Course: Theory of Probability I
Term: Fall 2013
Instructor: Gordan Zitkovic

Lecture 10
Conditional Expectation

The definition and existence of conditional expectation

For events A, B with $\mathbb{P}[B] > 0$, we recall the familiar object

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

We say that $\mathbb{P}[A|B]$ the **conditional probability of A, given B.** It is important to note that the condition $\mathbb{P}[B] > 0$ is crucial. When X and Y are random variables defined on the same probability space, we often want to give a meaning to the expression $\mathbb{P}[X \in A|Y = y]$, even though it is usually the case that $\mathbb{P}[Y = y] = 0$. When the random vector (X, Y) admits a joint density $f_{X,Y}(x, y)$, and $f_Y(y) > 0$, the concept of conditional density $f_{X|Y=y}(x) = f_{X,Y}(x, y)/f_Y(y)$ is introduced and the quantity $\mathbb{P}[X \in A|Y = y]$ is given meaning via $\int_A f_{X|Y=y}(x, y)\, dx$. While this procedure works well in the restrictive case of absolutely continuous random vectors, we will see how it is encompassed by a general concept of a conditional expectation. Since probability is simply an expectation of an indicator, and expectations are linear, it will be easier to work with expectations and no generality will be lost.

Two main conceptual leaps here are: 1) we condition with respect to a σ-algebra, and 2) we view the conditional expectation itself as a random variable. Before we illustrate the concept in discrete time, here is the definition.

Definition 10.1. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $X \in \mathcal{L}^1$ be a random variable. We say that the random variable ξ is (a version of) the **conditional expectation of X with respect to $\mathcal{G}$** - and denote it by $\mathbb{E}[X|\mathcal{G}]$ - if

1. $\xi \in \mathcal{L}^1$.
2. ξ is $\mathcal{G}$-measurable,
3. $\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[X \mathbf{1}_A]$, for all $A \in \mathcal{G}$.

Example 10.2. Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space where $\Omega = \{a, b, c, d, e, f\}$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}$ is uniform. Let X, Y and Z be random variables given by (in the obvious notation)

$$X \sim \begin{pmatrix} a & b & c & d & e & f \\ 1 & 3 & 3 & 5 & 5 & 7 \end{pmatrix},$$

$$Y \sim \begin{pmatrix} a & b & c & d & e & f \\ 2 & 2 & 1 & 1 & 7 & 7 \end{pmatrix} \text{ and } Z \sim \begin{pmatrix} a & b & c & d & e & f \\ 3 & 3 & 3 & 3 & 2 & 2 \end{pmatrix}$$

We would like to think about $\mathbb{E}[X|\mathcal{G}]$ as the average of $X(\omega)$ over all ω which are consistent with the our current information (which is $\mathcal{G}$). For example, if $\mathcal{G} = \sigma(Y)$, then the information contained in $\mathcal{G}$ is exactly the information about the exact value of Y. Knowledge of the fact that $Y = y$ does not necessarily reveal the "true" ω, but certainly rules out all those ω for which $Y(\omega) \neq y$.

In our specific case, if we know that $Y = 2$, then $\omega = a$ or $\omega = b$, and the expected value of X, given that $Y = 2$, is $\frac{1}{2}X(a) + \frac{1}{2}X(b) = 2$. Similarly, this average equals 4 for $Y = 1$, and 6 for $Y = 7$. Let us show that the random variable ξ defined by this average, i.e.,

$$\xi \sim \begin{pmatrix} a & b & c & d & e & f \\ 2 & 2 & 4 & 4 & 6 & 6 \end{pmatrix},$$

satisfies the definition of $\mathbb{E}[X|\sigma(Y)]$, as given above. The integrability is not an issue (we are on a finite probability space), and it is clear that ξ is measurable with respect to $\sigma(Y)$. Indeed, the atoms of $\sigma(Y)$ are $\{a, b\}$, $\{c, d\}$ and $\{e, f\}$, and ξ is constant over each one of those. Finally, we need to check that

$$\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[X \mathbf{1}_A], \text{ for all } A \in \sigma(Y),$$

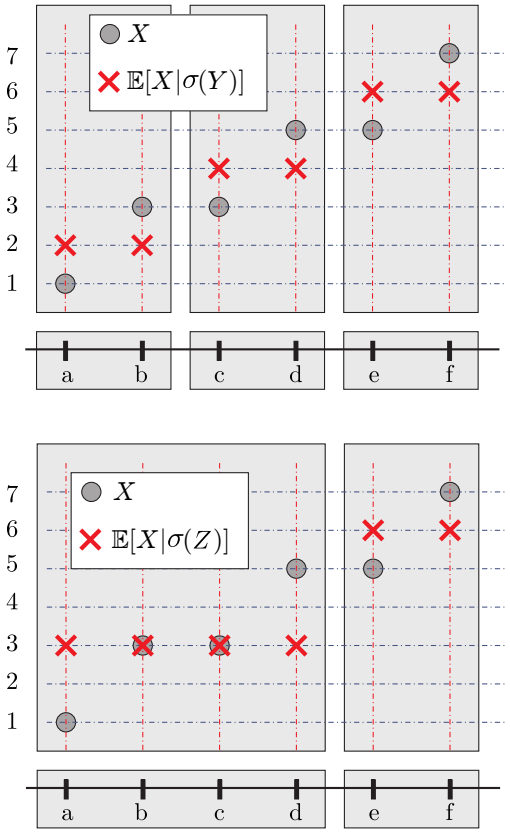
which for an atom A translates into

$$\xi(\omega) = \tfrac{1}{\mathbb{P}[A]}\mathbb{E}[X\mathbf{1}_A] = \sum_{\omega' \in A} X(\omega')\mathbb{P}[\{\omega'\}|A], \text{ for all } \omega \in A.$$

The moral of the story is that when A is an atom, part 3. of Definition 10.1 translates into a requirement that ξ be constant on A with value equal to the expectation of X over A with respect to the conditional probability $\mathbb{P}[\cdot|A]$. In the general case, when there are no atoms, 3. still makes sense and conveys the same message.

Btw, since the atoms of $\sigma(Z)$ are $\{a, b, c, d\}$ and $\{e, f\}$, it is clear that

$$\mathbb{E}[X|\sigma(Z)](\omega) = \begin{cases} 3, & \omega \in \{a, b, c, d\}, \\ 6, & \omega \in \{e, f\}. \end{cases}$$



Look at the illustrations above and convince yourself that

$$\mathbb{E}[\mathbb{E}[X|\sigma(Y)]|\sigma(Z)] = \mathbb{E}[X|\sigma(Z)].$$

A general result along the same lines - called the *tower property of conditional expectation* - will be stated and proved below.

Our first task is to prove that conditional expectations always exist. When Ω is finite (as explained above) or countable, we can always construct them by averaging over atoms. In the general case, a different argument is needed. In fact, here are two:

Proposition 10.3. *Let $\mathcal{G}$ be a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. Then*

1. *there exists a conditional expectation $\mathbb{E}[X|\mathcal{G}]$ for any $X \in \mathcal{L}^1$, and*
2. *any two conditional expectations of $X \in \mathcal{L}^1$ are equal $\mathbb{P}$-a.s.*

Proof. (*Uniqueness*): Suppose that ξ and ξ' both satisfy 1., 2. and 3. of Definition 10.1. Then

$$\mathbb{E}[\xi \mathbf{1}_A] = \mathbb{E}[\xi' \mathbf{1}_A], \text{ for all } A \in \mathcal{G}.$$

For $A_n = \{\xi' - \xi \geq \frac{1}{n}\}$, we have $A_n \in \mathcal{G}$ and so

$$\mathbb{E}[\xi \mathbf{1}_{A_n}] = \mathbb{E}[\xi' \mathbf{1}_{A_n}] \geq \mathbb{E}[(\xi + \tfrac{1}{n})\mathbf{1}_{A_n}] = \mathbb{E}[\xi \mathbf{1}_{A_n}] + \tfrac{1}{n}\mathbb{P}[A_n].$$

Consequently, $\mathbb{P}[A_n] = 0$, for all $n \in \mathbb{N}$, so that $\mathbb{P}[\xi' > \xi] = 0$. By a symmetric argument, we also have $\mathbb{P}[\xi' < \xi] = 0$.

