Chapter IV

Conditioning

This chapter is a continuation of Chapter II. We start with conditional expectations, namely, estimation of a random variable in the presence of partial information. Then, the notion of conditional independence follows as a natural generalization of independence, where the role played by the expectation operator is now played by the conditional expectation operator. Finally, we discuss various ways of constructing random variables and stochastic processes.

1 Conditional Expectations

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$, and X an $\bar{\mathbb{R}}$-valued random variable. As usual, we regard $\mathcal{F}$ both as a collection of events and as the collection of all $\mathcal{F}$-measurable random variables. Moreover, we think heuristically and regard $\mathcal{F}$ as a body of information, namely, the information that determines the values $V(\omega)$ for all V in $\mathcal{F}$ and all ω in Ω. Recall from Chapter II, Section 4, that knowing the value $V(\omega)$ is much less than knowing ω but, rather, it affords us to infer various properties of ω.

Heuristically, the conditional expectation of X given $\mathcal{F}$ is a random variable $\bar{X}$ such that, for every possibility ω, $\bar{X}(\omega)$ is the "best" estimate of $X(\omega)$ based on the information $\mathcal{F}$. So, $\bar{X}$ must be determined by $\mathcal{F}$, and among all random variables determined by $\mathcal{F}$ it must be the "closest" to X, the terms "best" and "closest" having a definite meaning at least when X is square integrable, that is, when $\mathbb{E}X^2 < \infty$. Then, $\mathbb{E}(X - Y)^2$ can be regarded as a measure of the error committed by using Y as an estimate of X, and we want $\mathbb{E}(X - \bar{X})^2 \leq \mathbb{E}(X - Y)^2$ for all Y in $\mathcal{F}$.

Preparatory steps

These are to illustrate the thought processes involved and to motivate the formal definitions to come. Recall that $\mathcal{H}$ is the collection of all events

E. Çınlar, *Probability and Stochastics*, Graduate Texts in Mathematics 261, DOI 10.1007/978-0-387-87859-1_4,

and is also the collection of all $\bar{\mathbb{R}}$-valued random variables, $\mathcal{H}_+$ being the subcollection consisting of the positive ($\bar{\mathbb{R}}_+$-valued) variables. Similarly, for the sub-σ-algebra $\mathcal{F}$, we regard $\mathcal{F}$ as the collection of all $\mathcal{F}$-measurable random variables and $\mathcal{F}_+$ as the sub-collection of positive ones. To simplify the discussion here, we assume $X \in \mathcal{H}_+$.

Let H be an event. Fix ω in Ω. Suppose that all that is known about ω is that it is in H. Based on this information, our best estimate of $X(\omega)$ should be the "average" of X over H, namely, the number

$$\mathbb{E}_H X = \frac{1}{\mathbb{P}(H)} \int_H X d\mathbb{P} = \frac{1}{\mathbb{P}(H)} \mathbb{E} X 1_H \tag{1.1}$$

assuming that $\mathbb{P}(H) > 0$. If $\mathbb{P}(H) = 0$ then so is the integral over H and we allow $\mathbb{E}_H X$ to be any positive number desired. This estimate is best in the same sense that $\mathbb{E}X$ is the best number estimating $X(\omega)$ when nothing whatsoever is known about ω, which corresponds to the case $H = \Omega$ (see Exercise II.2.15). The number $\mathbb{E}_H X$ is called the *conditional expectation* of X *given the event* H.

Suppose, next, that $\mathcal{F}$ is generated by a measurable partition (H_n) of Ω. For fixed ω again, consider what our estimate of $X(\omega)$ should be if we were given the information $\mathcal{F}$. Given $\mathcal{F}$, we shall be able to tell which one of the events $H_1, H_2, \ldots$ includes ω, and if it were H_n that included ω then our estimate $\bar{X}(\omega)$ would be $\mathbb{E}_{H_n} X$. In other words,

$$\bar{X}(\omega) = \sum_n (\mathbb{E}_{H_n} X) 1_{H_n}(\omega). \tag{1.2}$$

Doing this thinking for each possibility ω, we arrive at a random variable $\bar{X}$, which we denote by $\mathbb{E}_{\mathcal{F}} X$ and call the *conditional expectation of* X *given* $\mathcal{F}$. Figure 3 is for the case where $\Omega = (0,1)$, $\mathcal{H}$ is the Borel σ-algebra, and $\mathbb{P}$ is the Lebesgue measure on Ω, and $\mathcal{F} = \sigma\{H_1, H_2, H_3\}$.

To see the proper generalization for arbitrary $\mathcal{F}$ we mark two properties of $\bar{X} = \mathbb{E}_{\mathcal{F}} X$ for the special $\mathcal{F}$ above. First, $\bar{X}$ belongs to $\mathcal{F}$, that is, $\bar{X}$ is determined by the information $\mathcal{F}$. Second, $\mathbb{E}\ VX = \mathbb{E}\ V\bar{X}$ for every V in $\mathcal{F}_+$; this is obvious from 1.1 and 1.2 when $V = 1_{H_n}$ for some fixed n, and it extends to arbitrary V in $\mathcal{F}_+$ through the monotone convergence theorem, because every such V has the form $V = \sum a_n 1_{H_n}$ for this $\mathcal{F}$. We use these two properties to define conditional expectations and proceed to show their existence, uniqueness, and various properties.

Definition of conditional expectations

1.3 Definition. *Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. The* conditional expectation of X given $\mathcal{F}$, *denoted by $\mathbb{E}_{\mathcal{F}} X$, is defined in two steps: For X in $\mathcal{H}_+$, it is defined to be any random variable $\bar{X}$ that satisfies*

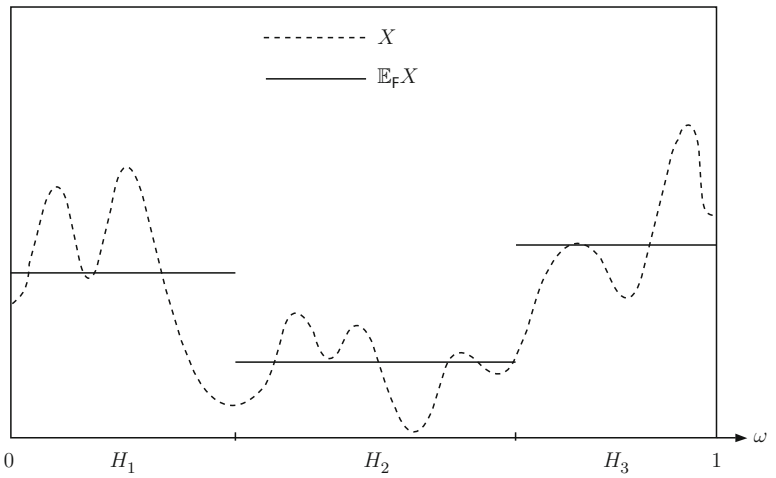

Figure 3: The conditional expectation $\mathbb{E}_{\mathcal{F}}X$ is that $\mathcal{F}$-measurable random variable that is closest to X.

1.4 a) *measurability:* $\bar{X}$ belongs to $\mathcal{F}_+$, and
b) *projection property:* $\mathbb{E}\, VX = \mathbb{E}\, V\bar{X}$ for every V in $\mathcal{F}_+$,

and then we write $\mathbb{E}_{\mathcal{F}}X = \bar{X}$ *and call* $\bar{X}$ *a version of* $\mathbb{E}_{\mathcal{F}}X$. *For arbitrary* X *in* $\mathcal{H}$, *if* $\mathbb{E}X$ *exists, then we define*

1.5 $$\mathbb{E}_{\mathcal{F}}X = \mathbb{E}_{\mathcal{F}}X^+ - \mathbb{E}_{\mathcal{F}}X^-;$$

otherwise, if $\mathbb{E}X^+ = \mathbb{E}X^- = +\infty$, *then* $\mathbb{E}_{\mathcal{F}}X$ *is left undefined.*

1.6 Remarks. a) *Projection property.* For X in $\mathcal{H}_+$, the projection property 1.4b for $\bar{X} = \mathbb{E}_{\mathcal{F}}X$ is equivalent to the condition that

$$\mathbb{E}\, 1_H X = \mathbb{E}\, 1_H \bar{X}, \qquad H \in \mathcal{F}.$$

This is immediate from the monotone convergence theorem used on both sides of the equality to extend it from indicators to simple variables and from simple to arbitrary positive V in $\mathcal{F}$.

b) *Tests for equality.* This is to recall an elementary point made in Remark II.2.3d. If Y and Z are in $\mathcal{F}_+$ and if

$$\mathbb{E}\, VY \le \mathbb{E}\, VZ \qquad \text{for every } V \text{ in } \mathcal{F}_+$$

or for every indicator 1_H in $\mathcal{F}_+$, then $Y \le Z$ almost surely (and the almost sure set $\{Y \le Z\}$ is in $\mathcal{F}$). If the inequality is in fact an equality for all V in $\mathcal{F}_+$, then $Y = Z$ almost surely.

c) *Uniqueness of conditional expectations.* Let $\bar{X}$ and $\bar{\bar{X}}$ be versions of $\mathbb{E}_{\mathcal{F}}X$ for $X \geq 0$. Then they are both in $\mathcal{F}_+$ and $\mathbb{E}\ V\bar{X} = \mathbb{E}\ VX = \mathbb{E}\ V\bar{\bar{X}}$ for every V in $\mathcal{F}_+$. It follows from the preceding remark that $\bar{X} = \bar{\bar{X}}$ almost surely. Conversely, if $\mathbb{E}_{\mathcal{F}}X = \bar{X}$, and if $\bar{\bar{X}} \in \mathcal{F}_+$ and $\bar{\bar{X}} = \bar{X}$ almost surely, then $\bar{\bar{X}}$ satisfies 1.4 and hence is another version of $\mathbb{E}_{\mathcal{F}}X$. This uniqueness up to equivalence extends to $\mathbb{E}_{\mathcal{F}}X$ for arbitrary X for which $\mathbb{E}X$ exists; see also (f) below.

d) *Language.* In view of the uniqueness up to equivalence, the definite article in "the conditional expectation ..." is only a slight abuse of language. For the same reason, $\mathbb{E}_{\mathcal{F}}X$ should be regarded as an all-purpose notation for each and every version $\bar{X}$. Some authors take the extra logical step and define "the conditional expectation ..." to be the equivalence class of all versions, and then use $\mathbb{E}_{\mathcal{F}}X$ as a representative of that class, and write "$\mathbb{E}_{\mathcal{F}}X = \bar{X}$ almost surely" to mean that $\bar{X}$ is a version.

e) *Integrability.* If $X \in \mathcal{H}_+$ then $\mathbb{E}\ X = \mathbb{E}\ \mathbb{E}_{\mathcal{F}}X$ in view of the projection property with $V = 1$; thus, if X is integrable then so is $\mathbb{E}_{\mathcal{F}}X$. This remark applies to X^+ and X^- separately for arbitrary X and, thus, if X is integrable then so is $\mathbb{E}_{\mathcal{F}}X$. Hence, if X is integrable and $\mathbb{E}_{\mathcal{F}}X = \bar{X}$, the projection property can be expressed as

$$\mathbb{E}\ V(X - \bar{X}) = 0 \qquad \text{for every bounded } V \text{ in } \mathcal{F}.$$

In general, if X is not integrable, this expression fails to have meaning because the expectation involved might fail to exist.

f) *Definition for arbitrary X.* If X^+ is integrable, then so is $\mathbb{E}_{\mathcal{F}}X^+$ which implies that $\mathbb{E}_{\mathcal{F}}X^+$ is real-valued almost surely, and a similar statement holds for X^-. Thus, if $\mathbb{E}X$ exists, there is an almost sure event Ω_0 in $\mathcal{F}$ such that, on Ω_0, at least one of the random variables $\mathbb{E}_{\mathcal{F}}X^+$ and $\mathbb{E}_{\mathcal{F}}X^-$ is real-valued, and therefore, $\bar{X} = \mathbb{E}_{\mathcal{F}}X^+ - \mathbb{E}_{\mathcal{F}}X^-$ is well-defined on Ω_0. By putting $\bar{X}$ equal to 0 outside Ω_0 for definiteness, we obtain a well-defined version $\bar{X}$ of $\mathbb{E}_{\mathcal{F}}X$.

g) *Heuristics.* Suppose that X is integrable; then, so is $\bar{X} = \mathbb{E}_{\mathcal{F}}X$ and so is $\tilde{X} = X - \bar{X}$. We thus have a decomposition

$$X = \bar{X} + \tilde{X}$$

where $\bar{X}$ is determined by the information $\mathcal{F}$, and $\tilde{X}$ is orthogonal to $\mathcal{F}$ (that is, $\mathbb{E}\ 1_H\tilde{X} = 0$ for all events H in $\mathcal{F}$; see the remark (e) above). For this reason, we may call $\bar{X}$ the orthogonal projection of X onto $\mathcal{F}$, and the defining property 1.4b is named "projection property" to suggest this picture. We shall return to this theme and make it rigorous in Theorem 1.11.

Existence of conditional expectations

The following proposition uses the Radon-Nikodym theorem to show the existence of conditional expectations. We limit the proposition to positive X, since the extension to arbitrary X is covered by Remark 1.6f above. See 1.15 below for another proof which is independent of the Radon-Nikodym theorem.

