

## PART C: DEPENDENCE

## Contents

<b>16 Martingales</b>	<b>33</b>
16a Basic definitions . . . . .	33
16b Gambling strategy, martingale transform, stopping . . . . .	34
16c Positive martingales . . . . .	36
<b>17 Conditioning</b>	<b>38</b>
17a What is the problem . . . . .	38
17b Discrete case . . . . .	40
17c Conditional expectation . . . . .	41
17d A convergence theorem . . . . .	43
17e Conditional measures . . . . .	43

## 16 Martingales

## 16a Basic definitions

**16a1 Definition.** (a) A filtration on a probability space  $(\Omega, \mathcal{F}, P)$  is an increasing sequence of sub- $\sigma$ -fields:  $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots \subset \mathcal{F}$ ;

(b) An adapted (to the given filtration) process is a sequence  $(X_0, X_1, \dots)$  of random variables such that for each  $k$ ,  $X_k$  is  $\mathcal{F}_k$ -measurable.

A filtered probability space is a probability space endowed with a filtration.

Assume for now that  $\Omega$  is (finite or) countable (the discrete framework).

**16a2 Definition.** An adapted process  $(X_n)_n$  such that  $\forall n \mathbb{E}|X_n| < \infty$  is

- (a) a martingale, if  $\forall n \mathbb{E}(X_{n+1} | \mathcal{F}_n) = X_n$  a.s.;
- (b) a supermartingale, if  $\forall n \mathbb{E}(X_{n+1} | \mathcal{F}_n) \leq X_n$  a.s.;
- (c) a submartingale, if  $\forall n \mathbb{E}(X_{n+1} | \mathcal{F}_n) \geq X_n$  a.s.

We generalize it to arbitrary  $\Omega$  as follows.<sup>1</sup>

**16a3 Definition.** An adapted process  $(X_n)_n$  such that  $\forall n \mathbb{E}|X_n| < \infty$  is

- (a) a martingale, if  $\forall n \forall Y \in L_\infty(\mathcal{F}_n) \mathbb{E}((X_{n+1} - X_n)Y) = 0$ ;
- (b) a supermartingale, if  $\forall n \forall Y \in L_\infty^+(\mathcal{F}_n) \mathbb{E}((X_{n+1} - X_n)Y) \leq 0$ ;
- (c) a submartingale, if  $\forall n \forall Y \in L_\infty^+(\mathcal{F}_n) \mathbb{E}((X_{n+1} - X_n)Y) \geq 0$ .

---

<sup>1</sup>See also Sect. 17c.

Equivalently:  $\mathbb{E}((X_{n+k} - X_n)Y) = 0$  (or  $\leq 0$ , or  $\geq 0$ ). Indeed,  $Y \in L_\infty(\mathcal{F}_n) \subset L_\infty(\mathcal{F}_{n+1})$ ;  $\mathbb{E}((X_{n+2} - X_n)Y) = \mathbb{E}((X_{n+2} - X_{n+1})Y) + \mathbb{E}((X_{n+1} - X_n)Y) = 0$  (and so on).

In particular,  $Y = \mathbf{1}$  gives  $\mathbb{E} X_n = \mathbb{E} X_0$  (a necessary condition).

Assume again the discrete framework.

Every  $X \in L_1$  leads to a martingale  $M_n = \mathbb{E}(X | \mathcal{F}_n)$ . “Accumulating data”, “revising prediction”... Locally it is the general form of a martingale; globally — not.

Here is an explanation of the terms “supermartingale” and “submartingale”. Let  $(X_n)_{n=1}^N$  be adapted; introduce a martingale  $M_n = \mathbb{E}(X_N | \mathcal{F}_n)$ ; then:

if  $(X_n)$  is a martingale then  $X_n = M_n$ ;

if  $(X_n)$  is a supermartingale then  $X_n \geq M_n$ ;

if  $(X_n)$  is a submartingale then  $X_n \leq M_n$ .

Conditional Jensen inequality gives: if  $(M_n)$  is a martingale and  $f$  is convex (“sublinear”) then  $(f(M_n))$  is a submartingale.

**16a4 Example.** The one-dimensional simple random walk  $(S_n)$  is a martingale;  $(S_n^2)$  is a submartingale.

Functions on a tree...

**Proof of 6a1.** Denote by  $M_n$  the total mass of A-monsters at time  $n$ , then  $(M_n)_n$  is a martingale ( $\mathcal{F}_n$  being the whole past...) since  $b_1 \cdot \frac{a_1}{a_1+b_1} + (-a_1) \cdot \frac{b_1}{a_1+b_1} = 0$ . Thus,  $\mathbb{E} M_{m+n} = \mathbb{E} M_0 = A$ ; we note that  $M_{m+n}$  takes on two values only, 0 and  $A + B$ .  $\square$

## 16b Gambling strategy, martingale transform, stopping

**16b1 Definition.** (a) A previsible (with respect to the given filtration) process is a sequence  $(C_1, C_2, \dots)$  of random variables such that for each  $k$ ,  $C_k$  is  $\mathcal{F}_{k-1}$ -measurable.<sup>1</sup>

(b) Given a previsible process  $(C_n)_n$  and adapted process  $(X_n)_n$  (on the same filtered probability space), we define an adapted process  $C \bullet X$  by

$$\begin{aligned}(C \bullet X)_0 &= 0, \\(C \bullet X)_n &= (C \bullet X)_{n-1} + C_n(X_n - X_{n-1}).\end{aligned}$$

<sup>1</sup>In discrete time it look strange, but in continuous time it does not...

