(A) = \sum_1^n I\!P(A/B_i)I\!P(B_i).$$

As a further example, consider the problem of defining $E[X/Y = y]$ for two random variables X and Y. If Y is not discrete, we can not resort to definition (1), because we may have $I\!P(Y = y) = 0$ for some values y in the support of the law of Y. Elementary probability handles the definition in the case that (X, Y) is jointly continuous with probability density $f(x, y)$ as follows. Y has a density f_Y which can be computed as the marginal of f:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx.$$

Let y be such that $f_Y(y) > 0$. Then a simple formal calculation shows that

$$\lim_{h\to 0} I\!P(X \le z/Y \in (y - h, y + h)) = \int_{-\infty}^{z} \frac{f(x, y)}{f_Y(y)}\, dx,$$

Taking $\displaystyle\int_{-\infty}^{z} \frac{f(x, y)}{f_Y(y)}\, dx$ as the *definition* of $I\!P(X \le z/Y = y)$ motivates defining the conditional density of X given $Y = y$ by as

$$f_{X|Y}(x/y) := \frac{f(x, y)}{f_Y(y)},$$

Then the conditional expectation of X given $Y = y$ is,

$$E[X/Y = y] := \int x f_{X|Y}(x/y)\, dx. \tag{8}$$

Notice that this conditional expectation is a function on the range of Y. (Properly speaking, $E[X/Y = y]$ is defined only on the set of y such that $f_Y(y) > 0$. Since the probability that Y takes on values in the set $\{y\,;\, f_Y(y) = 0\}$ is zero, there is no need to extend the domain of $E[X/Y = y]$; if we choose to extend the domain, it does not matter how the extension is defined.)

We can also approach the problem of defining the conditional expectation X given Y using the abstract approach. For this we simply define

$$E[X/Y] := E[X/\sigma(Y)],$$

where $\sigma(Y)$ is the σ-algebra generated by Y, without any need to define a conditional density or any restriction on the law of (X, Y), except of course, for the integrability of X. We emphasize that in this definition $E[X/Y]$ is a random variable, not a function. Here is how the two definition of $E[X/Y]$ and $E[X/Y=y]$ are connected. Assume that (X, Y) has a joint density, and, for clarity of notation, write $c_X(y) = E[X/Y = y]$. Then

$$c_X(Y(\omega)) \text{ is a version of } E[X/Y](\omega)$$

The proof is simple, using Lemma A.1. We must check that $c_X(Y)$ is $\sigma(Y)$–measurable and that condition (ii) of Lemma A.1 holds. The $\sigma(Y)$ measurability is trivial. As for condition (ii), take an arbitrary $A \in \sigma(Y)$. Then there is a Borel set U such that $\mathbf{1}_A = \mathbf{1}_U(Y)$. Hence

$$\begin{aligned} E[c_X(Y)\mathbf{1}_A] = E[c_X(Y)\mathbf{1}_U(Y)] &= \int c_X(y)\mathbf{1}_U(y) f_Y(y)\, dy \\ &= \int \mathbf{1}_U(y) \int x \frac{f(x,y)}{f_Y(y)}\, dx\, f_Y(y)\, dy \\ &= \int\int x\mathbf{1}_U(y) f(x,y)\, dx\, dy \\ &= E[X\mathbf{1}_U(Y)] = E[X\mathbf{1}_A]. \end{aligned}$$

These examples show that the rather abstruse, general definition of conditional expectation is really the right one. In each case, the random variable created by evaluating the elementary definition of conditional expectation at a random outcome is uniquely characterized by conditions (i) and (ii) of Lemma A.1, which identify a Radon-Nikodym derivative. The general definition takes the Radon-Nikodym characterization as its starting point. The generality and flexibility of the abstract definition are essential in advanced probability and stochastic process theory. Existence of the conditional expectation in all generality is guaranteed by the Radon-Nikodym theorem; there is no need to worry about how to construct the conditional expectation on zero measure atoms of $\mathcal{G}$, or to invoke a special hypothesis like the existence of joint densities. Thus, so long as X is integrable $E[X/Y] := E[X/\sigma(Y)]$ makes sense for any random variable Y, whether (X, Y) has a joint density or not. The general definition applies also to any sub–σ-algebras $\mathcal{G}$, not just those generated by random variables. For example, without further ado, we can condition on the outcome of any family $\{Y_j\,;\, j \in J\}$ of random variables simply by defining

