Conditional Expectations without tears

by Philipp Düren

Universität Augsburg

Abstract

Following [1], we give a review over the standard theory of expectation values with a decisive focus on an intuitive view on this topic.

1 Revision of basic theory: Conditioning on regular events and σ-Algebras of simplest type

We assume that the reader is already familiar with basic conditional probability theory (e.g. "Given the information that the sum of two dice is 9, what is the probability for the first dice to show a 5?") and that he has an understanding of the following topics:

Theorem 1. *Law of total probability*

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $(B_i)_{i\in I}$ an at-most countably collection of disjoint sets with $\mathbb{P}\big(\biguplus_{i\in I} B_i\big) = 1$. Then for every event $A \in \mathfrak{A}$

$$\mathbb{P}(A) = \sum_{i\in I} \mathbb{P}(A|B_i)\cdot\mathbb{P}(B_i)$$

Theorem 2. *Bayes' theorem*

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space and $(B_i)_{i\in I}$ an at-most countably collection of disjoint sets with $\mathbb{P}\big(\biguplus_{i\in I} B_i\big) = 1$. Then for every event $A \in \mathfrak{A}$ having probability $\mathbb{P}(A) > 0$ and every $k \in I$

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\cdot\mathbb{P}(B_k)}{\sum_{i\in I}\mathbb{P}(A|B_i)\cdot\mathbb{P}(B_i)}$$

Definition 3. *Conditional expectation of random variables on regular events*

Let $X \in L^1(\mathbb{P})$ (i.e. X has a finite "ordinary" expectation) and $A \in \mathfrak{A}$ be an event with probability $\mathbb{P}(A) > 0$. Then we define

$$\mathbb{E}[X|A] \equiv \int X(\omega)\mathbb{P}(\mathrm{d}\omega|A) = \frac{\mathbb{E}[1_A X]}{\mathbb{P}(A)} = \frac{\mathbb{E}[1_A X]}{\mathbb{E}[1_A]} \tag{1}$$

For $A \in \mathfrak{A}$ with probability $\mathbb{P}(A) = 0$ we set $\mathbb{E}[X|A] = 0$.

According to the last term we can interpret the conditional expectation as the center of mass of X on A, just like the common expectation $\mathbb{E}[X]$ can be thought of as the center of "probability mass" on the whole probability space.

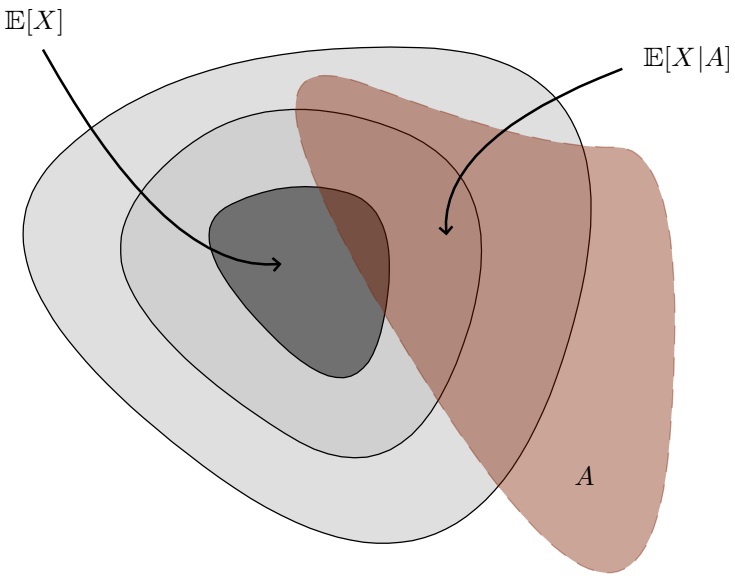

Figure 1. A visualization of conditional expectation: The contour plot in grey denotes contour lines of the density function. The expectation value will in this case be near the maximum of the density function as there is a lot of probability mass around it. The shade in red denotes a measurable set (event) A. The center of mass of X's probability distribution conditioned on A is depicted as well.

Having defined conditional expectations on specific events A by $\mathbb{E}[X\,|\,A]$ we could ask ourselves if we can generalize that notion to collections of sets A. In probability theory, those are σ-Algebras.

