# Understanding Conditional Expectation

Mohammad Dorgham
mohammad.dorgham@studium.uni-hamburg.de
Supervisor
Prof. Dr. Holger Drees

10.10.2017

**Abstract**

Conditional expectation notion is of great importance in probability theory and statistics. However the elementary definition of conditioning suffers from ambiguity and has led to some well-known paradoxes such as the Borel-Kolmogorov paradox. One approach that is developed to avoid these problems and to establish a rigorous treatment of conditioning is the one by Kolmogorov. Kolmogorov's approach had become a mainstream in probability theory, yet it is not easily digested to students at first encounter, and most explanations are brief and lacking a proper motivation to the concept. In this exposé we explain Kolmogorov's approach to handle conditioning with respect to sigma algebras. We will try to simplify the concepts in a way that combines both intuition and rigor. The existence of conditional expectation is proved by Radon-Nikodym theorem. Also the proofs of some of the important properties of conditional expectation are given and simplified.

## 1 Motivation

Suppose $X$ is a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$, and $B$ is a measurable subset of $\Omega$. The elementary definition of the conditional expectation of $X$ given $B$ is given by $E[X \mid B] = E(XI_B)/P(B)$, where $I_B$ is the indicator function of $B$, and $P(B) > 0$. The problem with that definition is that the conditional expectation is not defined when $P(B) = 0$. But it is often desirable to condition on events of probability zero, such as conditioning on a specific value of a continuous random variable. A natural approach to overcome this problem when faced with a set of probability 0 is to condition instead on an arbitrary small neighborhood and take the limit, for example instead of calculating $P(X = x)$, we calculate the probability density on a neighborhood ($|X - x| < \varepsilon$) and take the limit as $\varepsilon$ goes to 0. However, this approach stays unreliable as it is not guaranteed that the limit will always exist [Breiman p. 68][2], which makes this approach suffers from ambiguity in general cases. Another important issue that arises with that definition, is that

defining the conditional expectation with respect to a single null event might lead to unexpected results as the next example demonstrates.

**Example 1.1** (**Borel–Kolmogorov paradox**)**.** Suppose a random variable is distributed uniformly on the surface of a unit sphere. We are interested in getting the probability distribution of this random variable given the knowledge that it lies on a specific great circle. By the great circle we mean the circle resulting from the intersection of the sphere and a plane that passes through the center point of the sphere. Due to the symmetry of the sphere one's intuition might expect that the conditional distribution would be the same regardless of the choice of the great circle, i.e. that no matter which great circle we choose the random variable will always have the same uniform distribution. However it turns out that one could obtain two different distributions for two different great circles.

Before establishing the two example circles, we need to introduce the relation between the Cartesian and spherical coordinates. We assume that the center point of the sphere is located at $(0, 0, 0)$ point in the Cartesian coordinates where the $x, y, z$ axes form a right-handed coordinate system as shown in figure 1. A point on the unit sphere is represented in spherical coordinates by the latitude and longitude coordinates, where the latitude $\phi$ represents the angle between the projection of the point into the $yz$-plane and the $y$-axis, and the longitude $\lambda$ represents the angle between the projection of the point into the $xy$-plane and the $x$-axis. The latitude usually ranges over $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and the longitude usually ranges over $[-\pi, \pi)$. But for the sake of the current example we will restrict the longitude to range over $[0, \pi)$, so that the choice of a longitude corresponds to a complete meridian circle (not semicircle), and compensate by letting the latitude to range over $[-\pi, \pi)$. In standard spherical coordinates we know that the surface area element of the unit sphere is equal to $\cos\phi \; d\phi \; d\lambda$, but within the current boundaries we obtain $|\cos\phi| \; d\phi \; d\lambda$, where the absolute value is to guarantee that the total area of the sphere integrates to $4\pi$.

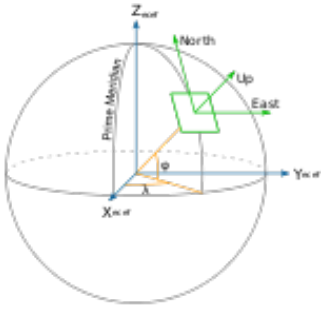

Figure 1: Geographic coordinate system. From wikipedia

We now introduce random variables $\Phi$, $\Lambda$ that measure latitude and longitude respectively. We define the joint density function of $\Phi$, $\Lambda$ such that a uniform distribution on the sphere is maintained:

$$f_{\Phi,\Lambda}(\phi, \lambda) = \frac{1}{4\pi} |\cos\phi|$$

where $4\pi$ is a normalizing factor to ensure $\int_0^\pi \int_{-\pi}^\pi f_{\Phi,\Lambda}(\phi, \lambda) \; d\phi \; d\lambda = 1$. We can then easily calculate the marginal density of each random variable from the joint

density by integration

$$f_\Phi(\phi) = \frac{1}{4}|\cos\phi|,$$
$$f_\Lambda(\lambda) = \frac{1}{\pi}\ .$$

