Contents

7 Conditioning **61**
7a What is the problem 61
7b Discrete case 63
7c Conditional expectation 64
7d A convergence theorem 66
7e Conditional measures 66
7f Markov property as conditional independence 74

7 Conditioning

7a What is the problem

First, three results announced. They are rather trivial is such special cases as discrete distributions or absolutely continuous distributions, but nontrivial for singular (and mixed) distributions. We'll see that the general theory holds for *all* distributions, and generalizes to high and even infinite dimension.

7a1 Proposition. Every 2-dimensional random variable (X, Y) is distributed like $\big(f(U), g(U, V)\big)$ for some Lebesgue measurable functions $f : (0,1) \to \mathbb{R}$ and $g : (0,1) \times (0,1) \to \mathbb{R}$ such that f is increasing on $(0,1)$, and $g(u, \cdot)$ is increasing on $(0,1)$ for each $u \in (0,1)$; here U, V are independent random variables distributed uniformly on $(0,1)$.

Such f is called the quantile function of X; and $g(u, \cdot)$ is the conditional quantile function of Y given $X = f(u)$.

7a2 Proposition. Let (X_1, Y_1) be a 2-dimensional random variable (on some probability space), and (Y_2, Z_2) another 2-dimensional random variable (on another probability space) such that Y_1 and Y_2 are identically distributed. Then there exists (on some probability space) a 3-dimensional random variable (X, Y, Z) such that (X, Y) is distributed like (X_1, Y_1), and (Y, Z) is distributed like (Y_2, Z_2).

The idea is simple: X and Z are conditionally independent given Y. But what exactly does it mean?

7a3 Theorem (disintegration of measure)**.** For every probability measure μ on $\mathbb{R}^2$ there exist a probability measure ν on $\mathbb{R}$ and a family $(\mu_x)_{x \in \mathbb{R}}$ of probability measures μ_x on $\mathbb{R}$ such that

$$\mu(A) = \int \mu_x(A_x)\, \nu(\mathrm{d}x)$$

for all Borel sets $A \subset \mathbb{R}^2$; here $A_x = \{y : (x,y) \in A\}$. (The function $x \mapsto \mu_x(A_x)$ is thus claimed to be ν-measurable.)

Now, the problem in general.

The elementary definition

$$\mathbb{P}\big(A\,\big|\,B\big) = \frac{\mathbb{P}\big(A \cap B\big)}{\mathbb{P}\big(B\big)}$$

works if and only if $\mathbb{P}\big(B\big) \neq 0$. However, in many cases a limiting procedure gives us a useful result when $\mathbb{P}\big(B\big) = 0$.

7a4 Example. Let $(S_n)_n$ be the simple one-dimensional random walk and $B = \{\forall n \;\; S_n > -10\}$ (a zero-probability event). We introduce $B_n = \{S_1 > -10, \dots, S_n > -10\}$, observe that $B_n \downarrow B$ and let $\mathbb{P}\big(A\,\big|\,B\big) = \lim_n \mathbb{P}\big(A\,\big|\,B_n\big)$ for "simple" A. In fact, we get a Markov chain with the transition probability $p_{k,k-1} = \frac{k+9}{2k+20}$, $p_{k,k+1} = \frac{k+11}{2k+20}$. However, the formula $\mathbb{P}\big(A\,\big|\,B\big) = \lim_n \mathbb{P}\big(A\,\big|\,B_n\big)$ should not be used for all A; otherwise, trying $A = B$, we get a paradox: $\mathbb{P}\big(B\,\big|\,B\big) = 0$.

Similarly we may define the self-avoiding random walk on $\mathbb{Z}^2$ (assuming convergence); in fact, no one knows the joint distribution of the first two moves (even roughly)!

Sometimes different "reasonable" sequences $B_n \downarrow B$ lead to different results, which is known as Borel's paradox or Borel-Kolmogorov paradox. For example,

$$\lim_{\varepsilon\to 0+} \mathbb{P}\big(X \le 1\big|\, -\varepsilon < Y < \varepsilon\big) \neq \lim_{\varepsilon\to 0+} \mathbb{P}\big(X \le 1\big|\, -\varepsilon|X| < Y < \varepsilon|X|\big)\,.$$

Also, meridians (lines of longitude) and parallels (circles of latitude) on a sphere.

Sometimes conditioning is *really* impossible.

7a5 Example. Let $(S_n)_n$ be the simple one-dimensional random walk and $B = \{S_n \to +\infty\}$ (a zero-probability event). We note that $B = \{S_n - S_{10} \to +\infty\}$, "therefore" B is independent of $S_1, \dots, S_{10}$; conditionally, given B, the walk behaves as usual, and we get the paradox, $\mathbb{P}\big(B\,\big|\,B\big) = 0$, once again.

7a6 Example. Let $X \sim U(0,1)$ and $B = \{X \in \mathbb{Q}\}$. We consider $\mathrm{e}^{2\pi \mathrm{i} X}$; by symmetry, all rational points "must" get equal probabilities, which is impossible.

The conditional density formula

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

works well whenever the joint distribution of X, Y is absolutely continuous. Neither this formula, nor the similar discrete formula

$$\mathbb{P}\big(Y = y \,\big|\, X = x\big) = \frac{\mathbb{P}\big(X = x, Y = y\big)}{\mathbb{P}\big(X = x\big)}$$

covers the (practically important) case of discrete Y but absolutely continuous X.[1] A still more complicated case is, conditioning of Y on $X = g(Y)$ when Y is absolutely continuous and g is not one-to-one (especially when g behaves like the Weierstrass function).

Bad news: we have no conditioning theory that covers under a single umbrella all "good" cases listed above. Good news (for those who do not fear of measure theory): we have a conditioning theory that covers conditioning of Y on X for an *arbitrary* joint distribution of random variables (or random vectors) X, Y, and this theory includes both the discrete case and the absolutely continuous case.

