Randal Douc and Sylvain Le Corff

Bayesian Learning for partially observed dynamical systems

September 23, 2019

M2 Data Science

Contents

Plan of action

This is a 21 hours course schedule. The plan of action is the following.

1. Markov chains: kernel, invariant measures. Examples including Observation-Driven Models: lecture: 1H45, tutorials: 1H45. Randal Cours. Sylvain TD
2. Bayesian inference, asymptotic properties of the MLE for Markov chains : lecture: 1H45, tutorials: 1H45. Sylvain Cours et Td.
3. Markov Chains Monte Carlo (MCMC) algorithms: lecture: 1H45, computer sessions: 1H45. Randal Cours, Td.
4. Some properties of the MCMC algorithms: lecture: 1H45, tutorials: 1H45. Randal Cours.
5. Pseudo marginal MCMC and applications: lecture: 1H45, computer sessions: 1H45. Sylvain Cours, td.
6. Hamiltonian Monte Carlo algorithms: lecture: 1H30, tutorials and computer sessions: 1H45 Sylvain Cours, td.

Chapter 1

Preliminaries

Contents

These lecture notes are built from various sources:

(i) The monograph *Nonlinear Time Series: Theory, Methods, and Applications with R Examples*, by R. Douc, E. Moulines and D. Stoffer. Chapmann and Hall editors.
(ii) *Handbook of Markov Chain Monte Carlo* by Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. Chapmann and Hall editors.

Do not hesitate to point out the errors or typos that still remain and to propose any improvements on the content of this course. Email contact:

- `randal.douc@telecom-sudparis.eu`
- `sylvain.lecorff@telecom-sudparis.eu`

We start with **some notation**. In what follows,

- i.i.d means independent and identically distributed.
- r.v. means random variables.
- for $r,s \in \mathbb{N}$ such that $r \leq s$, we write $[r:s] = \{r, r+1, \dots, s\}$,
- $X \perp\!\!\!\perp Y$ means X and Y are independent random variables,
- $X \overset{\mathscr{L}}{=} Y$ means X and Y have the same law.
- Let $(\mathsf{Z}, \mathscr{Z})$ and $(\mathsf{X}, \mathscr{X})$ two measurable spaces. Assume that for all $w \in \mathsf{Z}$, the two random vectors X and $Y(w)$ take values in X and assume the existence of a third random variable W taking values in Z, then the notation: $X|_{W=w} \overset{\mathscr{L}}{=} Y(w)$ means that for all $A \in \mathscr{X}$,

$$\mathbb{P}(X \in A|W)|_{W=w} = \mathbb{P}(Y(w) \in A)$$

  In words, the distribution of X conditionally on W taken on $W = w$ is the same as the unconditonal distribution of $Y(w)$.
- $\liminf_n a_n = \lim_{n\to\infty} (\inf_{k\geq n} a_k)$ and similarly, $\limsup_n a_n = \lim_{n\to\infty} \left(\sup_{k\geq n} a_k\right)$. Moreover, $\lim_n a_n$ exists if and only if $\liminf_n a_n = \limsup_n a_n$.
- for any $a \in \mathbb{R}$, $a^+ = \max(a,0)$ and $a^- = \max(-a,0) = -\min(a,0)$ and we have $|a| = a^+ + a^-$ and $a = a^+ - a^-$.

Moreover, the following notions of convergence for random variables is used throughout these lecture notes.

- $\boxed{X_n \overset{w}{\Rightarrow} X}$ means *convergence in distribution* (or "convergence en loi" in French). It is equivalent to any of the following statements.

  (a) for all bounded continuous functions h, we have $\lim_n \mathbb{E}[h(X_n)] = \mathbb{E}[h(X)]$.
  (b) for all $A \in \mathscr{B}(\mathbb{R})$ such that $\mathbb{P}(X \in \partial A) = 0$, we have $\lim_n \mathbb{P}(X_n \in A) = \mathbb{P}(X \in A)$.
  (c) for all $x \in \mathbb{R}$ such that $\mathbb{P}(X = x) = 0$, we have $\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$.
  (d) for all $u \in \mathbb{R}$, we have $\lim_n \mathbb{E}\left[e^{iuX_n}\right] = \mathbb{E}\left[e^{iuX}\right]$

  By abuse of terminology, we may also say that $\boxed{X_n \textbf{ weakly converges to } X}$ instead of saying the distribution of X_n converges weakly to the distribution of X. An equivalent formulation is $\boxed{X_n \overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} X}$.

- $\boxed{X_n \overset{\mathbb{P}\text{-prob}}{\longrightarrow} X}$ means *convergence in probability*: for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

- $\boxed{X_n \overset{\mathbb{P}\text{-a.s.}}{\longrightarrow} X}$ means *almost sure convergence*:

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1\ .$$

$\boxed{\text{Almost sure convergence implies convergence in probability}}$. We recall the following properties.

(i) If $X_n \overset{w}{\Rightarrow} X$ then for all continuous functions f, $f(X_n) \overset{w}{\Rightarrow} f(X)$. Note that this property holds, when f is continuous (and not necessarily bounded), for example $f(u) = u^2$ so that $X_n^2 \overset{w}{\Rightarrow} X^2$.

(ii) **The Slutsky Lemma** If $X_n \overset{\mathbb{P}\text{-prob}}{\longrightarrow} c$ where c is a constant and if $Y_n \overset{w}{\Rightarrow} Y$, then $(X_n, Y_n) \overset{w}{\Rightarrow} (c, Y)$ that is for all continuous functions f, $f(X_n, Y_n) \overset{w}{\Rightarrow} f(c, Y)$.

(iii) $X \sim \mathrm{N}(0,1)$ iff for all $u \geq 0$, $\mathbb{E}\left[e^{iuX}\right] = e^{-u^2/2}$. Moreover, $X \sim \mathrm{N}(\mu, \sigma^2)$ iff for all $u \geq 0$, $\mathbb{E}\left[e^{iuX}\right] = e^{-u^2\mathbb{V}\mathrm{ar}(X)/2 + iu\mathbb{E}(X)}$ and in that case, $\sigma^2 = \mathbb{V}\mathrm{ar}(X)$ and $\mu = \mathbb{E}(X)$.

1.1 Refresher on Martingales

- "$\{\mathscr{F}_t,\ t \in \mathbb{N}\}$ is a filtration" means that $(\mathscr{F}_t)$ is an increasing sequence of σ-fields, that is $\mathscr{F}_t$ is a σ-field and $\mathscr{F}_t \subset \mathscr{F}_{t+1}$ for all $t \in \mathbb{N}$.
- "$(\Omega, \mathscr{F}, \{\mathscr{F}_t,\ t \in \mathbb{N}\}, \mathbb{P})$ is a filtered probability space" means that $\mathscr{F}$ is a σ-field on Ω, $\mathbb{P}$ is a probability on $(\Omega, \mathscr{F})$ and $\{\mathscr{F}_t,\ t \in \mathbb{N}\}$ is a filtration such that $\mathscr{F}_t \subset \mathscr{F}$ for all $t \in \mathbb{N}$.

Moreover, the sequence $(M_t)_{t\in\mathbb{N}}$ is a martingale wrt the filtration $\{\mathscr{F}_t,\ t \in \mathbb{N}\}$, if for all $t \geq 0$, M_t is a random variable on $(\Omega, \mathscr{F}, \mathbb{P})$ and

$$\mathbb{E}[M_{t+1}|\mathscr{F}_t] = M_t\ .$$

The following property on martingales will be important in this course.

Theorem 1.1. *Let $(M_t)_{t\in\mathbb{N}}$ be a martingale wrt the filtration $\{\mathscr{F}_t,\ t \in \mathbb{N}\}$ and denote $\Delta_t = M_t - M_{t-1}$ for $t \geq 1$. Assume that $\{\Delta_k,\ k \in \mathbb{N}\}$ is a strict-sense stationary and ergodic process such that $\mathbb{E}\left[M_1^2\right]$ is finite. Then,*

$$n^{-1/2} M_n \overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0, \mathbb{E}\left[\Delta_1^2\right])\ .$$

Another interesting property is actually a byproduct of the martingale theory since $M_n = \mathbb{E}[\mathbb{1}_A|\mathscr{F}_n]$ is a nonnegative martingale.

Lemma 1.2 *Let $A \in \mathscr{F} = \sigma(\cup_{k=0}^{\infty} \mathscr{F}_k)$. Then,*

$$\lim_{n\to\infty} \mathbb{E}[\mathbb{1}_A|\mathscr{F}_n] = \mathbb{1}_A \quad \mathbb{P} - \text{a.s.}$$

1.2 Some usual distributions

Name	Acronym	Parameter	density function: $f_X(x)$	cdf: $F_X(x)=\int_{-\infty}^{x} f_X(u)\mathrm{d}u$	Other properties
Gaussian	$\mathrm{N}(\mu,\sigma^2)$	(μ,σ^2)	$\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$	No explicit expression	$\mathbb{E}[X]=\mu,\ \mathbb{V}\mathrm{ar}X=\sigma^2$
Exponential	$\exp(\lambda)$	$\lambda>0$	$\lambda\mathrm{e}^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x)$	$\left(1-\mathrm{e}^{-\lambda x}\right)\mathbb{1}_{\mathbb{R}^+}(x)$	$\mathbb{E}[X]=1/\lambda,\ \mathbb{V}\mathrm{ar}X=1/\lambda^2$
Gamma	$\Gamma(k,\theta)$	$(k,\theta)\in\left(\mathbb{R}_+^*\right)^2$	$\frac{x^{k-1}\mathrm{e}^{-x/\theta}}{\Gamma(k)\theta^k}$	$\frac{\Gamma_{x/\theta}(k)}{\Gamma(k)}$	

In the above description,

(i) if $X_i\sim\Gamma(k_i,\theta)$ and (X_i) are independent, then $\sum_{i=1}^n X_i\sim\Gamma(\sum_{i=1}^n k_i,\theta)$.
(ii)

$$\Gamma(k)=\begin{cases}\int_0^\infty t^{k-1}\mathrm{e}^{-t}\mathrm{d}t & \text{if } k\in\mathbb{R}_+^* \\ k! & \text{if } k\in\mathbb{N}.\end{cases}\quad(\blacktriangleright\textsc{Gamma function})$$

$$\Gamma_x(k)=\int_0^x t^{k-1}\mathrm{e}^{-t}\mathrm{d}t\ (\blacktriangleright\textsc{incomplete Gamma function})$$

1.3 Abbreviations

$\mathbb{F}(\mathsf{X})$	The set of measurable functions on $(\mathsf{X},\mathscr{X})$
$\mathbb{F}_+(\mathsf{X})$	The set of nonnegative measurable functions on $(\mathsf{X},\mathscr{X})$
$\mathbb{M}_1(\mathscr{X})$	The set of probability measures on $(\mathsf{X},\mathscr{X})$
$\mathbb{M}_+(\mathscr{X})$	The set of (nonnegative) measures on $(\mathsf{X},\mathscr{X})$

Chapter 2

Introduction to Markov Chains

Contents

There are two rather different approaches to constructing a Markov chain. The first is through the transition laws of the chain. The second is through the use of iterating functions. This is particularly suited to the use of Markov chains in a time series context, where we wish to construct sample paths of the chain in an explicit manner rather than work with the distributions of the chain. This can also be carried out in a great degree of generality.

For both of these approaches we will discuss the existence of invariant measures (sometimes called stationary distributions) for the chain. This is the key property for most models because it implies (and is in some sense equivalent to) the existence of strict-sense stationary solutions. More precisely, when a Markov chain admits an invariant distribution, then we may construct strictly stationary versions of Markov chains; these stationary versions play a key role when considering the inference of Markov chains.

We will illustrate these constructions with several important time series models. It is worthwhile to note that, while many models in the nonlinear time series context might be viewed as (most of the time straightforward) extensions of ARMA models, finding necessary and sufficient conditions for the existence of an invariant distribution is a difficult task that will be covered in some generality in subsequent chapters.

Some of the material in this chapter will seem, to those interested in applications, to be both technical and exotic. However, we use the formal structures to define, in a rigorous manner, different time-series models. We encourage the reader to go through this material, perhaps skipping the proofs in a first pass.

A Markov chain is a discrete time stochastic process $X = \{X_t,\ t \in \mathbb{N}\}$, i.e., a countable collection of random variables defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$. In this definition, t is thought of as a time index and the set of times is taken by convention as $\mathbb{N}$.

The chain X evolves on a *state space* X. Although we let X be a general set (not assuming a priori any topological structure), in practice, it is most likely that we will be considering the set of real numbers or, more generally, the d-dimensional Euclidean space, so that $\mathsf{X} = \mathbb{R}^d$, or some countable or uncountable subset of $\mathbb{R}^d$. In order to define probabilities, we use $\mathscr{X}$ to denote a countably generated σ-field on X. When $\mathsf{X} = \mathbb{R}^d$, then $\mathscr{X}$ will be taken as the Borel σ-field $\mathscr{X} = \mathscr{B}(\mathsf{X})$. When X is countable, $\mathscr{X}$ contains all the subsets, that is $\mathscr{X} = \mathscr{P}(\mathsf{X})$, the set of all parts of X. All random variables are assumed measurable individually with respect to $\mathscr{B}(\mathsf{X})$, and we shall in general denote elements of X by letters $x,\ y,\ z,\ \ldots$ and elements of $\mathscr{X}$ by $A,\ B,\ C$.

The distinguishing feature of a Markov chain as opposed to an arbitrary stochastic process is that the future, given the present, is forgetful of the past. This is reflected in the conditional distributions of the future random variable X_{t+1} given its present, X_t, and its past $\{X_0,\dots,X_{t-1}\}$.

Definition 2.1 (Markov chain). Let $(\Omega,\mathscr{F},\{\mathscr{F}_t,\ t\in\mathbb{N}\},\mathbb{P})$ be a filtered probability space. An adapted stochastic process $\{(X_t,\mathscr{F}_t),\ t\in\mathbb{N}\}$ is called a Markov chain of *order* m (or with *memory* m) if, for all $t\geq m-1$ and $A\in\mathscr{X}$, $\mathbb{P}\left[X_{t+1}\in A|\mathscr{F}_t\right]=\mathbb{P}\left[X_{t+1}\in A|X_t,\dots,X_{t-m+1}\right]$, $\mathbb{P}-$a.s.

When the order is one, we simply say that $\{(X_t,\mathscr{F}_t),\ t\in\mathbb{N}\}$ is a Markov chain. This definition is equivalent to assuming that for all $f\in\mathbb{F}_+(\mathsf{X})$,

$$\mathbb{E}\left[f(X_t)|\mathscr{F}_{t-1}\right]=\mathbb{E}\left[f(X_t)|X_{t-1}\right],\quad \mathbb{P}-\text{a.s.}$$

Note that a Markov chain $\{(X_t,\mathscr{F}_t),\ t\in\mathbb{N}\}$ is always a Markov chain with respect to the *natural filtration* $\{\mathscr{F}_t^X,\ t\in\mathbb{N}\}$, where $\mathscr{F}_t^X=\sigma(X_0,\dots,X_t)$.

2.1 Kernels

The construction of Markov chains is based on the definition of kernels.

Definition 2.2. Let $(\mathsf{X},\mathscr{X})$ and $(\mathsf{Y},\mathscr{Y})$ be two measurable spaces.

- A *kernel* is a mapping N from $\mathsf{X}\times\mathscr{Y}$ into $[0,\infty]$ satisfying the following conditions:
  (i) for every $x\in\mathsf{X}$, the mapping $N(x,\cdot):A\mapsto N(x,\ A)$ is a measure on $\mathscr{Y}$,
  (ii) for every $A\in\mathscr{Y}$ the mapping $N(\cdot,A):x\to N(x,A)$ is a measurable function from $(\mathsf{X},\mathscr{X})$ to $[0,\infty]$.
- The kernel N is said to be *finite* if $N(x,\mathsf{Y})<\infty$ for all $x\in\mathsf{X}$.
- The kernel N is said to be *bounded* if $\sup_{x\in\mathsf{X}}N(x,\mathsf{Y})<\infty$.
- A kernel N is said to be *Markovian* if $N(x,\mathsf{Y})=1$, for all $x\in\mathsf{X}$.

Example 2.3 (Discrete state-space kernel) *Assume that* X *and* Y *are countable sets. Each element* $x\in\mathsf{X}$ *is then called a state. A kernel* N *on* $\mathsf{X}\times\mathscr{P}(\mathsf{Y})$*, where* $\mathscr{P}(\mathsf{Y})$ *is the set of all parts of* Y*, is specified by a (possibly infinite) matrix* $N=(N(x,y):x,y\in\mathsf{X}\times\mathsf{Y})$ *with nonnegative entries. Each row* $(N(x,y):y\in\mathsf{Y})$ *is a measure on* $(\mathsf{Y},\mathscr{P}(\mathsf{Y}))$ *defined by*

$$N(x,A)=\sum_{y\in A}N(x,y)\ ,$$

for $A\subset\mathsf{Y}$*. The matrix* N *is said to be Markovian if every row* $(N(x,y):y\in\mathsf{Y})$ *is a probability measure on* $(\mathsf{Y},\mathscr{P}(\mathsf{Y}))$*, i.e.* $\sum_{y\in\mathsf{Y}}N(x,y)=1$ *for all* $x\in\mathsf{X}$*.*

Example 2.4 (Kernel density) *Let* $\lambda\in\mathbb{M}_1(\mathscr{Y})$ *and* $n:\mathsf{X}\times\mathsf{Y}\to\mathbb{R}_+$ *be a nonnegative function, measurable with respect to the product* σ*-algebra* $\mathscr{X}\otimes\mathscr{Y}$*. Then, the application* N *defined on* $\mathsf{X}\times\mathscr{Y}$ *by*

$$N(x,A)\ =\int_A n(x,y)\lambda(\mathrm{d}y)\ ,$$

is a kernel. The function n *is called the density of the kernel* N *with respect to the measure* λ*. The kernel* N *is Markovian if and only if* $\int_{\mathsf{Y}}n(x,y)\lambda(\mathrm{d}y)=1$ *for all* $x\in\mathsf{X}$*.*

Let N be a kernel and $f \in \mathbb{F}_+(\mathsf{Y})$. A function $Nf : \mathsf{X} \to \mathbb{R}^+$ is defined by setting, for any $x \in \mathsf{X}$,

$$\boxed{Nf : x \mapsto \int_{\mathsf{Y}} N(x, \mathrm{d}y) f(y) \, .}$$

Proposition 2.5 *Let N be a kernel on $\mathsf{X} \times \mathscr{Y}$. For any $f \in \mathbb{F}_+(\mathsf{Y})$, $Nf \in \mathbb{F}_+(\mathsf{X})$.*

With a slight abuse of notation, we will use the same symbol N for the kernel and the associated operator $N : \mathbb{F}_+(\mathsf{Y}) \to \mathbb{F}_+(\mathsf{X})$, $f \mapsto Nf$. By defining $Nf = Nf^+ - Nf^-$, we may extend the application $f \mapsto Nf$ to all functions f of $\mathbb{F}(\mathsf{Y})$ such that Nf^+ and Nf^- are not both infinite. We will sometimes write $N(x, f)$ for $Nf(x)$ and make use of the notations $N(x, \mathbb{1}_A)$ and $N\mathbb{1}_A(x)$ for $N(x, A)$.

Let μ be a positive measure on $(\mathsf{X}, \mathscr{X})$ and for $A \in \mathscr{Y}$, define

$$\boxed{\mu N(A) = \int_{\mathsf{X}} \mu(\mathrm{d}x) \, N(x, \, A) \, .}$$

Proposition 2.6 *Let N be a kernel on $\mathsf{X} \times \mathscr{Y}$ and $\mu \in \mathbb{M}_+(\mathscr{X})$. Then $\mu N \in \mathbb{M}_+(\mathscr{Y})$.*

Let $(\mathsf{X}, \mathscr{X})$, $(\mathsf{Y}, \mathscr{Y})$ and $(\mathsf{Z}, \mathscr{Z})$ be measurable spaces and let M and N be two kernels on $\mathsf{X} \times \mathscr{Y}$ and $\mathsf{Y} \times \mathscr{Z}$. For any $A \in \mathscr{Z}$, $y \mapsto N(y, A)$ is a measurable function, and by Proposition 2.5, $x \mapsto \int_{\mathsf{Y}} M(x, \mathrm{d}y) N(y, A)$ is a measurable function. For any $x \in \mathsf{X}$, $M(x, \cdot)$ is a measure on $(\mathsf{Y}, \mathscr{Y})$ and by Proposition 2.6, $A \mapsto \int M(x, \mathrm{d}y) N(y, A)$ is a measure on $(\mathsf{Z}, \mathscr{Z})$. Hence, the function $MN : (x, A) \mapsto \int_{\mathsf{X}} M(x, \mathrm{d}y) N(y, A)$ is a kernel on $\mathsf{X} \times \mathscr{Z}$.

The *composition* or *product* of the kernels M on $\mathsf{X} \times \mathscr{Y}$ and N on $\mathsf{Y} \times \mathscr{Z}$ is the kernel MN defined for $x \in \mathsf{X}$ and $A \in \mathscr{Z}$ by

$$\boxed{MN(x, A) = \int_{\mathsf{Y}} M(x, \mathrm{d}y) N(y, A) \, .} \tag{2.1}$$

Since MN is a kernel on $\mathsf{X} \times \mathscr{Z}$, for any $f \in \mathbb{F}_+(\mathsf{Z})$, we may define the function $MNf : x \mapsto MNf(x)$, which by Proposition 2.5 belongs to $\mathbb{F}_+(\mathsf{X})$. On the other hand, $Nf : y \mapsto Nf(y)$ is a function belonging to $\mathbb{F}_+(\mathsf{Y})$, and since M is a kernel on $\mathsf{X} \times \mathscr{Y}$, we may consider the function $x \mapsto M[Nf](x)$. As shown in the following proposition, these two quantities coincide.

Proposition 2.7 *Let M be a kernel on $\mathsf{X} \times \mathscr{Y}$ and N be a kernel on $\mathsf{Y} \times \mathscr{Z}$. Then, for each $x \in \mathsf{X}$, and $f \in \mathbb{F}_+(\mathsf{Z})$,*

$$MNf(x) = M[Nf](x) \, . \tag{2.2}$$

For $x \in \mathsf{X}$ and $A \in \mathscr{X}$, we set $N^0(x, \, A) = \delta_x(A)$ and for $n \geq 1$, we define inductively N^n by

$$\boxed{N^n(x, \, A) = \int_{\mathsf{X}} N(x, \mathrm{d}y) N^{n-1}(y, \, A) \, .} \tag{2.3}$$

We finally define the tensor products of kernels.

Proposition 2.8 *Let M be a kernel on $\mathsf{X} \times \mathscr{Y}$ and N be a kernel on $\mathsf{Y} \times \mathscr{Z}$. Then, there exists a kernel $M \otimes N$ on $\mathsf{X} \times (\mathscr{Y} \otimes \mathscr{Z})$, called the tensor product of M and N, such that, for all $f \in \mathbb{F}_+(\mathsf{Y} \times \mathsf{Z}, \mathscr{Y} \otimes \mathscr{Z})$,*

$$M \otimes Nf(x) = \int_{\mathsf{Y}} M(x, \mathrm{d}y) \int_{\mathsf{Z}} f(y, z) N(y, \mathrm{d}z) \, . \tag{2.4}$$

For $n \geq 1$, the n-th tensorial power $P^{\otimes n}$ of a kernel P on $X \times \mathscr{Y}$ is the kernel on $(\mathsf{X}, \mathscr{X}^{\otimes n})$ defined by

$$P^{\otimes n} f(x) = \int_{\mathsf{X}^n} f(x_1, \ldots, x_n) P(x, \mathrm{d}x_1) P(x_1, \mathrm{d}x_2) \cdots P(x_{n-1}, \mathrm{d}x_n) \, .$$

2.2 Homogeneous Markov chain

Definition 2.9 (Homogeneous Markov chain). Let $(\Omega, \mathscr{F}, \{\mathscr{F}_t,\, t\in\mathbb{N}\}, \mathbb{P})$ be a filtered probability space. Let $m \geq 1$ be an integer and P a Markov kernel on $\mathsf{X}^m \times \mathscr{X}$. An adapted stochastic process $\{(X_t, \mathscr{F}_t),\, t\in\mathbb{N}\}$ is called an *homogeneous Markov chain of order* m with Markov kernel P and initial distribution $\nu \in \mathbb{M}_1(\mathscr{X}^{\otimes m})$ if

(i) for all $A \in \mathscr{X}^{\otimes m}$, $\mathbb{P}((X_0,\ldots,X_{m-1}) \in A) = \nu(A)$,
(ii) for all $t \geq m-1$ $\mathbb{P}[X_{t+1} \in A | \mathscr{F}_t] = P(X_t,\ldots,X_{t-m+1}, A)$, $\quad \mathbb{P}-$a.s.

By construction, for any function $f_0 \in \mathbb{F}_+(\mathsf{X})$, $\mathbb{E}[f_0(X_0)] = \int \nu(\mathrm{d}x_0) f_0(x_0) = \nu(f_0)$. For all functions $f_0, f_1 \in \mathbb{F}_+(\mathsf{X})$, we may write similarly

$$\begin{aligned}\mathbb{E}[f_0(X_0)f_1(X_1)] = \mathbb{E}[f_0(X_0)\mathbb{E}[f_1(X_1)|\mathscr{F}_0]] &= \mathbb{E}[f_0(X_0)Pf_1(X_0)] \\ &= \iint \nu(\mathrm{d}x_0)P(x_0,\mathrm{d}x_1)f_0(x_0)f_1(x_1) = \nu\otimes P(f_0\otimes f_1)\,,\end{aligned}$$

where $f_0\otimes f_1(x,x') = f_0(x)f_1(x')$. Assume that for some $t \geq 1$, and for all functions $f_0,\ldots,f_t \in \mathbb{F}_+(\mathsf{X})$,

$$\begin{aligned}\mathbb{E}[f_0(X_0)f_1(X_1)\ldots f_t(X_t)] &= \int \nu(\mathrm{d}x_0)\prod_{s=1}^{t} P(x_{s-1},\mathrm{d}x_s)\prod_{s=0}^{t} f_s(x_s) \\ &= \nu\otimes P^{\otimes t}(f_0\otimes f_1\otimes\cdots\otimes f_t)\,.\end{aligned} \tag{2.5}$$

Then, for any $f_{t+1} \in \mathbb{F}_+(\mathsf{X})$,

$$\begin{aligned}\mathbb{E}[f_0(X_0)\ldots f_t(X_t)f_{t+1}(X_{t+1})] &= \mathbb{E}[f_0(X_0)\ldots f_t(X_t)\mathbb{E}[f_{t+1}(X_{t+1})|\mathscr{F}_t]] \\ &= \mathbb{E}[f_0(X_0)\ldots f_t(X_t)Pf_{t+1}(X_{t+1})] = \nu\otimes P^{\otimes(t+1)}(f_0\otimes f_1\otimes\cdots\otimes f_{t+1})\,,\end{aligned}$$

showing that (2.5) is still true with t replaced by $t+1$, and hence is satisfied for all $t\in\mathbb{N}$. Since the cylinder sets generate the product σ-algebra $\mathscr{X}^{\otimes\mathbb{N}}$, the initial distribution ν on $(\mathsf{X},\mathscr{X})$ and the Markov kernel P allows to define a unique distribution on $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$.

Conversely, assume that for all $t\in\mathbb{N}$, (2.5) is satisfied. Then, using the Tonelli-Fubini theorem, we may write

$$\begin{aligned}&\mathbb{E}[f_0(X_0)\ldots f_{t-1}(X_{t-1})f_t(X_t)] \\ &\quad= \int \nu(\mathrm{d}x_0)\prod_{s=1}^{t-1} P(x_{s-1},\mathrm{d}x_s)\prod_{s=0}^{t-1} f_s(x_s)\int P(x_{t-1},\mathrm{d}x_t)f_t(x_t) \\ &\quad= \mathbb{E}[f_0(X_0)\ldots f_{t-1}(X_{t-1})Pf_t(X_{t-1})]\,,\end{aligned}$$

showing that for any $\mathscr{F}_{t-1}$-measurable random variable $Y = f_0(X_0)\ldots f_{t-1}(X_{t-1})$, $\mathbb{E}[Yf_t(X_t)] = \mathbb{E}[YPf_t(X_{t-1})]$, which implies that $\mathbb{E}[f_t(X_t)|\mathscr{F}_{t-1}] = Pf_t(X_{t-1})$. Hence, an adapted stochastic process satisfying (2.5) is an homogeneous Markov chain with initial distribution ν and Markov kernel P. We summarize these conclusions in the following theorem.

Theorem 2.10. *Let $\{(X_t,\mathscr{F}_t),\, t\in\mathbb{N}\}$ be a Markov chain on $(\mathsf{X},\mathscr{X})$ with initial distribution ν and Markov kernel P. For any $f\in\mathbb{F}_b(\mathsf{X}^{t+1},\mathscr{X}^{\otimes(t+1)})$, and $t\in\mathbb{N}$, we have*

$$\mathbb{E}[f(X_0,\ldots,X_t)] = \nu\otimes P^{\otimes t}(f)\,. \tag{2.6}$$

Conversely, let $\{X_t,\, t\in\mathbb{N}\}$ be a stochastic process on $(\mathsf{X},\mathscr{X})$ satisfying (2.6) *for some probability ν and a Markov kernel P on $\mathsf{X}\times\mathscr{X}$. Then, $\{X_t,\, t\in\mathbb{N}\}$ is a Markov chain with initial distribution ν and transition probability P.*

PROOF. The essence of the proof is presented above; technical details are filled in Exercise 8.10 and Exercise 8.11. ■

Example 2.11 (Autoregressive process of order 1) *Consider the AR(*1*) process*

$$X_{t+1} = \phi X_t + Z_{t+1}\,, \quad t\in\mathbb{N}\,, \tag{2.7}$$

where $\{Z_t,\, t\in\mathbb{N}^*\}$ is a sequence of zero-mean i.i.d. random variables independent from X_0. For $t\geq 0$, define $\mathscr{F}_t=\sigma(X_0,Z_s,s\leq t)$. The process $\{(X_t,\mathscr{F}_t),\, t\in\mathbb{N}\}$ given by (2.7) *is a Markov chain. For any $x\in\mathsf{X}$ and $A\in\mathscr{B}(\mathbb{R})$, the Markov kernel of this chain is given by $P(x,A)=\mathbb{E}\left[\mathbb{1}_A(\phi x+Z_0)\right]=\mu(A-\phi x)$, where μ is the distribution of Z_0. If the law of Z_0 has a density q, with respect to Lebesgue's measure on $\mathbb{R}$, then the Markov kernel also has a density $p(x,y)=q(y-\phi x)$, i.e.*

$$P(x,A)=\int_A q(y-\phi x)\,\mathrm{Leb}(\mathrm{d}y)\,.$$

Example 2.12 (ARCH(1) process) *An autoregressive conditional heteroscedastic (ARCH) model of order* 1 *is defined as*

$$X_t=\sigma_t Z_t\,,\quad \sigma_t^2=\alpha_0+\alpha_1 X_{t-1}^2\,,\quad t\geq 1\,, \tag{2.8}$$

where the coefficients α_0,α_1 are positive and $\{Z_t,\, t\in\mathbb{N}\}$ is an i.i.d. sequence such that $\mathbb{E}\left[Z_0\right]=0$, $\mathbb{E}\left[Z_0^2\right]=1$, and $\{Z_t,\, t\in\mathbb{N}\}$ is independent of X_0. Equation (2.8) *may be equivalently rewritten as*

$$X_t=\sqrt{\alpha_0+\alpha_1 X_{t-1}^2}Z_t\,,$$

which shows that $\{X_t,\, t\in\mathbb{N}\}$ is an homogeneous Markov chain. The Markov kernel of this chain is given by $P(x,A)=\mathbb{P}(\sqrt{\alpha_0+\alpha_1 x^2}Z\in A)$, for $x\in\mathbb{R}$ and $A\in\mathscr{B}(\mathbb{R})$. If the distribution of Z_0 has a density q with respect to the Lebesgue measure on $(\mathbb{R},\mathscr{B}(\mathbb{R}))$, then the Markov kernel P has a density p given by

$$p(x,x')=\frac{1}{\sqrt{\alpha_0+\alpha_1 x^2}}q\left(\frac{x'}{\sqrt{\alpha_0+\alpha_1 x^2}}\right)\,.$$

The square volatility $\{\sigma_t^2, t\geq 1\}$ satisfies the recursion: for $t\geq 2$,

$$\sigma_t^2=\alpha_0+\alpha_1 Z_{t-1}^2\sigma_{t-1}^2$$

and $\sigma_1^2=\alpha_0+\alpha_1 X_0^2$. This is again a Markov chain with state space $\mathbb{R}^+$ and Markov kernel Q defined by $Q(s,A)=\mathbb{P}(\alpha_0+\alpha_1 sZ^2\in A)$ for all $s\in\mathbb{R}^+$ and $A\in\mathscr{B}(\mathbb{R}^+)$.

Example 2.13 (A simple Markov bilinear process) *Consider the following bilinear process*

$$X_t=aX_{t-1}+bZ_tX_{t-1}+Z_t\,,$$

where $\{Z_t,\, t\in\mathbb{N}\}$ is an i.i.d. sequence independent from X_0. The Markov kernel of the chain $\{X_t,\, t\in\mathbb{N}\}$ is given by $P(x,A)=\mathbb{P}(ax+(1+bx)Z\in A)$ for $x\in\mathbb{R}$ and $A\in\mathscr{B}(\mathbb{R})$. If the distribution of Z_0 has a density q with respect to the Lebesgue measure and if $x\neq -1/b$, then the Markov kernel P has a density, denoted p, given by

$$p(x,x')=\frac{1}{|1+bx|}q\left(\frac{x'-ax}{1+bx}\right)\,.$$

If $x=-1/b$, then $P(-1/b,\cdot)=\delta_{-a/b}$.

2.3 Canonical representation

In this section, we show that, given an initial distribution $\nu \in \mathbb{M}_1(\mathscr{X})$ and a Markov kernel P on $\mathsf{X} \times \mathscr{X}$, we can construct a filtered probability space and a Markov chain with initial distribution ν and transition kernel P. A first construction has been given by Theorem 2.37, under some topological restriction. The following construction imposes no additional assumption on $(\mathsf{X}, \mathscr{X})$. Let $\mathsf{X}^{\mathbb{N}}$ be the set of X-valued sequences $w = (w_0, \dots, w_n, \dots)$ endowed with the product σ-field $\mathscr{X}^{\otimes \mathbb{N}}$. We define the coordinate process $\{X_t,\, t \in \mathbb{N}\}$ by

$$X_t(w) = w_t \; . \tag{2.9}$$

The natural filtration of the coordinate process is denoted by $\{\mathscr{F}_t^X,\, t \in \mathbb{N}\}$; by construction, $\bigvee_{n \geq 0} \mathscr{F}_n^X = \mathscr{X}^{\otimes \mathbb{N}}$.

Theorem 2.14. *Let P be a Markov kernel P on $\mathsf{X} \times \mathscr{X}$ and $\nu \in \mathbb{M}_1(\mathscr{X})$. Then, there exists a unique probability measure $\mathbb{P}_\nu$ on $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes \mathbb{N}})$ such that the coordinate process is a Markov chain with initial distribution ν and Markov kernel (or transition probability) P.*

PROOF. Set for $t \in \mathbb{N}$ and $f \in \mathbb{F}_+(\mathsf{X}^{t+1}, \mathscr{X}^{\otimes (t+1)})$,

$$\mu_t(f) = \nu \otimes P^{\otimes t}(f) = \int \nu(\mathrm{d}x_0) \int P(x_0, \mathrm{d}x_1) \dots \int P(x_{t-1}, \mathrm{d}x_t) f(x_0, x_1, \dots, x_t) \; .$$

By Proposition 2.8, $\mu_t \in \mathbb{M}_1(\mathscr{X}^{\otimes (t+1)})$ and the family of finite dimensional distributions $\{\mu_t,\, t \in \mathbb{N}\}$ satisfies the usual consistency conditions. We conclude by applying the Kolmogorov extension theorem. ∎

For $x \in \mathsf{X}$, we use the shorthand notation $\mathbb{P}_x = \mathbb{P}_{\delta_x}$. We denote by $\mathbb{E}_\nu$ the expectation associated to $\mathbb{P}_\nu$. For all $A \in \mathscr{F}$, the function $x \mapsto \mathbb{P}_x(A)$ is $\mathscr{X}$-measurable. For all $\nu \in \mathbb{M}_1(\mathscr{X})$ and $A \in \mathscr{F}$, $\mathbb{P}_\nu(A) = \int_{\mathsf{X}} \mathbb{P}_x(A) \nu(\mathrm{d}x)$.