Now consider the first great circle to be that one lying on the equator (latitude $\Phi = 0$). We would like to know the distribution of $\Lambda$ within the chosen great circle. The classical Bayes formula $P(\Lambda \in A \mid \Phi = 0) = \frac{P(\Lambda \in A, \Phi = 0)}{P(\Phi = 0)}$ won't help us since the event $\{\Phi = \phi\}$ has probability 0 for any value of $\phi$. So we will resort to the limit trick and condition instead on the event $\{\phi \le \Phi \le \phi + \delta\phi\}$:

$$\begin{aligned}
P(\Lambda \in A \mid \Phi = \phi) &= \lim_{\delta\phi \to 0} P(\Lambda \in A \mid \phi \le \Phi \le \phi + \delta\phi) \\
&= \lim_{\delta\phi \to 0} \frac{P(\Lambda \in A, \phi \le \Phi \le \phi + \delta\phi)}{P(\phi \le \Phi \le \phi + \delta\phi)} \\
&= \lim_{\delta\phi \to 0} \frac{\int_A \int_\phi^{\phi+\delta\phi} f_{\Phi,\Lambda}(u,v)dudv}{\int_\phi^{\phi+\delta\phi} f_\Phi(u)du} \\
&= \lim_{\delta\phi \to 0} \frac{\int_A \int_\phi^{\phi+\delta\phi} \frac{1}{4\pi}|\cos u|dudv}{\int_\phi^{\phi+\delta\phi} \frac{1}{4}|\cos u|du} \\
&= \frac{1}{\pi}\int_A dv \quad =:\ \mu_{[0,\pi]}(A)
\end{aligned}$$

Therefore $\mu_{[0,\pi]}$ is the conditional distribution of $\Lambda$ given $\Phi$ which is equivalent to the unconditional distribution of $\Lambda$ with the density

$$f_{\Lambda|\Phi}(\lambda \mid \phi) = f_\Lambda(\lambda) = \frac{1}{\pi}. \tag{1}$$

Consider the second great circle to be that one lying on the prime meridian (longitude $\Lambda = 0$). We would follow the same limiting approach as above. We then compute the conditional density:

$$f_{\Phi|\Lambda}(\phi \mid \lambda) = f_\Phi(\phi) = \frac{1}{4}|\cos\phi|. \tag{2}$$

The first distribution is uniform while the second is not! This means that when we choose a point randomly it is more likely to lie near the equator than the poles, but it is equally likely to lie on any longitude. To intuitively see why the choice of different great circles has led to different distributions, take first the first scenario. We took the limit over a small neighborhood around $\{\Phi = 0\}$, notice that this neighborhood consists of parallel rings of latitudes around the equator, each ring

keeps the same distance from the equator everywhere. While in the second scenario, the neighborhood around a meridian consists of the parallel longitudes forming the shape of lunes, so the points near the equator has more density than the points near the poles (see figure 2). □

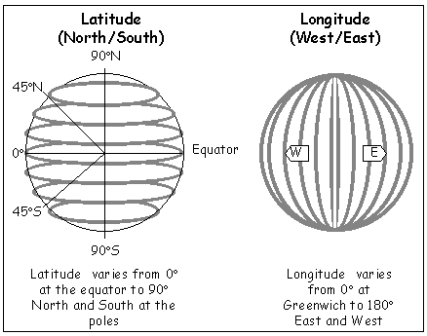


Figure 2: Illustrating parallels of latitudes and longitudes. From wikipedia

**Remark 1.2.** What seems paradoxical in the previous example is that we are describing the same null event (the points lying on a great circle) with two different parameterizations, hence we are expecting to get the same distribution in both cases. However, the extension of the elementary definition via the limit gave us a different answer when we parameterized the event differently[1], but our common sense says that the conditional probability should be invariant under different parameterizations. It turns out that considering a single null event to find the conditional probability is not enough to make the problem well posed. In Kolmogorov's words: "the concept of a conditional probability with regard to an isolated given hypothesis whose probability equals zero is inadmissible." [4][P. 51]. It is inadmissible because the uniqueness of the solution is not guaranteed [2] as the problem is ill posed or not completely specified [Rao p. 441][6]. In section 3 we will see Kolmogorov's approach to make the problem well posed by involving the knowledge of extra information while conditioning. In Kolmogorov's theory the concept of conditioning involves only measure-theoretic notions and does not depend on the parametrization of the problem.