Thus,

$$\begin{aligned}(C \bullet X)_n &= C_1(X_1 - X_0) + C_2(X_2 - X_1) + \cdots + C_n(X_n - X_{n-1}) = \\ &= -C_1X_0 - (C_2 - C_1)X_1 - \cdots - (C_n - C_{n-1})X_{n-1} + C_nX_n.\end{aligned}$$

**16b2 Proposition.** Let  $(M_n)_n$  be a martingale,  $(C_n)_n$  previsible, and  $C_n(M_n - M_{n-1}) \in L_1$  for all  $n$ . Then  $C \bullet M$  is a martingale.

*Proof.* If  $Y \in L_\infty(\mathcal{F}_n)$  then  $\mathbb{E}(((C \bullet M)_{n+1} - (C \bullet M)_n)Y) = \mathbb{E}(C_{n+1}(M_{n+1} - M_n)Y)$ , and it vanishes if  $C_{n+1}Y \in L_\infty$ ; otherwise apply it to  $Y_k = Y \cdot \mathbf{1}_{[-k, k]}(C_{n+1})$  and note that  $\sup_k |C_{n+1}(M_{n+1} - M_n)Y_k| \leq |C_{n+1}(M_{n+1} - M_n)Y| \leq \|C_{n+1}Y\|_\infty \cdot |M_{n+1} - M_n|$ .  $\square$

A sufficient condition:  $\forall n \ C_n \in L_\infty$ .

An important special case:

$$(C^\tau)_n = \mathbf{1}_{n \leq \tau} = \begin{cases} 1 & \text{for } n \leq \tau, \\ 0 & \text{for } n > \tau, \end{cases}$$

where  $\tau$  is a stopping time as defined below.

**16b3 Definition.** A stopping time is a map  $\tau : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}$  such that  $\{\tau \leq n\} \in \mathcal{F}_n$  for all  $n$ .

Note that

$$\{C_n = 0\} = \{\tau < n\} = \{\tau \leq n - 1\} \in \mathcal{F}_{n-1}.$$

In terms of a tree, it is just a subtree...

The corresponding martingale transform is the stopped process,

$$(C^\tau \bullet X)_n = X_{\tau \wedge n} - X_0.$$

**16b4 Corollary.** If  $(M_n)_n$  is a martingale and  $\tau$  a stopping time then the stopped process  $(M_{\tau \wedge n})_n$  is also a martingale.

**Proof of 6b1.** The process  $M$  defined by  $M_n = S_n^2 - n$  is a martingale (think, why). On the other hand,  $T$  is a stopping time (think, why). Thus, the stopped process  $M_{T \wedge n}$  is a martingale. Therefore  $\mathbb{E} M_{T \wedge n} = 0$ , that is,  $\mathbb{E} S_{T \wedge n}^2 = \mathbb{E}(T \wedge n)$ . We get  $\mathbb{E}(T \wedge n) \leq 100$  for all  $n$ , and therefore  $T < \infty$  a.s.,  $\mathbb{E} T \leq 100$ ; thus  $S_T$  is well-defined,  $S_{T \wedge n} \rightarrow S_T$  a.s., and the bounded convergence theorem gives  $\mathbb{E} S_{T \wedge n}^2 \rightarrow \mathbb{E} S_T^2 = 100$ ; therefore  $\mathbb{E} T = 100$ .  $\square$

By the way,  $\mathbb{E} S_T = 0$ ;  $\mathbb{P}(S_T = -10) = 0.5 = \mathbb{P}(S_T = 10)$ .

By the way,  $\sup_n |S_n| = \infty$  a.s. Kolmogorov's 0-1 law gives  $\inf_n S_n = -\infty$  and  $\sup_n S_n = \infty$  a.s.

However, do not think that  $\mathbb{E} M_\tau$  must vanish! Think about  $\tau = \min\{n : S_n = +10\}$ .

## 16c Positive martingales

Let  $(M_n)_n$  be a positive martingale, that is,  $M_n \geq 0$  a.s. for every  $n$ .