(*Existence*): By linearity, it will be enough to prove that the conditional expectation exists for $X \in \mathcal{L}^1_+$.

1. *A Radon-Nikodym argument*. Suppose, first, that $X \geq 0$ and $\mathbb{E}[X] = 1$, as the general case follows by additivity and scaling. Then the prescription

$$\mathbb{Q}[A] = \mathbb{E}[X \mathbf{1}_A],$$

defines a probability measure on $(\Omega, \mathcal{F})$, which is absolutely continuous with respect to $\mathbb{P}$. Let $\mathbb{Q}^{\mathcal{G}}$ be the restriction of $\mathbb{Q}$ to $\mathcal{G}$; it is trivially absolutely continuous with respect to the restriction $\mathbb{P}^{\mathcal{G}}$ of $\mathbb{P}$ to $\mathcal{G}$. The Radon-Nikodym theorem - applied to the measure space $(\Omega, \mathcal{G}, \mathbb{P}^{\mathcal{G}})$ and the measure $\mathbb{Q}^{\mathcal{G}} \ll \mathbb{P}^{\mathcal{G}}$ - guarantees the existence of the Radon-Nikodym derivative

$$\xi = \frac{d\mathbb{Q}^{\mathcal{G}}}{d\mathbb{P}^{\mathcal{G}}} \in \mathbb{L}^1_+(\Omega, \mathcal{G}, \mathbb{P}^{\mathcal{G}}).$$

For $A \in \mathcal{G}$, we thus have

$$\mathbb{E}[X \mathbf{1}_A] = \mathbb{Q}[A] = \mathbb{Q}^{\mathcal{G}}[A] = \mathbb{E}^{\mathbb{P}^{\mathcal{G}}}[\xi \mathbf{1}_A] = \mathbb{E}[\xi \mathbf{1}_A].$$

where the last equality follows from the fact that $\xi \mathbf{1}_A$ is $\mathcal{G}$-measurable. Therefore, ξ is (a version of) the conditional expectation $\mathbb{E}[X|\mathcal{G}]$.

1. *An $\mathcal{L}^2$-argument*. Suppose, first, that $X \in \mathcal{L}^2$. Let H be the family of all $\mathcal{G}$-measurable elements in $\mathcal{L}^2$. Let $\bar{H}$ denote the closure of H in the topology induced by $\mathcal{L}^2$-convergence. Being a closed and convex (why?) subset of $\mathcal{L}^2$, $\bar{H}$ satisfies all the conditions of Problem **??** so that there exists $\xi \in \bar{H}$ at the minimal $\mathcal{L}^2$-distance from X (when $X \in \bar{H}$, we take $\xi = X$). The same problem states that ξ has the following property:

$$\mathbb{E}[(\eta - \xi)(X - \xi)] \geq 0 \text{ for all } \eta \in \bar{H},$$

and, since $\bar{H}$ is a linear space, we have

$$\mathbb{E}[(\eta - \xi)(X - \xi)] = 0, \text{for all } \eta \in \bar{H}.$$

It remains to pick η of the form $\eta = \xi + \mathbf{1}_A \in \bar{H}$, $A \in \mathcal{G}$, to conclude that

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\xi \mathbf{1}_A], \text{ for all } A \in \mathcal{G}.$$

Our next step is to show that ξ is $\mathcal{G}$-measurable (after a modification on a null set, perhaps). Since $\xi \in \bar{H}$, there exists a sequence $\{\xi_n\}_{n\in\mathbb{N}}$ such that $\xi_n \to \xi$ in $\mathcal{L}^2$. By Corollary **??**, $\xi_{n_k} \overset{a.s.}{\to} \xi$, for some subsequence $\{\xi_{n_k}\}_{k\in\mathbb{N}}$ of $\{\xi_n\}_{n\in\mathbb{N}}$. Set $\xi' = \liminf_{k\in\mathbb{N}} \xi_{n_k} \in \mathcal{L}^0([-\infty, \infty], \mathcal{G})$ and $\hat{\xi} = \xi' \mathbf{1}_{\{|\xi'|<\infty\}}$, so that $\hat{\xi} = \xi$, a.s., and $\hat{\xi}$ is $\mathcal{G}$-measurable.

We still need to remove the restriction $X \in \mathcal{L}^2_+$. We start with a general $X \in \mathcal{L}^1_+$ and define $X_n = \min(X, n) \in \mathcal{L}^\infty_+ \subseteq \mathcal{L}^2_+$. Let $\xi_n = \mathbb{E}[X_n|\mathcal{G}]$, and note that $\mathbb{E}[\xi_{n+1}\mathbf{1}_A] = \mathbb{E}[X_{n+1}\mathbf{1}_A] \geq \mathbb{E}[X_n\mathbf{1}_A] = \mathbb{E}[\xi_n\mathbf{1}_A]$. It follows (just like in the proof of uniqueness above) that $\xi_n \leq \xi_{n+1}$, a.s. We define $\xi = \sup_n \xi_n$, so that $\xi_n \nearrow \xi$, a.s. Then, for $A \in \mathcal{G}$, the monotone-convergence theorem implies that

$$\mathbb{E}[X\mathbf{1}_A] = \lim_n \mathbb{E}[X_n\mathbf{1}_A] = \lim_n \mathbb{E}[\xi_n\mathbf{1}_A] = \mathbb{E}[\xi\mathbf{1}_A],$$

and it is easy to check that $\xi\mathbf{1}_{\{\xi<\infty\}} \in \mathcal{L}^1(\mathcal{G})$ is a version of $\mathbb{E}[X|\mathcal{G}]$. □

Remark 10.4. There is no canonical way to choose "the version" of the conditional expectation. We follow the convention started with Radon-Nikodym derivatives, and interpret a statement such at $\xi \leq \mathbb{E}[X|\mathcal{G}]$, a.s., to mean that $\xi \leq \xi'$, a.s., for any version ξ' of the conditional expectation of X with respect to $\mathcal{G}$.

If we use the symbol $\mathbb{L}^1$ to denote the set of all a.s.-equivalence classes of random variables in $\mathcal{L}^1$, we can write:

$$\mathbb{E}[\cdot|\mathcal{G}] : \mathcal{L}^1(\mathcal{F}) \to \mathbb{L}^1(\mathcal{G}),$$

but $\mathbb{L}^1(\mathcal{G})$ cannot be replaced by $\mathcal{L}^1(\mathcal{G})$ in a natural way. Since $X = X'$, a.s., implies that $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X'|\mathcal{G}]$, a.s. (why?), we consider conditional expectation as a map from $\mathbb{L}^1(\mathcal{F})$ to $\mathbb{L}^1(\mathcal{G})$

$$\mathbb{E}[\cdot|\mathcal{G}] : \mathbb{L}^1(\mathcal{F}) \to \mathbb{L}^1(\mathcal{G}).$$

Properties

Conditional expectation inherits many of the properties from the "ordinary" expectation. Here are some familiar and some new ones:

Proposition 10.5. *Let $X, Y, \{X_n\}_{n\in\mathbb{N}}$ be random variables in $\mathcal{L}^1$, and let $\mathcal{G}$ and $\mathcal{H}$ be sub-σ-algebras of $\mathcal{F}$. Then*

