# Inference in Hidden Markov Models

Olivier Cappé, Eric Moulines and Tobias Rydén

June 17, 2009

# Contents

# Chapter 1

# 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.

## 1.1 Markov Chains

### 1.1.1 Transition Kernels

**Definition 1** (Transition Kernel)**.** *Let* $(\mathsf{X}, \mathcal{X})$ *and* $(\mathsf{Y}, \mathcal{Y})$ *be two measurable spaces. An* unnormalized transition kernel *from* $(\mathsf{X}, \mathcal{X})$ *to* $(\mathsf{Y}, \mathcal{Y})$ *is a function* $Q : \mathsf{X} \times \mathcal{Y} \to [0, \infty]$ *that satisfies*

*(i) for all* $x \in \mathsf{X}$, $Q(x, \cdot)$ *is a positive measure on* $(\mathsf{Y}, \mathcal{Y})$*;*

*(ii) for all* $A \in \mathcal{Y}$*, the function* $x \mapsto Q(x, A)$ *is measurable.*

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

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

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

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

*When* $\mathsf{X}$ *and* $\mathsf{Y}$ *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* $\mathsf{X}$ *and* $\mathsf{Y}$ *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 $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$ and from $(\mathsf{Y}, \mathcal{Y})$ to $(\mathsf{Z}, \mathcal{Z})$, respectively. The product $QR$, defined by

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

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

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

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

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

$$Q^{n+m}(x, A) = \int Q^n(x, dy)\, Q^m(y, A) \; . \tag{1.1}$$

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

$$q_n(x, y) = \int_{\mathsf{X}^{n-1}} q(x, x_1) \cdots q(x_{n-1}, y)\, \mu(dx_1) \cdots \mu(dx_{n-1}) \; . \tag{1.2}$$

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

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

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

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

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 $\mathsf{Y}$. The real measurable function $Qf$ on $\mathsf{X}$ is defined by

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

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 $\mu Qf$ instead of $(\mu Q)(f)$ or $\mu(Qf)$.

**Definition 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 $\mathsf{X} \times \mathsf{Y}$,*

$$\iint_{\mathsf{X}\times\mathsf{Y}} f(x,y)\,\nu(dx)Q(x,dy) = \iint_{\mathsf{X}\times\mathsf{Y}} f(x,y)\,\nu Q(dy)\,\overleftarrow{Q}_\nu(y,dx)\;. \tag{1.3}$$

The reverse kernel does not necessarily exist and is not uniquely defined. Nevertheless, if $\overleftarrow{Q}_{\nu,1}$ and $\overleftarrow{Q}_{\nu,2}$ satisfy (1.3), 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. 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

$$\overleftarrow{Q}_\nu(y,dx) = \frac{q(x,y)\,\nu(dx)}{\int_{\mathsf{X}} q(z,y)\,\nu(dz)}\;. \tag{1.4}$$

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 $\nu Q(y) \neq 0$,

$$\overleftarrow{Q}_\nu(y,x) = \frac{\nu(x)Q(x,y)}{\nu Q(y)}\;. \tag{1.5}$$

### 1.1.2 Homogeneous Markov Chains

Let $(\Omega, \mathcal{F}, \mathrm{P})$ 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; $(\Omega, \mathcal{F}, \mathbb{F}, \mathrm{P})$ is called a *filtered probability space*. For any filtration $\mathbb{F} = \{\mathcal{F}_n\}_{n\geq 0}$, we denote by $\mathcal{F}_\infty = \vee_{n=0}^{\infty} \mathcal{F}_n$ 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 3** (Markov Chain)**.** *Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathrm{P})$ be a filtered probability space and let $Q$ be a Markov transition kernel on a measurable space $(\mathsf{X}, \mathcal{X})$. An $\mathsf{X}$-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 $A \in \mathcal{X}$,*

$$\mathrm{P}\left(X_{k+1} \in A \,|\, \mathcal{F}_k\right) = Q(X_k, A)\;. \tag{1.6}$$

*The distribution of $X_0$ is called the* initial distribution *of the chain, and $\mathsf{X}$ 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 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 $\mathsf{X}^{(k+1)}$,*

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

In the following, we will use the generic notation $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Z})$ to denote the fact that $f$ is a measurable bounded function on $(\mathsf{Z},\mathcal{Z})$. In the case of Proposition 4 for instance, one considers functions $f$ that are in $\mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{(k+1)}\right)$. More generally, we will usually describe measures and transition kernels on $(\mathsf{Z},\mathcal{Z})$ by specifying the way they operate on the functions of $\mathcal{F}_{\mathrm{b}}(\mathsf{Z})$.

### Canonical Version

Let $(\mathsf{X},\mathcal{X})$ be a measurable space. The *canonical space* associated to $(\mathsf{X},\mathcal{X})$ is the infinite-dimensional product space $(\mathsf{X}^{\mathbb{N}},\mathcal{X}^{\otimes\mathbb{N}})$. The *coordinate process* is the $\mathsf{X}$-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 $\mathbb{F}^{\mathsf{X}}$ of the coordinate process.

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

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

The iterates of the shift operator are defined inductively by $\theta^0 = \mathrm{Id}$ (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 5.** *Let $(\mathsf{X},\mathcal{X})$ be a measurable set, $\nu$ a probability measure on $(\mathsf{X},\mathcal{X})$, and $Q$ a transition kernel on $(\mathsf{X},\mathcal{X})$. Then there exists an unique probability measure on $(\mathsf{X}^{\mathbb{N}},\mathcal{X}^{\otimes\mathbb{N}})$, denoted by $\mathrm{P}_\nu$, 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 $x\in\mathsf{X}$, let $\mathrm{P}_x$ be an alternative simplified notation for $\mathrm{P}_{\delta_x}$. Then for all $A\in\mathcal{X}^{\otimes\mathbb{N}}$, the mapping $x\to\mathrm{P}_x(A)=Q(x,A)$ is $\mathcal{X}$-measurable, and for any probability measure $\nu$ on $(\mathsf{X},\mathcal{X})$,*

$$\mathrm{P}_\nu(A) = \int \nu(dx)\;\mathrm{P}_x(A)\;. \tag{1.7}$$

The Markov chain defined in Theorem 5 is referred to as the *canonical version* of the Markov chain. The probability $\mathrm{P}_\nu$ defined on $(\mathsf{X}^{\mathbb{N}},\mathcal{X}^{\otimes\mathbb{N}})$ depends on $\nu$ and on the transition kernel $Q$. Nevertheless, the dependence with respect to $Q$ is traditionally omitted in the notation. The relation (1.7) implies that $x\to\mathrm{P}_x$ is a regular version of the conditional probability $\mathrm{P}_\nu\left(\cdot\,|\,X_k=x\right)$ in the sense that one can rewrite (1.6) as

$$\mathrm{P}_\nu\left(X_{k+1}\in A\,|\,\mathcal{F}_k^X\right) = \mathrm{P}_\nu\left(X_1\circ\theta^k\in A\,\middle|\,\mathcal{F}_k^X\right) = \mathrm{P}_{X_k}(X_1\in A)\qquad \mathrm{P}_\nu\text{-a.s.}$$

**Markov Properties**

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

$$\mathrm{E}_\nu[Y \circ \theta^k \,|\, \mathcal{F}_k^X] = \mathrm{E}_{X_k}[Y] \qquad \mathrm{P}_\nu\text{-a.s.} \tag{1.8}$$

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 $\tau : \Omega \to \bar{\mathbb{N}}$ 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 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 $\Psi$,*

$$\mathrm{E}_\nu[\mathbb{1}_{\{\tau<\infty\}} \Psi \circ \theta^\tau \,|\, \mathcal{F}_\tau^X] = \mathbb{1}_{\{\tau<\infty\}} \mathrm{E}_{X_\tau}[\Psi] \qquad \mathrm{P}_\nu\text{-a.s.} \tag{1.9}$$

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 $X_2 + X_3^2$.

### 1.1.3 Non-homogeneous Markov Chains

**Definition 7** (Non-homogeneous Markov Chain)**.** *Let $(\Omega, \mathcal{F}, \mathbb{F}, \mathrm{P})$ be a filtered probability space and let $\{Q_k\}_{k\geq 0}$ be a family of transition kernels on a measurable space $(\mathsf{X}, \mathcal{X})$. An $\mathsf{X}$-valued stochastic process $\{X_k\}_{k\geq 0}$ is said to be a* non-homogeneous Markov chain *under $\mathrm{P}$, 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 $A \in \mathcal{X}$,*

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

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 8** (Reverse Chain)**.** *Let $Q$ be a Markov kernel on some space $\mathsf{X}$, 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 $k = 0, 1, \ldots, n$ and transition kernels*

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

*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 $(\mathsf{X}, \mathcal{X})$, then $Q_k$ also admits a density with respect to the same measure $\mu$, namely

$$h_k(y,x) = \frac{\int q_{n-k-1}(z,x) q(x,y)\, \nu(dz)}{\int q_{n-k}(z,y)\, \nu(dz)} \,. \tag{1.10}$$

Here, $q_l$ is the transition density function of $Q^l$ with respect to $\mu$ as defined in (1.2). If the state space is countable, then

$$Q_k(y,x) = \frac{\nu Q^{n-k-1}(x) Q(x,y)}{\nu Q^{n-k}(y)} \,. \tag{1.11}$$

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

$$\iint f(x,x')\, \nu(dx)\, Q(x,dx') = \iint f(x,x')\, \nu(dx')\, Q(x',dx) \,. \tag{1.12}$$

The relation (1.12) is referred to as the *local balance equations* (or *detailed balance equations*). If the state space is countable, these equations hold if for all $x, x' \in \mathsf{X}$,

$$\nu(x) Q(x,x') = \nu(x') Q(x',x) \,. \tag{1.13}$$

Upon choosing a function $f$ that only depends on the second variable in (1.12), it is easily seen that $\nu Q(f) = \nu(f)$ for all functions $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$. 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 10** (Stationary Process)**.** *A stochastic process $\{X_k\}$ is said to be* stationary *(under $\mathrm{P}$) if its finite-dimensional distributions are translation invariant, that is, if for all $k, n \geq 1$ and all $n_1, \ldots, n_k$, the distribution of the random vector $(X_{n_1+n}, \ldots, X_{n_k+n})$ 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 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 7.

## 1.2 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.

### 1.2.1 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. It turns out that conditional independence is mathematically more difficult to define in general settings (in particular, when the state space $\mathsf{X}$ 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 15 below).

**Definition 11** (Hidden Markov Model)**.** *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces and let $Q$ and $G$ denote, respectively, a Markov transition kernel on $(\mathsf{X}, \mathcal{X})$ and a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Consider the Markov transition kernel defined on the product space $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ by*

$$T\left[(x, y), C\right] = \iint_C Q(x, dx')\, G(x', dy')\ , \qquad (x, y) \in \mathsf{X} \times \mathsf{Y},\, C \in \mathcal{X} \otimes \mathcal{Y}\ . \quad (1.14)$$

*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 $(\mathsf{X}, \mathcal{X})$, 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 $\mathrm{P}_\nu$ and $\mathrm{E}_\nu$ the probability measure and corresponding expectation associated with the process $\{X_k, Y_k\}_{k\geq 0}$ on the canonical space $\left((\mathsf{X} \times \mathsf{Y})^{\mathbb{N}}, (\mathcal{X} \otimes \mathcal{Y})^{\otimes \mathbb{N}}\right)$. Notice that this constitutes a slight departure from the Markov notations introduced previously, as $\nu$ is a probability measure on $\mathsf{X}$ only and not on the state space $\mathsf{X} \times \mathsf{Y}$ of the joint process. This slight abuse of notation is justified by the special structure of the model considered here. Equation (1.14) 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 $\mathsf{X}$ and $\mathsf{Y}$ are countable, the hidden Markov model is said to be *discrete*, which is the case originally considered by Baum and Petrie (1966).

**Definition 12** (Partially Dominated Hidden Markov Model)**.** *The model of Definition 11 is said to be partially dominated if there exists a probability measure $\mu$ on $(\mathsf{Y}, \mathcal{Y})$ such that for all $x \in \mathsf{X}$, $G(x, \cdot)$ 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*

$$T\left[(x, y), C\right] = \iint_C Q(x, dx') g(x', y')\, \mu(dy')\ C \in \mathcal{X} \otimes \mathcal{Y}\ . \quad (1.15)$$

In the third part of the book (Chapter 5 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 13** (Fully Dominated Hidden Markov Model)**.** *If, in addition to the requirements of Definition 12, there exists a probability measure $\lambda$ on $(\mathsf{X}, \mathcal{X})$ such that $\nu \ll \lambda$ and, for all $x \in \mathsf{X}$, $Q(x, \cdot) \ll \lambda(\cdot)$ 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*

$$t\left[(x, y), (x', y')\right] \stackrel{\text{def}}{=} q(x, x')g(x', y') \; . \tag{1.16}$$

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.

### 1.2.2 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 11.

**Proposition 14.** *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* (1.14)*. Then, for any integer $p$, any ordered set $\{k_1 < \cdots < k_p\}$ of indices and all functions $f_1, \ldots, f_p \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y})$,*

$$\mathrm{E}_\nu \left[ \prod_{i=1}^{p} f_i(Y_{k_i}) \,\middle|\, X_{k_1}, \ldots, X_{k_p} \right] = \prod_{i=1}^{p} \int_{\mathsf{Y}} f_i(y)\, G(X_{k_i}, dy) \; . \tag{1.17}$$

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

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

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

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

□

**Corollary 15.**

*(i) For any integer $p$ and any ordered set $\{k_1 < \cdots < k_p\}$ of indices, the random variables $Y_{k_1}, \ldots, Y_{k_p}$ are $\mathrm{P}_\nu$-conditionally independent given $(X_{k_1}, X_{k_2}, \ldots, X_{k_p})$.*

*(ii) For any integers $k$ and $p$ and any ordered set $\{k_1 < \cdots < k_p\}$ of indices such that $k \notin \{k_1, \ldots, k_p\}$, the random variables $Y_k$ and $(X_{k_1}, \ldots, X_{k_p})$ are $\mathrm{P}_\nu$-conditionally independent given $X_k$.*

*Proof.* Part (i) is an immediate consequence of Proposition 14. To prove (ii), note that for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y})$ and $h \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}^p)$,

$$\begin{aligned}
&\mathrm{E}_\nu\left[f(Y_k)h(X_{k_1}, \ldots, X_{k_p}) \,|\, X_k\right] \\
&\quad = \mathrm{E}_\nu\left[\mathrm{E}_\nu[f(Y_k) \,|\, X_{k_1}, \ldots, X_{k_p}, X_k]h(X_{k_1}, \ldots, X_{k_p}) \,\middle|\, X_k\right] \\
&\quad = \mathrm{E}_\nu\left[f(Y_k) \,|\, X_k\right] \mathrm{E}_\nu[h(X_{k_1}, \ldots, X_{k_p}) \,|\, X_k] \ .
\end{aligned}$$

□

The conditional independence of the observations given the underlying sequence of states implies that for any integers $p$ and $p'$, any indices $k_1 < \cdots < k_p$ and $k'_1 < \cdots < k'_{p'}$ such that $\{k_1, \ldots, k_p\} \cap \{k'_1, \ldots, k'_{p'}\} = \emptyset$ and any function $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y}^p)$,

$$\begin{aligned}
&\mathrm{E}_\nu[f(Y_{k_1}, \ldots, Y_{k_p}) \,|\, X_{k_1}, \ldots, X_{k_p}, X_{k'_1}, \ldots, X_{k'_p}, Y_{k'_1}, \ldots Y_{k'_p}] \\
&\qquad\qquad = \mathrm{E}_\nu[f(Y_{k_1}, \ldots, Y_{k_p}) \,|\, X_{k_1}, \ldots, X_{k_p}] \ . \quad (1.18)
\end{aligned}$$

Indeed, in terms of conditional independence of the variables,

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

and

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

Hence,

$$(Y_{k_1}, \ldots, Y_{k_p}) \perp\!\!\!\perp (X_{k'_1}, \ldots, X_{k'_{p'}}, Y_{k'_1}, \ldots, Y_{k'_p}) \,|\, (X_{k_1}, \ldots, X_{k_p}) \ [\mathrm{P}_\nu] \ ,$$

which implies (1.18).

# Part I

# State Inference

# Chapter 2

# Filtering and Smoothing Recursions

This chapter deals with a fundamental issue in hidden Markov modeling: given a fully specified model and some observations $Y_0, \ldots, Y_n$, what can be said about the corresponding unobserved state sequence $X_0, \ldots, X_n$? More specifically, we shall be concerned with the evaluation of the conditional distributions of the state at index $k$, $X_k$, given the observations $Y_0, \ldots, Y_n$, a task that is generally referred to as *smoothing*. There are of course several options available for tackling this problem (Anderson and Moore, 1979, Chapter 7) and we focus, in this chapter, on the *fixed-interval smoothing* paradigm in which $n$ is held fixed and it is desired to evaluate the conditional distributions of $X_k$ for all indices $k$ between 0 and $n$. Note that only the general mechanics of the smoothing problem are dealt with in this chapter. In particular, most formulas will involve integrals over $\mathsf{X}$. We shall not, for the moment, discuss ways in which these integrals can be effectively evaluated, or at least approximated, numerically.

The driving line of this chapter is the existence of a variety of smoothing approaches that involve a number of steps that only increase linearly with the number of observations. This is made possible by the fact (to be made precise in Section 2.3) that conditionally on the observations $Y_0, \ldots, Y_n$, the state sequence still is a Markov chain, albeit a non-homogeneous one.

From a historical perspective, it is interesting to recall that most of the early references on smoothing, which date back to the 1960s, focused on the specific case of Gaussian linear state-space models, following the pioneering work by Kalman and Bucy (1961). The classic book by Anderson and Moore (1979) on *optimal filtering*, for instance, is fully devoted to linear state-space models—see also Chapter 10 of the recent book by Kailath *et al.* (2000) for a more exhaustive set of early references on the smoothing problem. Although some authors such as (for instance) Ho and Lee (1964) considered more general state-space models, it is fair to say that the Gaussian linear state-space model was the dominant paradigm in the automatic control community[1]. In contrast, the work by Baum and his colleagues on hidden Markov models (Baum *et al.*, 1970) dealt with the case where the state space $\mathsf{X}$ of the hidden state is finite. These two streams of research (on Gaussian linear models and finite state space models) remained largely separated. Approximately at the same time, in the field of probability theory, the seminal work by Stratonovich (1960) stimulated a number of contributions that were to compose a body of work generally referred to

[1]Interestingly, until the early 1980s, the works that *did not* focus on the linear state-space model were usually advertised by the use of the words "Bayes" or "Bayesian" in their title—see, e.g., Ho and Lee (1964) or Askar and Derin (1981).

as *filtering theory*. The object of filtering theory is to study inference about partially observable Markovian processes in *continuous time*. A number of early references in this domain indeed consider some specific form of discrete state space continuous-time equivalent of the HMM (Shiryaev, 1966; Wonham, 1965)—see also Lipster and Shiryaev (2001), Chapter 9. Working in continuous time, however, implies the use of mathematical tools that are definitely more complex than those needed to tackle the discrete-time model of Baum *et al.* (1970). As a matter of fact, filtering theory and hidden Markov models evolved as two mostly independent fields of research. A poorly acknowledged fact is that the pioneering paper by Stratonovich (1960) (translated from an earlier Russian publication) describes, in its first section, an equivalent to the forward-backward smoothing approach of Baum *et al.* (1970). It turns out, however, that the formalism of Baum *et al.* (1970) generalizes well to models where the state space is *not* discrete anymore, in contrast to that of Stratonovich (1960).

## 2.1 Basic Notations and Definitions

### 2.1.1 Likelihood

The joint probability of the unobservable states and observations up to index $n$ is such that for any function $f \in \mathcal{F}_{\mathrm{b}}\left(\{\mathsf{X} \times \mathsf{Y}\}^{n+1}\right)$,

$$\mathrm{E}_\nu[f(X_0, Y_0, \ldots, X_n, Y_n)] = \int \cdots \int f(x_0, y_0, \ldots, x_n, y_n) \times \nu(dx_0) g(x_0, y_0) \prod_{k=1}^{n} \{Q(x_{k-1}, dx_k) g(x_k, y_k)\} \, \mu_n(dy_0, \ldots, dy_n) \;, \quad (2.1)$$

where $\mu_n$ denotes the product distribution $\mu^{\otimes(n+1)}$ on $(\mathsf{Y}^{n+1}, \mathcal{Y}^{\otimes(n+1)})$. Marginalizing with respect to the unobservable variables $X_0, \ldots, X_n$, one obtains the marginal distribution of the observations only,

$$\mathrm{E}_\nu[f(Y_0, \ldots, Y_n)] = \int \cdots \int f(y_0, \ldots, y_n) \, \mathrm{L}_{\nu,n}(y_0, \ldots, y_n) \, \mu_n(dy_0, \ldots, dy_n) \;, \quad (2.2)$$

where $\mathrm{L}_{\nu,n}$ is an important quantity which we define below for future reference.

**Definition 16** (Likelihood)**.** *The* likelihood *of the observations is the probability density function of $Y_0, Y_1, \ldots, Y_n$ with respect to $\mu_n$ defined, for all $(y_0, \ldots, y_n) \in \mathsf{Y}^{n+1}$, by*

$$\mathrm{L}_{\nu,n}(y_0, \ldots, y_n) = \int \cdots \int \nu(dx_0) g(x_0, y_0) Q(x_0, dx_1) g(x_1, y_1) \cdots Q(x_{n-1}, dx_n) g(x_n, y_n) \;. \quad (2.3)$$

*In addition,*

$$\ell_{\nu,n} \stackrel{\mathrm{def}}{=} \log \mathrm{L}_{\nu,n} \;, \quad (2.4)$$

*is referred to as the* log-likelihood function*.*

**Remark 17** (Concise Notation for Sub-sequences)**.** For the sake of conciseness, we will use in the following the notation $Y_{l:m}$ to denote the collection of consecutively indexed variables $Y_l, \ldots, Y_m$ wherever possible (proceeding the same way for the unobservable sequence $\{X_k\}$). In quoting (2.3) for instance, we shall write $\mathrm{L}_{\nu,n}(y_{0:n})$ rather than $\mathrm{L}_{\nu,n}(y_0, \ldots, y_n)$. By transparent convention, $Y_{k:k}$ refers to the single variable $Y_k$, although the second notation $(Y_k)$ is to be preferred in this particular

case. In systematic expressions, however, it may be helpful to understand $Y_{k:k}$ as a valid replacement of $Y_k$. For similar reasons, we shall, when needed, accept $Y_{k+1:k}$ as a valid empty set. The latter convention should easily be recalled by programmers, as instructions of the form "for i equals k+1 to k, do..." , which do nothing, constitute a well-accepted ingredient of most programming idioms.

### 2.1.2 Smoothing

We first define generically what is meant by the word *smoothing* before deriving the basic results that form the core of the techniques discussed in the rest of the chapter.

**Definition 18** (Smoothing, Filtering, Prediction)**.** *For positive indices $k$, $l$, and $n$ with $l \geq k$, denote by $\phi_{\nu,k:l|n}$ the conditional distribution of $X_{k:l}$ given $Y_{0:n}$, that is*

*(a) $\phi_{\nu,k:l|n}$ is a transition kernel from $\mathsf{Y}^{(n+1)}$ to $\mathsf{X}^{(l-k+1)}$:*

- *for any given set $A \in \mathcal{X}^{\otimes(l-k+1)}$, $y_{0:n} \mapsto \phi_{\nu,k:l|n}(y_{0:n}, A)$ is a $\mathcal{Y}^{\otimes(n+1)}$-measurable function,*
- *for any given sub-sequence $y_{0:n}$, $A \mapsto \phi_{\nu,k:l|n}(y_{0:n}, A)$ is a probability distribution on $(\mathsf{X}^{l-k+1}, \mathcal{X}^{\otimes(l-k+1)})$.*

*(b) $\phi_{\nu,k:l|n}$ satisfies, for any function $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{l-k+1}\right)$,*

$$\mathrm{E}_\nu\left[f(X_{k:l}) \,|\, Y_{0:n}\right] = \int \cdots \int f(x_{k:l})\, \phi_{\nu,k:l|n}(Y_{0:n}, dx_{k:l}) \;,$$

*where the equality holds $\mathrm{P}_\nu$-almost surely. Specific choices of $k$ and $l$ give rise to several particular cases of interest:*

**Joint Smoothing:** *$\phi_{\nu,0:n|n}$, for $n \geq 0$;*

**(Marginal) Smoothing:** *$\phi_{\nu,k|n}$ for $n \geq k \geq 0$;*

**Prediction:** *$\phi_{\nu,n+1|n}$ for $n \geq 0$; In describing algorithms, it will be convenient to extend our notation to use $\phi_{\nu,0|-1}$ as a synonym for the initial distribution $\nu$;*

**$p$-step Prediction:** *$\phi_{\nu,n+p|n}$ for $n, p \geq 0$.*

**Filtering:** *$\phi_{\nu,n|n}$ for $n \geq 0$; Because the use of filtering will be preeminent in the following, we shall most often abbreviate $\phi_{\nu,n|n}$ to $\phi_{\nu,n}$.*

In more precise terms, $\phi_{\nu,k:l|n}$ is *a version* of the conditional distribution of $X_{k:l}$ given $Y_{0:n}$. It is however not obvious that such a quantity indeed exists in great generality. The proposition below complements Definition 18 by a constructive approach to defining the smoothing quantities from the elements of the hidden Markov model.

**Proposition 19.** *Consider a hidden Markov model compatible with Definition 12, let $n$ be a positive integer and $y_{0:n} \in \mathsf{Y}^{n+1}$ a sub-sequence such that $\mathrm{L}_{\nu,n}(y_{0:n}) > 0$. The joint smoothing distribution $\phi_{\nu,0:n|n}$ then satisfies*

$$\phi_{\nu,0:n|n}(y_{0:n}, f) = \mathrm{L}_{\nu,n}(y_{0:n})^{-1} \int \cdots \int f(x_{0:n}) \times \nu(dx_0) g(x_0, y_0) \prod_{k=1}^{n} Q(x_{k-1}, dx_k) g(x_k, y_k) \quad (2.5)$$

*for all functions* $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+1}\right)$. *Likewise, for indices* $p \geq 0$,

$$\phi_{\nu,0:n+p|n}(y_{0:n}, f) = \int \cdots \int f(x_{0:n+p}) \times \phi_{\nu,0:n|n}(y_{0:n}, dx_{0:n}) \prod_{k=n+1}^{n+p} Q(x_{k-1}, dx_k) \quad (2.6)$$

*for all functions* $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+p+1}\right)$.

*Proof.* Equation (2.5) defines $\phi_{\nu,0:n|n}$ in a way that obviously satisfies part (a) of Definition 18. To prove the (b) part of the definition, consider a function $h \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{Y}^{n+1}\right)$. By (2.1),

$$\mathrm{E}_\nu[h(Y_{0:n})f(X_{0:n})] = \int \cdots \int h(y_{0:n})f(x_{0:n}) \times \nu(dx_0)g(x_0, y_0) \left[\prod_{k=1}^{n} Q(x_{k-1}, dx_k)g(x_k, y_k)\right] \mu_n(dy_{0:n}) \,.$$

Using Definition 16 of the likelihood $\mathrm{L}_{\nu,n}$ and (2.5) for $\phi_{\nu,0:n|n}$ yields

$$\begin{aligned} \mathrm{E}_\nu[h(Y_{0:n})f(X_{0:n})] &= \int \cdots \int h(y_{0:n})\, \phi_{\nu,0:n|n}(y_{0:n}, f)\mathrm{L}_{\nu,n}(y_{0:n})\, \mu_n(dy_{0:n}) \\ &= \mathrm{E}_\nu[h(Y_{0:n})\phi_{\nu,0:n|n}(Y_{0:n}, f)] \,. \end{aligned} \quad (2.7)$$

Hence $\mathrm{E}_\nu[f(X_{0:n}) \,|\, Y_{0:n}]$ equals $\phi_{\nu,0:n|n}(Y_{0:n}, f)$, $\mathrm{P}_\nu$-a.e., for any function $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+1}\right)$.

For (2.6), proceed similarly and consider two functions $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+p+1}\right)$ and $h \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{Y}^{n+1}\right)$. First apply (2.1) to obtain

$$\begin{aligned} \mathrm{E}_\nu[h(Y_{0:n})f(X_{0:n+p})] = \int \cdots \int f(x_{0:n+p}) \\ \times \nu(dx_0)g(x_0, y_0) \left[\prod_{k=1}^{n} Q(x_{k-1}, dx_k)g(x_k, y_k)\right] h(y_{0:n}) \\ \times \left[\prod_{l=n+1}^{n+p} Q(x_{l-1}, dx_l)g(x_l, y_l)\right] \mu_{n+p}(dy_{0:n+p}) \,. \end{aligned}$$

When integrating with respect to the subsequence $y_{n+1:n+p}$, the third line of the previous equation reduces to $\prod_{l=n+1}^{n+p} Q(x_{l-1}, dx_l)\mu_n(dy_{0:n})$. Finally use (2.3) and (2.5) to obtain

$$\begin{aligned} \mathrm{E}_\nu[h(Y_{0:n})f(X_{0:n+p})] = \int \cdots \int h(y_{0:n})f(x_{0:n+p}) \\ \phi_{\nu,0:n|n}(y_{0:n}, dx_{0:n}) \left[\prod_{k=n+1}^{n+p} Q(x_{k-1}, dx_k)\right] \mathrm{L}_{\nu,n}(y_{0:n})\mu_n(dy_{0:n}) \,, \end{aligned} \quad (2.8)$$

which concludes the proof. □

Proposition 19 also implicitly defines all other particular cases of smoothing kernels mentioned in Definition 18, as these are obtained by marginalization. For instance, the marginal smoothing kernel $\phi_{\nu,k|n}$ for $0 \leq k \leq n$ is such that for any $y_{0:n} \in \mathsf{Y}^{n+1}$ and $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$\phi_{\nu,k|n}(y_{0:n}, f) \stackrel{\text{def}}{=} \int \cdots \int f(x_k)\, \phi_{\nu,0:n|n}(y_{0:n}, dx_{0:n}) \,, \quad (2.9)$$

where $\phi_{\nu,0:n|n}$ is defined by (2.5).

Likewise, for any given $y_{0:n} \in \mathsf{Y}^{n+1}$, the $p$-step predictive distribution $\phi_{\nu,n+p|n}(y_{0:n}, \cdot)$ may be obtained by marginalization of the joint distribution $\phi_{\nu,0:n+p|n}(y_{0:n}, \cdot)$ with respect to all variables $x_k$ except the last one (the one with index $k = n + p$). A closer examination of (2.6) together with the use of the Chapman-Kolmogorov equations introduced in (1.1) (cf. Chapter 7) directly shows that $\phi_{\nu,n+p|n}(y_{0:n}, \cdot) = \phi_{\nu,n}(y_{0:n}, \cdot)Q^p$, where $\phi_{\nu,n}$ refers to the filter (conditional distribution of $X_n$ given $Y_{0:n}$).

### 2.1.3 The Forward-Backward Decomposition

Replacing $\phi_{\nu,0:n|n}$ in (2.9) by its expression given in (2.5) shows that it is always possible to rewrite $\phi_{\nu,k|n}(y_{0:n}, f)$, for functions $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, as

$$\phi_{\nu,k|n}(y_{0:n}, f) = \mathrm{L}_{\nu,n}(y_{0:n})^{-1} \int f(x)\, \alpha_{\nu,k}(y_{0:k}, dx) \beta_{k|n}(y_{k+1:n}, x)\ , \tag{2.10}$$

where $\alpha_{\nu,k}$ and $\beta_{k|n}$ are defined below in (2.11) and (2.12), respectively. In simple terms, $\alpha_{\nu,k}$ correspond to the factors in the multiple integral that are to be integrated with respect to the state variables $x_l$ with indices $l \leq k$ while $\beta_{k|n}$ gathers the remaining factors (which are to be integrated with respect to $x_l$ for $l > k$). This simple splitting of the multiple integration in (2.9) constitutes the forward-backward decomposition.

**Definition 20** (Forward-Backward "Variables")**.** *For $k \in \{0, \ldots, n\}$, define the following quantities.*
**Forward Kernel** *$\alpha_{\nu,k}$ is the non-negative finite kernel from $(\mathsf{Y}^{k+1}, \mathcal{Y}^{\otimes(k+1)})$ to $(\mathsf{X}, \mathcal{X})$ such that*

$$\alpha_{\nu,k}(y_{0:k}, f) = \int \cdots \int f(x_k)\, \nu(dx_0) g(x_0, y_0) \prod_{l=1}^{k} Q(x_{l-1}, dx_l) g(x_l, y_l)\ , \tag{2.11}$$

*with the convention that the rightmost product term is empty for $k = 0$.*
**Backward Function** *$\beta_{k|n}$ is the non-negative measurable function on $\mathsf{Y}^{n-k} \times \mathsf{X}$ defined by*

$$\beta_{k|n}(y_{k+1:n}, x) = \int \cdots \int Q(x, dx_{k+1}) g(x_{k+1}, y_{k+1}) \prod_{l=k+2}^{n} Q(x_{l-1}, dx_l) g(x_l, y_l)\ , \tag{2.12}$$

*for $k \leq n - 1$ (with the same convention that the rightmost product is empty for $k = n - 1$); $\beta_{n|n}(\cdot)$ is set to the constant function equal to 1 on $\mathsf{X}$.*

The term "forward and backward variables" as well as the use of the symbols $\alpha$ and $\beta$ is part of the HMM credo and dates back to the seminal work of Baum and his colleagues (Baum *et al.*, 1970, p. 168). It is clear however that for a general model as given in Definition 12, these quantities as defined in (2.11) and (2.12) are very different in nature, and indeed sufficiently so to prevent the use of the loosely defined term "variable". In the original framework studied by Baum and his coauthors where $\mathsf{X}$ is a finite set, both the forward measures $\alpha_{\nu,k}(y_{0:k}, \cdot)$ and the backward functions $\beta_{k|n}(y_{k+1:n}, \cdot)$ can be represented by vectors with non-negative entries. Indeed, in this case $\alpha_{\nu,k}(y_{0:k}, x)$ has the interpretation $\mathrm{P}_\nu(Y_0 = y_0, \ldots, Y_k = y_k, X_k = x)$ while $\beta_{k|n}(y_{k+1:n}, x)$ has the interpretation $\mathrm{P}(Y_{k+1} = y_{k+1}, \ldots, Y_n = y_n \,|\, X_k = x)$. This way of thinking of $\alpha_{\nu,k}$ and $\beta_{k|n}$ may be extended to general state

spaces: $\alpha_{\nu,k}(y_{0:k}, dx)$ is then the joint density (with respect to $\mu_{k+1}$) of $Y_0, \ldots, Y_k$ and distribution of $X_k$, while $\beta_{k|n}(y_{k+1:n}, x)$ is the conditional joint density (with respect to $\mu_{n-k}$) of $Y_{k+1}, \ldots, Y_n$ given $X_k = x$. Obviously, these entities may then not be represented as vectors of finite length, as when $\mathsf{X}$ is finite; this situation is the exception rather than the rule.

Let us simply remark at this point that while the forward kernel at index $k$ is defined irrespectively of the length $n$ of the observation sequence (as long as $n \geq k$), the same is not true for the backward functions. The sequence of backward functions clearly depends on the index where the observation sequence stops. In general, for instance, $\beta_{k|n-1}$ differs from $\beta_{k|n}$ even if we assume that the same sub-observation sequence $y_{0:n-1}$ is considered in both cases. This is the reason for adding the terminal index $n$ to the notation used for the backward functions. This notation also constitutes a departure from HMM traditions in which the backward functions are simply indexed by $k$. For $\alpha_{\nu,k}$, the situation is closer to standard practice and we simply add the subscript $\nu$ to recall that the forward kernel $\alpha_{\nu,k}$, in contrast with the backward measure, does depend on the distribution $\nu$ postulated for the initial state $X_0$.

### 2.1.4 Implicit Conditioning (Please Read This Section!)

We now pause to introduce a convention that will greatly simplify the exposition of the material contained in the first part of the book (from this chapter on, starting with the next section), both from terminological and notational points of view. This convention would however generate an acute confusion in the mind of a hypothetical reader who, having read Chapter 2 up to now, would decide to skip our friendly encouragement to read what follows carefully.

In the rest of Part I (with the notable exception of Section 3), we focus on the evaluation of quantities such as $\phi_{\nu,0:n|n}$ or $\phi_{\nu,k|n}$ for a given value of the observation sequence $y_{0:n}$. In this context, we *expunge from our notations the fact that all quantities depend on* $y_{0:n}$. In particular, we rewrite (2.5) for any $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{n+1}\right)$ more concisely as

$$\phi_{\nu,0:n|n}(f) = \mathrm{L}_{\nu,n}^{-1} \int \cdots \int f(x_{0:n})\, \nu(dx_0) g_0(x_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i) g_i(x_i)\ , \qquad (2.13)$$

where $g_k$ are the data-dependent functions on $\mathsf{X}$ defined by $g_k(x) \stackrel{\text{def}}{=} g(x, y_k)$ for the particular sequence $y_{0:n}$ under consideration. The sequence of functions $\{g_k\}$ is about the only new notation that is needed as we simply re-use the previously defined quantities omitting their explicit dependence on the observations. For instance, in addition to writing $\mathrm{L}_{\nu,n}$ instead of $\mathrm{L}_{\nu,n}(y_{0:n})$, we will also use $\phi_n(\cdot)$ rather than $\phi_n(y_{0:n}, \cdot)$, $\beta_{k|n}(\cdot)$ rather than $\beta_{k|n}(y_{k+1:n}, \cdot)$, etc. This notational simplification implies a corresponding terminological adjustment. For instance, $\alpha_{\nu,k}$ will be referred to as the *forward measure* at index $k$ and considered as a positive finite measure on $(\mathsf{X}, \mathcal{X})$. In all cases, the conversion should be easy to do mentally, as in the case of $\alpha_{\nu,k}$, for instance, what is meant is really "the measure $\alpha_{\nu,k}(y_{0:k}, \cdot)$, for a particular value of $y_{0:k} \in \mathsf{Y}^{k+1}$".

At first sight, omitting the observations may seem a weird thing to do in a statistically oriented book. However, for *posterior state inference* in HMMs, one indeed works conditionally on a given fixed sequence of observations. Omitting the observations from our notation will thus allow more concise expressions in most parts of the book. There are of course some properties of the hidden Markov model for which dependence with respect to the distribution of the observations does matter (hopefully!) This is in particular the case of Section 3 on forgetting

and Chapter 6, which deals with statistical properties of the estimates for which we will make the dependence with respect to the observations explicit.

## 2.2 Forward-Backward

The forward-backward decomposition introduced in Section 2.1.3 is just a rewriting of the multiple integral in (2.9) such that for $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$\phi_{\nu,k|n}(f) = \mathrm{L}_{\nu,n}^{-1} \int f(x)\, \alpha_{\nu,k}(dx)\beta_{k|n}(x)\ , \tag{2.14}$$

where

$$\alpha_{\nu,k}(f) = \int \cdots \int f(x_k)\, \nu(dx_0) g_0(x_0) \prod_{l=1}^{k} Q(x_{l-1}, dx_l) g_l(x_l) \tag{2.15}$$

and

$$\beta_{k|n}(x) = \int \cdots \int Q(x, dx_{k+1}) g_{k+1}(x_{k+1}) \prod_{l=k+2}^{n} Q(x_{l-1}, dx_l) g_l(x_l)\ . \tag{2.16}$$

The last expression is, by convention, equal to 1 for the final index $k = n$. Note that we are now using the implicit conditioning convention discussed in the previous section.

### 2.2.1 The Forward-Backward Recursions

The point of using the forward-backward decomposition for the smoothing problem is that both the forward measures $\alpha_{\nu,k}$ and the backward functions $\beta_{k|n}$ can be expressed *recursively* rather than by their integral representations (2.15) and (2.1.4). This is the essence of the *forward-backward algorithm* proposed by Baum *et al.* (1970, p. 168), which we now describe.

**Proposition 21** (Forward-Backward Recursions)**.** *The forward measures defined by* (2.15) *may be obtained, for all $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, recursively for $k = 1, \ldots, n$ according to*

$$\alpha_{\nu,k}(f) = \int f(x') \int \alpha_{\nu,k-1}(dx) Q(x, dx') g_k(x') \tag{2.17}$$

*with initial condition*

$$\alpha_{\nu,0}(f) = \int f(x) g_0(x)\, \nu(dx)\ . \tag{2.18}$$

*Similarly, the backward functions defined by* (2.16) *may be obtained, for all $x \in \mathsf{X}$, by the recursion*

$$\beta_{k|n}(x) = \int Q(x, dx') g_{k+1}(x') \beta_{k+1|n}(x') \tag{2.19}$$

*operating on decreasing indices $k = n - 1$ down to 0; the initial condition is*

$$\beta_{n|n}(x) = 1\ . \tag{2.20}$$

*Proof.* The proof of this result is straightforward and similar for both recursions. For $\alpha_{\nu,k}$ for instance, simply rewrite (2.15) as

$$\alpha_{\nu,k}(f) = \int_{x_k \in \mathsf{X}} f(x_k) \int_{x_{k-1} \in \mathsf{X}} \left[ \int \cdots \int_{x_0 \in \mathsf{X}, \ldots, x_{k-2} \in \mathsf{X}} \nu(dx_0) g_0(x_0) \prod_{l=1}^{k-1} Q(x_{l-1}, dx_l) g_l(x_l) \right] Q(x_{k-1}, dx_k) g_k(x_k) \ ,$$

where the term in brackets is recognized as $\alpha_{\nu,k-1}(dx_{k-1})$. □

**Remark 22** (Concise Markov Chain Notations)**.** In the following, we shall often quote the above results using the concise Markov chain notations introduced in Chapter 1. For instance, instead of (2.17) and (2.19) one could write more simply $\alpha_{\nu,k}(f) = \alpha_{\nu,k-1} Q(f g_k)$ and $\beta_{k|n} = Q(g_{k+1} \beta_{k+1|n})$. Likewise, the decomposition (2.14) may be rewritten as

$$\phi_{\nu,k|n}(f) = \mathrm{L}_{\nu,n}^{-1} \alpha_{\nu,k} \left( f \beta_{k|n} \right) \ .$$

The main shortcoming of the forward-backward representation is that the quantities $\alpha_{\nu,k}$ and $\beta_{k|n}$ do not have an immediate probabilistic interpretation. Recall, in particular, that the first one is a finite (positive) measure but certainly not a probability measure, as $\alpha_{\nu,k}(1) \neq 1$ (in general). There is however an important solidarity result between the forward and backward quantities $\alpha_{\nu,k}$ and $\beta_{k|n}$, which is summarized by the following proposition.

**Proposition 23.** *For all indices $k \in \{0, \ldots, n\}$,*

$$\alpha_{\nu,k}(\beta_{k|n}) = \mathrm{L}_{\nu,n}$$

*and*

$$\alpha_{\nu,k}(1) = \mathrm{L}_{\nu,k} \ ,$$

*where $\mathrm{L}_{\nu,k}$ refers to the likelihood of the observations up to index $k$ (included) only, under $\mathrm{P}_\nu$.*

*Proof.* Because (2.14) must hold in particular for $f = 1$ and the marginal smoothing distribution $\phi_{\nu,k|n}$ is a probability measure,

$$\phi_{\nu,k|n}(1) \stackrel{\text{def}}{=} 1 = \mathrm{L}_{\nu,n}^{-1} \alpha_{\nu,k} \left( \beta_{k|n} \right) \ .$$

For the final index $k = n$, $\beta_{n|n}$ is the constant function equal to 1 and hence $\alpha_{\nu,n}(1) = \mathrm{L}_{\nu,n}$. This observation is however not specific to the final index $n$, as $\alpha_{\nu,k}$ only depends on the observations up to index $k$ and thus any particular index may be selected as a potential final index (in contrast to what happens for the backward functions). □

### 2.2.2 Filtering and Normalized Recursion

The forward and backward quantities $\alpha_{\nu,k}$ and $\beta_{k|n}$, as defined in previous sections, are unnormalized in the sense that their scales are largely unknown. On the other hand, we know that $\alpha_{\nu,k}(\beta_{k|n})$ is equal to $\mathrm{L}_{\nu,n}$, the likelihood of the observations up to index $n$ under $\mathrm{P}_\nu$.

The long-term behavior of the likelihood $\mathrm{L}_{\nu,n}$, or rather its logarithm, is a result known as the asymptotic equipartition property, or AEP (Cover and Thomas, 1991)

in the information theoretic literature and as theShannon-McMillan-Breiman theorem in the statistical literature. For HMMs, Proposition 112 (Chapter 6) shows that under suitable mixing conditions on the underlying unobservable chain $\{X_k\}_{k\geq 0}$, the AEP holds in that $n^{-1}\log \mathrm{L}_{\nu,n}$ converges $\mathrm{P}_\nu$-a.s. to a limit as $n$ tends to infinity. The likelihood $\mathrm{L}_{\nu,n}$ will thus either grow to infinity or shrink to zero, depending on the sign of the limit, exponentially fast in $n$.

The famous tutorial by Rabiner (1989) coined the term *scaling* to describe a practical solution to this problem. Interestingly, scaling also partly answers the question of the probabilistic interpretation of the forward and backward quantities.

Scaling as described by Rabiner (1989) amounts to normalizing $\alpha_{\nu,k}$ and $\beta_{k|n}$ by positive real numbers to keep the numeric values needed to represent $\alpha_{\nu,k}$ and $\beta_{k|n}$ within reasonable bounds. There are clearly a variety of options available, especially if one replaces (2.14) by the equivalent auto-normalized form

$$\phi_{\nu,k|n}(f) = [\alpha_{\nu,k}(\beta_{k|n})]^{-1} \int \alpha_{\nu,k}(f\beta_{k|n}) \;, \tag{2.21}$$

assuming that $\alpha_{\nu,k}(\beta_{k|n})$ is indeed finite and non-zero.

In our view, the most natural scaling scheme (developed below) consists in replacing the measure $\alpha_{\nu,k}$ and the function $\beta_{k|n}$ by scaled versions $\bar{\alpha}_{\nu,k}$ and $\bar{\beta}_{k|n}$ of these quantities, satisfying both

(i) $\bar{\alpha}_{\nu,k}(1) = 1$, and

(ii) $\bar{\alpha}_{\nu,k}(\bar{\beta}_{k|n}) = 1$.

Item (i) implies that the normalized forward measures $\bar{\alpha}_{\nu,k}$ are probability measures that have a probabilistic interpretation given below. Item (ii) implies that the normalized backward functions are such that $\phi_{\nu,k|n}(f) = \int f(x)\bar{\beta}_{k|n}(x)\,\bar{\alpha}_{\nu,k}(dx)$ for all $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, without the need for a further renormalization. We note that this scaling scheme differs slightly from the one described by Rabiner (1989).

To derive the probabilistic interpretation of $\bar{\alpha}_{\nu,k}$, observe that (2.14) and Proposition 23, instantiated for the final index $k = n$, imply that the filtering distribution $\phi_{\nu,n}$ at index $n$ (recall that $\phi_{\nu,n}$ is used as a simplified notation for $\phi_{\nu,n|n}$) may be written $[\alpha_{\nu,n}(1)]^{-1}\alpha_{\nu,n}$. This finding is of course not specific to the choice of the index $n$ as already discussed when proving the second statement of Proposition 23. Thus, the normalized version $\bar{\alpha}_{\nu,k}$ of the forward measure $\alpha_{\nu,k}$ coincides with the filtering distribution $\phi_{\nu,k}$ introduced in Definition 18. This observation together with Proposition 23 implies that there is a unique choice of scaling scheme that satisfies the two requirements of the previous paragraph, as

$$\begin{aligned}\int f(x)\,\phi_{\nu,k|n}(dx) &= \mathrm{L}_{\nu,n}^{-1} \int f(x)\,\alpha_{\nu,k}(dx)\beta_{k|n}(x) \\ &= \int f(x) \underbrace{\mathrm{L}_{\nu,k}^{-1}\alpha_{\nu,k}(dx)}_{\bar{\alpha}_{\nu,k}(dx)} \underbrace{\mathrm{L}_{\nu,n}^{-1}\mathrm{L}_{\nu,k}\beta_{k|n}(x)}_{\bar{\beta}_{k|n}(x)}\end{aligned}$$

must hold for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$. The following definition summarizes these conclusions, using the notation $\phi_{\nu,k}$ rather than $\bar{\alpha}_{\nu,k}$, as these two definitions refer to the same object—the filtering distribution at index $k$.

**Definition 24** (Normalized Forward-Backward Variables)**.** *For $k \in \{0,\ldots,n\}$, the normalized forward measure $\bar{\alpha}_{\nu,k}$ coincides with the filtering distribution $\phi_{\nu,k}$ and satisfies*

$$\phi_{\nu,k} = [\alpha_{\nu,k}(1)]^{-1}\alpha_{\nu,k} = \mathrm{L}_{\nu,k}^{-1}\alpha_{\nu,k} \;.$$

*The normalized backward functions $\bar{\beta}_{k|n}$ are defined by*

$$\bar{\beta}_{k|n} = \frac{\alpha_{\nu,k}(1)}{\alpha_{\nu,k}(\beta_{k|n})}\beta_{k|n} = \frac{\mathrm{L}_{\nu,k}}{\mathrm{L}_{\nu,n}}\beta_{k|n} \ .$$

The above definition would be pointless if computing $\alpha_{\nu,k}$ and $\beta_{k|n}$ was indeed necessary to obtain the normalized variables $\phi_{\nu,k}$ and $\bar{\beta}_{k|n}$. The following result shows that this is not the case.

**Proposition 25** (Normalized Forward-Backward Recursions)**. Forward Filtering Recursion** *The filtering measures may be obtained, for all $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, recursively for $k = 1, \ldots, n$ according to*

$$\begin{aligned} c_{\nu,k} &= \int\int \phi_{\nu,k-1}(dx)Q(x,dx')g_k(x') \ , \\ \phi_{\nu,k}(f) &= c_{\nu,k}^{-1} \int f(x') \int \phi_{\nu,k-1}(dx)Q(x,dx')g_k(x') \ , \end{aligned} \tag{2.22}$$

*with initial condition*

$$\begin{aligned} c_{\nu,0} &= \int g_0(x)\nu(dx) \ , \\ \phi_{\nu,0}(f) &= c_{\nu,0}^{-1} \int f(x)g_0(x)\,\nu(dx) \ . \end{aligned}$$

**Normalized Backward Recursion** *The normalized backward functions may be obtained, for all $x \in \mathsf{X}$, by the recursion*

$$\bar{\beta}_{k|n}(x) = c_{\nu,k+1}^{-1} \int Q(x,dx')g_{k+1}(x')\bar{\beta}_{k+1|n}(x') \tag{2.23}$$

*operating on decreasing indices $k = n-1$ down to 0; the initial condition is $\bar{\beta}_{n|n}(x) = 1$.*

*Once the two recursions above have been carried out, the smoothing distribution at any given index $k \in \{0, \ldots, n\}$ is available via*

$$\phi_{\nu,k|n}(f) = \int f(x)\ \bar{\beta}_{k|n}(x)\phi_{\nu,k}(dx) \tag{2.24}$$

*for all $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$.*

*Proof.* Proceeding by forward induction for $\phi_{\nu,k}$ and backward induction for $\beta_{k|n}$, it is easily checked from (2.22) and (2.23) that

$$\phi_{\nu,k} = \left(\prod_{l=0}^{k} c_{\nu,l}\right)^{-1} \alpha_{\nu,k} \qquad \text{and} \qquad \bar{\beta}_{k|n} = \left(\prod_{l=k+1}^{n} c_{\nu,l}\right)^{-1} \beta_{k|n} \ . \tag{2.25}$$

Because $\phi_{\nu,k}$ is normalized,

$$\phi_{\nu,k}(1) \stackrel{\mathrm{def}}{=} 1 = \left(\prod_{l=0}^{k} c_{\nu,l}\right)^{-1} \alpha_{\nu,k}(1) \ .$$

Proposition 23 then implies that for any integer $k$,

$$\mathrm{L}_{\nu,k} = \prod_{l=0}^{k} c_{\nu,l} \ . \tag{2.26}$$

In other words, $c_{\nu,0} = \mathrm{L}_{\nu,0}$ and for subsequent indices $k \geq 1$, $c_{\nu,k} = \mathrm{L}_{\nu,k}/\mathrm{L}_{\nu,k-1}$. Hence (2.25) coincides with the normalized forward and backward variables as specified by Definition 24. □

We now pause to state a series of remarkable consequences of Proposition 25.

**Remark 26.** The forward recursion in (2.22) may also be rewritten to highlight a two-step procedure involving both the predictive and filtering measures. Recall our convention that $\phi_{\nu,0|-1}$ refers to the predictive distribution of $X_0$ when no observation is available and is thus an alias for $\nu$, the distribution of $X_0$. For $k \in \{0, 1, \ldots, n\}$ and $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, (2.22) may be decomposed as

$$\begin{aligned} c_{\nu,k} &= \phi_{\nu,k|k-1}(g_k) \,, \\ \phi_{\nu,k}(f) &= c_{\nu,k}^{-1} \, \phi_{\nu,k|k-1}(f g_k) \,, \\ \phi_{\nu,k+1|k} &= \phi_{\nu,k} Q \,. \end{aligned} \tag{2.27}$$

The equivalence of (2.27) with (2.22) is straightforward and is a direct consequence of the remark that $\phi_{k+1|k} = \phi_{\nu,k}Q$, which follows from Proposition 19 in Section 2.1.2. In addition, each of the two steps in (2.27) has a very transparent interpretation.

***Predictor to Filter*:** The first two equations in (2.27) may be summarized as

$$\phi_{\nu,k}(f) \propto \int f(x) \, g(x, Y_k) \, \phi_{\nu,k|k-1}(dx) \,, \tag{2.28}$$

where the symbol $\propto$ means "up to a normalization constant" (such that $\phi_{\nu,k}(1) = 1$) and the full notation $g(x, Y_k)$ is used in place of $g_k(x)$ to highlight the dependence on the current observation $Y_k$. Equation (2.28) is recognized as Bayes' rule applied to a very simple equivalent Bayesian pseudo-model in which

- $X_k$ is distributed *a priori* according to the predictive distribution $\phi_{\nu,k|k-1}$,
- $g$ is the conditional probability density function of $Y_k$ given $X_k$.

The filter $\phi_{\nu,k}$ is then interpreted as the posterior distribution of $X_k$ given $Y_k$ in this simple equivalent Bayesian pseudo-model.

***Filter to Predictor*:** The last equation in (2.27) simply means that the updated predicting distribution $\phi_{\nu,k+1|k}$ is obtained by applying the transition kernel $Q$ to the current filtering distribution $\phi_{\nu,k}$. We are thus left with the very basic problem of determining the one-step distribution of a Markov chain given its initial distribution.

**Remark 27.** In many situations, using (2.27) to determine $\phi_{\nu,k}$ is indeed the goal rather than simply a first step in computing smoothed distributions. In particular, for sequentially observed data, one may need to take actions based on the observations gathered so far. In such cases, filtering (or prediction) is the method of choice for inference about the unobserved states, a topic that will be developed further in Chapter 4.

**Remark 28.** Another remarkable fact about the filtering recursion is that (2.26) together with (2.27) provides a method for evaluating the likelihood $\mathrm{L}_{\nu,k}$ of the observations up to index $k$ recursively in the index $k$. In addition, as $c_{\nu,k} = \mathrm{L}_{\nu,k}/\mathrm{L}_{\nu,k-1}$ from (2.26), $c_{\nu,k}$ may be interpreted as the conditional likelihood of $Y_k$ given the previous observations $Y_{0:k-1}$. However, as discussed at the beginning of Section 2.2.2, using (2.26) directly is generally impracticable for numerical reasons. In order to avoid numerical under- or overflow, one can equivalently compute the log-likelihood $\ell_{\nu,k}$. Combining (2.26) and (2.27) gives the important formula

$$\ell_{\nu,k} \stackrel{\mathrm{def}}{=} \log \mathrm{L}_{\nu,k} = \sum_{l=0}^{k} \log \phi_{\nu,l|l-1}(g_l) \,, \tag{2.29}$$

where $\phi_{\nu,l|l-1}$ is the one-step predictive distribution computed according to (2.27) (recalling that by convention, $\phi_{\nu,0|-1}$ is used as an alternative notation for $\nu$).

**Remark 29.** The normalized backward function $\bar{\beta}_{k|n}$ does not have a simple probabilistic interpretation when isolated from the corresponding filtering measure. However, (2.24) shows that the marginal smoothing distribution, $\phi_{\nu,k|n}$, is dominated by the corresponding filtering distribution $\phi_{\nu,k}$ and that $\bar{\beta}_{k|n}$ is by definition the Radon-Nikodym derivative of $\phi_{\nu,k|n}$ with respect to $\phi_{\nu,k}$,

$$\bar{\beta}_{k|n} = \frac{d\phi_{\nu,k|n}}{d\phi_{\nu,k}}$$

As a consequence,

$$\inf\left\{M \in \mathbb{R} : \phi_{\nu,k}(\{\bar{\beta}_{k|n} \geq M\}) = 0\right\} \geq 1$$

and

$$\sup\left\{M \in \mathbb{R} : \phi_{\nu,k}(\{\bar{\beta}_{k|n} \leq M\}) = 0\right\} \leq 1\,,$$

with the conventions $\inf \emptyset = \infty$ and $\sup \emptyset = -\infty$. As a consequence, all values of $\bar{\beta}_{k|n}$ cannot get simultaneously large or close to zero as was the case for $\beta_{k|n}$, although one cannot exclude the possibility that $\bar{\beta}_{k|n}$ still has important dynamics without some further assumptions on the model.

The normalizing factor $\prod_{l=k+1}^{n} c_{\nu,l} = \mathrm{L}_{\nu,n}/\mathrm{L}_{\nu,k}$ by which $\bar{\beta}_{k|n}$ differs from the corresponding unnormalized backward function $\beta_{k|n}$ may be interpreted as the conditional likelihood of the future observations $Y_{k+1:n}$ given the observations up to index $k$, $Y_{0:k}$.

## 2.3 Markovian Decompositions

The forward-backward recursions (Proposition 21) and their normalized versions (Proposition 25) were probably already well-known to readers familiar with the hidden Markov model literature. A less widely observed fact is that the smoothing distributions may also be expressed using Markov transitions. In contrast to the forward-backward algorithm, this second approach will already be familiar to readers working with dynamic (or state-space) models (Kailath *et al.*, 2000, Chapter 10). Indeed, the method to be described in Section 2.3.2, when applied to the specific case of Gaussian linear state-space models, is known as Rauch-Tung-Striebel (sometimes, abbreviated to RTS) smoothing after Rauch *et al.* (1965). The important message here is that $\{X_k\}_{k\geq 0}$ (as well as the index-reversed version of $\{X_k\}_{k\geq 0}$, although greater care is needed to handle this second case) is a *non-homogeneous* Markov chain when conditioned on some observed values $\{Y_k\}_{0\leq k\leq n}$. The use of this approach for HMMs with finite state spaces as an alternative to the forward-backward recursions is due to Askar and Derin (1981)—see also (Ephraim and Merhav, 2002, Section V) for further references.

### 2.3.1 Forward Decomposition

Let $n$ be a given positive index and consider the finite-dimensional distributions of $\{X_k\}_{k\geq 0}$ given $Y_{0:n}$. Our goal will be to show that the distribution of $X_k$ given $X_{0:k-1}$ and $Y_{0:n}$ reduces to that of $X_k$ given $X_{k-1}$ only and $Y_{0:n}$, this for any positive index $k$. The following definition will be instrumental in decomposing the joint posterior distributions $\phi_{\nu,0:k|n}$.

**Definition 30** (Forward Smoothing Kernels)**.** *Given $n \geq 0$, define for indices $k \in \{0, \ldots, n-1\}$ the transition kernels*

$$\mathrm{F}_{k|n}(x, A) \stackrel{\text{def}}{=} \begin{cases} [\beta_{k|n}(x)]^{-1} \int_A Q(x, dx') g_{k+1}(x') \beta_{k+1|n}(x') & \text{if } \beta_{k|n}(x) \neq 0 \\ 0 & \text{otherwise} , \end{cases} \tag{2.30}$$

*for any point $x \in \mathsf{X}$ and set $A \in \mathcal{X}$. For indices $k \geq n$, simply set*

$$\mathrm{F}_{k|n} \stackrel{\text{def}}{=} Q , \tag{2.31}$$

*where $Q$ is the transition kernel of the unobservable chain $\{X_k\}_{k \geq 0}$.*

Note that for indices $k \leq n-1$, $\mathrm{F}_{k|n}$ depends on the future observations $Y_{k+1:n}$ through the backward variables $\beta_{k|n}$ and $\beta_{k+1|n}$ only. The subscript $n$ in the $\mathrm{F}_{k|n}$ notation is meant to underline the fact that, like the backward functions $\beta_{k|n}$, the forward smoothing kernels $\mathrm{F}_{k|n}$ depend on the final index $n$ where the observation sequence ends. The backward recursion of Proposition 21 implies that $[\beta_{k|n}(x)]^{-1}$ is the correct normalizing constant. Thus, for any $x \in \mathsf{X}$, $A \mapsto \mathrm{F}_{k|n}(x, A)$ is a probability measure on $\mathcal{X}$. Because the functions $x \mapsto \beta_{k|n}(x)$ are measurable on $(\mathsf{X}, \mathcal{X})$, for any set $A \in \mathcal{X}$, $x \mapsto \mathrm{F}_{k|n}(x, A)$ is $\mathcal{X}/\mathcal{B}(\mathbb{R})$-measurable. Therefore, $\mathrm{F}_{k|n}$ is indeed a Markov transition kernel on $(\mathsf{X}, \mathcal{X})$. The next proposition provides a probabilistic interpretation of this definition in terms of the posterior distribution of the state at time $k+1$, given the observations up to time $n$ and the state sequence up to time $k$.

**Proposition 31.** *Given $n$, for any index $k \geq 0$ and function $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,*

$$\mathrm{E}_\nu[f(X_{k+1}) \,|\, X_{0:k}, Y_{0:n}] = \mathrm{F}_{k|n}(X_k, f) ,$$

*where $\mathrm{F}_{k|n}$ is the forward smoothing kernel defined by* (2.30) *for indices $k \leq n-1$ and* (2.31) *for indices $k \geq n$.*

*Proof.* First consider an index $0 \leq k \leq n$ and let $f$ and $h$ denote functions in $\mathcal{F}_{\mathrm{b}}(\mathsf{X})$ and $\mathcal{F}_{\mathrm{b}}(\mathsf{X}^{k+1})$, respectively. Then

$$\mathrm{E}_\nu[f(X_{k+1}) h(X_{0:k}) \,|\, Y_{0:n}] = \int \cdots \int f(x_{k+1}) h(x_{0:k}) \, \phi_{\nu, 0:k+1|n}(dx_{0:k+1}) ,$$

which, using (2.13) and the definition (2.16) of the backward function, expands to

$$\begin{aligned} \mathrm{L}_{\nu,n}^{-1} \int \cdots \int h(x_{0:k}) \, \nu(dx_0) g_0(x_0) \prod_{i=1}^{k} Q(x_{i-1}, dx_i) g_i(x_i) \\ \times \int Q(x_k, dx_{k+1}) f(x_{k+1}) g_{k+1}(x_{k+1}) \\ \times \underbrace{\int \cdots \int \prod_{i=k+2}^{n} Q(x_{i-1}, dx_i) g_i(x_i)}_{\beta_{k+1|n}(x_{k+1})} . \end{aligned} \tag{2.32}$$

From Definition 30, $\int Q(x_k, dx_{k+1}) f(x_{k+1}) g_{k+1}(x_{k+1}) \beta_{k+1|n}(x_{k+1})$ is equal to $\mathrm{F}_{k|n}(x_k, f) \beta_{k|n}(x_k)$. Thus, (2.32) may be rewritten as

$$\begin{aligned} \mathrm{E}_\nu[f(X_{k+1}) h(X_{0:k}) \,|\, Y_{0:n}] = \mathrm{L}_{\nu,n}^{-1} \int \cdots \int \mathrm{F}_{k|n}(x_k, f) h(x_{0:k}) \\ \times \nu(dx_0) g_0(x_0) \left[ \prod_{i=1}^{k} Q(x_{i-1}, dx_i) g_i(x_i) \right] \beta_{k|n}(x_k) . \end{aligned} \tag{2.33}$$

Using the definition (2.16) of $\beta_{k|n}$ again, this latter integral is easily seen to be similar to (2.32) except for the fact that $f(x_{k+1})$ has been replaced by $\mathrm{F}_{k|n}(x_k, f)$. Hence

$$\mathrm{E}_\nu[f(X_{k+1})h(X_{0:k}) \,|\, Y_{0:n}] = \mathrm{E}_\nu[\mathrm{F}_{k|n}(X_k, f)h(X_{0:k}) \,|\, Y_{0:n}] \;,$$

for all functions $h \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{k+1}\right)$ as requested.

For $k \geq n$, the situation is simpler because (2.6) implies that $\phi_{\nu,0:k+1|n} = \phi_{\nu,0:k|n}Q$. Hence,

$$\begin{aligned}\mathrm{E}_\nu[f(X_{k+1})h(X_{0:k}) \,|\, Y_{0:n}] \\ = \int \cdots \int h(x_{0:k})\, \phi_{\nu,0:k|n}(dx_{0:k}) \int Q(x_k, dx_{k+1}) f(x_{k+1}) \;,\end{aligned}$$

and thus

$$\begin{aligned}\mathrm{E}_\nu[f(X_{k+1})h(X_{0:k}) \,|\, Y_{0:n}] &= \int \cdots \int h(x_{0:k}) \phi_{\nu,0:k|n}(dx_{0:k}) Q(x_k, f) \;, \\ &= \mathrm{E}_\nu[Q(X_k, f)h(X_{0:k}) \,|\, Y_{0:n}] \;.\end{aligned}$$

□

**Remark 32.** A key ingredient of the above proof is (2.32), which gives a representation of the joint smoothing distribution of the state variables $X_{0:k}$ given the observations up to index $n$, with $n \geq k$. This representation, which states that

$$\begin{aligned}&\phi_{\nu,0:k|n}(f) \\ &\quad = \mathrm{L}_{\nu,n}^{-1} \int \cdots \int f(x_{0:k})\, \nu(dx_0) g_0(x_0) \left[\prod_{i=1}^{k} Q(x_{i-1}, dx_i) g_i(x_i)\right] \beta_{k|n}(x_k) \end{aligned} \quad (2.34)$$

for all $f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{k+1}\right)$, is a generalization of the marginal forward-backward decomposition as stated in (2.14).

Proposition 31 implies that, *conditionally on* the observations $Y_{0:n}$, the state sequence $\{X_k\}_{k\geq 0}$ is a non-homogeneous Markov chain associated with the family of Markov transition kernels $\{\mathrm{F}_{k|n}\}_{k\geq 0}$ and initial distribution $\phi_{\nu,0|n}$. The fact that the Markov property of the state sequence is preserved when conditioning sounds surprising because the (marginal) smoothing distribution of the state $X_k$ depends on both past and future observations. There is however nothing paradoxical here, as the Markov transition kernels $\mathrm{F}_{k|n}$ indeed depend (and depend only) on the future observations $Y_{k+1:n}$.

As a consequence of Proposition 31, the joint smoothing distributions may be rewritten in a form that involves the forward smoothing kernels using the Chapman-Kolmogorov equations (1.1).

**Proposition 33.** *For any integers $n$ and $m$, function $f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{m+1}\right)$ and initial probability $\nu$ on $(\mathsf{X}, \mathcal{X})$,*

$$\begin{aligned}\mathrm{E}_\nu[f(X_{0:m})) \,|\, Y_{0:n}] = \\ \int \cdots \int f(x_{0:m})\, \phi_{\nu,0|n}(dx_0) \prod_{i=1}^{m} \mathrm{F}_{i-1|n}(x_{i-1}, dx_i) \;,\end{aligned} \quad (2.35)$$

*where $\{\mathrm{F}_{k|n}\}_{k\geq 0}$ are defined by* (2.30) *and* (2.31) *and $\phi_{\nu,0|n}$ is the marginal smoothing distribution defined, for any $A \in \mathcal{X}$, by*

$$\phi_{\nu,0|n}(A) = [\nu(g_0\beta_{0|n})]^{-1} \int_A \nu(dx) g_0(x) \beta_{0|n}(x) \;. \quad (2.36)$$

If one is only interested in computing the fixed point marginal smoothing distributions, (2.35) may also be used as the second phase of a smoothing approach which we recapitulate below.

**Corollary 34** (Alternative Smoothing Algorithm)**. Backward Recursion** *Compute the backward variables $\beta_{n|n}$ down to $\beta_{0|n}$ by backward recursion according to* (2.19) *in Proposition 21.*

**Forward Smoothing** *$\phi_{\nu,0|n}$ is given by* (2.36) *and for $k \geq 0$,*

$$\phi_{\nu,k+1|n} = \phi_{\nu,k|n} \mathrm{F}_{k|n} \ ,$$

*where $\mathrm{F}_{k|n}$ are the forward kernels defined by* (2.30)*.*

For numerical implementation, Corollary 34 is definitely less attractive than the normalized forward-backward approach of Proposition 25 because the backward pass cannot be carried out in normalized form without first determining the forward measures $\alpha_{\nu,k}$.

On the other hand, Proposition 33 provides a general decomposition of the joint smoothing distribution that will be instrumental in establishing some form of ergodicity of the Markov chain that corresponds to the unobservable states $\{X_k\}_{k\geq 0}$, conditional on some observations $Y_{0:n}$ (see Section 3).

### 2.3.2 Backward Decomposition

In the previous section it was shown that, conditionally on the observations up to index $n$, $Y_{0:n}$, the state sequence $\{X_k\}_{k\geq 0}$ is a Markov chain, with transition kernels $\mathrm{F}_{k|n}$. We now turn to the so-called *time-reversal* issue: is it true in general that the unobserved chain *with the indices in reverse order*, forms a non-homogeneous Markov chain, conditionally on some observations $Y_{0:n}$?

We already discussed time-reversal for Markov chains in Section 1.1 where it has been argued that the main technical difficulty consists in guaranteeing that the reverse kernel does exist. For this, we require somewhat stronger assumptions on the nature of $\mathsf{X}$ by assuming for the rest of this section that $\mathsf{X}$ is a Polish space and that $\mathcal{X}$ is the associated Borel $\sigma$-field. From the discussion in Section 1.1 (see Definition 2 and comment below), we then know that the reverse kernel does exist although we may not be able to provide a simple closed-form expression for it. The reverse kernel does have a simple expression, however, as soon as one assumes that the kernel to be reversed and the initial distribution admit densities with respect to some measure on $\mathsf{X}$.

Let us now return to the smoothing problem. For positive indices $k$ such that $k \leq n-1$, the posterior distribution of $(X_k, X_{k+1})$ given the observations up to time $k$ satisfies

$$\mathrm{E}_\nu[f(X_k, X_{k+1}) \,|\, Y_{0:k}] = \iint f(x_k, x_{k+1}) \, \phi_{\nu,k}(dx_k) \, Q(x_k, dx_{k+1}) \tag{2.37}$$

for all $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X} \times \mathsf{X})$. From the previous discussion, there exists a Markov transition kernel $\mathrm{B}_{\nu,k}$ which satisfies Definition 2, that is

$$\mathrm{B}_{\nu,k} \stackrel{\mathrm{def}}{=} \{\mathrm{B}_{\nu,k}(x, A), x \in \mathsf{X}, A \in \mathcal{X}\}$$

such that for any function $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X} \times \mathsf{X})$,

$$\mathrm{E}_\nu[f(X_k, X_{k+1}) \,|\, Y_{0:k}] = \iint f(x_k, x_{k+1}) \, \phi_{\nu,k+1|k}(dx_{k+1}) \, \mathrm{B}_{\nu,k}(x_{k+1}, dx_k) \ , \tag{2.38}$$

where $\phi_{\nu,k+1|k} = \phi_{\nu,k} Q$ is the one-step predictive distribution.

**Proposition 35.** *Given a strictly positive index $n$, initial distribution $\nu$, and index $k \in \{0, \dots, n-1\}$,*

$$\mathrm{E}_\nu[f(X_k) \,|\, X_{k+1:n}, Y_{0:n}] = \mathrm{B}_{\nu,k}(X_{k+1}, f)$$

*for any $f \in \mathcal{F}_\mathrm{b}(\mathsf{X})$. Here, $\mathrm{B}_{\nu,k}$ is the backward smoothing kernel defined in* (2.38)*.*

Before giving the proof of this result, we make a few remarks to provide some intuitive understanding of the backward smoothing kernels.

**Remark 36.** Contrary to the forward kernel, the backward transition kernel is only defined implicitly through the equality of the two representations (2.37) and (2.38). This limitation is fundamentally due to the fact that the backward kernel implies a non-trivial time-reversal operation.

Proposition 35 however allows a simple interpretation of the backward kernel: Because $\mathrm{E}_\nu[f(X_k) \,|\, X_{k+1:n}, Y_{0:n}]$ is equal to $\mathrm{B}_{\nu,k}(X_{k+1}, f)$ and thus depends neither on $X_l$ for $l > k+1$ nor on $Y_l$ for $l \geq k+1$, the tower property of conditional expectation implies that not only is $\mathrm{B}_{\nu,k}(X_{k+1}, f)$ equal to $\mathrm{E}_\nu[f(X_k) \,|\, X_{k+1}, Y_{0:n}]$ but also coincides with $\mathrm{E}_\nu[f(X_k) \,|\, X_{k+1}, Y_{0:k}]$, for any $f \in \mathcal{F}_\mathrm{b}(\mathsf{X})$. In addition, the distribution of $X_{k+1}$ given $X_k$ and $Y_{0:k}$ reduces to $Q(X_k, \cdot)$ due to the particular form of the transition kernel associated with a hidden Markov model (see Definition 11). Recall also that the distribution of $X_k$ given $Y_{0:k}$ is denoted by $\phi_{\nu,k}$. Thus, $\mathrm{B}_{\nu,k}$ can be interpreted as a Bayesian posterior in the equivalent pseudo-model where

- $X_k$ is distributed *a priori* according to the filtering distribution $\phi_{\nu,k}$,
- The conditional distribution of $X_{k+1}$ given $X_k$ is $Q(X_k, \cdot)$.

$\mathrm{B}_{\nu,k}(X_{k+1}, \cdot)$ is then interpreted as the posterior distribution of $X_k$ given $X_{k+1}$ in this equivalent pseudo-model.

In particular, for HMMs that are "fully dominated" in the sense of Definition 13, $Q$ has a transition probability density function $q$ with respect to a measure $\lambda$ on $\mathsf{X}$. This is then also the case for $\phi_{\nu,k}$, which is a marginal of (2.13). In such cases, we shall use the slightly abusive but unambiguous notation $\phi_{\nu,k}(dx) = \phi_{\nu,k}(x)\,\lambda(dx)$ (that is, $\phi_{\nu,k}$ denotes the probability density function with respect to $\lambda$ rather than the probability distribution). The backward kernel $\mathrm{B}_{\nu,k}(x_{k+1}, \cdot)$ then has a probability density function with respect to $\lambda$, which is given by Bayes' formula,

$$\mathrm{B}_{\nu,k}(x_{k+1}, x) = \frac{\phi_{\nu,k}(x) q(x, x_{k+1})}{\int_\mathsf{X} \phi_{\nu,k}(x) q(x, x_{k+1})\,\lambda(dx)} \,. \tag{2.39}$$

Thus, in many cases of interest, the backward transition kernel $\mathrm{B}_{\nu,k}$ can be written straightforwardly as a function of $\phi_{\nu,k}$ and $Q$. In these situations, Proposition 38 is the method of choice for smoothing, as it only involves normalized quantities, whereas Corollary 34 is not normalized and thus can generally not be implemented as it stands.

*of Proposition 35.* Let $k \in \{0, \dots, n-1\}$ and $h \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{n-k}\right)$. Then

$$\mathrm{E}_\nu[f(X_k) h(X_{k+1:n}) \,|\, Y_{0:n}] = \int \cdots \int f(x_k) h(x_{k+1:n})\, \phi_{\nu,k:n|n}(dx_{k:n}) \,. \tag{2.40}$$

Using the definition (2.13) of the joint smoothing distribution $\phi_{\nu,k:n|n}$ yields

$$\begin{aligned}
&\mathrm{E}_\nu[f(X_k)h(X_{k+1:n}) \,|\, Y_{0:n}] \\
&\quad = \mathrm{L}_{\nu,n}^{-1} \int \cdots \int \nu(dx_0) g_0(x_0) \prod_{i=1}^{k} Q(x_{i-1}, dx_i) g_i(x_i) f(x_k) \\
&\qquad\qquad \times \left[ \prod_{i=k+1}^{n} Q(x_{i-1}, dx_i) g_i(x_i) \right] h(x_{k+1:n}) \,, \\
&\quad = \frac{\mathrm{L}_{\nu,k}}{\mathrm{L}_{\nu,n}} \iint \phi_{\nu,k|n}(dx_k) Q(x_k, dx_{k+1}) f(x_k) g_{k+1}(x_{k+1}) \\
&\qquad\qquad \times \int \cdots \int \left[ \prod_{i=k+2}^{n} Q(x_{i-1}, dx_i) g_i(x_i) \right] h(x_{k+1:n}) \,,
\end{aligned} \tag{2.41}$$

which implies, by the definition (2.38) of the backward kernel, that

$$\begin{aligned}
&\mathrm{E}_\nu[f(X_k)h(X_{k+1:n}) \,|\, Y_{0:n}] \\
&\quad = \frac{\mathrm{L}_{\nu,k}}{\mathrm{L}_{\nu,n}} \iint \mathrm{B}_{\nu,k}(x_{k+1}, dx_k) f(x_k) \phi_{\nu,k+1|k}(dx_{k+1}) g_{k+1}(x_{k+1}) \\
&\qquad\qquad \times \int \cdots \int \left[ \prod_{i=k+2}^{n} Q(x_{i-1}, dx_i) g_i(x_i) \right] h(x_{k+1:n}) \,.
\end{aligned} \tag{2.42}$$

Taking $f \equiv 1$ shows that for any function $h' \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{n-k}\right)$,

$$\begin{aligned}
\mathrm{E}_\nu[h'(X_{k+1:n}) \,|\, Y_{0:n}] = \frac{\mathrm{L}_{\nu,k}}{\mathrm{L}_{\nu,n}} \int \cdots \int h'(x_{k+1:n}) \\
\times \phi_{\nu,k+1|k}(dx_{k+1}) g_{k+1}(x_{k+1}) \prod_{i=k+2}^{n} Q(x_{i-1}, dx_i) g_i(x_i) \,.
\end{aligned}$$

Identifying $h'$ with $h(x_{k+1:n}) \int f(x) \, \mathrm{B}_{\nu,k}(x_{k+1}, dx)$, we find that (2.42) may be rewritten as

$$\begin{aligned}
&\mathrm{E}_\nu[f(X_k)h(X_{k+1:n}) \,|\, Y_{0:n}] \\
&\qquad\qquad = \mathrm{E}_\nu \left[ h(X_{k+1:n}) \int \mathrm{B}_{\nu,k}(X_{k+1}, dx) f(x) \,\middle|\, Y_{0:n} \right],
\end{aligned}$$

which concludes the proof. □

The next result is a straightforward consequence of Proposition 35, which reformulates the joint smoothing distribution $\phi_{\nu,0:n|n}$ in terms of the backward smoothing kernels.

**Corollary 37.** *For any integer $n > 0$ and initial probability $\nu$,*

$$\mathrm{E}_\nu[f(X_{0:n}) \,|\, Y_{0:n}] = \int \cdots \int f(x_{0:m}) \, \phi_{\nu,n}(dx_n) \prod_{k=0}^{n-1} \mathrm{B}_{\nu,k}(x_{k+1}, dx_k) \tag{2.43}$$

*for all $f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}^{n+1}\right)$. Here, $\{\mathrm{B}_{\nu,k}\}_{0 \le k \le n-1}$ are the backward smoothing kernels defined in* (2.38) *and $\phi_{\nu,n}$ is the marginal filtering distribution corresponding to the final index $n$.*

It follows from Proposition 35 and Corollary 37 that, conditionally on $Y_{0:n}$, the joint distribution of the index-reversed sequence $\{\bar{X}_k\}_{0\le k\le n}$, with $\bar{X}_k = X_{n-k}$, is that of a non-homogeneous Markov chain with initial distribution $\phi_{\nu,n}$ and transition kernels $\{\mathrm{B}_{\nu,n-k}\}_{1\le k\le n}$. This is an exact analog of the forward decomposition where the ordering of indices has been reversed, starting from the end of the observation sequence and ending with the first observation. Three important differences versus the forward decomposition should however be kept in mind.

(i) The backward smoothing kernel $\mathrm{B}_{\nu,k}$ depends on the initial distribution $\nu$ and on the observations up to index $k$ but it depends neither on the future observations nor on the index $n$ where the observation sequence ends. As a consequence, the sequence of backward transition kernels $\{\mathrm{B}_{\nu,k}\}_{0\le k\le n-1}$ may be computed by forward recurrence on $k$, irrespectively of the length of the observation sequence. In other terms, the backward smoothing kernel $\mathrm{B}_{\nu,k}$ depends only on the filtering distribution $\phi_{\nu,k}$, whereas the forward smoothing kernel $\mathrm{F}_{k|n}$ was to be computed from the backward function $\beta_{k|n}$.

(ii) Because $\mathrm{B}_{\nu,k}$ depends on $\phi_{\nu,k}$ rather than on the unnormalized forward measure $\alpha_{\nu,k}$, its computation involves only properly normalized quantities (Remark 36). The backward decomposition is thus more adapted to the actual computation of the smoothing probabilities than the forward decomposition. The necessary steps are summarized in the following result.

**Proposition 38** (Forward Filtering/Backward Smoothing)**. Forward Filtering** *Compute, forward in time, the filtering distributions $\phi_{\nu,0}$ to $\phi_{\nu,n}$ using the recursion* (2.22)*. At each index $k$, the backward transition kernel $\mathrm{B}_{\nu,k}$ may be computed according to* (2.38)*.*

**Backward Smoothing** *From $\phi_{\nu,n}$, compute, for $k = n-1, n-2, \ldots, 0$,*

$$\phi_{\nu,k|n} = \phi_{\nu,k+1|n}\,\mathrm{B}_{\nu,k}\ ,$$

*recalling that $\phi_{\nu,n|n} \stackrel{\mathrm{def}}{=} \phi_{\nu,n}$.*

(iii) A more subtle difference between the forward and backward Markovian decompositions is the observation that Definition 30 does provide an expression of the forward kernels $\mathrm{F}_{k|n}$ for any $k \ge 0$, that is, also for indices *after* the end of the observation sequence. Hence, the process $\{X_k\}_{k\ge 0}$, when conditioned on some observations $Y_{0:n}$, really forms a non-homogeneous Markov chain whose finite-dimensional distributions are defined by Proposition 33. In contrast, the backward kernels $\mathrm{B}_{\nu,k}$ are defined for indices $k \in \{0, \ldots, n-1\}$ only, and thus the index-reversed process $\{X_{n-k}\}$ is also defined, by Proposition 35, for indices $k$ in the range $\{0, \ldots, n\}$ only. In order to define the index-reversed chain for negative indices, a minimal requirement is that the underlying chain $\{X_k\}$ also be well defined for $k < 0$. Defining Markov chains $\{X_k\}$ with indices $k \in \mathbb{Z}$ is only meaningful in the stationary case, that is when $\nu$ is the stationary distribution of $Q$. As both this stationarization issue and the forward and backward Markovian decompositions play a key role in the analysis of the statistical properties of the maximum likelihood estimator, we postpone further discussion of this point to Chapter 6.

# Chapter 3

# Forgetting of the initial condition and filter stability

Recall from previous chapters that in a partially dominated HMM model (see Definition 12), we denote by

- $\mathrm{P}_\nu$ the probability associated to the Markov chain $\{X_k, Y_k\}_{k\geq 0}$ on the canonical space $\left((\mathsf{X}\times\mathsf{Y})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{Y})^{\otimes\mathbb{N}}\right)$ with initial probability measure $\nu$ and transition kernel $T$ defined by (1.15);
- $\phi_{\nu,k|n}$ the distribution of the hidden state $X_k$ conditionally on the observations $Y_{0:n}$, under the probability measure $\mathrm{P}_\nu$.

*Forgetting properties* pertain to the dependence of $\phi_{\nu,k|n}$ with respect to the initial distribution $\nu$. A typical question is to ask whether $\phi_{\nu,k|n}$ and $\phi_{\nu',k|n}$ are close (in some sense) for large values of $k$ and arbitrary choices of $\nu$ and $\nu'$. This issue will play a key role both when studying the convergence of sequential Monte Carlo methods (Chapter **??**) and when analyzing the asymptotic behavior of the maximum likelihood estimator (Chapter 6).

In the following, it is shown more precisely that, under appropriate conditions on the kernel $Q$ of the hidden chain and on the transition density function $g$, the total variation distance $\left\|\phi_{\nu,k|n} - \phi_{\nu',k|n}\right\|_{\mathrm{TV}}$ converges to zero as $k$ tends to infinity. Remember that, following the implicit conditioning convention (Section 2.1.4), we usually omit to indicate explicitly that $\phi_{\nu,k|n}$ indeed depends on the observations $Y_{0:n}$. In this section however we cannot use this convention anymore, as we will meet both situations in which, say, $\|\phi_{\nu,n} - \phi_{\nu',n}\|_{\mathrm{TV}}$ converges to zero (as $n$ tends to infinity) *for all possible values of* the sequence $\{y_n\}_{n\geq 0} \in \mathsf{Y}^{\mathbb{N}}$ (*uniform forgetting*) and cases where $\|\phi_{\nu,n} - \phi_{\nu',n}\|_{\mathrm{TV}}$ can be shown to converge to zero almost surely only when $\{Y_k\}_{k\geq 0}$ is assumed to be distributed under a specific distribution (typically $\mathrm{P}_{\nu_\star}$ for some initial distribution $\nu_\star$). In this section, we thus make dependence with respect to the observations explicit by indicating the relevant subset of observation between brackets, using, for instance, $\phi_{\nu,k|n}[y_{0:n}]$ rather than $\phi_{\nu,k|n}$.

