8

Conditional Expectations

If there is partial information on the outcome of a random experiment, the probabilities for the possible events may change. The concept of conditional probabilities and conditional expectations formalises the corresponding calculus.

8.1 Elementary Conditional Probabilities

Example 8.1. We throw a die and consider the events

$$A := \{\text{the face shows three or smaller}\},$$
$$B := \{\text{the face shows an odd number}\}.$$

Clearly, $\mathbf{P}[A] = \frac{1}{2}$ and $\mathbf{P}[B] = \frac{1}{2}$. However, what is the probability that B occurs if we already know that A occurs?

We model the experiment on the probability space $(\Omega, \mathcal{A}, \mathbf{P})$, where $\Omega = \{1, \ldots, 6\}$, $\mathcal{A} = 2^{\Omega}$ and $\mathbf{P}$ is the uniform distribution on Ω. Then

$$A = \{1,\, 2,\, 3\} \quad \text{and} \quad B = \{1,\, 3,\, 5\}.$$

If we know that A has occurred, it is plausible to assume the uniform distribution on the remaining possible outcomes; that is, on $\{1,\, 2,\, 3\}$. Thus we define a new probability measure $\mathbf{P}_A$ on $(A, 2^A)$ by

$$\mathbf{P}_A[C] = \frac{\#C}{\#A} \quad \text{for } C \subset A.$$

By assigning the points in $\Omega \setminus A$ probability zero (since they are impossible if A has occurred), we can extend $\mathbf{P}_A$ to a measure on Ω:

$$\mathbf{P}[C \,|\, A] \;:=\; \mathbf{P}_A[C \cap A] \;=\; \frac{\#(C \cap A)}{\#A} \quad \text{for } C \subset \Omega.$$

In this way, we get $\mathbf{P}[B\,|\,A] = \dfrac{\#\{1,\,3\}}{\#\{1,\,2,\,3\}} = \dfrac{2}{3}$. ◇

Motivated by this example, we make the following definition.

Definition 8.2 (Conditional probability). *Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $A \in \mathcal{A}$. We define the **conditional probability given** A for any $B \in \mathcal{A}$ by*

$$\mathbf{P}[B\,|\,A] = \begin{cases} \dfrac{\mathbf{P}[A \cap B]}{\mathbf{P}[A]}, & \text{if } \mathbf{P}[A] > 0, \\ 0, & \text{else}. \end{cases} \tag{8.1}$$

Remark 8.3. The specification in (8.1) for the case $\mathbf{P}[A] = 0$ is arbitrary and is of no importance. ◇

Theorem 8.4. *If $\mathbf{P}[A] > 0$, then $\mathbf{P}[\,\cdot\,|\,A]$ is a probability measure on $(\Omega, \mathcal{A})$.*

Proof. This is obvious. □

Theorem 8.5. *Let $A, B \in \mathcal{A}$ with $\mathbf{P}[A]$, $\mathbf{P}[B] > 0$. Then*

$$A, B \text{ are independent} \iff \mathbf{P}[B\,|\,A] = \mathbf{P}[B] \iff \mathbf{P}[A\,|\,B] = \mathbf{P}[A].$$

Proof. This is trivial! □

Theorem 8.6 (Summation formula). *Let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint sets with $\mathbf{P}\left[\biguplus_{i\in I} B_i\right] = 1$. Then, for any $A \in \mathcal{A}$,*

$$\mathbf{P}[A] = \sum_{i\in I} \mathbf{P}[A\,|\,B_i]\,\mathbf{P}[B_i]. \tag{8.2}$$

Proof. Due to the σ-additivity of $\mathbf{P}$, we have

$$\mathbf{P}[A] = \mathbf{P}\left[\biguplus_{i\in I}(A \cap B_i)\right] = \sum_{i\in I} \mathbf{P}[A \cap B_i] = \sum_{i\in I} \mathbf{P}[A\,|\,B_i]\mathbf{P}[B_i]. \qquad \square$$

Theorem 8.7 (Bayes' formula). *Let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint sets with $\mathbf{P}\left[\biguplus_{i\in I} B_i\right] = 1$. Then, for any $A \in \mathcal{A}$ with $\mathbf{P}[A] > 0$ and any $k \in I$,*

$$\mathbf{P}[B_k\,|\,A] = \frac{\mathbf{P}[A\,|\,B_k]\,\mathbf{P}[B_k]}{\sum_{i\in I} \mathbf{P}[A\,|\,B_i]\,\mathbf{P}[B_i]}. \tag{8.3}$$

Proof. We have

$$\mathbf{P}[B_k\,|\,A] = \frac{\mathbf{P}[B_k \cap A]}{\mathbf{P}[A]} = \frac{\mathbf{P}[A\,|\,B_k]\,\mathbf{P}[B_k]}{\mathbf{P}[A]}.$$

Now use the expression in (8.2) for $\mathbf{P}[A]$. □

Example 8.8. In the production of certain electronic devices, a fraction of 2% of the production is defective. A quick test detects a defective device with probability 95%; however, with probability 10% it gives a false alarm for an intact device.

If the test gives an alarm, what is the probability that the device just tested is indeed defective?

We formalise the description given above. Let

$$A := \{\text{device is declared as defective}\},$$
$$B := \{\text{device is defective}\},$$

and

$$\mathbf{P}[B] = 0.02, \qquad \mathbf{P}[B^c] = 0.98,$$
$$\mathbf{P}[A\,|\,B] = 0.95, \qquad \mathbf{P}[A\,|\,B^c] = 0.1.$$

Bayes' formula yields

$$\mathbf{P}[B\,|\,A] = \frac{\mathbf{P}[A\,|\,B]\,\mathbf{P}[B]}{\mathbf{P}[A\,|\,B]\,\mathbf{P}[B] + \mathbf{P}[A\,|\,B^c]\,\mathbf{P}[B^c]}$$
$$= \frac{0.95 \cdot 0.02}{0.95 \cdot 0.02 + 0.1 \cdot 0.98} = \frac{19}{117} \approx 0.162.$$

On the other hand, the probability that a device that was not classified as defective is in fact defective is

$$\mathbf{P}[B\,|\,A^c] = \frac{0.05 \cdot 0.02}{0.05 \cdot 0.02 + 0.9 \cdot 0.98} = \frac{1}{883} \approx 0.00113. \qquad \diamond$$

Now let $X \in \mathcal{L}^1(\mathbf{P})$. If $A \in \mathcal{A}$, then clearly also $\mathbb{1}_A X \in \mathcal{L}^1(\mathbf{P})$. We define

$$\mathbf{E}[X;\,A] := \mathbf{E}[\mathbb{1}_A\,X]. \tag{8.4}$$

If $\mathbf{P}[A] > 0$, then $\mathbf{P}[\,\cdot\,|\,A]$ is a probability measure. Since $\mathbb{1}_A X \in \mathcal{L}^1(\mathbf{P})$, we have $X \in \mathcal{L}^1(\mathbf{P}[\,\cdot\,|\,A])$. Hence we can define the expectation of X with respect to $\mathbf{P}[\,\cdot\,|\,A]$.

Definition 8.9. *Let $X \in \mathcal{L}^1(\mathbf{P})$ and $A \in \mathcal{A}$. Then we define*

$$\mathbf{E}[X\,|\,A] := \int X(\omega)\,\mathbf{P}[d\omega\,|\,A] = \begin{cases} \dfrac{\mathbf{E}[\mathbb{1}_A X]}{\mathbf{P}[A]}, & \text{if } \mathbf{P}[A] > 0, \\ 0, & \text{else.} \end{cases} \tag{8.5}$$

Clearly, $\mathbf{P}[B\,|\,A] = \mathbf{E}[\mathbb{1}_B\,|\,A]$ for all $B \in \mathcal{A}$.

Consider now the situation that we studied with the summation formula for conditional probabilities. Hence, let I be a countable set and let $(B_i)_{i\in I}$ be pairwise disjoint events with $\biguplus_{i\in I} B_i = \Omega$. We define $\mathcal{F} := \sigma(B_i,\, i \in I)$. For $X \in \mathcal{L}^1(\mathbf{P})$, we define a map $\mathbf{E}[X\,|\,\mathcal{F}] : \Omega \to \mathbb{R}$ by

$$\mathbf{E}[X\,|\,\mathcal{F}](\omega) = \mathbf{E}[X\,|\,B_i] \quad \Longleftrightarrow \quad B_i \ni \omega. \tag{8.6}$$

Lemma 8.10. *The map $\mathbf{E}[X\,|\,\mathcal{F}]$ has the following properties.*

(i) $\mathbf{E}[X\,|\,\mathcal{F}]$ is $\mathcal{F}$-measurable.

(ii) $\mathbf{E}[X\,|\,\mathcal{F}] \in \mathcal{L}^1(\mathbf{P})$, and for any $A \in \mathcal{F}$, we have $\displaystyle\int_A \mathbf{E}[X\,|\,\mathcal{F}]\,d\mathbf{P} = \int_A X\,d\mathbf{P}$.