1. (linearity) $\mathbb{E}[\alpha X + \beta Y|\mathcal{G}] = \alpha\mathbb{E}[X|\mathcal{G}] + \beta\mathbb{E}[Y|\mathcal{G}]$, *a.s.*
2. (monotonicity) $X \leq Y$, *a.s., implies* $\mathbb{E}[X|\mathcal{G}] \leq \mathbb{E}[Y|\mathcal{G}]$, *a.s.*
3. (identity on $\mathbb{L}^1(\mathcal{G})$) *If X is $\mathcal{G}$-measurable, then $X = \mathbb{E}[X|\mathcal{G}]$, a.s. In particular, $c = \mathbb{E}[c|\mathcal{G}]$, for any constant $c \in \mathbb{R}$.*
4. (conditional Jensen's inequality) *If $\psi : \mathbb{R} \to \mathbb{R}$ is convex and $\mathbb{E}[|\psi(X)|] < \infty$ then*
$$\mathbb{E}[\psi(X)|\mathcal{G}] \geq \psi(\mathbb{E}[X|\mathcal{G}]), \text{ a.s.}$$
5. ($\mathcal{L}^p$-nonexpansivity) *If $X \in \mathcal{L}^p$, for $p \in [1, \infty]$, then $\mathbb{E}[X|\mathcal{G}] \in \mathbb{L}^p$ and*
$$||\mathbb{E}[X|\mathcal{G}]||_{\mathcal{L}^p} \leq ||X||_{\mathcal{L}^p}.$$
In particular,
$$\mathbb{E}[|X|\,|\mathcal{G}] \geq |\mathbb{E}[X|\mathcal{G}]| \text{ a.s.}$$
6. (pulling out what's known) *If Y is $\mathcal{G}$-measurable and $XY \in \mathcal{L}^1$, then*
$$\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}], \text{ a.s.}$$
7. ($\mathbb{L}^2$-projection) *If $X \in \mathcal{L}^2$, then $\xi^* = \mathbb{E}[X|\mathcal{G}]$ minimizes $\mathbb{E}[(X-\xi)^2]$ over all $\mathcal{G}$-measurable random variables $\xi \in \mathcal{L}^2$.*
8. (tower property) *If $\mathcal{H} \subseteq \mathcal{G}$, then*
$$\mathbb{E}[\mathbb{E}[X|\mathcal{G}]|\mathcal{H}] = \mathbb{E}[X|\mathcal{H}], \text{ a.s..}$$
9. (irrelevance of independent information) *If $\mathcal{H}$ is independent of $\sigma(\mathcal{G}, \sigma(X))$ then*
$$\mathbb{E}[X|\sigma(\mathcal{G}, \mathcal{H})] = \mathbb{E}[X|\mathcal{G}], \text{ a.s.}$$
In particular, if X is independent of $\mathcal{H}$, then $\mathbb{E}[X|\mathcal{H}] = \mathbb{E}[X]$, a.s.

10. (conditional monotone-convergence theorem) *If* $0 \leq X_n \leq X_{n+1}$, *a.s., for all* $n \in \mathbb{N}$ *and* $X_n \to X \in \mathcal{L}^1$, *a.s., then*

$$\mathbb{E}[X_n|\mathcal{G}] \nearrow \mathbb{E}[X|\mathcal{G}], \text{ a.s.}$$

11. (conditional Fatou's lemma) *If* $X_n \geq 0$, *a.s., for all* $n \in \mathbb{N}$, *and* $\liminf_n X_n \in \mathcal{L}^1$, *then*

$$\mathbb{E}[\liminf_n X|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}], \text{ a.s.}$$

12. (conditional dominated-convergence theorem) *If* $|X_n| \leq Z$, *for all* $n \in \mathbb{N}$ *and some* $Z \in \mathcal{L}^1$, *and if* $X_n \to X$, *a.s., then*

$$\mathbb{E}[X_n|\mathcal{G}] \to \mathbb{E}[X|\mathcal{G}], \text{ a.s. and in } \mathcal{L}^1.$$

Note: Some of the properties are proved in detail. The others are only commented upon, since they are either similar to the other ones or otherwise not hard.

Proof.

1. (linearity) $\mathbb{E}[(\alpha X + \beta Y)\mathbf{1}_A] = \mathbb{E}[(\alpha\mathbb{E}[X|\mathcal{G}] + \beta\mathbb{E}[Y|\mathcal{G}])\mathbf{1}_A]$, for $A \in \mathcal{G}$.
2. (monotonicity) Use $A = \{\mathbb{E}[X|\mathcal{G}] > \mathbb{E}[Y|\mathcal{G}]\} \in \mathcal{G}$ to obtain a contradiction if $\mathbb{P}[A] > 0$.
3. (identity on $\mathbb{L}^1(\mathcal{G})$) Check the definition.
4. (conditional Jensen's inequality) Use the result of Lemma **??** which states that $\psi(x) = \sup_{n\in\mathbb{N}}(a_n + b_n x)$, where $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ are sequences of real numbers.
5. ($\mathbb{L}^p$-nonexpansivity) For $p \in [1, \infty)$, apply conditional Jensen's inequality with $\psi(x) = |x|^p$. The case $p = \infty$ follows directly.
6. (pulling out what's known) For Y $\mathcal{G}$-measurable and $XY \in \mathcal{L}^1$, we need to show that

$$\mathbb{E}[XY\mathbf{1}_A] = \mathbb{E}[Y\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A], \text{ for all } A \in \mathcal{G}. \tag{10.1}$$

Let us prove a seemingly less general statement:

$$\mathbb{E}[ZX] = \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]], \text{ for all } \mathcal{G}\text{-measurable } Z \text{ with } ZX \in \mathcal{L}^1. \tag{10.2}$$

The statement (10.1) will follow from it by taking $Z = Y\mathbf{1}_A$. For $Z = \sum_{k=1}^n \alpha_k \mathbf{1}_{A_k}$, (10.2) is a consequence of the definition of conditional expectation and linearity. Let us assume that both Z and X are nonnegative and $ZX \in \mathcal{L}^1$. In that case we can find a nondecreasing sequence $\{Z_n\}_{n\in\mathbb{N}}$ of non-negative simple random variables with $Z_n \nearrow Z$. Then $Z_n X \in \mathcal{L}^1$ for all $n \in \mathbb{N}$ and the monotone convergence theorem implies that

$$\mathbb{E}[ZX] = \lim_n \mathbb{E}[Z_n X] = \lim_n \mathbb{E}[Z_n\mathbb{E}[X|\mathcal{G}]] = \mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]].$$

Our next task is to relax the assumption $X \in \mathcal{L}^1_+$ to the original one $X \in \mathcal{L}^1$. In that case, the $\mathcal{L}^p$-nonexpansivity for $p = 1$ implies that

$$|\mathbb{E}[X|\mathcal{G}]| \leq \mathbb{E}[|X|\,|\mathcal{G}] \text{ a.s.},$$

and so

$$|Z_n\mathbb{E}[X|\mathcal{G}]| \leq Z_n\mathbb{E}[|X|\,|\mathcal{G}] \leq Z\mathbb{E}[|X|\,|\mathcal{G}].$$

We know from the previous case that

$$\mathbb{E}[Z\mathbb{E}[|X|\,|\mathcal{G}]] = \mathbb{E}[Z\,|X|], \text{ so that } Z\mathbb{E}[|X|\,|\mathcal{G}] \in \mathcal{L}^1.$$

We can, therefore, use the dominated convergence theorem to conclude that

$$\mathbb{E}[Z\mathbb{E}[X|\mathcal{G}]] = \lim_n \mathbb{E}[Z_n\mathbb{E}[X|\mathcal{G}]] = \lim_n \mathbb{E}[Z_nX] = \mathbb{E}[ZX].$$

Finally, the case of a general Z follows by linearity.