Given  $0 < a < b < \infty$ , we define stopping times

$$\begin{aligned}\sigma_1 &= \inf\{n \geq 0 : M_n \leq a\}, & \tau_1 &= \inf\{n > \sigma_1 : M_n \geq b\}, \\ \sigma_2 &= \inf\{n > \tau_1 : M_n \leq a\}, & \tau_2 &= \inf\{n > \sigma_2 : M_n \geq b\},\end{aligned}$$

and so on. (As usual,  $\inf \emptyset = \infty$ .) Now we define the (random) number of upcrossings:

$$U = \sup\{k : \tau_k < \infty\}; \quad U : \Omega \rightarrow \{0, 1, 2, \dots\} \cup \{\infty\}.$$

**16c1 Proposition** (Dubins's inequality).

$$\mathbb{P}(U \geq k) \leq \left(\frac{a}{b}\right)^k \quad \text{for } k = 0, 1, 2, \dots$$

*Proof.* It is sufficient to prove that  $\mathbb{P}(\tau_k < \infty) \leq \frac{a}{b}\mathbb{P}(\sigma_k < \infty)$ . We have

$$\begin{aligned}\mathbb{E} M_{\sigma_k \wedge n} &= \mathbb{E} M_{\tau_k \wedge n}; \\ \underbrace{\mathbb{E} M_{\sigma_k \wedge n}}_{= \mathbb{E} M_0} &= \underbrace{\mathbb{E}(M_{\sigma_k}; \sigma_k \leq n)}_{\leq a\mathbb{P}(\sigma_k \leq n)} + \mathbb{E}(M_n; \sigma_k > n); \\ \underbrace{\mathbb{E} M_{\tau_k \wedge n}}_{= \mathbb{E} M_0} &= \underbrace{\mathbb{E}(M_{\tau_k}; \tau_k \leq n)}_{\geq b\mathbb{P}(\tau_k \leq n)} + \mathbb{E}(M_n; \tau_k > n); \\ a\mathbb{P}(\sigma_k \leq n) &\geq \mathbb{E} M_0 - \mathbb{E}(M_n; \sigma_k > n) \geq \\ &\geq b\mathbb{P}(\tau_k \leq n) + \underbrace{\mathbb{E}(M_n; \tau_k > n) - \mathbb{E}(M_n; \sigma_k > n)}_{= \mathbb{E}(M_n; \sigma_k \leq n < \tau_k) \geq 0};\end{aligned}$$

take  $n \rightarrow \infty$ . □

The same holds for supermartingales.

**16c2 Theorem.** Every positive martingale converges a.s. to an integrable random variable.

*Proof.* By Dubins's inequality, the martingale  $(M_n)_n$  cannot cross  $(a, b)$  infinitely many times. Almost surely, for all *rational*  $a < b$ , it crosses  $(a, b)$  finitely many times, which excludes the case  $\liminf M_n < a < b < \limsup M_n$ . It means that  $\liminf M_n = \limsup M_n$  a.s. Integrability of the limit follows from Fatou lemma. □

We turn to branching. Let  $(Z_n)_n$  be the simple branching process introduced in Sect. 6c.

Given  $\mathcal{F}_n$  we have  $\frac{1}{2}Z_{n+1} \sim \text{Binom}(Z_n, p)$ , thus  $\mathbb{E}(Z_{n+1} | \mathcal{F}_n) = 2pZ_n$ , which shows that  $M_n = \frac{1}{(2p)^n}Z_n$  is a (positive) martingale. By 16c2,  $M_n \rightarrow M_\infty$  a.s.,  $\mathbb{E} M_\infty \leq 1$ .

*Proof of 6c1.* Case  $p < 0.5$ : we have  $\sup_n \frac{1}{(2p)^n}Z_n < \infty$ , that is,  $Z_n = O((2p)^n)$ , a.s., which ultimately excludes the case  $Z_n \geq 1$ ; extinction.

Case  $p = 0.5$ :  $Z_n \rightarrow M_\infty$  a.s.; we have to prove that  $M_\infty = 0$  a.s. Assuming the contrary we take  $k > 0$  such that  $\mathbb{P}(M_\infty = k) > 0$ , then

$$\begin{aligned} \mathbf{1}_{Z_n=k} &\rightarrow \mathbf{1}_{M_\infty=k}; \\ \mathbf{1}_{Z_n=k, Z_{n+1}=k} &\rightarrow \mathbf{1}_{M_\infty=k}; \\ \mathbb{P}(Z_n = k) &\rightarrow \mathbb{P}(M_\infty = k); \\ \mathbb{P}(Z_n = k, Z_{n+1} = k) &\rightarrow \mathbb{P}(M_\infty = k); \\ \mathbb{P}(Z_{n+1} = k | Z_n = k) &\rightarrow 1, \end{aligned}$$

that is, the distribution  $\text{Binom}(k, 0.5)$  is concentrated at  $0.5k$ , — a contradiction.  $\square$