Consider a common dice with six sides. We choose the probability space canonically: $\Omega = \{1, 2, ..., 6\}$ with elementary probabilities $\mathbb{P}(\{1\}) = \cdots = \mathbb{P}(\{6\}) = \frac{1}{6}$. As σ-Algebras we take $\mathfrak{A} = \{\emptyset, \{1, 2\}, \{3, ..., 6\}, \Omega\}$, a σ-Algebra "unable to make distinctions" for example between 1 and 2. There are two nontrivial conditional expectations. Denote $A_1 = \{1, 2\}$ and $A_2 = \{3, ..., 6\}$ for brevity:

$$\begin{aligned}\mathbb{E}[X\,|\,A_1] &= \frac{\mathbb{E}[1_{A_1}X]}{\mathbb{P}(A_1)} = \frac{\frac{1}{6}\cdot(1+2)}{\frac{2}{6}} = \frac{3}{2}\\ \mathbb{E}[X\,|\,A_2] &= \frac{\mathbb{E}[1_{A_2}X]}{\mathbb{P}(A_2)} = \frac{\frac{1}{6}\cdot(3+4+5+6)}{\frac{4}{6}} = \frac{9}{2}\end{aligned}$$

The two remaining expectations are $\mathbb{E}[X\,|\,\emptyset] = 0$ and $\mathbb{E}[X\,|\,\Omega] = \mathbb{E}[X] = \frac{7}{2}$.

This leads us to consider conditional expectations as being dependent from chance. Formalized, this gives rise to the following definition:

Definition 4. *Conditional expectation with respect to a countable collection of events*

Let $(B_i)_{i\in I}$ be an ***at-most countably collection of disjoint sets*** *$B_i \subset \Omega$ with $\biguplus_{i\in I} B_i = \Omega$. We construct the σ-Algebra generated by all unions and intersections of B_i sets $\mathcal{F} \equiv \sigma(\{B_i\}_{i\in I})$. Consider some probability measure $\mathbb{P}$ to complete the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

Let $X \in L^1(\mathbb{P})$ be a random variable. We define the conditional expectation of X given $\mathcal{F}$ *as the* ***random variable***

$$\mathbb{E}[X\,|\,\mathcal{F}](\omega) = \mathbb{E}[X\,|\,B_i] \quad \Leftrightarrow \quad \omega \in B_i.$$

Lemma 5.

The random variable from Definition 4 has the following properties:

- $\mathbb{E}[X\,|\,\mathcal{F}]$ *is measurable with respect to $\mathcal{F}$.*

- $\mathbb{E}[X\,|\mathcal{F}] \in L^1(\mathbb{P})$ *and for every* $A \in \mathcal{F}$

$$\int_A \mathbb{E}[X\,|\mathcal{F}]\mathrm{d}\mathbb{P} = \int_A X\mathrm{d}\mathbb{P}.$$

In particular, $\mathbb{E}[\mathbb{E}[X\,|\mathcal{F}]] = \mathbb{E}[X]$

Remark 6. Measurability and the integral condition will be defining properties for a more general notion of conditional expectations in the next section. The last property can be taken as "The mean value of all centers of masses of disjoint subsets is equal to the actual center of mass".

Remark 7. Note that we took the following order of steps on defining conditional probabilities:

1. Define conditional expectations $\mathbb{E}[X\,|A]$ on individual events $A \in \mathfrak{A}$.
2. Generalize to conditional expectations $\mathbb{E}[X\,|\mathcal{F}]$ on a (certain type of) σ-Algebra.

This is a natural way of introducing conditional expectations on "simple" events as the expectations $\mathbb{E}[X\,|A]$ are easily defined but the progression will be reversed for more general types of conditional expectations: The value of $\mathbb{E}[X\,|Y = y]$ for singular events $\{Y = y\}$ needs to be derived from the notion of conditional expectation on σ-Algebras.

This can lead to a lot of misunderstandings if ignored.