2.4 Invariant measures

Stochastic processes are *strict-sense stationary* if, for any integer k, the distribution of the random vector $(X_t, \dots, X_{t+k})$ does not depend on the time-shift t. In general, a Markov chain will not be stationary. For example, suppose we define an AR(1) model $\{X_t,\, t \in \mathbb{N}\}$, as $X_t = \phi X_{t-1} + Z_t$ where $|\phi| < 1$ and $Z_t \sim_{\text{iid}} \mathrm{N}(0, \sigma^2)$. To initialize the process, set for example $X_0 = Z_0$. In this case, $X_0 \sim \mathrm{N}(0, \sigma^2)$, but $X_1 = \phi X_0 + Z_1 \sim \mathrm{N}(0, \sigma^2 (1 + \phi^2))$, so that X_0 and X_1 do not have the same distribution. A simple fix is to put $X_0 = Z_0 / \sqrt{1 - \phi^2}$, in which case

$$X_t = \phi^t X_0 + \sum_{s=0}^{t-1} \phi^s Z_{t-s} \sim \mathrm{N}(0, \sigma^2 / (1 - \phi^2)) \; ,$$

for all $t \geq 0$.

Under appropriate conditions on the Markov kernel P, it is possible to produce a stationary process with a proper choice of the initial distribution π. Assuming that such a distribution exists, the stationarity of the marginal distribution implies that $\mathbb{E}_\pi[\mathbb{1}_A(X_0)] = \mathbb{E}_\pi[\mathbb{1}_A(X_1)]$ for any $A \in \mathscr{X}$. This can equivalently be written as $\pi(A) = \pi P(A)$, or $\pi = \pi P$. We can iterate to give, for any $h \in \mathbb{N}$, $\pi P^h = \pi$. For all integers h and n, and all $A_0, \dots, A_n \in \mathscr{X}$,

$$\begin{aligned} \mathbb{P}_\pi(X_h \in A_0, X_{h+1} \in A_1, \dots, X_{h+n} \in A_n) &= \int \cdots \int \pi P^h(\mathrm{d}x_0) \prod_{i=1}^{n} P(x_{i-1},\, \mathrm{d}x_i)\, \mathbb{1}_{A_i}(x_i) \\ &= \int \cdots \int \pi(\mathrm{d}x_0) \prod_{i=1}^{n} P(x_{i-1},\, \mathrm{d}x_i)\, \mathbb{1}_{A_i}(x_i) = \mathbb{P}_\pi(X_0 \in A_0,\, X_1 \in A_1,\, \dots,\, X_n \in A_n) \; , \end{aligned}$$

which shows that, for any integers h and n, the random vectors $(X_h, X_{h+1}, \ldots, X_{h+n})$ and $(X_0, X_1, \ldots, X_n)$ have the same distributions. Therefore, the Markov property implies that all finite-dimensional distributions of $\{X_t,\, t \in \mathbb{N}\}$ are also invariant under translation in time. These considerations lead to the definition of the *invariant measure*.

Definition 2.15 (Invariant measure). Given a Markov kernel P, a σ-finite measure π on $(\mathsf{X}, \mathscr{X})$ with the property

$$\pi(A) = \pi P(A)\,, \quad A \in \mathscr{X}\,,$$

will be called *invariant* (with respect to P).

If an invariant measure is finite, it may be normalized to an *invariant probability measure*. In general, there may exist more than one invariant measure, and if X is not finite, an invariant measure may not exist. As a trivial example, consider $\mathsf{X} = \mathbb{N}$ and $P(x, x+1) = 1$.

It turns out that we can extend a stationary Markov chain to have time t take on negative values, as well.

Proposition 2.16 *Let $(\mathsf{X}, \mathscr{X})$ be a measurable space. Let P a Markov kernel on $(\mathsf{X}, \mathscr{X})$ admitting π has an invariant distribution. Then, there exists on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ a stochastic process $\{X\}$ such that, for any integer p, any p-tuple $t_1 < t_2 < \cdots < t_p$, and any $A \in \mathscr{X}^{\otimes p}$,*

$$\begin{aligned}\mathbb{P}((X_{t_1}, X_{t_2}, \ldots, X_{t_p}) \in A) \\ = \int \cdots \int \pi(\mathrm{d}x_1) P^{t_2 - t_1}(x_1, \mathrm{d}x_2) \ldots P^{t_p - t_{p-1}}(x_{p-1}, \mathrm{d}x_p) \mathbb{1}_A(x_1, \ldots, x_p)\,. \quad (2.10)\end{aligned}$$

PROOF. This is a direct consequence of the Kolmogorov extension theorem. ■

Remark 2.17 *We may take $\Omega = \prod_{t=-\infty}^{\infty} \mathsf{X}_t$ and $\mathscr{F} = \bigvee_{t=-\infty}^{\infty} \mathscr{X}_t$, where $(\mathsf{X}_t, \mathscr{X}_t)$ is a copy of $(\mathsf{X}, \mathscr{X})$. For $t \in \mathbb{Z}$, X_t is the t-th coordinate mapping, i.e. for $\{\omega\}$ an element of Ω, $X_t(\omega) = \omega_t$. This is the canonical two-sided extension of a stationary Markov chain.*

Example 2.18 (Markov chain over a finite state-space) *Consider a Markov chain on a finite state space $\mathsf{X} = \{1, \ldots, n\}$ with transition kernel P. A probability measure is a vector ξ with nonnegative entries summing to 1, $\sum_x \xi(x) = 1$. If ξ is the initial distribution, then after one step, the distribution of the chain is ξP; after t steps, the distribution is ξP^t. The probability π is stationary if and only if $\pi P = \pi$. This means that 1 should be an eigenvalue of P and in this case π is the left-eigenvector of P associated to the eigenvalue 1. It may be shown that, provided there exists an integer m such that $P^m(x, x') > 0$ for all $(x, x') \in \mathsf{X} \times \mathsf{X}$ (the Markov kernel P is said to be irreducible in such a case), such distribution exists and is unique.*

Example 2.19 (Gaussian AR(1) processes) *Consider a Gaussian AR(1) process, $X_t = \mu + \phi X_{t-1} + \sigma Z_t$, where $\{Z_t,\, t \in \mathbb{N}\}$ is an i.i.d. sequence of standard Gaussian random variables, independent of X_0. Assume that $|\phi| < 1$ and that X_0 is Gaussian with mean μ_0 and variance γ_0^2. Then X_1 is Gaussian with mean $\mu + \phi\mu_0$, and variance $\phi^2\gamma_0^2 + \sigma^2$. If we choose*

$$\begin{cases} \mu + \phi\mu_0 = \mu_0 \\ \phi^2\gamma_0^2 + \sigma^2 = \gamma_0^2 \end{cases} \quad \Rightarrow \quad \begin{cases} \mu_0 = \mu/(1-\phi) \\ \gamma_0^2 = \sigma^2/(1-\phi^2) \end{cases}$$

then X_1 and X_0 have the same distribution. Therefore, the Gaussian distribution with mean $\mu/(1-\phi)$ and variance $\sigma^2/(1-\phi^2)$ is a stationary distribution. We will show later that this distribution is unique.

Example 2.20 *Consider a Markov chain whose state space* $\mathsf{X} = (0,1)$ *is the open unit interval. If the chain is at x, it picks one of the two intervals* $(0,\ x)$ *or* $(x,\ 1)$ *with equal probability* $1/2$*, and then moves to a point y which is uniformly distributed in the chosen interval. This Markov chain has a transition density with respect to Lebesgue measure on the interval* $(0,1)$*, which is given by*

$$k(x,y) = \frac{1}{2}\frac{1}{x}\mathbb{1}_{(0,x)}(y) + \frac{1}{2}\frac{1}{1-x}\mathbb{1}_{(x,1)}(y)\ . \tag{2.11}$$

The first term in the sum corresponds to a move from x to the interval $(0,x)$*; the second, to a move from x to the interval* $(x,1)$*.*

This Markov chain can be equivalently represented as an iterated random sequence. Let $\{U_t,\ t \in \mathbb{N}\}$ *be a sequence of i.i.d. random variable uniformly distributed on the interval* $(0,1)$*. Let* $\{\varepsilon_t,\ t \in \mathbb{N}\}$ *be a sequence of i.i.d. random Bernoulli variables distributed with probability of success* $1/2$*, independent of* $\{U_t,\ t \in \mathbb{N}\}$*. Let* X_0*, the initial state, be distributed according to some initial distribution* ξ *on* $(0,1)$*, and be independent of* $\{U_t,\ t \in \mathbb{N}\}$ *and* $\{\varepsilon_t,\ t \in \mathbb{N}\}$*. Define the sequence* $\{X_t,\ t \in \mathbb{N}\}$ *for* $t \geq 1$ *as follows*

$$X_t = \varepsilon_t\left[X_{t-1}U_t\right] + (1-\varepsilon_t)\left[X_{t-1} + U_t(1-X_{t-1})\right]\ , \tag{2.12}$$

Begin by assuming that the stationary distribution has a density p with respect to Lebesgue measure. From (2.11)*,*

$$p(y) = \int_0^1 k(x,\ y)p(x)\mathrm{d}x = \frac{1}{2}\int_y^1 \frac{p(x)}{x}\mathrm{d}x + \frac{1}{2}\int_0^y \frac{p(x)}{1-x}\mathrm{d}x\ . \tag{2.13}$$

Differentiation gives

$$p'(x) = -\frac{1}{2}\frac{p(x)}{x} + \frac{1}{2}\frac{p(x)}{1-x} \qquad or \qquad \frac{p'(x)}{p(x)} = \frac{1}{2}\left(-\frac{1}{x} + \frac{1}{1-x}\right).$$

The solutions of this first-order non-linear differential equation are given by

$$f_C(x) = \frac{C}{\sqrt{x(1-x)}}\ , \tag{2.14}$$

where $C \in \mathbb{R}$ *is a constant. Note that for* $x \in (0,1)$ *and* $C \geq 0$*,* $f_C(x) \geq 0$*. For all* $z \in (0,1)$*,*

$$\int_0^z f_C(y)\mathrm{d}y = 2C\arcsin(\sqrt{z})\ . \tag{2.15}$$

Choosing $C = 1/\pi$ *makes* $\int_0^1 f_{1/\pi}(y)\mathrm{d}y = 1$*. Therefore, the* arcsine density $p = f_{1/\pi}$ *is a stationary distribution. It is the unique stationary distribution admitting a density with respect to Lebesgue measure. We will show later that p is indeed the unique stationary distribution.*

2.5 Observation-driven models

Of course there are many time series models that are not Markov chains. However, in many useful situations, Markov chains are not very far from the surface.

Definition 2.21 (Observation-driven model). Let $(\mathsf{X},\mathscr{X})$ and $(\mathsf{Y},\mathscr{Y})$ be a measurable space, Q be Markov kernel on $\mathsf{X}\times\mathscr{Y}$ and $(x,y) \mapsto f_y(x)$ a measurable function from $(\mathsf{X}\times\mathsf{Y}, \mathscr{X}\otimes\mathscr{Y})$ to $(\mathsf{X},\mathscr{X})$. Let $(\Omega,\mathscr{F},\{\mathscr{F}_t\},\mathbb{P})$ be a filtered probability space.

An *observation-driven time series model* is a stochastic process $\{(X_t,Y_t),\ t \in \mathbb{N}\}$ adapted to $(\Omega,\mathscr{F},\{\mathscr{F}_t\},\mathbb{P})$ taking values in $\mathsf{X}\times\mathsf{Y}$ satisfying the following recursions: for all $t \in \mathbb{N}^*$,

$$\mathbb{P}[Y_t \in A | \mathscr{F}_{t-1}] = \mathbb{P}[Y_t \in A | X_{t-1}] = Q(X_{t-1}, A) , \quad \text{for any } A \in \mathscr{Y} , \tag{2.16}$$

$$X_t = f_{Y_t}(X_{t-1}) . \tag{2.17}$$

The name *observation-driven* models was introduced in Cox (1981), but the definition used here is slightly different from the one used in this original contribution. For any integer $t > 1$, $X_{t-1} = f_{Y_{t-1}} \circ f_{Y_{t-2}} \circ \cdots \circ f_{Y_1}(X_0)$: therefore X_{t-1} is a function of the trajectory, up to time $t-1$ and the initial condition X_0. The state X_{t-1} summarizes the information available on the conditional distribution of Y_t given $X_0, Y_1, \ldots, Y_{t-1}$. For any $C \in \mathscr{X} \otimes \mathscr{Y}$, and $t \geq 1$, we get

$$\mathbb{E}[\mathbb{1}_C(X_t, Y_t) | \mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{1}_C(f_{Y_t}(X_{t-1})) | \mathscr{F}_{t-1}] = \int Q(X_{t-1}, \mathrm{d}y) \mathbb{1}_C(f_y(X_{t-1})) ,$$

showing that $\{((X_t, Y_t), \mathscr{F}_t), t \geq 0\}$ is a Markov chain on the product space $\mathsf{X} \times \mathsf{Y}$, with transition kernel $P((x,y), C) = \int Q(x, \mathrm{d}y') \mathbb{1}_C(f_{y'}(x), y')$, for any $C \in \mathscr{X} \otimes \mathscr{Y}$. Note that this transition kernel depends only upon $x \in \mathsf{X}$. For any $A \in \mathscr{X}$, we may write similarly,

$$\mathbb{E}[\mathbb{1}_A(X_t) | \mathscr{F}_{t-1}] = \mathbb{E}[\mathbb{1}_A(f_{Y_t}(X_{t-1})) | \mathscr{F}_{t-1}] = \int Q(X_{t-1}, \mathrm{d}y) \mathbb{1}_A(f_y(X_{t-1})) ,$$

showing that $\{(X_t, \mathscr{F}_t), t \in \mathbb{N}\}$ is also a Markov chain on X, with transition kernel $H(x, A) = \int Q(x, \mathrm{d}y) \mathbb{1}_A(f_y(x))$, for any $x \in \mathsf{X}$ and $A \in \mathscr{X}$.

Example 2.22 (ARMA(1,1) model) *Consider for example an ARMA(1,1) model,*

$$Y_t - \phi_1 Y_{t-1} = Z_t + \theta_1 Z_{t-1} , \quad t \geq 1 , \tag{2.18}$$

where $\{Z_t, t \in \mathbb{N}\}$ is a sequence of i.i.d. random variables with density q with respect to the Lebesgue measure on $\mathbb{R}$, and $\{Z_t, t \in \mathbb{N}\}$ is independent of Y_0, which has distribution ξ. The process of interest, $\{Y_t, t \in \mathbb{N}\}$ is referred to as the observations. *Consider $X_t = [Y_t, Z_t]' \in \mathsf{X} = \mathbb{R}^2$. Since $Y_t = [\phi_1, \theta_1] X_{t-1} + Z_t$, the conditional distribution of Y_t given $\mathscr{F}_{t-1} = \sigma((X_s, Y_s), 0 \leq s \leq t-1)$ is given for any $A \in \mathscr{B}(\mathbb{R})$ by*

$$\mathbb{P}[Y_t \in A | \mathscr{F}_{t-1}] = \int_A q(y - [\phi_1, \theta_1] X_{t-1}) \lambda(\mathrm{d}y) ,$$

and

$$X_t = \begin{bmatrix} Y_t \\ Z_t \end{bmatrix} = \begin{bmatrix} Y_t \\ Y_t \end{bmatrix} - \begin{pmatrix} 0 & 0 \\ \phi_1 & \theta_1 \end{pmatrix} X_{t-1} .$$

Example 2.23 (GARCH(1,1) model) *The GARCH(1,1) model is defined as*

$$Y_t = \sigma_t \varepsilon_t , \quad \sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 Y_{t-1}^2 , \quad t \geq 1 , \tag{2.19}$$

where the driving noise $\{\varepsilon_t, t \in \mathbb{N}\}$ is an i.i.d. sequence, σ_0^2 is a random variable distributed according to some initial distribution ξ on $(\mathbb{R}^+, \mathscr{B}(\mathbb{R}^+))$ assumed to be independent of $\{\varepsilon_t, t \in \mathbb{N}\}$. The case $\beta_1 = 0$ corresponds to an ARCH(1) model.

The GARCH(1,1) model can be cast into the framework of observation-driven models by setting $X_t = \sigma_{t+1}^2$ and $f_y(x) = \alpha_0 + \alpha_1 y^2 + \beta_1 x$.

The square volatility $\{\sigma_t^2, t \in \mathbb{N}\}$ satisfies the following stochastic recurrence equation

$$\sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 \sigma_{t-1}^2 \varepsilon_{t-1}^2 = \alpha_0 + (\beta_1 + \alpha_1 \varepsilon_{t-1}^2) \sigma_{t-1}^2 , \quad t \geq 1 , \tag{2.20}$$

By construction, $\{\sigma_t^2, t \in \mathbb{N}\}$ is a Markov chain. By iterating the relation $\sigma_t^2 = \alpha_0 + \beta_1 \sigma_{t-1}^2 + \alpha_1 Y_{t-1}^2$ backwards in time, we get

$$\sigma_t^2 = \sum_{j=0}^{t-1} \beta_1^j (\alpha_0 + \alpha_1 Y_{t-j-1}^2) + \beta_1^t \sigma_0^2 , \tag{2.21}$$

showing that the state σ_t^2 at time t is a deterministic function of the lagged observations $Y_0,\dots,Y_{t-1}$ and of the initial state σ_0^2.

2.6 Iterated random functions

2.6.1 Strict stationarity

Let (X,d) be a Polish space and denote by $\mathscr{X}$ the associated Borel σ-field. Consider now $\{X_t,\ t\in\mathbb{Z}\}$ a stochastic process on $(\mathsf{X},\mathscr{X})$ satisfying the following recurrence equation: for all $t\in\mathbb{Z}$,

$$X_t = f_{Z_t}(X_{t-1})\,, \tag{2.22}$$

where $\{Z_t,\ t\in\mathbb{Z}\}$ is a stochastic process on $(\mathsf{Z}^{\mathbb{Z}},\mathscr{Z}^{\otimes\mathbb{Z}})$ where $(\mathsf{Z},\mathscr{Z})$ a measurable space. Moreover, throughout this section, the function $(z,x)\mapsto f_z(x)$ is assumed to be $(\mathscr{Z}\otimes\mathscr{X},\mathscr{X})$-measurable. We will need the following additional assumptions.

H 2.24 *The sequence $\{Z\}$ is strict-sense stationary and ergodic.*

H 2.25 *There exists a measurable function $z\mapsto K_z$ such that*

$$d(f_z(x),f_z(y))\le K_z d(x,y)\,, \tag{2.23}$$

for all $(x,y,z)\in\mathsf{X}\times\mathsf{X}\times\mathsf{Z}$ and

$$\mathbb{E}\left[\ln^+(K_{Z_0})\right]<\infty \quad \textit{and} \quad \mathbb{E}\left[\ln(K_{Z_0})\right]<0\,. \tag{2.24}$$

H 2.26 *There exists $x_0\in\mathsf{X}$ such that*

$$\mathbb{E}\left[\ln^+ d(x_0,f_{Z_0}(x_0))\right]<\infty\,. \tag{2.25}$$

Remark 2.27 *Assume **H** 2.25: for all $(x_0,x_0')\in\mathsf{X}\times\mathsf{X}$ and $z\in\mathsf{Z}$, we have*

$$d(x_0',f_z(x_0'))\le(1+K_z)d(x_0,x_0')+d(x_0,f_z(x_0))\,,$$

so that

$$1\wedge d(x_0',f_z(x_0'))\le 1\wedge[(1+K_z)d(x_0,x_0')]+1\wedge d(x_0,f_z(x_0))\,.$$

*Taking the logarithm function on both sides of the inequality and using that $\ln(x+y)\le\ln(x)+\ln(y)$ for all $x,y\ge 2$, we obtain that if **H** 2.26 holds for some $x_0\in\mathsf{X}$, then it also holds for all $x_0\in\mathsf{X}$.*

Theorem 2.28. *Assume **H** 2.24, **H** 2.25 and **H** 2.26. Then, for all $x_0\in\mathsf{X}$,*

$$X_{k,t}(x_0)=f_{Z_t}\circ\cdots\circ f_{Z_{t-k+1}}(x_0)\,,\quad t\in\mathbb{Z}\,, \tag{2.26}$$

converges $\mathbb{P}-$a.s. *to a random variable* $\tilde{X}_t$ *which does not depend on* x_0. *Moreover,* $\{\tilde{X}_t, t\in\mathbb{Z}\}$ *is the only strict-sense stationary solution of* (2.22). *Furthermore, for an arbitrary random variable* U,

$$\lim_{t\to\infty} d(X_t(U),\tilde{X}_t)=0\,,\quad \mathbb{P}-\text{a.s.}\ ,$$

where $X_t(U)$ *satisfies* (2.22) *for all* $t\in\mathbb{N}$ *and* $X_0(U)=U$. *In particular, the random variable* $X_t(U)$ *converges in distribution to* $\tilde{X}_0$, *i.e., as* $t\to\infty$, $X_t(U)\stackrel{\mathscr{L}_{\mathbb{R}}}{\Longrightarrow}\tilde{X}_0$.

A technical lemma is first needeed.

Lemma 2.29 *Let* $\{(A_t,B_t),\, t\in\mathbb{Z}\}$ *be a strict-sense stationary and ergodic real-valued stochastic process. Assume that*

$$\mathbb{E}[\ln^+|A_0|]<\infty,\quad -\infty\le\mathbb{E}[\ln|A_0|]<0,\quad \textit{and}\quad \mathbb{E}[\ln^+|B_0|]<\infty\,, \tag{2.27}$$

where $x^+=\max(0,\,x)$ *for* $x\in\mathbb{R}$.

$$\sum_{j=0}^{\infty}\left|\left(\prod_{i=t-j}^{t-1}A_i\right)B_{t-1-j}\right|<\infty\,,\quad \mathbb{P}-\text{a.s.} \tag{2.28}$$

PROOF. We will show that

$$\limsup_{j\to\infty}(|A_{t-1}\cdots A_{t-j}|\,|B_{t-1-j}|)^{1/j}<1\,,\quad \mathbb{P}-\text{a.s.} \tag{2.29}$$

The proof then follows by application of the Cauchy root criterion for the absolute convergence of series.

We now prove (2.29) for a fixed $t\in\mathbb{Z}$. Note that the ergodicity of the process $\{(A,B),\in\}$ implies that

$$\limsup_{j\to\infty}\frac{1}{j}\sum_{i=1}^{j}\ln|A_{t-i}|<0\,,\quad \mathbb{P}-\text{a.s.} \tag{2.30}$$

Furthermore, note that by stationarity, for any $\delta>0$,

$$\sum_{j=1}^{\infty}\mathbb{P}(j^{-1}\ln^+|B_{t-1-j}|\ge\delta)=\sum_{j=1}^{\infty}\mathbb{P}(\delta^{-1}\ln^+|B_0|\ge j)$$
$$=\mathbb{E}\left(\sum_{j=1}^{\infty}\mathbb{1}\{j\le\delta^{-1}\ln^+|B_0|\}\right)\le\delta^{-1}\mathbb{E}(\ln^+|B_0|)<\infty\,,$$

and the Borel-Cantelli Lemma therefore shows that $\limsup_{j\to\infty}j^{-1}\ln^+|B_{t-1-j}|=0$, $\mathbb{P}-$a.s. Combining this with (2.30), we get

$$\limsup_{j\to\infty}\ln\left(|A_{t-1}|\cdots|A_{t-j}||B_{t-1-j}|\right)^{1/j}<0\,,\quad \mathbb{P}-\text{a.s.}$$

which implies (2.29) by exponentiating the two terms. ■

PROOF. [of Theorem 2.28] For all $x\in\mathsf{X}$ and all $(k,t)\in\mathbb{N}^*\times\mathbb{Z}$, define

$$X_{k,t}(x):=f_{Z_t}\circ\cdots\circ f_{Z_{t-k+1}}(x)\,,$$

with the convention $X_{0,t}(x)=x$. According to (2.23), we have for all $x,x'\in\mathsf{X}$ and all $(k,t)\in\mathbb{N}\times\mathbb{Z}$,

$$d(X_{k,t}(x),X_{k,t}(x'))\le K_{Z_t}d(X_{k-1,t-1}(x),X_{k-1,t-1}(x'))\,,$$

A straightforward induction yields

$$d(X_{k,t}(x),X_{k,t}(x'))\le d(x,x')\prod_{s=t-k+1}^{t}K_{Z_s}\,. \tag{2.31}$$

Let $x_0\in\mathsf{X}$ such that $\mathbb{E}[\ln^+ d(x_0,f_{Z_0}(x_0))]<\infty$. By setting $x=x_0$ and $x'=f_{Z_{t-k}}(x_0)$ in (2.31), we obtain

$$d(X_{k,t}(x_0),X_{k+1,t}(x_0))\le d(x_0,f_{Z_{t-k}}(x_0))\prod_{s=t-k+1}^{t}K_{Z_s}\,. \tag{2.32}$$

Assumptions **H** 2.24, **H** 2.25 and **H** 2.26 allow to apply Lemma 2.29. Thus, the series $\sum_{k\geq 0} d(X_{k,t}(x_0), X_{k+1,t}(x_0))$ converges $\mathbb{P}-$a.s. and since (X, d) is complete, there exists a $\mathbb{P}-$a.s. finite random variable $\tilde{X}_t(x_0)$ such that for all $t \in \mathbb{N}$,

$$\lim_{k\to\infty} X_{k,t}(x_0) = \tilde{X}_t(x_0), \quad \mathbb{P}-\text{a.s.} \tag{2.33}$$

Moreover, using again (2.31) and applying again Lemma 2.29, we obtain that $\tilde{X}_t(x_0)$ does not depend on x_0 provided that $\mathbb{E}[\ln^+ d(x_0, f_{Z_0}(x_0))] < \infty$ and thus, we can write $\tilde{X}_t$ instead of $\tilde{X}_t(x_0)$. Since $X_{k,t}(x_0) = f_{Z_t}(X_{k-1,t-1}(x_0))$, we have

$$\begin{aligned} d(\tilde{X}_t, f_{Z_t}(\tilde{X}_{t-1})) &\leq d(\tilde{X}_t, X_{k,t}(x_0)) + d(X_{k,t}(x_0), f_{Z_t}(\tilde{X}_{t-1})) \\ &\leq d(\tilde{X}_t, X_{k,t}(x_0)) + K_{Z_t} d(X_{k-1,t-1}(x_0), \tilde{X}_{t-1}) \, . \end{aligned}$$

Using (2.33), we obtain that the right-hand side converges $\mathbb{P}-$a.s. to 0 as k goes to infinity. Thus, for all $t \in \mathbb{Z}$,

$$\tilde{X}_t = f_{Z_t}(\tilde{X}_{t-1}), \quad \mathbb{P}-\text{a.s.}$$

Moreover, let $h, t_1, \ldots, t_p \in \mathbb{N}$. Then, by stationarity of the process $\{A_t, t \in \mathbb{Z}\}$, we have that for all $k \in \mathbb{N}$, $(X_{k,t_1}(x_0), \ldots, X_{k,t_p}(x_0))$ has the same distribution as $(X_{k,t_1+h}(x_0), \ldots, X_{k,t_p+h}(x_0))$ so that by letting k go to infinity, $(\tilde{X}_{t_1}, \ldots, \tilde{X}_{t_p})$ has the same distribution as $(\tilde{X}_{t_1+h}, \ldots, \tilde{X}_{t_p+h})$, showing that the process $\{\tilde{X}_t, t \in \mathbb{N}\}$ is strict-sense stationary.

Let $\{\check{X}_t, t \in \mathbb{Z}\}$ be another strict-sense stationary solution of (2.22). Then, applying again (2.31) ,

$$d(\tilde{X}_t, \check{X}_t) \leq d(\tilde{X}_{t-k}, \check{X}_{t-k}) \prod_{\ell=t-k+1}^{t} K_{Z_\ell} \, ,$$

so that for all $M > 0$ and all $x_0 \in \mathsf{X}$,

$$\begin{aligned} \mathbb{P}(d(\tilde{X}_t, \check{X}_t) > \varepsilon) &\leq \mathbb{P}(d(\tilde{X}_{t-k}, \check{X}_{t-k}) > M) + \mathbb{P}\left(M \prod_{\ell=t-k+1}^{t} K_{Z_\ell} > \varepsilon \right) \\ &\leq \mathbb{P}(d(\tilde{X}_0, x_0) > M/2) + \mathbb{P}(d(x_0, \check{X}_0) > M/2) + \mathbb{P}\left(M \prod_{\ell=t-k+1}^{t} K_{Z_\ell} > \varepsilon \right) , \end{aligned}$$

where we have used the triangle inequality and the stationarity of the processes $\{\tilde{X}_t, t \in \mathbb{N}\}$ and $\{\check{X}_t, t \in \mathbb{Z}\}$. Now again, using Lemma 2.29, we let k go to infinity to obtain

$$\mathbb{P}(d(\tilde{X}_t, \check{X}_t) > \varepsilon) \leq \mathbb{P}(d(\tilde{X}_0, x_0) > M/2) + \mathbb{P}(d(x_0, \check{X}_0) > M/2) \, .$$

Finally, since M and ε are arbitrary, $\tilde{X}_t$ is equal to $\check{X}_t$ $\mathbb{P}-$a.s., so that $\{\tilde{X}_t, t \in \mathbb{Z}\}$ is the only strict-sense stationary solution of (2.22). Now, if $X_t(U)$ satisfies (2.22) for all $t \in \mathbb{N}$ and $X_0(U) = U$, then,

$$d(X_t(U), \tilde{X}_t) \leq d(U, \tilde{X}_0) \prod_{\ell=0}^{t} K_{Z_\ell} \, ,$$

and thus $d(X_t(U), \tilde{X}_t) \xrightarrow{\mathbb{P}\text{-a.s.}} 0$ as t goes to infinity. The proof is completed. ■

We now generalize Theorem 2.28. The proof closely follows the lines of the proof of Theorem 2.28 and is left as an exercise for the reader.

Theorem 2.30. *Let $(z_1, \ldots, z_p) \mapsto K_{z_{1:p}}$ be a $(\mathscr{Z}^{\otimes p}, \mathscr{B}(\mathbb{R}))$-measurable function from Z^p to $\mathbb{R}$ and let $z \mapsto L_z$ be a $(\mathscr{Z}, \mathscr{B}(\mathbb{R}))$-measurable function from Z to $\mathbb{R}$. Assume that for all $x, x' \in \mathsf{X}$ and all $z_{1:p} = (z_1, \ldots, z_p) \in \mathsf{Z}^p$,*

$$d(f_{z_p} \circ \cdots \circ f_{z_1}(x), f_{z_p} \circ \cdots \circ f_{z_1}(x')) \leq K_{z_{1:p}} \, d(x, x') \tag{2.34}$$

$$d(f_z(x), f_z(x')) \leq L_z \, d(x, x') \, . \tag{2.35}$$

In addition, assume that $\mathbb{E}[\ln^+ L_{Z_0}] < \infty$, $\mathbb{E}[\ln^+(K_{Z_{0:p-1}})] < \infty$, $\mathbb{E}[\ln(K_{Z_{0:p-1}})] < 0$, and $\mathbb{E}[\ln^+ d(x_0, f_{Z_0}(x_0))] < \infty$ for some $x_0 \in \mathsf{X}$. Then, the conclusions of Theorem 2.28 hold.

Example 2.31 (Random coefficient autoregression: the vector case) *Set $\mathsf{X} = \mathbb{R}^d$ and define the sequence $\{\mathbf{Y}\}$ by the following linear recursion:*

$$\mathbf{Y}_{t+1} = A_t \mathbf{Y}_t + \mathbf{B}_t \, , \quad t \in \mathbb{Z} \, , \tag{2.36}$$

where $\{(A_t, \mathbf{B}_t),\ t \in \mathbb{Z}\}$ *is a sequence of i.i.d. random elements defined on* $(\Omega, \mathscr{F},\ \mathbb{P})$ *taking values in* $\mathbb{M}_d(\mathbb{R}) \times \mathbb{R}^d$*, where* $\mathbb{M}_d(\mathbb{R})$ *is the set of* $d \times d$ *matrices with real coefficients. Let* $|||\cdot|||$ *be a matrix norm on* $\mathbb{R}^d$ *associated to a vector norm on* $\mathbb{R}^d$ *denoted by* $\|\cdot\|$*. Suppose that*

$$\mathbb{E}[\log^+ |||A_t|||] < \infty, \quad \mathbb{E}[\log^+ \|\mathbf{B}_t\|] < \infty\,. \tag{2.37}$$

The existence and uniqueness of a stationary solution of (2.36) *is obtained under the condition that* γ*, the top-Lyapunov exponent satisfies*

$$\gamma = \inf_{n>0} \frac{1}{n} \mathbb{E}[\log |||A_1 \ldots A_n|||] < 0\,. \tag{2.38}$$

This result may also be obtained from Theorem 2.30. Write (2.36) *as* $\mathbf{Y}_{t+1} = f_{Z_t}(\mathbf{Y}_t)$ *with* $Z_t = (A_t, \mathbf{B}_t)$ *and* $f_{(A,\mathbf{B})}(y) = Ay + \mathbf{B}$*. Then,*

$$f_{Z_p} \circ \cdots \circ f_{Z_1}(y) = A_p \ldots A_1 y + \sum_{j=1}^{p} \left(A_{j-1} \cdots A_1\right) \mathbf{B}_j\,,$$

showing that

$$\|f_{Z_p} \circ \cdots \circ f_{Z_1}(y) - f_{Z_p} \circ \cdots \circ f_{Z_1}(y')\| \leq |||A_p \ldots A_1|||\ \|y - y'\|\,.$$

Then, Theorem 2.30 applies on the condition that for some $p > 0$*,*

$$\mathbb{E}[\log |||A_p \ldots A_1|||] < 0\,,$$

which is exactly equivalent to the fact that the top-Lyapunov exponent γ *is strictly less that* 0*.*

2.6.2 Weak stationarity

We now consider conditions upon which (2.22) admits moments of order $p \geq 1$.