1.7 THEOREM. *Let $X \in \mathcal{H}_+$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then $\mathbb{E}_{\mathcal{F}}X$ exists and is unique up to equivalence.*

Proof. For each event H in $\mathcal{F}$, define

$$P(H) = \mathbb{P}(H), \qquad Q(H) = \int_H \mathbb{P}(d\omega)X(\omega).$$

On the measurable space $(\Omega, \mathcal{F})$, then, P is a probability measure, and Q is a measure that is absolutely continuous with respect to P. Hence, by I.5.11, the Radon-Nikodym theorem, there exists $\bar{X}$ in $\mathcal{F}_+$ such that

$$\int_\Omega Q(d\omega)V(\omega) = \int_\Omega \mathbb{P}(d\omega)\bar{X}(\omega)V(\omega)$$

for every V in $\mathcal{F}_+$. This shows that $\bar{X}$ is a version of $\mathbb{E}_{\mathcal{F}}X$. For its uniqueness up to almost sure equality, see Remark 1.6c. □

Properties similar to expectations

The following properties are the same as those for expectations. They are easy to show and will not be proved. Throughout, $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{H}$, all the random variables are in $\mathcal{H}$ (that is, they are arbitrary $\bar{\mathbb{R}}$-valued variables), the constants a, b, c are real numbers, *and* it is assumed that all conditional expectations involved exist. Of course, all these conditional expectations exist if all the random variables are positive or integrable.

1.8 *Monotonicity:* $X \le Y \;\Rightarrow\; \mathbb{E}_{\mathcal{F}}X \le \mathbb{E}_{\mathcal{F}}Y$.

Linearity: $\mathbb{E}_{\mathcal{F}}(aX + bY + c) = a\,\mathbb{E}_{\mathcal{F}}X + b\,\mathbb{E}_{\mathcal{F}}Y + c$.

Monotone convergence theorem: $X_n \ge 0, X_n \nearrow X \;\Rightarrow\; \mathbb{E}_{\mathcal{F}}X_n \nearrow \mathbb{E}_{\mathcal{F}}X$.

Fatou's lemma: $X_n \ge 0 \;\Rightarrow\; \mathbb{E}_{\mathcal{F}} \liminf X_n \le \liminf \mathbb{E}_{\mathcal{F}}X_n$.

Dominated convergence theorem: $X_n \to X$, $|X_n| \le Y$, Y integrable $\Rightarrow\; \mathbb{E}_{\mathcal{F}}X_n \to \mathbb{E}_{\mathcal{F}}X$.

Jensen's inequality: f convex $\;\Rightarrow\; \mathbb{E}_{\mathcal{F}}f(X) \ge f(\mathbb{E}_{\mathcal{F}}X)$.

1.9 REMARK. There are occasions when the properties above require careful interpretation, because conditional expectations are unique only up to equivalence. To illustrate, we now re-state the monotone convergence theorem for them in precise terms and prove it:

> Suppose that $(X_n) \subset \mathcal{H}_+$ and $X_n \nearrow X$. Then there are versions $\bar{X}_n$ of $\mathbb{E}_{\mathcal{F}}X_n$ and $\bar{X}$ of $\mathbb{E}_{\mathcal{F}}X$ such that $\bar{X}_n \nearrow \bar{X}$.

To show this, let X'_n be a version of $\mathbb{E}_{\mathcal{F}}X_n$ for each n. By monotonicity, for each n, $X'_n \le X'_{n+1}$ almost surely. Thus, there is an almost sure event

Ω_0 in $\mathcal{F}$ such that, for every ω in Ω_0, we have $X_1'(\omega) \le X_2'(\omega) \le \cdots$. Define $\bar{X}_n(\omega) = X_n'(\omega)$ for ω in Ω_0 and put $\bar{X}_n(\omega) = 0$ for ω outside Ω_0. Then, each $\bar{X}_n$ is a version of $\mathbb{E}_{\mathcal{F}} X_n$, and $(\bar{X}_n)$ is an increasing sequence in $\mathcal{F}_+$; let $\bar{X}$ be the limit. Then, $\bar{X} \in \mathcal{F}_+$ and, for $V \in \mathcal{F}_+$,

$$\mathbb{E}\, V\bar{X} = \lim \mathbb{E}\, V\bar{X}_n = \lim \mathbb{E}\, VX_n = \mathbb{E}\, VX$$

where the projection property 1.4b justifies the middle equality sign, and the ordinary monotone convergence theorem the other two. Thus, $\bar{X}$ is a version of $\mathbb{E}_{\mathcal{F}} X$.

Special properties

The following theorem summarizes the properties special to conditional expectations. Heuristically, conditional determinism is that, if W is determined by $\mathcal{F}$ then it should be treated as if it is a deterministic number. For the repeated conditioning, think of $\mathcal{F}$ as the information a fool has, and $\mathcal{G}$ as that a genius has: the genius cannot improve on the fool's estimate, but the fool has no difficulty worsening the genius's. In repeated conditioning, fools win all the time.

1.10 Theorem. *Let $\mathcal{F}$ and $\mathcal{G}$ be sub-σ-algebras of $\mathcal{H}$. Let W and X be random variables (in $\mathcal{H}$) such that $\mathbb{E}X$ and $\mathbb{E}WX$ exist. Then, the following hold:*

a) Conditional determinism*:* $W \in \mathcal{F} \;\Rightarrow\; \mathbb{E}_{\mathcal{F}} WX = W\mathbb{E}_{\mathcal{F}} X$.
b) Repeated conditioning*:* $\mathcal{F} \subset \mathcal{G} \;\Rightarrow\; \mathbb{E}_{\mathcal{F}}\mathbb{E}_{\mathcal{G}} X = \mathbb{E}_{\mathcal{G}}\mathbb{E}_{\mathcal{F}} X = \mathbb{E}_{\mathcal{F}} X$.

Proof. We give the proofs for W and X positive; the general cases follow by easy considerations.

a) Suppose $X \in \mathcal{H}_+$ and $W \in \mathcal{F}_+$. Then $\bar{X} = \mathbb{E}_{\mathcal{F}} X$ is in $\mathcal{F}_+$ and

$$\mathbb{E}\, V \cdot (WX) = \mathbb{E}\, (V \cdot W)X = \mathbb{E}\, (V \cdot W)\bar{X} = \mathbb{E}\, V \cdot (W\bar{X})$$

for every V in $\mathcal{F}_+$, where the crucial middle equality sign is justified by the projection property for $\bar{X}$ and the fact that $VW \in \mathcal{F}_+$. Hence, $W\bar{X} = W\mathbb{E}_{\mathcal{F}} X$ is a version of $\mathbb{E}_{\mathcal{F}}(WX)$.

b) Let $\mathcal{F} \subset \mathcal{G}$ and $X \in \mathcal{H}_+$. Since $W = \mathbb{E}_{\mathcal{F}} X$ is in $\mathcal{F}_+$ by definition, we have that $W \in \mathcal{G}_+$ and hence $\mathbb{E}_{\mathcal{G}} W = W$ by the conditional determinism property proved above. This proves the second equality in the statement of repeated conditioning. There remains to show that, with $Y = \mathbb{E}_{\mathcal{G}} X$ and $\bar{X} = \mathbb{E}_{\mathcal{F}} X$,

$$\mathbb{E}_{\mathcal{F}} Y = \bar{X}.$$

Obviously, $\bar{X} \in \mathcal{F}_+$. To check the projection property, let $V \in \mathcal{F}_+$. By the definition of $\bar{X}$, we have $\mathbb{E}\, V\bar{X} = \mathbb{E}\, VX$. By the definition of Y, we have $\mathbb{E}\, VY = \mathbb{E}\, VX$, since $V \in \mathcal{G}_+$, which is in turn because $V \in \mathcal{F}_+$ and $\mathcal{F} \subset \mathcal{G}$. Hence, $\mathbb{E}\, V\bar{X} = \mathbb{E}\, VY$ as needed. □

Conditioning as projection

We return to the beginnings. Recall the heuristic remarks about $\mathbb{E}_{\mathcal{F}}X$, its geometric interpretation as a projection, and its interpretation as an $\mathcal{F}$-determined estimate that minimizes the expected value of the squared error caused by such estimates of X. The following theorem is the rigorous statement justifying such remarks. As a corollary, we obtain a second proof of the existence and uniqueness Theorem 1.7, this time without recourse to the unproved Radon-Nikodym theorem.

Since we shall be dealing with "expected value of squared error," we are necessarily limited to square-integrable random variables in the next theorem. Accordingly, if $\mathcal{F}$ is a sub-σ-algebra of $\mathcal{H}$, we write $L^2(\mathcal{F})$ for the collection of all square-integrable random variables in $\mathcal{F}$; then, $L^2(\mathcal{H})$ is the L^2-space introduced in Chapter II.

1.11 Theorem. *For every X in $L^2(\mathcal{H})$ there exists a unique (up to equivalence) $\bar{X}$ in $L^2(\mathcal{F})$ such that*

$$\mathbb{E}\ |X-\bar{X}|^2 = \inf_{Y\in L^2(\mathcal{F})}\mathbb{E}\ |X-Y|^2 \tag{1.12}$$

Moreover, $X-\bar{X}$ is orthogonal to $L^2(\mathcal{F})$, that is, for every V in $L^2(\mathcal{F})$,

$$\mathbb{E}\ V\cdot(X-\bar{X}) = 0 \tag{1.13}$$

1.14 Remarks. Obviously, $\bar{X} = \mathbb{E}_{\mathcal{F}}X$. With $\mathbb{E}XY$ as the inner product of X and Y, the space $L^2(\mathcal{H})$ is a complete Hilbert space, and $L^2(\mathcal{F})$ is a subspace of it. In the terminology of such spaces, $\bar{X}$ is the orthogonal projection of the vector X onto $L^2(\mathcal{F})$, and we have the decomposition $X = \bar{X}+\tilde{X}$ where $\bar{X}$ is in $L^2(\mathcal{F})$ and $\tilde{X}$ is orthogonal to $L^2(\mathcal{F})$.

Proof. It is convenient to use the L^2-norm introduced in Chapter II, section 3, but omitting the subscripts: Thus, $\|X\| = \|X\|_2 = \sqrt{\mathbb{E}X^2}$. Fix X in $L^2(\mathcal{H})$. Define

$$\delta = \inf_{Y\in L^2(\mathcal{F})}\|X-Y\|$$

and let $(Y_n)\subset L^2(\mathcal{F})$ such that $\delta_n = \|X-Y_n\|\to\delta$ as $n\to\infty$.

We show now that (Y_n) is Cauchy for convergence in $L^2(\mathcal{F})$. Note that

$$|Y_n-Y_m|^2 = 2\ |X-Y_m|^2 + 2\ |X-Y_n|^2 - 4\ |X-\tfrac{1}{2}(Y_m+Y_n)|^2$$

and take expectations to get

$$\mathbb{E}|Y_n-Y_m|^2 \le 2\delta_m^2 + 2\delta_n^2 - 4\delta^2,$$

since $\|X-Y\|\ge\delta$ for every Y in $L^2(\mathcal{F})$ and in particular for $Y=\frac{1}{2}(Y_m+Y_n)$.

Hence, (Y_n) is Cauchy, and by Theorem III.4.6 and III.4.13, there exists $\bar{X}$ in $L^2(\mathcal{F})$ such that $\|Y_n-\bar{X}\|\to 0$ as $n\to\infty$. Of course, $\bar{X}$ is unique up

to almost sure equality. Since $\bar{X} \in L^2(\mathcal{F})$, we have $\|X - \bar{X}\| \geq \delta$; and by Minkowski's inequality (see Theorem II.3.6),

$$\|X - \bar{X}\| \leq \|X - Y_n\| + \|Y_n - \bar{X}\| \longrightarrow \delta + 0 = \delta.$$

It follows that $\|X - \bar{X}\| = \delta$ as needed to complete the proof of the first statement.

Moreover, for V in $L^2(\mathcal{F})$ and a real, since $\mathbb{E}(X - \bar{X})^2 = \delta^2$,

$$a^2\ \mathbb{E}\ V^2 - 2a\ \mathbb{E}\ V \cdot (X - \bar{X}) + \delta^2 = \|aV - (X - \bar{X})\|^2 = \|X - (aV + \bar{X})\|^2 \geq \delta^2$$

because $aV + \bar{X} \in L^2(\mathcal{F})$. Thus, $a^2\ \mathbb{E}\ V^2 - 2a\ \mathbb{E}\ V \cdot (X - \bar{X}) \geq 0$ for every real number a, which is impossible unless 1.13 holds. □

1.15 Second Proof For 1.7. Let $X \in \mathcal{H}_+$. Define $X_n = X \wedge n$, which is in $L^2(\mathcal{H})$. Thus, by the preceding theorem, there is $\bar{X}_n$ in $L^2(\mathcal{F})$ such that, for every V in $L^2(\mathcal{F})$,

1.16 $$\mathbb{E}\ VX_n = \mathbb{E}\ V\bar{X}_n$$

Now fix $V = 1_H$ for some event H in $\mathcal{H}$. Then, since $X_n \nearrow X$, 1.16 implies that $(\bar{X}_n)$ increases almost surely to some $\bar{X}$ in $\mathcal{F}_+$, and the monotone convergence theorem shows that $\mathbb{E}VX = \mathbb{E}V\bar{X}$. □

Conditional expectations given random variables

Let Y be a random variable taking values in some measurable space. Recall that σY denotes the σ-algebra generated by Y, which consists of numerical random variables of the form $f \circ Y$ for some measurable function f; see II.4.1 *et seq.* For X in $\mathcal{H}$, the *conditional expectation of X given Y* is defined to be $\mathbb{E}_{\sigma Y} X$. Similarly, if $\{Y_t : t \in T\}$ is a collection of random variables taking values in some measurable spaces, the *conditional expectation of X given $\{Y_t : t \in T\}$* is $\mathbb{E}_{\mathcal{F}} X$ with $\mathcal{F} = \sigma\{Y_t : t \in T\}$. Of course, these two definitions coincide when the collection $\{Y_t : t \in T\}$ is identified with $Y = (Y_t)_{t \in T}$. The following is an immediate consequence of the definitions here and Theorem II.4.4.