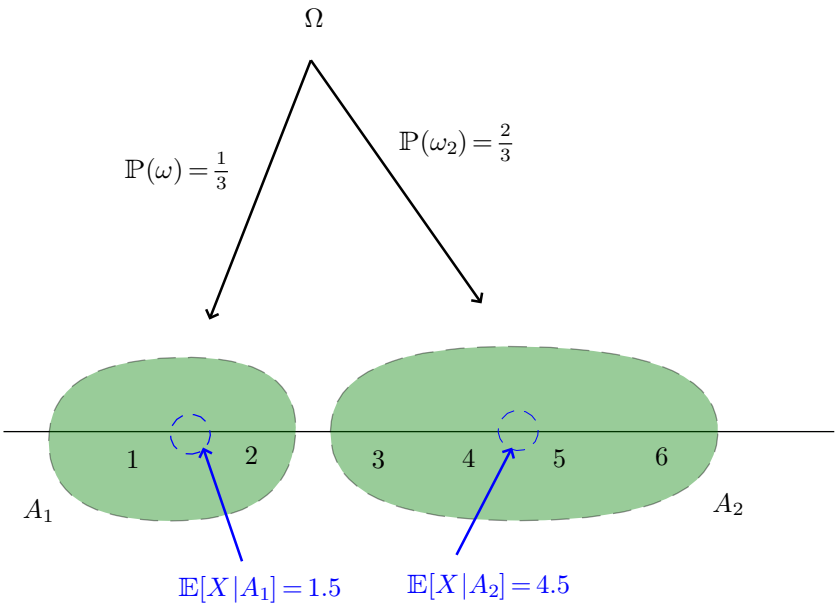

Figure 2. Conditional expectation with respect to a σ-Algebra as a random variable.

2 Conditional expectation with respect to a σ-Algebra

2.1 Why do we need all that?

Consider the following example: The bias B of a bent coin is unknown to us, we model it by a uniform probability distribution on $[0, 1]$, i.e. every "bentness parameter" is equally possible. Now denote the result of the coin toss as X. What is the probability of seeing "Heads", i.e. $\mathbb{P}[X = H]$? And more concretely, what is

$$\mathbb{P}[X = H\,|B = 0.3]$$

Intuitively, the last probability needs to be 0.3. We can't solve that problem with our current machinery, though:

$$\mathbb{P}(X = H \,|B = 0.3) = \mathbb{E}[1_{\{X=H\}}|B = 0.3|]$$

Tempted to use (1), we would obtain an invalid expression: $\{B = 0.3\}$ is a *singular event* in our case, so its probability is 0. We will derive a better notion of conditional expectation for singular events in the next section, but first we need conditional expectations on σ-Algebras, as announced in Remark 7.

2.2 Let's dive in

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be a probability space, $\mathcal{F} \subset \mathfrak{A}$ be a σ-Algebra and $X \in L^1(\Omega, \mathfrak{A}, \mathbb{P})$. Following our intuition from Lemma 5, we give the following definition:

Definition 8.

The random variable Y is called conditional expectation of X given $\mathcal{F}$, *in symbols* $Y = \mathbb{E}[X\,|\mathcal{F}]$ *if*

i. Y is measurable with respect to $\mathcal{F}$ and

ii. For every $A \in \mathcal{F}$ one has $\mathbb{E}[1_A X] = \mathbb{E}[1_A Y]$.

For $B \in \mathcal{A}$ we call $\mathbb{P}[B\,|\mathcal{F}] \equiv \mathbb{E}[1_B\,|\mathcal{F}]$ the conditional probability of B given $\mathcal{F}$.

For a random variable Z we call $\mathbb{E}[X\,|Z] \equiv \mathbb{E}[X\,|\sigma(Z)]$ the conditional expectation of X given Z.

Theorem 9. *$\mathbb{E}[X\,|\mathcal{F}]$ exists and is unique a.s.*

Proof. For a proof see for example [1] □

Remark 10. We can argue how our definition of $\mathbb{E}[X\,|\mathcal{F}]$ fits in the framework of the last section: We were able to derive the quantity $\mathbb{E}[X\,|\mathcal{G}]$ for $\mathcal{G} = \sigma(\{B_i\}_{i\in I})$ and $\biguplus_{i\in I} B_i = \Omega$. This intuitive notion of conditional expectation fulfills Definition 8 and is by uniqueness thus identically to the more general version (that justifies our "method overloading").