More detailed information on the branching process can be obtained using the generating functions

$$f_n(\theta) = \mathbb{E} \theta^{Z_n}.$$

We have  $f_0(\theta) = \theta$ ;  $f_1(\theta) = p\theta^2 + 1 - p$ ;  $\mathbb{E}(\theta^{Z_{n+1}} | Z_n = k) = (f_1(\theta))^k$  (think, why); thus  $f_{n+1}(\theta) = \mathbb{E}(f_1(\theta))^{Z_n} = f_n(f_1(\theta))$ , that is,

$$f_n = f_1 \circ \cdots \circ f_1 \quad (n \text{ times}).$$

Iterations for  $f_n(0) = \mathbb{P}(Z_n = 0)$  converge (draw a picture!) to the first root of the equation  $f_1(\theta) = \theta$ . Taking into account that  $f_1(1) = 1$  we solve the equation easily:  $\theta = (1 - p)/p$ . We get

$$\mathbb{P}(Z_n = 0) \rightarrow \frac{1-p}{p} = \mathbb{P}(M_\infty = 0).$$

In order to prove 6c2 it remains to prove that  $\mathbb{E} M_\infty = 1$ . It is sufficient to prove that  $M_n \rightarrow M_\infty$  in  $L_1$ , or in  $L_2$ , or just convergence of  $M_n$  in  $L_2$  (to whatever).

**16c3 Lemma.** If a martingale  $(M_n)_n$  satisfies  $\mathbb{E} M_n^2 < \infty$  for all  $n$  then random variables<sup>1</sup>  $M_{n+1} - M_n$  are mutually orthogonal.

<sup>1</sup>So-called martingale differences.

*Proof.* We have  $\mathbb{E}((M_{n+1} - M_n)Y) = 0$  for all  $Y \in L_\infty(\mathcal{F}_n)$  and therefore (by approximation) for all  $Y \in L_2(\mathcal{F}_n)$ ; apply it to  $Y = M_{k+1} - M_k$  for  $k < n$ .  $\square$

Thus,  $\|M_{n+k} - M_n\|^2 = \|M_{n+k} - M_{n+k-1}\|^2 + \dots + \|M_{n+1} - M_n\|^2$ ; convergence of  $(M_n)_n$  in  $L_2$  is equivalent to convergence of  $\sum \|M_{n+1} - M_n\|^2$ , that is, to  $\sup_n \|M_n\|^2 < \infty$ . Here is the conclusion.

**16c4 Proposition.** A positive<sup>1</sup> martingale bounded in  $L_2$  converges both in  $L_2$  and almost surely.

In order to prove 6c2 it remains to prove that  $(M_n)_n$  is bounded in  $L_2$ . We have

$$\begin{aligned} \text{Var } Z_1 &= 4p(1-p); \\ \text{Var}(Z_{n+1} | Z_n = k) &= k \text{Var } Z_1 = 4p(1-p)k; \\ \text{Var}(Z_{n+1} | Z_n) &= 4p(1-p)Z_n; \\ \text{Var}(M_{n+1} | M_n) &= \frac{1}{((2p)^{n+1})^2} \text{Var}(Z_{n+1} | Z_n) = \frac{4p(1-p)(2p)^n}{(2p)^{2n+2}} M_n; \\ \text{Var } M_{n+1} &= \mathbb{E} \text{Var}(M_{n+1} | M_n) + \text{Var} \mathbb{E}(M_{n+1} | M_n); \\ \text{Var } M_{n+1} - \text{Var } M_n &= \mathbb{E} \text{Var}(M_{n+1} | M_n) = \dots \end{aligned}$$

But why  $\mathbb{P}(M_\infty = 0) = \mathbb{P}(Z_n \rightarrow 0)$ ? (“ $\geq$ ” is evident.) Is the event  $1 \leq Z_n = o((2p)^n)$  negligible for  $p > 0.5$ ?

First,  $\mathbb{P}(M_\infty = 0 | \mathcal{F}_n) = (\mathbb{P}(M_\infty = 0))^{Z_n}$  (independent subtrees...).

Second,  $\mathbb{P}(M_\infty = 0 | \mathcal{F}_n) \rightarrow \mathbf{1}_{M_\infty=0}$  a.s., as we’ll see in 17d2.

Thus, on the event  $1 \leq Z_n = o((2p)^n)$  we have  $(\mathbb{P}(M_\infty = 0))^{Z_n} \rightarrow 1$ , therefore  $\mathbb{P}(M_\infty = 0) = 1$  in contradiction to  $\mathbb{E} M_\infty = 1$ .

The proof of 6c2 is thus finished (except for one claim postponed to “Conditioning”).

(In fact,  $\mathbb{P}(M_\infty = 0) = 1$  if and only if  $\mathbb{E}(X \ln X) = \infty$ ...)

## 17 Conditioning

### 17a What is the problem

The elementary definition

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

<sup>1</sup>The same holds for non-positive martingales, as we’ll see in 17c4.

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

**17a1 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  $\mathbb{P}(A|B) = \lim_n \mathbb{P}(A|B_n)$  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  $\mathbb{P}(A|B) = \lim_n \mathbb{P}(A|B_n)$  should not be used for all  $A$ ; otherwise, trying  $A = B$ , we get a paradox:  $\mathbb{P}(B|B) = 0$ .

