# 7 Conditional distributions and kernels

## 7.1 Conditional expectation and distribution

If not specified, in all this section, we consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two measurable spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$. Similarly, any random variables are defined on $(\Omega, \mathcal{F}, \mathbb{P})$.

We start by a few reminders on the conditional expectations (see [Pol02, Chapter 5]):

**Theorem 7.1.** *Let $p \geq 1$ and $\mathcal{G} \subset \mathcal{F}$ be a sub-$\sigma$-field and $Z$ be a random variable on $\Omega$. Then, there exists a unique $W \in \mathrm{L}^p(\Omega, \mathcal{G}, \mathbb{P})$ such that for any representative, still denoted by $W$, $\mathcal{G}/\mathcal{B}(\mathbb{R})$-measurable for any $\mathsf{A} \in \mathcal{G}$, $\mathbb{E}[\mathbb{1}_{\mathsf{A}} Z] = \mathbb{E}[\mathbb{1}_{\mathsf{A}} W]$.*

*In addition, it satisfies $\mathbb{E}[|W|^p] \leq \mathbb{E}[|Z|^p]$.*

**Remark 7.1.** We can in fact also define the conditional expectations of random variable $Z$ which are supposed to be only non-negative almost surely.

### 7.1.1 A bit of history: conditional distribution in the sense of Kolmogorov

This section is inspired by the very instructive survey [CP97]. Here we consider a random variable $X$ valued in the measurable space $(\mathsf{X}, \mathcal{X})$.

From the conditional expectation, we can define an application:

$$\mathsf{A} \in \mathcal{X} \mapsto K(\mathsf{A}),$$

where $K(\mathsf{A})$ is a version of $\mathbb{E}[\mathbb{1}_{\mathsf{A}}(X)|\mathcal{G}]$. This leads to an application:

$$\begin{aligned} &\Omega \times \mathcal{X} \to [0,1] \\ &(\omega, \mathsf{A}) \mapsto K(\omega, \mathsf{A}). \end{aligned} \tag{7.1}$$

**Definition 7.1.** We call conditional distribution of $X$ given $\mathcal{G}$ in the sense of Kolmogorov any application of the form (7.1) and satisfies for any $\mathsf{A} \in \mathcal{X}$, $\mathbb{P}$-almost surely,

$$K_{\mathsf{A}} = \mathbb{E}[\mathbb{1}_{\mathsf{A}}(X)|\mathcal{G}]\,.$$

From this definition and since we know that conditional distributions in the sense of Kolmogorov exist, we aim to define a "random" distribution in the sense that $\mathbb{P}$-almost surely, $\mathsf{A} \mapsto K(\omega, \mathsf{A})$ is a probability measure and for any $\mathsf{A} \in \mathcal{X}$,

$$\omega \mapsto K(\omega, \mathsf{A})\,,$$

is measurable from $\Omega$ in the space of probability measure on $(\mathsf{X}, \mathcal{X})$ endowed with the weak topology. The rationale behind the introduction of such ideas is to be able to write conditional expectation as expectation with respect to this random measure: i.e., for any measurable function $f : \mathsf{X} \to \mathbb{R}_+$,

$$\mathbb{E}[f(X)|\mathcal{G}](\omega) = \int f(x)K(\omega, \mathrm{d}x)\,. \tag{7.2}$$

However to be able to write (7.2), we have to verify:

(a) $\mathbb{P}$-almost surely, $\mathsf{A} \mapsto K(\omega, \mathsf{A})$ is a probability measure;

(b) $\omega \mapsto K(\omega, \mathsf{A})$ is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable for any $\mathsf{A} \in \mathcal{X}$.

While (b) is easily verified, (a) is in general false. Indeed, we have to verify that

$$\begin{aligned}&\mathbb{P}\text{-almost surely for any sequence of disjoint measurable subsets } (\mathsf{A}_n)_{n\in\mathsf{I}}\,,\ \mathsf{I} \subset \mathbb{N}, \\ &K(\omega, \cup_n \mathsf{A}_n) = \sum_{n\in\mathsf{I}} K(\omega, \mathsf{A}_n)\,.\end{aligned} \tag{7.3}$$

Yet the definition of a conditional distribution in the sense of Kolmogorov only ensures that general false. Indeed, we only can conclude that

$$\begin{aligned}&\text{for any sequence of disjoint measurable subsets } (\mathsf{A}_n)_{n\in\mathsf{I}},\ \mathsf{I} \subset \mathbb{N},\ \mathbb{P}\text{-almost surely}, \\ &K(\omega, \cup_n \mathsf{A}_n) = \sum_{n\in\mathsf{I}} K(\omega, \mathsf{A}_n)\,.\end{aligned} \tag{7.4}$$

Note the subtle difference between (7.3) and (7.4). In fact, if the set of disjoint measurable subsets of $\mathsf{X}$ is countable[(1)], then the equivalence holds but in general we cannot conclude that (7.4) implies (7.3). Therefore, we need to impose some additional conditions on conditional distribution in the sense of Kolmogorov for (7.2) to holds. This leads to the concept of regular conditional distribution.

[(1)]which is the case if for example $\mathsf{X}$ is finite

### 7.1.2 Regular conditional distribution

**Regular conditional distribution with respect to a sub-$\sigma$-field**

**Definition 7.2.** Let $\mathcal{G} \subset \mathcal{F}$ be a sub-$\sigma$-field and $X$ be random variable valued in $(\mathsf{X}, \mathcal{X})$ A conditional distribution $K$ (in the sense of Kolmogorov) of $X$ given $\mathcal{G}$ is said to be regular if

(a) $\omega \mapsto K(\omega, \mathsf{A})$ is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable for any $\mathsf{A} \in \mathcal{X}$.

(b) $\mathsf{A} \mapsto K(\omega, \mathsf{A})$ is a probability measure on $(\mathsf{X}, \mathcal{X})$ for any $\omega$.

**Remark 7.2.** Note that we can be a bit general in the definition just requiring the second condition $\mathbb{P}$-almost surely. But in that case, a modification of the conditional distribution on the set of null probability where it is untrue gives a conditional distribution satisfying (b) for any $\omega$.

For regular conditional distribution, we drop the terminology "in the sense of Kolmogorov". We first show that equipped with a regular conditional distribution, (7.2) holds.

**Theorem 7.2.** *Let $X$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathsf{X}, \mathcal{X})$. Let $\mathcal{G} \subset \mathcal{F}$ be a $\sigma$-algebra and let $K$ be a regular conditional distribution of $X$ given $\mathcal{G}$. Further, let $f : E \to \mathbb{R}$ be measurable and $\mathbb{E}[|f(X)|] < \infty$ or $f$ non-negative. Then*

$$\mathbb{E}[f(X) \mid \mathcal{G}](\omega) = \int f(x) K(\omega, \mathrm{d}x) \text{ for } \mathbb{P}\text{-almost all } \omega\,. \tag{7.5}$$

*Proof.* That $\omega \mapsto \int f(x) K(\omega, \mathrm{d}x)$ is measurable is a consequence of Proposition 7.12 below.

It remains to show that the right-hand side in (7.5) has the properties of the conditional expectation Theorem 7.1.

It is enough to consider the case $f \geq 0$. Let $(g_n)_{n\in\mathbb{N}}$ be a sequence of non-decreasing simple functions that converges pointwise to $f$:

$$g_n := \sum_{i=1}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i}\,, \mathsf{A}_i \in \mathcal{X}\,, \alpha_i \in \mathbb{R}_+\,.$$

Now, for any $n \in \mathbb{N}$ and $\mathsf{B} \in \mathcal{G}$,

$$\begin{aligned}
\mathbb{E}[g_n(X)\mathbb{1}_{\mathsf{B}}] &= \sum_{i=1}^{n} \alpha_i \mathbb{P}[\{X \in \mathsf{A}_i\} \cap \mathsf{B}] = \sum_{i=1}^{n} \alpha_i \int_{\mathsf{B}} \mathbb{P}[\{X \in \mathsf{A}_i\} \mid \mathcal{G}] \mathbb{P}(\mathrm{d}\omega) \\
&= \sum_{i=1}^{n} \alpha_i \int_{\mathsf{B}} K(\omega, \mathsf{A}_i) \mathbb{P}(\mathrm{d}\omega) = \int_{\mathsf{B}} \sum_{i=1}^{n} \alpha_i K(\omega, \mathsf{A}_i) \mathbb{P}(\mathrm{d}\omega) = \int_{\mathsf{B}} \left( \int g_n(x) K(\omega, \mathrm{d}x) \right) \mathbb{P}(\mathrm{d}\omega) \\
&= \mathbb{E}[\phi_n \mathbb{1}_{\mathsf{B}}]\,, \quad \phi_n(\omega) = \int g_n(x) K(\omega, \mathrm{d}x)\,.
\end{aligned} \tag{7.6}$$