Remark 11. Interpretation of conditional expectations w.r.t. a σ-Algebra

The condition $\mathbb{E}[1_A X] = \mathbb{E}[1_A \mathbb{E}[X\,|\mathcal{F}]]$ on measurable sets $A \in \mathcal{F}$ can be interpreted as follows: $\mathbb{E}[X\,|\mathcal{F}]$ and X carry the same probability mass on "event chunks" $A \in \mathcal{F}$. The measurability criterion on the conditional expectation means: When we aks ourselves the question, "What's the cause for $\mathbb{E}[X\,|\mathcal{F}] \in M$, $M \in \mathcal{B}(\mathbb{R})$?", we get an answer that's only in $\mathcal{F}$, as

$$(\mathbb{E}[X\,|\mathcal{F}])^{-1}(M) \in \mathcal{F}$$

So if we interpret $\mathcal{F}$ as being the information we have and can use, the conditional expectation gives us a "guess" on X with the same probability chunk but only in terms of events of $\mathcal{F}$.

Theorem 12. *Properties of conditional expectation*

Let $(\Omega, \mathfrak{A}, \mathbb{P})$ be as above, $\mathcal{G} \subset \mathcal{F} \subset \mathfrak{A}$ be σ-Algebras and $Y \in L^1(\Omega, \mathfrak{A}, \mathbb{P})$. Then

i. ***(linearity):*** $\mathbb{E}[\lambda X + Y\,|\mathcal{F}] = \lambda \mathbb{E}[X\,|\mathcal{F}] + \mathbb{E}[Y\,|\mathcal{F}]$.

ii. ***(monotonicity):*** *For $X \geqslant Y$ a.s.,* $\mathbb{E}[X\,|\mathcal{F}] \geqslant \mathbb{E}[Y\,|\mathcal{F}]$.

iii. ***(on measurable random variables):*** *For Y measurable w.r.t. $\mathcal{F}$ and $\mathbb{E}[|XY|] < \infty$,*

$$\mathbb{E}[XY|\mathcal{F}] = Y\,\mathbb{E}[X\,|\mathcal{F}] \quad \textit{and} \quad \mathbb{E}[Y\,|\mathcal{F}] = \mathbb{E}[Y\,|Y] = Y.$$

iv. ***(stacking property):*** $\mathbb{E}[\mathbb{E}[X\,|\mathcal{F}]|\mathcal{G}] = \mathbb{E}[\mathbb{E}[X\,|\mathcal{G}]|\mathcal{F}] = \mathbb{E}[X\,|\mathcal{G}]$.

v. ***(△-inequality):*** $\mathbb{E}[|X|\,|\,\mathcal{F}] \geqslant |\mathbb{E}[X|\mathcal{F}]$.

vi. ***(on independent random variables):*** *For* X *independent from* $\mathcal{F}$, $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X]$.

vii. ***(on bounded convergent sequences):*** *For* $Y \geqslant 0$ *and a sequence* $X_n \to X$ *a.s. with* $|X_n| \leqslant Y$,

$$\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{F}] = \mathbb{E}[X|\mathcal{F}] \quad \text{a.s. and in } L^1(\mathbb{P}).$$

Remark 13. Another interpretation

Intuitively, $\mathbb{E}[X|\mathcal{F}]$ is the best prediction we can make about the value of X if we only know information on the events in the smaller σ-Algebra $\mathcal{F}$.

Property *iii.* then means that if all preimages of Y are in the (known) σ-Algebra $\mathcal{F}$, then our guess is very precise, i.e. $\mathbb{E}[X|\mathcal{F}] = X$.

Property *vi.* states that if the random variable X is independent from our knowledge $\mathcal{F}$, our best guess is just the ordinary expectation value.

For L^2-integrable random variables we have another intuitive way of thinking about conditional expectations:

Theorem 14. *Conditional expectation as a projection*

Let $\mathcal{F} \subset \mathfrak{A}$ *be a* σ*-Algebra and* X *be a random variable with finite variance* $\mathbb{E}[X^2] < \infty$. *Then* $\mathbb{E}[X|\mathcal{F}]$ *is the orthogonal projection of* X *on* $L^2(\Omega, \mathcal{F}, \mathbb{P})$. *This means that for every* $\mathcal{F}$*-measurable* Y *with finite variance* $\mathbb{E}[Y^2] < \infty$,

$$\mathbb{E}[(X-Y)^2] \geqslant \mathbb{E}[(X - \mathbb{E}[X|\mathcal{F}])^2]$$

with equality if and only if $Y = \mathbb{E}[X|\mathcal{F}]$.