H 2.32 $\{Z\}$ *is a sequence of i.i.d. random elements.*

H 2.33 *There exists* $\alpha \in (0,1)$*,* $p \geq 1$ *such that*

$$\left\|d(f_{Z_0}(x), f_{Z_0}(y))\right\|_p \leq \alpha d(x,y)\,, \quad \textit{for all } (x,y) \in \mathsf{X} \times \mathsf{X}\,. \tag{2.39}$$

H 2.34 *There exists* $x_0 \in \mathsf{X}$ *such that*

$$\left\|d(x_0, f_{Z_0}(x_0))\right\|_p < \infty\,. \tag{2.40}$$

Remark 2.35 *Note that, under* ***H*** *2.33, if* (2.40) *holds for one* $x_0 \in \mathsf{X}$*, then it holds for all* $x_0' \in \mathsf{X}$*. Indeed, by Minkowski's inequality, we have*

$$\begin{aligned}\left\|d(x_0', f_{Z_0}(x_0'))\right\|_p &\leq d(x_0, x_0') + \left\|d(f_{Z_0}(x_0), f_{Z_0}(x_0'))\right\|_p + \left\|d(x_0, f_{Z_0}(x_0))\right\|_p \\ &\leq (1+\alpha) d(x_0, x_0') + \left\|d(x_0, f_{Z_0}(x_0))\right\|_p < \infty\,.\end{aligned}$$

Theorem 2.36. *Assume **H** 2.32, **H** 2.33 and **H** 2.34. Then, for all $x_0 \in \mathsf{X}$,*

$$X_{k,t}(x_0) = f_{Z_t} \circ \cdots \circ f_{Z_{t-k+1}}(x_0)\,, \quad t \in \mathbb{Z}\,, \tag{2.41}$$

converges to a random variable $\tilde{X}_t$, $\mathbb{P}-$a.s., which does not depend on x_0 and $\lim_{k\to\infty} \left\| d(X_{k,t}(x_0), \tilde{X}_t) \right\|_p = 0$. Moreover, $\{\tilde{X}_t\,, t \in \mathbb{Z}\}$ is the only strict-sense stationary solution of (2.22) *satisfying $\mathbb{E}\left[d^p(x_0, \tilde{X}_0)\right] < \infty$ for all $x_0 \in \mathsf{X}$. Furthermore, for an arbitrary random variable U such that $\mathbb{E}\left[d^p(x_0, U)\right] < \infty$ for all $x_0 \in \mathsf{X}$,*

$$\lim_{t\to\infty} d(X_t(U), \tilde{X}_t) = 0\,, \quad \mathbb{P}-\text{a.s.}\,,$$

where $X_t(U)$ satisfies (2.22) *for all $t \in \mathbb{N}$ and $X_0(U) = U$. In particular, the random variable $X_t(U)$ converges in distribution to $\tilde{X}_0$, i.e., as $t \to \infty$, $X_t(U) \xRightarrow{\mathscr{L}_{\mathbb{P}}} \tilde{X}_0$.*

PROOF. For all $x \in \mathsf{X}$ and all $(k,t) \in \mathbb{N} \times \mathbb{Z}$, define

$$X_{k,t}(x) := f_{Z_t} \circ \cdots \circ f_{Z_{t-k+1}}(x)\,,$$

with the convention $X_{0,t}(x) = x$. According to **H** 2.32 and **H** 2.33, we have for all $x, x' \in \mathsf{X}$ and all $(k,t) \in \mathbb{N} \times \mathbb{Z}$,

$$\left\| d(X_{k,t}(x), X_{k,t}(x')) \right\|_p \le \alpha \left\| d(X_{k-1,t-1}(x), X_{k-1,t-1}(x')) \right\|_p\,,$$

A straightforward induction yields

$$\left\| d(X_{k,t}(x), X_{k,t}(x')) \right\|_p \le \left\| d(x, x') \right\|_p \alpha^k\,. \tag{2.42}$$

By setting $x = x_0$ and $x' = f_{Z_{t-k}}(x_0)$ in (2.42), we obtain

$$\left\| d(X_{k,t}(x_0), X_{k+1,t}(x_0)) \right\|_p \le \left\| d(x_0, f_{Z_{t-k}}(x_0)) \right\|_p \alpha^k\,. \tag{2.43}$$

Using **H** 2.34, the series $\sum_{k\ge 0} \left\| d(X_{k,t}(x_0), X_{k+1,t}(x_0)) \right\|_p$ is converging and thus, there exists a random variable $\tilde{X}_t(x_0)$ in L^p such that for all $t \in \mathbb{N}$,

$$\lim_{k\to\infty} \left\| d(X_{k,t}(x_0), \tilde{X}_t(x_0)) \right\|_p = 0, \quad \mathbb{P}-\text{a.s.} \tag{2.44}$$

Moreover, using again (2.42), we obtain that $\tilde{X}_t(x_0)$ does not depend on x_0 and thus, we can write $\tilde{X}_t$ instead of $\tilde{X}_t(x_0)$. Moreover, since

$$\mathbb{E}^{1/p}\left[\left(\sum_{k=0}^{\infty} d(X_{k,t}(x_0), X_{k+1,t}(x_0)) \right)^p \right] \le \sum_{k=0}^{\infty} \left\| d(X_{k,t}(x_0), X_{k+1,t}(x_0)) \right\|_p < \infty\,,$$

the series $\sum_{k\ge 0} d(X_{k,t}(x_0), X_{k+1,t}(x_0))$ converges $\mathbb{P}-$a.s.and thus, there exists a $\mathbb{P}-$a.s. random variable $X_{\infty,t}(x_0)$ such that $\lim_{k\to\infty} d(X_{k,t}(x_0), X_{\infty,t}(x_0)) = 0$. Now, by Fatou's Lemma,

$$\mathbb{E}\left[d(X_{\infty,t}(x_0), \tilde{X}_t)\right] = \mathbb{E}\left[\liminf_{k\to\infty} d(X_{k,t}(x_0), \tilde{X}_t)^p \right] \le \liminf_{k\to\infty} \mathbb{E}\left[d(X_{k,t}(x_0), \tilde{X}_t)^p\right] = 0\,,$$

which implies that $X_{\infty,t}(x_0) = \tilde{X}_t$.

Since $X_{k,t}(x_0) = f_{Z_t}(X_{k-1,t-1}(x_0))$, we have

$$\begin{aligned} \left\| d(\tilde{X}_t, f_{Z_t}(\tilde{X}_{t-1})) \right\|_p &\le \left\| d(\tilde{X}_t, X_{k,t}(x_0)) \right\|_p + \left\| d(X_{k,t}(x_0), f_{Z_t}(\tilde{X}_{t-1})) \right\|_p \\ &\le \left\| d(\tilde{X}_t, X_{k,t}(x_0)) \right\|_p + \alpha \left\| d(X_{k-1,t-1}(x_0), \tilde{X}_{t-1}) \right\|_p\,. \end{aligned}$$

Using (2.44), we obtain that the right-hand side converges to 0 as k goes to infinity. Thus, for all $t \in \mathbb{Z}$,

$$\tilde{X}_t = f_{Z_t}(\tilde{X}_{t-1})\,, \quad \mathbb{P}-\text{a.s.}$$

Moreover, let $h, t_1, \ldots, t_p \in \mathbb{N}$. Then, by stationarity of the process $\{Z_t\,, t \in \mathbb{Z}\}$, we have that for all $k \in \mathbb{N}$, $(X_{k,t_1}(x_0), \ldots, X_{k,t_p}(x_0))$ has the same distribution as $(X_{k,t_1+h}(x_0), \ldots, X_{k,t_p+h}(x_0))$ so that by letting k go to infinity, $(\tilde{X}_{t_1}, \ldots, \tilde{X}_{t_p})$ has the same distribution as $(\tilde{X}_{t_1+h}, \ldots, \tilde{X}_{t_p+h})$ and the process $\{\tilde{X}_t\,, t \in \mathbb{N}\}$ is thus strict-sense stationary.

Let $\{\check{X}_t\,, t \in \mathbb{Z}\}$ be another strict-sense stationary solution of (2.22) such that $\mathbb{E}\left[d^p(x_0, \check{X}_0)\right] < \infty$ for all $x_0 \in \mathsf{X}$. Then, applying again (2.31) ,

$$\left\|d(\tilde{X}_t,\check{X}_t)\right\|_p \le \left\|d(\tilde{X}_{t-k},\check{X}_{t-k})\right\|_p \alpha^k$$
$$\le \alpha^k \left(\left\|d(\tilde{X}_{t-k},x_0)\right\|_p + \left\|d(x_0,\check{X}_{t-k})\right\|_p\right) = \alpha^k \left(\left\|d(\tilde{X}_0,x_0)\right\|_p + \left\|d(x_0,\check{X}_0)\right\|_p\right) ,$$

By letting k go to infinity, we obtain that $\tilde{X}_t$ is $\mathbb{P}-$a.s. equal to X_t so that $\{\tilde{X}_t, t\in\mathbb{Z}\}$ is the only strict-sense stationary solution of (2.22) such that $\mathbb{E}\left[d^p(x_0,\check{X}_0)\right] < \infty$ for all $x_0 \in \mathsf{X}$. Now, if $X_t(U)$ satisfies (2.22) for all $t\in\mathbb{N}$ and $X_0(U) = U$, then,

$$d(X_t(U),\tilde{X}_t) \le d(U,\tilde{X}_0)\prod_{\ell=0}^{t} K_{Z_\ell} .$$

so that $d(X_t(U),\tilde{X}_t) \xrightarrow{\mathbb{P}\text{-a.s.}} 0$ as t goes to infinity. The proof is completed. ■

2.6.3 Iterated random functions

Under weak conditions on the structure of the state space, every homogeneous Markov chain $\{X_t,\, t\in\mathbb{N}\}$ may be represented as a functional autoregressive process, i.e., $X_{t+1} = f_{Z_{t+1}}(X_t)$ where $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise and X_0 is independent of $\{Z_t,\, t\in\mathbb{N}\}$ and is distributed according to some initial probability ν. For simplicity, consider the case of a real valued Markov chain $\{X_t,\, t\in\mathbb{N}\}$ with initial distribution ν and Markov kernel P. Let X be a real-valued random variable and let $F(x) = \mathbb{P}(X\le x)$ be the cumulative distribution function of X. Let F^{-1} be the quantile function, defined as the generalized inverse of F by

$$F^{-1}(u) = \inf\{x\in\mathbb{R} :\, F(x)\ge u\} . \tag{2.45}$$

The right continuity of F implies that $u\le F(x) \Leftrightarrow F^{-1}(u)\le x$. Therefore, if U is uniformly distributed on $[0,1]$, $F^{-1}(U)$ has the same distribution as X, since $\mathbb{P}(F^{-1}(U)\le t) = \mathbb{P}(U\le F(t)) = F(t) = \mathbb{P}(X\le t)$.

Define $F_0(t) = \nu((-\infty,t])$ and $g = F_0^{-1}$. Consider the function F from $\mathbb{R}\times\mathbb{R}$ to $[0,1]$ defined by $F(x,x') = P(x,(-\infty,x'])$. Then, for all $x\in\mathbb{R}$, $F(x,\cdot)$ is a cumulative distribution function; the associated quantile function $f(x,\cdot)$ is

$$f(x,u) = f_u(x) = \inf\left\{x'\in\mathbb{R} :\, F(x,x')\ge u\right\} . \tag{2.46}$$

The function $(x,u)\mapsto f_u(x)$ is a Borel function since $(x,x')\mapsto F(x,x')$ is itself a Borel function; see Exercise 8.25. If U is uniformly distributed on $[0,1]$, then, for all $x\in\mathbb{R}$ and $A\in\mathscr{B}(\mathbb{R})$, we have

$$\mathbb{P}\left(f_U(x)\in A\right) = P(x,A) .$$

Let $\{U_t,\, t\in\mathbb{N}\}$ be a sequence of i.i.d. random variables, uniformly distributed on $[0,1]$. Define a sequence of random variables $\{X_t,\, t\in\mathbb{N}\}$ by $X_0 = g(U_0)$ and for $t\ge 0$,

$$X_{t+1} = f_{U_{t+1}}(X_t) .$$

Then, $\{X_t,\, t\in\mathbb{N}\}$ is a Markov chain with Markov kernel P and initial distribution ν.

Theorem 2.37. *Assume that $(\mathsf{X},\mathscr{X})$ is a measurable space and that $\mathscr{X}$ is countably generated. Let P be a Markov kernel and ν be a probability on $(\mathsf{X},\mathscr{X})$. Let $\{U_t,\, t\in\mathbb{N}\}$ be a sequence of i.i.d. random variables uniformly distributed on $[0,1]$. There exists a measurable application g from $([0,1],\mathscr{B}([0,1]))$ to $(\mathsf{X},\mathscr{X})$ and a measurable application f from $(\mathsf{X}\times[0,1],\mathscr{X}\otimes\mathscr{B}([0,1]))$ to $(\mathsf{X},\mathscr{X})$ such that the sequence $\{X_t, t\in\mathbb{N}\}$ defined by $X_0 = g(U_0)$ and $X_{t+1} = f_{U_{t+1}}(X_t))$ for $t\ge 0$, is a Markov chain with initial distribution ν and transition probability P defined on $\mathsf{X}\times\mathscr{X}$ by*

$$P(x,A) = \mathbb{P}\left(f_U(x)\in A\right) .$$

We consider the Markov chain $\{X_t,\, t\in\mathbb{N}\}$ defined by the following recurrence equation

$$X_t = f(X_{t-1},Z_t) = f_{Z_t}(X_{t-1})\ ,\ t\geq 1\ , \tag{2.47}$$

where $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise independent of the initial condition X_0. As shown in Theorem 2.37, most Markov chains can be represented in this way, but the function f will in general not display any useful property. We assume in this section that the function f has some contraction properties. We will establish, under these conditions, the existence and uniqueness of the invariant measure.

For $x_0\in\mathsf{X}$, define the *forward iteration* starting from $X_0^{x_0}=x_0$ by

$$X_t^{x_0} = f_{Z_t}(X_{t-1}^{x_0}) = f_{Z_t}\circ\cdots\circ f_{Z_1}(x_0)\ ;$$

this is just a rewrite of equation (2.47). Now, define the *backward iteration* as

$$Y_t^{x_0} = f_{Z_1}\circ\cdots\circ f_{Z_t}(x_0)\ . \tag{2.48}$$

Of course, $Y_t^{x_0}$ has the same distribution as $X_t^{x_0}$ for each $t\in\mathbb{N}$. We will show that, under **H** 2.25-**H** 2.26, the forward process $\{X_t,\, t\in\mathbb{N}\}$ has markedly different behavior from the backward process $\{Y_t^{x_0},\, t\in\mathbb{N}\}$: the forward process moves *ergodically* in X, while the backward process converges to a limit $\mathbb{P}-$a.s. The distribution of this limit, which does not depend on the initial state x_0, is the unique stationary distribution.

Lemma 2.38 *Assume that there exists a $\mathbb{P}-$a.s. finite random variable Y_∞ such that for all $x_0\in\mathsf{X}$, the backward process $\{Y_t^{x_0},\, t\in\mathbb{N}\}$ defined in* (2.48) *satisfies*

$$\lim_{t\to\infty} Y_t^{x_0} = Y_\infty\,,\quad \mathbb{P}-\text{a.s.} \tag{2.49}$$

Then, the Markov chain $\{X_t,\, t\in\mathbb{N}\}$ defined in (2.47) *with Markov kernel P implicitly defined by* (2.47)*, admits a unique invariant distribution π. In addition, π is the distribution of Y_∞ and for any $\xi\in\mathbb{M}_1(\mathscr{X})$, the sequence of probability measures $\{\xi P^t, t\in\mathbb{N}\}$ converges weakly to π.*

PROOF. Denote by π the distribution of Y_∞. Since $\{Z_t\}$ is an i.i.d. sequence, it holds that

$$Y_{t+1}^{x_0} = f_{Z_1}\circ\cdots\circ f_{Z_{t+1}}(x_0) \overset{\text{law}}{=} f_{Z_0}\circ\cdots\circ f_{Z_t}(x_0) = f_{Z_0}(Y_t^{x_0})\ .$$

Since f_{Z_0} is continuous, passing to the limit implies that $Y_\infty \overset{\text{law}}{=} f_{Z_0}(Y_\infty)$, hence π is an invariant probability measure for P.

We now prove that π is the unique invariant probability measure of P. For any $x\in\mathsf{X}$, the distribution of $X_t^x = f_{Z_t}\circ\cdots\circ f_{Z_1}(x)$ is $\delta_x P^t$. Since $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise, for each $t\in\mathbb{N}$, $X_t^x = f_{Z_t}\circ\cdots\circ f_{Z_1}(x)$ has the same distribution as $Y_t^x = f_{Z_1}\circ\cdots\circ f_{Z_t}(x)$. Thus the sequence $\{X_t^x,\, t\in\mathbb{N}\}$ converges weakly to π, i.e., for all $h\in\mathrm{C}_b(\mathsf{X})$ and $x\in\mathsf{X}$, we get

$$\lim_{t\to\infty}\delta_x P^t h = \lim_{t\to\infty}\mathbb{E}\left[h(X_t^x)\right] = \lim_{n\to\infty}\mathbb{E}\left[h(Y_t^x)\right] = \pi(h)\ .$$

By dominated convergence, this yields, for any probability measure ξ and $h\in\mathrm{C}_b(\mathsf{X})$,

$$\lim_{t\to\infty}\xi P^t h = \lim_{t\to\infty}\int_{\mathsf{X}} P^t h(x)\xi(\mathrm{d}x) = \pi(h)\ . \tag{2.50}$$

If π' is an invariant distribution for P, then $\pi' P^t=\pi'$ for all $t\in\mathbb{N}$, so that (2.50) with $\xi=\pi'$ yields $\pi'(h)=\lim_{t\to\infty}\pi' P^t(h) = \pi(h)$ for any $h\in\mathrm{C}_b(\mathsf{X})$, showing that $\pi'=\pi$. ■

Example 2.39 *Consider the model*

$$X_t = \varepsilon_t U_t X_{t-1} + (1-\varepsilon_t)[X_{t-1} + U_t(1-X_{t-1})]\ , \tag{2.51}$$

where $\{U_t,\, t\in\mathbb{N}\}$ is i.i.d. uniform on $[0,1]$, $\{\varepsilon_t,\, t\in\mathbb{N}\}$ is i.i.d. Bernoulli with probability of success 1/2, $\{U_t\}$, $\{\varepsilon_t\}$ and X_0 are independent. Set $\mathsf{T}=[0,1]\times\{0,1\}$ and $\mathscr{T}=\mathscr{B}([0,1])\otimes\mathscr{P}\{0,1\}$. Then, $X_t = f(X_{t-1},Z_t)$ with $Z_t=(U_t,\varepsilon_t)$ and $f_{u,\varepsilon}(x)(u) = xu\varepsilon + (1-\varepsilon)[x+u(1-x)]$. For any $(x,y)\in[0,1]\times[0,1]$, $|f_{u,\varepsilon}(x)-f_{u,\varepsilon}(y)|\leq K_{(u,\varepsilon)}|x-y|$ with

$$K_{u,\varepsilon} = \varepsilon u + (1-\varepsilon)(1-u)\ . \tag{2.52}$$

Figure 2.1 illustrates the difference between the backward process (right hand panel, convergence) and the forward process (left hand panel, ergodic behavior). Both processes start from $x_0 = 0.3$, and use the same random functions to move. The order in which the functions are composed is the only difference. In the right panel, the limit 0.2257 is random because it depends on the functions being iterated; we will see in Theorem 2.40 that this limiting value does not depend on the initial condition x_0.

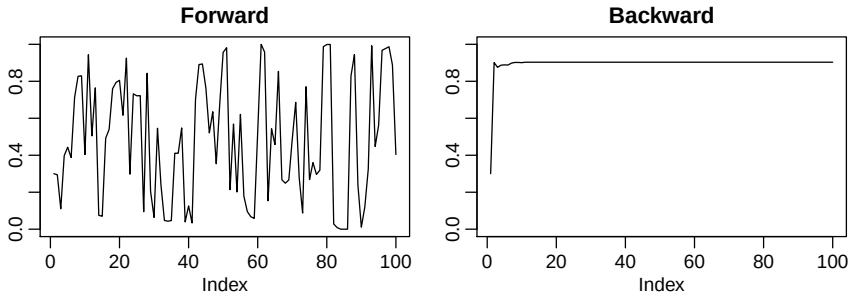

Fig. 2.1 Display for Example 2.39. The left hand panel shows ergodic behavior by the forward process; the right hand side panel shows convergence of the backward process.

We now state the main result of this section, i.e., conditions upon which the Markov chain (2.47) admits a unique stationary distribution.

Theorem 2.40. *Assume **H** 2.25 and **H** 2.26. Then, the Markov chain $\{X_t,\, t \in \mathbb{N}\}$ defined in* (2.47) *admits a unique invariant probability π, which is the distribution of the almost-sure limit of the sequence of random variables $\{Y_t^{x_0},\, t \in \mathbb{N}\}$, where $Y_t^{x_0} = f_{Z_1} \circ f_{Z_2} \circ \cdots \circ f_{Z_t}(x_0)$. In addition, for any $\xi \in \mathbb{M}_1(\mathscr{X})$, the sequence of probability measures $\{\xi P^t, t \in \mathbb{N}\}$ converges weakly to π.*

PROOF. According to Lemma 2.38, we only need to show that the backward process $\{Y_t^{x_0},\, t \in \mathbb{N}\}$ converges $\mathbb{P}-$a.s. to a random variable Y_∞, which is $\mathbb{P}-$a.s. finite and does not depend on x_0. First note that for all $x_0, x \in \mathsf{X}$,

$$\begin{aligned} d(Y_t^{x_0}, Y_t^{x}) &= d(f_{Z_1}[f_{Z_2} \circ \cdots \circ f_{Z_t}(x_0)], f_{Z_1}[f_{Z_2} \circ \cdots \circ f_{Z_t}(x)]) \\ &\leq K_{Z_1}\, d(f_{Z_2} \circ \cdots \circ f_{Z_t}(x_0), f_{Z_2} \circ \cdots \circ f_{Z_t}(x))\,. \end{aligned}$$

Thus, by induction,

$$d(Y_t^{x_0}, Y_t^{x}) \leq d(x_0, x) \prod_{i=1}^{t} K_{Z_i}\,. \tag{2.53}$$

Noting that $Y_{t+1}^{x_0} = Y_t^{x}$ with $x = f_{Z_{t+1}}(x_0)$, we obtain

$$d(Y_t^{x_0}, Y_{t+1}^{x_0}) \leq d(x_0, f_{Z_{t+1}}(x_0)) \prod_{i=1}^{t} K_{Z_i}.$$

Since $\mathbb{E}\left[\log K_{Z_0}\right] < 0$ (possibly $\mathbb{E}\left[\log K_{Z_0}\right] = -\infty$), the strong law of large numbers implies that $\limsup_{t\to\infty} t^{-1} \sum_{i=1}^{t} \log(K_{Z_t}) < 0$, $\mathbb{P}-$a.s. Thus,

$$\limsup_{t\to\infty} \left\{ \prod_{i=1}^{t} K_{Z_i} \right\}^{1/t} = \exp\left\{ \limsup_{t\to\infty} t^{-1} \sum_{i=1}^{t} \log(K_{Z_i}) \right\} < 1 \quad \mathbb{P}-\text{a.s.} \tag{2.54}$$

Next, since $\mathbb{E}\left[\log_+(d(x_0, f_{Z_0}(x_0)))\right] < \infty$, by the strong law of large numbers, it holds that $\lim_{t\to\infty} t^{-1} \sum_{k=1}^{t} \log_+ d(x_0, f_{Z_k}(x_0)) < \infty$ $\mathbb{P}-$a.s. which implies that $\lim_{t\to\infty} t^{-1} \log_+ d(x_0, f_{Z_t}(x_0)) = 0$, $\mathbb{P}-$a.s.[1] Thus $\limsup_{t\to\infty} t^{-1} \log d(x_0, f_{Z_t}(x_0)) \leq 0$ $\mathbb{P}-$a.s. which, with (2.54), yields

[1] Recall that if $\{u_k,\, k \in \mathbb{N}\}$ is such that $U_n/n = \sum_{k=1}^{n} u_k/n$ converges as n goes to infinity, then $u_n/n = U_n/n - [(n-1)/n]U_{n-1}/(n-1) \to 0$.

$$\limsup_{t\to\infty}\left\{d(Y_t^{x_0},Y_{t+1}^{x_0})\right\}^{1/t}=\limsup_{t\to\infty}\left\{d(x_0,f(x_0,Z_{t+1}))\prod_{i=1}^{t}K_{Z_i}\right\}^{1/t}<1\quad\mathbb{P}-\text{a.s.}$$

By the Cauchy root test, this implies that the series $\sum_{t=1}^{\infty} d(Y_t^{x_0},Y_{t+1}^{x_0})$ is almost surely absolutely convergent. Since (X,d) is complete, this in turn implies that $\{Y_t^{x_0}\}$ is almost surely convergent to a $\mathbb{P}-$a.s. finite random variable. Denote by $Y_\infty^{x_0}$ the almost sure limit.

Using again (2.53) and $\lim_{t\to\infty}\prod_{i=1}^{t}K_{Z_i}=0$ $\mathbb{P}-$a.s., we get that for any $x_0,x\in\mathsf{X}$, $\lim_{t\to\infty}d(Y_t^{x_0},Y_t^{x})=0$ $\mathbb{P}-$a.s. so that $Y_\infty^{x_0}=Y_\infty^{x}$. The proof then follows by applying Lemma 2.38. ■

Example 2.41 (GARCH(1,1) model) *$\{\sigma_t^2\}$ is a Markov chain and, for $t\geq 1$,*

$$\sigma_t^2=\alpha_0+\alpha_1\sigma_{t-1}^2Z_{t-1}^2+\beta_1\sigma_{t-1}^2=f(\sigma_{t-1}^2,Z_{t-1}^2)\,,$$

where $f(x,z)=\alpha_0+(\alpha_1z^2+\beta_1)x$ and $\{Z_t,\,t\in\mathbb{N}\}$ is a sequence of i.i.d. random variables. Thus the GARCH(1,1) model satisfies (2.23) *with $K_z=\alpha_1z^2+\beta_1$. Consequently, a sufficient condition for the existence and uniqueness of an invariant distribution is*

$$\mathbb{E}\left[\ln(\alpha_1Z_1^2+\beta_1)\right]<0\,. \tag{2.55}$$

Example 2.42 (Bilinear model) *Consider the bilinear process $\{X_t,\,t\in\mathbb{N}\}$ defined by*

$$X_t=aX_{t-1}+bZ_tX_{t-1}+Z_t\,, \tag{2.56}$$

where $\{Z_t,\,t\in\mathbb{N}\}$ is a sequence of integrable i.i.d. random variables and a and b are constants.

? *studies this process and concludes that if $\{Z_t\}$ is a Gaussian white noise, a sufficient condition for the existence of a stationary distribution is $\mathbb{E}\left[|a+bZ_0|\right]<1$.* **?** *relax the Gaussian assumption and simply assume that the law Z_0 is absolutely continuous with respect to Lebesgue's measure on $\mathbb{R}$, and its density is positive and lower semicontinuous. We will show that none of these assumptions on the distribution of Z_0 are necessary.*

Define $f(x,z)=(a+bz)x+z$. The bilinear model (2.56) *may be written as $X_t=f(X_{t-1},Z_t)$. For any $(x,y,z)\in\mathbb{R}$, $|f(x,z)-f(y,z)|\leq|a+bz||x-y|$, which implies that **H** 2.25 is satisfied with $K_z=|a+bz|$ as soon as $\mathbb{E}\left[\ln(|a+bZ_0|)\right]<0$. If in addition, $\mathbb{E}\left[\ln_+(|Z_0|)\right]<\infty$, then **H** 2.26 is also satisfied. Theorem 2.40 shows that the bilinear model* (2.56) *has a unique stationary distribution π, and that, starting from any initial distribution, the distribution of the iterates of the chain converge to π.*

Example 2.43 (?? cont.) *Using* (2.52)*, we get*

$$\mathbb{E}\left[\log(K_Z)\right]=(1/2)\mathbb{E}\left[\log(U_0)\right]+(1/2)\mathbb{E}\left[\log(1-U_0)\right]=\mathbb{E}\left[\log(U_0)\right]=-1\,.$$

*Therefore, **H** 2.25 is satisfied and the Markov chain $\{X_t,\,t\in\mathbb{N}\}$ defined by* (2.51) *has a unique stationary distribution.*

Instead of assuming that the contraction condition is satisfied pathwise, we may alternatively try to consider a contraction condition in the p-th norm.

Theorem 2.44. *Assume **H** 2.33 and **H** 2.34. Then, the Markov chain $\{X_t,\,t\in\mathbb{N}\}$ defined in* (2.47) *admits a unique invariant probability π, which is the distribution of the almost-sure limit of the sequence of random variables $Y_t^{x_0}=f_{Z_1}\circ f_{Z_2}\circ\cdots\circ f_{Z_t}(x_0)$. In addition,*

$$\int_{\mathsf{X}}d^p(x_0,x)\pi(\mathrm{d}x)<\infty\,.$$

For any $\xi\in\mathbb{M}_1(\mathscr{X})$, the sequence of probability measures $\{\xi P^n,n\in\mathbb{N}\}$ converges weakly to π.

PROOF. Using Lemma 2.38, we will show that the backward process $\{Y_t^{x_0},\, t\in\mathbb{N}\}$ converges $\mathbb{P}-$a.s. to a random variable Y_∞ which is $\mathbb{P}-$a.s. finite and does not depend on x_0. Since $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise, we have

$$\begin{aligned}(Y_t^{x_0}, Y_{t+1}^{x_0}) &= (f_{Z_1}[f_{Z_2}\circ\cdots\circ f_{Z_t}(x_0)], f_{Z_1}[f_{Z_2}\circ\cdots\circ f_{Z_{t+1}}(x_0)])\\ &\overset{\text{law}}{=} (f_{Z_0}[f_{Z_1}\circ\cdots\circ f_{Z_{t-1}}(x_0)], f_{Z_0}[f_{Z_1}\circ\cdots\circ f_{Z_t}(x_0)])\\ &= (f_{Z_0}(Y_{t-1}^{x_0}), f_{Z_0}(Y_t^{x_0}))\,.\end{aligned}$$

Applying (2.39), we obtain

$$\left\|d(Y_{t+1}^{x_0}, Y_t^{x_0})\right\|_p = \left\|d(f_{Z_0}(Y_t^{x_0}), f_{Z_0}(Y_{t-1}^{x_0}))\right\|_p \le \alpha\left\|d(Y_t^{x_0}, Y_{t-1}^{x_0})\right\|_p\,.$$

Thus, by induction, $\left\|d(Y_{t+1}^{x_0}, Y_t^{x_0})\right\|_p \le \alpha^t\left\|d(x_0, f_{Z_0}(x_0))\right\|_p$. By the Markov inequality, for all $\varepsilon>0$,

$$\mathbb{P}(\beta^t d(Y_{t+1}^{x_0}, Y_t^{x_0})>\varepsilon) \le \frac{\beta^{tp}}{\varepsilon^p}(\left\|d(Y_{t+1}^{x_0}, Y_t^{x_0})\right\|_p)^p \le \frac{(\alpha\beta)^{tp}}{\varepsilon^p}(\left\|d(x_0, f_{Z_0}(x_0))\right\|_p)^p\,,$$

where β is chosen in $(1, 1/\alpha)$. The Borel-Cantelli Lemma then implies that $\lim_{t\to\infty}\beta^t d(Y_{t+1}^{x_0}, Y_t^{x_0})=0$, $\mathbb{P}-$a.s. and since $\beta>1$, this in its turn yields that the sequence $\{Y_t^{x_0},\, t\in\mathbb{N}\}$ is $\mathbb{P}-$a.s. a Cauchy sequence in (X, d). Since (X, d) is complete, the sequence $\{Y_t^{x_0},\, t\in\mathbb{N}\}$ converges $\mathbb{P}-$a.s. to a random variable $Y_\infty^{x_0}$, which is $\mathbb{P}-$a.s. finite. We will now prove that $Y_\infty^{x_0}$ does not depend on x_0. For $x\in\mathsf{X}$ and $t\in\mathbb{N}$, define $Y_t^x = f_{Z_1}\circ\cdots\circ f_{Z_t}(x)$. Again, applying (2.39), we obtain

$$\left\|d(Y_t^{x_0}, Y_t^x)\right\|_p = \left\|d(f_{Z_0}(Y_{t-1}^{x_0}), f_{Z_0}(Y_{t-1}^x))\right\|_p \le \alpha\left\|d(Y_{t-1}^{x_0}, Y_{t-1}^x)\right\|_p\,,$$

which implies $\left\|d(Y_t^{x_0}, Y_t^x)\right\|_p \le \alpha^t\left\|d(x_0, x)\right\|_p$. As above, this implies that $d(Y_\infty^{x_0}, Y_\infty^x)=0$, $\mathbb{P}-$a.s. Thus, $Y_\infty^{x_0}$ does not depend on x_0 and we can thus set $Y_\infty = Y_\infty^{x_0}$. Moreover, using $Y_\infty = \lim_{t\to\infty} Y_t^{x_0}$, $\mathbb{P}-$a.s., we obtain, by Fatou's lemma,

$$\begin{aligned}\left\|d(x_0, Y_\infty)\right\|_p = \left\|\liminf_{t\to\infty} d(x_0, Y_t^{x_0})\right\|_p &\le \liminf_{t\to\infty}\left\|d(x_0, Y_t^{x_0})\right\|_p\\ &\le \liminf_{t\to\infty}\sum_{n=0}^{t-1}\left\|d(Y_n^{x_0}, Y_{n+1}^{x_0})\right\|_p \le (1-\alpha)^{-1}\left\|d(x_0, f_{Z_0}(x_0))\right\|_p < \infty\,.\end{aligned}$$

Thus,

$$\int_{\mathsf{X}} d^p(x_0, x)\pi(\mathrm{d}x) = \left(\left\|d(x_0, Y_\infty)\right\|_p\right)^p < \infty\,.$$

■

Example 2.45 (GARCH(1,1); cont.) *By Theorem 2.44, a sufficient condition for $\mathbb{E}_\pi[\sigma_1^2]<\infty$ is $\mathbb{E}\left[\alpha_1 Z_1^2+\beta_1\right]<1$. Since we have assumed that $\mathbb{E}\left[Z_1^2\right]=1$, this condition boils down to $\alpha_1+\beta_1<1$. It is easily seen that this is also a necessary condition, since the condition $\mathbb{E}_\pi[\sigma_1^2]<\infty$ implies that*

$$\mathbb{E}_\pi[\sigma_t^2] = \alpha_0 + (\alpha_1+\beta_1)\mathbb{E}_\pi[\sigma_{t-1}^2]\,,\quad \mathbb{E}_\pi[\sigma_1^2] = \mathbb{E}_\pi[\sigma_t^2] = \mathbb{E}_\pi[\sigma_{t-1}^2]\,,$$

which admits a finite solution only if $\alpha_1+\beta_1<\infty$. If, moreover,

$$\mathbb{E}_\pi[(\alpha_1 Z_1^2+\beta_1)^p] < 1\,, \tag{2.57}$$

for some $p\ge 1$, then the stationary distribution of σ_t^2 has a finite moment of order p. For $p=2$, this condition becomes $\alpha_1^2\mathbb{E}\left[Z_1^4\right]+2\alpha_1\beta_1+\beta_1^2<1$.

Example 2.46 (Bilinear model; cont.) *Assume that there exists $p\ge 1$ such that $\mathbb{E}\left[Z_0^p\right]<\infty$, then,*

$$\left\|f(x, Z_0)-f(y, Z_0)\right\|_p = \left\|(a+bZ_0)\right\|_p|x-y|\,.$$

Thus, Assumption (2.39) holds if $\mathbb{E}\left[|a+bZ_0|^p\right]<1$. In that case, (2.40) also holds and thus there exists an unique invariant probability measure π such that $\mathbb{E}\left[|X_0|^p\right]<\infty$ if the distribution of X_0 is π.