We start by recalling some elementary facts and results about the total variation norm of a signed measure, providing in particular useful characterizations of the total variation as an operator norm over appropriately defined function spaces. We then discuss the contraction property of Markov kernels, using the measure-theoretic approach introduced in an early paper by Dobrushin (1956) and recently revisited and extended by Del Moral *et al.* (2003). We finally present the applications of these results to establish forgetting properties of the smoothing and filtering recursions and discuss the implications of the technical conditions required to obtain these results.

### 3.0.3 Total Variation

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and let $\xi$ be a signed measure on $(\mathsf{X}, \mathcal{X})$. Then there exists a measurable set $H \in \mathcal{X}$, called a *Jordan set*, such that

(i) $\xi(A) \geq 0$ for each $A \in \mathcal{X}$ such that $A \subseteq H$;

(ii) $\xi(A) \leq 0$ for each $A \in \mathcal{X}$ such that $A \subseteq \mathsf{X} \setminus H$.

The set $H$ is not unique, but any other such set $H' \in \mathcal{X}$ satisfies $\xi(H \cap H') = 1$. Hence two Jordan sets differ by at most a set of zero measure. If $\mathsf{X}$ is finite or countable and $\mathcal{X} = \mathcal{P}(\mathsf{X})$ is the collection of all subsets of $\mathsf{X}$, then $H = \{x : \xi(x) \geq 0\}$ and $H' = \{x : \xi(x) > 0\}$ are two Jordan sets. As another example, if $\xi$ is absolutely continuous with respect to a measure $\nu$ on $(\mathsf{X}, \mathcal{X})$ with Radon-Nikodym derivative $f$, then $\{f \geq 0\}$ and $\{f > 0\}$ are two Jordan sets. We define two measures on $(\mathsf{X}, \mathcal{X})$ by

$$\xi_+(A) = \xi(H \cap A) \quad \text{and} \quad \xi_-(A) = -\xi(H^c \cap A) \,, \qquad A \in \mathcal{X} \,.$$

The measures $\xi_+$ and $\xi_-$ are referred to as the *positive* and *negative variations* of the signed measure $\xi$. By construction, $\xi = \xi_+ - \xi_-$. This decomposition of $\xi$ into its positive and negative variations is called the *Hahn-Jordan decomposition* of $\xi$. The definition of the positive and negative variations above is easily shown to be independent of the particular Jordan set chosen.

**Definition 39** (Total Variation of a Signed Measure)**.** *Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and let $\xi$ be a signed measure on $(\mathsf{X}, \mathcal{X})$. The* total variation norm *of $\xi$ is defined as*

$$\|\xi\|_{\mathrm{TV}} = \xi_+(\mathsf{X}) + \xi_-(\mathsf{X}) \,,$$

*where $(\xi_+, \xi_-)$ is the Hahn-Jordan decomposition of $\xi$.*

If $\mathsf{X}$ is finite or countable and $\xi$ is a signed measure on $(\mathsf{X}, \mathcal{P}(\mathsf{X}))$, then $\|\xi\|_{\mathrm{TV}} = \sum_{x \in \mathsf{X}} |\xi(x)|$. If $\xi$ has a density $f$ with respect to a measure $\lambda$ on $(\mathsf{X}, \mathcal{X})$, then $\|\xi\|_{\mathrm{TV}} = \int |f(x)| \, \lambda(dx)$.

**Definition 40** (Total Variation Distance)**.** *Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and let $\xi$ and $\xi'$ be two measures on $(\mathsf{X}, \mathcal{X})$. The* total variation distance *between $\xi$ and $\xi'$ is the total variation norm of the signed measure $\xi - \xi'$.*

Denote by $\mathrm{M}(\mathsf{X}, \mathcal{X})$ the set of finite signed measures on the measurable space $(\mathsf{X}, \mathcal{X})$, by $\mathrm{M}_1(\mathsf{X}, \mathcal{X})$ the set of probability measures on $(\mathsf{X}, \mathcal{X})$ and by $\mathrm{M}_0(\mathsf{X}, \mathcal{X})$ the set of finite signed measures $\xi$ on $(\mathsf{X}, \mathcal{X})$ satisfying $\xi(\mathsf{X}) = 0$. $\mathrm{M}(\mathsf{X}, \mathcal{X})$ is a Banach space with respect to the total variation norm. In this Banach space, the subset $\mathrm{M}_1(\mathsf{X}, \mathcal{X})$ is closed and convex.

Let $\mathcal{F}_{\mathrm{b}}(\mathsf{X})$ denote the set of bounded measurable real functions on $\mathsf{X}$. This set embedded with the supremum norm $\|f\|_\infty = \sup\{f(x) : x \in \mathsf{X}\}$ also is a Banach space. For any $\xi \in \mathrm{M}(\mathsf{X}, \mathcal{X})$ and $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$, we may define $\xi(f) = \int f \, d\xi$. Therefore any finite signed measure $\xi$ in $\mathrm{M}(\mathsf{X}, \mathcal{X})$ defines a linear functional on the Banach space $(\mathcal{F}_{\mathrm{b}}(\mathsf{X}), \|\cdot\|_\infty)$. We will use the same notation for the measure and for the functional. The following lemma shows that the total variation of the signed measure $\xi$ agrees with the operator norm of $\xi$.

**Lemma 41.**

*(i) For any $\xi \in \mathrm{M}(\mathsf{X}, \mathcal{X})$ and $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,*

$$\left| \int f \, d\xi \right| \leq \|\xi\|_{\mathrm{TV}} \ \|f\|_\infty \ .$$

*(ii) For any* $\xi \in \mathrm{M}(\mathsf{X}, \mathcal{X})$*,*

$$\|\xi\|_{\mathrm{TV}} = \sup\left\{\xi(f) : f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}, \mathcal{X}\right), \|f\|_{\infty} = 1\right\} \ .$$

*(iii) For any* $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$*,*

$$\|f\|_{\infty} = \sup\left\{\xi(f) : \xi \in \mathrm{M}(\mathsf{X}, \mathcal{X}), \|\xi\|_{\mathrm{TV}} = 1\right\} \ .$$

*Proof.* Let $H$ be a Hahn-Jordan set of $\xi$. Then $\xi_+(H) = \xi(H)$ and $\xi_-(H^c) = -\xi(H^c)$. For $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$,

$$|\xi(f)| \leq |\xi_+(f)| + |\xi_-(f)| \leq \|f\|_{\infty}\left(\xi_+(\mathsf{X}) + \xi_-(\mathsf{X})\right) = \|f\|_{\infty}\|\xi\|_{\mathrm{TV}} \ ,$$

showing (i). It also shows that the suprema in (ii) and (iii) are no larger than $\|\xi\|_{\mathrm{TV}}$ and $\|f\|_{\infty}$, respectively. To establish equality in these relations, first note that $\|\mathbb{1}_H - \mathbb{1}_{H^c}\|_{\infty} = 1$ and $\xi\left(\mathbb{1}_H - \mathbb{1}_{H^c}\right) = \xi(H) - \xi(H^c) = \|\xi\|_{\mathrm{TV}}$. This proves (ii). Next pick $f$ and let let $\{x_n\}$ be a sequence in $\mathsf{X}$ such that $\lim_{n\to\infty}|f(x_n)| = \|f\|_{\infty}$. Then $\|f\|_{\infty} = \lim_{n\to\infty}|\delta_{x_n}(f)|$, proving (iii). □

The set $\mathrm{M}_0(\mathsf{X}, \mathcal{X})$ possesses some interesting properties that will prove useful in the sequel. Let $\xi$ be in this set. Because $\xi(\mathsf{X}) = 0$, for any $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$ and any real $c$ it holds that $\xi(f) = \xi(f - c)$. Therefore by Lemma 41(i), $|\xi(f)| \leq \|\xi\|_{\mathrm{TV}}\|f - c\|_{\infty}$, which implies that

$$|\xi(f)| \leq \|\xi\|_{\mathrm{TV}} \inf_{c\in\mathbb{R}} \|f - c\|_{\infty} \ .$$

It is easily seen that for any $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$, $\inf_{c\in\mathbb{R}}\|f - c\|_{\infty}$ is related to the oscillation semi-norm of $f$, also called the global modulus of continuity,

$$\mathrm{osc}\left(f\right) \stackrel{\mathrm{def}}{=} \sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} |f(x) - f(x')| = 2\inf_{c\in\mathbb{R}} \|f - c\|_{\infty} \ . \tag{3.1}$$

The lemma below provides some additional insight into this result.

**Lemma 42.** *For any* $\xi \in \mathrm{M}(\mathsf{X}, \mathcal{X})$ *and* $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$*,*

$$|\xi(f)| \leq \sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} |\xi_+(\mathsf{X})f(x) - \xi_-(\mathsf{X})f(x')| \ , \tag{3.2}$$

*where* $(\xi_+, \xi_-)$ *is the Hahn-Jordan decomposition of* $\xi$*. In particular, for any* $\xi \in \mathrm{M}_0(\mathsf{X}, \mathcal{X})$ *and* $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$*,*

$$|\xi(f)| \leq \frac{1}{2}\|\xi\|_{\mathrm{TV}}\ \mathrm{osc}\left(f\right) \ , \tag{3.3}$$

*where* $\mathrm{osc}\left(f\right)$ *is given by* (3.1)*.*

*Proof.* First note that

$$\begin{aligned}\xi(f) &= \int f(x)\,\xi_+(dx) - \int f(x)\,\xi_-(dx) \\ &= \frac{\iint f(x)\,\xi_+(dx)\,\xi_-(dx')}{\xi_-(\mathsf{X})} - \frac{\iint f(x')\,\xi_+(dx)\,\xi_-(dx')}{\xi_+(\mathsf{X})} \ .\end{aligned}$$

Therefore

$$\begin{aligned}|\xi(f)| &\leq \iint |f(x)/\xi_-(\mathsf{X}) - f(x')/\xi_+(\mathsf{X})|\,\xi_+(dx)\,\xi_-(dx') \\ &\leq \sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} |f(x)/\xi_-(\mathsf{X}) - f(x')/\xi_+(\mathsf{X})|\,\xi_+(\mathsf{X})\xi_-(\mathsf{X}) \ ,\end{aligned}$$

which shows (3.2). If $\xi(\mathsf{X}) = 0$, then $\xi_+(\mathsf{X}) = \xi_-(\mathsf{X}) = \frac{1}{2}\|\xi\|_{\mathrm{TV}}$, showing (3.3). □

Therefore, for $\xi \in \mathrm{M}_0(\mathsf{X}, \mathcal{X})$, $\|\xi\|_{\mathrm{TV}}$ is the operator norm of $\xi$ considered as an operator over the space $\mathcal{F}_{\mathrm{b}}(\mathsf{X})$ equipped with the oscillation semi-norm (3.1). As a direct application of this result, if $\xi$ and $\xi'$ are two probability measures on $(\mathsf{X}, \mathcal{X})$, then $\xi - \xi' \in \mathrm{M}_0(\mathsf{X}, \mathcal{X})$ which implies that for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$|\xi(f) - \xi'(f)| \leq \frac{1}{2} \|\xi - \xi'\|_{\mathrm{TV}} \operatorname{osc}(f) \ . \tag{3.4}$$

This inequality is sharper than the bound $|\xi(f) - \xi'(f)| \leq \|\xi - \xi'\|_{\mathrm{TV}} \|f\|_\infty$ provided by Lemma 41(i), because $\operatorname{osc}(f) \leq 2\|f\|_\infty$.

We conclude this section by establishing some alternative expressions for the total variation distance between two probability measures.

**Lemma 43.** *For any $\xi$ and $\xi'$ in $\mathrm{M}_1(\mathsf{X}, \mathcal{X})$,*

$$\frac{1}{2}\|\xi - \xi'\|_{\mathrm{TV}} = \sup_A |\xi(A) - \xi'(A)| \tag{3.5}$$

$$= 1 - \sup_{\nu \leq \xi, \xi'} \nu(\mathsf{X}) \tag{3.6}$$

$$= 1 - \inf \sum_{p=1}^{n} \xi(A_i) \wedge \xi'(A_i) \ . \tag{3.7}$$

*Here the supremum in* (3.5) *is taken over all measurable subsets of $\mathsf{X}$, the supremum in* (3.6) *is taken over all finite signed measures $\nu$ on $(\mathsf{X}, \mathcal{X})$ satisfying $\nu \leq \xi$ and $\nu \leq \xi'$, and the infimum in* (3.7) *is taken over all finite measurable partitions $A_1, \ldots, A_n$ of $\mathsf{X}$.*

*Proof.* To prove (3.5), first write $\xi(A) - \xi'(A) = (\xi - \xi')\mathbb{1}_A$ and note that $\operatorname{osc}(\mathbb{1}_A) = 1$. Thus (3.4) shows that the supremum in (3.5) is no larger than $(1/2)\|\xi - \xi'\|_{\mathrm{TV}}$. Now let $H$ be a Jordan set of the signed measure $\xi - \xi'$. The supremum is bounded from below by $\xi(H) - \xi'(H) = (\xi - \xi')_+(\mathsf{X}) = (1/2)\|\xi - \xi'\|_{\mathrm{TV}}$. This establishes equality in (3.5).

We now turn to (3.6). For any $p, q \in \mathbb{R}$, $|p - q| = p + q - 2(p \wedge q)$. Therefore for any $A \in \mathcal{X}$,

$$\frac{1}{2}|\xi(A) - \xi'(A)| = \frac{1}{2}(\xi(A) + \xi'(A)) - \xi(A) \wedge \xi'(A) \ .$$

Applying this relation to the sets $H$ and $H^c$, where $H$ is as above, shows that

$$\frac{1}{2}(\xi - \xi')(H) = \frac{1}{2}[\xi(H) + \xi'(H)] - \xi(H) \wedge \xi'(H) \ ,$$
$$\frac{1}{2}(\xi' - \xi)(H^c) = \frac{1}{2}[\xi(H^c) + \xi'(H^c)] - \xi(H^c) \wedge \xi'(H^c) \ .$$

For any measure $\nu$ such that $\nu \leq \xi$ and $\nu \leq \xi'$, it holds that $\nu(H) \leq \xi(H) \wedge \xi'(H)$ and $\nu(H^c) \leq \xi(H^c) \wedge \xi'(H^c)$, showing that

$$\frac{1}{2}(\xi - \xi')(H) + \frac{1}{2}(\xi' - \xi)(H^c) = \frac{1}{2}\|\xi - \xi'\|_{\mathrm{TV}} \leq 1 - \nu(\mathsf{X}) \ .$$

Thus (3.6) is no smaller than the left-hand side. To show equality, let $\nu$ be the measure defined by

$$\nu(A) = \xi(A \cap H^c) + \xi'(A \cap H) \ . \tag{3.8}$$

By the definition of $H$, $\xi(A \cap H^c) \leq \xi'(A \cap H^c)$ and $\xi'(A \cap H) \leq \xi(A \cap H)$ for any $A \in \mathcal{X}$. Therefore $\nu(A) \leq \xi(A)$ and $\nu(A) \leq \xi'(A)$. In addition, $\nu(H) = \xi'(H) =$

$\xi(H) \wedge \xi'(H)$ and $\nu(H^c) = \xi(H^c) = \xi(H^c) \wedge \xi'(H^c)$, showing that $\frac{1}{2} \|\xi - \xi'\|_{\mathrm{TV}} = 1 - \nu(\mathsf{X})$ and concluding the proof of (3.6).

Finally, because $\nu(\mathsf{X}) = \xi(H) \wedge \xi'(H) + \xi(H^c) \wedge \xi'(H^c)$ we have

$$\sup_{\nu \leq \xi, \xi'} \nu(\mathsf{X}) \geq \inf \sum_{i=1}^{n} \xi(A_i) \wedge \xi'(A_i) \; .$$

Conversely, for any measure $\nu$ satisfying $\nu \leq \xi$ and $\nu \leq \xi'$, and any partition $A_1, \ldots, A_n$,

$$\nu(\mathsf{X}) = \sum_{i=1}^{n} \nu(A_i) \leq \sum_{i=1}^{n} \xi(A_i) \wedge \xi'(A_i) \; ,$$

showing that

$$\sup_{\nu \leq \xi, \xi'} \nu(\mathsf{X}) \leq \inf \sum_{i=1}^{n} \xi(A_i) \wedge \xi'(A_i) \; .$$

The supremum and the infimum thus agree, and the proof of (3.7) follows from (3.6). □

### 3.0.4 Lipshitz Contraction for Transition Kernels

In this section, we study the contraction property of transition kernels with respect to the total variation distance. Such results have been discussed in a seminal paper by Dobrushin (1956) (see Del Moral, 2004, Chapter 4, for a modern presentation and extensions of these results to a general class of distance-like entropy criteria). Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces and let $K$ be a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$ (see Definition 1). The kernel $K$ is canonically associated to two linear mappings:

(i) a mapping $\mathrm{M}(\mathsf{X}, \mathcal{X}) \to \mathrm{M}(\mathsf{Y}, \mathcal{Y})$ that maps any $\xi$ in $\mathrm{M}(\mathsf{X}, \mathcal{X})$ to a (possibly signed) measure $\xi K$ given by $\xi K(A) = \int_{\mathsf{X}} \xi(dx)\, K(x, A)$ for any $A \in \mathcal{Y}$;

(ii) a mapping $\mathcal{F}_{\mathrm{b}}(\mathsf{Y}) \to \mathcal{F}_{\mathrm{b}}(\mathsf{X})$ that maps any $f$ in $\mathcal{F}_{\mathrm{b}}(\mathsf{Y})$ to the function $Kf$ given by $Kf(x) = \int K(x, dy)\, f(y)$.

Here again, with a slight abuse in notation, we use the same notation $K$ for these two mappings. If we equip the spaces $\mathrm{M}(\mathsf{X}, \mathcal{X})$ and $\mathrm{M}(\mathsf{Y}, \mathcal{Y})$ with the total variation norm and the spaces $\mathcal{F}_{\mathrm{b}}(\mathsf{X})$ and $\mathcal{F}_{\mathrm{b}}(\mathsf{Y})$ with the supremum norm, a first natural problem is to compute the operator norm(s) of the kernel $K$.

**Lemma 44.** *Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces and let $K$ be a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Then*

$$\begin{aligned} 1 &= \sup \{ \|\xi K\|_{\mathrm{TV}} : \xi \in \mathrm{M}(\mathsf{X}, \mathcal{X}), \|\xi\|_{\mathrm{TV}} = 1 \} \\ &= \sup \{ \|Kf\|_{\infty} : f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y}), \|f\|_{\infty} = 1 \} \; . \end{aligned}$$

*Proof.* By Lemma 41,

$$\begin{aligned} &\sup \{ \|\xi K\|_{\mathrm{TV}} : \xi \in \mathrm{M}(\mathsf{X}, \mathcal{X}), \|\xi\|_{\mathrm{TV}} = 1 \} \\ &\quad = \sup \{ \xi K f : \xi \in \mathrm{M}(\mathsf{X}, \mathcal{X}), f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y}), \|f\|_{\infty} = 1, \|\xi\|_{\mathrm{TV}} = 1 \} \\ &\quad = \sup \{ \|Kf\|_{\infty} : f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y}, \mathcal{Y}), \|f\|_{\infty} = 1 \} \leq 1 \; . \end{aligned}$$

If $\xi$ is a probability measure then so is $\xi K$. Because the total variation of any probability measure is one, we see that the left-hand side of this display is indeed equal to one. Thus all members equate to one, and the proof is complete. □

To get sharper results, we will have to consider $K$ as an operator acting on a smaller set of finite measures than $\mathrm{M}(\mathsf{X}, \mathcal{X})$. Of particular interest is the subset $\mathrm{M}_0(\mathsf{X}, \mathcal{X})$ of signed measures with zero total mass. Note that if $\xi$ lies in this subset, then $\xi K$ is in $\mathrm{M}_0(\mathsf{Y}, \mathcal{Y})$. Below we will bound the operator norm of the restriction of the operator $K$ to $\mathrm{M}_0(\mathsf{X}, \mathcal{X})$.

**Definition 45** (Dobrushin Coefficient)**.** *Let $K$ be a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Its* Dobrushin coefficient *$\delta(K)$ is given by*

$$\begin{aligned}\delta(K) &= \frac{1}{2} \sup_{(x,x') \in \mathsf{X}\times\mathsf{X}} \|K(x,\cdot) - K(x',\cdot)\|_{\mathrm{TV}} \\ &= \sup_{(x,x') \in \mathsf{X}\times\mathsf{X}, x\neq x'} \frac{\|K(x,\cdot) - K(x',\cdot)\|_{\mathrm{TV}}}{\|\delta_x - \delta_{x'}\|_{\mathrm{TV}}} \ .\end{aligned}$$

We remark that as $K(x,\cdot)$ and $K(x',\cdot)$ are probability measures, it holds that $\|K(x,\cdot)\|_{\mathrm{TV}} = \|K(x',\cdot)\|_{\mathrm{TV}} = 1$. Hence $\delta(K) \leq \frac{1}{2}(1+1) = 1$, so that the Dobrushin coefficient satisfies $0 \leq \delta(K) \leq 1$.

**Lemma 46.** *Let $\xi$ be a finite signed measure on $(\mathsf{X}, \mathcal{X})$ and let $K$ be a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. Then*

$$\|\xi K\|_{\mathrm{TV}} \leq \delta(K) \|\xi\|_{\mathrm{TV}} + (1 - \delta(K)) |\xi(\mathsf{X})| \ . \tag{3.9}$$

*Proof.* Pick $\xi \in \mathrm{M}(\mathsf{X}, \mathcal{X})$ and let, as usual, $\xi_+$ and $\xi_-$ be its positive and negative part, respectively. If $\xi_-(\mathsf{X}) = 0$ ($\xi$ is a measure), then $\|\xi\|_{\mathrm{TV}} = \xi(\mathsf{X})$ and (3.9) becomes $\|\xi K\|_{\mathrm{TV}} \leq \|\xi\|_{\mathrm{TV}}$; this follows from Lemma 44. If $\xi_+(\mathsf{X}) = 0$, an analogous argument applies.

Thus assume that both $\xi_+$ and $\xi_-$ are non-zero. In view of Lemma 41(ii), it suffices to prove that for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y})$ with $\|f\|_\infty = 1$,

$$|\xi K f| \leq \delta(K)(\xi_+(\mathsf{X}) + \xi_-(\mathsf{X})) + (1 - \delta(K))|\xi_+(\mathsf{X}) - \xi_-(\mathsf{X})| \ . \tag{3.10}$$

We shall suppose that $\xi_+(\mathsf{X}) \geq \xi_-(\mathsf{X})$, if not, replace $\xi$ by $-\xi$ and (3.10) remains the same. Then as $|\xi_+(\mathsf{X}) - \xi_-(\mathsf{X})| = \xi_+(\mathsf{X}) - \xi_-(\mathsf{X})$, (3.10) becomes

$$|\xi K f| \leq 2\xi_-(\mathsf{X})\delta(K) + \xi_+(\mathsf{X}) - \xi_-(\mathsf{X}) \ . \tag{3.11}$$

Now, by Lemma 42, for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{Y})$ it holds that

$$\begin{aligned}|\xi K f| &\leq \sup_{(x,x') \in \mathsf{X}\times\mathsf{X}} |\xi_+(\mathsf{X}) K f(x) - \xi_-(\mathsf{X}) K f(x')| \\ &\leq \sup_{(x,x') \in \mathsf{X}\times\mathsf{X}} \|\xi_+(\mathsf{X}) K(x,\cdot) - \xi_-(\mathsf{X}) K(x',\cdot)\|_{\mathrm{TV}} \ \|f\|_\infty \ .\end{aligned}$$

Finally (3.11) follows upon noting that

$$\begin{aligned}&\|\xi_+(\mathsf{X}) K(x,\cdot) - \xi_-(\mathsf{X}) K(x',\cdot)\|_{\mathrm{TV}} \\ &\quad\leq \ \xi_-(\mathsf{X}) \|K(x,\cdot) - K(x',\cdot)\|_{\mathrm{TV}} + [\xi_+(\mathsf{X}) - \xi_-(\mathsf{X})] \|K(x,\cdot)\|_{\mathrm{TV}} \\ &\quad= \ 2\xi_-(\mathsf{X})\delta(K) + \xi_+(\mathsf{X}) - \xi_-(\mathsf{X}) \ .\end{aligned}$$

□

**Corollary 47.**

$$\delta(K) = \sup \{\|\xi K\|_{\mathrm{TV}} : \xi \in \mathrm{M}_0(\mathsf{X}, \mathcal{X}), \ \|\xi\|_{\mathrm{TV}} \leq 1\} \ . \tag{3.12}$$

*Proof.* If $\xi(\mathsf{X}) = 0$, then (3.9) becomes $\|\xi K\|_{\mathrm{TV}} \leq \delta(K)\,\|\xi\|_{\mathrm{TV}}$, showing that

$$\sup\{\|\xi K\|_{\mathrm{TV}} : \xi \in \mathrm{M}_0(\mathsf{X},\mathcal{X}),\ \|\xi\|_{\mathrm{TV}} \leq 1\} \leq \delta(K)\ .$$

The converse inequality is obvious, as

$$\delta(K) = \sup\left\{(x,x') \in \mathsf{X}\times\mathsf{X}, \left\|\frac{1}{2}(\delta_x - \delta_{x'})K\right\|_{\mathrm{TV}}\right\} \leq \sup\{\|\xi K\|_{\mathrm{TV}} : \xi \in \mathrm{M}_0(\mathsf{X},\mathcal{X}),\ \|\xi\|_{\mathrm{TV}} = 1\}\ .$$

□

If $\xi$ and $\xi'$ are two probability measures on $(\mathsf{X},\mathcal{X})$, Corollary 47 implies that

$$\|\xi K - \xi' K\|_{\mathrm{TV}} \leq \delta(K)\,\|\xi - \xi'\|_{\mathrm{TV}}\ .$$

Thus the Dobrushin coefficient is the norm of $K$ considered as a linear operator from $\mathrm{M}_0(\mathsf{X},\mathcal{X})$ to $\mathrm{M}_0(\mathsf{Y},\mathcal{Y})$.

**Proposition 48.** *The Dobrushin coefficient is sub-multiplicative. That is, if $K : (\mathsf{X},\mathcal{X}) \to (\mathsf{Y},\mathcal{Y})$ and $R : (\mathsf{Y},\mathcal{Y}) \to (\mathsf{Z},\mathcal{Z})$ are two transition kernels, then $\delta(KR) \leq \delta(K)\delta(R)$.*

*Proof.* This is a direct consequence of the fact that the Dobrushin coefficient is an operator norm. By Corollary 47, if $\xi \in \mathrm{M}_0(\mathsf{X},\mathcal{X})$, then $\xi K \in \mathrm{M}_0(\mathsf{Y},\mathcal{Y})$ and $\|\xi K\|_{\mathrm{TV}} \leq \delta(K)\,\|\xi\|_{\mathrm{TV}}$. Likewise, $\|\nu R\|_{\mathrm{TV}} \leq \delta(R)\,\|\nu\|_{\mathrm{TV}}$ holds for any $\nu \in \mathrm{M}_0(\mathsf{Y},\mathcal{Y})$. Thus

$$\|\xi KR\|_{\mathrm{TV}} = \|(\xi K)R\|_{\mathrm{TV}} \leq \delta(R)\,\|\xi K\|_{\mathrm{TV}} \leq \delta(K)\delta(R)\,\|\xi\|_{\mathrm{TV}}$$

□

### 3.0.5 The Doeblin Condition and Uniform Ergodicity

Anticipating results on general state-space Markov chains presented in Chapter 7, we will establish, using the contraction results developed in the previous section, some ergodicity results for a class of Markov chains $(\mathsf{X},\mathcal{X})$ satisfying the so-called Doeblin condition.

**Assumption 49** (Doeblin Condition)**.** *There exist an integer $m \geq 1$, $\epsilon \in (0,1)$, and a transition kernel $\nu = \{\nu_{x,x'},\ (x,x') \in \mathsf{X}\times\mathsf{X}\}$ from $(\mathsf{X}\times\mathsf{X}, \mathcal{X}\otimes\mathcal{X})$ to $(\mathsf{X},\mathcal{X})$ such that for all $(x,x') \in \mathsf{X}\times\mathsf{X}$ and $A \in \mathcal{X}$,*

$$Q^m(x,A) \wedge Q^m(x',A) \geq \epsilon\nu_{x,x'}(A)\ .$$

We will frequently consider a strengthened version of this assumption.

**Assumption 50** (Doeblin Condition Reinforced)**.** *There exist an integer $m \geq 1$, $\epsilon \in (0,1)$, and a probability measure $\nu$ on $(\mathsf{X},\mathcal{X})$ such that for any $x \in \mathsf{X}$ and $A \in \mathcal{X}$,*

$$Q^m(x,A) \geq \epsilon\nu(A)\ .$$

By Lemma 43, the Dobrushin coefficient of $Q^m$ may be equivalently written as

$$\delta(Q^m) = 1 - \inf\sum_{i=1}^{n} Q^m(x,A_i) \wedge Q^m(x',A_i)\ , \tag{3.13}$$

where the infimum is taken over all $(x,x') \in \mathsf{X}\times\mathsf{X}$ and all finite measurable partitions $A_1,\ldots,A_n$ of $\mathsf{X}$ of $\mathsf{X}$. Under the Doeblin condition, the sum in this display is bounded from below by $\epsilon\sum_{i=1}^{n}\nu_{x,x'}(A_i) = \epsilon$. Hence the following lemma is true.

**Lemma 51.** *Under Assumption 49, $\delta(Q^m) \leq 1 - \epsilon$.*

Stochastic processes that are such that for any $k$, the distribution of the random vector $(X_n, \ldots, X_{n+k})$ does not depend on $n$ are called *stationary* (see Definition 10). It is clear that in general a Markov chain will not be stationary. Nevertheless, given a transition kernel $Q$, it is possible that with an appropriate choice of the initial distribution $\nu$ we may produce a stationary process. Assuming that such a distribution exists, the stationarity of the marginal distribution implies that $\mathrm{E}_\nu[\mathbb{1}_A(X_0)] = \mathrm{E}_\nu[\mathbb{1}_A(X_1)]$ for any $A \in \mathcal{X}$. This can equivalently be written as $\nu(A) = \nu Q(A)$, or $\nu = \nu Q$. In such a case, the Markov property implies that all finite-dimensional distributions of $\{X_k\}_{k\geq 0}$ are also invariant under translation in time. These considerations lead to the definition of *invariant measure*.

**Definition 52** (Invariant Measure)**.** *If $Q$ is a Markov kernel on $(\mathsf{X}, \mathcal{X})$ and $\pi$ is a $\sigma$-finite measure satisfying $\pi Q = \pi$, then $\pi$ is called an* invariant measure*.*

If an invariant measure is finite, it may be normalized to an *invariant probability measure*. In practice, this is the main situation of interest. If an invariant measure has infinite total mass, its probabilistic interpretation is much more difficult. In general, there may exist more than one invariant measure, and if $\mathsf{X}$ is not finite, an invariant measure may not exist. As a trivial example, consider $\mathsf{X} = \mathbb{N}$ and $Q(x, x+1) = 1$.

Invariant probability measures are important not merely because they define stationary processes. Invariant probability measures also define the long-term or *ergodic* behavior of a stationary Markov chain. Assume that for some initial measure $\nu$, the sequence of probability measures $\{\nu Q^n\}_{n\geq 0}$ converges to a probability measure $\gamma_\nu$ in total variation norm. This implies that for any function $f \in \mathcal{F}_\mathrm{b}(\mathsf{X})$, $\lim_{n\to\infty} \nu Q^n(f) = \gamma_\nu(f)$. Therefore

$$\begin{aligned}\gamma_\nu(f) = \lim_{n\to\infty} \iint \nu(dx)\, Q^n(x, dx')\, f(x') \\ = \lim_{n\to\infty} \iint \nu(dx)\, Q^{n-1}(x, dx')\, Qf(x') = \gamma_\nu(Qf)\,.\end{aligned}$$

Hence, if a limiting distribution exists, it is an invariant probability measure, and if there exists a unique invariant probability measure, then the limiting distribution $\gamma_\nu$ will be independent of $\nu$, whenever it exists. These considerations lead to the following definitions.

**Definition 53.** *Let $Q$ be a Markov kernel admitting a unique invariant probability measure $\pi$. The chain is said to be* ergodic *if for all $x$ in a set $A \in \mathcal{X}$ such that $\pi(A) = 1$, $\lim_{n\to\infty} \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} = 0$. It is said to be* uniformly ergodic *if $\lim_{n\to\infty} \sup_{x\in\mathsf{X}} \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} = 0$.*

Note that when a chain is uniformly ergodic, it is indeed uniformly geometrically ergodic because $\lim_{n\to\infty} \sup_{x\in\mathsf{X}} \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} = 0$ implies that there exists an integer $m$ such that $\frac{1}{2}\sup_{(x,x')\in\mathsf{X}\times\mathsf{X}} \|Q^m(x,\cdot) - Q^m(x',\cdot)\|_{\mathrm{TV}} < 1$ by the triangle inequality. Hence the Dobrushin coefficient $\delta(Q^m)$ is strictly less than 1, and $Q^m$ is contractive with respect to the total variation distance by Lemma 46. Thus there exist constants $C < \infty$ and $\rho \in [0,1)$ such that $\sup_{x\in\mathsf{X}} \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq C\rho^n$ for all $n$.

The following result shows that if a power $Q^m$ of the Markov kernel $Q$ satisfies Doeblin's condition, then the chain admits a unique invariant probability and is uniformly ergodic.

**Theorem 54.** *Under Assumption 49, $Q$ admits a unique invariant probability measure $\pi$. In addition, for any $\xi \in \mathrm{M}_1(\mathsf{X}, \mathcal{X})$,*

$$\|\xi Q^n - \pi\|_{\mathrm{TV}} \leq (1-\epsilon)^{\lfloor n/m \rfloor} \ \|\xi - \pi\|_{\mathrm{TV}} \ ,$$

*where $\lfloor u \rfloor$ is the integer part of $u$.*

*Proof.* Let $\xi$ and $\xi'$ be two probability measures on $(\mathsf{X}, \mathcal{X})$. Corollary 47, Proposition 48, and Lemma 51 yield that for all $k \geq 1$,

$$\left\|\xi Q^{km} - \xi' Q^{km}\right\|_{\mathrm{TV}} \leq \delta^k(Q^m) \, \|\xi - \xi'\|_{\mathrm{TV}} \leq (1-\epsilon)^k \, \|\xi - \xi'\|_{\mathrm{TV}} \ . \tag{3.14}$$

Taking $\xi' = \xi Q^{pm}$, we find that

$$\left\|\xi Q^{km} - \xi Q^{(k+p)m}\right\|_{\mathrm{TV}} \leq (1-\epsilon)^k \ ,$$

showing that $\{\xi Q^{km}\}$ is a Cauchy sequence in $\mathrm{M}_1(\mathsf{X}, \mathcal{X})$ endowed with the total variation norm. Because this metric space is complete, there exists a probability measure $\pi$ such that $\xi Q^{km} \to \pi$. In view of the discussion above, $\pi$ is invariant for $Q^m$. Moreover, by (3.14) this limit does not depend on $\xi$. Thus $Q^m$ admits $\pi$ as unique invariant probability measure. The Chapman-Kolmogorov equations imply that $(\pi Q)Q^m = (\pi Q^m)Q = \pi Q$, showing that $\pi Q$ is also invariant for $Q^m$ and hence that $\pi Q = \pi$ as claimed. □

**Remark 55.** Classical uniform convergence to equilibrium for Markov processes has been studied during the first half of the 20th century by Doeblin, Kolmogorov, and Doob under various conditions. Doob (1953) gave a unifying form to these conditions, which he named *Doeblin type conditions*. More recently, starting in the 1970s, an increasing interest in non-uniform convergence of Markov processes has arisen. An explanation for this interest is that many useful processes do not converge uniformly to equilibrium, while they do satisfy weaker properties such as a geometric convergence. It later became clear that non-uniform convergence relates to *local* Doeblin type condition and to hitting times for so-called *small sets*. These types of conditions are detailed in Chapter 7.

### 3.0.6 Forgetting Properties

Recall from Chapter 2 that the smoothing probability $\phi_{\nu,k|n}[Y_{0:n}]$ is defined by

$$\phi_{\nu,k|n}[Y_{0:n}](f) = \mathrm{E}_\nu[f(X_k) \,|\, Y_{0:n}] \ , \qquad f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}) \ .$$

Here, $k$ and $n$ are integers, and $\nu$ is the initial probability measure on $(\mathsf{X}, \mathcal{X})$. The filtering probability is defined by $\phi_{\nu,n}[Y_{0:n}] = \phi_{\nu,n|n}[Y_{0:n}]$. In this section, we will establish that under appropriate conditions on the transition kernel $Q$ and on the function $g$, the sequence of filtering probabilities satisfies a property referred to in the literature as "forgetting of the initial condition". This property can be formulated as follows: given two probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,

$$\lim_{n\to\infty} \|\phi_{\nu,n}[Y_{0:n}] - \phi_{\nu',n}[Y_{0:n}]\|_{\mathrm{TV}} = 0 \quad \mathrm{P}_{\nu_\star}\text{-a.s.} \tag{3.15}$$

where $\nu_\star$ is the initial probability measure that defines the law of the observations $\{Y_k\}$. Forgetting is also a concept that applies to the smoothing distributions, as it is often possible to extend the previous results showing that

$$\lim_{k\to\infty} \sup_{n\geq 0} \left\|\phi_{\nu,k|n}[Y_{0:n}] - \phi_{\nu',k|n}[Y_{0:n}]\right\|_{\mathrm{TV}} = 0 \quad \mathrm{P}_{\nu_\star}\text{-a.s.} \tag{3.16}$$

Equation (3.16) can also be strengthened by showing that, under additional conditions, the forgetting property is uniform with respect to the observed sequence $Y_{0:n}$ in the sense that there exists a deterministic sequence $\{\rho_k\}$ satisfying $\rho_k \to 0$ and

$$\sup_{y_{0:n} \in \mathsf{Y}^{n+1}} \sup_{n \geq 0} \left\| \phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}] \right\|_{\mathrm{TV}} \leq \rho_k \ .$$

Several of the results to be proven in the sequel are of this latter type (uniform forgetting).

As shown in (2.5), the smoothing distribution is defined as the ratio

$$\phi_{\nu,k|n}[y_{0:n}](f) = \frac{\int \cdots \int f(x_k)\, \nu(dx_0)\, g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i)\, g(x_i, y_i)}{\int \cdots \int \nu(dx_0)\, g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i)\, g(x_i, y_i)} \ .$$

Therefore, the mapping associating the probability measure $\nu \in \mathrm{M}_1(\mathsf{X}, \mathcal{X})$ to the probability measure $\phi_{\nu,k|n}[y_{0:n}]$ is non-linear. The theory developed above allows one to separately control the numerator and the denominator of this quantity but does not lend a direct proof of the forgetting properties (3.15) or (3.16). To achieve this, we use the alternative representation of the smoothing probability $\phi_{\nu,k|n}[y_{0:n}]$ introduced in Proposition 33, which states that

$$\begin{aligned} \phi_{\nu,k|n}[y_{0:n}](f) &= \int \cdots \int \phi_{\nu,0|n}[y_{0:n}](dx_0) \prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}](x_{i-1}, dx_i)\, f(x_k) \\ &= \phi_{\nu,0|n}[y_{0:n}] \prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] f \ . \end{aligned} \tag{3.17}$$

Here we have used the following notations and definitions from Chapter 2.

(i) $\mathrm{F}_{i|n}[y_{i+1:n}]$ are the *forward smoothing kernels* (see Definition 30) given for $i = 0, \ldots, n-1$, $x \in \mathsf{X}$ and $A \in \mathcal{X}$, by

$$\begin{aligned} \mathrm{F}_{i|n}[y_{i+1:n}](x, A) &\stackrel{\mathrm{def}}{=} \left(\beta_{i|n}[y_{i+1:n}](x)\right)^{-1} \\ &\qquad \times \int_A Q(x, dx_{i+1})\, g(x_{i+1}, y_{i+1}) \beta_{i+1|n}[y_{i+2:n}](x_{i+1}) \ , \end{aligned} \tag{3.18}$$

where $\beta_{i|n}[y_{i+1:n}](x)$ are the backward functions (see Definition 20)

$$\begin{aligned} &\beta_{i|n}[y_{i+1:n}](x) = \\ &\qquad \int Q(x, dx_{i+1})\, g(x_{i+1}, y_{i+1}) \beta_{i+1|n}[y_{i+2:n}](x_{i+1}) \ . \end{aligned} \tag{3.19}$$

Recall that, by Proposition 31, $\{\mathrm{F}_{i|n}\}_{i \geq 0}$ are the transition kernels of the non-homogeneous Markov chain $\{X_k\}$ conditionally on $Y_{0:n}$,

$$\mathrm{E}_\nu[f(X_{i+1}) \,|\, X_{0:i}, Y_{0:n}] = \mathrm{F}_{i|n}[Y_{i+1:n}](X_i, f) \ .$$

(ii) $\phi_{\nu,0|n}[y_{0:n}]$ is the posterior distribution of the state $X_0$ conditionally on $Y_{0:n} = y_{0:n}$, defined for any $A \in \mathcal{X}$ by

$$\phi_{\nu,0|n}[y_{0:n}](A) = \frac{\int_A \nu(dx_0)\, g(x_0, y_0) \beta_{0|n}[y_{1:n}](x_0)}{\int \nu(dx_0)\, g(x_0, y_0) \beta_{0|n}[y_{1:n}](x_0)} \ . \tag{3.20}$$

We see that the non-linear mapping $\nu \mapsto \phi_{\nu,k|n}[y_{0:n}]$ is the composition of two mappings on $\mathcal{M}_1(\mathsf{X}, \mathcal{X})$.

(i) The mapping $\nu \mapsto \phi_{\nu,0|n}[y_{0:n}]$, which associates to the initial distribution $\nu$ the posterior distribution of the state $X_0$ given $Y_{0:n} = y_{0:n}$. This mapping consists in applying Bayes' formula, which we write as

$$\phi_{\nu,0|n}[y_{0:n}] = \mathrm{B}[g(\cdot, y_0)\beta_{0|n}[y_{1:n}](\cdot), \nu] \; .$$

Here

$$\mathrm{B}[\phi, \xi](f) = \frac{\int f(x)\phi(x)\,\xi(dx)}{\int \phi(x)\,\xi(dx)} \; , \qquad f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right) \; , \tag{3.21}$$

for any probability measure $\xi$ on $(\mathsf{X}, \mathcal{X})$ and any non-negative measurable function $\phi$ on $\mathsf{X}$. Note that $\mathrm{B}[\phi, \xi]$ is a probability measure on $(\mathsf{X}, \mathcal{X})$. Because of the normalization, this step is non-linear.

(ii) The mapping $\xi \mapsto \xi \prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}]$, which is a linear mapping being defined as product of Markov transition kernels.

For two initial probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$, the difference of the associated smoothing distributions may thus be expressed as

$$\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}] = \\ \left(\mathrm{B}[g(\cdot, y_0)\beta_{0|n}[y_{1:n}], \nu] - \mathrm{B}[g(\cdot, y_0)\beta_{0|n}[y_{1:n}], \nu']\right) \prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \; . \tag{3.22}$$

Note that the function $g(x, y_0)\beta_{0|n}[y_{1:n}](x)$ defined for $x \in \mathsf{X}$ may also be interpreted as the likelihood of the observation $\mathrm{L}_{\delta_x,n}[y_{0:n}]$ when starting from the initial condition $X_0 = x$ (Proposition 23). In the sequel, we use the likelihood notation whenever possible, writing, in addition, $\mathrm{L}_{x,n}[y_{0:n}]$ rather than $\mathrm{L}_{\delta_x,n}[y_{0:n}]$ and $\mathrm{L}_{\bullet,n}[y_{0:n}]$ when referring to the whole function.

Using Corollary 47, (3.22) implies that

$$\left\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu,k|n}[y_{0:n}]\right\|_{\mathrm{TV}} \leq \\ \left\|\mathrm{B}[\mathrm{L}_{\bullet,n}[y_{0:n}], \nu] - \mathrm{B}[\mathrm{L}_{\bullet,n}[y_{0:n}], \nu']\right\|_{\mathrm{TV}} \delta\left(\prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}]\right) \; , \tag{3.23}$$

where the final factor is a Dobrushin coefficient. Because Bayes operator B returns probability measures, the total variation distance in the right-hand side of this display is always bounded by 2. Although this bound may be sufficient, it is often interesting to relate the total variation distance between $\mathrm{B}[\phi, \xi]$ and $\mathrm{B}[\phi, \xi']$ to the total variation distance between $\xi$ and $\xi'$. The following lemma is adapted from (Künsch, 2000)—see also (Del Moral, 2004, Theorem 4.3.1).

**Lemma 56.** *Let $\xi$ and $\xi'$ be two probability measures on $(\mathsf{X}, \mathcal{X})$ and let $\phi$ be a non-negative measurable function such that $\xi(\phi) > 0$ or $\xi'(\phi) > 0$. Then*

$$\left\|\mathrm{B}[\phi, \xi] - \mathrm{B}[\phi, \xi']\right\|_{\mathrm{TV}} \leq \frac{\|\phi\|_\infty}{\xi(\phi) \vee \xi'(\phi)} \left\|\xi - \xi'\right\|_{\mathrm{TV}} \; . \tag{3.24}$$

*Proof.* We may assume, without loss of generality, that $\xi(\phi) \geq \xi'(\phi)$. For any $f \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}\right)$,

$$\begin{aligned} &\mathrm{B}[\phi, \xi](f) - \mathrm{B}[\phi, \xi'](f) \\ &\quad = \frac{\int f(x)\phi(x)\,(\xi - \xi')(dx)}{\int \phi(x)\,\xi(dx)} + \frac{\int f(x)\phi(x)\,\xi'(dx)}{\int \phi(x)\,\xi'(dx)} \frac{\int \phi(x)\,(\xi' - \xi)(dx)}{\int \phi(x)\,\xi(dx)} \\ &\quad = \frac{1}{\xi(\phi)} \int (\xi - \xi')(dx)\,\phi(x)(f(x) - \mathrm{B}[\phi, \xi'](f)) \; . \end{aligned}$$

By Lemma 43,

$$\left| \int (\xi - \xi')(dx)\, \phi(x)(f(x) - \mathrm{B}[\phi, \xi'](f)) \right| \leq \|\xi - \xi'\|_{\mathrm{TV}} \times \frac{1}{2} \sup_{(x,x') \in \mathsf{X} \times \mathsf{X}} |\phi(x)(f(x) - \mathrm{B}[\phi, \xi'](f)) - \phi(x')(f(x') - \mathrm{B}[\phi, \xi'](f))| \ .$$

Because $|\mathrm{B}[\phi, \xi'](f)| \leq \|f\|_\infty$ and $\phi \geq 0$, the supremum on the right-hand side of this display is bounded by $2\,\|\phi\|_\infty\,\|f\|_\infty$. This concludes the proof. □

As mentioned by Künsch (2000), the Bayes operator may be non-contractive: the numerical factor in the right-hand side of (3.24) is sometimes larger than one and the bound may be shown to be tight on particular examples. The intuition that the posteriors should at least be as close as the priors if the same likelihood (the same data) is applied is thus generally wrong.

Equation (3.17) also implies that for any integer $j$ such that $j \leq k$,

$$\begin{aligned} \phi_{\nu,k|n}[y_{0:n}] &= \phi_{\nu,0|n}[y_{0:n}] \prod_{i=1}^{j} \mathrm{F}_{i-1|n}[y_{i:n}] \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \\ &= \phi_{\nu,j|n}[y_{0:n}] \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \ . \end{aligned} \tag{3.25}$$

This decomposition and Corollary 47 shows that for any $0 \leq j \leq k$, any initial distributions $\nu$ and $\nu'$ and any sequence $y_{0:n}$ such that $\mathrm{L}_{\nu,n}[y_{0:n}] > 0$ and $\mathrm{L}_{\nu',n}[y_{0:n}] > 0$,

$$\begin{aligned} &\left\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}]\right\|_{\mathrm{TV}} \\ &\qquad \leq \delta \left( \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \right) \left\|\phi_{\nu,j|n}[y_{0:n}] - \phi_{\nu',j|n}[y_{0:n}]\right\|_{\mathrm{TV}} \ . \end{aligned}$$

Because the Dobrushin coefficient of a Markov kernel is bounded by one, this relation implies that the total variation distance between the smoothing distributions associated with two different initial distributions is non-expanding. To summarize this discussion, we have obtained the following result.

**Proposition 57.** *Let $\nu$ and $\nu'$ be two probability measures on $(\mathsf{X}, \mathcal{X})$. For any non-negative integers $j$, $k$, and $n$ such that $j \leq k$ and any sequence $y_{0:n} \in \mathsf{Y}^{n+1}$ such that $\mathrm{L}_{\nu,n}[y_{0:n}] > 0$ and $\mathrm{L}_{\nu',n}[y_{0:n}] > 0$,*

$$\begin{aligned} &\left\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}]\right\|_{\mathrm{TV}} \\ &\qquad \leq \delta \left( \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \right) \left\|\phi_{\nu,j|n}[y_{0:n}] - \phi_{\nu',j|n}[y_{0:n}]\right\|_{\mathrm{TV}} \ , \end{aligned} \tag{3.26}$$

$$\begin{aligned} &\left\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}]\right\|_{\mathrm{TV}} \\ &\qquad \leq \frac{\|\mathrm{L}_{\bullet,n}[y_{0:n}]\|_\infty}{\mathrm{L}_{\nu,n}[y_{0:n}] \vee \mathrm{L}_{\nu',n}[y_{0:n}]} \delta \left( \prod_{i=1}^{k} \mathrm{F}_{i-1|n}[y_{i:n}] \right) \|\nu - \nu'\|_{\mathrm{TV}} \ . \end{aligned} \tag{3.27}$$

Along the same lines, we can compare the posterior distribution of the state $X_k$ given observations $Y_{j:n}$ for different values of $j$. To avoid introducing new notations, we will simply denote these conditional distributions by $\mathrm{P}_\nu\,(X_k \in \cdot \,|\, Y_{j:n} = y_{j:n})$.

As mentioned in the introduction of this chapter, it is sensible to expect that $\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{j:n}\right)$ gets asymptotically close to $\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{0:n}\right)$ as $k-j$ tends to infinity. Here again, to establish this alternative form of the forgetting property, we will use a representation of $\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{j:n}\right)$ similar to (3.17).

Because $\{(X_k, Y_k)\}$ is a Markov chain, and assuming that $k \geq j$,

$$\mathrm{P}_\nu\left(X_k \in \cdot \,|\, X_j, Y_{j:n}\right) = \mathrm{P}_\nu\left(X_k \in \cdot \,|\, X_j, Y_{0:n}\right) \ .$$

Moreover, we know that conditionally on $Y_{0:n}$, $\{X_k\}$ is a non-homogeneous Markov chain with transition kernels $\mathrm{F}_{k|n}[Y_{k+1:n}]$ where $\mathrm{F}_{i|n} = Q$ for $i \geq n$ (Proposition 31). Therefore the Chapman-Kolmogorov equations show that for any function $f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}\right)$,

$$\begin{aligned}\mathrm{E}_\nu\left[f(X_k) \,|\, Y_{j:n}\right] &= \mathrm{E}_\nu\left[\mathrm{E}_\nu\left[f(X_k) \,|\, X_j, Y_{j:n}\right] \,|\, Y_{j:n}\right] \\ &= \mathrm{E}_\nu\left[\prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[Y_{i:n}] f(X_j) \,\middle|\, Y_{j:n}\right] = \tilde{\phi}_{\nu,j|n}[Y_{j:n}] \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[Y_{i:n}] f \ ,\end{aligned}$$

cf. (3.25), where the probability measure $\tilde{\phi}_{\nu,j|n}[Y_{j:n}(f)]$ is defined by

$$\tilde{\phi}_{\nu,j|n}[Y_{j:n}](f) = \mathrm{E}_\nu[f(X_j) \,|\, Y_{j:n}] \ , \qquad f \in \mathcal{F}_\mathrm{b}\left(\mathsf{X}\right) \ .$$

Using (3.25) as well, we thus find that the difference between $\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{j:n}\right)$ and $\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{0:n}\right)$ may be expressed by

$$\mathrm{E}_\nu[f(X_k) \,|\, Y_{j:n}] - \mathrm{E}_\nu[f(X_k) \,|\, Y_{0:n}] = (\tilde{\phi}_{\nu,j|n} - \phi_{\nu,j|n}) \prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[Y_{i:n}] f \ .$$

Proceeding like in Proposition 57, we may thus derive a bound on the total variation distance between these probability measures.

**Proposition 58.** *For any integers $j$, $k$, and $n$ such that $0 \leq j \leq k$ and any probability measure $\nu$ on $(\mathsf{X}, \mathcal{X})$,*

$$\left\|\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{0:n}\right) - \mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{j:n}\right)\right\|_\mathrm{TV} \leq 2\delta \left(\prod_{i=j+1}^{k} \mathrm{F}_{i-1|n}[Y_{i:n}]\right) \ . \tag{3.28}$$

### 3.0.7 Uniform Forgetting Under Strong Mixing Conditions

In light of the discussion above, establishing forgetting properties amounts to determining non-trivial bounds on the Dobrushin coefficient of products of forward transition kernels and, if required, on ratio of likelihoods $\mathrm{L}_{x,n}(y_{0:n})/(\mathrm{L}_{\nu,n}(y_{0:n}) \vee \mathrm{L}_{\nu',n}(y_{0:n}))$. To do so, we need to impose additional conditions on $Q$ and $g$. We consider in this section the following assumption, which was introduced by Le Gland and Oudjane (2004, Section 2).

**Assumption 59** (Strong Mixing Condition)**.** *There exist a transition kernel $K : (\mathsf{Y}, \mathcal{Y}) \to (\mathsf{X}, \mathcal{X})$ and measurable functions $\varsigma^-$ and $\varsigma^+$ from $\mathsf{Y}$ to $(0, \infty)$ such that for any $A \in \mathcal{X}$ and $y \in \mathsf{Y}$,*

$$\varsigma^-(y) K(y, A) \leq \int_A Q(x, dx')\, g(x', y) \leq \varsigma^+(y) K(y, A) \ . \tag{3.29}$$

We first show that under this condition, one may derive a non-trivial upper bound on the Dobrushin coefficient of the forward smoothing kernels.

**Lemma 60.** *Under Assumption 59, the following hold true.*

*(i) For any non-negative integers $k$ and $n$ such that $k < n$ and $x \in \mathsf{X}$,*

$$\prod_{j=k+1}^{n} \varsigma^-(y_j) \le \beta_{k|n}[y_{k+1:n}](x) \le \prod_{j=k+1}^{n} \varsigma^+(y_j) \,. \tag{3.30}$$

*(ii) For any non-negative integers $k$ and $n$ such that $k < n$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,*

$$\frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})} \le \frac{\int_{\mathsf{X}} \nu(dx)\, \beta_{k|n}[y_{k+1:n}](x)}{\int_{\mathsf{X}} \nu'(dx)\, \beta_{k|n}[y_{k+1:n}](x)} \le \frac{\varsigma^+(y_{k+1})}{\varsigma^-(y_{k+1})} \,.$$

*(iii) For any non-negative integers $k$ and $n$ such that $k < n$, there exists a transition kernel $\lambda_{k,n}$ from $(\mathsf{Y}^{n-k}, \mathcal{Y}^{\otimes(n-k)})$ to $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$, $A \in \mathcal{X}$, and $y_{k+1:n} \in \mathsf{Y}^{n-k}$,*

$$\begin{aligned} \frac{\varsigma^-(y_{k+1})}{\varsigma^+(y_{k+1})} \lambda_{k,n}(y_{k+1:n}, A) &\le \mathrm{F}_{k|n}[y_{k+1:n}](x, A) \\ &\le \frac{\varsigma^+(y_{k+1})}{\varsigma^-(y_{k+1})} \lambda_{k,n}(y_{k+1:n}, A) \,. \end{aligned} \tag{3.31}$$

*(iv) For any non-negative integers $k$ and $n$, the Dobrushin coefficient of the forward smoothing kernel $\mathrm{F}_{k|n}[y_{k+1:n}]$ satisfies*

$$\delta(\mathrm{F}_{k|n}[y_{k+1:n}]) \le \begin{cases} \rho_0(y_{k+1}) & k < n \,, \\ \rho_1 & k \ge n \,, \end{cases}$$

*where for any $y \in \mathsf{Y}$,*

$$\rho_0(y) \stackrel{\text{def}}{=} 1 - \frac{\varsigma^-(y)}{\varsigma^+(y)} \quad \text{and} \quad \rho_1 \stackrel{\text{def}}{=} 1 - \int \varsigma^-(y)\, \mu(dy) \,. \tag{3.32}$$

*Proof.* Take $A = \mathsf{X}$ in Assumption 59 to see that $\int_{\mathsf{X}} Q(x, dx')\, g(x', y)$ is bounded from above and below by $\varsigma^+(y)$ and $\varsigma^-(y)$, respectively. Part (i) then follows from (2.16).

Next, (2.19) shows that

$$\begin{aligned} &\int \nu(dx)\, \beta_{k|n}[y_{k+1:n}](x) \\ &\qquad = \iint \nu(dx)\, Q(x, dx_{k+1})\, g(x_{k+1}, y_{k+1}) \beta_{k+1|n}[y_{k+2:n}](x_{k+1}) \,. \end{aligned}$$

This expression is bounded from above by

$$\varsigma^+(y_{k+1}) \int K(y_{k+1}, dx_{k+1})\, \beta_{k+1|n}[y_{k+2:n}](x_{k+1}) \,,$$

and similarly a lower bound, with $\varsigma^-(y_{k+1})$ rather than $\varsigma^+(y_{k+1})$, holds too. These bounds are independent of $\nu$, and (ii) follows.

We turn to part (iii). Using the definition (2.30), the forward kernel $\mathrm{F}_{k|n}[y_{k+1:n}]$ may be expressed as

$$\mathrm{F}_{k|n}[y_{k+1:n}](x, A) = \frac{\int_A Q(x, dx_{k+1})\, g(x_{k+1}, y_{k+1}) \beta_{k+1|n}[y_{k+2:n}](x_{k+1})}{\int_{\mathsf{X}} Q(x, dx_{k+1})\, g(x_{k+1}, y_{k+1}) \beta_{k+1|n}[y_{k+2:n}](x_{k+1})} \,.$$

Using arguments as above, (3.31) holds with

$$\lambda_{k,n}(y_{k+1:n}, A) \stackrel{\text{def}}{=} \frac{\int_A K(y_{k+1}, dx_{k+1})\, \beta_{k+1|n}[y_{k+2:n}](x_{k+1})}{\int_{\mathsf{X}} K(y_{k+1}, dx_{k+1})\, \beta_{k+1|n}[y_{k+2:n}](x_{k+1})} \ .$$

Finally, part (iv) for $k < n$ follows from part (iii) and Lemma 51. In the opposite case, recall from (2.31) that $\mathrm{F}_{k|n} = Q$ for indices $k \geq n$. Integrating (3.29) with respect to $\mu$ and using $\int g(x,y)\,\mu(dy) = 1$, we find that for any $A \in \mathcal{X}$ and any $x \in \mathsf{X}$,

$$Q(x,A) \geq \int \varsigma^-(y) K(y,A)\,\mu(dy) = \int \varsigma^-(y)\,\mu(dy) \times \frac{\int \varsigma^-(y) K(y,A)\,\mu(dy)}{\int \varsigma^-(y)\,\mu(dy)} \ ,$$

where the ratio on the right-hand side is a probability measure. The proof of part (iv) again follows from Lemma 51. □

The final part of the above lemma shows that under Assumption 59, the Dobrushin coefficient of the transition kernel $Q$ satisfies $\delta(Q) \leq 1 - \epsilon$ for some $\epsilon > 0$. This is in fact a rather stringent assumption, which fails to be satisfied in many of the examples considered in Chapter **??**. When $\mathsf{X}$ is finite, this condition is satisfied if $Q(x,x') \geq \epsilon$ for any $(x,x') \in \mathsf{X} \times \mathsf{X}$. When $\mathsf{X}$ is countable, $\delta(Q) < 1$ is satisfied under the Doeblin condition 49 with $n = 1$. When $\mathsf{X} \subseteq \mathbb{R}^d$ or more generally is a topological space, $\delta(Q) < 1$ typically requires that $\mathsf{X}$ is compact, which is, admittedly, a serious limitation.

**Proposition 61.** *Under 59 the following hold true.*

*(i) For any non-negative integers $k$ and $n$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,*

$$\begin{aligned} &\big\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}]\big\|_{\mathrm{TV}} \\ &\qquad \leq \prod_{j=1}^{k\wedge n} \rho_0(y_j) \times \rho_1^{k-k\wedge n} \big\|\phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}]\big\|_{\mathrm{TV}} \ , \end{aligned}$$

*where $\rho_0$ and $\rho_1$ are defined in* (3.32).

*(ii) For any non-negative integer $n$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$ such that $\int \nu(dx_0)\, g(x_0, y_0) > 0$ and $\int \nu'(dx_0)\, g(x_0, y_0) > 0$,*

$$\begin{aligned} &\big\|\phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}]\big\|_{\mathrm{TV}} \\ &\qquad \leq \frac{\varsigma^+(y_1)}{\varsigma^-(y_1)} \frac{\|g\|_\infty}{\nu(g(\cdot, y_0)) \vee \nu'(g(\cdot, y_0))} \|\nu - \nu'\|_{\mathrm{TV}} \ . \end{aligned}$$

*(iii) For any non-negative integers $j$, $k$, and $n$ such that $j \leq k$ and any probability measure $\nu$ on $(\mathsf{X}, \mathcal{X})$,*

$$\begin{aligned} &\|\mathrm{P}_\nu\,(X_k \in \cdot \,|\, Y_{0:n} = y_{0:n}) - \mathrm{P}_\nu(X_k \in \cdot \,|\, Y_{j:n} = y_{j:n})\|_{\mathrm{TV}} \\ &\qquad \leq 2 \prod_{i=j\wedge n+1}^{k\wedge n} \rho_0(y_i) \times \rho_1^{k-j-(k\wedge n - j\wedge n)} \ . \end{aligned}$$

*Proof.* Using Lemma 60(iv) and Proposition 48, we find that for $j \leq k$,

$$\delta(\mathrm{F}_{j|n}[y_{j+1:n}] \cdots \mathrm{F}_{k|n}[y_{k+1:n}]) \leq \prod_{i=j\wedge n+1}^{k\wedge n} \rho_0(y_i) \times \rho_1^{k-j-(k\wedge n - j\wedge n)} \ .$$

Parts (i) and (iii) then follow from Propositions 57 and 58, respectively. Next we note that (3.20) shows that

$$\phi_{\nu,0|n}[y_{0:n}] = \mathrm{B}\left[\beta_{0|n}[y_{1:n}](\cdot), \mathrm{B}[g(\cdot, y_0), \nu]\right] \ .$$

Apply Lemma 56 twice to this form to arrive at a bound on the total variation norm of the difference $\phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}]$ given by

$$\frac{\left\|\beta_{0|n}[y_{1:n}]\right\|_\infty}{\mathrm{B}[g(\cdot, y_0), \nu](\beta_{0|n}[y_{1:n}])} \times \frac{\|g(\cdot, y_0)\|_\infty}{\nu(g(\cdot, y_0)) \vee \nu'(g(\cdot, y_0))} \|\nu - \nu'\|_{\mathrm{TV}} \ .$$

Finally, bound the first ratio of this display using Lemma 60(ii); the supremum norm is obtained by taking one of the initial measures as an atom at some point $x \in \mathsf{X}$. This completes the proof of part (ii). □

From the above it is clear that forgetting properties stem from properties of the product

$$\prod_{i=j\wedge n+1}^{k\wedge n} \rho_0(Y_i)\rho_1^{k-j-(k\wedge n-j\wedge n)} \ . \tag{3.33}$$

The situation is elementary when the factors of this product are (non-trivially) upper-bounded uniformly with respect to the observations $Y_{0:n}$. To obtain such bounds, we consider the following strengthening of the strong mixing condition, first introduced by Atar and Zeitouni (1997).

**Assumption 62** (Strong Mixing Reinforced)**.** *(i) There exist two positive real numbers $\sigma^-$ and $\sigma^+$ and a probability measure $\kappa$ on $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$ and $A \in \mathcal{X}$,*

$$\sigma^-\kappa(A) \le Q(x, A) \le \sigma^+\kappa(A) \ .$$

*(ii) For all $y \in \mathsf{Y}$, $0 < \int_{\mathsf{X}} \kappa(dx)\, g(x, y) < \infty$.*

It is easily seen that this implies Assumption 59.

**Lemma 63.** *Assumption 62 implies Assumption 59 with $\varsigma^-(y) = \sigma^- \int_{\mathsf{X}} \kappa(dx)\, g(x, y)$, $\varsigma^+(y) = \sigma^+ \int_{\mathsf{X}} \kappa(dx)\, g(x, y)$, and*

$$K(y, A) = \frac{\int_A \kappa(dx)\, g(x, y)}{\int_{\mathsf{X}} \kappa(dx)\, g(x, y)} \ .$$

*In particular, $\varsigma^-(y)/\varsigma^+(y) = \sigma^-/\sigma^+$ for any $y \in \mathsf{Y}$.*

*Proof.* The proof follows immediately upon observing that

$$\sigma^- \int_A \kappa(dx')\, g(x', y) \le \int_A Q(x, dx')\, g(x', y) \le \sigma^+ \int_A \kappa(dx')\, g(x', y) \ .$$

□

Replacing Assumption 59 by Assumption 62, Proposition 61 may be strengthened as follows.

**Proposition 64.** *Under Assumption 62, the following hold true.*

*(i) For any non-negative integers $k$ and $n$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,*

$$\begin{aligned}&\left\|\phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}]\right\|_{\mathrm{TV}} \\ &\qquad \le \left(1 - \frac{\sigma^-}{\sigma^+}\right)^{k\wedge n} (1 - \sigma^-)^{k-k\wedge n} \left\|\phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}]\right\|_{\mathrm{TV}} \ .\end{aligned}$$

*(ii) For any non-negative integer $n$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$ such that $\int \nu(dx_0)\, g(x_0, y_0) > 0$ and $\int \nu'(dx_0)\, g(x_0, y_0) > 0$,*

$$\left\|\phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}]\right\|_{\mathrm{TV}} \leq \frac{\sigma^+}{\sigma^-} \frac{\|g\|_\infty}{\nu[g(\cdot, y_0)] \vee \nu'[g(\cdot, y_0)]} \|\nu - \nu'\|_{\mathrm{TV}} \ .$$

*(iii) For any non-negative integers $j$, $k$, and $n$ such that $j \leq k$ and any probability measure $\nu$ on $(\mathsf{X}, \mathcal{X})$,*

$$\|\mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{0:n} = y_{0:n}\right) - \mathrm{P}_\nu\left(X_k \in \cdot \,|\, Y_{j:n} = y_{j:n}\right)\|_{\mathrm{TV}} \leq 2\left(1 - \frac{\sigma^-}{\sigma^+}\right)^{k\wedge n - j\wedge n} \left(1 - \sigma^-\right)^{k-j-(k\wedge n - j\wedge n)} \ .$$

Thus, under Assumption 62 the filter and the smoother forget their initial conditions exponentially fast, uniformly with respect to the observations. This property, which holds under rather stringent assumptions, plays a key role in the sequel (see for instance Chapters **??** and 6).

Of course, the product (3.33) can be shown to vanish asymptotically under conditions that are less stringent than Assumption 62. A straightforward adaptation of Lemma 63 shows that the following result is true.

**Lemma 65.** *Assume 59 and that there exists a set $C \in \mathcal{Y}$ and constants $0 < \sigma^- \leq \sigma^+ < \infty$ satisfying $\mu(C) > 0$ and, for all $y \in C$, $\sigma^- \leq \varsigma^-(y) \leq \varsigma^+(y) \leq \sigma^+$. Then, $\rho_0(y) \leq 1 - \sigma^-/\sigma^+$, $\rho_1 \geq 1 - \sigma^-\mu(C)$ and*

$$\prod_{i=j\wedge n+1}^{k\wedge n} \rho_0(Y_i)\rho_1^{k-j-(k\wedge n - j\wedge n)} \leq \left(1 - \sigma^-/\sigma^+\right)^{\sum_{i=j\wedge n+1}^{k\wedge n} \mathbb{1}_C(Y_i)} \left[1 - \sigma^-\mu(C)\right]^{k-j-(k\wedge n - j\wedge n)} \ . \quad (3.34)$$

In words, forgetting is guaranteed to occur when $\{Y_k\}$ visits a given set $C$ infinitely often in the long run. Of course, such a property cannot hold true for all possible sequences of observations but it may hold with probability one under appropriate assumptions on the law of $\{Y_k\}$, assuming in particular that the observations are distributed under the model, perhaps with a different initial distribution $\nu_\star$. To answer whether this happens or not requires additional results from the general theory of Markov chains, and we postpone this discussion to Section 7.3 (see in particular Proposition 208 on the recurrence of the joint chain in HMMs).

### 3.0.8 Forgetting Under Alternative Conditions

Because Assumptions 59 and 62 are not satisfied in many contexts of interest, it is worthwhile to consider ways in which these assumptions can be weakened. This happens to raise difficult mathematical challenges that largely remain unsolved today. Perhaps surprisingly, despite many efforts in this direction, there is up to now no truly satisfactory assumption that covers a reasonable fraction of the situations of practical interest. The problem really is more complicated than appears at first sight. In particular, Example 66 below shows that the forgetting property does not necessarily hold under assumptions that imply that the underlying Markov chain is uniformly ergodic. This last section on forgetting is more technical and requires some knowledge of Markov chain theory as can be found in Chapter 7.

**Example 66.** This example was first discussed by Kaijser (1975) and recently worked out by Chigansky and Lipster (2004). Let $\{X_k\}$ be a Markov chain on $\mathsf{X} = \{0, 1, 2, 3\}$, defined by the recurrence equation $X_k = (X_{k-1} + U_k) \bmod 4$, where $\{U_k\}$ is an i.i.d. binary sequence with $\mathrm{P}(B_k = 0) = p$ and $\mathrm{P}(B_k = 1) = 1 - p$ for some $0 < p < 1$. For any $(x, x') \in \mathsf{X} \times \mathsf{X}$, $Q^4(x, x') > 0$, which implies that $\delta(Q^4) < 1$ and, by Theorem 54, that the chain is uniformly geometrically ergodic. The observations $\{Y_k\}$ are a deterministic binary function of the chain, namely

$$Y_k = \mathbb{1}_{\{0,2\}}(X_k) \; .$$

The function mapping $X_k$ to $Y_k$ is not injective, but knowledge of $Y_k$ indicates two possible values of $X_k$. The filtering distribution is given recursively by

$$\begin{aligned}
\phi_{\nu,k}[y_{0:k}](0) &= y_k \left\{\phi_{\nu,k-1}[y_{0:k-1}](0) + \phi_{\nu,k-1}[y_{0:k-1}](3)\right\} \; , \\
\phi_{\nu,k}[y_{0:k}](1) &= (1 - y_k) \left\{\phi_{\nu,k-1}[y_{0:k-1}](1) + \phi_{\nu,k-1}[y_{0:k-1}](0)\right\} \; , \\
\phi_{\nu,k}[y_{0:k}](2) &= y_k \left\{\phi_{\nu,k-1}[y_{0:k-1}](2) + \phi_{\nu,k-1}[y_{0:k-1}](1)\right\} \; , \\
\phi_{\nu,k}[y_{0:k}](3) &= (1 - y_k) \left\{\phi_{\nu,k-1}[y_{0:k-1}](3) + \phi_{\nu,k-1}[y_{0:k-1}](2)\right\} \; .
\end{aligned}$$

In particular, either one of the two sets $\{0, 2\}$ and $\{1, 3\}$ has null probability under $\phi_{\nu,k}[y_{0:k}]$, depending on the value of $y_k$, and irrespectively of the choice of $\nu$. We also notice that

$$\begin{aligned}
y_k \, \phi_{\nu,k}[y_{0:k}](j) &= \phi_{\nu,k}[y_{0:k}](j) \; , && \text{for } j = 0, 2, \\
(1 - y_k) \, \phi_{\nu,k}[y_{0:k}](j) &= \phi_{\nu,k}[y_{0:k}](j) \; , && \text{for } j = 1, 3.
\end{aligned} \tag{3.35}$$

In addition, it is easily checked that, except when $\nu(\{0, 2\})$ or $\nu(\{1, 3\})$ equals 1 (which rules out one of the two possible values for $y_0$), the likelihood $\mathrm{L}_{\nu,n}[y_{0:n}]$ is strictly positive for any integer $n$ and any sequence $y_{0:n} \in \{0, 1\}^{n+1}$.

Dropping the dependence on $y_{0:k}$ for notational simplicity and using (3.35) we obtain

$$\begin{aligned}
&|\phi_{\nu,k}(0) - \phi_{\nu',k}(0)| \\
&\quad = y_k |\phi_{\nu,k-1}(0) - \phi_{\nu',k-1}(0) + \phi_{\nu,k-1}(3) - \phi_{\nu',k-1}(3)| \\
&\quad = y_k \left\{y_{k-1} |\phi_{\nu,k-1}(0) - \phi_{\nu',k-1}(0)| + (1 - y_{k-1}) |\phi_{\nu,k-1}(3) - \phi_{\nu',k-1}(3)|\right\} \; .
\end{aligned}$$

Proceeding similarly, we also find that

$$\begin{aligned}
&|\phi_{\nu,k}(1) - \phi_{\nu',k}(1)| = \\
&(1 - y_k) \left\{(1 - y_{k-1}) |\phi_{\nu,k-1}(1) - \phi_{\nu',k-1}(1)| + y_{k-1} |\phi_{\nu,k-1}(0) - \phi_{\nu',k-1}(0)|\right\} \; , \\
&|\phi_{\nu,k}(2) - \phi_{\nu',k}(2)| = \\
&y_k \left\{y_{k-1} |\phi_{\nu,k-1}(2) - \phi_{\nu',k-1}(2)| + (1 - y_{k-1}) |\phi_{\nu,k-1}(1) - \phi_{\nu',k-1}(1)|\right\} \; , \\
&|\phi_{\nu,k}(3) - \phi_{\nu',k}(3)| = \\
&(1 - y_k) \left\{(1 - y_{k-1}) |\phi_{\nu,k-1}(3) - \phi_{\nu',k-1}(3)| + y_{k-1} |\phi_{\nu,k-1}(2) - \phi_{\nu',k-1}(2)|\right\} \; .
\end{aligned}$$

Adding the above equalities using (3.35) again shows that for any $k = 1, \ldots, n$,

$$\begin{aligned}
\|\phi_{\nu,k}[y_{0:k}] - \phi_{\nu',k}[y_{0:k}]\|_{\mathrm{TV}} &= \|\phi_{\nu,k-1}[y_{0:k-1}] - \phi_{\nu',k-1}[y_{0:k-1}]\|_{\mathrm{TV}} \\
&= \|\phi_{\nu,0}[y_0] - \phi_{\nu',0}[y_0]\|_{\mathrm{TV}} \; .
\end{aligned}$$

By construction, $\phi_{\nu,0}[y_0](j) = y_0 \, \nu(j)/(\nu(0)+\nu(2))$ for $j = 0$ and 2, and $\phi_{\nu,0}[y_0](j) = (1-y_0) \, \nu(j)/(\nu(1)+\nu(3))$ for $j = 1$ and 3. This implies that $\|\phi_{\nu,0}[y_0] - \phi_{\nu',0}[y_0]\|_{\mathrm{TV}} \neq 0$ if $\nu \neq \nu'$.

In this model, the hidden Markov chain $\{X_k\}$ is uniformly ergodic, but the filtering distributions $\phi_{\nu,k}[y_{0:k}]$ never forget the influence of the initial distribution $\nu$, whatever the observed sequence.

In the above example, the kernel $Q$ does not satisfy Assumption 62 with $m = 1$ (one-step minorization), but the condition is verified for a power $Q^m$ (here for $m = 4$). This situation is the rule rather than the exception. In particular, a Markov chain on a finite state space has a unique invariant probability measure and is ergodic if and only if there exists an integer $m > 0$ such that $Q^m(x, x') > 0$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$ (but the condition may not hold for $m = 1$). This suggests considering the following assumption (see for instance Del Moral, 2004, Chapter 4).

**Assumption 67.**

*(i) There exist an integer $m$, two positive real numbers $\sigma^-$ and $\sigma^+$, and a probability measure $\kappa$ on $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$ and $A \in \mathcal{X}$,*

$$\sigma^- \kappa(A) \leq Q^m(x, A) \leq \sigma^+ \kappa(A) \; .$$

*(ii) There exist two measurable functions $g^-$ and $g^-$ from $\mathsf{Y}$ to $(0, \infty)$ such that for any $y \in \mathsf{Y}$,*

$$g^-(y) \leq \inf_{x \in \mathsf{X}} g(x, y) \leq \sup_{x \in \mathsf{X}} g(x, y) \leq g^+(y) \; .$$

Compared to Assumption 62, the condition on the transition kernel has been weakened, but at the expense of strengthening the assumption on the function $g$. Note in particular that part (ii) is *not* satisfied in Example 66.

Using (3.17) and writing $k = jm+r$ with $0 \leq r < m$, we may express $\phi_{\nu,k|n}[y_{0:n}]$ as

$$\phi_{\nu,k|n}[y_{0:n}] = \phi_{\nu,0|n}[y_{0:n}] \prod_{u=0}^{j-1} \left( \prod_{i=um}^{(u+1)m-1} \mathrm{F}_{i|n}[y_{i+1:n}] \right) \prod_{i=jm}^{k-1} \mathrm{F}_{i|n}[y_{i+1:n}] \; .$$

This implies, using Corollary 47, that for any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$ and any sequence $y_{0:n}$ satisfying $\mathrm{L}_{\nu,n}[y_{0:n}] > 0$ and $\mathrm{L}_{\nu',n}[y_{0:n}] > 0$,

$$\begin{aligned} &\left\| \phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}] \right\|_{\mathrm{TV}} \\ &\qquad \leq \prod_{u=0}^{j-1} \delta \left( \prod_{i=um}^{(u+1)m-1} \mathrm{F}_{i|n}[y_{i+1:n}] \right) \left\| \phi_{\nu,0|n}[y_{0:n}] - \phi_{\nu',0|n}[y_{0:n}] \right\|_{\mathrm{TV}} \; . \end{aligned} \tag{3.36}$$

This expression suggest computing a bound on $\delta(\prod_{i=um}^{um+m-1} \mathrm{F}_{i|n}[y_{i+1:n}])$ rather than a bound on $\delta(\mathrm{F}_{i|n})$. The following result shows that such a bound can be derived under Assumption 67.

**Lemma 68.** *Under Assumption 67, the following hold true.*

*(i) For any non-negative integers $k$ and $n$ such that $k < n$ and $x \in \mathsf{X}$,*

$$\prod_{j=k+1}^{n} g^-(y_j) \leq \beta_{k|n}[y_{k+1:n}](x) \leq \prod_{j=k+1}^{n} g^+(y_j) \; , \tag{3.37}$$

*where $\beta_{k|n}$ is the backward function (2.16).*

*(ii) For any non-negative integers $u$ and $n$ such that $0 \leq u < \lfloor n/m \rfloor$ and any probability measures $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,*

$$\frac{\sigma^-}{\sigma^+} \prod_{i=um+1}^{(u+1)m} \frac{g^-(y_i)}{g^+(y_i)} \leq \frac{\int_{\mathsf{X}} \nu(dx) \, \beta_{um|n}[y_{um+1:n}](x)}{\int_{\mathsf{X}} \nu'(dx) \, \beta_{um|n}[y_{um+1:n}](x)} \leq \frac{\sigma^+}{\sigma^-} \prod_{i=um+1}^{(u+1)m} \frac{g^+(y_i)}{g^-(y_i)} \; .$$

*(iii) For any non-negative integers $u$ and $n$ such that $0 \le u < \lfloor n/m \rfloor$, there exists a transition kernel $\lambda_{u,n}$ from $\left(\mathsf{Y}^{(n-(u+1)m)}, \mathcal{Y}^{\otimes(n-(u+1)m)}\right)$ to $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$, $A \in \mathcal{X}$ and $y_{um+1:n} \in \mathsf{Y}^{(n-um)}$,*

$$\frac{\sigma^-}{\sigma^+} \prod_{i=um+1}^{(u+1)m} \frac{g^-(y_i)}{g^+(y_i)} \lambda_{u,n}(y_{(u+1)m+1:n}, A) \le \prod_{i=um}^{(u+1)m-1} \mathrm{F}_{i|n}[y_{i+1:n}](x, A) \le \frac{\sigma^+}{\sigma^-} \prod_{i=um+1}^{(u+1)m} \frac{g^+(y_i)}{g^-(y_i)} \lambda_{u,n}(y_{(u+1)m+1:n}, A) \; . \quad (3.38)$$

*(iv) For any non-negative integers $u$ and $n$,*

$$\delta\left(\prod_{i=um}^{(u+1)m-1} \mathrm{F}_{i|n}[y_{i+1:n}]\right) \le \begin{cases} \rho_0(y_{um+1:(u+1)m}) & u < \lfloor n/m \rfloor \; , \\ \rho_1 & u \ge \lceil n/m \rceil \; , \end{cases}$$

*where for any $y_{um+1:(u+1)m} \in \mathsf{Y}^m$,*

$$\rho_0(y_{um+1:(u+1)m}) \stackrel{\text{def}}{=} 1 - \frac{\sigma^-}{\sigma^+} \prod_{i=um+1}^{(u+1)m} \frac{g^-(y_i)}{g^+(y_i)} \quad \text{and} \quad \rho_1 \stackrel{\text{def}}{=} 1 - \sigma^- \; . \quad (3.39)$$

*Proof.* Part (i) can be proved using an argument similar to the one used for Lemma 60(i).

Next notice that for $0 \le u < \lfloor n/m \rfloor$,

$$\beta_{um|n}[y_{um+1:n}](x_{um}) = \int \cdots \int \prod_{i=um+1}^{(u+1)m} Q(x_{i-1}, dx_i)\, g(x_i, y_i)\, \beta_{(u+1)m|n}[y_{(u+1)m+1:n}](x_{(u+1)m}) \; .$$

Under Assumption 67, dropping the dependence on the $y$s for notational simplicity, the right-hand side of this display is bounded from above by

$$\prod_{i=um+1}^{(u+1)m} g^+(y_i) \int \cdots \int \prod_{i=um+1}^{(u+1)m} Q(x_{i-1}, dx_i)\, \beta_{(u+1)m|n}(x_{(u+1)m}) \le \sigma^+ \prod_{i=um+1}^{(u+1)m} g^+(y_i) \int \beta_{(u+1)m|n}(x_{(u+1)m})\, \kappa(dx_{(u+1)m}) \; .$$

In a similar fashion, a lower bound may be obtained, containing $\sigma^-$ and $g^-$ rather than $\sigma^+$ and $g^+$. Thus part (ii) follows.

For part (iii), we use (2.30) to write

$$\prod_{i=um}^{(u+1)m-1} \mathrm{F}_{i|n}[y_{i+1:n}](x_{um}, A) = \frac{\int \cdots \int \prod_{i=um+1}^{(u+1)m} Q(x_{i-1}, x_i)\, g(x_i, y_i)\, \mathbb{1}_A(x_{(u+1)m}) \beta_{(u+1)m|n}(x_{(u+1)m})}{\int \cdots \int \prod_{i=um+1}^{(u+1)m} Q(x_{i-1}, x_i)\, g(x_i, y_i) \beta_{(u+1)m|n}(x_{(u+1)m})} \; .$$

The right-hand side is bounded from above by

$$\frac{\sigma^+}{\sigma^-} \prod_{i=um+1}^{(u+1)m} \frac{g^+(y_i)}{g^-(y_i)} \times \frac{\int_A \kappa(dx)\, \beta_{(u+1)m|n}[y_{(u+1)m+1:n}](x)}{\int \kappa(dx)\, \beta_{(u+1)m|n}[y_{(u+1)m+1:n}](x)} \; .$$

We define $\lambda_{u,n}$ as the second ratio of this expression. Again a corresponding lower bound is obtained similarly, proving part (iii).

Part (iv) follows from part (iii) and Lemma 51. □

Using this result together with (3.36), we may obtain statements analogous to Proposition 61. In particular, if there exist positive real numbers $\gamma^-$ and $\gamma^+$ such that for all $y \in \mathsf{Y}$,

$$\gamma^- \leq g^-(y) \leq g^+(y) \leq \gamma^+ ,$$

then the smoothing and the filtering distributions both forget uniformly the initial distribution.

Assumptions 62 and 67 are still restrictive and fail to hold in many interesting situations. In both cases, we assume that either the one-step or the $m$-step transition kernel is uniformly bounded from above and below. The following weaker condition is a first step toward handling more general settings.

**Assumption 69.** *Let $Q$ be dominated by a probability measure $\kappa$ on $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$ and $A \in \mathcal{X}$, $Q(x, A) = \int_A q_\kappa(x, x')\, \kappa(dx')$ for some transition density function $q_\kappa$. Assume in addition that*

*(i) There exists a set $C \in \mathcal{X}$, two positive real numbers $\sigma^-$ and $\sigma^+$ such that for all $x \in C$ and $x' \in \mathsf{X}$,*

$$\sigma^- \leq q_\kappa(x, x') \leq \sigma^+ .$$

*(ii) For all $y \in \mathsf{Y}$ and all $x \in \mathsf{X}$, $\int_C q_\kappa(x, x')\, g(x', y)\, \kappa(dx') > 0$;*

*(iii) There exists a (non-identically null) function $\alpha : \mathsf{Y} \to [0, 1]$ such that for any $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $y \in \mathsf{Y}$,*

$$\frac{\int_C \rho[x, x'\,; y](x'')\, \kappa(dx'')}{\int_{\mathsf{X}} \rho[x, x'\,; y](x'')\, \kappa(dx'')} \geq \alpha(y) ,$$

*where for $(x, x', x'') \in \mathsf{X}^3$ and $y \in \mathsf{Y}$,*

$$\rho[x, x'\,; y](x'') \stackrel{\text{def}}{=} q_\kappa(x, x'')g(x'', y)q_\kappa(x'', x') . \tag{3.40}$$

Part (i) of this assumption implies that the set $C$ is 1-small for the kernel $Q$ (see Definition 155). It it shown in Section 7.2.2 that such small sets do exist under conditions that are weak and generally simple to check. Assumption 69 is trivially satisfied under Assumption 62 using the whole state space $\mathsf{X}$ as the state $C$: in that case, their exists a transition density function $q_\kappa(x, x')$ that is bounded from above and below for all $(x, x') \in \mathsf{X}^2$. It is more interesting to consider cases in which the hidden chain is not uniformly ergodic. One such example, first addressed by Budhiraja and Ocone (1997), is a Markov chain observed in noise with bounded support.

