Conditional expectation

François Chapon

Contents

1. Introduction 1
1.1. Conditioning by an event 1
1.2. Discrete conditioning 2
1.3. Conditional density 3
2. Conditional expectation 4
2.1. Definition for L^2 random variables 4
2.2. Extension to non-negative random variables 5
2.3. Extension to L^1 random variables 6
3. Properties of conditional expectations 7
3.1. Main properties 7
3.2. Convergence theorems 11
3.3. Inequalities 12
4. Conditional distributions 13

1. Introduction

It is natural to want to estimate a random variable on which we only have partial information. The concept of conditional expectation will formalize this idea. We start by looking at examples of discrete conditioning and conditional density before formally introducing the notion of conditional expectation.

1.1. Conditioning by an event. Recall the well known definition of conditioning given some event of positive probability.

Definition 1.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let B be an event with $\mathbb{P}(B) > 0$. The conditional probability of A given B is defined by*

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

The map

$$\mathbb{P}(\cdot \,|\, B)\colon A \mapsto \mathbb{P}(A \,|\, B)$$

is thus a probability measure on $(\Omega, \mathcal{F})$. The conditional expectation given B of a non-negative or integrable random variable X, is then defined using this conditional probability measure:

$$\mathbb{E}(X \,|\, B) = \int_\Omega X(\omega)\, \mathbb{P}(d\omega \,|\, B) = \frac{\mathbb{E}(X \mathbb{1}_B)}{\mathbb{P}(B)}.$$

This is the expectation of X under the distribution $\mathbb{P}(\cdot \,|\, B)$, corresponding to the average value of X when the event B has occurred.

We want to generalize this notion by conditioning given another random variable and even given a σ-algebra. Let us start with the example of discrete random variables.

1.2. Discrete conditioning. Let X be a discrete random variable with values in E, i.e. E is countable. Define $E' = \{x \in E \,|\, \mathbb{P}(X = x) > 0\}$. Let Y be another discrete random variable. As above, define the conditional expectation of Y given the event $\{X = x\}$ by

$$\mathbb{E}(Y \,|\, X = x) = \frac{\mathbb{E}(Y \mathbb{1}_{\{X=x\}})}{\mathbb{P}(X = x)},$$

for $x \in E'$. We now define the conditional expectation of Y given X as **the random variable** defined by

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

where φ is the function defined by

$$\varphi(x) = \begin{cases} \mathbb{E}(Y \,|\, X = x) & \text{if } x \in E', \\ 0 & \text{if } x \in E \setminus E'. \end{cases}$$

In other words, one has

$$\mathbb{E}(Y \,|\, X) = \sum_{x \in E} \mathbb{1}_{\{X=x\}}\, \mathbb{E}(Y \,|\, X = x).$$

Note that the value 0 of φ on $E \setminus E'$ is arbitrary since $\mathbb{P}(E \setminus E') = 0$. Changing the definition of φ on $E \setminus E'$ would give another "version" of the conditional expectation which would be equal to $\mathbb{E}(Y \,|\, X)$ almost surely.

Example 1.1. We roll a 6 sided dice. Consider $\Omega = \{1, \dots, 6\}$ equipped with the uniform distribution. Let X be the random variable defined by

$$X = \begin{cases} 1, & \text{if the outcome of the dice is even,} \\ 0, & \text{if the outcome of the dice is odd.} \end{cases}$$

Let Y be the random variable given by the outcome of the dice. Then

$$\mathbb{E}(Y \,|\, X)(\omega) = \begin{cases} 4, & \text{if } \omega \in \{2, 4, 6\}, \\ 3, & \text{if } \omega \in \{1, 3, 5\}. \end{cases}$$

Hence, the random variable $\mathbb{E}(Y \,|\, X)$ takes the values 3 or 4, each with probability $\frac{1}{2}$.

1.3. Conditional density. Consider a random couple (X, Y) with joint distribution given by the density $f_{(X,Y)}$, that is for all bounded measurable function f,

$$\mathbb{E}(f(X,Y)) = \int_{\mathbb{R}^2} f(x,y) f_{(X,Y)}(x,y) dxdy.$$

Let h be a bounded Borel function, we want to compute the conditional expectation of $h(Y)$ given X, denoted by $\mathbb{E}(h(Y) \,|\, X)$.

The marginal distribution of X is given by the density

$$f_X(x) = \int_{\mathbb{R}} f_{(X,Y)}(x,y) dy.$$

Let g be a bounded measurable function. We write

$$\begin{aligned}
\mathbb{E}(h(Y)g(X)) &= \int_{\mathbb{R}^2} h(y)g(x) f_{(X,Y)}(x,y) dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(y) f_{(X,Y)}(x,y) dy \right) g(x) dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(y) \frac{f_{(X,Y)}(x,y)}{f_X(x)} dy \right) g(x) f_X(x) \mathbb{1}_{\{f_X > 0\}} dx \\
&= \int_{\mathbb{R}} \varphi(x) g(x) f_X(x) dx,
\end{aligned}$$

where the function φ is defined by

$$\varphi(x) = \int_{\mathbb{R}} h(y) \frac{f_{(X,Y)}(x,y)}{f_X(x)} dy \mathbb{1}_{\{f_X > 0\}}.$$

Thus, we have

$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X)\varphi(X)).$$

Now define the conditional density of Y given $X = x$ by

$$f_{Y|X=x}(y) = \frac{f_{(X,Y)}(x,y)}{f_X(x)} \mathbb{1}_{\{f_X > 0\}}(x).$$

Beware that it is an abuse of notation since $\{X = x\}$ has measure 0. Now, we can defined the conditional expectation of $h(Y)$ given $X = x$ by

$$\mathbb{E}(h(Y) \,|\, X = x) = \varphi(x) = \int_{\mathbb{R}} h(y) f_{Y|X=x}(y) dy.$$

The conditional expectation of $h(Y)$ given X is then defined as the **random variable**

$$\mathbb{E}(h(Y) \,|\, X) = \varphi(X).$$

Example 1.2. Let (X, Y) be a point uniformly drawn on the square $[0,1]^2$, that is (X, Y) has uniform distribution on $[0,1]^2$, so X and Y are independent with uniform distribution on $[0,1]$. Let $S = X + Y$. We want to compute the conditional distribution of X given S. Consider the $\mathcal{C}^1$-diffeomorphism Ψ defined by $\Psi(x,y) = (x, x+y)$. Then the change of variables formula gives that the density of (X, S) is

$$f_{(X,S)}(x,s) = \mathbb{1}_{[0,1]}(x) \mathbb{1}_{[0,1]}(s-x),$$

and the density of S is the triangular distribution given by

$$f_S(s) = \int \mathbb{1}_{[0,1]}(x) \mathbb{1}_{[0,1]}(s-x) dx = s \mathbb{1}_{[0,1]}(s) + (2-s) \mathbb{1}_{[1,2]}(s).$$

The conditional density of X given $S = s$ is thus

$$f_{X|S=s}(x) = \begin{cases} \frac{1}{s} \mathbb{1}_{[0,s]}(x) & \text{if } s \in [0,1], \\ \frac{1}{2-s} \mathbb{1}_{[s-1,1]}(x) & \text{if } s \in [1,2]. \end{cases}$$

Eventually, we find that the conditional distribution of the random variable X given $S = s$ is thus the uniform distribution on $[0, s]$ if $s \in [0, 1]$, and the uniform distribution on $[s - 1, 1]$ if $s \in [1, 2]$.