Example 2.47 (?? cont.) *Using* (2.52)*, for any $p\ge 1$, we have $\mathbb{E}\left[K_{U,\varepsilon}^p\right]\le 2\mathbb{E}\left[U^p\right]=2/(p+1)$, therefore,* (2.39) *is satisfied for any $p\ge 1$.* (2.44) *is satisfied showing that Theorem 2.44 is satisfied for any $p\ge 1$. We have already shown that the probability density $p(x)=\sqrt{x(1-x)}/\pi$ is a stationary distribution;*

Theorem 2.44 shows that this distribution is unique. In **??**, *we have sampled* 10^4 *independent trajectories of the process started from* $X_0 = 0.01$. *We have displayed the marginal stationary distribution after 3, 5, 10 and 20 iterations, together with the limiting distribution. In this case, the convergence to the stationary distribution is very quick.*

Example 2.48 (Log-linear Poisson autoregression) *Consider the following log-linear autoregressive models introduced by* **?**, *defined as follows:*

$$\mathscr{L}(Y_t|\mathscr{F}_{t-1}^Y) = \text{Poisson}(\exp(U_t)) \tag{2.58a}$$

$$U_t = a + bU_{t-1} + c\ln(1+Y_{t-1}),\ t \geq 1\,, \tag{2.58b}$$

where $\mathscr{F}_t^Y = \sigma(U_0, Y_s, s \leq t)$ *and d, a, b are real-valued parameters. Let* $\{N_k, k \geq 1\}$ *be a sequence of independent unit rate homogeneous Poisson process on the real line, independent of* U_0. *Then* $\{U_n,\ n \in \mathbb{N}\}$ *can be expressed as* $U_k = F(U_{k-1}, N_k)$, *where* F *is the function defined on* $\mathbb{R} \times \mathbb{N}^{\mathbb{R}}$ *by*

$$F(u,N) = a + bu + c\ln\{1 + N(\mathrm{e}^u)\}\,.$$

The transition kernel P *of the Markov chain* $\{U_t,\ t \in \mathbb{N}\}$ *can be expressed as*

$$Pf(v) = \mathbb{E}\left[f(a + bv + c\ln\{1 + N(\mathrm{e}^v)\})\right]\,,$$

where N *is a unit rate homogeneous Poisson process.*

Given the coefficients (a,b,c) *and the initial intensity* U_0, *the log-intensity* U_t *can be expressed explicitly from the lagged responses by expanding* (2.58b)*:*

$$U_t = a\frac{1-b^t}{1-b} + b^t U_0 + c\sum_{i=0}^{t-1} b^i \ln(1 + Y_{t-i-1})\,.$$

Hence this model belongs to the class of observation-driven models. See **?**. *In this log-linear model, the lagged observation* Y_t *is fed into the autoregressive equation for* U_t *via the term* $\ln(Y_{t-1}+1)$. *Adding one to the integer valued observation is a standard way to avoid potential problems with zero counts. In addition, both the intensity and the counts* Y_t *are transformed onto the same logarithmic scale. Covariates can be included in the right-hand side of* (2.58b)*.*

Let us check that ***H*** *2.33 holds for this model. By Minkowski's inequality, we have, for* $v \geq u \geq 0$,

$$\|F(u,N) - F(v,N)\|_1 \leq |b||u-v| + |c|\left\|\ln\left(\frac{1+N(\mathrm{e}^v)}{1+N(\mathrm{e}^u)}\right)\right\|_1\,.$$

We will prove below that for $v \geq u \geq 0$,

$$\mathbb{E}\left[\ln\left(\frac{1+N(\mathrm{e}^v)}{1+N(\mathrm{e}^u)}\right)\right] \leq v - u\,. \tag{2.59}$$

This yields, for $u, v \geq 0$,

$$\|F(v,N) - F(u,N)\|_1 \leq (|b| + |c|)|v-u|\,.$$

The contraction property ***H*** *2.33 holds if* $|b| + |c| < 1$.

We now prove (2.59). Since N *has independent increments, we can write* $(1+N(e^v))/(1+N(e^u)) = 1 + W/(1+U)$, *where* W *and* U *are independent Poisson random variable with respective means* $\mathrm{e}^v - \mathrm{e}^u$ *and* e^u. *The function* $x \mapsto \ln(1+x)$ *is concave, thus, by Jensen's inequality, we obtain*

$$\begin{aligned}\mathbb{E}\left[\ln\left(\frac{1+N(\mathrm{e}^{v})}{1+N(\mathrm{e}^{u})}\right)\right] &= \mathbb{E}\left[\ln\left(1+\frac{W}{1+U}\right)\right]\\ &\leq \ln\left(1+\mathbb{E}\left[\frac{W}{1+U}\right]\right)\\ &= \ln\left(1+(\mathrm{e}^{v}-\mathrm{e}^{u})\mathbb{E}\left[\frac{1}{1+U}\right]\right)\ .\end{aligned}$$

Note now that

$$\mathbb{E}\left[\frac{1}{1+U}\right] = \mathrm{e}^{-\mathrm{e}^{u}}\sum_{k=0}^{+\infty}\frac{1}{1+k}\frac{\mathrm{e}^{ku}}{k!} = \mathrm{e}^{-u}\mathrm{e}^{-\mathrm{e}^{u}}\sum_{k=1}^{\infty}\frac{\mathrm{e}^{ku}}{k!} \leq \mathrm{e}^{-u}\ .$$

This yields

$$\mathbb{E}\left[\ln\left(\frac{1+N(\mathrm{e}^{v})}{1+N(\mathrm{e}^{u})}\right)\right] \leq \ln\left(1+(\mathrm{e}^{v}-\mathrm{e}^{u})\mathrm{e}^{-u}\right) = v-u\ .$$

Chapter 3

Inference for Markovian Models

3.1 Likelihood inference

Assume that $(X_1,\ldots,X_n)$ is an observation from a collection of distributions $(\mathbb{P}_\theta, \theta \in \Theta)$ that depends on a parameter θ ranging over a set Θ. A popular method for finding an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1,\ \ldots,\ X_n)$ is to maximize a criterion of the type $\theta \mapsto M_n(\theta)$ over the parameter set Θ. For notational simplicity, the dependence of M_n in the observations is implicit. Such an estimator is often called an *M-estimator*. When $(X_1,\ldots,X_n)$ are i.i.d., the criterion $\theta \mapsto M_n(\theta)$ is often chosen to be the sample average of some known functions $m^\theta : \mathsf{X} \to \mathbb{R}$,

$$M_n(\theta) = n^{-1}\sum_{t=1}^{n} m^\theta(X_t)\ . \tag{3.1}$$

Sometimes, the maximizing value is computed by setting a derivative (or the set of partial derivatives in the multidimensional case) equal to zero. In such a case, the estimator $\hat{\theta}_n$ is defined as the solution of a system of equations of the type $\Psi_n(\theta) = 0$. For instance, if θ is d-dimensional, then Ψ_n typically has d coordinate functions $\Psi_n(\theta) = (\Psi_n^{(1)}(\theta),\ldots,\Psi_n^{(d)}(\theta))$. Such estimators are often referred to as Z-estimators. In the i.i.d. case, $\Psi_n(\theta)$ is often chosen to be $\Psi_n(\theta) = n^{-1}\sum_{t=1}^n \psi^\theta(X_t)$ where $\psi^\theta := (\psi_1^\theta,\ \ldots,\ \psi_d^\theta)$. In such a case, $\Psi_n(\theta) = 0$ is shorthand for the system of equations

$$n^{-1}\sum_{t=1}^{n} \psi_j^\theta(X_t) = 0,\ j = 1,2,\ \ldots,\ d. \tag{3.2}$$

In many examples ψ_j^θ is taken to be the j-th partial derivative of the function m_θ, $\psi_j^\theta = \partial m^\theta/\partial\theta_j$.

In this section, we consider *maximum likelihood estimators*. Suppose first that $X_1,\ \ldots,\ X_n$ are i.i.d. with a common density $x \mapsto \mathrm{p}^\theta(x)$. Then the maximum likelihood estimator maximizes the likelihood or, equivalently, the log-likelihood, given by

$$\theta \mapsto n^{-1}\sum_{t=1}^{n} \ln \mathrm{p}^\theta(X_t)\ .$$

Thus, a maximum likelihood estimator is an M-estimator with $m^\theta = \ln p^\theta$. If the density is partially differentiable with respect to θ for each $x \in \mathsf{X}$, then the maximum likelihood estimator also solves (3.2), with

$$\psi^\theta(x) = \nabla \ln \mathrm{p}^\theta(x) = \left(\frac{\partial \ln \mathrm{p}^\theta(x)}{\partial\theta_1},\ldots,\frac{\partial \ln \mathrm{p}^\theta(x)}{\partial\theta_d}\right)'\ .$$

This approach extends directly to the Markov chain context. Let (X, d) be a Polish space equipped with its Borel sigma-field $\mathscr{X}$ and p be a positive integer. Consider $\{Q^\theta, \theta \in \Theta\}$, a family of Markov kernels on $\mathsf{X}^p \times \mathscr{X}$ indexed by $\theta \in \Theta$ where (Θ, d) is a compact metric space. Assume that all $(\theta, x) \in \Theta \times \mathsf{X}^p$, $Q^\theta(x;\cdot)$

is dominated by some σ-finite measure μ on $(\mathsf{X}, \mathscr{X})$ and denote by $q^\theta(x;\cdot)$ its Radon-Nikodym derivative: $q^\theta(x;y) = \mathrm{d}Q^\theta(x;\cdot)/\mathrm{d}\mu(y)$. If $\{X_t,\ t \in \mathbb{N}\}$ is a Markov chain of order p associated to the Markov kernel Q^θ, then, the conditional distribution of the observations $(X_p,\ldots,X_n)$ given $X_0,\ldots,X_{p-1}$ has a density with respect to the product measure $\mu^{\otimes(n-p+1)}$ given by

$$x_{0:n} \mapsto \mathrm{p}^\theta(x_{p:n}|x_{0:p-1}) = \prod_{t=p}^{n} q^\theta(x_{t-p},\ldots,x_{t-1};x_t)\ . \tag{3.3}$$

The conditional Maximum Likelihood Estimator $\hat{\theta}_n$, based on the observations $X_{0:n}$ and on the family of likelihood functions (3.3) is defined by

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \ln \mathrm{p}^\theta(X_{p:n}|X_{0:p-1})\ . \tag{3.4}$$

Starting at θ_0, optimization algorithms generate a sequence $\{\theta_k,\ k \in \mathbb{N}\}$ that terminate when either no progress can be made or when it is apparent that a solution point has been approximated with a sufficient accuracy. Prior knowledge of the model and the data set may be used to choose θ_0 to be a reasonable estimate of the solution; otherwise the starting point must be set by the algorithm, either by a systematic approach, or in some arbitrary manner. In deciding how to move from one iterate θ_k to the other, the algorithm will use information about the likelihood, and possibly information gathered by earlier iterates $\theta_0, \theta_1, \ldots, \theta_{k-1}$.

A classical strategy is to use a line search: at each iteration, the algorithm chooses a direction δ_k, and searches along this direction from the current iterate θ_k for a new iterate with a higher function value. The distance to move along the search direction δ_k can be found by solving approximately the following one-dimensional optimization problem

$$\max_{\alpha>0} \ln \mathrm{p}^{\theta_k+\alpha\delta_k}(X_{p:n}|X_{0:p-1})\ .$$

By solving exactly this optimization problem, we would derive the maximum benefit from moving along the direction δ_k, but an exact maximization may be quite expensive and is usually worthless. Instead, the line search algorithm will typically generate a limited number of trial step lengths, until it finds one that loosely approximates the maximum.

The most obvious choice for search direction is the *steepest ascent* function $\theta \mapsto \nabla \ln \mathrm{p}^\theta(X_{p:n}|X_{0:p-1})$:

$$\delta_k = \nabla \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})\ .$$

The gradient $\nabla \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})$ is the *score* vector computed at $\theta = \theta_k$. Along all the directions we could move from θ_k, it is the one along which $\theta \mapsto \ln \mathrm{p}^\theta(X_{p:n}|X_{0:p-1})$ increases most rapidly. One of the advantage of the steepest ascent algorithm is that it requires only the calculation of the score function; however, it is known to be slow on difficult problems. Line search methods may use other search directions than the gradient. In general, any direction that makes a positive angle with $\nabla \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})$ is guaranteed to produce an increase in the likelihood, provided that the step-size is chosen appropriately. Another important search direction (and perhaps the most important one of all), is the *Newton direction*. This direction is derived from the second order Taylor series approximation of log-likelihood $\mathrm{p}^\theta(X_{p:n}|X_{0:p-1})$ in the neighborhood of the current fit θ_k. Assuming that $\nabla^2 \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})$ is positive definite, the Newton direction is

$$\delta_k^N = -\left[\nabla^2 \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})\right]^{-1} \nabla \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})\ . \tag{3.5}$$

Unlike the steepest ascent algorithm, there is a "natural" step length of 1 associated with the Newton direction. Most line search implementations of Newton's method use the unit step $\alpha = 1$ where it is possible to adjust α only when it does not produce a satisfactory increase in the value of the likelihood.

When $\nabla^2 \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})$ is not positive definite, the Newton direction may not even be defined, since $\left(\nabla^2 \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})\right)^{-1}$ may not exist. Even when it is defined, it may not satisfy the descent property $\nabla \ln \mathrm{p}^{\theta_k}(X_{p:n}|X_{0:p-1})'\delta_k^N < 0$, in which case it is unsuitable as a search direction. In these situations, line

search methods modify the definition of the search direction δ_k^N to make it satisfy the descent condition while retaining the benefit of the second-order information.

Methods that use the Newton direction have a fast rate of local convergence, typically quadratic (at least when the score and the Hessian can be estimated exactly). After a neighborhood of the solution is reached, convergence to a high accuracy solution occurs most often in a few iterations.

The main drawback of the Newton direction is the need for the Hessian (also called the *information matrix*). Explicit computation of this matrix of second derivatives can sometimes be cumbersome. Finite-difference and automatic differentiation techniques may be useful in avoiding the need to calculate second derivatives. *Quasi-Newton* search directions provide an attractive alternative to Newton's method in that they do not require computation of the Hessian and yet still attain a superlinear rate of convergence. In place of the true Hessian, Quasi-Newton strategies use an approximation that is updated after each step. This update uses the changes in the gradient to gain information about the second derivatives along the search direction.

Before going further, consider some examples:

Example 3.1 (Discrete-valued Markov chain) *In this case, the set X is at most countable; the dominating measure ν on X is chosen to be the counting measure. If $\mathsf{X} = \{1,\dots,K\}$ is finite and if the parameters $\theta = (\theta_{i,j})_{K\times K}$ are the transition probabilities, $\theta_{i,j} = q^\theta(i,j)$ for all $(i,j) \in \{1,\dots,K\}^2$, then the maximum likelihood estimator is given by*

$$\hat{\theta}_{n,i,j} = \frac{\sum_{t=1}^{n-1} \mathbb{1}_{i,j}(X_t, X_{t+1})}{\sum_{t=1}^{n} \mathbb{1}_i(X_t)} \,.$$

Example 3.2 (AR models) *Consider the AR(p) model, $X_t = \sum_{i=1}^p \phi_i X_{t-i} + \sigma Z_t$, where $\{Z_t,\ t \in \mathbb{N}\}$ is a strong white Gaussian noise. Here $\theta = (\phi_1,\dots,\phi_p,\sigma^2)$ and Θ is a compact subset of $\mathbb{R}^p \times \mathbb{R}_+$. The conditional log-likelihood of the observations may be written as*

$$\ln \mathrm{p}^\theta(X_{p:n}|X_{0:p-1}) = -\frac{n-p+1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=p}^{n}(X_t - \sum_{j=1}^{p}\phi_j X_{t-j})^2 \,.$$

The conditional likelihood function is therefore a quadratic function in the regression parameter and a convex function in the innovation variance. The maximum likelihood estimator can be computed in such a case explicitly as follows:

$$\begin{pmatrix} \hat{\phi}_{n,1} \\ \hat{\phi}_{n,2} \\ \vdots \\ \hat{\phi}_{n,p} \end{pmatrix} = \hat{\Gamma}_n^{-1} \begin{pmatrix} n^{-1}\sum_{t=p}^n X_t X_{t-1} \\ n^{-1}\sum_{t=p}^n X_t X_{t-2} \\ \vdots \\ n^{-1}\sum_{t=p}^n X_t X_{t-p} \end{pmatrix} \tag{3.6}$$

where $\hat{\Gamma}_n$ is the $(p\times p)$ empirical covariance matrix for which the i,j-th element is defined by $\hat{\Gamma}_n(i,j) = n^{-1}\sum_{t=p}^n X_{t-i}X_{t-j}$. The maximum likelihood estimator for the innovation variance is given by

$$\hat{\sigma}_n^2 = \frac{1}{n-p+1}\sum_{t=p}^{n}\left(X_t - \sum_{j=1}^{p}\hat{\phi}_{n,j}X_{t-j}\right)^2 \,. \tag{3.7}$$

Example 3.3 (Threshold autoregressive models) *Consider a two-regime threshold autoregressive TAR model*

$$X_t = \left\{\phi_{1,0} + \sum_{j=1}^{p_1}\phi_{1,j}X_{t-j} + \sigma_1 Z_t\right\}\mathbb{1}_{\{X_{t-d}\le r\}} + \left\{\phi_{2,0} + \sum_{j=1}^{p_2}\phi_{2,j}X_{t-j} + \sigma_2 Z_t\right\}\mathbb{1}_{\{X_{t-d}>r\}} \,, \tag{3.8}$$

where $\{Z_t,\ t\in\mathbb{N}\}$ is a strong Gaussian white noise with zero mean and variance 1. While the delay d may be theoretically larger than the maximum autoregressive order p, this is seldom the case in practice; hence we assume that $d \le p$. The normal error assumption implies that for every $t \ge 0$, the conditional distribution of X_t given $X_{t-1},\dots,X_{t-p}$ is normal, where $p = \max(p_1,p_2)$. To estimate the parameters, we use the likelihood conditional to the p initial values. Assume first that the threshold parameter r and the delay parameter

d are known. In such a case, the observations may be split into two parts according to whether or not $X_{t-d} \leq r$. Denote by $n_1(r,d)$ the number of observations in the lower regimes. With the observations in the lower regime, we can regress X_t on $X_{t-1},\ldots,X_{t-p_1}$, to find estimates of the autoregressive coefficients $\{\hat{\phi}_{1,j}(r,d)\}_{j=1}^{p_1}$ and the noise variance estimate (see (3.6) *and* (3.7)*)*

$$\sigma_1^2(r,d) = \frac{1}{n_1(r,d)} \sum_{t=p+1}^{n} \left(X_t - \sum_{j=1}^{p_1} \hat{\phi}_{1,j}(r,d) X_{t-j} \right)^2 \mathbb{1}\{X_{t-d} \leq r\} \,. \tag{3.9}$$

Similarly, using the $n_2(r,d)$ observations in the upper regime (note that $n-p = n_1(r,d)+n_2(r,d)$), we can find estimates of the autoregressive coefficients $\{\hat{\phi}_{2,j}(r,d)\}_{j=1}^{p_2}$ and the noise variance, $\sigma_2^2(r,d)$. To estimate (r,d), we consider the profile likelihood

$$\ell(r,d) = -\frac{n-p+1}{2}\{1+\ln(2\pi)\} - \sum_{i=1}^{2} \frac{n_i(r,d)}{2} \ln(\hat{\sigma_i}^2(r,d)) \,, \tag{3.10}$$

which is maximized over r and d. The optimization need only be performed with r over the observations $X_{p+1},\ldots,X_n$ and $d \in \{1,\ldots,p\}$. Note indeed that, for a given d, the functions are constant if r lies in between two consecutive observations. To avoid inconsistent estimators, we typically restrict the search of the threshold to be between two predetermined quantiles of $X_{p+1},\ldots,X_t$.

Example 3.4 (Functional autoregressive models) *Consider a functional autoregressive model:*

$$X_t = a^\theta(X_{t-1},\ldots,X_{t-p}) + b^\theta(X_{t-1},\ldots,X_{t-q}) Z_t \,, \tag{3.11}$$

where $\{Z_t,\, t \in \mathbb{N}\}$ is a strong Gaussian noise with zero mean and unit variance, $\theta \in \Theta \subset \mathbb{R}^d$ and for each $\theta \in \Theta$. This model includes many of the popular non linear autoregressive models such as

(a) ARCH(p) $a^\theta(x_1,\ldots,x_p) = 0$ *and*

$$b^\theta(x_1,\ldots,x_q) = \sqrt{\alpha_0 + \alpha_1 x_1^2 + \cdots + \alpha_q x_q^2} \,, \tag{3.12}$$

with $\theta = (\alpha_0,\alpha_1,\ldots,\alpha_q) \in \Theta$ a compact subset of $\mathbb{R}_+^* \times \mathbb{R}_+^q$.

(b) Autoregressive models with ARCH errors $a^\theta(x_1,\ldots,x_p) = \phi_1 x_1 + \cdots + \phi_p x_p$ *and*

$$b^\theta(x_1,\ldots,x_q) = \sqrt{\alpha_0 + \sum_{j=1}^{\ell} \alpha_j (x_j - a^\theta(x_{j+1},x_{j+2},\ldots,x_{j+p}))^2} \,, \tag{3.13}$$

where $\theta = (\phi_1,\phi_2,\ldots,\phi_p,\alpha_0,\alpha_1,\ldots,\alpha_\ell) \in \mathbb{R}^p \times \mathbb{R}_+^{\ell+1}$, $\alpha_0 > 0$ and $q = p+\ell$.

(c) Logistic Smooth Transition AR(p) $b^\theta(x_1,\ldots,x_p) = 1$ *and*

$$a^\theta(x_1,\ldots,x_p) = \mu_1 + \sum_{j=1}^{p} \phi_{1,j} x_j + \left(\mu_2 + \sum_{j=1}^{p} \phi_{2,j} x_j \right) (1+\exp(-\gamma(x_d - c)))^{-1} \tag{3.14}$$

where $d \in \{1,\ldots,j\}$ and $\theta = (\mu_i,\phi_{i,j}, i \in \{1,2\}, j = \{1,\ldots,p\},\gamma,c) \in \Theta$, a compact subset of $\mathbb{R}^{2(p+1)+2}$.

(d) Exponential Smooth Transition AR(p) $b^\theta(x_1,\ldots,x_p) = 1$ *and*

$$a^\theta(x_1,\ldots,x_p) = \mu_1 + \sum_{j=1}^{p} \phi_{2,j} x_j + \left(\mu_2 + \sum_{j=1}^{p} \phi_{2,j} x_j \right) \left(1 - \exp(-\gamma(x_d - c)^2)\right) \tag{3.15}$$

where $d \in \{1,\ldots,j\}$ and $\theta = (\mu_i,\phi_{i,j}, i \in \{1,2\}, j = \{1,\ldots,p\},\gamma,c) \in \Theta$, a compact subset of $\mathbb{R}^{2(p+1)+2}$.

The functional autoregressive models also cover situations for which the regression function is not regular, like the 2 *regimes scalar threshold autoregressive model,* $b^\theta(x_1,\dots,x_p) = \sigma_1^2 \mathbb{1}_{\{x_d \le c\}} + \sigma_2^2 \mathbb{1}_{\{x_d > c\}}$, $d \in \{1,\dots,p\}$ *and*

$$a^\theta(x_1,\dots,x_p) = \mathbb{1}_{\{x_d \le c\}} \sum_{j=1}^{p} \phi_{1,j} x_j + \mathbb{1}_{\{x_d > c\}} \sum_{j=1}^{p} \phi_{2,j} x_j \,, \tag{3.16}$$

where $\theta = (\phi_{i,j}, \sigma_i^2, i \in \{1,2\}, j \in \{1,\dots,p\}) \in \Theta \subset \mathbb{R}^{2(p+1)}$.

Assume that for any $\theta \in \Theta$ *and* $(x_1,\dots,x_p) \in \mathbb{R}^p$, $b^\theta(x_1,\dots,x_p) > 0$. *In such a case, the conditional likelihood of the observations may be written as*

$$\ln \mathrm{p}_\theta(X_{p:n}|X_{0:p-1}) = -\frac{1}{2}\sum_{t=p}^{n} \ln\left[2\pi (b^\theta(X_{t-1},\dots,X_{t-p}))^2\right] - \frac{1}{2}\sum_{t=p}^{n} \frac{\{X_t - a^\theta(X_{t-1},X_{t-2},\dots,X_{t-p})\}^2}{(b^\theta(X_{t-1},X_{t-2},\dots,X_{t-p}))^2} \,.$$

The solution of this maximization problem cannot be in general obtained in closed form; a numerical optimization technique is therefore the only option. The choice of "good" initial points is in general mandatory for the optimization algorithm to converge to a sensible optimum. There are several ways to obtain preliminary estimators for these models, but these are in general model dependent.

3.2 Consistency and asymptotic normality of the MLE

3.2.1 Consistency

We consider first the inference of the parameter θ for *misspecified models*. We postulate a model $\{\mathrm{p}^\theta(\cdot) : \theta \in \Theta\}$ for the observations $(X_0,\dots,X_n)$. However, the model is misspecified in that the true underlying distribution does not belong to the model. We use the postulated model anyway, and obtain an estimate $\hat{\theta}_n$ from maximizing the log-likelihood (3.4) where

$$\mathrm{p}^\theta(x_{p:n}|x_{0:p-1}) = \prod_{t=p}^{n} q^\theta(x_{t-p:t-1}; x_t) \,,$$

and $\{q^\theta : \theta \in \Theta\}$ is a set of Markov kernel densities (with respect to some dominating measure μ) associated to a parametric family of Markov chains of order p. We derive in this section the asymptotic behavior of $\hat{\theta}_n$. Perhaps surprisingly, $\hat{\theta}_n$ does not behave erratically despite the use of a wrong family of models. First, we show that $\hat{\theta}_n$ is asymptotically consistent, i.e., converges to a value $\theta_\star$ that maximizes $\theta \mapsto \mathbb{E}[\ln q^\theta(X_{0:p-1}; X_p)]$ (or a set of value if the solution of this problem is not unique), where the expectation is taken under the true underlying distribution. The density $q^{\theta_\star}$ can be viewed as the "projection" of the true underlying distribution on the model using the Kullback-Leibler divergence, which is defined as $\mathbb{E}[\ln q^\theta(X_{0:p-1}; X_p)]$, i.e., as a "distance" measure: $q^{\theta_\star}$ optimizes this quantity over all transition densities in the model. Consider the following assumptions:

H 3.5 *For any* $n \in \mathbb{N}$, *the vector of observations* $(X_0, X_1, \dots, X_n)$ *is a realization of a (strict-sense) stationary and ergodic process* $\{X_t,\, t \in \mathbb{Z}\}$.

We denote by $\mathbb{P}$ the probability induced on $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}})$ by $\{X_t,\, t \in \mathbb{Z}\}$ and by $\mathbb{E}$ the associated expectation. In particular, $\{X_t,\, t \in \mathbb{Z}\}$ is a stationary process but it is not necessarily a stationary Markov chain.

H 3.6

(a) $\mathbb{P}-$a.s., *the function* $\theta \mapsto q^{\theta}(X_{0:p-1};X_p)$ *is continuous.*
(b) $\mathbb{E}\left[\sup_{\theta\in\Theta} \ln^{+} q^{\theta}(X_{0:p-1};X_p)\right] < \infty$

Theorem 3.7. *Assume* ***H*** *3.5 and* ***H*** *3.6. Then, any estimator* $\hat{\theta}_n$ *belonging to the set*

$$\arg\max_{\theta\in\Theta} \ln \mathrm{p}^{\theta}(X_{p:n}|X_{0:p-1})$$

is strongly consistent in the sense that:

$$\lim_{n\to\infty} \mathrm{d}(\hat{\theta}_n, \Theta_\star) = 0,\ \mathbb{P}-\text{a.s.} \tag{3.17}$$

where

$$\Theta_\star := \arg\max_{\theta\in\Theta} \mathbb{E}[\ln q^{\theta}(X_{0:p-1};X_p)]\,. \tag{3.18}$$

PROOF. We have for $n \geq p$,

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} \bar{\mathsf{L}}_n^{\theta}(X_{p:n}|X_{0:p-1})\,,$$

where

$$\bar{\mathsf{L}}_n^{\theta}(X_{p:n}|X_{0:p-1}) := (n-p+1)^{-1}\left(\sum_{t=p}^{n} \ln q^{\theta}(X_{t-p:t-1};X_t)\right)\,.$$

The proof follows from Theorem 3.42 since

(a) $\mathbb{E}\left[\sup_{\theta\in\Theta} \ln^{+} q^{\theta}(X_{0:p-1};X_p)\right] < \infty$,
(b) $\mathbb{P}-$a.s., the function $\theta \mapsto \ln q^{\theta}(X_{0:p-1};X_p)$ is continuous.

■

Thus, in misspecified models, the MLE strongly converges to the set of parameters that maximize the *relative entropy* (or Kullback-Leibler divergence) between the true distribution and the family of postulated likelihoods. We now consider well-specified models, that is, we assume that $\{X_t,\, t\in\mathbb{N}\}$ is the observation process of a Markov chain of order p associated to the Markov kernel Q^{θ} with $\theta = \theta_\star \in \Theta$. In well-specified models, we stress the dependence in $\theta_\star$ by using the notations

$$\mathbb{P}^{\theta_\star} := \mathbb{P}, \quad \mathbb{E}^{\theta_\star} := \mathbb{E}\,. \tag{3.19}$$

In this situation, the consistency of the sequence of conditional MLE $\{\hat{\theta}_n\}_{n\geq 0}$ follows from Theorem 3.7 provided the set $\Theta_\star$ is reduced to the singleton $\{\theta_\star\}$, that is: $\theta_\star$ is the only parameter θ satisfying

$$\mathbb{E}^{\theta_\star}\left[\ln \frac{q^{\theta}(X_{0:p-1};X_p)}{q^{\theta_\star}(X_{0:p-1};X_p)}\right] = 0\,.$$

H 3.8 $Q^{\theta}(X_{0:p-1};\cdot) = Q^{\theta_\star}(X_{0:p-1};\cdot)$, $\mathbb{P}^{\theta_\star}-$a.s. *if and only if* $\theta = \theta_\star$.

The following Corollary is immediate.

Corollary 3.9 *Under **H** 3.6 and **H** 3.8, if $\{X_t,\, t\in\mathbb{N}\}$ is a strict-sense stationary and ergodic sequence under $\mathbb{P}^{\theta_\star}$, then*

$$\lim_{n\to\infty}\hat{\theta}_n=\theta_\star\,,\quad \mathbb{P}^{\theta_\star}-\text{a.s.}$$

PROOF. According to Theorem 3.7, it is sufficient to show that $\Theta_\star=\{\theta_\star\}$. Now, by the tower property, we have for all $\theta\in\Theta$,

$$\mathbb{E}^{\theta_\star}\left[\ln\frac{q^{\theta_\star}(X_{0:p-1};X_p)}{q^{\theta}(X_{0:p-1};X_p)}\right]=\mathbb{E}^{\theta_\star}\left[\mathbb{E}^{\theta_\star}\left[\ln\frac{q^{\theta_\star}(X_{0:p-1};X_p)}{q^{\theta}(X_{0:p-1};X_p)}\,\middle|\,X_{0:p-1}\right]\right]\,. \tag{3.20}$$

The RHS is always nonnegative as the expectation of a conditional Kullback-Leibler divergence, which implies that $\theta_\star\in\Theta_\star$. Moreover, using that $\ln(u)=u-1$ if and only if $u=1$, (3.20) also shows that if $\theta\in\Theta_\star$, then

$$q^{\theta_\star}(X_{0:p-1};X_p)=q^{\theta}(X_{0:p-1};X_p)\,,\quad \mathbb{P}^{\theta_\star}-\text{a.s.}$$

which concludes the proof under **H** 3.8. ■

3.2.2 Asymptotic normality

Suppose a sequence of estimators $\{\hat{\theta}_n\}_{n\geq 0}$ is consistent for a parameter $\theta_\star$ which belongs to an open subset of a Euclidean space. A question of interest relates to the order at which the error $\hat{\theta}_n-\theta_\star$ converges to zero. The answer of course depends on the specificity of the model, but for i.i.d. samples and *regular* statistical models, the order for sensible estimators based on n observations is $n^{1/2}$. Multiplying the estimation error $\hat{\theta}_n-\theta_\star$ by $n^{1/2}$ produces a proper tradeoff so that, most often, the sequence $\sqrt{n}(\hat{\theta}_n-\theta_\star)$ converges in distribution to a Gaussian random variable with zero-mean and a covariance that can be computed explicitly. This property extends to the Markov chain case without much technical trouble (but of course the expression of the covariance is more involved, except when the model is well-specified). The convergence of $\sqrt{n}(\hat{\theta}_n-\theta_\star)$ is interesting not only from a theoretical point of view but also in practice, since it makes it possible to construct asymptotic confidence regions.

Before going further, just recall briefly the outline of the standard proof in the i.i.d. case. The proof in the Markov case follows almost exactly along the same lines. Let $X_1,\,\ldots,\,X_n$ be a sample from some distribution $\mathbb{P}$, and let the random and the "true" criterion function be of the form:

$$M_n(\theta)\equiv\frac{1}{n}\sum_{i=1}^{n}m^{\theta}(X_i),\quad M(\theta)=\mathbb{E}[m^{\theta}(X_0)]\,.$$

Assume that the estimator $\hat{\theta}_n$ is a maximizer of $M_n(\theta)$ and converges in probability to an element $\theta_\star$ of $\arg\max_\theta M(\theta)$ such that $\theta_\star$ is in the interior of Θ. Assume for simplicity that θ is one-dimensional. Because $\hat{\theta}_n\to_{\mathbb{P}}\theta_\star$, it makes sense to expand $\dot{M}_n(\hat{\theta}_n)$ in a Taylor series around $\theta_\star$. Then,

$$0=\dot{M}_n(\hat{\theta}_n)=\dot{M}_n(\theta_\star)+(\hat{\theta}_n-\theta_\star)\ddot{M}_n(\tilde{\theta}_n)\,,$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and $\theta_\star$. This can be rewritten as

$$\sqrt{n}(\hat{\theta}_n-\theta_\star)=\frac{-\sqrt{n}\dot{M}_n(\theta_\star)}{\ddot{M}_n(\tilde{\theta}_n)}\,. \tag{3.21}$$

If, in this case, $\mathbb{E}\left[(\dot{m}^{\theta_\star})^2(X_0)\right]$ is finite, then by the central limit theorem, the numerator $-\sqrt{n}\dot{M}_n(\theta_\star)=-n^{-1/2}\sum_{i=1}^{n}\dot{m}^{\theta_\star}(X_i)$ is asymptotically normal with zero-mean $\mathbb{E}\left[\dot{m}^{\theta_\star}(X_0)\right]=\dot{M}(\theta_\star)=0$ and variance $\mathbb{E}\left[(\dot{m}^{\theta_\star})^2(X_0)\right]$.

Next consider the denominator of (3.21). The denominator $\ddot{M}_n(\tilde{\theta}_n)$ is an average and its limit can be found by using some "uniform" version of the law of large numbers: $\ddot{M}_n(\tilde{\theta}_n)\to_{\mathbb{P}}\mathbb{E}\left[\ddot{m}^{\theta_\star}(X_0)\right]$, provided

the expectation exists. The difficulty here stems from the fact that $\tilde{\theta}_n$ is itself a random variable: this is why it is required to use a law of large number that is valid uniformly over at least a vanishing neighborhood of $\theta_\star$. Together with Slutsky's lemma, these two conclusions yield

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}\left(0, \frac{\mathbb{E}\left[(\dot{m}^{\theta_\star})^2(X_0)\right]}{(\mathbb{E}\left[\ddot{m}^{\theta_\star}(X_0)\right])^2}\right) . \tag{3.22}$$

The preceding derivation can be made rigorous by imposing appropriate technical conditions (on the regularity of the function $\theta \mapsto \dot{M}_n(\theta)$, the validity of interchanging derivation and expectation), often called *regularity conditions* in the i.i.d. case. Perhaps surprisingly, the most challenging part of the proof consists of showing that $\ddot{M}_n(\tilde{\theta}_n)$ is converging to $\mathbb{E}\left[\ddot{m}^{\theta_\star}(X_0)\right]$ (see above); all the other steps are direct consequences of classical limit theorems.

The derivation can be extended to higher-dimensional parameters. For a d-dimensional parameter, we use d estimating equations $\dot{M}_n(\theta) = 0$, where $\dot{M}_n : \mathbb{R}^d \mapsto \mathbb{R}^d$. The derivatives $\ddot{M}_n(\theta_\star)$ are $(d \times d)$-matrices that converge to the $(d \times d)$ matrix $\Gamma_\star := \mathbb{E}\left[\ddot{m}^{\theta_\star}(X_0)\right]$ with entries $\mathbb{E}\left[\partial^2 m^{\theta_\star}(X_0)/\partial\theta_i\partial\theta_j\right]$. The limiting distribution (see (3.22)) may be expressed as

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}_d(0, \Gamma_\star^{-1}\Sigma_\star\Gamma_\star^{-1}) , \tag{3.23}$$

where $\Sigma_\star = \mathbb{E}\left[\dot{m}^{\theta_\star}(X_0)(\dot{m}^{\theta_\star}(X_0))'\right]$ is the covariance matrix of the "pseudo-scores." Note that the matrix $\Gamma_\star$ should be nonsingular.

In this section we extend these results to Markov chains. We consider first the case of (possibly) misspecified models. Assume that $\Theta \subset \mathbb{R}^d$ and recall that Θ^o is the interior of Θ.