## 2 Thinking more general

Kolmogorov's suggestion to solve the problem of conditioning is to condition with respect to a collection of events instead of an isolated event. But to understand which collection of events one should consider and why this choice does make sense,

[1]For more examples of similar problems when conditioning on null events see [5, 6, 7].

[2]Except in the case of discrete probability spaces.

we need to motivate some ideas before the formal definition is given in the next section.

Given a probability space $(\Omega, \mathcal{F}, P)$, one can intuitively view $\mathcal{F}$ as a collection that represents all possible information (outcomes) in that probability space. When an observer performs an experiment, the experiment usually yields partial information. These partial information that represent all possible outcomes of the experiment constitute a sigma algebra, call it $\mathcal{G}$, that is a sub-sigma algebra of $\mathcal{F}$. The observer is usually interested in measuring some random variable which cannot be measured directly, so he/she uses the available partial information to estimate the best prediction of the value of the random variable of interest.

To view the expectation in the context of partial information, it might be useful to think of the expectation in general from another perspective. Suppose $X$ is a random variable defined on $(\Omega, \mathcal{F}, P)$ with $E[X^2] < \infty$. Instead of thinking of the expectation $E[X]$ as just a numerical average of the values of the random variable $X$, another perspective is to think of it as the best predictor of the random variable $X$ in the sense that it minimizes $E[(X - E[X])^2]$. That is if you are not able to obtain any information about $X$ then your best prediction for $X$ would be the mean $E[X]$. If you could do an experiment that would yield some partial information about $X$, then you know that your prediction will be more representative of $X$. If another experiment could yield a wider range of information, then the prediction will become better.

An important difference to note though between regular expectation and conditional expectation is that the former is a constant numerical value, while the later is a random variable since the conditional expectation depends on the conditioning random variable and its value changes with the value the random variable takes. In other words, if $X, Y$ are random variables that map from $(\Omega, \mathcal{F})$ into $(S, \mathcal{C})$, $E[Y \mid X]$ is a function of $X$ taking on constant value for each value of $X$, therefore $E[Y \mid X]$ is measurable with respect to $\sigma(X)$ where $\sigma(X)$ consists of the sets $\{\omega : X(\omega) \in C\}$ for $C \in \mathcal{C}$.

Back to our discussion about expectations as best predictors. It might be of benefit to consider defining the conditional expectation with respect to the whole sigma algebra $\mathcal{G}$ corresponding to an experiment instead of just one event in $\mathcal{G}$. In fact there is a good reason for considering this choice. If we define the conditional expectation of an integrable random variable $X$ that is $\mathcal{F}$-measurable, with respect to the sigma algebra $\mathcal{G}$, one would expect that $E[X \mid\mid \mathcal{G}]$ will be the best predictor of $X$ amongst all $\mathcal{G}$-measurable random variables (provided $E[X^2] < \infty$). That being said, we would like to know whether such an object exists, and to what extent we can guarantee its uniqueness.

## 3 Existence of conditional expectation

Let $X$ be an integrable random variable on $(\Omega, \mathcal{F}, P)$, and let $\mathcal{G}$ be a $\sigma$-algebra in $\mathcal{F}$. An essential and intuitive requirement that should be satisfied by conditional expectation is that $E[E[X \mid\mid \mathcal{G}]] = E[X]$, but this is a weak condition since there could exist infinitely many distributions that have the same mean. So to strengthen this condition we should require that $E[E[X \mid\mid \mathcal{G}]I_G] = E[XI_G]$ for every subset $G$ of $\mathcal{G}$. We therefore define the conditional expectation as follows:

**Definition 3.1.** *The conditional expectation $E[X \mid\mid \mathcal{G}]$ of an integrable random variable $X$ is any random variable satisfying the following properties:*

*(i) $E[X \mid\mid \mathcal{G}]$ is measurable with respect to $\mathcal{G}$,*

*(ii) $E[X \mid\mid \mathcal{G}]$ satisfies the functional requirement*

$$\int_G E[X \mid\mid \mathcal{G}]dP = \int_G XdP, \qquad for\ all\ G \in \mathcal{G}. \tag{3}$$