Proof. **(i)** Let f be the map $f : \Omega \to I$ with

$$f(\omega) = i \qquad \Longleftrightarrow \qquad B_i \ni \omega.$$

Further, let $g : I \to \mathbb{R}$, $i \mapsto \mathbf{E}[X\,|\,B_i]$. Since I is discrete, g is measurable. Since f is $\mathcal{F}$-measurable, $\mathbf{E}[X\,|\,\mathcal{F}] = g \circ f$ is also $\mathcal{F}$-measurable.

(ii) Let $A \in \mathcal{F}$ and $J \subset I$ with $A = \biguplus_{j\in J} B_j$. Let $J' := \{i \in J :\, \mathbf{P}[B_i] > 0\}$. Hence

$$\int_A \mathbf{E}[X\,|\,\mathcal{F}]\,d\mathbf{P} \;=\; \sum_{i\in J'} \mathbf{P}[B_i]\,\mathbf{E}[X\,|\,B_i] \;=\; \sum_{i\in J'} \mathbf{E}[\mathbb{1}_{B_i} X] \;=\; \int_A X\,d\mathbf{P}. \qquad \square$$

Exercise 8.1.1 (Lack of memory of the exponential distribution). Let X be a non-negative random variable and let $\theta > 0$. Show that X is exponentially distributed if and only if

$$\mathbf{P}[X > t + s\,|\,X > s] = \mathbf{P}[X > t] \quad \text{ for all } \; s, t \geq 0.$$

In particular, $X \sim \exp_\theta$ if and only if $\mathbf{P}[X > t + s\,|\,X > s] = e^{-\theta t}$ for all $s, t \geq 0$. ♣

Exercise 8.1.2. Consider a theatre with n seats that is fully booked for this evening. Each of the n people entering the theatre (one by one) has a seat reservation. However, the first person is absent-minded and takes a seat at random. Any subsequent person takes his or her reserved seat if it is free and otherwise picks a free seat at random.

(i) What is the probability that the last person gets his or her reserved seat?

(ii) What is the probability that the kth person gets his or her reserved seat? ♣

8.2 Conditional Expectations

Let X be a random variable that is uniformly distributed on $[0,1]$. Assume that if we know the value $X = x$, the random variables $Y_1, \ldots, Y_n$ are independent and Ber_x-distributed. So far, with our machinery we can only deal with conditional probabilities of the type $\mathbf{P}[\,\cdot\,|\,X \in [a,b]]$, $a < b$ (since $X \in [a,b]$ has positive probability). How about $\mathbf{P}[Y_1 = \ldots = Y_n = 1 \,\big|\, X = x]$? Intuitively, this should be x^n. We thus need a notion of conditional probabilities that allows us to deal with conditioning on events with probability zero and that is consistent with our intuition. In the next section, we will see that in the current example this can be done using transition kernels. First, however, we have to consider a more general situation.

In the following, $\mathcal{F} \subset \mathcal{A}$ will be a sub-σ-algebra and $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbf{P})$. In analogy with Lemma 8.10, we make the following definition.

Definition 8.11 (Conditional expectation). *A random variable Y is called a **conditional expectation** of X given $\mathcal{F}$, symbolically $\mathbf{E}[X\,|\,\mathcal{F}] := Y$, if:*

(i) Y is $\mathcal{F}$-measurable.

(ii) For any $A \in \mathcal{F}$, we have $\mathbf{E}[X\mathbb{1}_A] = \mathbf{E}[Y\mathbb{1}_A]$.

*For $B \in \mathcal{A}$, $\mathbf{P}[B|\mathcal{F}] := \mathbf{E}[\mathbb{1}_B\,|\,\mathcal{F}]$ is called a **conditional probability** of B given the σ-algebra $\mathcal{F}$.*

Theorem 8.12. *$\mathbf{E}[X\,|\,\mathcal{F}]$ exists and is unique (up to equality almost surely).*

Since conditional expectations are defined only up to equality a.s., all equalities with conditional expectations are understood as equalities a.s., even if we do not say so explicitly.

Proof. Uniqueness. Let Y and Y' be random variables that fulfil (i) and (ii). Let $A = \{Y > Y'\} \in \mathcal{F}$. Then, by (ii),

$$0 \;=\; \mathbf{E}[Y\mathbb{1}_A] - \mathbf{E}[Y'\mathbb{1}_A] \;=\; \mathbf{E}[(Y - Y')\,\mathbb{1}_A].$$

Since $(Y - Y')\,\mathbb{1}_A \geq 0$, we have $\mathbf{P}[A] = 0$; hence $Y \leq Y'$ almost surely. Similarly, we get $Y \geq Y'$ almost surely.

Existence. Let $X^+ = X \vee 0$ and $X^- = X^+ - X$. By

$$Q^{\pm}(A) := \mathbf{E}[X^{\pm}\,\mathbb{1}_A] \quad \text{for all } A \in \mathcal{F},$$

we define two finite measures on $(\Omega, \mathcal{F})$. Clearly, $Q^{\pm} \ll \mathbf{P}$; hence the Radon-Nikodym theorem (Corollary 7.34) yields the existence of densities $Y^{\pm}$ such that

$$Q^{\pm}(A) = \int_A Y^{\pm}\, d\mathbf{P} = \mathbf{E}[Y^{\pm}\, \mathbb{1}_A].$$

Now define $Y = Y^+ - Y^-$. □

Definition 8.13. *If Y is a random variable and $X \in \mathcal{L}^1(\mathbf{P})$, then we define $\mathbf{E}[X\,|\,Y] := \mathbf{E}[X\,|\,\sigma(Y)]$.*

Theorem 8.14 (Properties of the conditional expectation). *Let $(\Omega, \mathcal{A}, \mathbf{P})$ and let X be as above. Let $\mathcal{G} \subset \mathcal{F} \subset \mathcal{A}$ be σ-algebras and let $Y \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbf{P})$. Then:*

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

(ii) ***(Monotonicity)*** *If $X \geq Y$ a.s., then $\mathbf{E}[X\,|\,\mathcal{F}] \geq \mathbf{E}[\,Y\,|\,\mathcal{F}]$.*

(iii) If $\mathbf{E}[|XY|] < \infty$ and Y is measurable with respect to $\mathcal{F}$, then

$$\mathbf{E}[XY\,|\,\mathcal{F}] = Y\,\mathbf{E}[X\,|\,\mathcal{F}] \qquad \text{and} \qquad \mathbf{E}[\,Y\,|\,\mathcal{F}] = \mathbf{E}[\,Y\,|\,Y] = Y.$$

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

(v) ***(Triangle inequality)*** $\mathbf{E}[|X|\,|\,\mathcal{F}] \geq \big|\mathbf{E}[X\,|\,\mathcal{F}]\big|$.

(vi) ***(Independence)*** *If $\sigma(X)$ and $\mathcal{F}$ are independent, then $\mathbf{E}[X\,|\,\mathcal{F}] = \mathbf{E}[X]$.*

(vii) If $\mathbf{P}[A] \in \{0, 1\}$ for any $A \in \mathcal{F}$, then $\mathbf{E}[X\,|\,\mathcal{F}] = \mathbf{E}[X]$.

(viii) ***(Dominated convergence)*** *Assume $Y \in \mathcal{L}^1(\mathbf{P})$, $Y \geq 0$ and $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables with $|X_n| \leq Y$ for $n \in \mathbb{N}$ and such that $X_n \overset{n\to\infty}{\longrightarrow} X$ a.s. Then*

$$\lim_{n\to\infty} \mathbf{E}[X_n\,|\,\mathcal{F}] = \mathbf{E}[X\,|\,\mathcal{F}] \quad \text{a.s. and in } L^1(\mathbf{P}). \tag{8.7}$$

Proof. **(i)** The right hand side is $\mathcal{F}$-measurable; hence, for $A \in \mathcal{F}$,

$$\begin{aligned} \mathbf{E}\big[\mathbb{1}_A\,\big(\lambda \mathbf{E}[X\,|\,\mathcal{F}] + \mathbf{E}[Y\,|\,\mathcal{F}]\big)\big] &= \lambda \mathbf{E}\big[\mathbb{1}_A\,\mathbf{E}[X\,|\,\mathcal{F}]\big] + \mathbf{E}\big[\mathbb{1}_A\,\mathbf{E}[Y\,|\,\mathcal{F}]\big] \\ &= \lambda \mathbf{E}[\mathbb{1}_A\,X] + \mathbf{E}[\mathbb{1}_A\,Y] \\ &= \mathbf{E}\big[\mathbb{1}_A\,(\lambda X + Y)\big]. \end{aligned}$$

(ii) Let $A = \{\mathbf{E}[X\,|\,\mathcal{F}] < \mathbf{E}[Y\,|\,\mathcal{F}]\} \in \mathcal{F}$. Since we have $X \geq Y$, we get $\mathbf{E}[\mathbb{1}_A\,(X - Y)] \geq 0$ and thus $\mathbf{P}[A] = 0$.