H 3.10 *There exists $\theta_\star \in \Theta^o$ such that $\hat{\theta}_n \stackrel{\mathbb{P}\text{-}prob}{\longrightarrow} \theta_\star$.*

H 3.11
*(a) The function $\theta \mapsto q^\theta(X_{0:p-1}; X_p)$ is $\mathbb{P}-$*a.s.* twice continuously differentiable in an open neighborhood of $\theta_\star$.*
(b) There exists $\rho \in (0,1)$ such that for all indexes $i, j \in \{1, \ldots, d\}$,

$$\mathbb{E}\left[\sup_{\theta \in \mathrm{B}(\theta_\star, \rho)} \left| \frac{\partial^2 \ln q^\theta}{\partial\theta_i\partial\theta_j}(X_{0:p-1}; X_p)\right|\right] < \infty .$$

H 3.12 *There exists a $d \times d$ nonsingular matrix $\Sigma_\star$ such that*

$$n^{-1/2}\sum_{t=p}^{n} (\nabla \ln q^{\theta_\star}(X_{t-p:t-1}; X_t) - \mathbb{E}[\nabla \ln q^{\theta_\star}(X_{0:p-1}; X_p)]) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}_d(0, \Sigma_\star) . \tag{3.24}$$

Theorem 3.13. *Assume **H** 3.10, **H** 3.11 and **H** 3.12. In addition, assume that $\{X_t,\, t \in \mathbb{N}\}$ is a strict-sense stationary and ergodic sequence under $\mathbb{P}$ and that $\Gamma_\star := \mathbb{E}[\nabla^2 \ln q^{\theta_\star}(X_{0:p-1}; X_p)]$ is nonsingular. Then,*

$$n^{1/2}(\hat{\theta}_n - \theta_\star) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0, \Gamma_\star^{-1}\Sigma_\star\Gamma_\star^{-1})$$

where $\Sigma_\star$ is defined in (3.24).

We preface the proof by the following technical Lemma, which will be useful in misspecified and well-specified models as well.

Lemma 3.14 *Assume that **H** 3.11 hold for some $\theta_\star \in \Theta$. Assume in addition that $\{X_t,\, t \in \mathbb{N}\}$ is a strict-sense stationary and ergodic sequence under $\mathbb{P}$ and let $\{\theta_n,\, n \in \mathbb{N}\}$ be a sequence of random vectors such that $\theta_n \to_{\mathbb{P}} \theta_\star$. Then, for all $i, j \in \{1, \ldots, d\}$,*

$$n^{-1} \sum_{t=p}^{n} \frac{\partial^2 \ln q^{\theta_n}(X_{t-p:t-1}; X_t)}{\partial \theta_i \partial \theta_j} \xrightarrow{\mathbb{P}\text{-}prob} \mathbb{E}\left[\frac{\partial^2 \ln q^{\theta_\star}(X_{0:p-1}; X_p)}{\partial \theta_i \partial \theta_j} \right] .$$

PROOF. Denote $A_t(\theta) = \frac{\partial^2 \ln q^{\theta}(X_{t-p:t-1}; X_t)}{\partial \theta_i \partial \theta_j}$. Since by the Birkhoff ergodic theorem, $n^{-1} \sum_{t=p}^{n} A_t(\theta_\star) \to_{\mathbb{P}} \mathbb{E}[A(\theta_\star)]$, we only need to show that

$$n^{-1} \sum_{t=p}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \xrightarrow{\mathbb{P}\text{-prob}} 0 . \tag{3.25}$$

Let $\varepsilon > 0$ and choose $0 < \eta < \rho$ such that

$$\mathbb{E}\left(\sup_{\theta \in \mathrm{B}(\theta_\star, \eta)} |A_p(\theta_\star) - A_p(\theta)| \right) < \varepsilon . \tag{3.26}$$

The existence of such η follows from the $\mathbb{P}-$a.s. continuity of $\theta \mapsto A_t(\theta)$ under **H** 3.11-((a)) and by the Lebesgue convergence theorem under **H** 3.11-((b)). We then have

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-1} \sum_{t=p}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \geq \varepsilon ,\ \theta_n \in \mathrm{B}(\theta_\star, \eta) \right) \leq \limsup_{n\to\infty} \mathbb{P}\left(n^{-1} \sum_{t=p}^{n} \sup_{\theta \in \mathrm{B}(\theta_\star, \eta)} |A_t(\theta_\star) - A_t(\theta)| \geq \varepsilon \right) = 0 ,$$

where the last equality follows from (3.26) and the Birkhoff ergodic theorem. Moreover, since $\hat{\theta}_n \xrightarrow{\mathbb{P}\text{-prob}} \theta_\star$, $\lim_{n\to\infty} \mathbb{P}(\theta_n \notin \mathrm{B}(\theta_\star, \eta)) = 0$ so that finally,

$$\lim_{n\to\infty} \mathbb{P}\left(n^{-1} \sum_{t=p}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \geq \varepsilon \right) = 0 .$$

Thus, (3.25) holds and the proof follows. ■

PROOF. [of Theorem 3.13] Since $\hat{\theta}_n$ cancels the derivatives of the log-likelihood, a Taylor expansion at the point $\theta = \theta_\star$ with an integral form of the remainder yields

$$n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\hat{\theta}_n}(X_{t-p:t-1}; X_t) = 0 = n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\theta_\star}(X_{t-p:t-1}; X_t) + n^{-1} \sum_{t=p}^{n} \left(\int_0^1 \nabla^2 \ln q^{\theta_{n,s}}(X_{t-p:t-1}; X_t) \mathrm{d}s \right) \sqrt{n}(\hat{\theta}_n - \theta_\star) , \tag{3.27}$$

where $\theta_{n,s} = s\hat{\theta}_n + (1-s)\theta_\star$. According to **H** 3.12, we have

$$n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\theta_\star}(X_{t-p:t-1}; X_t) \xRightarrow{\mathscr{L}_{\mathbb{P}^{\theta_\star}}} \mathrm{N}(0, \Sigma_\star) .$$

We now turn to the asymptotic properties of the second term of the right-hand side in (3.27). More precisely, since by **H** 3.10, $\hat{\theta}_n \to_{\mathbb{P}^{\theta_\star}} \theta_\star$, Lemma 3.14 shows that for all $i, j \in \{1, \ldots, d\}$,

$$n^{-1} \sum_{t=p}^{n} \int_0^1 \frac{\partial^2 \ln q^{\theta_{n,s}}}{\partial \theta_i \partial \theta_j}(X_{t-p:t-1}; X_t) \mathrm{d}s \xrightarrow{\mathbb{P}^{\theta_\star}\text{-prob}} \mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}}{\partial \theta_i \partial \theta_j}(X_{0:p-1}; X_p) \right] .$$

This implies that

$$n^{-1}\sum_{t=p}^{n}\left(\int_0^1 \nabla^2 \ln q^{\theta_{n,s}}(X_{t-p:t-1};X_t)\mathrm{d}s\right) \overset{\mathbb{P}^{\theta_\star}\text{-prob}}{\longrightarrow} \Gamma_\star := \mathbb{E}[\nabla^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)]\,.$$

The proof then follows by applying the Slutsky Lemma to (3.27). ■

We now turn to well-specified models, that is, the observation process $\{X_n,\ n\in\mathbb{N}\}$ is the realization of a Markov chain associated to an unknown parameter $\theta_\star$.

Before stating the result, we need some additional assumptions. The parameter set Θ is now a compact subset of $\mathbb{R}^d$ with a non-empty interior and we assume that $\theta_\star \in \Theta^o$, where Θ^o is the interior of Θ. Moreover,

H 3.15
(a) The function $\theta \mapsto q^\theta(X_{0:p-1};X_p)$ is, $\mathbb{P}^{\theta_\star}-$a.s., twice continuously differentiable in an open neighborhood of $\theta_\star$,
(b) there exists $\rho>0$ such that for all $i,j\in\{1,\dots,d\}$,

$$\mathbb{E}^{\theta_\star}\left[\sup_{\theta\in \mathrm{B}(\theta_\star,\rho)}\left|\frac{\partial^2 \ln q^\theta}{\partial\theta_i\partial\theta_j}(X_{0:p-1};X_p)\right|\right]<\infty\,.$$

H 3.15 is similar to **H** 3.11. In well-specified models, it is possible to obtain the weak convergence of the score function using a martingale argument (see Exercise 3.43). Some additional assumptions are needed.

H 3.16 *There exists $\rho>0$ such that*

$$\mathbb{E}^{\theta_\star}\left[\left|\frac{\partial \ln q^{\theta_\star}}{\partial\theta_i}(X_{0:p-1};X_p)\right|^2\right]<\infty\,, \tag{3.28}$$

$$\int \sup_{\theta\in \mathrm{B}(\theta_\star,\rho)}\left|\frac{\partial q^\theta}{\partial\theta_i}(X_{0:p-1};x_p)\right|\mathrm{d}\mu(x_p)<\infty,\quad \mathbb{P}^{\theta_\star}-\text{a.s.}\,, \tag{3.29}$$

$$\int \sup_{\theta\in \mathrm{B}(\theta_\star,\rho)}\left|\frac{\partial^2 q^\theta}{\partial\theta_i\partial\theta_j}(X_{0:p-1};x_p)\right|\mathrm{d}\mu(x_p)<\infty,\quad \mathbb{P}^{\theta_\star}-\text{a.s.}\,, \tag{3.30}$$

and the Fisher information matrix $\mathscr{J}(\theta_\star)$ defined by

$$\begin{aligned}\mathscr{J}(\theta_\star) &:= -\mathbb{E}^{\theta_\star}\left[\nabla^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)\right]\\ &= \mathbb{E}^{\theta_\star}\left[\nabla \ln q^{\theta_\star}(X_{0:p-1};X_p)\nabla \ln q^{\theta_\star}(X_{0:p-1};X_p)^T\right]\end{aligned} \tag{3.31}$$

is nonsingular.

Note that the Fisher information matrix has two expressions given in (3.31). These two expressions are classically obtained under the assumption (3.30). We now state and prove the weak convergence of the score function using a martingale-type approach.

Lemma 3.17 *Under **H** 3.16,*

$$n^{-1/2}\sum_{t=p}^{n}\nabla \ln q^{\theta_\star}(X_{t-p:t-1};X_t) \overset{\mathscr{L}_{\mathbb{P}^{\theta_\star}}}{\Longrightarrow} \mathrm{N}(0,\mathscr{J}(\theta_\star))\,,$$

where $\mathscr{J}(\theta_\star)$ is defined in (3.31).

PROOF. Let $\mathscr{F}$ be the filtration $\mathscr{F} = (\mathscr{F}_n)_{n\in\mathbb{N}}$ where $\mathscr{F}_n = \sigma(X_0,\dots,X_n)$ and let

$$M_n := \sum_{t=p}^{n} \nabla \ln q^{\theta_\star}(X_{t-p:t-1};X_t)\,.$$

Note that according to (3.28), $\mathbb{E}(|M_t|^2) < \infty$ and that

$$\mathbb{E}^{\theta_\star}\left[\nabla \ln q^{\theta_\star}(X_{t-p:t-1};X_t)\,\middle|\,\mathscr{F}_{t-1}\right] = \int \nabla q^{\theta_\star}(X_{t-p:t-1};x_t)\mathrm{d}x_t$$
$$= \nabla \int q^{\theta}(X_{t-p:t-1};x_t)\mathrm{d}x_t\Big|_{\theta=\theta_\star} = 0\,,$$

where the assumption (3.29) allows us to interchange $\int$ and ∇. Finally, $\{M_t\,, t \geq p\}$ is a square integrable $\mathscr{F}$-martingale with stationary increments. The proof follows by applying Theorem A.7. ■

We now have all the ingredients for obtaining the central limit theorem of the MLE in well-specified models.

Theorem 3.18. *Assume that* $\hat{\theta}_n \overset{\mathbb{P}^{\theta_\star}\text{-prob}}{\longrightarrow} \theta_\star$ *and that* ***H*** *3.15 and* ***H*** *3.16 hold. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0, \mathscr{J}^{-1}(\theta_\star))\,,$$

where $\mathscr{J}(\theta_\star)$ *is defined in* (3.31).

PROOF. See Exercise 3.48 ■

Note that $\mathscr{J}(\theta_\star)$ is the asymptotic Fisher information matrix in the sense that

$$\mathscr{J}(\theta_\star) = -\lim_{n\to\infty} n^{-1}\mathbb{E}^{\theta_\star}\left[\nabla^2 \ln\left(\prod_{k=p}^{n} q^{\theta}(X_{k-p:k-1};X_k)\right)\Big|_{\theta=\theta_\star}\right]\,.$$

Example 3.19 (Autoregressive models) *Let* $(\phi_{*,1},\phi_{*,2},\dots,\phi_{*,p}) \in \mathbb{R}^p$ *be coefficients such that* $\phi_*(z) = 1 - \sum_{j=1}^{p}\phi_{*,j}z^j \neq 0$ *for* $|z| \leq 1$*, where* $p > 0$ *is an integer. Assume that* $\{X_t,\, t\in\mathbb{Z}\}$ *is the causal autoregressive process:* $X_t = \sum_{j=1}^{p}\phi_{*,j}X_{t-j} + \sigma_* Z_t$ *where* $\{Z_t,\, t\in\mathbb{Z}\}$ *is white Gaussian noise with unit-variance. Set* $\theta = (\phi_1,\dots,\phi_p,\sigma^2)$*,* $\theta_\star = (\phi_{*,1},\dots,\phi_{*,p},\sigma_*^2)$ *and*

$$\ln q^{\theta}(x_{0:p-1},x_p) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(x_p - \sum_{j=1}^{p}\phi_j x_{p-j}\right)^2\,.$$

By differentiating twice with respect to σ^2 *and* ϕ_i*, we get*

$$\frac{\partial^2 \ln q^{\theta}(X_{0:p-1};X_p)}{\partial\phi_i\partial\phi_j} = -\frac{1}{\sigma^2}X_{p-i}X_{p-j}\,, \qquad (i,j)\in\{1,\dots,p\}^2$$
$$\frac{\partial^2 \ln q^{\theta}(X_{0:p-1};X_p)}{\partial\phi_j\partial\sigma^2} = -\frac{1}{\sigma^4}\left(X_p - \sum_{j=1}^{p}\phi_j X_{p-j}\right)X_{p-j}\,, \qquad j\in\{1,\dots,p\}$$
$$\frac{\partial^2 \ln q^{\theta}(X_{0:p-1};X_p)}{\partial\sigma^2\partial\sigma^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}\left(X_p - \sum_{j=1}^{p}\phi_j X_{p-j}\right)^2\,.$$

Since, for any $i\in\{1,\dots,p\}$*,*

$$\mathbb{E}^{\theta_\star}\left[\left(X_p - \sum_{j=1}^{p} \phi_{*,j} X_{p-j}\right) X_{p-i}\right] = \mathbb{E}^{\theta_\star}[Z_p X_{p-i}] = 0\,,$$

and, similarly

$$\mathbb{E}^{\theta_\star}\left[\left(X_p - \sum_{j=1}^{p} \phi_{*,j} X_{p-j}\right)^2\right] = \mathbb{E}^{\theta_\star}[Z_p^2] = \sigma_*^2\,,$$

we get that $\mathscr{J}(\theta_\star)$ is a block diagonal matrix, with block diagonal elements equal to $1/2\sigma_*^4$ and $\sigma_*^{-2}\Gamma_\star$, where $\Gamma_\star$ is the $p \times p$ covariance matrix, $[\Gamma_\star]_{i,j} = \mathrm{Cov}^{\theta_\star}(X_0, X_{i-j})$, $1 \le i, j \le p$. If $p = 1$, then $\Gamma_\star = \sigma_*^2/(1-\phi_{*,1}^2)$, and the asymptotic variance for the autoregressive parameter is equal to $1/(1-\phi_{*,1}^2)$. For $p = 2$, the reader can verify (Exercise 3.46) that

$$\sigma_*^{-2}\Gamma_\star = \begin{bmatrix} 1-\phi_{*,2}^2 & -\phi_{*,1}(1+\phi_{*,2}) \\ -\phi_{*,1}(1+\phi_{*,2}) & 1-\phi_{*,2}^2 \end{bmatrix}. \tag{3.32}$$

Note that the asymptotic variance of the estimate of $\phi_{*,1}$ depends only on $\phi_{*,2}$.

Example 3.20 (Threshold autoregressive models) *We use the notations and definitions of Example 3.3. Denote by $\theta = [\eta_1, \eta_2, r, d]$ with $\eta_i = [(\phi_{i,j})_{j=0}^{p_1}, \sigma_i^2]$, $i = 1, 2$, the unknown parameters. In this case the log-likelihood may be expressed as*

$$\begin{aligned}\ln q^\theta(x_{0:p-1}; x_p) &= \left\{ -\frac{1}{2}\ln(2\pi\sigma_1^2) - \frac{1}{2\sigma_1^2}\left(X_p - \phi_{1,0} - \sum_{j=1}^{p_1}\phi_{1,j}X_{p-j}\right)^2 \right\} \mathbb{1}_{\{X_{t-d} \le r\}} \\ &\quad + \left\{ -\frac{1}{2}\ln(2\pi\sigma_2^2) - \frac{1}{2\sigma_2^2}\left(X_p - \phi_{2,0} - \sum_{j=1}^{p_2}\phi_{2,j}X_{p-j}\right)^2 \right\} \mathbb{1}_{\{X_{t-d} > r\}}\,,\end{aligned}$$

where $p = \max(p_1, p_2)$. When the delay d and the threshold r are known, Theorem 3.7 shows that, provided that $\{X_t,\ t \in \mathbb{Z}\}$ is an ergodic sequence, then the regression parameters $\hat{\eta}_i(r,d) = [\hat{\phi}_{i,j}(r,d), \hat{\sigma}_i(r,d)]_{j=1}^{p_i}$, $i = 1, 2$ are consistent in the sense of (3.17)*. It has been established by* **?***, Theorem 1 that the same remains true if the delay and the threshold are estimated as the maximum of the profile likelihood.*

Assume now that $\{X_t,\ t \in \mathbb{Z}\}$ is a strict sense stationary geometrically ergodic SETAR process

$$\begin{aligned}X_t &= \left\{ \phi_{*,1,0} + \sum_{j=1}^{p_1}\phi_{*,1,j}X_{t-j} + \sigma_{*,1}Z_t \right\} \mathbb{1}_{\{X_{t-d_*} \le r_*\}} \\ &\quad + \left\{ \phi_{*,2,0} + \sum_{j=1}^{p_2}\phi_{*,2,j}X_{t-j} + \sigma_{*,2}Z_t \right\} \mathbb{1}_{\{X_{t-d_*} > r_*\}}\,,\end{aligned} \tag{3.33}$$

where $\{Z_t,\ t \in \mathbb{Z}\}$ is a strong white Gaussian noise with zero-mean and unit-variance. Such condition is satisfied, for example, if $\max_{i=1,2}\sum_{j=1}^{p_i}|\phi_{*,i,j}| < 1$ (see **??***, condition (**??**)). Assume first that the delay d and the threshold r are both known: $r = r_*$ and $d = d_*$. We may then apply Theorem 3.18 to show that the distribution of the parameters in the two regimes are asymptotically normal,*

$$\sqrt{n}\,(\hat{\eta}_1(r_*, d_*) - \eta_{*,1}, \hat{\eta}_2(r_*, d_*) - \eta_{*,2}) \xRightarrow{\mathscr{L}_{\mathbb{P}}} \mathrm{N}(0, \mathscr{J}^{-1}(\theta_\star))$$

where $\theta_\star = [\eta_{*,1}, \eta_{*,2}, r_*, d_*]$, $\eta_{*,i} = [(\phi_{*,i,j})_{j=0}^{p_1}, \sigma_{*,i}^2]$, $i = 1, 2$. Note that, for any $(i, j) \in \{1, 2\}$, $i \ne j$, and any $(k, \ell) \in \{1, \ldots, p_i\}^2$, we get

$$\frac{\partial^2 \ln q^\theta(x_{0:p-1}; x_p)}{\partial\phi_{i,k}\partial\phi_{j,\ell}} = 0\,, \quad \frac{\partial^2 \ln q^\theta(x_{0:p-1}; x_p)}{\partial\sigma_i^2\partial\phi_{j,k}} = 0\,,$$

showing that the matrix $\mathscr{J}(\theta_\star)$ is block diagonal and therefore that the estimators in the two regimes are asymptotically independent. The entries of the Fisher information matrix are given by

$$\mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)}{\partial\phi_{1,k}\partial\phi_{1,\ell}}\right] = \frac{1}{\sigma_1^2}\mathbb{E}^{\theta_\star}\left[X_{p-k}X_{p-\ell}\mathbb{1}_{\{X_{p-d}\le r\}}\right], \qquad (k,\ell)\in\{1,\dots,p_1\}$$

$$\mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)}{\partial\phi_{1,0}\partial\phi_{1,k}}\right] = \frac{1}{\sigma_1^2}\mathbb{E}^{\theta_\star}\left[X_{p-k}\mathbb{1}_{\{X_{p-d}\le r\}}\right], \qquad k\in\{1,\dots,p_1\}$$

$$\mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)}{\partial\phi_{1,k}\partial\sigma_1^2}\right] = 0\,,$$

$$\mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}(X_{0:p-1};X_p)}{\partial\sigma_1^2\partial\sigma_1^2}\right] = \frac{1}{2\sigma_1^2}\mathbb{P}^{\theta_\star}(X_0\le r)\,.$$

The Fisher information matrix for the parameters of the upper regime can be written similarly. When the threshold r and d are both unknown, the assumptions of Theorem 3.18 are no longer satisfied. It is shown in **?***, Theorem 2 that*

$$\left\{n(\hat{r}_n - r_*), \sqrt{n}\,(\hat{\eta}_1 - \eta_{*,1}, \hat{\eta}_2 - \eta_{*,2})\right\}$$

converges in distribution, that $n(\hat{r}_n - r_*)$ and $\sqrt{n}\,(\hat{\eta}_1 - \eta_{*,1}, \hat{\eta}_2 - \eta_{*,2})$ are asymptotically independent. **?** *identified the distribution of $n(\hat{r}_n - r_*)$: this is an interval on which a compound Poisson process (whose parameters depend on $\theta_\star$) attains its global minimum.*

In general, the problem of determining the set of parameters Θ_* for $\{X_t,\ t\in\mathbb{Z}\}$ to be geometrically ergodic is still an open problem. For $p_1 = p_2 = 1$ and $d = 1$, however, the problem is solved. The condition in this case is $\phi_{*,1,1} < 1$, $\phi_{*,2,1} < 1$ and $\phi_{*,1,1}\phi_{*,2,1} < 1$. For details, see **?**.

3.3 Observation-driven models

Let (X, d) and (Y, δ) be Polish spaces equipped with their Borel sigma-field $\mathscr{X}$ and $\mathscr{Y}$. Consider $\{Q^\theta, \theta\in\Theta\}$ a family of Markov kernels on $\mathsf{X}\times\mathscr{Y}$ indexed by $\theta\in\Theta$ where (Θ, d) is a compact metric space. Assume that all $(\theta, x)\in\Theta\times\mathsf{X}$, $Q^\theta(x,\cdot)$ is dominated by some σ-finite measure μ on $(\mathsf{Y},\mathscr{Y})$ and denote by $q^\theta(x,\cdot)$ its Radon-Nikodym derivative: $q^\theta(x,y) = \mathrm{d}Q^\theta(x,\cdot)/\mathrm{d}\mu(y)$. Finally, let $\left\{f^\theta\,:\,\theta\in\Theta\right\}$ be a family of measurable functions from $(\mathsf{X}\times\mathsf{Y}, \mathscr{X}\otimes\mathscr{Y})$ to $(\mathsf{X},\mathscr{X})$. We consider the following observation-driven model (see Theorem 2.21)

$$\mathbb{P}^\theta\left[Y_t\in A\,|\,\mathscr{F}_{t-1}\right] = Q^\theta(X_{t-1}, A) = \int_A q^\theta(X_{t-1},y)\mu(\mathrm{d}y)\,, \quad \text{for any } A\in\mathscr{Y}\,,$$
$$X_t = f^\theta_{Y_t}(X_{t-1})\,.$$

With these notations, the distribution of $(Y_1,\dots,Y_n)$ conditionally on $X_0 = x$ has a density with respect to the product measure $\mu^{\otimes n}$ given by

$$y_{1:n}\mapsto\prod_{t=1}^n q^\theta(f^\theta_{y_{1:t-1}}(x), y_t)\,, \tag{3.34}$$

where we have set for all $s\le t$ and all $y_{s:t}\in\mathsf{Y}^{t-s+1}$,

$$f^\theta_{y_{s:t}} = f^\theta_{y_t}\circ f^\theta_{y_{t-1}}\circ\cdots\circ f^\theta_{y_s}\,, \tag{3.35}$$

with the convention $f^\theta_{y_{1:0}}(x_0) = x_0$. Note that, for any $t\ge 0$, X_t is a deterministic function of $Y_{1:t}$ and X_0, i.e.,

$$X_t = f^\theta_{Y_{1:t}}(X_0) = f^\theta_{Y_t}\circ f^\theta_{Y_{t-1}}\circ\cdots\circ f^\theta_{Y_1}(X_0)\,. \tag{3.36}$$

In this section, we study the asymptotic properties of $\hat{\theta}_{n,x}$, the conditional Maximum Likelihood Estimator (MLE) of the parameter θ based on the observations $(Y_1,\ldots,Y_n)$ and associated to the parametric family of likelihood functions given in (3.34), that is, we consider

$$\hat{\theta}_{n,x} \in \arg\max_{\theta\in\Theta} \mathsf{L}^{\theta}_{n,x}(Y_{1:n}) \,, \tag{3.37}$$

where

$$\mathsf{L}^{\theta}_{n,x}(y_{1:n}) := n^{-1} \ln \left(\prod_{t=1}^{n} q^{\theta}(f^{\theta}_{y_{1:t-1}}(x), y_t) \right) . \tag{3.38}$$

We are especially interested here in inference for *misspecified models*, that is, we *do not assume* that the distribution of the observations belongs to the set of distributions where the maximization occurs. In particular, $\{Y_t,\, t\in\mathbb{Z}\}$ are not necessarily the observation process associated to the recursion (3.36).

Nevertheless, consider the following assumptions

H 3.21 *$\{Y\}$ is a strict-sense stationary and ergodic stochastic process.*

Under **H** 3.21, denote by $\mathbb{P}$ the distribution of $\{Y\}$ on $(\mathsf{Y}^{\mathbb{Z}}, \mathscr{Y}^{\mathbb{Z}})$. Write $\mathbb{E}$, the associated expectation.

H 3.22 *For all $(x,y)\in\mathsf{X}\times\mathsf{Y}$, the functions $\theta\mapsto f^{\theta}_{y}(x)$ and $(y,\theta)\mapsto q^{\theta}(x,y)$ are continuous.*

H 3.23 *There exists a family of $\mathbb{P}-$a.s. finite random variables*

$$\left\{ f^{\theta}_{Y_{-\infty:t}} \,:\, (\theta,t)\in\Theta\times\mathbb{Z} \right\}$$

such that for all $x\in\mathsf{X}$,

(a) $\lim_{m\to\infty} \sup_{\theta\in\Theta} d(f^{\theta}_{Y_{-m:0}}(x), f^{\theta}_{Y_{-\infty:0}}) = 0\,,\ \mathbb{P}-$a.s.,

(b) $\lim_{t\to\infty} \sup_{\theta\in\Theta} |\ln q^{\theta}(f^{\theta}_{Y_{1:t-1}}(x), Y_t) - \ln q^{\theta}(f^{\theta}_{Y_{-\infty:t-1}}, Y_t)| = 0\,,\ \mathbb{P}-$a.s.,

(c) $\mathbb{E}\left[\sup_{\theta\in\Theta} \left(\ln q^{\theta}(f^{\theta}_{Y_{-\infty:t-1}}, Y_t) \right)_+ \right] < \infty\,.$

In the following, we set for all $(\theta,t)\in\Theta\times\mathbb{N}$,

$$\bar{\ell}^{\theta}(Y_{-\infty:t}) := \ln q^{\theta}(f^{\theta}_{Y_{-\infty:t-1}}, Y_t)\,. \tag{3.39}$$

Remark 3.24 *When checking **H** 3.23, we usually introduce $f^{\theta}_{Y_{-\infty:0}}$ by showing that for all $(\theta,x)\in\Theta\times\mathsf{X}$, $f^{\theta}_{Y_{-m:0}}(x)$ converges, $\mathbb{P}-$a.s., as m goes to infinity to a limit that does not depend on x. We can therefore denote by $f^{\theta}_{Y_{-\infty:0}}$ this limit. With this definition, we then check **H** 3.23-((a))-((b))-((c)).*

Note that under **H** 3.22, $\hat{\theta}_{n,x}$ is well-defined. The following theorem establishes the consistency of the sequence of estimators $\{\hat{\theta}_{n,x}\,,\, n\in\mathbb{N}\}$.

Theorem 3.25. *Assume **H** 3.21, **H** 3.22 and **H** 3.23. Then, for all $x\in\mathsf{X}$,*

$$\lim_{n\to\infty} \mathsf{d}(\hat{\theta}_{n,x}, \Theta_\star) = 0\,, \quad \mathbb{P}-\text{a.s.}$$

where $\Theta_\star := \arg\max_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^\theta(Y_{-\infty:1})]$ *and* $\bar{\ell}^\theta(Y_{-\infty:1})$ *is defined in* (3.39).

PROOF. The proof directly follows from Theorem 3.42 provided that

$$\mathbb{E}[\sup_{\theta\in\Theta}(\bar{\ell}^\theta(Y_{-\infty:1}))_+] < \infty, \tag{3.40}$$

$$\text{the function } \theta \mapsto \bar{\ell}^\theta(Y_{-\infty:1}) \text{ is upper-semicontinuous, } \mathbb{P}-\text{a.s.}, \tag{3.41}$$

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |\mathsf{L}^\theta_{n,x}(Y_{1:n}) - \bar{\mathsf{L}}^\theta_n(Y_{-\infty:n})| = 0, \mathbb{P}-\text{a.s.}, \tag{3.42}$$

where $\bar{\mathsf{L}}^\theta_n(Y_{-\infty:n}) = n^{-1}\sum_{t=1}^n \bar{\ell}^\theta(Y_{-\infty:t})$.

But, (3.40) follows from **H** 3.23-((c)), (3.41) follows by combining **H** 3.23-((a)) and **H** 3.22 since a uniform limit of continuous functions is continuous and (3.42) is direct from **H** 3.23-((b)) and the definitions of $\mathsf{L}^\theta_{n,x}(Y_{1:n})$ and $\bar{\mathsf{L}}^\theta_n(Y_{-\infty:n})$. The proof is completed. ■

Example 3.26 (GARCH(1,1)) *Assume that the observations* $\{Y_t,\, t\in\mathbb{Z}\}$ *are a strict-sense stationary ergodic process. We fit to the observations a GARCH*$(1,1)$ *model with Student* t*-innovations,*

$$\begin{cases} Y_t = \sigma_t \left(\frac{\nu-2}{\nu}\right)^{1/2} \varepsilon_t \\ \sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \quad t\in\mathbb{Z} \end{cases} \tag{3.43}$$

where $\{\varepsilon_t,\, t\in\mathbb{N}\}$ *is an i.i.d. sequence of* t_ν*-distributed random variables (where* ν *is the number of degrees of freedom) and* $\theta = (\nu, \alpha_0, \alpha_1, \beta_1) \in \Theta$ *is a compact subset of*

$$\{(\nu, \alpha_0, \alpha_1, \beta_1), \nu > 2, \alpha_0 > 0, \alpha_1 > 0, \beta_1 > 0, \alpha_1 + \beta_1 < 1\}\,.$$

The restriction on the degrees of freedom parameter ν *ensures the conditional variance to be finite. The variance of* ε_1 *is equal to* $\nu/(\nu-2)$*, therefore,* $((\nu-2)/\nu)^{1/2}\varepsilon_1$ *has unit-variance. Recall that the density of a Student* t*-distribution with* ν *degrees of freedom is given by*

$$t_\nu(z) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})} \nu^{-1/2}[1+z^2/\nu]^{-(\nu+1)/2}\,.$$

Setting $X_t = \sigma_{t+1}^2$, (3.43) *defines an observation-driven model with* $q^\theta(x,y) = t_\nu(y/\sqrt{x})$ *and* $f_y^\theta(x) = \alpha_0 + \alpha_1 y^2 + \beta_1 x$*. Given initial values* Y_0 *and* σ_0^2 *to be specified below, the conditional log-likelihood may be expressed as*

$$\mathsf{L}^\theta_{n,\sigma_0}(Y_{1:n}) = \sum_{t=1}^n \log t_\nu(Y_t/\sigma_t)\,,$$

where σ_t^2 *are computed recursively using* (3.43)*. For a given value of* θ*, the unconditional variance (corresponding to the stationary value of the variance) is a sensible choice for the unknown initial values* σ_0^2

$$\sigma_0^2 = \frac{\alpha_0}{1-\alpha_1-\beta_1}\,. \tag{3.44}$$

In this case, for any integer m*, we have (see* (2.21)*)*

$$f^\theta_{Y_{-m:0}}(x) = \beta_1^{m+1}x + \sum_{j=0}^m \beta_1^j(\alpha_0 + \alpha_1 Y_{-j}^2)\,.$$

Assume that $\mathbb{E}_*[Y_0^2] < \infty$. *Since* Θ *is compact, there exist* $b_1 < 1$, $a_0 > 0$ *and* $a_1 < 1$ *such that, for all* $\theta \in \Theta$, $0 \le \beta_1 \le b_1$, $\alpha_0 \le a_0$ *and* $\alpha_1 \le a_1$. *The series* $\sum b_1^j(a_0 + a_1 Y_{-j}^2)$ *converges* $\mathbb{P}_* - $a.s. *Define*

$$f^\theta_{Y_{-\infty:0}} = \sum_{j=0}^{\infty} \beta_1^j (\alpha_0 + \alpha_1 Y_{-j}^2) \, .$$

With these definitions, we get

$$\sup_{\theta\in\Theta} |f^\theta_{Y_{-m:0}}(x) - f^\theta_{Y_{-\infty:0}}| \le b_1^{m+1} x + \sum_{j=m+1}^{\infty} b_1^j (a_0 + a_1 Y_{-j}^2) \overset{\mathbb{P}_*\text{-a.s.}}{\longrightarrow} 0 \, ,$$

showing ***H*** *3.23-*((a)). *Note similarly that*

$$\left|\log q^\theta(f^\theta_{Y_{1:t-1}}(x), Y_t) - \log q^\theta(f^\theta_{Y_{-\infty:t-1}}, Y_t)\right| = \frac{\nu+1}{2} \left| \ln\left(1 + \frac{Y_t^2}{\nu(f^\theta_{Y_{1:t-1}}(x))^2}\right) - \ln\left(1 + \frac{Y_t^2}{\nu(f^\theta_{Y_{-\infty:t-1}})^2}\right)\right| \, .$$

Since for any $a > 0$, $b > 0$ *and* $z > 0$, $|\log(1+z^2/a^2) - \log(1+z^2/b^2)| \le 2|a-b|/(a \wedge b)$ *and* $f^\theta_{Y_{1:t-1}}(x) > \alpha_0$, $f^\theta_{Y_{-\infty:t-1}} > \alpha_0$, *there exists a constant* $K > 0$ *such that*

$$\sup_{\theta\in\Theta} \left|\log q^\theta(f^\theta_{Y_{1:t-1}}(x), Y_t) - \log q^\theta(f^\theta_{Y_{-\infty:t-1}}, Y_t)\right| \le K \left(b_1^{m+1} x + \sum_{j=m+1}^{\infty} b_1^j (a_0 + a_1 Y_{-j}^2) \right) \overset{\mathbb{P}_*\text{-a.s.}}{\longrightarrow} 0 \, ,$$

showing ***H*** *3.23-*((b)). *The proof of* ***H*** *3.23-*((c)) *is along the same lines. Theorem 3.25 shows the consistency of the estimator.*

Of course, in well-specified models where the observations process is associated to a parameter $\theta_\star \in \Theta$, Theorem 3.25 allows us to obtain the convergence of the MLE to $\theta_\star$ on the condition that $\Theta_\star$ is reduced to the singleton $\{\theta_\star\}$.