To prove the existence of the conditional expectation, suppose first that $X$ is nonnegative. Define a measure $\nu$ on $\mathcal{G}$ by

$$\nu(G) = E[XI_G] = \int_G XdP.$$

Then $\nu$ is finite since $X$ is integrable. It is clear that $P(G) = 0$ implies $\nu(G) = 0$ (i.e. $\nu$ is dominated by $P_{|\mathcal{G}}$. i.e. the restriction of $P$ to $\mathcal{G}$), therefore Radon-Nikodym theorem (see Appendix A) implies that there exists a function $f$ measurable with respect to $\mathcal{G}$ and integrable such that $\nu(G) = \int_G fdP$ for all $G \in \mathcal{G}$, and this $f$ is unique up to a set of measure 0 (with respect to $P$ restricted to $\mathcal{G}$). We denote this function $f$ by $E[X \mid\mid \mathcal{G}]$. Finally if $X$ is not nonnegative, $E[X^+ \mid\mid \mathcal{G}] - E[X^- \mid\mid \mathcal{G}]$ achieves the same result. □

Since the Radon-Nikodym theorem implies only almost sure uniqueness, $E[X \mid\mid \mathcal{G}]$ refers to a family of random variables, a specific such random variable is called a version of the conditional expectation. Obviously any two versions are equal with probability 1.

We note that Definition 3.1 is not a constructive definition, in the sense that it does not give us a ready-to-use formula to directly compute the conditional expectation. Rather it is an abstract definition that gives the required properties that a conditional expectation must satisfy. We will see in the next section some examples

for developing the concrete conditional expectations such that they agree with the abstract definition.

The next theorem rigorously formulates and confirms the intuition we have introduced in the previous section about the conditional expectations as best predictors.

**Theorem 3.2.** *If $E[X^2]$ is finite, then the conditional expectation $E[X \mid\mid \mathcal{G}]$ is the orthogonal projection of $X$ onto $\mathcal{L}^2(\Omega, \mathcal{G}, P)$, therefore $E[X \mid\mid \mathcal{G}]$ is the least squares best predictor of $X$ between all $\mathcal{G}$-measurable functions, that is, $E[X \mid\mid \mathcal{G}]$ minimizes $E[(X - E[X \mid\mid \mathcal{G}])^2]$.* □

**Remark 3.3.** Before ending this section we would like to return to the Borel-Kolmogorov paradox to see how the new approach has solved the contradiction. In Kolmogorov's approach one does not lose sight of the relative sigma algebra containing the conditioning event. By keeping that in mind when you look at the paradox, you realize that the conditioning event $\{\Phi = 0\}$ belongs to the sigma algebra $\sigma(\Phi)$, which is independent of the sigma algebra $\sigma(\Lambda)$ containing the event $\{\Lambda = 0\}$. That means they are not the same thing as our intuition had led us, the seemingly contradiction disappears now. Kolmogorov's framework gives a precise meaning to the concept of conditioning that is enough to remove any ambiguity and make problems well posed. In Kolmogorov's theory different parameterizations do not necessarily give the same answer as they are not necessarily equivalent under the given definition.

# 4 Examples

In the following examples assume that $X$ is integrable and $\mathcal{F}$-measurable.

**Example 4.1.** If $\mathcal{G}$ is the trivial sigma algebra $\{\emptyset, \Omega\}$, that means we have no specific information about $X$, so one would expect $E[X \mid\mid \{\emptyset, \Omega\}] = E[X]$ *a.s*, and in fact that is true since the constant random variable $E[X]$ is satisfying the two properties in Definition 3.1, it is $\mathcal{G}$-measurable and integrable, and is satisfying the functional requirement. □

**Example 4.2.** If $\mathcal{G}$ is equal to $\mathcal{F}$ that means we have all possible information, so if you knew which sets in $\mathcal{F}$ contain $\omega$ you can know for sure what is the value of $\omega$. So one would expect that $E[X \mid\mid \mathcal{F}] = X$ *a.s*. And again $X$ clearly satisfies the required properties. (Also if $\mathcal{G} = \sigma(X)$ the same argument applies). □

One can describe finding the correct value of the conditional expectation as a guess-and-verify game [3], you guess the correct value depending on the settings you have in the problem, and then verify whether your guess matches the definition or not.

**Remark 4.3.** Note that the same concepts (of conditioning w.r.t sigma algebras) carry on when we want to condition with respect to a random variable. So for given random variables $X, Y$ that map from $(\Omega, \mathcal{F})$ into $(S, \mathcal{C})$, the conditional expectation of $Y$ given $X$ is defined as $E[Y \mid\mid \sigma(X)]$ and is denoted $E[Y \mid\mid X]$ for short where $\sigma(X)$ consists of the sets $\{\omega : X(\omega) \in C\}$ for $C \in \mathcal{C}$.