7b Discrete case

Let Ω be (at most) countable, $\mathcal{F} = 2^\Omega$, and $\mathcal{F}_1 \subset \mathcal{F}$ a sub-σ-algebra. Clearly, $\mathcal{F}_1 = \sigma(X)$ for some $X : \Omega \to \mathbb{R}$ (X just indexes the equivalence classes with some real numbers). Here is the elementary conditioning:

$$\mathbb{P}\big(A \,\big|\, X = x\big) = \frac{\mathbb{P}\big(A \cap X^{-1}(x)\big)}{\mathbb{P}\big(X^{-1}(x)\big)} = \frac{\sum_{\omega \in A \cap X^{-1}(x)} p(\omega)}{\sum_{\omega \in X^{-1}(x)} p(\omega)} = P_x(A) = f(x)\,,$$
$$\mathbb{P}\big(A \,\big|\, \mathcal{F}_1\big) = \mathbb{P}\big(A \,\big|\, X\big) = f(X) : \Omega \to \mathbb{R}\,;$$
$$\mathbb{E}\big(Y \,\big|\, X = x\big) = \frac{\sum_{\omega \in A \cap X^{-1}(x)} Y(\omega) p(\omega)}{\sum_{\omega \in X^{-1}(x)} p(\omega)} = \int Y \,\mathrm{d}P_x = g(x)\,,$$
$$\mathbb{E}\big(Y \,\big|\, \mathcal{F}_1\big) = \mathbb{E}\big(Y \,\big|\, X\big) = g(X) : \Omega \to \mathbb{R}\,;$$

each *conditional measure* P_x is a probability measure on Ω, concentrated on $X^{-1}(x)$; the map $x \to P_x$ depends on the choice of X, but the map $P_X : \omega \mapsto P_{X(\omega)}$ does not. Similarly, *regression functions* f and g depend on the choice of X, but the random variables $f(X)$ and $g(X)$ do not. The conditional probability is a special case of the conditional expectation: $Y = \mathbb{1}_A$.

[1] See 7e18.

Convergence of the series is ensured if Y is integrable (except for negligible x, if any).

Note that $\mathbb{E}\big(Y\big|\mathcal{F}\big) = Y$ and $\mathbb{P}\big(A\big|\mathcal{F}\big) = \mathbb{1}_A$. On the other extreme, $\mathbb{E}\big(Y\big|\{\emptyset,\Omega\}\big) = \mathbb{E}\,Y$ (a constant function), and $\mathbb{P}\big(A\big|\{\emptyset,\Omega\}\big) = \mathbb{P}\big(A\big)$.

The *total probability formula* and *total expectation formula*

$$\mathbb{E}\left(\mathbb{P}\big(A\big|\mathcal{F}_1\big)\right) = \mathbb{P}\big(A\big)\,,$$
$$\mathbb{E}\left(\mathbb{E}\big(Y\big|\mathcal{F}_1\big)\right) = \mathbb{E}\,Y$$

boil down to the *decomposition of measure,*

$$P = \sum_x \mathbb{P}\big(X=x\big)\cdot P_x = \sum_\omega p(\omega) P_{X(\omega)} = \mathbb{E}\,P_X\,,$$

the latter expectation being taken in the linear space of signed measures...[1]

The function $a \mapsto \mathbb{E}\,\big((Y-a)^2\big) = a^2 - 2a\mathbb{E}\,Y + \mathbb{E}\,(Y^2)$ reaches its minimum at $a = \mathbb{E}\,Y$ (assuming $Y \in L_2$, which evidently holds for $Y = \mathbb{1}_A$). That is, $\mathbb{E}\,Y$ is the orthogonal projection of Y to the one-dimensional space of constants, in $L_2(\Omega,\mathcal{F},P)$. The same holds in each $L_2(P_x)$, thus, the regression function g minimizes $\mathbb{E}\,\big(Y-g(X)\big)^2 = \mathbb{E}\,\big(\mathbb{E}\,\big((Y-g(X))^2\big|X\big)\big)$. That is, $\mathbb{E}\big(Y\big|\mathcal{F}_1\big)$ is the orthogonal projection of Y to $L_2(\Omega,\mathcal{F}_1,P) \subset L_2(\Omega,\mathcal{F},P)$.

7c Conditional expectation

We turn to the general case: $(\Omega,\mathcal{F},P)$ is an arbitrary probability space, and $\mathcal{F}_1 \subset \mathcal{F}$ a sub-σ-algebra. We assume that all null sets belong to $\mathcal{F}$ and also to $\mathcal{F}_1$.

The Hilbert space $L_2(\mathcal{F}_1) = L_2(\Omega,\mathcal{F}_1,P)$ is a subspace of $L_2(\mathcal{F}) = L_2(\Omega,\mathcal{F},P)$. We consider the orthogonal projection $L_2(\mathcal{F}) \to L_2(\mathcal{F}_1)$ and *denote* it $Y \mapsto \mathbb{E}\big(Y\big|\mathcal{F}_1\big)$. Note that $\mathbb{E}\big(Y\big|\mathcal{F}_1\big)$ is an equivalence class. Orthogonality means that $\langle Y - \mathbb{E}\big(Y\big|\mathcal{F}_1\big), X\rangle = 0$, that is,

$$\mathbb{E}\left(X\cdot\mathbb{E}\big(Y\big|\mathcal{F}_1\big)\right) = \mathbb{E}\,(XY) \quad \text{for all } X \in L_2(\mathcal{F}_1)\,;$$

this property characterizes $\mathbb{E}\big(Y\big|\mathcal{F}_1\big)$ among $L_2(\mathcal{F}_1)$. In particular,

$$\mathbb{E}\left(\mathbb{E}\big(Y\big|\mathcal{F}_1\big)\right) = \mathbb{E}\,(Y)\,,$$

the total expectation formula (not a characterization, of course). Moreover,

$$\mathbb{E}\left(\mathbb{E}\big(Y\big|\mathcal{F}_1\big); B\right) = \mathbb{E}\,(Y; B) \quad \text{for all } B \in \mathcal{F}_1\,; \tag{7c1}$$

[1] Recall Sect. 5c.