An important subclass of misspecified models corresponds to the case where the kernel density $q^\theta = q$ does not depend on θ and where the observation process is assumed to follow the recursions:

$$\begin{aligned} &\mathbb{P}[Y_t \in A | \mathscr{F}_{t-1}] = Q^\star(X_{t-1}, A) = \int_A q^\star(X_{t-1}, y)\mu(\mathrm{d}y) \, , \quad \text{for any } A \in \mathscr{Y} \, , \\ &X_t = f^{\theta_\star}_{Y_t}(X_{t-1}) \, , \quad t \in \mathbb{Z} \, . \end{aligned}$$

The MLE is based on the likelihood functions $y_{1:n} \mapsto \mathsf{L}^\theta_{n,x}(y_{1:n})$ defined (3.38) whose expression includes a kernel density $q \neq q^\star$ and a family of functions $\{f^\theta : \theta \in \Theta\}$. Since $\theta_\star$ is assumed to be in Θ^o, the interior of Θ but $q^\star \neq q$, this situation falls into the misspecified models framework. In such cases, Maximum Likelihood Estimators $\{\hat\theta_{n,x}\}$ defined in (3.37) are called Quasi Maximum Likelihood estimators (QMLE) and $\theta_\star$ is not the true value of the parameter in the sense that the distribution of the observation process is not characterized by $\theta_\star$ only. Nevertheless, perhaps surprisingly, it can be shown that, under some additional assumptions, the QMLE $\{\hat\theta_{n,x}\}$ are consistent and asymptotically normal with respect to the parameter $\theta_\star$. For simplicity, we assume here that $\mathsf{X} = \mathbb{R}$.

The main assumption that links $q^\star$ and q is the following:

H 3.27 *For all* $x^\star \in \mathbb{R}$,

$$\arg\max_{x\in\mathbb{R}} \int Q^\star(x^\star, \mathrm{d}y) \ln q(x, y) = \{x^\star\} \, . \tag{3.45}$$

Theorem 3.28. *Assume **H** 3.21, **H** 3.22, **H** 3.23 and **H** 3.27. Moreover, assume that $f^{\theta}_{Y_{-\infty:0}} = f^{\theta_\star}_{Y_{-\infty:0}}$, $\mathbb{P}-$a.s. implies that $\theta = \theta_\star$. Then, for all $x \in \mathsf{X}$,*

$$\lim_{n\to\infty} \mathrm{d}(\hat{\theta}_{n,x}, \theta_\star) = 0\,, \quad \mathbb{P}-\text{a.s.}$$

PROOF. See Exercise 3.52. ■

Example 3.29 (Normal innovations) *Assume that the observations $\{Y_t,\, t \in \mathbb{Z}\}$ is a strict-sense stationary ergodic process associated to*

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \middle| \mathscr{F}_{t-1}\right] = Q^\star(\sigma^2_{t-1}, A)\,, \quad \textit{for any } A \in \mathscr{Y}\,, \\ &\sigma^2_t = f^{\theta_\star}_{Y_t}(\sigma^2_{t-1})\,, \quad t \in \mathbb{Z}\,. \end{aligned} \tag{3.46}$$

We fit to the observations an observation-driven model with normal innovations,

$$\begin{aligned} &Y_t = \sigma_t \varepsilon_t \\ &\sigma^2_t = f^{\theta}_{Y_t}(\sigma^2_{t-1})\,, \quad (t,\theta) \in \mathbb{N} \times \Theta\,, \end{aligned} \tag{3.47}$$

where $\{\varepsilon_t,\, t \in \mathbb{N}\} \sim_{\text{iid}} \mathrm{N}(0,1)$. This model typically covers the GARCH(1,1) model with normal innovations. Now, the function

$$x \mapsto \int Q^\star(x^\star, \mathrm{d}y) \ln q(x,y) = \int Q^\star(x^\star, \mathrm{d}y) \left(-\frac{y^2}{2x} - \frac{1}{2}\ln(2\pi x)\right) = \left(-\frac{\int Q^\star(x^\star, \mathrm{d}y) y^2}{2x} - \frac{1}{2}\ln(2\pi x)\right)\,,$$

*is maximized at $x = \int Q^\star(x^\star, \mathrm{d}y) y^2$ by straightforward algebra. Thus, **H** 3.27 implies*

$$\int Q^\star(x^\star, \mathrm{d}y) y^2 = x^\star\,.$$

Plugging this equality into (3.46)*,we obtain that the observations $\{Y_t,\, t \in \mathbb{Z}\}$ is a strict-sense stationary ergodic process associated to*

$$\begin{aligned} &Y_t \sim \sigma_t \varepsilon^\star_t \\ &\sigma^2_t = f^{\theta_\star}_{Y_t}(\sigma^2_{t-1})\,, \quad t \in \mathbb{Z}\,, \end{aligned}$$

where $\{\varepsilon^\star_t,\, t \in \mathbb{Z}\}$ is an i.i.d. sequence of random variables with potentially any unknown distribution provided that $\mathbb{E}[(\varepsilon^\star_t)^2] = 1$.

Example 3.30 (Exponential families) *Assume that the observations $\{Y_t,\, t \in \mathbb{Z}\}$ are a strict-sense stationary ergodic process associated to*

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \middle| \mathscr{F}_{t-1}\right] = Q^\star(X_{t-1}, A) = \int_A q^\star(X_{t-1}, y)\mu(\mathrm{d}y)\,, \quad \textit{for any } A \in \mathscr{Y}\,, \\ &X_t = f^{\theta_\star}_{Y_t}(X_{t-1})\,, \quad t \in \mathbb{Z}\,. \end{aligned}$$

We fit to the observations the following observation-driven model

$$\begin{aligned} &\mathbb{P}\left[Y_t \in A \middle| \mathscr{F}_{t-1}\right] = Q(X_{t-1}, A)\,, \quad \textit{for any } A \in \mathscr{Y}\,, \\ &X_t = f^{\theta}_{Y_t}(X_{t-1})\,, \quad (t,\theta) \in \mathbb{Z} \times \Theta\,. \end{aligned}$$

where $Q(x,\cdot)$ is assumed to belong to the class of exponential family distributions. More precisely, we assume that for all $(x,y) \in \mathsf{X}\times\mathsf{Y}$, $q(x,y) = \exp(xy - A(x))h(y)$ for some twice differentiable function $A : \mathsf{X} \to \mathbb{R}$ and some measurable function $h : \mathsf{Y} \to \mathbb{R}^+$. Using that

$$\int Q(x,\mathrm{d}y)\frac{\partial^2 \ln q(x,y)}{\partial x^2} \leq 0\,,$$

it can be readily checked that $A'' \geq 0$ so that A is concave. Thus, the function

$$x \mapsto \int Q^\star(x^\star,\mathrm{d}y)\ln q(x,y) = \int Q^\star(x^\star,\mathrm{d}y)\,(xy - A(x) + \ln h(y)) \\ = x\int Q^\star(x^\star,\mathrm{d}y)y - A(x) + \int Q^\star(x^\star,\mathrm{d}y)\ln h(y)\,,$$

*is convex. The maximum of this function can thus be obtained by cancelling the derivatives with respect to x, which yields $\int Q^\star(x^\star,\mathrm{d}y)y - A'(x) = 0$. Then, **H** 3.27 implies that*

$$\int Q^\star(x^\star,\mathrm{d}y)y = A'(x^\star)\,. \tag{3.48}$$

For example in the log-linear Poisson autoregression model (see Example 2.48),

$$q(x,y) = \exp(xy - \mathrm{e}^x)/y!$$

so that $A(x) = \exp(x)$. Thus, (3.48) *yields that $\int Q^\star(x^\star,\mathrm{d}y)y = \exp(x^\star)$.*

Remark 3.31 *In well-specified models, $Q^\star = Q$. Since the Kullback-Leibler divergence is nonnegative, we obtain*

$$\int Q^\star(x^\star,\mathrm{d}y)\ln q^\star(x,y) \leq \int Q^\star(x^\star,\mathrm{d}y)\ln q^\star(x^\star,y)\,,$$

and, provided that $x \mapsto Q(x,\cdot)$ is one-to-one, the equality holds if and only if $x = x^\star$. Thus, (3.45) *in **H** 3.27 is satisfied.*

If for all $y \in \mathsf{Y}$, $x \mapsto q(x,y)$ is twice differentiable, we may define for all vector $\boldsymbol{u}$ in $\mathbb{R}^d$, all matrix Γ of size $d \times d$ with real entries and all $(x,y) \in \mathbb{R}\times\mathsf{Y}$,

$$\varphi(\boldsymbol{u},x,y) := \boldsymbol{u}\frac{\partial \ln q}{\partial x}(x,y)\,, \tag{3.49}$$

$$\psi(\Gamma,\boldsymbol{u},x,y) := \Gamma\frac{\partial \ln q}{\partial x}(x,y) + \boldsymbol{u}\boldsymbol{u}'\frac{\partial^2 \ln q}{\partial x^2}(x,y)\,. \tag{3.50}$$

These functions appear naturally when differentiating $\theta \mapsto \ln q(f(\theta),y)$ where $\theta \mapsto f(\theta)$ is a twice differentiable function. More precisely, we have in such a case by straightforward algebra,

$$\nabla \ln q(f(\theta),y) = \varphi(\nabla f(\theta), f(\theta), y)\,,$$
$$\nabla^2 \ln q(f(\theta),y) = \psi(\nabla^2 f(\theta), \nabla f(\theta), f(\theta), y)\,.$$

H 3.32 *For all $y \in \mathsf{Y}$, the function $x \mapsto q(x,y)$ is twice continuously differentiable. Moreover, there exist $\rho > 0$ and a family of $\mathbb{P}-$a.s. finite random variables*

$$\left\{f^\theta_{Y_{-\infty:t}} :\ (\theta,t) \in \Theta\times\mathbb{Z}\right\}$$

such that $\theta \mapsto f^\theta_{Y_{-\infty:0}}$ is, $\mathbb{P}-$a.s., twice continuously differentiable on some ball $\mathrm{B}(\theta_\star,\rho)$ and for all $x \in \mathsf{X}$,

(a) $\mathbb{P}-$a.s. ,

$$\lim_{t\to\infty} \|\varphi(\nabla f^{\theta_\star}_{Y_{1:t-1}}(x), f^{\theta_\star}_{Y_{1:t-1}}(x), Y_t) - \varphi(\nabla f^{\theta_\star}_{Y_{-\infty:t-1}}, f^{\theta_\star}_{Y_{-\infty:t-1}}, Y_t)\| = 0\,,$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^d$,

(b) $\mathbb{P}-$a.s. ,

$$\lim_{t\to\infty} \sup_{\theta\in \mathrm{B}(\theta_\star,\rho)} \|\psi(\nabla^2 f^{\theta}_{Y_{1:t-1}}(x), \nabla f^{\theta}_{Y_{1:t-1}}(x), f^{\theta}_{Y_{1:t-1}}(x), Y_t) - \psi(\nabla^2 f^{\theta}_{Y_{-\infty:t-1}}, \nabla f^{\theta}_{Y_{-\infty:t-1}}, f^{\theta}_{Y_{-\infty:t-1}}, Y_t)\| = 0\,,$$

where by abuse of notation, we use again $\|\cdot\|$ to denote any norm on the set of $d\times d$-matrices with real entries,

(c)

$$\mathbb{E}\left[\|\varphi(\nabla f^{\theta_\star}_{Y_{-\infty:0}}, f^{\theta_\star}_{Y_{-\infty:0}}, Y_1)\|^2\right] < \infty\,, \tag{3.51}$$

$$\mathbb{E}\big[\sup_{\theta\in \mathrm{B}(\theta_\star,\rho)} \|\psi(\nabla^2 f^{\theta}_{Y_{-\infty:0}}, \nabla f^{\theta}_{Y_{-\infty:0}}, f^{\theta}_{Y_{-\infty:0}}, Y_1)\|\big] < \infty\,. \tag{3.52}$$

Moreover, the matrix $\mathscr{J}(\theta_\star)$ defined by

$$\mathscr{J}(\theta_\star) := \mathbb{E}\left[(\nabla f^{\theta_\star}_{Y_{-\infty:0}})(\nabla f^{\theta_\star}_{Y_{-\infty:0}})'\left(\frac{\partial^2 \ln q}{\partial x^2}(f^{\theta_\star}_{Y_{-\infty:0}}, y)\right)\right]\,, \tag{3.53}$$

is nonsingular.

Theorem 3.33. *Assume **H** 3.21, **H** 3.27 and **H** 3.32. Assume in addition that $\hat\theta_{n,x} \to_{\mathbb{P}} \theta_\star$. Then,*

$$\sqrt{n}(\hat\theta_{n,x} - \theta_\star) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0, \mathscr{J}(\theta_\star)^{-1}\mathscr{I}(\theta_\star)\mathscr{J}(\theta_\star)^{-1})\,,$$

where $\mathscr{J}(\theta_\star)$ is given by (3.53) *and $\mathscr{I}(\theta_\star)$ is defined by*

$$\mathscr{I}(\theta_\star) := \mathbb{E}\left[(\nabla f^{\theta_\star}_{Y_{-\infty:0}})(\nabla f^{\theta_\star}_{Y_{-\infty:0}})'\left(\frac{\partial \ln q}{\partial x}(\nabla f^{\theta_\star}_{Y_{-\infty:0}}, y)\right)^2\right]\,.$$

We preface the proof by two Lemmas, whose proofs are given as exercises.

Lemma 3.34 *Assume **H** 3.27 and **H** 3.32. Then,*

$$n^{-1/2}\sum_{t=1}^{n} \nabla \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}}, Y_t) \stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0, \mathscr{I}(\theta_\star))\,.$$

PROOF. See Exercise 3.53. ■

Lemma 3.35 *Assume **H** 3.27 and **H** 3.32. Let $\{\theta_n,\ n\in\mathbb{N}\}$ be a sequence of random vectors such that $\theta_n \to_{\mathbb{P}} \theta_\star$. Then, for all $i,j\in\{1,\dots,d\}$,*

$$n^{-1}\sum_{t=1}^{n} \frac{\partial^2 \ln q(f^{\theta_n}_{Y_{-\infty:t-1}}, Y_t)}{\partial\theta_i\partial\theta_j} \stackrel{\mathbb{P}\text{-}prob}{\longrightarrow} \mathbb{E}\left[\frac{\partial^2 \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}}, Y_t)}{\partial\theta_i\partial\theta_j}\right]\,.$$

PROOF. See Exercise 3.54. ■

PROOF. [of Theorem 3.33] Since the MLE $\hat{\theta}_{n,x}$ cancels the derivatives of the log-likelihood, a Taylor expansion at $\theta = \theta_\star$ with an integral form of the remainder yields

$$n^{-1/2}\sum_{t=1}^{n}\nabla \ln q(f^{\hat{\theta}_{n,x}}_{Y_{1:t-1}}(x),Y_t) = 0 = n^{-1/2}\sum_{t=1}^{n}\nabla \ln q(f^{\theta_\star}_{Y_{1:t-1}}(x),Y_t) + n^{-1}\sum_{t=1}^{n}\left(\int_0^1 \nabla^2 \ln q(f^{\theta_{n,x,s}}_{Y_{1:t-1}}(x),Y_t)\mathrm{d}s\right)\sqrt{n}(\hat{\theta}_n-\theta_\star)\,, \quad (3.54)$$

where $\theta_{n,x,s} = s\hat{\theta}_{n,x} + (1-s)\theta_\star$. The proof of Theorem 3.33 then follows from (3.54) and the Slutsky Lemma, provided we show that for all $\theta_n \to_{\mathbb{P}} \theta_\star$,

$$n^{-1/2}\sum_{t=1}^{n}\nabla \ln q(f^{\theta_\star}_{Y_{1:t-1}}(x),Y_t) \overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0,\mathscr{J}(\theta_\star))\,, \quad (3.55)$$

$$n^{-1}\sum_{t=1}^{n}\frac{\partial^2 \ln q(f^{\theta_n}_{Y_{1:t-1}}(x),Y_t)}{\partial\theta_i\partial\theta_j} \overset{\mathbb{P}\text{-prob}}{\longrightarrow} \mathbb{E}\left[\frac{\partial^2 \ln q(f^{\theta_\star}_{Y_{-\infty:0}},Y_1)}{\partial\theta_i\partial\theta_j}\right]\,, \quad (3.56)$$

$$\mathscr{J}(\theta_\star) = \mathbb{E}\left[\psi(\nabla^2 f^{\theta}_{Y_{-\infty:0}},\nabla f^{\theta}_{Y_{-\infty:0}},f^{\theta}_{Y_{-\infty:0}},Y_1)\right]\,. \quad (3.57)$$

By noting that

$$\nabla \ln q(f^{\theta_\star}_{Y_{1:t-1}}(x),Y_t) = \varphi(\nabla f^{\theta_\star}_{Y_{1:t-1}}(x), f^{\theta_\star}_{Y_{1:t-1}}(x),Y_t)\,,$$
$$\nabla^2 \ln q(f^{\theta}_{Y_{1:t-1}}(x),Y_t) = \psi(\nabla^2 f^{\theta}_{Y_{1:t-1}}(x),\nabla f^{\theta}_{Y_{1:t-1}}(x), f^{\theta}_{Y_{1:t-1}}(x),Y_t)\,,$$

we obtain, according to **H** 3.32-((a))-((b)), that showing (3.55)-(3.56) is equivalent to showing that for all $\theta_n \to_{\mathbb{P}} \theta_\star$,

$$n^{-1/2}\sum_{t=1}^{n}\nabla \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}},Y_t) \overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow} \mathrm{N}(0,\mathscr{J}(\theta_\star))\,, \quad (3.58)$$

$$n^{-1}\sum_{t=1}^{n}\frac{\partial^2 \ln q(f^{\theta_n}_{Y_{-\infty:t-1}},Y_t)}{\partial\theta_i\partial\theta_j} \overset{\mathbb{P}\text{-prob}}{\longrightarrow} \mathbb{E}\left[\frac{\partial^2 \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}},Y_t)}{\partial\theta_i\partial\theta_j}\right]\,. \quad (3.59)$$

But this follows from Lemma 3.34 and Lemma 3.35. It remains to show (3.57). Since under **H** 3.27,

$$\mathbb{E}\left[\nabla^2 f^{\theta_\star}_{Y_{-\infty:0}}\frac{\partial \ln q}{\partial x}(\nabla f^{\theta_\star}_{Y_{-\infty:0}},Y_1)\right] = 0\,,$$

we have, using (3.50),

$$\mathscr{J}(\theta_\star) = \mathbb{E}\left[\psi(\nabla^2 f^{\theta_\star}_{Y_{-\infty:0}},\nabla f^{\theta_\star}_{Y_{-\infty:0}},f^{\theta_\star}_{Y_{-\infty:0}},Y_1)\right]\,.$$

The proof is completed. ■

3.4 Bayesian inference

In addition to classical or likelihood inference, there is another approach termed Bayesian inference, wherein prior belief is combined with data to obtain posterior distributions on which statistical inference is based. Although we focus primarily on likelihood based inference in this text, it is also worthwhile to consider Bayesian inference for some problems. Except for some simple cases, Bayesian inference can be computationally intensive and may rely on computational techniques.

The basic idea in Bayesian analysis is that a parameter vector, say $\theta \in \Theta$, is unknown to a researcher, so a *prior* distribution, $\pi(\theta)$, is put on the parameter vector. The researcher also proposes a model or likelihood, $\mathrm{p}(x \mid \theta)$, that describes how the data X depend on the parameter vector. Inference about θ is then based on the *posterior* distribution, which is obtained via Bayes's theorem,

$$\pi(\theta \mid X = x) \propto \pi(\theta)\,\mathrm{p}(x \mid \theta).$$

In some simple cases, the prior and the likelihood are *conjugate* distributions that may be combined easily. For example, in n fixed repeated (i.i.d.) Bernoulli experiments with probability of success θ, a *Beta-Binomial* conjugate pair is taken. In this case the prior is Beta(a,b): $\pi(\theta) \propto \theta^a(1-\theta)^b$; the values $a,b > -1$ are called hyperparameters. The likelihood in this example is Binomial(n,θ): $\mathrm{p}(x \mid \theta) \propto \theta^x(1-\theta)^{n-x}$, from which we easily deduce that the posterior is also Beta, $\pi(\theta \mid X = x) \propto \theta^{x+a}(1-\theta)^{n+b-x}$, and from which inference may easily be achieved. As previously mentioned, in more complex experiments, the posterior distribution is often difficult to obtain by direct calculation, so MCMC techniques are employed. The main idea is that we may not be able to explicitly display the posterior, but we may be able to simulate from the posterior.

In the remaining part of this section, we give three examples of the Bayesian analysis of time series models. The first example is a simple linear AR(p) model. Then, we discuss two nonlinear examples, a threshold model and a GARCH model.

Example 3.36 (AR(p) Model) *We now discuss the Bayesian approach for fitting a normal autoregressive model of order p. We write the model as* $Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sigma Z_t$ *where* $\{Z_t,\, t \in \mathbb{N}\} \sim_{\text{iid}} \mathrm{N}(0,1)$*; replace* Y_s *by* $Y_s - \mu$ *if the mean is not zero. Define*

$$\mathbf{Y} = \begin{pmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{pmatrix},\ \boldsymbol{\phi} = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{pmatrix},\ \boldsymbol{Z} = \begin{pmatrix} Z_{p+1} \\ Z_{p+2} \\ \vdots \\ Z_n \end{pmatrix},$$

$$X = \begin{pmatrix} Y_p & Y_{p-1} & \dots & Y_1 \\ Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & & \vdots \\ Y_n & Y_{n-1} & \dots & Y_{n-p+1} \end{pmatrix}.$$

The model can then be written as

$$\mathbf{Y} = X\boldsymbol{\phi} + \sigma\boldsymbol{Z},$$

which implies that $\mathbf{Y} \sim \mathrm{N}(X\phi,\ \sigma^2\mathrm{I})$*, with X being* $(n-p) \times p$ *matrix; for simplicity, X is assumed to be of full column rank p. It is easier to reparametrize the model in terms of a precision parameter,* $\tau = 1/\sigma^2$*. Then the likelihood function can be written as, up to a constant,*

$$\mathrm{L}_{\boldsymbol{\phi},\,\tau}\mathbf{Y} = \mathrm{p}(\mathbf{Y} \mid \boldsymbol{\phi},\ \tau) \propto \tau^{(n-p)/2} \exp(-\tau(\mathbf{Y} - X\boldsymbol{\phi})'(\mathbf{Y} - X\boldsymbol{\phi})/2)\ . \tag{3.60}$$

In a Bayesian framework, there are two ways for considering the prior settings of the ARMA parameters. One is to constrain the prior to be non-zero only over the region defined by the stationarity conditions, such as in **?** *and* **?**. *The second method is to simply ignore the stationary assumptions and proceed to use, e.g., the normal-gamma family as a conjugate prior distribution, such as in* **?**. *In this example, we apply the latter method, i.e., no restriction for stationarity or invertibility is made in the prior distribution. We use the following specification*

$$\begin{cases} \mathbf{Y} \mid (\boldsymbol{\phi},\ \tau) \sim \mathrm{N}(X\boldsymbol{\phi},\ \mathrm{I}/\tau) \\ \boldsymbol{\phi} \mid \tau \sim \mathrm{N}_p(\boldsymbol{\phi}_0,\ V_0/\tau) \\ \tau \sim \Gamma(a,\ b) \end{cases}$$

where $\boldsymbol{\phi}_0$ *and* V_0 *are the prior mean and covariance matrix, and a and b are the shape and scale of the prior of the precision parameter. The matrix* V_0 *is assumed to be invertible (which is generally chosen to be* $V_0 = \gamma_0^2\mathrm{I}$*, and we can take* γ_0^2 *large if the prior is meant to be noninformative). After straightforward simplification, the posterior joint density of* $(\boldsymbol{\phi},\ \tau)$ *becomes*

$$\begin{aligned} \pi(\boldsymbol{\phi},\ \tau \mid \mathbf{Y}) &\propto \tau^{\tilde{a}-1} \exp(-\tau(b + (1/2)\boldsymbol{\phi}_0' V_0^{-1}\boldsymbol{\phi}_0 + (1/2)\|\boldsymbol{y} - X\boldsymbol{\phi}\|^2)) \\ &\propto \tau^{\tilde{a}-1} \exp(-\tau\tilde{b}) \exp\left(-(\tau/2)\cdot(\boldsymbol{\phi} - \overline{\boldsymbol{\phi}})'(X'X + V_0^{-1})(\boldsymbol{\phi} - \overline{\boldsymbol{\phi}})\right) \end{aligned}$$

where

$$\tilde{a} = \frac{n}{2} + a$$
$$\tilde{b} = b + \frac{\mathbf{Y}'\mathbf{Y} - (X'\mathbf{Y} + V_0^{-1}\boldsymbol{\phi}_0)'(X'X + V_0^{-1})^{-1}(X'\mathbf{Y} + V_0^{-1}\boldsymbol{\phi}_0)}{2}$$
$$\overline{\boldsymbol{\phi}} = (X'X + V_0^{-1})^{-1}(X'\mathbf{Y} + V_0^{-1}\boldsymbol{\phi}_0).$$

From the joint posterior density $f(\boldsymbol{\phi},\ \tau \mid \mathbf{Y})$, it can be seen that the conditional posterior density of $(\boldsymbol{\phi} \mid \tau,\ \mathbf{Y})$ is multivariate normal with mean $\overline{\boldsymbol{\phi}}$ and covariance matrix $\tau^{-1}(X'X + V_0^{-1})^{-1}$, i.e.,

$$\boldsymbol{\phi} \mid (\tau,\ \mathbf{Y}) \sim \mathrm{N}_p(\overline{\boldsymbol{\phi}},\ \tau^{-1}(X'X + V_0^{-1})^{-1}) .$$

The conditional posterior density of $(\tau \mid \boldsymbol{\phi},\ \mathbf{Y})$ is $\Gamma(\tilde{a},\ \tilde{b}(\boldsymbol{\phi}))$,

$$\tilde{b}(\boldsymbol{\phi}) := b + (1/2)\boldsymbol{\phi}_0' V_0^{-1}\boldsymbol{\phi}_0 + (1/2)\|\boldsymbol{Y} - X\boldsymbol{\phi}\|^2 .$$

As an example, we use our R *script* `arp.mcmc` *to analyze a simulated AR(2) model with $\phi_1 = 1$, $\phi_2 = -.9$ and $\sigma^2 = 1$.*

```
> x = arima.sim(list(order=c(2,0,0), ar=c(1,-.9)), n=1000)  # generate an ar2
> arp.mcmc(x, porder=2, n.iter=1000, n.warmup=100)    # sample from posterior
      phi.Series 1     phi.Series 2         tauinv
    Min.   :0.9831   Min.   :-0.9598   Min.   :0.8068
    1st Qu.:1.0142   1st Qu.:-0.9292   1st Qu.:0.9117
    Median :1.0228   Median :-0.9219   Median :0.9419
    Mean   :1.0226   Mean   :-0.9217   Mean   :0.9430
    3rd Qu.:1.0311   3rd Qu.:-0.9138   3rd Qu.:0.9739
    Max.   :1.0672   Max.   :-0.8744   Max.   :1.0886
```

A summary as well as trace plots are produced by the script; see Figure 3.1.

Example 3.37 (SETAR model for the lynx series) *A two-regime SETAR model is considered, and denoted by SETAR (2; $p_1,\ p_2$):*

$$Y_t = \begin{cases} \phi_0^{(1)} + \sum_{i=1}^{p_1} \phi_i^{(1)} Y_{t-i} + Z_t^{(1)} & Y_{t-d} \le r \\ \phi_0^{(2)} + \sum_{i=1}^{p_2} \phi_i^{(2)} Y_{t-i} + Z_t^{(2)} & Y_{t-d} > r \end{cases} \tag{3.61}$$

where $Z_t^{(k)} = \sigma_k Z_t$, $k = 1,2$, and $\{Z_t,\ t \in \mathbb{N}\} \sim_{\text{iid}} \mathrm{N}(0,\ 1)$. This is a piecewise AR model, where regime switching is driven by lagged values of the series and a threshold value r. The unknown parameters in the model are $\boldsymbol{\phi}_k = (\phi_0^{(k)},\ \phi_1^{(k)},\ \ldots,\ \phi_{p_k}^{(k)})'$, $k = 1,2$, the threshold value r, the delay lag d and the regime error variances are $\sigma_k^2,\ k = 1,2$.

Likelihood approaches for estimating threshold models usually involve setting the threshold r and the delay d (e.g., $r = 0$ and $d = 1$), or choosing r, d via some information criterion. The likelihood inference for the parameters is then carried out conditionally upon these choices, but ignore the associated uncertainty in those parameters. The uncertainty in estimating r and d is explicitly captured via a Bayesian approach.

For SETAR models, the stationary and invertible conditions in the literature are rather involved. As in Example 3.36, we do not take these constraints into account in the Bayesian analysis. We use the following specification for the priors. The AR parameters are assumed to be normal with zero mean and covariance V_k, $k = 1,2$, i.e., $\boldsymbol{\phi}_k \sim \mathrm{N}(0, V_k)$, where V_k is the prior covariance matrix. We may for example set $V_k = \gamma_v^2 \mathrm{I}_{g_k}$, where γ_v^2 is the prior variance (generally taken to be large). An inverse-gamma prior is assumed for σ_k^2, i.e., $1/\sigma_k^2 \sim \Gamma(a_k, b_k)$, for $k = 1,2$, where the hyper-parameters a_k and b_k are known constants. The prior for the threshold value r is uniformly distributed between $[r_{\min}, r_{\max}]$, where $r_{\min}$ and $r_{\max}$ are generally specified as percentiles of the observations (e.g., 10 and 90 percentiles), for identification purposes in each regime. A discrete uniform prior on $\{1,\ 2,\ \ldots,\ d_{\max}\}$ is employed for the delay d.

Let $\boldsymbol{Y} = (Y_{p+1}, \ldots, Y_n)$ be the observed values of the TAR model, where $p = \max(p_1, p_2)$. the conditional likelihood function of the model is:

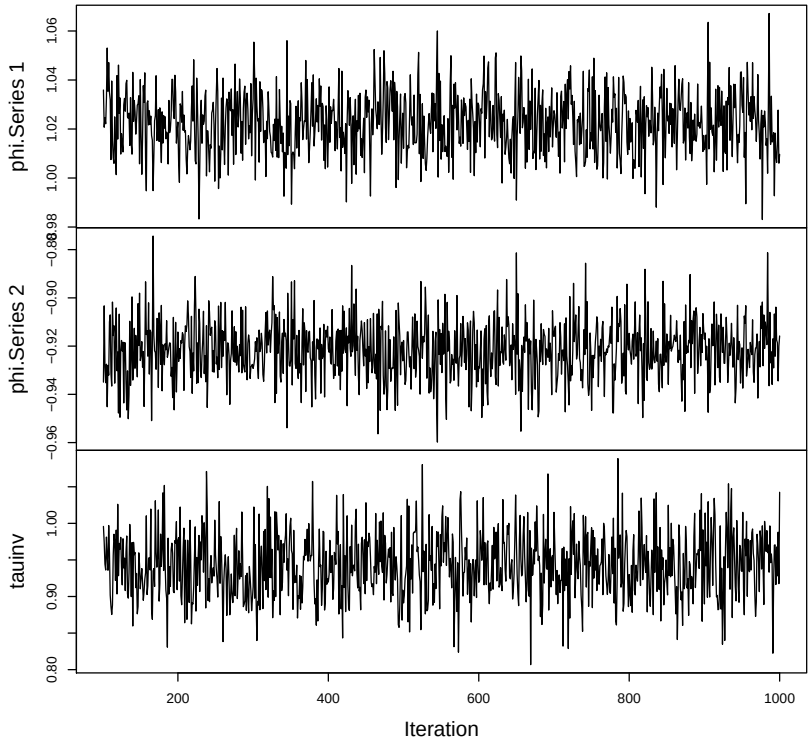

Fig. 3.1 Output of `arp.mcmc` for Example 3.36. Each plot shows the trace, after burnin, of the draws for each component.

$$L(\boldsymbol{y} \mid \boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\ \sigma_1^2,\ \sigma_2^2,\ r,\ d) \propto \sigma_1^{-n_1}\sigma_2^{-n_2} \times \exp\left(\sum_{t=p+1}^{n}\sum_{k=1}^{2} -\frac{1}{2\sigma_k^2}\left\{Y_t - \phi_0^{(k)} - \sum_{i=1}^{p_k}\phi_j^{(k)}Y_{t-i}\right\}^2 I_t^{(k)}\right)$$

where $n_1 := \sum_{t=p+1}^{n} \mathbb{1}\{Y_{t-d} \le r\}$, $n_2 = n - n_1$ *and* $I_t^{(k)}$ *is the indicator variable* $\mathbb{1}\{r_{k-1} \le Y_{t-d} < r_k\}$ *with* $r_0 = -\infty$, $r_1 = r$ *and* $r_2 = \infty$. *For notational convenience, let the full parameter vector be* $\boldsymbol{\theta} = \{\boldsymbol{\phi}^{(1)},\ \boldsymbol{\phi}^{(2)},\ \sigma_1^2,\ \sigma_2^2,\ r,\ d\}$. *Under the prior specification, the k-th regime AR parameter* $\boldsymbol{\phi}^{(k)} \mid (\boldsymbol{Y}, \boldsymbol{\theta} \setminus \{\boldsymbol{\phi}^{(k)}\})$ *is Gaussian with mean* $\bar{\boldsymbol{\phi}}^{(k)}$ *and covariance* $\bar{V}_k$ *given by*

$$\begin{aligned}\bar{\boldsymbol{\phi}}^{(k)} &= \sigma_k^{-2}\bar{V}_k X_k' \boldsymbol{Y}_k \\ \bar{V}_k &= (\sigma_k^{-2}X_k'X_k + V_k^{-1})^{-1}\,,\end{aligned}$$

where the design matrices X_1 *and* X_2 *are given by*

$$X_1 = \begin{bmatrix} 1 & Y_{\ell_1-1} & \cdots & \cdots & Y_{\ell_1-p_1} \\ 1 & Y_{\ell_2-1} & \cdots & \cdots & Y_{\ell_2-p_1} \\ \vdots & \vdots & & & \vdots \\ 1 & Y_{\ell_{n_1}-1} & \cdots & \cdots & Y_{\ell_{n_1}-p_1} \end{bmatrix}$$

and

$$X_2 = \begin{bmatrix} 1 & Y_{u_1-1} & \cdots & \cdots & Y_{u_1-p_2} \\ 1 & Y_{u_2-1} & \cdots & \cdots & Y_{u_2-p_2} \\ \vdots & \vdots & & & \vdots \\ 1 & Y_{u_{n_1}-1} & \cdots & \cdots & Y_{u_{n_2}-p_2} \end{bmatrix}$$

where ℓ_i and u_j are the time indices of the i-th and j-th observations in the "lower" and "upper" regimes: ℓ_i is the index for regime 1 and u_j is for regime 2. The vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are set as $\boldsymbol{Y}_1 = (Y_{\ell_1}, \ldots, Y_{\ell_{n_1}})'$ and $\boldsymbol{Y}_2 = (Y_{u_1}, \ldots, Y_{u_{n_2}})'$.