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,

$$\lim_{\varepsilon \rightarrow 0^+} \mathbb{P}(X \leq 1 \mid -\varepsilon < Y < \varepsilon) \neq \lim_{\varepsilon \rightarrow 0^+} \mathbb{P}(X \leq 1 \mid -\varepsilon|X| < Y < \varepsilon|X|).$$

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

Sometimes conditioning is *really* impossible.

**17a2 Example.** Let  $(S_n)_n$  be the simple one-dimensional random walk and  $B = \{S_n \rightarrow +\infty\}$  (a zero-probability event). We note that  $B = \{S_n - S_{10} \rightarrow +\infty\}$ , “therefore”  $B$  is independent of  $S_1, \dots, S_{10}$ ; conditionally, given  $B$ , the walk behaves as usual, and we get the paradox,  $\mathbb{P}(B|B) = 0$ , once again.

**17a3 Example.** Let  $X \sim U(0, 1)$  and  $B = \{X \in \mathbb{Q}\}$ . We consider  $e^{2\pi i X}$ ; 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

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

covers the (practically important) case of discrete  $Y$  but absolutely continuous  $X$ .<sup>1</sup> 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.

### 17b Discrete case

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

$$\begin{aligned} \mathbb{P}(A|X = x) &= \frac{\mathbb{P}(A \cap X^{-1}(x))}{\mathbb{P}(X^{-1}(x))} = \frac{\sum_{\omega \in A \cap X^{-1}(x)} p(\omega)}{\sum_{\omega \in X^{-1}(x)} p(\omega)} = P_x(A) = f(x), \\ \mathbb{P}(A|\mathcal{F}_1) &= \mathbb{P}(A|X) = f(X) : \Omega \rightarrow \mathbb{R}; \\ \mathbb{E}(Y|X = x) &= \frac{\sum_{\omega \in A \cap X^{-1}(x)} Y(\omega)p(\omega)}{\sum_{\omega \in X^{-1}(x)} p(\omega)} = \int Y \, dP_x = g(x), \\ \mathbb{E}(Y|\mathcal{F}_1) &= \mathbb{E}(Y|X) = g(X) : \Omega \rightarrow \mathbb{R}; \end{aligned}$$

each *conditional measure*  $P_x$  is a probability measure on  $\Omega$ , concentrated on  $X^{-1}(x)$ ; the map  $x \rightarrow P_x$  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:  $Y = \mathbf{1}_A$ . Convergence of the series is ensured if  $Y$  is integrable (except for negligible  $x$ , if any).

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

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

$$\begin{aligned} \mathbb{E}(\mathbb{P}(A|\mathcal{F}_1)) &= \mathbb{P}(A), \\ \mathbb{E}(\mathbb{E}(Y|\mathcal{F}_1)) &= \mathbb{E}Y \end{aligned}$$

---

<sup>1</sup>See 17e15.

boil down to the *decomposition of measure*,

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

the latter expectation being taken in the linear space of signed measures...<sup>1</sup>

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

### 17c 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- $\sigma$ -field. 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  $L_2(\mathcal{F}) \rightarrow L_2(\mathcal{F}_1)$  and denote it  $Y \mapsto \mathbb{E}(Y | \mathcal{F}_1)$ . Note that  $\mathbb{E}(Y | \mathcal{F}_1)$  is an equivalence class. Orthogonality means that  $\langle Y - \mathbb{E}(Y | \mathcal{F}_1), X \rangle = 0$ , that is,

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

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

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

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

$$(17c1) \quad \mathbb{E}(\mathbb{E}(Y | \mathcal{F}_1); B) = \mathbb{E}(Y; B) \quad \text{for all } B \in \mathcal{F}_1;$$

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

$$\begin{aligned} \mathbb{E}(aX | \mathcal{F}_1) &= a\mathbb{E}(X | \mathcal{F}_1), \\ \mathbb{E}(X + Y | \mathcal{F}_1) &= \mathbb{E}(X | \mathcal{F}_1) + \mathbb{E}(Y | \mathcal{F}_1), \\ \text{if } X_n \rightarrow X \text{ in } L_2 \text{ then } \mathbb{E}(X_n | \mathcal{F}_1) &\rightarrow \mathbb{E}(X | \mathcal{F}_1) \text{ in } L_2. \end{aligned}$$

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

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

---

<sup>1</sup>Recall Sect. 15c.

It follows easily that the projection is positive:

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

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”

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

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

**17c2 Example.** Let  $Y \sim U(0, 1)$  and  $X = f(Y)$ ,

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

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

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

**17c3 Exercise.** Do it twice. Namely, (a) check it via (17c1); (b) derive it by minimization.

#### MORE ON MARTINGALES

Now we may use Definition 16a2 (rather than 16a3) in full generality.