7. ($\mathbb{L}^2$-projection) It is enough to show that $X - \mathbb{E}[X|\mathcal{G}]$ is orthogonal to all $\mathcal{G}$-measurable $\xi \in \mathcal{L}^2$. for that we simply note that for $\xi \in \mathcal{L}^2$, $x \in \mathcal{G}$, we have

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}[X|\mathcal{G}])\xi] &= \mathbb{E}[\xi X] - \mathbb{E}[\xi\mathbb{E}[X|\mathcal{G}]]\\ &= \mathbb{E}[\xi X] - \mathbb{E}[\mathbb{E}[\xi X|\mathcal{G}]] = 0.\end{aligned}$$

8. (tower property) Use the definition.

9. (irrelevance of independent information) We assume $X \geq 0$ and show that

$$\mathbb{E}[X\mathbf{1}_A] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_A], \text{ a.s. for all } A \in \sigma(\mathcal{G},\mathcal{H}). \tag{10.3}$$

Let $\mathcal{L}$ be the collection of all $A \in \sigma(\mathcal{G},\mathcal{H})$ such that (10.3) holds. It is straightforward that $\mathcal{L}$ is a λ-system, so it will be enough to establish (10.3) for some π-system that generates $\sigma(\mathcal{G},\mathcal{H})$. One possibility is $\mathcal{P} = \{G \cap H : G \in \mathcal{G}, H \in \mathcal{H}\}$, and for $G \cap H \in \mathcal{P}$ we use independence of $\mathbf{1}_H$ and $\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G$, as well as the independence of $\mathbf{1}_H$ and $X\mathbf{1}_G$ to get

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_{G\cap H}] &= \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G\mathbf{1}_H] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H]\\ &= \mathbb{E}[X\mathbf{1}_G]\mathbb{E}[\mathbf{1}_H] = \mathbb{E}[X\mathbf{1}_{G\cap H}]\end{aligned} \tag{10.4}$$

10. (conditional monotone-convergence theorem) By monotonicity, we have $\mathbb{E}[X_n|\mathcal{G}] \nearrow \xi \in \mathcal{L}^0_+(\mathcal{G})$, a.s. The monotone convergence theorem implies that, for each $A \in \mathcal{G}$,

$$\mathbb{E}[\xi\mathbf{1}_A] = \lim_n \mathbb{E}[\mathbf{1}_A\mathbb{E}[X_n|\mathcal{G}]] = \lim_n \mathbb{E}[\mathbf{1}_AX_n] = \mathbb{E}[\mathbf{1}_AX].$$

11. (conditional Fatou's lemma) Set $Y_n = \inf_{k\geq n} X_k$, so that $Y_n \nearrow Y = \liminf_k X_k$. By monotonicity,

$$\mathbb{E}[Y_n|\mathcal{G}] \leq \inf_{k\geq n} \mathbb{E}[X_k|\mathcal{G}], \text{ a.s.},$$

and the conditional monotone-convergence theorem implies that

$$\mathbb{E}[Y|\mathcal{G}] = \lim_{n\in\mathbb{N}} \mathbb{E}[Y_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}], \text{ a.s.}$$

12. (conditional dominated-convergence theorem) By the conditional Fatou's lemma, we have

$$\mathbb{E}[Z + X|\mathcal{G}] \leq \liminf_n \mathbb{E}[Z + X_n|\mathcal{G}],$$

as well as

$$\mathbb{E}[Z - X|\mathcal{G}] \leq \liminf_n \mathbb{E}[Z - X_n|\mathcal{G}], \text{ a.s.},$$

and the a.s.-statement follows. □

Problem 10.1.

1. Show that the condition $\mathcal{H} \subseteq \mathcal{G}$ is necessary for the tower property to hold in general. *Hint:* Take $\Omega = \{a, b, c\}$.
2. For $X, Y \in \mathcal{L}^2$ and a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$, show that the following self-adjointness property holds

$$\mathbb{E}[X\,\mathbb{E}[Y|\mathcal{G}]] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,Y] = \mathbb{E}[\mathbb{E}[X|\mathcal{G}]\,\mathbb{E}[Y|\mathcal{G}]].$$

3. Let $\mathcal{H}$ and $\mathcal{G}$ be two sub-σ-algebras of $\mathcal{F}$. Is it true that

$$\mathcal{H} = \mathcal{G} \text{ if and only if } \mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X|\mathcal{H}], \text{ a.s., for all } X \in \mathcal{L}^1?$$

4. Construct two random variables X and Y in $\mathcal{L}^1$ such that $\mathbb{E}[X|\sigma(Y)] = \mathbb{E}[X]$, a.s., but X and Y are not independent.

Regular conditional distributions

Once we have a the notion of conditional expectation defined and analyzed, we can use it to define other, related, conditional quantities. The most important of those is the conditional probability:

Definition 10.6. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The **conditional probability** of $A \in \mathcal{F}$, given $\mathcal{G}$ - denoted by $\mathbb{P}[A|\mathcal{G}]$ - is defined by

$$\mathbb{P}[A|\mathcal{G}] = \mathbb{E}[\mathbf{1}_A|\mathcal{G}].$$

It is clear (from the conditional version of the monotone-convergence theorem) that

$$\mathbb{P}[\cup_{n\in\mathbb{N}} A_n|\mathcal{G}] = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n|\mathcal{G}], \text{ a.s.} \tag{10.5}$$

We can, therefore, think of the conditional probability as a countably-additive map from events to (equivalence classes of) random variables $A \mapsto \mathbb{P}[A|\mathcal{G}]$. In fact, this map has the structure of a vector measure:

Definition 10.7. Let $(B, ||\cdot||)$ be a Banach space, and let $(S, \mathcal{S})$ be a measurable space. A map $\mu : \mathcal{S} \to B$ is called a **vector measure** if

1. $\mu(\emptyset) = 0$, and
2. for each pairwise-disjoint sequence $\{A_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}$, we have

$$\mu(\cup_n A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$$

(where the series in B converges absolutely).

Proposition 10.8. *The conditional probability $A \mapsto \mathbb{P}[A|\mathcal{G}] \in \mathbb{L}^1$ is a vector measure with values in $B = \mathbb{L}^1$.*

Proof. Clearly $\mathbb{P}[0|\mathcal{G}] = 0$, a.s. Let $\{A_n\}_{n\in\mathbb{N}}$ be a pairwise-disjoint sequence in $\mathcal{F}$. Then

$$\Big|\Big|\mathbb{P}[A_n|\mathcal{G}]\Big|\Big|_{\mathbb{L}^1} = \mathbb{E}[|\mathbb{E}[\mathbf{1}_{A_n}|\mathcal{G}]|] = \mathbb{E}[\mathbf{1}_{A_n}] = \mathbb{P}[A_n],$$

and so

$$\sum_{n\in\mathbb{N}} \Big|\Big|\mathbb{P}[A_n|\mathcal{G}]\Big|\Big|_{\mathbb{L}^1} = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n] = \mathbb{P}[\cup_n A_n] \leq 1 < \infty,$$

which implies that $\sum_{n\in\mathbb{N}} \mathbb{P}[A_n|\mathcal{G}]$ converges absolutely in $\mathbb{L}^1$. Finally, for $A = \cup_{n\in\mathbb{N}} A_n$, we have

$$\begin{aligned}\Big|\Big|\mathbb{P}[A|\mathcal{G}] - \sum_{n=1}^{N} \mathbb{P}[A_n|\mathcal{G}]\Big|\Big|_{\mathbb{L}^1} &= \Big|\Big|\mathbb{E}[\sum_{n=N+1}^{\infty} \mathbf{1}_{A_n}|\mathcal{G}]\Big|\Big|_{\mathbb{L}^1} \\ &= \mathbb{P}[\cup_{n=N+1}^{\infty} A_n] \to 0 \text{ as } N \to \infty. \quad \square\end{aligned}$$

It is tempting to try to interpret the map $A \mapsto \mathbb{P}[A|\mathcal{G}](\omega)$ as a probability measure for a fixed $\omega \in \Omega$. It will not work in general; the reason is that $\mathbb{P}[A|\mathcal{G}]$ is defined only a.s., and the exceptional sets pile up when uncountable families of events A are considered. Even if we fixed versions $\mathbb{P}[A|\mathcal{G}] \in \mathcal{L}^0_+$, for each $A \in \mathcal{F}$, the countable additivity relation (10.5) holds only almost surely so there is no guarantee that, for a fixed $\omega \in \Omega$, $\mathbb{P}[\cup_{n\in\mathbb{N}} A_n|\mathcal{G}](\omega) = \sum_{n\in\mathbb{N}} \mathbb{P}[A_n|\mathcal{G}](\omega)$, for all pairwise disjoint sequences $\{A_n\}_{n\in\mathbb{N}}$ in $\mathcal{F}$.