$$E\left[X/\{Y_j\,;\, j \in J\}\right] := E\left[X/\sigma\left(\{Y_j\,;\, j \in J\}\right)\right]. \tag{9}$$

We make one more remark on $E[X/\sigma(Y)]$. In the case when (X, Y) has a joint density, we were able to give a meaning to $E[X/Y = y]$ for a given value of Y. What about the general case, when no joint density may exist? There is a theorem that says that, if H is a $\sigma(Y)$–measurable random variable, there exists a Borel measurable function h such that $H = h(Y)$ almost-surely; obviously, h is uniquely defined up to sets of $I\!F_Y$-measure zero, where $I\!F_Y$ is the law of Y. Thus, for any pair of random variables X, Y, where X is integrable, there is a Borel function c_X such that $c_X(Y) = E[X/\sigma(Y)]$. We may not have an explicit formula like (8) for c_X, but we know it exists. It is natural to interpret $E[X/Y = y]$ as $c_X(y)$.

B. Properties of conditional expectation

We first compute the conditional expectation in some simple, special cases, as an exercise in using the definition.

Consider first a random variable X and a σ-algebra $\mathcal{G}$ that are independent of one another. Then for any $A \in \mathcal{G}$,

$$E[\mathbf{1}_A X] = E[X]E[\mathbf{1}_A] = E\big[\mathbf{1}_A E[X]\big].$$

Certainly, the random variable identically equal to $E[X]$ is $\mathcal{G}$-measurable. Thus the conditions of Lemma A.1 are satisfied by the constant $E[X]$.

Proposition B.1 If $\mathcal{G}$ and X are independent and X is integrable, then

$$E[X/\mathcal{G}] = E[X].$$

A σ-algebra $\mathcal{G}$ is called trivial if every event in $\mathcal{G}$ has either probability 0 or probability 1. If $\mathcal{G}$ is trivial, then $\mathcal{G}$ is independent from any other σ-algebra , and, in particular from any random variable, because $I\!P(A \cap B) = I\!P(A)I\!P(B)$ if A has either probability 0 or probability 1. Thus, if $\mathcal{G}$ is trivial and X is any integrable random variable, we have from Proposition B.2 that $E[X/\mathcal{G}] = E[X]$.

One of the most important properties of conditional expectation is the so-called tower property.

Proposition B.2 Let X be an integrable random variable and let $\mathcal{G} \subset \mathcal{H}$. Then

$$E\left[E[X/\mathcal{H}]/\mathcal{G}\right] = E\left[X/\mathcal{G}\right]. \tag{1}$$

Proof: Let A be an arbitrary event in $\mathcal{G}$. Then, since $A \in \mathcal{H}$ also, the definition of conditional expectation, applied first to $\mathcal{H}$ and then to $\mathcal{G}$, implies

$$E\left[\mathbf{1}_A E[X/\mathcal{H}]\right] = E[\mathbf{1}_A X] = E\left[\mathbf{1}_A E[X/\mathcal{G}]\right].$$

Thus $E[X/\mathcal{G}]$ satisfies condition (ii) of Lemma A.1 with X replaced by $E[X/\mathcal{H}]$. Since $E[X/\mathcal{G}]$ is $\mathcal{G}$-measurable, it must equal $E[E[X/\mathcal{H}]/\mathcal{G}]$. $\diamond$

It is useful to remember this rule by thinking about the case in which $\mathcal{G}$ is the σ-algebra generated by a finite partition Π of Ω, and $\mathcal{H}$ is the σ-algebra generated by a refinement of Π. Then we know from section A that $E[X/\mathcal{G}]$ is computed by averaging over the atoms of $\mathcal{G}$— here, "atom" means a set of the partition—and similarly for $\mathcal{H}$. Since an atom of $\mathcal{G}$ is a finite union of atoms of $\mathcal{H}$, the average of $E[X/\mathcal{H}]$ over an atom A of $\mathcal{G}$ equals the average of X over A.