(iii) First assume $X \geq 0$ and $Y \geq 0$. For $n \in \mathbb{N}$, define $Y_n = 2^{-n}\lfloor 2^n Y \rfloor$. Then $Y_n \uparrow Y$ and $Y_n\,\mathbf{E}[X\,|\,\mathcal{F}] \uparrow Y\,\mathbf{E}[X\,|\,\mathcal{F}]$ (since $\mathbf{E}[X\,|\,\mathcal{F}] \geq 0$ by (ii)). By the monotone convergence theorem (Lemma 4.6(ii)),

$$\mathbf{E}\big[\mathbb{1}_A\,Y_n\,\mathbf{E}[X\,|\,\mathcal{F}]\big] \overset{n\to\infty}{\longrightarrow} \mathbf{E}\big[\mathbb{1}_A\,Y\,\mathbf{E}[X\,|\,\mathcal{F}]\big].$$

On the other hand,

$$\begin{aligned}\mathbf{E}\big[\mathbb{1}_A\, Y_n\, \mathbf{E}[X\,|\,\mathcal{F}]\big] &= \sum_{k=1}^{\infty} \mathbf{E}\big[\mathbb{1}_A\, \mathbb{1}_{\{Y_n = k\,2^{-n}\}}\, k\,2^{-n}\, \mathbf{E}[X\,|\,\mathcal{F}]\big] \\ &= \sum_{k=1}^{\infty} \mathbf{E}\big[\mathbb{1}_A\, \mathbb{1}_{\{Y_n = k\,2^{-n}\}}\, k\,2^{-n}\, X\big] \\ &= \mathbf{E}\big[\mathbb{1}_A\, Y_n X\big] \overset{n\to\infty}{\longrightarrow} \mathbf{E}[\mathbb{1}_A\, Y X].\end{aligned}$$

Hence $\mathbf{E}[\mathbb{1}_A\, Y\, \mathbf{E}[X\,|\,\mathcal{F}]] = \mathbf{E}[\mathbb{1}_A\, Y X]$. In the general case, write $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ and exploit the linearity of the conditional expectation.

(iv) The second equality follows from (iii) with $Y = \mathbf{E}[X\,|\,\mathcal{G}]$ and $X = 1$. Now let $A \in \mathcal{G}$. Then, in particular, $A \in \mathcal{F}$; hence

$$\mathbf{E}\big[\mathbb{1}_A \mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]\,|\,\mathcal{G}]\big] \;=\; \mathbf{E}\big[\mathbb{1}_A \mathbf{E}[X\,|\,\mathcal{F}]\big] \;=\; \mathbf{E}[\mathbb{1}_A\, X] \;=\; \mathbf{E}\big[\mathbb{1}_A\, \mathbf{E}[X\,|\,\mathcal{G}]\big].$$

(v) This follows from (i) and (ii) with $X = X^+ - X^-$.

(vi) Trivially, $\mathbf{E}[X]$ is measurable with respect to $\mathcal{F}$. Let $A \in \mathcal{F}$. Then X and $\mathbb{1}_A$ are independent; hence $\mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]\,\mathbb{1}_A] = \mathbf{E}[X\,\mathbb{1}_A] = \mathbf{E}[X]\,\mathbf{E}[\mathbb{1}_A]$.

(vii) For any $A \in \mathcal{F}$ and $B \in \mathcal{A}$, we have $\mathbf{P}[A \cap B] = 0$ if $\mathbf{P}[A] = 0$, and $\mathbf{P}[A \cap B] = \mathbf{P}[B]$ if $\mathbf{P}[A] = 1$. Hence $\mathcal{F}$ and $\mathcal{A}$ are independent and thus $\mathcal{F}$ is independent of any sub-σ-algebra of $\mathcal{A}$. In particular, $\mathcal{F}$ and $\sigma(X)$ are independent. Hence the claim follows from (vi).

(viii) Let $|X_n| \le Y$ for any $n \in \mathbb{N}$ and $X_n \overset{n\to\infty}{\longrightarrow} X$ almost surely. Define $Z_n := \sup_{k\ge n} |X_k - X|$. Then $0 \le Z_n \le 2Y$ and $Z_n \overset{\text{a.s.}}{\longrightarrow} 0$. By Corollary 6.26 (dominated convergence), we have $\mathbf{E}[Z_n] \overset{n\to\infty}{\longrightarrow} 0$; hence, by the triangle inequality,

$$\mathbf{E}\big[\big|\mathbf{E}[X_n\,|\,\mathcal{F}] - \mathbf{E}[X\,|\,\mathcal{F}]\big|\big] \le \mathbf{E}[\mathbf{E}[|X_n - X|\,\big|\,\mathcal{F}]] = \mathbf{E}[|X_n - X|] \le \mathbf{E}[Z_n] \overset{n\to\infty}{\longrightarrow} 0.$$

However, this is the $L^1(\mathbf{P})$-convergence in (8.7). Let $Z := \limsup_{n\to\infty} \mathbf{E}[Z_n\,\big|\,\mathcal{F}]$. By Fatou's lemma,

$$\mathbf{E}[Z] \;\le\; \lim_{n\to\infty} \mathbf{E}[Z_n] \;=\; 0.$$

Hence $Z = 0$ and thus $\mathbf{E}[Z_n\,\big|\,\mathcal{F}] \overset{n\to\infty}{\longrightarrow} 0$ almost surely. However, by (v),

$$\big|\mathbf{E}[X_n\,\big|\,\mathcal{F}] - \mathbf{E}[X\,\big|\,\mathcal{F}]\big| \;\le\; \mathbf{E}[Z_n]. \qquad \square$$

Remark 8.15. Intuitively, $\mathbf{E}[X\,|\,\mathcal{F}]$ is the best prediction we can make for the value of X if we only have the information of the σ-algebra $\mathcal{F}$. For example, if $\sigma(X) \subset \mathcal{F}$ (that is, if we know X already), then $\mathbf{E}[X\,|\,\mathcal{F}] = X$, as shown in (iii). At the other end of the spectrum is the case where X and $\mathcal{F}$ are independent; that is, where knowledge of $\mathcal{F}$ does not give any information on X. Here the best prediction for X is its mean; hence $\mathbf{E}[X] = \mathbf{E}[X\,|\,\mathcal{F}]$, as shown in (vii).

What exactly do we mean by "best prediction"? For square integrable random variables X, by the best prediction for X we will understand the $\mathcal{F}$-measurable random

variable that minimises the L^2-distance from X. The next corollary shows that the conditional expectation is in fact this minimiser. $\diamond$

Corollary 8.16 (Conditional expectation as projection). *Let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra and let X be a random variable with $\mathbf{E}[X^2] < \infty$. Then $\mathbf{E}[X\,|\,\mathcal{F}]$ is the orthogonal projection of X on $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$. That is, for any $\mathcal{F}$-measurable Y with $\mathbf{E}[Y^2] < \infty$,*

$$\mathbf{E}\big[(X-Y)^2\big] \geq \mathbf{E}\big[(X - \mathbf{E}[X\,|\,\mathcal{F}])^2\big]$$

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

Proof. First assume that $\mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]^2] < \infty$. (In Theorem 8.19, we will see that we have $\mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]^2] \leq \mathbf{E}[X^2]$, but here we want to keep the proof self-contained.) Let Y be $\mathcal{F}$-measurable and assume $\mathbf{E}[Y^2] < \infty$. Then, by the Cauchy-Schwarz inequality, we have $\mathbf{E}[|XY|] < \infty$. Thus, using the tower property, we infer $\mathbf{E}[XY] = \mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]Y]$ and $\mathbf{E}\big[X\mathbf{E}[X\,|\,\mathcal{F}]\big] = \mathbf{E}\big[\mathbf{E}[X\mathbf{E}[X\,|\,\mathcal{F}]\,\big|\,\mathcal{F}]\big] = \mathbf{E}\big[\mathbf{E}[X\,|\,\mathcal{F}]^2\big]$. Summing up, we have

$$\begin{aligned}
&\mathbf{E}\big[(X-Y)^2\big] - \mathbf{E}\Big[\big(X - \mathbf{E}[X\,|\,\mathcal{F}]\big)^2\Big] \\
&\qquad = \mathbf{E}\Big[X^2 - 2XY + Y^2 - X^2 + 2X\mathbf{E}[X\,|\,\mathcal{F}] - \mathbf{E}[X\,|\,\mathcal{F}]^2\Big] \\
&\qquad = \mathbf{E}\Big[Y^2 - 2Y\,\mathbf{E}[X\,|\,\mathcal{F}] + \mathbf{E}[X\,|\,\mathcal{F}]^2\Big] \\
&\qquad = \mathbf{E}\Big[\big(Y - \mathbf{E}[X\,|\,\mathcal{F}]\big)^2\Big] \geq 0.
\end{aligned}$$