There is a way out of this predicament in certain situations, and we start with a description of an abstract object that corresponds to a well-behaved conditional probability:

Definition 10.9. Let $(R, \mathcal{R})$ and $(S, \mathcal{S})$ be measurable spaces. A map $\nu : R \times \mathcal{S} \to \mathbb{R}$ is called a **(measurable) kernel** if

1. $x \mapsto \nu(x, B)$ is $\mathcal{R}$-measurable for each $B \in \mathcal{S}$, and
2. $B \mapsto \nu(x, B)$ is a measure on $\mathcal{S}$ for each $x \in R$.

Definition 10.10. Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, let $(S, \mathcal{S})$ be a measurable space, and let $e : \Omega \to S$ be a random element in S. A kernel $\mu_{e|\mathcal{G}} : \Omega \times \mathcal{S} \to [0, 1]$ is called the **regular conditional distribution of e, given $\mathcal{G}$**, if

$$\mu_{e|\mathcal{G}}(\omega, B) = \mathbb{P}[e \in B|\mathcal{G}](\omega), \text{ a.s., for all } B \in \mathcal{S}.$$

Remark 10.11.

1. When $(S, \mathcal{S}) = (\Omega, \mathcal{F})$, and $e(\omega) = \omega$, the regular conditional distribution of e (if it exists) is called the **regular conditional probability**. Indeed, in this case, $\mu_{e|\mathcal{G}}(\cdot, B) = \mathbb{P}[e \in B|\mathcal{G}] = \mathbb{P}[B|\mathcal{G}]$, a.s.
2. It can be shown that regular conditional distributions not need to exist in general if S is "too large".

When $(S, \mathcal{S})$ is "small enough", however, regular conditional distributions can be constructed. Here is what we mean by "small enough":

Definition 10.12. A measurable space $(S, \mathcal{S})$ is said to be a **Borel space** (or a **nice space**) if it is isomorphic to a Borel subset of $\mathbb{R}$, i.e., if there one-to-one map $\rho : S \to \mathbb{R}$ such that both ρ and ρ^{-1} are measurable.

Problem 10.2. Show that $\mathbb{R}^n$, $n \in \mathbb{N}$ (together with their Borel σ-algebras) are Borel spaces.

Hint: Show, first, that there is a measurable bijection $\rho : [0, 1] \to [0, 1] \times [0, 1]$ such that ρ^{-1} is also measurable. Use binary (or decimal, or ...) expansions.

Remark 10.13. It can be show that any Borel subset of any complete and separable metric space is a Borel space. In particular, the coin-toss space is a Borel space.

Proposition 10.14. *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $(S, \mathcal{S})$ be a Borel space. Any random element $e : \Omega \to S$ admits a regular conditional distribution.*

Proof. Let us, first, deal with the case $S = \mathbb{R}$, so that $e = X$ is a random variable. Let Q be a countable dense set in $\mathbb{R}$. For $q \in Q$, consider the

random variable P^q, defined as an arbitrary version of

$$P^q = \mathbb{P}[X \leq q|\mathcal{G}].$$

By redefining each P^q on a null set (and aggregating the countably many null sets - one for each $q \in Q$), we may suppose that $P^q(\omega) \leq P^r(\omega)$, for $q \leq r$, $q, r \in Q$, for *all* $\omega \in \Omega$ and that $\lim_{q\to\infty} P^q(\omega) = 1$ and $\lim_{q\to-\infty} P^q(\omega) = 0$, for *all* $\omega \in \Omega$. For $x \in \mathbb{R}$, we set

$$F(\omega, x) = \inf_{q\in Q, q>x} P^q(\omega),$$

so that, for each $\omega \in \Omega$, $F(\omega, \cdot)$ is a right-continuous non-decreasing function from $\mathbb{R}$ to $[0,1]$, which satisfies $\lim_{x\to\infty} F(\omega, x) = 1$ and $\lim_{x\to-\infty} F(\omega, x) = 0$, for all $\omega \in \Omega$. Moreover, as an infimum of countably many random variables, the map $\omega \mapsto F(\omega, x)$ is a random variable for each $x \in \mathbb{R}$.

By (the proof of) Proposition **??**, for each $\omega \in \Omega$, there exists a unique probability measure $\mu_{e|\mathcal{G}}(\omega, \cdot)$ on $\mathbb{R}$ such that $\mu_{e|\mathcal{G}}(\omega, (-\infty, x]) = F(\omega, x)$, for all $x \in \mathbb{R}$. Let $\mathcal{L}$ denote the set of all $B \in \mathcal{B}$ such that

1. $\omega \mapsto \mu_{e|\mathcal{G}}(\omega, B)$ is a random variable, and
2. $\mu_{e|\mathcal{G}}(\cdot, B)$ is a version of $\mathbb{P}[X \in B|\mathcal{G}]$.

It is not hard to check that $\mathcal{L}$ is a λ-system, so we need to prove that 1. and 2. hold for all B in some π-system which generates $\mathcal{B}(\mathbb{R})$. A convenient π-system to use is $\mathcal{P} = \{(-\infty, x] : x \in \mathbb{R}\}$. For $B = (-\infty, x] \in \mathcal{P}$, we have $\mu_{e|\mathcal{G}}(\omega, B) = F(\omega, x)$, so that 1. holds. To check 2., we need to show that $F(x, \omega) = \mathbb{P}[X \leq x|\mathcal{G}]$, a.s. This follows from the fact that

$$F(\cdot, x) = \inf_{q>x} P^q = \lim_{q\searrow x} P^q = \lim_{q\searrow x} \mathbb{P}[X \leq q|\mathcal{G}] = \mathbb{P}[X \leq x|\mathcal{G}], \text{ a.s.},$$

by the conditional dominated convergence theorem.