We will give a general notion of conditional distribution at the end of this chapter, after having introduced the notion of conditional expectation.

2. Conditional expectation

In the following, $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space, and $\mathcal{B}$ denotes a sub-σ-algebra of $\mathcal{A}$. All random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$ will take values in $\mathbb{R}$, but everything can be generalized to $\mathbb{R}^d$. We say that a random variable X is $\mathcal{B}$-measurable, if X is measurable from $(\Omega, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra.

2.1. Definition for L^2 random variables. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. Recall that $L^2(\Omega, \mathcal{A}, \mathbb{P})$ is an Hilbert space for the inner product

$$\langle X, Y \rangle = \mathbb{E}(XY) = \int_\Omega X(\omega) Y(\omega)\, \mathbb{P}(d\omega).$$

Let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{A}$. The subspace $L^2(\Omega, \mathcal{B}, \mathbb{P})$ is also complete, hence closed in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Recall the following theorem from Hilbert space theory.

THEOREM 2.1 (Hilbert projection theorem). *Let $(H, \langle \cdot, \cdot \rangle)$ be an Hilbert space. Let S be a closed subspace of H. For any $y \in H$, there exists a unique $p(y) \in S$ such that*

$$||y - p(y)|| = \inf_{u \in S} ||y - u||.$$

Moreover, $p(y)$ is characterized as the unique element in S such that $y - p(y)$ is orthogonal to S, that is: for all $u \in S$,

$$\langle y - p(y), u \rangle = 0.$$

The map $p\colon H \to S$ is called the orthogonal projection onto S.

Using this theorem in our context of the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$, we make the following definition.

DEFINITION 2.1. *Let $Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. Let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{A}$. The conditional expectation of Y given $\mathcal{B}$, denoted by*

$$\mathbb{E}(Y \,|\, \mathcal{B}),$$

is defined as the orthogonal projection of Y onto the subspace $L^2(\Omega, \mathcal{B}, \mathbb{P})$. It is characterized by

(i) $\mathbb{E}(Y \,|\, \mathcal{B}) \in L^2(\Omega, \mathcal{B}, \mathbb{P})$;
(ii) For all Z in $L^2(\Omega, \mathcal{B}, \mathbb{P})$, we have

$$\mathbb{E}(ZY) = \mathbb{E}(Z\, \mathbb{E}\,(Y \,|\, \mathcal{B}))\,.$$

Condition (ii) will be called the characteristic property of the conditional expectation.

REMARK 2.1. The conditional expectation $\mathbb{E}(Y \,|\, \mathcal{B})$ is thus **a random variable** which is $\mathcal{B}$-measurable and uniquely defined modulo the equivalent relation "almost surely" (as an element of $L^2(\Omega, \mathcal{B}, \mathbb{P})$). The random variable $\mathbb{E}(Y \,|\, \mathcal{B})$ is called a "version" of the conditional expectation, and any other random variable defined by the two above conditions will be equal to $\mathbb{E}(Y \,|\, \mathcal{B})$ almost surely. As an ease of notation, we will not always write the a.s. in properties involving the conditional expectation.

REMARK 2.2. The condition expectation has thus the following interpretation: if Y is square integrable, then $\mathbb{E}(Y \,|\, \mathcal{B})$ is the best approximation of Y (for the L^2 norm) by a $\mathcal{B}$-measurable random variable.

REMARK 2.3. By a classical density argument of simple functions into $L^2(\Omega, \mathcal{A}, \mathbb{P})$, the characteristic property of the conditional expectation may also be written: for all $B \in \mathcal{B}$,

$$\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}\left(\mathbb{1}_B \, \mathbb{E}(Y \,|\, \mathcal{B})\right),$$

that is to say:

$$\int_B Y d\,\mathbb{P} = \int_B \mathbb{E}(Y \,|\, \mathcal{B}) d\,\mathbb{P}\,.$$

REMARK 2.4. When $\mathcal{B} = \sigma(X)$, we will write

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

for the conditional expectation of Y given $\sigma(X)$, i.e. $\mathbb{E}(Y \,|\, \sigma(X))$. It is thus a measurable function of X. Indeed, any random variable Z which is $\sigma(X)$-measurable, is of the form $Z = h(X)$ where h is a measurable function. This is true if $Z = \mathbb{1}_A$, since then $A \in \sigma(X)$, so by definition there exists a measurable set B such that $A = X^{-1}(B)$, giving that

$$Z = \mathbb{1}_A = \mathbb{1}_{X^{-1}(B)} = \mathbb{1}_B(X).$$

For general Z, the assertion follows using the density of simple functions into Borel functions.

PROPOSITION 2.1. *Let $Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$. Then,*

(i) if Y is $\mathcal{B}$-measurable, then

$$\mathbb{E}(Y \,|\, \mathcal{B}) = Y;$$

(ii) $\mathbb{E}\left(\mathbb{E}(Y \,|\, \mathcal{B})\right) = \mathbb{E}(Y)$;
(iii) $\mathbb{E}(\cdot \,|\, \mathcal{B})$ is a linear map;
(iv) if $Y \geq 0$, then $\mathbb{E}(Y \,|\, \mathcal{B}) \geq 0$. In particular, $\mathbb{E}(\cdot \,|\, \mathcal{B})$ is a non-decreasing map.

PROOF. The first three items follow immediately from the fact the conditional expectation is an orthogonal projection.

For the last item, consider the random variable $Z = \mathbb{1}_{\{\mathbb{E}(Y \,|\, \mathcal{B}) < 0\}}$ (which is clearly $\mathcal{B}$-measurable). From the definition, one has

$$0 \leq \mathbb{E}(ZY) = \mathbb{E}\left(Z \, \mathbb{E}(Y \,|\, \mathcal{B})\right) \leq 0,$$

where the last inequality follows from the definition of Z. Hence, $\mathbb{E}\left(\mathbb{1}_{\{\mathbb{E}(Y \,|\, \mathcal{B}) < 0\}} \, \mathbb{E}(Y \,|\, \mathcal{B})\right) = 0$. Since $\mathbb{E}(Y \,|\, \mathcal{B}) \mathbb{1}_{\{\mathbb{E}(Y \,|\, \mathcal{B}) < 0\}} \leq 0$, we get

$$\mathbb{E}(Y \,|\, \mathcal{B}) \mathbb{1}_{\{\mathbb{E}(Y \,|\, \mathcal{B}) < 0\}} = 0 \quad \text{a.s.}$$

which implies that $\mathbb{E}(Y \,|\, \mathcal{B}) \geq 0$ a.s. □

2.2. Extension to non-negative random variables. We start by extending the definition of the conditional expectation to non-negative random variables, using the monotone convergence theorem.

THEOREM-DEFINITION 2.1. *Let Y be a non-negative random variable. There exists an almost surely unique non-negative random variable $\mathbb{E}(Y \,|\, \mathcal{B})$ such that:*

(i) $\mathbb{E}(Y \,|\, \mathcal{B})$ is $\mathcal{B}$-measurable;
(ii) for any $\mathcal{B}$-measurable random variable Z which is non-negative,

$$\mathbb{E}(ZY) = \mathbb{E}\left(Z \, \mathbb{E}(Y \,|\, \mathcal{B})\right).$$

The random variable $\mathbb{E}(Y \,|\, \mathcal{B})$ is called (a version of) the conditional expectation of Y given $\mathcal{B}$.