A martingale bounded in  $L_2$  converges in  $L_2$  (recall Lemma 16c3 and the paragraph after it),  $M_n \rightarrow M_\infty$  in  $L_2$ . We have  $M_\infty = M_\infty^+ - M_\infty^-$  and  $M_n = \mathbb{E}(M_\infty | \mathcal{F}_n) = \mathbb{E}(M_\infty^+ | \mathcal{F}_n) - \mathbb{E}(M_\infty^- | \mathcal{F}_n)$ , the difference between two  $L_2$ -bounded positive martingales. Proposition 16c4 is thus generalized.

**17c4 Proposition.** A martingale bounded in  $L_2$  converges both in  $L_2$  and almost surely.

## 17d A convergence theorem

In the end of Sect. 16c, when proving  $\mathbb{P}(1 \leq Z_n = o((2p)^n)) = 0$ , we used the relation  $\mathbb{P}(A | \mathcal{F}_n) \rightarrow \mathbf{1}_A$  a.s. (for  $A \in \mathcal{F}_\infty$ ). This relation is proved below.

Recall Lemma 13b5: if  $\mathcal{F}_n \uparrow \mathcal{F}_\infty$  then  $\cup_n \mathcal{F}_n$  is dense in  $\mathcal{F}_\infty$ .

**17d1 Lemma.** If  $\mathcal{F}_n \uparrow \mathcal{F}_\infty$  then  $\cup_n L_2(\mathcal{F}_n)$  is dense in  $L_2(\mathcal{F}_\infty)$ .

*Proof.* First,  $\cup_n L_2(\mathcal{F}_n)$  contains  $\mathbf{1}_A$  for all  $A \in \cup_n \mathcal{F}_n$ . Second, the closure of  $\cup_n L_2(\mathcal{F}_n)$  contains  $\mathbf{1}_A$  for all  $A \in \mathcal{F}_\infty$ . Third, linear combinations of these  $\mathbf{1}_A$  are dense in  $L_2(\mathcal{F}_\infty)$ .  $\square$

It follows that  $\mathbb{E}(X | \mathcal{F}_n) \rightarrow X$  in  $L_2$  for all  $X \in L_2(\mathcal{F}_\infty)$ , and  $\mathbb{E}(X | \mathcal{F}_n) \rightarrow \mathbb{E}(X | \mathcal{F}_\infty)$  in  $L_2$  for all  $X \in L_2 = L_2(\mathcal{F})$ . But this is a martingale, and the difference of two positive martingales (since  $X = X^+ - X^-$ ); 16c4 ensures a.s. convergence, and we get the following.

**17d2 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- $\sigma$ -fields such that  $\mathcal{F}_n \uparrow \mathcal{F}_\infty$ . Then

$$\begin{aligned} \mathbb{E}(X | \mathcal{F}_n) &\rightarrow X \text{ a.s. and in } L_2 \text{ for all } X \in L_2(\mathcal{F}_\infty); \\ \mathbb{E}(X | \mathcal{F}_n) &\rightarrow \mathbb{E}(X | \mathcal{F}_\infty) \text{ a.s. and in } L_2 \text{ for all } X \in L_2(\mathcal{F}); \\ \mathbb{P}(A | \mathcal{F}_n) &\rightarrow \mathbf{1}_A \text{ a.s. and in } L_2 \text{ for all } A \in \mathcal{F}_\infty; \\ \mathbb{P}(A | \mathcal{F}_n) &\rightarrow \mathbb{P}(A | \mathcal{F}_\infty) \text{ a.s. and in } L_2 \text{ for all } A \in \mathcal{F}. \end{aligned}$$

## 17e Conditional measures

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

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

satisfies

$$\begin{aligned} 0 &\leq \mathbb{P}(A | \mathcal{F}_1) \leq 1 \quad \text{a.s.}, \\ \mathbb{P}(A_1 \uplus A_2 \uplus \dots | \mathcal{F}_1) &= \mathbb{P}(A_1 | \mathcal{F}_1) + \mathbb{P}(A_2 | \mathcal{F}_1) + \dots \quad \text{a.s.}, \\ \forall B \in \mathcal{F}_1 \quad \mathbb{P}(B | \mathcal{F}_1) &= \mathbf{1}_B; \\ \mathbb{E}(\mathbb{P}(A | \mathcal{F}_1)) &= \mathbb{P}(A), \end{aligned}$$

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

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

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

It is a *disintegration* of  $P$  into probability measures  $P_\omega$  localized on corresponding parts of the partition.

In general it does not go...

**17e1 Definition.** Let  $(\Omega, \mathcal{F}, P)$  be a probability space and  $\mathcal{F}_1 \subset \mathcal{F}$  a sub- $\sigma$ -field. A *regular conditional probability* (given  $\mathcal{F}_1$ ) is a family  $(P_\omega)_{\omega \in \Omega}$  of probability measures  $P_\omega$  on  $(\Omega, \mathcal{F})$  such that for every  $A \in \mathcal{F}$  the function  $\omega \mapsto P_\omega(A)$  belongs to the equivalence class  $\mathbb{P}(A | \mathcal{F}_1)$ .

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 17e3).

