

## Main Definitions and Notations

---

We now formally describe hidden Markov models, setting the notations that will be used throughout the book. We start by reviewing the basic definitions and concepts pertaining to Markov chains.

### 2.1 Markov Chains

#### 2.1.1 Transition Kernels

**Definition 2.1.1 (Transition Kernel).** *Let  $(X, \mathcal{X})$  and  $(Y, \mathcal{Y})$  be two measurable spaces. An unnormalized transition kernel from  $(X, \mathcal{X})$  to  $(Y, \mathcal{Y})$  is a function  $Q : X \times \mathcal{Y} \rightarrow [0, \infty]$  that satisfies*

- (i) *for all  $x \in X$ ,  $Q(x, \cdot)$  is a positive measure on  $(Y, \mathcal{Y})$ ;*
- (ii) *for all  $A \in \mathcal{Y}$ , the function  $x \mapsto Q(x, A)$  is measurable.*

*If  $Q(x, Y) = 1$  for all  $x \in X$ , then  $Q$  is called a transition kernel, or simply a kernel. If  $X = Y$  and  $Q(x, X) = 1$  for all  $x \in X$ , then  $Q$  will also be referred to as a Markov transition kernel on  $(X, \mathcal{X})$ .*

*An (unnormalized) transition kernel  $Q$  is said to admit a density with respect to the positive measure  $\mu$  on  $Y$  if there exists a non-negative function  $q : X \times Y \rightarrow [0, \infty]$ , measurable with respect to the product  $\sigma$ -field  $\mathcal{X} \otimes \mathcal{Y}$ , such that*

$$Q(x, A) = \int_A q(x, y) \mu(dy), \quad A \in \mathcal{Y}.$$

*The function  $q$  is then referred to as an (unnormalized) transition density function.*

*When  $X$  and  $Y$  are countable sets it is customary to write  $Q(x, y)$  as a shorthand notation for  $Q(x, \{y\})$ , and  $Q$  is generally referred to as a transition matrix (whether or not  $X$  and  $Y$  are finite sets).*

We summarize below some key properties of transition kernels, introducing important pieces of notation that are used in the following.

- Let  $Q$  and  $R$  be unnormalized transition kernels from  $(X, \mathcal{X})$  to  $(Y, \mathcal{Y})$  and from  $(Y, \mathcal{Y})$  to  $(Z, \mathcal{Z})$ , respectively. The product  $QR$ , defined by

$$QR(x, A) \stackrel{\text{def}}{=} \int Q(x, dy) R(y, A), \quad x \in X, A \in \mathcal{Z},$$

is then an unnormalized transition kernel from  $(X, \mathcal{X})$  to  $(Z, \mathcal{Z})$ . If  $Q$  and  $R$  are transition kernels, then so is  $QR$ , that is,  $QR(x, Z) = 1$  for all  $x \in X$ .

- If  $Q$  is an (unnormalized) Markov transition kernel on  $(X, \mathcal{X})$ , its iterates are defined inductively by

$$Q^0(x, \cdot) = \delta_x \text{ for } x \in X \text{ and } Q^k = QQ^{k-1} \text{ for } k \geq 1.$$

These iterates satisfy the *Chapman-Kolmogorov* equation:  $Q^{n+m} = Q^n Q^m$  for all  $n, m \geq 0$ . That is, for all  $x \in X$  and  $A \in \mathcal{X}$ ,

$$Q^{n+m}(x, A) = \int Q^n(x, dy) Q^m(y, A). \quad (2.1)$$

If  $Q$  admits a density  $q$  with respect to the measure  $\mu$  on  $(X, \mathcal{X})$ , then for all  $n \geq 2$  the kernel  $Q^n$  is also absolutely continuous with respect to  $\mu$ . The corresponding transition density is

$$q_n(x, y) = \int_{X^{n-1}} q(x, x_1) \cdots q(x_{n-1}, y) \mu(dx_1) \cdots \mu(dx_{n-1}). \quad (2.2)$$

- Positive measures operate on (unnormalized) transition kernels in two different ways. If  $\mu$  is a positive measure on  $(X, \mathcal{X})$ , the positive measure  $\mu Q$  on  $(Y, \mathcal{Y})$  is defined by

$$\mu Q(A) \stackrel{\text{def}}{=} \int \mu(dx) Q(x, A), \quad A \in \mathcal{Y}.$$

Moreover, the measure  $\mu \otimes Q$  on the product space  $(X \times Y, \mathcal{X} \otimes \mathcal{Y})$  is defined by

$$\mu \otimes Q(C) \stackrel{\text{def}}{=} \iint_C \mu(dx) Q(x, dy), \quad C \in \mathcal{X} \otimes \mathcal{Y}.$$

If  $\mu$  is a probability measure and  $Q$  is a transition kernel, then  $\mu Q$  and  $\mu \otimes Q$  are probability measures.