1.17 Theorem. *Let $X \in \mathcal{H}_+$. Let Y be a random variable taking values in some measurable space $(E, \mathcal{E})$. Then, every version of $\mathbb{E}_{\sigma Y} X$ has the form $f \circ Y$ for some f in $\mathcal{E}_+$. Conversely, $f \circ Y$ is a version of $\mathbb{E}_{\sigma Y} X$ if and only if*

1.18 $$\mathbb{E}\ f \circ Y\ h \circ Y = \mathbb{E}\ X \cdot h \circ Y \qquad \text{for every } h \text{ in } \mathcal{E}_+.$$

In the preceding theorem, the hypothesis that X be positive is inessential, except for ensuring that the conditional expectation exists. If X is integrable, the claim remains the same for f in $\mathcal{E}$ and bounded h in $\mathcal{E}_+$. Also, when X is integrable, 1.18 can be replaced, via a monotone class – monotone convergence argument, with the requirement that it hold for $h = 1_E$ and $h = 1_A$ for every A in some p-system $\mathcal{E}_0$ that generates $\mathcal{E}$.

Notation

The traditional notation for $\mathbb{E}_{\mathcal{F}}X$ is $\mathbb{E}(X|\mathcal{F})$. The notation $\mathbb{E}_{\mathcal{F}}X$ is sharper, conveys better its meaning as a projection, is clearer in expressions like $\mathbb{E}_{\mathcal{F}}\mathbb{E}_{\mathcal{G}}$, and conforms to ordinary practices in notations for operators. The other, $\mathbb{E}(X|\mathcal{F})$ is no better than a shorthand, but has the advantage of being more linear and more convenient when $\mathcal{F}$ has to be replaced by a cumbersome expression. For instance, it is usual to write $\mathbb{E}(X|Y_t : t \in T)$ for $\mathbb{E}_{\mathcal{F}}X$ when $\mathcal{F} = \sigma\{Y_t : t \in T\}$.

The notations $\mathbb{E}(X|Y)$ and $\mathbb{E}_{\sigma Y}X$ are used often to denote $\mathbb{E}_{\mathcal{F}}X$ with $\mathcal{F} = \sigma Y$. Similarly, $\mathbb{E}(X|H)$ is the traditional notation for $\mathbb{E}_H X$ and works better when H is expressed in terms of other things, for example, if $H = \{Y = y\}$.

Finally, the notation $\mathbb{E}(X|Y = y)$ is used and read "the conditional expectation of X given that $Y = y$" despite its annoying ambiguity. It is reasonable, when $\mathbb{P}\{Y = y\} > 0$, as a notation for $\mathbb{E}_H X = \mathbb{E}(X|H)$ with $H = \{Y = y\}$. It is also used when $\mathbb{P}\{Y = y\} = 0$ for all y, and the proper interpretation then is that it is a notation for $f(y)$ when $f \circ Y = \mathbb{E}_{\sigma Y}X$.

Examples

1.19 *Remaining lifetime.* A device is installed at time 0. It is known that the device has not failed during $[\,0\,,\,t\,]$. We would like to estimate its remaining lifetime. Let X represent the length of its life. What we know is that the event $H = \{X > t\}$ has occurred (that is, outcome is a point in H). We want to compute $\mathbb{E}_H(X - t) = \mathbb{E}_H X - t$:

Let μ be the distribution of X. Then, with $H = \{X > t\}$,

$$\mathbb{E}_H X = \frac{1}{\mathbb{P}(H)}\mathbb{E}\, X1_H = \frac{1}{\mu(t,\infty)}\int_{\mathbb{R}_+} \mu(ds)\; s\; 1_{(t,\infty)}(s) = \frac{1}{\mu(t,\infty)}\int_{(t,\infty)} \mu(ds)s.$$

1.20 *Effect of independence.* Suppose that X and Y are independent and take values in $(E, \mathcal{E})$ and $(D, \mathcal{D})$ respectively. If $f \in \mathcal{E}_+$ and $g \in \mathcal{D}_+$, then

$$\mathbb{E}_{\sigma Y}\; f\circ X\; g\circ Y = g\circ Y\; \mathbb{E}\; f\circ X \tag{1.21}$$

This adds meaning to independence: information obtained by observing Y determines $g \circ Y$, but is worthless for estimating $f \circ X$. To see 1.21, first observe that the right side is a constant multiple of $g\circ Y$ and hence satisfies the measurability condition that it be in σY. To check for the projection property, let V be positive and in σY; then $V = h\circ Y$ for some h in $\mathcal{D}_+$, and the required equality, namely,

$$\mathbb{E}\; f\circ X\; g\circ Y\; h\circ Y = \mathbb{E}\; h\circ Y\; g\circ Y\; \mathbb{E}\; f\circ X$$

follows from the definition of independence (see Proposition II.5.6).

1.22 *Continuation.* Let X and Y be as in 1.20 above. Then, for every positive h in $\mathcal{E} \otimes \mathcal{D}$,

$$\mathbb{E}_{\sigma Y}\ h(X,Y) = \bar{h} \circ Y, \tag{1.23}$$

where $\bar{h}(y) = \mathbb{E}\ h(X,y)$ for each y in D. First, we note that $\bar{h} \in \mathcal{D}_+$ by Fubini's theorem, and thus $\bar{h} \circ Y$ has the measurability required for it to be a conditional expectation given Y. As to the projection property, we observe that the collection of h in $\mathcal{E} \otimes \mathcal{D}$ for which

$$\mathbb{E}\ V \cdot h(X,Y) = \mathbb{E}\ V \cdot \bar{h} \circ Y \tag{1.24}$$

for some fixed V in $(\sigma Y)_+$ is a monotone class, and it includes all such h of the form $h(x,y) = f(x)g(y)$ by the preceding example. Thus, by the monotone class theorem, 1.24 holds for all positive h in $\mathcal{E} \otimes \mathcal{D}$ for which the conditional expectation is defined (that is, if $\mathbb{E}\ h(X,Y)$ exists).

1.25 *Conditional expectation of a part given the whole.* Let X and Y be independent and gamma distributed with the respective shape indices a and b, and the same scale parameter c. Then we have seen in Example II.2.11 that $Z = X + Y$ and $U = X/(X+Y)$ are independent, and some easy computations give $\mathbb{E}X = a/c$ and $\mathbb{E}Z = (a+b)/c$. Now, $\mathbb{E}\ UZ = \mathbb{E}U \cdot \mathbb{E}Z$ by independence, from which we solve for $\mathbb{E}U$ to get $\mathbb{E}\ U = a/(a+b)$ since $UZ = X$. By the same token, using 1.21, we have

$$\mathbb{E}_{\sigma Z} X = \mathbb{E}_{\sigma Z} UZ = Z\ \mathbb{E}U = \frac{a}{a+b} Z.$$

Exercises

1.26 *Relationship between $\mathbb{E}_{\mathcal{F}}X$ and $\mathbb{E}_H X$.* Let H be an event and let $\mathcal{F} = \sigma H = \{\emptyset, H, H^c, \Omega\}$. Show that $\mathbb{E}_{\mathcal{F}} X(\omega) = \mathbb{E}_H X$ for all ω in H.

1.27 *Remaining lifetime.* In Example 1.19, suppose that μ is the exponential distribution with parameter c. Show that, for $H = \{X > t\}$,

$$\mathbb{E}_H X = t + \mathbb{E}X.$$

Heuristically, then, if we know that $X(\omega) > t$ but nothing further about ω, our estimate of the lifetime $X(\omega)$ is $t + 1/c$. That is, the remaining lifetime at time t is as if the device is new at time t. This property characterizes the exponential distribution: if μ is absolutely continuous and $\mathbb{E}(X|X > t) = t + a$ for some constant a and all $t \geq 0$, then μ is the exponential distribution with parameter $c = 1/a$. Prove this.

1.28 *Conditional probabilities–elementary setup.* Let H and G be events. The conditional probability of G given H, denoted by $\mathbb{P}_H(G)$ or by $\mathbb{P}(G|H)$, is defined to be $\mathbb{E}_H 1_G$. Show that it satisfies

$$\mathbb{P}(H \cap G) = \mathbb{P}(H)\mathbb{P}_H(G).$$

If $\mathbb{P}(H) > 0$, this defines $\mathbb{P}_H(G)$ uniquely. If $\mathbb{P}(H) = 0$, then so is $\mathbb{P}(H \cap G)$, and $\mathbb{P}_H(G)$ can be taken to be any convenient number in $[0, 1]$.

1.29 *Expectations given an event.* Let $X \in \mathcal{H}_+$ and let H be an event with $\mathbb{P}(H) > 0$. Show that, similar to the result of Exercise II.2.14,

$$\mathbb{E}_H X = \int_{\mathbb{R}_+} dt\ \mathbb{P}_H\{X > t\}.$$

1.30 *Expectations given discrete variables.* Let $X \in \mathcal{H}_+$. Let Y be a random variable taking values in some countable set D. Show that, then, $\mathbb{E}_{\sigma Y} X = f \circ Y$, where f is defined by

$$f(a) = \mathbb{E}(X|Y = a) = \int_{\mathbb{R}_+} dt\ \mathbb{P}\{X > t|Y = a\}, \quad a \in D.$$

1.31 *Bounds.* If $X \le b$ for some constant b then $\mathbb{E}_{\mathcal{F}} X \le b$. If X takes values in $[a, b]$, then so does $\mathbb{E}_{\mathcal{F}} X$.

1.32 *Conditional expectation operator.* The mapping $\mathbb{E}_{\mathcal{F}} : X \mapsto \mathbb{E}_{\mathcal{F}} X$ maps $L^p(\mathcal{H})$ into $L^p(\mathcal{F})$ for every p in $[1, \infty]$. For $p = 1$ see Remark 1.6e, for $p = 2$ see Theorem 1.11, and for $p = +\infty$ see 1.31 above.

1.33 *Continuation.* If $X_n \to X$ in L^p for some p in $[1, \infty]$, then $\mathbb{E}_{\mathcal{F}} X_n \to \mathbb{E}_{\mathcal{F}} X$ in L^p with the same p. Show.

1.34 *Continuation.* If (X_n) is uniformly integrable and converges to X in probability, then $\mathbb{E}_{\mathcal{F}} X_n \to \mathbb{E}_{\mathcal{F}} X$ in L^1. Hint: see Theorem III.4.6 and the preceding exercise.

2 Conditional Probabilities and Distributions

Let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. For each event H in $\mathcal{H}$,

2.1 $$\mathbb{P}_{\mathcal{F}} H = \mathbb{E}_{\mathcal{F}} 1_H$$

is called the *conditional probability of H given $\mathcal{F}$*. In more elementary settings, for events G and H, the *conditional probability of H given G* is defined to be any number $\mathbb{P}_G(H)$ in $[0, 1]$ satisfying

2.2 $$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}_G(H);$$

of course, it is unique when $\mathbb{P}(G) > 0$.

Regular versions

Let $Q(H)$ be a version of $\mathbb{P}_{\mathcal{F}}H$ for each H in $\mathcal{H}$. We may, and do, assume that $Q(\emptyset) = 0$ and $Q(\Omega) = 1$. Of course, $Q(H)$ is a random variable in the σ-algebra $\mathcal{F}$; let $Q_\omega(H)$ denote its value at the point ω of Ω.

At first, the mapping $Q : (\omega, H) \mapsto Q_\omega(H)$ looks like a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$: By the definition of conditional expectations, the mapping $\omega \mapsto Q_\omega(H)$ is $\mathcal{F}$-measurable for each H in $\mathcal{H}$; and by the monotone convergence property of 1.8, for every disjointed sequence (H_n) in $\mathcal{H}$,

$$Q_\omega(\cup_n H_n) = \sum_n Q_\omega(H_n) \tag{2.3}$$

for almost every ω in Ω. It is this "almost" that keeps Q from being a transition kernel, and it is a serious limitation in this case.

Generally, the almost sure event Ω_h of all ω for which 2.3 holds depends on the sequence $h = (H_n)$. The set Ω_0 of all ω for which $H \mapsto Q_\omega(H)$ is a probability measure is equal to $\cap\Omega_h$, where the intersection is over all disjointed sequences h. We need Ω_0 to be almost sure before we can fix Q to become a kernel. But, usually, there are uncountably many disjointed sequences h and, hence, the intersection Ω_0 is generally a miserable object: Ω_0 might fail to belong to $\mathcal{F}$ or even to $\mathcal{H}$, and even if Ω_0 is in $\mathcal{F}$, we might have $\mathbb{P}(\Omega_0) < 1$ or even $\mathbb{P}(\Omega_0) = 0$.

Nevertheless, it is often possible to pick versions of $Q(H)$ such that $\Omega_0 = \Omega$. Such versions are highly prized.

2.4 Definition. *Let $Q(H)$ be a version of $\mathbb{P}_{\mathcal{F}}H$ for every H in $\mathcal{H}$. Then $Q : (\omega, H) \mapsto Q_\omega(H)$ is said to be a* regular version *of the conditional probability $\mathbb{P}_{\mathcal{F}}$ provided that Q be a transition probability kernel from $(\Omega, \mathcal{F})$ into $(\Omega, \mathcal{H})$.*

The reason for the popularity of regular versions is the following:

2.5 Proposition. *Suppose that $\mathbb{P}_{\mathcal{F}}$ has a regular version Q. Then,*

$$QX : \omega \mapsto Q_\omega X = \int_\Omega Q_\omega(d\omega')X(\omega')$$

is a version of $\mathbb{E}_{\mathcal{F}}X$ for every random variable X whose expectation exists.