also a characterization, since indicators and their linear combinations are dense in L_2. A projection operator is always linear:

$$\mathbb{E}\big(aX\,\big|\,\mathcal{F}_1\big) = a\mathbb{E}\big(X\,\big|\,\mathcal{F}_1\big)\,,$$
$$\mathbb{E}\big(X+Y\,\big|\,\mathcal{F}_1\big) = \mathbb{E}\big(X\,\big|\,\mathcal{F}_1\big) + \mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)\,,$$
$$\text{if } X_n \to X \text{ in } L_2 \quad \text{then} \quad \mathbb{E}\big(X_n\,\big|\,\mathcal{F}_1\big) \to \mathbb{E}\big(X\,\big|\,\mathcal{F}_1\big) \text{ in } L_2\,.$$

But the subspace $L_2(\mathcal{F}_1)$ is special:

$$\text{if } X \in L_2(\mathcal{F}_1) \quad \text{then} \quad X^+ \in L_2(\mathcal{F}_1)\,.$$

It follows easily that the projection is positive:

$$\text{if } X \ge 0 \text{ a.s.} \quad \text{then} \quad \mathbb{E}\big(X\,\big|\,\mathcal{F}_1\big) \ge 0 \text{ a.s.}$$

Thus, the projection operator is continuous (of norm 1) also in the L_1 norm (apply the total expectation formula to X^+ and X^-), and therefore extends to $L_1(\Omega, \mathcal{F}, P)$ by continuity. It is still positive.

The "tower property"

$$\mathbb{E}\big(X\,\big|\,\mathcal{F}_1\big) = \mathbb{E}\big(\mathbb{E}\big(X\,\big|\,\mathcal{F}_2\big)\,\big|\,\mathcal{F}_1\big) \text{ a.s.} \quad \text{if } \mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}$$

holds in L_2 for a simple geometric reason, and extends to L_1 by continuity.

7c2 Example. Let $Y \sim U(0,1)$ and $X = f(Y)$,

$$f(y) = \begin{cases} 3y & \text{for } 0 \le y \le 1/3, \\ 1.5(1-y) & \text{for } 1/3 \le y \le 2/3, \\ 0.5 & \text{for } 2/3 \le y \le 1. \end{cases}$$

Then $\mathbb{E}\big(Y\,\big|\,X\big) = g(X)$,

$$g(x) = \begin{cases} x/3 & \text{for } 0 < x < 0.5, \\ 5/6 & \text{for } x = 0.5, \\ (2-x)/3 & \text{for } 0.5 < x < 1. \end{cases}$$

7c3 Exercise. Do it twice. Namely, (a) check it via (7c1); (b) derive it by minimization.

7c4 Exercise. Two σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$ are independent if and only if $\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big) = \mathbb{E}\,Y$ a.s. for all $Y \in L_2(\mathcal{F}_2)$.

Formulate it accurately, and prove.

7c5 Exercise. Let $\mathbb{E}\big(Y\big|\mathcal{F}_1\big)=\mathbb{E}\,Y$ a.s. for a given $Y\in L_2(\mathcal{F})$.
(a) Does it follow that Y is independent of $\mathcal{F}_1$?
(b) Reconsider (a) assuming in addition that Y takes on only 3 values.
(c) The same for only two values.

More on martingales

Now we may use Definition 6a3 (rather than 6a4) in full generality.

A martingale bounded in L_2 converges in L_2 (recall Lemma 6c5 and the paragraph after it), $M_n\to M_\infty$ in L_2. We have $M_\infty=M_\infty^+-M_\infty^-$ and $M_n=\mathbb{E}\big(M_\infty\big|\mathcal{F}_n\big)=\mathbb{E}\big(M_\infty^+\big|\mathcal{F}_n\big)-\mathbb{E}\big(M_\infty^-\big|\mathcal{F}_n\big)$, the difference between two L_2-bounded positive martingales. Proposition 6c6 is thus generalized.

7c6 Proposition. A martingale bounded in L_2 converges both in L_2 and almost surely.

7d A convergence theorem

In the end of Sect. 6c, when proving $\mathbb{P}\big(1\le Z_n=o((2p)^n)\big)=0$, we used the relation $\mathbb{P}\big(A\big|\mathcal{F}_n\big)\to\mathbb{1}_A$ a.s. (for $A\in\mathcal{F}_\infty$). This relation is proved below.

Recall 3b13: if $\mathcal{F}_n\uparrow\mathcal{F}_\infty$ then $\cup_n L_2(\mathcal{F}_n)$ is dense in $L_2(\mathcal{F}_\infty)$. It follows that $\mathbb{E}\big(X\big|\mathcal{F}_n\big)\to X$ in L_2 for all $X\in L_2(\mathcal{F}_\infty)$, and $\mathbb{E}\big(X\big|\mathcal{F}_n\big)\to\mathbb{E}\big(X\big|\mathcal{F}_\infty\big)$ in L_2 for all $X\in L_2=L_2(\mathcal{F})$. But this is a martingale, bounded in L_2; 7c6 ensures a.s. convergence, and we get the following.

7d1 Theorem. Let $(\Omega,\mathcal{F},P)$ be a probability space and $\mathcal{F}_1\subset\mathcal{F}_2\subset\dots\subset\mathcal{F}_\infty\subset\mathcal{F}$ sub-σ-algebras such that $\mathcal{F}_n\uparrow\mathcal{F}_\infty$. Then

$$\mathbb{E}\big(X\big|\mathcal{F}_n\big)\to X \text{ a.s. and in } L_2 \quad \text{for all } X\in L_2(\mathcal{F}_\infty)\,;$$
$$\mathbb{E}\big(X\big|\mathcal{F}_n\big)\to\mathbb{E}\big(X\big|\mathcal{F}_\infty\big) \text{ a.s. and in } L_2 \quad \text{for all } X\in L_2(\mathcal{F})\,;$$
$$\mathbb{P}\big(A\big|\mathcal{F}_n\big)\to\mathbb{1}_A \text{ a.s. and in } L_2 \quad \text{for all } A\in\mathcal{F}_\infty\,;$$
$$\mathbb{P}\big(A\big|\mathcal{F}_n\big)\to\mathbb{P}\big(A\big|\mathcal{F}_\infty\big) \text{ a.s. and in } L_2 \quad \text{for all } A\in\mathcal{F}\,.$$