Another simple case worthy of mention is when X is $\mathcal{G}$ measurable. Since X itself satisfies condition (ii) of Lemma A.1, we see that $E[X/\mathcal{G}] = E[X]$ for $\mathcal{G}$-measurable X. More is true:

Proposition B.3 Let X be $\mathcal{G}$-measurable and suppose Y is a second random variable satisfying $E[|Y|] < \infty$ and $E[|XY|] < \infty$. Then

$$E[XY/\mathcal{G}] = XE[Y/\mathcal{G}]. \tag{2}$$

Proof: The strategy is to first prove (2) for simple $\mathcal{G}$-measurable functions X and then to pass to the general case by taking limits. To prove (2) for simple functions X it suffices to consider the case $X = \mathbf{1}_B$ where $B \in \mathcal{G}$. Now $\mathbf{1}_B E[Y/\mathcal{G}]$ is $\mathcal{G}$-measurable. Moreover, if $A \in \mathcal{G}$, so is $A \cap B$, and so by definition of conditional expectation

$$E\left[\mathbf{1}_A \mathbf{1}_B E[Y/\mathcal{G}]\right] = E\left[\mathbf{1}_A (\mathbf{1}_B Y)\right].$$

Thus, $E[\mathbf{1}_B Y/\mathcal{G}] = \mathbf{1}_B E[Y/\mathcal{G}]$.

Now suppose X is a non-negative $\mathcal{G}$-measurable random variable and that Y also is non-negative. Let $\{X_n\}$ be an increasing sequence of simple $\mathcal{G}$ measurable functions with limit X. Then by monotone convergence, for any $A \in \mathcal{G}$

$$\begin{aligned} E[\mathbf{1}_A XY] &= \lim_{n\to\infty} E[\mathbf{1}_A X_n Y] = \lim_{n\to\infty} E\left[\mathbf{1}_A X_n E[Y/\mathcal{G}]\right] \\ &= E\left[\mathbf{1}_A X E[Y/\mathcal{G}]\right]. \end{aligned} \tag{3}$$

The monotone convergence theorem was applied to derive the last equality, using the fact that $E[Y/\mathcal{G}]$ is non-negative for non-negative Y. We prove this in the next proposition. From (3) it follows that $E[XY/\mathcal{G}] = XE[Y/\mathcal{G}]$.

We know now that (2) is true for non-negative random variables. To prove the general case, decompose X and Y into their positive and negative parts. $\diamond$

The operation of conditional expectation behaves very much like that of expectation. In particular all the usual rules for the interchange of limit and expectation continue to hold.

Proposition B.4 (In this proposition, all random variables are integrable.)

(a) (Positivity) If $X \geq 0$ a.s., then $E[X/\mathcal{G}] \geq 0$ a.s., as well.

(b) (Monotone convergence) If $\{X_n\}$ is an almost surely increasing sequence of random variables, meaning $X_{n+1} \geq X_n$, a.s. for each n, and $X = \text{a.s.-}\lim X_n$ is integrable, then $\lim_{n\to\infty} E[X_n/\mathcal{G}] = E[X/\mathcal{G}]$, a.s.

(c) (Fatou's lemma) If $\{X_n\}$ is a sequence of non-negative random variables, then $E[\liminf X_n/\mathcal{G}] \leq \liminf E[X_n/\mathcal{G}]$.

(d) (Dominated convergence) Suppose that there is an integrable random variable Y such that $|X_n| \leq Y$ a.s., for all n. Suppose that $X = \lim X_n$ a.s. Then $\lim E[X_n/\mathcal{G}] = E[X/\mathcal{G}]$ a.s.