Proof. It is sufficient to prove this for X in $\mathcal{H}_+$. For such X, by (Fubini's) Theorem I.6.3 applied to the transition kernel Q and function X, we see that $QX \in \mathcal{F}_+$. It is thus enough to check the projection property, namely that, for V in $\mathcal{F}_+$,

$$\mathbb{E}\, VX = \mathbb{E}\, V\, QX.$$

Fix V. For $X = 1_H$, this is immediate from the definition of $Q(H) = Q1_H$ as a version of $\mathbb{P}_{\mathcal{F}}H = \mathbb{E}_{\mathcal{F}}1_H$. This extends first to simple random variables X and then to arbitrary positive X by the linearity of, and the monotone convergence theorem for, the operators $X \mapsto QX$ and $Z \mapsto \mathbb{E}Z$. □

Existence of a regular version for $\mathbb{P}_{\mathcal{F}}$ requires conditions either on $\mathcal{F}$ or on $\mathcal{H}$. For instance, if $\mathcal{F}$ is generated by a partition (G_n) of Ω, which is the case if $\mathcal{F}$ is generated by a random variable taking values in a countable space, then

$$2.6 \qquad Q_\omega(H) = \sum_n (\mathbb{P}_{G_n} H) 1_{G_n}(\omega) \ , \qquad \omega \in \Omega, H \in \mathcal{H},$$

defines a regular version, since 2.2 yields a probability measure $H \mapsto \mathbb{P}_G H$ for each G. When $\mathcal{F}$ is arbitrary, the approach is to define $Q(H)$ for all H from a judiciously chosen collection of versions $Q(H_n)$ for some sequence (H_n) in $\mathcal{H}$, and this requires conditions or limitations on $\mathcal{H}$. The following is the general result on existence. We shall give the proof afterward in Remark 2.11; see Chapter I, Section 2 for standard spaces.

2.7 THEOREM. *If $(\Omega, \mathcal{H})$ is a standard measurable space, then $\mathbb{P}_{\mathcal{F}}$ has a regular version.*

Conditional distributions

Let Y be a random variable taking values in some measurable space $(E, \mathcal{E})$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$. Then, the *conditional distribution of Y given $\mathcal{F}$* is any transition probability kernel $L : (\omega, B) \mapsto L_\omega(B)$ from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$ such that

$$2.8 \qquad \mathbb{P}_{\mathcal{F}}\{Y \in B\} = L(B), \qquad B \in \mathcal{E}.$$

If $\mathbb{P}_{\mathcal{F}}$ has a regular version Q, then

$$2.9 \qquad L_\omega(B) = Q_\omega\{Y \in B\}, \qquad \omega \in \Omega, \ B \in \mathcal{E},$$

defines a version L of the conditional distribution of Y given $\mathcal{F}$. In general, the problem is equivalent to finding a regular version of $\mathbb{P}_{\mathcal{F}}$ restricted to the σ-algebra generated by Y. The following is the standard result.

2.10 THEOREM. *If $(E, \mathcal{E})$ is a standard measurable space, then there exists a version of the conditional distribution of Y given $\mathcal{F}$.*

2.11 REMARK. Theorem 2.7 is a straightforward corollary of the preceding theorem: Suppose that $(\Omega, \mathcal{H})$ is standard. Define $Y(\omega) = \omega$ for all ω in Ω. Then, Theorem 2.10 applies with $(E, \mathcal{E}) = (\Omega, \mathcal{H})$, and the conditional distribution of Y given $\mathcal{F}$ is precisely the regular version of $\mathbb{P}_{\mathcal{F}}$ in view of 2.8 and the fact that $H = \{Y \in H\}$ for every H in $\mathcal{H} = \mathcal{E}$.

Proof of 2.10. First, we give the proof for $E = \bar{\mathbb{R}}$ and $\mathcal{E} = \mathcal{B}(\bar{\mathbb{R}})$. For each rational number q, let

$$C_q = \mathbb{P}_{\mathcal{F}}\{Y \le q\}.$$

We shall construct the conditional distribution L of Y given $\mathcal{F}$ from these countably many random variables C_q.

Since $\{Y \le q\} \subset \{Y \le r\}$ for $q < r$, the monotonicity of conditional expectations implies that the event $\Omega_{qr} = \{C_q \le C_r\}$ in $\mathcal{F}$ is almost sure for every pair of rationals q and r with $q < r$. Let Ω_0 be the intersection of all those Ω_{qr}; it belongs to $\mathcal{F}$ and is almost sure. Fix ω in Ω_0. The mapping $q \mapsto C_q(\omega)$ from the rationals into $[0,1]$ is increasing, and thus, for each t in $\mathbb{R}$, the limit $\bar{C}_t(\omega)$ of $C_q(\omega)$ over all rationals $q > t$ exists. The resulting function $t \mapsto \bar{C}_t(\omega)$ is a cumulative distribution function on $\mathbb{R}$, and there is a unique probability measure $\bar{L}_\omega$ on $(E, \mathcal{E})$ that admits $t \mapsto \bar{C}_t(\omega)$ as its distribution function. We define

$$L_\omega(B) = 1_{\Omega_0}(\omega)\bar{L}_\omega(B) + 1_{\Omega\setminus\Omega_0}(\omega)\delta_0(B), \quad \omega \in \Omega,\ B \in \mathcal{E},$$

where δ_0 is Dirac at 0. We proceed to show that L is as desired.

a) For each ω in Ω, L_ω is a probability measure on $(E, \mathcal{E})$.

b) Let $\mathcal{D}$ be the collection of all B in $\mathcal{E}$ for which $L(B) : \omega \mapsto L_\omega(B)$ is in $\mathcal{F}_+$. It is checked easily that $\mathcal{D}$ is a d-system. Thus, in order to show that $\mathcal{D} = \mathcal{E}$ via the monotone class theorem, it is enough to show that $[-\infty, t] \in \mathcal{D}$ for every t in $\mathbb{R}$. Fix t such, let $B = [-\infty, t]$, and note that

$$L(B) = 1_{\Omega_0} \cdot \bar{C}_t + 1_{\Omega\setminus\Omega_0}\delta_0(B) = \lim_n 1_{\Omega_0} \cdot C_{r_n} + 1_{\Omega\setminus\Omega_0}\delta_0(B),$$

where (r_n) is a sequence of rationals strictly decreasing to t. On the right side, $\Omega_0 \in \mathcal{F}$ and $\Omega \setminus \Omega_0 \in \mathcal{F}$ and $\delta_0(B)$ is a constant and every C_{r_n} is in $\mathcal{F}_+$ by choice. Hence, $L(B) \in \mathcal{F}_+$ and, therefore, $B \in \mathcal{D}$. In other words, $\omega \mapsto L_\omega(B)$ is in $\mathcal{F}_+$ for every B in $\mathcal{E}$.

c) We have shown that L is a transition probability kernel from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$. To show that it is the conditional distribution of Y given $\mathcal{F}$, there remains to show the projection property for 2.8, that is, we need to show that

$$\mathbb{P}(H \cap \{Y \in B\}) = \mathbb{E}\ 1_H L(B) \tag{2.12}$$

for H in $\mathcal{F}$ and B in $\mathcal{E}$. By the same monotone class argument as in part (b) above, it is enough to check this for $B = [-\infty, t]$ with t in $\mathbb{R}$. Fix t and let (r_n) be a sequence of rationals strictly decreasing to t. Then, by the way C_q are chosen,

$$\mathbb{P}(H \cap \{Y \in B\}) = \lim_n \mathbb{P}(H \cap \{Y \le r_n\}) = \lim_n \mathbb{E}\ 1_H C_{r_n}. \tag{2.13}$$

On the other hand, $1_H = 1_{H\cap\Omega_0}$ almost surely, and $C_{r_n}(\omega) \to \bar{C}_t(\omega) = L_\omega(B)$ for ω in Ω_0. Thus,

$$\lim \mathbb{E}\ 1_H C_{r_n} = \mathbb{E}\ 1_{H\cap\Omega_0} L(B) = \mathbb{E}\ 1_H L(B). \tag{2.14}$$

Now, putting 2.13 and 2.14 together yields 2.12 and completes the proof for the case $E = \bar{\mathbb{R}}$.

Finally, we extend the proof to the general case where $(E, \mathcal{E})$ is a standard measurable space. Then, there is an isomorphism g from E onto some Borel subset $\hat{E}$ of $[0,1]$; let $h : \hat{E} \mapsto E$ be the functional inverse of g. The preceding part applies to the real-valued random variable $g \circ Y$ and shows the existence of the conditional distribution $\hat{L} : (\omega, B) \mapsto \hat{L}_\omega(B)$ from $(\Omega, \mathcal{F})$ into $(\hat{E}, \hat{\mathcal{E}})$ for $g \circ Y$ given $\mathcal{F}$. We now put

$$L_\omega(A) = \hat{L}_\omega(h^{-1}A), \qquad \omega \in \Omega, A \in \mathcal{E}.$$

It is obvious that L is a transition probability kernel from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$, and observing that, for H in $\mathcal{F}$ and A in $\mathcal{E}$,

$$\mathbb{P}(H \cap \{Y \in A\}) = \mathbb{P}(H \cap \{g \circ Y \in h^{-1}A\}) = \mathbb{E}\ 1_H \cdot \hat{L}(h^{-1}A) = \mathbb{E}\ 1_H L(A)$$

completes the proof that L is the conditional distribution of Y given $\mathcal{F}$. □

Disintegrations

The usual method of constructing measures over a product space was discussed in Chapter I, Section 6. Here, we treat the converse problem of disintegrating a given measure to its components. In the next subsection, we shall provide probabilistic meanings to such constructions and disintegrations. We start with a brief recall.

Let $(D, \mathcal{D})$ and $(E, \mathcal{E})$ be measurable spaces. Let μ be a probability measure on $(D, \mathcal{D})$ and let K be a transition probability kernel from $(D, \mathcal{D})$ into $(E, \mathcal{E})$. Then, according to Theorem I.6.11, the following formula for positive f in $\mathcal{D} \otimes \mathcal{E}$ defines a probability measure π on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$:

$$2.15 \qquad \pi f = \int_{D \times E} \pi(dx, dy) f(x, y) = \int_D \mu(dx) \int_E K(x, dy) f(x, y).$$

Indeed, π is the unique measure (on the product space) that satisfies

$$2.16 \qquad \pi(A \times B) = \int_A \mu(dx) K(x, B), \qquad A \in \mathcal{D},\ B \in \mathcal{E}$$

In keeping with the short notation system mentioned in I.6.22, we write

$$2.17 \qquad \pi(dx, dy) = \mu(dx) K(x, dy), \qquad x \in D,\ y \in E,$$

to represent π defined by 2.15 and/or 2.16. This is the usual method of constructing a measure π on the product space from the measure μ and kernel K. In the next subsection, we shall give a probabilistic meaning to 2.17: if π is the joint distribution of X and Y, then $\mu(dx)$ is "the probability that X is in the small set dx" and $K(x, dy)$ is "the conditional probability that Y is in the small set dy given that X is equal to x."

The problem of disintegration is the converse to the construction described: given a probability measure π on the product space, find μ and K such that 2.17 holds (or, equivalently, 2.15 or 2.16 holds). The following is the general result; note that this is an exact converse to Theorem I.6.11 but for the condition that $(E, \mathcal{E})$ be a standard measurable space.

2.18 THEOREM. *Let π be a probability measure on the product space $(D \times E, \mathcal{D} \otimes \mathcal{E})$. Suppose that $(E, \mathcal{E})$ is standard. Then, there exist a probability measure μ on $(D, \mathcal{D})$ and a transition probability kernel K from $(D, \mathcal{D})$ into $(E, \mathcal{E})$ such that 2.17 holds.*

Proof. We cast the problem into a special case of Theorem 2.10. Let $W = D \times E$, $\mathcal{W} = \mathcal{D} \otimes \mathcal{E}$, $P = \pi$. On the probability space $(W, \mathcal{W}, P)$, define random variables X and Y by putting $X(w) = x$ and $Y(w) = y$ for $w = (x, y)$ in W. Let μ be the distribution of X, that is, $\mu(A) = \pi(A \times E), A \in \mathcal{D}$. Since Y takes values in the standard measurable space $(E, \mathcal{E})$, by Theorem 2.10, there is a regular version L of the conditional distribution of Y given $\mathcal{F} = \sigma X$. Note that $\mathcal{F}$ consists of measurable rectangles of the form $A \times E$; thus, a random variable V is in $\mathcal{F}_+$ if and only if $V(x, y) = v(x)$, free of y, for some function v in $\mathcal{D}_+$. It follows that $L_w(B) = K(X(w), B)$, where K is a transition probability kernel from $(D, \mathcal{D})$ into $(E, \mathcal{E})$. Now, the projection property for $L(B)$ yields, if $A \in \mathcal{D}$ and $B \in \mathcal{E}$, writing E for the integration under $P = \pi$,

$$\pi(A \times B) = E\ 1_A \circ X\ 1_B \circ Y = E\ 1_A \circ X\ K(X, B) = \int_D \mu(dx) 1_A(x) K(x, B).$$

This shows 2.15 for $f = 1_{A \times B}$, and the general case follows from a monotone class argument. □

Conditional distribution of Y given X

We return to the general setup of an arbitrary probability space $(\Omega, \mathcal{H}, \mathbb{P})$. Let X and Y be random variables taking values in the measurable spaces $(D, \mathcal{D})$ and $(E, \mathcal{E})$ respectively. By the *conditional distribution of Y given X* is meant the conditional distribution of Y given $\mathcal{F}$, where $\mathcal{F} = \sigma X$, the σ-algebra generated by X. The following is the description of such. Note that its condition is fulfilled at least when $(E, \mathcal{E})$ is standard; see the preceding theorem.