Remark 15. A last attempt on interpretation

Theorem 14 thus states that conditional expectation is really the best guess among all $\mathcal{F}$-measurable random variables Y in the sense that it is the one with the smallest distance to X, if we interpret variance as a distance (which is legitimate given L^2 is a normed vector space)

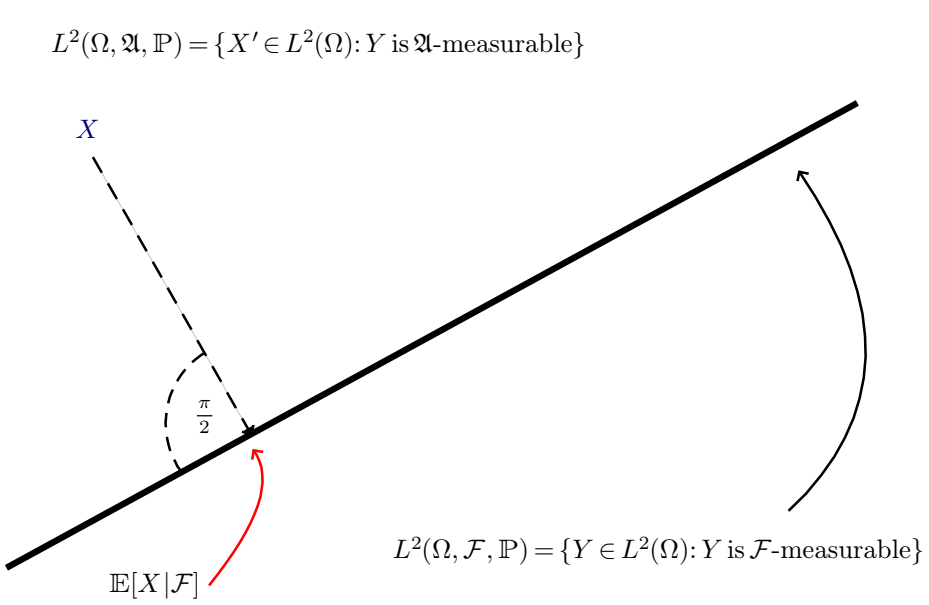

Figure 3. Conditional expectation as a projection

3 Conditional expectation with respect to singular events

Lemma 16. *Factorization lemma*

Let $(\Omega, \mathfrak{A})$ *and* $(\Omega', \mathfrak{A}')$ *be two measure spaces and* $f: \Omega \to \Omega'$ *be a map.*

For any $g: \Omega \to \bar{\mathbb{R}} \equiv \mathbb{R} \cup \{\infty\}$ *we have the following equivalence:*

$$g \text{ is } \sigma(f)\text{-measurable} \quad \Leftrightarrow \quad \text{There is a measurable map } \varphi: (\Omega', \mathfrak{A}') \to (\bar{\mathbb{R}}, \mathcal{B}(\bar{\mathbb{R}})) \text{ with } g = \varphi \circ f$$