(e) (Jensen's Inequality) If f is a convex function and $f(X)$ is integrable, then $f(E[X/\mathcal{G}]) \leq E[f(X)/\mathcal{G}]$.

Proof (sketch): To prove (a), let X be non-negative and set $B = \{E[X/\mathcal{G}] < 0\}$. Then $0 \leq E[\mathbf{1}_B X]$ and, since B is $\mathcal{G}$-measurable,

$$0 \leq E[\mathbf{1}_B X] = E\left[\mathbf{1}_B E[X/\mathcal{G}]\right].$$

It follows that $I\!P(B) = 0$.

Going on to (b), we see from (a) that $E[X_{n+1}/\mathcal{G}] \geq E[X_n/\mathcal{G}]$ a.s. By modifying eac $E[X_n/\mathcal{G}]$ on $\mathcal{G}$-measurable sets of measure zero, we can produce a sequence of versions $\{E[X_n/\mathcal{G}]\}$ that is increasing eveywhere, and hence $Z = \lim E[X_n/\mathcal{G}]$ defines an (extended) random variable Z everywhere. Then by applying the dominated convergence theorem twice,

$$\begin{aligned} E[\mathbf{1}_A X] &= \lim E[\mathbf{1}_A X_n] = \lim E\left[\mathbf{1}_A E[X_n/\mathcal{G}]\right] \\ &= E[\mathbf{1}_A Z], \end{aligned}$$

for every $A \in \mathcal{G}$. It follows that $Z = E[X/\mathcal{G}]$ a.s.

Fatou's lemma and the dominated convergence theorem follow from (b) by the same arguments used in proving them in integration theory.

The proof (e) also just follows the proof of Jensen's inequality in integration theory. $\diamond$

Conditional expectations are important in the theory of statistical estimation. Given two random variable X and Y, a function c_X such that $c_X(Y) = E[X/Y]$ is called a regression function for X on Y. We can think of $c_X(Y) = E[X/Y]$ as an estimate of X using the information available from observing Y. How good is this estimate? The following theorem, stated in a more general setting, says that it is optimal in the sense of minimizing the mean-square error. In the statement of the result we use the notation $L^2(\Omega, \mathcal{G}, I\!P)$ to denote the space of all square integrable, sG-measurable random variables.

Proposition B.5 Let X be a random variable with finite variance and let $\mathcal{G}$ be a sub-σ-algebra. Then

$$E\left[\left(X - E[X/\mathcal{G}]\right)^2\right] = \inf\left\{E\left[(X-Z)^2\right] \;;\; Z \in L^2(\Omega, \mathcal{G}, I\!P)\right\}.$$

In effect, conditional expecation on $\mathcal{G}$ acts as a projection operator from $L^2(\Omega, \mathcal{F}, I\!P)$ to $L^2(\Omega, \mathcal{G}, I\!P)$.

Proof: For any $W \in L^2(\Omega, \mathcal{G}, I\!P)$, $W(X - E[X/\mathcal{G}])$ is integrable by the Cauchy-Schwartz inequality and the fact that $E[(E[X/\mathcal{G}])^2] \leq E[E[X^2/\mathcal{G}]] = E[X^2] < \infty$, which follows from using the conditional Jensen inequality. Taking conditional expectations,

$$E\left[W(X - E[X/\mathcal{G}])/\mathcal{G}\right] = WE\left[(X - E[X/\mathcal{G}])/\mathcal{G}\right] = 0.$$

Thus, for any $Z \in L^2(\Omega, \mathcal{G}, I\!P)$,

$$\begin{aligned} E\left[(X - Z)^2\right] &= E\left[(X - E[X/\mathcal{G}])^2\right] + 2E\left[(X - E[X/\mathcal{G}])(E[X/\mathcal{G}] - Z)\right] \\ &\qquad + E\left[(E[X/\mathcal{G}] - Z)^2\right] \\ &\geq E\left[(X - E[X/\mathcal{G}])^2\right] \qquad \diamond \end{aligned}$$