2.19 THEOREM. *Suppose that the joint distribution π of X and Y has the representation 2.17. Then, the kernel L defined by*

$$L_\omega(B) = K(X(\omega), B), \qquad \omega \in \Omega, B \in \mathcal{E},$$

is a version of the conditional distribution of Y given $\mathcal{F} = \sigma X$, and for every positive f in $\mathcal{D} \otimes \mathcal{E}$,

$$\mathbb{E}_{\mathcal{F}} f(X, Y) = \int_E K(X, dy) f(X, y).$$

Proof. The statement about L is immediate from Theorem 2.10 and the observation that $\mathcal{F} = \sigma X$ consists of $\mathcal{D}$-measurable functions of X. Then, by the meaning of L, if $h \in \mathcal{E}_+$,

$$\mathbb{E}_{\mathcal{F}} h(Y) = \int_E K(X, dy) h(y).$$

Thus, if $f = g \times h$ for some g in $\mathcal{D}_+$ and h in $\mathcal{E}_+$, the conditional determinism property allows $g(X)$ to come out of $\mathbb{E}_{\mathcal{F}}$, and we have

$$\mathbb{E}_{\mathcal{F}} f(X, Y) = g(X) \mathbb{E}_{\mathcal{F}} h(Y) = \int_E K(X, dy) f(X, y)$$

as claimed. Since measurable rectangles generate $\mathcal{D} \otimes \mathcal{E}$, the monotone class theorem completes the proof. □

The claims of the preceding theorem form the exact meaning of the phrase "given that $X = x$, the conditional probability that $Y \in B$ is equal to $K(x, B)$." The intuitive meaning, of course, is that transition probability kernels represent conditional probabilities, and that conditional probabilities are often the primary data in the construction of probability measures on product spaces.

Returning to the representation 2.17, which holds usually by construction or by a result like Theorem 2.18, we add that it holds trivially if X and Y are independent, and then $K(x, B) = \nu(B)$ is free of x. The following provides another construction leading to it.

Conditional densities

This is to mention a situation that is encountered often, especially in elementary probability. In the setup of the preceding subsection, suppose that the joint distribution π of X and Y has the form

$$\pi(dx, dy) = \mu_0(dx) \nu_0(dy) p(x, y), \qquad x \in D, y \in E, \tag{2.20}$$

where μ_0 and ν_0 are σ-finite measures on $(D, \mathcal{D})$ and $(E, \mathcal{E})$ respectively, and p is a positive function in $\mathcal{D} \otimes \mathcal{E}$; often, $D = E = \mathbb{R}^d$ and $\mu_0 = \nu_0 =$ Lebesgue. This π can be put in the form 2.17:

$$\pi(dx, dy) = [\mu_0(dx) m(x)][\nu_0(dy) k(x, y)],$$

where

$$m(x) = \int_E \nu_0(dy) p(x, y), \qquad k(x, y) = \begin{cases} p(x, y)/m(x) & \text{if } m(x) > 0, \\ \int_D \mu_0(dx') p(x', y) & \text{if } m(x) = 0. \end{cases}$$

Then the function $y \mapsto k(x, y)$ is called the conditional density (with respect to ν_0) of Y given that $X = x$.

2.21 EXAMPLE. Let Y and Z be independent and have the standard Gaussian distribution on $\mathbb{R}$. As an illustration of the computations discussed above, we now derive the conditional distribution of Y given $X = Y + Z$.

First, we find the joint distribution π of X and Y. To that end, we repeat that Y is standard Gaussian, whereas the conditional distribution of X given Y is Gaussian with mean Y and variance 1, because Z is independent of Y and X is the sum of Z and the "known" quantity Y given. Thus,

$$\pi(dx, dy) = dy \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\ dx \frac{1}{\sqrt{2\pi}} e^{-(x-y)^2/2}.$$

We also know that the distribution μ of $X = Y + Z$ is Gaussian with mean 0 and variance 2, that is,

$$\mu(dx) = dx \frac{1}{\sqrt{4\pi}} e^{-x^2/4}.$$

It follows that the conditional distribution $K(x, \cdot)$ of Y given $X = x$ is

$$K(x, dy) = \frac{\pi(dx, dy)}{\mu(dx)} = dy \ \cdot\ \frac{\sqrt{4\pi}}{\sqrt{2\pi}\sqrt{2\pi}}\ \exp\left[-\frac{y^2}{2} - \frac{(x-y)^2}{2} + \frac{x^2}{4}\right]$$

$$= dy\ \frac{1}{\sqrt{\pi}}\ \exp\left[-(y - \frac{1}{2}x)^2\right],$$

which we recognize as the Gaussian distribution with mean $x/2$ and variance $1/2$. To re-iterate, given the sum $X = Y + Z$, the conditional distribution of Y is $B \mapsto K(X, B)$, the Gaussian distribution with mean $\frac{1}{2}X$ and variance $1/2$.

Exercises

2.22 *Conditional distributions.* Let Y take values is $(E, \mathcal{E})$. Let $\mathcal{F}$ be a sub-σ-algebra of $\mathcal{H}$, let L be a transition probability kernel from $(\Omega, \mathcal{F})$ into $(E, \mathcal{E})$. As usual, for g in $\mathcal{E}_+$,

$$Lg(\omega) = \int_E L_\omega(dy) g(y), \qquad \omega \in \Omega.$$

Show that, if L is the conditional distribution of Y given $\mathcal{F}$, then, for every g in $\mathcal{E}_+$,

$$\mathbb{E}_{\mathcal{F}}\ g \circ Y = Lg.$$

2.23 *Continuation.* Let $\mathcal{F} = \sigma X$ for some random variable X with values in $(D, \mathcal{D})$. Suppose that $L_\omega(B) = K(X(\omega), B)$ for some transition probability kernel K from $(D, \mathcal{D})$ into $(E, \mathcal{E})$. Then, in the measure-kernel-function notation, show that

$$\mathbb{E}\ f \circ X\ g \circ Y = \mathbb{E}\ f \circ X\ Kg(X) = \int_D \mu(dx) f(x) \int_E K(x, dy) g(y).$$

for every f in $\mathcal{D}_+$ and g in $\mathcal{E}_+$. In particular, then,

$$\mathbb{E}_{\mathcal{F}}\ g \circ Y = Kg(X).$$

2.24 *Gamma variables.* Let Y and Z be independent gamma distributed variables with shape indices a and b respectively, and the same scale parameter c. Let $X = Y + Z$. Find the kernel K such that the conditional distribution of Y given X is $K(X, \cdot)$. In particular, for $a = b = 1$, show that $K(x, dy) = \frac{1}{x} dy$ for y in $(0, x)$.

2.25 *Gaussian with gamma variance.* Let X, Y, Z be as in Exercise II.2.30. Show that the conditional distribution of Z given X is $K(X, \cdot)$ where

$$K(x, dz) = dz \frac{1}{\sqrt{2\pi x}} e^{-z^2/2x}, \qquad x > 0, z \in \mathbb{R}.$$

Show that

$$\mathbb{E}_X \; e^{irZ} = e^{-r^2 X/2},$$

and that

$$\mathbb{E} \; e^{irZ} = \mathbb{E} \; \mathbb{E}_X \; e^{irZ} = \mathbb{E} \; e^{-r^2 X/2} = \left(\frac{2c}{2c + r^2} \right)^a .$$

This should explain the workings of II.2.30.

2.26 *Independence.* Let X and Y be independent and taking values in $(D, \mathcal{D})$ and $(E, \mathcal{E})$ respectively. Let $Z = h(X, Y)$ for some h in $\mathcal{D} \otimes \mathcal{E}$. Then, show that, the conditional distribution of Z given X is given as $K(X, \cdot)$ where

$$K(x, B) = \mathbb{P}\{Z \in B \mid X = x\} = \mathbb{P}\{h(x, Y) \in B\}.$$

Moral: Given that $X = x$, in most situations, we are allowed to replace X with x in our computations. Of course, to repeat the point of Example 1.22, then, $\mathbb{E}_{\sigma X} Z = \bar{h} \circ X$ where

$$\bar{h}(x) = \mathbb{E}[Z \mid X = x] = \mathbb{E} \; h(x, Y), \qquad x \in D.$$

2.27 *Stochastic process at a random time.* This is a far-fetched corollary of the preceding. Let $Y = (Y_t)_{t \in \mathbb{R}_+}$ be a stochastic process with state space $(E, \mathcal{E})$. Suppose that the mapping $(t, \omega) \mapsto Y_t(\omega)$ from $\mathbb{R}_+ \times \Omega$ into E is measurable relative to $\mathcal{B}_{\mathbb{R}_+} \otimes \mathcal{H}$ and $\mathcal{E}$; this condition is fulfilled automatically if $t \mapsto Y_t(\omega)$ is right continuous for every ω, assuming that E is topological. We think of $\mathbb{R}_+$ as the time-set and of Y_t as the state of some system at the fixed time t. Now, let T be a random time, that is, a random variable taking values in $\mathbb{R}_+$. We are interested in the state of the system at that random time, namely, Y_T.

a) Show that $Y_T \; : \; \omega \mapsto Y_{T(\omega)}(\omega)$ is a random variable taking values in $(E, \mathcal{E})$.

b) Assume that T and Y are independent. Show that, for f in $\mathcal{E}_+$,

$$\mathbb{E}_{\sigma T} \; f \circ Y_T = g \circ T, \qquad \mathbb{E} \; f \circ Y_T = \mathbb{E} \; g \circ T,$$

where

$$g(t) = \mathbb{E} \; f \circ Y_t.$$

3 Conditional Independence

This is an important generalization of the concept of independence, and it is reduced to independence when the conditioning σ-algebra is trivial.

Let $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ be sub-σ-algebras of $\mathcal{H}$. Then $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are said to be *conditionally independent given* $\mathcal{F}$ if

$$\mathbb{E}_{\mathcal{F}}\ V_1 \cdots V_n = \mathbb{E}_{\mathcal{F}}\ V_1 \cdots \mathbb{E}_{\mathcal{F}}\ V_n \tag{3.1}$$

for all positive random variables $V_1, \ldots, V_n$ in $\mathcal{F}_1, \ldots, \mathcal{F}_n$ respectively.

This definition compares to the definition II.5.1 of independence except for the substitution of $\mathbb{E}_{\mathcal{F}}$ for $\mathbb{E}$. Hence, all the results for independence have their counterparts for conditional independence given $\mathcal{F}$, and these counterparts are obtained by replacing $\mathbb{E}$ with $\mathbb{E}_{\mathcal{F}}$ throughout. Of course, if $\mathcal{F}$ is trivial, that is, if $\mathcal{F} = \{\emptyset, \Omega\}$, then $\mathbb{E}_{\mathcal{F}} = \mathbb{E}$ and conditional independence given $\mathcal{F}$ is the same as independence.

Heuristically, independence of $\mathcal{F}_1$ and $\mathcal{F}_2$ meant that information $\mathcal{F}_1$ is useless as far as estimating the quantities determined by $\mathcal{F}_2$. Similarly for conditional independence given $\mathcal{F}$: given the information $\mathcal{F}$, the further information provided by $\mathcal{F}_1$ is useless in estimating quantities determined by $\mathcal{F}_2$. Here is the precise version of this remark.

3.2 Proposition. *The following are equivalent:*

a) $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *are conditionally independent given* $\mathcal{F}$.
b) $\mathbb{E}_{\mathcal{F}\vee\mathcal{F}_1} V_2 = \mathbb{E}_{\mathcal{F}} V_2$ *for every positive* V_2 *in* $\mathcal{F}_2$.
c) $\mathbb{E}_{\mathcal{F}\vee\mathcal{F}_1} V_2 \in \mathcal{F}$ *for every positive* V_2 *in* $\mathcal{F}_2$.

Proof. Throughout V, V_1, V_2 are positive and in $\mathcal{F}, \mathcal{F}_1, \mathcal{F}_2$ respectively. Consider (a). It is equivalent to having

$$\mathbb{E}_{\mathcal{F}}\ V_1 V_2 = (\mathbb{E}_{\mathcal{F}}\ V_1)(\mathbb{E}_{\mathcal{F}}\ V_2) = \mathbb{E}_{\mathcal{F}}\ (V_1\ \mathbb{E}_{\mathcal{F}}\ V_2),$$

where the last equality is justified by the conditional determinism property. This is in turn equivalent to having, by definition,

$$\mathbb{E}\ V V_1 V_2 = \mathbb{E}\ V V_1\ \mathbb{E}_{\mathcal{F}}\ V_2,$$

which is equivalent to (b), since random variables of the form VV_1 generate the σ-algebra $\mathcal{F} \vee \mathcal{F}_1$. Thus, (a) $\iff$ (b). It is obvious that (b) $\implies$ (c). Conversely, if (c) holds, then

$$\mathbb{E}_{\mathcal{F}\vee\mathcal{F}_1} V_2 = \mathbb{E}_{\mathcal{F}}\ \mathbb{E}_{\mathcal{F}\vee\mathcal{F}_1} V_2 = \mathbb{E}_{\mathcal{F}}\ V_2$$

by the conditional determinism property followed by repeated conditioning. Hence, (c) $\implies$ (b). □

3.3 Remark. The definition of conditional independence and the preceding theorem are stated in terms of positive random variables. As usual, this is because we want to avoid the trite but annoying considerations involved with arbitrary variables.