**17e2 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  $\sigma$ -field;  $P$ -null sets are *not* added; (c)  $\mathcal{F}_1 \subset \mathcal{F}$  is an arbitrary  $\sigma$ -field.

**17e3 Example.** There exists<sup>1</sup>  $Z \subset [0, 1]$  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  $P((A \cap Z) \uplus (B \setminus Z)) = 0.5 \text{ mes } A + 0.5 \text{ mes } B$ .

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

**17e4 Exercise.** Show that  $\mathbb{P}(Z | \mathcal{F}_1) = 0.5$  a.s.

---

<sup>1</sup>Using the choice axiom, of course.

<sup>2</sup>W.r.t. Lebesgue measure.

## PROOF OF THEOREM 17E2

The uniqueness part of Theorem 17e2 is easy: almost every  $x$  satisfy  $P_x(I) = P'_x(I)$  for all rational intervals  $I$ , which implies  $P_x = P'_x$ .

The existence part needs more effort.

**17e5 Definition.** A measurable space is a pair  $(\Omega, \mathcal{F})$  of a set  $\Omega$  and a  $\sigma$ -field  $\mathcal{F}$  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).

An embedding of one measurable space into another is, by definition, an isomorphism between the former and a part of the latter.

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

**17e6 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- $\sigma$ -field of  $\mathcal{F}$ .

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).

**17e7 Lemma.** The Cantor set (with its Borel  $\sigma$ -field) is a good measurable space.

**17e8 Lemma.** The real line (with its Borel  $\sigma$ -field) is embeddable (as a measurable space) into the Cantor set (with its Borel  $\sigma$ -field).

**17e9 Lemma.** A part of a good measurable space is a good measurable space. That is, 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- $\sigma$ -fields,  $\mathcal{F}'_1 \subset \mathcal{F}'$  on  $\Omega'$  and  $\mathcal{F}''_1 \subset \mathcal{F}''$  on  $\Omega''$ , we get the corresponding sub- $\sigma$ -field  $\mathcal{F}_1 \subset \mathcal{F}$  on  $\Omega$  (namely,  $\mathcal{F}_1 = \{A \uplus B : A \in \mathcal{F}'_1, B \in \mathcal{F}''_1\}$ ).

*Proof of 17e9.* 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- $\sigma$ -field  $\mathcal{F}'_1 \subset \mathcal{F}'$ , we choose some sub- $\sigma$ -field  $\mathcal{F}''_1 \subset \mathcal{F}''$  (no matter which one) and get the corresponding sub- $\sigma$ -field  $\mathcal{F}_1 \subset \mathcal{F}$  on  $\Omega$  such that  $L_2(\Omega, \mathcal{F}_1) = L_2(\Omega', \mathcal{F}'_1)$  and therefore  $\mathbb{E}(\cdot | \mathcal{F}_1) = \mathbb{E}(\cdot | \mathcal{F}'_1)$ .

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  $P_\omega(\Omega') = \mathbb{P}(\Omega' | \mathcal{F}_1)(\omega) = \mathbf{1}_{\Omega'}(\omega) = 1$ . Thus we may treat  $P_\omega$  as a probability measure on  $(\Omega', \mathcal{F}')$ , getting a regular conditional probability for  $\mathcal{F}'_1$ .  $\square$

*Proof of 17e8.* 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.  $\square$

*Proof of 17e7.* The Borel  $\sigma$ -field of the Cantor set is generated by the countable algebra  $\mathcal{A}$  of clopen (that is, both closed and open) sets. Every finitely additive set function on this algebra is automatically  $\sigma$ -additive, due to compactness: if  $A = A_1 \uplus A_2 \uplus \dots$  then  $A_k = \emptyset$  for all  $k$  large enough. By Caratheodory theorem, every finitely additive set function on this algebra extends to a measure.

We define  $P_\omega$  by

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

choosing a function in each equivalence class (countably many choices) and extend  $P_\omega$  to a probability measure. Additivity holds a.s. (countably many equalities!) and is easily ensured everywhere.

Let  $B \in \mathcal{F}$ ; we have to prove that  $\mathbb{P}(B | \mathcal{F}_1) = P_\bullet(B)$  a.s. (denoting the function  $\omega \mapsto P_\omega(\dots)$  by  $P_\bullet(\dots)$ ). Given  $\varepsilon > 0$  we take  $A_n, C_n \in \mathcal{A}$  and  $A, C \in \mathcal{F}$  such that  $A_n \downarrow A$ ,  $C_n \uparrow C$ ,  $A \subset B \subset C$  and  $P(C) - P(A) \leq \varepsilon$  (“sandwich”). Then we have  $P_\bullet(A_n) = \mathbb{P}(A_n | \mathcal{F}_1)$  a.s., thus  $P_\bullet(A) = \mathbb{P}(A | \mathcal{F}_1)$  a.s. (monotone convergence); the same for  $C$ ; and so,