For the case $\mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]^2] < \infty$, we are done. Hence, it suffices to show that this condition follows from the assumption $\mathbf{E}[X^2] < \infty$. For $N \in \mathbb{N}$, define the truncated random variables $|X| \wedge N$. Clearly, we have $\mathbf{E}[\mathbf{E}[|X| \wedge N\,|\,\mathcal{F}]^2] \leq N^2$. By what we have shown already (with X replaced by $|X| \wedge N$ and with $Y = 0 \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$), and using the elementary inequality $a^2 \leq 2(a-b)^2 + 2b^2$, $a, b \in \mathbb{R}$, we infer

$$\begin{aligned}
\mathbf{E}\Big[\mathbf{E}\big[|X| \wedge N\,\big|\,\mathcal{F}\big]^2\Big] &\leq 2\mathbf{E}\Big[\big((|X| \wedge N) - \mathbf{E}[|X| \wedge N\,\big|\,\mathcal{F}]\big)^2\Big] + 2\mathbf{E}\big[(|X| \wedge N)^2\big] \\
&\leq 4\mathbf{E}\big[(|X| \wedge N)^2\big] \leq 4\mathbf{E}[X^2].
\end{aligned}$$

By Theorem 8.14(ii) and (viii), we get $\mathbf{E}[|X| \wedge N\,\big|\,\mathcal{F}] \uparrow \mathbf{E}[|X|\,\big|\,\mathcal{F}]$ for $N \to \infty$. By the triangle inequality (Theorem 8.14(v)) and the monotone convergence theorem (Theorem 4.20), we conclude

$$\mathbf{E}\big[\mathbf{E}[X\,|\,\mathcal{F}]^2\big] \leq \mathbf{E}\big[\mathbf{E}[|X|\,\big|\,\mathcal{F}]^2\big] = \lim_{N\to\infty} \mathbf{E}\big[\mathbf{E}[|X| \wedge N\,\big|\,\mathcal{F}]^2\big] \leq 4\mathbf{E}[X^2] < \infty.$$

This completes the proof. $\square$

Example 8.17. Let $X, Y \in \mathcal{L}^1(\mathbf{P})$ be independent. Then

$$\mathbf{E}[X+Y\,|\,Y] \;=\; \mathbf{E}[X\,|\,Y] + \mathbf{E}[Y\,|\,Y] \;=\; \mathbf{E}[X] + Y. \qquad \diamond$$

Example 8.18. Let $X_1, \ldots, X_N$ be independent with $\mathbf{E}[X_i] = 0$, $i = 1, \ldots, N$. For $n = 1, \ldots, N$, define $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$ and $S_n := X_1 + \ldots + X_n$. Then, for $n \geq m$,

$$\begin{aligned} \mathbf{E}[S_n \,|\, \mathcal{F}_m] &= \mathbf{E}[X_1 \,|\, \mathcal{F}_m] + \ldots + \mathbf{E}[X_n \,|\, \mathcal{F}_m] \\ &= X_1 + \ldots + X_m + \mathbf{E}[X_{m+1}] + \ldots + \mathbf{E}[X_n] \\ &= S_m. \end{aligned}$$

By Theorem 8.14(iv), since $\sigma(S_m) \subset \mathcal{F}_m$, we have

$$\mathbf{E}[S_n \,|\, S_m] \;=\; \mathbf{E}\big[\mathbf{E}[S_n \,|\, \mathcal{F}_m] \,\big|\, S_m\big] \;=\; \mathbf{E}[S_m \,|\, S_m] \;=\; S_m. \qquad \diamond$$

Next we show Jensen's inequality for conditional expectations.

Theorem 8.19 (Jensen's inequality). *Let $I \subset \mathbb{R}$ be an interval, let $\varphi : I \to \mathbb{R}$ be convex and let X be an I-valued random variable on $(\Omega, \mathcal{A}, \mathbf{P})$. Further, let $\mathbf{E}[|X|] < \infty$ and let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra. Then*

$$\infty \;\geq\; \mathbf{E}[\varphi(X) \,|\, \mathcal{F}] \;\geq\; \varphi(\mathbf{E}[X \,|\, \mathcal{F}]).$$

Proof. (Recall from Definition 1.68 the jargon words "almost surely on A".) Note that $X = \mathbf{E}[X \,|\, \mathcal{F}]$ on the event $\{\mathbf{E}[X \,|\, \mathcal{F}]$ is a boundary point of $I\}$; hence here the claim is trivial. Indeed, without loss of generality, assume 0 is the left boundary of I and $A := \{\mathbf{E}[X \,|\, \mathcal{F}] = 0\}$. As X assumes values in $I \subset [0, \infty)$, we have $0 \leq \mathbf{E}[X \, \mathbb{1}_A] = \mathbf{E}[\mathbf{E}[X \,|\, \mathcal{F}] \, \mathbb{1}_A] = 0$; hence $X \mathbb{1}_A = 0$. The case of a right boundary point is similar.

Hence now consider the event $B := \{\mathbf{E}[X \,|\, \mathcal{F}]$ is an interior point of $I\}$. For every interior point $x \in I$, let $D^+\varphi(x)$ be the maximal slope of a tangent of φ at x; i.e., the maximal number t with $\varphi(y) \geq (y - x)t + \varphi(x)$ for all $y \in I$ (see Theorem 7.7).

For each $x \in I^\circ$, there exists a $\mathbf{P}$-null set N_x such that, for every $\omega \in B \setminus N_x$, we have

$$\begin{aligned} \mathbf{E}\big[\varphi(X) \,|\, \mathcal{F}\big](\omega) &\geq \varphi(x) + \mathbf{E}\big[D^+\varphi(x)\,(X - x) \,\big|\, \mathcal{F}\big](\omega) \\ &= \varphi(x) + D^+\varphi(x)\,\big(\mathbf{E}[X \,|\, \mathcal{F}](\omega) - x\big) =: \psi_\omega(x). \end{aligned} \tag{8.8}$$

Let $V := \mathbb{Q} \cap I^\circ$. Then $N := \bigcup_{x \in V} N_x$ is a $\mathbf{P}$-null set and (8.8) holds for every $\omega \in B \setminus N$ and every $x \in V$.

The map $x \mapsto D^+\varphi(x)$ is right continuous (by Theorem 7.7(iv)). Therefore $x \mapsto \psi_\omega(x)$ is also right continuous. Hence, for every $\omega \in B \setminus N$, we have

$$\begin{aligned} \varphi\big(\mathbf{E}[X \,|\, \mathcal{F}](\omega)\big) &= \psi_\omega\big(\mathbf{E}[X \,|\, \mathcal{F}](\omega)\big) \\ &\leq \sup_{x \in I^\circ} \psi_\omega(x) \;=\; \sup_{x \in V} \psi_\omega(x) \;\leq\; \mathbf{E}\big[\varphi(X) \,|\, \mathcal{F}\big](\omega). \end{aligned}$$

$\square$

Corollary 8.20. *Let $p \in [1,\infty]$ and let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra. Then the map*

$$\mathcal{L}^p(\Omega,\mathcal{A},\mathbf{P}) \to \mathcal{L}^p(\Omega,\mathcal{F},\mathbf{P}), \qquad X \mapsto \mathbf{E}[X\,|\,\mathcal{F}],$$

is a contraction (that is, $\|\mathbf{E}[X\,|\,\mathcal{F}]\|_p \le \|X\|_p$) and thus continuous. Hence, for $X, X_1, X_2, \ldots \in \mathcal{L}^p(\Omega,\mathcal{A},\mathbf{P})$ with $\|X_n - X\|_p \overset{n\to\infty}{\longrightarrow} 0$,

$$\big\|\mathbf{E}[X_n\,|\,\mathcal{F}] - \mathbf{E}[X\,|\,\mathcal{F}]\big\|_p \overset{n\to\infty}{\longrightarrow} 0.$$

Proof. For $p \in [1,\infty)$, use Jensen's inequality with $\varphi(x) = |x|^p$. For $p = \infty$, note that $|\mathbf{E}[X\,|\,\mathcal{F}]| \le \mathbf{E}[|X|\,|\,\mathcal{F}] \le \mathbf{E}[\|X\|_\infty \,\big|\, \mathcal{F}] = \|X\|_\infty$. □

Corollary 8.21. *Let $(X_i,\ i \in I)$ be uniformly integrable and let $(\mathcal{F}_j,\ j \in J)$ be a family of sub-σ-algebras of $\mathcal{A}$. Define $X_{i,j} := \mathbf{E}[X_i \,\big|\, \mathcal{F}_j]$. Then $(X_{i,j},\ (i,j) \in I\times J)$ is uniformly integrable. In particular, for $X \in \mathcal{L}^1(\mathbf{P})$, the family $(\mathbf{E}[X\,|\,\mathcal{F}_j],\ j \in J)$ is uniformly integrable.*