By the monotone convergence theorem, for almost all $\omega$ (in fact all), $\phi_n(\omega)$ converges to $\int f(x)K(\omega,\mathrm{d}x)$ since $g_n$ converges to $f$ and is non-decreasing. In addition for any $\omega$, $\phi_n(\omega) \leq \phi_{n+1}(\omega)$. Applying the monotone convergence theorem once more regarding the left-hand side of (7.6) and with respect to the sequence $(\phi_n)_{n\in\mathbb{N}}$, we get

$$\mathbb{E}[f(X)\mathbb{1}_{\mathsf{B}}] = \lim_{n\to+\infty} \mathbb{E}[g_n(X)\mathbb{1}_{\mathsf{B}}] = \lim_{n\to+\infty} \mathbb{E}\left[\varphi_n \mathbb{1}_{\mathsf{B}}\right] = \int_{\mathsf{B}} \int f(x)K(\omega,\mathrm{d}x)\mathbb{P}(\mathrm{d}\omega)\,.$$

□

**A first introduction to kernels**

The definition of a regular conditional distribution motivates the following concept.

**Definition 7.3.** Let $(\mathsf{Y},\mathcal{Y})$ and $(\mathsf{X},\mathcal{X})$ be two measurable spaces. We say that $K : \mathsf{Y}\times\mathcal{X} \to \overline{\mathbb{R}}_+$ is a kernel if

(a) $y \mapsto K(y,\mathsf{A})$ is $\mathcal{Y}/\mathcal{B}(\mathbb{R})$-measurable for any $\mathsf{A} \in \mathcal{X}$.

(b) $\mathsf{A} \mapsto K(y,\mathsf{A})$ is a $\sigma$-finite measure on $(\mathsf{X},\mathcal{X})$ for any $y \in \mathsf{Y}$.

We say that $K$ is finite if $K(y,\mathsf{X}) < \infty$ for any $y \in \mathsf{Y}$. We say that $K$ is stochastic (resp. substochastic) kernel if $K(y,\mathsf{X}) = 1$ (resp. $K(y,\mathsf{X}) \leq 1$) for any $y \in \mathsf{Y}$. If $\mathsf{Y} = \mathsf{X}$, a (sub)stochastic kernel is said to be a (sub)Markov kernel.

**Remark 7.3.** From Definition 7.3, an equivalent formulation of Definition 7.2 is that there exists a stochastic kernel $K$ on $\Omega\times\mathcal{X}$ which is a conditional distribution of $X$ given $\mathcal{G}$, i.e., such that for any $\mathsf{A} \in \mathcal{X}$, $\mathbb{P}$-almost surely:

$$\mathbb{E}[\mathbb{1}_{\mathsf{A}}(X)|\mathcal{G}](\omega) = K(\omega,\mathsf{A})\,.$$

**Example 7.1** (Identity kernel)**.** The map $(y,\mathsf{A}) \in \mathsf{X}\times\mathcal{X} \mapsto \delta_y(\mathsf{A}) = \mathbb{1}_{\mathsf{A}}(y)$ is a Markov kernel on $(\mathsf{X},\mathcal{X})$. It will denoted by Id for simplicity and referred to as the Identity kernel. This notation will be justified in Remark 7.7.

**Example 7.2** (Stochastic matrices and discrete kernels)**.** Stochastic matrices $(y,x) \mapsto \mathrm{M}(y,x)$ on a countable space $\mathsf{D}$, i.e., function on $\mathsf{D}^2 \to [0,1]$ satisfying $\sum_{y\in\mathsf{D}} \mathrm{M}(y,x) = 1$ naturally defines a Markov kernel by

$$\tilde{\mathrm{M}}(y,\mathsf{A}) = \sum_{x\in\mathsf{A}} \mathrm{M}(y,x)\,, \quad \mathsf{A} \subset \mathsf{D}\,.$$

Conversely, a Markov kernel $\tilde{\mathrm{M}}$ on $\mathsf{D}$ is said to be discrete and defined a stochastic matrix:

$$\mathrm{M}(y,x) = \tilde{\mathrm{M}}(y,\{x\})\,.$$

In fact the former example can be encompassed in the following result:

**Proposition 7.3.** *Let* $(x,y) \mapsto \mathrm{k}(x,y)$ *be a non-negative measurable function on* $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$ *endowed with a* $\sigma$*-finite measure* $\lambda_X \otimes \lambda_Y$[2]*. In addition, suppose that there exists a sequence* $(\mathsf{A}_n)_{n\in\mathbb{N}} \in \mathcal{X}^{\mathbb{N}}$*, such that for any* $n \in \mathbb{N}$*,*

$$\int \mathbb{1}_{\mathsf{A}_n}(x)\mathrm{k}(y,x)\mathrm{d}\lambda_X(x) < +\infty\,, \text{ for any } n \in \mathbb{N}\,. \tag{7.7}$$

*Then, for any* $y \in \mathsf{Y}$, $\mathsf{A} \in \mathcal{X}$*,*

$$K(y,\mathsf{A}) = \int \mathbb{1}_{\mathsf{A}}(x)\mathrm{k}(y,x)\mathrm{d}\lambda_X(x)\,,$$

*defines a kernel. on* $\mathsf{Y}\times\mathcal{X}$*.*

*In addition, if for any* $y \in \mathsf{Y}$*,*

$$\int \mathrm{k}(y,x)\mathrm{d}\lambda_X(x) = 1\,,$$

*we say that k is a conditional density associated with the kernel K. Then K is a stochastic kernel and if* $\mathsf{Y}=\mathsf{X}$*, K is a Markov kernel.*

*Proof.* For any $\mathsf{A}$, $y \mapsto K(y,\mathsf{A})$ is well-defined and measurable by Fubini theorems Theorem 6.5. In addition, for any $y\in\mathsf{Y}$, $\mathsf{A}\mapsto K(y,\mathsf{A})$ is a $\sigma$-finite measure by the condition (7.7) and an easy application of the monotone convergence theorem. □

In our next result, we show that the condition (a) can only verified on a $\pi$-system.

**Theorem 7.4.** *Let* $(\mathsf{Y},\mathcal{Y})$ *and* $(\mathsf{X},\mathcal{X})$ *be two measurable spaces. Suppose that* $\mathcal{X}=\sigma(\mathcal{C})$*. The map* $K:\mathsf{Y}\times\mathcal{X}\to\overline{\mathbb{R}}_+$ *is a kernel if and only if (b) in 7.3 holds and for any* $\mathsf{A}\in\mathcal{C}$*,* $y\mapsto K(y,\mathsf{A})$ *is* $\mathcal{Y}$*-measurable.*

*Proof.* The set

$$\tilde{\mathcal{A}} = \{\mathsf{A}\in\mathcal{X} : y\mapsto K(y,\mathsf{A})\}\,,$$

is $\lambda$-system Definition 2.7 (details are left to the reader). Since it contains $\mathcal{C}$, it also contains $\mathcal{X}$ and the proof is completed. □

### Regular conditional distribution with respect to a random variable

We consider now the particular case where $\mathcal{G}=\sigma(Y)$ where $Y$ is random variable valued in $(\mathsf{Y},\mathcal{Y})$. We could apply and consider the same concepts as previously to be able to write for any measurable function $f:\mathsf{X}\to\mathbb{R}_+$,

$$\mathbb{E}[f(X)|Y](\omega) = \int f(x)K(\omega,\mathrm{d}x)\,, \tag{7.8}$$

[2]Note that $\lambda_Y$ is a bit artificial here and could be $\lambda_Y=\delta_{y_0}$ for $y_0\in\mathsf{Y}$

for some stochastic kernel $K$ on $\Omega \times \mathcal{X}$.

However, from a result from measure theory Theorem 7.7, it holds that for any non-negative real random variable $Z$:

$$\mathbb{E}[Z|Y] = \varphi(Y),$$

for some measurable function $\varphi : \mathsf{Y} : \mathbb{R}$. Consequently, in defining a conditional distribution $X$ with respect to a random variable $Y$, we would like to have a map $Q : \mathsf{Y} \times \mathcal{X} \to \bar{\mathbb{R}}_+$ so that we may write for any non-negative Borel function:

$$\mathbb{E}[f(X) \mid Y](\omega) = \int f(x) Q(Y(\omega), \mathrm{d}x) \text{ for } \mathbb{P}\text{-almost all } \omega. \tag{7.9}$$

Note the difference between (7.8) and (7.9) is that the kernel $K$ and $Q$ are defines on $\Omega \times \mathcal{X}$ and $\mathsf{Y} \times \mathcal{X}$ respectively.