7e Conditional measures

Let $(\Omega,\mathcal{F},P)$ be a probability space, and $\mathcal{F}_1\subset F$ a sub-σ-algebra. The *conditional probability*

$$\mathbb{P}\big(A\big|\mathcal{F}_1\big)=\mathbb{E}\big(\mathbb{1}_A\big|\mathcal{F}_1\big)$$

satisies

$$0 \le \mathbb{P}\big(A\big|\mathcal{F}_1\big) \le 1 \quad \text{a.s.},$$
$$\mathbb{P}\big(A_1 \uplus A_2 \uplus \dots \,\big|\mathcal{F}_1\big) = \mathbb{P}\big(A_1\big|\mathcal{F}_1\big) + \mathbb{P}\big(A_2\big|\mathcal{F}_1\big) + \dots \quad \text{a.s.},$$
$$\forall B \in \mathcal{F}_1 \;\; \mathbb{P}\big(B\big|\mathcal{F}_1\big) = \mathbb{1}_B\,;$$
$$\mathbb{E}\left(\mathbb{P}\big(A\big|\mathcal{F}_1\big)\right) = \mathbb{P}\big(A\big)\,,$$

which, however, does not mean that we can define conditional measures just by $P_\omega(A) = \mathbb{P}\big(A\big|\mathcal{F}_1\big)(\omega)$.

In the discrete case (at most countable Ω) we may get rid of all negligible points (if any) and define $P_\omega(A) = \mathbb{P}\big(A\big|\mathcal{F}_1\big)(\omega)$, getting

$$0 \le P_\omega(A) \le 1\,;$$
$$P_\omega(A_1 \uplus A_2 \uplus \dots) = P_\omega(A_1) + P_\omega(A_2) + \dots\,;$$
$$\forall B \in \mathcal{F}_1 \;\; P_\omega(B) = \mathbb{1}_B(\omega)\,; \quad \text{especially, } P_\omega(\Omega) = 1\,;$$
$$\forall A \in \mathcal{F} \;\; \int P_\omega(A)\,P(\mathrm{d}\omega) = P(A) \quad \left(\text{in this sense, } \int P_\omega\,P(\mathrm{d}\omega) = P\right).$$

It is a *disintegration* of P into probability measures P_ω localized on corresponding parts of the partition.

In general it does not go...

7e1 Definition. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{F}_1 \subset \mathcal{F}$ a sub-σ-algebra. A *regular conditional probability* (given $\mathcal{F}_1$) is a family $(P_\omega)_{\omega\in\Omega}$ of probability measures P_ω on $(\Omega, \mathcal{F})$ such that for every $A \in \mathcal{F}$ the function $\omega \mapsto P_\omega(A)$ belongs to the equivalence class $\mathbb{P}\big(A\big|\mathcal{F}_1\big)$.

The function $\omega \mapsto P_\omega(A)$ is not required to be $\mathcal{F}_1$-measurable; rather, it must be equal a.s. to some $\mathcal{F}_1$-measurable function (of the given equivalence class).

It is usual to write "$P_\omega(A) = \mathbb{P}\big(A\big|\mathcal{F}_1\big)(\omega)$ a.s." treating $\mathbb{P}\big(A\big|\mathcal{F}_1\big)$ as an (arbitrary) element of the equivalence class.

Only the equivalence class of the map $\omega \mapsto P_\omega$ matters; but the exceptional set should not depend on A.

Generally, a regular conditional probability need not exist (see 7e3).

7e2 Theorem. For $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$ a regular conditional probability exists and is unique (up to equivalence).

Note that (a) P is an *arbitrary* Borel probability measure on $\mathbb{R}$; (b) $\mathcal{F}$ is the Borel σ-algebra; P-null sets are *not* added; (c) $\mathcal{F}_1 \subset \mathcal{F}$ is an arbitrary sub-σ-algebra.

7e3 Example. There exists[1] $Z \subset [0,1]$ of interior Lebesgue measure 0 and exterior Lebesgue measure 1. We take $\Omega = [0,1]$, $\mathcal{F} = \{(A \cap Z) \uplus (B \setminus Z) : A, B \in \mathcal{B}\}$, $\mathcal{F}_1 = \mathcal{B}$, and define a probability measure P on $(\Omega, \mathcal{F})$ by $P\big((A \cap Z) \uplus (B \setminus Z)\big) = 0.5 \operatorname{mes} A + 0.5 \operatorname{mes} B$.

If $(P_\omega)_\omega$ is a regular conditional probability, then for almost every[2] $x \in [0,1]$ the measure P_x must be equal to δ_x (the atom at x), since P_x is concentrated on every rational interval containing x. Thus $P_x(Z) = \mathbb{1}_Z(x)$ a.s., which contradicts to its measurability w.r.t. $\mathcal{F}_1$.

7e4 Exercise. Show that $\mathbb{P}\big(Z \,\big|\, \mathcal{F}_1\big) = 0.5$ a.s.

Proof of Theorem 7e2

The uniqueness part of Theorem 7e2 is easy: almost every x satisfies $P_x(I) = P'_x(I)$ for all rational intervals I, which implies $P_x = P'_x$.

The existence part needs more effort.

7e5 Definition. A measurable space is a pair $(\Omega, \mathcal{F})$ of a set Ω and a σ-algebra $\mathcal{F}$ on it.

Do not confuse "measure space" and "measurable space"! A probability measure on a measurable space turns it into a probability space.

Elements of $\mathcal{F}$ are called measurable sets. A map between two measurable spaces is called measurable, if the inverse image of every measurable set is measurable. Two measurable spaces are called isomorphic, if there exists an isomorphism between them, that is, a measurable bijection with measurable inverse.

The disjoint union of two measurable spaces is a measurable space, $(\Omega, \mathcal{F}) = (\Omega', \mathcal{F}') \uplus (\Omega'', \mathcal{F}'')$.

A measurable part of a measurable space is itself a measurable space (and the original measurable space becomes the disjoint union).

By an embedding of one measurable space into another we mean an isomorphism between the former and a measurable part of the latter.

The following technical definition is introduced temporarily, for this proof only.

7e6 Definition. A measurable space $(\Omega, \mathcal{F})$ is *good,* if a regular conditional probability exists for every probability measure on $(\Omega, \mathcal{F})$ and every sub-σ-algebra of $\mathcal{F}$.