PROOF. If Y is non-negative and in $L^2(\Omega, \mathcal{A}, \mathbb{P})$, we define $\mathbb{E}(Y \,|\, \mathcal{B})$ using the definition of the conditional expectation in the L^2-case. Let Z be a non-negative $\mathcal{B}$-measurable random variable. We have to verify that the characteristic property of $\mathbb{E}(Y \,|\, \mathcal{B})$ still holds for Z which is only non-negative. We introduce, for every n, the truncated random variable

$$Z_n = Z \wedge n = \inf(Z, n).$$

Hence, the sequence $(Z_n)_n$ is non-negative and in $L^2(\Omega, \mathcal{B}, \mathbb{P})$, and $Z_n \nearrow Z$ a.s. Using the monotone convergence theorem twice, one has:

$$\mathbb{E}(ZY) = \lim_n \mathbb{E}(Z_n Y) = \lim_n \mathbb{E}\left(Z_n \,\mathbb{E}(Y \,|\, \mathcal{B})\right) = \mathbb{E}\left(\lim_n Z_n \,\mathbb{E}(Y \,|\, \mathcal{B})\right) = \mathbb{E}\left(Z \,\mathbb{E}(Y \,|\, \mathcal{B})\right).$$

Now assume that Y is only non-negative. We introduce for every n, $Y_n = Y \wedge n$, which is in $L^2(\Omega, \mathcal{A}, \mathbb{P})$. Thus, define:

$$\mathbb{E}(Y \,|\, \mathcal{B}) = \lim_{n \to \infty} \mathbb{E}(Y_n \,|\, \mathcal{B}),$$

which is well defined a.s. (possibly $+\infty$) since $\left(\mathbb{E}(Y \wedge n \,|\, \mathcal{B})\right)_n$ is non-decreasing. Note that $\mathbb{E}(Y \,|\, \mathcal{B})$ is then $\mathcal{B}$-measurable as a limit of $\mathcal{B}$-measurable random variables. Let Z be a non-negative $\mathcal{B}$-measurable random variable. By the monotone convergence theorem (again twice) and the fact that the characteristic property of the conditional expectation holds for Z non-negative, we have:

$$\mathbb{E}(ZY) = \lim_n \mathbb{E}(ZY_n) = \lim_n \mathbb{E}\left(Z \,\mathbb{E}(Y_n \,|\, \mathcal{B})\right) = \mathbb{E}\left(Z \lim_n \mathbb{E}(Y_n \,|\, \mathcal{B})\right) = \mathbb{E}\left(Z \,\mathbb{E}(Y \,|\, \mathcal{B})\right).$$

This proves the existence. Moreover, if $0 \leq Y \leq Y'$ a.s., then $0 \leq Y \wedge n \leq Y' \wedge n$, hence

$$0 \leq \mathbb{E}(Y \wedge n \,|\, \mathcal{B}) \leq \mathbb{E}(Y' \wedge n \,|\, \mathcal{B}),$$

as a positive operator in $L^2(\Omega, \mathcal{A}, \mathbb{P})$, and taking the limit we get that

$$0 \leq \mathbb{E}(Y \,|\, \mathcal{B}) \leq \mathbb{E}(Y' \,|\, \mathcal{B}).$$

Now we show the unicity a.s. Let U and V be two versions of $\mathbb{E}(Y \,|\, \mathcal{B})$. Introduce the $\mathcal{B}$-measurable set $B = \{U \leq a < b \leq V\}$. By the characteristic property of $\mathbb{E}(Y \,|\, \mathcal{B})$, one has

$$\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}(\mathbb{1}_B U) = \mathbb{E}(\mathbb{1}_B V).$$

Since $\mathbb{E}(\mathbb{1}_B U) \leq a\,\mathbb{P}(B)$ and $\mathbb{E}(\mathbb{1}_B V) \geq b\,\mathbb{P}(B)$, we get

$$a\,\mathbb{P}(B) \geq b\,\mathbb{P}(B),$$

hence $\mathbb{P}(B) = 0$ since $a < b$. By considering the union over positive rationals a and b, we get

$$\mathbb{P}(U < V) = \mathbb{P}\left(\bigcup_{a,b \in \mathbb{Q}_+} \{U \leq a < b \leq V\}\right) = 0.$$

Hence $U \geq V$ a.s., and by swapping the roles of U and V, we eventually get that $U = V$ a.s. □

2.3. Extension to L^1 random variables. We now extend the definition to L^1 random variables.

THEOREM-DEFINITION 2.2. *Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. There exists an almost surely unique random variable $\mathbb{E}(Y \,|\, \mathcal{B})$ such that:*

(i) $\mathbb{E}(Y \,|\, \mathcal{B}) \in L^1(\Omega, \mathcal{B}, \mathbb{P})$;

(ii) for any $\mathcal{B}$-measurable set B,

$$\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}\left(1_B \,\mathbb{E}(Y \,|\, \mathcal{B})\right).$$

More generally, for any bounded random variable Z which is $\mathcal{B}$-measurable,

$$\mathbb{E}(ZY) = \mathbb{E}\left(Z\,\mathbb{E}(Y\,|\,\mathcal{B})\right).$$

The random variable $\mathbb{E}(Y\,|\,\mathcal{B})$ is called (a version of) the conditional expectation of Y given $\mathcal{B}$.

Proof. Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ and suppose that $Y \geq 0$ a.s. Then, $\mathbb{E}(Y\,|\,\mathcal{B}) \geq 0$ a.s. and using the characteristic property of $\mathbb{E}(Y\,|\,\mathcal{B})$ with $Z = 1$, one has

$$\mathbb{E}\left(\mathbb{E}(Y\,|\,\mathcal{B})\right) = \mathbb{E}(Y) < \infty$$

since $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Hence, $\mathbb{E}(Y\,|\,\mathcal{B})$ is integrable and a.s. finite. Now if Y is not non-negative, let Y_+ and Y_- be the positive and negative part of Y respectively (that is, $Y_+ = \max(0, Y)$ and $Y_- = \max(0, -Y)$), and define

$$\mathbb{E}(Y\,|\,\mathcal{B}) = \mathbb{E}(Y_+\,|\,\mathcal{B}) - \mathbb{E}(Y_-\,|\,\mathcal{B}),$$

which is well defined since $\mathbb{E}(Y_+\,|\,\mathcal{B})$ and $\mathbb{E}(Y_-\,|\,\mathcal{B})$ are a.s. finite. Using the triangular inequality and taking the expectation, one also gets that

$$\mathbb{E}\left(|\mathbb{E}(Y\,|\,\mathcal{B})|\right) \leq \mathbb{E}\left(\mathbb{E}(Y_+\,|\,\mathcal{B})\right) + \mathbb{E}\left(\mathbb{E}(Y_-\,|\,\mathcal{B})\right) = \mathbb{E}(Y_+) + \mathbb{E}(Y_-) = \mathbb{E}\left(|Y|\right) < \infty,$$

hence $\mathbb{E}(Y\,|\,\mathcal{B}) \in L^1(\Omega, \mathcal{B}, \mathbb{P})$. Hence, for any $\mathcal{B}$-measurable set B,

$$\begin{aligned}\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}(\mathbb{1}_B Y_+) - \mathbb{E}(\mathbb{1}_B Y_-) &= \mathbb{E}(\mathbb{1}_B\,\mathbb{E}(Y_+\,|\,\mathcal{B})) - \mathbb{E}(\mathbb{1}_B\,\mathbb{E}(Y_-\,|\,\mathcal{B}))\\ &= \mathbb{E}\left(\mathbb{1}_B\,\mathbb{E}(Y_+\,|\,\mathcal{B}) - \mathbb{1}_B\,\mathbb{E}(Y_-\,|\,\mathcal{B})\right)\\ &= \mathbb{E}\left(\mathbb{1}_B\,\mathbb{E}(Y\,|\,\mathcal{B})\right).\end{aligned}$$