This then motivates the following definition.

**Definition 7.4.** Let $X, Y$ be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ respectively. We say that $Q : \mathsf{Y} \times \mathcal{X} \to \bar{\mathbb{R}}_+$ is a regular conditional distribution of $X$ given $Y$ if it is a stochastic kernel and if for any $\mathsf{A} \in \mathcal{X}$, $\mathbb{P}$-almost surely

$$\mathbb{E}[\mathbb{1}_{\mathsf{A}}(X)|Y](\omega) = Q(Y(\omega), \mathsf{A}).$$

We first verify that (7.9) holds.

**Theorem 7.5.** *Let $X, Y$ be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ respectively. Let $Q$ be a regular conditional distribution of $X$ given $Y$. Further, let $f : \mathsf{X} \to \mathbb{R}$ be measurable and $\mathbb{E}[|f(X)|] < \infty$. Then*

$$\mathbb{E}[f(X) \mid Y](\omega) = \int f(x) Q(Y(\omega), \mathrm{d}x) \textit{ for } \mathbb{P}\textit{-almost all } \omega.$$

*Proof.* The proof follows the same line as Theorem 7.2 and is omitted. □

We specify the relation between the regular conditional distribution given $Y$ and the regular conditional distribution given $\sigma(Y)$, in particular their existence. Note that the existence of a regular conditional distribution given $Y$ easily implies the existence of a conditional distribution given $\sigma(Y)$. The following result gives the converse.

**Theorem 7.6.** *Let $X, Y$ be two random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ respectively. Suppose that there exists a regular conditional distribution given $\sigma(Y)$ then there exists a a regular conditional distribution given $Y$.*

The proof relies on the following important result from measure theory.

**Theorem 7.7.** *Let* $(\mathsf{X},\mathcal{X})$ *and* $(\mathsf{Y},\mathcal{Y})$ *be two measurable spaces and* $f:\mathsf{X}\to\mathsf{Y}$*. Then,* $h:\mathsf{X}\to\mathbb{R}$ *is* $\sigma(f)/\mathcal{B}(\mathbb{R})$*-measurable if and only there exists a measurable function* $\varphi:\mathsf{Y}\to\mathbb{R}$*,* $\mathcal{Y}/\mathcal{B}(\mathbb{R})$*-measurable such that*

$$h=\varphi\circ f\,.$$

*Proof.* Denote by $K$ the regular conditional distribution given $\sigma(Y)$ which is supposed to exist. Then, for any $\mathsf{A}\in\mathcal{X}$, there exists $h_{\mathsf{A}}$ by Theorem 7.7 such that for any $\omega\in\Omega$,

$$K(\omega,\mathsf{A})=h_{\mathsf{A}}\circ Y(\omega)\,,$$

since $\omega\mapsto K(\omega,\mathsf{A})$ is by construction $\sigma(Y)$-measurable. Then, defining

$$Q(y,\mathsf{A})=h_{\mathsf{A}}(y)\,,$$

is almost the object we are looking for. In the sense that $\mathsf{A}\mapsto h_{\mathsf{A}}(y)$ is a measure only for $y\in Y(\Omega)$ and therefore for $\mathbb{P}_Y$-almost every $y$. Denote by $\mathsf{B}\in\mathcal{Y}$ the set of $y$ such that $\mathsf{A}\mapsto h_{\mathsf{A}}(y)$ is not a measure. Then, $Y(\Omega)\subset\mathsf{B}^{\complement}$ and $\mathbb{P}(Y\not\in\mathsf{B})=0$. We then define

$$\tilde{Q}(y,\mathsf{A})=\begin{cases}Q(y,\mathsf{A}) & \text{if } y\not\in\mathsf{B}\\ \delta_{x_0}(\mathsf{A}) & \text{otherwise}\,.\end{cases}$$

Then, it is straightforward to see that $\tilde{Q}$ is a regular conditional distribution of $X$ given $Y$. □

**Existence of regular conditional distribution**

**Theorem 7.8** (Regular conditional distributions in $\mathbb{R}$)**.** *Let $X$ be a real random variable on the probability space $(\Omega,\mathcal{F},\mathbb{P})$ and let $\mathcal{G}\subset\mathcal{F}$ be a sub-$\sigma$-field. Then, there exists a regular conditional distribution $K$ of $X$ given $\mathcal{G}$.*

**Example 7.3.** Let $Z_1,Z_2$ be independent Poisson random variables with parameters $\lambda_1,\lambda_2\geq 0$. One can show (exercise!) that (with $Y=Z_1$ and $X=Z_1+Z_2$ )

$$\mathbf{P}[Z_1=k\mid Z_1+Z_2=n]=b_{n,p}(k)\text{ for }k=0,\dots,n,$$

where $p=\frac{\lambda_1}{\lambda_1+\lambda_2}$.

**Definition 7.5.** • Two measurable spaces $(\mathsf{E},\mathcal{E})$ and $(\mathsf{E}',\mathcal{E}')$ are called isomorphic if there exists a bijective map $\varphi:\mathsf{E}\to\mathsf{E}'$ such that $\varphi$ is $\mathcal{E}-\mathcal{E}'$-measurable and the inverse map $\varphi^{-1}$ is $\mathcal{E}'-\mathcal{E}$-measurable. Then we say that $\varphi$ is an isomorphism of measurable spaces.

• If in addition $\mu$ and $\mu'$ are measures on $(\mathsf{E},\mathcal{E})$ and $(\mathsf{E}',\mathcal{E}')$ and if $\mu'=\mu\circ\varphi^{-1}$, then $\varphi$ is an isomorphism of measure spaces, and the measure spaces $(\mathsf{E},\mathcal{E},\mu)$ and $(\mathsf{E}',\mathcal{E}',\mu')$ are called isomorphic.

**Definition 7.6.** A measurable space $(\mathsf{E},\mathcal{E})$ is called a Borel space if there exists a Borel set $\mathsf{B}\in\mathcal{B}(\mathbb{R})$ such that $(\mathsf{E},\mathcal{E})$ and $(\mathsf{B},\mathcal{B}(\mathsf{B}))$ are isomorphic.

Borel spaces are precisely the ones for which we can construct regular conditional distributions.

**Theorem 7.9.** *Let $\mathcal{G}\subset\mathcal{F}$ be a sub- $\sigma$-algebra . Let $X$ be a random variable with values in a Borel space $(\mathsf{E},\mathcal{E})$ (hence, for example, $\mathsf{E}$ Polish, $\mathsf{E}=\mathbb{R}^d, \mathsf{E}=\mathbb{R}^{\mathbb{N}}, \mathsf{E}=C([0,1])$, etc.). Then there exists a regular conditional distribution $K$ of $X$ given $\mathcal{G}$.*

*Proof.* Let $\mathsf{B}\in\mathcal{B}(\mathbb{R})$ and let $\varphi:\mathsf{E}\to\mathsf{B}$ be an isomorphism of measurable spaces. By Theorem 7.8, we obtain the regular conditional distribution $\kappa_{Y',\mathcal{G}}$ of the real random variable $Y'=\varphi\circ Y$. Now define $\kappa_{Y,\mathcal{G}}(\omega,\mathsf{A})=\kappa_{Y',\mathcal{G}}(\omega,\varphi(\mathsf{A}))$ for $\mathsf{A}\in\mathcal{E}$. □

Among Borel spaces are Polish spaces. Recall that a complete separable metric space is called a Polish space. In particular, $\mathbb{R}^d, \mathbb{Z}^d, \mathbb{R}^{\mathbb{N}}, (C([0,1]), \|\cdot\|_\infty)$ and so forth are Polish. Closed subsets of Polish spaces are again Polish. Without proof, we present the following topological result (see, e.g., [RS94]).

**Theorem 7.10** (Kuratowski theorem)**.** *Let $\mathsf{E}$ be a Polish space with Borel $\sigma$-algebra $\mathcal{E}$. Then $(\mathsf{E},\mathcal{E})$ is a Borel space. More precisely, it is isomorphic to either $\mathbb{R}$, $\mathbb{Z}$ or a finite set.*

### 7.1.3 Computation of conditional distribution/kernels in practice

To compute the conditional distribution of $X$ given either a $\sigma$-field $\mathcal{G}$ or another random variable $Y$, the first road with there is no explicit densities involved to get back to the definition Definition 7.2 and Definition 7.4.