[1] Using the choice axiom, of course.

[2] W.r.t. Lebesgue measure.

Existence of regular conditional probability on $(\mathbb{R}, \mathcal{B})$ becomes the claim that $(\mathbb{R}, \mathcal{B})$ is good. It follows from the three lemmas below (since a measurable space isomorphic to a good one is good).

7e7 Lemma. The Cantor set (with its Borel σ-algebra) is a good measurable space.

7e8 Lemma. The real line (with its Borel σ-algebra) is embeddable (as a measurable space) into the Cantor set (with its Borel σ-algebra).

7e9 Lemma. A measurable part of a good measurable space is a good measurable space. In other words, if $(\Omega, \mathcal{F}) = (\Omega', \mathcal{F}') \uplus (\Omega'', \mathcal{F}'')$ is good then $(\Omega', \mathcal{F}')$ is good.

Given $(\Omega, \mathcal{F}) = (\Omega', \mathcal{F}') \uplus (\Omega'', \mathcal{F}'')$ and two sub-σ-algebras, $\mathcal{F}'_1 \subset \mathcal{F}'$ on Ω' and $\mathcal{F}''_1 \subset \mathcal{F}''$ on Ω'', we get the corresponding sub-σ-algebra $\mathcal{F}_1 \subset \mathcal{F}$ on Ω (namely, $\mathcal{F}_1 = \{A \uplus B : A \in \mathcal{F}'_1, B \in \mathcal{F}''_1\}$).

Proof of 7e9. Given a probability measure P on $(\Omega', \mathcal{F}')$, we extend it to a probability measure on $(\Omega, \mathcal{F})$ ($P(\Omega'') = 0$, necessarily). Further, given a sub-σ-algebra $\mathcal{F}'_1 \subset \mathcal{F}'$, we choose some sub-$\sigma$-algebra $\mathcal{F}''_1 \subset \mathcal{F}''$ (no matter which one) and get the corresponding sub-σ-algebra $\mathcal{F}_1 \subset \mathcal{F}$ on Ω such that $L_2(\Omega, \mathcal{F}_1) = L_2(\Omega', \mathcal{F}'_1)$ and therefore $\mathbb{E}\big(\cdot \big| \mathcal{F}_1\big) = \mathbb{E}\big(\cdot \big| \mathcal{F}'_1\big)$.

We take a regular conditional probability $(P_\omega)_{\omega\in\Omega}$ for $\mathcal{F}_1$ and restrict it to $(P_\omega)_{\omega\in\Omega'}$. We have $P_\omega(\Omega') = 1$ for almost all $\omega \in \Omega'$, since $P_\omega(\Omega') = \mathbb{P}\big(\Omega' \big| \mathcal{F}_1\big)(\omega) = \mathbb{1}_{\Omega'}(\omega) = 1$. Thus we may treat P_ω as a probability measure on $(\Omega', \mathcal{F}')$, getting a regular conditional probability for $\mathcal{F}'_1$. □

Proof of 7e8. First, $\mathbb{R}$ is isomorphic to $(0, 1)$ (as a topological space, the more so, as a measurable space).

Second, we embed $(0, 1)$ into the Cantor set via binary digits and observe that the image is a Borel set. □

7e10 Lemma. If the function $\omega \mapsto P_\omega(A)$ belongs to the equivalence class $\mathbb{P}\big(A \big| \mathcal{F}_1\big)$ for all A of an algebra $\mathcal{E}$ that generates $\mathcal{F}$ then it holds for all $A \in \mathcal{F}$ (and therefore $(P_\omega)_{\omega\in\Omega}$ is a regular conditional probability).

Proof. Recall Sect. 1b: $P^*(A) = P(A)$ for all $A \in \mathcal{F}$, where P^* is the outer measure (defined via $\mathcal{E}$).

Let $B \in \mathcal{F}$; we have to prove that $\mathbb{P}\big(B \big| \mathcal{F}_1\big) = P_\bullet(B)$ a.s. (denoting the function $\omega \mapsto P_\omega(\dots)$ by $P_\bullet(\dots)$). Given $\varepsilon > 0$, we take $A_1, A_2, \dots \in \mathcal{E}$ such that $\cup_n A_n \supset B$ and $\sum_n P(A_n) \le P(B) + \varepsilon$. Introducing a measurable function (and equivalence class) $h_\varepsilon = \sum_n P_\bullet(A_n) = \sum_n \mathbb{P}\big(A_n \big| \mathcal{F}_1\big)$ we get, on one hand, $P_\bullet(B) \le h_\varepsilon$ (since $P_\omega\big(\cup_n A_n\big) \le \sum_n P_\omega(A_n)$) and

$\mathbb{E}\, h_\varepsilon \le P(B) + \varepsilon$ (since $\mathbb{E}\, P_\bullet(A_n) = \mathbb{E}\, \mathbb{P}\big(A_n \big| \mathcal{F}_1\big) = P(A_n)$), and on the other hand, $\mathbb{P}\big(B \big| \mathcal{F}_1\big) \le h_\varepsilon$ (since $\mathbb{1}_B \le \sum_n \mathbb{1}_{A_n}$).

Applying the same argument to $\Omega \setminus B$ we get a measurable function g_ε such that $\mathbb{E}\, g_\varepsilon \ge P(B) - \varepsilon$, $P_\bullet(B) \ge g_\varepsilon$ and $\mathbb{P}\big(B \big| \mathcal{F}_1\big) \ge g_\varepsilon$.

We have $|\mathbb{P}\big(B \big| \mathcal{F}_1\big) - P_\bullet(B)| \le h_\varepsilon - g_\varepsilon$ and $\mathbb{E}\,(h_\varepsilon - g_\varepsilon) \le 2\varepsilon$ for arbitrary ε; therefore $\mathbb{P}\big(B \big| \mathcal{F}_1\big) = P_\bullet(B)$ a.s. □

Proof of 7e7. The Borel σ-algebra of the Cantor set is generated by the countable algebra $\mathcal{E}$ of clopen (that is, both closed and open) sets. Every finitely additive set function on this algebra is automatically σ-additive, due to compactness: if $A = A_1 \uplus A_2 \uplus \dots$ then $A_k = \emptyset$ for all k large enough. By Theorem 1b3, every finitely additive set function on this algebra extends to a measure.