- (Unnormalized) transition kernels operate on functions. Let  $f$  be a real measurable function on  $Y$ . The real measurable function  $Qf$  on  $X$  is defined by

$$Qf(x) \stackrel{\text{def}}{=} \int Q(x, dy) f(y), \quad x \in X,$$

provided the integral is well-defined. It will sometimes be more convenient to use the alternative notation  $Q(x, f)$  instead of  $Qf(x)$ . In particular,

for  $x \in \mathsf{X}$  and  $A \in \mathcal{Y}$ ,  $Q(x, A)$ ,  $\delta_x Q(A)$ ,  $Q\mathbb{1}_A(x)$ , and  $Q(x, \mathbb{1}_A)$ , where  $\mathbb{1}_A$  denotes the indicator function of the set  $A$ , are four equivalent ways of denoting the same quantity. In general, we prefer using the  $Q(x, \mathbb{1}_A)$  and  $Q(x, A)$  variants, which are less prone to confusion in complicated expressions.

- For any positive measure  $\mu$  on  $(\mathsf{X}, \mathcal{X})$  and any real measurable function  $f$  on  $(\mathsf{Y}, \mathcal{Y})$ ,

$$(\mu Q)(f) = \mu(Qf) = \iint \mu(dx) Q(x, dy) f(y) ,$$

provided the integrals are well-defined. We may thus use the simplified notation  $\nu Qf$  instead of  $(\nu Q)(f)$  or  $\nu(Qf)$ .

**Definition 2.1.2 (Reverse Kernel).** *Let  $Q$  be a transition kernel from  $(\mathsf{X}, \mathcal{X})$  to  $(\mathsf{Y}, \mathcal{Y})$  and let  $\nu$  be a probability measure on  $(\mathsf{X}, \mathcal{X})$ . The reverse kernel  $\overleftarrow{Q}_\nu$  associated to  $\nu$  and  $Q$  is a transition kernel from  $(\mathsf{Y}, \mathcal{Y})$  to  $(\mathsf{X}, \mathcal{X})$  such that for all bounded measurable functions  $f$  defined on  $\mathsf{X} \times \mathsf{Y}$ ,*

$$\iint_{\mathsf{X} \times \mathsf{Y}} f(x, y) \nu(dx) Q(x, dy) = \iint_{\mathsf{X} \times \mathsf{Y}} f(x, y) \nu Q(dy) \overleftarrow{Q}_\nu(y, dx) . \quad (2.3)$$

The reverse kernel does not necessarily exist and is not uniquely defined. Nevertheless, if  $\overleftarrow{Q}_{\nu,1}$  and  $\overleftarrow{Q}_{\nu,2}$  satisfy (2.3), then for all  $A \in \mathcal{X}$ ,  $\overleftarrow{Q}_{\nu,1}(y, A) = \overleftarrow{Q}_{\nu,2}(y, A)$  for  $\nu Q$ -almost every  $y$  in  $\mathsf{Y}$ . The reverse kernel does exist if  $\mathsf{X}$  and  $\mathsf{Y}$  are Polish spaces endowed with their Borel  $\sigma$ -fields (see Appendix A.1 for details). If  $Q$  admits a density  $q$  with respect to a measure  $\mu$  on  $(\mathsf{Y}, \mathcal{Y})$ , then  $\overleftarrow{Q}_\nu$  can be defined for all  $y$  such that  $\int_{\mathsf{X}} q(z, y) \nu(dz) \neq 0$  by

$$\overleftarrow{Q}_\nu(y, dx) = \frac{q(x, y) \nu(dx)}{\int_{\mathsf{X}} q(z, y) \nu(dz)} . \quad (2.4)$$

The values of  $\overleftarrow{Q}_\nu$  on the set  $\{y \in \mathsf{Y} : \int_{\mathsf{X}} q(z, y) \nu(dz) = 0\}$  are irrelevant because this set is  $\nu Q$ -negligible. In particular, if  $\mathsf{X}$  is discrete and  $\mu$  is counting measure, then for all  $(x, y) \in \mathsf{X} \times \mathsf{Y}$  such that  $\nu Q(y) \neq 0$ ,

$$\overleftarrow{Q}_\nu(y, x) = \frac{\nu(x) Q(x, y)}{\nu Q(y)} . \quad (2.5)$$

### 2.1.2 Homogeneous Markov Chains

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and let  $(\mathsf{X}, \mathcal{X})$  be a measurable space. An  $\mathsf{X}$ -valued (discrete index) *stochastic process*  $\{X_n\}_{n \geq 0}$  is a collection of  $\mathsf{X}$ -valued random variables. A *filtration* of  $(\Omega, \mathcal{F})$  is a non-decreasing sequence  $\{\mathcal{F}_n\}_{n \geq 0}$  of sub- $\sigma$ -fields of  $\mathcal{F}$ . A *filtered space* is a triple  $(\Omega, \mathcal{F}, \mathbb{F})$ , where  $\mathbb{F}$  is a filtration;  $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$  is called a *filtered probability space*. For any filtration

$\mathbb{F} = \{\mathcal{F}_n\}_{n \geq 0}$ , we denote by  $\mathcal{F}_\infty = \bigvee_{n=0}^\infty \mathcal{F}_n$  the  $\sigma$ -field generated by  $\mathbb{F}$  or, in other words, the minimal  $\sigma$ -field containing  $\mathbb{F}$ . A stochastic process  $\{X_n\}_{n \geq 0}$  is *adapted* to  $\mathbb{F} = \{\mathcal{F}_n\}_{n \geq 0}$ , or simply  *$\mathbb{F}$ -adapted*, if  $X_n$  is  $\mathcal{F}_n$ -measurable for all  $n \geq 0$ . The *natural filtration* of a process  $\{X_n\}_{n \geq 0}$ , denoted by  $\mathbb{F}^X = \{\mathcal{F}_n^X\}_{n \geq 0}$ , is the smallest filtration with respect to which  $\{X_n\}$  is adapted.