However, if $(X,Y)$ has some density with respect to some reference measure, the following result make the connection between conditional densities and conditional distribution clear.

First recall the definition of conditional densities:

**Definition 7.7** (Conditional density)**.** Let $(X,Y)$ be two random elements admitting a density $f$ with respect to $\lambda_X\otimes\lambda_Y$ on $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$. Then the function $(x,y)\mapsto f(y|x)$ defined for any $(x,y)\in\mathsf{X}\times\mathsf{Y}$ by

$$f_{X|Y}(y,x)=\begin{cases}\int f(x,y)\mathrm{d}\lambda_Y(y)\,, \text{ if } \int f(x',y)\mathrm{d}\lambda_X(x')=0 \text{ or } =+\infty\\ f(x,y)/\int f(x',y)\mathrm{d}\lambda_X(x') \text{ otherwise}\,,\end{cases}$$

is called the conditional density of $X$ given $Y$.

Observe that the denominator is the density of $Y$ with respect to $\lambda_Y$ applied to $y$. Hence, it satisfies

$$0<\int f(x',y)\mathrm{d}\lambda_X(x')<\infty \quad \text{for } \mathbb{P}^Y\text{-almost every } y\,.$$

For $x$'s such that this is not true, we simply set $f_{X|Y}(y,x)$ the marginal density of $X$, but we could have set $x \mapsto f_{X|Y}(y,x)$ to be any other arbitrary density.

In the case where $(X,Y)$ satisfies the assumptions of Definition 7.7, the conditional distribution of $Y$ given $X$ is given by the following result, whose proof is left as an exercise.

**Theorem 7.11.** *Let $(X,Y)$ be two random elements admitting a density $f : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}_+$ with respect to $\lambda_X \otimes \lambda_Y$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then a regular conditional distribution of $X$ given $Y$ is given by*

$$Q(y,\mathsf{A}) = \int_{\mathsf{A}} f_{X|Y}(y,x)\lambda_X(\mathrm{d}x) \quad \text{for any } y \in \mathsf{Y} \text{ and } \mathsf{A} \in \mathcal{Y},$$

*where $(x,y) \mapsto f_{X|Y}(y,x)$ is the conditional density of $X$ given $Y$.*

### 7.1.4 Operation with transition/stochastic kernels

We have shown in Theorem 7.5 that given a regular conditional distribution $Q$ of $X$ given $Y$, we can characterize conditional expectations of random variables $f(X)$ given $Y$ as

$$\mathbb{E}[f(X)|Y] = \int f(x)Q(Y,\mathrm{d}x).$$

However in the proof of this result, we have overlooked the existence and construction of the right-hand side, in particular the measurability of $y \mapsto \int f(x)Q(Y,\mathrm{d}x)$. In this section, we will deal with this problem.

**Proposition 7.12.** *Let $K$ be a finite transition kernel from $(\mathsf{Y},\mathcal{Y})$ to $(\mathsf{X},\mathcal{X})$ and let $f : \mathsf{Y} \times \mathsf{X} \to [0,\infty]$ be measurable with respect to $\mathcal{Y} \otimes \mathcal{X}/\mathcal{B}([0,\infty])$. Then the map*

$$\begin{aligned} Kf : \mathsf{Y} &\to [0,\infty], \\ y &\mapsto \int f(y,x)K(y,\mathrm{d}x) \end{aligned}$$

*is well-defined and $\mathcal{Y}$-measurable.*

*Proof.* First note that for a fixed $y$, $x \mapsto f(y,x)$ is measurable with respect to $\mathcal{X}$ by Theorem 6.2. Therefore, $Kf(y) = \int f_y(x)K(y,\mathrm{d}x)$ is well-defined. Hence, it remains to show measurability of this function. To this end, we first consider the case $f = \mathbb{1}_{\mathsf{A}}$ using the $\pi$-$\lambda$ theorem Theorem 2.10. We will then conclude using a sequence of simple functions converging pointwise to $f$.

If $g = \mathbb{1}_{\mathsf{A}_1 \times \mathsf{A}_2}$ for some $\mathsf{A}_1 \in \mathcal{Y}$ and $\mathsf{A}_2 \in \mathcal{X}$, then $Kg(y) = \mathbb{1}_{\mathsf{A}_1}(y)K(y,\mathsf{A}_2)$ is measurable. Now let

$$\mathcal{D} = \{\mathsf{A} \in \mathcal{Y} \otimes \mathcal{X} : K\mathbb{1}_{\mathsf{A}} \text{ is } \mathcal{Y}\text{-measurable}\}.$$

We show that $\mathcal{D}$ is a $\lambda$-system: (i) Evidently, $\mathsf{Y} \times \mathsf{X} \in \mathcal{D}$. (ii) If $\mathsf{A},\mathsf{B} \in \mathcal{D}$ with $\mathsf{A} \subset \mathsf{B}$, then $K\mathbb{1}_{\mathsf{B}\setminus\mathsf{A}} = K\mathbb{1}_{\mathsf{B}} - K\mathbb{1}_{\mathsf{A}}$ is measurable, where we used the fact that $K$ is finite; hence $\mathsf{B}\setminus\mathsf{A} \in \mathcal{D}$.

(iii) If $\mathsf{A}_1, \mathsf{A}_2, \ldots \in \mathcal{D}$ are pairwise disjoint and $\mathsf{A} := \bigcup_{n=1}^{\infty} \mathsf{A}_n$, then $K\mathbb{1}_{\mathsf{A}} = \sum_{n=1}^{\infty} K\mathbb{1}_{\mathsf{A}_n}$ is measurable; hence $\mathsf{A} \in \mathcal{D}$.

Summarizing, $\mathcal{D}$ is a $\lambda$-system that contains the $\pi$-system $\{\mathsf{A}_1 \times \mathsf{A}_2 : \mathsf{A}_1 \in \mathcal{Y}, \mathsf{A}_2 \in \mathcal{X}\}$ that generates $\mathcal{Y} \otimes \mathcal{X}$. Hence, by the $\pi$-$\lambda$ theorem Theorem 2.10, $\mathcal{D} = \mathcal{Y} \otimes \mathcal{X}$. This complete the proof for $f = \mathbb{1}_{\mathsf{A}}$. Taking a sequence of non-decreasing simple functions $\{f_n : n \in \mathbb{N}\}$ converging to an arbitrary $f$ pointwise and using the monotone convergence theorem, we get $Kf = \lim_{n \to +\infty} Kf_n$. Then $Kf$ is measurable as a limit of measurable functions. □

In addition, we aim to explore the use of conditional distribution to characterize now the distribution of $(X, Y)$ and extend this result to more than two random variables. More precisely, the distribution of a pair $(X, Y)$ characterize their regular conditional distribution. Indeed, it is not hard to see that if $Q_1, Q_2$ are two regular conditional distribution of $X$ given $Y$, then for any $\mathsf{A} \in \mathcal{Y}$, $\mathsf{B} \in \mathcal{X}$, it holds $\mathbb{E}[\mathbb{1}_{\mathsf{A}}(Y)(Q_1(Y, \mathsf{B}) - Q_2(Y, \mathsf{B}))] = 0$ which implies that for any $\mathsf{B} \in \mathcal{X}$, for $\mathbb{P}_Y$-almost every $y$, $Q_1(Y, \mathsf{B}) = Q_2(Y, \mathsf{B})$. Now we would like to the other way around, i.e., defining the joint distribution of $(X, Y)$ from a conditional distribution, i.e., the data of a kernel.

Let $X, Y, Z$ three random variables. We start by a result that characterizes the conditional distribution of a pair $(X, Y)$ given $Z$ from the conditional distribution of $Y$ given $Z$ and the one of $X$ given $(Z, Y)$. To this end, we need the definition of the tensor product of kernels.