$$\begin{aligned} \mathbb{P}(A | \mathcal{F}_1) &= P_\bullet(A) \leq P_\bullet(B) \leq P_\bullet(C) = \mathbb{P}(C | \mathcal{F}_1), \\ \mathbb{P}(A | \mathcal{F}_1) &\leq \mathbb{P}(B | \mathcal{F}_1) \leq \mathbb{P}(C | \mathcal{F}_1); \\ |P_\bullet(B) - \mathbb{P}(B | \mathcal{F}_1)| &\leq \mathbb{P}(C | \mathcal{F}_1) - \mathbb{P}(A | \mathcal{F}_1); \\ \mathbb{E}(\mathbb{P}(C | \mathcal{F}_1) - \mathbb{P}(A | \mathcal{F}_1)) &= \mathbb{P}(C) - \mathbb{P}(A) \leq \varepsilon; \end{aligned}$$

we take  $\varepsilon_n \rightarrow 0$  and get  $|P_\bullet(B) - \mathbb{P}(B | \mathcal{F}_1)| \leq \inf_n (\mathbb{P}(C_{\varepsilon_n} | \mathcal{F}_1) - \mathbb{P}(A_{\varepsilon_n} | \mathcal{F}_1)) = 0$  a.s.  $\square$

Theorem 17e2 is thus proved.

The same holds for  $\mathbb{R}^n$  and many other spaces.<sup>1</sup>

The requirement that the function  $\omega \mapsto P_\omega(A)$  belongs to the equivalence class  $\mathbb{P}(A|\mathcal{F}_1)$  can be reformulated using (17c1):

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

Or, equivalently,

$$(17e10) \quad P_\omega(B) = \mathbf{1}_B(\omega) \quad \text{for all } B \in \mathcal{F}_1,$$

$$(17e11) \quad \int_\Omega P_\omega(A) P(d\omega) = P(A) \quad \text{for all } A \in \mathcal{F}.$$

Indeed, (17e10) implies  $\int_B P_\omega(A) P(d\omega) = \int_\Omega P_\omega(A \cap B) P(d\omega)$ .

**Proof of Theorem 7.1.** Theorem 17e2 (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

$$\mu(A) = \int P_{x,y}(A) \mu(dx dy).$$

Measurability of  $P_{x,y}$  w.r.t.  $\mathcal{F}_1$  means that  $P_{x,y} = P_x$ . Property (17e10) 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  $\mu$  is  $\nu$ .  $\square$

**17e12 Exercise.** Prove uniqueness of  $(\mu_x)_{x \in \mathbb{R}}$  up to a  $\nu$ -negligible set.

**17e13 Exercise.** Disintegrate the joint distribution  $\mu$  of the random variables  $X, Y$  of Example 17c2. Namely, guess the measures  $\mu_x$  and check the equality  $\mu(A) = \int \mu_x(A_x) \nu(dx)$ .

**17e14 Exercise.** Disintegrate a measure that has a density (w.r.t. the 2-dimensional Lebesgue measure). Namely, guess the measures  $\mu_x$  and check the equality  $\mu(A) = \int \mu_x(A_x) \nu(dx)$ .

**17e15 Exercise.** Derive formulas for conditioning of a discrete random variable  $Y : \Omega \rightarrow \mathbb{Z}$  on a continuous random variable  $X : \Omega \rightarrow \mathbb{R}$  that has a density  $f_X$ .

---

<sup>1</sup>In fact, for all Polish spaces.

## CONDITIONAL MEASURES AND CONDITIONAL EXPECTATIONS

**17e16 Proposition.** Let  $(\Omega, \mathcal{F}, P)$  be a probability space,  $\mathcal{F}_1 \subset \mathcal{F}$  a sub- $\sigma$ -field,  $(P_\omega)_{\omega \in \Omega}$  a regular conditional probability (given  $\mathcal{F}_1$ ), and  $Y \in L_1(\Omega, \mathcal{F}, P)$ . Then the function  $\omega \mapsto \int Y dP_\omega$  belongs to the equivalence class  $\mathbb{E}(Y | \mathcal{F}_1)$ .

*Proof.* The relation holds for indicators, and (by linearity) for their linear combinations. Let it hold for each  $Y_n$ , and  $0 \leq Y_n \uparrow Y$  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  $Y_n \rightarrow Y$  in  $L_1$ , therefore  $\mathbb{E}(Y_n | \mathcal{F}_1) \rightarrow \mathbb{E}(Y | \mathcal{F}_1)$  in  $L_1$ . On the other hand,  $\int Y_n dP_\omega \uparrow \int Y dP_\omega \in [0, \infty]$  for each  $\omega$  by the monotone convergence theorem; the rest is easy.  $\square$

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

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

“Taking out what is known”:

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

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

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

Conditional Jensen’s inequality:

$$\mathbb{E}(h(Y) | \mathcal{F}_1) \geq h(\mathbb{E}(Y | \mathcal{F}_1))$$

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

$$\|\mathbb{E}(Y | \mathcal{F}_1)\|_p \leq \|Y\|_p.$$

Conditional Cauchy-Schwartz inequality:

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

And so on.

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