This proves the existence. For the unicity, suppose that there exists two $\mathcal{B}$-measurable random variables U and V such that

$$\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}(\mathbb{1}_B U) = \mathbb{E}(\mathbb{1}_B V),$$

for all $B \in \mathcal{B}$. Choosing $B = \{U > V\}$ which is $\mathcal{B}$-measurable, one gets

$$\mathbb{E}((U - V)\mathbb{1}_{\{U>V\}}) = 0,$$

hence $U \leq V$ a.s., and by symmetry $U = V$ a.s. The last claim follows from a classical density argument of simple functions into L^∞. □

3. Properties of conditional expectations

3.1. Main properties.

Proposition 3.1. *Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then, one has:*

(i) the map $\mathbb{E}(\cdot\,|\,\mathcal{B})$ is linear;
(ii) $\mathbb{E}\left(\mathbb{E}(Y\,|\,\mathcal{B})\right) = \mathbb{E}(Y)$;
(iii) $|\mathbb{E}(Y\,|\,\mathcal{B})| \leq \mathbb{E}\left(|Y|\,|\,\mathcal{B}\right)$;
(iv) if Y is $\mathcal{B}$-measurable, $\mathbb{E}(Y\,|\,\mathcal{B}) = Y$ a.s.;
(v) ("pulling out what is known") if X is a bounded $\mathcal{B}$-measurable random variable, one has

$$\mathbb{E}(XY\,|\,\mathcal{B}) = X\,\mathbb{E}(Y\,|\,\mathcal{B}), \quad \text{a.s.}$$

Remark 3.1. By the first three items, the map $\mathbb{E}(\cdot\,|\,\mathcal{B})$ is thus a bounded operator on $L^1(\Omega, \mathcal{A}, \mathbb{P})$ with norm 1. Moreover, it is positive in the sense that $Y \geq 0$ a.s. implies that $\mathbb{E}(Y\,|\,\mathcal{B}) \geq 0$ a.s.

PROOF. (i) Let Y' be another random variable in $L^1(\Omega, \mathcal{A}, \mathbb{P})$ and α some constant. If Z is a bounded $\mathcal{B}$-measurable random variable,

$$\begin{aligned}\mathbb{E}\left(Z\left(\alpha Y + Y'\right)\right) &= \alpha\,\mathbb{E}(ZY) + \mathbb{E}(ZY') \\ &= \alpha\,\mathbb{E}\left(Z\,\mathbb{E}(Y \,|\, \mathcal{B})\right) + \mathbb{E}\left(Z\,\mathbb{E}(Y' \,|\, \mathcal{B})\right) \\ &= \mathbb{E}\left(Z\left(\alpha\,\mathbb{E}(Y \,|\, \mathcal{B}) + \mathbb{E}(Y' \,|\, \mathcal{B})\right)\right),\end{aligned}$$

hence, since $\alpha\,\mathbb{E}(Y \,|\, \mathcal{B}) + \mathbb{E}(Y' \,|\, \mathcal{B})$ is $\mathcal{B}$-measurable, it is a version of the conditional expectation $\mathbb{E}(\alpha Y + Y' \,|\, \mathcal{B})$.

(ii) Take $Z = 1$ is the characteristic property of $\mathbb{E}(Y \,|\, \mathcal{B})$.

(iii) As seen in the proof of the existence of $\mathbb{E}(Y \,|\, \mathcal{B})$ in the L^1 case, one has

$$|\mathbb{E}(Y \,|\, \mathcal{B})| = |\mathbb{E}(Y_+ \,|\, \mathcal{B}) - \mathbb{E}(Y_- \,|\, \mathcal{B})| \leq \mathbb{E}(Y_+ \,|\, \mathcal{B}) + \mathbb{E}(Y_- \,|\, \mathcal{B}) = \mathbb{E}(|Y| \,|\, \mathcal{B}).$$

(iv) If Y is $\mathcal{B}$-measurable, Y is obviously a version of $\mathbb{E}(Y \,|\, \mathcal{B})$ by definition.

(v) Let Z be a bounded $\mathcal{B}$-measurable random variable. Then, ZX is bounded and $\mathcal{B}$-measurable, hence

$$\mathbb{E}(ZXY) = \mathbb{E}\left(ZX\,\mathbb{E}(Y \,|\, \mathcal{B})\right),$$

by the characteristic property of $\mathbb{E}(Y \,|\, \mathcal{B})$. Hence, $X\,\mathbb{E}(Y \,|\, \mathcal{B})$ is a version of $\mathbb{E}(XY \,|\, \mathcal{B})$, since it is obviously $\mathcal{B}$-measurable.

□

The following proposition is called the tower property of conditional expectation.

PROPOSITION 3.2 (Tower property). *Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ or non-negative. Let $\mathcal{B}_1$ and $\mathcal{B}_2$ two sub-σ-algebras of $\mathcal{A}$, such that $\mathcal{B}_1 \subset \mathcal{B}_2$. Then, one has:*

$$\mathbb{E}(Y \,|\, \mathcal{B}_1) = \mathbb{E}\left(\mathbb{E}(Y \,|\, \mathcal{B}_2) \,|\, \mathcal{B}_1\right). \tag{1}$$

REMARK 3.2. When $Y \in L^2(\Omega, \mathcal{A}, \mathbb{P})$, this is a direct consequence of a property of orthogonal projections: projecting onto $L^2(\Omega, \mathcal{B}_2, \mathbb{P})$ then onto $L^2(\Omega, \mathcal{B}_1, \mathbb{P})$ is indeed the same that projecting directly onto $L^2(\Omega, \mathcal{B}_1, \mathbb{P})$, since $L^2(\Omega, \mathcal{B}_1, \mathbb{P})$ is a closed subset of $L^2(\Omega, \mathcal{B}_2, \mathbb{P})$. From a probabilistic point of view, it says that if we know $\mathcal{B}_1$ the surplus of information provided by $\mathcal{B}_2$ is of no use.

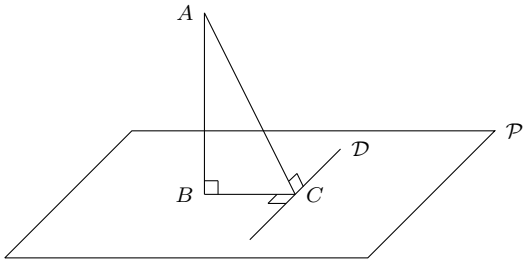

FIGURE 1. Three perpendicular theorem: if AB is perpendicular to a plane $\mathcal{P}$, and a straight line BC is drawn perpendicular to any straight line $\mathcal{D}$ in the plane, then AC is also perpendicular to $\mathcal{D}$.