Turning to the case of a general random element e which takes values in a Borel space $(S, \mathcal{S})$, we pick a one-to-one measurable map $f : S \to \mathbb{R}$ whose inverse ρ^{-1} is also measurable. Then $X = \rho(e)$ is a random variable, and so, by the above, there exists a kernel $\mu_{X|\mathcal{G}} : \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ such that

$$\mu_{X|\mathcal{G}}(\cdot, A) = \mathbb{P}[\rho(e) \in A|\mathcal{G}], \text{ a.s.}$$

We define the kernel $\mu_{e|\mathcal{G}} : \Omega \times \mathcal{S} \to [0,1]$ by

$$\mu_{e|\mathcal{G}}(\omega, B) = \mu_{X|\mathcal{G}}(\omega, \rho(B)).$$

Then, $\mu_{e|\mathcal{G}}(\cdot, B)$ is a random variable for each $B \in \mathcal{S}$ and for a pairwise disjoint sequence $\{B_n\}_{n\in\mathbb{N}}$ in $\mathcal{S}$, we have

$$\begin{aligned}\mu_{e|\mathcal{G}}(\omega, \cup_n B_n) &= \mu_{X|\mathcal{G}}(\omega, \rho(\cup_n B_n)) = \mu_{X|\mathcal{G}}(\omega, \cup_n \rho(B_n)) \\ &= \sum_{n\in\mathbb{N}} \mu_{X|\mathcal{G}}(\omega, \rho(B_n)) = \sum_{n\in\mathbb{N}} \mu_{e|\mathcal{G}}(\omega, B_n),\end{aligned}$$

which shows that $\mu_{e|\mathcal{G}}$ is a kernel; we used the measurability of ρ^{-1} to conclude that $\rho(B_n) \in \mathcal{B}(\mathbb{R})$ and the injectivity of ρ to ensure that $\{\rho(B_n)\}_{n\in\mathbb{N}}$ is pairwise disjoint. Finally, we need to show that $\mu_{e|\mathcal{G}}(\cdot, B)$ is a version of the conditional probability $\mathbb{P}[e \in B|\mathcal{G}]$. By injectivity of ρ, we have

$$\mathbb{P}[e \in B|\mathcal{G}] = \mathbb{P}[\rho(e) \in \rho(B)|\mathcal{G}] = \mu_{X|\mathcal{G}}(\cdot, \rho(B)) = \mu_{e|\mathcal{G}}(\cdot, B), \text{ a.s.} \quad \square$$

Remark 10.15. Note that the conditional distribution, even in its regular version, is not unique in general. Indeed, we can redefine it arbitrarily (as long as it remains a kernel) on a set of the form $N \times \mathcal{S} \subseteq \Omega \times \mathcal{S}$, where $\mathbb{P}[N] = 0$, without changing any of its defining properties. This will, in these notes, never be an issue.

One of the many reasons why regular conditional distributions are useful is that they sometimes allow non-conditional thinking to be transferred to the conditional case:

Proposition 10.16. *Let $\boldsymbol{X}$ be an $\mathbb{R}^n$-valued random vector, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $g : \mathbb{R}^n \to \mathbb{R}$ be a Borel function with the property $g(\boldsymbol{X}) \in \mathbb{L}^1$. Then $\int_{\mathbb{R}^n} g(\boldsymbol{x})\mu_{\boldsymbol{X}|\mathcal{G}}(\cdot, d\boldsymbol{x})$ is a $\mathcal{G}$-measurable random variable and*

$$\mathbb{E}[g(\boldsymbol{X})|\mathcal{G}] = \int_{\mathbb{R}^n} g(\boldsymbol{x})\mu_{\boldsymbol{X}|\mathcal{G}}(\cdot, d\boldsymbol{x}), \text{ a.s.}$$

Proof. When $g = \mathbf{1}_B$, for $B \in \mathbb{R}^n$, the statement follows by the very definition of the regular condition distribution. For the general case, we simply use the standard machine. $\square$

Just like we sometimes express the distribution of a random variable or a vector in terms of its density, cdf or characteristic function, we can talk about the conditional density, conditional cdf or the conditional characteristic function. All of those will correspond to the case covered in Proposition 10.14 and all conditional distributions will be assumed to be regular. For $\boldsymbol{x} = (x_1, \dots, x_n)$ and $\boldsymbol{y} = (y_1, \dots, y_n)$, $\boldsymbol{y} \leq_n \boldsymbol{x}$ means $y_1 \leq x_1, \dots, y_n \leq x_n$.

Definition 10.17. Let $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ be a random vector, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$, and let $\mu_{\boldsymbol{X}|\mathcal{G}} : \Omega \times \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ be the regular conditional distribution of $\boldsymbol{X}$ given $\mathcal{G}$.

1. The **(regular) conditional cdf of $\boldsymbol{X}$, given $\mathcal{G}$** is the map $F : \Omega \times \mathbb{R}^n \to [0, 1]$, given by

$$F(\omega, \boldsymbol{x}) = \mu_{\boldsymbol{X}|\mathcal{G}}(\omega, \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} \leq_n \boldsymbol{x}\}), \text{ for } \boldsymbol{x} \in \mathbb{R}^n,$$

2. A map $f_{\boldsymbol{X}|\mathcal{G}} : \Omega \times \mathbb{R}^n \to [0, \infty)$ is called the **conditional density of $\boldsymbol{X}$ with respect to $\mathcal{G}$** if

(a) $f_{\boldsymbol{X}|\mathcal{G}}(\omega, \cdot)$ is Borel measurable for all $\omega \in \Omega$,

(b) $f_{\boldsymbol{X}|\mathcal{G}}(\cdot, \boldsymbol{x})$ is $\mathcal{G}$-measurable for each $\boldsymbol{x} \in \mathbb{R}^n$, and

(c) $\int_B f_{\boldsymbol{X}|\mathcal{G}}(\omega, \boldsymbol{x})\, d\boldsymbol{x} = \mu_{\boldsymbol{X}|\mathcal{G}}(\omega, B)$, for all $\omega \in \Omega$ and all $B \in \mathcal{B}(\mathbb{R}^n)$,

3. The **conditional characteristic function of $\boldsymbol{X}$, given $\mathcal{G}$** is the map $\varphi_{\boldsymbol{X}|\mathcal{G}} : \Omega \times \mathbb{R}^n \to \mathbb{C}$, given by

$$\varphi_{\boldsymbol{X}|\mathcal{G}}(\omega, \boldsymbol{t}) = \int_{\mathbb{R}^n} e^{i\boldsymbol{t}\cdot\boldsymbol{x}}\, \mu_{\boldsymbol{X}|\mathcal{G}}(\omega, d\boldsymbol{x}), \text{ for } \boldsymbol{t} \in \mathbb{R}^n \text{ and } \omega \in \Omega.$$

To illustrate the utility of the above concepts, here is a versatile result (see Example 10.20 below):

Proposition 10.18. *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$, and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The following two statements are equivalent:*

1. *There exists a (deterministic) function $\varphi : \mathbb{R}^n \to \mathbb{C}$ such that for $\mathbb{P}$-almost all $\omega \in \Omega$,*
$$\varphi_{\boldsymbol{X}|\mathcal{G}}(\omega, \boldsymbol{t}) = \varphi(\boldsymbol{t}), \text{ for all } \boldsymbol{t} \in \mathbb{R}^n.$$
2. *$\sigma(\boldsymbol{X})$ is independent of $\mathcal{G}$.*

Moreover, whenever the two equivalent statements hold, φ is the characteristic function of $\boldsymbol{X}$.

Proof. 1. $\to$ 2.. By Proposition 10.16, we have $\varphi_{\boldsymbol{X}|\mathcal{G}}(\cdot, \boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}|\mathcal{G}]$, a.s. If we replace $\varphi_{\boldsymbol{X}|\mathcal{G}}$ by φ, multiplying both sides by a bounded $\mathcal{G}$-measurable random variable Y and take expectations, we get

$$\varphi(\boldsymbol{t})\mathbb{E}[Y] = \mathbb{E}[Ye^{i\boldsymbol{t}\cdot\boldsymbol{X}}].$$

In particular, for $Y = 1$ we get $\varphi(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}]$, so that

$$\mathbb{E}[Ye^{i\boldsymbol{t}\cdot\boldsymbol{X}}] = \mathbb{E}[Y]\mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}], \tag{10.5}$$

for all $\mathcal{G}$-measurable and bounded Y, and all $\boldsymbol{t} \in \mathbb{R}^n$. For Y of the form $Y = e^{isZ}$, where Z is a $\mathcal{G}$-measurable random variable, relation (10.5) and (a minimal extension of) part 1. of Problem **??**, we conclude that $\boldsymbol{X}$ and Z are independent. Since Z is arbitrary and $\mathcal{G}$-measurable, $\boldsymbol{X}$ and $\mathcal{G}$ are independent.