**Example 70** (Markov Chain in Additive Bounded Noise)**.** We consider real states $\{X_k\}$ and observations $\{Y_k\}$, assuming that the states form a Markov chain with a transition density $q(x, x')$ with respect to Lebesgue measure. Furthermore we assume the following.

(i) $Y_k = X_k + V_k$, where $\{V_k\}$ is an i.i.d. sequence of satisfying $\mathrm{P}(|V| \geq M) = 0$ for some finite $M$ (the essential supremum of the noise sequence is bounded). In addition, $V_k$ has a probability density $g$ with respect to Lebesgue measure.

(ii) The transition density satisfies $q(x, x') > 0$ for all $(x, x')$ and there exists a positive constant $A$, a probability density $h$ and positive constants $\sigma^-$ and $\sigma^+$ such that for all $x \in C = [-A - M, A + M]$,

$$\sigma^- h(x') \leq q(x, x') \leq \sigma^+ h(x') \; .$$

The results below can readily be extended to cover the case $Y_k = \psi(X_k) + V_k$, provided that the level sets $\{x \in \mathbb{R} : |\psi(x)| \leq K\}$ of the function $\psi$ are compact. This is equivalent to requiring $|\psi(x)| \to \infty$ as $|x| \to \infty$. Likewise extensions to multivariate states and/or observations are obvious.

Under (ii), Assumption 69(i) is satisfied with $C$ as above and $\kappa(dx) = h(x)\, dx$. Denote by $\phi$ the probability density of the random variables $V_k$. Then $g(x, y) = \phi(y - x)$. The density $\phi$ may be chosen such that $\mathrm{supp}\phi \subseteq [-M, +M]$, so that $g(x, y) > 0$ if and only if $x \in [y - M, y + M]$. To verify Assumption 69(iii), put $\Gamma = [-A, A]$. For $y \in \Gamma$, we then have $g(x, y) = 0$ if $x \notin [-A - M, A + M]$, and thus

$$\int q(x, x'')g(x'', y)q(x'', x')\, dx'' = \int_{-A-M}^{A+M} q(x, x'')g(x'', y)q(x'', x')\, dx'' \; .$$

This implies that for all $(x, x') \in \mathsf{X} \times \mathsf{X}$,

$$\frac{\int_C q(x, x'')g(x'', y)q(x'', x')\, dx''}{\int_{\mathsf{X}} q(x, x'')g(x'', y)q(x'', x')\, dx''} = 1 \; .$$

The bounded noise case is of course very specific, because an observation $Y_k$ allows locating the corresponding state $X_k$ within a bounded set.

Under assumption 69, the lemma below establishes that the set $C$ is a 1-small set for the forward transition kernels $\mathrm{F}_{k|n}[y_{k+1:n}]$ and that it is also uniformly accessible from the whole space $\mathsf{X}$ (for the same kernels).

**Lemma 71.** *Under Assumption 69, the following hold true.*

*(i) For any initial probability measure $\nu$ on $(\mathsf{X}, \mathcal{X})$ and any sequence $y_{0:n} \in \mathsf{Y}^{n+1}$ satisfying $\int_C \nu(dx_0)\, g(x_0, y_0) > 0$,*

$$\mathrm{L}_{\nu,n}(y_{0:n}) > 0 \; .$$

*(ii) For any non-negative integers $k$ and $n$ such that $k < n$ and any $y_{0:n} \in \mathsf{Y}^{n+1}$, the set $C$ is a 1-small set for the transitions kernels $\mathrm{F}_{k|n}$. Indeed there exists a transition kernel $\lambda_{k,n}$ from $(\mathsf{Y}^{(n-k)}, \mathcal{Y}^{\otimes(n-k)})$ to $(\mathsf{X}, \mathcal{X})$ such that for all $x \in C$, $y_{k+1:n} \in \mathsf{Y}^{n-k}$ and $A \in \mathcal{X}$,*

$$\mathrm{F}_{k|n}[y_{k+1:n}](x, A) \geq \frac{\sigma^-}{\sigma^+} \lambda_{k,n}[y_{k+1:n}](A) \; .$$

*(iii) For any non-negative integers $k$ and $n$ such that $n \geq 2$ and $k < n - 1$, and any $y_{k+1:n} \in \mathsf{Y}^{n-k}$,*

$$\inf_{x \in \mathsf{X}} \mathrm{F}_{k|n}[y_{k+1:n}](x, C) \geq \alpha(y_{k+1}) \; .$$

*Proof.* Write

$$\begin{aligned}
\mathrm{L}_{\nu,n}(y_{0:n}) &= \int \cdots \int \nu(dx_0)\, g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i)\, g(x_i, y_i) \\
&\geq \int \cdots \int \nu(dx_0)\, g(x_0, y_0) \prod_{i=1}^{n} Q(x_{i-1}, dx_i)\, g(x_i, y_i) \mathbb{1}_C(x_{i-1}) \\
&\geq \int_C \nu(dx_0)\, g(x_0, y_0) \left(\sigma^-\right)^n \prod_{i=1}^{n} \int_C g(x_i, y_i)\, \kappa(dx_i) \; ,
\end{aligned}$$

showing part (i). The proof of (ii) is similar to that of Lemma 60(iii). For (iii), write

$$\begin{aligned}
&\mathrm{F}_{k|n}[y_{k+1:n}](x, C) \\
&\quad = \frac{\iint \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1}) \mathbb{1}_C(x_{k+1}) \varphi[y_{k+2:n}](x_{k+2})\, \kappa(dx_{k+1:k+2})}{\iint \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1}) \varphi[y_{k+2:n}](x_{k+2})\, \kappa(dx_{k+1:k+2})} \\
&\quad = \frac{\iint \Phi[y_{k+1}](x, x_{k+2}) \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1}) \varphi[y_{k+2:n}](x_{k+2})\, \kappa(dx_{k+1:k+2})}{\iint \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1}) \varphi[y_{k+2:n}](x_{k+2})\, \kappa(dx_{k+1:k+2})} \,.
\end{aligned}$$

where $\rho$ is defined in (3.40) and

$$\begin{aligned}
\varphi[y_{k+2:n}](x_{k+2}) &= g(x_{k+2}, y_{k+2}) \beta_{k+2|n}[y_{k+3:n}](x_{k+2}) \,, \\
\Phi[y_{k+1}](x, x_{k+2}) &= \frac{\int \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1}) \mathbb{1}_C(x_{k+1})\, \kappa(dx_{k+1})}{\int \rho[x, x_{k+2}\,; y_{k+1}](x_{k+1})\, \kappa(dx_{k+1})} \,.
\end{aligned}$$

Under Assumption 69, $\Phi(x, x'\,; y) \geq \alpha(y)$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $y \in \mathsf{Y}$, which concludes the proof. □

The corollary below then shows that the whole set $\mathsf{X}$ is a 1-small set for the composition $\mathrm{F}_{k|n}[y_{k+1:n}]\mathrm{F}_{k+1|n}[y_{k+2:n}]$. This generalizes a well-known result for homogeneous Markov chains (see Proposition 157).

**Corollary 72.** *Under Assumption 69, for positive indices $2 \leq k \leq n$,*

$$\left\| \phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}] \right\|_{\mathrm{TV}} \leq 2 \prod_{j=0}^{\lfloor k/2 \rfloor - 1} \left[ 1 - \frac{\sigma^-}{\sigma^+} \alpha(y_{2j+1}) \right] \,.$$

*Proof.* Because of Lemma 71(i), we may use the decomposition in (3.26) with $j = 0$ bounding the total variation distance by 2 to obtain

$$\left\| \phi_{\nu,k|n}[y_{0:n}] - \phi_{\nu',k|n}[y_{0:n}] \right\|_{\mathrm{TV}} \leq 2 \prod_{j=0}^{k-1} \delta\left( \mathrm{F}_{j|n}[y_{j+1:n}] \right) \,.$$

Now, using assertions (ii) and (iii) of Lemma 71,

$$\begin{aligned}
&\mathrm{F}_{j|n}[y_{j+1:n}]\mathrm{F}_{j+1|n}[y_{j+2:n}](x, A) \\
&\qquad \geq \int_C \mathrm{F}_{j|n}[y_{j+1:n}](x, dx') \mathrm{F}_{j+1|n}[y_{j+2:n}](x', A) \\
&\qquad\qquad \geq \alpha(y_{j+1}) \frac{\sigma^-}{\sigma^+} \lambda_{j+1,n}[y_{j+2:n}](A) \,,
\end{aligned}$$

for all $x \in \mathsf{X}$ and $A \in \mathcal{X}$. Hence the composition $\mathrm{F}_{j|n}[y_{j+1:n}]\mathrm{F}_{j+1|n}[y_{j+2:n}]$ satisfies Doeblin's condition (Assumption 50) and the proof follows by Application of Lemma 51. □

Corollary 72 is only useful in cases where the function $\alpha$ is such that the obtained bound indeed decreases as $k$ and $n$ grow. In Example 70, one could set $\alpha(y) = \mathbb{1}_\Gamma(y)$, for an interval $\Gamma$. In such a case, it suffices that the joint chain $\{X_k, Y_k\}_{k \geq 0}$ be recurrent under $\mathrm{P}_{\nu_\star}$—which was the case in Example 70—to guarantee that $\mathbb{1}_\Gamma(Y_k)$ equals one infinitely often and thus that $\left\| \phi_{\nu,k|n}[Y_{0:n}] - \phi_{\nu',k|n}[Y_{0:n}] \right\|_{\mathrm{TV}}$ tends to zero $\mathrm{P}_{\nu_\star}$-almost surely as $k, n \to \infty$. The following example illustrates a slightly more complicated situation in which Assumption 69 still holds.

**Example 73** (Non-Gaussian Autoregressive Process in Gaussian Noise)**.** In this example, we consider a first-order non-Gaussian autoregressive process, observed in Gaussian noise. This is a practically relevant example for which there is apparently no results on forgetting available in the literature. The model is thus

$$\begin{aligned} X_{k+1} &= \phi X_k + U_k \ , \qquad X_0 \sim \nu \ , \\ Y_k &= X_k + V_k \ , \end{aligned}$$

where

(i) $\{U_k\}_{k\geq 0}$ is an i.i.d. sequence of random variables with Laplace (double exponential) distribution with scale parameter $\lambda$;

(ii) $\{V_k\}_{k\geq 0}$ is an i.i.d. sequence of Gaussian random variable with zero mean and variance $\sigma^2$.

We will see below that the fact that the tails of the $X$s are heavier than the tails of the observation noise is important for the derivations that follow. It is assumed that $|\phi| < 1$, which implies that the chain $\{X_k\}$ is positive recurrent, that is, admits a single invariant probability measure $\pi$. It may be shown (see Chapter 7) that although the Markov chain $\{X_k\}$ is geometrically ergodic, that is, $\|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} \to 0$ geometrically fast, it is not uniformly ergodic as $\liminf_{n\to\infty} \sup_{x\in\mathbb{R}} \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} > 0$. We will nevertheless see that the forward smoothing kernel is uniformly geometrically ergodic.

Under the stated assumptions,

$$\begin{aligned} q(x,x') &= \frac{1}{2\lambda} \exp\left(-\lambda|x' - \phi x|\right) \ , \\ g(x,y) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-x)^2}{2\sigma^2}\right] \ . \end{aligned}$$

Here we set, for some $M > 0$ to be specified later, $C = [-M - 1/2, M + 1/2]$, and we let $y \in [-1/2, +1/2]$. Note that

$$\begin{aligned} &\frac{\int_{-M-1/2}^{M+1/2} \exp(-\lambda|u - \phi x| - |y-u|^2/2\sigma^2 - \lambda|x' - \phi u|)\, du}{\int_{-\infty}^{\infty} \exp(-\lambda|u - \phi x| - |y-u|^2/2\sigma^2 - \lambda|x' - \phi u|)\, du} \\ &\qquad\qquad \geq \frac{\int_{-M}^{M} \exp(-\lambda|u - x| - u^2/2\sigma^2 - \phi\lambda|x' - u|)\, du}{\int_{-\infty}^{\infty} \exp(-\lambda|u - x| - u^2/2\sigma^2 - \phi\lambda|x' - u|)\, du} \ , \end{aligned}$$

and to show Assumption 69(iii) it suffices to show that the right-hand side is bounded from below. This in turn is equivalent to showing that $\sup_{(x,x')\in\mathbb{R}\times\mathbb{R}} R(x,x') < 1$, where

$$R(x,x') = \frac{\left(\int_{-\infty}^{-M} + \int_{M}^{\infty}\right) \exp(-\alpha|u - x| - \beta u^2 - \gamma|x' - u|)\, du}{\int_{-\infty}^{\infty} \exp(-\alpha|u - x| - \beta u^2 - \gamma|x' - u|)\, du} \tag{3.41}$$

with $\alpha = \lambda$, $\beta = 1/2\sigma^2$ and $\gamma = \phi\lambda$.

To do this, first note that any $M > 0$ we have $\sup\{R(x,x') : |x| \leq M, |x'| \leq M\} < 1$, and we thus only need to study the behavior of this quantity when $x$ and/or $x'$ become large. We first show that

$$\limsup_{M\to\infty} \sup_{x\geq M,\, |x'|\leq M} R(x,x') < 1 \ . \tag{3.42}$$

For this we note that for $|x'| \le M$ and $x \ge M$, it holds that

$$\left(\int_M^x + \int_x^\infty\right) \exp\left[-\alpha|x-u| - \beta u^2 - \gamma(u-x')\right] du \\ \le \mathrm{e}^{-\alpha x}\mathrm{e}^{\gamma M}\,\frac{\exp[-\beta M^2 + (\alpha-\gamma)M]}{2\beta M - (\alpha-\gamma)} + \mathrm{e}^{\gamma M}\,\frac{\exp(-\beta x^2 - \gamma x)}{2\beta x + (\gamma+\alpha)}\ ,$$

where we used the bound

$$\int_y^\infty \exp(\lambda u - \beta u^2)\, du \le (2\beta y - \lambda)\exp(-\beta y^2 + \lambda y)\ ,$$

which holds as soon as $2\beta y - \lambda \ge 0$. Similarly, we have

$$\int_{-\infty}^{-M} \exp\left[-\alpha(x-u) - \beta u^2 - \gamma(x'-u)\right] du \\ \le \mathrm{e}^{-\alpha x}\mathrm{e}^{\gamma M}\,\frac{\exp[-\beta M^2 - (\gamma+\alpha)M]}{2\beta M + (\gamma+\alpha)}\ ,$$

$$\int_{-M}^{M} \exp\left[-\alpha(x-u) - \beta u^2 - \gamma|u-x'|\right] du \\ \ge \mathrm{e}^{-2\gamma M}\mathrm{e}^{-\alpha x}\int_{-M}^{M}\exp(-\beta u^2 + \alpha u)\, du\ .$$

Thus, (3.41) is bounded by

$$\mathrm{e}^{3\gamma M}\,\frac{\dfrac{2\exp[-\beta M^2 + (\alpha-\gamma)M]}{2\beta M + \gamma - \alpha} + \sup\limits_{x\ge M}\dfrac{\exp[-\beta x^2 + (\alpha-\gamma)x]}{\beta x + (\gamma+\alpha)}}{\int_{-M}^{M}\exp(-\beta u^2 + \alpha u)\, du}$$

proving (3.42).

Next we show that

$$\limsup_{M\to\infty}\ \sup_{x\ge M,\, x'\ge M} R(x,x') < 1\ . \tag{3.43}$$

We consider the case $M \le x \le x'$; the other case can be handled similarly. The denominator in (3.41) is then bounded by

$$\mathrm{e}^{-\alpha x - \gamma x'}\int_{-M}^{M}\exp(-\beta u^2 + (\alpha+\gamma)u)\, du\ .$$

The two terms in the numerator are bounded by, respectively,

$$\int_{-\infty}^{-M} \exp\left[-\alpha(x-u) - \beta u^2 - \gamma(x'-u)\right] du \\ \le \mathrm{e}^{-\alpha x - \gamma x'}\,\frac{\exp[-\beta M^2 - (\alpha+\gamma)M]}{2\beta M + \alpha + \gamma}$$

and

$$\int_M^\infty \exp\left(-\alpha|x-u| - \beta u^2 - \gamma|x'-u|\right) du \\ \le \mathrm{e}^{-\alpha x - \gamma x'}\,\frac{\exp[-\beta M^2 + (\alpha+\gamma)M]}{2\beta M - \alpha - \gamma} \\ + \frac{\exp(-\beta x^2 + \gamma x - \gamma x')}{2\beta x - \gamma + \alpha} + \frac{\exp[-\beta(x')^2 + \alpha x - \alpha x']}{2\beta x' + \alpha + \gamma}\ ,$$

and (3.43) follows by combining the previous bounds.

We finally have to check that

$$\limsup_{M\to\infty} \sup_{x'\le -M,\, x\ge M} R(x,x') < 1 \; .$$

This can be done along the same lines.

# Chapter 4

# Sequential Monte Carlo Methods

The use of Monte Carlo methods for non-linear filtering can be traced back to the pioneering contributions of Handschin and Mayne (1969) and Handschin (1970). These early attempts were based on sequential versions of the *importance sampling* paradigm, a technique that amounts to simulating samples under an instrumental distribution and then approximating the target distributions by weighting these samples using appropriately defined *importance weights*. In the non-linear filtering context, importance sampling algorithms can be implemented sequentially in the sense that, by defining carefully a sequence of instrumental distributions, it is not needed to regenerate the population of samples from scratch upon the arrival of each new observation. This algorithm is called *sequential importance sampling*, often abbreviated SIS. Although the SIS algorithm has been known since the early 1970s, its use in non-linear filtering problems was rather limited at that time. Most likely, the available computational power was then too limited to allow convincing applications of these methods. Another less obvious reason is that the SIS algorithm suffers from a major drawback that was not clearly identified and properly cured until the seminal paper by Gordon *et al.* (1993). As the number of iterations increases, the importance weights tend to degenerate, a phenomenon known as *sample impoverishment* or *weight degeneracy*. Basically, in the long run most of the samples have very small normalized importance weights and thus do not significantly contribute to the approximation of the target distribution. The solution proposed by Gordon *et al.* (1993) is to allow rejuvenation of the set of samples by duplicating the samples with high importance weights and, on the contrary, removing samples with low weights.

The *particle filter* of Gordon *et al.* (1993) was the first successful application of sequential Monte Carlo techniques to the field of non-linear filtering. Since then, sequential Monte Carlo (or SMC) methods have been applied in many different fields including computer vision, signal processing, control, econometrics, finance, robotics, and statistics (Doucet *et al.*, 2001; Ristic *et al.*, 2004). This chapter reviews the basic building blocks that are needed to implement a sequential Monte Carlo algorithm, starting with concepts related to the importance sampling approach. More specific aspects of sequential Monte Carlo techniques will be further discussed in Chapter **??**, while convergence issues will be dealt with in Chapter **??**.

## 4.1 Importance Sampling and Resampling

### 4.1.1 Importance Sampling

Importance sampling is a method that dates back to, at least, Hammersley and Handscomb (1965) and that is commonly used in several fields (for general references on importance sampling, see Glynn and Iglehart, 1989, Geweke, 1989, Evans and Swartz, 1995, or Robert and Casella, 2004.)

Throughout this section, $\mu$ will denote a probability measure of interest on a measurable space $(\mathsf{X}, \mathcal{X})$, which we shall refer to as the *target distribution*. As in Chapter **??**, the aim is to approximate integrals of the form $\mu(f) = \int_{\mathsf{X}} f(x)\,\mu(dx)$ for real-valued measurable functions $f$. The Monte Carlo approach exposed in Section **??** consists in drawing an i.i.d. sample $\xi^1, \ldots, \xi^N$ from the probability measure $\mu$ and then evaluating the sample mean $N^{-1}\sum_{i=1}^N f(\xi^i)$. Of course, this technique is applicable only when it is possible (and reasonably simple) to sample from the target distribution $\mu$.

Importance sampling is based on the idea that in certain situations it is more appropriate to sample from an *instrumental distribution* $\nu$, and then to apply a change-of-measure formula to account for the fact that the instrumental distribution is different from the target distribution. More formally, assume that the target probability measure $\mu$ is absolutely continuous with respect to an *instrumental probability measure* $\nu$ from which sampling is easily feasible. Denote by $d\mu/d\nu$ the Radon-Nikodym derivative of $\mu$ with respect to $\nu$. Then for any $\mu$-integrable function $f$,

$$\mu(f) = \int f(x)\,\mu(dx) = \int f(x)\,\frac{d\mu}{d\nu}(x)\,\nu(dx)\;. \tag{4.1}$$

In particular, if $\xi^1, \xi^2, \ldots$ is an i.i.d. sample from $\nu$, (4.1) suggests the following estimator of $\mu(f)$:

$$\tilde{\mu}^{\mathrm{IS}}_{\nu,N}(f) = N^{-1}\sum_{i=1}^N f(\xi^i)\frac{d\mu}{d\nu}(\xi^i)\;. \tag{4.2}$$

Because this estimator is the sample mean of independent random variables, there is a range of results to assess the quality of $\tilde{\mu}^{\mathrm{IS}}_{\nu,N}(f)$ as an estimator of $\mu(f)$. First of all, the strong law of large number implies that $\tilde{\mu}^{\mathrm{IS}}_{\nu,N}(f)$ converges to $\mu(f)$ almost surely as $N$ tends to infinity. In addition, the central limit theorem for i.i.d. variables (or deviation inequalities) may serve as a guidance for selecting the proposal distribution $\nu$, beyond the obvious requirement that it should dominate the target distribution $\mu$. We postpone this issue and, more generally, considerations that pertain to the behavior of the approximation for large values of $N$ to Chapter **??**.

In many situations, the target probability measure $\mu$ or the instrumental probability measure $\nu$ is known only up to a normalizing factor. As already discussed in Remark **??**, this is particularly true when applying importance sampling ideas to HMMs and, more generally, in Bayesian statistics. The Radon-Nikodym derivative $d\mu/d\nu$ is then known up to a (constant) scaling factor only. It is however still possible to use the importance sampling paradigm in that case, by adopting the self-normalized form of the importance sampling estimator,

$$\widehat{\mu}^{\mathrm{IS}}_{\nu,N}(f) = \frac{\sum_{i=1}^N f(\xi^i)\frac{d\mu}{d\nu}(\xi^i)}{\sum_{i=1}^N \frac{d\mu}{d\nu}(\xi^i)}\;. \tag{4.3}$$

This quantity is obviously free from any scale factor in $d\mu/d\nu$. The self-normalized importance sampling estimator $\widehat{\mu}^{\mathrm{IS}}_{\nu,N}(f)$ is defined as a ratio of the sample means of the functions $f_1 = f \times (d\mu/d\nu)$ and $f_2 = d\mu/d\nu$. The strong law of large numbers thus implies that $N^{-1}\sum_{i=1}^N f_1(\xi^i)$ and $N^{-1}\sum_{i=1}^N f_2(\xi^i)$ converge almost surely, to

$\mu(f_1)$ and $\nu(d\mu/d\nu) = 1$, respectively, showing that $\widehat{\mu}^{\mathrm{IS}}_{\nu,N}(f)$ is a consistent estimator of $\mu(f)$. Again, more precise results on the behavior of this estimator will be given in Chapter **??**. In the following, the term *importance sampling* usually refers to the self-normalized form (4.3) of the importance sampling estimate.

### 4.1.2 Sampling Importance Resampling

Although importance sampling is primarily intended to overcome difficulties with direct sampling from $\mu$ when approximating integrals of the form $\mu(f)$, it can also be used for (approximate) sampling from the distribution $\mu$. The latter can be achieved by the *sampling importance resampling* (or SIR) method due to Rubin (1987, 1988). Sampling importance resampling is a two-stage procedure in which importance sampling as discussed below is followed by an additional random sampling step. In the first stage, an i.i.d. sample $(\tilde{\xi}^1, \ldots, \tilde{\xi}^M)$ is drawn from the instrumental distribution $\nu$, and one computes the normalized version of the importance weights,

$$\omega^i = \frac{\frac{d\mu}{d\nu}(\tilde{\xi}^i)}{\sum_{i=1}^M \frac{d\mu}{d\nu}(\tilde{\xi}^i)} \ , \ i = 1, \ldots, M \ . \tag{4.4}$$

In the second stage, the resampling stage, a sample of size $N$ denoted by $\xi^1, \ldots, \xi^N$ is drawn from the intermediate set of points $\tilde{\xi}^1, \ldots, \tilde{\xi}^M$, taking into account the weights computed in (4.4). The rationale is that points $\tilde{\xi}^i$ for which $\omega^i$ in (4.4) is large are most likely under the target distribution $\mu$ and should thus be selected with higher probability during the resampling than points with low (normalized) importance weights. This principle is illustrated in Figure 4.1.

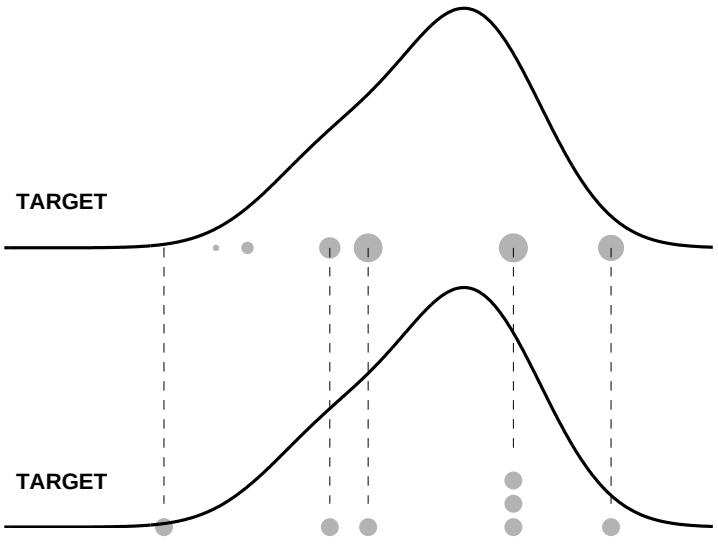



Figure 4.1: Principle of resampling. Top plot: the sample drawn from $\nu$ with associated normalized importance weights depicted by bullets with radii proportional to the normalized weights (the target density corresponding to $\mu$ is plotted in solid line). Bottom plot: after resampling, all points have the same importance weight, and some of them have been duplicated ($M = N = 7$).

There are several ways of implementing this basic idea, the most obvious approach being sampling with replacement with probability of sampling each $\xi^i$ equal

to the importance weight $\omega^i$. Hence the number of times $N^i$ each particular point $\tilde{\xi}^i$ in the first-stage sample is selected follows a binomial $\mathrm{Bin}(N, \omega^i)$ distribution. The vector $(N^1, \ldots, N^M)$ is distributed from $\mathrm{Mult}(N, \omega^1, \ldots, \omega^M)$, the multinomial distribution with parameter $N$ and probabilities of success $(\omega^1, \ldots, \omega^M)$. In this resampling step, the points in the first-stage sample that are associated with small normalized importance weights are most likely to be discarded, whereas the best points in the sample are duplicated in proportion to their importance weights. In most applications, it is typical to choose $M$, the size of the first-stage sample, larger (and sometimes much larger) than $N$. The SIR algorithm is summarized below.

**Algorithm 74** (SIR: Sampling Importance Resampling)**. Sampling:** Draw an i.i.d. sample $\tilde{\xi}^1, \ldots, \tilde{\xi}^M$ from the instrumental distribution $\nu$.

**Weighting:** Compute the (normalized) importance weights

$$\omega^i = \frac{\frac{d\mu}{d\nu}(\tilde{\xi}^i)}{\sum_{j=1}^M \frac{d\mu}{d\nu}(\tilde{\xi}^j)} \quad \text{for } i = 1, \ldots, M \ .$$

**Resampling:**

- Draw, conditionally independently given $(\tilde{\xi}^1, \ldots, \tilde{\xi}^M)$, $N$ discrete random variables $(I^1, \ldots, I^N)$ taking values in the set $\{1, \ldots, M\}$ with probabilities $(\omega^1, \ldots, \omega^M)$, i.e.,

$$\mathrm{P}(I^1 = j) = \omega^j \ , \quad j = 1, \ldots, M \ . \tag{4.5}$$

- Set, for $i = 1, \ldots, N$, $\xi^i = \tilde{\xi}^{I^i}$.

The set $(I^1, \ldots, I^N)$ is thus a multinomial trial process. Hence, this method of selection is known as the *multinomial* resampling scheme.

At this point, it may not be obvious that the sample $\xi^1, \ldots, \xi^N$ obtained from Algorithm 74 is indeed (approximately) i.i.d. from $\mu$ in any suitable sense. In Chapter **??**, it will be shown that the sample mean of the draws obtained using the SIR algorithm,

$$\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f) = \frac{1}{N} \sum_{i=1}^N f(\xi^i) \ , \tag{4.6}$$

is a consistent estimator of $\mu(f)$ for all functions $f$ satisfying $\mu(|f|) < \infty$. The resampling step might thus be seen as a means to transform the weighted importance sampling estimate $\widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f)$ defined by (4.3) into an unweighted sample average. Recall that $N^i$ is the number of times that the element $\tilde{\xi}^i$ is resampled. Rewriting

$$\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f) = \frac{1}{N} \sum_{i=1}^N f(\xi^i) = \sum_{i=1}^M \frac{N^i}{N} f(\tilde{\xi}^i) \ ,$$

it is easily seen that the sample mean $\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f)$ of the SIR sample is, conditionally on the first-stage sample $(\tilde{\xi}^1, \ldots, \tilde{\xi}^M)$, equal to the importance sampling estimator $\widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f)$ defined in (4.3),

$$\mathrm{E}\left[ \hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f) \,\middle|\, \tilde{\xi}^1, \ldots, \tilde{\xi}^M \right] = \widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f) \ .$$

As a consequence, the mean squared error of the SIR estimator $\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f)$ is always larger than that of the importance sampling estimator (4.3) due to the well-known

variance decomposition

$$\mathrm{E}\left[\left(\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f)-\mu(f)\right)^2\right] = \mathrm{E}\left[\left(\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f)-\widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f)\right)^2\right]+\mathrm{E}\left[\left(\widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f)-\mu(f)\right)^2\right].$$

The variance $\mathrm{E}[(\hat{\mu}^{\mathrm{SIR}}_{\nu,M,N}(f)-\widehat{\mu}^{\mathrm{IS}}_{\nu,M}(f))^2]$ may be interpreted as the price to pay for converting the weighted importance sampling estimate into an unweighted approximation.

Showing that the SIR estimate (4.6) is a consistent and asymptotically normal estimator of $\mu(f)$ is not a trivial task, as $\xi^1, \ldots, \xi^N$ are no more independent due to the normalization of the weights followed by resampling. As such, the elementary i.i.d. convergence results that underlie the theory of the importance sampling estimator are of no use, and we refer to Section **??** for the corresponding proofs.

**Remark 75.** A closer examination of the numerical complexity of Algorithm 74 reveals that whereas all steps of the algorithm have a complexity that grows in proportion to $M$ and $N$, this is not quite true for the multinomial sampling step whose numerical complexity is, *a priori*, growing faster than $N$ (about $N \log_2 M$—see Section 4.4.1 below for details). This is very unfortunate, as we know from elementary arguments discussed in Section **??** that Monte Carlo methods are most useful when $N$ is large (or more appropriately that the quality of the approximation improves rather slowly as $N$ grows).

A clever use of elementary probabilistic results however makes it possible to devise methods for sampling $N$ times from a multinomial distribution with $M$ possible outcomes using a number of operations that grows only linearly with the maximum of $N$ and $M$. In order not to interrupt our exposition of sequential Monte Carlo, the corresponding algorithms are discussed in Section 4.4.1 at the end of this chapter. Note that we are here only discussing implementations issues. There are however different motivations, also discussed in Section 4.4.2, for adopting sampling schemes other than multinomial sampling.

## 4.2 Sequential Importance Sampling

### 4.2.1 Sequential Implementation for HMMs

We now specialize the sampling techniques considered above to hidden Markov models. As in previous chapters, we adopt the hidden Markov model as specified by Definition 12 where $Q$ denotes the Markov transition kernel of the hidden chain, $\nu$ is the distribution of the initial state $X_0$, and $g(x,y)$ (for $x \in \mathsf{X}, y \in \mathsf{Y}$) denotes the transition density function of the observation given the state, with respect to the measure $\mu$ on $(\mathsf{Y}, \mathcal{Y})$. To simplify the mathematical expressions, we will also use the shorthand notation $g_k(\cdot) = g(\cdot, Y_k)$ introduced in Section 2.1.4. We denote the joint smoothing distribution by $\phi_{0:k|k}$, omitting the dependence with respect to the initial distribution $\nu$, which does not play an important role here. According to (**??**), the joint smoothing distribution may be updated recursively in time according to the relations

$$\phi_0(f) = \frac{\int f(x_0)\, g_0(x_0)\, \nu(dx_0)}{\int g_0(x_0)\, \nu(dx_0)} \qquad \text{for all } f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}),$$

$$\phi_{0:k+1|k+1}(f_{k+1}) = \int \cdots \int f_{k+1}(x_{0:k+1})\, \phi_{0:k|k}(dx_{0:k}) T^{\mathrm{u}}_k(x_k, dx_{k+1}) \qquad \text{for all } f_{k+1} \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+2}\right), \quad (4.7)$$

where $T_k^{\mathrm{u}}$ is the transition kernel on $(\mathsf{X}, \mathcal{X})$ defined by

$$T_k^{\mathrm{u}}(x, f) = \left(\frac{\mathrm{L}_{k+1}}{\mathrm{L}_k}\right)^{-1} \int f(x')\, Q(x, dx') g_{k+1}(x')$$

$$\text{for all } x \in \mathsf{X}, f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}) . \quad (4.8)$$

The superscript "u" (for "unnormalized") in the notation $T_k^{\mathrm{u}}$ is meant to highlight the fact that $T_k^{\mathrm{u}}$ is not a probability transition kernel. This distinction is important here because the normalized version $T_k = T_k^{\mathrm{u}}/T_k^{\mathrm{u}}(1)$ of the kernel will play an important role in the following. Note that except in some special cases discussed in Chapter **??**, the likelihood ratio $\mathrm{L}_{k+1}/\mathrm{L}_k$ can generally not be computed in closed form, rendering analytic evaluation of $T_k^{\mathrm{u}}$ or $\phi_{0:k|k}$ hopeless. The rest of this section reviews importance sampling methods that make it possible to approximate $\phi_{0:k|k}$ recursively in $k$.

First, because importance sampling can be used when the target distribution is known only up to a scaling factor, the presence of non-computable constants such as $\mathrm{L}_{k+1}/\mathrm{L}_k$ does not preclude the use of the algorithm. Next, it is convenient to choose the instrumental distribution as the probability measure associated with a possibly non-homogeneous Markov chain on $\mathsf{X}$. As seen below, this will make it possible to derive a sequential version of the importance sampling technique. Let $\{R_k\}_{k\geq 0}$ denote a family of Markov transition kernels on $(\mathsf{X}, \mathcal{X})$ and let $\rho_0$ denote a probability measure on $(\mathsf{X}, \mathcal{X})$. Further denote by $\{\rho_{0:k}\}_{k\geq 0}$ the family of probability measures associated with the inhomogeneous Markov chain with initial distribution $\rho_0$ and transition kernels $\{R_k\}_{k\geq 0}$,

$$\rho_{0:k}(f_k) \stackrel{\mathrm{def}}{=} \int \cdots \int f_k(x_{0:k})\, \rho_0(dx_0) \prod_{l=0}^{k-1} R_l(x_l, dx_{l+1}) .$$

In this context, the kernels $R_k$ will be referred to as the *instrumental kernels*. The term *importance kernel* is also used. The following assumptions will be adopted in the sequel.

**Assumption 76** (Sequential Importance Sampling)**.** *1. The target distribution $\phi_0$ is absolutely continuous with respect to the instrumental distribution $\rho_0$.*

*2. For all $k \geq 0$ and all $x \in \mathsf{X}$, the measure $T_k^{\mathrm{u}}(x, \cdot)$ is absolutely continuous with respect to $R_k(x, \cdot)$.*

Then for any $k \geq 0$ and any function $f_k \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+1}\right)$,

$$\phi_{0:k|k}(f_k) = \int \cdots \int f_k(x_{0:k}) \frac{d\phi_0}{d\rho_0}(x_0) \left\{ \prod_{l=0}^{k-1} \frac{dT_l^{\mathrm{u}}(x_l, \cdot)}{dR_l(x_l, \cdot)}(x_{l+1}) \right\} \rho_{0:k}(dx_{0:k}) , \quad (4.9)$$

which implies that the target distribution $\phi_{0:k|k}$ is absolutely continuous with respect to the instrumental distribution $\rho_{0:k}$ with Radon-Nikodym derivative given by

$$\frac{d\phi_{0:k|k}}{d\rho_{0:k}}(x_{0:k}) = \frac{d\phi_0}{d\rho_0}(x_0) \prod_{l=0}^{k-1} \frac{dT_l^{\mathrm{u}}(x_l, \cdot)}{dR_l(x_l, \cdot)}(x_{l+1}) . \quad (4.10)$$

It is thus legitimate to use $\rho_{0:k}$ as an instrumental distribution to compute importance sampling estimates for integrals with respect to $\phi_{0:k|k}$. Denoting by $\xi_{0:k}^1, \ldots, \xi_{0:k}^N$ $N$ i.i.d. random sequences with common distribution $\rho_{0:k}$, the importance sampling estimate of $\phi_{0:k|k}(f_k)$ for $f_k \in \mathcal{F}_{\mathrm{b}}\left(\mathsf{X}^{k+1}\right)$ is defined as

$$\hat{\phi}_{0:k|k}^{\mathrm{IS}}(f_k) = \frac{\sum_{i=1}^N \omega_k^i f_k(\xi_{0:k}^i)}{\sum_{i=1}^N \omega_k^i} , \quad (4.11)$$

where $\omega_k^i$ are the unnormalized importance weights defined recursively by

$$\omega_0^i = \frac{d\phi_0}{d\rho_0}(\xi_0^i) \qquad \text{for } i = 1, \ldots, N \,, \tag{4.12}$$

and, for $k \geq 0$,

$$\omega_{k+1}^i = \omega_k^i \frac{dT_k^{\mathrm{u}}(\xi_k^i, \cdot)}{dR_k(\xi_k^i, \cdot)}(\xi_{k+1}^i) \qquad \text{for } i = 1, \ldots, N \,. \tag{4.13}$$

The multiplicative decomposition of the (unnormalized) importance weights in (4.13) implies that these weights may be computed recursively in time as successive observations become available. In the sequential Monte Carlo literature, the update factor $dT_k^{\mathrm{u}}/dR_k$ is often called the *incremental weight*. As discussed previously in Section 4.1.1, the estimator in (4.11) is left unmodified if the weights, or equivalently the incremental weights, are evaluated up to a constant only. In particular, one may omit the problematic scaling factor $\mathrm{L}_{k+1}/\mathrm{L}_k$ that we met in the definition of $T_k^{\mathrm{u}}$ in (4.8). The practical implementation of sequential importance sampling thus goes as follows.

**Algorithm 77** (SIS: Sequential Importance Sampling)**. Initial State:** Draw an i.i.d. sample $\xi_0^1, \ldots, \xi_0^N$ from $\rho_0$ and set

$$\omega_0^i = g_0(\xi_0^i) \frac{d\nu}{d\rho_0}(\xi_0^i) \qquad \text{for } i = 1, \ldots, N \,.$$

**Recursion:** For $k = 0, 1, \ldots,$

- Draw $(\xi_{k+1}^1, \ldots, \xi_{k+1}^N)$ conditionally independently given $\{\xi_{0:k}^j \,, j = 1, \ldots, N\}$ from the distribution $\xi_{k+1}^i \sim R_k(\xi_k^i, \cdot)$. Append $\xi_{k+1}^i$ to $\xi_{0:k}^i$ to form $\xi_{0:k+1}^i = (\xi_{0:k}^i, \xi_{k+1}^i)$.
- Compute the updated importance weights

$$\omega_{k+1}^i = \omega_k^i \times g_{k+1}(\xi_{k+1}^i) \frac{dQ(\xi_k^i, \cdot)}{dR_k(\xi_k^i, \cdot)}(\xi_{k+1}^i), \qquad i = 1, \ldots, N \,.$$

At any iteration index $k$ importance sampling estimates may be evaluated according to (4.11).

An important feature of Algorithm 77, which corresponds to the method originally proposed in Handschin and Mayne (1969) and Handschin (1970), is that the $N$ trajectories $\xi_{0:k}^1, \ldots, \xi_{0:k}^N$ are independent and identically distributed for all time indices $k$. Following the terminology in use in the non-linear filtering community, we shall refer to the sample at time index $k$, $\xi_k^1, \ldots, \xi_k^N$, as the population (or system) of *particles* and to $\xi_{0:k}^i$ for a specific value of the particle index $i$ as the history (or trajectory) of the $i$th particle. The principle of the method is illustrated in Figure 4.2.

### 4.2.2 Choice of the Instrumental Kernel

Before discussing in Section 4.3 a serious drawback of Algorithm 77 that needs to be fixed in order for the method to be applied to any problem of practical interest, we examine strategies that may be helpful in selecting proper instrumental kernels $R_k$ in several models (or families of models) of interest.

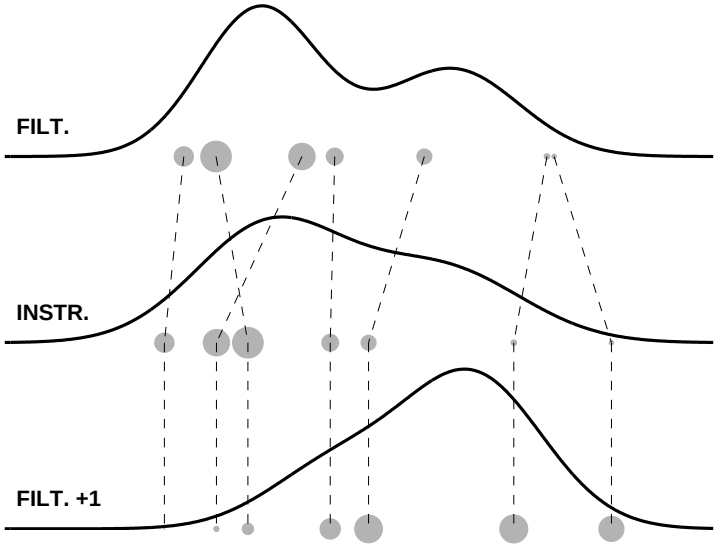



Figure 4.2: Principle of sequential importance sampling (SIS). Upper plot: the curve represents the filtering distribution, and the particles with weights are represented along the axis by bullets, the radii of which being proportional to the normalized weight of the particle. Middle plot: the instrumental distribution with resampled particle positions. Bottom plot: filtering distribution at the next time index with particle updated weights. The case depicted here corresponds to the choice $R_k = Q$.

### Prior Kernel

The first obvious and often very simple choice of instrumental kernel $R_k$ is that of setting $R_k = Q$ (irrespectively of $k$). In that case, the instrumental kernel simply corresponds to the prior distribution of the new state in the absence of the corresponding observation. The incremental weight then simplifies to

$$\frac{dT_k^{\mathrm{u}}(x,\cdot)}{dQ(x,\cdot)}(x') = \frac{\mathrm{L}_k}{\mathrm{L}_{k+1}} g_{k+1}(x') \propto g_{k+1}(x') \qquad \text{for all } (x,x') \in \mathsf{X}^2 \,. \tag{4.14}$$

A distinctive feature of the prior kernel is that the incremental weight in (4.14) *does not depend on* $x$, that is, on the previous position. The use of the prior kernel $R_k = Q$ is popular because sampling from the prior kernel $Q$ is often straightforward, and computing the incremental weight simply amounts to evaluating the conditional likelihood of the new observation given the current particle position. The prior kernel also satisfies the minimal requirement of importance sampling as stated in Assumption 76. In addition, because the importance function reduces to $g_{k+1}$, it is upper-bounded as soon as one can assume that $\sup_{x\in\mathsf{X},y\in\mathsf{Y}} g(x,y)$ is finite, which (often) is a very mild condition (see also Section **??**). Despite these appealing properties, the use of the prior kernel can sometimes lead to poor performance, often manifesting itself as a lack of robustness with respect to the values taken by the observed sequence $\{Y_k\}_{k\geq 0}$. The following example illustrates this problem in a very simple situation.

**Example 78** (Noisy AR(1) Model)**.** To illustrate the potential problems associated with the use of the prior kernel, Pitt and Shephard (1999) consider the simple model where the observations arise from a first-order linear autoregression observed in

noise,

$$X_{k+1} = \phi X_k + \sigma_U U_k \;, \qquad U_k \sim \mathrm{N}(0,1) \;,$$
$$Y_k = X_k + \sigma_V V_k \;, \qquad V_k \sim \mathrm{N}(0,1) \;,$$

where $\phi = 0.9$, $\sigma_U^2 = 0.01$, $\sigma_V^2 = 1$ and $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are independent Gaussian white noise processes. The initial distribution $\nu$ is the stationary distribution of the Markov chain $\{X_k\}_{k\geq 0}$, that is, normal with zero mean and variance $\sigma_U^2/(1-\phi^2)$.

In the following, we assume that $n = 5$ and simulate the first five observations from the model, whereas the sixth observation is set to the arbitrary value 20. The observed series is

$$(-0.652,\, -0.345,\, -0.676,\, 1.142,\, 0.721,\, 20) \;.$$

The last observation is located 20 standard deviations away from the mean (zero) of the stationary distribution, which definitively corresponds to an aberrant value from the model's point of view. In a practical situation however, we would of course like to be able to handle also data that does not necessarily come from the model under consideration. Note also that in this toy example, one can evaluate the exact smoothing distributions by means of the Kalman filtering recursion discussed in Section **??**.

Figure 4.3 displays box and whisker plots for the SIS estimate of the posterior mean of the final state $X_5$ as a function of the number $N$ of particles when using the prior kernel. These plots have been obtained from 125 independent replications of the SIS algorithm. The vertical line corresponds to the true posterior mean of $X_5$ given $Y_{0:5}$, computed using the Kalman filter. The figure shows that the SIS algorithm with the prior kernel grossly underestimates the values of the state even when the number of particles is very large. This is a case where there is a conflict between the prior distribution and the posterior distribution: under the instrumental distribution, all particles are proposed in a region where the conditional likelihood function $g_5$ is extremely low. In that case, the renormalization of the weights used to compute the filtered mean estimate according to (4.11) may even have unexpectedly adverse consequences: a weight close to 1 does not necessarily correspond to a simulated value that is important for the distribution of interest. Rather, it is a weight that is large relative to other, even smaller weights (of particles even less important for the filtering distribution). This is a logical consequence of the fact that the weights must sum to one.

### Optimal Instrumental Kernel

The mismatch between the instrumental distribution and the posterior distribution observed in the previous example is the type of problem that one should try to alleviate by a proper choice of the instrumental kernel. An interesting choice to address this problem is the kernel

$$T_k(x, f) = \frac{\int f(x')\, Q(x, dx') g_{k+1}(x')}{\int Q(x, dx') g_{k+1}(x')} \qquad \text{for } x \in \mathsf{X}, f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X}), \tag{4.15}$$

which is just $T_k^{\mathrm{u}}$ defined in (4.8) properly normalized to correspond to a Markov transition kernel (that is, $T_k(x, 1) = 1$ for all $x \in \mathsf{X}$). The kernel $T_k$ may be interpreted as a regular version of the conditional distribution of the hidden state $X_{k+1}$ given $X_k$ and the current observation $Y_{k+1}$. In the sequel, we will refer to this kernel as the *optimal kernel*, following the terminology found in the sequential importance sampling literature. This terminology dates back probably to Zaritskii *et al.* (1975)

and Akashi and Kumamoto (1977) and is largely adopted by authors such as Liu and Chen (1995), Chen and Liu (2000), Doucet *et al.* (2000), Doucet *et al.* (2001) and Tanizaki (2003). The word "optimal" is somewhat misleading, and we refer to Chapter **??** for a more precise discussion of optimality of the instrumental distribution in the context of importance sampling (which generally has to be defined for a specific choice of the function $f$ of interest). The main property of $T_k$ as defined in (4.15) is that

$$\frac{dT_k^{\mathrm{u}}(x,\cdot)}{dT_k(x,\cdot)}(x') = \frac{\mathrm{L}_k}{\mathrm{L}_{k+1}}\gamma_k(x) \propto \gamma_k(x) \qquad \text{for } (x,x') \in \mathsf{X}^2, \tag{4.16}$$

where $\gamma_k(x)$ is the denominator of $T_k$ in (4.15):

$$\gamma_k(x) \stackrel{\text{def}}{=} \int Q(x,dx')g_{k+1}(x') \; . \tag{4.17}$$

Equation (4.16) means that the incremental weight in (4.13) now depends on the previous position of the particle only (and not on the new position proposed at index $k+1$). This is the exact opposite of the situation observed previously for the prior kernel. The optimal kernel (4.15) is attractive because it incorporates information both on the state dynamics and on the current observation: the particles move "blindly" with the prior kernel, whereas they tend to cluster into regions where the current local likelihood $g_{k+1}$ is large when using the optimal kernel. There are however two problems with using $T_k$ in practice. First, drawing from this kernel is usually not directly feasible. Second, calculation of the incremental importance weight $\gamma_k$ in (4.17) may be analytically intractable. Of course, the optimal kernel takes a simple form with easy simulation and explicit evaluation of (4.17) in the particular cases discussed in Chapter **??**. It turns out that it can also be evaluated for a slightly larger class of non-linear Gaussian state-space models, as soon as the observation equation is linear (Zaritskii *et al.*, 1975). Indeed, consider the state-space model with non-linear state evolution equation

$$X_{k+1} = A(X_k) + R(X_k)U_k \; , \qquad U_k \sim \mathrm{N}(0,I) \; , \tag{4.18}$$

$$Y_k = BX_k + SV_k \; , \qquad V_k \sim \mathrm{N}(0,I) \; , \tag{4.19}$$

where $A$ and $R$ are matrix-valued functions of appropriate dimensions. By application of Proposition **??**, the conditional distribution of the state vector $X_{k+1}$ given $X_k = x$ and $Y_{k+1}$ is multivariate Gaussian with mean $m_{k+1}(x)$ and covariance matrix $\Sigma_{k+1}(x)$, given by

$$\begin{aligned}
K_{k+1}(x) &= R(x)R^t(x)B^t \left[BR(x)R^t(x)B^t + SS^t\right]^{-1} \; , \\
m_{k+1}(x) &= A(x) + K_{k+1}(x)\left[Y_{k+1} - BA(x)\right] \; , \\
\Sigma_{k+1}(x) &= \left[I - K_{k+1}(x)B\right] R(x)R^t(x) \; .
\end{aligned}$$

Hence new particles $\xi_{k+1}^i$ need to be simulated from the distribution

$$\mathrm{N}\left(m_{k+1}(\xi_k^i), \Sigma_{k+1}(\xi_k^i)\right) \; , \tag{4.20}$$

and the incremental weight for the optimal kernel is proportional to

$$\begin{aligned}
\gamma_k(x) = \int q(x,x')g_{k+1}(x')\,dx' \propto \\
|\Gamma_{k+1}(x)|^{-1/2} \exp\left\{-\frac{1}{2}\left[Y_{k+1} - BA(x)\right]^t \Gamma_{k+1}^{-1}(x)\left[Y_{k+1} - BA(x)\right]\right\}
\end{aligned}$$

where

$$\Gamma_{k+1}(x) = BR(x)R^t(x)B^t + SS^t \ .$$

In other situations, sampling from the kernel $T_k$ and/or computing the normalizing constant $\gamma_k$ is a difficult task. There is no general recipe to solve this problem, but rather a set of possible solutions that should be considered.

**Example 79** (Noisy AR(1) Model, Continued)**.** We consider the noisy AR(1) model of Example 78 again using the optimal importance kernel, which corresponds to the particular case where all variables are scalar and $A$ and $R$ are constant in (4.18)–(4.19) above. Thus, the optimal instrumental transition density is given by

$$t_k(x, \cdot) = \mathrm{N}\left(\frac{\sigma_U^2\sigma_V^2}{\sigma_U^2 + \sigma_V^2}\left\{\frac{\phi x}{\sigma_U^2} + \frac{Y_k}{\sigma_V^2}\right\}, \ \frac{\sigma_U^2\sigma_V^2}{\sigma_U^2 + \sigma_V^2}\right)$$

and the incremental importance weights are proportional to

$$\gamma_k(x) \propto \exp\left[-\frac{1}{2}\frac{(Y_k - \phi x)^2}{\sigma_U^2 + \sigma_V^2}\right] \ .$$

Figure 4.4 is the exact analog of Figure 4.3, also obtained from 125 independent runs of the algorithm, for this new choice of instrumental kernel. The figure shows that whereas the SIS estimate of posterior mean is still negatively biased, the optimal kernel tends to reduce the bias compared to the prior kernel. It also shows that as soon as $N = 400$, there are at least some particles located around the true filtered mean of the state, which means that the method should not get entirely lost as subsequent new observations arrive.

To illustrate the advantages of the optimal kernel with respect to the prior kernel graphically, we consider the model (4.18)–(4.19) again with $\phi = 0.9$, $\sigma_u^2 = 0.4$, $\sigma_v^2 = 0.6$, and $(0, 2.6, 0.6)$ as observed series (of length 3). The initial distribution is a mixture $0.6\,\mathrm{N}(-1, 0.3) + 0.4\,\mathrm{N}(1, 0.4)$ of two Gaussians, for which it is still possible to evaluate the exact filtering distributions as the mixture of two Kalman filters using, respectively, $\mathrm{N}(-1, 0.3)$ and $\mathrm{N}(1, 0.4)$ as the initial distribution of $X_0$. We use only seven particles to allow for an interpretable graphical representation. Figures 4.5 and 4.6 show the positions of the particles propagated using the prior kernel and the optimal kernel, respectively. At time 1, there is a conflict between the prior and the posterior as the observation does not agree with the particle approximation of the predictive distribution. With the prior kernel (Figure 4.5), the mass becomes concentrated on a single particle with several particles lost out in the left tail of the distribution with negligible weights. In contrast, in Figure 4.6 most of the particles stay in high probability regions through the iterations with several distinct particles having non-negligible weights. This is precisely because the optimal kernel "pulls" particles toward regions where the current local likelihood $g_k(x) = g_k(x, Y_k)$ is large, whereas the prior kernel does not.

### Accept-Reject Algorithm

Because drawing from the optimal kernel $T_k$ is most often not feasible, a first natural idea consists in trying the accept-reject method (Algorithm **??**), which is a versatile approach to sampling from general distributions. To sample from the optimal importance kernel $T_k(x, \cdot)$ defined by (4.15), one needs an instrumental kernel $R_k(x, \cdot)$ from which it is easy to sample and such that there exists $M$ satisfying $\frac{dQ(x,\cdot)}{dR_k(x,\cdot)}(x')g_k(x') \leq M$ (for all $x \in \mathsf{X}$). Note that because it is generally impossible to evaluate the normalizing constant $\gamma_k$ of $T_k$, we must resort here to the unnormalized version of the accept-reject algorithm. The algorithm consists in

generating pairs $(\xi, U)$ of independent random variables with $\xi \sim R_k(x, \cdot)$ and $U$ uniformly distributed on $[0, 1]$ and accepting $\xi$ if

$$U \leq \frac{1}{M} \frac{dQ(x, \cdot)}{dR_k(x, \cdot)}(\xi) g_k(\xi) \ .$$

Recall that the distribution of the number of simulations required is geometric with parameter

$$p(x) = \frac{\int Q(x, dx') g_k(x')}{M} \ .$$

The strength of the accept-reject technique is that, using any instrumental kernel $R_k$ satisfying the domination condition, one can obtain independent samples from the optimal importance kernel $T_k$. When the conditional likelihood of the observation $g_k(x)$—viewed as a function of $x$—is bounded, one can for example use the prior kernel $Q$ as the instrumental distribution. In that case

$$\frac{dT_k(x, \cdot)}{dQ(x, \cdot)}(x') = \frac{g_k(x')}{\int g_k(u)\, Q(x, du)} \leq \frac{\sup_{x' \in \mathsf{X}} g_k(x')}{\int g_k(u)\, Q(x, du)} \ .$$

The algorithm then consists in drawing $\xi$ from the prior kernel $Q(x, \cdot)$, $U$ uniformly on $[0, 1]$ and accepting the draw if $U \leq g_k(\xi) / \sup_{x \in \mathsf{X}} g_k(x)$. The acceptance rate of this algorithm is then given by

$$p(x) = \frac{\int_{\mathsf{X}} Q(x, dx') g_k(x')}{\sup_{x' \in \mathsf{X}} g_k(x')} \ .$$

Unfortunately, it is not always possible to design an importance kernel $R_k(x, \cdot)$ that is easy to sample from, for which the bound $M$ is indeed finite, *and* such that the acceptance rate $p(x)$ is reasonably large.

### Local Approximation of the Optimal Importance Kernel

A different option consists in trying to approximate the optimal kernel $T_k$ by a simpler proposal kernel $R_k$ that is handy for simulating. Ideally, $R_k$ should be such that $R_k(x, \cdot)$ both has heavier tails than $T_k(x, \cdot)$ and is close to $T_k(x, \cdot)$ around its modes, with the aim of keeping the ratio $\frac{dT_k(x, \cdot)}{dR_k(x, \cdot)}(x')$ as small as possible. To do so, authors such as Pitt and Shephard (1999) and Doucet *et al.* (2000) suggest to first locate the high-density regions of the optimal distribution $T_k(x, \cdot)$ and then use an over-dispersed (that is, with sufficiently heavy tails) approximation of $T_k(x, \cdot)$. The first part of this program mostly applies to the case where the distribution $T_k(x, \cdot)$ is known to be unimodal with a mode that can be located in some way. The overall procedure will need to be repeated $N$ times with $x$ corresponding in turn to each of the current particles. Hence the method used to construct the approximation should be reasonably simple if the potential advantages of using a "good" proposal kernel are not to be offset by an unbearable increase in computational cost.

A first remark of interest is that there is a large class of state-space models for which the distribution $T_k(x, \cdot)$ can effectively be shown to be unimodal using convexity arguments. In the remainder of this section, we assume that $\mathsf{X} = \mathbb{R}^d$ and that the hidden Markov model is fully dominated (in the sense of Definition 13), denoting by $q$ the transition density function associated with the hidden chain. Recall that for a certain form of non-linear state-space models given by (4.18)–(4.19), we were able to derive the optimal kernel and its normalization constant explicitly. Now consider the case where the state evolves according to (4.18), so that

$$q(x, x') \propto \exp\left[ -\frac{1}{2} (x' - A(x))^t \left\{ R(x) R^t(x) \right\}^{-1} (x' - A(x)) \right] \ ,$$

and $g(x, y)$ is simply constrained to be a log-concave function of its $x$ argument. This of course includes the linear Gaussian observation model considered previously in (4.19) but also many other cases like the non-linear observation considered below in Example 80. Then the optimal transition density $t_k^{\mathrm{u}}(x, x') = (\mathrm{L}_{k+1}/\mathrm{L}_k)^{-1} q(x, x') g_k(x')$ is also a log-concave function of its $x'$ argument, as its logarithm is the sum of two concave functions (and a constant term). This implies in particular that $x' \mapsto t_k^{\mathrm{u}}(x, x')$ is unimodal and that its mode may be located using computationally efficient techniques such as Newton iterations.

The instrumental transition density function is usually chosen from a parametric family $\{r_\theta\}_{\theta \in \Theta}$ of densities indexed by a finite-dimensional parameter $\theta$. An obvious choice is the multivariate Gaussian distribution with mean $m$ and covariance matrix $\Gamma$, in which case $\theta = (\mu, \Gamma)$. A better choice is a multivariate $t$-distribution with $\eta$-degrees of freedom, location $m$, and scale matrix $\Gamma$. Recall that the density of this distribution is proportional to $r_\theta(x) \propto [\eta + (x - m)^t \Gamma^{-1} (x - m)]^{(-\eta+d)/2}$. The choice $\eta = 1$ corresponds to a Cauchy distribution. This is a conservative choice that ensures over-dispersion, but if $\mathsf{X}$ is high-dimensional, most draws from a multivariate Cauchy might be too far away from the mode to reasonably approximate the target distribution. In most situations, values such as $\eta = 4$ (three finite moments) are more reasonable, especially if the underlying model does not feature heavy-tailed distributions. Recall also that simulation from the multivariate $t$-distribution with $\eta$ degrees of freedom, location $m$, and scale $\Sigma$ can easily be achieved by first drawing from a multivariate Gaussian distribution with mean $m$ and covariance $\Gamma$ and then dividing the outcome by the square root of an independent chi-square draw with $\eta$ degrees of freedom divided by $\eta$.

To choose the parameter $\theta$ of the instrumental distribution $r_\theta$, one should try to minimize the supremum of the importance function,

$$\min_{\theta \in \Theta} \sup_{x' \in \mathsf{X}} \frac{q(x, x') g_k(x')}{r_\theta(x')} \,. \tag{4.21}$$

This is a minimax guarantee by which $\theta$ is chosen to minimize an upper bound on the importance weights. Note that if $r_\theta$ was to be used for sampling from $t_k(x, \cdot)$ by the accept-reject algorithm, the value of $\theta$ for which the minimum is achieved in (4.21) is also the one that would make the acceptance probability maximal. In practice, solving the optimization problem in (4.21) is often too demanding, and a more generic strategy consists in locating the mode of $x' \mapsto t_k(x, x')$ by an iterative algorithm and evaluating the Hessian of its logarithm at the mode. The parameter $\theta$ is then selected in the following way.

**Multivariate normal:** fit the mean of the normal distribution to the mode of $t_k(x, \cdot)$ and fit the covariance to minus the inverse of the Hessian of $\log t_k(x, \cdot)$ at the mode.

**Multivariate $t$-distribution:** fit the location and scale parameters as the mean and covariance parameters in the normal case; the number of degrees of freedom is usually set arbitrarily (and independently of $x$) based on the arguments discussed above.

We discuss below an important model for which this strategy is successful.

**Example 80** (Stochastic Volatility Model)**.** The state-space equations that define the model is

$$\begin{aligned} X_{k+1} &= \phi X_k + \sigma U_k \,, \\ Y_k &= \beta \exp(X_k/2) V_k \,, \end{aligned}$$

We directly obtain

$$q(x, x') = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x' - \phi x)^2}{2\sigma^2}\right] ,$$
$$g_k(x') = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left[-\frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{1}{2}x'\right] .$$

Simulating from the optimal transition kernel $t_k(x, x')$ is difficult, but the function $x' \mapsto \log(q(x, x')g_k(x'))$ is indeed (strictly) concave. The mode $m_k(x)$ of $x' \mapsto t_k(x, x')$ is the unique solution of the non-linear equation

$$-\frac{1}{\sigma^2}(x' - \phi x) + \frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{1}{2} = 0 , \tag{4.22}$$

which can be found using Newton iterations. Once at the mode, the (squared) scale $\sigma_k^2(x)$ is set as minus the inverse of the second-order derivative of $x' \mapsto (\log q(x, x')g_k(x'))$ evaluated at the mode $m_k(x)$. The result is

$$\sigma_k^2(x) = \left\{\frac{1}{\sigma^2} + \frac{Y_k^2}{2\beta^2}\exp\left[-m_k(x)\right]\right\}^{-1} . \tag{4.23}$$

In this example, a $t$-distribution with $\eta = 5$ degrees of freedom was used, with location $m_k(x)$ and scale $\sigma_k(x)$ obtained as above. The incremental importance weight is then given by

$$\frac{\exp\left[-\frac{(x'-\phi x)^2}{2\sigma^2} - \frac{Y_k^2}{2\beta^2}\exp(-x') - \frac{x'}{2}\right]}{\sigma_k^{-1}(x)\left\{\eta + \frac{[x'-m_k(x)]^2}{\sigma_k^2(x)}\right\}^{-(\eta+1)/2}} .$$

Figure 4.7 shows a typical example of the type of fit that can be obtained for the stochastic volatility model with this strategy using 1,000 particles. Note that although the data used is the same as in Figure **??**, the estimated distributions displayed in both figures are not directly comparable, as the MCMC method in Figure **??** approximates the marginal smoothing distribution, whereas the sequential importance sampling approach used for Figure 4.7 provides a (recursive) approximation to the *filtering* distributions.

When there is no easy way to implement the local linearization technique, a natural idea explored by Doucet *et al.* (2000) and Van der Merwe *et al.* (2000) consists in using classical non-linear filtering procedures to approximate $t_k$. These include in particular the so-called extended Kalman filter (EKF), which dates back to the 1970s (Anderson and Moore, 1979, Chapter 10), as well as the unscented Kalman filter (UKF) introduced by Julier and Uhlmann (1997)—see, for instance, Ristic *et al.* (2004, Chapter 2) for a recent review of these techniques. We illustrate below the use of the extended Kalman filter in the context of sequential importance sampling.

We now consider the most general form of the state-space model with Gaussian noises:

$$X_{k+1} = a(X_k, U_k) , \qquad U_k \sim \mathrm{N}(0, I) , \tag{4.24}$$
$$Y_k = b(X_k, V_k) , \qquad V_k \sim \mathrm{N}(0, I) , \tag{4.25}$$

where $a$, $b$ are vector-valued measurable functions. It is assumed that $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are independent white Gaussian noises. As usual, $X_0$ is assumed to be $\mathrm{N}(0, \Sigma_\nu)$ distributed and independent of $\{U_k\}$ and $\{V_k\}$. The extended Kalman

filter proceeds by approximating the non-linear state-space equations (4.24)–(4.25) by a non-linear Gaussian state-space model with linear measurement equation. We are then back to a model of the form (4.18)–(4.19) for which the optimal kernel may be determined exactly using Gaussian formulas. We will adopt the approximation

$$X_k \approx a(X_{k-1}, 0) + R(X_{k-1})U_{k-1} \ , \tag{4.26}$$
$$Y_k \approx b\left[a(X_{k-1}, 0), 0\right] + B(X_{k-1})\left[X_k - a(X_{k-1}, 0)\right] + S(X_{k-1})V_k \ , \tag{4.27}$$

where

- $R(x)$ is the $d_x \times d_u$ matrix of partial derivatives of $a(x, u)$ with respect to $u$ and evaluated at $(x, 0)$,

$$\left[R(x)\right]_{i,j} \stackrel{\text{def}}{=} \frac{\partial \left[a(x,0)\right]_i}{\partial u_j} \qquad \text{for } i = 1, \ldots, d_x \text{ and } j = 1, \ldots, d_u \ ;$$

- $B(x)$ and $S(x)$ are the $d_y \times d_x$ and $d_y \times d_v$ matrices of partial derivatives of $b(x, v)$ with respect to $x$ and $v$ respectively and evaluated at $(a(x, 0), 0)$,

$$\left[B(x)\right]_{i,j} = \frac{\partial \left\{b\left[a(x,0),0\right]\right\}_i}{\partial x_j} \qquad \text{for } i = 1, \ldots, d_y \text{ and } j = 1, \ldots, d_x \ ,$$
$$\left[S(x)\right]_{i,j} = \frac{\partial \left\{b\left[a(x,0),0\right]\right\}_i}{\partial v_j} \qquad \text{for } i = 1, \ldots, d_y \text{ and } j = 1, \ldots, d_v \ .$$

It should be stressed that the measurement equation in (4.27) differs from (4.19) in that it depends both on the current state $X_k$ and on the previous one $X_{k-1}$. The approximate model specified by (4.26)–(4.27) thus departs from the HMM assumptions. On the other hand, *when conditioning* on the value of $X_{k-1}$, the structure of both models, (4.18)–(4.19) and (4.26)–(4.27), are exactly similar. Hence the posterior distribution of the state $X_k$ given $X_{k-1} = x$ and $Y_k$ is a Gaussian distribution with mean $m_k(x)$ and covariance matrix $\Gamma_k(x)$, which can be evaluated according to

$$K_k(x) = R(x)R^t(x)B^t(x)\left[B(x)R(x)R^t(x)B^t(x) + S(x)S^t(x)\right]^{-1} \ ,$$
$$m_k(x) = a(x,0) + K_k(x)\ \left\{Y_k - b\left[a(x,0),0\right]\right\} \ ,$$
$$\Gamma(x) = \left[I - K_k(x)B(x)\right]R(x)R^t(x) \ .$$

The Gaussian distribution with mean $m_k(x)$ and covariance $\Gamma_k(x)$ may then be used as a proxy for the optimal transition kernel $T_k(x, \cdot)$. To improve the robustness of the method, it is safe to increase the variance, that is, to use $c\Gamma_k(x)$ as the simulation variance, where $c$ is a scalar larger than one. A perhaps more recommendable option consists in using as previously a proposal distribution with tails heavier than the Gaussian, for instance, a multivariate $t$-distribution with location $m_k(x)$, scale $\Gamma_k(x)$, and four or five degrees of freedom.

**Example 81** (Growth Model)**.** We consider the univariate growth model discussed by Kitagawa (1987) and Polson *et al.* (1992) given, in state-space form, by

$$X_k = a_{k-1}(X_{k-1}) + \sigma_u U_{k-1} \ , \qquad U_k \sim \mathrm{N}(0,1) \ , \tag{4.28}$$
$$Y_k = bX_k^2 + \sigma_v V_k \ , \qquad V_k \sim \mathrm{N}(0,1) \ , \tag{4.29}$$

where $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are independent white Gaussian noise processes and

$$a_{k-1}(x) = \alpha_0 x + \alpha_1 \frac{x}{1+x^2} + \alpha_2 \cos\left[1.2(k-1)\right] \tag{4.30}$$

with $\alpha_0 = 0.5$, $\alpha_1 = 25$, $\alpha_2 = 8$, $b = 0.05$, and $\sigma_v^2 = 1$ (the value of $\sigma_u^2$ will be discussed below). The initial state is known deterministically and set to $X_0 = 0.1$. This model is non-linear both in the state and in the measurement equation. Note that the form of the likelihood adds an interesting twist to the problem: whenever $Y_k \leq 0$, the conditional likelihood function

$$g_k(x) \stackrel{\text{def}}{=} g(x; Y_k) \propto \exp\left[-\frac{b^2}{2\sigma_v^2}\left(x^2 - Y_k/b\right)^2\right]$$

is unimodal and symmetric about 0; when $Y_k > 0$ however, the likelihood $g_k$ is symmetric about 0 with two modes located at $\pm(Y_k/b)^{1/2}$.

The EKF approximation to the optimal transition kernel is a Gaussian distribution with mean $m_k(x)$ and variance $\Gamma_k(x)$ given by

$$\begin{aligned} K_k(x) &= 2\sigma_u^2 b a_{k-1}(x)\left[4\sigma_u^2 b^2 a_{k-1}^2(x) + \sigma_v^2\right]^{-1} , \\ m_k(x) &= a_{k-1}(x) + K_k(x)\left[Y_k - b a_{k-1}^2(x)\right] , \\ \Gamma_k(x) &= \frac{\sigma_v^2 \sigma_u^2}{4\sigma_u^2 b^2 a_{k-1}^2(x) + \sigma_v^2} . \end{aligned}$$

In Figure 4.8, the optimal kernel, the EKF approximation to the optimal kernel, and the prior kernel for two different values of the state variance are compared. This figure corresponds to the time index one, and $Y_1$ is set to 6 (recall that the initial state $X_0$ is equal to 0.1). In the case where $\sigma_u^2 = 1$ (left plot in Figure 4.8), the prior distribution of the state, $\mathrm{N}(a_0(X_0), \sigma_u^2)$, turns out to be more informative (more peaky, less diffuse) than the conditional likelihood $g_1$. In other words, the observed $Y_1$ does not carry a lot of information about the state $X_1$, compared to the information provided by $X_0$; this is because the measurement variance $\sigma_v^2$ is not small compared to $\sigma_u^2$. The optimal transition kernel, which does take $Y_1$ into account, is then very close to the prior kernel, and the differences between the three kernels are minor. In such a situation, one should not expect much improvement with the EKF approximation compared to the prior kernel.

In the case shown in the right plot of Figure 4.8 ($\sigma_u^2 = 10$), the situation is reversed. Now $\sigma_v^2$ is relatively small compared to $\sigma_u^2$, so that the information about $X_1$ contained in $g_1$ is large to that provided by the prior information on $X_0$. This is the kind of situation where we expect the optimal kernel to improve considerably on the prior kernel. Indeed, because $Y_1 > 0$, the optimal kernel is bimodal, with the second mode far smaller than the first one (recall that the plots are on log-scale); the EKF kernel correctly picks the dominant mode. Figure 4.8 also illustrates the fact that, in contrast to the prior kernel, the EKF kernel does not necessarily dominate the optimal kernel in the tails; hence the need to simulate from an over-dispersed version of the EKF approximation as discussed above.

## 4.3 Sequential Importance Sampling with Resampling

Despite quite successful results for short data records, as was observed in Example 80, it turns out that the sequential importance sampling approach discussed so far is bound to fail in the long run. We first substantiate this claim with a simple illustrative example before examining solutions to this shortcoming based on the concept of resampling introduced in Section 4.1.2.

### 4.3.1 Weight Degeneracy

The intuitive interpretation of the importance sampling weight $\omega_k^i$ is as a measure of the adequacy of the simulated trajectory $\xi_{0:k}^i$ to the target distribution $\phi_{0:k|n}$. A small importance weight implies that the trajectory is drawn far from the main body of the posterior distribution $\phi_{0:k|n}$ and will contribute only moderately to the importance sampling estimates of the form (4.11). Indeed, a particle such that the associated weight $\omega_k^i$ is orders of magnitude smaller than the sum $\sum_{i=1}^N \omega_k^i$ is practically ineffective. If there are too many ineffective particles, the particle approximation becomes both computationally and statistically inefficient: most of the computing effort is put on updating particles and weights that do not contribute significantly to the estimator; the variance of the resulting estimator will not reflect the large number of terms in the sum but only the small number of particles with non-negligible normalized weights.

Unfortunately, the situation described above is the rule rather than the exception, as the importance weights will (almost always) degenerate as the time index $k$ increases, with most of the normalized importance weights $\omega_k^i / \sum_{j=1}^N \omega_k^j$ close to 0 except for a few ones. We consider below the case of i.i.d. models for which it is possible to show using simple arguments that the large sample variance of the importance sampling estimate can only increase with the time index $k$.

**Example 82** (Weight Degeneracy in the I.I.D. Case)**.** The simplest case of application of the sequential importance sampling technique is when $\mu$ is a probability distribution on $(\mathsf{X}, \mathcal{X})$ and the sequence of target distributions corresponds to the product distributions, that is, the sequence of distributions on $(\mathsf{X}^{k+1}, \mathcal{X}^{\otimes(k+1)})$ defined recursively by $\mu_0 = \mu$ and $\mu_k = \mu_{k-1} \otimes \mu$, for $k \geq 1$. Let $\nu$ be another probability distribution on $(\mathsf{X}, \mathcal{X})$ and assume that $\mu$ is absolutely continuous with respect to $\nu$ and that

$$\int \left[\frac{d\mu}{d\nu}(x)\right]^2 \nu(dx) < \infty \,. \tag{4.31}$$

Finally, let $f$ be a bounded measurable function that is not ($\mu$-a.s.) constant such that its variance under $\mu$, $\mu(f^2) - \mu^2(f)$, is strictly positive.

Consider the sequential importance sampling estimate given by

$$\widehat{\mu}_{k,N}^{\mathrm{IS}}(f) = \sum_{i=1}^N f(\xi_k^i) \frac{\prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_l^i)}{\sum_{j=1}^N \prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_l^j)} \,, \tag{4.32}$$

where the random variables $\{\xi_l^j\}$, $l = 1, \ldots, k$, $j = 1, \ldots, N$ are i.i.d. with common distribution $\nu$. As discussed in Section 4.2, the unnormalized importance weights may be computed recursively and hence (4.32) really corresponds to an estimator of the form (4.11) in the particular case of a function $f_k$ that depends on the last component only. This is of course a rather convoluted and very inefficient way of constructing an estimate of $\mu(f)$ but still constitutes a valid instance of the sequential importance sampling approach (in a very particular case).

Now let $k$ be fixed and write

$$N^{1/2}\left\{\widehat{\mu}_{k,N}^{\mathrm{IS}}(f) - \mu(f)\right\} = \frac{N^{-1/2}\sum_{i=1}^N \prod_{l=0}^k \left\{f(\xi_k^i) - \mu(f)\right\} \frac{d\mu}{d\nu}(\xi_l^i)}{N^{-1}\sum_{i=1}^N \prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_l^i)} \,. \tag{4.33}$$

Because

$$\mathrm{E}\left[\prod_{l=0}^k \frac{d\mu}{d\nu}(\xi_l^i)\right] = 1 \,,$$

the weak law of large numbers implies that the denominator of the right-hand side of (4.33) converges to 1 in probability as $N$ increases. Likewise, under (4.31), the

central limit theorem shows that the numerator of the right-hand side of (4.33) converges in distribution to the normal $\mathrm{N}(0, \sigma_k^2(f))$ distribution, where

$$\sigma_k^2(f) = \mathrm{E}\left(\left\{\prod_{l=0}^{k}\left[f(\xi_k^1) - \mu(f)\right]^2 \frac{d\mu}{d\nu}(\xi_l^1)\right\}^2\right) \tag{4.34}$$

$$= \left[\int \left(\frac{d\mu}{d\nu}(x)\right)^2 \nu(dx)\right]^k \int \left[\frac{d\mu}{d\nu}(x)\right]^2 \left[f(x) - \mu(f)\right]^2 \nu(dx) \; .$$

Slutsky's lemma then implies that (4.33) also converges in distribution to the same $\mathrm{N}(0, \sigma_k^2(f))$ limit as $N$ grows. Now Jensen's inequality implies that

$$1 = \left[\int \frac{d\mu}{d\nu}(x)\nu(dx)\right]^2 \leq \int \left[\frac{d\mu}{d\nu}(x)\right]^2 \nu(dx) \; ,$$

with equality if and only if $\mu = \nu$. Therefore, if $\mu \neq \nu$, the asymptotic variance $\sigma_k^2(f)$ grows exponentially with the iteration index $k$ for all functions $f$ such that

$$\int \left[\frac{d\mu}{d\nu}(x)\right]^2 \left[f(x) - \mu(f)\right]^2 \nu(dx) = \int \frac{d\mu}{d\nu}(x) \left[f(x) - \mu(f)\right]^2 \mu(dx) \neq 0 \; .$$

Because $\mu$ is absolutely continuous with respect to $\nu$, $\mu\{x \in \mathsf{X} : d\mu/d\nu(x) = 0\} = 0$ and the last integral is null if and only if $f$ has zero variance under $\mu$.

Thus in the i.i.d. case, the asymptotic variance of the importance sampling estimate (4.32) increases exponentially with the time index $k$ as soon as the proposal and target differ (except for constant functions).

It is more difficult to characterize the degeneracy of the weights for general target and instrumental distributions. There have been some limited attempts to study more formally this phenomenon in some specific scenarios. In particular, Del Moral and Jacod (2001) have shown the degeneracy of the sequential importance sampling estimator of the posterior mean in Gaussian linear models when the instrumental kernel is the prior kernel. Such results are in general difficult to derive (even in the Gaussian linear models where most of the derivations can be carried out explicitly) and do not provide much additional insight. Needless to say, in practice, weight degeneracy is a prevalent and serious problem making the vanilla sequential importance sampling method discussed so far almost useless. The degeneracy can occur after a very limited number of iterations, as illustrated by the following example.

**Example 83** (Stochastic Volatility Model, Continued)**.** Figure 4.9 displays the histogram of the base 10 logarithm of the normalized importance weights after 1, 10, and 100 time indices for the stochastic volatility model considered in Example 80 (using the same instrumental kernel). The number of particles is set to 1,000. Figure 4.9 shows that, despite the choice of a reasonably good approximation to the optimal importance kernel, the normalized importance weights quickly degenerate as the number of iterations of the SIS algorithm increases. Clearly, the results displayed in Figure 4.7 still are reasonable for $k = 20$ but would be disastrous for larger time horizons such as $k = 100$.

Because the weight degeneracy phenomenon is so detrimental, it is of great practical significance to set up tests that can detect this phenomenon. A simple criterion is the coefficient of variation of the normalized weights used by Kong *et al.* (1994), which is defined by

$$\mathrm{CV}_N = \left[\frac{1}{N}\sum_{i=1}^{N}\left(N\frac{\omega^i}{\sum_{j=1}^{N}\omega^j} - 1\right)^2\right]^{1/2} . \tag{4.35}$$

The coefficient of variation is minimal when the normalized weights are all equal to $1/N$, and then $\mathrm{CV}_N = 0$. The maximal value of $\mathrm{CV}_N$ is $\sqrt{N-1}$, which corresponds to one of the normalized weights being one and all others being null. Therefore, the coefficient of variation is often interpreted as a measure of the number of ineffective particles (those that do not significantly contribute to the estimate). A related criterion with a simpler interpretation is the so-called *effective sample size* $N_{\mathrm{eff}}$ (Liu, 1996), defined as

$$N_{\mathrm{eff}} = \left[ \sum_{i=1}^{N} \left( \frac{\omega^i}{\sum_{j=1}^{N} \omega^j} \right)^2 \right]^{-1} , \tag{4.36}$$

which varies between 1 (all weights null but one) and $N$ (equal weights). It is straightforward to verify the relation

$$N_{\mathrm{eff}} = \frac{N}{1 + \mathrm{CV}_N^2} .$$

Some additional insights and heuristics about the coefficient of variation are given by Liu and Chen (1995).

Yet another possible measure of the weight imbalance is the Shannon entropy of the importance weights,

$$\mathrm{Ent} = -\sum_{i=1}^{N} \frac{\omega^i}{\sum_{j=1}^{N} \omega^j} \log_2 \left( \frac{\omega^i}{\sum_{j=1}^{N} \omega^j} \right) . \tag{4.37}$$

When all the normalized importance weights are null except for one of them, the entropy is null. On the contrary, if all the weights are equal to $1/N$, then the entropy is maximal and equal to $\log_2 N$.

**Example 84** (Stochastic Volatility Model, Continued)**.** Figure 4.10 displays the coefficient of variation (left) and Shannon entropy (right) as a function of the time index $k$ under the same conditions as for Figure 4.9, that is, for the stochastic volatility model of 80. The figure shows that the distribution of the weights steadily degenerates: the coefficient of variation increases and the entropy of the importance weights decreases. After 100 iterations, there are less than 50 particles (out 1,000) significantly contributing to the importance sampling estimator. Most particles have importance weights that are zero to machine precision, which is of course a tremendous waste in computational resource.

### 4.3.2 Resampling

The solution proposed by Gordon *et al.* (1993) to reduce the degeneracy of the importance weights is based on the concept of *resampling* already discussed in the context of importance sampling in Section 4.1.2. The basic method consists in resampling in the current population of particles using the normalized weights as probabilities of selection. Thus, trajectories with small importance weights are eliminated, whereas those with large importance weights are duplicated. After resampling, all importance weights are reset to one. Up to the first instant when resampling occurs, the method can really be interpreted as an instance of the sampling importance resampling (SIR) technique discussed in Section 4.1.2. In the context of sequential Monte Carlo, however, the main motivation for resampling is to avoid future weight degeneracy by reseting (periodically) the weights to equal values. The resampling step has a drawback however: as emphasized in Section 4.1.2, resampling

introduces additional variance in Monte Carlo approximations. In some situations, the additional variance may be far from negligible: when the importance weights already are nearly equal for instance, resampling can only reduce the number of distinct particles, thus degrading the accuracy of the Monte Carlo approximation. The one-step effect of resampling is thus negative but, in the long term, resampling is required to guarantee a stable behavior of the algorithm. This interpretation suggests that it may be advantageous to restrict the use of resampling to cases where the importance weights are becoming very uneven. The criteria defined in (4.35), (4.36), or (4.37) are of course helpful for that purpose. The resulting algorithm, which is generally known under the name of *sequential importance sampling with resampling* (SISR), is summarized below.

**Algorithm 85** (SISR: Sequential Importance Sampling with Resampling)**.** Initialize the particles as in Algorithm 77, optionally applying the resampling step below. For subsequent time indices $k \geq 0$, do the following.

**Sampling:**

- Draw $(\tilde{\xi}_{k+1}^1, \ldots, \tilde{\xi}_{k+1}^N)$ conditionally independently given $\{\xi_{0:k}^j, j = 1, \ldots, N\}$ from the instrumental kernel: $\tilde{\xi}_{k+1}^i \sim R_k(\xi_k^i, \cdot)$, $i = 1, \ldots, N$.
- Compute the updated importance weights

$$\omega_{k+1}^i = \omega_k^i g_{k+1}(\tilde{\xi}_{k+1}^i) \frac{dQ(\xi_k^i, \cdot)}{dR_k(\xi_k^i, \cdot)}(\tilde{\xi}_{k+1}^i) , \qquad i = 1, \ldots, N .$$

**Resampling (Optional):**

- Draw, conditionally independently given $\{(\xi_{0:k}^i, \tilde{\xi}_{k+1}^j),\ i, j = 1, \ldots, N\}$, the multinomial trial $(I_{k+1}^1, \ldots I_{k+1}^N)$ with probabilities of success

$$\frac{\omega_{k+1}^1}{\sum_j^N \omega_{k+1}^j}, \ldots, \frac{\omega_{k+1}^N}{\sum_j^N \omega_{k+1}^j} .$$

- Reset the importance weights $\omega_{k+1}^i$ to a constant value for $i = 1, \ldots, N$.

If resampling is not applied, set for $i = 1, \ldots, N$, $I_{k+1}^i = i$.