The posterior distribution of the inverse of the variance $\sigma_k^{-2} \mid (\boldsymbol{Y}, \boldsymbol{\theta} \setminus \{\sigma_k^{-2}\})$ is Gamma with shape and scale

$$\tilde{a}_k = a_k + n_k/2 \ , \ \tilde{b}_k = b_k + (1/2)\|\boldsymbol{Y}_k - X_k \boldsymbol{\phi}_k\|^2 \ .$$

The posterior distribution for the threshold is a bit more involved. By simple multiplication of the likelihood and prior for r, the posterior density of $r \mid (\boldsymbol{Y}, \boldsymbol{\theta} \setminus \{r\})$ is given by (up to a normalization constant)

$$\sigma_1^{-n_1} \sigma_2^{-n_2} \exp\left(-\frac{1}{2} \sum_{k=1}^{2} \frac{\|\boldsymbol{Y}_k - X_k \boldsymbol{\phi}_k\|^2}{\sigma_k^2} \right) \mathbb{1}_{\{r \in [r_{\min}, r_{\max}]\}}$$

where $n_1, n_2, \boldsymbol{Y}_1, \boldsymbol{Y}_2, X_1, X_2$ depend on r. This posterior distribution is not a known or standard form and therefore we cannot use the elementary Gibbs sampler. To solve this problem, we use a hybrid of the ideas behind the Gibbs sampler and the Metropolis-Hastings algorithm. The algorithm merely replaces the step that samples from the full-posterior by a Metropolis-Hastings step, which leaves the full-posterior stationary. Now by simulating sufficiently many of these Metropolis-Hastings steps before proceeding to the next component, we will approximately mimic the elementary Gibbs sampler itself. However, there is often a far more efficient procedure that only updates the component once (or a small finite number of times) before moving on to other coordinates. Like the Gibbs sampler, this algorithm can be thought of as a special case of the Metropolis-Hastings algorithm. In this special case, **?** *use a random walk Metropolis algorithm with a Gaussian proposal density with zero mean and variance σ^2, to sample this parameter. Finally, the posterior distribution of the delay $d \mid (\boldsymbol{Y}, \boldsymbol{\theta} \setminus \{d\})$ is a multinomial trial with probabilities*

$$\frac{L(\boldsymbol{Y} \mid \boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\ \sigma_1^2,\ \sigma_2^2,\ r, j)}{\sum_{i=1}^{d_{\max}} L(\boldsymbol{Y} \mid \boldsymbol{\phi}^{(1)}, \boldsymbol{\phi}^{(2)}\ \sigma_1^2,\ \sigma_2^2,\ r, i)} \ , \ j = 1, 2, \ldots, d_{\max} \ .$$

The sampling scheme amounts to updating parameters in turn, which leads to the algorithm implemented in the R *package* `BAYSTAR`*. However, sampling the delay distribution this way is not optimal and may lead to slow mixing.*

We analyzed the lynx data set presented in **??**. *We used* `BAYSTAR` *to perform a Bayesian inference of a two-regime threshold model, which was considered in* **??**. *In particular, our goal was to fit the same model, an AR(2) for each regime, with the threshold delay being 2. In this example, we set the maximum delay, $d_{\max}$, equal to 2. In this case, the highest posterior probability for the delay lag was indeed 2. We also note that the variance of the proposal density, σ^2, which is called* `step.thv`*, must be specified. If the scaling parameter is too large, the acceptance ratio of the algorithm will be small. If it is too small, then the acceptance ratio is high but the algorithm mixes slowly (see Example 4.4 for a discussion of this issue). This can be monitored by observing the output produced by the package.*

The R *code and a partial output is given below. In addition to the output shown below,* `BAYSTAR` *exhibits various diagnostic plots, which we do not display.*

```
> require(BAYSTAR)
> lynx.out = BAYSTAR(log10(lynx), lagp1=1:2, lagp2=1:2, Iteration=10000, Burnin=2000, d0=2,
    step.thv=.5)
              mean  median     s.d.   lower   upper
phi1.0      0.5620  0.5701   0.1683  0.2039  0.8733
phi1.1      1.2614  1.2607   0.0724  1.1251  1.4022
phi1.2     -0.4143 -0.4174   0.0919 -0.5867 -0.2274
phi2.0      1.4567  1.5353   1.0021 -0.6318  3.2400
```

```
phi2.1      1.5694  1.5643   0.1317  1.3221  1.8346
phi2.2     -1.0623 -1.0759   0.2859 -1.5988 -0.4590
sigma^2 1   0.0363  0.0356   0.0067  0.0258  0.0513
simga^2 2   0.0557  0.0535   0.0143  0.0349  0.0899
r           3.2057  3.2810   0.1818  2.5784  3.3848
acceptance rate of r =  23.76 %
The highest posterior prob. of lag is at :  2
```

We note that the result of the MCMC analysis is similar to the frequentist analysis presented in **??**.

Example 3.38 (Bayesian GARCH) *In this example, we consider a GARCH(1,1) model with Student-t innovations for log-returns, i.e.,*

$$Y_t = \sigma_t \left(\frac{\nu-2}{\nu}\right)^{1/2} \varepsilon_t \tag{3.62}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{3.63}$$

where $\{Z_t,\, t \in \mathbb{N}\}$ *are i.i.d. t-distribution with* $\nu > 2$ *degrees of freedom,* $\alpha_0 > 0,\ \alpha_1, \beta_1 \geq 0$ *and* $\nu > 2$. *To perform the Bayesian analysis of this model, it is easier to use data augmentation; see* **?** *and* **?**. *The idea is to represent the t-distribution as a scale mixture of Gaussians as suggested by* **?**. *A random variable Z is a Gaussian scale mixture if it can be expressed as the product of a standard Gaussian G and an independent positive scalar random variable* $H^{1/2}$*: Z* $= H^{1/2}G$. *The variable H is the* multiplier *or the* scale. *If H has finite support, then Z is a finite mixture of Gaussians, whereas if H has a density (with respect to Lebesgue measure) on* $\mathbb{R}_+$*, then Z is a continuous mixture of Gaussian variables. Gaussian scale mixtures are symmetric, zero mean, and have leptokurtic marginal densities (tails heavier than those of a Gaussian distribution). Assume that the distribution of H is inverse gamma (denoted* IG*), with shape and scale parameters both equal to* $\nu/2$*; then the distribution of H has the density*

$$p_\nu(h) = \left(\frac{\nu}{2}\right)^{\nu/2} \frac{h^{-(\nu+2)/2}}{\Gamma(\nu/2)} \exp\left(-\frac{\nu}{2h}\right),\ h \geq 0.$$

If $G \sim \mathrm{N}(0,1)$*, then the distribution of* $Z = H^{1/2}G$ *has a density given by*

$$f(z) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma(\nu/2)\sqrt{2\pi}} \int_0^\infty h^{-(\nu+3)/2} \exp\left(-\frac{z^2+\nu}{2h}\right) \mathrm{d}h\,.$$

Using the result,

$$\int_0^\infty x^{-a/2} \exp\left(-\frac{b}{2x}\right) \mathrm{d}x = \left(\frac{2}{b}\right)^{(a-2)/2} \Gamma\left(\frac{a-2}{2}\right),$$

which derives from a simple change variable in the definition of the complete gamma function, we obtain that

$$f(z) = \frac{\left(\frac{\nu}{2}\right)^{\nu/2} \Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{2\pi}} \left(\frac{2}{z^2+\nu}\right)^{(\nu+1)/2},$$

which is the density of a Student-t distribution with ν *degrees of freedom. Using this representation of the t-distribution, we may rewrite* (3.62) *as follows,*

$$\begin{aligned}
Y_t &= \sigma_t \left(\frac{\nu-2}{\nu}\right)^{1/2} H_t^{1/2} G_t \\
G_t &\sim_{\text{iid}} \mathrm{N}(0,1) \\
H_t &\sim_{\text{iid}} \mathrm{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) \\
\sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2\,.
\end{aligned}$$

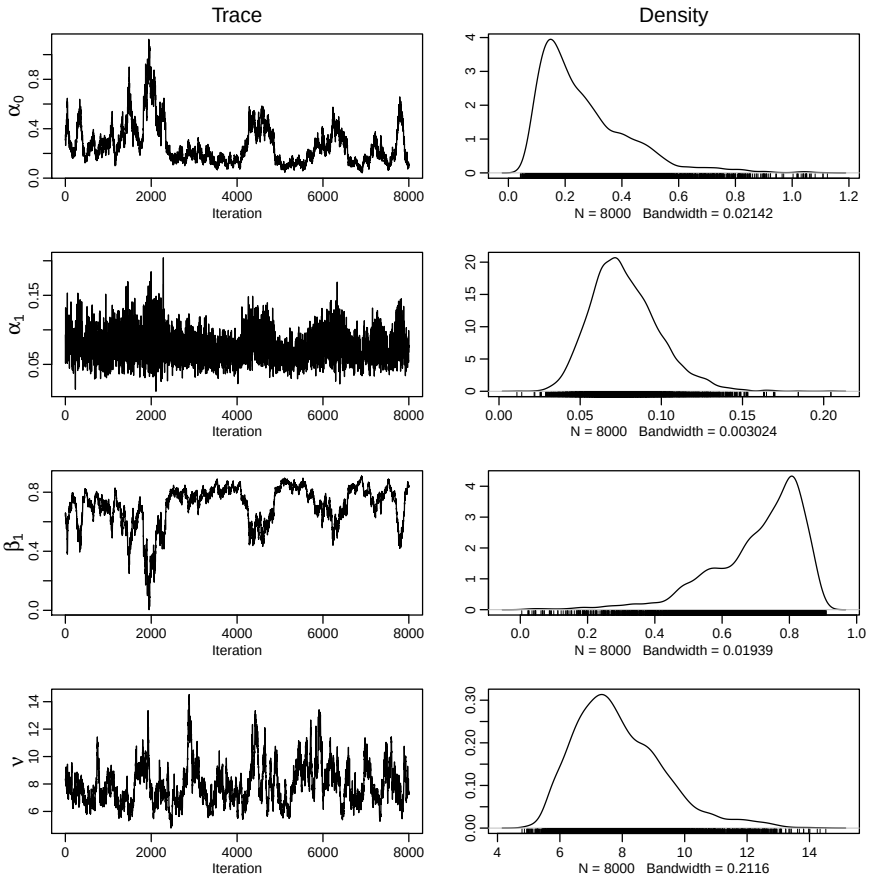

Fig. 3.2 Output of `bayesGARCH` for Example 3.38

Assume that we have n observations from this model. In order to write the likelihood function, we define the vectors, $\boldsymbol{H} = (H_1, \ldots, H_n)'$ *and* $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$*. The model parameters are collected into the vector* $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \beta, \nu)$*. Define the* $n \times n$ *diagonal matrix*

$$\Sigma = \Sigma(\boldsymbol{\theta}, \boldsymbol{H}) = \operatorname{diag}\left(\left\{ H_t \frac{\nu - 2}{\nu} \sigma_t^2(\boldsymbol{\alpha}, \beta_1) \right\}_{i=1}^{n} \right) ,$$

where $\sigma_t^2(\boldsymbol{\alpha}, \beta_1) = \alpha_0 + \alpha_1 Y_{t-1}^2 + \beta_1 \sigma_{t-1}^2(\boldsymbol{\alpha}, \beta_1)$*, and* $\sigma_0^2(\boldsymbol{\alpha}, \beta_1) = 0$*. We can express the conditional likelihood*

$$L(\boldsymbol{Y} \mid \boldsymbol{\theta}, \boldsymbol{H}) \propto (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{Y}' \Sigma^{-1} \boldsymbol{Y} \right) .$$

The prior distribution for the GARCH parameters $\boldsymbol{\alpha}$ *and* β_1 *are chosen to be independent and distributed according to the truncated normal distribution (to ensure the positivity of the coefficients)*

$$p(\boldsymbol{\alpha}) \propto (\det \Sigma_{\boldsymbol{\alpha}})^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha)' \Sigma_\alpha^{-1} (\boldsymbol{\alpha} - \boldsymbol{\mu}_\alpha) \right) \mathbb{1}_{\mathbb{R}_+^2}(\boldsymbol{\alpha})$$

$$p(\beta_1) \propto \sigma_\beta^{-1} \exp\left(-\frac{1}{2\sigma_\beta^2} (\beta_1 - \mu_\beta)^2 \right) \mathbb{1}_{\mathbb{R}^+}(\beta_1),$$

where μ_α*,* Σ_α*,* μ_β*,* σ_β *are the hyperparameters. Only positivity constraints are taken into account in the prior specification; no stationarity conditions are imposed.*

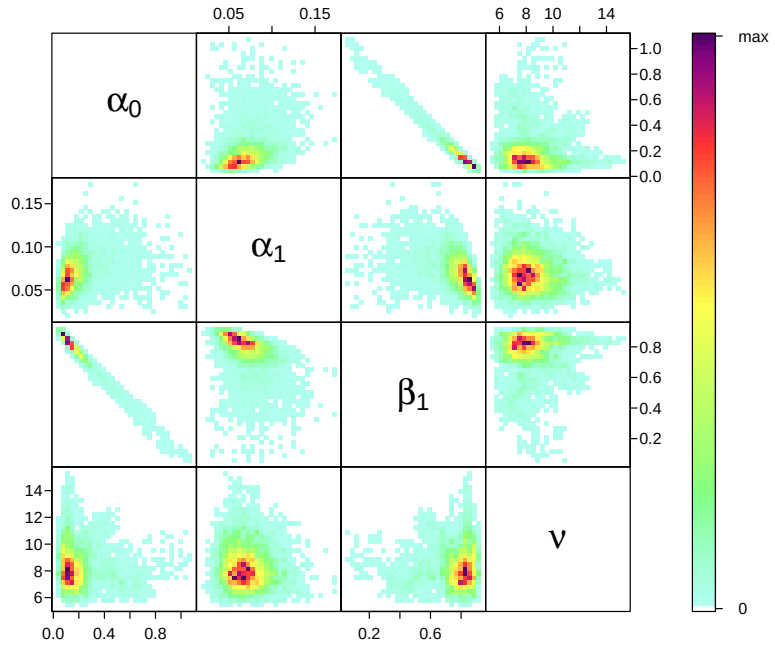

Fig. 3.3 Pairwise sampling distributions for Example 3.38.

Since the components H_t are i.i.d. from the inverse gamma distribution, the conditional distribution of the vector $\boldsymbol{H}$ given ν has a density given by

$$p(\boldsymbol{h} \mid \nu) = \left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}} \left[\Gamma(\frac{\nu}{2})\right]^{-n} \left(\prod_{t=1}^{n} h_t\right)^{-\frac{\nu}{2}-1} \exp\left(-\frac{1}{2}\sum_{t=1}^{n}\frac{\nu}{h_t}\right) .$$

The prior distribution on the degrees of freedom ν is chosen to be a translated exponential with parameters $\lambda > 0$ and $\delta \geq 2$

$$p(\nu) \ = \lambda \exp\left[-\lambda(\nu-\delta)\right] \mathbb{1}\{\nu \geq \delta\} .$$

For large values of λ, most of the mass of the prior is concentrated in the neighborhood of δ and a constraint on the degrees of freedom can be imposed in this manner. A light tail prior may be imposed by taking both δ and λ large. The joint prior distribution is then obtained by assuming that the priors are independent, i.e., $p(\bar{\theta},\bar{h}) = p(\bar{\alpha})p(\beta_1)p(h \mid \nu)p(\nu)$. The joint posterior and the full conditional densities cannot be expressed in closed form. Therefore, we cannot use the elementary Gibbs sampler and need to rely on a Metropolis-within-Gibbs strategy to approximate the posterior density. In this case, there is an R *package called* `bayesGARCH` *that can be used to perform the analysis. Implementations details are given in* **?**. *For this example, we use part of the* `CAC` *returns from the* R *dataset* `EuStockMarkets`. *The graphical output of the package is displayed in Figure 3.2, which shows the traces for each parameter (after burnin) and the corresponding posterior densities. We also use the* R *package* `IDPmisc` *to graph the pairwise results; see Figure 3.3.*

3.5 Some proofs

Let (X, d) be a Polish space equipped with its Borel sigma-field $\mathscr{X}$. Let $\theta : \mathsf{X}^{\mathbb{N}} \to \mathsf{X}^{\mathbb{N}}$ and $\tilde{\theta} : \mathsf{X}^{\mathbb{Z}} \to \mathsf{X}^{\mathbb{Z}}$ be the shift operators defined by: for all $\mathbf{x} = (x_t)_{t\in\mathbb{N}} \in \mathsf{X}^{\mathbb{N}}$ and all $\tilde{\mathbf{x}} = (\tilde{x}_t)_{t\in\mathbb{Z}} \in \mathsf{X}^{\mathbb{Z}}$,

$$\theta(\mathbf{x}) = (y_t)_{t\in\mathbb{N}}, \quad \text{where} \quad y_t = x_{t+1}, \quad \forall t \in \mathbb{N}\,, \tag{3.64}$$

$$\tilde{\theta}(\tilde{\mathbf{x}}) = (\tilde{y}_t)_{t\in\mathbb{Z}}, \quad \text{where} \quad \tilde{y}_t = \tilde{x}_{t+1}, \quad \forall t \in \mathbb{Z}\,. \tag{3.65}$$

Assume that $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}}, \mathbb{P}, \tilde{\theta})$ is a measure-preserving ergodic dynamical system. Denote by $\mathbb{E}$ the expectation operator associated to $\mathbb{P}$.

Let $(\bar{\ell}^{\theta}\,, \theta \in \Theta)$ be a family of measurable functions, $\bar{\ell}^{\theta} : \mathsf{X}^{\mathbb{Z}} \to \mathbb{R}$, indexed by $\theta \in \Theta$ where (Θ, d) is a compact metric space and denote $\bar{\mathsf{L}}_n^{\theta} := n^{-1}\sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k$. Moreover, consider $(\mathsf{L}_n^{\theta}\,, n \in \mathbb{N}^*, \theta \in \Theta)$ a family of upper-semicontinuous functions $\mathsf{L}_n^{\theta} : \mathsf{X}^{\mathbb{Z}} \to \mathbb{R}$ indexed by $n \in \mathbb{N}^*$ and $\theta \in \Theta$. Consider the following assumptions:

H 3.39 $\mathbb{E}[\sup_{\theta\in\Theta} \bar{\ell}_+^{\theta}] < \infty$,

H 3.40 $\mathbb{P}-$a.s., *the function* $\theta \mapsto \bar{\ell}^{\theta}$ *is upper-semicontinuous,*

H 3.41 $\lim_{n\to\infty} \sup_{\theta\in\Theta} |\mathsf{L}_n^{\theta} - \bar{\mathsf{L}}_n^{\theta}| = 0\,, \quad \mathbb{P}-$a.s.

Let $\left\{\bar{\theta}_n\,:\, n \in \mathbb{N}^*\right\} \subset \Theta$ and $\left\{\hat{\theta}_n\,:\, n \in \mathbb{N}^*\right\} \subset \Theta$ such that for all $n \geq 1$,

$$\bar{\theta}_n \in \arg\max_{\theta\in\Theta} \bar{\mathsf{L}}_n^{\theta}\,, \quad \hat{\theta}_n \in \arg\max_{\theta\in\Theta} \mathsf{L}_n^{\theta}\,.$$

The proofs are adapted from **?**.

Theorem 3.42. *Assume* ***H*** *3.39–****H*** *3.40. Then*

$$\lim_{n\to\infty} \mathsf{d}(\bar{\theta}_n, \Theta_\star) = 0,\ \mathbb{P}-\text{a.s. } \textit{where } \Theta_\star := \arg\max_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^{\theta}]. \tag{3.66}$$

If, in addition, ***H*** *3.41 holds, then* $\lim_{n\to\infty} \mathsf{d}(\hat{\theta}_n, \Theta_\star) = 0\,,\ \mathbb{P}-$a.s. *and,*

$$\lim_{n\to\infty} \mathsf{L}_n^{\hat{\theta}_n} = \sup_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^{\theta}]\,, \quad \mathbb{P}-\text{a.s.} \tag{3.67}$$

$$\textit{for all } \theta \in \Theta, \quad \lim_{n\to\infty} \mathsf{L}_n^{\theta} = \mathbb{E}[\bar{\ell}^{\theta}]\,, \quad \mathbb{P}-\text{a.s.} \tag{3.68}$$

PROOF. First note that according to the Birkhoff ergodic theorem (**??**) and **H** 3.39, for all $\theta \in \Theta$, $\lim_{n\to\infty} \bar{\mathsf{L}}_n^{\theta}$ exists $\mathbb{P}-$a.s., and

$$\lim_{n\to\infty} \bar{\mathsf{L}}_n^{\theta} = \lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k = \mathbb{E}[\bar{\ell}^{\theta}], \quad \mathbb{P}-\text{a.s.} \tag{3.69}$$

Let K be a compact subset of Θ. For all $\theta_0 \in K$, $\mathbb{P}-$a.s.,

$$\begin{aligned} &\limsup_{\rho\to 0} \limsup_{n\to\infty} \sup_{\theta\in\mathrm{B}(\theta_0,\rho)} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k \\ &\leq \limsup_{\rho\to 0} \limsup_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} \sup_{\theta\in\mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta} \circ \tilde{\theta}^k = \limsup_{\rho\to 0} \mathbb{E}[\sup_{\theta\in\mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta}]\,, \end{aligned} \tag{3.70}$$

where the last equality follows from **H** 3.39 and the Birkhoff ergodic theorem (**??**). Moreover, by the monotone convergence theorem applied to the nondecreasing function $\rho \mapsto \sup_{\theta\in \mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta}$, we have

$$\limsup_{\rho\to 0} \mathbb{E}[\sup_{\theta\in \mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta}] = \mathbb{E}[\limsup_{\rho\to 0} \sup_{\theta\in \mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta}] \le \mathbb{E}[\bar{\ell}^{\theta_0}] \,, \tag{3.71}$$

where the last inequality follows from **H** 3.40. Combining (3.70) and (3.71), we obtain that for all $\eta > 0$ and $\theta_0 \in K$, there exists $\rho^{\theta_0} > 0$ satisfying

$$\limsup_{n\to\infty} \sup_{\theta\in \mathrm{B}(\theta_0,\rho^{\theta_0})} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k \le \mathbb{E}[\bar{\ell}^{\theta_0}] + \eta \le \sup_{\theta\in K} \mathbb{E}[\bar{\ell}^{\theta}] + \eta \,, \quad \mathbb{P}-\text{a.s.}$$

Since K is a compact subset of Θ, we can extract a finite subcover of K from $\bigcup_{\theta_0\in K} \mathrm{B}(\theta_0,\rho^{\theta_0})$, so that

$$\limsup_{n\to\infty} \sup_{\theta\in K} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k \le \sup_{\theta\in K} \mathbb{E}[\bar{\ell}^{\theta}] + \eta \,, \quad \mathbb{P}-\text{a.s.} \tag{3.72}$$

Since η is arbitrary, we obtain

$$\limsup_{n\to\infty} \sup_{\theta\in K} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k \le \sup_{\theta\in K} \mathbb{E}[\bar{\ell}^{\theta}] \,, \quad \mathbb{P}-\text{a.s.} \tag{3.73}$$

Moreover, (3.71) implies

$$\limsup_{\rho\to 0} \sup_{\theta\in \mathrm{B}(\theta_0,\rho)} \mathbb{E}[\bar{\ell}^{\theta}] \le \limsup_{\rho\to 0} \mathbb{E}[\sup_{\theta\in \mathrm{B}(\theta_0,\rho)} \bar{\ell}^{\theta}] \le \mathbb{E}[\bar{\ell}^{\theta_0}] \,,$$

This shows that $\theta \mapsto \mathbb{E}[\bar{\ell}^{\theta}]$ is upper-semicontinuous. As a consequence, $\Theta_\star := \arg\max_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^{\theta}]$ is a closed and nonempty subset of Θ and therefore, for all $\varepsilon > 0$, $K_\varepsilon := \{\theta \in \Theta; \mathsf{d}(\theta,\Theta_\star) \ge \varepsilon\}$ is a compact subset of Θ. Using again the upper-semicontinuity of $\theta \mapsto \mathbb{E}[\bar{\ell}^{\theta}]$, there exists $\theta_\varepsilon \in K_\varepsilon$ such that for all $\theta_\star \in \Theta_\star$,

$$\sup_{\theta\in K_\varepsilon} \mathbb{E}[\bar{\ell}^{\theta}] = \mathbb{E}[\bar{\ell}^{\theta_\varepsilon}] < \mathbb{E}[\bar{\ell}^{\theta_\star}] \,.$$

Finally, combining this inequality with (3.73), we obtain that $\mathbb{P}-$a.s.,

$$\begin{aligned}\limsup_{n\to\infty} \sup_{\theta\in K_\varepsilon} \bar{\mathsf{L}}_n^{\theta} = \limsup_{n\to\infty} \sup_{\theta\in K_\varepsilon} n^{-1} \sum_{k=0}^{n-1} \bar{\ell}^{\theta} \circ \tilde{\theta}^k \le \sup_{\theta\in K_\varepsilon} \mathbb{E}[\bar{\ell}^{\theta}] \\ < \mathbb{E}[\bar{\ell}^{\theta_\star}] \overset{(1)}{=} \lim_{n\to\infty} \bar{\mathsf{L}}_n^{\theta_\star} \le \liminf_{n\to\infty} \bar{\mathsf{L}}_n^{\bar{\theta}_n} \,,\end{aligned} \tag{3.74}$$

where (1) follows from (3.69). This inequality ensures that $\bar{\theta}_n \notin K_\varepsilon$ for all n larger to some $\mathbb{P}-$a.s. finite integer-valued random variable. This completes the proof of (3.66) since ε is arbitrary.

Eq. (3.68) follows from (3.69) and **H** 3.41. Let $\theta_\star$ be any point in $\Theta_\star$. Then, $\mathbb{P}-$a.s.,

$$\begin{aligned}\mathbb{E}[\bar{\ell}^{\theta_\star}] \overset{(1)}{=} \liminf_{n\to\infty} \bar{\mathsf{L}}_n^{\theta_\star} \overset{(2)}{\le} \liminf_{n\to\infty} \bar{\mathsf{L}}_n^{\bar{\theta}_n} \le \limsup_{n\to\infty} \bar{\mathsf{L}}_n^{\bar{\theta}_n} \\ = \limsup_{n\to\infty} \sup_{\theta\in\Theta} \bar{\mathsf{L}}_n^{\theta} \overset{(3)}{\le} \sup_{\theta\in\Theta} \mathbb{E}[\bar{\ell}^{\theta}] = \mathbb{E}[\bar{\ell}^{\theta_\star}] \,,\end{aligned}$$

where (1) follows from (3.69), (2) follows from the definition of $\bar{\theta}_n$ and (3) is obtained by applying (3.73) with $K = \Theta$. Thus,

$$\lim_{n\to\infty} \bar{\mathsf{L}}_n^{\bar{\theta}_n} = \mathbb{E}[\bar{\ell}^{\theta_\star}] \,, \quad \mathbb{P}-\text{a.s.} \tag{3.75}$$

Denote $\delta_n := \sup_{\theta\in\Theta} |\mathsf{L}_n^{\theta} - \bar{\mathsf{L}}_n^{\theta}|$. We get

$$\bar{\mathsf{L}}_n^{\bar{\theta}_n} - \delta_n \overset{(1)}{\le} \mathsf{L}_n^{\bar{\theta}_n} \overset{(2)}{\le} \mathsf{L}_n^{\hat{\theta}_n} \overset{(1)}{\le} \bar{\mathsf{L}}_n^{\hat{\theta}_n} + \delta_n \overset{(3)}{\le} \bar{\mathsf{L}}_n^{\bar{\theta}_n} + \delta_n \,. \tag{3.76}$$

where (1) follows from the definition of δ_n, (2) from the definition of $\hat{\theta}_n$ and (3) from the definition of $\bar{\theta}_n$. Combining the above inequalities with (3.75) and **H** 3.41 yields (3.67). (3.76) also implies that

$$\lim_{n\to\infty} \bar{\mathsf{L}}_n^{\hat{\theta}_n} = \mathbb{E}[\bar{\ell}^{\theta_\star}] \,, \quad \mathbb{P}-\text{a.s.}$$

which yields, using (3.74),

$$\limsup_{n\to\infty} \sup_{\theta\in K_\varepsilon} \bar{\mathsf{L}}_n^\theta < \liminf_{n\to\infty} \bar{\mathsf{L}}_n^{\hat{\theta}_n} = \limsup_{n\to\infty} \bar{\mathsf{L}}_n^{\hat{\theta}_n} = \mathbb{E}[\bar{\ell}^{\theta_\star}], \quad \mathbb{P}-\text{a.s.}$$

where $K_\varepsilon := \{\theta \in \Theta; \mathsf{d}(\theta, \Theta_\star) \geq \varepsilon\}$. Therefore, $\hat{\theta}_n \notin K_\varepsilon$ for all n larger to some $\mathbb{P}-$a.s.-finite integer-valued random variable. The proof is completed since ε is arbitrary. ■

Exercises

Exercise 3.43. Let $\{X_t,\, t \in \mathbb{N}\}$ be an X-valued stochastic process, where $(\mathsf{X}, \mathscr{X})$ is a measurable space. Assume that for each positive integer n, the random vector $(X_0, \ldots, X_{n-1})$ has a joint density $x_{0:n-1} \mapsto \mathrm{p}^\theta(x_{0:n-1})$ with respect to the product measure $\lambda^{\otimes n}$ on the product space $(\mathsf{X}^n, \mathscr{X}^{\otimes(n-1)})$, where $\lambda \in \mathbb{M}_1(\mathscr{X})$. For simplicity, it is assumed that for all $n \in \mathbb{N}$ and all $x_{0:n-1}$, $\mathrm{p}^\theta(x_{0:n-1}) > 0$. Denote by $\mathrm{p}^\theta(x_n|x_{0:n-1})$ the conditional density of X_n given $X_{0:n-1}$, $\mathrm{p}^\theta(x_n|x_{0:n-1}) = \mathrm{p}^\theta(x_{0:n})/\mathrm{p}^\theta(x_{0:n-1})$. Assume that, for all $n \in \mathbb{N}$, and $x_{0:n-1} \in \mathsf{X}^n$, $\theta \mapsto \mathrm{p}^\theta(x_{0:n-1})$ is differentiable and that

$$\nabla \mathrm{p}^\theta(x_{0:n-1}) = \int \nabla \mathrm{p}^\theta(x_{0:n}) \lambda(\mathrm{d}x_n)\,.$$

Prove that the score function (the gradient of the log-likelihood) $\sum_{k=1}^n \nabla \log \mathrm{p}^\theta(X_k|X_{0:k-1})$ is a martingale adapted to the natural filtration of the process.

Exercise 3.44 (Conditionally Gaussian MLE for an AR(1)). Consider a causal AR(1) model, $X_t = \phi X_{t-1} + \sigma Z_t$, where $\{Z_t,\, t \in \mathbb{N}\}$ is a strong white noise having a density f, which is symmetric around the origin and satisfies $\int_{-\infty}^{\infty} x^2 f(x)\mathrm{Leb}(\mathrm{d}x) < \infty$. We put $\theta = (\phi, \sigma) \in \Theta$, where Θ is a compact subset of $(-1,1) \times \mathbb{R}_*^+$. We denote by $\theta_\star = (\phi_*, \sigma_*^2) \in \Theta^o$ the true value of the parameters. We observe $(X_0, \ldots, X_n)$ and we are interested in estimating these parameters. We fit the model using the conditional *Gaussian* maximum likelihood estimator. We denote by $(\hat{\phi}_n, \hat{\sigma}_n^2)$ these estimators.

(a) Derive the expressions of $(\hat{\phi}_n, \hat{\sigma}_n^2)$.
(b) Show that this estimator is strongly consistent.
(c) Show that the estimator $\hat{\phi}_n$ is asymptotically normal. Compute the asymptotic variance of this estimator.
(d) Assume that f is t-distribution with $\nu > 2$ degrees of freedom. Compute the variance of $\hat{\phi}_n$ (as a function of ν) and construct confidence intervals. Discuss the result.
(e) Assume that $\int_{-\infty}^{\infty} x^4 f(x)\mathrm{Leb}(\mathrm{d}x) < \infty$. Show that $(\hat{\phi}_n, \hat{\sigma}_n^2)$ is asymptotically normal.
(f) Determine confidence regions for $(\hat{\phi}_n, \hat{\sigma}_n^2)$ when f is a Student t-distribution with $\nu > 4$ degrees of freedom.

Exercise 3.45 (Conditionally Gaussian MLE for an AR(1)). We use the same notations and definitions as in Exercise 3.44. We now fit the model using the conditional maximum likelihood estimator. We assume in this exercise that the model is well specified. We denote by $(\tilde{\phi}_n, \tilde{\sigma}_n^2)$ these estimators.

(a) Derive the estimating equations for $(\tilde{\phi}_n, \tilde{\sigma}_n^2)$.
(b) Derive explicitly the likelihood equations when f is a t-distribution with ν-degrees of freedom.
(c) Show that these estimators (assuming that the model is well-specified) are consistent and asymptotically normal.
(d) Compute the asymptotic variance of the resulting estimator.
(e) Compare the asymptotic variance with Exercise 3.44.

Exercise 3.46. Verify (3.32).

Exercise 3.47 (AR process with nonnegative innovation). Consider the nonnegative autoregressive process $X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t$, where $\{Z_t,\, t \in \mathbb{N}\}$ is a strong white noise with left endpoint of their common distributions being zero and $(\phi_1, \ldots, \phi_p) \in (0,1)^p$. This type of model is suitable for modelling

applications of data that are inherently nonnegative, such as streams flow in hydrology, interarrival times, etc. Based on the observations, $(X_0,\dots,X_n)$, we are interested in estimating the parameters. The parameters $(\phi_1,\dots,\phi_p)$ are estimated using the maximum likelihood estimator associated to the assumption that the common distribution of the Z's is unit exponential, so that $\mathbb{P}(Z_1 > x) = \mathrm{e}^{-x}$, $x \geq 0$. See **?**.

(a) Consider first the case $p = 1$. Show that the conditional likelihood function is given (up to a constant) by

$$\prod_{t=1}^{n} \mathbb{1}_{\{X_t - \phi_1 X_{t-1} \geq 0\}} \exp\left(\phi_1 \sum_{t=1}^{n} X_{t-1}\right) .$$

(b) Show that the conditional maximum likelihood estimator is given by

$$\hat{\phi}_{1,n} = \min_{1 \leq t \leq n} (X_t / X_{t-1}) .$$

(c) Consider now the general autoregressive case. Show that the likelihood function is given (up to a constant) by

$$\prod_{t=p}^{n} \mathbb{1}_{\{X_t - \phi X_{t-1} \geq 0\}} \exp\left(\phi_1 \sum_{t=p}^{n} X_{t-1} + \cdots + \phi_p \sum_{t=p}^{n} X_{t-p}\right) .$$

(d) Argue that the maximum likelihood estimator is approximately determined by solving the linear program

$$\max\left(\sum_{i=1}^{p} \phi_i\right) \quad \text{subject to} \quad X_t \geq \sum_{i=1}^{p} \phi_i X_{t-i} \text{ for } t \in \{1,\dots,n\} .$$

Determining the asymptotic properties of these estimators is difficult (see **?**); the rate of convergence can be in such a case faster than $n^{1/2}$.

Exercise 3.48 (Proof of Theorem 3.18).

(a) Show that

$$n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\hat{\theta}_n}(X_{t-p:t-1}; X_t) = 0 = n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\theta_\star}(X_{t-p:t-1}; X_t) + n^{-1} \sum_{t=p}^{n} \left(\int_0^1 \nabla^2 \ln q^{\theta_{n,s}}(X_{t-p:t-1}; X_t) \mathrm{d}s \right) \sqrt{n}(\hat{\theta}_n - \theta_\star) , \quad (3.77)$$

where $\theta_{n,s} = s\hat{\theta}_n + (1-s)\theta_\star$.

(b) Show that

$$n^{-1/2} \sum_{t=p}^{n} \nabla \ln q^{\theta_\star}(X_{t-p:t-1}; X_t) \stackrel{\mathscr{L}_{\mathbb{P}^{\theta_\star}}}{\Longrightarrow} \mathrm{N}(0, \mathscr{J}(\theta_\star)) .$$

(c) Show that, for all $i, j \in \{1,\dots,d\}$,

$$n^{-1} \sum_{t=p}^{n} \int_0^1 \frac{\partial^2 \ln q^{\theta_{n,s}}}{\partial \theta_i \partial \theta_j}(X_{t-p:t-1}; X_t) \mathrm{d}s \stackrel{\mathbb{P}^{\theta_\star}\text{-prob}}{\longrightarrow} \mathbb{E}^{\theta_\star}\left[\frac{\partial^2 \ln q^{\theta_\star}}{\partial \theta_i \partial \theta_j}(X_{0:p-1}; X_p) \right] .$$

(d) Show that

$$n^{-1} \sum_{t=p}^{n} \left(\int_0^1 \nabla^2 \ln q^{\theta_{n,s}}(X_{t-p:t-1}; X_t) \mathrm{d}s \right) \stackrel{\mathbb{P}^{\theta_\star}\text{-prob}}{\longrightarrow} \mathscr{J}(\theta_\star) .$$

(e) Conclude.