**Definition 2.1.3 (Markov Chain).** *Let  $(\Omega, \mathcal{F}, \mathbb{F}, P)$  be a filtered probability space and let  $Q$  be a Markov transition kernel on a measurable space  $(X, \mathcal{X})$ . An  $X$ -valued stochastic process  $\{X_k\}_{k \geq 0}$  is said to be a Markov chain under  $P$ , with respect to the filtration  $\mathbb{F}$  and with transition kernel  $Q$ , if it is  $\mathbb{F}$ -adapted and for all  $k \geq 0$  and  $A \in \mathcal{X}$ ,*

$$P(X_{k+1} \in A \mid \mathcal{F}_k) = Q(X_k, A) . \quad (2.6)$$

*The distribution of  $X_0$  is called the initial distribution of the chain, and  $X$  is called the state space.*

If  $\{X_k\}_{k \geq 0}$  is  $\mathbb{F}$ -adapted, then for all  $k \geq 0$  it holds that  $\mathcal{F}_k^X \subseteq \mathcal{F}_k$ ; hence a Markov chain with respect to a filtration  $\mathbb{F}$  is also a Markov chain with respect to its natural filtration. Hereafter, a Markov chain with respect to its natural filtration will simply be referred to as a Markov chain. When there is no risk of confusion, we will not mention the underlying probability measure  $P$ .

A fundamental property of a Markov chain is that its finite-dimensional distributions, and hence the distribution of the process  $\{X_k\}_{k \geq 0}$ , are entirely determined by the initial distribution and the transition kernel.

**Proposition 2.1.4.** *Let  $\{X_k\}_{k \geq 0}$  be a Markov chain with initial distribution  $\nu$  and transition kernel  $Q$ . For any  $k \geq 0$  and any bounded  $\mathcal{X}^{\otimes(k+1)}$ -measurable function  $f$  on  $X^{(k+1)}$ ,*

$$E[f(X_0, \dots, X_k)] = \int f(x_0, \dots, x_k) \nu(dx_0) Q(x_0, dx_1) \cdots Q(x_{k-1}, dx_k) .$$

In the following, we will use the generic notation  $f \in \mathcal{F}_b(Z)$  to denote the fact that  $f$  is a measurable bounded function on  $(Z, \mathcal{Z})$ . In the case of Proposition 2.1.4 for instance, one considers functions  $f$  that are in  $\mathcal{F}_b(X^{(k+1)})$ . More generally, we will usually describe measures and transition kernels on  $(Z, \mathcal{Z})$  by specifying the way they operate on the functions of  $\mathcal{F}_b(Z)$ .

### 2.1.2.1 Canonical Version

Let  $(X, \mathcal{X})$  be a measurable space. The *canonical space* associated to  $(X, \mathcal{X})$  is the infinite-dimensional product space  $(X^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$ . The *coordinate process* is the  $X$ -valued stochastic process  $\{X_k\}_{k \geq 0}$  defined on the canonical space by  $X_n(\omega) = \omega(n)$ . The canonical space will always be endowed with the natural filtration  $\mathbb{F}^X$  of the coordinate process.

Let  $(\Omega, \mathcal{F}) = (\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$  be the canonical space associated to the measurable space  $(\mathbf{X}, \mathcal{X})$ . The *shift operator*  $\theta : \Omega \rightarrow \Omega$  is defined by

$$\theta(\omega)(n) = \omega(n+1), \quad n \geq 0.$$

The iterates of the shift operator are defined inductively by  $\theta^0 = \text{Id}$  (the identity),  $\theta^1 = \theta$  and  $\theta^k = \theta \circ \theta^{k-1}$  for  $k \geq 1$ . If  $\{X_k\}_{k \geq 0}$  is the coordinate process with associated natural filtration  $\mathbb{F}^X$ , then for all  $k, n \geq 0$ ,  $X_k \circ \theta^n = X_{k+n}$ , and more generally for any  $\mathcal{F}_k^X$ -measurable random variable  $Y$ ,  $Y \circ \theta^n$  is  $\mathcal{F}_{n+k}^X$ -measurable.

The following theorem, which is a particular case of the Kolmogorov consistency theorem, states that it is always possible to define a Markov chain on the canonical space.

**Theorem 2.1.5.** *Let  $(\mathbf{X}, \mathcal{X})$  be a measurable set,  $\nu$  a probability measure on  $(\mathbf{X}, \mathcal{X})$ , and  $Q$  a transition kernel on  $(\mathbf{X}, \mathcal{X})$ . Then there exists a unique probability measure on  $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$ , denoted by  $P_\nu$ , such that the coordinate process  $\{X_k\}_{k \geq 0}$  is a Markov chain (with respect to its natural filtration) with initial distribution  $\nu$  and transition kernel  $Q$ .*