**Trajectory update:** for $i = 1, \ldots, N$,

$$\xi_{0:k+1}^i = \left( \xi_{0:k}^{I_{k+1}^i} , \tilde{\xi}_{k+1}^{I_{k+1}^i} \right) . \tag{4.38}$$

As discussed previously the resampling step in the algorithm above may be used systematically (for all indices $k$), but it is often preferable to perform resampling from time to time only. Usually, resampling is either used systematically but at a lower rate (for one index out of $m$, where $m$ is fixed) or at random instants based on the values of the coefficient of variation or the entropy criteria defined in (4.35) and (4.37), respectively. Note that in addition to arguments based on the variance of the Monte Carlo approximation, there is usually also a computational incentive for limiting the use of resampling; indeed, except in models where the evaluation of the incremental weights is costly (think of large-dimensional multivariate observations for instance), the computational cost of the resampling step is not negligible. Both Sections 4.4.1 and 4.4.2 discuss several implementations and variants of the resampling step that may render the latter argument less pregnant.

The term *particle filter* is often used to refer to Algorithm 85 although the terminology SISR is preferable, as particle filtering is sometimes also used more generically for any sequential Monte Carlo method. Gordon *et al.* (1993) actually proposed a specific instance of Algorithm 85 in which resampling is done systematically at each step and the instrumental kernel is chosen as the prior kernel $R_k = Q$. This particular algorithm, commonly known as the *bootstrap filter*, is most often very easy to implement because it only involves simulating from the transition kernel $Q$ of the hidden chain and evaluation of the conditional likelihood function $g$.

There is of course a whole range of variants and refinements of Algorithm 85, many of which will be covered in some detail in the next chapter. A simple remark though is that, as in the case of the simplest SIR method discussed in Section 4.1.2, it is possible to resample $N$ times from a larger population of $M$ intermediate samples. In practice, it means that Algorithm 85 should be modified as follows at indices $k$ for which resampling is to be applied.

**SIS:** For $i = 1, \ldots, N$, draw $\alpha$ candidates $\tilde{\xi}_{k+1}^{i,1}, \ldots, \tilde{\xi}_{k+1}^{i,\alpha}$ from each proposal distribution $R_k(\xi_k^i, \cdot)$.

**Resampling:** Draw $(N_{k+1}^{1,1}, \ldots, N_{k+1}^{1,\alpha}, \ldots, N_{k+1}^{N,1}, \ldots, N_{k+1}^{N,\alpha})$ from the multinomial distribution with parameter $N$ and probabilities

$$\frac{\omega_{k+1}^{i,j}}{\sum_{l=1}^{N}\sum_{m=1}^{\alpha}\omega_{k+1}^{l,m}} \qquad \text{for } i = 1, \ldots, N,\ j = 1, \ldots, \alpha \ .$$

**Remark 86** (Marginal Interpretation of SIS and SISR)**.** Both Algorithms 77 and 85 have been introduced as methods to simulate whole *trajectories* $\{\xi_{0:k}^i\}_{1\leq i\leq N}$ that approximate the joint smoothing distribution $\phi_{0:k|k}$. This was done quite easily in the case of sequential importance sampling (Algorithm 77), as the trajectories are simply extended independently of one another as new samples arrive. When using resampling however, the process is more involved because it becomes necessary to duplicate or discard some trajectories according to (4.38).

This presentation of the SIS and SISR methods has been adopted because it is the most natural way to introduce sequential Monte Carlo methods. It does not mean that, when implementing the SISR algorithm, storing the whole trajectories is required. Neither do we claim that for large $k$, the approximation of the complete joint distribution $\phi_{0:k|k}$ provided by the particle trajectories $\{\xi_{0:k}^i\}_{1\leq i\leq N}$ is accurate. Most often, Algorithm 85 is implemented storing only the current generation of particles $\{\xi_k^i\}_{1\leq i\leq N}$, and (4.38) simplifies to

$$\xi_{k+1}^i = \tilde{\xi}_{k+1}^{I_{k+1}^i} \qquad i = 1, \ldots, N \ .$$

In that case, the system of particles $\{\xi_k^i\}_{1\leq i\leq N}$ with associated weights $\{\omega_k^i\}_{1\leq i\leq N}$, provides an approximation to the filtering distribution $\phi_k$, which is the *marginal* of the joint smoothing distribution $\phi_{0:k|k}$.

The notation $\xi_k^i$ could be ambiguous when resampling is applied, as the first $k+1$ elements of the $i$th trajectory $\xi_{0:k+1}^i$ at time $k+1$ do not necessarily coincide with the $i$th trajectory $\xi_{0:k}^i$ at time $k$. By convention, $\xi_k^i$ *always refers to the last point in the ith trajectory, as simulated at index k*. Likewise, $\xi_{l:k}^i$ is the portion of the same trajectory that starts at index $l$ and ends at the last index (that is, $k$). When needed, we will use the notation $\xi_{0:k}^i(l)$ for the element of index $l$ in the $i$th particle trajectory at time $k$ to avoid ambiguity.

To conclude this section on the SISR algorithm, we briefly revisit two of the examples already considered previously to contrast the results obtained with the SIS and SISR approaches.

**Example 87** (Stochastic Volatility Model, Continued)**.** To illustrate the effectiveness of the resampling strategy, we consider once again the stochastic volatility model introduced in Example 80, for which the weight degeneracy phenomenon (in the basic SIS approach) was patent in Figures 4.9 and 4.10.

Figures 4.11 and 4.12 are the counterparts of Figs. 4.10 and 4.9, respectively, when resampling is applied whenever the coefficient of variation (4.35) of the normalized weights exceeds one. Note that Figure 4.11 displays the coefficient of variation and Shannon entropy computed, for each index $k$, *before resampling*, at indices for which resampling do occur. Contrary to what happened in plain importance sampling, the histograms of the normalized importance weights shown in Figure 4.12 are remarkably similar, showing that the weight degeneracy phenomenon is now under control. Another important remark in this example is that both criteria (the coefficient of variation and entropy) are strongly correlated. Triggering resampling whenever the entropy gets below, say 9.2, would thus be nearly equivalent with resampling occurring, on average, once every tenth time indices. The Shannon entropy of the normalized importance weights evolves between 10 and 9, suggesting that there are at least 500 particles that are significantly contributing to the importance sampling estimate (out of 1,000).

**Example 88** (Growth Model, Continued)**.** Consider again the non-linear state-space model of Example 81, with the variance $\sigma_u^2$ of the state noise set to 10; this makes the observations very informative relative to the prior distribution on the hidden states. Figures 4.13 and 4.14 display the filtering distributions estimated for the first 31 time indices when using the SIS method with the prior kernel $Q$ as instrumental kernel (Figure 4.13), and the corresponding SISR algorithm with systematic resampling—that is, the bootstrap filter—in Figure 4.14. Both algorithms use 500 particles.

For each time index, the top plots of Figures 4.13 and 4.14 show the highest posterior density (HPD) regions corresponding to the estimated filtering distribution, where the lighter grey zone contains 95% of the probability mass and the darker area corresponds to 50% of the probability mass. These HPD regions are based on a kernel density estimate (using the Epanechnikov kernel with bandwidth 0.2) computed from the weighted particles (that is, before resampling in the case of the bootstrap filter). Up to $k = 8$, the two methods yield very similar results. With the SIS algorithm however, the bottom panel of Figure 4.13 shows that the weights degenerate quickly. Remember that the maximal value of the coefficient of variation (4.35) is $\sqrt{N-1}$, that is about 22.3 in the case of Figure 4.13. Hence for $k = 6$ and for all indices after $k = 12$, the bottom panel of Figure 4.13 indeed means that almost all normalized weights but one are null: the filtered estimate is concentrated at one point, which sometimes severely departs from the actual state trajectory shown by the crosses. In contrast, the bootstrap filter (Figure 4.14) appears to be very stable and provides reasonable state estimates even at indices for which the filtering distribution is strongly bimodal (see Example 81 for an explanation of this latter feature).

## 4.4 Complements

As discussed above, resampling is a key ingredient of the success of sequential Monte Carlo techniques. We discuss below two separate aspects related to this issue. First, we show that there are several schemes based on clever probabilistic results that may be exploited to reduce the computational load associated with multinomial resampling. Next, we examine some variants of resampling that achieves lower

conditional variance than multinomial resampling. In this latter case, the aim is of course to be able to decrease the number of particles without losing too much on the quality of the approximation.

Throughout this section, we will assume that it is required to draw $N$ samples $\xi^1, \ldots, \xi^N$ out of a, usually larger, set $\{\tilde{\xi}^1, \ldots, \tilde{\xi}^M\}$ according to the *normalized* importance weights $\{\omega^1, \ldots, \omega^M\}$. We denote by $\mathcal{G}$ a $\sigma$-field such that both $\omega^1, \ldots, \omega^M$ and $\tilde{\xi}^1, \ldots, \tilde{\xi}^M$ are $\mathcal{G}$-measurable.

### 4.4.1 Implementation of Multinomial Resampling

Drawing from the multinomial distribution is equivalent to drawing $N$ random indices $I^1, \ldots, I^N$ conditionally independently given $\mathcal{G}$ from the set $\{1, \ldots, M\}$ and such that $\mathrm{P}(I^j = i \,|\, \mathcal{G}) = \omega^i$. This is of course the simplest example of use of the *inversion method*, and each index may be obtained by first simulating a random variable $U$ with uniform distribution on $[0, 1]$ and then determining the index $I$ such that $U \in (\sum_{j=1}^{I-1} \omega^j, \sum_{j=1}^{I} \omega^j]$ (see Figure 4.15). Determining the appropriate index $I$ thus requires on the average $\log_2 M$ comparisons (using a simple binary tree search). Therefore, the naive technique to implement multinomial resampling requires the simulation of $N$ independent uniform random variables and, on the average, of the order $N \log_2 M$ comparisons.

A nice solution to avoid the repeated sorting operations consists in pre-sorting the uniform variables. Because the resampling is to be repeated $N$ times, we need $N$ uniform random variables, which will be denoted by $U_1, \ldots, U_N$ and $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(N)}$ denoting the associated order statistics. It is easily checked that applying the inversion method from the *ordered* uniforms $\{U_{(i)}\}$ requires, in the worst case, only $M$ comparisons. The problem is that determining the order statistics from the unordered uniforms $\{U_i\}$ by sorting algorithms such as Heapsort or Quicksort is an operation that requires, at best, of the order $N \log_2 N$ comparisons (Press *et al.*, 1992, Chapter 8). Hence, except in cases where $N \ll M$, we have not gained anything yet by pre-sorting the uniform variables prior to using the inversion method. It turns out however that two distinct algorithms are available to sample directly the ordered uniforms $\{U_{(i)}\}$ with a number of operations that scales linearly with $N$.

Both of these methods are fully covered in by Devroye (1986, Chapter 5), and we only cite here the appropriate results, referring to Devroye (1986, pp. 207–215) for proofs and further references on the methods.

**Proposition 89** (Uniform Spacings)**.** *Let $U_{(1)} \leq \ldots \leq U_{(N)}$ be the order statistics associated with an i.i.d. sample from the* $\mathrm{U}([0, 1])$ *distribution. Then the increments*

$$S_i = U_{(i)} - U_{(i-1)} \,, \quad i = 1, \ldots, N \,, \tag{4.39}$$

*(where by convention $S_1 = U_{(1)}$) are called the uniform spacings and distributed as*

$$\frac{E_1}{\sum_{i=1}^{N+1} E_i}, \ldots, \frac{E_N}{\sum_{i=1}^{N+1} E_i} \,,$$

*where $E_1, \ldots, E_{N+1}$ is a sequence of i.i.d. exponential random variables.*

**Proposition 90** (Malmquist, 1950)**.** *Let $U_{(1)} \leq \ldots \leq U_{(N)}$ be the order statistics of $U_1, U_2, \ldots, U_N$—a sequence of i.i.d. uniform $[0, 1]$ random variables. Then $U_N^{1/N}$, $U_N^{1/N} U_{N-1}^{1/(N-1)}$, …, $U_N^{1/N} U_{N-1}^{1/(N-1)} \cdots U_1^{1/1}$ is distributed as $U_{(N)}$, …, $U_{(1)}$.*

The two sampling algorithms associated with these probabilistic results may be summarized as follows.

**Algorithm 91** (After Proposition 89)**.**

**For** $i = 1, \ldots, N+1$**:** Simulate $U_i \sim \mathrm{U}([0,1])$ and set $E_i = -\log U_i$.

**Set** $G = \sum_{i=1}^{N+1} E_i$ **and** $U_{(1)} = E_1/G$**.**

**For** $i = 2, \ldots, n$**:** $U_{(i)} = U_{(i-1)} + E_i/G$.

**Algorithm 92** (After Proposition 90)**.**

**Generate** $V_N \sim \mathrm{U}([0,1])$ and set $U_{(N)} = V_N^{1/N}$.

**For** $i = N-1$ **down to** 1**:** Generate $V_i \sim \mathrm{U}([0,1])$ and set $U_{(i)} = V_i^{1/i} U_{(i+1)}$.

Note that Devroye (1986) also discusses a third, slightly more complicated algorithm—the bucket sort method of Devroye and Klincsek (1981)—which also has an expected computation time of order $N$. Using any of these methods, the computational cost of multinomial resampling scales only linearly in $N$ and $M$, which makes the method practicable even when a large number of particles is used.

### 4.4.2 Alternatives to Multinomial Resampling

Instead of using the multinomial sampling scheme, it is also possible to use a different resampling (or reallocation) scheme. For $i = 1, \ldots, M$, denote by $N^i$ the number of times the $i$th element $\tilde{\xi}^i$ is selected. A resampling scheme will be said to be *unbiased with respect to* $\mathcal{G}$ if

$$\sum_{i=1}^{M} N^i = N \,, \tag{4.40}$$

$$\mathrm{E}\left[N^i \middle| \mathcal{G}\right] = N\omega^i, \qquad i = 1, \ldots, M \,. \tag{4.41}$$

We focus here on resampling techniques that keep the number of particles constant (see for instance Crisan *et al.*, 1999, for unbiased sampling with a random number of particles). There are many different conditions under which a resampling scheme is unbiased. The simplest unbiased scheme is multinomial resampling, for which $(N^1, \ldots, N^M)$, conditionally on $\mathcal{G}$, has the multinomimal distribution $\mathrm{Mult}(N, \omega^1, \ldots, \omega^N)$. Because $I^1, \ldots, I^M$ are conditionally i.i.d. given $\mathcal{G}$, it is easy to evaluate the conditional variance in the multinomial resampling scheme:

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{\xi}^{I^i}) \middle| \mathcal{G}\right] &= \frac{1}{N}\sum_{i=1}^{M} \omega^i \left[f(\tilde{\xi}^i) - \sum_{j=1}^{M} \omega^j f(\tilde{\xi}^j)\right]^2 \\ &= \frac{1}{N}\left\{\sum_{i=1}^{M} \omega^i f^2(\tilde{\xi}^i) - \left[\sum_{i=1}^{M} \omega^i f(\tilde{\xi}^i)\right]^2\right\} \,. \end{aligned} \tag{4.42}$$

A sensible objective is to try to construct resampling schemes for which the conditional variance $\mathrm{Var}(\sum_{i=1}^{N} \frac{N^i}{N} f(\tilde{\xi}^i) \,|\, \mathcal{G})$ is as small as possible and, in particular, smaller than (4.42), preferably for any choice of the function $f$.

#### Residual Resampling

Residual resampling, or *remainder resampling*, is mentioned by Whitley (1994) (see also Liu and Chen, 1998) as a simple means to decrease the variance incurred by the sampling step. In this scheme, for $i = 1, \ldots, M$ we set

$$N^i = \lfloor N\omega^i \rfloor + \bar{N}^i \,, \tag{4.43}$$

where $\bar{N}^1, \ldots, \bar{N}^M$ are distributed, conditionally on $\mathcal{G}$, according to the multinomial distribution $\mathrm{Mult}(N-R, \bar{\omega}^1, \ldots, \bar{\omega}^N)$ with $R = \sum_{i=1}^M \lfloor N\omega^i \rfloor$ and

$$\bar{\omega}^i = \frac{N\omega^i - \lfloor N\omega^i \rfloor}{N-R}, \quad i = 1, \ldots, M \,. \tag{4.44}$$

This scheme is obviously unbiased with respect to $\mathcal{G}$. Equivalently, for any measurable function $f$, the residual sampling estimator is

$$\frac{1}{N}\sum_{i=1}^N f(\xi^i) = \sum_{i=1}^M \frac{\lfloor N\omega^i \rfloor}{N} f(\tilde{\xi}^i) + \frac{1}{N}\sum_{i=1}^{N-R} f(\tilde{\xi}^{J^i}) \,, \tag{4.45}$$

where $J^1, \ldots, J^{N-R}$ are conditionally independent given $\mathcal{G}$ with distribution $\mathrm{P}(J^i = k \,|\, \mathcal{G}) = \bar{\omega}^k$ for $i = 1, \ldots, N-R$ and $k = 1, \ldots, M$. Because the residual resampling estimator is the sum of one term that, given $\mathcal{G}$, is deterministic and one term that involves conditionally i.i.d. labels, the variance of residual resampling is given by

$$\begin{aligned} \frac{1}{N^2}\,\mathrm{Var}\left[\sum_{i=1}^{N-R} f(\tilde{\xi}^{J^i}) \,\middle|\, \mathcal{G}\right] &= \frac{N-R}{N^2}\,\mathrm{Var}\left[f(\tilde{\xi}^{J^1}) \,\middle|\, \mathcal{G}\right] \qquad (4.46) \\ &= \frac{(N-R)}{N^2}\sum_{i=1}^M \bar{\omega}^i \left\{ f(\tilde{\xi}^i) - \sum_{j=1}^M \bar{\omega}^j f(\tilde{\xi}^j) \right\}^2 \\ &= \frac{1}{N}\sum_{i=1}^M \omega^i f^2(\tilde{\xi}^i) - \sum_{i=1}^M \frac{\lfloor N\omega^i \rfloor}{N^2} f^2(\tilde{\xi}^i) - \frac{N-R}{N^2}\left\{\sum_{i=1}^M \bar{\omega}^i f(\tilde{\xi}^i)\right\}^2 . \end{aligned}$$

Residual sampling dominates multinomial sampling also in the sense of having smaller conditional variance. Indeed, first write

$$\sum_{i=1}^M \omega^i f(\tilde{\xi}^i) = \sum_{i=1}^M \frac{\lfloor N\omega^i \rfloor}{N} f(\tilde{\xi}^i) + \frac{N-R}{N}\sum_{i=1}^M \bar{\omega}^i f(\tilde{\xi}^i) \,.$$

Then note that the sum of the $M$ numbers $\lfloor N\omega^i \rfloor / N$ plus $(N-R)/N$ equals one, whence this sequence of $M+1$ numbers can be viewed as a probability distribution. Thus Jensen's inequality applied to the square of the right-hand side of the above display yields

$$\left\{\sum_{i=1}^M \omega^i f(\tilde{\xi}^i)\right\}^2 \le \sum_{i=1}^M \frac{\lfloor N\omega^i \rfloor}{N} f^2(\tilde{\xi}^i) + \frac{N-R}{N}\left\{\sum_{i=1}^M \bar{\omega}^i f(\tilde{\xi}^i)\right\}^2 .$$

Combining with (4.46) and (4.42), this shows that the conditional variance of residual sampling is always smaller than that of multinomial sampling.

### Stratified Resampling

The inversion method for sampling a multinomial sequence of trials maps uniform $(0,1)$ random variables $U^1, \ldots, U^N$ into indices $I^1, \ldots, I^N$ through a deterministic function. For any function $f$,

$$\sum_{i=1}^N f(\tilde{\xi}^{I^i}) = \sum_{i=1}^N \Phi_f(U^i) \,,$$

where the function $\Phi_f$ (which depends on both $f$ and $\{\tilde{\xi}^i\}$) is defined, for any $u \in (0,1]$, by

$$\Phi_f(u) \stackrel{\text{def}}{=} f(\tilde{\xi}^{I(u)}), \qquad I(u) = \sum_{i=1}^{M} i \mathbb{1}_{(\sum_{j=1}^{i-1}\omega^j, \sum_{j=1}^{i}\omega^j]}(u) \; . \tag{4.47}$$

Note that, by construction, $\int_0^1 \Phi_f(u)\,du = \sum_{i=1}^M \omega^i f(\tilde{\xi}^i)$. To reduce the conditional variance of $\sum_{i=1}^N f(\tilde{\xi}^{I^i})$, we may change the way in which the sample $U^1, \ldots, U^N$ is drawn. A possible solution, commonly used in survey sampling, is based on *stratification* (see Kitagawa, 1996, and Fearnhead, 1998, Section 5.3, for discussion of the method in the context of particle filtering). The interval $(0,1]$ is partitioned into different *strata*, assumed for simplicity to be intervals $(0,1] = (0,1/N] \cup (1/N, 2/N] \cup \cdots \cup (\{N-1\}/N, 1]$. More general partitions could have been considered as well; in particular, the number of partitions does not have to equal $N$, and the interval lengths could be made dependent on the $\omega^i$. One then draws a sample $\tilde{U}^1, \ldots, \tilde{U}^N$ conditionally independently given $\mathcal{G}$ from the distribution $\tilde{U}^i \sim \mathrm{U}\left(\left(\{i-1\}/N, i/N\right]\right)$ (for $i = 1, \ldots, N$) and let $\tilde{I}^i = I(\tilde{U}^i)$ with $I$ as in (4.47) (see Figure 4.16). By construction, the difference between $\tilde{N}^i = \sum_{j=1}^N \mathbb{1}_{\{\tilde{I}^j = i\}}$ and the target (non-integer) value $N\omega^i$ is less than one in absolute value. It also follows that

$$\begin{aligned}
\mathrm{E}\left[\sum_{i=1}^N f(\tilde{\xi}^{\tilde{I}^i}) \,\middle|\, \mathcal{G}\right] &= \mathrm{E}\left[\sum_{i=1}^N \Phi_f(\tilde{U}^i) \,\middle|\, \mathcal{G}\right] \\
&= N \sum_{i=1}^N \int_{(i-1)/N}^{i/N} \Phi_f(u)\,du = N \int_0^1 \Phi_f(u)\,du = N \sum_{i=1}^M \omega^i f(\tilde{\xi}^i) \; ,
\end{aligned}$$

showing that the stratified sampling scheme is unbiased. Because $\tilde{U}^1, \ldots, \tilde{U}^N$ are conditionally independent given $\mathcal{G}$,

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^N f(\tilde{\xi}^{\tilde{I}^i}) \,\middle|\, \mathcal{G}\right] &= \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^N \Phi_f(\tilde{U}^i) \,\middle|\, \mathcal{G}\right] \\
&= \frac{1}{N^2}\sum_{i=1}^N \mathrm{Var}\left[\Phi_f(\tilde{U}^i) \,\middle|\, \mathcal{G}\right] \\
&= \frac{1}{N}\sum_{i=1}^M \omega^i f^2(\tilde{\xi}^i) - \frac{1}{N}\sum_{i=1}^N \left[N \int_{(i-1)/N}^{i/N} \Phi_f(u) du\right]^2 \; ;
\end{aligned}$$

here we used that $\int_0^1 \Phi_f^2(u)\,du = \int_0^1 \Phi_{f^2}(u)\,du = \sum_{i=1}^M \omega^i f^2(\tilde{\xi}^i)$. By Jensen's inequality,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^N \left[N \int_{(i-1)/N}^{i/N} \Phi_f(u) du\right]^2 &\geq \left[\sum_{i=1}^N \int_{(i-1)/N}^{i/N} \Phi_f(u) du\right]^2 \\
&= \left[\sum_{i=1}^M \omega^i f(\tilde{\xi}^i)\right]^2 \; ,
\end{aligned}$$

showing that the conditional variance of stratified sampling is always smaller than that of multinomial sampling.

**Remark 93.** Note that stratified sampling may be coupled with the residual sampling method discussed previously: the proof above shows that using stratified sampling on the $R$ residual indices that are effectively drawn randomly can only decrease the conditional variance.

**Systematic Resampling**

Stratified sampling aims at reducing the *discrepancy*

$$D_N^\star(U^1,\dots,U^N) \stackrel{\text{def}}{=} \sup_{a\in(0,1]} \left| \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{(0,a]}(U^i) - a \right|$$

of the sample $U$ from the uniform distribution function on $(0,1]$. This is simply the Kolmogorov-Smirnov distance between the empirical distribution function of the sample and the distribution function of the uniform distribution. The Koksma-Hlawka inequality (Niederreiter, 1992) shows that for any function $f$ having bounded variation on $[0,1]$,

$$\left| \frac{1}{N}\sum_{i=1}^N f(u^i) - \int_0^1 f(u)\,du \right| \le C(f) D_N^\star(u^1,\dots,u^N) \;,$$

where $C(f)$ is the variation of $f$. This inequality suggests that it is desirable to design random sequences $U^1,\dots,U^N$ whose expected discrepancy is as low as possible. This provides another explanation of the improvement brought by stratified resampling (compared to multinomial resampling).

Pursuing in this direction, it makes sense to look for sequences with even smaller average discrepancy. One such sequence is $U^i = U + (i-1)/N$, where $U$ is drawn from a uniform $\mathrm{U}((0,1/N])$ distribution. In survey sampling, this method is known as *systematic sampling*. It was introduced in the particle filter literature by Carpenter *et al.* (1999) but is mentioned by Whitley (1994) under the name of *universal* sampling. The interval $(0,1]$ is still divided into $N$ sub-intervals $(\{i-1\}/N, i/N]$ and one sample is taken from each of them, as in stratified sampling. However, the samples are no longer independent, as they have the same relative position within each stratum (see Figure 4.17). This sampling scheme is obviously still unbiased. Because the samples are not taken independently across strata, it is however not possible to obtain simple formulas for the conditional variance (Künsch, 2003). It is often conjectured that the conditional variance of systematic resampling is always lower than that of multinomial resampling. This is not correct, as demonstrated by the following example.

**Example 94.** Consider the case where the initial population of particles $\{\tilde{\xi}^i\}_{1\le i\le N}$ is composed of the interleaved repetition of only two distinct values $x_0$ and $x_1$, with identical multiplicities (assuming $N$ to be even). In other words,

$$\{\tilde{\xi}^i\}_{1\le i\le N} = \{x_0, x_1, x_0, x_1, \dots, x_0, x_1\} \;.$$

We denote by $2\omega/N$ the common value of the normalized weight $\omega^i$ associated to the $N/2$ particles $\tilde{\xi}^i$ that satisfy $\tilde{\xi}^i = x_1$, so that the remaining ones (which are such that $\tilde{\xi}^i = x_0$) share a common weight of $2(1-\omega)/N$. Without loss of generality, we assume that $1/2 \le \omega < 1$ and that the function of interest $f$ is such that $f(x_0) = 0$ and $f(x_1) = F$.

Under multinomial resampling, (4.42) shows that the conditional variance of the estimate $N^{-1}\sum_{i=1}^N f(\xi^i)$ is given by

$$\mathrm{Var}\left[ \frac{1}{N}\sum_{i=1}^N f(\xi^i_{\text{mult}}) \middle| \mathcal{G} \right] = \frac{1}{N}(1-\omega)\omega F^2 \;. \tag{4.48}$$

Because the value $2\omega/N$ is assumed to be larger than $1/N$, it is easily checked that systematic resampling deterministically sets $N/2$ of the $\xi^i$ to be equal to $x_1$.

| $\omega$ | 0.51 | 0.55 | 0.6 | 0.65 | 0.70 | 0.75 |
|---|---|---|---|---|---|---|
| Multinomial | 0.050 | 0.049 | 0.049 | 0.048 | 0.046 | 0.043 |
| Residual, stratified | 0.010 | 0.021 | 0.028 | 0.032 | 0.035 | 0.035 |
| Systematic | 0.070 | 0.150 | 0.200 | 0.229 | 0.245 | 0.250 |
| Systematic with prior random shuffling | 0.023 | 0.030 | 0.029 | 0.029 | 0.028 | 0.025 |

Table 4.1: Standard deviations of various resampling methods for $N = 100$ and $F = 1$. The bottom line has been obtained by simulations, averaging 100,000 Monte Carlo replications.

Depending on the draw of the initial shift, *all* the $N/2$ remaining particles are either set to $x_1$, with probability $2\omega - 1$, or to $x_0$, with probability $2(1 - \omega)$. Hence the variance is that of a *single* Bernoulli draw scaled by $N/2$, that is,

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} f(\xi^i_{\mathrm{syst}})\,\middle|\,\mathcal{G}\right] = (\omega - 1/2)(1 - \omega)F^2 \,.$$

Note that in this case, the conditional variance of systematic resampling is not only larger than (4.48) for most values of $\omega$ (except when $\omega$ is very close to $1/2$), but it does not even decrease to zero as $N$ grows! Clearly, this observation is very dependent on the order in which the initial population of particles is presented. Interestingly, this feature is common to the systematic and stratified sampling schemes, whereas the multinomial and residual approaches are unaffected by the order in which the particles are labelled. In this particular example, it is straightforward to verify that residual and stratified resampling are equivalent—which is not the case in general—and amount to deterministically setting $N/2$ particles to the value $x_1$, whereas the $N/2$ remaining ones are drawn by $N/2$ *conditionally independent* Bernoulli trials with probability of picking $x_1$ equal to $2\omega - 1$. Hence the conditional variance, for both the residual and stratified schemes, is equal to $N^{-1}(2\omega - 1)(1 - \omega)F^2$. It is hence always smaller than (4.48), as expected from the general study of these two methods.

Once again, the failure of systematic resampling in this example is entirely due to the specific order in which the particles are labeled: it is easy to verify, at least empirically, that the problem vanishes upon randomly permuting the initial particles before applying systematic resampling. Table 4.1 also shows that a common feature of both the residual, stratified, and systematic resampling procedures is to become very efficient in some particular configurations of the weights such as when $\omega = 0.51$ for which the probabilities of selecting the two types of particles are almost equal and the selection becomes quasi-deterministic. Note also that prior random shuffling does somewhat compromise this ability in the case of systematic resampling.

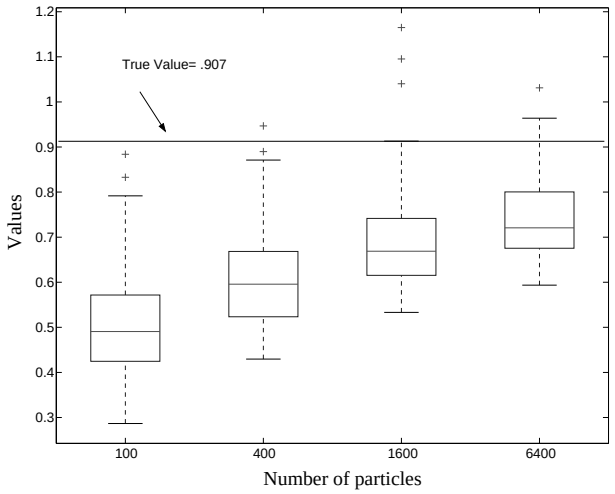



Figure 4.3: Box and whisker plot of the posterior mean estimate of $X_5$ obtained from 125 replications of the SIS filter using the prior kernel and increasing numbers of particles. The horizontal line represents the true posterior mean.

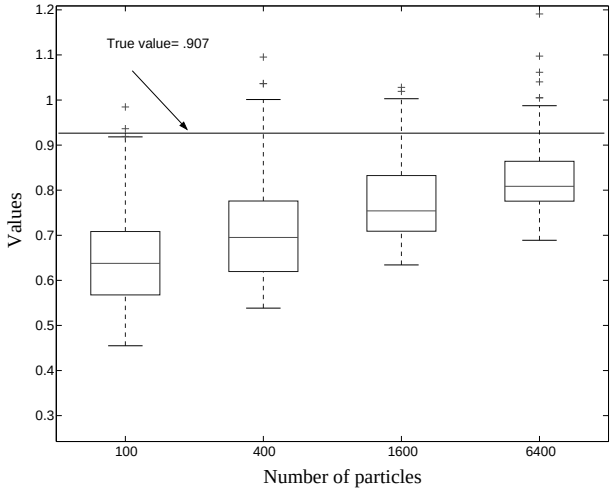



Figure 4.4: Box and whisker plot of the posterior mean estimate for $X_5$ obtained from 125 replications of the SIS filter using the optimal kernel and increasing numbers of particles. Same data and axes as Figure 4.3.

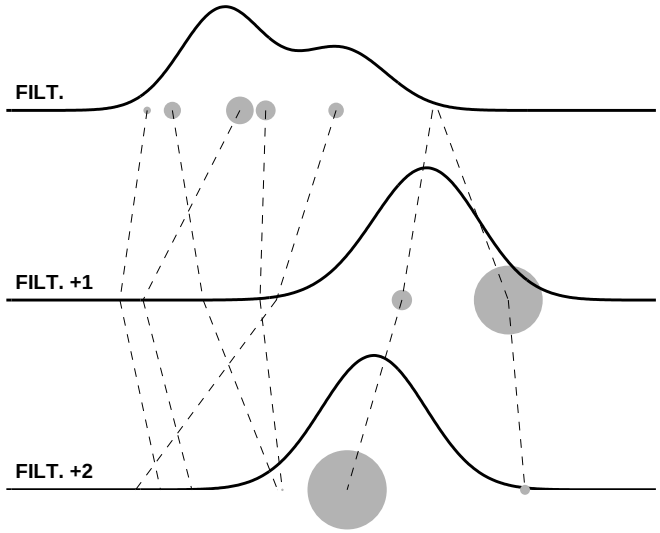



Figure 4.5: SIS using the prior kernel. The positions of the particles are indicated by circles whose radii are proportional to the normalized importance weights. The solid lines show the filtering distributions for three consecutive time indices.

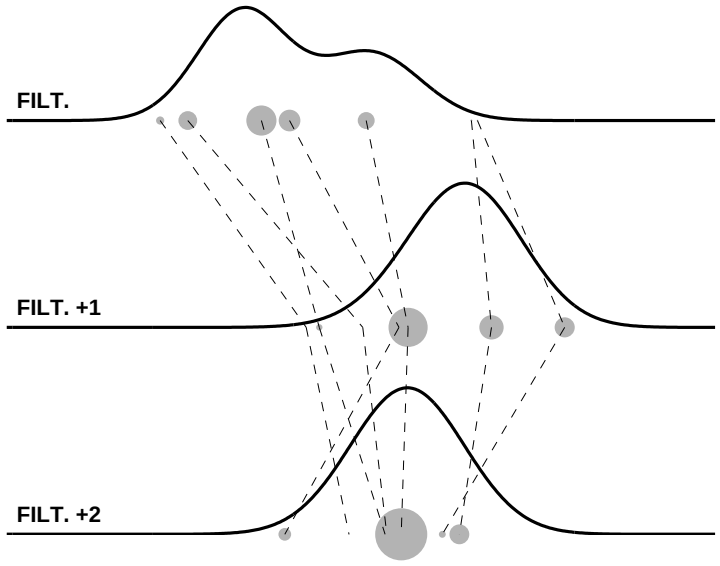



Figure 4.6: SIS using the optimal kernel (same data and display as in Figure 4.5).

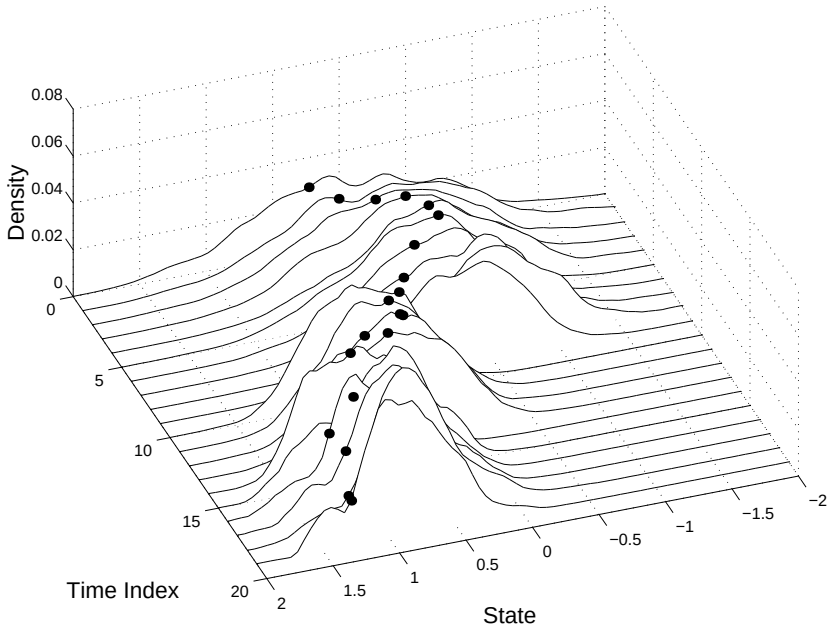



Figure 4.7: Waterfall representation of filtering distributions as estimated by SIS with $N = 1{,}000$ particles (densities estimated with Epanechnikov kernel, bandwidth 0.2). Data is the same as in Figure **??**.

.

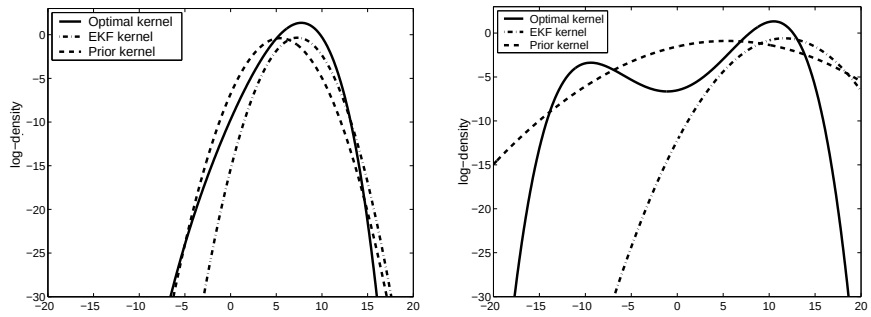



Figure 4.8: Log-density of the optimal kernel (solid line), EKF approximation of the optimal kernel (dashed-dotted line), and the prior kernel (dashed line) for two different values of the state noise variance $\sigma_u^2$: left, $\sigma_u^2 = 1$; right, $\sigma_u^2 = 10$.

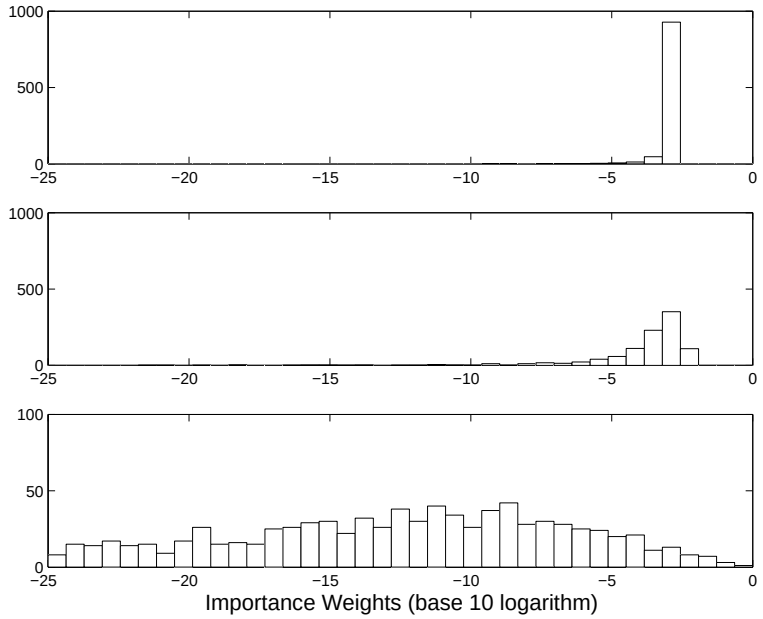



Figure 4.9: Histograms of the base 10 logarithm of the normalized importance weights after (from top to bottom) 1, 10, and 100 iterations for the stochastic volatility model of Example 80. Note that the vertical scale of the bottom panel has been multiplied by 10.

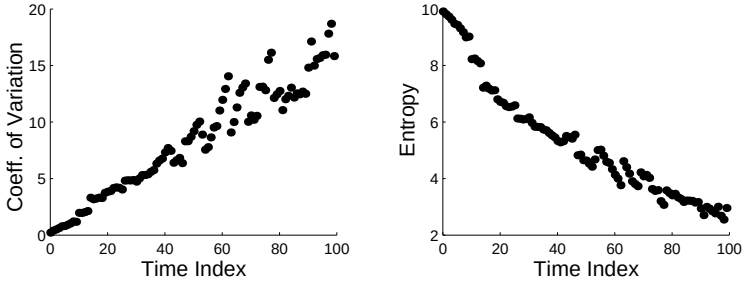



Figure 4.10: Coefficient of variation (left) and entropy of the normalized importance weights as a function of the number of iterations for the stochastic volatility model of Example 80. Same model and data as in Figure 4.9.

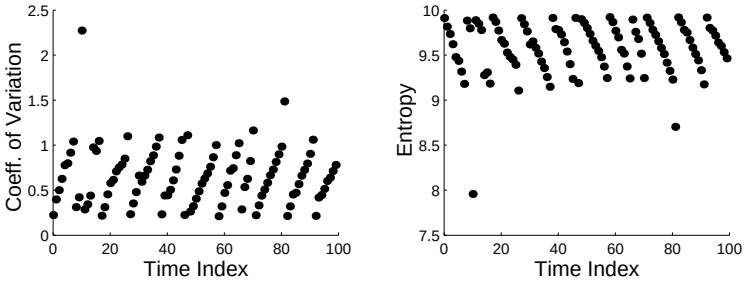



Figure 4.11: Coefficient of variation (left) and entropy of the normalized importance weights as a function of the number of iterations in the stochastic volatility model of Example 80. Same model and data as in Figure 4.10. Resampling occurs when the coefficient of variation gets larger than 1.

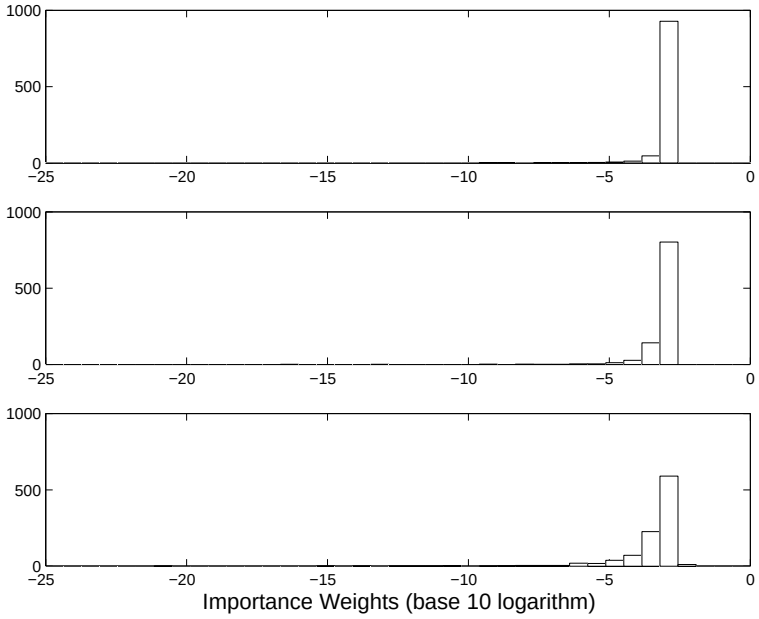



Figure 4.12: Histograms of the base 10 logarithm of the normalized importance weights after (from top to bottom) 1, 10, and 100 iterations in the stochastic volatility model of Example 80. Same model and data as in Figure 4.9. Resampling occurs when the coefficient of variation gets larger than 1.

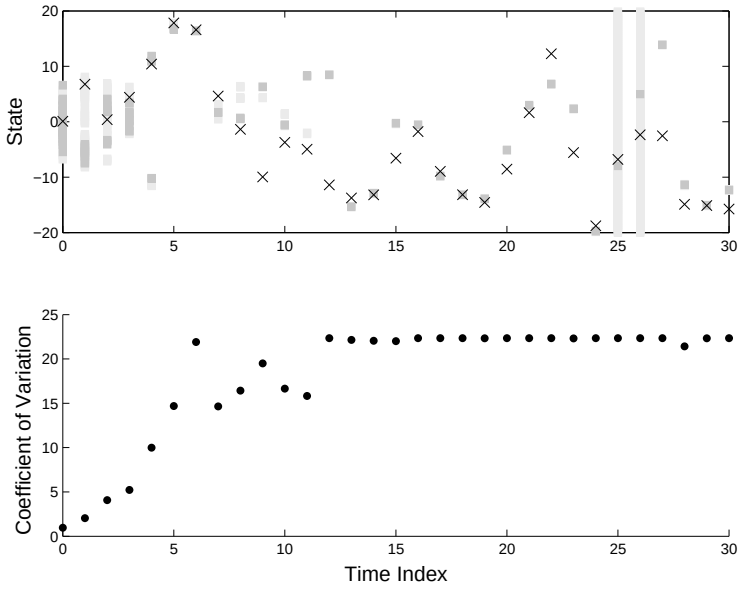
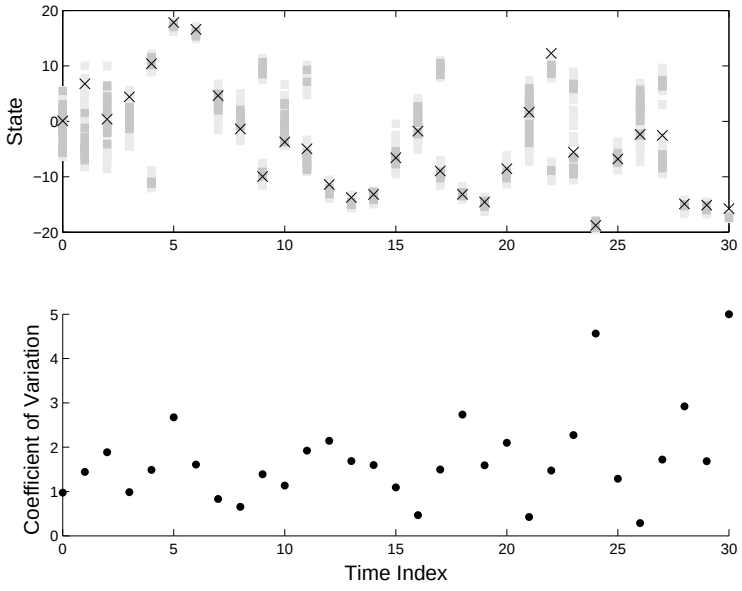



Figure 4.13: SIS estimates of the filtering distributions in the growth model with instrumental kernel being the prior one and 500 particles. Top: true state sequence (×) and 95%/50% HPD regions (light/dark grey) of estimated filtered distribution. Bottom: coefficient of variation of the normalized importance weights.




p]

Figure 4.14: Same legend for Figure 4.13, but with results for the corresponding bootstrap filter.

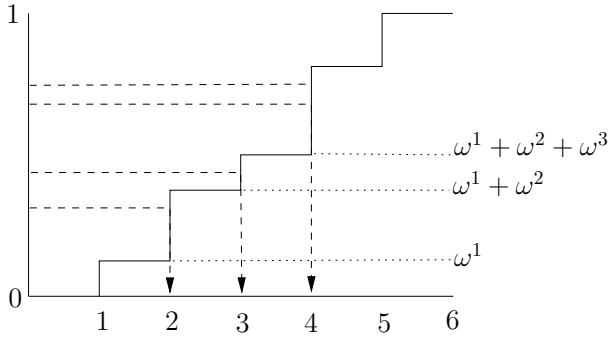



Figure 4.15: Multinomial sampling from uniform distribution by the inversion method.

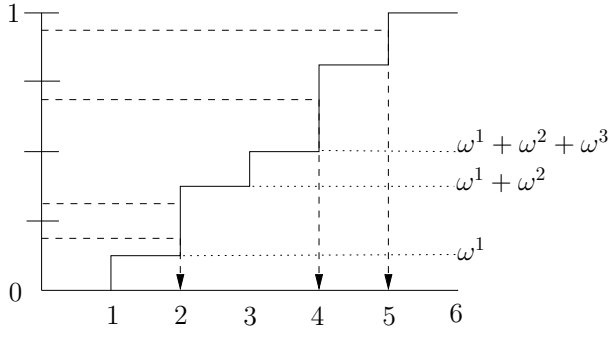



Figure 4.16: Stratified sampling: the interval $(0, 1]$ is divided into $N$ intervals $((i-1)/N, i/N]$. One sample is drawn uniformly from each interval, independently of samples drawn in the other intervals.

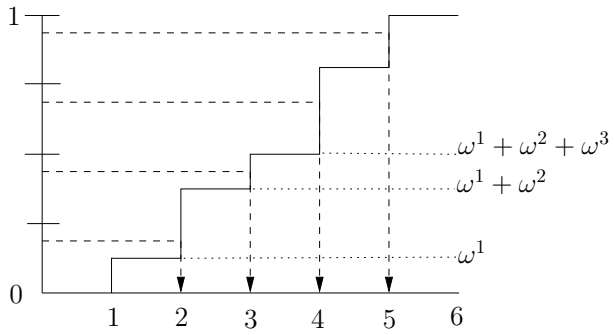



Figure 4.17: Systematic sampling: the unit interval is divided into $N$ intervals $((i-1)/N, i/N]$ and one sample is drawn from each of them. Contrary to stratified sampling, each sample has the same relative position within its stratum.

# Part II

# Parameter Inference

# Chapter 5

# Maximum Likelihood Inference, Part I: Optimization Through Exact Smoothing

In previous chapters, we have focused on structural results and methods for HMMs, considering in particular that the models under consideration were always perfectly known. In most situations, however, the model cannot be fully specified beforehand, and some of its parameters need to be calibrated based on observed data. Except for very simplistic instances of HMMs, the structure of the model is sufficiently complex to prevent the use of direct estimators such as those provided by moment or least squares methods. We thus focus in the following on computation of the *maximum likelihood estimator*.

Given the specific structure of the likelihood function in HMMs, it turns out that the key ingredient of any optimization method applicable in this context is the ability to compute smoothed functionals of the unobserved sequence of states. Hence the methods discussed in the second part of the book for evaluating smoothed quantities are instrumental in devising parameter estimation strategies.

This chapter only covers the class of HMMs for which the smoothing recursions may effectively be implemented on computers. For such models, the likelihood function is computable, and hence our main task will be to optimize a possibly complex but entirely known function. The topic of this chapter thus relates to the more general field of numerical optimization. For models that do not allow for exact numerical computation of smoothing distributions, this chapter provides a framework from which numerical approximations can be built.

## 5.1 Likelihood Optimization in Incomplete Data Models

To describe the methods as concisely as possible, we adopt a very general viewpoint in which we only assume that the likelihood function of interest may be written as the marginal of a higher dimensional function. In the terminology introduced by Dempster *et al.* (1977), this higher dimensional function is described as the *complete data* likelihood; in this framework, the term *incomplete data* refers to the actual observed data while the *complete data* is a (not fully observable) higher

dimensional random variable. In Section 5.2, we will exploit the specific structure of the HMM, and in particular the fact that it corresponds to a *missing data model* in which the observations simply are a subset of the complete data. We ignore these specifics for the moment however and consider the general likelihood optimization problem in incomplete data models.

### 5.1.1 Problem Statement and Notations

Given a $\sigma$-finite measure $\lambda$ on $(\mathsf{X}, \mathcal{X})$, we consider a family $\{f(\cdot;\theta)\}_{\theta\in\Theta}$ of non-negative $\lambda$-integrable functions on $\mathsf{X}$. This family is indexed by a parameter $\theta \in \Theta$, where $\Theta$ is a subset of $\mathbb{R}^{d_\theta}$ (for some integer $d_\theta$). The task under consideration is the maximization of the integral

$$\mathrm{L}(\theta) \stackrel{\text{def}}{=} \int f(x\,;\theta)\,\lambda(dx) \tag{5.1}$$

with respect to the parameter $\theta$. The function $f(\cdot\,;\theta)$ may be thought of as an *unnormalized probability density* with respect to $\lambda$. Thus $\mathrm{L}(\theta)$ is the normalizing constant for $f(\cdot\,;\theta)$. In typical examples, $f(\cdot\,;\theta)$ is a relatively simple function of $\theta$. In contrast, the quantity $\mathrm{L}(\theta)$ usually involves high-dimensional integration and is therefore sufficiently complex to prevent the use of simple maximization approaches; even the direct evaluation of the function might turn out to be non-feasible.

In Section 5.2, we shall consider more specifically the case where $f$ is the joint probability density function of two random variables $X$ and $Y$, the latter being observed while the former is not. Then $X$ is referred to as the *missing data*, $f$ is the *complete data likelihood*, and L is the density of $Y$ alone, that is, the *likelihood* available for estimating $\theta$. Note however that thus far, the dependence on $Y$ is not made explicit in the notation; this is reminiscent of the implicit conditioning convention discussed in Section 2.1.4 in that the observations do not appear explicitly. Having sketched these statistical ideas, we stress that we feel it is actually easier to understand the basic mechanisms at work without relying on the probabilistic interpretation of the above quantities. In particular, it is not required that L be a likelihood, as any function satisfying (5.1) is a valid candidate for the methods discussed here (cf. Remark **??**).

In the following, we will assume that $\mathrm{L}(\theta)$ is positive, and thus maximizing $\mathrm{L}(\theta)$ is equivalent to maximizing

$$\ell(\theta) \stackrel{\text{def}}{=} \log \mathrm{L}(\theta)\ . \tag{5.2}$$

In a statistical setting, $\ell$ is the *log-likelihood*. We also associate to each function $f(\cdot\,;\theta)$ the probability density function $p(\cdot\,;\theta)$ (with respect to the dominating measure $\lambda$) defined by

$$p(x\,;\theta) \stackrel{\text{def}}{=} f(x\,;\theta)/\mathrm{L}(\theta)\ . \tag{5.3}$$

In the statistical setting sketched above, $p(x;\theta)$ is the conditional density of $X$ given $Y$.

### 5.1.2 The Expectation-Maximization Algorithm

The most popular method for solving the general optimization problem outlined above is the EM (for *expectation-maximization*) algorithm introduced, in its full generality, by Dempster *et al.* (1977) in their landmark paper. Given the literature available on the topic, our aim is not to provide a comprehensive review of all the results related to the EM algorithm but rather to highlight some of its key features and properties in the context of hidden Markov models.

### The Intermediate Quantity of EM

The central concept in the framework introduced by Dempster *et al.* (1977) is an auxiliary function (or, more precisely, a family of auxiliary functions) known as the intermediate quantity of EM.

**Definition 95** (Intermediate Quantity of EM)**.** *The* intermediate quantity of EM *is the family* $\{\mathcal{Q}(\cdot\,;\theta')\}_{\theta'\in\Theta}$ *of real-valued functions on* $\Theta$*, indexed by* $\theta'$ *and defined by*

$$\mathcal{Q}(\theta\,;\theta') \stackrel{\text{def}}{=} \int \log f(x\,;\theta) p(x\,;\theta')\,\lambda(dx)\;. \tag{5.4}$$

**Remark 96.** To ensure that $\mathcal{Q}(\theta\,;\theta')$ is indeed well-defined for all values of the pair $(\theta,\theta')$, one needs regularity conditions on the family of functions $\{f(\cdot\,;\theta)\}_{\theta\in\Theta}$, which will be stated below (Assumption 97). To avoid trivial cases however, we use the convention $0\log 0 = 0$ in (5.4) and in similar relations below. In more formal terms, for every measurable set $N$ such that *both* $f(x\,;\theta)$ and $p(x\,;\theta')$ vanish $\lambda$-a.e. on $N$, set

$$\int_N \log f(x\,;\theta) p(x\,;\theta')\,\lambda(dx) \stackrel{\text{def}}{=} 0\;.$$

With this convention, $\mathcal{Q}(\theta\,;\theta')$ stays well-defined in cases where there exists a non-empty set $N$ such that *both* $f(x\,;\theta)$ and $f(x\,;\theta')$ vanish $\lambda$-a.e. on $N$.

The intermediate quantity $\mathcal{Q}(\theta\,;\theta')$ of EM may be interpreted as the expectation of the function $\log f(X\,;\theta)$ when $X$ is distributed according to the probability density function $p(\cdot\,;\theta')$ indexed by a, possibly different, value $\theta'$ of the parameter. Using (5.2) and (5.3), one may rewrite the intermediate quantity of EM in (5.4) as

$$\mathcal{Q}(\theta\,;\theta') = \ell(\theta) - \mathcal{H}(\theta\,;\theta')\;, \tag{5.5}$$

where

$$\mathcal{H}(\theta\,;\theta') \stackrel{\text{def}}{=} -\int \log p(x\,;\theta) p(x\,;\theta')\,\lambda(dx)\;. \tag{5.6}$$

Equation (5.5) states that the intermediate quantity $\mathcal{Q}(\theta\,;\theta')$ of EM differs from (the log of) the objective function $\ell(\theta)$ by a quantity that has a familiar form. Indeed, $\mathcal{H}(\theta'\,;\theta')$ is recognized as the *entropy* of the probability density function $p(\cdot\,;\theta')$ (see for instance Cover and Thomas, 1991). More importantly, the increment of $\mathcal{H}(\theta\,;\theta')$,

$$\mathcal{H}(\theta\,;\theta') - \mathcal{H}(\theta'\,;\theta') = -\int \log \frac{p(x\,;\theta)}{p(x\,;\theta')} p(x\,;\theta')\,\lambda(dx)\;, \tag{5.7}$$

is recognized as the *Kullback-Leibler divergence* (or *relative entropy*) between the probability density functions $p$ indexed by $\theta$ and $\theta'$, respectively.

The last piece of notation needed is the following: the gradient and Hessian of a function, say L, at $\theta'$ will be denoted by $\nabla_\theta \mathrm{L}(\theta')$ and $\nabla^2_\theta \mathrm{L}(\theta')$, respectively. To avoid ambiguities, the gradient of $\mathcal{H}(\cdot\,;\theta')$ with respect to its first argument, evaluated at $\theta''$, will be denoted by $\nabla_\theta \mathcal{H}(\theta\,;\theta')|_{\theta=\theta''}$ (where the same convention will also be used, if needed, for the Hessian).

We conclude this introductory section by stating a minimal set of assumptions that guarantee that all quantities introduced so far are indeed well-defined.

**Assumption 97.**

*(i) The parameter set* $\Theta$ *is an open subset of* $\mathbb{R}^{d_\theta}$ *(for some integer* $d_\theta$*).*

*(ii) For any* $\theta\in\Theta$*,* $\mathrm{L}(\theta)$ *is positive and finite.*

*(iii) For any* $(\theta,\theta')\in\Theta\times\Theta$*,* $\int |\nabla_\theta \log p(x\,;\theta)| p(x\,;\theta')\,\lambda(dx)$ *is finite.*

.

Assumption 97(iii) implies in particular that the probability distributions in the family $\{p(\cdot\,;\theta)\,d\lambda\}_{\theta\in\Theta}$ are all absolutely continuous with respect to one another. Any individual distribution $p(\cdot\,;\theta)\,d\lambda$ can only vanish on sets that are assigned null probability by all other probability distributions in the family. Thus both $\mathcal{H}(\theta\,;\theta')$ and $\mathcal{Q}(\theta\,;\theta')$ are well-defined for all pairs of parameters.

### The Fundamental Inequality of EM

We are now ready to state the fundamental result that justifies the standard construction of the EM algorithm.

**Proposition 98.** *Under Assumption 97, for any* $(\theta,\theta')\in\Theta\times\Theta$,

$$\ell(\theta)-\ell(\theta')\geq\mathcal{Q}(\theta\,;\theta')-\mathcal{Q}(\theta'\,;\theta')\;, \tag{5.8}$$

*where the inequality is strict unless* $p(\cdot\,;\theta)$ *and* $p(\cdot\,;\theta')$ *are equal* $\lambda$*-a.e.*

*Assume in addition that*

*(a)* $\theta\mapsto\mathrm{L}(\theta)$ *is continuously differentiable on* $\Theta$*;*

*(b) for any* $\theta'\in\Theta$, $\theta\mapsto\mathcal{H}(\theta\,;\theta')$ *is continuously differentiable on* $\Theta$.

*Then for any* $\theta'\in\Theta$, $\theta\mapsto\mathcal{Q}(\theta\,;\theta')$ *is continuously differentiable on* $\Theta$ *and*

$$\nabla_\theta\ell(\theta')=\nabla_\theta\mathcal{Q}(\theta\,;\theta')|_{\theta=\theta'}\quad. \tag{5.9}$$

*Proof.* The difference between the left-hand side and the right-hand side of (5.8) is the quantity defined in (5.7), which we already recognized as a Kullback-Leibler distance. Under Assumption 97(iii), this latter term is well-defined and known to be strictly positive (by direct application of Jensen's inequality) unless $p(\cdot\,;\theta)$ and $p(\cdot\,;\theta')$ are equal $\lambda$-a.e. (Cover and Thomas, 1991; Lehmann and Casella, 1998).

For (5.9), first note that $\mathcal{Q}(\theta\,;\theta')$ is a differentiable function of $\theta$, as it is the difference of two functions that are differentiable under the additional assumptions (a) and (b). Next, the previous discussion implies that $\mathcal{H}(\theta\,;\theta')$ is minimal for $\theta=\theta'$, although this may not be the only point where the minimum is achieved. Thus its gradient vanishes at $\theta'$, which proves (5.9). □

### The EM Algorithm

The essence of the EM algorithm, which is suggested by (5.5), is that $\mathcal{Q}(\theta\,;\theta')$ may be used as a surrogate for $\ell(\theta)$. Both functions are not necessarily comparable but, in view of (5.8), any value of $\theta$ such that $\mathcal{Q}(\theta\,;\theta')$ is increased over its baseline $\mathcal{Q}(\theta'\,;\theta')$ corresponds to an increase of $\ell$ (relative to $\ell(\theta')$) that is at least as large.

The EM algorithm as proposed by Dempster *et al.* (1977) consists in iteratively building a sequence $\{\theta^i\}_{i\geq1}$ of parameter estimates given an initial guess $\theta^0$. Each iteration is classically broken into two steps as follows.

E-Step: Determine $\mathcal{Q}(\theta\,;\theta^i)$;

M-Step: Choose $\theta^{i+1}$ to be the (or any, if there are several) value of $\theta\in\Theta$ that maximizes $\mathcal{Q}(\theta\,;\theta^i)$.

Proposition 98 provides the two decisive arguments behind the EM algorithm. First, an immediate consequence of (5.8) is that, by the very definition of the sequence $\{\theta^i\}$, the sequence $\{\ell(\theta^i)\}_{i\geq0}$ of log-likelihood values is non-decreasing. Hence EM is a monotone optimization algorithm. Second, if the iterations ever stop at a point

$\theta_\star$, then $\mathcal{Q}(\theta\,;\theta_\star)$ has to be maximal at $\theta_\star$ (otherwise it would still be possible to improve over $\theta_\star$), and hence $\theta_\star$ is such that $\nabla_\theta \mathrm{L}(\theta_\star) = 0$, that is, this is a *stationary point of the likelihood*.

Although this picture is largely correct, there is a slight flaw in the second half of the above intuitive reasoning in that the if part *(if the iterations ever stop at a point)* may indeed never happen. Stronger conditions are required to ensure that the sequence of parameter estimates produced by EM from any starting point indeed converges to a limit $\theta_\star \in \Theta$. However, it is actually true that when convergence to a point takes place, the limit has to be a stationary point of the likelihood. In order not to interrupt our presentation of the EM framework, convergence results pertaining to the EM algorithm are deferred to Section 5.5 at the end of this chapter; see in particular Theorems 105 and 106.

**EM in Exponential Families**

**Definition 99** (Exponential Family)**.** *The family* $\{f(\cdot\,;\theta)\}_{\theta\in\Theta}$ *defines an* exponential family *of positive functions on* $\mathsf{X}$ *if*

$$f(x\,;\theta) = \exp\{\psi(\theta)^t S(x) - c(\theta)\}h(x)\ , \tag{5.10}$$

*where* $S$ *and* $\psi$ *are vector-valued functions (of the same dimension) on* $\mathsf{X}$ *and* $\Theta$ *respectively,* $c$ *is a real-valued function on* $\Theta$ *and* $h$ *is a non-negative real-valued function on* $\mathsf{X}$*.*

Here $S(x)$ is known as the vector of *natural sufficient statistics*, and $\eta = \psi(\theta)$ is the *natural parameterization*. If $\{f(\cdot\,;\theta)\}_{\theta\in\Theta}$ is an exponential family and if $\int |S(x)| f(x\,;\theta)\,\lambda(dx)$ is finite for any $\theta \in \Theta$, the intermediate quantity of EM reduces to

$$\mathcal{Q}(\theta\,;\theta') = \psi(\theta)^t \left[\int S(x) p(x\,;\theta')\,\lambda(dx)\right] - c(\theta) + \int p(x\,;\theta')\log h(x)\,\lambda(dx)\ . \tag{5.11}$$

Note that the right-most term does not depend on $\theta$ and thus plays no role in the maximization. It may as well be ignored, and in practice it is not required to compute it. Except for this term, the right-hand side of (5.11) has an explicit form as soon as it is possible to evaluate the expectation of the vector of sufficient statistics $S$ under $p(\cdot\,;\theta')$. The other important feature of (5.11), ignoring the rightmost term, is that $\mathcal{Q}(\theta\,;\theta')$, viewed as a function of $\theta$, is similar to the logarithm of (5.10) for the particular value $S_{\theta'} = \int S(x) p(x\,;\theta')\,\lambda(dx)$ of the sufficient statistic.

In summary, if $\{f(\cdot\,;\theta)\}_{\theta\in\Theta}$ is an exponential family, the two above general conditions needed for the EM algorithm to be practicable reduce to the following.

E-Step: The expectation of the vector of sufficient statistics $S(X)$ under $p(\cdot\,;\theta')$ must be computable.

M-Step: Maximization of $\psi(\theta)^t s - c(\theta)$ with respect to $\theta \in \Theta$ must be feasible in closed form for any $s$ in the convex hull of $S(\mathsf{X})$ (that is, for any valid value of the expected vector of sufficient statistics).

### 5.1.3 Gradient-based Methods

A frequently ignored observation is that in any model where the EM strategy may be applied, it is also possible to evaluate derivatives of the objective function $\ell(\theta)$ with respect to the parameter $\theta$. This is obvious from (5.9), and we will expand on this matter below. As a consequence, instead of resorting to a specific algorithm such as EM, one may borrow tools from the (comprehensive and well-documented) toolbox of gradient-based optimization methods.

**Computing Derivatives in Incomplete Data Models**

A first remark is that in cases where the EM algorithm is applicable, the objective function $\ell(\theta)$ is actually computable: because the EM requires the computation of expectations under the conditional density $p(\cdot\,;\theta)$, it is restricted to cases where the normalizing constant $\mathrm{L}(\theta)$—and hence $\ell(\theta) = \log \mathrm{L}(\theta)$—is available. The two equalities below show that it is indeed also the case for the first- and second-order derivatives of $\ell(\theta)$.

**Proposition 100** (Fisher's and Louis' Identities)**.** *Assume 97 and that the following conditions hold.*

*(a)* $\theta \mapsto \mathrm{L}(\theta)$ *is twice continuously differentiable on* $\Theta$*.*

*(b) For any* $\theta' \in \Theta$*,* $\theta \mapsto \mathcal{H}(\theta\,;\theta')$ *is twice continuously differentiable on* $\Theta$*. In addition,* $\int |\nabla_\theta^k \log p(x\,;\theta)| p(x\,;\theta')\,\lambda(dx)$ *is finite for* $k = 1, 2$ *and any* $(\theta, \theta') \in \Theta \times \Theta$*, and*

$$\nabla_\theta^k \int \log p(x\,;\theta) p(x\,;\theta')\,\lambda(dx) = \int \nabla_\theta^k \log p(x\,;\theta) p(x\,;\theta')\,\lambda(dx)\ .$$

*Then the following identities hold:*

$$\nabla_\theta \ell(\theta') = \int \nabla_\theta \log f(x\,;\theta)|_{\theta=\theta'}\, p(x\,;\theta')\,\lambda(dx)\ , \tag{5.12}$$

$$\begin{aligned} -\nabla_\theta^2 \ell(\theta') = &- \int \nabla_\theta^2 \log f(x\,;\theta)\big|_{\theta=\theta'}\, p(x\,;\theta')\,\lambda(dx) \\ &+ \int \nabla_\theta^2 \log p(x\,;\theta)\big|_{\theta=\theta'}\, p(x\,;\theta')\,\lambda(dx)\ . \end{aligned} \tag{5.13}$$

*The second equality may be rewritten in the equivalent form*

$$\begin{aligned} \nabla_\theta^2 \ell(\theta') + \{\nabla_\theta \ell(\theta')\}\,\{\nabla_\theta \ell(\theta')\}^t = \int \Big[ &\nabla_\theta^2 \log f(x\,;\theta)\big|_{\theta=\theta'} \\ &+ \{\nabla_\theta \log f(x\,;\theta)|_{\theta=\theta'}\}\,\{\nabla_\theta \log f(x\,;\theta)|_{\theta=\theta'}\}^t \Big] p(x\,;\theta')\,\lambda(dx)\ . \end{aligned} \tag{5.14}$$

Equation (5.12) is sometimes referred to as *Fisher's identity* (see the comment by B. Efron in the discussion of Dempster *et al.*, 1977, p. 29). In cases where the function L may be interpreted as the likelihood associated with some statistical model, the left-hand side of (5.12) is the *score function* (gradient of the log-likelihood). Equation (5.12) shows that the score function may be evaluated by computing the expectation, under $p(\cdot\,;\theta')$, of the function $\nabla_\theta \log f(X\,;\theta)|_{\theta=\theta'}$. This latter quantity, in turn, is referred to as the *complete score function* in a statistical context, as $\log f(x\,;\theta)$ is the joint log-likelihood of the complete data $(X, Y)$; again we remark that at this stage, $Y$ is not explicit in the notation.

Equation (5.13) is usually called the *missing information principle* after Louis (1982) who first named it this way, although it was mentioned previously in a slightly different form by Orchard and Woodbury (1972) and implicitly used in Dempster *et al.* (1977). In cases where L is a likelihood, the left-hand side of (5.13) is the associated *observed information matrix*, and the second term on the right-hand side is easily recognized as the (negative of the) Fisher information matrix associated with the probability density function $p(\cdot\,;\theta')$.

Finally (5.14), which is here written in a form that highlights its symmetry, was also proved by Louis (1982) and is thus known as *Louis' identity*. Together with

(5.12), it shows that the first- and second-order derivatives of $\ell$ may be evaluated by computing expectations under $p(\cdot\,;\theta')$ of quantities derived from $f(\cdot\,;\theta)$. We now prove these three identities.

*of Proposition 100.* Equations (5.12) and (5.13) are just (5.5) where the right-hand side is differentiated once, using (5.9), and then twice under the integral sign.

To prove (5.14), we start from (5.13) and note that the second term on its right-hand side is the negative of an information matrix for the parameter $\theta$ associated with the probability density function $p(\cdot\,;\theta)$ and evaluated at $\theta'$. We rewrite this second term using the well-known information matrix identity

$$\int \nabla_\theta^2 \log p(x\,;\theta)\big|_{\theta=\theta'}\, p(x\,;\theta')\,\lambda(dx) \\ = -\int \{\,\nabla_\theta \log p(x\,;\theta)|_{\theta=\theta'}\}\,\{\,\nabla_\theta \log p(x\,;\theta)|_{\theta=\theta'}\}^t\, p(x\,;\theta')\,\lambda(dx)\;.$$

This is again a consequence of assumption (b) and the fact that $p(\cdot\,;\theta)$ is a probability density function for all values of $\theta$, implying that

$$\int \nabla_\theta \log p(x\,;\theta)|_{\theta=\theta'}\, p(x\,;\theta')\,\lambda(dx) = 0\;.$$

Now use the identity $\log p(x\,;\theta) = \log f(x\,;\theta) - \ell(\theta)$ and (5.12) to conclude that

$$\int \{\,\nabla_\theta \log p(x\,;\theta)|_{\theta=\theta'}\}\,\{\,\nabla_\theta \log p(x\,;\theta)|_{\theta=\theta'}\}^t\, p(x\,;\theta')\,\lambda(dx) \\ = \int \{\,\nabla_\theta \log f(x\,;\theta)|_{\theta=\theta'}\}\,\{\,\nabla_\theta \log f(x\,;\theta)|_{\theta=\theta'}\}^t\, p(x\,;\theta')\,\lambda(dx) \\ - \{\nabla_\theta \ell(\theta')\}\,\{\nabla_\theta \ell(\theta')\}^t\;,$$

which completes the proof. □

**The Steepest Ascent Algorithm**

We briefly discuss the main features of gradient-based iterative optimization algorithms, starting with the simplest, but certainly not most efficient, approach. We restrict ourselves to the case where the optimization problem is *unconstrained* in the sense that $\Theta = \mathbb{R}^{d_\theta}$, so that any parameter value produced by the algorithms below is valid. For an in-depth coverage of the subject, we recommend the monographs by Luenberger (1984) and Fletcher (1987).

The simplest method is the *steepest ascent* algorithm in which the current value of the estimate $\theta^i$ is updated by adding a multiple of the gradient $\nabla_\theta \ell(\theta^i)$, referred to as the *search direction*:

$$\theta^{i+1} = \theta^i + \gamma_i \nabla_\theta \ell(\theta^i)\;. \tag{5.15}$$

Here the multiplier $\gamma_i$ is a non-negative scalar that needs to be adjusted at each iteration to ensure, *a minima*, that the sequence $\{\ell(\theta^i)\}$ is non-decreasing—as was the case for EM. The most sensible approach consists in choosing $\gamma_i$ as to maximize the objective function in the search direction:

$$\gamma_i = \arg\max_{\gamma \geq 0} \ell[\theta^i + \gamma \nabla_\theta \ell(\theta^i)]\;. \tag{5.16}$$

It can be shown (Luenberger, 1984, Chapter 7) that under mild assumptions, the steepest ascent method with multipliers (5.16) is globally convergent, with a set of

limit points corresponding to the stationary points of $\ell$ (see Section 5.5 for precise definitions of these terms and a proof that this property holds for the EM algorithm).

It remains that the use of the steepest ascent algorithm is not recommended, particularly in large-dimensional parameter spaces. The reason for this is that its speed of convergence *linear* in the sense that if the sequence $\{\theta^i\}_{i\geq 0}$ converges to a point $\theta_\star$ such that the Hessian $\nabla^2_\theta \ell(\theta_\star)$ is negative definite (see Section 5.5.2), then

$$\lim_{i\to\infty} \frac{\left|\theta^{i+1}(k) - \theta_\star(k)\right|}{|\theta^i(k) - \theta_\star(k)|} = \rho_k < 1 \; ; \tag{5.17}$$

here $\theta(k)$ denotes the $k$th coordinate of the parameter vector. For large-dimensional problems it frequently occurs that, at least for some components $k$, the factor $\rho_k$ is close to one, resulting in very slow convergence of the algorithm. It should be stressed however that the same is true for the EM algorithm, which also exhibits speed of convergence that is linear, and often very poor (Dempster *et al.*, 1977; Jamshidian and Jennrich, 1997; Meng, 1994; Lange, 1995; Meng and Van Dyk, 1997). For gradient-based methods however, there exists a whole range of approaches, based on the second-order properties of the objective function, to guarantee faster convergence.

### Newton and Second-order Methods

The prototype of second-order methods is the Newton, or Newton-Raphson, algorithm:

$$\theta^{i+1} = \theta^i - H^{-1}(\theta^i)\nabla_\theta \ell(\theta^i) \; , \tag{5.18}$$

where $H(\theta^i) = \nabla^2_\theta \ell(\theta^i)$ is the Hessian of the objective function. The Newton iteration is based on the second-order approximation

$$\ell(\theta) \approx \ell(\theta') + \nabla\ell(\theta')\,(\theta - \theta') + \frac{1}{2}\,(\theta - \theta')^t\, H(\theta')\,(\theta - \theta') \; .$$

If the sequence $\{\theta^i\}_{i\geq 0}$ produced by the algorithm converges to a point $\theta_\star$ at which the Hessian is negative definite, the convergence is, at least, quadratic in the sense that for sufficiently large $i$ there exists a positive constant $\beta$ such that $\|\theta^{i+1} - \theta_\star\| \leq \beta\|\theta^i - \theta_\star\|^2$. Therefore the procedure can be very efficient.

The practical use of the Newton algorithm is however hindered by two serious difficulties. The first is analogous to the problem already encountered for the steepest ascent method: there is no guarantee that the algorithm meets the minimal requirement to provide a final parameter estimate that is at least as good as the starting point $\theta^0$. To overcome this difficulty, one may proceed as for the steepest ascent method and introduce a multiplier $\gamma_i$ controlling the step-length in the search direction, so that the method takes the form

$$\theta^{i+1} = \theta^i - \gamma_i H^{-1}(\theta^i)\nabla_\theta \ell(\theta^i) \; . \tag{5.19}$$

Again, $\gamma_i$ may be set to maximize $\ell(\theta^{i+1})$. In practice, it is most often impossible to obtain the exact maximum point called for by the ideal line-search, and one uses approximate directional maximization procedures. Generally speaking, a *line-search algorithm* is an algorithm to find a reasonable multiplier $\gamma_i$ in a step of the form (5.19). A frequently used algorithm consists in determining the (approximate) maximum based on a polynomial interpolation of $\ell(\theta)$ along the line-segment between the current point $\theta^i$ and the proposed update given by (5.18).

A more serious problem is that except in the particular case where the function $\ell(\theta)$ is strictly concave, the direct implementation of (5.18) is prone to numerical instabilities: there may well be whole regions of the parameter space where the

Hessian $H(\theta)$ is either non-invertible (or at least very badly conditioned) or not negative semi-definite (in which case $-H^{-1}(\theta^i)\nabla_\theta \ell(\theta^i)$ is not necessarily an ascent direction). To combat this difficulty, Quasi-Newton methods[1] use the modified recursion

$$\theta^{i+1} = \theta^i + \gamma_i W^i \nabla \ell(\theta^i) \; ; \tag{5.20}$$

here $W^i$ is a weight matrix that may be tuned at each iteration, just like the multiplier $\gamma_i$. The rationale is that if $W^i$ becomes close to $-H^{-1}(\theta^i)$ when convergence occurs, the modified algorithm will share the favorable convergence properties of the Newton algorithm. On the other hand, by using a weight matrix $W^i$ different from $-H^{-1}(\theta^i)$, numerical issues associated with the matrix inversion may be avoided. We again refer to Luenberger (1984) and Fletcher (1987) for a more precise discussion of the available approaches and simply mention here the fact that usually the methods only take profit of gradient information to construct $W^i$, for instance using finite difference calculations, without requiring the direct evaluation of the Hessian $H(\theta)$.

In some contexts, it may be possible to build explicit strategies that are not as good as the Newton algorithm—failing in particular to reach quadratic convergence rates—but yet significantly faster at converging than the basic steepest ascent approach. For incomplete data models, Lange (1995) suggested to use in (5.20) a weight matrix $I_c^{-1}(\theta^i)$ given by

$$I_c(\theta') = -\int \nabla_\theta^2 \log f(x\,;\theta)\big|_{\theta=\theta'} \, p(x\,;\theta') \, \lambda(dx) \; . \tag{5.21}$$

This is the first term on the right-hand side of (5.13). In many models of interest, this matrix is positive definite for all $\theta' \in \Theta$, and thus its inversion is not subject to numerical instabilities. Based on (5.13), it is also to be expected that in some circumstances, $I_c(\theta')$ is a reasonable approximation to the Hessian $\nabla_\theta^2 \ell(\theta')$ and hence that the weighted gradient algorithm converges faster than the steepest ascent or EM algorithms (see Lange, 1995, for further results and examples). In a statistical context, where $f(x\,;\theta)$ is the joint density of two random variables $X$ and $Y$, $I_c(\theta')$ is the conditional expectation given $Y$ of the observed information matrix of associated with this pair.

## 5.2 Application to HMMs

We now return to our primary focus and discuss the application of the previous methods to the specific case of hidden Markov models.

### 5.2.1 Hidden Markov Models as Missing Data Models

HMMs correspond to a sub-category of incomplete data models known as missing data models. In missing data models, the observed data $Y$ is a subset of some not fully observable *complete data* $(X, Y)$. We here assume that the joint distribution of $X$ and $Y$, for a given parameter value $\theta$, admits a joint probability density function $f(x, y\,;\theta)$ with respect to the product measure $\lambda \otimes \mu$. As mentioned in Section 5.1.1, the function $f$ is sometimes referred to as the *complete data likelihood*. It is important to understand that $f$ is a probability density function only when considered as a function of both $x$ and $y$. For a fixed value of $y$ and considered as a function of $x$ only, $f$ is a positive integrable function. Indeed, the actual *likelihood*

[1] *Conjugate gradient* methods are another alternative approach that we do not discuss here.

of the observation, which is defined as the probability density function of $Y$ with respect to $\mu$, is obtained by marginalization as

$$\mathrm{L}(y\,;\theta) = \int f(x,y\,;\theta)\,\lambda(dx)\;. \tag{5.22}$$

For a given value of $y$ this is of course a particular case of (5.1), which served as the basis for developing the EM framework in Section 5.1.2. In missing data models, the family of probability density functions $\{p(\cdot\,;\theta)\}_{\theta\in\Theta}$ defined in (5.3) may thus be interpreted as

$$p(x|y\,;\theta) = \frac{f(x,y\,;\theta)}{\int f(x,y\,;\theta)\,\lambda(dx)}\;, \tag{5.23}$$

the conditional probability density function of $X$ given $Y$.

In the last paragraph, slightly modified versions of the notations introduced in (5.1) and (5.3) were used to reflect the fact that the quantities of interest now depend on the observed variable $Y$. This is obviously mostly a change regarding terminology, with no impact on the contents of Section 5.1.2, except that we may now think of integrating with respect to $p(\cdot\,;\theta)\,d\lambda$ as taking the conditional expectation with respect to the *missing data* $X$, given the observed data $Y$, in the model indexed by the parameter value $\theta$.

### 5.2.2 EM in HMMs

We now consider more specifically hidden Markov models using the notations introduced in Section 1.2, assuming that observations $Y_0$ to $Y_n$ (or, in short, $Y_{0:n}$) are available. Because we only consider HMMs that are fully dominated in the sense of Definition 13, we will use the notations $\nu$ and $\phi_{k|n}$ to refer to the probability density functions of these distributions (of $X_0$ and of $X_k$ given $Y_{0:n}$) with respect to the dominating measure $\lambda$. The joint probability density function of the hidden states $X_{0:n}$ and associated observations $Y_{0:n}$, with respect to the product measure $\lambda^{\otimes(n+1)} \otimes \mu^{\otimes(n+1)}$, is given by

$$\begin{aligned} f_n(x_{0:n}, y_{0:n}\,;\theta) = \nu(x_0\,;\theta) g(x_0, y_0\,;\theta) q(x_0, x_1\,;\theta) g(x_1, y_1\,;\theta) \\ \cdots q(x_{n-1}, x_n\,;\theta) g(x_n, y_n\,;\theta)\;, \end{aligned} \tag{5.24}$$

where we used the same convention as above to indicate dependence with respect to the parameter $\theta$.

Because we mainly consider estimation of the HMM parameter vector $\theta$ from a single sequence of observations, it does not make much sense to consider $\nu$ as an independent parameter. There is no hope to estimate $\nu$ consistently, as there is only one random variable $X_0$ (that is not even observed!) drawn from this density. In the following, we shall thus consider that $\nu$ is either fixed (and known) or fully determined by the parameter $\theta$ that appears in $q$ and $g$. A typical example of the latter consists in assuming that $\nu$ is the stationary distribution associated with the transition function $q(\cdot,\cdot\,;\theta)$ (if it exists). This option is generally practicable only in very simple models (see Example **??** below for an example) because of the lack of analytical expressions relating the stationary distribution of $q(\cdot,\cdot\,;\theta)$ to $\theta$ for general parameterized hidden chains. Irrespective of whether $\nu$ is fixed or determined by $\theta$, it is convenient to omit dependence with respect to $\nu$ in our notations, writing, for instance, $\mathrm{E}_\theta$ for expectations under the model parameterized by $(\theta,\nu)$.

The likelihood of the observations $\mathrm{L}_n(y_{0:n}\,;\theta)$ is obtained by integrating (5.24) with respect to the $x$ (state) variables under the measure $\lambda^{\otimes(n+1)}$. Note that here we use yet another slight modification of the notations adopted in Section 5.1 to acknowledge that both the observations and the hidden states are indeed sequences

with indices ranging from 0 to $n$ (hence the subscript $n$). Upon taking the logarithm in (5.24),

$$\log f_n(x_{0:n}, y_{0:n}\,;\theta) = \log \nu(x_0\,;\theta) + \sum_{k=0}^{n-1} \log q(x_k, x_{k+1}\,;\theta) + \sum_{k=0}^{n} \log g(x_k, y_k\,;\theta)\;,$$

and hence the intermediate quantity of EM has the additive structure

$$\mathcal{Q}(\theta\,;\theta') = \mathrm{E}_{\theta'}[\log \nu(X_0\,;\theta)\,|\,Y_{0:n}] + \sum_{k=0}^{n-1} \mathrm{E}_{\theta'}[\log q(X_k, X_{k+1}\,;\theta)\,|\,Y_{0:n}] + \sum_{k=0}^{n} \mathrm{E}_{\theta'}[\log g(X_k, Y_k\,;\theta)\,|\,Y_{0:n}]\;.$$

In the following, we will adopt the "implicit conditioning" convention that we have used extensively from Section 2.1.4 and onwards, writing $g_k(x\,;\theta)$ instead of $g(x, Y_k\,;\theta)$. With this notation, the intermediate quantity of EM may be rewritten as

$$\mathcal{Q}(\theta\,;\theta') = \mathrm{E}_{\theta'}[\log \nu(X_0\,;\theta)\,|\,Y_{0:n}] + \sum_{k=0}^{n} \mathrm{E}_{\theta'}[\log g_k(X_k\,;\theta)\,|\,Y_{0:n}] + \sum_{k=0}^{n-1} \mathrm{E}_{\theta'}[\log q(X_k, X_{k+1}\,;\theta)\,|\,Y_{0:n}]\;. \quad (5.25)$$

Equation (5.25) shows that in great generality, evaluating the intermediate quantity of EM only requires the computation of expectations under the marginal $\phi_{k|n}(\cdot\,;\theta')$ and bivariate $\phi_{k:k+1|n}(\cdot\,;\theta')$ smoothing distributions, given the parameter vector $\theta'$. The required expectations may thus be computed using either any of the variants of the forward-backward approach presented in Chapter 2 or the recursive smoothing approach discussed in Section **??**. To make the connection with the recursive smoothing approach of Section **??**, we simply rewrite (5.25) as $\mathrm{E}_{\theta'}[t_n(X_{0:n}\,;\theta)\,|\,Y_{0:n}]$, where

$$t_0(x_0\,;\theta) = \log \nu(x_0\,;\theta) + \log g_0(x_0\,;\theta) \quad (5.26)$$

and

$$t_{k+1}(x_{0:k+1}\,;\theta) = t_k(x_{0:k}\,;\theta) + \{\log q(x_k, x_{k+1}\,;\theta) + \log g_{k+1}(x_{k+1}\,;\theta)\}\,. \quad (5.27)$$

Proposition **??** may then be applied directly to obtained the smoothed expectation of the sum functional $t_n$.