**Proposition 7.13.** *Let $K_1$ be a finite transition kernel $\mathsf{Z} \times \mathcal{Y}$ and let $K_2$ be a finite transition kernel from $\mathsf{Z} \times \mathsf{Y} \times \mathcal{X}$. Then the map*

$$
\begin{aligned}
K_1 \otimes K_2 : \mathsf{Z} \times (\mathcal{Y} \otimes \mathcal{X}) &\to \mathbb{R}_+ , \\
(z, \mathsf{A}) &\mapsto \int_{\mathsf{Y}} K_1(z, \mathrm{d}y) \int_{\mathsf{X}} K_2((z, y), \mathrm{d}x) \mathbb{1}_{\mathsf{A}}((y, x))
\end{aligned}
$$

*is well-defined and is a ($\sigma$-finite but not necessarily a finite) transition kernel on $\mathsf{Z} \times (\mathcal{Y} \times \mathcal{X})$. If $K_1$ and $K_2$ are stochastic, then $K_1 \otimes K_2$ is stochastic. We call $K_1 \otimes K_2$ the tensor product of $K_1$ and $K_2$.*

**Remark 7.4.** In the following, we often write $\int K(y, \mathrm{d}x) f(y, x)$ instead of $\int f(y, x) K(y, \mathrm{d}x)$ since for multiple integrals this notation allows us to write the integrator closer to the corresponding integral sign.

If $K_2$ is a kernel from $(\mathsf{Y}, \mathcal{Y})$ to $(\mathsf{X}, \mathcal{X})$, then we define the product $K_1 \otimes K_2$ similarly by formally understanding $K_2$ as a kernel from $(z \times \mathsf{Y}, \mathcal{F}_0 \otimes \mathcal{Y})$ to $(\mathsf{X}, \mathcal{X})$ that does not depend on the $z$-coordinate.

*Proof.* Let $\mathsf{A} \in \mathcal{Y} \otimes \mathcal{X}$. By Proposition 7.12, the map

$$
g_{\mathsf{A}} : (z, y) \mapsto \int K_2((z, y), \mathrm{d}x) \mathbb{1}_{\mathsf{A}}(y, x)
$$

is well-defined and $\mathcal{F}_0 \otimes \mathcal{Y}$-measurable. Thus, again by Proposition 7.12, the map

$$z \mapsto K_1 \otimes K_2(z, \mathsf{A}) = \int K_1(z, \mathrm{d}y) g_{\mathsf{A}}(z, y)$$

is well-defined and $\mathcal{F}_0$-measurable. For fixed $z$, by the monotone convergence theorem, the map $\mathsf{A} \mapsto K_1 \otimes K_2(z, \mathsf{A})$ is $\sigma$-additive and thus a measure.

For $z \in z$ and $n \in \mathbb{N}$, let $\mathsf{A}_{z,n} := \{y \in \mathsf{Y} : K_2((z,y), \mathsf{X}) < n\}$. Since $K_2$ is finite, we have $\bigcup_{n\geq 1} \mathsf{A}_{z,n} = \mathsf{Y}$ for any $z \in z$. Furthermore, $K_1 \otimes K_2(z\,,\, \mathsf{A}_n \times \mathsf{X}) \leq n \cdot K_1\big(z, \bar{\mathsf{A}}_n\big) < \infty$. Hence $K_1 \otimes K_2(z, \cdot)$ is $\sigma$-finite and is thus a transition kernel. The supplement is trivial. □

**Definition 7.8.** Let $K$ be a Markov kernel on $\mathsf{X} \times \mathcal{X}$, we define by induction on $n \geq 1$, the Markov Kernel on $\mathsf{X} \otimes \mathcal{X}^{\otimes n}$ by $P^{\otimes n+1} = P^{\otimes n} \otimes P$, with $P^{\otimes 1} = P$.

**Definition 7.9.** Let $n \in \mathbb{N}$ and $\{(\mathsf{X}_i, \mathcal{X}_i)\}_{i=0}^n$, be measurable spaces.

(1) For $i = 1, \ldots, n$, let $K_i$ be a stochastic kernel from $\big(\prod_{k=0}^{i-1} \mathsf{X}_k, \otimes_{k=0}^{i-1} \mathcal{X}_k\big)$ to $(\mathsf{X}_i, \mathcal{X}_i)$. Then the recursion $K_1 \otimes \ldots \otimes K_i := (K_1 \otimes \ldots \otimes K_{i-1}) \otimes K_i$ for any $i = 1, \ldots, n$ defines a stochastic kernel $\otimes_{k=1}^i K_k := K_1 \otimes \ldots \otimes K_i$ from $(\mathsf{X}_0, \mathcal{X}_0)$ to $\big(\prod_{k=1}^i \mathsf{X}_k, \bigotimes_{k=1}^i \mathcal{X}_k\big)$.

(2) If for $i = 1, \ldots, n$, $K_i$ is a stochastic kernel from from $\big(\prod_{j \in \mathsf{J}_i} \mathsf{X}_j, \otimes_{j \in \mathsf{J}_i} \mathcal{X}_j\big)$ to $(\mathsf{X}_i, \mathcal{X}_i)$ for $\mathsf{J} \subset \{0, \ldots, i-1\}$ (in particular we consider in general $\mathsf{J} = \{i-1\}$, then we define

$$\tilde{K}_i((x_0, \ldots, x_{i-1}), \mathsf{A}) = \kappa((x_j)_{j \in \mathsf{J}_i}, \mathsf{A})\,,$$

and set $\otimes_{k=1}^i K_k = \otimes_{k=1}^i \tilde{K}_k$.

The following consequence and probability interpretation of Proposition 7.13 is the following.

**Proposition 7.14.** *Let $X, Y, Z$ be three random variables valued in $(\mathsf{X}, \mathcal{X})$, $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{Z}, \mathcal{Z})$ respectively. Let $K_1$ and $K_2$ be regular distribution of $Y$ given $Z$ and $X$ given $(Y, Z)$ respectively. Then a regular conditional distribution of $(Y, X)$ given $Z$ is $K_1 \otimes K_2$.*

*Proof.* It is left as an exercise. □

**Corollary 7.15.** *Let $(\mathsf{Y}, \mathcal{Y}), (\mathsf{X}, \mathcal{X})$ be measurable spaces and $\mu$ a finite measure on $(\mathsf{Y}, \mathcal{Y})$. Let $K$ be a finite transition kernel from $\mathsf{Y}$ to $\mathsf{X}$. Then there exists a unique $\sigma$-finite measure $\mu \otimes K$ on $(\mathsf{Y} \times \mathsf{X}, \mathcal{Y} \otimes \mathcal{X})$ with*

$$\mu \otimes K(\mathsf{A}_1 \times \mathsf{A}_2) = \int_{\mathsf{Y}} \mathbb{1}_{\mathsf{A}_1}(y) K(y, \mathsf{A}_2) \mu(\mathrm{d}y) \quad \text{for any } \mathsf{A}_1 \in \mathcal{Y}, \mathsf{A}_2 \in \mathcal{X}\,.$$

*If $K$ is stochastic and if $\mu$ is a probability measure, then $\mu \otimes K$ is a probability measure.*

*Proof.* It is a simple consequence of Proposition 7.13 with $K_2 = K$ and $K_1(z, \cdot) = \mu$ for any $z$. □

**Definition 7.10.** Let $(\mathsf{Y},\mathcal{Y},\mu)$ be a finite measure space, let $(\mathsf{X},\mathcal{X})$ be a measurable space and let $K$ be a finite transition kernel from $\mathsf{Y}$ to $\mathsf{X}$. We define the $\sigma$-finite measure $\mu K$ for any $\mathsf{A}\in\mathcal{X}$ as $\mu K(\mathsf{A})=\mu\otimes K(\mathsf{Y}\times\mathsf{A})$. $\mu K$ is referred to as the composition of $\mu$ and $K$.

**Remark 7.5.** It is easy to verify that if $\mu$ is a probability measure and $K$ is a stochastic kernel, $\mu K$ is a probability measure.

**Proposition 7.16.** *Let $(X,Y)$ be a pair of random variables valued in $(\mathsf{X}\times\mathsf{Y},\mathcal{X}\otimes\mathcal{Y})$. Let $K$ be a regular conditional distribution of $X$ given $Y$ and let $\mu$ be the distribution of $Y$. Then $(X,Y)$ has distribution $\mu\otimes K$ and $X$ has distribution $\mu K$.*

**Remark 7.6.** It may be proven as an immediate consequence of Proposition 7.14 taking for $Z=c$ for a fixed constant $c$.