For  $x \in \mathbf{X}$ , let  $P_x$  be an alternative simplified notation for  $P_{\delta_x}$ . Then for all  $A \in \mathcal{X}^{\otimes \mathbb{N}}$ , the mapping  $x \rightarrow P_x(A) = Q(x, A)$  is  $\mathcal{X}$ -measurable, and for any probability measure  $\nu$  on  $(\mathbf{X}, \mathcal{X})$ ,

$$P_\nu(A) = \int \nu(dx) P_x(A). \quad (2.7)$$

The Markov chain defined in Theorem 2.1.5 is referred to as the *canonical version* of the Markov chain. The probability  $P_\nu$  defined on  $(\mathbf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes \mathbb{N}})$  depends on  $\nu$  and on the transition kernel  $Q$ . Nevertheless, the dependence with respect to  $Q$  is traditionally omitted in the notation. The relation (2.7) implies that  $x \rightarrow P_x$  is a regular version of the conditional probability  $P_\nu(\cdot | X_k = x)$  in the sense that one can rewrite (2.6) as

$$P_\nu(X_{k+1} \in A | \mathcal{F}_k^X) = P_\nu(X_1 \circ \theta^k \in A | \mathcal{F}_k^X) = P_{X_k}(X_1 \in A) \quad P_\nu\text{-a.s.}$$

### 2.1.2.2 Markov Properties

More generally, an induction argument easily yields the *Markov property*: for any  $\mathcal{F}_\infty^X$ -measurable random variable  $Y$ ,

$$E_\nu[Y \circ \theta^k | \mathcal{F}_k^X] = E_{X_k}[Y] \quad P_\nu\text{-a.s.} \quad (2.8)$$

The Markov property can be extended to a specific class of random times known as *stopping times*. Let  $\bar{\mathbb{N}} = \mathbb{N} \cup \{+\infty\}$  denote the extended integer set and let  $(\Omega, \mathcal{F}, \mathbb{F})$  be a filtered space. Then, a mapping  $\tau : \Omega \rightarrow \bar{\mathbb{N}}$  is said to be an  $\mathbb{F}$ -stopping time if  $\{\tau = n\} \in \mathcal{F}_n$  for all  $n \geq 0$ . Intuitively, this means that at any time  $n$  one should be able to tell, based on the information  $\mathcal{F}_n$

available at that time, if the stopping time occurs at this time  $n$  (or before then) or not. The class  $\mathcal{F}_\tau$  defined by

$$\mathcal{F}_\tau = \{B \in \mathcal{F}_\infty : B \cap \{\tau = n\} \in \mathcal{F}_n \text{ for all } n \geq 0\} ,$$

is a  $\sigma$ -field, referred to as the  $\sigma$ -field of the events occurring before  $\tau$ .

**Theorem 2.1.6 (Strong Markov Property).** *Let  $\{X_k\}_{k \geq 0}$  be the canonical version of a Markov chain and let  $\tau$  be an  $\mathbb{F}^X$ -stopping time. Then for any bounded  $\mathcal{F}_\infty^X$ -measurable function  $\Psi$ ,*

$$\mathbb{E}_\nu[\mathbb{1}_{\{\tau < \infty\}} \Psi \circ \theta^\tau \mid \mathcal{F}_\tau^X] = \mathbb{1}_{\{\tau < \infty\}} \mathbb{E}_{X_\tau}[\Psi] \quad \mathbb{P}_\nu\text{-a.s.} \quad (2.9)$$

We note that an  $\mathcal{F}_\infty^X$ -measurable function, or random variable,  $\Psi$ , is typically a function of potentially the whole trajectory of the Markov chain, although it may of course be a rather simple function like  $X_1$  or  $X_2 + X_3^2$ .

### 2.1.3 Non-homogeneous Markov Chains

**Definition 2.1.7 (Non-homogeneous Markov Chain).** *Let  $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$  be a filtered probability space and let  $\{Q_k\}_{k \geq 0}$  be a family of transition kernels on a measurable space  $(X, \mathcal{X})$ . An  $X$ -valued stochastic process  $\{X_k\}_{k \geq 0}$  is said to be a non-homogeneous Markov chain under  $\mathbb{P}$ , with respect to the filtration  $\mathbb{F}$  and with transition kernels  $\{Q_k\}$ , if it is  $\mathbb{F}$ -adapted and for all  $k \geq 0$  and  $A \in \mathcal{X}$ ,*

$$\mathbb{P}(X_{k+1} \in A \mid \mathcal{F}_k) = Q_k(X_k, A) .$$

For  $i \leq j$  we define

$$Q_{i,j} = Q_i Q_{i+1} \cdots Q_j .$$

With this notation, if  $\nu$  denotes the distribution of  $X_0$  (which we refer to as the initial distribution as in the homogeneous case), the distribution of  $X_n$  is  $\nu Q_{0,n-1}$ . An important example of a non-homogeneous Markov chain is the so-called reverse chain. The construction of the reverse chain is based on the observation that if  $\{X_k\}_{k \geq 0}$  is a Markov chain, then for any index  $n \geq 1$  the time-reversed (or, index-reversed) process  $\{X_{n-k}\}_{k=0}^n$  is a Markov chain too. The definition below provides its transition kernels.

**Definition 2.1.8 (Reverse Chain).** *Let  $Q$  be a Markov kernel on some space  $X$ , let  $\nu$  be a probability measure on this space, and let  $n \geq 1$  be an index. The reverse chain is the non-homogeneous Markov chain with initial distribution  $\nu Q^n$ , (time) index set  $k = 0, 1, \dots, n$  and transition kernels*

$$Q_k = \overleftarrow{Q}_{\nu Q^{n-k-1}} , \quad k = 0, \dots, n-1 ,$$

assuming that the reverse kernels are indeed well-defined.