PROOF. Let Z be a bounded $\mathcal{B}_1$-measurable random variable. Then Z is also $\mathcal{B}_2$-measurable since $\mathcal{B}_1 \subset \mathcal{B}_2$. Hence,

$$\mathbb{E}(ZY) = \mathbb{E}\left(Z\,\mathbb{E}(Y \,|\, \mathcal{B}_2)\right).$$

Using now the definition of the conditional expectation given $\mathcal{B}_1$, one has

$$\mathbb{E}\left(Z\,\mathbb{E}(Y \,|\, \mathcal{B}_2)\right) = \mathbb{E}\left(Z\,\mathbb{E}\left(\mathbb{E}(Y \,|\, \mathcal{B}_2) \,|\, \mathcal{B}_1\right)\right),$$

hence

$$\mathbb{E}(ZY) = \mathbb{E}\left(Z\,\mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{B}_2) \mid \mathcal{B}_1\right)\right),$$

so by characteristic property of the conditional expectation given $\mathcal{B}_1$, one has $\mathbb{E}(Y \mid \mathcal{B}_1) = \mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{B}_2) \mid \mathcal{B}_1\right)$ a.s. □

REMARK 3.3. Trivially, if $\mathcal{B}_1 \subset \mathcal{B}_2$, then $\mathbb{E}(Y \mid \mathcal{B}_1)$ is $\mathcal{B}_2$-measurable, hence

$$\mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{B}_1) \mid \mathcal{B}_2\right) = \mathbb{E}(Y \mid \mathcal{B}_1).$$

The next proposition asserts that independent information is irrelevant.

PROPOSITION 3.3. *Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. If Y is independent of $\mathcal{B}$, then*

$$\mathbb{E}(Y \mid \mathcal{B}) = \mathbb{E}(Y) \quad \textit{a.s.}$$

PROOF. Since $\mathbb{E}(Y)$ is a constant, it is $\mathcal{B}$-measurable. Let $B \in \mathcal{B}$. Then Y and $\mathbb{1}_B$ are independent, hence

$$\mathbb{E}(\mathbb{1}_B Y) = \mathbb{E}(\mathbb{1}_B)\,\mathbb{E}(Y) = \mathbb{E}\left(\mathbb{1}_B\,\mathbb{E}(Y)\right),$$

which proves that $\mathbb{E}(Y)$ is almost surely equal to $\mathbb{E}(Y \mid \mathcal{B})$. □

A more general statement is the following. For $\mathcal{F}$ and $\mathcal{G}$ two σ-algebras, we denote by $\sigma(\mathcal{F}, \mathcal{G})$ the σ-algebra generated by $\mathcal{F} \cup \mathcal{G}$, that is the smallest σ-algebra which contains $\mathcal{F}$ and $\mathcal{G}$[1].

PROPOSITION 3.4. *Let Y be a non-negative or integrable random variable. Let $\mathcal{F}$ and $\mathcal{G}$ two σ-algebras such that $\mathcal{G}$ is independent of $\sigma(\sigma(Y), \mathcal{F})$. Then*

$$\mathbb{E}(Y \mid \sigma(\mathcal{F}, \mathcal{G})) = \mathbb{E}(Y \mid \mathcal{F}).$$

PROOF. Note that $\mathbb{E}(Y \mid \mathcal{F})$ is clearly $\sigma(\mathcal{F}, \mathcal{G})$-measurable. Let A be of the form $A = F \cap G$ with $F \in \mathcal{F}$ and $G \in \mathcal{G}$. Then

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(Y \mid \mathcal{F})\mathbb{1}_{F\cap G}) &= \mathbb{E}\left(\mathbb{E}(Y\mathbb{1}_F \mid \mathcal{F})\mathbb{1}_G\right) \quad (\text{since } F \in \mathcal{F}) \\
&= \mathbb{E}\left(\mathbb{E}(Y\mathbb{1}_F \mid \mathcal{F})\right)\mathbb{E}(\mathbb{1}_G) \quad (\text{since } \mathcal{G} \text{ is independent of } \mathcal{F}) \\
&= \mathbb{E}(Y\mathbb{1}_F)\,\mathbb{E}(\mathbb{1}_G) \\
&= \mathbb{E}(Y\mathbb{1}_{F\cap G}) \quad (\text{since } \mathcal{G} \text{ is independent of } \sigma(\sigma(Y), \mathcal{F})).
\end{aligned}$$

Since $\sigma(\mathcal{F}, \mathcal{G})$ is generated by the π-system[2]

$$\mathcal{C} = \{F \cap G \mid F \in \mathcal{F}, G \in \mathcal{G}\},$$

the proposition follows by Dynkin's π-λ theorem: the class

$$\mathcal{M} = \{A \in \mathcal{A} \mid \mathbb{E}\left(\mathbb{E}(Y \mid \sigma(\mathcal{F}, \mathcal{G}))\mathbb{1}_A\right) = \mathbb{E}\left(\mathbb{E}(Y \mid \mathcal{F})\mathbb{1}_A\right)\}$$

is a λ-system[3] which contains the π-system $\mathcal{C}$, hence contains $\sigma(\mathcal{C})$. □

REMARK 3.4. The condition of Proposition 3.3 is not sufficient for Y to be independent of $\mathcal{B}$. For instance, let Y be a standard Gaussian random variable. Let $X = |Y|$, and let $\mathcal{B} = \sigma(X)$. Any $\sigma(X)$-measurable bounded random variable can be written $h(X)$ for some bounded measurable function h. Hence,

$$\mathbb{E}(h(X)Y) = \mathbb{E}(h(|Y|)Y) = \int_{-\infty}^{+\infty} h(|x|)x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = 0,$$

by symmetry. Hence, $\mathbb{E}(Y \mid X) = 0 = \mathbb{E}(Y)$, but X and Y are clearly not independent.

[1] Also denoted by $\mathcal{F} \vee \mathcal{G}$.

[2] Recall that a π-system is a collection of subsets closed under finite intersection.

[3] Recall that a λ-system is a collection of subsets which contains Ω and is closed under complements of subsets in supersets and under countable increasing unions.

But we have:

PROPOSITION 3.5. *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two sub-σ-algebras. Then, $\mathcal{B}_1$ and $\mathcal{B}_2$ are independent if and only if*

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

for all $\mathcal{B}_2$-measurable bounded random variable Y.

PROOF. The "only if" condition is just the previous proposition. To prove the sufficient condition, we have to prove that for all bounded $\mathcal{B}_1$-measurable Z and all bounded $\mathcal{B}_2$ -measurable Y, we have

$$\mathbb{E}(ZY) = \mathbb{E}(Z)\,\mathbb{E}(Y).$$

But, by definition

$$\mathbb{E}(ZY) = \mathbb{E}(Z\,\mathbb{E}(Y \,|\, \mathcal{B}_1)) = \mathbb{E}(Z)\,\mathbb{E}(Y),$$

by assumption. □

The following proposition is often useful for explicit computations of conditional expectations.