Although the exact form taken by the M-step will obviously depend on the way $g$ and $q$ depend on $\theta$, the EM update equations follow a very systematic scheme that does not change much with the exact model under consideration. For instance, all discrete state space models for which the transition matrix $q$ is parameterized by its $r \times r$ elements and such that $g$ and $q$ do not share common parameters (or parameter constraints) give rise to the same update equations for $q$, given in (5.34) below. Several examples of the EM update equations will be reviewed in Sections 5.3 and 5.4.

### 5.2.3 Computing Derivatives

Recall that the Fisher identity—(5.12)—provides an expression for the gradient of the log-likelihood $\ell_n(\theta)$ with respect to the parameter vector $\theta$, closely related to the intermediate quantity of EM. In the HMM context, (5.12) reduces to

$$\nabla_\theta \ell_n(\theta) = \mathrm{E}_\theta[\nabla_\theta \log \nu(X_0\,;\theta) \,|\, Y_{0:n}] + \sum_{k=0}^{n} \mathrm{E}_\theta[\nabla_\theta \log g_k(X_k\,;\theta) \,|\, Y_{0:n}] + \sum_{k=0}^{n-1} \mathrm{E}_\theta[\nabla_\theta \log q(X_k, X_{k+1}\,;\theta) \,|\, Y_{0:n}] \,. \quad (5.28)$$

Hence the gradient of the log-likelihood may also be evaluated using either the forward-backward approach or the recursive technique discussed in Chapter 3. For the latter, we only need to redefine the functional of interest, replacing (5.26) and (5.27) by their gradients with respect to $\theta$.

Louis' identity (5.14) gives rise to more complicated expressions, and we only consider here the case where $g$ does depend on $\theta$, whereas the state transition density $q$ and the initial distribution $\nu$ are assumed to be fixed and known (the opposite situation is covered in detail in a particular case in Section 5.3.3). In this case, (5.14) may be rewritten as

$$\begin{aligned} &\nabla^2_\theta \ell_n(\theta) + \{\nabla_\theta \ell_n(\theta)\}\,\{\nabla_\theta \ell_n(\theta)\}^t \qquad (5.29)\\ &\quad = \sum_{k=0}^{n} \mathrm{E}_\theta[\,\nabla^2_\theta \log g_k(X_k\,;\theta) \,\big|\, Y_{0:n}] \\ &\qquad + \sum_{k=0}^{n}\sum_{j=0}^{n} \mathrm{E}_\theta\left[\,\{\nabla_\theta \log g_k(X_k\,;\theta)\}\,\{\nabla_\theta \log g_j(X_j\,;\theta)\}^t \,\middle|\, Y_{0:n}\right] \,. \end{aligned}$$

The first term on the right-hand side of (5.29) is obviously an expression that can be computed proceeding as for (5.28), replacing first- by second-order derivatives. The second term is however more tricky because it (seemingly) requires the evaluation of the joint distribution of $X_k$ and $X_j$ given the observations $Y_{0:n}$ for all pairs of indices $k$ and $j$, which is not obtainable by the smoothing approaches based on some form of the forward-backward decomposition. The rightmost term of (5.29) is however easily recognized as a squared sum functional similar to (**??**), which can thus be evaluated recursively (in $n$) proceeding as in Example **??**. Recall that the trick consists in observing that if

$$\begin{aligned} \tau_{n,1}(x_{0:n}\,;\theta) &\stackrel{\mathrm{def}}{=} \sum_{k=0}^{n} \nabla_\theta \log g_k(x_k\,;\theta) \,, \\ \tau_{n,2}(x_{0:n}\,;\theta) &\stackrel{\mathrm{def}}{=} \left\{\sum_{k=0}^{n} \nabla_\theta \log g_k(x_k\,;\theta)\right\}\left\{\sum_{k=0}^{n} \nabla_\theta \log g_k(x_k\,;\theta)\right\}^t \,, \end{aligned}$$

then

$$\begin{aligned} \tau_{n,2}(x_{0:n}\,;\theta) = \tau_{n-1,2}(x_{0:n-1}\,;\theta) &+ \{\nabla_\theta \log g_n(x_n\,;\theta)\}\,\{\nabla_\theta \log g_n(x_n\,;\theta)\}^t \\ &+ \tau_{n-1,1}(x_{0:n-1}\,;\theta)\,\{\nabla_\theta \log g_n(x_n\,;\theta)\}^t \\ &+ \nabla_\theta \log g_n(x_n\,;\theta)\,\{\tau_{n-1,1}(x_{0:n-1}\,;\theta)\}^t \,. \end{aligned}$$

This last expression is of the general form given in Definition **??**, and hence Proposition **??** may be applied to update recursively in $n$

$$\mathrm{E}_\theta[\tau_{n,1}(X_{0:n}\,;\theta) \,|\, Y_{0:n}] \qquad \text{and} \qquad \mathrm{E}_\theta[\tau_{n,2}(X_{0:n}\,;\theta) \,|\, Y_{0:n}] \,.$$

To make this approach more concrete, we will describe below, in Section 5.3.3, its application to a very simple finite state space HMM.

## 5.3 The Example of Normal Hidden Markov Models

In order to make the general principles outlined in the previous section more concrete, we now work out the details on selected examples of HMMs. We begin with the case where the state space is finite and the observation transition function $g$ corresponds to a (univariate) Gaussian distribution. Only the most standard case where the parameter vector is split into two sub-components that parameterize, respectively, $g$ and $q$, is considered.

### 5.3.1 EM Parameter Update Formulas

In the widely used normal HMM, $\mathsf{X}$ is a finite set, identified with $\{1, \ldots, r\}$, $\mathsf{Y} = \mathbb{R}$, and $g$ is a Gaussian probability density function (with respect to Lebesgue measure) given by

$$g(x, y\,;\theta) = \frac{1}{\sqrt{2\pi\upsilon_x}} \exp\left\{-\frac{(y-\mu_x)^2}{2\upsilon_x}\right\} .$$

By definition, $g_k(x\,;\theta)$ is equal to $g(x, Y_k\,;\theta)$. We first assume that the initial distribution $\nu$ is known and fixed, before examining the opposite case briefly in Section 5.3.2 below. The parameter vector $\theta$ thus encompasses the transition probabilities $q_{ij}$ for $i, j = 1, \ldots, r$ as well as the means $\mu_i$ and variances $\upsilon_i$ for $i = 1, \ldots, r$. Note that in this section, because we will often need to differentiate with respect to $\upsilon_i$, it is simpler to use the variances $\upsilon_i = \sigma_i^2$ rather than the standard deviations $\sigma_i$ as parameters. The means and variances are unconstrained, except for the positivity of the latter, but the transition probabilities are subject to the equality constraints $\sum_{j=1}^r q_{ij} = 1$ for $i = 1, \ldots, r$ (in addition to the obvious constraint that $q_{ij}$ should be non-negative). When considering the parameter vector denoted by $\theta'$, we will denote by $\mu_i'$, $\upsilon_i'$, and $q_{ij}'$ its various elements.

For the model under consideration, (5.25) may be rewritten as

$$\begin{aligned}\mathcal{Q}(\theta\,;\theta') = C^{st} - \frac{1}{2}\sum_{k=0}^{n} \mathrm{E}_{\theta'}\left[\sum_{i=1}^{r} \mathbb{1}\{X_k = i\}\left(\log\upsilon_i + \frac{(Y_k-\mu_i)^2}{\upsilon_i}\right)\,\middle|\, Y_{0:n}\right] \\ + \sum_{k=1}^{n} \mathrm{E}_{\theta'}\left[\sum_{i=1}^{r}\sum_{j=1}^{r} \mathbb{1}\{(X_{k-1}, X_k) = (i,j)\}\log q_{ij}\,\middle|\, Y_{0:n}\right] ,\end{aligned}$$

where the leading term does not depend on $\theta$. Using the notations introduced in Section 2.1 for the smoothing distributions, we may write

$$\begin{aligned}\mathcal{Q}(\theta\,;\theta') = C^{st} - \frac{1}{2}\sum_{k=0}^{n}\sum_{i=1}^{r} \phi_{k|n}(i\,;\theta')\left[\log\upsilon_i + \frac{(Y_k-\mu_i)^2}{\upsilon_i}\right] \\ + \sum_{k=1}^{n}\sum_{i=1}^{r}\sum_{j=1}^{r} \phi_{k-1:k|n}(i,j\,;\theta')\log q_{ij} . \quad (5.30)\end{aligned}$$

Now, given the initial distribution $\nu$ and parameter $\theta'$, the smoothing distributions appearing in (5.30) can be evaluated by any of the variants of forward-backward smoothing discussed in Chapter 2. As already explained above, the E-step of EM thus reduces to solving the smoothing problem. The M-step is specific

and depends on the model parameterization: the task consists in finding a global optimum of $\mathcal{Q}(\theta\,;\theta')$ that satisfies the constraints mentioned above. For this, simply introduce the Lagrange multipliers $\lambda_1,\ldots,\lambda_r$ that correspond to the equality constraints $\sum_{j=1}^r q_{ij} = 1$ for $i = 1,\ldots,r$ (Luenberger, 1984, Chapter 10). The first-order partial derivatives of the Lagrangian

$$\mathfrak{L}(\theta,\lambda\,;\theta') = \mathcal{Q}(\theta\,;\theta') + \sum_{i=1}^r \lambda_i \left(1 - \sum_{j=1}^r q_{ij}\right)$$

are given by

$$\begin{aligned}
\frac{\partial}{\partial\mu_i}\mathfrak{L}(\theta,\lambda\,;\theta') &= \frac{1}{\upsilon_i}\sum_{k=0}^n \phi_{k|n}(i\,;\theta')(Y_k - \mu_i)\;,\\
\frac{\partial}{\partial\upsilon_i}\mathfrak{L}(\theta,\lambda\,;\theta') &= -\frac{1}{2}\sum_{k=0}^n \phi_{k|n}(i\,;\theta')\left[\frac{1}{\upsilon_i} - \frac{(Y_k-\mu_i)^2}{\upsilon_i^2}\right]\;,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial q_{ij}}\mathfrak{L}(\theta,\lambda\,;\theta') &= \sum_{k=1}^n \frac{\phi_{k-1:k|n}(i,j\,;\theta')}{q_{ij}} - \lambda_i\;,\\
\frac{\partial}{\partial\lambda_i}\mathfrak{L}(\theta,\lambda\,;\theta') &= 1 - \sum_{j=1}^r q_{ij}\;.
\end{aligned} \tag{5.31}$$

Equating all expressions in (5.31) to zero yields the parameter vector

$$\theta^* = \left[(\mu_i^*)_{i=1,\ldots,r}, (\upsilon_i^*)_{i=1,\ldots,r}, (q_{ij}^*)_{i,j=1,\ldots,r}\right]$$

which achieves the maximum of $\mathcal{Q}(\theta\,;\theta')$ under the applicable parameter constraints:

$$\mu_i^* = \frac{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')Y_k}{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')}\;, \tag{5.32}$$

$$\upsilon_i^* = \frac{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')(Y_k-\mu_i^*)^2}{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')}\;, \tag{5.33}$$

$$q_{ij}^* = \frac{\sum_{k=1}^n \phi_{k-1:k|n}(i,j\,;\theta')}{\sum_{k=1}^n\sum_{l=1}^r \phi_{k-1:k|n}(i,l\,;\theta')} \tag{5.34}$$

for $i,j = 1,\ldots,r$, where the last equation may be rewritten more concisely as

$$q_{ij}^* = \frac{\sum_{k=1}^n \phi_{k-1:k|n}(i,j\,;\theta')}{\sum_{k=1}^n \phi_{k-1|n}(i\,;\theta')}\;. \tag{5.35}$$

Equations (5.32)–(5.34) are emblematic of the intuitive form taken by the parameter update formulas derived though the EM strategy. These equations are simply the maximum likelihood equations for the *complete model* in which both $\{X_k\}_{0\le k\le n}$ and $\{Y_k\}_{0\le k\le n}$ would be observed, except that the functions $\mathbb{1}\{X_k = i\}$ and $\mathbb{1}\{X_{k-1} = i, X_k = j\}$ are replaced by their conditional expectations, $\phi_{k|n}(i\,;\theta')$ and $\phi_{k-1:k|n}(i,j\,;\theta')$, given the actual observations $Y_{0:n}$ and the available parameter estimate $\theta'$. As discussed in Section 5.1.2, this behavior is fundamentally due to the fact that the probability density functions associated with the complete model form an exponential family. As a consequence, the same remark holds more generally for all discrete HMMs for which the conditional probability density functions $g(i,\cdot\,;\theta)$ belong to an exponential family. A final word of warning about the way in which (5.33) is written: in order to obtain a concise and intuitively interpretable

expression, (5.33) features the value of $\mu_i^*$ as given by (5.32). It is of course possible to rewrite (5.33) in a way that only contains the current parameter value $\theta'$ and the observations $Y_{0:n}$ by combining (5.32) and (5.33) to obtain

$$\upsilon_i^* = \frac{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')Y_k^2}{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')} - \left[\frac{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')Y_k}{\sum_{k=0}^n \phi_{k|n}(i\,;\theta')}\right]^2 . \tag{5.36}$$

### 5.3.2 Estimation of the Initial Distribution

As mentioned above, in this chapter we generally assume that the initial distribution $\nu$, that is, the distribution of $X_0$, is fixed and known. There are cases when one wants to treat this as an unknown parameter however, and we briefly discuss below this issue in connection with the EM algorithm for the normal HMM. We shall assume that $\nu = (\nu_i)_{1\le i\le r}$ is an unknown probability vector (that is, with non-negative entries summing to unity), which we accommodate within the parameter vector $\theta$. The complete log-likelihood will then be as above, where the initial term

$$\log \nu_{X_0} = \sum_{i=1}^r \mathbb{1}\{X_0 = i\} \log \nu_i$$

goes into $\mathcal{Q}(\theta\,;\theta')$ as well, giving the additive contribution

$$\sum_{i=1}^r \phi_{0|n}(i\,;\theta') \log \nu_i$$

to (5.30). This sum is indeed part of (5.30) already, but hidden within $C^{st}$ when $\nu$ is not a parameter to be estimated. Using Lagrange multipliers as above, it is straightforward to show that the M-step update of $\nu$ is $\nu_i^* = \phi_{0|n}(i\,;\theta')$.

It was also mentioned above that sometimes it is desirable to link $\nu$ to $q_\theta$ as being the stationary distribution of $q_\theta$. Then there is an additive contribution to $\mathcal{Q}(\theta\,;\theta')$ as above, with the difference that $\nu$ can now not be chosen freely but is a function of $q_\theta$. As there is no simple formula for the stationary distribution of $q_\theta$, the M-step is no longer explicit. However, once the sums (over $k$) in (5.30) have been computed for all $i$ and $j$, we are left with an optimization problem over the $q_{ij}$ for which we have an excellent initial guess, namely the standard update (ignoring $\nu$) (5.34). A few steps of a standard numerical optimization routine (optimizing over the $q_{ij}$) is then often enough to find the maximum of $\mathcal{Q}(\cdot\,;\theta')$ under the stationarity assumption. Variants of the basic EM strategy, to be discussed in Section 5.5.3, may also be useful in this situation.

### 5.3.3 Computation of the Score and Observed Information

For reasons discussed above, computing the gradient of the log-likelihood is not a difficult task in finite state space HMMs and should preferably be done using smoothing algorithms based on the forward-backward decomposition. The only new requirement is to evaluate the derivatives with respect to $\theta$ that appear in (5.28). In the case of the normal HMM, we already met the appropriate expressions in (5.31), as Fisher's identity (5.12) implies that the gradient of the intermediate quantity at the current parameter estimate coincides with the gradient of the log-likelihood.

Hence

$$\frac{\partial}{\partial \mu_i}\ell_n(\theta) = \frac{1}{\upsilon_i}\sum_{k=0}^{n}\phi_{k|n}(i\,;\theta)(Y_k-\mu_i)\;,$$
$$\frac{\partial}{\partial \upsilon_i}\ell_n(\theta) = -\frac{1}{2}\sum_{k=0}^{n}\phi_{k|n}(i\,;\theta)\left[\frac{1}{\upsilon_i}-\frac{(Y_k-\mu_i)^2}{\upsilon_i^2}\right]\;,$$
$$\frac{\partial}{\partial q_{ij}}\ell_n(\theta) = \sum_{k=1}^{n}\frac{\phi_{k-1:k|n}(i,j\,;\theta)}{q_{ij}}\;.$$

We now focus on the computation of the derivatives of the log-likelihood in the model of Example **??** with respect to the transition parameters $\rho_0$ and $\rho_1$. As they play a symmetric role, it is sufficient to consider, say, $\rho_0$ only. The variance $\upsilon$ is considered as fixed so that the only quantities that depend on the parameter $\rho_0$ are the initial distribution $\nu$ and the transition matrix $Q$. We will, as usual, use the simplified notation $g_k(x)$ rather than $g(x,Y_k)$ to denote the Gaussian density function $(2\pi\upsilon)^{-1/2}\exp\{-(Y_k-x)^2/(2\upsilon)\}$ for $x\in\{0,1\}$. Furthermore, in order to simplify the expressions below, we also omit to indicate explicitly the dependence with respect to $\rho_0$ in the rest of this section. Fisher's identity (5.12) reduces to

$$\frac{\partial}{\partial \rho_0}\ell_n = \mathrm{E}\left[\frac{\partial}{\partial \rho_0}\log\nu(X_0)+\sum_{k=0}^{n-1}\frac{\partial}{\partial \rho_0}\log q_{X_kX_{k+1}}\,\middle|\, Y_{0:n}\right]\;,$$

where the notation $q_{ij}$ refers to the element in the $(1+i)$-th row and $(1+j)$-th column of the matrix $Q$ (in particular, $q_{00}$ and $q_{11}$ are alternative notations for $\rho_0$ and $\rho_1$). We are thus in the framework of Proposition **??** with a smoothing functional $t_{n,1}$ defined by

$$t_{0,1}(x) = \frac{\partial}{\partial \rho_0}\log\nu(x)\;,$$
$$s_{k,1}(x,x') = \frac{\partial}{\partial \rho_0}\log q_{xx'} \qquad \text{for } k\geq 0\;,$$

where the multiplicative functions $\{m_{k,1}\}_{k\geq 0}$ are equal to 1. Straightforward calculations yield

$$t_{0,1}(x) = (\rho_0+\rho_1)^{-1}\left[\frac{\rho_1}{\rho_0}\delta_0(x)-\delta_1(x)\right],$$
$$s_{k,1}(x,x') = \frac{1}{\rho_0}\delta_{(0,0)}(x,x')-\frac{1}{1-\rho_0}\delta_{(0,1)}(x,x')\;.$$

Hence a first recursion, following Proposition **??**.

**Algorithm 101** (Computation of the Score in Example **??**)**.** [Init:]

**Initialization:** Compute $c_0=\sum_{i=0}^{1}\nu(i)g_0(i)$ and, for $i=0,1$,

$$\phi_k(i) = c_0^{-1}\nu(i)g_0(i)\;,$$
$$\tau_{0,1}(i) = t_{0,1}(i)\phi_0(i)\;.$$

**Recursion:** For $k=0,1,\ldots,$ compute $c_{k+1}=\sum_{i=0}^{1}\sum_{j=0}^{1}\phi_k(i)q_{ij}g_k(j)$ and, for

$j = 0, 1$,

$$\phi_{k+1}(j) = c_{k+1}^{-1} \sum_{i=0}^{1} \phi_k(i) q_{ij} g_k(j) \ ,$$

$$\tau_{k+1,1}(j) = c_{k+1}^{-1} \Bigg\{ \sum_{i=0}^{1} \tau_{k,1}(i) q_{ij} g_{k+1}(j) + \phi_k(0) g_{k+1}(0) \delta_0(j) - \phi_k(0) g_{k+1}(1) \delta_1(j) \Bigg\} \ .$$

At each index $k$, the log-likelihood is available via $\ell_k = \sum_{l=0}^{k} \log c_l$, and its derivative with respect to $\rho_0$ may be evaluated as

$$\frac{\partial}{\partial \rho_0} \ell_k = \sum_{i=0}^{1} \tau_{k,1}(i) \ .$$

For the second derivative, Louis' identity (5.14) shows that

$$\frac{\partial^2}{\partial \rho_0^2} \ell_n + \left\{ \frac{\partial}{\partial \rho_0} \ell_n \right\}^2 = \mathrm{E} \left[ \frac{\partial^2}{\partial \rho_0^2} \log \nu(X_0) + \sum_{k=0}^{n-1} \frac{\partial^2}{\partial \rho_0^2} \log q_{X_k X_{k+1}} \,\middle|\, Y_{0:n} \right] + \mathrm{E} \left[ \left( \frac{\partial}{\partial \rho_0} \log \nu(X_0) + \sum_{k=0}^{n-1} \frac{\partial}{\partial \rho_0} \log q_{X_k X_{k+1}} \right)^2 \,\middle|\, Y_{0:n} \right] . \quad (5.37)$$

The first term on the right-hand side of (5.37) is very similar to the case of $\tau_{n,1}$ considered above, except that we now need to differentiate the functions twice, replacing $t_{0,1}$ and $s_{k,1}$ by $\frac{\partial}{\partial \rho_0} t_{0,1}$ and $\frac{\partial}{\partial \rho_0} s_{k,1}$, respectively. The corresponding smoothing functional $t_{n,2}$ is thus now defined by

$$t_{0,2}(x) = -\frac{\rho_1(2\rho_0 + \rho_1)}{\rho_0^2(\rho_0 + \rho_1)^2} \delta_0(x) + \frac{1}{(\rho_0 + \rho_1)^2} \delta_1(x) \ ,$$
$$s_{k,2}(x, x') = -\frac{1}{\rho_0^2} \delta_{(0,0)}(x, x') - \frac{1}{(1 - \rho_0)^2} \delta_{(0,1)}(x, x') \ .$$

The second term on the right-hand side of (5.37) is more difficult, and we need to proceed as in Example **??**: the quantity of interest may be rewritten as the conditional expectation of

$$t_{n,3}(x_{0:n}) = \left[ t_{0,1}(x_0) + \sum_{k=0}^{n-1} s_{k,1}(x_k, x_{k+1}) \right]^2 \ .$$

Expanding the square in this equation yields the update formula

$$t_{k+1,3}(x_{0:k+1}) = t_{k,3}(x_{0:k}) + s_{k,1}^2(x_k, x_{k+1}) + 2 t_{k,1}(x_{0:k}) s_{k,1}(x_k, x_{k+1}) \ .$$

Hence $t_{k,1}$ and $t_{k,3}$ jointly are of the form prescribed by Definition **??** with incremental additive functions $s_{k,3}(x, x') = s_{k,1}^2(x, x')$ and multiplicative updates $m_{k,3}(x, x') = 2 s_{k,1}(x, x')$. As a consequence, the following smoothing recursion holds.

**Algorithm 102** (Computation of the Observed Information in Example **??**)**.** [Init:]

**Initialization:** For $i = 0, 1$,

$$\tau_{0,2}(i) = t_{0,2}(i)\phi_0(i) \; .$$
$$\tau_{0,3}(i) = t_{0,1}^2(i)\phi_0(i) \; .$$

**Recursion:** For $k = 0, 1, \ldots$, compute for $j = 0, 1$,

$$\tau_{k+1,2}(j) = c_{k+1}^{-1}\left\{\sum_{i=0}^{1}\tau_{k,2}(i)q_{ij}g_{k+1}(j) - \frac{1}{\rho_0}\phi_k(0)g_{k+1}(0)\delta_0(j) - \frac{1}{(1-\rho_0)}\phi_k(0)g_{k+1}(1)\delta_1(j)\right\},$$

$$\tau_{0,3}(j) = c_{k+1}^{-1}\left\{\sum_{i=0}^{1}\tau_{k,3}(i)q_{ij}g_{k+1}(j) + 2\left[\tau_{k,1}(0)g_{k+1}(0)\delta_0(j) - \tau_{k,1}(0)g_{k+1}(1)\delta_1(j)\right] + \frac{1}{\rho_0}\phi_k(0)g_{k+1}(0)\delta_0(j) + \frac{1}{(1-\rho_0)}\phi_k(0)g_{k+1}(1)\delta_1(j)\right\}.$$

At each index $k$, the second derivative of the log-likelihood satisfies

$$\frac{\partial^2}{\partial\rho_0^2}\ell_k + \left(\frac{\partial}{\partial\rho_0}\ell_k\right)^2 = \sum_{i=0}^{1}\tau_{k,2}(i) + \sum_{i=0}^{1}\tau_{k,3}(i) \; ,$$

where the second term on the left-hand side may be evaluated in the same recursion, following Algorithm 101.

To illustrate the results obtained with Algorithms 101–102, we consider the model with parameters $\rho_0 = 0.95$, $\rho_1 = 0.8$, and $\upsilon = 0.1$ (using the notations introduced in Example **??**). Figure 5.1 displays the typical aspect of two sequences of length 200 simulated under slightly different values of $\rho_0$. One possible use of the output of Algorithms 101–102 consists in testing for changes in the parameter values. Indeed, under conditions to be detailed in Chapter 6 (and which hold here), the normalized score $n^{-1/2}\frac{\partial}{\partial\rho_0}\ell_n$ satisfies a central limit theorem with variance given by the limit of the normalized information $-n^{-1}(\partial^2/\partial\rho_0^2)\ell_n$. Hence it is expected that

$$\mathfrak{R}_n = \frac{\frac{\partial}{\partial\rho_0}\ell_n}{\sqrt{-\frac{\partial^2}{\partial\rho_0^2}\ell_n}}$$

be asymptotically $\mathrm{N}(0,1)$-distributed under the null hypothesis that $\rho_0$ is indeed equal to the value used for computing the score and information recursively with Algorithms 101–102.

Figure 5.2 displays the empirical quantiles of $\mathfrak{R}_n$ against normal quantiles for $n = 200$ and $n =$1,000. For the longer sequences ($n =$1,000), the result is clearly as expected with a very close fit to the normal quantiles. When $n = 200$, asymptotic normality is not yet reached and there is a significant bias toward high values of $\mathfrak{R}_n$. Looking back at Figure 5.1, even if $\upsilon$ was equal to zero—or in other words, if we were able to identify without ambiguity the 0 and 1 states from the data—there would not be much information about $\rho_0$ to be gained from runs of length 200: when $\rho_0 = 0.95$ and $\rho_1 = 0.8$, the average number of distinct runs of 0s that one can observe in 200 consecutive data points is only about $200/(20 + 5) = 8$. To construct a goodness of fit test from $\mathfrak{R}_n$, one can monitor values of $\mathfrak{R}_n^2$, which asymptotically has a chi-square distribution with one degree of freedom. Testing

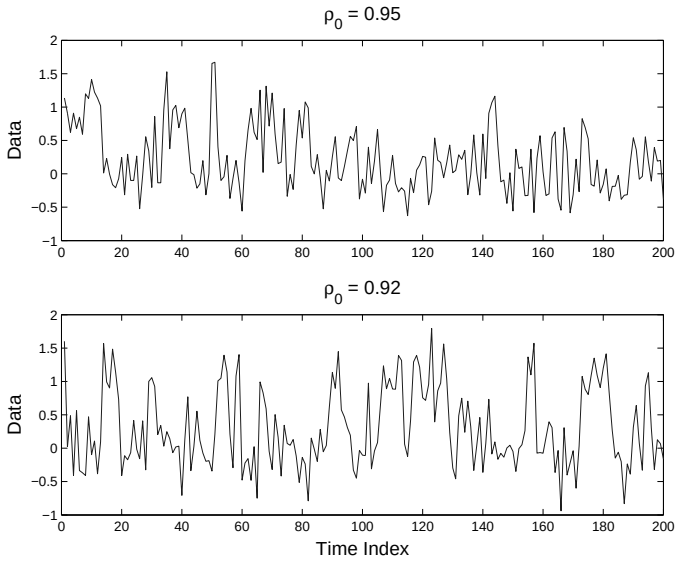



Figure 5.1: Two simulated trajectories of length $n = 200$ from the simplified ion channel model of Example **??** with $\rho_0 = 0.95$, $\rho_1 = 0.8$, and $\sigma^2 = 0.1$ (top), and $\rho_0 = 0.92$, $\rho_1 = 0.8$, and $\sigma^2 = 0.1$ (bottom).

the null hypothesis $\rho_0 = 0.95$ gives $p$-values of 0.87 and 0.09 for the two sequences in the top and bottom plots, respectively, of Figure 5.1. When testing at the 10% level, both sequences thus lead to the correct decision: no rejection and rejection of the null hypothesis, respectively. Interestingly, testing the other way around, that is, postulating $\rho_0 = 0.92$ as the null hypothesis, gives $p$-values of 0.20 and 0.55 for the top and bottom sequences of Figure 5.1, respectively. The outcome of the test is now obviously less clear-cut, which reveals an asymmetry in its discrimination ability: it is easier to detect values of $\rho_0$ that are smaller than expected than the converse. This is because smaller values of $\rho_0$ means more changes (on average) in the state sequence and hence more usable information about $\rho_0$ to be obtained from a fixed size record. This asymmetry is connected to the upward bias visible in the left plot of Figure 5.2.

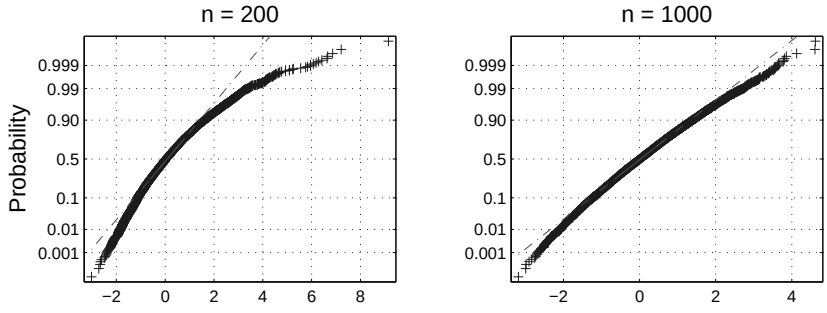



Figure 5.2: QQ-plot of empirical quantiles of the test statistic $\mathfrak{R}_n$ (abscissas) for the simplified ion channel model of Example **??** with $\rho_0 = 0.95$, $\rho_1 = 0.8$, and $\sigma^2 = 0.1$ vs. normal quantiles (ordinates). Samples sizes were $n = 200$ (left) and $n$ =1,000 (right), and 10,000 independent replications were used to estimate the empirical quantiles.

## 5.4 The Example of Gaussian Linear State-Space Models

We now consider more briefly the case of Gaussian linear state-space models that form the other major class of hidden Markov models for which the methods discussed in Section 5.1 are directly applicable. It is worth mentioning that Gaussian linear state-space models are perhaps the only important subclass of the HMM family for which there exist reasonable simple non-iterative parameter estimation algorithms not based on maximum likelihood arguments but are nevertheless useful in practical applications. These sub-optimal algorithms, proposed by Van Overschee and De Moor (1993), rely on the linear structure of the model and use only eigendecompositions of empirical covariance matrices—a general principle usually referred to under the denomination of *subspace methods* (Van Overschee and De Moor, 1996). Keeping in line with the general topic of this chapter, we nonetheless consider below only algorithms for maximum likelihood estimation in Gaussian linear state-space models.

The Gaussian linear state-space model is given by

$$\begin{aligned} X_{k+1} &= AX_k + RU_k \;, \\ Y_k &= BX_k + SV_k \;, \end{aligned}$$

where $X_0$, $\{U_k\}_{k\geq 0}$ and $\{V_k\}_{k\geq 0}$ are jointly Gaussian. The parameters of the model are the four matrices $A$, $R$, $B$, and $S$. Note that except for scalar models, it is not possible to estimate $R$ and $S$ because both $\{U_k\}$ and $\{V_k\}$ are unobservable and hence $R$ and $S$ are only identifiable up to an orthonormal matrix. In other words, multiplying $R$ or $S$ by any orthonormal matrix of suitable dimension does not modify the distribution of the observations. Hence the parameters that are identifiable are the covariance matrices $\Upsilon_R = RR^t$ and $\Upsilon_S = SS^t$, which we consider below. Likewise, the matrices $A$ and $B$ are identifiable up to a similarity transformation only. Indeed, setting $X'_k = TX_k$ for some invertible matrix $T$, that is, making a change of basis for the state process, it is straightforward to check that the joint process $\{(X'_k, Y_k)\}$ satisfies the model assumptions with $TAT^{-1}$, $BT^{-1}$, and $TR$ replacing $A$, $B$, and $R$, respectively. Nevertheless, we work with $A$ and $B$ in the algorithm below. If a unique representation is desired, one may use, for instance, the companion form of $A$ given its eigenvalues; this matrix may contain complex entries though. As in the case of finite state space HMMs (Section 5.2.2), it is not sensible to consider the initial covariance matrix $\Sigma_\nu$ as an independent parameter when using a single observed sequence. On the other hand, for such models it is very natural to assume that $\Sigma_\nu$ is associated with the stationary distribution of $\{X_k\}$. We shall also assume that both $\Upsilon_R$ and $\Upsilon_S$ are full rank covariance matrices so that all Gaussian distributions admit densities with respect to (multi-dimensional) Lebesgue measure.

### 5.4.1 The Intermediate Quantity of EM

With the previous notations, the intermediate quantity $\mathcal{Q}(\theta\,;\theta')$ of EM, defined in (5.25), may be expressed as

$$\begin{aligned} &-\frac{1}{2}\,\mathrm{E}_{\theta'}\left[ n\log|\Upsilon_R| + \sum_{k=0}^{n-1}(X_{k+1}-AX_k)^t\Upsilon_R^{-1}(X_{k+1}-AX_k) \,\middle|\, Y_{0:n}\right] \\ &\quad -\frac{1}{2}\,\mathrm{E}_{\theta'}\left[ (n+1)\log|\Upsilon_S| + \sum_{k=0}^{n}(Y_k-BX_k)^t\Upsilon_S^{-1}(Y_k-BX_k) \,\middle|\, Y_{0:n}\right], \end{aligned} \tag{5.38}$$

up to terms that do not depend on the parameters. In order to elicit the M-step equations or to compute the score, we differentiate (5.38) using elementary perturbation calculus as well as the identity $\nabla_C \log|C| = C^{-t}$ for an invertible matrix $C$—which is a consequence of the adjoint representation of the inverse (Horn and Johnson, 1985, Section 0.8.2):

$$\nabla_A \mathcal{Q}(\theta\,;\theta') = -\Upsilon_R^{-1}\,\mathrm{E}_{\theta'}\left[\sum_{k=0}^{n-1}(AX_kX_k^t - X_{k+1}X_k^t)\,\middle|\,Y_{0:n}\right]\,, \tag{5.39}$$

$$\begin{aligned}\nabla_{\Upsilon_R^{-1}} \mathcal{Q}(\theta\,;\theta') = -\frac{1}{2}\Bigg\{&-n\Upsilon_R \\ &+\mathrm{E}_{\theta'}\left[\sum_{k=0}^{n-1}(X_{k+1}-AX_k)(X_{k+1}-AX_k)^t\,\middle|\,Y_{0:n}\right]\Bigg\}\,,\end{aligned} \tag{5.40}$$

$$\nabla_B \mathcal{Q}(\theta\,;\theta') = -\Upsilon_S^{-1}\,\mathrm{E}_{\theta'}\left[\sum_{k=0}^{n}(BX_kX_k^t - Y_kX_k^t)\,\middle|\,Y_{0:n}\right]\,, \tag{5.41}$$

$$\begin{aligned}\nabla_{\Upsilon_S^{-1}} \mathcal{Q}(\theta\,;\theta') = -\frac{1}{2}\Bigg\{&-(n+1)\Upsilon_S \\ &+\mathrm{E}_{\theta'}\left[\sum_{k=0}^{n}(Y_k-BX_k)(Y_k-BX_k)^t\,\middle|\,Y_{0:n}\right]\Bigg\}\,.\end{aligned} \tag{5.42}$$

Note that in the expressions above, we differentiate with respect to the inverses of $\Upsilon_R$ and $\Upsilon_S$ rather than with respect to the covariance matrices themselves, which is equivalent, because we assume both of the covariance matrices to be positive definite, but yields simpler formulas. Equating all derivatives simultaneously to zero defines the EM update of the parameters. We will denote these updates by $A^*$, $B^*$, $\Upsilon_R^*$, and $\Upsilon_S^*$, respectively. To write them down, denote $\hat{X}_{k|n}(\theta') = \mathrm{E}_{\theta'}[X_k \,|\, Y_{0:n}]$ and $\Sigma_{k|n}(\theta') = \mathrm{E}_{\theta'}[X_kX_k' \,|\, Y_{0:n}] - \hat{X}_{k|n}(\theta')\hat{X}_{k|n}^t(\theta')$, where we now indicate explicitly that these first two smoothing moments indeed depend on the current estimates of the model parameters (they also depend on the initial covariance matrix $\Sigma_\nu$, but we ignore this fact here because this quantity is considered as being fixed). We also need to evaluate the conditional covariances

$$\begin{aligned}C_{k,k+1|n}(\theta') &\stackrel{\mathrm{def}}{=} \mathrm{Cov}_{\theta'}[X_k, X_{k+1} \,|\, Y_{0:n}] \\ &= \mathrm{E}_{\theta'}[X_kX_{k+1}^t \,|\, Y_{0:n}] - \hat{X}_{k|n}(\theta')\hat{X}_{k+1|n}^t(\theta')\,.\end{aligned}$$

With these notations, the EM update equations are given by

$$A^* = \left[\sum_{k=0}^{n-1} C_{k,k+1|n}(\theta') + \hat{X}_{k|n}(\theta')\hat{X}^t_{k+1|n}(\theta')\right]^t \left[\sum_{k=0}^{n-1} \Sigma_{k|n}(\theta') + \hat{X}_{k|n}(\theta')\hat{X}^t_{k|n}(\theta')\right]^{-1}, \tag{5.43}$$

$$\Upsilon^*_R = \frac{1}{n}\sum_{k=0}^{n-1}\left\{\left[\Sigma_{k+1|n}(\theta') + \hat{X}_{k+1|n}(\theta')\hat{X}^t_{k+1|n}(\theta')\right] - A^*\left[C_{k,k+1|n}(\theta') + \hat{X}_{k|n}(\theta')\hat{X}^t_{k+1|n}(\theta')\right]\right\}, \tag{5.44}$$

$$B^* = \left[\sum_{k=0}^{n} \hat{X}_{k|n}(\theta')Y_k^t\right]^t \left[\sum_{k=0}^{n} \Sigma_{k|n}(\theta') + \hat{X}_{k|n}(\theta')\hat{X}^t_{k|n}(\theta')\right]^{-1}, \tag{5.45}$$

$$\Upsilon^*_S = \frac{1}{n+1}\sum_{k=0}^{n}\left[Y_kY_k^t - B^*\hat{X}_{k|n}(\theta')Y_k^t\right] . \tag{5.46}$$

In obtaining the covariance update, we used the same remark that made it possible to rewrite, in the case of normal HMMs, (5.33) as (5.36).

## 5.5 Complements

To conclude this chapter, we briefly return to an issue mentioned in Section 5.1.2 regarding the conditions that ensure that the EM iterations indeed converge to stationary points of the likelihood.

### 5.5.1 Global Convergence of the EM Algorithm

As a consequence of Proposition 98, the EM algorithm described in Section 5.1.2 has the property that the log-likelihood function $\ell$ can never decrease in an iteration. Indeed,

$$\ell(\theta^{i+1}) - \ell(\theta^i) \geq \mathcal{Q}(\theta^{i+1};\theta^i) - \mathcal{Q}(\theta^i;\theta^i) \geq 0 .$$

This class of algorithms, sometimes referred to as *ascent algorithms* (Luenberger, 1984, Chapter 6), can be treated in a unified manner following a theory developed mostly by Zangwill (1969). Wu (1983) showed that this general theory applies to the EM algorithm as defined above, as well as to some of its variants that he calls generalized EM (or GEM). The main result is a strong stability guarantee known as *global convergence*, which we discuss below.

We first need a mathematical formalism that describes the EM algorithm. This is done by identifying any homogeneous (in the iterations) iterative algorithm with a specific choice of a mapping $M$ that associates $\theta^{i+1}$ to $\theta^i$. In the theory of Zangwill (1969), one indeed considers families of algorithms by allowing for *point-to-set* maps $M$ that associate a set $M(\theta') \subseteq \Theta$ to each parameter value $\theta' \in \Theta$. A specific algorithm in the family is such that $\theta^{i+1}$ is selected in $M(\theta^i)$. In the example of EM, we may define $M$ as

$$M(\theta') = \left\{\theta \in \Theta \,:\, \mathcal{Q}(\theta;\theta') \geq \mathcal{Q}(\tilde{\theta};\theta') \text{ for all } \tilde{\theta} \in \Theta\right\} , \tag{5.47}$$

that is, $M(\theta')$ is the set of values $\theta$ that maximize $\mathcal{Q}(\theta\,;\theta')$ over $\Theta$. Usually $M(\theta')$ reduces to a singleton, and the mapping $M$ is then simply a point-to-point map (a usual function from $\Theta$ to $\Theta$). But the use of point-to-set maps makes it possible to deal also with cases where the intermediate quantity of EM may have several global maxima, without going into the details of what is done in such cases. We next need the following definition before stating the main convergence theorem.

**Definition 103** (Closed Mapping)**.** *A map $T$ from points of $\Theta$ to subsets of $\Theta$ is said to be* closed *on a set $\mathcal{S} \subseteq \Theta$ if for any converging sequences $\{\theta^i\}_{i\geq 0}$ and $\{\tilde{\theta}^i\}_{i\geq 0}$, the conditions*

*(a) $\theta^i \to \theta \in \mathcal{S}$,*

*(b) $\tilde{\theta}^i \to \tilde{\theta}$ with $\tilde{\theta}^i \in T(\theta^i)$ for all $i \geq 0$,*

*imply that $\tilde{\theta} \in T(\theta)$.*

Note that for point-to-point maps, that is, if $T(\theta)$ is a singleton for all $\theta$, the definition above is equivalent to the requirement that $T$ be continuous on $\mathcal{S}$. Definition 103 is thus a generalization of continuity for general (point-to-set) maps. We are now ready to state the main result, which is proved in Zangwill (1969, p. 91) or Luenberger (1984, p. 187).

**Theorem 104** (Global Convergence Theorem)**.** *Let $\Theta$ be a subset of $\mathbb{R}^{d_\theta}$ and let $\{\theta^i\}_{i\geq 0}$ be a sequence generated by $\theta^{i+1} \in T(\theta^i)$ where $T$ is a point-to-set map on $\Theta$. Let $\mathcal{S} \subseteq \Theta$ be a given "solution" set and suppose that*

*(1) the sequence $\{\theta^i\}_{i\geq 0}$ is contained in a compact subset of $\Theta$;*

*(2) $T$ is closed over $\Theta \setminus \mathcal{S}$ (the complement of $\mathcal{S}$);*

*(3) there is a continuous "ascent" function $s$ on $\Theta$ such that $s(\theta) \geq s(\theta')$ for all $\theta \in T(\theta')$, with strict inequality for points $\theta'$ that are* not *in $\mathcal{S}$.*

*Then the limit of any convergent subsequence of $\{\theta^i\}$ is in the solution set $\mathcal{S}$. In addition, the sequence of values of the ascent function, $\{s(\theta^i)\}_{i\geq 0}$, converges monotonically to $s(\theta_\star)$ for some $\theta_\star \in \mathcal{S}$.*

The final statement of Theorem 104 should not be misinterpreted: that $\{s(\theta^i)\}$ converges to a value that is the image of a point in $\mathcal{S}$ is a simple consequence of the first and third assumptions. It does however not imply that the sequence of parameters $\{\theta^i\}$ is itself convergent in the usual sense, but only that the limit points of $\{\theta^i\}$ have to be in the solution set $\mathcal{S}$. An important property however is that because $\{s(\theta^{i(l)})\}_{l\geq 0}$ converges to $s(\theta_\star)$ for any convergent subsequence $\{\theta^{i(l)}\}$, all limit points of $\{\theta^i\}$ must be in the set $\mathcal{S}_\star = \{\theta \in \Theta : s(\theta) = s(\theta_\star)\}$ (in addition to being in $\mathcal{S}$). This latter statement means that the sequence of iterates $\{\theta^i\}$ will ultimately approach a set of points that are "equivalent" as measured by the ascent function $s$.

The following general convergence theorem following the proof by Wu (1983) is a direct application of the previous theory to the case of EM.

**Theorem 105.** *Suppose that in addition to the hypotheses of Proposition 98 (Assumptions 97 as well as parts (a) and (b) of Proposition 98), the following hold.*

*(i) $\mathcal{H}(\theta\,;\theta')$ is continuous in its second argument, $\theta'$, on $\Theta$.*

*(ii) For any $\theta^0$, the level set $\Theta^0 = \big\{\theta \in \Theta : \ell(\theta) \geq \ell(\theta^0)\big\}$ is compact and contained in the interior of $\Theta$.*

*Then all limit points of any instance $\{\theta^i\}_{i\geq 0}$ of an EM algorithm initialized at $\theta^0$ are in $\mathcal{L}^0 = \{\theta \in \Theta^0 : \nabla_\theta \ell(\theta) = 0\}$, the set of stationary points of $\ell$ with log-likelihood larger than that of $\theta^0$. The sequence $\{\ell(\theta^i)\}$ of log-likelihoods converges monotonically to $\ell_\star = \ell(\theta_\star)$ for some $\theta_\star \in \mathcal{L}^0$.*

*Proof.* This is a direct application of Theorem 104 using $\mathcal{L}^0$ as the solution set and $\ell$ as the ascent function. The first hypothesis of Theorem 104 follows from (ii) and the third one from Proposition 98. The closedness assumption (2) follows from Proposition 98 and (i): for the EM mapping $M$ defined in (5.47), $\tilde{\theta}^i \in M(\theta^i)$ amounts to the condition

$$\mathcal{Q}(\tilde{\theta}^i\,;\theta^i) \geq \mathcal{Q}(\theta\,;\theta^i) \quad \text{for all } \theta \in \Theta\,,$$

which is also satisfied by the limits of the sequences $\{\tilde{\theta}^i\}$ and $\{\theta^i\}$ (if these converge) by continuity of the intermediate quantity $\mathcal{Q}$, which follows from that of $\ell$ and $\mathcal{H}$ (note that it is here important that $\mathcal{H}$ be continuous with respect to both arguments). Hence the EM mapping is indeed closed on $\Theta$ as a whole and Theorem 105 follows. □

The assumptions of Proposition 98 as well as item (i) above are indeed very mild in typical situations. Assumption (ii) however may be restrictive, even for models in which the EM algorithm is routinely used. The practical implication of (ii) being violated is that the EM algorithm may fail to converge to the stationary points of the likelihood for some particularly badly chosen initial points $\theta^0$.

Most importantly, the fact that $\theta^{i+1}$ maximizes the intermediate quantity $\mathcal{Q}(\cdot\,;\theta^i)$ of EM does in no way imply that, ultimately, $\ell_\star$ is the global maximum of $\ell$ over $\Theta$. There is even no guarantee that $\ell_\star$ is a local maximum of the log-likelihood: it may well only be a saddle point (Wu, 1983, Section 2.1). Also, the convergence of the sequence $\ell(\theta^i)$ to $\ell_\star$ does not automatically imply the convergence of $\{\theta^i\}$ to a point $\theta_\star$.

Pointwise convergence of the EM algorithm requires more stringent assumptions that are difficult to verify in practice. As an example, a simple corollary of the global convergence theorem states that if the solution set $\mathcal{S}$ in Theorem 104 is a single point, $\theta_\star$ say, then the sequence $\{\theta^i\}$ indeed converges to $\theta_\star$ (Luenberger, 1984, p. 188). The sketch of the proof of this corollary is that every subsequence of $\{\theta^i\}$ has a convergent further subsequence because of the compactness assumption (1), but such a subsequence admits $s$ as an ascent function and thus converges to $\theta_\star$ by Theorem 104 itself. In cases where the solution set is composed of several points, further conditions are needed to ensure that the sequence of iterates indeed converges and does not cycle through different solution points.

In the case of EM, pointwise convergence of the EM sequence may be guaranteed under an additional condition given by Wu (1983) (see also Boyles, 1983, for an equivalent result), stated in the following theorem.

**Theorem 106.** *Under the hypotheses of Theorem 105, if*

*(iii)* $\|\theta^{i+1} - \theta^i\| \to 0$ *as* $i \to \infty$,

*then all limit points of $\{\theta^i\}$ are in a connected and compact subset of $\mathcal{L}_\star = \{\theta \in \Theta : \ell(\theta) = \ell_\star\}$, where $\ell_\star$ is the limit of the log-likelihood sequence $\{\ell(\theta^i)\}$.*

*In particular, if the connected components of $\mathcal{L}_\star$ are singletons, then $\{\theta^i\}$ converges to some $\theta_\star$ in $\mathcal{L}_\star$.*

*Proof.* The set of limit points of a bounded sequence $\{\theta^i\}$ with $\|\theta^{i+1} - \theta^i\| \to 0$ is connected and compact (Ostrowski, 1966, Theorem 28.1). The proof follows because under Theorem 104, the limit points of $\{\theta^i\}$ must belong to $\mathcal{L}_\star$. □

### 5.5.2 Rate of Convergence of EM

Even if one can guarantee that the EM sequence $\{\hat{\theta}^i\}$ converges to some point $\theta_\star$, this limiting point can be either a local maximum, a saddle point, or even a local minimum. The proposition below states conditions under which the stable stationary points of EM coincide with local maxima only (see also Lange, 1995, Proposition 1, for a similar statement). We here consider that the EM mapping $M$ is a point-to-point map, that is, that the maximizer in the M-step is unique.

To understand the meaning of the term "stable", consider the following approximation to the limit behavior of the EM sequence: it is sensible to expect that if the EM mapping $M$ is sufficiently regular in a neighborhood of the limiting fixed point $\theta_\star$, the asymptotic behavior of the EM sequence $\{\theta^i\}$ follows the tangent linear dynamical system

$$(\theta^{i+1} - \theta_\star) = M(\theta^i) - M(\theta_\star) \approx \nabla_\theta M(\theta_\star)(\theta^i - \theta_\star) \; . \tag{5.48}$$

Here $\nabla_\theta M(\theta_\star)$ is called the *rate matrix* (see for instance Meng and Rubin, 1991). A fixed point $\theta_\star$ is said to be *stable* if the spectral radius of $\nabla_\theta M(\theta_\star)$ is less than 1. In this case, the tangent linear system is asymptotically stable in the sense that the sequence $\{\zeta^i\}$ defined recursively by $\zeta^{i+1} = \nabla_\theta M(\theta_\star)\zeta^i$ tends to zero as $n$ tends to infinity (for any choice of $\zeta^0$). The linear *rate of convergence* of EM is defined as the largest moduli of the eigenvalues of $\nabla_\theta M(\theta_\star)$. This rate is an upper bound on the factors $\rho_k$ that appear in (5.17).

**Proposition 107.** *Under the assumptions of Theorem 100, assume that $\mathcal{Q}(\cdot\,;\theta)$ has a unique maximizer for all $\theta \in \Theta$ and that, in addition,*

$$H(\theta_\star) = -\int \nabla^2_\theta \log f(x\,;\theta)\big|_{\theta=\theta_\star} \, p(x\,;\theta_\star)\, \lambda(dx) \tag{5.49}$$

*and*

$$G(\theta_\star) = -\int \nabla^2_\theta \log p(x\,;\theta)\big|_{\theta=\theta_\star} \, p(x\,;\theta_\star)\, \lambda(dx) \tag{5.50}$$

*are positive definite matrices for all stationary points of EM (i.e., such that $M(\theta_\star) = \theta_\star$). Then for all such points, the following hold true.*

*(i) $\nabla_\theta M(\theta_\star)$ is diagonalizable and its eigenvalues are positive real numbers.*

*(ii) The point $\theta_\star$ is stable for the mapping $M$ if and only if it is a proper maximizer of $\ell(\theta)$ in the sense that all eigenvalues of $\nabla^2_\theta \ell(\theta_\star)$ are negative.*

*Proof.* The EM mapping is defined implicitly through the fact that $M(\theta')$ maximizes $\mathcal{Q}(\cdot\,;\theta')$, which implies that

$$\int \nabla_\theta \, \log f(x\,;\theta)|_{\theta=M(\theta')} \, p(x\,;\theta')\, \lambda(dx) = 0 \; ,$$

using assumption (b) of Theorem 100. Careful differentiation of this relation at a point $\theta' = \theta_\star$, which is such that $M(\theta_\star) = \theta_\star$ and hence $\nabla_\theta \, \ell(\theta)|_{\theta=\theta_\star} = 0$, gives (Dempster *et al.*, 1977; Lange, 1995, see also)

$$\nabla_\theta M(\theta_\star) = [H(\theta_\star)]^{-1} \left[H(\theta_\star) + \nabla^2_\theta \ell(\theta_\star)\right] \; ,$$

where $H(\theta_\star)$ is defined in (5.49). The missing information principle—or Louis' formula (see Proposition 100)—implies that $G(\theta_\star) = H(\theta_\star) + \nabla^2_\theta \ell(\theta_\star)$ is positive definite under our assumptions.

Thus $\nabla_\theta M(\theta_\star)$ is diagonalizable with positive eigenvalues that are the same (counting multiplicities) as those of the matrix $A_\star = I + B_\star$, where $B_\star = [H(\theta_\star)]^{-1/2} \nabla^2_\theta \ell(\theta_\star) [H(\theta_\star)]^{-1/2}$.

Thus $\nabla_\theta M(\theta_\star)$ is stable if and only if $B_\star$ has negative eigenvalues only. The Sylvester law of inertia (see for instance Horn and Johnson, 1985) shows that $B_\star$ has the same inertia (number of positive, negative, and zero eigenvalues) as $\nabla^2_\theta \ell(\theta_\star)$. Thus all of $B_\star$'s eigenvalues are negative if and only if the same is true for $\nabla^2_\theta \ell(\theta_\star)$, that is, if $\theta_\star$ is a proper maximizer of $\ell$. □

The proof above implies that when $\theta_\star$ is stable, the eigenvalues of $M(\theta_\star)$ lie in the interval $(0, 1)$.

### 5.5.3 Generalized EM Algorithms

As discussed above, the type of convergence guaranteed by Theorem 105 is rather weak but, on the other hand, this result is remarkable as it indeed covers not only the original EM algorithm proposed by Dempster *et al.* (1977) but a whole class of variants of the EM approach. One of the most useful extensions of EM is the ECM (for expectation conditional maximization) by Meng and Rubin (1993), which addresses situations where direct maximization of the intermediate quantity of EM is intractable. Assume for instance that the parameter vector $\theta$ consists of two sub-components $\theta_1$ and $\theta_2$, which are such that maximization of $\mathcal{Q}((\theta_1, \theta_2)\,;\theta')$ with respect to $\theta_1$ or $\theta_2$ only (the other sub-component being fixed) is easy, whereas joint maximization with respect to $\theta = (\theta_1, \theta_2)$ is problematic. One may then use the following algorithm for updating the parameter estimate at iteration $i$.

E-step: Compute $\mathcal{Q}((\theta_1, \theta_2)\,;(\theta_1^i, \theta_2^i))$;

CM-step: Determine

$$\theta_1^{i+1} = \arg\max_{\theta_1} \mathcal{Q}((\theta_1, \theta_2^i)\,;(\theta_1^i, \theta_2^i))\ ,$$

and then

$$\theta_2^{i+1} = \arg\max_{\theta_2} \mathcal{Q}((\theta_1^{i+1}, \theta_2)\,;(\theta_1^i, \theta_2^i))\ .$$

It is easily checked that for this algorithm, (5.8) is still verified and thus $\ell$ is an ascent function; this implies that Theorem 105 holds under the same set of assumptions.

The example above is only the simplest case where the ECM approach may be applied, and further extensions are discussed by Meng and Rubin (1993).

# Chapter 6

# Statistical Properties of the Maximum Likelihood Estimator

The maximum likelihood estimator (MLE) is one of the backbones of statistics, and as we have seen in previous chapters, it is very much appropriate also for HMMs, even though numerical approximations are required when the state space is not finite. A standard result in statistics says that, except for "atypical cases", the MLE is consistent, asymptotically normal with asymptotic (scaled) variance equal to the inverse Fisher information matrix, and efficient. The purpose of the current chapter is to show that these properties are indeed true for HMMs as well, provided some conditions of rather standard nature hold. We will also employ the asymptotic results obtained to verify the validity of certain likelihood-based tests.

Recall that the distribution (law) P of $\{Y_k\}_{k\geq 0}$ depends on a parameter $\theta$ that lies in a parameter space $\Theta$, which we assume is a subset of $\mathbb{R}^{d_\theta}$ for some $d_\theta$. Commonly, $\theta$ is a vector containing some components that parameterize the transition kernel of the hidden Markov chain—such as the transition probabilities if the state space $\mathsf{X}$ is finite—and other components that parameterize the conditional distributions of the observations given the states. Throughout the chapter, it is assumed that the HMM model is, for all $\theta$, fully dominated in the sense of Definition 13 and that the underlying Markov chain is positive (see Definition 171).

**Assumption 108.**

*(i) There exists a probability measure $\lambda$ on $(\mathsf{X}, \mathcal{X})$ such that for any $x \in \mathsf{X}$ and any $\theta \in \Theta$, $Q_\theta(x, \cdot) \ll \lambda$ with transition density $q_\theta$. That is, $Q_\theta(x, A) = \int q_\theta(x, x')\, \lambda(dx')$ for $A \in \mathcal{X}$.*

*(ii) There exists a probability measure $\mu$ on $(\mathsf{Y}, \mathcal{Y})$ such that for any $x \in \mathsf{X}$ and any $\theta \in \Theta$, $G_\theta(x, \cdot) \ll \mu$ with transition density function $g_\theta$. That is, $G_\theta(x, A) = \int g_\theta(x, y)\, \mu(dy)$ for $A \in \mathcal{Y}$.*

*(iii) For any $\theta \in \Theta$, $Q_\theta$ is positive, that is, $Q_\theta$ is phi-irreducible and admits a (necessarily unique) invariant distribution denoted by $\pi_\theta$.*

In this chapter, we will generally assume that $\Theta$ is compact. Furthermore, $\theta_\star$ is used to denote the true parameter, that is, the parameter corresponding to the data that we actually observe.

## 6.1 A Primer on MLE Asymptotics

The standard asymptotic properties of the MLE hinge on three basic results: a law of large numbers for the log-likelihood, a central limit theorem for the score function, and a law of large of numbers for the observed information. More precisely,

(i) for all $\theta \in \Theta$, $n^{-1}\ell_n(\theta) \to \ell(\theta)$ $\mathrm{P}_{\theta_\star}$-a.s. uniformly over compact subsets of $\Theta$, where $\ell_n(\theta)$ is the log-likelihood of the parameter $\theta$ given the first $n$ observations and $\ell(\theta)$ is a continuous deterministic function with a unique global maximum at $\theta_\star$;

(ii) $n^{-1/2}\nabla_\theta \ell_n(\theta_\star) \to \mathrm{N}(0, \mathcal{J}(\theta_\star))$ $\mathrm{P}_{\theta_\star}$-weakly, where $\mathcal{J}(\theta)$ is the Fisher information matrix at $\theta$ (we do not provide a more detailed definition at the moment);

(iii) $\lim_{\delta\to 0} \lim_{n\to\infty} \sup_{|\theta-\theta_\star|\leq\delta} \| - n^{-1}\nabla_\theta^2 \ell_n(\theta) - \mathcal{J}(\theta_\star)\| = 0$ $\mathrm{P}_{\theta_\star}$-a.s.

The function $\ell$ in (i) is sometimes referred to as the *contrast function*. We note that $-n^{-1}\nabla_\theta^2\ell_n(\theta)$ in (iii) is the observed information matrix, so that (iii) says that the observed information should converge to the Fisher information in a certain uniform sense. This uniformity may be replaced by conditions on the third derivatives of the log-likelihood, which is common in statistical textbooks, but as we shall see, it is cumbersome enough even to deal with second derivatives of the log-likelihood for HMMs, whence avoiding third derivatives is preferable.

Condition (i) assures strong consistency of the MLE, which can be shown using an argument that goes back to Wald (1949). The idea of the argument is as follows. Denote by $\widehat{\theta}_n$ the maximum the ML estimator; $\ell_n(\widehat{\theta}_n) \geq \ell_n(\theta)$ for any $\theta \in \Theta$. Because $\ell$ has a unique global maximum at $\theta_\star$, $\ell(\theta_\star) - \ell(\theta) \geq 0$ for any $\theta \in \Theta$ and, in particular, $\ell(\theta_\star) - \ell(\widehat{\theta}_n) \geq 0$. We now combine these two inequalities to obtain

$$\begin{aligned} 0 &\leq \ell(\theta_\star) - \ell(\widehat{\theta}_n) \\ &\leq \ell(\theta_\star) - n^{-1}\ell_n(\theta_\star) + n^{-1}\ell_n(\theta_\star) - n^{-1}\ell_n(\widehat{\theta}_n) + n^{-1}\ell_n(\widehat{\theta}_n) - \ell(\widehat{\theta}_n) \\ &\leq 2 \sup_{\theta\in\Theta} |\ell(\theta) - n^{-1}\ell_n(\theta)| \,. \end{aligned}$$

Therefore, by taking the compact subset in (i) above as $\Theta$ itself, $\ell(\widehat{\theta}_n) \to \ell(\theta_\star)$ $\mathrm{P}_{\theta_\star}$-a.s. as $n \to \infty$, which in turn implies, as $\ell$ is continuous with a unique global maximum at $\theta_\star$, that the MLE converges to $\theta_\star$ $\mathrm{P}_{\theta_\star}$-a.s.. In other words, the MLE is strongly consistent.

Provided strong consistency holds, properties (ii) and (iii) above yield asymptotic normality of the MLE. In fact, we must also assume that $\theta_\star$ is an interior point of $\Theta$ and that the Fisher information matrix $\mathcal{J}(\theta_\star)$ is non-singular. Then we can for sufficiently large $n$ make a Taylor expansion around $\theta_\star$, noting that the gradient of $\ell_n$ vanishes at the MLE $\widehat{\theta}_n$ because $\theta_\star$ is maximal there,

$$0 = \nabla_\theta \ell_n(\widehat{\theta}_n) = \nabla_\theta \ell_n(\theta_\star) + \left\{ \int_0^1 \nabla_\theta^2 \ell_n[\theta_\star + t(\widehat{\theta}_n - \theta_\star)]\, dt \right\} (\widehat{\theta}_n - \theta_\star) \,.$$

From this expansion we obtain

$$n^{1/2}(\widehat{\theta}_n - \theta_\star) = \left\{ -n^{-1} \int_0^1 \nabla_\theta^2 \ell_n[\theta_\star + t(\widehat{\theta}_n - \theta_\star)]\, dt \right\}^{-1} n^{-1/2} \nabla_\theta \ell_n(\theta_\star) \,.$$

Now $\widehat{\theta}_n$ converges to $\theta_\star$ $\mathrm{P}_{\theta_\star}$-a.s. and so, using (iii), the first factor on the right-hand side tends to $\mathcal{J}(\theta_\star)^{-1}$ $\mathrm{P}_{\theta_\star}$-a.s. The second factor converges weakly to $\mathrm{N}(0, \mathcal{J}(\theta_\star))$;

this is (ii). Cramér-Slutsky's theorem hence tells us that $n^{1/2}(\widehat{\theta}_n - \theta_\star)$ tends $\mathrm{P}_{\theta_\star}$-weakly to $\mathrm{N}(0, \mathcal{J}^{-1}(\theta_\star))$, and this is the standard result on asymptotic normality of the MLE.

In an entirely similar way properties (ii) and (iii) also show that for any $u \in \mathbb{R}^{d_\theta}$ (recall that $\Theta$ is a subset of $\mathbb{R}^{d_\theta}$),

$$\ell_n(\theta_\star + n^{-1/2}u) - \ell_n(\theta_\star) = n^{-1/2}u^T\nabla_\theta\ell_n(\theta_\star) + \frac{1}{2}u^T[-n^{-1}\nabla_\theta^2\ell_n(\theta_\star)]u + R_n(u) \; ,$$

where $n^{-1/2}\nabla_\theta\ell_n(\theta_\star)$ and $-n^{-1}\nabla_\theta^2\ell_n(\theta_\star)$ converge as described above, and where $R_n(u)$ tends to zero $\mathrm{P}_{\theta_\star}$-a.s. Such an expansion is known as *local asymptotic normality (LAN)* of the model, cf. Ibragimov and Hasminskii (1981, Definition II.2.1). Under this condition, it is known that so-called *regular* estimators (a property possessed by all "sensible" estimators) cannot have an asymptotic covariance matrix smaller than $\mathcal{J}^{-1}(\theta_\star)$ (Ibragimov and Hasminskii, 1981, p. 161). Because this limit is obtained by the MLE, this estimator is efficient.

Later on in this chapter, we will also exploit properties (i)–(iii) to derive asymptotic properties of likelihood ratio and other tests for lower dimensional hypotheses regarding $\theta$.

## 6.2 Stationary Approximations

In this section, we will introduce a way of obtaining properties (i)–(iii) for HMMs; more detailed descriptions are given in subsequent sections.

Before proceeding, we will be precise on the likelihood we shall analyze. In this chapter, we generally make the assumption that the sequence $\{X_k\}_{k\geq 0}$ is stationary; then $\{X_k, Y_k\}_{k\geq 0}$ is stationary as well. Then there is obviously a corresponding likelihood. However, it is sometimes convenient to work with a likelihood $\mathrm{L}_{x_0,n}(\theta)$ that is conditional on an initial state $x_0$,

$$\mathrm{L}_{x_0,n}(\theta) = \int g_\theta(x_0, Y_0) \prod_{i=1}^{n} q_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i) \, \lambda(dx_i) \; . \tag{6.1}$$

We could also want to replace the fixed initial state by an initial distribution $\nu$ on $(\mathsf{X}, \mathcal{X})$, giving

$$\mathrm{L}_{\nu,n}(\theta) = \int_{\mathsf{X}} \mathrm{L}_{x_0,n}(\theta) \, \nu(dx_0) \; .$$

The stationary likelihood is then $\mathrm{L}_{\pi_\theta,n}(\theta)$, which we will simply denote by $\mathrm{L}_n(\theta)$. The advantage of working with the stationary likelihood is of course that it is the correct likelihood for the model and may hence be expected to provide better finite-sample performance. The advantage of assuming a fixed initial state $x_0$—and hence adopting the likelihood $\mathrm{L}_{x_0,n}(\theta)$—is that the stationary distribution $\pi_\theta$ is not always available in closed form when $\mathsf{X}$ is not finite. It is however important that $g_\theta(x_0, Y_0)$ is positive $\mathrm{P}_{\theta_\star}$-a.s.; otherwise the log-likelihood may not be well-defined. In fact, we shall require that $g_\theta(x_0, Y_0)$ is, $\mathrm{P}_{\theta_\star}$-a.s., bounded away from zero. In the following, we always assume that this condition is fulfilled. A further advantage of $\mathrm{L}_{x_0,n}(\theta)$ is that the methods described in the current chapter may be extended to Markov-switching autoregressions (Douc *et al.*, 2004), and then the stationary likelihood is almost never computable, not even when $\mathsf{X}$ is finite. Throughout the rest of this chapter, we will work with $\mathrm{L}_{x_0,n}(\theta)$ unless noticed, where $x_0 \in \mathsf{X}$ is chosen to satisfy the above positivity assumption but otherwise arbitrarily. The MLE arising from this likelihood has the same asymptotic properties as has the MLE arising from $\mathrm{L}_n(\theta)$, provided the initial stationary distribution $\pi_\theta$ has smooth

second-order derivatives (cf. Bickel *et al.*, 1998), whence from an asymptotic point of view there is no loss in using the incorrect likelihood $\mathrm{L}_{x_0,n}(\theta)$.

We now return to the analysis of log-likelihood and items (i)–(iii) above. In the setting of i.i.d. observations, the log-likelihood $\ell_n(\theta)$ is a sum of i.i.d. terms, and so (i) and (iii) follow from uniform versions of the strong law of large numbers and (ii) is a consequence of the simplest central limit theorem. In the case of HMMs, we can write $\ell_{x_0,n}(\theta)$ as a sum as well:

$$\ell_{x_0,n}(\theta) = \sum_{k=0}^{n} \log \left[ \int g_\theta(x_k, Y_k)\, \phi_{x_0,k|k-1}[Y_{0:k-1}](dx_k\,;\theta) \right] \tag{6.2}$$

$$= \sum_{k=0}^{n} \log \left[ \int g_\theta(x_k, Y_k)\, \mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{0:k-1}, X_0 = x_0) \right] , \tag{6.3}$$

where $\phi_{x_0,k|k-1}[Y_{0:k-1}](\cdot\,;\theta)$ is the predictive distribution of the state $X_k$ given the observations $Y_{0:k-1}$ and $X_0 = x_0$. These terms do not form a stationary sequence however, so the law of large numbers—or rather the ergodic theorem—does not apply directly. Instead we must first approximate $\ell_{x_0,n}(\theta)$ by the partial sum of a stationary sequence.

When the joint Markov chain $\{X_k, Y_k\}$ has an invariant distribution, this chain is stationary provided it is started from its invariant distribution. In this case, we can (and will!) extend it to a stationary sequence $\{X_k, Y_k\}_{-\infty<k<\infty}$ with doubly infinite time, as we can do with any stationary sequence. Having done this extension, we can imagine a predictive distribution of the state $X_k$ given the infinite past $Y_{-\infty:k-1}$ of observations. A key feature of these variables is that they now form a stationary sequence, whence the ergodic theorem applies. Furthermore we can approximate $\ell_{x_0,n}(\theta)$ by

$$\ell^s_n(\theta) = \sum_{k=0}^{n} \log \left[ \int g_\theta(x_k, Y_k)\, \mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{-\infty:k-1}) \right] , \tag{6.4}$$

where superindex $s$ stands for "stationary". Heuristically, one would expect this approximation to be good, as observations far in the past do not provide much information about the current one, at least not if the hidden Markov chain enjoys good mixing properties. What we must do is thus to give a precise definition of the predictive distribution $\mathrm{P}_\theta(X_k \in \cdot\,|\,Y_{-\infty:k-1})$ given the infinite past, and then show that it approximates the predictive distribution $\phi_{x_0,k|k-1}(\cdot\,;\theta)$ well enough that the two sums (6.2) and (6.4), after normalization by $n$, have the same asymptotic behavior. We can treat the score function similarly by defining a sequence that forms a stationary martingale increment sequence; for sums of such sequences there is a central limit theorem.

The cornerstone in this analysis is the result on conditional mixing stated in Section 3. We will rephrase it here, but before doing so we state a first assumption. It is really a variation of Assumption 62, adapted to the dominated setting and uniform in $\theta$.

**Assumption 109.**

*(i) The transition density $q_\theta(x, x')$ of $\{X_k\}$ satisfies $0 < \sigma^- \le q_\theta(x, x') \le \sigma^+ < \infty$ for all $x, x' \in \mathsf{X}$ and all $\theta \in \Theta$, and the measure $\lambda$ is a probability measure.*

*(ii) For all $y \in \mathsf{Y}$, the integral $\int_{\mathsf{X}} g_\theta(x, y)\, \lambda(dx)$ is bounded away from 0 and $\infty$ on $\Theta$.*

Part (i) of this assumption often, but not always holds when the state space $\mathsf{X}$ is finite or compact. Note that Assumption 109 says that for all $\theta \in \Theta$, the whole

state space $\mathsf{X}$ is a 1-small set for the transition kernel $Q_\theta$, which implies that for all $\theta \in \Theta$, the chain is phi-irreducible and strongly aperiodic (see Section 7.2 for definitions). It also ensures that there exists a stationary distribution $\pi_\theta$ for $Q_\theta$. In addition, the chain is uniformly geometrically ergodic in the sense that for any $x \in \mathsf{X}$ and $n \geq 0$, $\|Q_\theta^n(x,\cdot) - \pi_\theta\|_{\mathrm{TV}} \leq (1-\sigma^-)^n$. Under Assumption 108, it holds that $\pi_\theta \ll \lambda$, and we use the same notation for this distribution and its density with respect to the dominating measure $\lambda$.

Using the results of Section 7.3, we conclude that the state space $\mathsf{X}\times\mathsf{Y}$ is 1-small for the joint chain $\{X_k, Y_k\}$. Thus the joint chain is also phi-irreducible and strongly aperiodic, and it admits a stationary distribution with density $\pi_\theta(x)g_\theta(x,y)$ with respect to the product measure $\lambda \otimes \mu$ on $(\mathsf{X}\times\mathcal{Y}, \mathcal{X}\otimes\mathcal{Y})$ The joint chain also is uniformly geometrically ergodic.

Put $\rho = 1 - \sigma^-/\sigma^+$; then $0 \leq \rho < 1$. The important consequence of Assumption 109 that we need in the current chapter is Proposition 64. It says that if Assumption 109 holds true, then for all $k \geq 1$, all $y_{0:n}$ and all initial distributions $\nu$ and $\nu'$ on $(\mathsf{X}, \mathcal{X})$,

$$\left\| \int_{\mathsf{X}} \mathrm{P}_\theta(X_k \in \cdot \,|\, X_0 = x, Y_{0:n} = y_{0:n}) \, [\nu(dx) - \nu'(dx)] \right\|_{\mathrm{TV}} \leq \rho^k \,. \tag{6.5}$$

## 6.3 Consistency

### 6.3.1 Construction of the Stationary Conditional Log-likelihood

We shall now construct $\mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{-\infty:k-1})$ and $\int g_\theta(x_k, Y_k)\, \mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{-\infty:k-1})$. The latter variable will be defined as the limit of

$$\mathrm{H}_{k,m,x}(\theta) \stackrel{\mathrm{def}}{=} \int g_\theta(x_k, Y_k)\, \mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{-m+1:k-1}, X_{-m} = x) \tag{6.6}$$

as $m \to \infty$. Note that $\mathrm{H}_{k,m,x}(\theta)$ is the conditional density of $Y_k$ given $Y_{-m+1:k-1}$ and $X_{-m} = x$, under the law $\mathrm{P}_\theta$. Put

$$h_{k,m,x}(\theta) \stackrel{\mathrm{def}}{=} \log \mathrm{H}_{k,m,x}(\theta) \tag{6.7}$$

and consider the following assumption.

**Assumption 110.** *$b^+ = \sup_\theta \sup_{x,y} g_\theta(x,y) < \infty$ and $\mathrm{E}_{\theta_\star} |\log b^-(Y_0)| < \infty$, where $b^-(y) = \inf_\theta \int_{\mathsf{X}} g_\theta(x,y)\, \lambda(dx)$.*

**Lemma 111.** *The following assertions hold true $\mathrm{P}_{\theta_\star}$-a.s. for all indices $k$, $m$ and $m'$ such that $k > -(m \wedge m')$:*

$$\sup_{\theta\in\Theta} \sup_{x,x'\in\mathsf{X}} |h_{k,m,x}(\theta) - h_{k,m',x'}(\theta)| \leq \frac{\rho^{k+(m\wedge m')-1}}{1-\rho} \,, \tag{6.8}$$

$$\sup_{\theta\in\Theta} \sup_{m\geq -(k-1)} \sup_{x\in\mathsf{X}} |h_{k,m,x}(\theta)| \leq |\log b^+| \vee |\log(\sigma^- b^-(Y_k))| \,. \tag{6.9}$$

*Proof.* Assume that $m' \geq m$ and write

$$\begin{aligned} \mathrm{H}_{k,m,x}(\theta) = \iint \left[ \int g_\theta(x_k, Y_k) q_\theta(x_{k-1}, x_k)\, \lambda(dx_k) \right] \\ \times \mathrm{P}_\theta(X_{k-1} \in dx_{k-1} \,|\, Y_{-m+1:k-1}, X_{-m} = x_{-m})\, \delta_x(dx_{-m}) \,, \end{aligned} \tag{6.10}$$

$$\begin{aligned}\mathrm{H}_{k,m',x'}(\theta) = \iint \left[\int g_\theta(x_k, Y_k) q_\theta(x_{k-1}, x_k)\, \lambda(dx_k)\right] \\ \times \mathrm{P}_\theta(X_{k-1} \in dx_{k-1} \,|\, Y_{-m+1:k-1}, X_{-m} = x_{-m}) \\ \times \mathrm{P}_\theta(X_{-m} \in dx_{-m} \,|\, Y_{-m'+1:k-1}, X_{-m'} = x') \,, \end{aligned} \tag{6.11}$$

and invoke (6.5) to see that

$$\begin{aligned}|\mathrm{H}_{k,m,x}(\theta) - \mathrm{H}_{k,m',x'}(\theta)| &\leq \rho^{k+m-1} \sup_{x_{k-1}} \int g_\theta(x_k, Y_k) q_\theta(x_{k-1}, x_k)\, \lambda(dx_k) \\ &\leq \rho^{k+m-1} \sigma^+ \int g_\theta(x_k, Y_k)\, \lambda(dx_k) \,. \end{aligned} \tag{6.12}$$

Note that the step from the total variation bound to the bound on the difference between the integrals does not need a factor "2", because the integrands are non-negative. Also note that (6.5) is stated for $m = m' = 0$, but its initial time index is of course arbitrary. The integral in (6.10) can be bounded from below as

$$\mathrm{H}_{k,m,x}(\theta) \geq \sigma^- \int g_\theta(x_k, Y_k)\, \lambda(dx_k) \,, \tag{6.13}$$

and the same lower bound holds for (6.11). Combining (6.12) with these lower bounds and the inequality $|\log x - \log y| \leq |x - y|/(x \wedge y)$ shows that

$$|h_{k,m,x}(\theta) - h_{k,m',x'}(\theta)| \leq \frac{\sigma^+}{\sigma^-} \rho^{k+m-1} = \frac{\rho^{k+m-1}}{1-\rho} \,,$$

which is the first assertion of the lemma. Furthermore note that (6.10) and (6.13) yield

$$\sigma^- b^-(Y_k) \leq \mathrm{H}_{k,m,x}(\theta) \leq b^+ \,, \tag{6.14}$$

which implies the second assertion. □

Equation (6.8) shows that for any given $k$ and $x$, $\{h_{k,m,x}(\theta)\}_{m \geq -(k-1)}$ is a uniform (in $\theta$) Cauchy sequence as $m \to \infty$, $\mathrm{P}_{\theta_\star}$-a.s., whence there is a $\mathrm{P}_{\theta_\star}$-a.s. limit. Moreover, again by (6.8), this limit does not depend on $x$, so we denote it by $h_{k,\infty}(\theta)$. Our interpretation of this limit is as $\log \mathrm{E}_\theta\,[\, g_\theta(X_k, Y_k) \,|\, Y_{-\infty:k-1}]$. Furthermore (6.9) shows that provided Assumption 110 holds, $\{h_{k,m,x}(\theta)\}_{m \geq -(k-1)}$ is uniformly bounded in $L^1(\mathrm{P}_{\theta_\star})$, so that $h_{k,\infty}(\theta)$ is in $L^1(\mathrm{P}_{\theta_\star})$ and, by the dominated convergence theorem, the limit holds in this mode as well. Finally, by its definition $\{h_{k,\infty}(\theta)\}_{k \geq 0}$ is a stationary process, and it is ergodic because $\{Y_k\}_{-\infty < k < \infty}$ is. We summarize these findings.

**Proposition 112.** *Assume 108, 109, and 110 hold. Then for each $\theta \in \Theta$ and $x \in \mathsf{X}$, the sequence $\{h_{k,m,x}(\theta)\}_{m \geq -(k-1)}$ has, $\mathrm{P}_{\theta_\star}$-a.s., a limit $h_{k,\infty}(\theta)$ as $m \to \infty$. This limit does not depend on $x$. In addition, for any $\theta \in \Theta$, $h_{k,\infty}(\theta)$ belongs to $L^1(\mathrm{P}_{\theta_\star})$, and $\{h_{k,m,x}(\theta)\}_{m \geq -(k-1)}$ also converges to $h_{k,\infty}(\theta)$ in $L^1(\mathrm{P}_{\theta_\star})$ uniformly over $\theta \in \Theta$ and $x \in \mathsf{X}$.*

Having come thus far, we can quantify the approximation of the log-likelihood $\ell_{x_0,n}(\theta)$ by $\ell^s_n(\theta)$.

**Proposition 113.** *For all $n \geq 0$ and $\theta \in \Theta$,*

$$|\ell_{x_0,n}(\theta) - \ell^s_n(\theta)| \leq |\log g_\theta(x_0, Y_0)| + h_{0,\infty}(\theta) + \frac{1}{(1-\rho)^2} \quad \mathrm{P}_{\theta_\star}\text{-a.s.}$$

*Proof.* Letting $m' \to \infty$ in (6.8) we obtain $|h_{k,0,x_0}(\theta) - h_{k,\infty}(\theta)| \le \rho^{k-1}/(1-\rho)$ for $k \ge 1$. Therefore, $\mathrm{P}_{\theta_\star}$-a.s.,

$$\begin{aligned} |\ell_{x_0,n}(\theta) - \ell_n^s(\theta)| &= \left| \sum_{k=0}^{n} h_{k,0,x_0}(\theta) - \sum_{k=0}^{n} h_{k,\infty}(\theta) \right| \\ &\le |\log g_\theta(x_0, Y_0)| + h_{0,\infty}(\theta) + \sum_{k=1}^{n} \frac{\rho^{k-1}}{1-\rho} \, . \end{aligned}$$

□

### 6.3.2 The Contrast Function and Its Properties

Because $h_{k,\infty}(\theta)$ is in $L^1(\mathrm{P}_{\theta_\star})$ under the assumptions made above, we can define the real-valued function $\ell(\theta) \stackrel{\text{def}}{=} \mathrm{E}_{\theta_\star}[h_{k,\infty}(\theta)]$. It does not depend on $k$, by stationarity. This is the contrast function $\ell(\theta)$ referred to above. By the ergodic theorem $n^{-1}\ell_n^s(\theta) \to \ell(\theta)$ $\mathrm{P}_{\theta_\star}$-a.s., and by Proposition 113, $n^{-1}\ell_{x_0,n}(\theta) \to \ell(\theta)$ $\mathrm{P}_{\theta_\star}$-a.s. as well. As noted above, however, we require this convergence to be uniform in $\theta$, which is not guaranteed so far. In addition, we require $\ell(\theta)$ to be continuous and possess a unique global maximum at $\theta_\star$; the latter is an identifiability condition. In the rest of this section, we address continuity and convergence; identifiability is addressed in the next one.

To ensure continuity we need a natural assumption on continuity of the building blocks of the likelihood.

**Assumption 114.** *For all $(x, x') \in \mathsf{X} \times \mathsf{X}$ and $y \in \mathsf{Y}$, the functions $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(x, y)$ are continuous.*

The following result shows that $h_{k,\infty}(\theta)$ is then continuous in $L^1(\mathrm{P}_{\theta_\star})$.

**Proposition 115.** *Assume 108, 109, 110, and 114. Then for any $\theta \in \Theta$,*

$$\mathrm{E}_{\theta_\star}\left[ \sup_{\theta' \in \Theta:\, |\theta' - \theta| \le \delta} |h_{0,\infty}(\theta') - h_{0,\infty}(\theta)| \right] \to 0 \quad \text{as } \delta \to 0 \, ,$$

*and $\theta \mapsto \ell(\theta)$ is continuous on $\Theta$.*

*Proof.* Recall that $h_{0,\infty}(\theta)$ is the limit of $h_{0,m,x}(\theta)$ as $m \to \infty$. We first prove that for any $x \in \mathsf{X}$ and any $m > 0$, the latter quantity is continuous in $\theta$ and then use this to show continuity of the limit. Recall the interpretation of $\mathrm{H}_{0,m,x}(\theta)$ as a conditional density and write

$$\mathrm{H}_{0,m,x}(\theta) = \frac{\int \cdots \int \prod_{i=-m+1}^{0} q_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i) \, \lambda(dx_{-m+1}) \cdots \lambda(dx_0)}{\int \cdots \int \prod_{i=-m+1}^{-1} q_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i) \, \lambda(dx_{-m+1}) \cdots \lambda(dx_{-1})} \tag{6.15}$$

The integrand in the numerator is, by assumption, continuous and bounded by $(\sigma^+ b^+)^m$, whence dominated convergence shows that the numerator is continuous with respect to $\theta$ (recall that $\lambda$ is assumed finite). Likewise the denominator is continuous, and it is bounded from below by $(\sigma^-)^{m-1} \prod_{-m+1}^{-1} b^-(Y_i) > 0$ $\mathrm{P}_{\theta_\star}$-a.s. Thus $\mathrm{H}_{0,m,x}(\theta)$ and $h_{0,m,x}(\theta)$ are continuous as well. Because $h_{0,m,x}(\theta)$ converges to $h_{0,\infty}(\theta)$ uniformly in $\theta$ as $m \to \infty$, $\mathrm{P}_{\theta_\star}$-a.s., $h_{0,\infty}(\theta)$ is continuous $\mathrm{P}_{\theta_\star}$-a.s. The uniform bound (6.9) assures that we can invoke dominated convergence to obtain the first part of the proposition.

The second part is a corollary of the first one, as

$$\begin{aligned}\sup_{\theta' \colon |\theta'-\theta|\le\delta} |\ell(\theta') - \ell(\theta)| &= \sup_{\theta' \colon |\theta'-\theta|\le\delta} |\,\mathrm{E}_{\theta_\star}[h_{0,\infty}(\theta') - h_{0,\infty}(\theta)]| \\ &\le \mathrm{E}_{\theta_\star}\left[\sup_{\theta' \colon |\theta'-\theta|\le\delta} |h_{0,\infty}(\theta') - h_{0,\infty}(\theta)|\right].\end{aligned}$$

□

We can now proceed to show uniform convergence of $n^{-1}\ell_{x_0,n}(\theta)$ to $\ell(\theta)$.