If the transition kernel  $Q$  admits a transition density function  $q$  with respect to a measure  $\mu$  on  $(\mathbf{X}, \mathcal{X})$ , then  $Q_k$  also admits a density with respect to the same measure  $\mu$ , namely

$$h_k(y, x) = \frac{\int q_{n-k-1}(z, x) q(x, y) \nu(dz)}{\int q_{n-k}(z, y) \nu(dz)}. \quad (2.10)$$

Here,  $q_l$  is the transition density function of  $Q^l$  with respect to  $\mu$  as defined in (2.2). If the state space is countable, then

$$Q_k(y, x) = \frac{\nu Q^{n-k-1}(x) Q(x, y)}{\nu Q^{n-k}(y)}. \quad (2.11)$$

An interesting question is in what cases the kernels  $Q_k$  do not depend on the index  $k$  and are in fact all equal to the forward kernel  $Q$ . A Markov chain with this property is said to be *reversible*. The following result gives a necessary and sufficient condition for reversibility.

**Theorem 2.1.9.** *Let  $\mathbf{X}$  be a Polish space. A Markov kernel  $Q$  on  $\mathbf{X}$  is reversible with respect to a probability measure  $\nu$  if and only if for all bounded measurable functions  $f$  on  $\mathbf{X} \times \mathbf{X}$ ,*

$$\iint f(x, x') \nu(dx) Q(x, dx') = \iint f(x, x') \nu(dx') Q(x', dx). \quad (2.12)$$

The relation (2.12) is referred to as the *local balance equations* (or *detailed balance equations*). If the state space is countable, these equations hold if for all  $x, x' \in \mathbf{X}$ ,

$$\nu(x) Q(x, x') = \nu(x') Q(x', x). \quad (2.13)$$

Upon choosing a function  $f$  that only depends on the second variable in (2.12), it is easily seen that  $\nu Q(f) = \nu(f)$  for all functions  $f \in \mathcal{F}_b(\mathbf{X})$ . We can also write this as  $\nu = \nu Q$ . This equation is referred to as the *global balance equations*. By induction, we find that  $\nu Q^n = \nu$  for all  $n \geq 0$ . The left-hand side of this equation is the distribution of  $X_n$ , which thus does not depend on  $n$  when global balance holds. This is a form of stationarity, obviously implied by local balance. We shall tie this form of stationarity to the following customary definition.

**Definition 2.1.10 (Stationary Process).** *A stochastic process  $\{X_k\}$  is said to be stationary (under  $P$ ) if its finite-dimensional distributions are translation invariant, that is, if for all  $k, n \geq 1$  and all  $n_1, \dots, n_k$ , the distribution of the random vector  $(X_{n_1+n}, \dots, X_{n_k+n})$  does not depend on  $n$ .*

A stochastic process with index set  $\mathbb{N}$ , stationary but otherwise general, can always be extended to a process with index set  $\mathbb{Z}$ , having the same finite-dimensional distributions (and hence being stationary). This is a consequence of Kolmogorov's existence theorem for stochastic processes.

For a Markov chain, any multi-dimensional distribution can be expressed in terms of the initial distribution and the transition kernel—this is Proposition 2.1.4—and hence the characterization of stationarity becomes much simpler than above. Indeed, a Markov chain is stationary if and only if its initial distribution  $\nu$  and transition kernel  $Q$  satisfy  $\nu Q = \nu$ , that is, satisfy global balance. Much more will be said about stationary distributions of Markov chains in Chapter 14.

## 2.2 Hidden Markov Models

A hidden Markov model is a doubly stochastic process with an underlying stochastic process that is not directly observable (it is “hidden”) but can be observed only through another stochastic process that produces the sequence of observations. As shown in the introduction, the scope of HMMs is large and covers a variety of situations. To accommodate these conceptually different models, we now define formally a hidden Markov model.

### 2.2.1 Definitions and Notations

In simple cases such as fully discrete models, it is common to define hidden Markov models by using the concept of conditional independence. Indeed, this was the view taken in Chapter 1, where an HMM was defined as a bivariate process  $\{(X_k, Y_k)\}_{k \geq 0}$  such that

- $\{X_k\}_{k \geq 0}$  is a Markov chain with transition kernel  $Q$  and initial distribution  $\nu$ ;
- Conditionally on the state process  $\{X_k\}_{k \geq 0}$ , the observations  $\{Y_k\}_{k \geq 0}$  are independent, and for each  $n$  the conditional distribution of  $Y_n$  depends on  $X_n$  only.

It turns out that conditional independence is mathematically more difficult to define in general settings (in particular, when the state space  $\mathbf{X}$  of the Markov chain is not countable), and we will adopt a different route to define general hidden Markov models. The HMM is defined as a bivariate Markov chain, only partially observed though, whose transition kernel has a special structure. Indeed, its transition kernel should be such that both the joint process  $\{X_k, Y_k\}_{k \geq 0}$  and the marginal unobservable (or hidden) chain  $\{X_k\}_{k \geq 0}$  are Markovian. From this definition, the usual conditional independence properties of HMMs will then follow (see Corollary 2.2.5 below).