PROPOSITION 3.6. *Let X and Y be two random variables. Suppose that X is independent of $\mathcal{B}$ and Y is $\mathcal{B}$-measurable. Then, for any measurable positive function f, one has*

$$\mathbb{E}(f(X,Y) \,|\, \mathcal{B}) = \varphi(Y),$$

where the function φ is defined by

$$\varphi(y) = \mathbb{E}(f(X,y)) = \int f(x,y)\,\mathbb{P}_X(dx).$$

PROOF. Since Y is $\mathcal{B}$-measurable, $\varphi(Y)$ is obviously $\mathcal{B}$-measurable. We have to prove that for any $\mathcal{B}$-measurable positive Z, one has

$$\mathbb{E}(Zf(X,Y)) = \mathbb{E}(Z\varphi(Y)).$$

Since X is independent of $\mathcal{B}$, it is independent of (Y,Z). The distribution of (X,Y,Z) is thus

$$\mathbb{P}_{(X,Y,Z)} = \mathbb{P}_X \otimes \mathbb{P}_{(Y,Z)}\,.$$

Hence, using Fubini's theorem,

$$\begin{aligned}
\mathbb{E}(Zf(X,Y)) &= \int z f(x,y)\,\mathbb{P}_{(X,Y,Z)}(dx,dy,dz) \\
&= \int z f(x,y)\,\mathbb{P}_X(dx)\,\mathbb{P}_{(Y,Z)}(dy,dz) \\
&= \int z\,\mathbb{E}\left(f(X,y)\right)\mathbb{P}_{(Y,Z)}(dy,dz) \\
&= \int z\varphi(y)\,\mathbb{P}_{(Y,Z)}(dy,dz) \\
&= \mathbb{E}(Z\varphi(Y)),
\end{aligned}$$

giving the result. □

EXAMPLE 3.1. Recall that a random vector $X = (X_1,\ldots,X_d)$ in $\mathbb{R}^d$ is called a Gaussian random vector if every linear combination of its coefficients has a Gaussian distribution.

Let $(X_1,\ldots,X_d,Y)$ be a centered Gaussian random vector in $\mathbb{R}^{d+1}$. The conditional expectation

$$\mathbb{E}(Y \,|\, X_1,\ldots,X_d)$$

coincides with the orthogonal projection of Y onto $\mathrm{Vect}(X_1,\dots,X_d)$. Note that in general $L^2(\Omega,\sigma(X_1,\dots,X_d),\mathbb{P})$ is an infinite dimensional space while $\mathrm{Vect}(X_1,\dots,X_d)$ is finite dimensional, thus apart from the Gaussian case, there is no reason for this projection to coincide with the conditional expectation. So, let

$$\widehat{Y}=\sum_{i=1}^{d}\lambda_i X_i$$

be the orthogonal projection of Y onto $\mathrm{Vect}(X_1,\dots,X_d)$, where the λ_i's are real coefficients. Then, for all $i=1,\dots,d$,

$$\mathrm{Cov}\left(Y-\widehat{Y},X_i\right)=\mathbb{E}\left((Y-\widehat{Y})X_i\right)=0$$

by definition of the orthogonal projection P_Y. Hence, since $(X_1,\dots,X_d,Y-\widehat{Y})$ is a Gaussian vector, one has that $Y-\widehat{Y}$ is independent of $(X_1,\dots,X_d)$ (recall that for Gaussian vectors, pairwise independence implies independence).

Hence,

$$\mathbb{E}(Y\,|\,X_1,\dots,X_d)=\mathbb{E}(Y-\widehat{Y}\,|\,X_1,\dots,X_d)+\widehat{Y}=\mathbb{E}(Y-\widehat{Y})+\widehat{Y}=\widehat{Y},$$

since $\widehat{Y}$ is $\sigma(X_1,\dots,X_d)$-measurable, $Y-\widehat{Y}$ is independent of $\sigma(X_1,\dots,X_d)$ and the variables are centered.

Let h be a bounded Borel function. Write

$$\mathbb{E}\left(h(Y)\,|\,X_1,\dots,X_d\right)=\mathbb{E}\left(h(Y-\widehat{Y}+\widehat{Y})\,|\,X_1,\dots,X_d\right).$$

Since $Y-\widehat{Y}$ is independent of $(X_1,\dots,X_d)$ and $\widehat{Y}$ is $\sigma(X_1,\dots,X_d)$-measurable, we find, using the previous proposition, that

$$\mathbb{E}\left(h(Y)\,|\,X_1,\dots,X_d\right)=\varphi(\widehat{Y}),$$

where

$$\varphi(z)=\mathbb{E}\left(h(Y-\widehat{Y}+z)\right).$$

Since $Y-\widehat{Y}$ has distribution $\mathcal{N}(0,\sigma^2)$, where $\sigma^2=\mathbb{E}((Y-\widehat{Y})^2)$, we have

$$\varphi(z)=\int_{\mathbb{R}}h(y)\frac{1}{\sqrt{2\pi}}e^{-(y-z)^2/2\sigma^2}dy.$$

3.2. Convergence theorems. We start with a monotone convergence theorem for conditional expectations:

THEOREM 3.1 (Conditional monotone theorem). *Let $(Y_n)_n$ be a non-decreasing sequence of non-negative random variables. Then,*

$$\sup_n\mathbb{E}(Y_n\,|\,\mathcal{B})=\mathbb{E}\left(\sup_n Y_n\,|\,\mathcal{B}\right)\quad a.s.$$

PROOF. Since $\mathbb{E}(\cdot\,|\,\mathcal{B})$ is a monotone map, the sequence $\left(\mathbb{E}(Y_n\,|\,\mathcal{B})\right)_n$ is non-decreasing, hence converges a.s. (in $\overline{\mathbb{R}}$). Put $Y=\sup_n Y_n$. Then, for any random variable Z which is $\mathcal{B}$-measurable and non-negative,

$$\mathbb{E}(ZY)=\lim_n\mathbb{E}(ZY_n)=\lim_n\mathbb{E}\left(Z\,\mathbb{E}(Y_n\,|\,\mathcal{B})\right)=\mathbb{E}\left(Z\lim_n\mathbb{E}(Y_n\,|\,\mathcal{B})\right),$$

using twice the monotone convergence theorem. This gives the result as $\lim_n\mathbb{E}(Y_n\,|\,\mathcal{B})$ is $\mathcal{B}$-measurable. □

We now state a version of Fatou's lemma for conditional expectations.

LEMMA 3.1 (Conditional Fatou's lemma). *Let $(Y_n)_n$ be a sequence of non-negative random variables. Then,*

$$\mathbb{E}(\liminf_n Y_n \,|\, \mathcal{B}) \leq \liminf_n \mathbb{E}(Y_n \,|\, \mathcal{B}) \quad a.s.$$

PROOF. The proof follows the same lines as in the classical Fatou's lemma, and is based on the monotone theorem for conditional expectations. Define the sequence $(X_n)_n$ by $X_n = \inf_{k\geq n} Y_k$. Then X_n is non-decreasing and goes towards $\sup_n \inf_{k\geq n} Y_k = \liminf_n Y_n$. By the conditional monotone theorem, one deduces that

$$\sup_n \mathbb{E}(X_n \,|\, \mathcal{B}) = \mathbb{E}(\liminf_n Y_n \,|\, \mathcal{B}).$$

On the other hand, since $X_n \leq Y_k$, for all $k \geq n$, one has, since $\mathbb{E}(\cdot \,|\, \mathcal{B})$ is non-decreasing, that for all $k \geq n$, $\mathbb{E}(X_n \,|\, \mathcal{B}) \leq \mathbb{E}(Y_k \,|\, \mathcal{B})$, hence

$$\mathbb{E}(X_n \,|\, \mathcal{B}) \leq \inf_{k\geq n} \mathbb{E}(Y_k \,|\, \mathcal{B})$$

Taking the supremum over n gives the lemma. □

Now we pass to the dominated convergence theorem for conditional expectations.