**Proposition 116.** *Assume 108, 109, 110, and 114. Then*

$$\sup_{\theta\in\Theta} |n^{-1}\ell_{x_0,n}(\theta) - \ell(\theta)| \to 0 \quad \mathrm{P}_{\theta_\star}\text{-a.s. as } n\to\infty.$$

*Proof.* First note that because $\Theta$ is compact, it is sufficient to prove that for all $\theta \in \Theta$,

$$\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} |n^{-1}\ell_{x_0,n}(\theta') - \ell(\theta)| = 0 \quad \mathrm{P}_{\theta_\star}\text{-a.s.}$$

Now write

$$\begin{aligned}&\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} |n^{-1}\ell_{x_0,n}(\theta') - \ell(\theta)| \\ &\quad= \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} |n^{-1}\ell_{x_0,n}(\theta') - n^{-1}\ell^s_n(\theta)| \\ &\quad\le \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} n^{-1}|\ell_{x_0,n}(\theta') - \ell^s_n(\theta')| \\ &\qquad+ \limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} n^{-1}|\ell^s_n(\theta') - \ell^s_n(\theta)|\,.\end{aligned}$$

The first term on the right-hand side vanishes by Proposition 113 (note that Lemma 111 shows that $\sup_{\theta'} |h_{0,\infty}(\theta')|$ is in $L^1(\mathrm{P}_{\theta_\star})$ and hence finite $\mathrm{P}_{\theta_\star}$-a.s.). The second term is bounded by

$$\begin{aligned}&\limsup_{\delta\to 0}\limsup_{n\to\infty}\sup_{\theta' \colon |\theta'-\theta|\le\delta} n^{-1}\left|\sum_{k=0}^{n}(h_{k,\infty}(\theta') - h_{k,\infty}(\theta))\right| \\ &\quad\le \limsup_{\delta\to 0}\limsup_{n\to\infty} n^{-1}\sum_{k=0}^{n}\sup_{\theta' \colon |\theta'-\theta|\le\delta}|h_{k,\infty}(\theta') - h_{k,\infty}(\theta)| \\ &\quad= \limsup_{\delta\to 0}\mathrm{E}_{\theta_\star}\left[\sup_{\theta' \colon |\theta'-\theta|\le\delta}|h_{0,\infty}(\theta') - h_{0,\infty}(\theta)|\right] = 0\,,\end{aligned}$$

with convergence $\mathrm{P}_{\theta_\star}$-a.s. The two final steps follow by the ergodic theorem and Proposition 115 respectively. The proof is complete. □

At this point, we thus know that $n^{-1}\ell_{x_0,n}$ converges uniformly to $\ell$. The same conclusion holds when other initial distributions $\nu$ are put on $X_0$, provided $\sup_\theta |\log \int g_\theta(x, Y_0)\,\nu(dx)|$ is finite $\mathrm{P}_{\theta_\star}$-a.s. When $\nu$ is the stationary distribution $\pi_\theta$, uniform convergence can in fact be proved without this extra regularity assumption by conditioning on the previous state $X_{-1}$ to get rid of the first two terms in the bound of Proposition 113; cf. Douc *et al.* (2004).

The uniform convergence of $n^{-1}\ell_{x_0,n}(\theta)$ to $\ell(\theta)$ can be used—with an argument entirely similar to the one of Wald outlined in Section 6.1—to show that the MLE converges a.s. to the set, $\Theta_\star$ say, of global maxima of $\ell$. Because $\ell$ is continuous,

we know that $\Theta_\star$ is closed and hence also compact. More precisely, for any (open) neighborhood of $\Theta_\star$, the MLE will be in that neighborhood for large $n$, $\mathrm{P}_{\theta_\star}$-a.s. We say that the MLE converges to $\Theta_\star$ *in the quotient topology*. This way of describing convergence was used, in the context of HMMs, by Leroux (1992). The purpose of the identifiability constraint, that $\ell(\theta)$ has a *unique* global maximum at $\theta_\star$, is thus to ensure that $\Theta_\star$ consists of the single point $\theta_\star$ so that the MLE indeed converges to the point $\theta_\star$.

## 6.4 Identifiability

As became obvious in the previous section, the set of global maxima of $\ell$ is of intrinsic importance, as this set constitutes the possible limit points of the MLE. The definition of $\ell(\theta)$ as a limit is however usually not suitable for extracting relevant information about the set of maxima, and the purpose of this section is to derive a different characterization of the set of global maxima of $\ell$.

### 6.4.1 Equivalence of Parameters

We now introduce the notion of *equivalence of parameters*.

**Definition 117.** *Two points $\theta, \theta' \in \Theta$ are said to be* equivalent *if they govern identical laws for the process $\{Y_k\}_{k\geq 0}$, that is, if $\mathrm{P}_\theta = \mathrm{P}_{\theta'}$.*

We note that, by virtue of Kolmogorov's extension theorem, $\theta$ and $\theta'$ are equivalent if and only if the finite-dimensional distributions $\mathrm{P}_\theta(Y_1 \in \cdot, Y_2 \in \cdot, \ldots, Y_n \in \cdot)$ and $\mathrm{P}_{\theta'}(Y_1 \in \cdot, Y_2 \in \cdot, \ldots, Y_n \in \cdot)$ agree for all $n \geq 1$.

We will show that a parameter $\theta \in \Theta$ is a global maximum point of $\ell$ if and only if $\theta$ is equivalent to $\theta_\star$. This implies that the limit points of the MLE are those points $\theta$ that govern the same law for $\{Y_k\}_{k\geq 0}$ as does $\theta_\star$. This is the best we can hope for because there is no way—even with an infinitely large sample of $Y$s!—to distinguish between the true parameter $\theta_\star$ and a different but equivalent parameter $\theta$. Naturally we would like to conclude that no parameter other than $\theta_\star$ itself is equivalent to $\theta_\star$. This is not always the case however, in particular when $\mathsf{X}$ is finite and we can number the states arbitrarily. We will discuss this matter further after proving the following result.

**Theorem 118.** *Assume 108, 109, and 110. Then a parameter $\theta \in \Theta$ is a global maximum of $\ell$ if and only if $\theta$ is equivalent to $\theta_\star$.*

An immediate implication of this result is that $\theta_\star$ is a global maximum of $\ell$.

*Proof.* By the definition of $\ell(\theta)$ and Proposition 112,

$$\begin{aligned}
\ell(\theta_\star) - \ell(\theta) &= \mathrm{E}_{\theta_\star}\left[\lim_{m\to\infty} h_{1,m,x}(\theta_\star)\right] - \mathrm{E}_{\theta_\star}\left[\lim_{m\to\infty} h_{1,m,x}(\theta)\right] \\
&= \lim_{m\to\infty} \mathrm{E}_{\theta_\star}\left[h_{1,m,x}(\theta_\star)\right] - \lim_{m\to\infty} \mathrm{E}_{\theta_\star}\left[h_{1,m,x}(\theta)\right] \\
&= \lim_{m\to\infty} \mathrm{E}_{\theta_\star}\left[h_{1,m,x}(\theta_\star) - h_{1,m,x}(\theta)\right] ,
\end{aligned}$$

where $h_{k,m,x}(\theta)$ is given in (6.7). Next, write

$$\begin{aligned}
&\mathrm{E}_{\theta_\star}\left[h_{1,m,x}(\theta_\star) - h_{1,m,x}(\theta)\right] \\
&\qquad = \mathrm{E}_{\theta_\star}\left\{\mathrm{E}_{\theta_\star}\left[\log\frac{\mathrm{H}_{1,m,x}(\theta_\star)}{\mathrm{H}_{1,m,x}(\theta)} \,\middle|\, Y_{-m+1:0}, X_{-m} = x\right]\right\} ,
\end{aligned}$$

where $\mathrm{H}_{k,m,x}(\theta)$ is given in (6.6). Recalling that $\mathrm{H}_{1,m,x}(\theta)$ is the conditional density of $Y_1$ given $Y_{-m+1:0}$ and $X_{-m} = x$, we see that the inner (conditional) expectation on the right-hand side is a Kullback-Leibler divergence and hence non-negative. Thus the outer expectation and the limit $\ell(\theta_\star) - \ell(\theta)$ are non-negative as well, so that $\theta_\star$ is a global mode of $\ell$.

Now pick $\theta \in \Theta$ such that $\ell(\theta) = \ell(\theta_\star)$. Throughout the remainder of the proof, we will use the letter $p$ to denote (possibly conditional) densities of random variables, with the arguments of the density indicating which random variables are referred to. For any $k \geq 1$,

$$\begin{aligned}
&\mathrm{E}_{\theta_\star}[\log p_\theta(Y_{1:k}|Y_{-m+1:0}, X_{-m} = x)] \\
&\quad = \sum_{i=1}^{k} \mathrm{E}_{\theta_\star}[\log p_\theta(Y_i|Y_{-m+1:i-1}, X_{-m} = x)] \\
&\quad = \sum_{i=1}^{k} \mathrm{E}_{\theta_\star}[h_{i,m,x}(\theta)]
\end{aligned}$$

so that, employing stationarity,

$$\lim_{m\to\infty} \mathrm{E}_{\theta_\star}[\log p_\theta(Y_{1:k}|Y_{-m+1:0}, X_{-m} = x)] = k\ell(\theta) \; .$$

Thus for any positive integer $n < k$,

$$\begin{aligned}
0 &= k(\ell(\theta_\star) - \ell(\theta)) \\
&= \lim_{m\to\infty} \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{1:k}|Y_{-m+1:0}, X_{-m} = x)}{p_\theta(Y_{1:k}|Y_{-m+1:0}, X_{-m} = x)}\right] \\
&= \lim_{m\to\infty} \left\{ \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{k-n+1:k}|Y_{-m+1:0}, X_{-m} = x)}{p_\theta(Y_{k-n+1:k}|Y_{-m+1:0}, X_{-m} = x)}\right] \right. \\
&\qquad \left. + \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{1:k-n}|Y_{k-n+1:k}, Y_{-m+1:0}, X_{-m} = x)}{p_\theta(Y_{1:k-n}|Y_{k-n+1:k}, Y_{-m+1:0}, X_{-m} = x)}\right] \right\} \\
&\geq \limsup_{m\to\infty} \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{1:n}|Y_{n-k-m+1:n-k}, X_{n-k-m} = x)}{p_\theta(Y_{1:n}|Y_{n-k-m+1:n-k}, X_{n-k-m} = x)}\right] \; ,
\end{aligned}$$

where the inequality follows by using stationarity for the first term and noting that the second term is non-negative as an expectation of a (conditional) Kullback-Leibler divergence as above. Hence we have inserted a gap between the variables $Y_{1:n}$ whose density we examine and the variables $Y_{n-k-m+1:n-k}$ and $X_{n-k-m}$ that appear as a condition. The idea is now to let this gap tend to infinity and to show that in the limit the condition has no effect. Next we shall thus show that

$$\begin{aligned}
\lim_{k\to\infty} \sup_{m\geq k} \left| \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x)}{p_\theta(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x)}\right] \right. \\
\left. - \mathrm{E}_{\theta_\star}\left[\log \frac{p_{\theta_\star}(Y_{1:n})}{p_\theta(Y_{1:n})}\right] \right| = 0 \; . \qquad (6.16)
\end{aligned}$$

Combining (6.16) with the previous inequality, it is clear that if $\ell(\theta) = \ell(\theta_\star)$, then $\mathrm{E}_{\theta_\star}\{\log[p_{\theta_\star}(Y_{1:n})/p_\theta(Y_{1:n})]\} = 0$, that is, the Kullback-Leibler divergence between the $n$-dimensional densities $p_{\theta_\star}(y_{1:n})$ and $p_\theta(y_{1:n})$ vanishes. This implies, by the information inequality, that these densities coincide except on a set with $\mu^{\otimes n}$-measure zero, so that the $n$-dimensional laws of $\mathrm{P}_{\theta_\star}$ and $\mathrm{P}_\theta$ agree. Because $n$ was arbitrary, we find that $\theta_\star$ and $\theta$ are equivalent.

What remains to do is thus to prove (6.16). To that end, put $U_{k,m}(\theta) = \log p_\theta(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x)$ and $U(\theta) = \log p_\theta(Y_{1:n})$. Obviously, it is enough

to prove that for all $\theta \in \Theta$,

$$\lim_{k\to\infty} \mathrm{E}_{\theta_\star}\left[\sup_{m\geq k} |U_{k,m}(\theta) - U(\theta)|\right] = 0\,. \tag{6.17}$$

To do that we write

$$\begin{aligned} p_\theta(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x) &= \iint p_\theta(Y_{1:n}|X_0 = x_0)\, Q_\theta^k(x_{-k}, dx_0) \\ &\quad \times \mathrm{P}_\theta(X_{-k} \in dx_{-k}\,|\,Y_{-m+1:-k}, X_{-m} = x) \end{aligned}$$

and

$$p_\theta(Y_{1:n}) = \iint p_\theta(Y_{1:n}|X_0 = x_0)\, Q_\theta^k(x_{-k}, dx_0)\, \pi_\theta(dx_{-k})\,,$$

where $\pi_\theta$ is the stationary distribution of $\{X_k\}$. Realizing that $p_\theta(Y_{1:n}|X_0 = x_0)$ is bounded from above by $(b^+)^n$ (condition on $X_{1:n}$!) and that the transition kernel $Q_\theta$ satisfies the Doeblin condition (see Definition 50) and is thus uniformly geometrically ergodic (see Definition 53 and Lemma 51), we obtain

$$\sup_{m\geq k} |p_\theta(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x) - p_\theta(Y_{1:n})| \leq (b^+)^n(1-\sigma^-)^k \tag{6.18}$$

$\mathrm{P}_{\theta_\star}$-a.s.. Moreover, the bound

$$\begin{aligned} p_\theta(Y_{1:n}|X_0 = x_0) &= \int\cdots\int \prod_{i=1}^n q_\theta(x_{i-1}, x_i) g_\theta(x_i, Y_i)\, \lambda(dx_i) \\ &\geq (\sigma^-)^n \prod_{i=1}^n b^-(Y_i) \end{aligned}$$

implies that $p_\theta(Y_{1:n}|Y_{-m+1:-k}, X_{-m} = x)$ and $p_\theta(Y_{1:n})$ both obey the same lower bound. Combined with the observation $b^-(Y_i) > 0$ $\mathrm{P}_{\theta_\star}$-a.s., which follows from Assumption 110, and the bound $|\log(x) - \log(y)| \leq |x-y|/x \wedge y$, (6.18) shows that

$$\lim_{k\to\infty} \sup_{m\geq k} |U_{k,m}(\theta) - U(\theta)| \to 0 \quad \mathrm{P}_{\theta_\star}\text{-a.s.}$$

Now (6.17) follows from dominated convergence provided

$$\mathrm{E}_\theta\left[\sup_k \sup_{m\geq k} U_{k,m}(\theta)\right] < \infty\,.$$

Using the aforementioned bounds, we conclude that this expectation is indeed finite. □

We remark that the basic structure of the proof is potentially applicable also to models other than HMMs. Indeed, using the notation of the proof, we may define $\ell$ as $\ell(\theta) = \lim_{m\to\infty} \mathrm{E}_{\theta_\star}[\log p_\theta(Y_1|Y_{-m:1})]$, a definition that does not exploit the HMM structure. Then the first part of the proof, up to (6.16), does not use the HMM structure either, so that all that is needed, in a more general framework, is to verify (6.16) (or, more precisely, a version thereof not containing $X_{-m}$). For particular other processes, this could presumably be carried out using, for instance, suitable mixing properties.

The above theorem shows that the points of global maxima of $\ell$—forming the set of possible limit points of the MLE—are those that are statistically equivalent to $\theta_\star$. This result, although natural and important (but not trivial!), is however yet of a somewhat "high level" character, that is, not verifiable in terms of "low level" conditions. We would like to provide some conditions, expressed directly in terms of the Markov chain and the conditional distributions $g_\theta(x, y)$, that give information about parameters that are equivalent to $\theta_\star$ and, in particular, when there is no other such parameter than $\theta_\star$. We will do this using the framework of mixtures of distributions.

### 6.4.2 Identifiability of Mixture Densities

We first define what is meant by a mixture density.

**Definition 119.** *Let $f_\phi(y)$ be a parametric family of densities on $\mathsf{Y}$ with respect to a common dominating measure $\mu$ and parameter $\phi$ in some set $\Phi$. If $\pi$ is a probability measure on $\Phi$, then the density*

$$f_\pi(y) = \int_\Phi f_\phi(y)\, \pi(d\phi)$$

*is called a* mixture density*; the distribution $\pi$ is called the* mixing distribution*.*

*We say that* the class of (all) mixtures of $(f_\phi)$ is identifiable *if $f_\pi = f_{\pi'}$ $\mu$-a.e. if and only if $\pi = \pi'$.*

*Furthermore we say that* the class of finite mixtures of $(f_\phi)$ is identifiable *if for all probability measures $\pi$ and $\pi'$ with finite support, $f_\pi = f_{\pi'}$ $\mu$-a.e. if and only if $\pi = \pi'$.*

In other words, the class of all mixtures of $(f_\phi)$ is identifiable if the two distributions with densities $f_\pi$ and $f_{\pi'}$ respectively agree only when $\pi = \pi'$. Yet another way to put this property is to say that identifiability means that the mapping $\pi \mapsto f_\pi$ is one-to-one (injective). A way, slightly Bayesian, of thinking of a mixture distribution that is often intuitive and fruitful is the following. Draw $\phi \in \Phi$ with distribution $\pi$ and then $Y$ from the density $f_\phi$. Then, $Y$ has density $f_\pi$.

Many important and commonly used parametric classes of densities are identifiable. We mention the following examples.

(i) The Poisson family (Feller, 1943). In this case, $\mathsf{Y} = \mathbb{Z}_+$, $\Phi = \mathbb{R}_+$, $\phi$ is the mean of the Poisson distribution, $\mu$ is counting measure, and $f_\phi(y) = \phi^y \mathrm{e}^{-\phi}/y!$.

(ii) The Gamma family (Teicher, 1961), with the mixture being either on the scale parameter (with a fixed form parameter) or on the form parameter (with a fixed scale parameter). The class of joint mixtures over both parameters is not identifiable however, but the class of joint *finite* mixtures is identifiable.

(iii) The normal family (Teicher, 1960), with the mixture being either on the mean (with fixed variance) or on the variance (with fixed mean). The class of joint mixtures over both mean and variance is not identifiable however, but the class of joint *finite* mixtures is identifiable.

(iv) The Binomial family $\mathrm{Bin}(N, p)$ (Teicher, 1963), with the mixture being on the probability $p$. The class of finite mixtures is identifiable, provided the number of components $k$ of the mixture satisfies $2k - 1 \leq N$.

Further reading on identifiability of mixtures is found, for instance, in Titterington *et al.* (1985, Section 3.1).

A very useful result on mixtures, taking identifiability in one dimension into several dimensions, is the following.

**Theorem 120** (Teicher, 1967)**.** *Assume that the class of all mixtures of the family $(f_\phi)$ of densities on $\mathsf{Y}$ with parameter $\phi \in \Phi$ is identifiable. Then the class of all mixtures of the $n$-fold product densities $f_\phi^{(n)}(y) = f_{\phi_1}(y_1) \cdots f_{\phi_n}(y_n)$ on $y \in \mathsf{Y}^n$ with parameter $\phi \in \Phi^n$ is identifiable. The same conclusion holds true when "all mixtures" is replaced by "finite mixtures".*

### 6.4.3 Application of Mixture Identifiability to Hidden Markov Models

Let us now explain how identifiability of mixture densities applies to HMMs. Assume that $\{X_k, Y_k\}$ is an HMM such that the conditional densities $g_\theta(x, y)$ all belong to a single parametric family. Then given $X_k = x$, $Y_k$ has conditional density $g_{\phi(x)}$ say, where $\phi(x)$ is a function mapping the current state $x$ into the parameter space $\Phi$ of the parametric family of densities. Now assume that the class of all mixtures of this family of densities is identifiable, and that we are given a true parameter $\theta_\star$ of the model as well as an equivalent other parameter $\theta$. Associated with these two parameters are two mappings $\phi_\star(x)$ and $\phi(x)$, respectively, as above. As $\theta_\star$ and $\theta$ are equivalent, the $n$-dimensional restrictions of $\mathrm{P}_{\theta_\star}$ and $\mathrm{P}_\theta$ coincide; that is, $\mathrm{P}_{\theta_\star}(Y_{1:n} \in \cdot)$ and $\mathrm{P}_\theta(Y_{1:n} \in \cdot)$ agree. Because the class of all mixtures of $(g_\phi)$ is identifiable, Theorem 120 tells us that the $n$-dimensional distributions of the processes $\{\phi_\star(X_k)\}$ and $\{\phi(X_k)\}$ agree. That is, for all subsets $A \subseteq \Phi^n$,

$$\begin{aligned} &\mathrm{P}_{\theta_\star}\{(\phi_\star(X_1), \phi_\star(X_2), \ldots, \phi_\star(X_n)) \in A\} \\ &\quad = \mathrm{P}_\theta\{(\phi(X_1), \phi(X_2), \ldots, \phi(X_n)) \in A\} \, . \end{aligned}$$

This condition is often informative for concluding $\theta = \theta_\star$.

**Example 121** (Normal HMM)**.** Assume that $\mathsf{X}$ is finite, say $\mathsf{X} = \{1, 2, \ldots, r\}$, and that $Y_k | X_k = i \sim \mathrm{N}(\mu_i, \sigma^2)$. The parameters of the model are the transition probabilities $q_{ij}$ of $\{X_k\}$, the $\mu_i$ and $\sigma^2$. We thus identify $\phi(x) = \mu_x$. If $\theta_\star$ and $\theta$ are two equivalent parameters, the laws of the processes $\{\mu_{\star X_k}\}$ and $\{\mu_{X_k}\}$ are thus the same, and in addition $\sigma_\star^2 = \sigma^2$. Here $\mu_{\star i}$ denotes the $\mu_i$-component of $\theta_\star$, etc. Assuming the $\mu_{\star i}$ to be distinct, this can only happen if the *sets* $\{\mu_{\star 1}, \ldots, \mu_{\star r}\}$ and $\{\mu_1, \ldots, \mu_r\}$ are identical. We may thus conclude that the sets of means must be the same for both parameters, but they need not be enumerated in the same order. Thus there is a permutation $\{c(1), c(2), \ldots, c(r)\}$ of $\{1, 2, \ldots, r\}$ such that $\mu_{c(i)} = \mu_{\star i}$ for all $i \in \mathsf{X}$. Now because the laws of $\{\mu_{\star X_k}\}$ under $\mathrm{P}_{\theta_\star}$ and $\{\mu_{c(X_k)}\}$ under $\mathrm{P}_\theta$ coincide with the $\mu_i$s being distinct, we conclude that the laws of $\{X_k\}$ under $\mathrm{P}_{\theta_\star}$ and of $\{c(X_k)\}$ under $\mathrm{P}_\theta$ also agree, which in turn implies $q_{\star ij} = q_{c(i),c(j)}$ for all $i, j \in \mathsf{X}$.

Hence any parameter $\theta$ that is equivalent to $\theta_\star$ is in fact identical, up to a permutation of state indices. Sometimes the parameter space is restricted by, for instance, requiring the means $\mu_i$ to be sorted: $\mu_1 < \mu_2 < \ldots < \mu_r$, which removes the ambiguity.

In the current example, we could also have allowed the variance $\sigma^2$ to depend on the state, $Y_k | X_k = i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, reaching the same conclusion. The assumption of conditional normality is of course not crucial either; any family of distributions for which finite mixtures are identifiable would do.

**Example 122** (General Stochastic Volatility)**.** In this example, we consider a stochastic volatility model of the form $Y_k | X_k = x \in \mathrm{N}(0, \sigma^2(x))$, where $\sigma^2(x)$ is a mapping from $\mathsf{X}$ to $\mathbb{R}_+$. Thus, we identify $\phi(x) = \sigma^2(x)$. Again assume that we are given a true parameter $\theta_\star$ as well as another parameter $\theta$, which is equivalent to $\theta_\star$. Because all variance mixtures of normal distributions are identifiable, the laws of $\{\sigma_\star^2(X_k)\}$ under $\mathrm{P}_{\theta_\star}$ and of $\{\sigma^2(X_k)\}$ under $\mathrm{P}_\theta$ agree. Assuming for instance that $\sigma_\star^2(x) = \sigma^2(x) = x$ (and hence also $\mathsf{X} \subseteq \mathbb{R}_+$), we conclude that the laws of $\{X_k\}$ under $\mathrm{P}_{\theta_\star}$ and $\mathrm{P}_\theta$, respectively, agree. For particular models of the transition kernel $Q$ of $\{X_k\}$, such as the finite case of the previous example, we may then be able to show that $\theta = \theta_\star$, possibly up to a permutation of state indices.

**Example 123.** Sometimes a model with finite state space is identifiable even though the conditional densities $g(x, \cdot)$ are identical for several $x$. For instance,

consider a model on the state space $\mathsf{X} = \{0, 1, 2\}$ with $Y_k | X_k = i \sim \mathrm{N}(\mu_i, \sigma^2)$, the constraints $\mu_0 = \mu_1 < \mu_2$, and transition probability matrix

$$Q = \begin{pmatrix} q_{00} & q_{01} & 0 \\ q_{10} & q_{11} & q_{12} \\ 0 & q_{21} & q_{22} \end{pmatrix} .$$

The Markov chain $\{X_k\}$ is thus a (discrete-time) birth-and-death process in the sense that it can change its state index by at most one in each step. This model is similar to models used in modeling ion channel dynamics (cf. Fredkin and Rice, 1992). Because $\mu_1 < \mu_2$, we could then think of states 0 and 1 as "closed" and of state 2 as "open".

Now assume that $\theta$ is equivalent to $\theta_\star$. Just as in Example 121, we may then conclude that the law of $\{\mu_{\star X_k}\}$ under $\mathrm{P}_{\theta_\star}$ and that of $\{\mu_{X_k}\}$ under $\mathrm{P}_\theta$ agree, and hence, because of the constraints on the $\mu$s, that the laws of $\{\mathbb{1}(X_k \in \{0,1\}) + \mathbb{1}(X_k = 2)\}$ under $\mathrm{P}_{\theta_\star}$ and $\mathrm{P}_\theta$ agree. In other words, after lumping states 0 and 1 of the Markov chain we obtain processes with identical laws. This in particular implies that the distributions under $\mathrm{P}_{\theta_\star}$ and $\mathrm{P}_\theta$ of the sojourn times in the state aggregate $\{0, 1\}$ coincide. The probability of such a sojourn having length 1 is $q_{12}$, whence $q_{12} = q_{\star 12}$ must hold. For length 2, the corresponding probability is $q_{11}q_{12}$, whence $q_{11} = q_{\star 11}$ follows and then also $q_{10} = q_{\star 10}$ as rows of $Q$ sum up to unity. For length 3, the probability is $q_{11}^2 q_{12} + q_{10}q_{01}q_{12}$, so that finally $q_{01} = q_{\star 01}$ and $q_{00} = q_{\star 00}$. We may thus conclude that $\theta = \theta_\star$, that is, the model is identifiable. The reason that identifiability holds despite the means $\mu_i$ being non-distinct is the special structure of $Q$. For further reading on identifiability of lumped Markov chains, see Ito *et al.* (1992).

## 6.5 Asymptotic Normality of the Score and Convergence of the Observed Information

We now turn to asymptotic properties of the score function and the observed information. The score function will be discussed in some detail, whereas for the information matrix we will just state the results.

### 6.5.1 The Score Function and Invoking the Fisher Identity

Define the score function

$$\nabla_\theta \ell_{x_0,n}(\theta) = \sum_{k=0}^{n} \nabla_\theta \log \left[ \int g_\theta(x_k, Y_k)\, \mathrm{P}_\theta(X_k \in dx_k \,|\, Y_{0:k-1}, X_0 = x_0) \right] . \quad (6.19)$$

To make sure that this gradient indeed exists and is well-behaved enough for our purposes, we make the following assumptions.

**Assumption 124.** *There exists an open neighborhood $\mathcal{U} = \{\theta : |\theta - \theta_\star| < \delta\}$ of $\theta_\star$ such that the following hold.*

*(i) For all $(x, x') \in \mathsf{X} \times \mathsf{X}$ and all $y \in \mathsf{Y}$, the functions $\theta \mapsto q_\theta(x, x')$ and $\theta \mapsto g_\theta(x, y)$ are twice continuously differentiable on $\mathcal{U}$.*

*(ii)*

$$\sup_{\theta \in \mathcal{U}} \sup_{x,x'} \|\nabla_\theta \log q_\theta(x, x')\| < \infty$$

*and*

$$\sup_{\theta \in \mathcal{U}} \sup_{x,x'} \|\nabla_\theta^2 \log q_\theta(x, x')\| < \infty .$$

*(iii)*

$$\mathrm{E}_{\theta_\star}\left[\sup_{\theta\in\mathcal{U}}\sup_x \|\nabla_\theta \log g_\theta(x, Y_1)\|^2\right] < \infty$$

*and*

$$\mathrm{E}_{\theta_\star}\left[\sup_{\theta\in\mathcal{U}}\sup_x \|\nabla^2_\theta \log g_\theta(x, Y_1)\|\right] < \infty\ .$$

*(iv) For $\mu$-almost all $y \in \mathsf{Y}$, there exists a function $f_y : \mathsf{X} \to \mathbb{R}_+$ in $L^1(\lambda)$ such that $\sup_{\theta\in\mathcal{U}} g_\theta(x, y) \le f_y(x)$.*

*(v) For $\lambda$-almost all $x \in \mathsf{X}$, there exist functions $f^1_x : \mathsf{Y} \to \mathbb{R}_+$ and $f^2_x : \mathsf{Y} \to \mathbb{R}_+$ in $L^1(\mu)$ such that $\|\nabla_\theta g_\theta(x, y)\| \le f^1_x(y)$ and $\|\nabla^2_\theta g_\theta(x, y)\| \le f^2_x(y)$ for all $\theta \in \mathcal{U}$.*

These assumptions assure that the log-likelihood is twice continuously differentiable, and also that the score function and observed information have finite moments of order two and one, respectively, under $\mathrm{P}_{\theta_\star}$. The assumptions are natural extensions of standard assumptions that are used to prove asymptotic normality of the MLE for i.i.d. observations. The asymptotic results to be derived below are valid also for likelihoods obtained using a distribution $\nu_\theta$ for $X_0$ (such as the stationary one), provided this distribution satisfies conditions similar to the above ones: for all $x \in \mathsf{X}$, $\theta \mapsto \nu_\theta(x)$ is twice continuously differentiable on $\mathcal{U}$, and the first and second derivatives of $\theta \mapsto \log \nu_\theta(x)$ are bounded uniformly over $\theta \in \mathcal{U}$ and $x \in \mathsf{X}$.

We shall now study the score function and its asymptotics in detail. Even though the log-likelihood is differentiable, one must take some care to arrive at an expression for the score function that is useful. A tool that is often useful in the context of models with incompletely observed data is the so-called *Fisher identity*, which we encountered in Section 5.1.3. Invoking this identity, which holds in a neighborhood of $\theta_\star$ under Assumption 124, we find that (cf. (5.28))

$$\nabla_\theta \ell_{x_0,n}(\theta) = \nabla_\theta \log g_\theta(x_0, Y_0) + \mathrm{E}_\theta\left[\sum_{k=1}^n \phi_\theta(X_{k-1}, X_k, Y_k)\,\middle|\, Y_{0:n}, X_0 = x_0\right]\ , \tag{6.20}$$

where $\phi_\theta(x, x', y') = \nabla_\theta \log[q_\theta(x, x')g_\theta(x', y')]$. However, just as when we obtained a law of large numbers for the normalized log-likelihood, we want to express the score function as a sum of increments, conditional scores. For that purpose we write

$$\nabla_\theta \ell_{x_0,n}(\theta) = \nabla_\theta \ell_{x_0,0}(\theta) + \sum_{k=1}^n \{\nabla_\theta \ell_{x_0,k}(\theta) - \nabla_\theta \ell_{x_0,k-1}(\theta)\} = \sum_{k=0}^n \dot{h}_{k,0,x_0}(\theta)\ , \tag{6.21}$$

where $\dot{h}_{0,0,x_0} = \nabla_\theta \log g_\theta(x_0, Y_0)$ and, for $k \ge 1$,

$$\begin{aligned}\dot{h}_{k,0,x}(\theta) \;=\; & \mathrm{E}_\theta\left[\sum_{i=1}^k \phi_\theta(X_{i-1}, X_i, Y_i)\,\middle|\, Y_{0:k}, X_0 = x\right] \\ & - \mathrm{E}_\theta\left[\sum_{i=1}^{k-1} \phi_\theta(X_{i-1}, X_i, Y_i)\,\middle|\, Y_{0:k-1}, X_0 = x\right]\ .\end{aligned}$$

Note that $\dot{h}_{k,0,x}(\theta)$ is the gradient with respect to $\theta$ of the conditional log-likelihood $h_{k,0,x}(\theta)$ as defined in (6.7). It is a matter of straightforward algebra to check that (6.20) and (6.21) agree.

### 6.5.2 Construction of the Stationary Conditional Score

We can extend, for any integers $k \geq 1$ and $m \geq 0$, the definition of $\dot{h}_{k,0,x}(\theta)$ to

$$\begin{aligned}\dot{h}_{k,m,x}(\theta) &= \mathrm{E}_\theta\left[\sum_{i=-m+1}^{k} \phi_\theta(X_{i-1}, X_i, Y_i) \,\middle|\, Y_{-m+1:k}, X_{-m} = x\right] \\ &\quad - \mathrm{E}_\theta\left[\sum_{i=-m+1}^{k-1} \phi_\theta(X_{i-1}, X_i, Y_i) \,\middle|\, Y_{-m+1:k-1}, X_{-m} = x\right]\end{aligned}$$

with the aim, just as before, to let $m \to \infty$. This will yield a definition of $\dot{h}_{k,\infty}(\theta)$; the dependence on $x$ will vanish in the limit. Note however that the construction below does not show that this quantity is in fact the gradient of $h_{k,\infty}(\theta)$, although one can indeed prove that this is the case.

As noted in Section 6.1, we want to prove a central limit theorem (CLT) for the score function evaluated at the true parameter. A quite general way to do that is to recognize that the corresponding score increments form, under reasonable assumptions, a martingale increment sequence with respect to the filtration generated by the observations. This sequence is not stationary though, so one must either use a general martingale CLT or first approximate the sequence by a stationary martingale increment sequence. We will take the latter approach, and our approximating sequence is nothing but $\{\dot{h}_{k,\infty}(\theta_\star)\}$.

We now proceed to the construction of $\dot{h}_{k,\infty}(\theta)$. First write $\dot{h}_{k,m,x}(\theta)$ as

$$\begin{aligned}\dot{h}_{k,m,x}(\theta) &= \mathrm{E}_\theta[\phi_\theta(X_{k-1}, X_k, Y_k) \,|\, Y_{-m+1:k}, X_{-m} = x] \\ &\quad + \sum_{i=-m+1}^{k-1} (\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x] \\ &\qquad - \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k-1}, X_{-m} = x]) \,. \qquad (6.22)\end{aligned}$$

The following result shows that it makes sense to take the limit as $m \to \infty$ in the previous display.

**Proposition 125.** *Assume 108, 109, and 124 hold. Then for any integers $1 \leq i \leq k$, the sequence $\{\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]\}_{m \geq 0}$ converges $\mathrm{P}_{\theta_\star}$-a.s. and in $L^2(\mathrm{P}_{\theta_\star})$, uniformly with respect to $\theta \in \mathcal{U}$ and $x \in \mathsf{X}$, as $m \to \infty$. The limit does not depend on $x$.*

We interpret and write this limit as $\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-\infty:k}]$.

*Proof.* The proof is entirely similar to that of Proposition 112. For any $(x, x') \in \mathsf{X} \times \mathsf{X}$ and non-negative integers $m' \geq m$,

$$\begin{aligned}&\big|\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x] \\ &\qquad - \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m'+1:k}, X_{-m'} = x']\big| \\ &= \left|\iiint \phi_\theta(x_{i-1}, x_i, Y_i)\, Q_\theta(x_{i-1}, dx_i)\right. \\ &\qquad \times \mathrm{P}_\theta(X_{i-1} \in dx_{i-1} \,|\, Y_{-m+1:k}, X_{-m} = x_{-m}) \\ &\qquad \left.\times [\delta_x(dx_{-m}) - \mathrm{P}_\theta(X_{-m} \in dx_{-m} \,|\, Y_{-m'+1:k}, X_{-m'} = x')]\right| \\ &\leq 2 \sup_{x,x'} \|\phi_\theta(x, x', Y_i)\| \rho^{(i-1)+m} \,, \qquad (6.23)\end{aligned}$$

where the inequality stems from (6.5). Setting $x = x'$ in this display shows that $\{\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]\}_{m\geq 0}$ is a Cauchy sequence, thus converging $\mathrm{P}_{\theta_\star}$-a.s. The inequality also shows that the limit does not depend on $x$. Moreover, because for any non-negative integer $m$, $x \in \mathsf{X}$ and $\theta \in \mathcal{U}$,

$$\|\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]\| \leq \sup_{x,x'} \|\phi_\theta(x, x', Y_i)\|$$

with the right-hand side belonging to $L^2(\mathrm{P}_{\theta_\star})$. The inequality (6.23) thus also shows that $\{\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]\}_{m\geq 0}$ is a Cauchy sequence in $L^2(\mathrm{P}_{\theta_\star})$ and hence converges in $L^2(\mathrm{P}_{\theta_\star})$. □

With the sums arranged as in (6.22), we can let $m \to \infty$ and define, for $k \geq 1$,

$$\begin{aligned}\dot{h}_{k,\infty}(\theta) &= \mathrm{E}_\theta[\phi_\theta(X_{k-1}, X_k, Y_k) \,|\, Y_{-\infty:k}] \\ &\quad + \sum_{i=-\infty}^{k-1} \left(\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-\infty:k}] - \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-\infty:k-1}]\right) .\end{aligned}$$

The following result gives an $L^2$-bound on the difference between $\dot{h}_{k,m,x}(\theta)$ and $\dot{h}_{k,\infty}(\theta)$.

**Lemma 126.** *Assume 108, 109, 110, and 124 hold. Then for $k \geq 1$,*

$$\begin{aligned}&(\mathrm{E}_\theta \,\|\dot{h}_{k,m,x}(\theta) - \dot{h}_{k,\infty}(\theta)\|^2)^{1/2} \\ &\qquad \leq \quad 12 \left(\mathrm{E}_\theta \left[\sup_{x,x'\in\mathsf{X}} \|\phi_\theta(x, x', Y_1)\|^2\right]\right)^{1/2} \frac{\rho^{(k+m)/2-1}}{1-\rho} .\end{aligned}$$

*Proof.* The idea of the proof is to match, for each index $i$ of the sums expressing $\dot{h}_{k,m,x}(\theta)$ and $\dot{h}_{k,\infty}(\theta)$, pairs of terms that are close. To be more precise, we match

1. The first terms of $\dot{h}_{k,m,x}(\theta)$ and $\dot{h}_{k,\infty}(\theta)$;
2. For $i$ close to $k$,
$$\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]$$
and
$$\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-\infty:k}] ,$$
and similarly for the corresponding terms conditioned on $Y_{-m+1:k-1}$ and $Y_{-\infty:k-1}$, respectively;
3. For $i$ far from $k$,
$$\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x]$$
and
$$\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k-1}, X_{-m} = x] ,$$
and similarly for the corresponding terms conditioned on $Y_{-\infty:k}$ and $Y_{-\infty:k-1}$, respectively.

We start with the second kind of matches (of which the first terms are a special case). Taking the limit in $m' \to \infty$ in (6.23), we see that

$$\begin{aligned}&\|\,\mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x] - \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-\infty:k}]\| \\ &\qquad\qquad \leq 2 \sup_{x,x'\in\mathsf{X}} \|\phi_\theta(x, x', Y_i)\| \rho^{(i-1)+m} .\end{aligned}$$

This bound remains the same if $k$ is replaced by $k-1$. Obviously, it is small if $i$ is far away from $m$, that is, close to $k$.

For the third kind of matches, we need a total variation bound that works "backwards in time". Such a bound reads

$$\|\mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k}, X_{-m} = x) - \mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k-1}, X_{-m} = x)\|_{\mathrm{TV}} \le \rho^{k-1-i} .$$

The proof of this bound is similar to that of Proposition 61 and uses the time-reversed process. We postpone the proof to the end of this section. We may also let $m \to \infty$ and omit the condition on $X_{-m}$ without affecting the bound. As a result of these bounds, we have

$$\begin{aligned} &\| \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k}, X_{-m} = x] \\ &\qquad - \mathrm{E}_\theta[\phi_\theta(X_{i-1}, X_i, Y_i) \,|\, Y_{-m+1:k-1}, X_{-m} = x]\| \\ &\le \; 2 \sup_{x,x' \in \mathsf{X}} \|\phi_\theta(x, x', Y_i)\| \rho^{k-1-i} , \end{aligned}$$

with the same bound being valid if the conditioning is on $Y_{-\infty:k}$ and $Y_{-\infty:k-1}$, respectively. This bound is small if $i$ is far away from $k$.

Combining these two kinds of bounds and using Minkowski's inequality for the $L^2$-norm, we find that $(\mathrm{E}_\theta \, \|\dot{h}_{k,m,x}(\theta) - \dot{h}_{k,\infty}(\theta)\|^2)^{1/2}$ is bounded by

$$\begin{aligned} &2\rho^{k+m-1} + 2 \times 2 \sum_{i=-m+1}^{k-1} (\rho^{k-i-1} \wedge \rho^{i+m-1}) + 2 \sum_{i=-\infty}^{-m} \rho^{k-i-1} \\ &\le \; 4\frac{\rho^{k+m-1}}{1-\rho} + 4 \sum_{-\infty < i \le (k-m)/2} \rho^{k-i-1} + 4 \sum_{(k-m)/2 \le i < \infty} \rho^{i+m-1} \\ &\le \; 12 \frac{\rho^{(k+m)/2-1}}{1-\rho} \end{aligned}$$

up to the factor $(\mathrm{E}_\theta \sup_{x,x' \in \mathsf{X}} \|\phi_\theta(x, x', Y_i)\|^2)^{1/2}$. The proof is complete. □

We now establish the "backwards in time" uniform forgetting property, which played a key role in the above proof.

**Proposition 127.** *Assume 108, 109, and 110 hold. Then for any integers $i$, $k$, and $m$ such that $m \ge 0$ and $-m < i < k$, any $x_{-m} \in \mathsf{X}$, $y_{-m+1:k} \in \mathsf{Y}^{k+m}$, and $\theta \in \mathcal{U}$,*

$$\|\mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k} = y_{-m+1:k}, X_{-m} = x_{-m}) - \mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k-1} = y_{-m+1:k-1}, X_{-m} = x_{-m})\|_{\mathrm{TV}} \le \rho^{k-1-i} .$$

*Proof.* The cornerstone of the proof is the observation that conditional on $Y_{-m+1:k}$ and $X_{-m}$, the time-reversed process $X$ with indices from $k$ down to $-m$ is a non-homogeneous Markov chain satisfying a uniform mixing condition. We shall indeed use a slight variant of the backward decomposition developed in Section 2.3.2. For any $j = -m+1, \ldots, k-1$, we thus define the backward kernel (cf. (2.39)) by

$$\mathrm{B}_{x_{-m},j}[y_{-m+1:j}](x, f) = \frac{\int \cdots \int \prod_{u=-m+1}^{j} q(x_{u-1}, x_u) g(x_u, y_u) \, \lambda(dx_u) \, f(x_j) q(x_j, x)}{\int \cdots \int \prod_{u=-m+1}^{j} q(x_{u-1}, x_u) g(x_u, y_u) \, \lambda(dx_u) \, q(x_j, x)} \tag{6.24}$$

for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$. For brevity, we do not indicate the dependence of the quantities involved on $\theta$. We note that the integral of the denominator of this display is bounded from below by $(\sigma^-)^{m+j} \prod_{-m+1}^{j} \int g_\theta(x_u, y_u)\, \lambda(dx_u)$, and is hence positive $\mathrm{P}_{\theta_\star}$-a.s. under Assumption 110.

It is trivial that for any $x \in \mathsf{X}$,

$$\int \cdots \int \prod_{u=-m+1}^{j} q(x_{u-1}, x_u) g(x_u, y_u)\, \lambda(dx_u)\, f(x_j) q(x_j, x) = \int \cdots \int \prod_{u=-m+1}^{j} q(x_{u-1}, x_u) g(x_u, y_u)\, \lambda(dx_u)\, q(x_j, x) \mathrm{B}_{x_{-m},j}[y_{-m+1:j}](x, f)\ ,$$

which implies that

$$\mathrm{E}_\theta[f(X_j) \,|\, X_{j+1:k}, Y_{-m+1:k} = y_{-m+1:k}, X_{-m} = x] = \mathrm{B}_{x_{-m},j}[y_{-m+1:j}](X_{j+1}, f)\ .$$

This is the desired Markov property referred to above.

Along the same lines as in the proof of Proposition 64, we can show that the backward kernels satisfy a Doeblin condition,

$$\frac{\sigma^-}{\sigma^+} \nu_{x_{-m},j}[y_{-m+1:j}] \leq \mathrm{B}_{x_{-m},j}[y_{-m+1:j}](x, \cdot) \leq \frac{\sigma^+}{\sigma^-} \nu_{x_{-m},j}[y_{-m+1:j}]\ ,$$

where for any $f \in \mathcal{F}_{\mathrm{b}}(\mathsf{X})$,

$$\nu_{x_{-m},j}[y_{-m+1:j}](f) = \frac{\int \cdots \int \prod_{u=-m+1}^{j} q_\theta(x_{u-1}, x_u) g_\theta(x_u, y_u)\, \lambda(dx_u)\, f(x_j)}{\int \cdots \int \prod_{u=-m+1}^{j} q_\theta(x_{u-1}, x_u) g_\theta(x_u, y_u)\, \lambda(dx_u)}\ .$$

Thus Lemma 51 shows that the Dobrushin coefficient of each backward kernel is bounded by $\rho = 1 - \sigma^-/\sigma^+$.

Finally

$$\begin{aligned}
&\mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k-1} = y_{-m+1:k-1}, X_{-m} = x_{-m}) \\
&\quad = \int \mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k-1} = y_{-m+1:k-1}, X_{-m} = x_{-m}, X_{k-1} = x_{k-1}) \\
&\qquad \times \mathrm{P}_\theta(X_{k-1} \in dx_{k-1} \,|\, Y_{-m+1:k-1} = y_{-m+1:k-1}, X_{-m} = x_{-m})
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k} = y_{-m+1:k}, X_{-m} = x_{-m}) \\
&\quad = \int \mathrm{P}_\theta(X_i \in \cdot \,|\, Y_{-m+1:k-1} = y_{-m+1:k-1}, X_{-m} = x_{-m}, X_{k-1} = x_{k-1}) \\
&\qquad \times \mathrm{P}_\theta(X_{k-1} \in dx_{k-1} \,|\, Y_{-m+1:k} = y_{-m+1:k}, X_{-m} = x_{-m})\ ,
\end{aligned}$$

so that the two distributions on the left-hand sides can be considered as the result of running the above-described reversed conditional Markov chain from index $k-1$ down to index $i$, using two different initial conditions. Therefore, by Proposition 48, they differ by at most $\rho^{k-1-i}$ in total variation distance. The proof is complete. □

### 6.5.3 Weak Convergence of the Normalized Score

We now return to the question of a weak limit of the normalized score $n^{-1/2}\sum_{k=0}^{n}\dot{h}_{k,0,x_0}(\theta_\star)$. Using Lemma 126 and Minkowski's inequality, we see that

$$\left[\mathrm{E}_{\theta_\star}\left\|n^{-1/2}\sum_{k=0}^{n}(\dot{h}_{k,0,x_0}(\theta_\star)-\dot{h}_{k,\infty}(\theta_\star))\right\|^2\right]^{1/2}$$
$$\leq \quad n^{-1/2}\sum_{k=0}^{n}\left[\mathrm{E}_{\theta_\star}\,\|\dot{h}_{k,0,x_0}(\theta_\star)-\dot{h}_{k,\infty}(\theta_\star)\|^2\right]^{1/2}\to 0 \quad \text{as } n\to\infty\,,$$

whence the limiting behavior of the normalized score agrees with that of $n^{-1/2}\sum_{k=0}^{n}\dot{h}_{k,\infty}(\theta_\star)$. Now define the filtration $\mathcal{F}$ by $\mathcal{F}_k=\sigma(Y_i,\,-\infty<i\leq k)$ for all integer $k$. By conditional dominated convergence,

$$\mathrm{E}_{\theta_\star}\left[\sum_{i=-\infty}^{k-1}(\mathrm{E}_{\theta_\star}[\phi_{\theta_\star}(X_{i-1},X_i,Y_i)\,|\,Y_{-\infty:k}]\right.$$
$$\left.-\;\mathrm{E}_{\theta_\star}[\phi_{\theta_\star}(X_{i-1},X_i,Y_i)\,|\,Y_{-\infty:k-1}])\,|\,\mathcal{F}_{k-1}\right]=0\,,$$

and Assumption 124 implies that

$$\mathrm{E}_{\theta_\star}[\phi_{\theta_\star}(X_{k-1},X_k,Y_k)\,|\,Y_{-\infty:k-1}]$$
$$=\mathrm{E}_{\theta_\star}[\mathrm{E}_{\theta_\star}[\phi_{\theta_\star}(X_{k-1},X_k,Y_k)\,|\,Y_{-\infty:k-1},X_{k-1}]\,|\,\mathcal{F}_{k-1}]=0\,.$$

It is also immediate that $h_{k,\infty}(\theta_\star)$ is $\mathcal{F}_k$-measurable. Hence the sequence $\{h_{k,\infty}(\theta_\star)\}_{k\geq 0}$ is a $\mathrm{P}_{\theta_\star}$-martingale increment sequence with respect to the filtration $\{\mathcal{F}_k\}_{k\geq 0}$ in $L^2(\mathrm{P}_{\theta_\star})$. Moreover, this sequence is stationary because $\{Y_k\}_{-\infty<k<\infty}$ is. Any stationary martingale increment sequence in $L^2(\mathrm{P}_{\theta_\star})$ satisfies a CLT (Durrett, 1996, p. 418), that is, $n^{-1/2}\sum_0^n\dot{h}_{k,\infty}(\theta_\star)\to\mathrm{N}(0,\mathcal{J}(\theta_\star))$ $\mathrm{P}_{\theta_\star}$-weakly, where

$$\mathcal{J}(\theta_\star)\stackrel{\mathrm{def}}{=}\mathrm{E}_{\theta_\star}[\dot{h}_{1,\infty}(\theta_\star)\dot{h}_{1,\infty}^t(\theta_\star)] \tag{6.25}$$

is the limiting Fisher information.

Because the normalized score function has the same limiting behavior, the following result is immediate.

**Theorem 128.** *Under Assumptions 108, 109, 110, and 124,*

$$n^{-1/2}\nabla_\theta\ell_{x_0,n}(\theta_\star)\to\mathrm{N}(0,\mathcal{J}(\theta_\star))\quad \mathrm{P}_{\theta_\star}\text{-}\textit{weakly}$$

*for all $x_0\in\mathsf{X}$, where $\mathcal{J}(\theta_\star)$ is the limiting Fisher information as defined above.*

We remark that above, we have normalized sums with indices from 0 to $n$, that is, with $n+1$ terms, by $n^{1/2}$ rather than by $(n+1)^{1/2}$. This of course does not affect the asymptotics. However, if $\mathcal{J}(\theta_\star)$ is estimated for the purpose of making a confidence interval for instance, then one may well normalize it using the number $n+1$ of observed data.

### 6.5.4 Convergence of the Normalized Observed Information

We shall now very briefly discuss the asymptotics of the observed information matrix, $-\nabla_\theta^2\ell_{x_0,n}(\theta)$. To handle this matrix, one can employ the so-called *missing information principle* (see Section 5.1.3 and (5.29)). Because the complete information matrix, just as the complete score, has a relatively simple form, this principle

allows us to study the asymptotics of the observed information in a fashion similar to what was done above for the score function. The analysis becomes more difficult however, as covariance terms, arising from the conditional variance of the complete score, also need to be accounted for. In addition, we need the convergence to be uniform in a certain sense. We state the following theorem, whose proof can be found in Douc *et al.* (2004).

**Theorem 129.** *Under Assumptions 108, 109, 110, and 124,*

$$\lim_{\delta\to 0}\lim_{n\to\infty}\sup_{|\theta-\theta_\star|\leq\delta}\|(-n^{-1}\nabla_\theta^2\ell_{x_0,n}(\theta))-\mathcal{J}(\theta_\star)\|=0 \quad \mathrm{P}_{\theta_\star}\text{-a.s.}$$

*for all* $x_0 \in \mathsf{X}$.

### 6.5.5 Asymptotics of the Maximum Likelihood Estimator

The general arguments in Section 6.1 and the theorems above prove the following result.

**Theorem 130.** *Assume 108, 109, 110, 114, and 124, and that* $\theta_\star$ *is identifiable, that is,* $\theta$ *is equivalent to* $\theta_\star$ *only if* $\theta=\theta_\star$ *(possibly up to a permutation of states if* $\mathsf{X}$ *is finite). Then the following hold true.*

*(i) The MLE* $\widehat{\theta}_n=\widehat{\theta}_{x_0,n}$ *is strongly consistent:* $\widehat{\theta}_n\to\theta_\star$ $\mathrm{P}_{\theta_\star}$*-a.s. as* $n\to\infty$.

*(ii) If the Fisher information matrix* $\mathcal{J}(\theta_\star)$ *defined above is non-singular and* $\theta_\star$ *is an interior point of* $\Theta$*, then the MLE is asymptotically normal:*

$$n^{1/2}(\widehat{\theta}_n-\theta_\star)\to\mathrm{N}(0,\mathcal{J}(\theta_\star)^{-1}) \quad \mathrm{P}_{\theta_\star}\text{-weakly as } n\to\infty$$

*for all* $x_0 \in \mathsf{X}$.

*(iii) The normalized observed information at the MLE is a strongly consistent estimator of* $\mathcal{J}(\theta_\star)$*:*

$$-n^{-1}\nabla_\theta^2\ell_{x_0,n}(\widehat{\theta}_n)\to\mathcal{J}(\theta_\star) \quad \mathrm{P}_{\theta_\star}\text{-a.s. as } n\to\infty.$$

As indicated above, the MLE $\widehat{\theta}_n$ depends on the initial state $x_0$, but that dependence will generally not be included in the notation.

The last part of the result is important, as is says that confidence intervals or regions and hypothesis tests based on the estimate $-(n+1)^{-1}\nabla_\theta^2\ell_{x_0,n}(\widehat{\theta}_n)$ of $\mathcal{J}(\theta_\star)$ will asymptotically be of correct size. In general, there is no closed-form expression for $\mathcal{J}(\theta_\star)$, so that it needs to be estimated in one way or another. The observed information is obviously one way to do that, while another one is to simulate data $Y_{1:N}^*$ from the HMM, using the MLE, and then computing $-(N+1)^{-1}\nabla_\theta^2\ell_{x_0,N}(\widehat{\theta}_n)$ for this set of simulated data and some $x_0$. An advantage of this approach is that $N$ can be chosen arbitrarily large. Yet another approach, motivated by (6.25), is to estimate the Fisher information by the empirical covariance matrix of the conditional scores of (6.19) at the MLE, that is, by $(n+1)^{-1}\sum_0^n[S_{k|k-1}(\widehat{\theta}_n)-\bar{S}(\widehat{\theta}_n)][S_{k|k-1}(\widehat{\theta}_n)-\bar{S}(\widehat{\theta}_n)]^t$ with $S_{k|k-1}(\theta)=\nabla_\theta\log\int g_\theta(x,Y_k)\,\phi_{x_0,k|k-1}[Y_{0:k-1}](dx\,;\theta)$ and $\bar{S}(\theta)=(n+1)^{-1}\sum_0^n S_{k|k-1}(\theta)$. This estimate can of course also be computed from estimated data, then using an arbitrary sample size. The conditional scores may be computed as $S_{k|k-1}(\theta)=\nabla_\theta\ell_{x_0,k}(\theta)-\nabla_\theta\ell_{x_0,k-1}(\theta)$, where the scores are computed using any of the methods of Section 5.2.3.

## 6.6 Applications to Likelihood-based Tests

The asymptotic properties of the score function and observed information have immediate implications for the asymptotics of the MLE, as has been described in previous sections. However, there are also other conclusions that can be drawn from these convergence results.

One such application is the validity of some classical procedures for testing whether $\theta_\star$ lies in some subset, $\Theta_0$ say, of the parameter space $\Theta$. Suppose that $\Theta_0$ is an $(d_\theta - s)$-dimensional subset that may be expressed in terms of constraints $R_i(\theta) = 0$, $i = 1, 2, \ldots, s$, and that there is an equivalent formulation $\theta_i = b_i(\gamma)$, $i = 1, 2, \ldots, d_\theta$, where $\gamma$ is the "constrained parameter" lying in a subset $\Gamma$ of $\mathbb{R}^{d_\theta - s}$. We also let $\gamma_\star$ be a point such that $\theta_\star = b(\gamma_\star)$. Each function $R_i$ and $b_i$ is assumed to be continuously differentiable and such that the matrices

$$C_\theta = \left( \frac{\partial R_i}{\partial \theta_j} \right)_{s \times d_\theta} \qquad \text{and} \qquad D_\gamma = \left( \frac{\partial b_i}{\partial \gamma_j} \right)_{d_\theta \times (d_\theta - s)}$$

have full rank ($s$ and $d_\theta - s$ respectively) in a neighborhood of $\theta_\star$ and $\gamma_\star$, respectively.

Perhaps the simplest example is when we want to test a simple (point) null hypothesis $\theta_\star = \theta_0$ versus the alternative $\theta_\star \neq \theta_0$. Then, we take $R_i(\theta) = \theta_i - \theta_{0i}$ and $b_i(\gamma) = \theta_{i0}$ for $i = 1, 2, \ldots, d_\theta$. In this case, $\gamma$ is void as $s = d_\theta$ and hence $d_\theta - s = 0$. Furthermore, $C$ is the identity matrix and $D$ is void.

Now suppose that we want to test the equality $\theta_i = \theta_{i0}$ only for $i$ in a subset $K$ of the $d_\theta$ coordinates of $\theta$, where $K$ has cardinality $s$. The constraints we employ are then $R_i(\theta) = \theta_i - \theta_{0i}$ for $i \in K$; furthermore, $\gamma$ comprises $\theta_i$ for $i \notin K$ and, using the $d_\theta - s$ indices not in $K$ for $\gamma$, $b_i(\gamma) = \theta_{0i}$ for $i \in K$ and $b_i(\gamma) = \gamma_i$ otherwise. Again it is easy to check that $C$ and $D$ are constant and of full rank.

**Example 131** (Normal HMM)**.** A slightly more involved example concerns the Gaussian hidden Markov model with finite state space $\{1, 2, \ldots, r\}$ and conditional distributions $Y_k | X_k = i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Suppose that we want to test for equality of all of the $r$ component-wise conditional variances $\sigma_i^2$: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_r^2$. Then, the $R$-functions are for instance $\sigma_i^2 - \sigma_r^2$ for $i = 1, 2, \ldots, r-1$. The parameter $\gamma$ is obtained by removing from $\theta$ all $\sigma_i^2$ and then adding a common conditional variance $\sigma^2$; those $b$-functions referring to any of the $\sigma_i^2$ evaluate to $\sigma^2$. The matrices $C$ and $D$ are again constant and of full rank.

A further application, to test the structure of conditional covariance matrices in a conditionally Gaussian HMM with multivariate output, can be found in Giudici *et al.* (2000).

There are many different tests available for testing the null hypothesis $\theta_\star \in \Theta_0$ versus the alternative $\theta_\star \in \Theta \setminus \Theta_0$. One is the generalized likelihood ratio test, which uses the test statistic

$$\lambda_n = 2 \left\{ \sup_{\theta \in \Theta} \ell_{x_0,n}(\theta) - \sup_{\theta \in \Theta_0} \ell_{x_0,n}(\theta) \right\} .$$

Another one is the Wald test, which uses the test statistic

$$W_n = n R(\widehat{\theta}_n)^t [C_{\widehat{\theta}_n} \mathcal{J}_n(\widehat{\theta}_n)^{-1} C_{\widehat{\theta}_n}^t]^{-1} R(\widehat{\theta}_n) ,$$

where $R(\theta)$ is the $s \times 1$ vector of $R$-functions evaluated at $\theta$, and $\mathcal{J}_n(\theta) = -n^{-1} \nabla_\theta^2 \ell_{x_0,n}(\theta)$ is the observed information evaluated at $\theta$. Yet another test is based on the Rao statistic, defined as

$$V_n = n^{-1} S_n(\widehat{\theta}_n^0) \mathcal{J}_n(\widehat{\theta}_n^0)^{-1} S_n(\widehat{\theta}_n^0)^t ,$$

where $\widehat{\theta}_n^0$ is the MLE over $\Theta_0$, that is, the point where $\ell_{x_0,n}(\theta)$ is maximized subject to the constraint $R_i(\theta) = 0$, $1 \leq i \leq s$, and $S_n(\theta) = \nabla_\theta \ell_{x_0,n}(\theta)$ is the score function at $\theta$. This test is also known under the names *efficient score test* and *Lagrange multiplier test*. The Wald and Rao test statistics are usually defined using the true Fisher information $\mathcal{J}(\theta)$ rather than the observed one, but as $\mathcal{J}(\theta)$ is generally infeasible to compute for HMMs, we replace it by the observed counterpart.

Statistical theory for i.i.d. data suggests that the likelihood ratio, Wald and Rao test statistics should all converge weakly to a $\chi^2$ distribution with $s$ degrees of freedom provided $\theta_\star \in \Theta_0$ holds true, so that an approximate $p$-value of the test of this null hypothesis can be computed by evaluating the complementary distribution function of the $\chi_s^2$ distribution at the point $\lambda_n$, $W_n$, or $V_n$, whichever is preferred. We now state formally that this procedure is indeed correct.

**Theorem 132.** *Assume 108, 109, 110, 114, and 124 as well as the conditions stated on the functions $R_i$ and $b_i$ above. Also assume that $\theta_\star$ is identifiable, that is, $\theta$ is equivalent to $\theta_\star$ only if $\theta = \theta_\star$ (possibly up to a permutation of states if $\mathsf{X}$ is finite), that $\mathcal{J}(\theta_\star)$ is non-singular, and that $\theta_\star$ and $\gamma_\star$ are interior points of $\Theta$ and $\Gamma$, respectively. Then if $\theta_\star \in \Theta_0$ holds true, each of the test statistics $\lambda_n$, $W_n$, and $V_n$ converges $\mathrm{P}_{\theta_\star}$-weakly to the $\chi_s^2$ distribution as $n \to \infty$.*

The proof of this result follows, for instance, Serfling (1980, Section 4.4). The important observation is that the validity of the proof does not hinge on independence of the data but on asymptotic properties of the score function and the observed information, properties that have been established for HMMs in this chapter.

It is important to realize that a key assumption for Theorem 132 to hold is that $\theta_\star$ is identifiable, so that $\widehat{\theta}_n$ converges to a unique point $\theta_\star$. As a result, the theorem does not apply to the problem of testing the number of components of a finite state HMM. In the normal HMM for instance, with $Y_k|X_k = i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, one can indeed effectively remove one component by invoking the constraints $\mu_1 - \mu_2 = 0$ and $\sigma_1^2 - \sigma_2^2 = 0$, say. In this way, within $\Theta_0$, components 1 and 2 collapse into a single one. However, any $\theta \in \Theta_0$ is then non-identifiable as the transition probabilities $q_{12}$ and $q_{21}$, among others, can be chosen arbitrarily without changing the dynamics of the model.

# Part III

# Background and Complements

# Chapter 7

# Elements of Markov Chain Theory

## 7.1 Chains on Countable State Spaces

We review the key elements of the mathematical theory developed for studying the limiting behavior of Markov chains. In this first section, we restrict ourselves to the case where the state space $\mathsf{X}$ is countable, which is conceptually simpler. On our way, we will also meet a number of important concepts to be used in the next section when dealing with Markov chains on general state spaces.

### 7.1.1 Irreducibility

Let $\{X_k\}_{k\geq 0}$ be a Markov chain on a countable state space $\mathsf{X}$ with transition matrix $Q$. For any $x \in X$, we define the first hitting time $\sigma_x$ on $x$ and the return time $\tau_x$ to $x$ respectively as

$$\sigma_x = \inf\{n \geq 0 : X_n = x\}\,, \tag{7.1}$$
$$\tau_x = \inf\{n \geq 1 : X_n = x\}\,, \tag{7.2}$$

where, by convention, $\inf \emptyset = +\infty$. The successive hitting times $\sigma_x^{(n)}$ and return times $\tau_x^{(n)}$, $n \geq 0$, are defined inductively by

$$\sigma_x^{(0)} = 0,\ \sigma_x^{(1)} = \sigma_x,\ \sigma_x^{(n+1)} = \inf\{k > \sigma_x^{(n)} :\ X_k = x\}\,,$$
$$\tau_x^{(0)} = 0,\ \tau_x^{(1)} = \tau_x,\ \tau_x^{(n+1)} = \inf\{k > \tau_x^{(n)} :\ X_k = x\}\,.$$

For two states $x$ and $y$, we say that state $x$ *leads to* state $y$, which we write $x \to y$, if $\mathrm{P}_x(\sigma_y < \infty) > 0$. In words, $x$ leads to $y$ if the state $y$ can be reached from $x$. An alternative, equivalent definition is that there exists some integer $n \geq 0$ such that the $n$-step transition probability $Q^n(x, y) > 0$. If both $x$ leads to $y$ and $y$ leads to $x$, then we say that the $x$ and $y$ *communicate*, which we write $x \leftrightarrow y$.

**Theorem 133.** *The relation "$\leftrightarrow$" is an equivalence relation on* $\mathsf{X}$.

*Proof.* We need to prove that the relation $\leftrightarrow$ is reflexive, symmetric, and transitive. The first two properties are immediate because, by definition, for all $x, y \in \mathsf{X}$, $x \leftrightarrow x$ (reflexivity), and $x \leftrightarrow y$ if and only if $y \leftrightarrow x$ (symmetry).

For any pairwise distinct $x, y, z \in \mathsf{X}$, $\{\sigma_y + \sigma_z \circ \theta^{\sigma_y} < \infty\} \subset \{\sigma_z < \infty\}$ (if the chain reaches $y$ at some time and later $z$, it certainly reaches $z$). The strong Markov

property (Theorem 6) implies that

$$\begin{aligned}\mathrm{P}_x(\sigma_z < \infty) &\geq \mathrm{P}_x(\sigma_y + \sigma_z \circ \theta^{\sigma_y} < \infty) = \mathrm{E}_x[\mathbb{1}_{\{\sigma_y<\infty\}}\mathbb{1}_{\{\sigma_z<\infty\}} \circ \theta^{\sigma_y}] \\ &= \mathrm{E}_x[\mathbb{1}_{\{\sigma_y<\infty\}}\, \mathrm{P}_{X_{\sigma_y}}(\sigma_z < \infty)] = \mathrm{P}_x(\sigma_y < \infty)\,\mathrm{P}_y(\sigma_z < \infty)\,.\end{aligned}$$

In words, if the chain can reach $y$ from $x$ and $z$ from $y$, it can reach $z$ from $x$ by going through $y$. Hence if $x \to y$ and $y \to z$, then $x \to z$ (transitivity). □

For $x \in \mathsf{X}$, we denote the equivalence class of $x$ with respect to the relation "$\leftrightarrow$" by $C(x)$. Because "$\leftrightarrow$" is an equivalence relation, there exists a collection $\{x_i\}$ of states, which may be finite or infinite, such that the classes $\{C(x_i)\}$ form a partition of the state space $\mathsf{X}$.

**Definition 134** (Irreducibility)**.** *If $C(x) = \mathsf{X}$ for some $x \in \mathsf{X}$ (and then for all $x \in \mathsf{X}$), the Markov chain is called* irreducible*.*

### 7.1.2 Recurrence and Transience

When a state is visited by the Markov chain, it is natural to ask how often the state is visited in the long-run. Define the *occupation time* of the state $x$ as

$$\eta_x \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} \mathbb{1}_x(X_n) = \sum_{j=1}^{\infty} \mathbb{1}_{\{\sigma_x^{(j)}<\infty\}}\,.$$

If the expected number of visits to $x$ starting from $x$ is finite, that is, if $\mathrm{E}_x[\eta_x] < \infty$, then the state $x$ is called *transient*. Otherwise, if $\mathrm{E}_x[\eta_x] = \infty$, $x$ is said to be *recurrent*. When $\mathsf{X}$ is countable, the recurrence or transience of a state $x$ can be expressed in terms of the probability $\mathrm{P}_x(\tau_x < \infty)$ that the chain started in $x$ ever returns to $x$.

**Proposition 135.** *For any $x \in \mathsf{X}$ the following hold true,*

*(i) If $x$ is recurrent, then $\mathrm{P}_x(\eta_x = \infty) = 1$ and $\mathrm{P}_x(\tau_x < \infty) = 1$.*

*(ii) If $x$ is transient, then $\mathrm{P}_x(\eta_x < \infty) = 1$ and $\mathrm{P}_x(\tau_x < \infty) < 1$.*

*(iii) $\mathrm{E}_x[\eta_x] = 1/[1 - \mathrm{P}_x(\tau_x < \infty)]$, with $1/0 = \infty$.*

*Proof.* By construction,

$$\mathrm{E}_x[\eta_x] = \sum_{k=1}^{\infty} \mathrm{P}_x(\eta_x \geq k) = \sum_{k=1}^{\infty} \mathrm{P}_x(\sigma_x^{(k)} < \infty)\,.$$

Applying strong Markov property (Theorem 6) for $n > 1$, we obtain

$$\begin{aligned}\mathrm{P}_x(\sigma_x^{(n)} < \infty) &= \mathrm{P}_x(\sigma_x^{(n-1)} < \infty,\, \tau_x \circ \theta^{\sigma_x^{(n-1)}} < \infty) \\ &= \mathrm{E}_x[\mathbb{1}_{\{\sigma_x^{(n-1)}<\infty\}}\, \mathrm{P}_{X_{\sigma_x^{(n-1)}}}(\tau_x < \infty)]\,.\end{aligned}$$

If $\sigma_x^{(n-1)} < \infty$, then $X_{\sigma_x^{(n-1)}} = x$ $\mathrm{P}_x$-a.s., so that

$$\mathrm{P}_x(\sigma_x^{(n)} < \infty) = \mathrm{P}_x(\tau_x < \infty)\,\mathrm{P}_x(\sigma_x^{(n-1)} < \infty)\,.$$

By definition $\mathrm{P}_x(\sigma_x < \infty) = 1$, whence $\mathrm{P}_x(\sigma_x^{(n)} < \infty) = [\mathrm{P}_x(\tau_x < \infty)]^{n-1}$ and

$$\mathrm{E}_x[\eta_x] = \sum_{n=1}^{\infty} [\mathrm{P}_x(\tau_x < \infty)]^{n-1}\,.$$

This proves part (iii).

Now assume $x$ is recurrent. Then by definition $\mathrm{E}_x[\eta_x] = \infty$, and hence $\mathrm{P}_x(\tau_x < \infty) = 1$ and $\mathrm{P}_x(\tau_x^{(n)} < \infty) = 1$ for all $n \geq 1$. Thus $\eta_x = \infty$ $\mathrm{P}_x$-a.s.

If $x$ is transient then $\mathrm{E}_x[\eta_x] < \infty$, which implies $\mathrm{P}_x(\tau_x < \infty) < 1$. □

For a recurrent state $x$, the occupation time of $x$ is infinite with probability one under $\mathrm{P}_x$; essentially, once the chain started from $x$ returns to $x$ with probability one, it returns a second time with probability one, and so on. Thus the occupation time of a state has a remarkable property, not shared by all random variables: if the expectation of the occupation time is infinite, then the actual number of returns is infinite with probability one. The mean of the occupation time of a state obeys the so-called maximum principle.

**Proposition 136.** *For all $x$ and $y$ in $\mathsf{X}$,*

$$\mathrm{E}_x[\eta_y] = \mathrm{P}_x(\sigma_y < \infty)\, \mathrm{E}_y[\eta_y] \,, \tag{7.3}$$

*with the convention $0 \times \infty = 0$.*

*Proof.* It follows from the definition that $\eta_y \mathbb{1}_{\{\sigma_y = \infty\}} = 0$ and $\eta_y \mathbb{1}_{\{\sigma_y < \infty\}} = \eta_y \circ \theta^{\sigma_y} \mathbb{1}_{\{\sigma_y < \infty\}}$. Thus, applying the strong Markov property,

$$\begin{aligned} \mathrm{E}_x[\eta_y] &= \mathrm{E}_x[\mathbb{1}_{\{\sigma_y < \infty\}} \eta_y] = \mathrm{E}_x[\mathbb{1}_{\{\sigma_y < \infty\}}\, \eta_y \circ \theta^{\sigma_y}] \\ &= \mathrm{E}_x[\mathbb{1}_{\{\sigma_y < \infty\}}\, \mathrm{E}_{X_{\sigma_y}}[\eta_y]] = \mathrm{P}_x(\sigma_y < \infty)\, \mathrm{E}_y[\eta_y] \,. \end{aligned}$$

□

**Corollary 137.** *If $\mathrm{E}_x[\eta_y] = \infty$ for some $x$, then $y$ is recurrent. If $\mathsf{X}$ is finite, then there exists at least one recurrent state.*

*Proof.* By Proposition 136, $\mathrm{E}_y[\eta_y] \geq \mathrm{E}_x[\eta_y]$, so that $\mathrm{E}_x[\eta_y] = \infty$ implies that $\mathrm{E}_y[\eta_y] = \infty$, that is, $y$ is recurrent.

Next, obviously $\sum_{y \in \mathsf{X}} \eta_y = \infty$ and thus for all $x \in \mathsf{X}$, $\sum_{y \in \mathsf{X}} \mathrm{E}_x[\eta_y] = \infty$. Hence if $\mathsf{X}$ is finite, given $x \in \mathsf{X}$ there necessarily exists at least one $y \in \mathsf{X}$ such that $\mathrm{E}_x[\eta_y] = \infty$, which implies that $y$ is recurrent. □

Our next result shows that a recurrent state can only lead to another recurrent state.

**Proposition 138.** *Let $x$ be a recurrent state. Then for $y \in \mathsf{X}$, either of the following two statements holds true.*

*(i) $x$ leads to $y$, $\mathrm{E}_x[\eta_y] = \infty$, $y$ is recurrent and leads to $x$, and $\mathrm{P}_x(\tau_y < \infty) = \mathrm{P}_y(\tau_x < \infty) = 1$;*

*(ii) $x$ does not lead to $y$ and $\mathrm{E}_x[\eta_y] = 0$.*

*Proof.* Assume that $x$ leads to $y$. Then there exists an integer $k$ such that $Q^k(x, y) > 0$. Applying the Chapman-Kolmogorov equations, we obtain $Q^{n+k}(x, y) \geq Q^n(x, x) Q^k(x, y)$ for all $n$. Hence

$$\mathrm{E}_x[\eta_y] \geq \sum_{n=1}^{\infty} Q^{n+k}(x, y) \geq \sum_{n=1}^{\infty} Q^n(x, x) Q^k(x, y) = \mathrm{E}_x[\eta_x] Q^k(x, y) = \infty \,.$$

Thus $y$ is also recurrent by Corollary 137. Because $x$ is recurrent, the strong Markov property implies that

$$\begin{aligned} 0 = \mathrm{P}_x(\tau_x = \infty) &\geq \mathrm{P}_x(\tau_y < \infty,\, \tau_x = \infty) \\ &= \mathrm{P}_x(\tau_y < \infty,\, \tau_x \circ \theta^{\tau_y} = \infty) = \mathrm{P}_x(\tau_y < \infty)\, \mathrm{P}_y(\tau_x = \infty) \,. \end{aligned}$$

Because $x$ leads to $y$, $\mathrm{P}_x(\tau_y < \infty) > 0$, whence $\mathrm{P}_y(\tau_x = \infty) = 0$. Thus $y$ leads to $x$ and moreover $\mathrm{P}_y(\tau_x < \infty) = 1$. By symmetry, $\mathrm{P}_x(\tau_y < \infty) = 1$.

If $x$ does not lead to $y$ then Proposition 136 shows that $\mathrm{E}_x[\eta_y] = 0$. □

For a recurrent state $x$, the equivalence class $C(x)$ (with respect to the relation of communication defined in Section 7.1.1) may thus be equivalently defined as

$$C(x) = \{y \in \mathsf{X} : \mathrm{E}_x[\eta_y] = \infty\} = \{y \in \mathsf{X} : \mathrm{P}_x(\tau_y < \infty) = 1\} \,. \tag{7.4}$$

If $y \notin C(x)$, then $\mathrm{P}_x(\eta_y = 0) = 1$, which implies that $\mathrm{P}_x(X_n \in C(x) \text{ for all } n \geq 0) = 1$. In words, the chain started from the recurrent state $x$ forever stays in $C(x)$ and visits each state of $C(x)$ infinitely many times.

The behavior of a Markov chain can thus be described as follows. If a chain is not irreducible, there may exist several equivalence classes of communication. Some of them contain only transient states, and some contain only recurrent states. The latter are then called recurrence classes. If a chain starts from a recurrent state, then it remains in its recurrence class forever. If it starts from a transient state, then either it stays in the class of transient states forever, which implies that there exist infinitely many transient states, or it reaches a recurrent state and then remains in its recurrence class forever.

In contrast, if the chain is irreducible, then all the states are either transient or recurrent. This is called the *solidarity property* of an irreducible chain. We now summarize the previous results.

**Theorem 139.** *Consider an irreducible Markov chain on a countable state space* $\mathsf{X}$*. Then every state is either transient, and the chain is called* transient*, or every state is recurrent, and the chain is called* recurrent*. Moreover, either of the following two statements holds true for all* $x$ *and* $y$ *in* $\mathsf{X}$*.*

*(i)* $\mathrm{P}_x(\tau_y < \infty) = 1$*,* $\mathrm{E}_x[\eta_y] = \infty$ *and the chain is recurrent.*

*(ii)* $\mathrm{P}_x(\tau_x < \infty) < 1$*,* $\mathrm{E}_x[\eta_y] < \infty$ *and the chain is transient.*

**Remark 140.** Note that in the transient case, we do not necessarily have $\mathrm{P}_x(\tau_y < \infty) < 1$ for all $x$ and $y$ in $\mathsf{X}$. For instance, if $Q$ is a transition matrix on $\mathbb{N}$ such that $Q(n, n+1) = 1$ for all $n$, then $\mathrm{P}_k(\tau_n < \infty) = 1$ for all $k < n$. Nevertheless all states are obviously transient because $X_n = X_0 + n$.

### 7.1.3 Invariant Measures and Stationarity

For many purposes, we might want the marginal distribution of $\{X_k\}$ not to depend on $k$. If this is the case, then by the Markov property it follows that the finite-dimensional distributions of $\{X_k\}$ are invariant under translation in time, and $\{X_k\}$ is thus a stationary process. Such considerations lead us to invariant distributions. A non-negative vector $\{\pi(x)\}_{x \in \mathsf{X}}$ with the property

$$\pi(y) = \sum_{x \in \mathsf{X}} \pi(x) Q(x, y) \,, \qquad y \in \mathsf{X} \,,$$

will be called *invariant*. If the invariant vector $\pi$ is summable, then we assume it is a probability distribution, that is, it sums to one. Such distributions are also called *stationary distributions* or *stationary probability measures*. The key result concerning the existence of invariant vectors is the following.

**Theorem 141.** *Consider an irreducible recurrent Markov chain* $\{X_k\}_{k \geq 0}$ *on a countable state space* $\mathsf{X}$*. Then there exists a unique (up to a scaling factor) invariant*

*measure $\pi$. Moreover $0 < \pi(x) < \infty$ for all $x \in \mathsf{X}$. This measure is summable if and only if there exists a state $x$ such that*

$$\mathrm{E}_x[\tau_x] < \infty \,. \tag{7.5}$$

*In this case, $\mathrm{E}_y[\tau_y] < \infty$ for all $y \in \mathsf{X}$ and the unique invariant probability measure is given by*

$$\pi(x) = 1/\,\mathrm{E}_x[\tau_x] \,, \qquad x \in \mathsf{X} \,. \tag{7.6}$$

*Proof.* Let $Q$ be the transition matrix of the chain. Pick an arbitrary state $x \in \mathsf{X}$ and define the measure $\lambda_x$ by

$$\lambda_x(y) = \mathrm{E}_x \left[ \sum_{k=0}^{\tau_x - 1} \mathbb{1}_y(X_k) \right] = \mathrm{E}_x \left[ \sum_{k=1}^{\tau_x} \mathbb{1}_y(X_k) \right] \,. \tag{7.7}$$

That is, $\lambda_x(y)$ is the expected number of visits to the state $y$ before the first return to $x$, given that the chain starts in $x$. Let $f$ be a non-negative function on $\mathsf{X}$. Then

$$\lambda_x(f) = \mathrm{E}_x \left[ \sum_{k=0}^{\tau_x - 1} f(X_k) \right] = \sum_{k=0}^{\infty} \mathrm{E}_x \left[ \mathbb{1}_{\{\tau_x > k\}} f(X_k) \right] \,.$$

Using this identity and the fact that $Qf(X_k) = \mathrm{E}_x[f(X_{k+1}) \,|\, \mathcal{F}_k^X]$ $\mathrm{P}_x$-a.s. for all $k \geq 1$, we find that

$$\begin{aligned} \lambda_x(Qf) &= \sum_{k=0}^{\infty} \mathrm{E}_x[\mathbb{1}_{\{\tau_x > k\}}\, Qf(X_k)] = \sum_{k=0}^{\infty} \mathrm{E}_x\{\mathbb{1}_{\{\tau_x > k\}}\, \mathrm{E}_x[f(X_{k+1}) \,|\, \mathcal{F}_k^X]\} \\ &= \sum_{k=0}^{\infty} \mathrm{E}_x[\mathbb{1}_{\{\tau_x > k\}} f(X_{k+1})] = \mathrm{E}_x \left[ \sum_{k=1}^{\tau_x} f(X_k) \right] \,, \end{aligned}$$

showing that $\lambda_x(Qf) = \lambda_x(f) - f(x) + \mathrm{E}_x[f(X_{\tau_x})] = \lambda_x(f)$. Because $f$ was arbitrary, we see that $\lambda_x Q = \lambda_x$; the measure $\lambda_x$ is invariant. For any other state $y$, the chain may reach $y$ before returning to $x$ when starting in $x$, as it is irreducible. This proves that $\lambda_x(y) > 0$. Moreover, again by irreducibility, we can pick an $m > 0$ such that $Q^m(y, x) > 0$. By invariance $\lambda_x(x) = \sum_{z \in \mathsf{X}} \lambda_x(z) Q^m(z, x) \geq \lambda_x(y) Q^m(y, x)$, and as $\lambda_x(x) = 1$, we see that $\lambda_x(y) < \infty$

We now prove that the invariant measure is unique up to a scaling factor. The first step consists in proving that if $\pi$ is an invariant measure such that $\pi(x) = 1$, then $\pi \geq \lambda_x$. It suffices to show that, for any $y \in \mathsf{X}$ and any integer $n$,

$$\pi(y) \geq \sum_{k=1}^{n} \mathrm{E}_x[\mathbb{1}_y(X_k) \mathbb{1}_{\{\tau_x \geq k\}}] \,. \tag{7.8}$$

The proof is by induction. The inequality is immediate for $n = 1$. Assume that (7.8) holds for some $n \geq 1$. Then

$$\begin{aligned} \pi(y) &= Q(x, y) + \sum_{z \neq x} \pi(z) Q(z, y) \\ &\geq Q(x, y) + \sum_{k=1}^{n} \mathrm{E}_x[Q(X_k, y) \mathbb{1}_{\{x\}^c}(X_k) \mathbb{1}_{\{\tau_x \geq k\}}] \\ &\geq Q(x, y) + \sum_{k=1}^{n} \mathrm{E}_x[\mathbb{1}_y(X_{k+1}) \mathbb{1}_{\{\tau_x \geq k+1\}}] \\ &= \sum_{k=1}^{n+1} \mathrm{E}_x[\mathbb{1}_{\{y\}}(X_k) \mathbb{1}_{\{\tau_x \geq k\}}] \,, \end{aligned}$$

showing the induction. We will now show that $\pi = \lambda_x$. The proof is by contradiction. Assume that $\pi(z) > \lambda_x(z)$ for some $z \in \mathsf{X}$. Then

$$1 = \pi(x) = \pi Q(x) = \sum_{z \in \mathsf{X}} \pi(z) Q(z,x) > \sum_{z \in \mathsf{X}} \lambda_x(z) Q(z,x) = \lambda_x(x) = 1 \,,$$

which cannot be true.

The measure $\lambda_x$ is summable if and only if

$$\infty > \sum_{y \in \mathsf{X}} \lambda_x(y) = \sum_{y \in \mathsf{X}} \mathrm{E}_x \left[ \sum_{k=0}^{\tau_x - 1} \mathbb{1}_{\{X_k = y\}} \right] = \mathrm{E}_x[\tau_x] \,.$$

Thus the unique invariant measure is summable if and only if a state $x$ satisfying this relation exists. On the other hand, if such a state $x$ exists then, by uniqueness of the invariant measure, $\mathrm{E}_y[\tau_y] < \infty$ must hold for all states $y$. In this case, the invariant probability measure, $\pi$ say, satisfies $\pi(x) = \lambda_x(x)/\lambda_x(\mathsf{X}) = 1/\mathrm{E}_x[\tau_x]$. Because the reference state $x$ was in fact arbitrary, we find that $\pi(y) = 1/\mathrm{E}_x[\tau_y]$ for all states $y$. □

It is natural to ask what can be inferred from the knowledge that a chain possesses an invariant probability measure. The next proposition gives a partial answer.