Proof. By Theorem 6.19, there exists a monotone increasing convex function f with the property that $f(x)/x \to \infty$, $x \to \infty$ and $L := \sup_{i\in I} \mathbf{E}[f(|X_i|)] < \infty$. Then $x \mapsto f(|x|)$ is convex; hence, by Jensen's inequality,

$$\mathbf{E}\big[f(|X_{i,j}|)\big] \;=\; \mathbf{E}\big[f\big(\big|\mathbf{E}[X_i\,|\,\mathcal{F}_j]\big|\big)\big] \;\le\; L \;<\; \infty.$$

Thus $(X_{i,j},\ (i,j) \in I \times J)$ is uniformly integrable by Theorem 6.19. □

Example 8.22. Let μ and ν be finite measures with $\nu \ll \mu$. Let $f = d\nu/d\mu$ be the Radon-Nikodym derivative and let $I = \{\mathcal{F} \subset \mathcal{A} :\ \mathcal{F} \text{ is a } \sigma\text{-algebra}\}$. Consider the measures $\mu\big|_{\mathcal{F}}$ and $\nu\big|_{\mathcal{F}}$ that are restricted to $\mathcal{F}$. Then $\nu\big|_{\mathcal{F}} \ll \mu\big|_{\mathcal{F}}$ (since in $\mathcal{F}$ there are fewer μ-null sets); hence the Radon-Nikodym derivative $f_{\mathcal{F}} := d\nu\big|_{\mathcal{F}}\big/d\mu\big|_{\mathcal{F}}$ exists. Then $(f_{\mathcal{F}} :\ \mathcal{F} \in I)$ is uniformly integrable (with respect to μ). (For finite σ-algebras $\mathcal{F}$, this was shown in Example 7.39.) Indeed, let $\mathbf{P} = \mu/\mu(\Omega)$ and $\mathbf{Q} = \nu/\mu(\Omega)$. Then $f_{\mathcal{F}} = d\mathbf{Q}\big|_{\mathcal{F}}\big/d\mathbf{P}\big|_{\mathcal{F}}$. For any $F \in \mathcal{F}$, we thus have $\mathbf{E}[f_{\mathcal{F}}\,\mathbb{1}_F] = \int_F f_{\mathcal{F}}\,d\mathbf{P} = \mathbf{Q}(F) = \int_F f\,d\mathbf{P} = \mathbf{E}[f\,\mathbb{1}_F]$; hence $f_{\mathcal{F}} = \mathbf{E}[f\,|\,\mathcal{F}]$. By the preceding corollary, $(f_{\mathcal{F}} :\ \mathcal{F} \in I)$ is uniformly integrable with respect to $\mathbf{P}$ and thus also with respect to μ. ◇

Exercise 8.2.1 (Bayes' formula). Let $A \in \mathcal{A}$ and $B \in \mathcal{F}$. Show that

$$\mathbf{P}[B\,|\,A] \;=\; \frac{\int_B \mathbf{P}[A\,|\,\mathcal{F}]\,d\mathbf{P}}{\int \mathbf{P}[A\,|\,\mathcal{F}]\,d\mathbf{P}}.$$

If $\mathcal{F}$ is generated by pairwise disjoint sets $B_1, B_2, \ldots$, then this is exactly Bayes' formula of Theorem 8.7. ♣

Exercise 8.2.2. Give an example for $\mathbf{E}[\mathbf{E}[X\,|\,\mathcal{F}]\,|\,\mathcal{G}] \neq \mathbf{E}[\mathbf{E}[X\,|\,\mathcal{G}]\,|\,\mathcal{F}]$. ♣

Exercise 8.2.3. Show the conditional Markov inequality: For monotone increasing $f : [0, \infty) \to [0, \infty)$ and $\varepsilon > 0$ with $f(\varepsilon) > 0$,

$$\mathbf{P}\big[|X| \geq \varepsilon \,|\, \mathcal{F}\big] \;\leq\; \frac{\mathbf{E}\big[f(|X|)\,\big|\,\mathcal{F}\big]}{f(\varepsilon)}. \qquad \clubsuit$$

Exercise 8.2.4. Show the conditional Cauchy-Schwarz inequality: For square integrable random variables X, Y,

$$\mathbf{E}[XY \,|\, \mathcal{F}]^2 \;\leq\; \mathbf{E}[X^2 \,|\, \mathcal{F}]\; \mathbf{E}[Y^2 \,|\, \mathcal{F}]. \qquad \clubsuit$$

Exercise 8.2.5. Let $X_1, \ldots, X_n$ be integrable i.i.d. random variables. Let $S_n = X_1 + \ldots + X_n$. Show that

$$\mathbf{E}[X_i \,|\, S_n] \;=\; \frac{1}{n}\, S_n \quad \text{ for every } \; i = 1, \ldots, n. \qquad \clubsuit$$

Exercise 8.2.6. Let X_1 and X_2 be independent and exponentially distributed with parameter $\theta > 0$. Compute $\mathbf{E}[X_1 \wedge X_2 \,|\, X_1]$. $\clubsuit$

Exercise 8.2.7. Let X and Y be real random variables with joint density f and let $h : \mathbb{R} \to \mathbb{R}$ be measurable with $\mathbf{E}[|h(X)|] < \infty$. Denote by λ the Lebesgue measure on $\mathbb{R}$.

(i) Show that almost surely

$$\mathbf{E}[h(X)|Y] \;=\; \frac{\int h(x) f(x, Y)\, \lambda(dx)}{\int f(x, Y)\, \lambda(dx)}.$$

(ii) Let X and Y be independent and $\exp_\theta$-distributed for some $\theta > 0$. Compute $\mathbf{E}[X \,|\, X + Y]$ and $\mathbf{P}[X \leq x \,|\, X + Y]$ for $x \geq 0$. $\clubsuit$

8.3 Regular Conditional Distribution

Let X be a random variable with values in a measurable space $(E, \mathcal{E})$. With our machinery, so far we can define the conditional probability $\mathbf{P}[A \,|\, X]$ for *fixed* $A \in \mathcal{A}$ only. However, we would like to define *for every* $x \in E$ a probability measure $\mathbf{P}[\,\cdot\,|\, X = x]$ such that for any $A \in \mathcal{A}$, we have $\mathbf{P}[A \,|\, X] = \mathbf{P}[A \,|\, X = x]$ on $\{X = x\}$. In this section, we show how to do this.

For example, we are interested in a two-stage random experiment. At the first stage, we manipulate a coin *at random* such that the probability of a success (i.e., "head") is X. At the second stage, we toss the coin n times independently with outcomes $Y_1, \ldots, Y_n$. Hence the "conditional distribution of $(Y_1, \ldots, Y_n)$ given $\{X = x\}$" should be $(\mathrm{Ber}_x)^{\otimes n}$.

Let X be as above and let Z be a $\sigma(X)$-measurable real random variable. By the factorisation lemma (Corollary 1.97 with $f = X$ and $g = Z$), there is a map $\varphi : E \to \mathbb{R}$ such that

$$\varphi \text{ is } \mathcal{E}\text{–}\mathcal{B}(\mathbb{R})\text{-measurable} \quad \text{and} \quad \varphi(X) = Z. \tag{8.9}$$

If X is surjective, then φ is determined uniquely. In this case, we denote $Z \circ X^{-1} := \varphi$ (even if the inverse map X^{-1} itself does not exist).

Definition 8.23. *Let $Y \in \mathcal{L}^1(\mathbf{P})$ and $X : (\Omega, \mathcal{A}) \to (E, \mathcal{E})$. We define the conditional expectation of Y given $X = x$ by $\mathbf{E}[Y \,|\, X = x] := \varphi(x)$, where φ is the function from* (8.9) *with $Z = \mathbf{E}[Y \,|\, X]$.*

Analogously, define $\mathbf{P}[A \,|\, X = x] = \mathbf{E}[\mathbb{1}_A \,\big|\, X = x]$ for $A \in \mathcal{A}$.

For a fixed set $B \in \mathcal{A}$ with $\mathbf{P}[B] > 0$, the conditional probability $\mathbf{P}[\,\cdot\,|\,B]$ is a probability measure. Is this true also for $\mathbf{P}[\,\cdot\,|\,X = x]$? The question is a bit tricky since for every given $A \in \mathcal{A}$, the expression $\mathbf{P}[A|X = x]$ is defined for almost all x only; that is, up to x in a null set that may, however, depend on A. Since there are uncountably many $A \in \mathcal{A}$ in general, we could not simply unite all the exceptional sets for any A. However, if the σ-algebra $\mathcal{A}$ can be approximated by countably many A sufficiently well, then there is hope.

Our first task is to give precise definitions. Then we present the theorem that justifies our hope.