THEOREM 3.2 (Conditional dominated convergence theorem). *Let $(X_n)_n$ be a sequence of random variables such that*

(i) $(X_n)_n$ converges a.s. to some random variable X;
(ii) there exists $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$, such that, for all $n \geq 0$, $|X_n| \leq Y$ a.s.

Then, one has

$$\lim_n \mathbb{E}(X_n \,|\, \mathcal{B}) = \mathbb{E}(X \,|\, \mathcal{B}) \quad a.s.$$

PROOF. Its suffices to apply the conditional Fatou lemma to $Z_n = 2Y - |X - X_n|$. □

3.3. Inequalities.

PROPOSITION 3.7 (Conditional Jensen inequality). *Let $X \in L^1(\Omega, \mathcal{A}, \mathbb{P})$ with values in an interval I and let $\varphi\colon I \to \mathbb{R}$ be a convex function such that $\varphi(X) \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then,*

$$\varphi\left(\mathbb{E}(X \,|\, \mathcal{B})\right) \leq \mathbb{E}(\varphi(X) \,|\, \mathcal{B}).$$

PROOF. We need the following fact from convex function theory. Let

$$L(\varphi) = \{g\colon I \to \mathbb{R} \text{ is affine linear and } g \leq \varphi\}.$$

Then $L(\varphi)$ is nonempty and $\varphi = \sup_{L(\varphi)} g$. Indeed, consider the subderivative of φ

$$D^+\varphi(x) = \lim_{y \searrow x} \frac{\varphi(y) - \varphi(x)}{y - x}.$$

By convexity, $D^+\varphi(x)$ is well defined and finite for all $x \in \mathring{I}$ ($D^+\varphi(x)$ is the maximal slope of a tangent at x). Hence, for all $x_0 \in \mathbb{R}$, the map

$$x \mapsto \varphi(x_0) + (x - x_0)D^+\varphi(x_0)$$

is in $L(\varphi)$. So let $g \in L(\varphi)$. By linearity,

$$g(\mathbb{E}(X \,|\, \mathcal{B}) = \mathbb{E}(g(X) \,|\, \mathcal{B}) \leq \mathbb{E}(\varphi(X) \,|\, \mathcal{B}).$$

Taking the supremum over $L(\varphi)$ gives the inequality. □

PROPOSITION 3.8 (Conditional Markov inequality). *Let $Y \in L^1(\Omega, \mathcal{A}, \mathbb{P})$. Then, for all $a > 0$,*

$$\mathbb{P}(|Y| > a \,|\, \mathcal{B}) \leq \frac{\mathbb{E}(|Y| \,|\, \mathcal{B})}{a},$$

where $\mathbb{P}(|Y| > a \,|\, \mathcal{B}) = \mathbb{E}(\mathbb{1}_{\{|Y|>a\}} \,|\, \mathcal{B})$.

PROOF. Exercise. □

4. Conditional distributions

We have seen how to define a conditional distribution in the case of discrete random variables and in the case of random variables having a density. We now give a general definition of the notion of conditional distribution.

DEFINITION 4.1. *Let $(E, \mathcal{E})$ and $(F, \mathcal{F})$ two measurable spaces. A transition kernel, or Markov kernel, from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ is a map $\nu \colon E \times \mathcal{F} \to [0, 1]$ such that:*

(i) for all $x \in E$, the map $\nu(x, \cdot)$ is a probability measure on $(F, \mathcal{F})$;
(ii) for all $A \in \mathcal{F}$, the map $\nu(\cdot, A)$ is $\mathcal{E}$-measurable.

PROPOSITION 4.1. *Let ν be a transition kernel on $E \times F$.*

(i) If h is a positive (or bounded) measurable function on $(F, \mathcal{F})$, then

$$x \mapsto \int_F h(y)\nu(x, dy)$$

is a positive (or bounded) measurable function on $(E, \mathcal{E})$.

(ii) If λ is a probability measure on $(E, \mathcal{E})$, then

$$A \mapsto \int_E \lambda(dx)\nu(x, A)$$

is a probability measure on $(F, \mathcal{F})$.

PROOF. (i) This is clear for simple functions. The statement follows by the monotone convergence theorem.
(ii) It is easy to verify the axioms of a probability measure (exercise).

□

DEFINITION 4.2. *Let X and Y be two random variables. We define (a version of) the conditional distribution of Y given X as any transition kernel ν such that for any bounded Borel function h,*

$$\mathbb{E}\left(h(Y) \,|\, X\right) = \int h(y)\nu(X, dy) \quad a.s.$$

For any Borel set A, we define the conditional probability of $Y \in A$ given X as

$$\mathbb{P}\left(Y \in A \,|\, X\right) = \mathbb{E}(\mathbb{1}_A(Y) \,|\, X) = \nu(X, A) \quad a.s.$$

As seen in the introduction section, we have the following two examples:

EXAMPLE 4.1. If X is a discrete random variable with values in E, such that $\mathbb{P}(X = x) > 0$ for all $x \in E$, then the transition kernel ν is

$$\nu(x, dy) = \mathbb{P}(Y \in dy \,|\, X = x) = \frac{\mathbb{P}(Y \in dy, X = x)}{\mathbb{P}(X = x)}.$$

EXAMPLE 4.2. If (X, Y) has density $f_{(X,Y)}$ the conditional distribution of Y given X is given by $\nu(X, dy)$ where the kernel ν is

$$\nu(x, dy) = f_{Y|X=x}(y)dy.$$

EXAMPLE 4.3. We return to the Gaussian conditioning of example 3.1. Let $(X_1, \ldots, X_d, Y)$ be a centered Gaussian random vector, and $\widehat{Y} = \mathbb{E}(Y \,|\, X_1, \ldots, X_d)$. Recall that there exists $\lambda_1, \ldots, \lambda_d \in \mathbb{R}$ such that $\widehat{Y} = \sum_{i=1}^d \lambda_i X_i$. We have seen that, for all bounded Borel function h,

$$\mathbb{E}(h(Y) \,|\, X_1, \ldots, X_d) = \varphi(\widehat{Y})$$

where

$$\varphi(z) = \int_{\mathbb{R}} h(y) \frac{1}{\sqrt{2\pi}} e^{-(y-z)^2/2\sigma^2} dy,$$

where $\sigma^2 = \mathbb{E}((Y - \widehat{Y})^2)$. In terms of conditional distribution, it says that the conditional distribution of Y given $X_1, \ldots, X_d$ is given by the kernel $\nu(\widehat{Y}, dy)$ where

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

EXAMPLE 4.4. Let $(X_n)_{n\geq 0}$ be a Markov chain on a countable state space E with Markov kernel Q, that is for all $n \geq 0$, and all $x_0, x_1, \ldots, x_{n+1}$ in E, such that $\mathbb{P}(X_0 = x_0, \ldots, X_n = x_n) > 0$,

$$\mathbb{P}(X_{n+1} = x_{n+1} \,|\, X_n = x_n, \ldots, X_1 = x_1, X_0 = x_0) = Q(x_n, x_{n+1}).$$

In other terms, one has that the conditional distribution of X_{n+1} given $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n)$ is given by the Markov kernel $Q(X_n, dy)$, that is:

$$\mathbb{E}\left(f(X_{n+1}) \,|\, \mathcal{F}_n\right) = \mathbb{E}(f(X_{n+1}) \,|\, X_n) = \int f(y) Q(X_n, dy) = Qf(X_n).$$

The distribution of X_0 is called the initial distribution of the chain. If $X_0 = x$ a.s., we denote by $\mathbb{P}_x$ the conditional probability measure $\mathbb{P}(\cdot \,|\, X_0 = x)$. For ν a probability measure on E, the formula $\mathbb{P}_\nu = \sum_{x\in E} \nu(x)\, \mathbb{P}_x$ defines another probability measure on Ω, which corresponds to the distribution of the chain when $X_0 \sim \nu$.