The following example recalls the density formula of the conditional expectation, and illustrates that the new definition agrees with the traditional usage.

**Example 4.4.** Let $X, Y$ be random variables mapping from $(\Omega, \mathcal{F})$ into $(S, \mathcal{C})$, whose joint distribution has the density function $f_{X,Y}$. We define $f_{Y|X}$ as follows

$$f_{Y|X}(y|x) := \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)} & \text{if } f_X(x) \neq 0 \\ 0 & \text{if } f_X(x) = 0 \end{cases}$$

where $f_X(x) = \int_R f_{X,Y}(x, y)dy$.

We can write $E[Y \mid\mid X]$ as a function $g(X)$ of $X$ for some $\mathcal{C}$-measurable function $g$ where $g$ maps from $(S, \mathcal{C})$ into $(\mathbb{R}, \mathcal{B})$. We set $E[Y \mid\mid X = x] := g(x)$, and we call it the factorized conditional expectation. We define $g(x)$ as follows

$$g(x) := \frac{\int_R y f_{X,Y}(x, y)dy}{\int_R f_{X,Y}(x, y)dy} \tag{4}$$

Now we want to prove that the right hand side of (4) is a version of the factorized conditional expectation.

We note that $g(.) \circ X$ maps from $(\Omega, \sigma(X))$ into $(\mathbb{R}, \mathcal{B})$, hence $g$ is $\sigma(X)$-measurable. Next, to prove the functional requirement we note that if $G \in \sigma(X)$ then $G = \{X^{-1}(g^{-1}(B))\}$ for some $B \in \mathcal{B}$. Also we have

$$\int (g(.).I_B) \circ X dP = \int_B g(x)P^X(dx)$$

where $P^X(B) = P(X \in B)$.

We need to show that

$$L := \int_B g(x)P^X(dx) = \int_{\{X \in B\}} Y dP =: R, \qquad B \in \mathcal{B}$$

We begin with $L$ :

$$\begin{aligned} L &= \int \frac{\int y f_{X,Y}(x, y)dy}{\int f_{X,Y}(x, y)dy} \, I_B(x) f_X(x)dx \\ &= \int \int y I_B(x) f_{X,Y}(x, y)dydx \end{aligned} \tag{5}$$

Then we proceed with $R$ :

$$\begin{aligned} R &= \int y I_B(x) P^{X,Y}(d(x,y)) \\ &= \int\int y I_B(x) f_{X,Y}(x,y) dxdy \end{aligned} \tag{6}$$

The equality of (5) and (6) follows from Fubini's theorem (see Appendix B). □

# 5 Properties of conditional expectation

The next theorem shows that the conditional expectation shares some properties with the unconditional one, it keeps on the same properties such as linearity, monotonicity, convergence, etc.

**Theorem 5.1.** *Suppose $X, Y, X_n$ are integrable. The following equalities and inequalities hold with probability 1:*

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

*(ii) (Linearity) $E[aX + bY \mid\mid \mathcal{G}] = aE[X \mid\mid \mathcal{G}] + bE[Y \mid\mid \mathcal{G}]$, for constants $a$ and $b$.*

*(iii) (Monotonicity) If $X \leq Y$ with probability 1, then $E[X \mid\mid \mathcal{G}] \leq E[Y \mid\mid \mathcal{G}]$.*

*(iv) $|E[X \mid\mid \mathcal{G}]| \leq E[|X| \mid\mid \mathcal{G}]$.*

*(v) (Dominated convergence) If $\lim_n X_n = X$ with probability 1, $|X_n| \leq Y$, $Y$ is integrable, then $\lim_n E[X_n \mid\mid \mathcal{G}] = E[X \mid\mid \mathcal{G}]$ with probability 1.*

*Proof.*

(i) This is immediate from the definition.

(ii) The right-hand side $aE[X \mid\mid \mathcal{G}] + bE[Y \mid\mid \mathcal{G}]$ is a linear combination of integrable and $\mathcal{G}$-measurable random variables, therefore it is integrable and $\mathcal{G}$-measurable. Also we have

$$\begin{aligned} \int_G aE[X \mid\mid \mathcal{G}] + bE[Y \mid\mid \mathcal{G}] dP &= a\int_G E[X \mid\mid \mathcal{G}] dP + b\int_G E[Y \mid\mid \mathcal{G}] dP \\ &= a\int_G X dP + b\int_G Y dP \\ &= \int_G (aX + bY) dP \end{aligned}$$

for all $G \in \mathcal{G}$.