**Definition 2.2.1 (Hidden Markov Model).** *Let  $(\mathbf{X}, \mathcal{X})$  and  $(\mathbf{Y}, \mathcal{Y})$  be two measurable spaces and let  $Q$  and  $G$  denote, respectively, a Markov transition kernel on  $(\mathbf{X}, \mathcal{X})$  and a transition kernel from  $(\mathbf{X}, \mathcal{X})$  to  $(\mathbf{Y}, \mathcal{Y})$ . Consider the Markov transition kernel defined on the product space  $(\mathbf{X} \times \mathbf{Y}, \mathcal{X} \otimes \mathcal{Y})$  by*



$$T[(x, y), C] = \iint_C Q(x, dx') G(x', dy') , \quad (x, y) \in \mathbf{X} \times \mathbf{Y}, C \in \mathcal{X} \otimes \mathcal{Y} . \quad (2.14)$$

The Markov chain  $\{X_k, Y_k\}_{k \geq 0}$  with Markov transition kernel  $T$  and initial distribution  $\nu \otimes G$ , where  $\nu$  is a probability measure on  $(\mathbf{X}, \mathcal{X})$ , is called a hidden Markov model.

Although the definition above concerns the joint process  $\{X_k, Y_k\}_{k \geq 0}$ , the term *hidden* is only justified in cases where  $\{X_k\}_{k \geq 0}$  is not observable. In this respect,  $\{X_k\}_{k \geq 0}$  can also be seen as a fictitious intermediate process that is useful only in defining the distribution of the observed process  $\{Y_k\}_{k \geq 0}$ . We shall denote by  $P_\nu$  and  $E_\nu$  the probability measure and corresponding expectation associated with the process  $\{X_k, Y_k\}_{k \geq 0}$  on the canonical space  $((\mathbf{X} \times \mathbf{Y})^\mathbb{N}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}})$ . Notice that this constitutes a slight departure from the Markov notations introduced previously, as  $\nu$  is a probability measure on  $\mathbf{X}$  only and not on the state space  $\mathbf{X} \times \mathbf{Y}$  of the joint process. This slight abuse of notation is justified by the special structure of the model considered here. Equation (2.14) shows that whatever the distribution of the initial joint state  $(X_0, Y_0)$ , even if it were not of the form  $\nu \times G$ , the law of  $\{X_k, Y_k\}_{k \geq 1}$  only depends on the marginal distribution of  $X_0$ . Hence it makes sense to index probabilities and expectations by this marginal initial distribution only.

If both  $\mathbf{X}$  and  $\mathbf{Y}$  are countable, the hidden Markov model is said to be *discrete*, which is the case originally considered by Baum and Petrie (1966). Many of the examples given in the introduction (those of Section 1.3.2 for instance) correspond to cases where  $\mathbf{Y}$  is uncountable and is a subset of  $\mathbb{R}^d$  for some  $d$ . In such cases, we shall generally assume that the following holds true.

**Definition 2.2.2 (Partially Dominated Hidden Markov Model).** *The model of Definition 2.2.1 is said to be partially dominated if there exists a probability measure  $\mu$  on  $(\mathbf{Y}, \mathcal{Y})$  such that for all  $x \in \mathbf{X}$ ,  $G(x, \cdot)$  is absolutely continuous with respect to  $\mu$ ,  $G(x, \cdot) \ll \mu(\cdot)$ , with transition density function  $g(x, \cdot)$ . Then, for  $A \in \mathcal{Y}$ ,  $G(x, A) = \int_A g(x, y) \mu(dy)$  and the joint transition kernel  $T$  can be written as*

$$T[(x, y), C] = \iint_C Q(x, dx') g(x', y') \mu(dy') \quad C \in \mathcal{X} \otimes \mathcal{Y} . \quad (2.15)$$

In the third part of the book (Chapter 10 and following) where we consider statistical estimation for HMMs with unknown parameters, we will require even stronger conditions and assume that the model is fully dominated in the following sense.

**Definition 2.2.3 (Fully Dominated Hidden Markov Model).** *If, in addition to the requirements of Definition 2.2.2, there exists a probability measure  $\lambda$  on  $(\mathbf{X}, \mathcal{X})$  such that  $\nu \ll \lambda$  and, for all  $x \in \mathbf{X}$ ,  $Q(x, \cdot) \ll \lambda(\cdot)$  with transition density function  $q(x, \cdot)$ . Then, for  $A \in \mathcal{X}$ ,  $Q(x, A) = \int_A q(x, x') \lambda(dx')$*

and the model is said to be fully dominated. The joint Markov transition kernel  $T$  is then dominated by the product measure  $\lambda \otimes \mu$  and admits the transition density function

$$t[(x, y), (x', y')] \stackrel{\text{def}}{=} q(x, x')g(x', y') . \quad (2.16)$$

Note that for such models, we will generally re-use the notation  $\nu$  to denote the *probability density function* of the initial state  $X_0$  (with respect to  $\lambda$ ) rather than the distribution itself.

### 2.2.2 Conditional Independence in Hidden Markov Models

In this section, we will show that the “intuitive” way of thinking about an HMM, in terms of conditional independence, is justified by Definition 2.2.1. Readers unfamiliar with conditioning in general settings may want to read more on this topic in Appendix A.4 before reading the rest of this section.