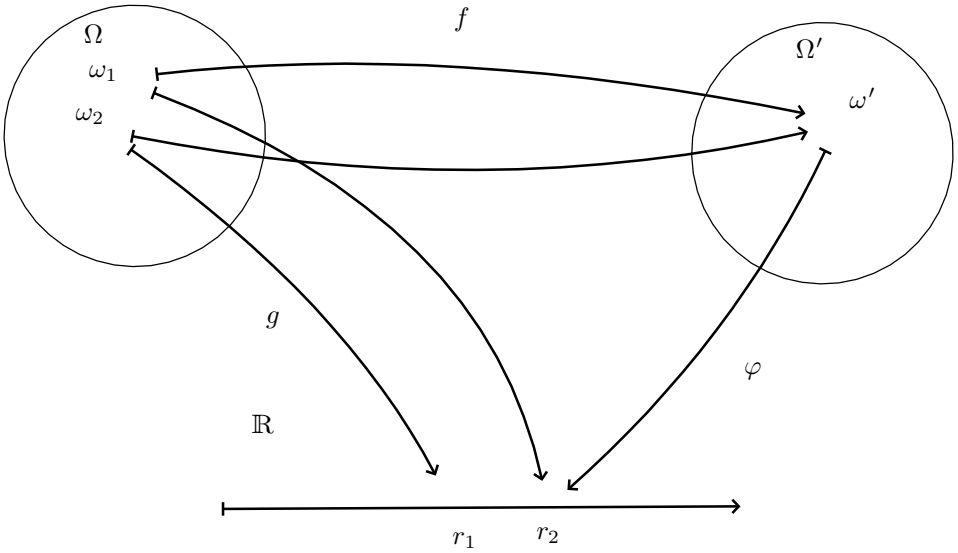

Figure 4. A counterexample for the Factorization Lemma: Assume $\Omega = \{\omega_1, \omega_2\}$ and $\Omega' = \{\omega\}$. Choose standard σ-Algebras $\mathfrak{A} = \mathcal{P}(\Omega)$, $\mathfrak{A}' = \mathcal{P}(\Omega')$ and $\mathcal{B}(\mathbb{R})$. This set of mappings does not fulfill the requirements of the Factorization Lemma: The concatenation $\varphi \circ f$ is not equal to g, as $\varphi \circ f(\omega_2) = \varphi(\omega') = r_2$, whereas $g(\omega_2) = r_1$. This is due to the fact that g is not $\sigma(f)$-measurable: $\sigma(f) = \{f^{-1}(A') | A' \in \mathfrak{A}'\} = \{\emptyset, \{\omega_1, \omega_2\}\}$. Now for small ε, the set $R = (r_1 - \varepsilon, r_1 + \varepsilon)$ is open but $g^{-1}(R) = \{\omega_1\} \notin \sigma(f)$. Intuitively, the problem is that g and f "cluster" events in Ω differently: For g, both single events have different results whereas f groups them together.

Existence of conditional expectations w.r.t. a singular event can be proven (non-constructivistically) by this lemma:

Assume $X: (\Omega, \mathfrak{A}, \mathbb{P}) \to (E, \mathcal{E})$ is a random variable into an measurable space E and $Z = \mathbb{E}[Y|\sigma(X)]: (\Omega, \mathfrak{A}, \mathbb{P}) \to \mathbb{R}$ be the conditional expectation of a random variable Y. According to the factorization lemma, there exists a map $\varphi: E \to \mathbb{R}$ such that φ is $(\mathcal{E}, \mathcal{B}(\mathbb{R}))$-measurable and $\varphi(X) = \mathbb{E}[Y|\sigma(X)]$. If X is surjective, φ is uniquely defined. In this case we write $\varphi \equiv Z \circ X^{-1}$, for $Z \equiv \mathbb{E}[Y|\sigma(X)]$ even though the inverse of X doesn't exist.

Definition 17. *Conditional expectation with respect to a continous random variable's results*

Let $Y \in L^1(\mathbb{P})$ *and* $X: (\Omega, \mathfrak{A}) \to (E, \mathcal{E})$. *Then we call the function* φ *from above for* $Z = \mathbb{E}[Y|X]$ *as the* conditional expectation of Y given $X = x$, *in terms* $\mathbb{E}[Y|X = x]$. *By analogy, we write* $\mathbb{P}(A|X = x) = \mathbb{E}[1_A|X = x]$ *for* $A \in \mathfrak{A}$.

This means that $\mathbb{E}[Y|X = x] \equiv \mathbb{E}[Y|X](\{X = x\})$.

Beware that $\varphi = Z \circ X^{-1}$ *only a.e. The exception set for this equality depends on the function* Y *or, in the conditional probability case, on the set* A.

Remark 18. Why can we reasonably call φ by $\mathbb{E}[Y|X=\cdot]$? In the regular case (discrete X), the construction is parallel to the regular $\mathbb{E}[Y|X=\cdot]$ (see Definition 4) and for the irregular case (continous X), this mapping extracts the right portion out of $\mathbb{E}[Y|X]$.