Definition 8.24 (Transition kernel, Markov kernel). *Let $(\Omega_1, \mathcal{A}_1)$, $(\Omega_2, \mathcal{A}_2)$ be measurable spaces. A map $\kappa : \Omega_1 \times \mathcal{A}_2 \to [0, \infty]$ is called a (σ-)finite **transition kernel** (from Ω_1 to Ω_2) if:*

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

(ii) $A_2 \mapsto \kappa(\omega_1, A_2)$ is a (σ-)finite measure on $(\Omega_2, \mathcal{A}_2)$ for any $\omega_1 \in \Omega_1$.

*If in (ii) the measure is a probability measure for all $\omega_1 \in \Omega_1$, then κ is called a **stochastic kernel** or a **Markov kernel**. If in (ii) we also have $\kappa(\omega_1, \Omega_2) \leq 1$ for any $\omega_1 \in \Omega_1$, then κ is called sub-Markov or substochastic.*

Remark 8.25. It is sufficient to check property (i) in Definition 8.24 for sets A_2 from a π-system $\mathcal{E}$ that generates $\mathcal{A}_2$ and that either contains Ω_2 or a sequence $E_n \uparrow \Omega_2$. Indeed, in this case,

$$\mathcal{D} := \big\{A_2 \in \mathcal{A}_2 : \omega_1 \mapsto \kappa(\omega_1, A_2) \text{ is } \mathcal{A}_1\text{-measurable}\big\}$$

is a λ-system (exercise!). Since $\mathcal{E} \subset \mathcal{D}$, by the π–λ theorem (Theorem 1.19), $\mathcal{D} = \sigma(\mathcal{E}) = \mathcal{A}_2$. ◇

Example 8.26. (i) Let $(\Omega_1, \mathcal{A}_1)$ and $(\Omega_2, \mathcal{A}_2)$ be discrete measurable spaces and let $(K_{ij})_{\substack{i\in\Omega_1\\ j\in\Omega_2}}$ be a matrix with nonnegative entries and finite row sums

$$K_i := \sum_{j\in\Omega_2} K_{ij} < \infty \quad \text{for } i \in \Omega_1.$$

Then we can define a finite transition kernel from Ω_1 to Ω_2 by $\kappa(i, A) = \sum_{j\in A} K_{ij}$. This kernel is stochastic if $K_i = 1$ for all $i \in \Omega_1$. It is substochastic if $K_i \le 1$ for all $i \in \Omega_1$.

(ii) If μ_2 is a finite measure on Ω_2, then $\kappa(\omega_1, \cdot\,) \equiv \mu_2$ is a finite transition kernel.

(iii) $\kappa(x, \cdot\,) = \mathrm{Poi}_x$ is a stochastic kernel from $[0, \infty)$ to $\mathbb{N}_0$ (note that $x \mapsto \mathrm{Poi}_x(A)$ is continuous and hence measurable for all $A \subset \mathbb{N}_0$).

(iv) Let μ be a distribution on $\mathbb{R}^n$ and let X be a random variable with $\mathbf{P}_X = \mu$. Then $\kappa(x, \cdot\,) = \mathbf{P}[X + x \in \cdot\,] = \delta_x * \mu$ defines a stochastic kernel from $\mathbb{R}^n$ to $\mathbb{R}^n$. Indeed, the sets $(-\infty, y]$, $y \in \mathbb{R}^n$ form an $\cap$-stable generator of $\mathcal{B}(\mathbb{R}^n)$ and $x \mapsto \kappa(x, (-\infty, y]) = \mu((-\infty, y - x])$ is left continuous and hence measurable. Hence, by Remark 8.25, $x \mapsto \kappa(x, A)$ is measurable for all $A \in \mathcal{B}(\mathbb{R}^n)$. $\diamond$

Definition 8.27. *Let Y be a random variable with values in a measurable space $(E, \mathcal{E})$ and let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra. A stochastic kernel $\kappa_{Y,\mathcal{F}}$ from $(\Omega, \mathcal{F})$ to $(E, \mathcal{E})$ is called a* ***regular conditional distribution*** *of Y given $\mathcal{F}$ if*

$$\kappa_{Y,\mathcal{F}}(\omega, B) = \mathbf{P}[\{Y \in B\}\,|\,\mathcal{F}](\omega)$$

for $\mathbf{P}$-almost all $\omega \in \Omega$ and for all $B \in \mathcal{E}$.

Consider the special case where $\mathcal{F} = \sigma(X)$ for a random variable X (with values in an arbitrary measurable space $(E', \mathcal{E}')$). Then the stochastic kernel

$$(x, A) \mapsto \kappa_{Y,X}(x, A) \;=\; \mathbf{P}[\{Y \in A\}\,|\,X = x] \;=\; \kappa_{Y,\sigma(X)}\big(X^{-1}(x), A\big)$$

(the function from the factorisation lemma with an arbitrary value for $x \notin X(\Omega)$) is called a regular conditional distribution of Y given X.

Theorem 8.28 (Regular conditional distributions in $\mathbb{R}$). *Let $Y : (\Omega, \mathcal{A}) \to \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ be real-valued. Then there exists a regular conditional distribution $\kappa_{Y,\mathcal{F}}$ of Y given $\mathcal{F}$.*

Proof. The strategy of the proof consists in constructing a measurable version of the distribution function of the conditional distribution of Y by first defining it for rational values (up to a null set) and then extending it to the real numbers.

For $r \in \mathbb{Q}$, let $F(r, \cdot\,)$ be a version of the conditional probability $\mathbf{P}[Y \in (-\infty, r]\,|\,\mathcal{F}]$. For $r \le s$, clearly $\mathbb{1}_{\{Y\in(-\infty,r]\}} \le \mathbb{1}_{\{Y\in(-\infty,s]\}}$. Hence, by Theorem 8.14(ii) (monotonicity of the conditional expectation), there is a null set $A_{r,s} \in \mathcal{F}$ with

$$F(r,\omega) \le F(s,\omega) \quad \text{for all } \omega \in \Omega \setminus A_{r,s}. \tag{8.10}$$

By Theorem 8.14(viii) (dominated convergence), there are null sets $(B_r)_{r\in\mathbb{Q}} \in \mathcal{F}$ and $C \in \mathcal{F}$ such that

$$\lim_{n\to\infty} F\left(r+\frac{1}{n},\omega\right) = F(r,\omega) \quad \text{for all } \omega \in \Omega \setminus B_r \tag{8.11}$$

as well as

$$\inf_{n\in\mathbb{N}} F(-n,\omega) = 0 \quad \text{and} \quad \sup_{n\in\mathbb{N}} F(n,\omega) = 1 \quad \text{for all } \omega \in \Omega \setminus C. \tag{8.12}$$

Let $N := \big(\bigcup_{r,s\in\mathbb{Q}} A_{r,s}\big) \cup \big(\bigcup_{r\in\mathbb{Q}} B_r\big) \cup C$. For $\omega \in \Omega \setminus N$, define

$$\tilde{F}(z,\omega) := \inf\big\{F(r,\omega) :\ r \in \mathbb{Q},\, r > z\big\} \quad \text{for all } z \in \mathbb{R}.$$

By construction, $\tilde{F}(\,\cdot\,,\omega)$ is monotone increasing and right continuous. By (8.10) and (8.11), we have

$$\tilde{F}(z,\omega) = F(z,\omega) \quad \text{for all } z \in \mathbb{Q} \text{ and } \omega \in \Omega \setminus N. \tag{8.13}$$

Therefore, by (8.12), $\tilde{F}(\,\cdot\,,\omega)$ is a distribution function for any $\omega \in \Omega \setminus N$. For $\omega \in N$, define $\tilde{F}(\,\cdot\,,\omega) = F_0$, where F_0 is an arbitrary but fixed distribution function.

For any $\omega \in \Omega$, let $\kappa(\omega,\,\cdot\,)$ be the probability measure on $(\Omega,\mathcal{A})$ with distribution function $\tilde{F}(\,\cdot\,,\omega)$. Then, for $r \in \mathbb{Q}$ and $B = (-\infty, r]$,

$$\omega \mapsto \kappa(\omega,B) = F(r,\omega)\,\mathbb{1}_{N^c}(\omega) + F_0(r)\,\mathbb{1}_N(\omega) \tag{8.14}$$

is $\mathcal{F}$-measurable. Now $\{(-\infty,r],\, r \in \mathbb{Q}\}$ is a π-system that generates $\mathcal{B}(\mathbb{R})$. By Remark 8.25, measurability holds for all $B \in \mathcal{B}(\mathbb{R})$ and hence κ is identified as a stochastic kernel.