**Proposition 2.2.4.** *Let  $\{X_k, Y_k\}_{k \geq 0}$  be a Markov chain over the product space  $\mathsf{X} \times \mathsf{Y}$  with transition kernel  $T$  given by (2.14). Then, for any integer  $p$ , any ordered set  $\{k_1 < \dots < k_p\}$  of indices and all functions  $f_1, \dots, f_p \in \mathcal{F}_b(\mathsf{Y})$ ,*

$$\mathbb{E}_\nu \left[ \prod_{i=1}^p f_i(Y_{k_i}) \middle| X_{k_1}, \dots, X_{k_p} \right] = \prod_{i=1}^p \int_{\mathsf{Y}} f_i(y) G(X_{k_i}, dy) . \quad (2.17)$$

*Proof.* For any  $h \in \mathcal{F}_b(\mathsf{X}^p)$ , it holds that

$$\begin{aligned} & \mathbb{E}_\nu \left[ \prod_{i=1}^p f_i(Y_{k_i}) h(X_{k_1}, \dots, X_{k_p}) \right] \\ &= \int \cdots \int \nu(dx_0) G(x_0, dy_0) \left[ \prod_{i=1}^{k_p} Q(x_{i-1}, dx_i) G(x_i, dy_i) \right] \\ & \quad \times \left[ \prod_{i=1}^p f_i(y_{k_i}) \right] h(x_{k_1}, \dots, x_{k_p}) \\ &= \int \cdots \int \nu(dx_0) \prod_{i=1}^{k_p} Q(x_{i-1}, dx_i) h(x_{k_1}, \dots, x_{k_p}) \\ & \quad \int \cdots \int \left[ \prod_{i \notin \{k_1, \dots, k_p\}} G(x_i, dy_i) \right] \left[ \prod_{i \in \{k_1, \dots, k_p\}} \int f_i(y_i) G(x_i, dy_i) \right] . \end{aligned}$$

Because  $\int G(x_i, dy_i) = 1$ ,

$$\begin{aligned} \mathbb{E}_\nu \left[ \prod_{i=1}^p f_i(Y_{k_i}) h(X_{k_1}, \dots, X_{k_p}) \right] = \\ \mathbb{E}_\nu \left[ h(X_{k_1}, \dots, X_{k_p}) \prod_{i \in \{k_1, \dots, k_p\}} \int f_i(y_i) G(X_i, dy_i) \right]. \end{aligned}$$

□

**Corollary 2.2.5.**

- (i) For any integer  $p$  and any ordered set  $\{k_1 < \dots < k_p\}$  of indices, the random variables  $Y_{k_1}, \dots, Y_{k_p}$  are  $\mathbb{P}_\nu$ -conditionally independent given  $(X_{k_1}, X_{k_2}, \dots, X_{k_p})$ .
- (ii) For any integers  $k$  and  $p$  and any ordered set  $\{k_1 < \dots < k_p\}$  of indices such that  $k \notin \{k_1, \dots, k_p\}$ , the random variables  $Y_k$  and  $(X_{k_1}, \dots, X_{k_p})$  are  $\mathbb{P}_\nu$ -conditionally independent given  $X_k$ .

*Proof.* Part (i) is an immediate consequence of Proposition 2.2.4. To prove (ii), note that for any  $f \in \mathcal{F}_b(\mathcal{Y})$  and  $h \in \mathcal{F}_b(\mathcal{X}^p)$ ,

$$\begin{aligned} \mathbb{E}_\nu [f(Y_k) h(X_{k_1}, \dots, X_{k_p}) | X_k] \\ = \mathbb{E}_\nu [\mathbb{E}_\nu [f(Y_k) | X_{k_1}, \dots, X_{k_p}, X_k] h(X_{k_1}, \dots, X_{k_p}) | X_k] \\ = \mathbb{E}_\nu [f(Y_k) | X_k] \mathbb{E}_\nu [h(X_{k_1}, \dots, X_{k_p}) | X_k]. \end{aligned}$$

□

As a direct application of Propositions A.4.2 and A.4.3, the conditional independence of the observations given the underlying sequence of states implies that for any integers  $p$  and  $p'$ , any indices  $k_1 < \dots < k_p$  and  $k'_1 < \dots < k'_{p'}$  such that  $\{k_1, \dots, k_p\} \cap \{k'_1, \dots, k'_{p'}\} = \emptyset$  and any function  $f \in \mathcal{F}_b(\mathcal{Y}^p)$ ,

$$\begin{aligned} \mathbb{E}_\nu [f(Y_{k_1}, \dots, Y_{k_p}) | X_{k_1}, \dots, X_{k_p}, X_{k'_1}, \dots, X_{k'_{p'}}, Y_{k'_1}, \dots, Y_{k'_{p'}}] \\ = \mathbb{E}_\nu [f(Y_{k_1}, \dots, Y_{k_p}) | X_{k_1}, \dots, X_{k_p}]. \end{aligned} \quad (2.18)$$

Indeed, in terms of conditional independence of the variables,

$$(Y_{k_1}, \dots, Y_{k_p}) \perp\!\!\!\perp (Y_{k'_1}, \dots, Y_{k'_{p'}}) | (X_{k_1}, \dots, X_{k_p}, X_{k'_1}, \dots, X_{k'_{p'}}) \quad [\mathbb{P}_\nu]$$

and

$$(Y_{k_1}, \dots, Y_{k_p}) \perp\!\!\!\perp (X_{k'_1}, \dots, X_{k'_{p'}}) | (X_{k_1}, \dots, X_{k_p}) \quad [\mathbb{P}_\nu].$$

Hence, by the contraction property of Proposition A.4.3,

$$(Y_{k_1}, \dots, Y_{k_p}) \perp\!\!\!\perp (X_{k'_1}, \dots, X_{k'_{p'}}, Y_{k'_1}, \dots, Y_{k'_{p'}}) | (X_{k_1}, \dots, X_{k_p}) \quad [\mathbb{P}_\nu],$$

which implies (2.18).