**Proposition 142.** *Let $Q$ be a transition matrix and $\pi$ an invariant probability measure. Then every state $x$ such that $\pi(x) > 0$ is recurrent. If $Q$ is irreducible, then it is recurrent.*

*Proof.* Let $y \in \mathsf{X}$. If $\pi(y) > 0$ then $\sum_{n=0}^{\infty} \pi Q^n(y) = \sum_{n=0}^{\infty} \pi(y) = \infty$. On the other hand, by Proposition 136,

$$\begin{aligned} \sum_{n=0}^{\infty} \pi Q^n(y) &= \sum_{x \in \mathsf{X}} \pi(x) \sum_{n=0}^{\infty} Q^n(x,y) \\ &= \sum_{x \in \mathsf{X}} \pi(x) \,\mathrm{E}_x[\eta_y] \le \mathrm{E}_y[\eta_y] \sum_{x \in \mathsf{X}} \pi(x) = \mathrm{E}_y[\eta_y] \,. \end{aligned} \tag{7.9}$$

Thus $\pi(y) > 0$ implies $\mathrm{E}_y[\eta_y] = \infty$, that is, $y$ is recurrent. □

Let $\{X_k\}$ be an irreducible Markov chain. If there exists an invariant probability measure, the chain is called *positive recurrent*; otherwise it is called *null*. Note that null chains can be either null recurrent or transient. Transient chains are always null, though they may admit an invariant measure.

### 7.1.4 Ergodicity

A key result for positive recurrent irreducible chains is that the transition laws converge, in a suitable sense, to the invariant vector $\pi$. The classical result is the following.

**Proposition 143.** *Consider an irreducible and positive recurrent Markov chain on a countable state space. Then for any states $x$ and $y$,*

$$n^{-1} \sum_{i=1}^{n} Q^n(x,y) \to \pi(y) \quad \text{as } n \to \infty \,. \tag{7.10}$$

The use of the Césaro limit can be avoided if the chain is *aperiodic*. The simplest definition of aperiodicity is that a state $x$ is aperiodic if $Q^k(x,x) > 0$ for all $k$ sufficiently large or, equivalently, that the *period* of the state $x$ is one. The *period* of $x$ is defined as the greatest common divisor of the set $I(x) = \{n > 0 : Q^n(x,x) > 0\}$. For irreducible chains, the following result holds true.

**Proposition 144.** *If the chain is irreducible, then all states have the same period. If the transition matrix $Q$ is irreducible and aperiodic, then for all $x$ and $y$ in $\mathsf{X}$, there exists $n(x,y) \in \mathbb{N}$ such that $Q^k(x,y) > 0$ for all $k \geq n(x,y)$.*

Thus, an irreducible chain can be said to be aperiodic if the common period of all states is one.

The traditional pointwise convergence (7.10) of transition probabilities has been replaced in more recent research by convergence in *total variation* (see Definition 39). The convergence result may then be formulated as follows.

**Theorem 145.** *Consider an irreducible and aperiodic positive recurrent Markov chain on a countable state space $\mathsf{X}$ with transition matrix $Q$ and invariant probability distribution $\pi$. Then for all initial distributions $\xi$ and $\xi'$ on $\mathsf{X}$,*

$$\|\xi Q^n - \xi' Q^n\|_{\mathrm{TV}} \to 0 \quad \text{as } n \to \infty \,. \tag{7.11}$$

*In particular, for any $x \in \mathsf{X}$ we may set $\xi = \delta_x$ and $\xi' = \pi$ to obtain*

$$\|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} \to 0 \quad \text{as } n \to \infty \,. \tag{7.12}$$

The proof of this result, and indeed the focus on convergence in total variation, follows using of the coupling technique. We postpone the presentation of this technique to Section 7.2.4 because essentially the same ideas can be applied to Markov chains on general state spaces.

## 7.2 Chains on General State Spaces

In this section, we extend the concepts and results pertaining to countable state spaces to general ones. In the following, $\mathsf{X}$ is an arbitrary set, and we just require that it is equipped with a countably generated $\sigma$-field $\mathcal{X}$. By $\{X_k\}_{k\geq 0}$ we denote an $\mathsf{X}$-valued Markov chain with transition kernel $Q$. It is defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, and $\mathbb{F}^X = \{\mathcal{F}_k^X\}_{k\geq 0}$ denotes the natural filtration of $\{X_k\}$.

For any set $A \in \mathcal{X}$, we define the first *hitting time* $\sigma_A$ and *return time* $\tau_A$ respectively by

$$\sigma_A = \inf\{n \geq 0 : X_n \in A\} \,, \tag{7.13}$$
$$\tau_A = \inf\{n \geq 1 : X_n \in A\} \,, \tag{7.14}$$

where, by convention, $\inf \emptyset = +\infty$. The successive hitting times $\sigma_A^{(n)}$ and return times $\tau_A^{(n)}$, $n \geq 0$, are defined inductively by

$$\sigma_A^{(0)} = 0,\ \sigma_A^{(1)} = \sigma_A,\ \sigma_A^{(n+1)} = \inf\{k > \sigma_A^{(n)} :\ X_k \in A\} \,,$$
$$\tau_A^{(0)} = 0,\ \tau_A^{(1)} = \tau_A,\ \tau_A^{(n+1)} = \inf\{k > \tau_A^{(n)} :\ X_k \in A\} \,.$$

We again define the *occupation time* $\eta_A$ as the number of visits by $\{X_k\}$ to $A$,

$$\eta_A \stackrel{\mathrm{def}}{=} \sum_{k=0}^{\infty} \mathbb{1}_A(X_k) \,. \tag{7.15}$$

### 7.2.1 Irreducibility

The first step to develop a theory on general state spaces is to define a suitable concept of irreducibility. The definition of irreducibility adopted for countable state spaces does not extend to general ones, as the probability of reaching single point $x$ in the state space is typically zero.

**Definition 146** (Phi-irreducibility)**.** *The transition kernel $Q$, or the Markov chain $\{X_k\}_{k\geq 0}$ with transition kernel $Q$, is said to be* phi-irreducible *if there exists a measure $\phi$ on $(\mathsf{X}, \mathcal{X})$ such that for any $A \in \mathcal{X}$ with $\phi(A) > 0$, $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathsf{X}$. Such a measure is called an* irreducibility measure *for $Q$.*

Phi-irreducibility is a weaker property than irreducibility of a transition kernel on a countable state space. If a transition kernel on a countable state space is irreducible, then it is phi-irreducible, and any measure is an irreducibility measure. The converse is not true. For instance, the transition kernel

$$Q = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

on $\{0, 1\}$ is phi-irreducible ($\delta_1$ is an irreducibility measure for $Q$) but not irreducible.

In general, there are infinitely many irreducibility measures, and two irreducibility measures are not necessarily equivalent. For instance, if $\phi$ is an irreducibility measure and $\hat{\phi}$ is absolutely continuous with respect to $\phi$, then $\hat{\phi}$ is also an irreducibility measure. Nevertheless, as shown in the next result, there exist *maximal irreducibility measures* $\psi$, which are such that any irreducibility measure $\phi$ is absolutely continuous with respect to $\psi$.

**Theorem 147.** *Let $Q$ be a phi-irreducible transition kernel on $(\mathsf{X}, \mathcal{X})$. Then there exists an irreducibility measure $\psi$ such that all irreducibility measures are absolutely continuous with respect to $\psi$ and for all $A \in \mathcal{X}$,*

$$\psi(A) > 0 \;\Leftrightarrow\; \mathrm{P}_x(\tau_A < \infty) > 0 \text{ for all } x \in \mathsf{X}\,. \tag{7.16}$$

*Proof.* Let $\phi$ be an irreducibility measure and $\epsilon \in (0, 1)$. Let $\phi_\epsilon$ be the measure defined by $\phi_\epsilon = \phi K_\epsilon$, where $K_\epsilon$ is the *resolvent kernel* defined by

$$K_\epsilon(x, A) \stackrel{\text{def}}{=} (1-\epsilon) \sum_{k\geq 0} \epsilon^k Q^k(x, A)\,, \qquad x \in \mathsf{X}, A \in \mathcal{X}\,. \tag{7.17}$$

We will first show that $\phi_\epsilon$ is an irreducibility measure. Let $A \in \mathcal{X}$ be such that $\phi_\epsilon(A) > 0$ and define

$$\bar{A} = \{x \in \mathsf{X} :\, \mathrm{P}_x(\sigma_A < \infty) > 0\} = \{x \in \mathsf{X} :\, K_\epsilon(x, A) > 0\}\,. \tag{7.18}$$

By definition, $\phi_\epsilon(A) > 0$ implies that $\phi(\bar{A}) > 0$. Define $\bar{A}_m = \{x \in \mathsf{X} : \mathrm{P}_x(\sigma_A < \infty) \geq 1/m\}$. By construction, $\bar{A} = \bigcup_{m>0} \bar{A}_m$, and because $\phi(\bar{A}) > 0$, there exists $m$ such that $\phi(\bar{A}_m) > 0$. Because $\phi$ is an irreducibility measure, $\mathrm{P}_x(\tau_{\bar{A}_m} < \infty) > 0$ for all $x \in \mathsf{X}$. Hence by the strong Markov property, for all $x \in \mathsf{X}$,

$$\begin{aligned} \mathrm{P}_x(\tau_A < \infty) &\geq \mathrm{P}_x(\tau_{\bar{A}_m} + \sigma_A \circ \theta^{\tau_{\bar{A}_m}} < \infty\,, \tau_{\bar{A}_m} < \infty) \\ &= \mathrm{E}_x[\mathbb{1}_{\{\tau_{\bar{A}_m} < \infty\}} \mathrm{P}_{X_{\tau_{\bar{A}_m}}}(\sigma_A < \infty)] \geq \frac{1}{m} \mathrm{P}_x(\tau_{\bar{A}_m} < \infty) > 0\,, \end{aligned}$$

showing that $\phi_\epsilon$ is an irreducibility measure.

Now for $m \geq 0$ the Chapman-Kolmogorov equations imply

$$\int_{\mathsf{X}} \phi_\epsilon(dx)\, \epsilon^m Q^m(x, A) = (1-\epsilon) \int_{\mathsf{X}} \sum_{n=m}^{\infty} \epsilon^n Q^n(x, A)\, \phi(dx) \leq \phi_\epsilon(A)\,.$$

Therefore, if $\phi_\epsilon(A) = 0$ then $\phi_\epsilon K_\epsilon(A) = 0$, which in turn implies $\phi_\epsilon(\bar{A}) = 0$. Summarizing the results above, for any $A \in \mathcal{X}$,

$$\phi_\epsilon(A) > 0 \Leftrightarrow \phi_\epsilon\left(\{x \in \mathsf{X} :\, \mathrm{P}_x(\sigma_A < \infty) > 0\}\right) > 0\,. \tag{7.19}$$

This proves (7.16)

To conclude we must show that all irreducibility measures are absolutely continuous with respect to $\phi_\epsilon$. Let $\hat{\phi}$ be an irreducibility measure and let $C \in \mathcal{X}$ be such that $\hat{\phi}(C) > 0$. Then $\phi_\epsilon(\{x \in \mathsf{X} : \mathrm{P}_x(\sigma_C < \infty) > 0\}) = \phi_\epsilon(\mathsf{X}) > 0$, which, by (7.19), implies that $\phi_\epsilon(C) > 0$ . This exactly says that $\hat{\phi}$ is absolutely continuous with respect to $\phi_\epsilon$. □

A set $A \in \mathcal{X}$ is said to be *accessible for the kernel* $Q$ (or $Q$-*accessible*, or simply *accessible* if there is no risk of confusion) if $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathsf{X}$. The family of accessible sets is denoted $\mathcal{X}^+$. If $\psi$ is a maximal irreducibility measure the set $A$ is accessible if and only if $\psi(A) > 0$.

**Example 148** (Autoregressive Model)**.** The first-order autoregressive model on $\mathbb{R}$ is defined iteratively by $X_n = \phi X_{n-1} + U_n$, where $\phi$ is a real number and $\{U_n\}$ is an i.i.d. sequence. If $\Gamma$ is the probability distribution of the noise sequence $\{U_n\}$, the transition kernel of this chain is given by $Q(x, A) = \Gamma(A - \phi x)$. The autoregressive model is phi-irreducible provided that the noise distribution has an everywhere positive density with respect to Lebesgue measure $\lambda^{\mathrm{Leb}}$. If we take $\phi = \lambda^{\mathrm{Leb}}$, it is easy to see that whenever $\lambda^{\mathrm{Leb}}(A) > 0$, we have $\Gamma(A - \phi x) > 0$ for any $x$, and so $Q(x, A) > 0$ in just one step.

**Example 149.** For simplicity, we assume here that $\mathsf{X} = \mathbb{R}^d$, which we equip with the Borel $\sigma$-field $\mathcal{X} = \mathcal{B}(\mathbb{R}^d)$. Assume that we are given a probability density function $\pi$ on with respect to Lebesgue measure $\lambda^{\mathrm{Leb}}$. Let $r$ be a transition density kernel. Starting from $X_n = x$, a candidate transition $x'$ is generated from $r(x, \cdot)$ and accepted with probability

$$\alpha(x, x') = \frac{\pi(x')\, r(x', x)}{\pi(x)\, r(x, x')} \wedge 1 \,. \tag{7.20}$$

The transition kernel of the Metropolis-Hastings chain is given by

$$Q(x, A) = \int_A \alpha(x, x') r(x, x')\, \lambda^{\mathrm{Leb}}(dx') + \mathbb{1}_x(A) \int [1 - \alpha(x, x')] r(x, x')\, \lambda^{\mathrm{Leb}}(dx') \,. \tag{7.21}$$

There are various sufficient conditions for the Metropolis-Hastings algorithm to be phi-irreducible (Roberts and Tweedie, 1996; Mengersen and Tweedie, 1996). For the Metropolis-Hastings chain, it is simple to check that the chain is phi-irreducible if for $\lambda^{\mathrm{Leb}}$-almost all $x' \in \mathsf{X}$, the condition $\pi(x') > 0$ implies that $r(x, x') > 0$ for any $x \in \mathsf{X}$.

### 7.2.2 Recurrence and Transience

In view of the discussion above, it is not sensible to define recurrence and transience in terms of the expectation of the occupation measure of a state, but for phi-irreducible chains it makes sense to consider the occupation measure of accessible sets.

**Definition 150** (Uniform Transience and Recurrence)**.** *A set $A \in \mathcal{X}$ is called* uniformly transient *if* $\sup_{x \in A} \mathrm{E}_x[\eta_A] < \infty$. *A set $A \in \mathcal{X}$ is called* recurrent *if* $\mathrm{E}_x[\eta_A] = +\infty$ *for all $x \in A$.*

Obviously, if $\sup_{x \in \mathsf{X}} \mathrm{E}_x[\eta_A] < \infty$, then $A$ is uniformly transient. In fact the reverse implication holds true too, because if the chain is started outside $A$ it cannot

hit $A$ more times, on average, than if it is started at "the most favorable location" in $A$. Thus an alternative definition of a uniformly transient set is $\sup_{x\in\mathsf{X}} \mathrm{E}_x[\eta_A] < \infty$.

The main result on phi-irreducible transition kernels is the following recurrence/transience dichotomy, which parallels Theorem 139 for countable state-space Markov chains.

**Theorem 151.** *Let $Q$ be a phi-irreducible transition kernel (or Markov chain). Then either of the following two statements holds true.*

*(i) Every accessible set is recurrent, in which case we call $Q$ recurrent.*

*(ii) There is a countable cover of $\mathsf{X}$ with uniformly transient sets, in which case we call $Q$ transient.*

In the next section, we will prove Theorem 151 in the particular case where the chain possesses an *accessible atom* (see Definition 152); the proof is then very similar to that for countable state space. In the general case, the proof is more involved. It is necessary to introduce *small sets* and the so-called *splitting construction*, which relates the chain to one that does possess an accessible atom.

### Transience and Recurrence for Chains Possessing an Accessible Atom

**Definition 152** (Atom)**.** *A set $\alpha \in \mathcal{X}$ is called an* atom *if there exists a probability measure $\nu$ on $(\mathsf{X}, \mathcal{X})$ such that $Q(x, A) = \nu(A)$ for all $x \in \alpha$ and $A \in \mathcal{X}$.*

Atoms behave the same way as do individual states in the countable state space case. Although any singleton $\{x\}$ is an atom, it is not necessarily accessible, so that Markov chain theory on general state spaces differs from the theory of countable state space chains.

If $\alpha$ is an atom for $Q$, then for any $m \geq 1$ it is an atom for $Q^m$. Therefore we denote by $Q^m(\alpha, \cdot)$ the common value of $Q^m(x, \cdot)$ for all $x \in \alpha$. This implies that if the chain starts from within the atom, the distribution of the whole chain does not depend on the precise starting point. Therefore we will also use the notation $\mathrm{P}_\alpha$ instead of $\mathrm{P}_x$ for any $x \in \alpha$.

**Example 153** (Random Walk on the Half-Line)**.** The random walk on the half-line (RWHL) is defined by an initial condition $X_0 \geq 0$ and the recursion

$$X_{k+1} = (X_k + W_{k+1})^+ , \qquad k \geq 0 , \tag{7.22}$$

where $\{W_k\}_{k\geq 1}$ is an i.i.d. sequence of random variables, independent of $X_0$, with distribution function $\Gamma$ on $\mathbb{R}$. This process is a Markov chain with transition kernel $Q$ defined by

$$Q(x, A) = \Gamma(A - x) + \Gamma((-\infty\, , -x])\mathbb{1}_A(0) , \qquad x \in \mathbb{R}_+,\, A \in \mathcal{B}(\mathbb{R}_+) ,$$

where $A - x = \{y - x : y \in A\}$. The set $\{0\}$ is an atom, and it is accessible if and only if $\Gamma((-\infty\, , 0]) > 0$.

We now prove Theorem 151 when there exists an accessible atom.

**Proposition 154.** *Let $\{X_k\}_{k\geq 0}$ be a Markov chain that possesses an accessible atom $\alpha$, with associated probability measure $\nu$. Then the chain is phi-irreducible, $\nu$ is an irreducibility measure, and a set $A \in \mathcal{X}$ is accessible if and only if $\mathrm{P}_\alpha(\tau_A < \infty) > 0$.*

*Moreover, $\alpha$ is recurrent if and only if $\mathrm{P}_\alpha(\tau_\alpha < \infty) = 1$ and (uniformly) transient otherwise, and the chain is recurrent if $\alpha$ is recurrent and transient otherwise.*

*Proof.* For all $A \in \mathcal{X}$ and $x \in \mathsf{X}$, the strong Markov property yields

$$\begin{aligned}\mathrm{P}_x(\tau_A < \infty) &\geq \mathrm{P}_x(\tau_\alpha + \tau_A \circ \theta^{\tau_\alpha} < \infty,\, \tau_\alpha < \infty) \\ &= \mathrm{E}_x[\mathrm{P}_{X_{\tau_\alpha}}(\tau_A < \infty)\mathbb{1}_{\{\tau_\alpha < \infty\}}] \\ &= \mathrm{P}_\alpha(\tau_A < \infty)\,\mathrm{P}_x(\tau_\alpha < \infty) \\ &\geq \nu(A)\,\mathrm{P}_x(\tau_\alpha < \infty)\,.\end{aligned}$$

Because $\alpha$ is accessible, $\mathrm{P}_x(\tau_\alpha < \infty) > 0$ for all $x \in \mathsf{X}$. Thus for any $A \in \mathcal{X}$ satisfying $\nu(A) > 0$, it holds that $\mathrm{P}_x(\tau_A < \infty) > 0$ for all $x \in \mathsf{X}$, showing that $\nu$ is an irreducibility measure. The above display also shows that $A$ is accessible if and only if $\mathrm{P}_\alpha(\tau_A < \infty)$.

Now let $\sigma_\alpha^{(n)}$ be the successive hitting times of $\alpha$ (see (7.13)). The strong Markov property implies that for any $n > 1$,

$$\mathrm{P}_\alpha(\sigma_\alpha^{(n)} < \infty) = \mathrm{P}_\alpha(\tau_\alpha < \infty)\,\mathrm{P}_\alpha(\sigma_\alpha^{(n-1)} < \infty)\,.$$

Hence, as for discrete state spaces, $\mathrm{P}_\alpha(\sigma_\alpha^{(n)} < \infty) = [\mathrm{P}_\alpha(\tau_\alpha < \infty)]^{n-1}$ and $\mathrm{E}_\alpha[\eta_\alpha] = 1/[1 - \mathrm{P}_\alpha(\tau_\alpha < \infty)]$. This proves that $\alpha$ is recurrent if and only if $\mathrm{P}_\alpha(\tau_\alpha < \infty) = 1$.

Assume that $\alpha$ is recurrent. Because the atom $\alpha$ is accessible, for any $x \in \mathsf{X}$, there exists $r$ such that $Q^r(x, \alpha) > 0$. If $A \in \mathcal{X}^+$ there exists $s$ such that $Q^s(\alpha, A) > 0$. By the Chapman-Kolmogorov equations,

$$\sum_{n \geq 1} Q^{r+s+n}(x, A) \geq Q^r(x, \alpha) \left[\sum_{n \geq 1} Q^n(\alpha, \alpha)\right] Q^s(\alpha, A) = \infty\,.$$

Hence $\mathrm{E}_x[\eta_A] = \infty$ for all $x \in \mathsf{X}$ and $A$ is recurrent. Because $A$ was an arbitrary accessible set, the chain is recurrent.

Assume now that $\alpha$ is transient, in which case $\mathrm{E}_\alpha(\eta_\alpha) < \infty$. Then, following the same line of reasoning as in the discrete state space case (proof of Proposition 136), we obtain that for all $x \in \mathsf{X}$,

$$\mathrm{E}_x[\eta_\alpha] = \mathrm{P}_x(\tau_\alpha < \infty)\,\mathrm{E}_\alpha[\eta_\alpha] \leq \mathrm{E}_\alpha[\eta_\alpha]\,. \tag{7.23}$$

Define $B_j = \{x : \sum_{n=1}^j Q^n(x, \alpha) \geq 1/j\}$. Then $\cup_{j=1}^\infty B_j = \mathsf{X}$ because $\alpha$ is accessible. Applying the definition of the sets $B_j$ and the Chapman-Kolmogorov equations, we find that

$$\begin{aligned}\sum_{k=1}^\infty Q^k(x, B_j) &\leq \sum_{k=1}^\infty Q^k(x, B_j) \inf_{y \in B_j} j \sum_{\ell=1}^j Q^\ell(y, \alpha) \\ &\leq j \sum_{k=1}^\infty \sum_{\ell=1}^j \int_{B_j} Q^k(x, dy)\, Q^\ell(y, \alpha) \leq j^2 \sum_{k=1}^\infty Q^k(x, \alpha) = j^2\,\mathrm{E}_x[\eta_\alpha] < \infty\,.\end{aligned}$$

The sets $B_j$ are thus uniformly transient. The proof is complete. □

### Small Sets and the Splitting Construction

We now return to the general phi-irreducible case. In order to prove Theorem 151, we need to introduce the splitting technique. To do so, we need to define a class of sets (containing accessible sets) that behave the same way in many respects as do atoms. We shall see this in many of the results below, which exactly mimic the atomic case results they generalize. These sets are called *small sets*.

**Definition 155** (Small Set)**.** *Let $Q$ and $\nu$ be a transition kernel and a probability measure, respectively, on $(\mathsf{X}, \mathcal{X})$, let $m$ be a positive integer and $\epsilon \in (0, 1]$. A set $C \in \mathcal{X}$ is called a* $(m, \epsilon, \nu)$-small set *for $Q$, or simply a* small set*, if $\nu(C) > 0$ and for all $x \in C$ and $A \in \mathcal{X}$,*

$$Q^m(x, A) \geq \epsilon\nu(A) \; .$$

*If $\epsilon = 1$ then $C$ is an atom for the kernel $Q^m$.*

Trivially, any individual point is a small set, but small sets that are not accessible are of limited interest. If the state space is countable and $Q$ is irreducible, then every finite set is small. The minorization measure associated to an accessible small set provides an irreducibility measure.

**Proposition 156.** *Let $C$ be an accessible $(m, \epsilon, \nu)$-small set for the transition kernel $Q$ on $(\mathsf{X}, \mathcal{X})$. Then $\nu$ is an irreducibility measure.*

*Proof.* Let $A \in \mathcal{X}$ be such that $\nu(A) > 0$. The strong Markov property yields

$$\mathrm{P}_x(\tau_A < \infty) \geq \mathrm{P}_x(\tau_C < \infty,\, \tau_A \circ \theta^{\tau_C} < \infty) = \mathrm{E}_x[\mathbb{1}_{\{\tau_C < \infty\}} \mathrm{P}_{X_{\tau_C}}(\tau_A < \infty)] \; .$$

Because $C$ is a small set, for all $y \in C$ it holds that

$$\mathrm{P}_y(\tau_A < \infty) \geq \mathrm{P}_y(X_m \in A) = Q^m(y, A) \geq \epsilon\nu(A) \; .$$

Because $C$ is accessible and $\nu(A) > 0$, for all $x \in \mathsf{X}$ it holds that

$$\mathrm{P}_x(\tau_A < \infty) \geq \epsilon\nu(A)\, \mathrm{P}_x(\tau_C < \infty) > 0 \; .$$

Thus $A$ is accessible, whence $\nu$ is an irreducibility measure. □

An important result due to Jain and Jamison (1967) states that if the transition kernel is phi-irreducible, then small sets do exist. For a proof see Nummelin (1984, p. 16) or Meyn and Tweedie (1993, Theorem 5.2.2).

**Proposition 157.** *If the transition kernel $Q$ on $(\mathsf{X}, \mathcal{X})$ is phi-irreducible, then every accessible set contains an accessible small set.*

Given the existence of just one small set from Proposition 157, we may show that it is possible to cover $\mathsf{X}$ with a countable number of small sets in the phi-irreducible case.

**Proposition 158.** *Let $Q$ be a phi-irreducible transition kernel on $(\mathsf{X}, \mathcal{X})$.*

*(i) If $C \in \mathcal{X}$ is an $(m, \epsilon, \nu)$-small set and for any $x \in D$ we have $Q^n(x, C) \geq \delta$, then $D$ is $(m + n, \delta\epsilon, \nu)$-small set.*

*(ii) If $Q$ is phi-irreducible then there exists a countable collection of small sets $C_i$ such that $\mathsf{X} = \bigcup_i C_i$.*

*Proof.* Using the Chapman-Kolmogorov equations, we find that for any $x \in D$,

$$Q^{n+m}(x, A) \geq \int_C Q^n(x, dy)\, Q^m(y, A) \geq \epsilon Q^n(x, C)\nu(A) \geq \epsilon\delta\nu(A) \; ,$$

showing part (i). Because $Q$ is phi-irreducible, by Proposition 157 there exists an accessible $(m, \epsilon, \nu)$-small set $C$. Moreover, by the definition of phi-irreducibility, the sets $C(n, m) = \{x : Q^n(x, C) \geq 1/m\}$ cover $\mathsf{X}$ and, by part (i), each $C(n, m)$ is small. □

**Proposition 159.** *If $Q$ is phi-irreducible and transient, then every accessible small set is uniformly transient.*

*Proof.* Let $C$ be an accessible $(m, \epsilon, \nu)$-small set. If $Q$ is transient, there exists at least one $A \in \mathcal{X}^+$ that is uniformly transient. For $\delta \in (0,1)$, by the Chapman-Kolmogorov equations,

$$\begin{aligned}
\mathrm{E}_x[\eta_A] &= \sum_{k=0}^{\infty} Q^k(x,A) \geq (1-\delta) \sum_{p=0}^{\infty} \delta^p \sum_{k=0}^{\infty} Q^{k+m+p}(x,A) \\
&\geq (1-\delta) \sum_{p=0}^{\infty} \delta^p \sum_{k=0}^{\infty} \int_C Q^k(x,dx') \int Q^m(x',dx'')\, Q^p(x'',A) \\
&\geq \epsilon \sum_{k=0}^{\infty} Q^k(x,C) \times (1-\delta) \sum_{p=0}^{\infty} \delta^p \nu Q^p(A) = \epsilon\, \mathrm{E}_x[\eta_C]\, \nu K_\delta(A) \;,
\end{aligned}$$

where $K_\delta$ is the resolvent kernel (7.17). Because $C$ is an accessible small set, Proposition 156 shows that $\nu$ is an irreducibility measure. By Theorem 147, $\nu K_\delta$ is a maximal irreducibility measure, so that $\nu K_\delta(A) > 0$. Thus $\sup_{x \in \mathsf{X}} \mathrm{E}_x[\eta_C] < \infty$ and we conclude that $C$ is uniformly transient (see the remark following Definition 150). □

**Example 160** (Autoregressive Process, Continued)**.** Suppose that the noise distribution in Example 148 has an everywhere positive continuous density $\gamma$ with respect to Lebesgue measure $\lambda^{\mathrm{Leb}}$. If $C = [-M, M]$ and $\epsilon = \inf_{|x| \leq (1+\phi)M} \gamma(u)$, then for $A \subseteq C$,

$$Q(x,A) = \int_A \gamma(x' - \phi x)\, dx' \geq \epsilon \lambda^{\mathrm{Leb}}(A) \;.$$

Hence the compact set $C$ is small. Obviously $\mathbb{R}$ is covered by a countable collection of small sets and every accessible set (here sets with non-zero Lebesgue measure) contains a small set.

**Example 161** (Metropolis-Hastings Algorithm, Continued)**.** Similar results hold for the Metropolis-Hastings algorithm of Example 149 if $\pi(x)$ and $r(x,x')$ are positive and continuous for all $(x,x') \in \mathsf{X} \times \mathsf{X}$. Suppose that $C$ is compact with $\lambda^{\mathrm{Leb}}(C) > 0$. By positivity and continuity, we then have $d = \sup_{x \in C} \pi(x) < \infty$ and $\varepsilon = \inf_{(x,x') \in C \times C} q(x,x') > 0$. For any $A \subseteq C$, define

$$R_x(A) \stackrel{\mathrm{def}}{=} \left\{ x' \in A : \frac{\pi(x')q(x',x)}{\pi(x)q(x,x')} < 1 \right\} \;,$$

the region of possible rejection. Then for any $x \in C$,

$$\begin{aligned}
Q(x,A) &\geq \int_A q(x,x')\alpha(x,x')\, dx' \\
&\geq \int_{R_x(A)} \frac{q(x',x)}{\pi(x)} \pi(x')\, dx' + \int_{A \setminus R_x(A)} q(x,x')\, dx' \\
&\geq \frac{\varepsilon}{d} \int_{R_x(A)} \pi(x')\, dx' + \frac{\varepsilon}{d} \int_{A \setminus R_x(A)} \pi(x')\, dx' \\
&= \frac{\varepsilon}{d} \int_A \pi(x')\, dx' \;.
\end{aligned}$$

Thus $C$ is small and, again, $\mathsf{X}$ can be covered by a countable collection of small sets.

We now show that it is possible to define a Markov chain with an atom, the so-called *split chain*, whose properties are directly related to those of the original chain. This technique was introduced by Nummelin (1978) (Athreya and Ney,

1978, introduced, independently, a virtually identical concept) and allows extending results valid for Markov chain possessing an accessible atom to irreducible Markov chains that only possess small sets. The basic idea is as follows. Suppose the chain admits a $(1, \epsilon, \nu)$-small set $C$. Then as long as the chain does not enter $C$, the transition kernel $Q$ is used to generate the trajectory. However, as soon as the chain hits $C$, say $X_n \in C$, a zero-one random variable $d_n$ is drawn, independent of everything else. The probability that $d_n = 1$ is $\epsilon$, and hence $d_n = 0$ with probability $1 - \epsilon$. Then if $d_n = 1$, the next value $X_{n+1}$ is drawn from $\nu$; otherwise $X_{n+1}$ is drawn from the kernel

$$R(x, A) = [1 - \epsilon \mathbb{1}_C(x)]^{-1}[Q(x, A) - \epsilon \mathbb{1}_C(x)\nu(A)] ,$$

with $x = X_n$. It is immediate that $\epsilon\nu(A) + (1 - \epsilon)R(x, A) = Q(x, A)$ for all $x \in C$, so $X_{n+1}$ is indeed drawn from the correct (conditional) distribution. Note also that $R(x, \cdot) = Q(x, \cdot)$ for $x \notin C$. So, what is gained by this approach? What is gained is that whenever $X_n \in C$ and $d_n = 1$, the next value of the chain will be independent of $X_n$ (because it is drawn from $\nu$). This is often called a *regeneration time*, as the joint chain $\{(X_k, d_k)\}$ in a sense "restarts" and forgets its history. In technical terms, the state $C \times \{1\}$ in the extended state space is as atom, and it will be accessible provided $C$ is.

We now make this formal. Thus we define the so-called *extended state space* as $\check{\mathsf{X}} = \mathsf{X} \times \{0, 1\}$ and let $\check{\mathcal{X}}$ be the associated product $\sigma$-field. We associate to every measure $\mu$ on $(\mathsf{X}, \mathcal{X})$ the split measure $\mu^\star$ on $(\check{\mathsf{X}}, \check{\mathcal{X}})$ as the unique measure satisfying, for $A \in \mathcal{X}$,

$$\begin{aligned} \mu^\star(A \times \{0\}) &= (1 - \epsilon)\mu(A \cap C) + \mu(A \cap C^c) , \\ \mu^\star(A \times \{1\}) &= \epsilon\mu(A \cap C) . \end{aligned}$$

If $Q$ is a transition kernel on $(\mathsf{X}, \mathcal{X})$, we define the kernel $Q^\star$ on $\mathsf{X} \times \check{\mathcal{X}}$ by $Q^\star(x, \check{A}) = [Q(x, \cdot)]^\star(\check{A})$ for $x \in \mathsf{X}$ and $\check{A} \in \check{\mathcal{X}}$.

Assume now that $Q$ is a phi-irreducible transition kernel and let $C$ be a $(1, \epsilon, \nu)$-small set. We define the split transition kernel $\check{Q}$ on $\check{\mathsf{X}} \times \check{\mathcal{X}}$ as follows. For any $x \in \mathsf{X}$ and $\check{A} \in \check{\mathcal{X}}$,

$$\check{Q}((x, 0), \check{A}) = R^\star(x, \check{A}) , \tag{7.24}$$

$$\check{Q}((x, 1), \check{A}) = \nu^\star(\check{A}) . \tag{7.25}$$

Examining the above technicalities, we find that transitions into $C^c \times \{1\}$ have zero probability from everywhere, so that $d_n = 1$ can only occur if $X_n \in C$. Because $d_n = 1$ indicates a regeneration time, from within $C$, this is logical. Likewise we find that given a transition to some $y \in C$, the conditional probability that $d_n = 1$ is $\epsilon$, wherever the transition took place from. Thus the above split transition kernel corresponds to the following simulation scheme for $\{(X_k, d_k)\}$. Assume $(X_k, d_k)$ are given. If $X_k \notin C$, then draw $X_{k+1}$ from $Q(X_k, \cdot)$. If $X_k \in C$ and $d_n = 1$, then draw $X_{k+1}$ from $\nu$, otherwise from $R(X_k, \cdot)$. If the realized $X_{k+1}$ is not in $C$, then set $d_{k+1} = 0$; if $X_{k+1}$ is in $C$, then set $d_{k+1} = 1$ with probability $\epsilon$, and otherwise set $d_{k+1} = 0$.

Split measures operate on the split kernel in the following way. For any measure $\mu$ on $(\mathsf{X}, \mathcal{X})$,

$$\mu^\star\check{Q} = (\mu Q)^\star. \tag{7.26}$$

For any probability measure $\check{\mu}$ on $\check{\mathcal{X}}$, we denote by $\check{\mathrm{P}}_{\check{\mu}}$ and $\check{\mathrm{E}}_{\check{\mu}}$, respectively, the probability distribution and the expectation on the canonical space $(\check{\mathsf{X}}^{\mathbb{N}}, \check{\mathcal{X}}^{\otimes\mathbb{N}})$ such that the coordinate process, denoted $\{(X_k, d_k)\}_{k\geq 0}$, is a Markov chain with initial

probability measure $\check{\mu}$ and transition kernel $\check{Q}$. We also denote by $\{\check{\mathcal{F}}_k\}_{k\geq 0}$ the natural filtration of this chain and, as usual, by $\{\mathcal{F}_k^X\}_{k\geq 0}$ the natural filtration of $\{X_k\}_{k\geq 0}$.

**Proposition 162.** *Let $Q$ be a phi-irreducible transition kernel on $(\mathsf{X},\mathcal{X})$, let $C$ be an accessible $(1,\epsilon,\nu)$-small set for $Q$ and let $\mu$ be a probability measure on $(\mathsf{X},\mathcal{X})$. Then for any bounded $\mathcal{X}$-measurable function $f$ and any $k\geq 1$,*

$$\check{\mathrm{E}}_{\mu^\star}[f(X_k)\,|\,\mathcal{F}_{k-1}^X] = Qf(X_{k-1}) \quad \check{\mathrm{P}}_{\mu^\star}\text{-a.s.} \tag{7.27}$$

Before giving the proof, we discuss the implications of this result. It implies that under $\check{\mathrm{P}}_{\mu^\star}$, $\{X_k\}_{k\geq 0}$ is a Markov chain (with respect to its natural filtration) with transition kernel $Q$ and initial distribution $\mu$. By abuse of notation, we can identify $\{X_k\}$ with the coordinate process associated to the canonical space $\mathsf{X}^{\mathbb{N}}$. Denote by $\mathrm{P}_\mu$ the probability measure on $(\mathsf{X}^{\mathbb{N}},\mathcal{X}^{\otimes\mathbb{N}})$ such that $\{X_k\}_{k\geq 0}$ is a Markov chain with transition kernel $Q$ and initial distribution $\mu$ (see Section 1.1.2) and denote by $\mathrm{E}_\mu$ the associated expectation operator. Then Proposition 162 yields the following identity. For any bounded $\mathcal{F}_\infty^X$-measurable random variable $Y$,

$$\check{\mathrm{E}}_{\mu^\star}[Y] = \mathrm{E}_\mu[Y]\,. \tag{7.28}$$

*of Proposition 162.* We have, $\mu^\star$-a.s.,

$$\check{\mathrm{E}}_{\mu^\star}[f(X_k)\,|\,\check{\mathcal{F}}_{k-1}] = \mathbb{1}_{\{d_{k-1}=1\}}\nu(f) + \mathbb{1}_{\{d_{k-1}=0\}}Rf(X_{k-1})\,.$$

Because $\check{\mathrm{P}}_{\check{\mu}}(d_{k-1}=1\,|\,\mathcal{F}_{k-1}^X) = \epsilon\mathbb{1}_C(X_{k-1})$ $\check{\mathrm{P}}_{\mu^\star}$-a.s., it holds that

$$\begin{aligned}
\check{\mathrm{E}}_{\mu^\star}[f(X_k)\,|\,\mathcal{F}_{k-1}^X] &= \check{\mathrm{E}}_{\mu^\star}\{\check{\mathrm{E}}[f(X_k)\,|\,\check{\mathcal{F}}_{k-1}]\,|\,\mathcal{F}_{k-1}^X\} \\
&= \epsilon\mathbb{1}_C(X_{k-1})\nu(f) + [1-\epsilon\mathbb{1}_C(X_{k-1})]Rf(X_{k-1}) \\
&= Qf(X_{k-1})\,.
\end{aligned}$$

□

**Corollary 163.** *Under the assumptions of Proposition 162, $\mathsf{X}\times\{1\}$ is an accessible atom and $\nu^\star$ is an irreducibility measure for the split kernel $\check{Q}$. More generally, if $B\in\mathcal{X}$ is accessible for $Q$, then $B\times\{0,1\}$ is accessible for the split kernel.*

*Proof.* Because $\check{\alpha}=\mathsf{X}\times\{1\}$ is an atom for the split kernel $\check{Q}$, Proposition 154 shows that $\nu^\star$ is an irreducibility measure if $\check{\alpha}$ is accessible. Applying (7.28) we obtain for $x\in\mathsf{X}$,

$$\begin{aligned}
\check{\mathrm{P}}_{(x,1)}(\tau_{\check{\alpha}}<\infty) &= \check{\mathrm{P}}_{(x,1)}(d_n=1 \text{ for some } n\geq 1) \\
&\geq \check{\mathrm{P}}_{(x,1)}(d_1=1) = \epsilon\nu(C) > 0\,, \\
\check{\mathrm{P}}_{(x,0)}(\tau_{\check{\alpha}}<\infty) &= \check{\mathrm{P}}_{(x,0)}((X_n,d_n)\in C\times\{1\} \text{ for some } n\geq 1) \\
&\geq \check{\mathrm{P}}_{(x,0)}(\tau_{C\times\{0,1\}}<\infty\,, d_{\tau_{C\times\{0,1\}}}=1) = \epsilon\,\mathrm{P}_x(\tau_C<\infty) > 0\,.
\end{aligned}$$

Thus $\check{\alpha}$ is accessible and $\nu^\star$ is an irreducibility measure for $\check{Q}$. This implies, by Theorem 147, that for all $\eta\in(0,1)$, $\nu^\star\check{K}_\eta$ is a maximal irreducibility measure for the split kernel $\check{Q}$; here $K_\eta$ is the resolvent kernel (7.17) associated to $\check{Q}$. By straightforward applications of the definitions, it is easy to check that $\nu^\star\check{K}_\eta = (\nu K_\eta)^\star$. Moreover, $\nu$ is an irreducibility measure for $Q$, and $\nu K_\eta$ is a maximal irreducibility measure for $Q$ (still by Proposition 156 and Theorem 147). If $B$ is accessible, then $\nu K_\eta(B)>0$ and

$$\nu^\star\check{K}_\eta(B\times\{0,1\}) = (\nu K_\eta)^\star(B\times\{0,1\}) = \nu K_\eta(B) > 0.$$

Thus $B\times\{0,1\}$ is accessible for $\check{Q}$. □

**Transience/Recurrence Dichotomy for General Phi-irreducible Chains**

Using the splitting construction, we are now able to prove Theorem 151 for chains not possessing accessible atoms. We first consider the simple case in which the chain possesses a 1-small set.

**Proposition 164.** *Let $Q$ be a phi-irreducible transition kernel that admits an accessible $(1, \epsilon, \nu)$-small set $C$. Then $Q$ is either recurrent or transient. It is recurrent if and only if the small set $C$ is recurrent.*

*Proof.* Because the split chain possesses an accessible atom, by Proposition 154 the split chain is phi-irreducible and either recurrent or transient. Applying (7.28) we can write

$$\check{\mathrm{E}}_{\delta_x^\star}[\eta_{B\times\{0,1\}}] = \mathrm{E}_x[\eta_B] \; . \tag{7.29}$$

Assume first that the split chain is recurrent. Let $B$ be an accessible set for $Q$. By Proposition 162, $B\times\{0,1\}$ is accessible for the split chain. Hence $\check{\mathrm{E}}_{\delta_x^\star}[\eta_{B\times\{0,1\}}] = \infty$ for all $x \in B$, so that, by (7.29), $\mathrm{E}_x[\eta_B] = \infty$ for all $x \in B$.

Conversely, if the split chain is transient, then by Proposition 154 the atom $\check{\alpha}$ is transient. For $j \geq 1$, define $B_j = \{x : \sum_{l=1}^{j} \check{Q}^l((x,0), \check{\alpha}) \geq 1/j\}$. Because $\check{\alpha}$ is accessible, $\cup_{j=1}^{\infty} B_j = \mathsf{X}$. By the same argument as in the proof of Proposition 154, the sets $B_j \times \{0,1\}$ are uniformly transient for the split chain. Hence, by (7.29), the sets $B_j$ are uniformly transient for $Q$.

It remains to prove that if the small set $C$ is recurrent, then the chain is recurrent. We have just proved that $Q$ is recurrent if and only if $\check{Q}$ is recurrent and, by Proposition 154, this is true if and only if the atom $\check{\alpha}$ is recurrent. Thus we only need to prove that if $C$ is recurrent then $\check{\alpha}$ is recurrent. If $C$ is recurrent, then (7.29) yields for all $x \in C$,

$$\check{\mathrm{E}}_{\delta_x^\star}[\eta_{\check{\alpha}}] \geq \epsilon \check{\mathrm{E}}_{\delta_x^\star}[\eta_{C\times\{0,1\}}] = \epsilon \, \mathrm{E}_x[\eta_C] = \infty \; .$$

Using the definition of $\delta_x^\star$, this implies that there exists $\check{x} \in \check{\mathsf{X}}$ such that $\check{\mathrm{E}}_{\check{x}}[\eta_{\check{\alpha}}] = \infty$. This observation and (7.23) imply that $\check{\mathrm{E}}_{\check{\alpha}}[\eta_{\check{\alpha}}] = \infty$, that is, the atom is recurrent. □

Using the resolvent kernel, the previous results can be extended to the general case where an accessible small set exists, but not necessarily a 1-small one.

**Proposition 165.** *Let $Q$ be transition kernel.*

*(i) If $Q$ is phi-irreducible and admits an accessible $(m, \epsilon, \nu)$-small set $C$, then for any $\eta \in (0,1)$, $C$ is an accessible $(1, \epsilon', \nu)$-small set for the resolvent kernel $K_\eta = (1-\eta)\sum_{k=0}^{\infty} \eta^k Q^k$ with $\epsilon' = (1-\eta)\eta^m \epsilon$.*

*(ii) A set is recurrent (resp. uniformly transient) for $Q$ if and only if it is recurrent (resp. uniformly transient) for $K_\eta$ for some (hence for all) $\eta \in (0,1)$.*

*(iii) $Q$ is recurrent (resp. transient) if and only if $K_\eta$ is recurrent (resp. transient) for some (hence for all) $\eta \in (0,1)$.*

*Proof.* For any $\eta > 0$, $x \in C$, and $A \in \mathcal{X}$,

$$K_\eta(x, A) \geq (1-\eta)\eta^m Q^m(x, A) \geq (1-\eta)\eta^m \epsilon \nu(A) = \epsilon' \nu(A) \; .$$

Thus $C$ is a $(1, \epsilon', \nu)$-small set for $K_\eta$, showing part (i). The remaining claims follow from the identity

$$\sum_{n\geq 1} K_\eta^n = \frac{1-\eta}{\eta} \sum_{n \geq 0} Q^n \; .$$

□

### Harris Recurrence

As for countable state spaces, it is sometimes useful to consider stronger recurrence properties, expressed in terms of return probabilities rather than mean occupation times.

**Definition 166** (Harris Recurrence)**.** *A set $A \in \mathcal{X}$ is said to be* Harris recurrent *if* $\mathrm{P}_x(\tau_A < \infty) = 1$ *for any* $x \in \mathsf{X}$*. A phi-irreducible Markov chain is said to be* Harris (recurrent) *if any accessible set is Harris recurrent.*

It is intuitively obvious that, as for countable state spaces, Harris recurrence implies recurrence.

**Proposition 167.** *A Harris recurrent set is recurrent.*

*Proof.* Let $A$ be a Harris recurrent set. Because for $j \geq 1$, $\sigma_A^{(j+1)} = \tau_A \circ \theta^{\sigma_A^{(j)}}$ on the set $\{\sigma_A^{(j)} < \infty\}$, the strong Markov property implies that for any $x \in A$,

$$\mathrm{P}_x(\sigma_A^{(j+1)} < \infty) = \mathrm{E}_x\left[\mathrm{P}_{X_{\sigma_A^{(j)}}}(\tau_A < \infty)\mathbb{1}_{\{\sigma_A^{(j)} < \infty\}}\right] = \mathrm{P}_x(\sigma_A^{(j)} < \infty)\ .$$

Because $\mathrm{P}_x(\sigma_A^{(1)} < \infty) = 1$ for $x \in A$, we obtain that for all $x \in A$ and all $j \geq 1$, $\mathrm{P}_x(\sigma_A^{(j)} = 1)$ and $\mathrm{E}_x[\eta_A] = \sum_{j=1}^{\infty} \mathrm{P}_x(\sigma_A^{(j)} < \infty) = \infty$. □

Even though all transition kernels may not be Harris recurrent, the following theorem provides a very useful decomposition of the state space of a recurrent phi-irreducible transition kernel. For a proof of this result, see Meyn and Tweedie (1993, Theorem 9.1.5)

**Theorem 168.** *Let $Q$ be a phi-irreducible recurrent transition kernel on a state space $\mathsf{X}$ and let $\psi$ be a maximal irreducibility measure. Then $\mathsf{X} = \mathsf{N} \cup \mathsf{H}$, where $\mathsf{N}$ is covered by a countable family of uniformly transient sets, $\psi(\mathsf{N}) = 0$ and every accessible subset of $\mathsf{H}$ is Harris recurrent.*

As a consequence, if $A$ is an accessible set of a recurrent phi-irreducible chain, then there exists a set $A' \subseteq A$ such that $\psi(A \setminus A') = 0$ for any maximal irreducibility measure $\psi$, and $\mathrm{P}_x(\tau_{A'} < \infty) = 1$ for all $x \in A'$.

**Example 169.** To understand why a recurrent Markov chain can fail to be Harris, consider the following elementary example of a chain on $\mathsf{X} = \mathbb{N}$. Let the transition kernel $Q$ be given by $Q(0,0) = 1$ and for $x \geq 1$, $Q(x, x+1) = 1 - 1/x^2$ and $Q(x,0) = 1/x^2$. Thus the state 0 is absorbing. Because $Q(x,0) > 0$ for any $x \in \mathsf{X}$, $\delta_0$ is an irreducibility measure. In fact, by application of Theorem 147, this measure is maximal. The set $\{0\}$ is an atom and because $\mathrm{P}_0(\tau_{\{0\}} < \infty) = 1$, the chain is recurrent by Proposition 154.

The chain is not Harris recurrent, however. Indeed, for any $x \geq 1$ we have

$$\mathrm{P}_x(\tau_0 \geq k) = \mathrm{P}_x(X_1 \neq 0, \ldots, X_{k-1} \neq 0) = \prod_{j=x}^{x+k-1} (1 - 1/j^2)\ .$$

Because $\prod_{j=2}^{\infty}(1 - 1/j^2) > 0$, we obtain that $\mathrm{P}_x(\tau_0 = \infty) = \lim_{k\to\infty} \mathrm{P}_x(\tau_0 \geq k) > 0$ for any $x \geq 2$, so that the accessible state 0 is not certainly reached from such an initial state. Comparing to Theorem 168, we see that the decomposition of the state space is given by $\mathsf{H} = \{0\}$ and $\mathsf{N} = \{1, 2, \ldots\}$.

### 7.2.3 Invariant Measures and Stationarity

On general state spaces, we again further classify chains using *invariant measures*. A $\sigma$-finite measure $\mu$ is called *$Q$-sub-invariant* if $\mu \geq \mu Q$ and *$Q$-invariant* if $\mu = \mu Q$.

**Theorem 170.** *A phi-irreducible recurrent transition kernel (or Markov chain) admits a unique (up to a multiplicative constant) invariant measure which is also a maximal irreducibility measure.*

This result leads us to define the following classes of chains.

**Definition 171** (Positive and Null Chains)**.** *A phi-irreducible transition kernel (or Markov chain) is called* positive *if it admits an invariant probability measure; otherwise it is called* null.

We now prove the existence of an invariant measure when the chain admits an accessible atom. The invariant measure is defined as for countable state spaces, by replacing an individual state by the atom. Thus define the measure $\mu_\alpha$ on $\mathcal{X}$ by

$$\mu_\alpha(A) = \mathrm{E}_\alpha \left[ \sum_{n=1}^{\tau_\alpha} \mathbb{1}_A(X_n) \right], \qquad A \in \mathcal{X} \,. \tag{7.30}$$

**Proposition 172.** *Let $\alpha$ be an accessible atom for the transition kernel $Q$. Then $\mu_\alpha$ is $Q$-sub-invariant. It is invariant if and only if the atom $\alpha$ is recurrent. In that case, any $Q$-invariant measure $\mu$ is proportional to $\mu_\alpha$, and $\mu_\alpha$ is a maximal irreducibility measure.*

*Proof.* By the definition of $\mu_\alpha$ and the strong Markov property,

$$\begin{aligned} \mu_\alpha Q(A) &= \mathrm{E}_\alpha \left[ \sum_{k=1}^{\tau_\alpha} Q(X_k, A) \right] = \mathrm{E}_\alpha \left[ \sum_{k=2}^{\tau_\alpha+1} \mathbb{1}_A(X_k) \right] \\ &= \mu_\alpha(A) - \mathrm{P}_\alpha(X_1 \in A) + \mathrm{E}_\alpha[\mathbb{1}_A(X_{\tau_\alpha+1}) \mathbb{1}_{\{\tau_\alpha < \infty\}}] \,. \end{aligned}$$

Applying the strong Markov property once again yields

$$\begin{aligned} \mathrm{E}_\alpha[\mathbb{1}_A(X_{\tau_\alpha+1}) \mathbb{1}_{\{\tau_\alpha<\infty\}}] &= \mathrm{E}_\alpha\{\mathrm{E}_\alpha[\mathbb{1}_A(X_1) \circ \theta^{\tau_\alpha} \,|\, \mathcal{F}^X_{\tau_\alpha}] \mathbb{1}_{\{\tau_\alpha<\infty\}}\} \\ &= \mathrm{E}_\alpha[\mathrm{P}_{X_{\tau_\alpha}}(X_1 \in A) \mathbb{1}_{\{\tau_\alpha<\infty\}}] = \mathrm{P}_\alpha(X_1 \in A)\, \mathrm{P}_\alpha(\tau_\alpha < \infty) \,. \end{aligned}$$

Thus $\mu_\alpha Q(A) = \mu_\alpha(A) - \mathrm{P}_\alpha(X_1 \in A)[1 - \mathrm{P}_\alpha(\tau_\alpha < \infty)]$. This proves that $\mu_\alpha$ is sub-invariant, and invariant if and only if $\mathrm{P}_\alpha(\tau_\alpha < \infty) = 1$.

Now let $\mu$ be an invariant non-trivial measure and let $A$ be an accessible set such that $\mu(A) < \infty$. Then there exists an integer $n$ such that $Q^n(\alpha, A) > 0$. Because $\mu$ is invariant, it holds that $\mu = \mu Q^n$, so that

$$\infty > \mu(A) = \mu Q^n(A) \geq \mu(\alpha) Q^n(\alpha, A) \,.$$

This implies that $\mu(\alpha) < \infty$. Without loss of generality, we can assume $\mu(\alpha) > 0$; otherwise we replace $\mu$ by $\mu + \mu_\alpha$. Assuming $\mu(\alpha) > 0$, there is then no loss of generality in assuming $\mu(\alpha) = 1$.

The next step is to prove that if $\mu$ is an invariant measure such that $\mu(\alpha) = 1$, then $\mu \geq \mu_\alpha$. To prove this it suffices to prove that for all $n \geq 1$,

$$\mu(A) \geq \sum_{k=1}^{n} \mathrm{P}_\alpha(X_k \in A,\, \tau_\alpha \geq k) \,.$$

We prove this inequality by induction. For $n = 1$ we can write

$$\mu(A) = \mu Q(A) \geq \mu(\alpha) Q(\alpha, A) = Q(\alpha, A) = \mathrm{P}_\alpha(X_1 \in A) \; .$$

Now assume now that the inequality holds for some $n \geq 1$. Then

$$\begin{aligned} \mu(A) &= Q(\alpha, A) + \int_{\alpha^c} \mu(dy)\, Q(y, A) \\ &\geq Q(\alpha, A) + \sum_{k=1}^{n} \mathrm{E}_\alpha[Q(X_k, A) \mathbb{1}_{\{\tau_\alpha \geq k\}} \mathbb{1}_{\{X_k \notin \alpha\}}] \\ &\geq Q(\alpha, A) + \sum_{k=1}^{n} \mathrm{E}_\alpha[Q(X_k, A) \mathbb{1}_{\{\tau_\alpha \geq k+1\}}] \; . \end{aligned}$$

Because $\{\tau_\alpha \geq k+1\} \in \mathcal{F}_k^X$, the Markov property yields

$$\mathrm{E}_\alpha[Q(X_k, A) \mathbb{1}_{\{\tau_\alpha \geq k+1\}}] = \mathrm{P}_\alpha(X_{k+1} \in A,\, \tau_\alpha \geq k+1) \; ,$$

whence

$$\mu(A) \geq Q(\alpha, A) + \sum_{k=2}^{n+1} \mathrm{P}_\alpha(X_k \in A,\, \tau_\alpha \geq k) = \sum_{k=1}^{n+1} \mathrm{P}_\alpha(X_k \in A,\, \tau_\alpha \geq k) \; .$$

This completes the induction, and we conclude that $\mu \geq \mu_\alpha$.

Assume that there exists a set $A$ such that $\mu(A) > \mu_\alpha(A)$. It is straightforward that $\mu$ and $\mu_\alpha$ are both invariant for the resolvent kernel $K_\delta$ (see (7.17)), for any $\delta \in (0, 1)$. Because $\alpha$ is accessible, $K_\delta(x, \alpha) > 0$ for all $x \in \mathsf{X}$. Hence $\int_A \mu(dx)\, Q(x, \alpha) > \int_A \mu_\alpha(dx)\, Q(x, \alpha)$, which implies that

$$\begin{aligned} 1 = \mu(\alpha) = \mu K_\delta(\alpha) &= \int_A \mu(dx)\, K_\delta(x, \alpha) + \int_{A^c} \mu(dx)\, K_\delta(x, \alpha) \\ &> \int_A \mu_\alpha(dx)\, K_\delta(x, \alpha) + \int_{A^c} \mu_\alpha(dx)\, K_\delta(x, \alpha) = \mu_\alpha K_\delta(\alpha) = \mu_\alpha(\alpha) = 1. \end{aligned}$$

This contradiction shows that $\mu = \mu_\alpha$.

We finally prove that $\mu_\alpha$ is a maximal irreducibility measure. Let $\psi$ be a maximal irreducibility measure and assume that $\psi(A) = 0$. Then $\mathrm{P}_x(\tau_A < \infty) = 0$ for $\psi$-almost all $x \in \mathsf{X}$. This obviously implies that $\mathrm{P}_x(\tau_A < \infty) = 0$ for $\psi$-almost all $x \in \alpha$. Because $\mathrm{P}_x(\tau_A < \infty)$ is constant over $\alpha$, we find that $\mathrm{P}_x(\tau_A < \infty) = 0$ for all $x \in \alpha$, and this yields $\mu_\alpha(A) = 0$. Thus $\mu_\alpha$ is absolutely continuous with respect to $\psi$, hence an irreducibility measure. Let again $K_\delta$ be the resolvent kernel. By Theorem 147, $\mu_\alpha K_\delta$ is a maximal irreducibility measure. But, as noted above, $\mu_\alpha K_\epsilon = \mu_\alpha$, and therefore $\mu_\alpha$ is a maximal irreducibility measure. □

**Proposition 173.** *Let $Q$ be a recurrent phi-irreducible transition kernel that admits an accessible $(1, \epsilon, \nu)$-small set $C$. Then it admits a non-trivial invariant measure, unique up to multiplication by a constant and such that $0 < \pi(C) < \infty$, and any invariant measure is a maximal irreducibility measure.*

*Proof.* By (7.26), $(\mu Q)^\star = \mu^\star \check{Q}$, so that $\mu$ is $Q$-invariant if and only if $\mu^\star$ is $\check{Q}$-invariant. Let $\check{\mu}$ be a $\check{Q}$-invariant measure and define

$$\mu = \int_{C \times \{0\}} \check{\mu}(d\check{x})\, R(x, \cdot) + \int_{C^c \times \{0\}} \check{\mu}(d\check{x})\, Q(x, \cdot) + \check{\mu}(\mathsf{X} \times \{1\})\nu \; .$$

By application of the definition of the split kernel and measures, it can be checked that $\check{\mu}\check{Q} = \mu^\star$. Hence $\mu^\star = \check{\mu}\check{Q} = \check{\mu}$. We thus see that $\mu^\star$ is $\check{Q}$-invariant, which, as

noted above, implies that $\mu$ is $Q$-invariant. Hence we have shown that there exists a $Q$-invariant measure if and only if there exists a $\check{Q}$-invariant one.

If $Q$ is recurrent then $C$ is recurrent, and as appears in the proof of Proposition 173 this implies that the atom $\check{\alpha}$ is recurrent for the split chain $\check{Q}$. Thus, by Proposition 154 the kernel $\check{Q}$ is recurrent, and by Proposition 172 it admits an invariant measure that is unique up to a scaling factor. Hence $Q$ also admits an invariant measure, unique up to a scaling factor and such that $0 < \pi(C) < \infty$.

Let $\mu$ be $Q$-invariant. Then $\mu^\star$ is $\check{Q}$-invariant and hence, by Proposition 172, a maximal irreducibility measure. If $\mu(A) > 0$, then $\mu^\star(A \times \{0,1\}) = \mu(A) > 0$. Thus $A \times \{0,1\}$ is accessible, and this implies that $A$ is accessible. We conclude that $\mu$ is an irreducibility measure, and it is maximal because it is $K_\eta$-invariant. □

If the kernel $Q$ is phi-irreducible and admits an accessible $(m, \epsilon, \nu)$-small set $C$, then, by Proposition 165, for any $\eta \in (0,1)$ the set $C$ is an accessible $(1, \epsilon', \nu)$-small set for the resolvent kernel $K_\eta$. If $C$ is recurrent for $Q$, it is also recurrent for $K_\eta$ and therefore, by Proposition 164, $K_\eta$ has a unique invariant probability measure. The following result shows that this probability measure is invariant also for $Q$.

**Lemma 174.** *A measure $\mu$ on $(\mathsf{X}, \mathcal{X})$ is $Q$-invariant if and only if $\mu$ is $K_\eta$-invariant for some (hence for all) $\eta \in (0,1)$.*

*Proof.* If $\mu Q = \mu$, then obviously $\mu Q^n = \mu$ for all $n \geq 0$, so that $\mu K_\eta = \mu$. Conversely, assume that $\mu K_\eta = \mu$. Because $K_\eta = \eta Q K_\eta + (1-\eta)Q^0$ and $QK_\eta = K_\eta Q$, it holds that

$$\mu = \mu K_\eta = \eta \mu Q K_\eta + (1-\eta)\mu = \eta \mu K_\eta Q + (1-\eta)\mu = \eta \mu Q + (1-\eta)\mu \; .$$

Hence $\eta \mu Q = \eta \mu$, which concludes the proof. □

### Drift Conditions

We first give a sufficient condition for a chain to be positive, based on the expectation of the return time to an accessible small set.

**Proposition 175.** *Let $Q$ be a transition kernel that admits an accessible small set $C$ such that*

$$\sup_{x \in C} \mathrm{E}_x[\tau_C] < \infty \; . \tag{7.31}$$

*Then the chain is positive and the invariant probability measure $\pi$ satisfies, for all $A \in \mathcal{X}$,*

$$\pi(A) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} \mathbb{1}_A(X_k) \right] = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} \mathbb{1}_A(X_k) \right] \; . \tag{7.32}$$

*If $f$ is a non-negative measurable function such that*

$$\sup_{x \in C} \mathrm{E}_x \left[ \sum_{k=0}^{\tau_C - 1} f(X_k) \right] < \infty \; , \tag{7.33}$$

*then $f$ is integrable with respect to $\pi$ and*

$$\pi(f) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} f(X_k) \right] = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} f(X_k) \right] \; .$$

*Proof.* First note that by Proposition 156, $Q$ is phi-irreducible. Equation (7.31) implies that for all $\mathrm{P}_x(\tau_C < \infty) = 1$ $x \in C$, that is, $C$ is Harris recurrent. By Proposition 167, $C$ is recurrent, and so, by Proposition 164, $Q$ is recurrent. Let $\pi$ be an invariant measure such that $0 < \pi(C) < \infty$, the existence of which is given by Proposition 173. Then define a measure $\mu_C$ on $\mathcal{X}$ by

$$\mu_C(A) \stackrel{\text{def}}{=} \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} \mathbb{1}_A(X_k) \right] .$$

Because $\tau_C < \infty$ $\mathrm{P}_y$-a.s. for all $y \in C$, it holds that $\mu_C(C) = \pi(C)$. Then we can show that $\mu_C(A) = \pi(A)$ for all $A \in \mathcal{X}$. The proof is along the same lines as the proof of Proposition 172 and is therefore omitted. Thus, $\mu_C$ is invariant. In addition, we obtain that for any measurable set $A$,

$$\int_C \pi(dy)\, \mathrm{E}_y \left[ \mathbb{1}_A(X_0) \right] = \pi(A \cap C) = \mu_C(A \cap C) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \mathbb{1}_A(X_{\tau_C}) \right] ,$$

and this yields

$$\mu_C(A) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} \mathbb{1}_A(X_k) \right] = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} \mathbb{1}_A(X_k) \right] .$$

We thus obtain the following equivalent expressions for $\mu_C$:

$$\begin{aligned} \mu_C(A) &= \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} \mathbb{1}_A(X_k) \right] = \int_C \mu_C(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} \mathbb{1}_A(X_k) \right] \\ &= \int_C \mu_C(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} \mathbb{1}_A(X_k) \right] = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C} \mathbb{1}_A(X_k) \right] \\ &= \pi(A) . \end{aligned}$$

Hence

$$\pi(\mathsf{X}) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} \mathbb{1}_{\mathsf{X}}(X_k) \right] \le \pi(C) \sup_{y \in C} \mathrm{E}_y[\tau_C] < \infty ,$$

so that any invariant measure is finite and the chain is positive. Finally, under (7.33) we obtain that

$$\pi(f) = \int_C \pi(dy)\, \mathrm{E}_y \left[ \sum_{k=0}^{\tau_C - 1} f(X_k) \right] \le \pi(C) \sup_{y \in C} \mathrm{E}_y \left[ \sum_{k=1}^{\tau_C - 1} f(X_k) \right] < \infty .$$

□

Except in specific examples (where, for example, the invariant distribution is known in advance), it may be difficult to decide if a chain is positive or null. To check such properties, it is convenient to use *drift conditions.*

**Proposition 176.** *Assume that there exists a set $C \in \mathcal{X}$, two measurable functions $1 \le f \le V$, and a constant $b > 0$ such that*

$$QV \le V - f + b\mathbb{1}_C . \tag{7.34}$$

*Then*

$$\mathrm{E}_x[\tau_C] \le V(x) + b\mathbb{1}_C(x) , \tag{7.35}$$

$$\mathrm{E}_x \left[ V(X_{\tau_C}) \right] + \mathrm{E}_x \left[ \sum_{k=0}^{\tau_C - 1} f(X_k) \right] \le V(x) + b\mathbb{1}_C(x) . \tag{7.36}$$

*If $C$ is an accessible small set and $V$ is bounded on $C$, then the chain is positive recurrent and $\pi(f) < \infty$.*

*Proof.* Set for $n \geq 1$,

$$M_n = \left[ V(X_n) + \sum_{k=0}^{n-1} f(X_k) \right] \mathbb{1}_{\{\tau_C \geq n\}} .$$

Then

$$\begin{aligned} \mathrm{E}[M_{n+1} \,|\, \mathcal{F}_n] &= \left[ QV(X_n) + \sum_{k=0}^{n} f(X_k) \right] \mathbb{1}_{\{\tau_C \geq n+1\}} \\ &\leq \left[ V(X_n) - f(X_n) + b\mathbb{1}_C(X_n) + \sum_{k=0}^{n} f(X_k) \right] \mathbb{1}_{\{\tau_C \geq n+1\}} \\ &= \left[ V(X_n) + \sum_{k=0}^{n-1} f(X_k) \right] \mathbb{1}_{\{\tau_C \geq n+1\}} \leq M_n , \end{aligned}$$

as $\mathbb{1}_C(X_n)\mathbb{1}_{\{\tau_C \geq n+1\}} = 0$. Hence $\{M_n\}_{n \geq 1}$ is a non-negative super-martingale. For any integer $n$, $\tau_C \wedge n$ is a bounded stopping time, and Doob's optional stopping theorem shows that for any $x \in \mathsf{X}$,

$$\mathrm{E}_x \left[ M_{\tau_C \wedge n} \right] \leq \mathrm{E}_x \left[ M_1 \right] \leq V(x) + b\mathbb{1}_C(x) . \tag{7.37}$$

Applying this relation with $f \equiv 1$ yields for any $x \in \mathsf{X}$ and $n \geq 0$,

$$\mathrm{E}_x \left[ \tau_C \wedge n \right] \leq V(x) + b\mathbb{1}_C(x) ,$$

and (7.35) follows using monotone convergence. This implies in particular that $\mathrm{P}_x(\tau_C < \infty) = 1$ for any $x \in \mathsf{X}$. The proof of (7.36) follows similarly from (7.37) by the letting $n \to \infty$ and $\pi(f)$ is finite by (7.33). □

**Example 177** (Random Walk on the Half-Line, Continued)**.** Consider again the model of Example 153. Previously we have seen that sets of the form $[0, c]$ are small. If $\Gamma((-\infty\,, -c]) > 0$, then for $x \in [0, c]$,

$$Q(x, A) \geq \Gamma((-\infty\,, -c])\mathbb{1}_A(0) ;$$

otherwise there exists an integer $m$ such that $\Gamma^{*m}((-\infty\,, -c]) > 0$, whence

$$Q^m(x, A) \geq \Gamma^{*m}((-\infty\,, -c])\mathbb{1}_A(0) .$$

To prove recurrence for $\mu < 0$, we apply Proposition 176. Because $\mu < 0$, there exists $c > 0$ such that $\int_{-c}^{\infty} w\,\Gamma(dw) \leq \mu/2 < 0$. Thus taking $V(x) = x$ for $x > c$,

$$\begin{aligned} QV(x) - V(x) &= \int_{-\infty}^{\infty} [(x + w)_+ - x]\,\Gamma(dw) \\ &= -x\Gamma((-\infty\,, -x]) + \int_{-x}^{\infty} w\,\Gamma(dw) \leq \mu/2 . \end{aligned}$$

Hence the chain is positive recurrent.

Consider now the case $\mu > 0$. In view of Proposition 154, we have to show that the atom $\{0\}$ is transient. For any $n$, $X_n \geq X_0 + \sum_{i=1}^{n} W_i$. Define $C_n = \left\{ \left| n^{-1} \sum_{i=1}^{n} W_i - \mu \right| \geq \mu/2 \right\}$ and write $D_n$ for $\{X_n = 0\}$. The strong law of large numbers implies that $\mathrm{P}_0(D_n \text{ i.o.}) \leq \mathrm{P}_0(C_n \text{ i.o.}) = 0$. Hence the atom $\{0\}$ is transient, and so is the chain.

When $\mu = 0$, additional assumptions on $\Gamma$ are needed to prove the recurrence of the RWHL (see for instance Meyn and Tweedie, 1993, Lemma 8.5.2).

**Example 178** (Autoregressive Model, Continued)**.** Consider again the model of Example 148 and assume that the noise process has zero mean and finite variance. Choosing $V(x) = x^2$ we have

$$PV(x) = \mathrm{E}[(\phi x + U_1)^2] = \phi^2 V(x) + \mathrm{E}[U_1^2] \,,$$

so that (7.34) holds when $C = [-M, M]$ for some large enough $M$, provided $|\phi| < 1$. Because we know that every compact set is small if the noise process has an everywhere continuous positive density, Proposition 176 shows that the chain is positive recurrent. Note that this approach provides an existence result but does not help us to determine $\pi$. If $\{U_k\}$ are Gaussian with zero mean and variance $\sigma^2$, then one can check that the invariant distribution also is Gaussian with zero mean and variance $\sigma^2/(1-\phi^2)$.

Theorem 170 shows that if a chain is phi-irreducible and recurrent then the chain is positive, that is, it admits a unique invariant probability measure $\pi$. In certain situations, and in particular when dealing with MCMC procedures, it is known that $Q$ admits an invariant probability measure, but it is not known, *a priori*, that the chain is recurrent. The following result shows that positivity implies recurrence.

**Proposition 179.** *If the Markov kernel $Q$ is positive, then it is recurrent.*

*Proof.* Suppose that the chain is positive and let $\pi$ be an invariant probability measure. If $Q$ is transient, the state space $\mathsf{X}$ is covered by a countable family $\{A_j\}$ of uniformly transient subsets (see Theorem 151). For any $j$ and $k$,

$$k\pi(A_j) = \sum_{n=1}^{k} \pi Q^n(A_j) \le \int \pi(dx)\, \mathrm{E}_x[\eta_{A_j}] \le \sup_{x \in \mathsf{X}} \mathrm{E}_x[\eta_{A_j}] \,. \tag{7.38}$$

The strong Markov property implies that

$$\begin{aligned} \mathrm{E}_x[\eta_{A_j}] &= \mathrm{E}_x[\eta_{A_j} \mathbb{1}_{\{\sigma_{A_j} < \infty\}}] \\ &\le \mathrm{E}_x\{\mathbb{1}_{\{\sigma_{A_j} < \infty\}} \mathrm{E}_{X_{\sigma_{A_j}}}[\eta_{A_j}]\} \le \sup_{x \in A_j} \mathrm{E}_x[\eta_{A_j}]\, \mathrm{P}_x(\sigma_{A_j} < \infty) \,. \end{aligned}$$

Thus, the left-hand side of (7.38) is bounded as $k \to \infty$. This implies that $\pi(A_j) = 0$, and hence $\pi(\mathsf{X}) = 0$. This is a contradiction so the chain cannot be transient. □

### 7.2.4 Ergodicity

In this section, we study the convergence of iterates $Q^n$ of the transition kernel to the invariant distribution. As for discrete state spaces case, we first need to avoid periodic behavior that prevents the iterates to converge. In the discrete case, the period of a state $x$ is defined as the greatest common divisor of the set of time points $\{n \ge 0 : Q^n(x, x) > 0\}$. Of course this notion does not extend to general state spaces, but for phi-irreducible chains we may define the period of accessible small sets. More precisely, let $Q$ be a phi-irreducible transition kernel with maximal irreducibility measure $\psi$. By Theorem 156, there exists an accessible $(m, \epsilon, \psi)$-small set $C$. Because $\psi$ is a maximal irreducibility measure, $\psi(C) > 0$, so that when the chain starts from $C$ there is a positive probability that the it will return to $C$ at time $m$. Let

$$E_C \stackrel{\text{def}}{=} \{n \ge 1 : \text{the set } C \text{ is } (n, \epsilon_n, \psi)\text{-small for some } \epsilon_n > 0\} \tag{7.39}$$

be the set of time points for which $C$ is small with minorizing measure $\psi$. Note that for $n$ and $m$ in $E_C$, $B \in \mathcal{X}^+$ and $x \in C$,

$$Q^{n+m}(x, B) \geq \int_C Q^m(x, dx')\, Q^n(x', B) \geq \epsilon_m \epsilon_n \psi(C)\psi(B) > 0 \,,$$

showing that $E_C$ is closed under addition. There is thus a natural period for $E_C$, given by the greatest common divisor. Similar to the discrete case (see Proposition 144), this period $d$ may be shown to be independent of the particular choice of the small set $C$ (see for instance Meyn and Tweedie, 1993, Theorem 5.4.4).

**Proposition 180.** *Suppose that $Q$ is phi-irreducible with maximal irreducibility measure $\psi$. Let $C$ be an accessible $(m, \epsilon, \psi)$-small set and let $d$ be the greatest common divisor of the set $E_C$, defined in* (7.39)*. Then there exist disjoint sets $D_1, \ldots, D_d$ (a* $d$-cycle*) such that*

*(i) for $x \in D_i$, $Q(x, D_{i+1}) = 1$, $i = 0, \ldots, d-1$ (mod $d$);*

*(ii) the set $N = (\cup_{i=1}^d D_i)^c$ is $\psi$-null.*

*The $d$-cycle is maximal in the sense if $D'_1, \ldots, D'_{d'}$ is a $d'$-cycle, then $d'$ divides $d$, and if $d = d'$, then up to a permutation of indices $D'_i$ and $D_i$ are $\psi$-almost equal.*

It is obvious from the this theorem that the period $d$ does not depend on the choice of the small set $C$ and that any small set must be contained (up to $\psi$-null sets) inside one specific member of a $d$-cycle. This in particular implies that if there exists an accessible $(1, \epsilon, \psi)$-small set $C$, then $d = 1$. This suggests the following definition

**Definition 181** (Aperiodicity)**.** *Suppose that $Q$ is a phi-irreducible transition kernel with maximal irreducibility measure $\psi$. The largest $d$ for which a $d$-cycle exists is called the* period *of $Q$. When $d = 1$, the chain is called* aperiodic*. When there exists a $(1, \epsilon, \psi)$-small set $C$, the chain is called* strongly aperiodic*.*

In all the examples considered above, we have shown the existence of a 1-small set; therefore all these Markov chains are strongly aperiodic.

Now we can state the main convergence result, formulated and proved by Athreya *et al.* (1996). It parallels Theorem 145.

**Theorem 182.** *Let $Q$ be a phi-irreducible positive aperiodic transition kernel. Then for $\pi$-almost all $x$,*

$$\lim_{n \to \infty} \|Q^n(x, \cdot) - \pi\|_{\mathrm{TV}} = 0 \,. \tag{7.40}$$

*If $Q$ is Harris recurrent, the convergence occurs for all $x \in \mathsf{X}$.*

Although this result does not provide information on the rate of convergence to the invariant distribution, its assumptions are quite minimal. In fact, it may be shown that these assumptions are essentially necessary and sufficient. If $\|Q^n(x, \cdot) - \pi\|_{\mathrm{TV}} \to 0$ for any $x \in \mathsf{X}$, then by Nummelin (1984, Proposition 6.3), the chain is $\pi$-irreducible, aperiodic, positive Harris, and $\pi$ is an invariant distribution. This form of the ergodicity theorem is of particular interest in cases where the invariant distribution is explicitly known, as in Markov chain Monte Carlo. It provides conditions that are simple and easy to verify, and under which an MCMC algorithm converges to its stationary distribution.

Of course the exceptional null set for non-Harris recurrent chain is a nuisance. The example below however shows that there is no way of getting rid of it.

**Example 183.** In the model of Example 169, $\pi = \delta_0$ is an invariant probability measure. Because $Q^n(x,0) = \mathrm{P}_x(\tau_{\{0\}} \leq n)$ for any $n \geq 0$, $\lim_{n\to\infty} Q^n(x,0) = \mathrm{P}_x(\tau_{\{0\}} < \infty)$. We have previously shown that $\mathrm{P}_x(\tau_{\{0\}} < \infty) = 1 - \mathrm{P}_x(\tau_{\{0\}} = \infty) < 1$ for $x \geq 2$, whence $\limsup \|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}} \neq 0$ for such $x$.

Fortunately, in many cases it is not hard to show that a chain is Harris.

A proof of Theorem 182 from first principles is given by Athreya *et al.* (1996). We give here a proof due to Rosenthal (1995), based on pathwise coupling (see Rosenthal, 2001; Roberts and Rosenthal, 2004). The same construction is used to compute bounds on $\|Q^n(x,\cdot) - \pi\|_{\mathrm{TV}}$. Before proving the theorem, we briefly introduce the pathwise coupling construction for phi-irreducible Markov chains and present the associated Lindvall inequalities.

### Pathwise Coupling and Coupling Inequalities

Suppose that we have two probability measures $\xi$ and $\xi'$ on $(\mathsf{X}, \mathcal{X})$ that are such that $\frac{1}{2}\|\xi - \xi'\|_{\mathrm{TV}} \leq 1 - \epsilon$ for some $\epsilon \in (0,1]$ or, equivalently (see (3.6)), that there exists a probability measure $\nu$ such that $\epsilon\nu \leq \xi \wedge \xi'$. Because $\xi$ and $\xi'$ are probability measures, we may construct a probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and $\mathsf{X}$-valued random variables $X$ and $X'$ such that $\mathrm{P}(X \in \cdot) = \xi(\cdot)$ and $\mathrm{P}(X' \in \cdot) = \xi'$, respectively. By definition, for any $A \in \mathcal{X}$,

$$|\xi(A) - \xi'(A)| = |\,\mathrm{P}(X \in A) - \mathrm{P}(X' \in A)| = |\,\mathrm{E}[\mathbb{1}_A(X) - \mathbb{1}_A(X')]| \tag{7.41}$$

$$= |\,\mathrm{E}[(\mathbb{1}_A(X) - \mathbb{1}_A(X'))\mathbb{1}_{\{X \neq X'\}}]| \leq \mathrm{P}(X \neq X')\,, \tag{7.42}$$

so that the total variation distance between the laws of two random elements is bounded by the probability that they are unequal. Of course, this inequality is not in general sharp, but we can construct on an appropriately defined probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathrm{P}})$ two $\mathsf{X}$-valued random variables $X$ and $X'$ with laws $\xi$ and $\xi'$ such that $\tilde{\mathrm{P}}(X = X') \geq 1 - \epsilon$. The construction goes as follows. We draw a Bernoulli random variable $d$ with probability of success $\epsilon$. If $d = 0$, we then draw $X$ and $X'$ independently from the distributions $(1-\epsilon)^{-1}(\xi - \epsilon\nu)$ and $(1-\epsilon)^{-1}(\xi' - \epsilon\nu)$, respectively. If $d = 1$, we draw $X$ from $\nu$ and set $X = X'$. Note that for any $A \in \mathcal{X}$,

$$\begin{aligned}\tilde{\mathrm{P}}(X \in A) &= \tilde{\mathrm{P}}(X \in A \,|\, d = 0)\tilde{\mathrm{P}}(d = 0) + \tilde{\mathrm{P}}(X \in A \,|\, d = 1)\tilde{\mathrm{P}}(d = 1) \\ &= (1-\epsilon)\{(1-\epsilon)^{-1}[\xi(A) - \epsilon\nu(A)]\} = \xi(A)\end{aligned}$$

and, similarly, $\tilde{\mathrm{P}}(X' \in A) = \xi'(A)$. Thus, marginally the random variables $X$ and $X'$ are distributed according to $\xi$ and $\xi'$. By construction, $\tilde{\mathrm{P}}(X = X') \geq \mathrm{P}(d = 1) \geq \epsilon$, showing that $X$ and $X'$ are equal with probability at least $\epsilon$. Therefore the *coupling bound* (7.41) can be made sharp by using an appropriate construction. Note that this construction may be used to derive bounds on distances between probability measures that generalize the total variation; we will consider in the sequel the $V$-total variation.

**Definition 184** (V-Total Variation)**.** *Let $V : \mathsf{X} \to [1, \infty)$ be a measurable function. The $V$-*total variation distance *between two probability measures $\xi$ and $\xi'$ on $(\mathsf{X}, \mathcal{X})$ is*

$$\|\xi - \xi'\|_V \stackrel{\mathrm{def}}{=} \sup_{|f| \leq V} |\xi(f) - \xi'(f)|\,.$$

*If $V \equiv 1$, $\|\cdot\|_1$ is the total variation distance.*

When applied to Markov chains, the whole idea of coupling is to construct on an appropriately defined probability space two Markov chains $\{X_k\}$ and $\{X'_k\}$ with transition kernel $Q$ and initial distributions $\xi$ and $\xi'$, respectively, in such a way

that $X_n = X'_n$ for all indices $n$ after a random time $T$, referred to as the *coupling time*. The coupling procedure attempts to *couple* the two Markov chains when they simultaneously enter a coupling set.

**Definition 185** (Coupling Set)**.** *Let $\bar{C} \subseteq \mathsf{X} \times \mathsf{X}$, $\epsilon \in (0,1]$ and let $\boldsymbol{\nu} = \{\nu_{x,x'}, x, x' \in \mathsf{X}\}$ be transition kernels from $\bar{C}$ (endowed with the trace $\sigma$-field) to $(\mathsf{X}, \mathcal{X})$. The set $\bar{C}$ is a $(1, \epsilon, \boldsymbol{\nu})$*-coupling set *if for all $(x, x') \in \bar{C}$ and all $A \in \mathcal{X}$,*

$$Q(x, A) \wedge Q(x', A) \geq \epsilon \, \nu_{x,x'}(A) \; . \tag{7.43}$$

By applying Lemma 43, this condition can be stated equivalently as: there exists $\epsilon \in (0,1]$ such that for all $(x, x') \in \bar{C}$,

$$\frac{1}{2} \left\| Q(x, \cdot) - Q(x', \cdot) \right\|_{\mathrm{TV}} \leq 1 - \epsilon \; . \tag{7.44}$$

For simplicity, only one-step minorization is considered in this chapter. Adaptations to $m$-step minorization (replacing $Q$ by $Q^m$ in (7.43)) can be carried out as in Rosenthal (1995). Condition (7.43) is often satisfied by setting $\bar{C} = C \times C$ for a $(1, \epsilon, \nu)$-small set $C$. Indeed, in that case, for all $(x, x') \in C \times C$ and $A \in \mathcal{X}$,

$$Q(x, A) \wedge Q(x', A) \geq \epsilon \nu(A) \; .$$

The case $\epsilon = 1$ needs some consideration. If there exists an atom, say $\alpha$, i.e., there exists a probability measure $\nu$ such that for all $x \in \alpha$ and $A \in \mathcal{X}$, $Q(x, A) = \nu(A)$, then $\bar{C} = \alpha \times \alpha$ is a $(1, 1, \boldsymbol{\nu})$-coupling set with $\nu_{x,x'} = \nu$ for all $(x, x') \in \bar{C}$. Conversely, assume that $\bar{C}$ is a $(1, 1, \boldsymbol{\nu})$-coupling set. The alternative characterization (7.44) shows that $Q(x, \cdot) = Q(x', \cdot)$ for all $(x, x') \in \bar{C}$, that is, $\bar{C}$ is an atom. This also implies that the set $\bar{C}$ contains a set $\alpha_1 \times \alpha_2$, where $\alpha_1$ and $\alpha_2$ are atoms for $Q$.

We now introduce the coupling construction. Let $\bar{C}$ be a $(1, \epsilon, \boldsymbol{\nu})$-coupling set. Define $\bar{\mathsf{X}} = \mathsf{X} \times \mathsf{X}$ and $\bar{\mathcal{X}} = \mathcal{X} \otimes \mathcal{X}$. Let $\bar{Q}$ be a transition kernel on $(\bar{\mathsf{X}}, \bar{\mathcal{X}})$ given for all $A$ and $A'$ in $\mathcal{X}$ by

$$\begin{aligned} \bar{Q}(x, x'; A \times A') = {} & Q(x, A) Q(x', A') \mathbb{1}_{\bar{C}^c}(x, x') + \\ & (1 - \epsilon)^{-2} [Q(x, A) - \epsilon \nu_{x,x'}(A)][Q(x', A') - \epsilon \nu_{x,x'}(A')] \mathbb{1}_{\bar{C}}(x, x') \end{aligned} \tag{7.45}$$

if $\epsilon < 1$ and $\bar{Q} = Q \otimes Q$ if $\epsilon = 1$. For any probability measure $\bar{\mu}$ on $(\bar{\mathsf{X}}, \bar{\mathcal{X}})$, let $\bar{\mathrm{P}}_{\bar{\mu}}$ be the probability measure on the canonical space $(\bar{\mathsf{X}}^{\mathbb{N}}, \bar{\mathcal{X}}^{\mathbb{N}})$ such that the coordinate process $\{X_k\}$ is a Markov chain with respect to its natural filtration and with initial distribution $\bar{\mu}$ and transition kernel $\bar{Q}$. As usual, denote the associated expectation operator by $\bar{\mathrm{E}}_{\bar{\mu}}$.