We still have to show that κ is a version of the conditional distribution. For $A \in \mathcal{F}$, $r \in \mathbb{Q}$ and $B = (-\infty, r]$, by (8.14),

$$\int_A \kappa(\omega,B)\,\mathbf{P}[d\omega] \;=\; \int_A \mathbf{P}\big[Y \in B\,|\,\mathcal{F}\big]\,d\mathbf{P} \;=\; \mathbf{P}\big[A \cap \{Y \in B\}\big].$$

As functions of B, both sides are finite measures on $\mathcal{B}(\mathbb{R})$ that coincide on the $\cap$-stable generator $\big\{(-\infty,r],\, r \in \mathbb{Q}\big\}$. By the uniqueness theorem (Lemma 1.42), we thus have equality for all $B \in \mathcal{B}(\mathbb{R})$. Hence $\mathbf{P}$-a.s. $\kappa(\,\cdot\,,B) = \mathbf{P}[Y \in B\,|\,\mathcal{F}]$ and thus $\kappa = \kappa_{Y,\mathcal{F}}$. □

Example 8.29. Let Z_1, Z_2 be independent Poisson random variables with parameters $\lambda_1, \lambda_2 \ge 0$. One can show (exercise!) that (with $Y = Z_1$ and $X = Z_1 + Z_2$)

$$\mathbf{P}[Z_1 = k \,\big|\, Z_1 + Z_2 = n] = b_{n,p}(k) \quad \text{for } k = 0,\dots,n,$$

where $p = \frac{\lambda_1}{\lambda_1+\lambda_2}$. ◇

This example could still be treated by elementary means. The full strength of the result is displayed in the following examples.

Example 8.30. Let X and Y be real random variables with joint density f (with respect to Lebesgue measure λ^2 on $\mathbb{R}^2$). For $x \in \mathbb{R}$, define

$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,\lambda(dy).$$

Clearly, $f_X(x) > 0$ for $\mathbf{P}_X$-a.a. $x \in \mathbb{R}$ and f_X^{-1} is the density of the absolutely continuous part of the Lebesgue measure λ with respect to $\mathbf{P}_X$. The regular conditional distribution of Y given X has density

$$\frac{\mathbf{P}[Y \in dy \,|\, X = x]}{dy} \;=\; f_{Y|X}(x,y) \;:=\; \frac{f(x,y)}{f_X(x)} \quad \text{for } \mathbf{P}_X[dx]\text{-a.a. } x \in \mathbb{R}. \tag{8.15}$$

Indeed, by Fubini's theorem (Theorem 14.16), the map $x \mapsto \int_B f_{Y|X}(x,y)\,\lambda(dy)$ is measurable for all $B \in \mathcal{B}(\mathbb{R})$ and for $A, B \in \mathcal{B}(\mathbb{R})$, we have

$$\begin{aligned}
&\int_A \mathbf{P}[X \in dx] \int_B f_{Y|X}(x,y)\,\lambda(dy) \\
&\qquad = \int_A \mathbf{P}[X \in dx]\, f_X(x)^{-1} \int_B f(x,y)\,\lambda(dy) \\
&\qquad = \int_A \lambda(dx) \int_B f(x,y)\,\lambda(dy) \\
&\qquad = \int_{A\times B} f\, d\lambda^2 = \mathbf{P}[X \in A,\, Y \in B]. \qquad \diamond
\end{aligned}$$

Example 8.31. Let $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$ and let Z_1, Z_2 be independent and $\mathcal{N}_{\mu_i,\sigma_i^2}$-distributed ($i = 1, 2$). Then there exists a regular conditional distribution

$$\mathbf{P}[Z_1 \in \cdot \,|\, Z_1 + Z_2 = x] \quad \text{for } x \in \mathbb{R}.$$

If we define $X = Z_1 + Z_2$ and $Y = Z_1$, then $(X,Y) \sim \mathcal{N}_{\mu,\Sigma}$ is bivariate normally distributed with covariance matrix $\Sigma := \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix}$ and with $\mu := \begin{pmatrix} \mu_1 + \mu_2 \\ \mu_1 \end{pmatrix}$. Note that

$$\Sigma^{-1} \;=\; \left(\sigma_1^2\sigma_2^2\right)^{-1} \begin{pmatrix} \sigma_1^2 & -\sigma_1^2 \\ -\sigma_1^2 & \sigma_1^2 + \sigma_2^2 \end{pmatrix} \;=\; (\sigma_1^2\sigma_2^2)^{-1}\, B^T B,$$

where $B = \begin{pmatrix} \sigma_1 & -\sigma_1 \\ 0 & \sigma_2 \end{pmatrix}$. Hence (X,Y) has the density (see Example 1.105(ix))

$$\begin{aligned} f(x,y) &= \det(2\pi\,\Sigma)^{-1/2} \exp\left(-\frac{1}{2\sigma_1^2\sigma_2^2} \left\| B \begin{pmatrix} x-(\mu_1+\mu_2) \\ y-\mu_1 \end{pmatrix} \right\|^2 \right) \\ &= \left(4\pi^2\sigma_1^2\sigma_2^2\right)^{-1/2} \exp\left(-\frac{\sigma_1^2(y-(x-\mu_1))^2 + \sigma_2^2(y-\mu_2)^2}{2\sigma_1^2\sigma_2^2} \right) \\ &= C_x \ \exp\big(-(y-\mu_x)^2/2\sigma_x^2 \big). \end{aligned}$$

Here C_x is a normalising constant and

$$\mu_x = \mu_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}(x-\mu_1-\mu_2) \quad \text{and} \quad \sigma_x^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}.$$

By (8.15), $\mathbf{P}[Z_1 \in \cdot \,|\, Z_1+Z_2=x]$ has the density

$$y \mapsto f_{Y|X}(x,y) = \frac{C_x}{f_X(x)} \ \exp\left(-\frac{(y-\mu_x)^2}{2\sigma_x^2} \right),$$

hence

$$\mathbf{P}[Z_1 \in \cdot \,|\, Z_1+Z_2=x] = \mathcal{N}_{\mu_x,\sigma_x^2} \ \text{ for almost all } \ x \in \mathbb{R}. \qquad \diamond$$

Example 8.32. If X and Y are independent real random variables, then for $\mathbf{P}_X$-almost all $x \in \mathbb{R}$

$$\mathbf{P}[X+Y \in \cdot \,|\, X=x] \ = \ \delta_x * \mathbf{P}_Y. \qquad \diamond$$

The situation is not completely satisfying as we have made the very restrictive assumption that Y is real-valued. Originally we were also interested in the situation where Y takes values in $\mathbb{R}^n$ or in even more general spaces. We now extend the result to a larger class of ranges for Y.

Definition 8.33. *Two measurable spaces $(E,\mathcal{E})$ and $(E',\mathcal{E}')$ are called **isomorphic** if there exists a bijective map $\varphi : E \to E'$ such that φ is $\mathcal{E}$ – $\mathcal{E}'$-measurable and the inverse map φ^{-1} is $\mathcal{E}'$ – $\mathcal{E}$-measurable. Then we say that φ is an isomorphism of measurable spaces. If in addition μ and μ' are measures on $(E,\mathcal{E})$ and $(E',\mathcal{E}')$ and if $\mu' = \mu \circ \varphi^{-1}$, then φ is an isomorphism of measure spaces, and the measure spaces $(E,\mathcal{E},\mu)$ and $(E',\mathcal{E}',\mu')$ are called isomorphic.*

Definition 8.34. *A measurable space $(E,\mathcal{E})$ is called a **Borel space** if there exists a Borel set $B \in \mathcal{B}(\mathbb{R})$ such that $(E,\mathcal{E})$ and $(B,\mathcal{B}(B))$ are isomorphic.*

A separable topological space whose topology is induced by a complete metric is called a **Polish space**. In particular, $\mathbb{R}^d$, $\mathbb{Z}^d$, $\mathbb{R}^{\mathbb{N}}$, $(C([0,1]), \|\cdot\|_\infty)$ and so forth are Polish. Closed subsets of Polish spaces are again Polish. We come back to Polish spaces in the context of convergence of measures in Chapter 13. Without proof, we present the following topological result (see, e.g., [35, Theorem 13.1.1]).