We define P_ω by

$$P_\omega(A) = \mathbb{P}\big(A \big| \mathcal{F}_1\big)(\omega) \quad \text{for all } A \in \mathcal{A}\,,$$

choosing a function in each equivalence class (countably many choices) and extend P_ω to a probability measure. Additivity holds a.s. (countably many equalities!) and is easily ensured everywhere. It remains to use Lemma 7e10. □

Theorem 7e2 is thus proved.

7e11 Exercise. Generalize Theorem 7e2 to:

(a) $\mathbb{R}^n$ (with the Borel σ-algebra);

(b) $\mathbb{R}^\infty$ (all infinite sequences of reals with the σ-algebra generated by the coordinates).

Hint: only Lemma 7e8 needs to be generalized; embedding into the Cantor set ensures both existence and uniqueness.

Measurable spaces embeddable into the Cantor set are called *standard.* Observe that (a) Theorem 7e2 holds for all standard measurable spaces; (b) all $\mathbb{R}^n$ and $\mathbb{R}^\infty$ (and all their Borel subsets) are standard.[1]

The requirement that the function $\omega \mapsto P_\omega(A)$ belongs to the equivalence class $\mathbb{P}\big(A \big| \mathcal{F}_1\big)$ can be reformulated using (7c1):

$$\int_B P_\omega(A)\, P(\mathrm{d}\omega) = P(A \cap B) \quad \text{for all } B \in \mathcal{F}_1\,,\ A \in \mathcal{F}\,.$$

[1] In fact, all Polish spaces (with Borel σ-algebras), and all their Borel subsets, are standard.

Or, equivalently,

$$P_\omega(B) = \mathbb{1}_B(\omega) \text{ a.s.} \quad \text{for each } B \in \mathcal{F}_1 \text{ (separately)}, \tag{7e12}$$

$$\int_\Omega P_\omega(A)\, P(\mathrm{d}\omega) = P(A) \quad \text{for each } A \in \mathcal{F} \text{ (separately)}. \tag{7e13}$$

Indeed, (7e12) implies $\int_B P_\omega(A)\, P(\mathrm{d}\omega) = \int_\Omega P_\omega(A \cap B)\, P(\mathrm{d}\omega)$.

7e14 Exercise. A given $A \in \mathcal{F}$ is independent of $\mathcal{F}_1$ if and only if $P_\omega(A) = P(A)$ for almost all ω.

Prove it.

Proof of Theorem 7a3. Theorem 7e2 (the existence part), applied to $(\Omega, \mathcal{F}, P) = (\mathbb{R}^2, \mathcal{B}_2, \mu)$ and $\mathcal{F}_1$ generated by the first coordinate, gives

$$\mu(A) = \int P_{x,y}(A)\, \mu(\mathrm{d}x\mathrm{d}y)\,.$$

Measurability of $P_{x,y}$ w.r.t. $\mathcal{F}_1$ means that $P_{x,y} = P_x$. Property (7e12) means that P_x is concentrated on $\{x\} \times \mathbb{R}$. Thus, $P_{x,y}(A) = P_x(A) = \mu_x(A_x)$. The first projection of μ is ν. □

7e15 Exercise. Prove uniqueness of $(\mu_x)_{x\in\mathbb{R}}$ up to a ν-negligible set.

The same holds for many other spaces.

7e16 Exercise. Disintegrate the joint distribution μ of the random variables X, Y of Example 7c2. Namely, guess the measures μ_x and check the equality $\mu(A) = \int \mu_x(A_x)\, \nu(\mathrm{d}x)$.

7e17 Exercise. Disintegrate a measure that has a density (w.r.t. the 2-dimensional Lebesgue measure). Namely, guess the measures μ_x and check the equality $\mu(A) = \int \mu_x(A_x)\, \nu(\mathrm{d}x)$.

7e18 Exercise. Derive formulas for conditioning of a discrete random variable $Y : \Omega \to \mathbb{Z}$ on a continuous random variable $X : \Omega \to \mathbb{R}$ that has a density f_X.

7e19 Exercise. Every random variable X is distributed like $f(U)$ for some increasing function $f : (0,1) \to \mathbb{R}$; here U is distributed uniformly on $(0,1)$.

Prove it.

Hint: first, prove it for a discrete X; second, take discrete X_n such that $X_n \uparrow X$.

The function f is unique except for its values at discontinuity points (at most countable set).

Proof of Proposition 7a1. By 7e19, X is distributed like some $f(U)$. Theorem 7a3 gives us the conditional distributions μ_x of Y given $X = x$. By 7e19 (again), $\mu_{f(u)}$ is the distribution of some $g(u, V)$. For every y,

$$m\big(\{v : g(u,v) \le y\}\big) = \mu_{f(u)}\big((-\infty, y]\big)$$

is measurable in u, and

$$\{(u,v) : v < \mu_{f(u)}((-\infty,y])\} \subset \{(u,v) : g(u,v) \le y\} \subset \{(u,v) : v \le \mu_{f(u)}((-\infty,y])\}$$

by monotonicity of $g(u,\cdot)$; it follows that g is measurable. Finally,

$$\begin{aligned}\mathbb{P}\big((f(U), g(U,V)) \in A\big) &= \int_0^1 \mathrm{d}u \int_0^1 \mathrm{d}v\, \mathbb{1}_A(f(u), g(u,v)) = \\ &= \int_0^1 \mathrm{d}u \int_0^1 \mathrm{d}v\, \mathbb{1}_{A_{f(u)}}(g(u,v)) = \int_0^1 \mathrm{d}u\, \mu_{f(u)}\big(A_{f(u)}\big) = \\ &= \int \mu_x(A_x)\, \nu(\mathrm{d}x) = \mu(A) = \mathbb{P}\big((X,Y) \in A\big)\,.\end{aligned}$$