*Proof.* The proof is left as an exercise. □

**Theorem 7.17.** *Let $Y,X_1,\ldots,X_n$ be random variables such that $Y$ is valued in $(\mathsf{Y},\mathcal{Y})$ and $X_i$ is valued in $(\mathsf{X},\mathcal{X})$ for any $i\in\{1,\ldots,n\}$. Suppose that $X_1,\ldots,X_n$ given $Y$ are i.i.d., i.e., for any measurable and bounded functions $f_1,\ldots,f_n$,*

$$\mathbb{E}[\textstyle\prod_{i=1}^n f_i(X_i)|Y]=\prod_{i=1}^n\mathbb{E}[f_i(X_i)|Y]\,.$$

*Suppose that $K$ is a regular conditional distribution of $X_1$ given $Y$ and denote by $\mu$ the distribution of $Y$.*

*Then, for any $i\in\{1,\ldots,n\}$, $K$ is a regular conditional distribution of $X_i$ given $Y$ and the distribution of $(Y,X_1,\ldots,X_n)$ is*

$$\mathbb{P}\circ(Y,X_1,\ldots,X_n)^{-1}(\mathsf{A})=\int\mathbb{1}_{\mathsf{A}}(y,x_1,\ldots,x_n)\mu(\mathrm{d}y)\otimes_{k=1}^n[K(y,\cdot)](\mathrm{d}x_1\cdots\mathrm{d}x_n)\,,$$

*where $\otimes_{k=1}^n[K(\cdot,\cdot)]$ is defined according to Definition 7.9 with $\mathsf{J}_i=\{0\}$ for any $i$.*

**Example 7.4.** Let $Y$ be a random variable that is uniformly distributed on $[0,1]$. Let $X_1,\ldots,X_n$ i.i.d. given $Y$ with conditional distribution given $Y$, **Ber**$(Y)$, i.e.,

$$K(y,\mathsf{A})=y\delta_1(\mathsf{A})+(1-y)\delta_0(\mathsf{A})\,,\quad \mathsf{A}\in\{0,1\}\,.$$

Then, the distribution of $(Y,X_1,\ldots,X_n)$ is **Unif**$([0,1])\otimes K^{\otimes n}$.

**Definition 7.11.** Let $(\mathsf{X},\mathcal{X})$, $(\mathsf{Y},\mathcal{Y})$ and $(\mathsf{Z},\mathcal{Z})$ be a measurable spaces and let $K_1,K_2$ be a (sub)-stochastic kernel on $\mathsf{Z}\times\mathcal{Y}$ and $\mathsf{Y}\times\mathcal{X}$ respectively. The function

$$K_1K_2:z\times\mathcal{X}\to\mathbb{R}_+\,,$$
$$(z,\mathsf{A}_2)\mapsto K_1\otimes K_2(z,\mathsf{Y}\times\mathsf{A})=\int_{\mathsf{Y}}K_1(z,\mathrm{d}y)K_2(y,\mathsf{A}_2)\,.$$

is well defined and defined a (sub)-stochastic kernel on $\mathsf{Z} \times \mathcal{X}$ referred to as the composition of $K_1$ and $K_2$.

**Remark 7.7.** Note that for any Markov kernel $P$ on $\mathcal{X} \times \mathsf{X}$, $\mathrm{Id}P = P\mathrm{Id} = P$ where Id is the identity kernel defined in Example 7.1: $(y, \mathsf{A}) \in \mathsf{X} \times \mathcal{X} \mapsto \delta_y(\mathsf{A}) = \mathbb{1}_{\mathsf{A}}(y)$.

**Proposition 7.18.** *Let $X, Y, Z$ be three random variables valued in $(\mathsf{X}, \mathcal{X})$, $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{Z}, \mathcal{Z})$ respectively. Suppose $K_1$ and $K_2$ are stochastic kernel such that $K_1 \otimes K_2$ is the conditional distribution of $(Y, X)$ given $Z$. Then, $K_1K_2$ is a regular conditional distribution of $X$ given $Z$.*

**Proposition 7.19.** *Let $K_1, K_2, K_3$ be three stochastic kernels on $\mathsf{Y} \times \mathcal{X}$, $\mathsf{X} \times \mathcal{W}$ and $\mathsf{W} \times \mathcal{Z}$ respectively. Then $(K_1K_2)K_3 = K_1(K_2K_3)$.*

*Proof.* This is a simple application of Fubini theorems Theorem 6.5. □

**Theorem 7.20** (another Fubini theorem)**.** *Let $(\mathsf{X}, \mathcal{X})$, $(\mathsf{Y}, \mathcal{Y})$ be a measurable spaces. Let $\mu$ be a finite measure on $(\mathsf{Y}, \mathcal{Y})$ and let $K$ be a finite transition kernel from $\mathsf{Y}$ to $\mathsf{X}$. Assume that $f : \mathsf{Y} \times \mathsf{X} \to \overline{\mathbb{R}}$ is measurable with respect to $\mathcal{Y} \otimes \mathcal{X}$. If $f \geq 0$ or $f \in \mathrm{L}^1(\mu \otimes K)$, then*

$$\int_{\mathsf{Y} \times \mathsf{X}} f \, \mathrm{d}(\mu \otimes K) = \int_{\mathsf{Y}} \left\{ \int_{\mathsf{X}} f(y,x) K(y, \mathrm{d}x) \right\} \mu(\mathrm{d}y) \, .$$

**Remark 7.8.** Note that by Proposition 7.12, $Kf : y \mapsto \int_{\mathsf{X}} f(y,x)K(y,\mathrm{d}x)$ is measurable.

*Proof.* For $f = \mathbb{1}_{\mathsf{A}_1 \times \mathsf{A}_2}$ with $\mathsf{A}_1 \in \mathcal{Y}$ and $\mathsf{A}_2 \in \mathcal{X}$, the statement is true by definition; see Corollary 7.15.

Then for $f = \mathbb{1}_{\mathsf{A}}$ with $\mathsf{A} \in \mathcal{Y} \otimes \mathcal{Y}$, the result follows from an easy application of the $\pi$-$\lambda$ theorem Theorem 2.10.

The proof is then completes taking decomposing $f = f^+ - f^-$ and two sequences of simple functions which monotonically converges to $f^+, f^-$ pointwise and applying the convergence monotone theorem. □

**Corollary 7.21.** *Let $(\mathsf{X}, \mathcal{X})$, $(\mathsf{Y}, \mathcal{Y})$ be a measurable spaces. Let $\mu$ be a finite measure on $(\mathsf{Y}, \mathcal{Y})$ and let $K$ be a finite transition kernel from $\mathsf{Y}$ to $\mathsf{X}$. Let $f : \mathsf{X} \to \mathbb{R}$, $f \geq 0$ or $f \in \mathrm{L}^1(\mu \otimes K)$. Then $(\mu K)(f) = \mu(Kf)$ denoting for any measure $\nu$ and $\nu$-integrable function $f$, $\nu(f) = \int f \mathrm{d}\nu$.*

### 7.1.5 Applications to Bayesian statistics

Recall that Bayesian statistics suppose a probabilistic model on some observed data $z_{\mathrm{obs}}$ which is assumed to be a sample from a random variable $z_{\mathrm{obs}}$ valued in a measurable space $(\mathsf{Z}, \mathcal{Z})$. The observation may gather i.i.d. observations, in such case $z_{\mathrm{obs}} = (z_1, \ldots, z_n)$ where the $z_i$ are i.i.d. samples.

A generalization of **??** is the following using the concept of kernel.

**Definition 7.12.** A Bayesian statistical model is the data of $((\mathsf{Z},\mathcal{Z}),(\mathsf{T},\mathcal{T}),\mathcal{P}_\mathsf{T},\nu_\mathsf{T})$ where

(1) $(\mathsf{Z},\mathcal{Z})$ and $(\mathsf{T},\mathcal{T})$ are measurable spaces;

(2) $\mathcal{P}_\mathsf{T}$ is a set of probability measure defined by:

$$\mathcal{P}_\mathsf{T} = \{K(\vartheta,\cdot) \mid \vartheta \in \mathsf{T}\} ,$$

where $K$ is a stochastic kernel on $\mathsf{T}\times\mathcal{Z}$;

(3) a distribution on $\nu_\mathsf{T}$ on $(\mathsf{T},\mathcal{T})$ called the prior distribution.

In most applications, $\mathsf{T}$ is either discrete or a subset of $\mathbb{R}^d$ for $d \geq 1$. Assume that the model is dominated by a measure $\mu$ on $(\mathsf{Z},\mathcal{Z})$, i.e., $K$ admits a transition density with respect to $\mu$: there exists a measurable function $\mathrm{L}:\mathsf{T}\times\mathsf{Z}\to\mathbb{R}_+$ such that for any $\vartheta\in\mathsf{T}$ and $z_{\mathrm{obs}}\in\mathsf{Z}$,

$$\frac{\mathrm{d}K(\vartheta,\cdot)}{\mathrm{d}\mu}(z_{\mathrm{obs}}) = \mathrm{L}(\vartheta,z_{\mathrm{obs}}) . \tag{7.10}$$

The function $\mathrm{L}:(\vartheta,z_{\mathrm{obs}})\mapsto\mathrm{L}(\vartheta,z_{\mathrm{obs}})$ is called the likelihood function of the model.