(iii) If $X \leq Y$ $a.s$, then $\int_G X dP \leq \int_G Y dP$ for every $G \in \mathcal{G}$ by the monotonicity of the integral. Then by the linearity and Definition 3.1 we get

$$\int_G (E[Y \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}])dP = \int_G (Y - X)dP \geq 0, \quad G \in \mathcal{G} \tag{7}$$

Now if we have a set $G \in \mathcal{G}$ with $P(G) > 0$ such that $E[Y \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}] < 0$ on $G$ then the integral of $(E[Y \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}])$ over $G$ will be negative, which contradicts (7). Therefore $(E[Y \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}])$ must be nonnegative a.s.

(iv) The inequality follows from (Monotonicity) since we have $-|X| \leq X \leq |X|$ and thus $-E[|X| \mid\mid \mathcal{G}] \leq E[X \mid\mid \mathcal{G}] \leq E[|X| \mid\mid \mathcal{G}]$, hence $|E[X \mid\mid \mathcal{G}]| \leq E[|X| \mid\mid \mathcal{G}]$.

(v) Basically we want to show that $|E[X_n \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}]| \to 0$ with probability 1. We can do that by bounding $|E[X_n \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}]|$ by a term that converges to 0.
Let $Z_n := \sup_{k \geq n} |X_k - X|$ for $n \in N$, then $(Z_n)$ is nonnegative and $(Z_n) \downarrow 0$ with probability 1. By the definition of $(Z_n)$ we have

$$|X_n - X| \leq Z_n.$$

By (Monotonicity) we have

$$E[|X_n - X| \mid\mid \mathcal{G}] \leq E[Z_n \mid\mid \mathcal{G}].$$

But by (iv) we have

$$|E[X_n - X \mid\mid \mathcal{G}]| \leq E[|X_n - X| \mid\mid \mathcal{G}].$$

Then by (Linearity) we get

$$|E[X_n \mid\mid \mathcal{G}] - E[X \mid\mid \mathcal{G}]| \leq E[Z_n \mid\mid \mathcal{G}].$$

If we can prove $E[Z_n \mid\mid \mathcal{G}] \downarrow 0$ then we are done. Now since $(Z_n)$ is nonincreasing, by (Monotonicity) again $E[Z_n \mid\mid \mathcal{G}]$ is nonincreasing. Let $Z$ be the limit of the sequence $E[Z_n \mid\mid \mathcal{G}]$, we want to prove that $Z = 0$.

Because $Z$ is nonnegative, proving $E[Z] = 0$ is equivalent to proving $Z = 0$. And it is actually easier to obtain a statement about $E[Z]$, since we have $0 \leq Z \leq E[Z_n \mid\mid \mathcal{G}]$ which implies $Z$ integrable then by Definition 3.1 we obtain

$$\int E[Z \mid\mid \mathcal{G}]dP = \int Z dP = E[Z]. \tag{8}$$

Similarly we have $0 \leq Z_n \leq 2Y$, which implies that $Z_n$ integrable. And by Definition 3.1

$$\int E[Z_n \mid\mid \mathcal{G}]dP = \int Z_n dP = E[Z_n]. \tag{9}$$

From $Z \leq E[Z_n \mid\mid \mathcal{G}]$ we get

$$\int Z dP \leq \int E[Z_n \mid\mid \mathcal{G}] dP. \tag{10}$$

So from (8), (9) and (10) we have

$$E[Z] \leq E[Z_n].$$

But we have $Z_n \to 0$, and $Z_n \leq 2Y$, so by dominated convergence theorem $E[Z_n] \to 0$, and the claim follows.

□

As seen in Theorem 5.1(i), $E[X \mid\mid \mathcal{G}] = X$ if $X$ is $\mathcal{G}$-measurable. The next theorem is a generalization of this, it is commonly referred to as "taking out what's known".

**Theorem 5.2.** *If $X$ is $\mathcal{G}$-measurable, and $Y$ and $XY$ are integrable, then we have*

$$E[XY \mid\mid \mathcal{G}] = XE[Y \mid\mid \mathcal{G}] \quad a.s \tag{11}$$

*Proof.* We will proceed by showing that the right-hand side is a version of the left-hand side, using the common proof technique by beginning with the case when $X$ is the indicator function, then a simple function and finally a general $\mathcal{G}$-measurable function.