□

Proof of Proposition 7a2. We have the one-dimensional distribution ν of Y_1 (and Y_2 as well), the two-dimensional distribution μ of (Y_1, X_1) and the two-dimensional distribution ξ of (Y_2, Z_2) Theorem 7a3 gives us μ_y and ξ_y such that[1] $\mu(A) = \int \mu_y(A_y)\, \nu(\mathrm{d}y)$ and $\xi(B) = \int \xi_y(B_y)\, \nu(\mathrm{d}y)$ for all Borel sets $A, B \subset \mathbb{R}^2$; here $A_y = \{x : (y,x) \in A\}$ and $B_y = \{z : (y,z) \in B\}$. We introduce a measure η on $\mathbb{R}^3$ by[2]

$$\eta(C) = \int (\mu_y \times \xi_y)(C_y)\, \nu(\mathrm{d}y)$$

for all Borel sets $C \subset \mathbb{R}^3$; here $C_y = \{(x,z) : (x,y,z) \in C\}$. For a three-dimensional random variable (X, Y, Z) distributed η we have

$$\begin{aligned}\mathbb{P}\big((Y,X) \in A\big) = \eta\big(\{(x,y,z) : (y,x) \in A\}\big) &= \int (\mu_y \times \xi_y)(A_y \times \mathbb{R})\, \nu(\mathrm{d}y) = \\ &= \int \mu_y(A_y)\, \nu(\mathrm{d}y) = \mu(A) = \mathbb{P}\big((Y_1, X_1) \in A\big)\,;\end{aligned}$$

and the same for (Y, Z). □

[1] μ_y is the conditional distribution of X_1 given $Y_1 = y$; and ξ_y is the conditional distribution of Z_2 given $Y_2 = y$.

[2] Think, is the integrand measurable?

Conditional measures and conditional expectations

7e20 Proposition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{F}_1 \subset \mathcal{F}$ a sub-σ-algebra, $(P_\omega)_{\omega\in\Omega}$ a regular conditional probability (given $\mathcal{F}_1$), and $Y \in L_1(\Omega, \mathcal{F}, P)$. Then the function $\omega \mapsto \int Y \,\mathrm{d}P_\omega$ belongs to the equivalence class $\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)$.

Proof. The relation holds for indicators, and (by linearity) for their linear combinations. Let it hold for each Y_n, and $0 \le Y_n \uparrow Y$ pointwise (not just a.s.), and $Y \in L_1(\Omega, \mathcal{F}, P)$; it is sufficient to prove that the relation holds for Y. We have $Y_n \to Y$ in L_1, therefore $\mathbb{E}\big(Y_n\,\big|\,\mathcal{F}_1\big) \to \mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)$ in L_1. On the other hand, $\int Y_n \,\mathrm{d}P_\omega \uparrow \int Y \,\mathrm{d}P_\omega \in [0,\infty]$ for each ω by the monotone convergence theorem; the rest is easy. □

It is tempting to extend $\mathbb{E}\big(\cdot\,\big|\,\mathcal{F}_1\big)$ to all Y such that $\int |Y| \,\mathrm{d}P_\omega < \infty$ for almost all ω (even if $\int |Y| \,\mathrm{d}P = \infty$). Then, however, strange things happen. For example, it may be that $\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big) > 0$ a.s., but $\mathbb{E}\big(Y\,\big|\,\mathcal{F}_2\big) < 0$ a.s.[1]

If X is $\mathcal{F}_1$-measurable then P_ω is concentrated on $X^{-1}(X(\omega))$ for almost every ω. That is, $X(\omega') = X(\omega)$ for P_ω-almost all ω'.

"Taking out what is known":

$$\mathbb{E}\big(XY\,\big|\,\mathcal{F}_1\big) = X\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big) \quad \text{for } X \in L_\infty(\mathcal{F}_1),\, Y \in L_1(\mathcal{F})\,.$$

Conditional versions of many inequalities follow immediately from existence of regular conditional probability. Conditional Markov inequality:

$$\mathbb{P}\big(Y > X\,\big|\,\mathcal{F}_1\big) \le \frac{\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)}{X} \text{ a.s.} \quad \text{for } X \in L_0^+(\mathcal{F}_1),\, Y \in L_0^+(\mathcal{F})\,.$$

Conditional Jensen's inequality:

$$\mathbb{E}\big(h(Y)\,\big|\,\mathcal{F}_1\big) \ge h\big(\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)\big)$$

for convex $h(\cdot)$, provided that $\mathbb{E}\big(|X|\,\big|\,\mathcal{F}_1\big) < \infty$ a.s. Choosing $h(x) = |x|^p$ with $p \in (1,\infty)$ and taking the (unconditional) expectation we get

$$\|\mathbb{E}\big(Y\,\big|\,\mathcal{F}_1\big)\|_p \le \|Y\|_p\,.$$

Conditional Cauchy-Schwartz inequality:

$$|\mathbb{E}\big(YZ\,\big|\,\mathcal{F}_1\big)| \le \sqrt{\mathbb{E}\big(Y^2\,\big|\,\mathcal{F}_1\big)}\sqrt{\mathbb{E}\big(Z^2\,\big|\,\mathcal{F}_1\big)}\,.$$

And so on.

[1] A counterexample (sketch): $\mathbb{P}\big(X = n, Y = n+1\big) = \mathbb{P}\big(X = n+1, Y = n\big) = 0.5p^n(1-p)$ for $n = 0, 1, 2, \ldots$; then $\mathbb{E}\big(a^Y\,\big|\,X = x\big) = \frac{pa+a^{-1}}{1+p}a^x$ for $x = 1, 2, \ldots$; we take $ap > 1$ and get $\mathbb{E}\big(a^Y\,\big|\,X\big) > a^X$ a.s., but also $\mathbb{E}\big(a^X\,\big|\,Y\big) > a^Y$ a.s.

7f Markov property as conditional independence

The "shortest" Markov process consists of 3 random variables X, Y, Z such that for every Borel set $B \subset \mathbb{R}$,

$$\mathbb{P}\big(Z \in B \,\big|\, X, Y\big) = \mathbb{P}\big(Z \in B \,\big|\, Y\big) \quad \text{a.s.} \tag{7f1}$$

In other words: in order to predict the future (Z) knowing the past (X) and the present (Y), use only the present; the past is irrelevant.