The Markov property can now be stated as follows.

THEOREM 4.1 (Markov property). *Let $(X_n)_{n\geq 0}$ be a Markov chain on E. Then, for all non-negative or bounded measurable function f on $\mathbb{E}^{\mathbb{N}}$, we have for all $x \in E$ and all $n \geq 0$,*

$$\mathbb{E}_x\left[f(X_n, X_{n+1}, \ldots) \,|\, \mathcal{F}_n\right] = \mathbb{E}_{X_n}[f(X_0, X_1, \ldots)]\,.$$

Here, $\mathbb{E}_{X_n}[f(X_0, X_1, \ldots)]$ denotes the function $x \mapsto \mathbb{E}_x[f(X_0, X_1, \ldots)]$ applied to X_n. Thus, the Markov property says that given the past $\mathcal{F}_n$, the conditional expectation of all the future depends only on X_n. Furthermore, given $\{X_n = y\}$, $(X_{n+k})_{k\geq 0}$ is a Markov chain with transition kernel Q starting from y.

PROOF. We have to prove that for all $A \in \mathcal{F}_n$,

$$\mathbb{E}_x\left(f\left((X_{n+k})_{k\geq 0}\right) \mathbb{1}_A\right) = \mathbb{E}_x\left(\mathbb{E}_{X_n}(f\left((X_k)_{k\geq 0}\right)) \mathbb{1}_A\right),$$

and it suffices to take A of the form $A = \{X_0 = x_0, X_1 = x_1 \ldots, X_n = x_n\}$ and the function f of the form $f = \mathbb{1}_{\{y_0\}\times\{y_1\}\times\cdots\times\{y_p\}\times E\times\cdots}$ for all $p \geq 0$. On the one hand, we have,

$$\begin{aligned}
\mathbb{E}_x(f\left((X_k)_{k\geq 0}\right)) &= \mathbb{E}_x(\mathbb{1}_{\{X_0=y_0,\ldots,X_p=y_p\}}) \\
&= \mathbb{P}_x(X_0 = y_0, \ldots, X_p = y_p) \\
&= \mathbb{1}_{\{y_0=x\}} Q(y_0, y_1) \cdots Q(y_{p-1}, y_p),
\end{aligned}$$

by definition of a Markov chain, which gives that

$$\begin{aligned}
\mathbb{E}_x\left(\mathbb{E}_{X_n}(f\left((X_k)_{k\geq 0}\right)) \mathbb{1}_A\right) &= \mathbb{E}_x\left(\mathbb{1}_{\{X_n=y_0\}} \mathbb{1}_{\{X_0=x_0, X_1=x_1\ldots, X_n=x_n\}}\right) Q(y_0, y_1) \cdots Q(y_{p-1}, y_p) \\
&= \mathbb{1}_{\{y_0=x_n\}}\, \mathbb{P}_x\left(X_0 = x_0, X_1 = x_1 \ldots, X_n = x_n\right) Q(y_0, y_1) \cdots Q(y_{p-1}, y_p) \\
&= \mathbb{1}_{\{y_0=x_n\}} \mathbb{1}_{\{x_0=x\}} Q(x_0, x_1) \cdots Q(x_{n-1}, x_n) Q(x_n, y_1) \cdots Q(y_{p-1}, y_p).
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\mathbb{E}_x\left(f\left((X_{n+k})_{k\geq 0}\right)\mathbb{1}_A\right) &= \mathbb{E}_x\left(\mathbb{1}_{\{X_n=y_0,X_{n+1}=y_1,\ldots,X_{n+p}=y_p\}}\mathbb{1}_{\{X_0=x_0,X_1=x_1\ldots,X_n=x_n\}}\right)\\ &= \mathbb{1}_{\{y_0=x\}}\mathbb{P}_x\left(X_0=x_0,X_1=x_1\ldots,X_n=x_n,X_{n+1}=y_1,\ldots,X_{n+p}=y_p\right)\\ &= \mathbb{1}_{\{y_0=x\}}\mathbb{1}_{\{x_0=x\}}Q(x_0,x_1)\cdots Q(x_{n-1},x_n)Q(x_n,y_1)\cdots Q(y_{p-1},y_p),\end{aligned}$$

and the theorem follows. □

As a corollary, the Markov property implies conditional independence of $(X_{n+k})_{k\geq 0}$ and $\mathcal{F}_n$ given X_n:

COROLLARY 4.1. *For all non-negative or bounded measurable function f, and for all random variable Z which is $\mathcal{F}_n$-measurable, we have*

$$\mathbb{E}_x\left(f(X_n,X_{n+1},\ldots)Z\,|\,X_n\right)=\mathbb{E}_x(f(X_n,X_{n+1},\ldots)\,|\,X_n)\,\mathbb{E}_x(Z\,|\,X_n).$$

PROOF. To ease notation, denote $F=f(X_n,X_{n+1},\ldots)$. Then, by the tower property of conditional expectation and the Markov property,

$$\begin{aligned}\mathbb{E}_x(FZ\,|\,X_n) &= \mathbb{E}_x(\mathbb{E}_x(FZ\,|\,\mathcal{F}_n)\,|\,X_n)=\mathbb{E}_x(\mathbb{E}_x(F\,|\,\mathcal{F}_n)Z\,|\,X_n)\\ &= \mathbb{E}_x(\mathbb{E}_{X_n}(f(X_0,X_1,\ldots))Z\,|\,X_n)\\ &= \mathbb{E}_{X_n}(f(X_0,X_1,\ldots))\,\mathbb{E}_x(Z\,|\,X_n).\end{aligned}$$

□

In fact, one sees that the Markov property is equivalent to conditional independence of $(X_{n+k})_{k\geq 0}$ and $\mathcal{F}_n$ given X_n *and* homogeneity:

PROPOSITION 4.2. *The process $(X_n)_{n\geq 0}$ satisfies the Markov property if and only if for all $x\in E$, for all random variable Z which is $\mathcal{F}_n$-measurable,*

$$\mathbb{E}_x(f(X_n,X_{n+1},\ldots)Z\,|\,X_n)=\mathbb{E}_{X_n}(f(X_0,X_1,\ldots))\,\mathbb{E}_x(Z\,|\,X_n),$$

for all bounded measurable function f.