**Example 7.5.** A basic example is the Gaussian model where $Z$ is a vector of $\mathbb{R}^n$, for $n\geq 1$, whose components are independent and identically distributed (i.i.d.) according to the one dimensional Gaussian law with mean $\vartheta\in\mathbb{R}$ and some fixed variance $\varsigma^2>0$. Then the model is dominated by the Lebesgue measure on $\mathbb{R}^n$ and admits the density defined by for any $z_{\mathrm{obs}}=(z_1,\ldots,z_n)\in\mathbb{R}^n$ and $\vartheta\in\mathbb{R}$

$$\mathrm{L}(\theta,z_{\mathrm{obs}}) = \prod_{i=1}^{n}\exp\left(-\|z_i-\vartheta\|^2/(2\varsigma^2)\right) .$$

Then while in frequentist statistics, the parameter $\vartheta$ would be inferred by maximizing the likelihood function, Bayesian statistics consider that the parameter $\vartheta$ is itself a sample from a random variable $\theta$, whose the distribution is $\nu_\mathsf{T}$.

**Definition 7.13.** The joint distribution of the model is $\nu_\mathsf{T}\otimes K$: for any $\mathsf{A}\in\mathcal{T}\otimes\mathcal{Z}$

$$\nu_\mathsf{T}\otimes K(\mathsf{A}) = \int_{\mathsf{T}\times\mathsf{Z}} \mathbb{1}_\mathsf{A}(\vartheta,z_{\mathrm{obs}})K(\vartheta,\mathrm{d}x)\nu_\mathsf{T}(\mathrm{d}\vartheta) .$$

The marginal distribution $\nu_Z$ of $X$ is defined as $\nu_\mathsf{T}K$: for any $\tilde{\mathsf{A}}\in\mathcal{Z}$ by

$$\nu_Z(\tilde{\mathsf{A}}) = \int_{\mathsf{T}\times\mathsf{Z}} \mathbb{1}_{\mathsf{T}\times\tilde{\mathsf{A}}}(\vartheta,z_{\mathrm{obs}})K(\vartheta,\mathrm{d}x)\nu_\mathsf{T}(\mathrm{d}\vartheta) .$$

Note that if (7.11) holds true, the joint distribution admits a density with respect to $\nu_\mathsf{T}\otimes\mu$ given by $(\theta,x)\mapsto\mathrm{L}(\theta,x)$.

In addition, the marginal distribution also admits a density, with respect to $\mu$ given by for any $z_{\mathrm{obs}} \in \mathsf{Z}$

$$\mathrm{p}_Z(z_{\mathrm{obs}}) = \int_{\mathsf{T}} \mathrm{L}(\vartheta, z_{\mathrm{obs}}) \nu_{\mathsf{T}}(\mathrm{d}\vartheta) .$$

Bayes theorem [Sch95, Theorem 1.31] which is a simple consequence of Theorem 7.11 gives an expression of the conditional law of $\theta$ given $z_{\mathrm{obs}}$, called the posterior law, depending on the likelihood function and the prior distribution $\nu_{\mathsf{T}}$:

**Theorem 7.22.** *The conditional distribution of $\theta$ given $z_{\mathrm{obs}}$ denoted by $\pi_{\theta|Z}$ admits a conditional density with respect to $\nu_{\mathsf{T}}$ given for $\nu_{\mathsf{T}} \otimes K$-almost all $(\vartheta, x)$ by*

$$(x, \vartheta) \mapsto = \frac{\mathrm{L}(\vartheta, x)}{\mathrm{p}_Z(x)} .$$

**Definition 7.14.** Given a Bayesian model and some observation $z_{\mathrm{obs}}$, we define the posterior distribution of $\theta$ as the distribution with density with respect to $\nu_{\mathsf{T}}$,

$$\mathrm{p}_{\theta|Z}(\vartheta) = \frac{\mathrm{L}(\vartheta, z_{\mathrm{obs}})}{\mathrm{p}_Z(z_{\mathrm{obs}})} .$$

## 7.2 A bit of decision theory

Let $((\mathsf{Z}, \mathcal{Z}), (\mathsf{T}, \mathcal{T}), \mathcal{P}_{\mathsf{T}}, \nu_{\mathsf{T}})$ be a Bayesian model. Using the posterior distribution of Bayesian model, the parameter $\vartheta$ can be inferred as follows. Let $(\mathsf{D}, \mathcal{D})$ be a measurable space, called the decision/action space. For example for a test, the space $\mathsf{D} = \{0, 1\}$ would be chosen. If the interest is in estimating the parameter $\vartheta$, $\mathsf{D} = \mathsf{T}$.

**Definition 7.15.** (1) A decision rule $\delta$ is any measurable function from $(\mathsf{Z}, \mathcal{Z})$ to $(\mathsf{D}, \mathcal{D})$ and the set of all decision rules is denoted by $\mathbf{D}$.

(2) A loss function is any measurable function $\mathscr{L} : \mathsf{T} \times \mathsf{D} \to \mathbb{R}_+$.

(3) The risk function is defined for any decision rule $\delta$ and $\theta \in \mathsf{T}$ as

$$\mathrm{R}(\theta, \delta) = \int_{\mathsf{Z}} \mathscr{L}(\vartheta, \delta(z)) K(\theta, \mathrm{d}z) .$$

(4) The Bayes (or integrated) risk function is defined for any decision rule $\delta$ and prior $\nu_{\mathsf{T}}$ as

$$\mathscr{R}_B(\delta) = \int_{\mathsf{T}} \mathrm{R}(\vartheta, \delta) \nu_{\mathsf{T}}(\mathrm{d}\vartheta) .$$

Note that this function depends also implicitly on the prior $\nu_{\mathsf{T}}$.

Based on this definition, we aim to find a decision rule which minimize $\delta \mapsto \mathscr{R}_B(\delta)$. Decision rules which achieve are called Bayes rules. To this end, we introduce the concept of posterior risk

**Definition 7.16.** (1) Then, the posterior risk or partial Bayes risk associated with $\mathscr{L}$ is defined for any prior distribution $\nu_{\mathsf{T}}$ on $(\mathsf{T},\mathcal{T})$ and decision $a \in \mathsf{D}$ by

$$\rho(a|z_{\mathrm{obs}}) = \int_{\mathsf{T}} \mathscr{L}(\vartheta, a)\pi_{\theta|Z}(\mathrm{d}\vartheta|z_{\mathrm{obs}}) \,,$$

where $\pi_{\theta|Z}(\cdot|z_{\mathrm{obs}})$ is the posterior distribution of $\theta$ given $z_{\mathrm{obs}}$ associated with the prior distribution $\nu_{\mathsf{T}}$.

(2) The posterior risk associated with the decision rule $\delta$, also denoted by $\rho$, is given for any prior distribution $\nu_{\mathsf{T}}$ on $(\mathsf{T},\mathcal{T})$ by $\rho(\delta|z_{\mathrm{obs}}) = \rho(\delta(z_{\mathrm{obs}})|z_{\mathrm{obs}})$.

Decision rules which minimize the posterior risk, i.e., for $\nu_Z$-almost all $z_{\mathrm{obs}} \in \mathsf{Z}$, $\delta^\star_{\nu_{\mathsf{T}}}(z_{\mathrm{obs}}) \in \operatorname{argmin}_{a\in\mathsf{D}} \rho(\nu_{\mathsf{T}}, a|z_{\mathrm{obs}})$, are called partial Bayes rules.

We have the following result relating partial and non-partial Bayes rules.