7f2 Exercise. A finite Markov chain $(S_0, S_1, S_2, \dots)$ (recall Sect. 5b) satisfies

$$\mathbb{P}\big(S_{n+1} \in B \,\big|\, S_0, \dots, S_n\big) = \mathbb{P}\big(S_{n+1} \in B \,\big|\, S_n\big) \quad \text{a.s.}$$

for every n.

Prove it.

Note that 7f2 may be treated as (7f1) generalized to multi-dimensional $X = (S_0, \dots, S_{n-1})$.

It may seem that (7f1) is not time-symmetric; that is, this condition on (X, Y, Z) does not imply the same condition on (Z, Y, X). However, let us consider it in the light of conditional distributions. Denoting $\mathbb{P}_y(\dots) = \mathbb{P}\big(\dots \,\big|\, Y = y\big)$ we rewrite (7f1) as $\mathbb{P}_y(Z \in B \mid X) = \mathbb{P}_y(Z \in B)$ a.s., which is just independence of Z and X (recall 7e14)... but conditionally, given $Y = y$.

In order to make this argument rigorous we need the following.

7f3 Proposition. Let (Ω, F) be a standard measurable space, (Ω, F, P) a probability space, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}$ sub-σ-algebras; $(P^{(1)}_\omega)_{\omega\in\Omega}$ a regular conditional probability on $(\Omega, \mathcal{F}, P)$ given $\mathcal{F}_1$; and $(P^{(2)}_\omega)_{\omega\in\Omega}$ a regular conditional probability on $(\Omega, \mathcal{F}, P)$ given $\mathcal{F}_2$. Then for almost every ω_1, $(P^{(2)}_\omega)_{\omega\in\Omega}$ is also a regular conditional probability on $(\Omega, \mathcal{F}, P^{(1)}_{\omega_1})$ given $\mathcal{F}_2$.

Proof. By 7e10 (and standardness) it is sufficient to check (7e12) and (7e13) on $(\Omega, \mathcal{F}, P^{(1)}_{\omega_1})$ for almost all ω_1; the exceptional set of ω_1 may depend on A, B. That is, we need

$$P^{(2)}_\omega(B) = \mathbb{1}_B(\omega) \quad P^{(1)}_{\omega_1}\text{-a.s.} \quad \text{for each } B \in \mathcal{F}_2 \text{ (separately)}\,,$$

$$\int_\Omega P^{(2)}_\omega(A)\, P^{(1)}_{\omega_1}(\mathrm{d}\omega) = P^{(1)}_{\omega_1}(A) \quad P\text{-a.s.} \quad \text{for each } A \in \mathcal{F} \text{ (separately)}\,.$$

The former: the equality holds P-a.s. therefore it holds $P^{(1)}_{\omega_1}$-a.s. for P-almost every ω_1.

The latter: the function $f : \omega \to P^{(2)}_\omega(A)$ belongs to the equivalence class $\mathbb{P}\big(A\big|\mathcal{F}_2\big)$; thus, by 7e20, the function $\omega_1 \mapsto \int_\Omega f\,\mathrm{d}P^{(1)}_{\omega_1} = \int_\Omega P^{(2)}_\omega(A)\,P^{(1)}_{\omega_1}(\mathrm{d}\omega)$ belongs to the equivalence class $\mathbb{E}\big(f\big|\mathcal{F}_1\big) = \mathbb{E}\big(\mathbb{P}\big(A\big|\mathcal{F}_2\big)\big|\mathcal{F}_1\big) = \mathbb{P}\big(A\big|\mathcal{F}_1\big)$ (the tower property); and the function $\omega_1 \mapsto P^{(1)}_{\omega_1}(A)$ belongs to the same equivalence class. □

We may apply it to a probability measure μ on $\mathbb{R}^3$ and its disintegrations:

$$\mu(A) = \int \mu_y(A_y)\,\nu_1(\mathrm{d}y)\,,\quad A_y = \{(x,z) : (x,y,z) \in A\}\,;$$
$$\mu(A) = \int \mu_{x,y}(A_{x,y})\,\nu_2(\mathrm{d}x\mathrm{d}y)\,,\quad A_{x,y} = \{z : (x,y,z) \in A\}$$

for Borel $A \subset \mathbb{R}^3$. By 7f3, for ν_1-almost every y we have the disintegration

$$\mu_y(A) = \int \mu_{x,y}(A_x)\,\mu_y(\mathrm{d}x\mathrm{d}z) = \int \mu_{x,y}(A_x)\,\nu_y(\mathrm{d}x)\,,\quad A_x = \{z : (x,z) \in A\}$$

for Borel $A \subset \mathbb{R}^2$; here ν_y is the marginal of μ_y.

In other words, for arbitrary random variables X, Y, Z we have

$$\mathbb{P}\big((X,Z) \in A\big|Y = y\big) = \int \mathbb{P}\big((X,Z) \in A\big|X = x, Y = y\big)\,\nu_y(\mathrm{d}x) = \\ = \mathbb{E}\big(\mathbb{P}\big((X,Z) \in A\big|X, Y = y\big)\big|Y = y\big)$$

for Borel $A \subset \mathbb{R}^2$ (since ν_y is the conditional distribution of X given $Y = y$). We get the conditioned (on $Y = y$) version of the formula $\mathbb{P}\big(\dots\big) = \mathbb{E}\,\mathbb{P}\big(\dots\big|X\big)$ mentioned in Sect. 7e.

7f4 Definition. Random variables X, Z are *conditionally independent* given Y, if for P_Y-almost every y the conditional distribution of (X, Z) given $Y = y$ is a product,

$$P_{X,Z|Y=y} = P_{X|Y=y} \times P_{Z|Y=y}\,.$$

An equivalent definition without conditional distributions:

$$\mathbb{P}\big(X \in A, Z \in B\big|Y\big) = \mathbb{P}\big(X \in A\big|Y\big)\mathbb{P}\big(Z \in B\big|Y\big) \quad \text{a.s.}$$

for all Borel sets $A, B \subset \mathbb{R}$.

7f5 Proposition. The Markov property (7f1) is equivalent to the conditional independence of X, Z given Y.

Now we can define easily the Markov property for random functions on a graph, on the lattice $\mathbb{Z}^n$, etc.