### 2.2.3 Hierarchical Hidden Markov Models

In examples such as 1.3.16 and 1.3.15, we met hidden Markov models whose state variable naturally decomposes into two distinct sub-components. To accommodate such structures, we define a specific sub-class of HMMs for which the state  $X_k$  consists of two components,  $X_k = (C_k, W_k)$ . This additional structure will be used to introduce a level of hierarchy in the state variables. We call this class *hierarchical hidden Markov models*. In general, the hierarchical structure will be as follows.

- $\{C_k\}_{k \geq 0}$  is a Markov chain on a state space  $(C, \mathcal{C})$  with transition kernel  $Q_C$  and initial distribution  $\nu_C$ . Thus, for any  $f \in \mathcal{F}_b(C)$  and any  $k \geq 1$ ,

$$E[f(C_k) | C_{0:k-1}] = Q_C(C_{k-1}, f) \quad \text{and} \quad E_{\nu_C}[f(C_0)] = \nu_C(f) .$$

- Conditionally on  $\{C_k\}_{k \geq 0}$ ,  $\{W_k\}_{k \geq 0}$  is a Markov chain on  $(W, \mathcal{W})$ . More precisely, there exists a transition kernel  $Q_W : (X \times C) \times \mathcal{W} \rightarrow [0, 1]$  such that for any  $k \geq 1$  and any function  $f \in \mathcal{F}_b(W)$ ,

$$E[f(W_k) | W_{0:k-1}, C_{0:k}] = Q_W[(W_{k-1}, C_k), f] .$$

In addition, there exists a transition kernel  $\nu_W : C \times \mathcal{W} \rightarrow [0, 1]$  such that for any  $f \in \mathcal{F}_b(W)$ ,

$$E[f(W_0) | C_0] = \nu_W(C_0, f) .$$

We denote by  $X_k = (C_k, W_k)$  the composite state variable. Then,  $\{X_k\}_{k \geq 0}$  is a Markov chain on  $X = C \times W$  with transition kernel

$$Q[(c, w), A \times B] = \int_A \int_B Q_C(c, dc') Q_W[(w, c'), dw'] , \quad A \in \mathcal{C}, B \in \mathcal{W} ,$$

and initial distribution

$$\nu(A \times B) = \int_A \nu_C(dc) \nu_W(c, B) .$$

As before, we assume that  $\{Y_k\}_{k \geq 0}$  is conditionally independent of  $\{X_k\}_{k \geq 0}$  and such that the conditional distribution of  $Y_n$  depends on  $X_n$  only, meaning that (2.17) holds.

The distinctive feature of hierarchical HMMs is that it is often advantageous to consider that the state variables are  $\{C_k\}_{k \geq 0}$  rather than  $\{X_k\}_{k \geq 0}$ . Of course, the model is then no longer an HMM because the observation  $Y_k$  depends on all partial states  $C_l$  for  $l \leq k$  due to the marginalization of the intermediate component  $W_l$  (for  $l = 0, \dots, k$ ). Nonetheless, this point of view is often preferable, particularly in cases where the structure of  $\{C_k\}_{k \geq 0}$  is very simple, such as when  $C$  is finite. The most common example of hierarchical HMM is the conditionally Gaussian linear state-space model (CGLSSM), which we already met in Examples 1.3.9, 1.3.11, and 1.3.16. We now formally define this model.

**Definition 2.2.6 (Conditionally Gaussian Linear State-Space Model).**  
*A CGLSSM is a model of the form*

$$\begin{aligned} W_{k+1} &= A(C_{k+1})W_k + R(C_{k+1})U_k, & W_0 &\sim N(\mu_\nu, \Sigma_\nu), \\ Y_k &= B(C_k)W_k + S(C_k)V_k, \end{aligned} \quad (2.19)$$

*subject to the following conditions.*

- *The indicator process  $\{C_k\}_{k \geq 0}$  is a Markov chain with transition kernel  $Q_C$  and initial distribution  $\nu_C$ . Usually,  $\mathcal{C}$  is finite and then identified with the set  $\{1, \dots, r\}$ .*
- *The state (or process) noise  $\{U_k\}_{k \geq 0}$  and the measurement noise  $\{V_k\}_{k \geq 0}$  are independent multivariate Gaussian white noises with zero mean and identity covariance matrices. In addition, the indicator process  $\{C_k\}_{k \geq 0}$  is independent of both the state noise and of the measurement noise.*
- *$A$ ,  $B$ ,  $R$ , and  $S$  are known matrix-valued functions of appropriate dimensions.*



<http://www.springer.com/978-0-387-40264-2>

Inference in Hidden Markov Models

Cappé, O.; Moulines, E.; Ryden, T.

2005, XVII, 653 p., Hardcover

ISBN: 978-0-387-40264-2