**Proposition 7.23.** *Assume that there exists $\delta \in \mathbf{D}$ such that*

$$\int_{\mathsf{T}\times\mathsf{Z}} \mathscr{L}(\theta, \delta(z_{\mathrm{obs}}))K(\vartheta, \mathrm{d}z)\nu_{\mathsf{T}}(\mathrm{d}\vartheta) < +\infty \,.$$

*Then, the decision rule $\delta^\star_{\nu_{\mathsf{T}}}$ defined by minimization of the posterior risk, i.e., for any $z_{\mathrm{obs}} \in \mathsf{Z}$, $\delta^\star_{\nu_{\mathsf{T}}}(z_{\mathrm{obs}}) \in \operatorname{argmin}_{a\in\mathsf{D}} \rho(\nu_{\mathsf{T}}, a|z_{\mathrm{obs}})$ satisfies for almost all $z_{\mathrm{obs}} \in \mathsf{Z}$,*

$$\int_{\mathsf{T}\times\mathsf{Z}} \mathscr{L}(\vartheta, \delta^\star_{\nu_{\mathsf{T}}}(z_{\mathrm{obs}}))K(\vartheta, \mathrm{d}z)\nu_{\mathsf{T}}(\mathrm{d}\vartheta) = \inf_{\delta\in\mathbf{D}} \int_{\mathsf{T}\times\mathsf{Z}} \mathscr{L}(\vartheta, \delta(z_{\mathrm{obs}}))K(\vartheta, \mathrm{d}z)\nu_{\mathsf{T}}(\mathrm{d}\vartheta) \,,$$

The decision rule $\delta^\star_{\nu_{\mathsf{T}}}$ is said to be a Bayes rule with respect to $\nu_{\mathsf{T}}$. For example, if $\mathsf{T} \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, $\mathsf{D} = \mathsf{T}$, $\mathscr{L}(\vartheta, a) = \|\vartheta - a\|^2$, and $\pi_{\theta|Z}(\cdot)$ has a second moment, then by definition $\delta^\star_{\nu_{\mathsf{T}}} = \int_{\mathsf{T}} \vartheta \pi_{\theta|Z}(\mathrm{d}\vartheta)$ is the posterior mean. Bayes rules have optimality properties in statistical decision theory which make them very interesting; see [Sch95, chapitre 3] and [Rob07, chapitre 2] for an introduction to the subject.

## 7.3 Uncertainty quantification

To measure some uncertainty or make some tests, Bayesian statistics consider some confidence regions. Let $(\tilde{\mathsf{T}}, \tilde{\mathcal{T}})$ be a measurable space and $g : \mathsf{T} \to \tilde{\mathsf{T}}$ be a measurable function. We are interested in estimating a confident region for $g(\vartheta)$. Let $\alpha \in [0,1]$. A set $\mathsf{C}_\alpha \in \tilde{\mathcal{T}}$ is a $\alpha$-credible region are the sets such that

$$\pi_{\theta|Z}(\{\vartheta \in \mathsf{T} : g(\vartheta) \in \mathsf{C}_\alpha\}) \geq 1 - \alpha \,.$$

Assume that the distribution of $g(\theta)$ given $z_{\text{obs}}$ admits a density $f_{g(\theta)|Z}$ with respect to a reference measure $\nu_{g(\theta)}$ on $(\tilde{\mathsf{T}}, \tilde{\mathcal{T}})$. Then, we can define $\alpha$-credible HPD regions (Highest Posterior Density): $\mathsf{C}_\alpha \in \tilde{\mathcal{T}}$ is a $\alpha$-credible HPD region if

$$\left\{\tilde{\vartheta} \in \tilde{\mathsf{T}} : f_{g(\theta)|Z}(\tilde{\vartheta}) > \eta_\alpha\right\} \subset \mathsf{C}_\alpha \subset \left\{\tilde{\vartheta} \in \tilde{\mathsf{T}} : f_{g(\theta)|Z}(\tilde{\vartheta}) \geq \eta_\alpha\right\} ,$$

where

$$\eta_\alpha = \sup\left\{\eta \in \mathbb{R} | \pi_{\theta|Z}\left(\left\{\vartheta \in \mathsf{T} : f_{g(\theta)|Z}(g(\vartheta)) \geq \eta\right\}\right) \geq 1 - \alpha\right\} .$$

The definition of HPD regions is motivated by the fact that they are $\alpha$-credible regions with minimal volume for $\nu_{g(\theta)}$. However it is important to note that this definition of HPD regions depends on the dominating measure $\nu_{g(\theta)}$, see [Sch95, Section 5.2.4].

# [redacted]nces


[AB09] Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[Bak84] A. Baker. *A Concise Introduction to the Theory of Numbers*. Cambridge University Press, 1984. ISBN: 9780521286541. URL: https://books.google.fr/books?id=hAFfb8BBC0IC.

[Bar+98] Robert Barlovic, Ludger Santen, Andreas Schadschneider, and Michael Schreckenberg. "Metastable states in cellular automata for traffic flow". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 5 (1998), pp. 793–800.

[BFS87] Paul Bratley, Bennett L. Fox, and Linus E. Schrage. *A Guide to Simulation (2nd Ed.)* Berlin, Heidelberg: Springer-Verlag, 1987. ISBN: 0387964673.

[CP97] J. Chang and D. Pollard. "Conditioning as disintegration". In: *Statistica Neerlandica* 51.3 (1997), pp. 287–317.

[Fel+98] Anja Feldmann, Anna C Gilbert, Walter Willinger, and Tom G Kurtz. "The changing nature of network traffic: Scaling phenomena". In: *ACM SIGCOMM Computer Communication Review* 28.2 (1998), pp. 5–29.

[Geo11] Hans-Otto Georgii. *Gibbs measures and phase transitions*. Walter de Gruyter GmbH & Co. KG, Berlin, 2011.

[HD62] Thomas E Hull and Alan R Dobell. "Random number generators". In: *SIAM review* 4.3 (1962), pp. 230–254.

[HK12] Yo-Sub Han and Sang-Ki Ko. "Analysis of a cellular automaton model for car traffic with a junction". In: *Theoretical Computer Science* 450 (2012), pp. 54–67.

[Jam90] Frederick James. "A review of pseudorandom number generators". In: *Computer physics communications* 60.3 (1990), pp. 329–344.

[Knu97] Donald E. Knuth. *The Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Third. Boston: Addison-Wesley, 1997. ISBN: 0201896842 9780201896848.

[Kra06] Werner Krauth. *Statistical mechanics: algorithms and computations*. Vol. 13. OUP Oxford, 2006.

[Ler14] N. Lerner. *A Course on Integration Theory: including more than 150 exercises with detailed answers*. Springer Basel, 2014. ISBN: 9783034806947. URL: https://books.google.fr/books?id=j6-5BQAAQBAJ.

[LGM69] P. A. W. Lewis, A. S. Goodman, and J. M. Miller. "A pseudo-random number generator for the System/360". In: *IBM Systems Journal* 8.2 (1969), pp. 136–146. DOI: 10.1147/sj.82.0136.

[LP13] Sílvia R.C. Lopes and Taiane S. Prass. "Seasonal FIEGARCH processes". In: *Computational Statistics & Data Analysis* 68 (2013), pp. 262–295. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2013.07.001. URL: https://www.sciencedirect.com/science/article/pii/S0167947313002466.

[NS92] Kai Nagel and Michael Schreckenberg. "A cellular automaton model for freeway traffic". In: *Journal de physique I* 2.12 (1992), pp. 2221–2229.

[Pol02] D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2002. ISBN: 9780521002899. URL: https://books.google.fr/books?id=B7Ch-c2G21MC.

[Rob07] C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York, 2007. ISBN: 9780387715995.

[Rou16] F. Roueff. *Time series analysis, lecture notes*. 2016.

[RS94] BV Rao and SM Srivastava. "An elementary proof of the Borel isomorphism theorem". In: *Real Analysis Exchange* (1994), pp. 347–349.

[Rud06] Walter Rudin. *Real and complex analysis*. Tata McGraw-hill education, 2006.

[RW13] Ian W Renner and David I Warton. "Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology". In: *Biometrics* 69.1 (2013), pp. 274–281.

[Sch95] M. J. Schervish. *Theory of statistics*. Springer series in statistics. Corrected second printing : 1997. New York, Berlin, Heidelberg: Springer, 1995. ISBN: 0-387-94546-6.

[Sho09] Victor Shoup. *A computational introduction to number theory and algebra*. Cambridge university press, 2009.

[Sip13] Michael Sipser. *Introduction to the Theory of Computation*. Third. Boston, MA: Course Technology, 2013. ISBN: 113318779X.

[Tez12] Shu Tezuka. *Uniform random numbers: Theory and practice*. Vol. 315. Springer Science & Business Media, 2012.

[Vil09] C Villani. *Optimal transport*. Vol. 338. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Old and new. Springer-Verlag, Berlin, 2009, pp. xxii+973. ISBN: 978-3-540-71049-3. DOI: 10.1007/978-3-540-71050-9. URL: https://doi.org/10.1007/978-3-540-71050-9.