If $X = I_{G_0}$ for $G_0 \in \mathcal{G}$, then $I_{G_0} E[Y \mid\mid \mathcal{G}]$ clearly is $\mathcal{G}$-measurable. Then, it suffices to show that $\int_G I_{G_0} E[Y \mid\mid \mathcal{G}] dP = \int_G I_{G_0} Y dP$. But this is equivalent to $\int_{G \cap G_0} E[Y \mid\mid \mathcal{G}] dP = \int_{G \cap G_0} Y dP$ which holds by Definition 3.1. So (11) holds for indicator functions of sets in $\mathcal{G}$.

Next if $X$ is a simple function (i.e. $X = \sum_i x_i I_{G_i}$ for disjoint $(G_i)_{i \in I}$ in $\mathcal{G}$ where $(G_i)_{i \in I}$ is a finite partition of $\Omega$), then by linearity (Theorem 5.1(ii)) and by the previous result we obtain

$$\begin{aligned} E[(\sum_i x_i I_{G_i}) Y \mid\mid \mathcal{G}] &= \sum_i x_i E[I_{G_i} Y \mid\mid \mathcal{G}] \\ &= \sum_i x_i I_{G_i} E[Y \mid\mid \mathcal{G}] = XE[Y \mid\mid \mathcal{G}]. \end{aligned}$$

Next if $X$ is a general $\mathcal{G}$-measurable function, then $X$ is the limit of a sequence of simple functions $X_n$ where $|X_n| \leq |X|$, and hence we have $\lim_n X_n E[Y \mid\mid \mathcal{G}] = XE[Y \mid\mid \mathcal{G}]$. Note also that $\lim_n E[X_n Y \mid\mid \mathcal{G}] = E[XY \mid\mid \mathcal{G}]$ by the dominated convergence (Theorem 5.1(v)) since we have $|X_n Y| \leq |XY|$ and $|XY|$ is integrable. But we have $E[X_n Y \mid\mid \mathcal{G}] = X_n E[Y \mid\mid \mathcal{G}]$ by the previous result, therefore $E[XY \mid\mid \mathcal{G}] = XE[Y \mid\mid \mathcal{G}]$ $a.s$ in general. □

One might want to consider taking conditional expectation iteratively over different sigma algebras, a special case of interest is when one sigma algebra is subset of another, $\mathcal{G}_1 \subset \mathcal{G}_2$ say. It turns out that taking the conditional expectation with respect to the larger $\sigma$-algebra $\mathcal{G}_2$ first and then taking it with respect to the smaller one $\mathcal{G}_1$ is equivalent to taking it with respect to $\mathcal{G}_1$ from the beginning. To see why, remember that the conditional expectation is a projection operator (Theorem 3.2) that projects a random variable $X$ onto the space $\mathcal{L}^2(\Omega, \mathcal{G}, P)$. This projection may lose some precision depending on the resolution of $\mathcal{G}$. So it makes sense that the coarser $\sigma$-algebra defines the quality of the final projection. Also the same effect happens if we went the other way (i.e. taking conditional expectation w.r.t $\mathcal{G}_1$ first, then $\mathcal{G}_2$ second), it is also equivalent to taking conditional expectation with respect to $\mathcal{G}_1$ from the beginning. It may seem counter intuitive at first that one does not gain any advantage from the the finer $\sigma$-algebra that contains more information, but if you think in it from the projection viewpoint it will become very intuitive. That is because when you project $X$ onto the space defined by the coarser $\sigma$-algebra, and then take that projection (that had lost some precision) and project it onto the space defined by the finer $\sigma$-algebra, you cannot get a better precision anymore. So the conclusion is, the smaller sigma algebra always determines the iterative conditional expectation. Obviously, if $\mathcal{G}_1 = \mathcal{G}_2$, then in this case taking the conditional expectation with respect to the same $\sigma$-algebra multiple times is equivalent to doing it just once.

**Theorem 5.3.** *If $X$ is integrable and the $\sigma$-algebras $\mathcal{G}_1, \mathcal{G}_2$ are satisfying $\mathcal{G}_1 \subset \mathcal{G}_2$, then*

*(i)* $E[E[X \mid\mid \mathcal{G}_2] \mid\mid \mathcal{G}_1] = E[X \mid\mid \mathcal{G}_1]$

*(ii)* $E[E[X \mid\mid \mathcal{G}_1] \mid\mid \mathcal{G}_2] = E[X \mid\mid \mathcal{G}_1]$