We now define a transition kernel $\tilde{Q}$ on the space $\tilde{X} \stackrel{\text{def}}{=} \mathsf{X} \times \mathsf{X} \times \{0, 1\}$ endowed with the product $\sigma$-field $\tilde{\mathcal{X}}$ by, for any $x, x' \in \mathsf{X}$ and $A, A' \in \mathcal{X}$,

$$\tilde{Q}\left((x, x', 0), A \times A' \times \{0\}\right) = [1 - \epsilon \mathbb{1}_{\bar{C}}(x, x')] \bar{Q}((x, x'), A \times A') \; , \tag{7.46}$$

$$\tilde{Q}\left((x, x', 0), A \times A' \times \{1\}\right) = \epsilon \mathbb{1}_{\bar{C}}(x, x') \nu_{x,x'}(A \cap A') \; , \tag{7.47}$$

$$\tilde{Q}\left((x, x', 1), A \times A' \times \{1\}\right) = Q(x, A \cap A') \; . \tag{7.48}$$

For any probability measure $\tilde{\mu}$ on $(\tilde{\mathsf{X}}, \tilde{\mathcal{X}})$, let $\tilde{\mathrm{P}}_{\tilde{\mu}}$ be the probability measure on the canonical space $(\tilde{\mathsf{X}}^{\mathbb{N}}, \tilde{\mathcal{X}}^{\otimes \mathbb{N}})$ such that the coordinate process $\{\tilde{X}_k\}$ is a Markov chain with transition kernel $\tilde{Q}$ and initial distribution $\tilde{\mu}$. The corresponding expectation operator is denoted by $\tilde{\mathrm{E}}_{\tilde{\mu}}$.

The transition kernel $\tilde{Q}$ can be described algorithmically. Given $\tilde{X}_0 = (X_0, X'_0, d_0) = (x, x', d)$, $\tilde{X}_1 = (X_1, X'_1, d_1)$ is obtained as follows.

- If $d = 1$, then draw $X_1$ from $Q(x, \cdot)$ and set $X_1' = X_1$, $d_1 = 1$.
- If $d = 0$ and $(x, x') \in \bar{C}$, flip a coin with probability of heads $\epsilon$.
  - If the coin comes up heads, draw $X_1$ from $\nu_{x,x'}$ and set $X_1' = X_1$ and $d_1 = 1$.
  - If the coin comes up tails, draw $(X_1, X_1')$ from $\bar{Q}(x, x'; \cdot)$ and set $d_1 = 0$.
- If $d = 0$ and $(x, x') \notin \bar{C}$, draw $(X_1, X_1')$ from $\bar{Q}(x, x'; \cdot)$ and set $d_1 = 0$.

The variable $d_n$ is called the *bell variable*; it indicates whether coupling has occurred by time $n$ ($d_n = 1$) or not ($d_n = 0$). The first index $n$ at which $d_n = 1$ is the coupling time;

$$T = \inf\{k \geq 1 : d_k = 1\}.$$

If $d_n = 1$, then $X_k = X_k'$ for all $k \geq n$. The coupling construction is carried out in such a way that under $\tilde{\mathrm{P}}_{\xi \otimes \xi' \otimes \delta_0}$, $\{X_k\}$ and $\{X_k'\}$ are Markov chains with transition kernel $Q$ with initial distributions $\xi$ and $\xi'$, respectively.

The coupling construction allows deriving quantitative bounds on the ($V$-)total variation distance in terms of the tail probability of the coupling time.

**Proposition 186.** *Assume that the transition kernel $Q$ admits a $(1, \epsilon, \boldsymbol{\nu})$-coupling set. Then for any probability measures $\xi$ and $\xi'$ on $(\mathsf{X}, \mathcal{X})$ and any measurable function $V : \mathsf{X} \to [1, \infty)$,*

$$\|\xi Q^n - \xi' Q^n\|_{\mathrm{TV}} \leq 2\tilde{\mathrm{P}}_{\xi \otimes \xi' \otimes \delta_0}(T > n) \,, \tag{7.49}$$

$$\|\xi Q^n - \xi' Q^n\|_V \leq 2\tilde{\mathrm{E}}_{\xi \otimes \xi' \otimes \delta_0}[\bar{V}(X_n, X_n')\mathbb{1}_{\{T > n\}}] \,, \tag{7.50}$$

*where $\bar{V} : \mathsf{X} \times \mathsf{X} \to [1, \infty)$ is defined by $\bar{V}(x, x') = \{V(x) + V(x')\}/2$.*

*Proof.* We only need to prove (7.50) because (7.49) is obtained by setting $V \equiv 1$. Pick a function $f$ such that $|f| \leq V$ and note that $[f(X_n) - f(X_n')]\mathbb{1}_{\{d_n = 1\}} = 0$. Hence

$$\begin{aligned} |\xi Q^n f - \xi' Q^n f| &= |\tilde{\mathrm{E}}_{\xi \otimes \xi' \otimes \delta_0}[f(X_n) - f(X_n')]| \\ &= |\tilde{\mathrm{E}}_{\xi \otimes \xi' \otimes \delta_0}[(f(X_n) - f(X_n'))\mathbb{1}_{\{d_n = 0\}}]| \\ &\leq 2\tilde{\mathrm{E}}_{\xi \otimes \xi' \otimes \delta_0}[\bar{V}(X_n, X_n')\mathbb{1}_{\{d_n = 0\}}] \,. \end{aligned}$$

□

We now provide an alternative expression of the coupling inequality that only involves the process $\{\bar{X}_k\}$. Let $\sigma_{\bar{C}}$ be the hitting time on the coupling set $\bar{C}$ by this process, define $K_0(\epsilon) = 1$, and for all $n \geq 1$,

$$K_n(\epsilon) = \begin{cases} \mathbb{1}_{\{\sigma_{\bar{C}} \geq n\}} & \text{if } \epsilon = 1 \,; \\ \prod_{j=0}^{n-1}[1 - \epsilon \mathbb{1}_{\bar{C}}(\bar{X}_j)] & \text{if } \epsilon \in (0, 1) \,. \end{cases} \tag{7.51}$$

**Proposition 187.** *Assume that the transition kernel $Q$ admits a $(1, \epsilon, \boldsymbol{\nu})$-coupling set. Let $\xi$ and $\xi'$ be probability measures on $(\mathsf{X}, \mathcal{X})$ and let $V : \mathsf{X} \to [1, \infty)$ be a measurable function. Then*

$$\|\xi Q^n - \xi' Q^n\|_V \leq 2\bar{\mathrm{E}}_{\xi \otimes \xi'}[\bar{V}(X_n, X_n')K_n(\epsilon)] \,, \tag{7.52}$$

*with $\bar{V}(x, x') = [V(x) + V(x')]/2$.*

*Proof.* We show that for any probability measure $\bar{\mu}$ on $(\bar{\mathsf{X}}, \bar{\mathcal{X}})$,

$$\tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}[\bar{V}(X_n, X'_n)\mathbb{1}_{\{T>n\}}] = \bar{\mathrm{E}}_{\bar{\mu}}[\bar{V}(X_n, X'_n)K_n(\epsilon)] \; .$$

To do this, we shall prove by induction that for any $n \geq 0$ and any bounded $\bar{\mathcal{X}}$-measurable functions $\{f_j\}_{j\geq 0}$,

$$\tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}\left[\prod_{j=0}^{n} f_j(X_j, X'_j)\,\mathbb{1}_{\{T>n\}}\right] = \bar{\mathrm{E}}_{\bar{\mu}}\left[\prod_{j=0}^{n} f_j(X_j, \bar{X}_j)\,K_n(\epsilon)\right] \; . \tag{7.53}$$

This is obviously true for $n = 0$. For $n \geq 0$, put $\chi_n = \prod_{j=0}^{n} f_j(X_j, X'_j)$. The induction assumption and the identity $\{T > n+1\} = \{d_{n+1} = 0\}$ yield

$$\begin{aligned}
\tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}[\chi_{n+1}\mathbb{1}_{\{T>n+1\}}] &= \tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}[\chi_n\ f_{n+1}(X_{n+1}, X'_{n+1})\mathbb{1}_{\{d_{n+1}=0\}}] \\
&= \tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}\{\chi_n\ \tilde{\mathrm{E}}[f_{n+1}(X_{n+1}, X'_{n+1})\mathbb{1}_{\{d_{n+1}=0\}} \,|\, \tilde{\mathcal{F}}_n]\mathbb{1}_{\{d_n=0\}}\} \\
&= \tilde{\mathrm{E}}_{\bar{\mu}\otimes\delta_0}\{\chi_n[1 - \epsilon\mathbb{1}_{\bar{C}}(X_n, X'_n)]\bar{Q}f_{n+1}(X_n, X'_n)\mathbb{1}_{\{d_n=0\}}\} \\
&= \bar{\mathrm{E}}_{\bar{\mu}}[\chi_n\bar{Q}f_{n+1}(\bar{X}_n)K_{n+1}(\epsilon)] = \bar{\mathrm{E}}_{\bar{\mu}}[\chi_{n+1}K_{n+1}(\epsilon)] \; .
\end{aligned}$$

This concludes the induction and the proof. □

**Proof of Theorem 182**

We preface the proof of Theorem 182 by two technical lemmas that establish some elementary properties of a chain on the product space with transition kernel $Q \otimes Q$.

**Lemma 188.** *Suppose that $Q$ is a phi-irreducible aperiodic transition kernel. Then for any $n$, $Q^n$ is phi-irreducible and aperiodic.*

*Proof.* Propositions 156 and 157 show that there exists an accessible $(m, \epsilon, \nu)$-small set $C$ and that $\nu$ is an irreducibility measure. Because $Q$ is aperiodic, there exists a sequence $\{\epsilon_k\}$ of positive numbers and an integer $n_C$ such that for all $n \geq n_C$, $x \in C$, and $A \in \mathcal{X}$, $Q^n(x, A) \geq \epsilon_n\nu(A)$. In addition, because $C$ is accessible, there exists $p$ such that $Q^p(x, C) > 0$ for any $x \in \mathsf{X}$. Therefore for any $n \geq n_C$ and any $A \in \mathcal{X}$ such that $\nu(A) > 0$,

$$Q^{n+p}(x, A) \geq \int_C Q^p(x, dx')\,Q^n(x', A) \geq \epsilon_n\nu(A)Q^p(x, C) > 0 \; . \tag{7.54}$$

□

**Lemma 189.** *Let $Q$ be an aperiodic positive transition kernel with invariant probability measure $\pi$. Then $Q \otimes Q$ is phi-irreducible, $\pi \otimes \pi$ is $Q \otimes Q$-invariant, and $Q \otimes Q$ is positive. If $C$ is an accessible $(m, \epsilon, \nu)$-small set for $Q$, then $C \times C$ is an accessible $(m, \epsilon^2, \nu \otimes \nu)$-small set for $Q \otimes Q$.*

*Proof.* Because $Q$ is phi-irreducible and admits $\pi$ as an invariant probability measure, $\pi$ is a maximal irreducibility measure for $Q$. Let $C$ be an accessible $(m, \epsilon, \nu)$-small set for $Q$. Then for $(x, x') \in C \times C$ and $A \in \mathcal{X} \otimes \mathcal{X}$,

$$(Q \otimes Q)^m(x, x'; A) = \iint_A Q^m(x, dy)\,Q^m(x', dy') \geq \epsilon^2\nu \otimes \nu(A) \; .$$

Because $\nu \otimes \nu(C \times C) = [\nu(C)]^2 > 0$, this shows that $C \times C$ is a $(1, \epsilon^2, \nu \otimes \nu)$-small set for $Q \otimes Q$. By (7.54) there exists an integer $n_x$ such that for any $n \geq n_x$, $Q^n(x, C) > 0$. This implies that for any $(x, x') \in \mathsf{X} \times \mathsf{X}$ and any $n \geq n_x \vee n_{x'}$,

$$(Q \otimes Q)^n(x, x'\,; C \times C) = Q^n(x, C)Q^n(x', C) > 0 \; ,$$

showing that $C \times C$ is accessible. Because $C \times C$ is a small set, Proposition 156 shows that $Q \otimes Q$ is phi-irreducible. In addition, $\pi \otimes \pi$ is invariant for $Q \otimes Q$, so that $\pi \otimes \pi$ is a maximal irreducibility measure and $Q \otimes Q$ is positive. □

We have now all the necessary ingredients to prove Theorem 182.

*of Theorem 182.* By Lemma 188, $Q^m$ is phi-irreducible for any integer $m$. By Proposition 157, there exists an accessible $(m, \epsilon, \nu)$-small set $C$ with $\nu(C) > 0$. Lemma 46 shows that for all integers $n$,

$$\|Q^n(x,\cdot) - Q^n(x',\cdot)\|_{\mathrm{TV}} \leq \|Q^{m[n/m]}(x,\cdot) - Q^{m[n/m]}(x',\cdot)\|_{\mathrm{TV}} \,.$$

Hence it suffices to prove that (7.40) holds for $Q^m$ and we may thus without loss of generality assume that $m = 1$.

For any probability measure $\mu$ on $(\mathsf{X} \times \mathsf{X}, \mathcal{X} \otimes \mathcal{X})$, let $\mathrm{P}^\star_\mu$ denote the probability measure on the canonical space $((\mathsf{X}\times\mathsf{X})^{\mathbb{N}}, (\mathcal{X}\otimes\mathcal{X})^{\otimes\mathbb{N}})$ such that the canonical process $\{(X_k, X'_k)\}_{k\geq 0}$ is a Markov chain with transition kernel $Q\otimes Q$ and initial distribution $\mu$. By Lemma 189, $Q \otimes Q$ is positive, and it is recurrent by Proposition 179.

Because $\pi \otimes \pi(C \times C) = \pi^2(C) > 0$, by Theorem 168 there exist two measurable sets $\bar{C} \subseteq C \times C$ and $\bar{H} \subseteq \mathsf{X} \times \mathsf{X}$ such that $\pi \otimes \pi(C \times C \setminus \bar{C}) = 0$, $\pi \times \pi(H) = 1$, and for all $(x, x') \in \bar{H}$, $\mathrm{P}^\star_{x,x'}(\tau_{\bar{C}} < \infty) = 1$. Moreover, the set $\bar{C}$ is a $(1, \epsilon, \boldsymbol{\nu})$-coupling set with $\nu_{x,x'} = \nu$ for all $(x, x') \in \bar{C}$.

Let the transition kernel $\bar{Q}$ be defined by (7.45) if $\epsilon < 1$ and by $\bar{Q} = Q \otimes Q$ if $\epsilon = 1$. For $\epsilon = 1$, $\bar{\mathrm{P}}_{x,x'} = \mathrm{P}^\star_{x,x'}$. Now assume that $\epsilon \in (0, 1)$. For $(x, x') \notin \bar{C}$, $\bar{\mathrm{P}}_{x,x'}(\tau_{\bar{C}} = \infty) = \mathrm{P}^\star_{x,x'}(\tau_{\bar{C}} = \infty)$. For $(x, x') \in \bar{C}$, noting that $\bar{Q}(x, x', A) \leq (1-\epsilon)^{-2} Q \otimes Q(x, x', A)$ we obtain

$$\begin{aligned}
\bar{\mathrm{P}}_{x,x'}(\tau_{\bar{C}} = \infty) &= \bar{\mathrm{P}}_{x,x'}(\tau_{\bar{C}} = \infty \,|\, (X_1, X'_1) \notin C \times C)\, \bar{Q}(x, x', \bar{C}^c) \\
&\leq (1-\epsilon)^{-2} Q \otimes Q(x, x', \bar{C}^c)\, \mathrm{P}^\star_{x,x'}(\tau_{\bar{C}} = \infty \,|\, \bar{X}_1 \notin \bar{C}) \\
&= (1-\epsilon)^{-2}\, \mathrm{P}^\star_{x,x'}(\tau_{\bar{C}} = \infty) = 0 \,.
\end{aligned}$$

Thus, for all $\epsilon \in (0, 1]$ the set $\bar{C}$ is Harris-recurrent for the kernel $\bar{Q}$. This implies that $\lim_{n\to\infty} \bar{\mathrm{E}}_{x,x'}[K_n(\epsilon)] = 0$ for all $(x, x') \in \bar{H}$ and, using Proposition 187, we conclude that (7.40) is true. □

### 7.2.5 Geometric Ergodicity and Foster-Lyapunov Conditions

Theorem 182 implies forgetting of the initial distribution and convergence to stationarity but does not provide us with rates of convergence. In this section, we show how to adapt the construction above to derive explicit bounds on $\|\xi Q^n - \xi' Q^n\|_V$. We focus on conditions that imply geometric convergence.

**Definition 190** (Geometric Ergodicity)**.** *A positive aperiodic transition kernel $Q$ with invariant probability measure $\pi$ is said to be* $V$-geometrically ergodic *if there exist constants $\rho \in (0, 1)$ and $M < \infty$ such that*

$$\|Q^n(x,\cdot) - \pi\|_V \leq M V(x) \rho^n \quad \textit{for } \pi\textit{-almost all } x. \tag{7.55}$$

We now present conditions that ensure geometric ergodicity.

**Definition 191** (Foster-Lyapunov Drift Condition)**.** *A transition kernel $Q$ is said to satisfy a Foster-Lyapunov drift condition outside a set $C \in \mathcal{X}$ if there exists a measurable function $V : \mathsf{X} \to [1, \infty]$, bounded on $C$, and non-negative constants $\lambda < 1$ and $b < \infty$ such that*

$$QV \leq \lambda V + b \mathbb{1}_C \,. \tag{7.56}$$

If $Q$ is phi-irreducible and satisfies a Foster-Lyapunov condition outside a small set $C$, then $C$ is accessible and, writing $QV \leq V - (1-\lambda)V + b\mathbb{1}_C$, Proposition 176 shows that $Q$ is positive and $\pi(V) < \infty$.

**Example 192** (Random Walk on the Half-Line, Continued)**.** Assume that for the model of Example 153 there exists $z > 0$ such that $\mathrm{E}[\mathrm{e}^{zW_1}] < \infty$. Then because $\mu < 0$, there exists $z > 0$ such that $\mathrm{E}[\mathrm{e}^{zW_1}] < 1$. Define $z_0 = \arg\min_{z>0} \mathrm{E}[\mathrm{e}^{zW_1}]$ and $V(x) = \mathrm{e}^{z_0 x}$, and choose $x_0 > 0$ such that $\lambda = \mathrm{E}[\mathrm{e}^{z_0 W_1}] + \mathrm{P}(W_1 < -x_0) < 1$. Then for $x > x_0$,

$$QV(x) = \mathrm{E}[\mathrm{e}^{z_0(x+W_1)_+}] = \mathrm{P}(W_1 \leq -x) + \mathrm{e}^{z_0 x}\,\mathrm{E}[\mathrm{e}^{z_0 W_1}\mathbb{1}_{\{W_1 > -x\}}] \leq \lambda V(x)\ .$$

Hence the Foster-Lyapunov drift condition holds outside the small set $[0, x_0]$, and the RWHL is geometrically ergodic. For a sharper choice of the constants $z_0$ and $\lambda$, see Scott and Tweedie (1996, Theorem 4.1).

**Example 193** (Metropolis-Hastings Algorithm, Continued)**.** Consider the Metropolis-Hastings algorithm of Example 149 with random walk proposal kernel $r(x, x') = r(|x - x'|)$. Geometric ergodicity of the Metropolis-Hastings algorithm on $\mathbb{R}^d$ is largely a property of the tails of the stationary distribution $\pi$. Conditions for geometric ergodicity can be shown to be, essentially, that the tails are exponential or lighter (Mengersen and Tweedie, 1996) and that in higher dimensions the contours of $\pi$ are regular near $\infty$ (see for instance Jarner and Hansen, 2000). To understand how the tail conditions come into play, consider the case where $\pi$ is a probability density on $\mathsf{X} = \mathbb{R}^+$. We suppose that $\pi$ is log-concave in the upper tail, that is, that there exists $\alpha > 0$ and $M$ such that for all $x' \geq x \geq M$,

$$\log \pi(x) - \log \pi(x') \geq \alpha(x' - x)\ . \tag{7.57}$$

To simplify the proof, we assume that $\pi$ is non-increasing, but this assumption is unnecessary. Define $A_x = \{x' \in \mathbb{R}^+ : \pi(x') \leq \pi(x)\}$ and $R_x = \{x' \in \mathbb{R}^+, \pi(x) > \pi(x')\}$, the acceptance and (possible) rejection regions for the chain started from $x$. Because $\pi$ is non-increasing, these sets are simple: $A_x = [0, x]$ and $R_x = (x, \infty) \cup (-\infty, 0)$. If we relax the monotonicity conditions, the acceptance and rejection regions become more involved, but because $\pi$ is log-concave and thus in particular monotone in the upper tail, $A_x$ and $R_x$ are essentially intervals when $x$ is sufficiently large.

For any function $V : \mathbb{R}^+ \to [1, +\infty)$ and $x \in \mathbb{R}^+$,

$$\frac{QV(x)}{V(x)} = 1 + \int_{A_x} r(x' - x)\left[\frac{V(x')}{V(x)} - 1\right] dx' + \int_{R_x} r(x' - x)\frac{\pi(x')}{\pi(x)}\left[\frac{V(x')}{V(x)} - 1\right] dx'\ .$$

We set $V(x) = \mathrm{e}^{sx}$ for some $s \in (0, \alpha)$. Because $\pi$ is log-concave, $\pi(x')/\pi(x) \leq \mathrm{e}^{-\alpha(x'-x)}$ when $x' \geq x \geq M$. For $x \geq M$, it follows from elementary calculations that

$$\limsup_{x\to\infty} \frac{QV(x)}{V(x)} \leq 1 - \int_0^\infty r(u)(1 - \mathrm{e}^{-su})[1 - \mathrm{e}^{-(\alpha-s)u}]\, du < 1\ ,$$

showing that the random walk Metropolis-Hastings algorithm on the positive real line satisfies the Foster-Lyapunov condition when $\pi$ is monotone and log-concave in the upper tail.

The main result guaranteeing geometric ergodicity is the following.

**Theorem 194.** *Let $Q$ be a phi-irreducible aperiodic positive transition kernel with invariant distribution $\pi$. Also assume that $Q$ satisfies a Foster-Lyapunov drift condition outside a small set $C$ with drift function $V$. Then $\pi(V)$ is finite and $Q$ is $V$-geometrically ergodic.*

In fact, it follows from Meyn and Tweedie (1993, Theorems 15.0.1 and 16.0.1) that the converse is also true: if a phi-irreducible aperiodic kernel is $V$-geometrically ergodic, then there exists an accessible small set $C$ such that $V$ is a drift function outside $C$.

For the sake of brevity and simplicity, we now prove Theorem 194 under the additional assumption that the level sets of $V$ are all $(1, \epsilon, \nu)$-small. In that case, it is possible to define a coupling set $\bar{C}$ and a transition kernel $\bar{Q}$ that satisfies a (bivariate) Foster-Lyapunov drift condition outside $\bar{C}$. The geometric ergodicity of the transition kernel $Q$ is then proved under this assumption. This is the purpose of the following propositions.

**Proposition 195.** *Let $Q$ be a kernel that satisfies the Foster-Lyapunov drift condition (7.56) with respect to a $(1, \epsilon, \nu)$-small set $C$ and a function $V$ whose level sets are $(1, \epsilon, \nu)$-small. Then for any $d > 1$, the set $C' = C \cup \{x \in \mathsf{X} : V(x) \leq d\}$ is small, $C' \times C'$ is a $(1, \epsilon, \nu)$-coupling set, and the kernel $\bar{Q}$, defined as in (7.45), satisfies the drift condition (7.58) with $\bar{C} = C' \times C'$, $\bar{V}(x, x') = (1/2)[V(x) + V(x')]$, and $\bar{\lambda} = \lambda + b/(1 + d)$ provided $\bar{\lambda} < 1$.*

*Proof.* For $(x, x') \notin \bar{C}$ we have $(1 + d)/2 \leq \bar{V}(x, x')$. Therefore

$$\bar{Q}\bar{V}(x, x') \leq \lambda \bar{V}(x, x') + \frac{b}{2} \leq \left(\lambda + \frac{b}{1 + d}\right) \bar{V}(x, x') \; ,$$

and for $(x, x') \in \bar{C}$ it holds that

$$\begin{aligned} \bar{Q}\bar{V}(x, x') &= \frac{1}{2(1 - \epsilon)} [QV(x) + QV(x') - 2\epsilon\nu(V)] \\ &\leq \frac{\lambda}{(1 - \epsilon)} \bar{V}(x, x') + \frac{b - \epsilon\nu(V)}{1 - \epsilon} \; . \end{aligned}$$

□

**Proposition 196.** *Assume that $Q$ admits a $(1, \epsilon, \boldsymbol{\nu})$-coupling set $\bar{C}$ and that there exists a choice of the kernel $\bar{Q}$ for which there is a measurable function $\bar{V} : \bar{\mathsf{X}} \to [1, \infty)$, $\bar{\lambda} \in (0, 1)$ and $\bar{b} > 0$ such that*

$$\bar{Q}\bar{V} \leq \bar{\lambda}\bar{V} + \bar{b}\mathbb{1}_{\bar{C}} \; . \tag{7.58}$$

*Let $W : \mathsf{X} \to [1, \infty)$ be a measurable function such that $W(x) + W(x') \leq 2\bar{V}(x, x')$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$. Then there exist $\rho \in (0, 1)$ and $c > 0$ such that for all $(x, x') \in \mathsf{X} \times \mathsf{X}$,*

$$\|Q^n(x, \cdot) - Q^n(x', \cdot)\|_W \leq c\bar{V}(x, x')\rho^n \; . \tag{7.59}$$

*Proof.* By Proposition 186, proving (7.59) amounts to proving the requested bound for $\bar{\mathrm{E}}_{x,x'}[\bar{V}(\bar{X}_n)K_n(\epsilon)]$. We only consider the case $\epsilon \in (0, 1)$, the case $\epsilon = 1$ being easier. Write $\bar{x} = (x, x')$. By induction, the drift condition (7.58) implies that

$$\bar{\mathrm{E}}_{\bar{x}}[\bar{V}(\bar{X}_n)] = \bar{Q}^n\bar{V}(\bar{x}) \leq \bar{\lambda}^n\bar{V}(\bar{x}) + \bar{b}\sum_{j=0}^{n-1} \bar{\lambda}^j \leq \bar{V}(\bar{x}) + \bar{b}/(1 - \bar{\lambda}) \; . \tag{7.60}$$

Recall that $K_n(\epsilon) = (1-\epsilon)^{\eta_n(\bar{C})}$ for $\epsilon \in (0,1)$, where $\eta_n(\bar{C}) = \sum_0^{n-1} \mathbb{1}_{\bar{C}}(X_j)$ is the number of visits to the coupling set $\bar{C}$ before time $n$. Hence $K_n(\epsilon)$ is $\bar{\mathcal{F}}_{n-1}$-measurable. Let $j \le n+1$ be an arbitrary positive integer to be chosen later. Then (7.60) yields

$$\begin{aligned}\bar{\mathrm{E}}_{\bar{x}}[\bar{V}(\bar{X}_n)K_n(\epsilon)\mathbb{1}_{\{\eta_n(\bar{C})\ge j\}}] &\le (1-\epsilon)^j \bar{\mathrm{E}}_{\bar{x}}[\bar{V}(\bar{X}_n)]\mathbb{1}_{\{j\le n\}} \\ &\le [\bar{V}(\bar{x}) + \bar{b}/(1-\bar{\lambda})](1-\epsilon)^j \mathbb{1}_{\{j\le n\}} \,. \end{aligned} \tag{7.61}$$

Put $M = \sup_{\bar{x}\in\bar{C}} \bar{Q}\bar{V}(\bar{x})/V(\bar{x})$ and $B = 1 \vee [M(1-\epsilon)/\bar{\lambda}]$. For $k = 0, \dots, n$, define $Z_k = \bar{\lambda}^{-k}[(1-\epsilon)/B]^{\eta_k(\bar{C})}\bar{V}(\bar{X}_k)$. Because $\eta_n(\bar{C})$ is $\bar{\mathcal{F}}_{n-1}$-measurable, we obtain

$$\begin{aligned}\bar{\mathrm{E}}_{\bar{x}}[Z_n \,|\, \bar{\mathcal{F}}_{n-1}] &= \bar{\lambda}^{-n}\bar{Q}\bar{V}(\bar{X}_{n-1})[(1-\epsilon)/B]^{\eta_n(\bar{C})} \\ &\le \bar{\lambda}^{-n+1}\bar{V}(\bar{X}_{n-1})[(1-\epsilon)/B]^{\eta_n(\bar{C})}\mathbb{1}_{\bar{C}^c}(\bar{X}_{n-1}) \\ &\quad + \bar{\lambda}^{-n}M\bar{V}(\bar{X}_{n-1})[(1-\epsilon)/B]^{\eta_n(\bar{C})}\mathbb{1}_{\bar{C}}(\bar{X}_{n-1}) \,.\end{aligned}$$

Using the relations $\eta_n(\bar{C}) = \eta_{n-1}(\bar{C}) + 1_{\bar{C}}(\bar{X}_{n-1})$ and $M(1-\epsilon) \le B\bar{\lambda}$, we find that $\bar{\mathrm{E}}_{\bar{x}}[Z_n \,|\, \bar{\mathcal{F}}_{n-1}] \le Z_{n-1}$ and, by induction, $\bar{\mathrm{E}}_{\bar{x}}[Z_n] \le \bar{\mathrm{E}}_{\bar{x}}[Z_0] = \bar{V}(\bar{x})$. Hence, as $B \ge 1$,

$$\bar{\mathrm{E}}_{\bar{x}}[\bar{V}(\bar{X}_n)K_n(\epsilon)\mathbb{1}_{\{\eta_n(\bar{C})<j\}}] \le \bar{\lambda}^n B^j \bar{\mathrm{E}}_{\bar{x}}[Z_n] \le \bar{\lambda}^n B^j \; \bar{V}(\bar{x}) \,. \tag{7.62}$$

Gathering (7.61) and (7.62) yields

$$\bar{\mathrm{E}}_{\bar{x}}[\bar{V}(\bar{X}_n)K_n(\epsilon)] \le [\bar{V}(\bar{x}) + \bar{b}/(1-\bar{\lambda})]\,[(1-\epsilon)^j \mathbb{1}_{\{j\le n\}} + \bar{\lambda}^n B^j] \,.$$

If $B = 1$, choosing $j = n+1$ yields (7.59) with $\rho = \bar{\lambda}$, and if $B > 1$ then set $j = [\alpha n]$ with $\alpha \in (0,1)$ such that $\log(\bar{\lambda}) + \alpha \log(B) < 0$; this choice yields (7.59) with $\rho = (1-\epsilon)^\alpha \vee (\bar{\lambda}B^\alpha) < 1$. □

**Example 197** (Autoregressive Model, Continued)**.** In the model of Example 148, we have verified that $V(x) = 1 + x^2$ satisfies (7.56) when the noise variance is finite. We can deduce from Theorem 194 a variety of results: the stationary distribution has finite variance and the iterates $Q^n(x,\cdot)$ of the transition kernel converge to the stationary distribution $\pi$ geometrically fast in $V$-total variation distance. Thus there exist constants $C$ and $\rho < 1$ such that for any $x \in \mathsf{X}$, $\|Q^n(x,\cdot) - \pi\|_V \le C(1+x^2)\rho^n$. This implies in particular that for any $x \in \mathsf{X}$ and any function $f$ such that $\sup_{x\in\mathsf{X}}(1+x^2)^{-1}|f(x)| < \infty$, $\mathrm{E}_x[f(X_n)]$ converges to the limiting value

$$\mathrm{E}_\pi[f(X_n)] = \sqrt{\frac{1-\phi^2}{2\pi\sigma^2}} \int \exp\left[-\frac{(1-\phi^2)x^2}{2\sigma^2}\right] f(x)\,dx$$

geometrically fast. This applies for the mean, $f(x) = x$, and the second moment, $f(x) = x^2$ (though in this case convergence can be derived directly from the autoregression).

### 7.2.6 Limit Theorems

One of the most important problems in probability theory is the investigation of limit theorems for appropriately normalized sums of random variables. The case of independent random variables is fairly well understood, but less is known about dependent random variables such as Markov chains. The purpose of this section is to study several basic limit theorems for additive functionals of Markov chains.

**Law of Large Numbers**

Suppose that $\{X_k\}$ is a Markov chain with transition kernel $Q$ and initial distribution $\nu$. Assume that $Q$ is phi-irreducible and aperiodic and has a stationary distribution $\pi$. Let $f$ be a $\pi$-integrable function; $\pi(|f|) < \infty$. We say that the sequence $\{f(X_k)\}$ satisfies a law of large numbers (LLN) if for any initial distribution $\nu$ on $(\mathsf{X}, \mathcal{X})$, the sample mean $n^{-1} \sum_{k=1}^{n} f(X_k)$ converges to $\pi(f)$ $\mathrm{P}_\nu$-a.s.

For i.i.d. samples, classical theory shows that the LLN holds provided $\pi(|f|) < \infty$. The following theorem shows that the LLN holds for ergodic Markov chains; it does not require any conditions on the rate of convergence to the stationary distribution.

**Theorem 198.** *Let $Q$ be a positive Harris recurrent transition kernel with invariant distribution $\pi$. Then for any real $\pi$-integrable function $f$ on $\mathsf{X}$ and any initial distribution $\nu$ on $(\mathsf{X}, \mathcal{X})$,*

$$\lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} f(X_k) = \pi(f) \quad \mathrm{P}_\nu\text{-a.s.} \tag{7.63}$$

The LLN can be obtained from general ergodic theorems for stationary processes. An elementary proof can be given when the chain possesses an accessible atom. The basic technique is then the regeneration method, which consists in dividing the chain into blocks between the chain's successive returns to the atom. These blocks are independent (see Lemma 199 below) and standard limit theorems for i.i.d. random variables yield the desired result. When the chain has no atom, one may still employ this technique by replacing the atom by a suitably chosen small set and using the splitting technique (see for instance Meyn and Tweedie, 1993, Chapter 17).

**Lemma 199.** *Let $Q$ be a positive Harris recurrent transition kernel that admits an accessible atom $\alpha$. Define for any measurable function $f$,*

$$s_j(f) = \left( \sum_{k=1}^{\tau_\alpha} f(X_k) \right) \circ \theta^{\tau_\alpha^{(j-1)}} , \qquad j \geq 1 . \tag{7.64}$$

*Then for any initial distribution $\nu$ on $(\mathsf{X}, \mathcal{X})$, $k \geq 0$ and functions $\{\Psi_j\}$ in $\mathcal{F}_{\mathrm{b}}(\mathbb{R})$,*

$$\mathrm{E}_\nu \left[ \prod_{j=1}^{k} \Psi_j(s_j(f)) \right] = \mathrm{E}_\nu \left[ \Psi_1(s_1(f)) \right] \prod_{j=2}^{k} \mathrm{E}_\alpha \left[ \Psi_j(s_j(f)) \right] .$$

*Proof.* Because the atom $\alpha$ is accessible and the chain is Harris recurrent, $\mathrm{P}_x(\tau_\alpha^{(k)} < \infty) = 1$ for any $x \in \mathsf{X}$. By the strong Markov property, for any integer $k$,

$$\begin{aligned}
&\mathrm{E}_\nu[\Psi_1(s_1(f)) \cdots \Psi_k(s_k(f))] \\
&\quad = \mathrm{E}_\nu[\Psi_1(s_1(f)) \cdots \Psi_{k-1}(s_{k-1}(f)) \, \mathrm{E}_\alpha[\Psi_k(s_k(f)) \,|\, \mathcal{F}_{\tau_\alpha^{(k-1)}}] \mathbb{1}_{\{\tau_\alpha^{(k-1)} < \infty\}}] \\
&\quad = \mathrm{E}_\nu[\Psi_1(s_1(f)) \cdots \Psi_{k-1}(s_{k-1}(f))] \, \mathrm{E}_\alpha[\Psi_k(s_1(f))] .
\end{aligned}$$

The desired result in then obtained by induction. □

*of Theorem 198 when there is an accessible atom.* First assume that $f$ is non-negative. Denote the accessible atom by $\alpha$ and define

$$\eta_n = \sum_{k=1}^{n} \mathbb{1}_\alpha(X_k) , \tag{7.65}$$

the occupation time of the atom $\alpha$ up to time $n$. We now split the sum $\sum_{k=1}^{n} f(X_k)$ into sums over the excursions between successive visits to $\alpha$,

$$\sum_{k=1}^{n} f(X_k) = \sum_{j=1}^{\eta_n} s_j(f) + \sum_{k=\tau_\alpha^{(\eta_n)}+1}^{n} f(X_k) \; .$$

This decomposition shows that

$$\sum_{j=1}^{\eta_n} s_j(f) \le \sum_{k=1}^{n} f(X_k) \le \sum_{j=1}^{\eta_n+1} s_j(f) \; . \tag{7.66}$$

Because $Q$ is Harris recurrent and $\alpha$ is accessible, $\eta_n \to \infty$ $\mathrm{P}_\nu$-a.s. as $n \to \infty$. Hence $s_1(f)/\eta_n \to 0$ and $(\eta_n - 1)/\eta_n \to 1$ $\mathrm{P}_\nu$-a.s. By Lemma 199 the variables $\{s_j(f)\}_{j\ge 2}$ are i.i.d. under $\mathrm{P}_\nu$. In addition $\mathrm{E}_\nu[s_j(f)] = \mu_\alpha(f)$ for $j \ge 2$ with $\mu_\alpha$, defined in (7.30), being an invariant measure. Because all invariant measures are constant multiples of $\mu_\alpha$ and $\pi(|f|) < \infty$, $\mathrm{E}_\alpha[s_j(f)]$ is finite. Writing

$$\frac{1}{\eta_n} \sum_{j=1}^{\eta_n} s_j(f) = \frac{s_1(f)}{\eta_n} + \frac{\eta_n - 1}{\eta_n} \frac{1}{\eta_n - 1} \sum_{j=2}^{\eta_n} s_j(f) \; ,$$

the LLN for i.i.d. random variables shows that

$$\lim_{n\to\infty} \frac{1}{\eta_n} \sum_{j=1}^{\eta_n} s_j(f) = \mu_\alpha(f) \quad \mathrm{P}_\nu\text{-a.s.} \; ,$$

whence, by (7.66), the same limit holds for $\eta_n^{-1} \sum_1^n f(X_k)$. Because $\pi(1) = 1$, $\mu_\alpha(1)$ is finite too. Applying the above result with $f \equiv 1$ yields $n/\eta_n \to \mu_\alpha(1)$, so that $n^{-1} \sum_1^n f(X_k) \to \mu_\alpha(f)/\mu_\alpha(1) = \pi(f)$ $\mathrm{P}_\nu$-a.s. This is the desired result when $f \ge 0$. The general case is is handled by splitting $f$ into its positive and negative parts. □

### Central Limit Theorems

We say that $\{f(X_k)\}$ satisfies a central limit theorem (CLT) if there is a constant $\sigma^2(f) \ge 0$ such that the normalized sum $n^{-1/2} \sum_{k=1}^{n} \{f(X_k) - \pi(f)\}$ converges $\mathrm{P}_\nu$-weakly to a Gaussian distribution with zero mean and variance $\sigma^2(f)$ (we allow for the special case $\sigma^2(f) = 0$ corresponding to weak convergence to the constant 0). CLTs are essential for understanding the error occurring when approximating $\pi(f)$ by the sample mean $n^{-1} \sum_{k=1}^{n} f(X_k)$ and are thus a topic of considerable importance.

For i.i.d. samples, classical theory guarantees a CLT as soon as $\pi(|f|^2) < \infty$. This is not true in general for Markov chains; the CLTs that are available do require some additional assumptions on the rate of convergence and/or the existence of higher order moments of $f$ under the stationary distribution.

**Theorem 200.** *Let $Q$ be a phi-irreducible aperiodic positive Harris recurrent transition kernel with invariant distribution $\pi$. Let $f$ be a measurable function and assume that there exists an accessible small set $C$ satisfying*

$$\int_{x\in C} \pi(dx)\, \mathrm{E}_x \left[ \left( \sum_{k=1}^{\tau_C} |f|(X_k) \right)^2 \right] < \infty \quad \textit{and} \quad \int_C \pi(dx)\, \mathrm{E}_x[\tau_C^2] < \infty \; . \tag{7.67}$$

*Then $\pi(f^2) < \infty$ and $\{f(X_k)\}$ satisfies a CLT.*

*Proof.* To start with, it follows from the expression (7.32) for the stationary distribution that

$$\pi(f^2) = \int_C \pi(dx)\, \mathrm{E}_x \left[ \sum_{k=1}^{\tau_C} f^2(X_k) \right] \leq \int_C \pi(dx)\, \mathrm{E}_x \left[ \left( \sum_{k=1}^{\tau_C} |f(X_k)| \right)^2 \right] < \infty \, .$$

We now prove the CLT under the additional assumption that the chain admits an accessible atom $\alpha$. The proof in the general phi-irreducible case can be obtained using the splitting construction. The proof is along the same lines as for the LLN. Put $\bar{f} = f - \pi(f)$. By decomposing the sum $\sum_{k=1}^{n} \bar{f}(X_k)$ into excursions between successive visits to the atom $\alpha$, we obtain

$$n^{-1/2} \left| \sum_{k=1}^{n} \bar{f}(X_k) - \sum_{j=2}^{\eta_n} s_j(\bar{f}) \right| \leq n^{-1/2} s_1(|\bar{f}|) + n^{-1/2} s_{\eta_n+1}(|\bar{f}|) \, , \tag{7.68}$$

where $\eta_n$ and $s_j(f)$ are defined in (7.65) and (7.64). It is clear that the first term on the right-hand side of this display vanishes (in $\mathrm{P}_\nu$-probability) as $n \to \infty$. For the second one, the strong LLN (Theorem 198) shows that $n^{-1} \sum_1^n s_j^2(|\bar{f}|)$ has an $\mathrm{P}_\nu$-a.s. finite limit, whence, $\mathrm{P}_\nu$-a.s.,

$$\limsup_{n\to\infty} \frac{s_n^2(|\bar{f}|)}{n} = \limsup_{n\to\infty} \left[ \frac{1}{n} \sum_{j=1}^{n} s_j^2(|\bar{f}|) - \frac{n+1}{n} \frac{1}{n+1} \sum_{j=1}^{n+1} s_j^2(|\bar{f}|) \right] = 0 \, .$$

The strong LLN with $f = \mathbb{1}_\alpha$ also shows that $\eta_n/n \to \pi(\alpha)$ $\mathrm{P}_\nu$-a.s., so that $s^2_{\eta_n}(|\bar{f}|)/\eta_n \to 0$ and $n^{-1/2} s_{\eta_n+1}(|\bar{f}|) \to 0$ $\mathrm{P}_\nu$-a.s.

Thus $n^{-1/2} \sum_1^n \bar{f}(X_k)$ and $n^{-1/2} \sum_2^{\eta_n} s_j(\bar{f})$ have the same limiting behavior. By Lemma 199, the blocks $\{s_j^2(|\bar{f}|)\}_{j\geq 2}$ are i.i.d. under $\mathrm{P}_\nu$. Thus, by the CLT for i.i.d. random variables, $n^{-1/2} \sum_2^n s_j(\bar{f})$ converges $\mathrm{P}_\nu$-weakly to a Gaussian law with zero mean and some variance $\sigma^2 < \infty$; that the variance is indeed finite follows as above with the small set $C$ being the accessible atom $\alpha$. The so-called Ascombe's theorem (see for instance Gut, 1988, Theorem 3.1) then implies that $\eta_n^{-1/2} \sum_2^{\eta_n} \bar{f}(X_k)$ converges $\mathrm{P}_\nu$-weakly to the same Gaussian law. Thus we may conclude that $n^{-1/2} \sum_2^{\eta_n} \bar{f}(X_k) = (\eta_n/n)^{1/2} \eta_n^{-1/2} \sum_2^{\eta_n} \bar{f}(X_k)$ converges $\mathrm{P}_\nu$-weakly to a Gaussian law with zero mean and variance $\pi(\alpha)\sigma^2$. By (7.68), so does $n^{-1/2} \sum_1^n \bar{f}(X_k)$. □

The condition (7.67) is stated in terms of the second moment of the excursion between two successive visits to a small set and appears rather difficult to verify directly. More explicit conditions can be obtained, in particular if we assume that the chain is $V$-geometrically ergodic.

**Proposition 201.** *Let $Q$ be a phi-irreducible, aperiodic, positive Harris reccurrent kernel that $Q$ satisfies a Foster-Lyapunov drift condition (see Definition 191) outside an accessible small set $C$, with drift function $V$. Then any measurable function $f$ such that $|f|^2 \leq V$ satisfies a CLT.*

*Proof.* Minkovski's inequality implies that

$$\begin{aligned} \mathrm{E}_x \left[ \left( \sum_{k=0}^{\tau_C - 1} |f(X_k)| \right)^2 \right] &\leq \left\{ \sum_{k=0}^{\infty} \sqrt{\mathrm{E}_x[f^2(X_k) \mathbb{1}_{\{\tau_C > k\}}]} \right\}^{1/2} \\ &\leq \left\{ \sum_{k=0}^{\infty} \sqrt{\mathrm{E}_x[V(X_k) \mathbb{1}_{\{\tau_C > k\}}]} \right\}^{1/2} \, . \end{aligned}$$

Put $M_k = \lambda^{-k} V(X_k) \mathbb{1}_{\{\tau_C \geq k\}}$, where $\lambda$ is as in (7.56). Then for $k \geq 1$,

$$\begin{aligned} \mathrm{E}[M_{k+1} \,|\, \mathcal{F}_k] &\leq \lambda^{-(k+1)} \,\mathrm{E}[V(X_{k+1}) \,|\, \mathcal{F}_k] \mathbb{1}_{\{\tau_C \geq k+1\}} \\ &\leq \lambda^{-k} V(X_k) \mathbb{1}_{\{\tau_C \geq k+1\}} \leq M_k \;, \end{aligned}$$

showing that $\{M_k\}$ is a super-martingale. Thus $\mathrm{E}_x[M_k] \leq \mathrm{E}_x[M_1]$ for any $x \in C$, which implies that for $k \geq 1$,

$$\sup_{x \in C} \mathrm{E}_x[V(X_k) \mathbb{1}_{\{\tau_C \geq k\}}] \leq \lambda^k \left[ \sup_{x \in C} V(x) + b \right] \;.$$

□

## 7.3 Applications to Hidden Markov Models

As discussed in Section 1.2, an HMM is best defined as a Markov chain $\{X_k, Y_k\}_{k \geq 0}$ on the product space $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. The transition kernel of this joint chain has a simple structure reflecting the conditional independence assumptions that are imposed. Let $Q$ and $G$ denote, respectively, a Markov transition kernel on $(\mathsf{X}, \mathcal{X})$ and a transition kernel from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$. The transition kernel of the joint chain $\{X_k, Y_k\}_{k \geq 0}$ is given by, for any $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$T\left[(x, y), C\right] = \iint_C Q(x, dx')\, G(x', dy) \;, \qquad (x, y) \in \mathsf{X} \times \mathsf{Y},\, C \in \mathcal{X} \otimes \mathcal{Y} \;. \tag{7.69}$$

This chain is said to be hidden because only a component (here $\{Y_k\}_{k \geq 0}$) is observed. Of course, the process $\{Y_k\}$ is not a Markov chain, but nevertheless most of the properties of this process are inherited from stability properties of the hidden chain. In this section, we establish stability properties of the kernel $T$ of the joint chain.

### 7.3.1 Phi-irreducibility

Phi-irreducibility of the joint chain $T$ is inherited from irreducibility of the hidden chain, and the maximal irreducibility measures of the joint and hidden chains are related in a simple way. Before stating the precise result, we recall (see Section 1.1.1) that if $\phi$ is a measure on $(\mathsf{X}, \mathcal{X})$, we define the measure $\phi \otimes G$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$ by

$$\phi \otimes G(A) \stackrel{\text{def}}{=} \iint_A \mu(dx)\, G(x, dy) \;, \qquad A \in \mathcal{X} \otimes \mathcal{Y} \;.$$

**Proposition 202.** *Assume that $Q$ is phi-irreducible, and let $\phi$ be an irreducibility measure for $Q$. Then $\phi \otimes G$ is an irreducibility measure for $T$. If $\psi$ is a maximal irreducibility measure for $Q$, then $\psi \otimes G$ is a maximal irreducibility measure for $T$.*

*Proof.* Let $A \in \mathcal{X} \otimes \mathcal{Y}$ be a set such that $\phi \otimes G(A) > 0$. Denote by $\Psi_A$ the function $\Psi_A(x) = \int_{\mathsf{Y}} G(x, dy)\, \mathbb{1}_A(x, y)$ for $x \in \mathsf{X}$. By Fubini's theorem,

$$\phi \otimes G(A) = \iint \phi(dx)\, G(x, dy)\, \mathbb{1}_A(x, y) = \int \phi(dx)\, \Psi_A(x) \;,$$

and the condition $\phi \otimes G(A) > 0$ implies that $\phi\left(\{\Psi_A > 0\}\right) > 0$. Because $\{\Psi_A > 0\} = \bigcup_{m=0}^{\infty} \{\Psi_A \geq 1/m\}$, we have $\phi\left(\{\Psi_A \geq 1/m\}\right) > 0$ for some integer $m$. Because $\phi$

is an irreducibility measure, for any $x \in \mathsf{X}$ there exists an integer $k \geq 0$ such that $Q^k(x, \{\Psi_A \geq 1/m\}) > 0$. Therefore for any $y \in \mathsf{Y}$,

$$T^k[(x,y),A] = \iint Q^k(x,dx')\,G(x',dy')\,\mathbb{1}_A(x',y') = \int Q^k(x,dx')\,\Psi_A(x')$$
$$\geq \int_{\{\Psi_A \geq 1/m\}} Q^k(x,dx')\,\Psi_A(x') \geq \frac{1}{m} Q^k(x,\{\Psi_A \geq 1/m\}) > 0\,,$$

showing that $\phi \otimes G$ is an irreducibility measure for $T$.

Morever, using Theorem 147, we see that a maximal irreducibility measure $\psi_T$ for $T$ is given by, for any $\delta \in (0,1)$ and $A \in \mathcal{X} \otimes \mathcal{Y}$,

$$\begin{aligned}
\psi_T(A) &= \iint \phi(dx)\,G(x,dy)\,(1-\delta)\sum_{m=0}^{\infty} \delta^m T^m[(x,y),A] \\
&= \iint (1-\delta)\sum_{m=0}^{\infty} \delta^m \int \phi(dx)\,Q^m(x,dx')\,G(x',dy')\,\mathbb{1}_A(x',y') \\
&= \iint \psi(dx')\,G(x',dy')\,\mathbb{1}_A(x',y') = \psi \otimes G(A)\,,
\end{aligned}$$

where

$$\psi(B) = \int \phi(dx)\,(1-\delta)\sum_{m=0}^{\infty} \delta^m Q^m(x,B)\,, \qquad B \in \mathcal{X}\,.$$

By Theorem 147, $\psi$ is a maximal irreducibility measure for $Q$. In addition, if $\hat{\psi}$ is a maximal irreducibility measure for $Q$, then $\hat{\psi}$ is equivalent to $\psi$. Because for any $A \in \mathcal{X} \otimes \mathcal{Y}$,

$$\hat{\psi} \otimes G(A) = \iint \hat{\psi}(dx)\,G(x,dy)\,\mathbb{1}_A(x,y) = \iint \psi \otimes G(dx,dy)\,\frac{d\hat{\psi}}{d\psi}(x)\mathbb{1}_A(x,y)\,,$$

$\hat{\psi} \otimes G(A) = 0$ whenever $\psi \otimes G(A) = 0$. Thus $\hat{\psi} \otimes G \ll \psi \otimes G$. Exchanging $\psi$ and $\hat{\psi}$ shows that $\psi \otimes G$ and $\hat{\psi} \otimes G$ are indeed equivalent, which concludes the proof. □

**Example 203** (Stochastic Volatility Model)**.** The canonical stochastic volatility model (see Example **??**) is given by

$$\begin{aligned}
X_{k+1} &= \phi X_k + \sigma U_k\,, & U_k &\sim \mathrm{N}(0,1)\,, \\
Y_k &= \beta \exp(X_k/2) V_k\,, & V_k &\sim \mathrm{N}(0,1)\,,
\end{aligned}$$

We have established (see Example 148) that because $\{U_k\}$ has a positive density on $\mathbb{R}^+$, the chain $\{X_k\}$ is phi-irreducible and $\lambda^{\mathrm{Leb}}$ is an irreducibility measure. Therefore $\{X_k, Y_k\}$ is also phi-irreducible and $\lambda^{\mathrm{Leb}} \otimes \lambda^{\mathrm{Leb}}$ is a maximal irreducibility measure.

### 7.3.2 Atoms and Small Sets

It is possible to relate atoms and small sets of the joint chain to those of the hidden chain. Examples of HMMs possessing accessible atoms are numerous, even when the state space of the joint chain is general. They include in particular the Markov chains whose hidden state space $\mathsf{X}$ is finite.

**Example 204** (Normal HMM, Continued)**.** For the normal HMM (see Example **??**), it holds that $T[(x,y),\cdot] = T[(x,y'),\cdot]$ for any $(y,y') \in \mathbb{R} \times \mathbb{R}$. Hence $\{x\} \times \mathbb{R}$ is an atom for $T$.

When accessible atoms do not exist, it is important to determine small sets. Here again the small sets of the joint chain can easily be related to those of the hidden chain.

**Lemma 205.** *Let $m$ be a positive integer, $\epsilon > 0$ and let $\eta$ be a probability measure on $(\mathsf{X}, \mathcal{X})$. Let $C \in \mathcal{X}$ be an $(m, \epsilon, \eta)$-small set for the transition kernel $Q$, that is, $Q^m(x, A) \geq \epsilon \mathbb{1}_C(x)\eta(A)$ for all $x \in \mathsf{X}$ and $A \in \mathcal{X}$. Then $C \times \mathsf{Y}$ is an $(m, \epsilon, \eta \otimes G)$-small set for the transition kernel $T$ defined in* (1.14)*, that is,*

$$T^m[(x, y), A] \geq \epsilon \mathbb{1}_C(x)\, \eta \otimes G(A) \,, \qquad (x, y) \in \mathsf{X} \times \mathsf{Y},\, A \in \mathcal{X} \otimes \mathcal{Y} \,.$$

*Proof.* Pick $(x, y) \in C \times \mathsf{Y}$. Then

$$\begin{aligned} T^m[(x, y), A] &= \iint Q^m(x, dx')\, G(x', dy')\, \mathbb{1}_A(x', y') \\ &\geq \epsilon \iint \eta(dx')\, G(x', dy')\, \mathbb{1}_A(x', y') \,. \end{aligned}$$

□

If the Markov transition kernel $Q$ on $(\mathsf{X}, \mathcal{X})$ is phi-irreducible (with maximal irreducibility measure $\psi$), then we know from Proposition 157 that there exists an accessible small set $C$. That is, there exists a set $C \in \mathcal{X}$ with $\mathrm{P}_x(\tau_C < \infty) > 0$ for all $x \in \mathsf{X}$ and such that $C$ is $(m, \epsilon, \eta)$-small for some triple $(m, \epsilon, \eta)$ with $\eta(C) > 0$. Then Lemma 205 shows that $C \times \mathsf{Y}$ is an $(m, \epsilon, \eta \otimes G)$-small set for the transition kernel $T$.

**Example 206** (Stochastic Volatility Model, Continued)**.** We have shown in Example 148 that any compact set $K \subset \mathbb{R}$ is small for the first-order autoregression constituting the hidden chain of the stochastic volatility model of Example 203. Therefore any set $K \times \mathbb{R}$, where $K$ a compact subset of $\mathbb{R}$, is small for the joint chain $\{X_k, Y_k\}$.

The simple relations between the small sets of the joint chain and those of the hidden chain immediately imply that $T$ and $Q$ have the same period.

**Proposition 207.** *Suppose that $Q$ is phi-irreducible and has period $d$. Then $T$ is phi-irreducible and has the same period $d$. In particular, if $Q$ is aperiodic, then so is $T$.*

*Proof.* Let $C$ be an accessible $(m, \epsilon, \eta)$-small set for $Q$ with $\eta(C) > 0$. Define $E_C$ as the set of time indices for which $C$ is a small set with minorizing probability measure $\eta$,

$$E_C \stackrel{\text{def}}{=} \{n \geq 0 : C \text{ is } (n, \epsilon, \eta)\text{-small for some } \epsilon > 0\} \,.$$

The period of the set $C$ is given by the greatest common divisor of $E_C$. Proposition 180 shows that this value is in fact common to the chain as such and does not depend on the particular small set chosen. Lemma 205 shows that $C \times \mathsf{Y}$ is an $(m, \epsilon, \eta \otimes G)$-small set for the joint Markov chain with transition kernel $T$, and that $\eta \otimes G(C \times \mathsf{Y}) = \eta(C) > 0$. The set $E_{C \times \mathsf{Y}}$ of time indices for which $C \times \mathsf{Y}$ is a small set for $T$ with minorizing measure $\eta \otimes G$ is thus, using Lemma 205 again, equal to $E_C$. Thus the period of the set $C$ is also the period of the set $C \times \mathsf{Y}$. Because the period of $T$ does not depend on the choice of the small set $C \times \mathsf{Y}$, it follows that the periods of $Q$ and $T$ coincide. □

### 7.3.3 Recurrence and Positive Recurrence

As the following result shows, recurrence and transience of the joint chain follows directly from the corresponding properties of the hidden chain.

**Proposition 208.** *Assume that the hidden chain is phi-irreducible. Then the following statements hold true.*

*(i) The joint chain is transient (recurrent) if and only if the hidden chain is transient (recurrent).*

*(ii) The joint chain is positive if and only if the hidden chain is positive. In addition, if the hidden chain is positive with stationary distribution $\pi$, then $\pi \otimes G$ is the stationary distribution of the joint chain.*

*Proof.* First assume that the transition kernel $Q$ is transient, that is, that there is a countable cover $\mathsf{X} = \cup_i A_i$ of $\mathsf{X}$ with uniformly transient sets,

$$\sup_{x \in A_i} \mathrm{E}_x \left[ \sum_{n=1}^{\infty} \mathbb{1}_{A_i}(X_n) \right] < \infty \, .$$

Then the sets $\{A_i \times \mathsf{Y}\}_{i \geq 1}$ form a countable cover of $\mathsf{X} \times \mathsf{Y}$, and these sets are uniformly transient because

$$\mathrm{E}_x \left[ \sum_{n=1}^{\infty} \mathbb{1}_{A_i \times \mathsf{Y}}(X_n, Y_n) \right] = \mathrm{E}_x \left[ \sum_{n=1}^{\infty} \mathbb{1}_{A_i}(X_n) \right] \, . \tag{7.70}$$

Thus the joint chain is transient.

Conversely, assume that the joint chain is transient. Because the hidden chain is phi-irreducible, Proposition 158 shows that there is a countable cover $\mathsf{X} = \cup_i A_i$ of $\mathsf{X}$ with sets that are small for $Q$. At least one of these, say $A_1$, is accessible for $Q$. By Lemma 205, the sets $A_i \times \mathsf{Y}$ are small. By Proposition 202, $A_1 \times \mathsf{Y}$ is accessible and, because $T$ is transient, Proposition 159 shows that $A_1 \times \mathsf{Y}$ is uniformly transient. Equation (7.70) then shows that $A_1$ is uniformly transient, and because $A_1$ is accessible, we conclude that $Q$ is transient.

Thus the hidden chain is transient if and only if the joint chain is so. The transience/recurrence dichotomy (Theorem 151) then implies that the hidden chain is recurrent if and only if the joint chain is so, which completes the proof of (i).

We now turn to (ii). First assume that the hidden chain is positive recurrent, that is, that there exists a unique stationary probability measure $\pi$ satisfying $\pi Q = \pi$. Then the probability measure $\pi \otimes G$ is stationary for the transition kernel $T$ of the joint chain, because

$$\begin{aligned} (\pi \otimes G)T(A) &= \int \cdots \int \pi(dx)\, G(x, dy)\, Q(x, dx')\, G(x', dy')\, \mathbb{1}_A(x', y') \\ &= \iiint \pi(dx)\, Q(x, dx')\, G(x', dy')\, \mathbb{1}_A(x', y') \\ &= \iint \pi(dx')\, G(x', dy')\, \mathbb{1}_A(x', y') = \pi \otimes G(A) \, . \end{aligned}$$

Because the joint chain admits a stationary distribution it is positive, and by Proposition 179 it is recurrent.

Conversely, assume that the joint chain is positive. Denote by $\bar{\pi}$ the (unique) stationary probability measure of $T$. Thus for any $\bar{A} \in \mathcal{X} \otimes \mathcal{Y}$, we have

$$\begin{aligned} &\iint \bar{\pi}(dx, dy)\, Q(x, dx')\, G(x', dy')\, \mathbb{1}_{\bar{A}}(x', y') \\ &\qquad = \iint \bar{\pi}(dx, \mathsf{Y})\, Q(x, dx')\, G(x', dy')\, \mathbb{1}_{\bar{A}}(x', y') = \bar{\pi}(\bar{A}) \, . \end{aligned}$$

Setting $\bar{A} = A \times \mathsf{Y}$ for $A \in \mathcal{X}$, this display implies that

$$\int \bar{\pi}(dx, \mathsf{Y})\, Q(x, A) = \bar{\pi}(A \times \mathsf{Y}) \; .$$

This shows that $\pi(A) = \bar{\pi}(A \times \mathsf{Y})$ is a stationary distribution for the hidden chain. Hence the hidden chain is positive and recurrent. □

When the joint (or hidden) chain is positive, it is natural to study the rate at which it converges to stationarity.

**Proposition 209.** *Assume that the hidden chain satisfies a uniform Doeblin condition, that is, there exists a positive integer $m$, $\epsilon > 0$ and a family $\{\eta_{x,x'}, (x, x') \in \mathsf{X} \times \mathsf{X}\}$ of probability measures such that*

$$Q^m(x, A) \wedge Q^m(x', A) \geq \epsilon \eta_{x,x'}(A), \quad A \in \mathcal{X},\ (x, x') \in \mathsf{X} \times \mathsf{X} \; .$$

*Then the joint chain also satisfies a uniform Doeblin condition. Indeed, for all $(x, y)$ and $(x', y')$ in $\mathsf{X} \times \mathsf{Y}$ and all $\bar{A} \in \mathcal{X} \otimes \mathcal{Y}$,*

$$T^m[(x, y), \bar{A}] \wedge T^m[(x', y'), \bar{A}] \geq \epsilon \bar{\eta}_{x,x'}(\bar{A}) \; ,$$

*where*

$$\bar{\eta}_{x,x'}(\bar{A}) = \int \eta_{x,x'}(dx)\, G(x, dy)\, \mathbb{1}_{\bar{A}}(x, y) \; .$$

The proof is along the same lines as the proof of Lemma 205 and is omitted. This proposition in particular implies that the ergodicity coefficients for the kernels $T^m$ and $Q^m$ coincide; $\delta(T^m) = \delta(Q^m)$. A straightforward but useful application of this result is when the hidden Markov chain is defined on a finite state space. If the transition matrix $Q$ of this chain is primitive, that is, there exists a positive integer $m$ such that $Q^m(x, x') > 0$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$ (or, equivalently, if the chain $Q$ is irreducible and aperiodic), then the joint Markov chain satisfies a uniform Doeblin condition and the ergodicity coefficient of the joint chain is bounded as $\delta(T^m) \leq 1 - \epsilon$ with

$$\epsilon = \inf_{(x,x') \in \mathsf{X} \times \mathsf{X}} \sup_{x'' \in \mathsf{X}} [Q^m(x, x'') \wedge Q^m(x', x'')] \; .$$

A similar result holds when the hidden chain satisfies a Foster-Lyapunov drift condition instead of a uniform Doeblin condition. This result is of particular interest when dealing with hidden Markov models on state spaces that are not finite or bounded.

**Proposition 210.** *Assume that $Q$ is phi-irreducible, aperiodic, and satisfies a Foster-Lyapunov drift condition (Definition 191) with drift function $V$ outside a set $C$. Then the transition kernel $T$ also satisfies a Foster-Lyapunov drift condition with drift function $V$ outside the set $C \times \mathsf{Y}$,*

$$T[(x, y), V] \leq \lambda V(x) + b \mathbb{1}_{C \times \mathsf{Y}}(x, y) \; .$$

*Here on the left-hand side, we wrote $V$ also for a function on $\mathsf{X} \times \mathsf{Y}$ defined by $V(x, y) = V(x)$.*

The proof is straightforward. Proposition 195 yields an explicit bound on the rate of convergence of the iterates of the Markov chain to the stationary distribution. This result has a lot of interesting consequences.

**Proposition 211.** *Suppose that $Q$ is phi-irreducible, aperiodic, and satisfies a Foster-Lyapunov drift condition with drift function $V$ outside a small set $C$. Then the transition kernel $T$ is positive and aperiodic with invariant distribution $\pi \otimes G$, where $\pi$ is the invariant distribution of $Q$. In addition, for any measurable function $f : \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$, the following statements hold true.*

*(i) If $\sup_{x \in \mathsf{X}} [V(x)]^{-1} \int G(x, dy) \, |f(x, y)| < \infty$, then there exist $\rho \in (0, 1)$ and $K < \infty$ (not depending on $f$) such that for any $n \geq 0$ and $(x, y) \in \mathsf{X} \times \mathsf{Y}$,*

$$|T^n f(x, y) - \pi \otimes G(f)| \leq K \rho^n V(x) \sup_{x' \in \mathsf{X}} [V(x')]^{-1} \; intG(x', dy) \, |f(x', y)| \; .$$

*(ii) If $\sup_{x \in \mathsf{X}} [V(x)]^{-1} \int G(x, dy) \, f^2(x, y) < \infty$, then $\mathrm{E}_{\pi \otimes G}[f^2(X_0, Y_0)] < \infty$ and there exist $\rho \in (0, 1)$ and $K < \infty$ (not depending on $f$) such that for any $n \geq 0$,*

$$|\mathrm{Cov}_\pi[f(X_n, Y_n), f(X_0, Y_0)]| \leq K \rho^n \pi(V) \left\{ \sup_{x \in \mathsf{X}} [V(x)]^{-1/2} \int G(x, dy) \, |f(x, y)| \right\}^2 \; .$$

*Proof.* First note that

$$\begin{aligned} |T^n f(x, y) - \pi \otimes G(f)| &= \left| \iint [Q^n(x, dx') - \pi(dx')] G(x', dy') \, f(x', y') \right| \\ &\leq \|Q^n(x, \cdot) - \pi\|_V \sup_{x' \in \mathsf{X}} [V(x')]^{-1} \int G(x', dy) \, |f(x', y)| \; . \end{aligned}$$

Now part (i) follows from the geometric ergodicity of $Q$ (Theorem 194). Next, because $\pi(V) < \infty$,

$$\begin{aligned} \mathrm{E}_{\pi \otimes G}[f^2(X_0, Y_0)] &= \iint \pi(dx) \, G(x, dy) \, f^2(x, y) \\ &\leq \pi(V) \sup_{x \in \mathsf{X}} [V(x)]^{-1} \int G(x, dy) \, f^2(x, y) < \infty \; , \end{aligned}$$

implying that $|\,\mathrm{Cov}_\pi[|f(X_n, Y_n)|, |f(X_0, Y_0)|]| \leq \mathrm{Var}_\pi[f(X_0, Y_0)] < \infty$. In addition

$$\begin{aligned} &\mathrm{Cov}_\pi[f(X_n, Y_n), f(X_0, Y_0)] \\ &\quad = \mathrm{E}_\pi\{\mathrm{E}[f(X_n, Y_n) - \pi \otimes G(f) \,|\, \mathcal{F}_0] f(X_0, Y_0)\} \\ &\quad = \iint \pi \otimes G(dx, dy) \, f(x, y) \iint [Q^n(x, dx') - \pi(dx')] \, G(x', dy') \, f(x', y') \; . \end{aligned} \tag{7.71}$$

By Jensen's inequality $\int G(x, dy) \, |f(x, y)| \leq [\int G(x, dy) \, f^2(x, y)]^{1/2}$ and

$$QV^{1/2}(x) \leq [QV(x)]^{1/2} \leq [\lambda V(x) + b \mathbb{1}_C(x)]^{1/2} \leq \lambda^{1/2} V^{1/2}(x) + b^{1/2} \mathbb{1}_C(x) \; ,$$

showing that $Q$ also satisfies a Foster-Lyapunov condition outside $C$ with drift function $V^{1/2}$. By Theorem 194, there exists $\rho \in (0, 1)$ and a constant $K$ such that

$$\begin{aligned} &\left| \iint [Q^n(x, dx') - \pi(dx)] \, G(x', dy') \, f(x', y') \right| \\ &\qquad \leq \|Q^n(x, \cdot) - \pi\|_{V^{1/2}} \sup_{x' \in \mathsf{X}} V^{-1/2}(x) \int G(x', dy) \, |f(x', y)| \\ &\qquad \leq K \rho^n V^{1/2}(x) \sup_{x' \in \mathsf{X}} V^{-1/2}(x') \int G(x', dy) \, |f(x', y)| \; . \end{aligned}$$

Part (ii) follows by plugging this bound into (7.71). □

**Example 212** (Stochastic Volatility Model, Continued)**.** In the model of Example 203, we set $V(x) = \mathrm{e}^{x^2/2\delta^2}$ for $\delta > \sigma_U$. It is easily shown that

$$QV(x) = \frac{\rho}{\sigma_U} \exp\left[\frac{x^2}{2\delta^2}\frac{\phi^2(\rho^2+\delta^2)}{\delta^2}\right] ,$$

where $\rho^2 = \sigma_U^2\delta^2/(\delta^2-\sigma_U^2)$. We may choose $\delta$ large enough that $\phi^2(\rho^2+\delta^2)/\delta^2 < 1$. Then $\limsup_{|x|\to\infty} QV(x)/V(x) = 0$ so that $Q$ satisfies a Foster-Lyapunov condition with drift function $V(x) = \mathrm{e}^{x^2/2\delta^2}$ outside a compact set $[-M, +M]$. Because every compact set is small, the assumptions of Proposition 211 are satisfied, showing that the joint chain is positive. Set $f(x, y) = |y|$. Then $\int G(x, dy)\,|y| = \beta\mathrm{e}^{x/2}\sqrt{2/\pi}$. Proposition 211(ii) shows that $\mathrm{Var}_\pi(Y_0) < \infty$ and that the autocovariance function $\mathrm{Cov}(|Y_n|, |Y_0|)$ decreases to zero exponentially fast.

# Bibliography


Akashi, H. and Kumamoto, H. (1977) Random sampling approach to state estimation in switching environment. *Automatica*, **13**, 429–434.

Anderson, B. D. O. and Moore, J. B. (1979) *Optimal Filtering*. Prentice-Hall.

Askar, M. and Derin, H. (1981) A recursive algorithm for the Bayes solution of the smoothing problem. *IEEE Trans. Automat. Control*, **26**, 558–561.

Atar, R. and Zeitouni, O. (1997) Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincaré Probab. Statist.*, **33**, 697–725.

Athreya, K. B., Doss, H. and Sethuraman, J. (1996) On the convergence of the Markov chain simulation method. *Ann. Statist.*, **24**, 69–100.

Athreya, K. B. and Ney, P. (1978) A new approach to the limit theory of recurrent Markov chains. *Trans. Am. Math. Soc.*, **245**, 493–501.

Baum, L. E. and Petrie, T. P. (1966) Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, **37**, 1554–1563.

Baum, L. E., Petrie, T. P., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Bickel, P. J., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, **26**, 1614–1635.

Boyles, R. (1983) On the convergence of the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **45**, 47–50.

Budhiraja, A. and Ocone, D. (1997) Exponential stability of discrete-time filters for bounded observation noise. *Systems Control Lett.*, **30**, 185–193.

Campillo, F. and Le Gland, F. (1989) MLE for patially observed diffusions: Direct maximization vs. the EM algorithm. *Stoch. Proc. App.*, **33**, 245–274.

Cappé, O. (2001) Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation. *Monte Carlo Methods Appl.*, **7**, 81–92.

Cappé, O., Buchoux, V. and Moulines, E. (1998) Quasi-Newton method for maximum likelihood estimation of hidden Markov models. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2265–2268.

Carpenter, J., Clifford, P. and Fearnhead, P. (1999) An improved particle filter for non-linear problems. *IEE Proc., Radar Sonar Navigation*, **146**, 2–7.

Cérou, F., Le Gland, F. and Newton, N. (2001) Stochastic particle methods for linear tangent filtering equations. In *Optimal Control and PDE's - Innovations and Applications, in Honor of Alain Bensoussan's 60th Anniversary* (eds. J.-L. Menaldi, E. Rofman and A. Sulem), 231–240. IOS Press.

Chen, R. and Liu, J. S. (2000) Mixture Kalman filter. *J. Roy. Statist. Soc. Ser. B*, **62**, 493–508.

Chigansky, P. and Lipster, R. (2004) Stability of nonlinear filters in nonmixing case. *Ann. Appl. Probab.*, **14**, 2038–2056.

Collings, I. B. and Rydén, T. (1998) A new maximum likelihood gradient algorithm for on-line hidden Markov model identification. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, 2261–2264.

Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*. Wiley.

Crisan, D., Del Moral, P. and Lyons, T. (1999) Discrete filtering using branching and interacting particle systems. *Markov Process. Related Fields*, **5**, 293–318.

Del Moral, P. (2004) *Feynman-Kac Formulae. Genealogical and Interacting Particle Systems with Applications*. Springer.

Del Moral, P. and Jacod, J. (2001) Interacting particle filtering with discrete-time observations: Asymptotic behaviour in the Gaussian case. In *Stochastics in Finite and Infinite Dimensions: In Honor of Gopinath Kallianpur* (eds. T. Hida, R. L. Karandikar, H. Kunita, B. S. Rajput, S. Watanabe and J. Xiong), 101–122. Birkhäuser.

Del Moral, P., Ledoux, M. and Miclo, L. (2003) On contraction properties of Markov kernels. *Probab. Theory Related Fields*, **126**, 395–420.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38 (with discussion).

Devroye, L. (1986) *Non-Uniform Random Variate Generation*. Springer. URL `http://cgm.cs.mcgill.ca/~luc/rnbookindex.html`.

Devroye, L. and Klincsek, T. (1981) Average time behavior of distributive sorting algorithms. *Computing*, **26**, 1–7.

Dobrushin, R. (1956) Central limit theorem for non-stationary Markov chains. I. *Teor. Veroyatnost. i Primenen.*, **1**, 72–89.

Doob, J. L. (1953) *Stochastic Processes*. Wiley.

Douc, R., Moulines, E. and Rydén, T. (2004) Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, **32**, 2254–2304.

Doucet, A., De Freitas, N. and Gordon, N. (eds.) (2001) *Sequential Monte Carlo Methods in Practice*. Springer.

Doucet, A., Godsill, S. and Andrieu, C. (2000) On sequential Monte-Carlo sampling methods for Bayesian filtering. *Stat. Comput.*, **10**, 197–208.

Doucet, A. and Tadić, V. B. (2003) Parameter estimation in general state-space models using particle methods. *Ann. Inst. Statist. Math.*, **55**, 409–422.

Durrett, R. (1996) *Probability: Theory and Examples.* Duxbury Press, 2nd ed.

Elliott, R. J. and Krishnamurthy, V. (1999) New finite-dimensional filters for parameter estimation of discrete-time linear Gaussian models. *IEEE Trans. Automat. Control*, **44**.

Ephraim, Y. and Merhav, N. (2002) Hidden Markov processes. *IEEE Trans. Inform. Theory*, **48**, 1518–1569.

Evans, M. and Swartz, T. (1995) Methods for approximating integrals in Statistics with special emphasis on Bayesian integration problems. *Statist. Sci.*, **10**, 254–272.

— (2000) *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford University Press.

Fearnhead, P. (1998) *Sequential Monte Carlo methods in filter theory.* Ph.D. thesis, University of Oxford.

Feller, W. (1943) On a general class of "contagious" distributions. *Ann. Math. Statist.*, **14**, 389–399.

Fletcher, R. (1987) *Practical Methods of Optimization.* Wiley.

Fredkin, D. R. and Rice, J. A. (1992) Maximum-likelihood-estimation and identification directly from single-channel recordings. *Proc. Roy. Soc. London Ser. B*, **249**, 125–132.

Geweke, J. (1989) Bayesian inference in econometric models using Monte-Carlo integration. *Econometrica*, **57**, 1317–1339.

Giudici, P., Rydén, T. and Vandekerkhove, P. (2000) Likelihood-ratio tests for hidden Markov models. *Biometrics*, **56**, 742–747.

Glynn, P. W. and Iglehart, D. (1989) Importance sampling for stochastic simulations. *Management Science*, **35**, 1367–1392.

Gordon, N., Salmond, D. and Smith, A. F. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, **140**, 107–113.

Gupta, N. and Mehra, R. (1974) Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Trans. Automat. Control*, **19**, 774–783.

Gut, A. (1988) *Stopped Random Walks.* Springer.

Hammersley, J. M. and Handscomb, D. C. (1965) *Monte Carlo Methods.* Methuen & Co.

Handschin, J. (1970) Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica*, **6**, 555–563.

Handschin, J. and Mayne, D. (1969) Monte Carlo techniques to estimate the conditionnal expectation in multi-stage non-linear filtering. In *Int. J. Control*, vol. 9, 547–559.

Ho, Y. C. and Lee, R. C. K. (1964) A Bayesian approach to problems in stochastic estimation and control. *IEEE Trans. Automat. Control*, **9**, 333–339.

Horn, R. A. and Johnson, C. R. (1985) *Matrix Analysis*. Cambridge University Press.

Ibragimov, I. A. and Hasminskii, R. Z. (1981) *Statistical Estimation. Asymptotic Theory*. Springer.

Ito, H., Amari, S. I. and Kobayashi, K. (1992) Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Inform. Theory*, **38**, 324–333.

Jain, N. and Jamison, B. (1967) Contributions to Doeblin's theory of Markov processes. *Z. Wahrsch. Verw. Geb.*, **8**, 19–40.

Jamshidian, M. and Jennrich, R. J. (1997) Acceleration of the EM algorithm using quasi-Newton methods. *J. Roy. Statist. Soc. Ser. B*, **59**, 569–587.

Jarner, S. and Hansen, E. (2000) Geometric ergodicity of Metropolis algorithms. *Stoch. Proc. App.*, **85**, 341–361.

Julier, S. J. and Uhlmann, J. K. (1997) A new extension of the Kalman filter to nonlinear systems. In *AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*.

Kaijser, T. (1975) A limit theorem for partially observed Markov chains. *Ann. Probab.*, **3**, 677–696.

Kailath, T., Sayed, A. and Hassibi, B. (2000) *Linear Estimation*. Prentice-Hall.

Kalman, R. E. and Bucy, R. (1961) New results in linear filtering and prediction theory. *J. Basic Eng., Trans. ASME, Series D*, **83**, 95–108.

Kitagawa, G. (1987) Non-Gaussian state space modeling of nonstationary time series. *J. Am. Statist. Assoc.*, **82**, 1023–1063.

— (1996) Monte-Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.*, **1**, 1–25.

Kong, A., Liu, J. S. and Wong, W. (1994) Sequential imputation and Bayesian missing data problems. *J. Am. Statist. Assoc.*, **89**.

Künsch, H. R. (2000) State space and hidden Markov models. In *Complex Stochastic Systems* (eds. O. E. Barndorff-Nielsen, D. R. Cox and C. Kluppelberg). CRC Press.

— (2003) Recursive Monte-Carlo filters: algorithms and theoretical analysis. Preprint ETHZ, seminar für statistics.

Lange, K. (1995) A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **57**, 425–437.

Le Gland, F. and Mevel, L. (1997) Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, 3468–3473.

Le Gland, F. and Oudjane, N. (2004) Stability and uniform approximation of nonlinear filters using the hilbert metric and application to particle filters. *Ann. Appl. Probab.*, **14**, 144–187.

Lehmann, E. L. and Casella, G. (1998) *Theory of Point Estimation*. Springer, 2nd ed.

Leroux, B. G. (1992) Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. Appl.*, **40**, 127–143.

Lipster, R. S. and Shiryaev, A. N. (2001) *Statistics of Random Processes: I. General theory*. Springer, 2nd ed.

Liu, J. and Chen, R. (1995) Blind deconvolution via sequential imputations. *J. Am. Statist. Assoc.*, **430**, 567–576.

— (1998) Sequential Monte-Carlo methods for dynamic systems. *J. Am. Statist. Assoc.*, **93**, 1032–1044.

Liu, J. S. (1996) Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.*, **6**, 113–119.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **44**, 226–233.

Luenberger, D. G. (1984) *Linear and Nonlinear Programming*. Addison-Wesley, 2nd ed.

Meng, X.-L. (1994) On the rate of convergence of the ECM algorithm. *Ann. Statist.*, **22**, 326–339.

Meng, X.-L. and Rubin, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Am. Statist. Assoc.*, **86**, 899–909.

— (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

Meng, X.-L. and Van Dyk, D. (1997) The EM algorithm–an old folk song sung to a fast new tune. *J. Roy. Statist. Soc. Ser. B*, **59**, 511–567.

Mengersen, K. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*. Springer.

Niederreiter, H. (1992) *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM.

Nummelin, E. (1978) A splitting technique for Harris recurrent Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **4**, 309–318.

— (1984) *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press.

Orchard, T. and Woodbury, M. A. (1972) A missing information principle: Theory and applications. In *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics*, vol. 1, 697–715.

Ostrowski, A. M. (1966) *Solution of Equations and Systems of Equations*. Academic Press, 2nd ed.

Pitt, M. K. and Shephard, N. (1999) Filtering via simulation: Auxiliary particle filters. *J. Am. Statist. Assoc.*, **94**, 590–599.

Polson, N. G., Carlin, B. P. and Stoffer, D. S. (1992) A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *J. Am. Statist. Assoc.*, **87**, 493–500.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1992) *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2nd ed. URL `http://www.numerical-recipes.com/`.

Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–285.

Rauch, H., Tung, F. and Striebel, C. (1965) Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, **3**, 1445–1450.

Ristic, B., Arulampalam, M. and Gordon, A. (2004) *Beyond Kalman Filters: Particle Filters for Target Tracking*. Artech House.

Robert, C. P. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer, 2nd ed.

Roberts, G. O. and Rosenthal, J. S. (2004) General state space Markov chains and MCMC algorithms. *Probab. Surv.*, **1**, 20–71.

Roberts, G. O. and Tweedie, R. L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Rosenthal, J. S. (1995) Minorization conditions and convergence rates for Markov chain Monte Carlo. *J. Am. Statist. Assoc.*, **90**, 558–566.

— (2001) A review of asymptotic convergence for general state space Markov chains. *Far East J. Theor. Stat.*, **5**, 37–50.

Rubin, D. B. (1987) A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when the fraction of missing information is modest: the SIR algorithm (discussion of Tanner and Wong). *J. Am. Statist. Assoc.*, **82**, 543–546.

— (1988) Using the SIR algorithm to simulate posterior distribution. In *Bayesian Statistics 3* (eds. J. M. Bernardo, M. H. DeGroot, D. Lindley and A. Smith), 395–402. Clarendon Press.

Scott, D. J. and Tweedie, R. L. (1996) Explicit rates of convergence of stochastically ordered Markov chains. In *Athens Conference on Applied Probability and Time Series: Applied Probability in Honor of J. M. Gani*, vol. 114 of *Lecture Notes in Statistics*. Springer.

Segal, M. and Weinstein, E. (1989) A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix for linear dynamic systems. *IEEE Trans. Inform. Theory*, **35**, 682–687.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. Wiley.

Shiryaev, A. N. (1966) On stochastic equations in the theory of conditional Markov process. *Theory Probab. Appl.*, **11**, 179–184.

Stratonovich, R. L. (1960) Conditional Markov processes. *Theory Probab. Appl.*, **5**, 156–178.

Tanizaki, H. (2003) Nonlinear and non-Gaussian state-space modeling with Monte-Carlo techniques: a survey and comparative study. In *Handbook of Statistics 21. Stochastic processes: Modelling and Simulation* (eds. D. N. Shanbhag and C. R. Rao), 871–929. Elsevier.

Teicher, H. (1960) On the mixture of distributions. *Ann. Math. Statist.*, **31**, 55–73.

— (1961) Identifiability of mixtures. *Ann. Math. Statist.*, **32**, 244–248.

— (1963) Identifiability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.

— (1967) Identifiability of mixtures of product measures. *Ann. Math. Statist.*, **38**, 1300–1302.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985) *Statistical Analysis of Finite Mixture Distributions*. Wiley.

Van der Merwe, R., Doucet, A., De Freitas, N. and Wan, E. (2000) The unscented particle filter. In *Adv. Neural Inf. Process. Syst.* (eds. T. K. Leen, T. G. Dietterich and V. Tresp), vol. 13. MIT Press.

Van Overschee, P. and De Moor, B. (1993) Subspace algorithms for the stochastic identification problem. *Automatica*, **29**, 649–660.

— (1996) *Subspace Identification for Linear Systems. Theory, Implementation, Applications*. Kluwer.

Wald, A. (1949) Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595–601.

Weinstein, E., Oppenheim, A. V., Feder, M. and Buck, J. R. (1994) Iterative and sequential algorithms for multisensor signal enhancement. *IEEE Trans. Acoust., Speech, Signal Process.*, **42**, 846–859.

Whitley, D. (1994) A genetic algorithm tutorial. *Stat. Comput.*, **4**, 65–85.

Wonham, W. M. (1965) Some applications of stochastic differential equations to optimal nonlinear filtering. *SIAM J. Control*, **2**, 347–369.

Wu, C. F. J. (1983) On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.

Zangwill, W. I. (1969) *Nonlinear Programming: A Unified Approach*. Prentice-Hall.

Zaritskii, V., Svetnik, V. and Shimelevich, L. (1975) Monte-Carlo techniques in problems of optimal data processing. *Autom. Remote Control*, **12**, 2015–2022.

# Index