Remark 19. We note that the conditional expectation $\mathbb{E}[Y|X=\cdot]$ is not defined on a set of measure 0, i.e. $\mathbb{E}[Y|X=x]$ makes only sense for a.s. all $x\in E$ where the exception set is dependent of Y.

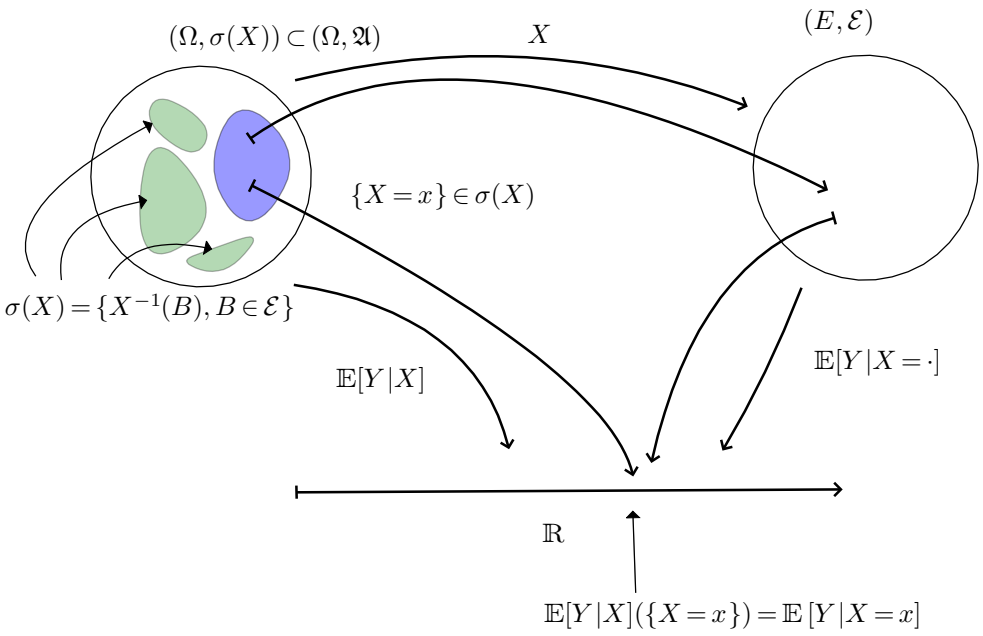

Figure 5. Conditional expectation w.r.t. singular events as concatenation of cond. exp. w.r.t. a random variable's σ-Algebra $\sigma(X)$ and the inverse image of X.

Why aren't we done? The factorization lemma yields a function (later called $\mathbb{E}[Y|X=\cdot]$) for every random variable Y. If we're only interested in single expectations, we're done. The definition of $\mathbb{P}(A|X=x)$ carries a hidden pitfall, though: The conditional probability of A given $X=x$ is defined via the conditional expectation of 1_A given $X=x$. In remark 19 we saw that the exception set of the definition of conditional expectations is dependent of the function Y. Hence, the definition of $\mathbb{P}(A|X=x)$ makes only sense for a A-dependent subset of x's. For different A we might fear that the exception sets amount to more than a set of measure 0, so we can't define a "joint measure" $\mathbb{P}(\cdot|X=x)$ but for a set of x's of measure 0. Thus, if we want to work with conditional probabilities, we first have to show that this cannot happen.

4 Regular Version of Conditional Probability

Intuitively, the problem is the "disconnectedness" of $(\Omega,\mathfrak{A})$ and $(E,\mathcal{E})$. Our definition of conditional expectations $\mathbb{E}[1_A|X=x]$ is non-uniform over the range of possible sets $A\in\mathfrak{A}$. The next definition incorporates the correct notion for making that connection uniformly:

Definition 20. *Markov Kernel*

For $(\Omega_1,\mathfrak{A}_1)$ and $(\Omega_2,\mathfrak{A}_2)$ measurable spaces we call a mapping $\kappa:\Omega_1\times\mathfrak{A}_2\to[0,\infty]$ a Markov Kernel from Ω_1 to Ω_2, *if*

i. $\omega_1\mapsto\kappa(\omega_1,A_2)$ is $\mathfrak{A}_1$-measurable for every $A_2\in\mathfrak{A}_2$,

ii. $A_2 \mapsto \kappa(\omega_1, A_2)$ *is a probability measure on* $(\Omega_2, \mathfrak{A}_2)$ *for every* $\omega_1 \in \Omega_1$.