*with probability 1.*

*Proof.*

(i) The left-hand side is $\mathcal{G}_1$-measurable. Hence it suffices to prove the functional requirement. By integrating the left-hand side over $G \in \mathcal{G}_1$ and resorting to Definition 3.1 we get

$$\int_G E[E[X \mid\mid \mathcal{G}_2] \mid\mid \mathcal{G}_1] dP = \int_G E[X \mid\mid \mathcal{G}_2] dP, \quad \forall\ G \in \mathcal{G}_1.$$

Note that by the definition also we have

$$\int_G E[X \mid\mid \mathcal{G}_2] dP = \int_G X dP, \quad \forall\ G \in \mathcal{G}_2.$$

But if $G \in \mathcal{G}_1$ then $G \in \mathcal{G}_2$ as well, therefore

$$\int_G E[X \mid\mid \mathcal{G}_2]dP = \int_G X dP, \quad \forall\ G \in \mathcal{G}_1. \tag{12}$$

Then the claim follows by (12) and the definition applied to the right-hand side of (i).

(ii) $E[X \mid\mid \mathcal{G}_1]$ is $\mathcal{G}_1$-measurable, and since $\mathcal{G}_1 \subset \mathcal{G}_2$, then $E[X \mid\mid \mathcal{G}_1]$ is $\mathcal{G}_2$-measurable. We therefore apply Theorem 5.1(i) to $E[X \mid\mid \mathcal{G}_1]$ and the claim follows.

□

**Example 5.4.** Consider the following two special cases of the previous theorem: First, if $\mathcal{G}_2 = \mathcal{F}$, then $E[X \mid\mid \mathcal{F}] = X$ and hence if one then condition w.r.t a smaller $\sigma$-agebra $\mathcal{G}$ then clearly $E[E[X \mid\mid \mathcal{F}] \mid\mid \mathcal{G}] = E[X \mid\mid \mathcal{G}]$. Second, if $\mathcal{G}_1 = \{\emptyset, \Omega\}$, then $E[E[X \mid\mid \mathcal{G}_2] \mid\mid \mathcal{G}_1] = E[E[X \mid\mid \mathcal{G}_2]] = E[X]$. □

## References


[1] Billingsley, Patrick. Probability and measure. John Wiley & Sons, 2008.

[2] Breiman, Leo. Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (1992).

[3] Durrett, Rick. Probability: theory and examples. Cambridge university press, 2010.

[4] Kolmogorov, Andreĭ Nikolaevich. "Foundations of the Theory of Probability." (1950).

[5] Proschan, Michael A., and Brett Presnell. "Expect the unexpected from conditional expectation." The American Statistician 52.3 (1998): 248-252.

[6] Rao, M. M. "Paradoxes in conditional probability." Journal of Multivariate Analysis 27.2 (1988): 434-446.

[7] Rescorla, Michael. "Some epistemological ramifications of the Borel–Kolmogorov paradox." Synthese 192.3 (2015): 735-767.

[8] Rosenthal, Jeffrey S. A first look at rigorous probability theory. World Scientific Publishing Co Inc, 2006.

[9] Whittle, Peter. Probability via expectation. Springer Science & Business Media, 2012.

[10] Williams, David. Probability with martingales. Cambridge university press, 1991.

# Appendices

## A Radon-Nikodym theorem

If $\mu$ and $\nu$ are $\sigma$-finite measures on $\mathcal{F}$ such that $\nu$ is dominated by $\mu$ (i.e. $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathcal{F}$), then there exists a nonnegative function $f$ such that $\nu(A) = \int_A f d\mu$ for every $A$ in $\mathcal{F}$. Moreover, if $g$ is another function satisfying $\nu(A) = \int_A g d\mu$, then $f = g$ $\mu$-almost everywhere.

## B Fubini's theorem

Let $(X, \mathcal{B}, \mu)$ and $(Y, \mathcal{C}, \nu)$ be $\sigma$-finite measure spaces, and let $\mathcal{B} \otimes \mathcal{C}$ be the product $\sigma$-algebra for the product space $X \times Y$. If a function $f$ is measurable w.r.t $\mathcal{B} \otimes \mathcal{C}$ and **integrable**, then

$$\int_X \left( \int_Y f(x,y) d\nu(y) \right) d\mu(x) = \int_Y \left( \int_X f(x,y) d\mu(x) \right) d\nu(y) = \int_{X \times Y} f d(\mu \times \nu)$$

where the partial integrals $\int_Y f(x,y) d\nu(y)$ and $\int_X f(x,y) d\mu(x)$ are defined for $\nu$-almost all $y$, and $\mu$-almost all $x$ respectively.