3.4 Warning. It is possible that $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent, but fail to be conditionally independent given $\mathcal{F}$. Here is an extreme example of this state of affairs: Let X and Y be independent and identically distributed positive random variables and let $Z = X + Y$. Then, $\mathcal{F}_1 = \sigma X$ and $\mathcal{F}_2 = \sigma Y$ are independent, but they are not independent given $\mathcal{F} = \sigma Z$. In fact, in this case, $\mathbb{E}_{\mathcal{F}\vee\mathcal{F}_1} Y = Y$ whereas $\mathbb{E}_{\mathcal{F}} Y = \frac{1}{2} Z$.

Conditional independence of random variables etc.

The definition of conditional independence is extended to various settings by following the conventions for independence. For example, for an arbitrary index set T, the sub-σ-algebras $\mathcal{F}_t,\ t \in T$, are said to be conditionally independent given $\mathcal{F}$ if $\mathcal{F}_{t_1}, \ldots, \mathcal{F}_{t_n}$ are so given $\mathcal{F}$ for all integers $n \geq 2$ and choices $t_1, \ldots, t_n$ in T. Random variables $X_t,\ t \in T$, are said to be conditionally independent given $\mathcal{F}$ if the σ-algebras they generate are so given $\mathcal{F}$. If $\mathcal{F} = \sigma X$, then "given $\mathcal{F}$"is replaced by "given X". And so on.

Exercises

It will be convenient to introduce a shorthand system for conditional independence. We propose the notation $a]c[b$ for "a and b are conditionally independent given c". The notation conveys our mental image that a and b are kept apart (to act independent of each other) by the force of c between them. The arguments a, b, c are σ-algebras ordinarily, but some or all could be random variables, events, collections, etc. If c is the trivial σ-algebra, then we omit it from notation and write $a][b$, which means that a and b are independent.

Throughout the following, letters like $\mathcal{F}$ and $\mathcal{G}$ and so on are sub-σ-algebras of $\mathcal{H}$, and X and Y and so on are random variables.

3.5 *Arithmetic of conditional independence.* Show the following:

$$\begin{aligned} \mathcal{F}_1]\mathcal{F}[\mathcal{F}_2 &\iff \mathcal{F}_2]\mathcal{F}[\mathcal{F}_1 \\ &\iff \mathcal{F}_1 \vee \mathcal{F}]\mathcal{F}[\mathcal{F}_2 \iff \mathcal{F}_1]\mathcal{F}[\mathcal{F} \vee \mathcal{F}_2 \\ &\iff \mathcal{G}_1]\mathcal{F}[\mathcal{G}_2 \qquad \text{for all } \mathcal{G}_1 \subset \mathcal{F}_1 \text{ and } \mathcal{G}_2 \subset \mathcal{F}_2. \end{aligned}$$

3.6 *Proposition 3.2.* Checks for conditional independence can always be reduced to the pairwise case of Proposition 3.2: Show that $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are conditionally independent given $\mathcal{F}$ if and only if $\mathcal{F}_1 \vee \cdots \vee \mathcal{F}_k]\mathcal{F}[\mathcal{F}_{k+1}$ for $k = 1, 2, \ldots, n-1$.

3.7 *Continuation.* As a corollary to Proposition 3.2, show the following. Supposing that $\mathcal{F} \subset \mathcal{F}_1$, we have $\mathcal{F}_1]\mathcal{F}[\mathcal{F}_2$ if and only if $\mathbb{E}_{\mathcal{F}_1} V_2 \in \mathcal{F}$ for all positive V_2 in $\mathcal{F}_2$.

3.8 *Uses of the monotone class theorems.* Show that the following are equivalent,

a) $\mathcal{F}_1]\mathcal{F}[\mathcal{F}_2$.

b) $\mathbb{P}_{\mathcal{F}\vee\mathcal{F}_1} H \in \mathcal{F}_+$ for every event H in some p-system that generates $\mathcal{F}_2$.

c) $\mathbb{P}_{\mathcal{F}}(H_1 \cap H_2) = (\mathbb{P}_{\mathcal{F}} H_1)(\mathbb{P}_{\mathcal{F}} H_2)$ for every H_1 in some p-system generating $\mathcal{F}_1$ and every H_2 in some p-system generating $\mathcal{F}_2$.

3.9 *Continuation.* Show that $\mathcal{F}_1]\mathcal{F}[\{X_t : t \in T\}$ if and only if $\mathcal{F}_1]\mathcal{F}[\{X_t : t \in T'\}$ for every finite $T' \subset T$. Here, each X_t is a random variable taking values in some measurable space $(E_t, \mathcal{E}_t)$.

3.10 *Continuation.* Let $X_1, \ldots, X_n$ be random variables taking values in $(E_1, \mathcal{E}_1)$, $\ldots$, $(E_n, \mathcal{E}_n)$. In view of Proposition 3.2, $\mathcal{F}_1]\mathcal{F}[(X_1, \ldots, X_n)$ if and only if $\mathbb{E}_{\mathcal{F}_1\vee\mathcal{F}} V$ is in $\mathcal{F}$ for every positive V in $\sigma(X_1, \ldots, X_n)$, that is, for every V having the form $V = f(X_1, \ldots, X_n)$ for some positive f in $\mathcal{E}_1 \otimes \cdots \otimes \mathcal{E}_n$. Show that, in fact, it is sufficient to check the condition for f having the form

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n), \qquad x_1 \in E_1, \ldots, x_n \in E_n,$$

where each f_k is in $\mathcal{E}_k$ and positive. Furthermore, each f_k can be taken to be bounded, or the indicator of an arbitrary set in $\mathcal{E}_k$, or the indicator of a set belonging to some p-system that generates $\mathcal{E}_k$, or, assuming that E_k is topological and $\mathcal{E}_k = \mathcal{B}(E_k)$, a bounded continuous function. Finally, if $E_k = \mathbb{R}^d$ with $\mathcal{E}_k = \mathcal{B}(\mathbb{R}^d)$, one can take $f_k(x) = \exp(ir_k \cdot x)$ with r_k in $\mathbb{R}^d$ and $r \cdot x$ denoting the inner product of r and x.

3.11 *Conditional independence as independence.* Suppose that there is a regular version Q of the conditional probability $\mathbb{P}_{\mathcal{F}}$. Suppose that $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent under the probability measure Q_ω, this being true for $\mathbb{P}$-almost every ω, that is, there exists an almost sure event Ω_0 such that

$$Q_\omega V_1 V_2 = (Q_\omega V_1)(Q_\omega V_2)$$

for every ω in Ω_0 and positive V_1 in $\mathcal{F}_1$ and positive V_2 in $\mathcal{F}_2$. Then, $\mathcal{F}_1]\mathcal{F}[\mathcal{F}_2$. The converse holds as well if, further, $\mathcal{F}_1$ and $\mathcal{F}_2$ are separable.

4 Construction of Probability Spaces

Our object is the construction of probability spaces and random variables corresponding to certain random experiments. We describe two basic constructions: Ionescu-Tulcea's and Kolmogorov's. Together, they show the existence of all the probability spaces that were ever needed. In particular, they yield the existence and construction of the Lebesgue measure on $(0, 1)$, and therefore the existence of all measures that were ever discussed.

To highlight the issue involved, recall from Chapter III our oft-repeated refrain, "let $(\Omega, \mathcal{H}, \mathbb{P})$ be a probability space and let $X_1, X_2, \ldots$ be independent random variables with distributions $\mu_1, \mu_2, \ldots$". Do such things exist? After all, related to each random variable there are at least as many events as there are points in $\mathbb{R}$, and there are infinitely many random variables, and $\mathcal{H}$ must include all those events and their complements and all countable unions and intersections of them. Now think of the conditions on $\mathbb{P}$: it must assign a probability to each event, and do it so that countable additivity condition is fulfilled for every disjointed sequence of events. Moreover, independence of the random variables requires that certain multiplicative rules be obeyed. What if the conditions are too onerous for $\mathbb{P}$ to bear, that is, what if there can be no such $\mathbb{P}$?

The first theorem below shows, as a special case, that such things do indeed exist. Proofs are not enlightening, but the constructions leading to the theorem clarifies many of the concepts discussed earlier. In particular, note the natural appearance of conditional probabilities as primary data, as things from which $\mathbb{P}$ is constructed rather than as things derived from $\mathbb{P}$.

Construction of chains: description of data and goal

The data for the problem consist of some measurable spaces $(E_0, \mathcal{E}_0)$, $(E_1, \mathcal{E}_1), \ldots$, some probability measure μ on $(E_0, \mathcal{E}_0)$, and some transition probability kernels $K_1, K_2, \ldots$ where, for each n in $\mathbb{N}$, the kernel K_{n+1} is from

$$(F_n^o, \mathcal{F}_n^o) = (E_0 \times \cdots \times E_n, \mathcal{E}_0 \otimes \cdots \otimes \mathcal{E}_n) \tag{4.1}$$

into $(E_{n+1}, \mathcal{E}_{n+1})$.

We regard the data as follows. A random experiment is being conducted. It consists of an infinite chain of trials. The set E_n is the space of all possible outcomes of the n^{th} trial. Our abilities to detect and discern are such that we can tell, for each A in $\mathcal{E}_n$, whether the n^{th} trial's outcome is in A. The law governing the initial trial is described by μ; for each A in $\mathcal{E}_0$, the probability is $\mu(A)$ that the outcome of the initial trial belongs to A. Having performed the trials up to and including the n^{th}, if the outcomes were $x_0, \ldots, x_n$ in $E_0, \ldots, E_n$ respectively, then the law governing the next trial is such that $K_{n+1}(x_0, \ldots, x_n; A)$ is the probability that the outcome belongs to the set A in $\mathcal{E}_{n+1}$.

The goal is to construct a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ that models the experiment described.

Construction and analysis

Each possible outcome of the experiment is a sequence $\omega = (x_0, x_1, \ldots)$ with x_n in E_n for each n in $\mathbb{N}$. Accordingly, we define the sample space Ω and the collection $\mathcal{H}$ of all events by

$$(\Omega, \mathcal{H}) = \otimes_{n \in \mathbb{N}} (E_n, \mathcal{E}_n). \tag{4.2}$$

We let $X_0, X_1, \dots$ be the coordinate variables: for each n,

$$X_n(\omega) = x_n \qquad \text{if } \omega = (x_0, x_1, \dots). \tag{4.3}$$

Obviously, X_n takes values in $(E_n, \mathcal{E}_n)$; for each outcome ω of the experiment, $X_n(\omega)$ is the result of the n^{th} trial. Similarly, $Y_n = (X_0, \dots, X_n)$ denotes the result of the trials up to and including the n^{th}; it is a random variable taking values in $(F_n^o, \mathcal{F}_n^o)$ defined by 4.1.

There remains to "construct" the probability measure $\mathbb{P}$ consistent with the data. We start with the properties it should have. The interpretation we gave to $\mu, K_1, K_2, \dots$ suggests that the distribution of Y_n be the probability measure π_n on $(F_n^o, \mathcal{F}_n^o)$ given by (see I.6.21 *et seq.*)

$$\begin{aligned} &\pi_n(dx_0, \dots, dx_n) \\ &\quad = \mu(dx_0) K_1(x_0, dx_1) K_2(x_0, x_1; dx_2) \cdots K_n(x_0, \dots, x_{n-1}; dx_n). \end{aligned} \tag{4.4}$$

Let $\mathcal{F}_n = \sigma Y_n$. Every H in $\mathcal{F}_n$ has the form

$$H = \{Y_n \in B\} = B \times E_{n+1} \times \cdots, \quad B \in \mathcal{F}_n^o, \tag{4.5}$$

and then H is said to be a *cylinder with base* B in $\mathcal{F}_n^o$. Since the measure $\mathbb{P}$ being sought must yield π_n as the distribution of Y_n, we must have

$$\mathbb{P}(H) = \pi_n(B) \qquad \text{if } H \in \mathcal{F}_n,\ H \text{ has base } B \in \mathcal{F}_n^o. \tag{4.6}$$

This specifies $\mathbb{P}(H)$ for every H in $\mathcal{H}^o = \cup_n \mathcal{F}_n$ in a consistent manner: if H is a cylinder with base B in $\mathcal{F}_n^o$ and, at the same time, with base A in $\mathcal{F}_m^o$ for some $m < n$, then $B = A \times E_{m+1} \times \cdots \times E_n$ and 4.4 implies that $\pi_n(B) = \pi_m(A) = \mathbb{P}(H)$ unambiguously. There remains to show that there is $\mathbb{P}$ satisfying 4.6.

Ionescu-Tulcea's theorem

4.7 THEOREM. *There exists a unique probability measure* $\mathbb{P}$ *on* $(\Omega, \mathcal{H})$ *such that 4.6 holds.*

4.8 REMARKS. a) Let $\mathbb{P}$ be as promised by the theorem. Then, 4.4 and 4.6 imply that

$$\mathbb{P}_{\mathcal{F}_m}\{X_{m+1} \in A\} = K_{m+1}(X_0, \dots, X_m; A), \qquad A \in \mathcal{E}_{m+1}.$$

Thus the essential data $K_1, K_2, \dots$ provide the conditional distributions of the next variable X_{m+1} given the past history $\mathcal{F}_m$ for each m in $\mathbb{N}$.

b) Fix n in $\mathbb{N}$. Fix H in $\mathcal{F}_n$. Let B be its base in $\mathcal{F}_n^o$. It follows from 4.4 and 4.6 that

$$\mathbb{P}_{\mathcal{F}_m}(H) = Q_m(X_0, \dots, X_m; H)$$

where the kernels Q_m from $(F_m^o, \mathcal{F}_m^o)$ into $(\Omega, \mathcal{H})$ satisfy

4.9 $$Q_m(x_0, \dots, x_m; H) = 1_B(x_0, \dots, x_n) = 1_H(x_0, x_1, \dots) \qquad \text{if } m \geq n,$$

4.10 $$Q_m(x_0, \dots, x_m; H) = \int_{E_{m+1}} K(x_0, \dots, x_m; dx_{m+1}) Q_{m+1}(x_0, \dots, x_{m+1}; H),$$

if $0 \leq m < n$. Moreover,

4.11 $$\mathbb{P}(H) = \int_{E_0} \mu(dx_0) Q_0(x_0; H).$$

Of these, 4.9 and 4.11 are obvious. To see 4.10, we use repeated conditioning $\mathbb{E}_{\mathcal{F}_m} = \mathbb{E}_{\mathcal{F}_m} \mathbb{E}_{\mathcal{F}_{m+1}}$ to get

$$Q_m(X_0, \dots, X_m; H) = \mathbb{E}_{\mathcal{F}_m} Q_{m+1}(X_0, \dots, X_{m+1}; H)$$

and evaluate the right side using part (a) above with Theorem 2.19.