2. $\to$ 1.. If $\sigma(X)$ is independent of $\mathcal{G}$, so is $e^{i\boldsymbol{t}\cdot\boldsymbol{X}}$, and so, the "irrelevance of independent information" property of conditional expectation implies that

$$\varphi(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}] = \mathbb{E}[e^{i\boldsymbol{t}\cdot\boldsymbol{X}}|\mathcal{G}] = \varphi_{\boldsymbol{X}|\mathcal{G}}(\cdot, \boldsymbol{t}), \text{ a.s.} \qquad \square$$

One of the most important cases used in practice is when a random vector $(X_1, \ldots, X_n)$ admits a density and we condition on the σ-algebra

generated by several of its components. To make the notation more intuitive, we denote the first d components $(X_1, \dots, X_d)$ by $\boldsymbol{X}^o$ (for *observed*) and the remaining $n-d$ components $(X_{d+1}, \dots, X_n)$ by $\boldsymbol{X}^u$ (for *unobserved*).

Proposition 10.19. *Suppose that the random vector*

$$\boldsymbol{X} = (\boldsymbol{X}^o, \boldsymbol{X}^u) = (\underbrace{X_1, \dots, X_d}_{\boldsymbol{X}^o}, \underbrace{X_{d+1}, \dots, X_n}_{\boldsymbol{X}^u})$$

admits a density $f_{\boldsymbol{X}} : \mathbb{R}^n \to [0, \infty)$ and that the σ-algebra $\mathcal{G} = \sigma(\boldsymbol{X}^o)$ is generated by the random vector $\boldsymbol{X}^o = (X_1, \dots, X_d)$, for some $d \in \{1, \dots, n-1\}$. Then, for $\boldsymbol{X}^u = (X_{d+1}, \dots, X_n)$, there exists a conditional density $f_{\boldsymbol{X}^u|\mathcal{G}} : \Omega \times \mathbb{R}^{n-d} \to [0, \infty)$, of $\boldsymbol{X}^u$ given $\mathcal{G}$, and (a version of it) is given by

$$f_{\boldsymbol{X}^u|\mathcal{G}}(\omega, \boldsymbol{x}^u) = \begin{cases} \frac{f_{\boldsymbol{X}}(\boldsymbol{X}^o(\omega), \boldsymbol{x}^u)}{\int_{\mathbb{R}^{n-d}} f_{\boldsymbol{X}}(\boldsymbol{X}^o(\omega), \boldsymbol{y})\, d\boldsymbol{y}}, & \int_{\mathbb{R}^{n-d}} f(\boldsymbol{X}^o, \boldsymbol{y})\, d\boldsymbol{y} > 0, \\ f_0(\boldsymbol{x}^u), & \text{otherwise}, \end{cases}$$

for $x \in \mathbb{R}^{n-d}$ and $\omega \in \Omega$, where $f_0 : \mathbb{R}^{n-d} \to \mathbb{R}$ is an arbitrary density function.

Proof. First, we note that $f_{\boldsymbol{X}^u|\mathcal{G}}$ is constructed from the jointly Borel-measurable function $f_{\boldsymbol{X}}$ and the random vector $\boldsymbol{X}^o$ in an elementary way, and is, thus, jointly measurable in $\mathcal{G} \times \mathcal{B}(\mathbb{R}^{n-d})$. It remains to show that

$$\int_A f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u)\, d\boldsymbol{x}^u \text{ is a version of } \mathbb{P}[\boldsymbol{X}^u \in A|\mathcal{G}], \text{ for all } A \in \mathcal{B}(\mathbb{R}^{n-d}).$$

Equivalently, we need to show that

$$\mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \int_{A^u} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u)\, d\boldsymbol{x}^u] = \mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \mathbf{1}_{\{\boldsymbol{X}^u \in A^u\}}],$$

for all $A^o \in \mathcal{B}(\mathbb{R}^d)$ and $A^u \in \mathcal{B}(\mathbb{R}^{n-d})$.

Fubini's theorem, and the fact that $f_{\boldsymbol{X}^o}(\boldsymbol{x}^o) = \int_{\mathbb{R}^{n-d}} f(\boldsymbol{x}^o, \boldsymbol{y})\, d\boldsymbol{y}$ is the density of $\boldsymbol{X}^o$ yield

$$\begin{aligned} \mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} \int_{A^u} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u)\, d\boldsymbol{x}^u] &= \int_{A^u} \mathbb{E}[\mathbf{1}_{\{\boldsymbol{X}^o \in A^o\}} f_{\boldsymbol{X}^u|\mathcal{G}}(\cdot, \boldsymbol{x}^u)]\, d\boldsymbol{x}^o \\ &= \int_{A^u} \int_{A^o} f_{\boldsymbol{X}^u|\mathcal{G}}(\boldsymbol{x}^o, \boldsymbol{x}^u) f_{\boldsymbol{X}^o}(\boldsymbol{x}^o)\, d\boldsymbol{x}^o\, d\boldsymbol{x}^u \\ &= \int_{A^u} \int_{A^o} f_{\boldsymbol{X}}(\boldsymbol{x}^o, \boldsymbol{x}^u)\, d\boldsymbol{x}^o\, d\boldsymbol{x}^u \\ &= \mathbb{P}[\boldsymbol{X}^o \in A^o, \boldsymbol{X}^u \in A^u]. \end{aligned}$$

□

The above result expresses a conditional density, given $\mathcal{G} = \sigma(\boldsymbol{X}^o)$, as a (deterministic) function of $\boldsymbol{X}^o$. Such a representation is possible even when there is no joint density. The core of the argument is contained in the following problem:

Problem 10.3. Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^d$, and let $\mathcal{G} = \sigma(\boldsymbol{X})$ be the σ-algebra generated by $\boldsymbol{X}$. Then, a random variable Z is $\mathcal{G}$-measurable if and only if there exists a Borel function $f : \mathbb{R}^d \to \mathbb{R}$ with the property that $Z = f(\boldsymbol{X})$.

Let $\boldsymbol{X}^o$ be a random vector in $\mathbb{R}^d$. For $X \in \mathcal{L}^1$ the conditional expectation $\mathbb{E}[X|\sigma(\boldsymbol{X}^o)]$ is $\sigma(\boldsymbol{X}^o)$-measurable, so there exists a Borel function $f : \mathbb{R}^d \to R$ such that $\mathbb{E}[X|\sigma(\boldsymbol{X}^o)] = f(\boldsymbol{X}^o)$, a.s. Note that f is uniquely defined only up to $\mu_{\boldsymbol{X}^o}$-null sets. The value $f(\boldsymbol{x}^o)$ at $\boldsymbol{x}^o \in \mathbb{R}^d$ is usually denoted by $\mathbb{E}[X|\boldsymbol{X}^o = \boldsymbol{x}^o]$.