Theorem 8.35. *Let E be a Polish space with Borel σ-algebra $\mathcal{E}$. Then $(E,\mathcal{E})$ is a Borel space.*

Theorem 8.36 (Regular conditional distribution). *Let $\mathcal{F} \subset \mathcal{A}$ be a sub-σ-algebra. Let Y be a random variable with values in a Borel space $(E, \mathcal{E})$ (hence, for example, E Polish, $E = \mathbb{R}^d$, $E = \mathbb{R}^\infty$, $E = C([0,1])$, etc.). Then there exists a regular conditional distribution $\kappa_{Y,\mathcal{F}}$ of Y given $\mathcal{F}$.*

Proof. Let $B \in \mathcal{B}(\mathbb{R})$ and let $\varphi : E \to B$ be an isomorphism of measurable spaces. By Theorem 8.28, we obtain the regular conditional distribution $\kappa_{Y',\mathcal{F}}$ of the real random variable $Y' = \varphi \circ Y$. Now define $\kappa_{Y,\mathcal{F}}(\omega, A) = \kappa_{Y',\mathcal{F}}(\omega, \varphi(A))$ for $A \in \mathcal{E}$. □

To conclude, we pick up again the example with which we started. Now we can drop the quotation marks from the statement and write it down formally. Hence, let X be uniformly distributed on $[0,1]$. Given $X = x$, let $(Y_1, \dots, Y_n)$ be independent and Ber_x-distributed. Define $Y = (Y_1, \dots, Y_n)$. By Theorem 8.36 (with $E = \{0,1\}^n \subset \mathbb{R}^n$), a regular conditional distribution exists:

$$\kappa_{Y,X}(x, \cdot) \;=\; \mathbf{P}[Y \in \cdot \,|\, X = x] \quad \text{for } x \in [0,1].$$

Indeed, for almost all $x \in [0,1]$,

$$\mathbf{P}[Y \in \cdot \,|\, X = x] \;=\; (\mathrm{Ber}_x)^{\otimes n}.$$

Theorem 8.37. *Let X be a random variable on $(\Omega, \mathcal{A}, \mathbf{P})$ with values in a Borel space $(E, \mathcal{E})$. Let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra and let $\kappa_{X,\mathcal{F}}$ be a regular conditional distribution of X given $\mathcal{F}$. Further, let $f : E \to \mathbb{R}$ be measurable and $\mathbf{E}[|f(X)|] < \infty$. Then*

$$\mathbf{E}[f(X)\,|\,\mathcal{F}](\omega) = \int f(x)\, \kappa_{Y,\mathcal{F}}(\omega, dx) \quad \textit{for } \mathbf{P}\textit{-almost all } \omega. \tag{8.16}$$

Proof. We check that the right hand side in (8.16) has the properties of the conditional expectation.

It is enough to consider the case $f \geq 0$. By approximating f by simple functions, we see that the right hand side in (8.16) is $\mathcal{F}$-measurable (see Lemma 14.20 for a formal argument). Hence, by Theorem 1.96, there exist sets $A_1, A_2, \dots \in \mathcal{E}$ and numbers $\alpha_1, \alpha_2, \dots \geq 0$ such that

$$g_n := \sum_{i=1}^{n} \alpha_i \, \mathbb{1}_{A_i} \;\overset{n\to\infty}{\longrightarrow}\; f.$$

Now, for any $n \in \mathbb{N}$ and $B \in \mathcal{F}$,

$$\begin{aligned}
\mathbf{E}[g_n(X)\,\mathbb{1}_B] &= \sum_{i=1}^{n} \alpha_i\, \mathbf{P}[\{X \in A_i\} \cap B] \\
&= \sum_{i=1}^{n} \alpha_i \int_B \mathbf{P}[\{X \in A_i\}\,|\,\mathcal{F}]\, \mathbf{P}[d\omega] \\
&= \sum_{i=1}^{n} \alpha_i \int_B \kappa_{X,\mathcal{F}}(\omega, A_i)\, \mathbf{P}[d\omega] \\
&= \int_B \sum_{i=1}^{n} \alpha_i\, \kappa_{X,\mathcal{F}}(\omega, A_i)\, \mathbf{P}[d\omega] \\
&= \int_B \left(\int g_n(x)\, \kappa_{X,\mathcal{F}}(\omega, dx) \right) \mathbf{P}[d\omega].
\end{aligned}$$

By the monotone convergence theorem, for almost all ω, the inner integral converges to $\int f(x)\kappa_{X,\mathcal{F}}(\omega, dx)$. Applying the monotone convergence theorem once more, we get

$$\mathbf{E}[f(X)\,\mathbb{1}_B] \;=\; \lim_{n\to\infty} \mathbf{E}[g_n(X)\,\mathbb{1}_B] \;=\; \int_B \int f(x)\, \kappa_{X,\mathcal{F}}(\omega, dx)\, \mathbf{P}[d\omega]. \qquad \square$$

Exercise 8.3.1. Let $(E, \mathcal{E})$ be a Borel space and let μ be an atom-free measure (that is, $\mu(\{x\}) = 0$ for any $x \in E$). Show that for any $A \in \mathcal{E}$ and any $n \in \mathbb{N}$, there exist pairwise disjoint sets $A_1, \ldots, A_n \in \mathcal{E}$ with $\biguplus_{k=1}^{n} A_k = A$ and $\mu(A_k) = \mu(A)/n$ for any $k = 1, \ldots, n$. ♣

Exercise 8.3.2. Let $p, q \in (1, \infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$ and let $X \in \mathcal{L}^p(\mathbf{P})$ and $Y \in \mathcal{L}^q(\mu)$. Let $\mathcal{F} \subset \mathcal{A}$ be a σ-algebra. Use the preceding theorem to show the conditional version of Hölder's inequality:

$$\mathbf{E}\big[|XY|\,\big|\,\mathcal{F}\big] \;\le\; \mathbf{E}\big[|X|^p\,\big|\,\mathcal{F}\big]^{1/p}\, \mathbf{E}\big[|Y|^q\,\big|\,\mathcal{F}\big]^{1/q} \quad \text{almost surely.} \qquad ♣$$

Exercise 8.3.3. Assume the random variable (X, Y) is uniformly distributed on the disc $B := \{(x, y) \in \mathbb{R}^2 :\, x^2 + y^2 \le 1\}$ and on $[-1, 1]^2$, respectively.

(i) In both cases, determine the conditional distribution of Y given $X = x$.

(ii) Let $R := \sqrt{X^2 + Y^2}$ and $\Theta = \arctan(Y/X)$. In both cases, determine the conditional distribution of Θ given $R = r$. ♣

Exercise 8.3.4. Let $A \subset \mathbb{R}^n$ be a Borel measurable set of finite Lebesgue measure $\lambda(A) \in (0, \infty)$ and let X be uniformly distributed on A (see Example 1.75). Let $B \subset A$ be measurable with $\lambda(B) > 0$. Show that the conditional distribution of X given $\{X \in B\}$ is the uniform distribution on B. ♣

Exercise 8.3.5 (Borel's paradox). Consider the earth as a ball (as widely accepted nowadays). Let X be a random point that is uniformly distributed on the surface. Let Θ be the longitude and let Φ be the latitude of X. A little differently from the usual convention, assume that Θ takes values in $[0, \pi)$ and Φ in $[-\pi, \pi)$. Hence, for fixed Θ, a complete great circle is described when Φ runs through its domain. Now, given Θ, is Φ uniformly distributed on $[-\pi, \pi)$? One could conjecture that any point on the great circle is equally likely. However, this is not the case! If we thicken the great circle slightly such that its longitudes range from Θ to $\Theta + \varepsilon$ (for a small ε), on the equator it is thicker (measured in metres) than at the poles. If we let $\varepsilon \to 0$, intuitively we should get the conditional probabilities as proportional to the thickness (in metres).

(i) Show that $\mathbf{P}[\{\Phi \in \cdot\} | \Theta = \theta]$ for almost all θ has the density $\frac{1}{4}|\cos(\phi)|$ for $\phi \in [-\pi, \pi)$.

(ii) Show that $\mathbf{P}[\{\Theta \in \cdot\} | \Phi = \phi] = \mathcal{U}_{[0,\pi)}$ for almost all ϕ.

Hint: Show that Θ and Φ are independent, and compute the distributions of Θ and Φ. ♣

Exercise 8.3.6 (Rejection sampling for generating random variables). Let E be a countable set and let P and Q be probability measures on E. Assume there is a $c > 0$ with

$$f(e) := \frac{Q(\{e\})}{P(\{e\})} \leq c \quad \text{for all } e \in E \text{ with } P(\{e\}) > 0.$$

Let $X_1, X_2, \ldots$ be independent random variables with distribution P. Let $U_1, U_2, \ldots$ be i.i.d. random variables that are independent of $X_1, X_2, \ldots$ and that are uniformly distributed on $[0, 1]$. Let N be the smallest (random) nonnegative integer n such that $U_n \leq f(X_n)/c$ and define $Y := X_N$.

Show that Y has distribution Q.

Remark. This method for generating random variables with a given distribution Q is called *rejection sampling*, as it can also be described as follows. The random variable X_1 is a proposal for the value of Y. This proposal is accepted with probability $f(X_1)/c$ and is rejected otherwise. If the first proposal is rejected, the game starts afresh with proposal X_2 and so on. ♣

Exercise 8.3.7. Let E be a Polish space and let $P, Q \in \mathcal{M}_1(\mathbb{R})$. Let $c > 0$ with $f := \frac{dQ}{dP} \leq c$ P-almost surely. Show the statement analogous to Exercise 8.3.6. ♣

Exercise 8.3.8. Show that $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\right)$ are isomorphic. Conclude that every Borel set $B \in \mathcal{B}(\mathbb{R}^n)$ is a Borel space. ♣