Proof. a) We start by noting that $\mathcal{H}^o = \cup \mathcal{F}_n$ is an algebra and that it generates $\mathcal{H}$. In view of 4.6, the mapping $H \mapsto \mathbb{P}(H)$ from $\mathcal{F}_n$ into $[0, 1]$ is a probability measure on $(\Omega, \mathcal{F}_n)$, and this is true for every n in $\mathbb{N}$. It follows that the mapping $\mathbb{P}$ from $\mathcal{H}^o$ into $[0, 1]$ is finitely additive: if G and H are in $\mathcal{H}^o$ and disjoint, then there is n such that both of them belong to $\mathcal{F}_n$, and therefore, $\mathbb{P}(G \cup H) = \mathbb{P}(G) + \mathbb{P}(H)$. We shall show below that

4.12 $$(H_k) \subset \mathcal{H}^o \text{ and } H_k \searrow \emptyset \quad \Longrightarrow \quad \mathbb{P}(H_k) \searrow 0.$$

Once this is shown, the existence of $\mathbb{P}$ on $(\Omega, \mathcal{H})$ satisfying 4.6 for all n will follow from the standard theorems on extensions of measures from algebras to σ-algebras (see Caratheodory's theorem, I.3.19). Uniqueness of $\mathbb{P}$ is immediate from Proposition I.3.7 since $\mathcal{H}^o$ is a p-system generating $\mathcal{H}$.

b) Each H in $\mathcal{H}^o$ is a cylinder with some base B in $\mathcal{F}_n^o$ for some n in $\mathbb{N}$; and, if so, we define $Q_m(x_0, \dots, x_m; H)$ for m in $\mathbb{N}$ and $(x_0, \dots, x_m)$ in F_m^o starting with 4.9 and continuing with 4.10 iteratively downward, and observe that 4.11 holds and that 4.10 holds for all $m < n$ and $m \geq n$.

To show 4.12, pick (H_k) as described. Then $\mathbb{P}(H_k)$ is well-defined for each k and decreases with k, since $\mathbb{P}$ is finitely additive on $\mathcal{H}^o$. Suppose for the moment that $\lim_k \mathbb{P}(H_k) > 0$; we shall show that this leads to a contradiction.

Replace H in 4.11 by H_k, take limits as $k \longrightarrow \infty$ on both sides, and pass the limit on the right side inside the integral with an appeal to the bounded convergence theorem. The limit on the left side is strictly positive by assumption; so there must exist x_0^* in E_0 such that $a_0 = \lim_k Q_0(x_0^*, H_k)$ is strictly positive. We now make the induction hypothesis that there exists $(x_0^*, \dots, x_m^*)$ in F_m^o such that

4.13 $$a_m = \lim_k Q_m(x_0^*, \dots, x_m^*; H_k) > 0.$$

Recall that 4.10 holds for all m, and by the bounded convergence theorem,

$$a_m = \int_{E_{m+1}} K_{m+1}(x_0^*, \dots, x_m^*; dx_{m+1}) \lim_k Q_{m+1}(x_0^*, \dots, x_m^*, x_{m+1}; H_k).$$

Since $a_m > 0$, there must exist x_{m+1}^* in E_{m+1} such that 4.13 holds for $m+1$ as well. Thus, in view of 4.9, we have shown the existence of a sequence $\omega^* = (x_0^*, x_1^*, \dots)$ in Ω such that $\lim_k 1_{H_k}(\omega^*)$ is strictly positive and, therefore, is equal to 1. This means that $\omega^* \in H_k$ for all k large enough and, thus, for all k since (H_k) is decreasing. In other words, $w^* \in \cap H_k$, which contradicts the hypothesis that $H_k \searrow \emptyset$. Hence, we must have $\lim_k \mathbb{P}(H_k) = 0$ as needed to complete the proof of 4.7. □

Initial distribution

It is often desirable to treat the initial distribution μ as a variable rather than as a part of the given data. To that end, it is customary to write $\mathbb{P}^\mu$ for the probability measure $\mathbb{P}$ of the preceding theorem. Of course, to each probability measure μ on $(E_0, \mathcal{E}_0)$ there corresponds a unique probability measure $\mathbb{P}^\mu$ on $(\Omega, \mathcal{H})$. In particular, let $\mathbb{P}^x$ denote $\mathbb{P}^\mu$ when $\mu = \delta_x$, Dirac measure sitting at the point x in E_0. It is clear from 4.4 and 4.6 that $x \mapsto \mathbb{P}^x(H)$ is $\mathcal{E}_0$-measurable for each H in $\mathcal{H}$. Thus, $(x, H) \mapsto \mathbb{P}^x(H)$ is a transition probability kernel from $(E_0, \mathcal{E}_0)$ into $(\Omega, \mathcal{H})$ and

$$\mathbb{P}^\mu(H) = \int_{E_0} \mu(dx)\mathbb{P}^x(H), \qquad H \in \mathcal{H}, \tag{4.14}$$

for every measure μ on $(E_0, \mathcal{E}_0)$. For fixed x in E_0, the measure $\mathbb{P}^x$ is called the *probability law* of the chain $X = (X_n)$ started at x, since

$$\mathbb{P}^x\{X_0 = x\} = 1 \tag{4.15}$$

provided that the singleton $\{x\}$ belong to $\mathcal{E}_0$ to ensure that $\{X_0 = x\}$ is an event. It can be also regarded as the conditional law of X given that $X_0 = x$.

Kolmogorov extension theorem

As opposed to a chain, we now consider the question of existence for a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ that can carry a process $\{X_t : t \in I\}$ where the index set I is arbitrary. In this case, Kolmogorov's theorem is the most general known, but requires that all the X_t take values in the same space $(E, \mathcal{E})$ and that $(E, \mathcal{E})$ be a standard measurable space.

Let I be an arbitrary index set. Let $(E, \mathcal{E})$ be a measurable space. The data are the probability measures π_J, one for each finite subset J of I, on the product space $(E^J, \mathcal{E}^J)$. The goal is to construct a probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and a stochastic process $X = (X_t)_{t \in I}$ over it such that π_J is the distribution of the random variable $X_J = (X_t)_{t \in J}$ for each finite $J \subset I$.

We start by letting $(\Omega, \mathcal{H}) = (E, \mathcal{E})^I$, that is, Ω is the collection of all functions $t \mapsto \omega(t)$ from I into E, and $\mathcal{H}$ is the σ-algebra generated by the finite-dimensional measurable rectangles. We define the X_t to be the coordinate variables; that is, $X_t(\omega) = \omega(t)$ for all t and ω. Obviously, each X_t is measurable with respect to $\mathcal{H}$ and $\mathcal{E}$, and in fact $\mathcal{H} = \sigma\{X_t : t \in I\}$.

For $I \supset J \supset K$, we let p_{JK} denote the natural projection from E^J onto E^K. For instance, if $J = (s, t, u)$ and $K = (u, t)$, then $p_{IJ}(\omega) = (\omega(s), \omega(t), \omega(u))$ and $p_{JK}(x, y, z) = (z, y)$ and $p_{IK}(\omega) = (\omega(u), \omega(t))$. We let $\mathcal{I}_f$ denote the collection of all finite sequences of elements of I, and $\mathcal{I}_c$ the collection of all infinite (countable) sequences.

The probability measure $\mathbb{P}$ we are seeking will be the probability law of X; accordingly, we want

$$\mathbb{P}\{X_J \in A\} = \pi_J(A), \qquad A \in \mathcal{E}^J,\ J \in \mathcal{I}_f. \tag{4.16}$$

This requires that the finite-dimensional distributions be consistent:

$$\pi_K = \pi_J \circ p_{JK}^{-1}, \qquad K \subset J \in \mathcal{I}_f, \tag{4.17}$$

since $X_K = p_{JK} \circ X_J$ for $K \subset J$. The following is the Kolmogorov extension theorem.

4.18 Theorem. *Suppose that $(E, \mathcal{E})$ is a standard measurable space and that $\{\pi_J : J \in \mathcal{I}_f\}$ satisfies the consistency condition 4.17. Then, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{H})$ such that 4.16 holds.*

Proof. We start by constructing a probability measure P_J on $(E, \mathcal{E})^J$ for every J in $\mathcal{I}_c$. Fix J so, say, $J = (t_0, t_1, \ldots)$, and let $J_n = (t_0, \ldots, t_n)$ and $\hat{\pi}_n = \pi_{J_n}$. Observe that the $\hat{\pi}_n$ have the representation 4.4: this is trivial for $n = 0$, and assuming it is true for n, it follows from the disintegration theorem 2.18 with $(D, \mathcal{D}) = (E, \mathcal{E})^{J_n}$ that it is true for $n + 1$ as well (this is where the standardness of $(E, \mathcal{E})$ gets used). Thus, Ionescu-Tulcea's theorem applies to show that there exists a probability measure P_J on $(E, \mathcal{E})^J$ such that the image of P_J under p_{JJ_n} is $\hat{\pi}_n$ for each n.

In fact, for $K \subset J \in \mathcal{I}_c$, the probability measures P_K and $P_J \circ p_{JK}^{-1}$ coincide over the finite-dimensional cylinders in $\mathcal{E}^K$ with bases in $\mathcal{E}^L$ for finite subsets L of K, and hence, since such cylinders generate $\mathcal{E}^K$, we have $P_K = P_J \circ p_{JK}^{-1}$.

By Proposition II.4.6, for every H in $\mathcal{H}$ there is J in $\mathcal{I}_c$ such that $H = \{X_J \in A\}$ for some A in $\mathcal{E}^J$; then we put $\mathbb{P}(H) = P_J(A)$. This assignment is without ambiguities: if $H = \{X_J \in A\} = \{X_K \in B\}$ for A in $\mathcal{E}^J$ and B in $\mathcal{E}^K$ for some J and K in $\mathcal{I}_c$, then $L = J \cup K$ is in $\mathcal{I}_c$, and

$$P_J(A) = P_L(p_{LJ}^{-1} A) = P_L(p_{LK}^{-1} B) = P_K(B).$$

We now show that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{H})$. Let $(H_n) \subset \mathcal{H}$ be disjointed and have union H. For each n, there is J_n in $\mathcal{I}_c$ such that $H_n = \{X_{J_n} \in A_n\}$ for some A_n belonging to $\mathcal{E}^{J_n}$. We may assume that the

J_n are all the same by replacing J_n with $J = \cup J_n$; and, then, $H = \{X_J \in A\}$ with $A = \cup A_n$, the A_n being disjoint. Hence, by the countable additivity of P_J,

$$\mathbb{P}(H) = P_J(A) = \sum_n P_J(A_n) = \sum_n \mathbb{P}(H_n).$$

So, $\mathbb{P}$ is countably additive and $\mathbb{P}(\Omega) = 1$ obviously. Finally, $\mathbb{P}$ is the unique probability measure satisfying 4.16 since the cylinders $\{X_J \in A\}$ form a p-system that generates $\mathcal{H}$. □

5 Special Constructions

This section is devoted to special cases of the probability spaces constructed in the preceding section, as well as certain alternatives to such constructions. The setup and notations are carried over from the preceding section.

Independent sequences

In the setup for chains, suppose that the transition kernels K_n have the following form: for each n in $\mathbb{N}$ there is a probability measure μ_n on $(E_n, \mathcal{E}_n)$ such that $\mu = \mu_0$ and

$$K_{n+1}(x_0, \ldots, x_n; A) = \mu_{n+1}(A), \quad A \in \mathcal{E}_{n+1}. \tag{5.1}$$

Heuristically, this corresponds to the situation where the law governing the $(n+1)^{\text{th}}$ trial is independent of the results $x_0, \ldots, x_n$ of the previous trials. Then, 4.4 becomes

$$\pi_n(dx_0, \ldots, dx_n) = \mu_0(dx_0)\mu_1(dx_1) \cdots \mu_n(dx_n), \tag{5.2}$$

that is, π_n is the product of $\mu_0, \ldots, \mu_n$. Equivalently, then, the random variables $X_0, \ldots, X_n$ are independent for each n, and hence, $(X_n)_{n\in\mathbb{N}}$ is an independency. In this case, the probability space $(\Omega, \mathcal{H}, \mathbb{P})$ is said to be the *product* of the probability spaces $(E_n, \mathcal{E}_n, \mu_n)$, $n \in \mathbb{N}$, and the following notation is used to express it:

$$(\Omega, \mathcal{H}, \mathbb{P}) = \otimes_{n\in\mathbb{N}}(E_n, \mathcal{E}_n, \mu_n). \tag{5.3}$$

Variations where $\mathbb{N}$ is replaced by $\mathbb{N}^* = \{1, 2, \ldots\}$ or by $\mathbb{Z} = \{\ldots, -1, 0, 1, 2, \ldots\}$ or by some other countable set are self-explanatory.