Definition 21. *Regular Version of Conditional Probability*

Let $Y\colon (\Omega, \mathfrak{A}) \to (E, \mathcal{E})$ *be a random variable and* $\mathcal{F} \subset \mathfrak{A}$ *be a sub-*σ*-Algebra. Assume we have a stochastic kernel* $\kappa_{Y,\mathcal{F}}$ *from* $(\Omega, \mathcal{F})$ *to* $(E, \mathcal{E})$ *fulfilling*

$$\int 1_B(Y) \cdot 1_A \mathrm{d}\mathbb{P} = \int \kappa_{Y,\mathcal{F}}(\cdot, B) \cdot 1_A \mathrm{d}\mathbb{P} \quad \text{for every } A \in \mathcal{F} \text{ and } B \in \mathcal{E},$$

i.e. $\kappa_{Y,\mathcal{F}}(\omega, B) = \mathbb{P}(\{Y \in B\}|\mathcal{F})(\omega)$ *for* $\mathbb{P}$*-a.s.* $\omega \in \Omega$ *and every* $B \in \mathcal{E}$.

Then we call $\kappa_{Y,\mathcal{F}}$ *a* regular version of the conditional probability of Y given $\mathcal{F}$.

For $\mathcal{F} = \sigma(X)$ *for some random variable* X*, we call* $\kappa_{Y,\mathcal{F}} = \kappa_{Y,\sigma(X)}$ *a regular version of the conditional probability of* Y *given* X.

Remark 22. This means that $\kappa_{Y,\mathcal{F}}$ is a conditional probability in the sense of definition 17 but also regular in the sense that there is a "good" correspondence between target sets of Y and events ω, also $\mathbb{P}(\{Y \in B\}|\mathcal{F})$ is now a probability measure for sets of the type $\{Y \in B\}$ for almost all $\omega \in \Omega$ (see the definition of Markov Kernels).

Theorem 23. *Existence of regular versions of conditional probabilities*

For $Y\colon (\Omega, \mathfrak{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *a real-valued random variable, there is a regular version of the conditional probability distribution* $\mathbb{P}(\{Y \in \cdot\}|\mathcal{F})$.

Proof. See [1]. □

Example 24. Most important example: Conditional densities

Let X, Y be real random variables with joint probability density f, i.e. $\mathbb{P}(X \in A, Y \in B) = \int_{A\times B} f(x, y)\, \mathrm{d}(x, y)$. The marginalization $\int_{\mathbb{R}} f(x, y)\mathrm{d}y = f_X(x)$ is the density of X. Then the regular version of the conditional probability of Y given X has a density given by

$$f_{Y|X}(x, y) \equiv \frac{f(x, y)}{f_X(x)}.$$

As a symbol we also write $f_{Y|X}(x, y) = \frac{\mathbb{P}(Y \in \mathrm{d}y|X = x)}{\mathrm{d}y}$.

Proof. We need to show that $\mathbb{P}(\{Y \in B\}|X = x) = \int_B f_{Y|X}(y, x)\mathrm{d}y$ in the sense of definition 21: Measurability of $\int_B f_{Y|X}(y, x)\mathrm{d}y$ is obtained by Fubini and

$$\begin{aligned}
\int_A \mathbb{P}(\{Y \in B\}|X = x)\, \mathbb{P}(X \in \mathrm{d}x) &= \int_A \int_B f_{Y|X}(x, y)\mathrm{d}y \mathbb{P}(X \in \mathrm{d}x) \\
&= \int_A \int_B f(x, y)\mathrm{d}y \frac{\mathbb{P}(X \in \mathrm{d}x)}{f_X(x)} \\
&= \int_A \int_B f(x, y)\mathrm{d}y\mathrm{d}x \\
&= \int_A 1_B(Y)\mathrm{d}x \\
&= \mathbb{P}(X \in A, Y \in B)
\end{aligned}$$

where the penultimate equality already constitutes the definition of the regular version of conditional probability. □

Bibliography

[1] Achim Klenke. *Probability Theory - A Comprehensive Course*. Springer, 2014.