Example 10.20 (Conditioning normals on their components)**.** Let $\boldsymbol{X} = (\boldsymbol{X}^o, \boldsymbol{X}^u) \in \mathbb{R}^d \times \mathbb{R}^{n-d}$ be a multivariate normal random vector with mean $\boldsymbol{\mu} = (\boldsymbol{\mu}^o, \boldsymbol{\mu}^u)$ and the variance-covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T]$, where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu}$. A block form of the matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{ou} \\ \boldsymbol{\Sigma}_{uo} & \boldsymbol{\Sigma}_{uu} \end{pmatrix},$$

Where

$$\begin{aligned}
\boldsymbol{\Sigma}_{oo} &= \mathbb{E}[\tilde{\boldsymbol{X}}^o(\tilde{\boldsymbol{X}}^o)^T] \in \mathbb{R}^{d\times d} \\
\boldsymbol{\Sigma}_{ou} &= \mathbb{E}[\tilde{\boldsymbol{X}}^o(\tilde{\boldsymbol{X}}^u)^T] \in \mathbb{R}^{d\times(n-d)} \\
\boldsymbol{\Sigma}_{uo} &= \mathbb{E}[\tilde{\boldsymbol{X}}^u(\tilde{\boldsymbol{X}}^o)^T] \in \mathbb{R}^{(n-d)\times d} \\
\boldsymbol{\Sigma}_{uu} &= \mathbb{E}[\tilde{\boldsymbol{X}}^u(\tilde{\boldsymbol{X}}^u)^T] \in \mathbb{R}^{(n-d)\times(n-d)}.
\end{aligned}$$

We assume that $\boldsymbol{\Sigma}_{oo}$ is invertible. Otherwise, we can find a subset of components of $\boldsymbol{X}^o$ whose variance-covariance matrix in invertible and which generate the same σ-algebra (why?). The matrix $A = \boldsymbol{\Sigma}_{uo}\boldsymbol{\Sigma}_{oo}^{-1}$ has the property that $\mathbb{E}[(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)(\tilde{\boldsymbol{X}}^o)^T] = 0$, i.e., that the random vectors $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$ and $\tilde{\boldsymbol{X}}^o$ are uncorrelated. We know, however, that $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}^o, \tilde{\boldsymbol{X}}^u)$ is a Gaussian random vector, so, by Problem **??**, part 3., $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$ is independent of $\tilde{\boldsymbol{X}}^o$. It follows from Proposition 10.18 that the conditional characteristic function of $\tilde{\boldsymbol{X}}^o - A\tilde{\boldsymbol{X}}^o$, given $\mathcal{G} = \sigma(\tilde{\boldsymbol{X}}^o)$ is deterministic and given by

$$\mathbb{E}[e^{i\boldsymbol{t}(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)}|\mathcal{G}] = \varphi_{\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o}(\boldsymbol{t}), \text{ for } \boldsymbol{t} \in \mathbb{R}^{n-d}.$$

Since $A\tilde{\boldsymbol{X}}^o$ is $\mathcal{G}$-measurable, we have

$$\mathbb{E}[e^{i\boldsymbol{t}\boldsymbol{X}^u}|\mathcal{G}] = e^{i\boldsymbol{t}\boldsymbol{\mu}^u}e^{i\boldsymbol{t}A\tilde{\boldsymbol{X}}^o}e^{-\frac{1}{2}\boldsymbol{t}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{t}}, \text{ for } \boldsymbol{t} \in \mathbb{R}^{n-d}.$$

where $\hat{\boldsymbol{\Sigma}} = \mathbb{E}[(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)(\tilde{\boldsymbol{X}}^u - A\tilde{\boldsymbol{X}}^o)^T]$. A simple calculation yields that, conditionally on $\mathcal{G}$, $\boldsymbol{X}^u$ is multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{X}^u|\mathcal{G}}$ and variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}^u|\mathcal{G}}$ given by

$$\boldsymbol{\mu}_{\boldsymbol{X}^u|\mathcal{G}} = \boldsymbol{\mu}^o + A(\boldsymbol{X}^o - \boldsymbol{\mu}^o), \ \boldsymbol{\Sigma}_{\boldsymbol{X}^u|\mathcal{G}} = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{uo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{ou}.$$

Note how the mean gets corrected by a multiple of the difference between the observed value $\boldsymbol{X}^o$ and its (unconditional) expected value. Similarly, the variance-covariance matrix gets corrected by $\boldsymbol{\Sigma}_{uo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{ou}$, but this quantity does not depend on the observation $\boldsymbol{X}^o$.

Problem 10.4. Let (X_1, X_2) be a bivariate normal vector with $\mathrm{Var}[X_1] > 0$. Work out the exact form of the conditional distribution of X_2, given X_1 in terms of $\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = \mathrm{Var}[X_i]$, $i = 1, 2$ and the correlation coefficient $\rho = \mathrm{corr}(X_1, X_2)$.

Additional Problems

Problem 10.5 (Conditional expectation for non-negative random variables)**.** A parallel definition of conditional expectation can be given for random variables in $\mathcal{L}_+^0$. For $X \in \mathcal{L}_+^0$, we say that the random variable Y is **a conditional expectation of** X **with respect to** $\mathcal{G}$ - and denote it by $\mathbb{E}[X|\mathcal{G}]$ - if

(a) Y is $\mathcal{G}$-measurable and $[0, \infty]$-valued, and

(b) $\mathbb{E}[Y\mathbf{1}_A] = \mathbb{E}[X\mathbf{1}_A] \in [0, \infty]$, for $A \in \mathcal{G}$.

Show that

1. $\mathbb{E}[X|\mathcal{G}]$ exists for each $X \in \mathcal{L}_+^0$.
2. $\mathbb{E}[X|\mathcal{G}]$ is unique a.s.
3. $\mathbb{E}[X|\mathcal{G}]$ no longer necessarily exists for all $X \in \mathcal{L}_+^0$ if we insist that $\mathbb{E}[X|\mathcal{G}] < \infty$, a.s., instead of $\mathbb{E}[X|\mathcal{G}] \in [0, \infty]$, a.s.

Hint: The argument in the proof of Proposition 10.3 needs to be modified before it can be used.

Problem 10.6 (How to deal with the independent component)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a bounded Borel-measurable function, and let X and Y be independent random variables. Define the function $g : \mathbb{R} \to \mathbb{R}$ by

$$g(y) = \mathbb{E}[f(X, y)].$$

Show that the function g is Borel-measurable, and that

$$\mathbb{E}[f(X, Y)|Y = y] = g(y),\ \mu_Y - a.s.$$

Problem 10.7 (Some exercises in conditional probability)**.**

1. Let X, Y_1, Y_2 be random variables. Show that the random vectors (X, Y_1) and (X, Y_2) have the same distribution if and only if $\mathbb{P}[Y_1 \in B|\sigma(X)] = \mathbb{P}[Y_2 \in B|\sigma(X)]$, for all $B \in \mathcal{B}(\mathbb{R})$.
2. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of non-negative integrable random variables, and let $\{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be sub-σ-algebras of $\mathcal{F}$. Show that $X_n \xrightarrow{\mathbb{P}} 0$ if $\mathbb{E}[X_n|\mathcal{F}_n] \xrightarrow{\mathbb{P}} 0$. Does the converse hold?

Hint: Prove that for $X_n \in \mathcal{L}_+^0$, we have $X_n \xrightarrow{\mathbb{P}} 0$ if and only if $\mathbb{E}[\min(X_n, 1)] \to 0$.

3. Let $\mathcal{G}$ be a complete sub-σ-algebra of $\mathcal{F}$. Suppose that for $X \in \mathcal{L}^1$, $\mathbb{E}[X|\mathcal{G}]$ and X have the same distribution. Show that X is $\mathcal{G}$-measurable.

Hint: Use the conditional Jensen's inequality.

Problem 10.8 (A characterization of $\mathcal{G}$-measurability)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Show that for a random variable $X \in \mathcal{L}^1$ the following two statements are equivalent:

1. X is $\mathcal{G}$-measurable.
2. For all $\xi \in \mathcal{L}^\infty$, $\mathbb{E}[X\xi] = \mathbb{E}[X\mathbb{E}[\xi|\mathcal{G}]]$.

Problem 10.9 (Conditioning a part with respect to the sum)**.** Let $X_1, X_2, \ldots$ be a sequence of iid r.v.'s with finite first moment, and let $S_n = X_1 + X_2 + \cdots + X_n$. Define $\mathcal{G} = \sigma(S_n)$.

1. Compute $\mathbb{E}[X_1|\mathcal{G}]$.
2. Supposing, additionally, that X_1 is normally distributed, compute $\mathbb{E}[f(X_1)|\mathcal{G}]$, where $f : \mathbb{R} \to \mathbb{R}$ is a Borel function with $f(X_1) \in \mathbb{L}^1$.