A further special case is where the spaces $(E_n, \mathcal{E}_n, \mu_n)$ are the same for all n, that is, say

$$(E_n, \mathcal{E}_n, \mu_n) = (E, \mathcal{E}, \mu), \quad n \in \mathbb{N}. \tag{5.4}$$

Then, instead of 5.3, it is usual to write

$$(\Omega, \mathcal{H}, \mathbb{P}) = (E, \mathcal{E}, \mu)^{\mathbb{N}}. \tag{5.5}$$

In this case, the coordinate variables $X_0, X_1, \ldots$ take values in the same measurable space, and, in addition to being independent, they are identically distributed.

Markov chains

We start with the most useful and the most common case: This is where all the trials are on the same space, that is,

5.6 $$(E_n, \mathcal{E}_n) = (E, \mathcal{E}), \quad n \in \mathbb{N},$$

and there is a Markov kernel P on $(E, \mathcal{E})$, that is, a transition probability kernel P from $(E, \mathcal{E})$into $(E, \mathcal{E})$, such that

5.7 $$K_{n+1}(x_0, \ldots, x_n; A) = P(x_n, A)$$

for all n in $\mathbb{N}$ and $x_0, \ldots, x_n$ in E and A in $\mathcal{E}$. In other words, the probability law governing the $(n+1)^{\text{th}}$ trial is independent of n and of all the previous results $x_0, \ldots, x_{n-1}$ except the result x_n of the n^{th} trial. Then, $X = (X_n)_{n\in\mathbb{N}}$ is said to be a *Markov chain* over $(\Omega, \mathcal{H}, \mathbb{P})$ with state space $(E, \mathcal{E})$ and initial distribution μ and transition kernel P.

Moreover, the kernel P is considered to be the only ingredient defining the law of X, and μ is treated as a variable by utilizing the notation $\mathbb{P}^\mu$ and $\mathbb{P}^x$ mentioned around 4.14.

The term "Markov chain" refers to the property, obvious from 5.7, that

5.8 $$\mathbb{P}\{X_{n+1} \in A \mid \mathcal{F}_n\} = P(X_n, A), \quad A \in \mathcal{E},$$

which displays $P(x, A)$ as the conditional probability that $X_{n+1} \in A$ given that $X_n = x$. Note the independence of this probability from n; if this last point needs to be emphasized, then X is said to be a *time-homogeneous* Markov chain.

5.9 *Non-canonical Markov chains.* Assuming 5.6 and 5.7, the construction given in the preceding section for $\Omega, \mathcal{H}, (\mathcal{F}_n), (X_n), \mathbb{P}$ is called the canonical setting of a Markov chain X with state space $(E, \mathcal{E})$, initial distribution μ, and transition kernel P. More generally, given a chain $X = (X_n)_{n\in\mathbb{N}}$ defined over some probability space $(\Omega, \mathcal{H}, \mathbb{P})$ and having some state space $(E, \mathcal{E})$, and given some filtration $(\mathcal{F}_n)$ on $(\Omega, \mathcal{H})$ such that X_n is measurable with respect to $\mathcal{F}_n$ and $\mathcal{E}$ for each n, the chain X is said to be a time-homogeneous Markov chain with transition kernel P provided that 5.8 hold for each n in $\mathbb{N}$. We shall see several such non-canonical examples below.

5.10 *Time-inhomogeneous Markov chains.* Returning to the constructions of the previous section, assume 5.6 and replace 5.7 with

5.11 $$K_{n+1}(x_0, \ldots, x_n; A) = P_{n+1}(x_n, A),$$

where $P_1, P_2, \ldots$ are Markov kernels on $(E, \mathcal{E})$. Then, we obtain the canonical construction for a time-inhomogeneous Markov chain X. Such chains can be made time-homogeneous by incorporating time into the state space: Let $\hat{E} = \mathbb{N} \times E, \hat{\mathcal{E}} = 2^{\mathbb{N}} \otimes \mathcal{E}$, and define the Markov kernel $\hat{P}$ on $(\hat{E}, \hat{\mathcal{E}})$ such that

$$\hat{P}(y, B) = P_{n+1}(x, A) \quad \text{if } y = (n, x),\ B = \{n+1\} \times A. \tag{5.12}$$

Then, putting $\hat{X}_n = (n, X_n)$, we note that $\hat{X}_n$ is measurable with respect to $\mathcal{F}_n$ and $\hat{\mathcal{E}}$, and the Markov property holds just as in 5.8:

$$\mathbb{P}\{\hat{X}_{n+1} \in B \mid \mathcal{F}_n\} = \hat{P}(\hat{X}_n, B).$$

Thus, $\hat{X} = (\hat{X}_n)$ is a time-homogeneous Markov chain over $(\Omega, \mathcal{H}, \mathbb{P})$ with state space $(\hat{E}, \hat{\mathcal{E}})$ and transition kernel $\hat{P}$ in the sense of 5.9 above. Note that this is not the canonical construction for $\hat{X}$.

5.13 *Periodicity in time.* This is a special case of time-inhomogeneity which is encountered often in applied work where seasonal effects or the day-of-the-week effects and the like affect the transition probabilities. For instance, if X_n is to denote the inventory level for some item at the end of day n, then we would expect $P_n(x, A)$ to depend on n only through whether n is a Monday or Tuesday and so on. In such cases, there is an integer $d \geq 2$ such that the sequence of transition kernels P_n in 5.11 has the form

$$(P_1, P_2, \ldots) = (P_1, P_2, \ldots, P_d, P_1, P_2, \ldots, P_d, \ldots). \tag{5.14}$$

The corresponding time-inhomogeneous chain X can be rendered time-homogeneous by incorporating periodicity into the state space: Let $D = \{1, 2, \ldots, d\}, \mathcal{D} = 2^D, \hat{E} = D \times E, \hat{\mathcal{E}} = \mathcal{D} \otimes \mathcal{E}$, and define $\hat{P}$ as a Markov kernel on $(\hat{E}, \hat{\mathcal{E}})$ that satisfies

$$\hat{P}(y, B) = P_j(x, A) \quad \text{if } y = (i, x),\ B = \{j\} \times A,\ j = (1 + i) \text{ modulo } d.$$

Then, $\hat{X}_n = (n \text{ modulo } d, X_n),\ n \in \mathbb{N}$, form a time-homogeneous Markov chain with state space $(\hat{E}, \hat{\mathcal{E}})$ and transition kernel $\hat{P}$. Again, this is not canonical.

5.15 *Cyclic Markov chains.* The concept is similar to periodicity but with the added twist that the space changes with time. Suppose X is a chain constructed as in the preceding section, but with

$$(E_0, E_1, \ldots) = (E_d, E_1, E_2, \ldots, E_d, E_1, E_2, \ldots, E_d, \ldots),$$

and similarly for the $\mathcal{E}_n$, and the conditions 5.11 and 5.14– note that P_j is a kernel from $(E_{j-1}, \mathcal{E}_{j-1})$ into $(E_j, \mathcal{E}_j), j = 1, \ldots, d$. Then, X is called a cyclic Markov chain in canonical form: if we think of X_n as the position of a particle after its n^{th} step, the particle moves from a point in space E_d into a random

point in E_1, and then to a random point in $E_2, \ldots$, and so on cyclically, with transition probabilities depending only on the space to be visited next. The chain X can be made time-homogeneous by the technique used on periodic chains but with a more careful construction for $(\hat{E}, \hat{\mathcal{E}})$.

5.16 *k-dependent Markov chains.* For the Markov chains introduced so far, the conditional distribution of X_{n+1} given all the past $\mathcal{F}_n$ depended only on the last state X_n. In some applications, for instance if X_n is to denote the weather on day n, it is desired to make the dependence on the past a little deeper: For fixed integer $k \geq 2$, suppose that

$$K_{n+1}(x_0, \ldots, x_n; A) = P(x_{n-k+1}, \ldots, x_n; A) \tag{5.17}$$

for each $n \geq k$ for all $x_0, \ldots, x_n$ in E and A in $\mathcal{E}$, where P is a transition probability kernel from $(E, \mathcal{E})^k$ to $(E, \mathcal{E})$. Then, $X = (X_n)$ is said to be a k-dependent time-homogeneous Markov chain with state space $(E, \mathcal{E})$.

Theoretically, such chains are no different from ordinary Markov chains. To convert such a chain to an ordinary one-dependent chain, one needs to re-define the term "state": Put $(\hat{E}, \hat{\mathcal{E}}) = (E, \mathcal{E})^k$, let $\hat{X}_n = (X_n, X_{n+1}, \ldots, X_{n+k-1})$, and define the Markov kernel $\hat{P}$ on $(\hat{E}, \hat{\mathcal{E}})$ such that

$$\hat{P}(y, B) = P(y_1, \ldots, y_k; A) \quad \text{if } B = \{y_2\} \times \cdots \times \{y_k\} \times A. \tag{5.18}$$

Then, $(\hat{X}_n)$ is an ordinary Markov chain with state space $(\hat{E}, \hat{\mathcal{E}})$ and transition kernel $\hat{P}$.

Markov processes, continuous time

These are the continuous time versions of Markov chains, that is, the parameter set is $\mathbb{R}_+$. We discuss only the most fundamental cases. In the setting of Kolmogorov extension theorem of the preceding section, let the index set be $I = \mathbb{R}_+$, let μ be a probability measure on $(E, \mathcal{E})$, and, for every pair (s, t) of times in $\mathbb{R}_+$ with $s < t$, let $P_{s,t}$ be a Markov kernel on $(E, \mathcal{E})$. We are to interpret μ as the distribution of X_0, and $P_{s,t}(x, A)$ as the conditional probability that $X_t \in A$ given that $X_s = x$. Accordingly, if $J = (t_0, t_1, \ldots, t_n)$ with $0 = t_0 < t_1 < \cdots < t_n$, then we specify the distribution π_J of $X_J = (X_{t_0}, X_{t_1}, \ldots, X_{t_n})$ by

$$\begin{aligned} &\pi_J(dx_0, dx_1, dx_2, \ldots, dx_{n-1}, dx_n) \\ &= \mu(dx_0) P_{t_0,t_1}(x_0, dx_1) P_{t_1,t_2}(x_1, dx_2) \cdots P_{t_{n-1},t_n}(x_{n-1}, dx_n). \end{aligned} \tag{5.19}$$

Assuming that

$$P_{s,t} P_{t,u} = P_{s,u}, \quad 0 \leq s < t < u < \infty, \tag{5.20}$$

that is, in more explicit notation,

$$\int_E P_{s,t}(x, dy) P_{t,u}(y, B) = P_{s,u}(x, B), \quad x \in E, B \in \mathcal{E}, \tag{5.21}$$

the consistency requirement 4.17 is satisfied. Then, according to Theorem 4.18, there exists a unique probability measure $\mathbb{P}^\mu$ on $(\Omega, \mathcal{H}) = (E, \mathcal{E})^{\mathbb{R}_+}$ such that 5.19 is the distribution of $(X_{t_0}, X_{t_1}, \ldots, X_{t_n})$, where $0 = t_0 < t_1 < \cdots < t_n$.

Let $\mathcal{F}_t$ be the σ-algebra generated by $\{X_s : 0 \le s \le t\}$, that is, the σ-algebra generated by the finite-dimensional cylinders of the form $\{X_{s_1} \in A_1, \ldots, X_{s_n} \in A_n\}$ with $0 \le s_1 < \cdots < s_n \le t$ and $A_1, \ldots, A_n$ in $\mathcal{E}$. Then, it follows from 5.19 and theorems on conditional expectations that

$$\mathbb{P}^\mu\{X_t \in A \mid \mathcal{F}_s\} = P_{s,t}(X_s, A), \quad A \in \mathcal{E}, \tag{5.22}$$

for all $0 \le s < t$; and, of course,

$$\mathbb{P}^\mu\{X_0 \in A\} = \mu(A), \quad A \in \mathcal{E}. \tag{5.23}$$

For these reasons, $X = (X_t)_{t \in \mathbb{R}_+}$ is called a *Markov process* with state space $(E, \mathcal{E})$ and transition kernels $(P_{s,t})_{0 \le s < t < \infty}$. This is the time-inhomogeneous case.

Mathematically more interesting is the *time-homogeneous* case where $P_{s,t}$ depends on s and t only through $t - s$. Thus, replacing $P_{s,t}$ with P_{t-s}, we see that 5.20 becomes

$$P_s P_t = P_{s+t}, \quad s, t \in \mathbb{R}_+, \tag{5.24}$$

and the collection $(P_t)_{t \in \mathbb{R}_+}$ is called the *transition semigroup* of the (time-homogeneous) Markov process $X = (X_t)_{t \in \mathbb{R}_+}$ with state space $(E, \mathcal{E})$. Incidentally, the semigroup property 5.24 as well as the parental 5.20 are called *Chapman-Kolmogorov equations*.

Random fields

Returning to Kolmogorov extension theorem 4.18, suppose that the index set I is $\mathbb{R}^d$ for some fixed dimension $d \ge 1$, and suppose that the state space $(E, \mathcal{E})$ is replaced with $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. The index set can no longer be thought as time, and we now write x for the generic element of $I = \mathbb{R}^d$. Assuming that Kolmogorov's theorem applies, we obtain a stochastic process $X = \{X(x) : x \in \mathbb{R}^d\}$, and we may view it as a random field on $\mathbb{R}^d$, that is, to each x in $\mathbb{R}^d$ we associate a real-valued random variable $X(x)$.

In the further case where the state space $(E, \mathcal{E})$ is taken to be $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the random variable $X(x)$ can be thought as the Eulerian velocity vector at x and the process X becomes a *random velocity field*. In such applications, one generally requires some smoothness (like continuity or differentiability) from the mapping $x \mapsto X(x, \omega)$ for each ω. Theorem 4.18 guarantees no such smoothness.