Exercise 3.49 (Autoregressive model with ARCH error). Consider the pth order autoregressive model with ARCH(1) error

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \ldots + \alpha_p \mathrm{Y}_{t-p} + \varepsilon_t , \quad (3.78)$$

where $\varepsilon_t|\mathscr{F}_{t-1} \sim \mathrm{N}(0,\ \beta_0+\beta_1\varepsilon_{t-1}^2)$ and $\{\mathscr{F}_t,\ t\in\mathbb{N}\}$ is the natural filtration of the process $\{Y_t,\ t\in\mathbb{N}\}$, i.e., for any nonnegative integer t, $\mathscr{F}_t=\sigma(Y_s,\ s\le t)$. It is assumed that β_0 is strictly positive and $0\le\beta_1<1$.

(a) Show that the unique strict-sense stationary solution to the recursion $\varepsilon_t=\sqrt{\beta_0+\beta_1\varepsilon_{t-1}^2}Z_t$, where $\{Z_t,\ t\in\mathbb{Z}\}$ is a strong Gaussian noise with zero-mean and unit-variance is given by

$$\varepsilon_t=\beta_0\sum_{l=0}^{\infty}\beta_1^l\left(\prod_{i=0}^{l}Z_{t-i}^2\right). \tag{3.79}$$

(b) Let r be a nonnegative integer. Show that $m_{2r}=\mathbb{E}\left[\varepsilon_t^{2r}\right]$ exists if and only if $\eta_r=\beta_1^r\prod_{j=1}^{r}(2j-1)<1$.

(c) Show that if $3\beta_1^2<1$, then $\mathrm{Var}(\varepsilon_t^2)=2\sigma^4(1-3\beta_1^2)^{-1}$, and $\mathrm{Cov}(\varepsilon_t^2,\ \varepsilon_{t-j}^2)=\beta_1^j\,\mathrm{Var}(\varepsilon_t^2)$ where $\sigma^2=\beta_0(1-\beta_1)^{-1}$ is the stationary unconditional variance.

(d) Show that, if $\eta_r<1$, then

$$n^{-1}\sum_{t=1}^{n}\varepsilon_t^{2r}\overset{\mathbb{P}\text{-a.s.}}{\longrightarrow}\mathbb{E}\left[\varepsilon_t^{2r}\right]\ ,\quad\text{and}\quad n^{-1}\sum_{t=1}^{n}\varepsilon_t^{2r-1}\overset{\mathbb{P}\text{-a.s.}}{\longrightarrow}0\ .$$

(e) Show that if $3\beta_1^2<1$, then for a fixed $j\neq 0$,

$$n^{-1}\sum_{t=1}^{n}\varepsilon_t\varepsilon_{t-j}\overset{\mathbb{P}\text{-a.s.}}{\longrightarrow}0\quad\text{and}\quad n^{-1}\sum_{t=1}^{n}\varepsilon_t^2\varepsilon_{t-j}^2\overset{\mathbb{P}\text{-a.s.}}{\longrightarrow}\mathbb{E}[\varepsilon_t^2\varepsilon_{t-j}^2]\ .$$

(f) Given $(Y_0,\ldots,Y_n)$ we want to estimate $\alpha=(\alpha_1,\ldots,\alpha_p)$ and $\beta=(\beta_0,\beta_1)$. We write $\theta=(\alpha,\beta)$. Show that the log-likelihood conditional on $(Y_p,Y_{p-1},\ldots,Y_0)$ is given by

$$\bar{\mathrm{L}}_n^{\theta}(Y_{p:n}|Y_{0:p-1})=(n-p)^{-1}\sum_{t=p+1}^{n}\ln \mathrm{p}^{\theta}(Y_t|Y_{t-1:t-p-1})$$

with

$$\begin{aligned}&\mathrm{p}^{\theta}(Y_t|Y_{t-1:t-p-1})=(2\pi h_t(\theta))^{-1/2}\exp[-(Y_t-\mathbf{X}_{t-1}'\alpha)^2/2h_t(\theta)]\ ,\\&h_t(\theta)=\beta_0+\beta_1(Y_{t-1}-\mathbf{X}_{t-2}'\alpha)^2\ ,\end{aligned}$$

where $\mathbf{X}_t=(Y_t,\ Y_{t-1},\ \ldots,\ Y_{t-p+1})$.

(g) We assume in the sequel that the model is well-specified. State sufficient conditions upon which the MLE estimator is strongly consistent.

(h) State sufficient conditions upon which the MLE estimator is asymptotically normal, i.e., $n^{1/2}(\hat{\theta}_n-\theta_\star)\overset{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow}\mathrm{N}(0,\ H^{-1}(\theta_\star))$.

(i) Show that the asymptotic covariance matrix $H(\theta_\star)=(H_{i,j}(\theta_\star))_{1\le i,j\le 2}$ is the $(p+2)\times(p+2)$ block matrix with $H_{1,2}(\theta_\star)=H_{2,1}(\theta_\star)=0$,

$$\begin{aligned}H_{1,1}(\theta_\star)=\mathbb{E}^{\theta_\star}[v_t^{-1}(\theta_\star)\mathbf{X}_{t-1}'\mathbf{X}_{t-1}]+2\beta_{\star,1}\mathbb{E}^{\theta_\star}[v_t^{-1}(\theta_\star)\mathbf{X}_{t-2}'\mathbf{X}_{t-2}]\\-2\beta_{\star,1}\beta_{\star,0}\mathbb{E}^{\theta_\star}[v_t^{-2}(\theta_\star)\mathbf{X}_{t-2}'\mathbf{X}_{t-2}]\ ,\end{aligned}$$

where $\{v_t(\theta_\star),t\in\mathbb{Z}\}$ is the unique strict sense stationary non anticipative solution of $v_t(\theta_\star)=\beta_{\star,0}+\beta_{\star,1}v_{t-1}(\theta_\star)Z_{t-1}^2$. Compute similarly $H_{2,2}(\theta_\star)$.

(j) Show that the matrix H may be estimated consistently by replacing the parameters by their estimates and expectations by sample averages in the above expression.

(k) Construct asymptotic confidence regions for the parameters.

Exercise 3.50 (EXPAR(1) model). Consider the EXPAR(1) model (see **??**)

$$X_t=\{\phi_0+\phi_1\exp(-\gamma X_{t-1}^2)\}X_{t-1}+\sigma Z_t\ , \tag{3.80}$$

where $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise with zero mean, unit variance independent of X_0. Assume that the distribution of Z_0 has a continuous density f with respect to the Lebesgue measure, which is everywhere positive and $\mathbb{E}\left[Z_0^{2q}\right]<\infty$ for some $q\in\mathbb{N}_*$. Denote by $\theta=(\phi_0,\phi_1,\gamma,\sigma^2)\in\Theta$ the unknown parameters, where Θ is a compact subset of

$$\left\{(\phi_0,\phi_1,\gamma,\sigma)\,:\,\gamma>0,\sigma^2>0,\max(|\phi_0|,|\phi_0+\phi_1|)<1\right\}\,.$$

For $\theta\in\Theta$, denote $a^\theta(x)=\phi_0+\phi_1\exp(-\gamma x^2)$ and $V_q(x)=1+|x|^{2q}$.

(a) Show that the Markov chain is V_q-geometrically ergodic.
(b) We observe $(X_0,X_1,\ldots,X_n)$ $(n+1)$ observations from the strict sense stationary solution of (3.80) associated to the "true" parameters $\theta_\star\in\Theta$. We fit the parameters of the model using the Gaussian likelihood. Show that for all $(x_0,x_1)\in\mathbb{R}^2$,

$$\ln q^\theta(x_0,x_1)=-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}(x_1-a^\theta(x_0))^2\,.$$

(c) Assume that $q=1$. Show that the sequence of estimators $\hat\theta_n\to_{\mathbb{P}_{\theta_\star}\text{-a.s.}}\theta_\star$.
(d) State conditions upon which the sequence of estimators $\{\hat\theta_n,n\in\mathbb{N}\}$ is asymptotically normal and compute the asymptotic covariance matrix.

Exercise 3.51 (Nonlinear autoregression). Consider the NLAR(1) process

$$X_t=f^\theta(X_{t-1})+Z_t\,, \tag{3.81}$$

where $\{Z_t,\, t\in\mathbb{N}\}$ is a strong white noise independent of X_0 and $\theta\in\Theta$ a compact subset of $\mathbb{R}$. Assume that for all $\theta\in\Theta$, $\lim_{|x|\to\infty}|f^\theta(x)|/|x|<1$, $f^\theta:\mathbb{R}\to\mathbb{R}$ is a twice continuously differentiable function, the distribution of Z_0 has a continuous density with respect to the Lebesgue measure, which is everywhere positive and that $\mathbb{E}\left[Z_0^4\right]<\infty$.

(a) Show that the conditional Gaussian likelihood is given by

$$\ln \mathrm{p}^\theta(X_{p:n}|X_{0:p-1})=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\sum_{t=1}^n\{X_t-f^\theta(X_{t-1})\}^2\,.$$

(b) Assume $\{X_t,\, t\in\mathbb{N}\}$ is strictly stationary such that $\mathbb{E}\left[\sup_{\theta\in\Theta}(X_1-f^\theta(X_0))^2\right]<\infty$. Then, show that the sequence of MLE estimators $\hat\theta_n$ is strongly consistent.
(c) Assume that the model is well-specified and denote by $\theta_\star\in\Theta$ the "true" values of the unknown parameters. Show that the model is identifiable if the condition $f^\theta(X_0)=f^{\theta_\star}(X_0)$ $\mathbb{P}^{\theta_\star}-$a.s. implies that $\theta=\theta_\star$.
(d) Give conditions upon which the sequence of estimators is asymptotically normal, and give the expressions of the asymptotic covariance matrix both in the well-specified and in the misspecified cases.

Exercise 3.52 (Proof of Theorem 3.28).

(a) Show that, under **H** 3.27, for all $\theta\in\Theta$,

$$\begin{aligned}
&\mathbb{E}[\bar\ell^\theta(Y_{-\infty:1})]\\
&\quad=\mathbb{E}\left[\mathbb{E}\left[\ln q(f^\theta_{Y_{-\infty:0}},Y_1)\,\middle|\,Y_s\,,\,s\le 0\right]\right]=\mathbb{E}\left[\int Q^\star(f^{\theta_\star}_{Y_{-\infty:0}},\mathrm{d}y)\ln q(f^\theta_{Y_{-\infty:0}},y)\right]\\
&\quad\le\mathbb{E}\left[\int Q^\star(f^{\theta_\star}_{Y_{-\infty:0}},\mathrm{d}y)\ln q(f^{\theta_\star}_{Y_{-\infty:0}},y)\right]=\mathbb{E}[\bar\ell^{\theta_\star}(Y_{-\infty:1})]\,.
\end{aligned}\tag{3.82}$$

(b) Show that (3.45) implies that $\theta=\theta_\star$.
(c) Conclude.

Exercise 3.53 (Proof of Lemma 3.34). Let $\mathscr{F}$ be the filtration $\mathscr{F} = (\mathscr{F}_n)_{n\in\mathbb{N}}$ where $\mathscr{F}_n = \sigma(Y_s,\ s \le n)$ and let

$$M_n := \sum_{t=1}^{n} \nabla \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}}, Y_t) = \sum_{t=1}^{n} \varphi(\nabla f^{\theta_\star}_{Y_{-\infty:t-1}}, f^{\theta_\star}_{Y_{1:t-1}}(x), Y_t)\ .$$

(a) Show that $\mathbb{E}(\|M_n\|^2) < \infty$ and

$$\mathbb{E}^{\theta_\star}\left[\nabla \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}}, Y_t)\,\middle|\,\mathscr{F}_{t-1}\right] = \nabla f^{\theta_\star}_{Y_{-\infty:t-1}} \int Q^\star(f^{\theta_\star}_{Y_{-\infty:t-1}}, \mathrm{d}y)\frac{\partial \ln q(f^{\theta_\star}_{Y_{-\infty:t-1}}, y)}{\partial x} = 0\ .$$

(b) Show that $\{M_t, t \ge 1\}$ is a square integrable $\mathscr{F}$-martingale with stationary and ergodic increments.
(c) Conclude.

Exercise 3.54 (Proof of Lemma 3.35). Denote by $A_t(\theta) = \frac{\partial^2 \ln q}{\partial\theta_i \partial\theta_j}(f^{\theta}_{Y_{-\infty:t-1}}, Y_t)$.

(a) Show that $n^{-1}\sum_{t=1}^{n} A_t(\theta_\star) \to_{\mathbb{P}} \mathbb{E}\,[A(\theta_\star)]$.
(b) Let $\varepsilon > 0$. Show that we can choose $0 < \eta < \rho$ such that

$$\mathbb{E}\left(\sup_{\theta\in \mathrm{B}(\theta_\star,\eta)} |A_1(\theta_\star) - A_1(\theta)|\right) < \varepsilon\ . \tag{3.83}$$

(c) Show that

$$\limsup_{n\to\infty} \mathbb{P}\left(n^{-1}\sum_{t=1}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \ge \varepsilon,\ \theta_n \in \mathrm{B}(\theta_\star,\eta)\right) \le \limsup_{n\to\infty} \mathbb{P}\left(n^{-1}\sum_{t=1}^{n} \sup_{\theta\in \mathrm{B}(\theta_\star,\eta)} |A_t(\theta_\star) - A_t(\theta)| \ge \varepsilon\right) = 0\ ,$$

(d) Show that

$$\lim_{n\to\infty} \mathbb{P}\left(n^{-1}\sum_{t=1}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \ge \varepsilon\right) = 0\ .$$

(e) Show that $n^{-1}\sum_{t=1}^{n} |A_t(\theta_\star) - A_t(\theta_n)| \overset{\mathbb{P}\text{-prob}}{\longrightarrow} 0$ and conclude.

Exercise 3.55 (Conditional least-squares estimators). Let p be an integer and $\left\{P^\theta :\ \theta = (\varphi,\gamma) \in \Theta\right\}$ be a parametric family of Markov kernels on $\mathsf{X}^p \times \mathscr{X}$, where Θ is a compact subset of $\mathbb{R}^d$ (here, φ is the parameter of interest and γ is a nuisance). Assume that for any $\theta \in \Theta$, P^θ has a unique stationary distribution satisfying $\mathbb{E}^\theta[|X_0|] < \infty$. Let $\theta_\star \in \Theta^o$. Assume that the vector of observations $(X_0, X_1, \ldots, X_n)$ is, for each n a realization of the stationary ergodic Markov chain $\{X_t,\ t \in \mathbb{Z}\}$ with transition kernel $P^{\theta_\star}$. Finally, assume that

- there exists $a^\varphi(X_{p-1},\ldots,X_0)$ a version of the conditional expectation $a^\varphi(X_{p-1},\ldots,X_0) = \mathbb{E}^\theta\left[X_p\middle|X_0,\ldots,X_{p-1}\right]$ with $\theta = (\varphi,\gamma)$ satisfying
$$\mathbb{E}^{\theta_\star}\left[\sup_{\varphi\in\Theta_\varphi} |a^\varphi(X_{p-1},\ldots,X_0)|^2\right] < \infty\ ,$$
where $\Theta_\varphi = \{\varphi, (\varphi,\gamma) \in \Theta\}$ is the canonical projection of Θ.
- for any $(x_0,\ldots,x_{p-1}) \in \mathsf{X}^p$, the function $\varphi \mapsto a^\varphi(x_{p-1},\ldots,x_0)$ is continuous.

We estimate the parameter $\varphi \in \Theta_\varphi$ by the *conditional least-squares*, which amounts to estimating the model by using the likelihood of

$$Y_t = a^\varphi(Y_{t-1},\ldots,Y_{t-p}) + \sigma^2 Z_t \tag{3.84}$$

where $\{Z_t,\, t\in\mathbb{Z}\}$ is a strong white noise with zero-mean and unit-variance Assume in addition that **??** is satisfied and $\mathbb{E}\left[Z_0^{2q}\right]<\infty$ for some $q\in\mathbb{N}_*$.

(a) Show that the log-likelihood for the misspecified model (3.84) is given by

$$\ln q^{\varphi}(x_{p-1:0};X_p)=-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\left\{x_p-a^{\varphi}(x_{p-1},\ldots,x_0)\right\}^2\,.$$

(b) Show that if $a^{\varphi_*}(X_{p-1},\ldots,X_0)=a^{\varphi}(X_{p-1},\ldots,X_0)\ \mathbb{P}_{\theta_\star}-$a.s. implies that $\varphi_=\varphi_*$, then the conditional least square estimator $\{\hat{\varphi}(n),n\in\mathbb{N}\}$ is strongly consistent.

(c) State conditions upon which the conditional least-squares estimator is asymptotically normal and compute its asymptotic covariance matrix.

Exercise 3.56. Consider an EXPAR(p) model

$$X_t=\{a_1+b_1\exp(-\lambda_1X_{t-d}^2)\}X_{t-1}+\ldots \\ +\{a_p+b_p\exp(-\lambda_pX_{t-p}^2)\}X_{t-p}+\sigma Z_t\,, \quad (3.85)$$

where $\{Z_t,\, t\in\mathbb{Z}\}$ is a strong white noise with zero-mean and unit-variance Assume in addition that **??** is satisfied and $\mathbb{E}\left[Z_0^{2q}\right]<\infty$ for some $q\in\mathbb{N}_*$. . We denote by $\theta=\{(a_i,b_i,\lambda_i),i=1,\ldots,p,d,\sigma^2\}\in\Theta$ the unknown parameters, where Θ is a compact subset of the set $\{(a_i,b_i,\lambda_i),i=1,\ldots,p,d,\sigma^2\}$ satisfying $\sigma^2>0$, $d\in\{0,\ldots,p\}$, $\lambda_i>0$, $i\in\{1,\ldots,p\}$ and $c(z)=1-c_1z-\cdots-c_pz^p$ where $c_i=\max(|a_i|,|a_i+b_i|)$. For $\theta\in\Theta$, the Markov chain $\mathbf{X}_t=[X_t,\ldots,X_{t-p+1}]'$ given by (3.85) is V_q-geometrically ergodic with $V_q(\mathbf{x})=1+\|\mathbf{x}\|^q$; see **??** and **??**. Denote by $\theta_\star\in\Theta$ the true parameter vector. Given the observations $(X_0,\ldots,X_n)$, we estimate the parameters using the conditional least square (Gaussian likelihood)

$$-\frac{n-p}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=p}^{n}(X_t-a^{\theta}(X_{t-1},\ldots,X_{t-p}))^2\,,$$

where

$$a^{\theta}(x_{p-1},\ldots,x_0)=\sum_{i=1}^{p}\{a_i+b_i\exp(-\lambda_ix_{p-i}^2)\}x_{p-i}\,.$$

(a) Compute the conditional least squares estimator of θ.

(b) Assume that $(X_0,X_1,\ldots,X_n)$ is a sample record of $\{X_t,\, t\in\mathbb{N}\}$ which is the ergodic solution of (3.85) associated to the true parameter vector $\theta_\star$. Show that if $q=1$, then $\hat{\theta}_n$ is strongly consistent.

(c) Show that if $q=2$, this estimator is asymptotically normal and compute the limiting covariance.

Exercise 3.57 (First order RCA; ?? cont.). We use the notations of **??**. We denote by $\theta=(\phi,\lambda,\sigma)$ the unknown parameters. We assume that the noise Z_t cannot take on only two values asymptotically.

(a) Show that $\mathbb{E}^{\theta}\left[X_t|\,\mathscr{F}_{t-1}^X\right]=\phi X_{t-1}$ and $\mathrm{Var}\,(\,[|\,\theta)]X_t\mathscr{F}_{t-1}^X=\lambda^2X_{t-1}^2+\sigma^2$, $\mathbb{P}^{\theta}-$a.s..

(b) Show that if $\{B_t,\, t\in\mathbb{Z}\}$ and $\{Z_t,\, t\in\mathbb{Z}\}$ are jointly Gaussian, then the conditional distribution of X_t given $\mathscr{F}_{t-1}^X$ is Gaussian with mean $\mathbb{E}^{\theta}\left[X_t|\,\mathscr{F}_{t-1}^X\right]$ and variance $\mathrm{Var}\,(\,[|\,\theta)]X_t\mathscr{F}_{t-1}^X$. Write the likelihood of this model (referred to as in the sequel as the Gaussian likelihood).

(c) Assume that $\theta_\star\in\Theta$, where Θ is a compact subset of

$$\left\{(\phi,\lambda,\sigma)\in\mathbb{R}\times\mathbb{R}_+^2\,:\,\phi^2+\lambda^2<1,\sigma>0\right\}\,.$$

Assume that the observations $(X_0,\ldots,X_n)$ is a sample record from the strict-sense stationary solution $\{X_t,\, t\in\mathbb{N}\}$ of (**??**) associated to some values of the parameters $\theta_\star\in\Theta$. Prove that the Gaussian MLE $\{\hat{\phi}_n\}_{n\geq1}$ is a strongly consistent sequence of estimator of ϕ_*

(d) Under which conditions is the Gaussian MLE $\{(\hat{\lambda}_n,\hat{\sigma}_n^2)\}_{n\geq0}$ a strongly consistent sequence of estimators of (λ,σ^2)?

(e) Assume that $\theta_\star\in\Theta^o$, where Θ is a compact subset of

$$\left\{(\phi,\lambda,\sigma)\in\mathbb{R}\times\mathbb{R}_+^2\,:\,\phi^4+6\lambda^2\phi^2+m_4\lambda^4<1,\sigma>0\right\}\,.$$

Prove that the Gaussian MLE $\{\hat{\phi}_n\}_{n\geq 0}$ is an asymptotically normal sequence of estimators of ϕ_*. Compute a confidence interval for this parameter.

(f) Under which conditions is the sequence $\{(\hat{\lambda}_n,\hat{\sigma}_n^2)\}_{n\geq 0}$ asymptotically normal?

Exercise 3.58. Consider the stochastic process

$$X_t = A_t X_{t-1} + B_t\,, \tag{3.86}$$

where $\{(A_t,B_t),\, t\in\mathbb{N}\}$ is a Gaussian white noise, independent of X_0, with mean zero and covariance matrix

$$\begin{pmatrix} \alpha^2 & \rho\alpha\beta \\ \rho\alpha\beta & \beta^2 \end{pmatrix}.$$

We denote by $\theta=(\alpha^2,\beta^2,\rho)\in\Theta$ the unknown parameters, where Θ is a compact subset of $(0,1)\times\mathbb{R}_*^+\times(-1,1)$.

(a) Show that for any $\theta\in\Theta$, (3.86) admits a unique strict sense stationary solution, satisfying $\mathbb{E}^\theta[X_0^2]<\infty$.
(b) Find a necessary and sufficient condition on θ under which (3.86) admits a unique strict sense stationary solution satisfying $\mathbb{E}^\theta[X_0^4]<\infty$.
(c) Write the conditional maximum likelihood estimator for the model (3.86).
(d) Write an algorithm in R to solve numerically this equation.
(e) Assume that the observation $(X_0,\ldots,X_n)$ is a realization of (3.86) for some parameter $\theta_\star\in\Theta$. Prove that $\hat{\theta}_n$ is a strongly consistent sequence of estimators of $\theta_\star$.
(f) State conditions upon which this estimator is asymptotically normal and compute the asymptotic covariance matrix.

Exercise 3.59 (APARCH(1,1); ?? cont.). The APARCH(1,1) model of **?** can be defined as follows:

$$Y_t = \sigma_t(\theta)\varepsilon_t \tag{3.87}$$

$$\sigma_t^\delta(\theta) = \alpha_0+\alpha_1(Y_{t-1}-\gamma Y_{t-1})^\delta+\beta_1\sigma_{t-1}^\delta(\theta)\,, \tag{3.88}$$

where $\{\varepsilon_t,\, t\in\mathbb{N}\}$ is a strong white Gaussian noise with zero-mean and unit variance and $\theta=(\alpha_0,\alpha_1,\beta_1,\gamma,\delta)\in\Theta$ a compact set of

$$\{(\alpha_0,\alpha_1,\beta_1,\gamma,\delta)\,:\,\alpha_0>0,\alpha_1>0,0<\gamma<1,\delta>0\}\,.$$

The parameter δ, ($\delta>0$) parameterizes a Box-Cox transformation of the conditional standard deviation $\sigma_t(\theta)$, while the parameters γ reflect the leverage effect.

(a) Show that the APARCH model is an observation-driven model. Determine the kernel q^θ and the function f_y^θ.
(b) Show that the maximum likelihood estimator is consistent.
(c) Compute the confidence intervals for the parameters.

Chapter 4

Markov chain Monte Carlo methods

Contents

Let π be a given probability measure on $(\mathsf{X}, \mathscr{X})$. The aim consists in constructing a Markov chain with invariant distribution π. To do so, we will construct Markov kernel P that are π-reversible. We first define π-reversibility and then show that if P is π-reversible then it admits π as invariant distribution.

Definition 4.1. Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$ and π a probability measure on $(\mathsf{X}, \mathscr{X})$. Then P is π-reversible iff for all $h \in \mathbb{F}_+(\mathsf{X} \times \mathsf{X})$,

$$\iint h(x,y)\pi(\mathrm{d}x)Q(x,\mathrm{d}y) = \iint h(x,y)\pi(\mathrm{d}y)Q(y,\mathrm{d}x)$$

Using infinitesimal notation (which is an abuse of notation), we can also say: P is π-reversible iff $\pi(\mathrm{d}x)Q(x,\mathrm{d}y) = \pi(\mathrm{d}y)Q(y,\mathrm{d}x)$.

Lemma 4.2 *If P is reversible with respect to π, then π is a stationary distribution for P.*

PROOF. Setting $f \equiv 1$, we indeed obtain for any function $g \in \mathbb{F}_+(\mathsf{X})$,

$$\iint \pi(\mathrm{d}x)P(x,\mathrm{d}y)g(y) = \iint \pi(\mathrm{d}x)P(x,\mathrm{d}y)g(x) = \int \pi(\mathrm{d}x)g(x)\,.$$

■

4.1 Metropolis-Hastings algorithm

Markov chain Monte Carlo is a general method for the simulation of distributions known up to a multiplicative constant. Let ν be a σ-finite measure on a state space $(\mathsf{X}, \mathscr{X})$ and let $h_\pi \in \mathbb{F}_+(\mathsf{X})$ such that $0 < \int_{\mathsf{X}} h_\pi(x)\nu(\mathrm{d}x) < \infty$. Typically X is an open subset of $\mathbb{R}^d$ and ν is the Lebesgue measure or X is countable and ν is the counting measure. This function is associated to a probability measure π on X defined by

$$\pi(A) = \frac{\int_A h_\pi(x)\nu(\mathrm{d}x)}{\int_{\mathsf{X}} h_\pi(x)\nu(\mathrm{d}x)} \,. \tag{4.1}$$

We want to approximate expectations of functions $f \in \mathbb{F}_+(\mathsf{X})$ with respect to π

$$\pi(f) = \frac{\int_{\mathsf{X}} f(x)h_\pi(x)\nu(\mathrm{d}x)}{\int_{\mathsf{X}} h_\pi(x)\nu(\mathrm{d}x)} \,.$$

If the state space X is high-dimensional and h_π is complex, direct numerical integration is not an option. The classical Monte Carlo solution to this problem is to simulate i.i.d. random variables $Z_0, Z_1, \ldots, Z_{n-1}$ with distribution π and then to estimate $\pi(f)$ by the sample mean

$$\hat{\pi}(f) = n^{-1} \sum_{i=0}^{n-1} f(Z_i) \,. \tag{4.2}$$

This gives an unbiased estimate with standard deviation of order $O(n^{-1/2})$ provided that $\pi(f^2) < \infty$. Furthermore, by the Central Limit Theorem, the normalized error $\sqrt{n}(\hat{\pi}(f) - \pi(f))$ has a limiting normal distribution, so that confidence intervals are easily obtained.

The problem often encountered in applications is that it might be very difficult to simulate i.i.d. random variables with distribution π. Instead, the Markov chain Monte Carlo (MCMC) solution is to construct a Markov chain on X which has π as invariant probability. The hope is that regardless of the initial distribution ξ, the law of large numbers will hold, i.e. $\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k) = \pi(f)$ $\mathbb{P}_\xi$ – a.s.

At first sight, it may seem even more difficult to find such a Markov chain than to estimate $\pi(f)$ directly. In the following subsections, we will exhibit several such constructions.

4.1.1 Metropolis-Hastings algorithms

Let Q be a Markov kernel having a density q with respect to ν i.e. $Q(x,A) = \int_A q(x,y)\nu(\mathrm{d}y)$ for every $x \in \mathsf{X}$ and $A \in \mathscr{X}$.

The Metropolis-Hastings algorithm proceeds in the following way. An initial starting value X_0 is chosen. Given X_k, a candidate move Y_{k+1} is sampled from $Q(X_k, \cdot)$. With probability $\alpha(X_k, Y_{k+1})$, it is accepted and the chain moves to $X_{k+1} = Y_{k+1}$. Otherwise the move is rejected and the chain remains at $X_{k+1} = X_k$. The probability $\alpha(X_k, Y_{k+1})$ of accepting the move is given by

$$\alpha(x,y) = \begin{cases} \min\left(\frac{h_\pi(y)}{h_\pi(x)}\frac{q(y,x)}{q(x,y)}, 1\right) & \text{if } h_\pi(x)q(x,y) > 0 \,, \\ 1 & \text{if } h_\pi(x)q(x,y) = 0 \,. \end{cases} \tag{4.3}$$

The acceptance probability $\alpha(x,y)$ only depends on the ratio $h_\pi(y)/h_\pi(x)$; therefore, we only need to know h_π up to a normalizing constant. In Bayesian inference, this property plays a crucial role.

This procedure produces a Markov chain, $\{X_k,\ k \in \mathbb{N}\}$, with Markov kernel P given by

$$P(x,A) = \int_A \alpha(x,y)q(x,y)\nu(\mathrm{d}y) + \bar{\alpha}(x)\delta_x(A) \,, \tag{4.4}$$

with

$$\bar{\alpha}(x) = \int_{\mathsf{X}} \{1 - \alpha(x,y)\} q(x,y) \nu(\mathrm{d}y) \,. \tag{4.5}$$

The quantity $\bar{\alpha}(x)$ is the probability of remaining at the same point.

Proposition 4.3 *The distribution π is reversible with respect to the Metropolis-Hastings kernel P.*

PROOF. Note first that for every $x, y \in \mathsf{X}$, it holds that

$$\begin{aligned} h_\pi(x)\alpha(x,y)q(x,y) &= \{h_\pi(x)q(x,y)\} \wedge \{h_\pi(y)q(y,x)\} \\ &= h_\pi(y)\alpha(y,x)q(y,x) \,. \end{aligned} \tag{4.6}$$

Thus for $C \in \mathscr{X} \times \mathscr{X}$,

$$\begin{aligned} &\iint h_\pi(x)\alpha(x,y)q(x,y)\mathbb{1}_C(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) \\ &\qquad = \iint h_\pi(y)\alpha(y,x)q(y,x)\mathbb{1}_C(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) \,. \end{aligned} \tag{4.7}$$

On the other hand,

$$\begin{aligned} &\iint h_\pi(x)\delta_x(\mathrm{d}y)\bar{\alpha}(x)\mathbb{1}_C(x,y)\nu(\mathrm{d}x) \\ &\quad = \int h_\pi(x)\bar{\alpha}(x)\mathbb{1}_C(x,x)\nu(\mathrm{d}x) = \int h_\pi(y)\bar{\alpha}(y)\mathbb{1}_C(y,y)\nu(\mathrm{d}y) \\ &\quad = \iint h_\pi(y)\delta_y(\mathrm{d}x)\bar{\alpha}(y)\mathbb{1}_C(x,y)\nu(\mathrm{d}y) \,. \end{aligned} \tag{4.8}$$

Hence, summing (4.7) and (4.8) we obtain

$$\iint h_\pi(x)P(x,\mathrm{d}y)\nu(\mathrm{d}x)\mathbb{1}_C(x,y) = \iint h_\pi(y)P(y,\mathrm{d}x)\mathbb{1}_C(x,y)\nu(\mathrm{d}y) \,.$$

This proves that π is reversible with respect to P. ■

From Lemma 4.2, we obtain that π is an invariant probability for the Markov kernel P.

Example 4.4 (Random walk Metropolis algorithm). This is a particular case of the Metropolis-Hasting algorithm, where the proposal transition density is symmetric, i.e. $q(x,y) = q(y,x)$, for every $(x,y) \in \mathsf{X} \times \mathsf{X}$. Furthermore, assume that $\mathsf{X} = \mathbb{R}^d$ and let $\bar{q}$ be a symmetric density with respect to 0, i.e. $\bar{q}(-y) = \bar{q}(y)$ for all $y \in \mathsf{X}$. Consider the transition density q defined by $q(x,y) = \bar{q}(y-x)$. This means that if the current state is X_k, an increment Z_{k+1} is drawn from $\bar{q}$ and the candidate $Y_{k+1} = X_k + Z_{k+1}$ is proposed.

The acceptance probability (4.3) for the random walk Metropolis algorithm is given by

$$\alpha(x,y) = 1 \wedge \frac{h_\pi(y)}{h_\pi(x)} \,. \tag{4.9}$$

If $h_\pi(Y_{k+1}) \geq h_\pi(X_k)$, then the move is accepted with probability one and if $h_\pi(Y_{k+1}) < h_\pi(X_k)$, then the move is accepted with a probability strictly less than one.

The choice of the incremental distribution is crucial for the efficiency of the algorithm. A classical choice for $\bar{q}$ is the multivariate normal distribution with zero-mean and covariance matrix Γ to be suitably chosen.

Example 4.5 (Independent Metropolis-Hastings sampler). Another possibility is to set the transition density to be $q(x,y) = \bar{q}(y)$, where $\bar{q}$ is a density on X. In this case, the next candidate is drawn independently of the current state of the chain. This yields the so-called independent sampler, which is closely related to the accept-reject algorithm for random variable simulation.

The acceptance probability (4.3) is given by

$$\alpha(x,\, y) = 1 \wedge \frac{h_\pi(y)\bar{q}(x)}{h_\pi(x)\bar{q}(y)} \,. \tag{4.10}$$

Candidate steps with a low weight $\bar{q}(Y_{k+1})/\pi(Y_{k+1})$ are rarely accepted, whereas candidates with a high weight are very likely to be accepted. Therefore the chain will remain at these states for several steps with a high probability, thus increasing the importance of these states within the constructed sample.

Assume for example that h is the standard Gaussian density and that q is the density of the Gaussian distribution with zero mean and variance σ^2, so that $q(x) = h(y/\sigma)/\sigma$. Assume that $\sigma^2 > 1$ so that the values being proposed are sampled from a distribution with heavier tails than the objective distribution h. Then the acceptance probability is

$$\alpha(x,y) = \begin{cases} 1 & |y| \leq |x| \,, \\ \exp(-(y^2-x^2)(1-\sigma^{-2})/2) & |y| > |x| \,. \end{cases}$$

Thus the algorithm accepts all moves which decrease the norm of the current state but only some of those which increase it.

If $\sigma^2 < 1$, the values being proposed are sampled from a lighter tailed distribution than h and the acceptance probability becomes

$$\alpha(x,y) = \begin{cases} \exp(-(x^2-y^2)(1-\sigma^{-2})/2) & |y| \leq |x| \,, \\ 1 & |y| > |x| \,. \end{cases}$$

It is natural to inquire whether heavy-tailed or light-tailed proposal distributions should be preferred. This question will be partially answered in **??**.

Example 4.6 (Langevin diffusion). More sophisticated proposals can be considered. Assuming that $x \mapsto \log h_\pi(x)$ is everywhere differentiable. Consider the Langevin diffusion defined by the stochastic differential equation (SDE)

$$\mathrm{d}X_t = \frac{1}{2}\nabla \log h_\pi(X_t)\mathrm{d}t + \mathrm{d}W_t \,,$$

where $\nabla \log h_\pi$ denotes the gradient of $\log h_\pi$. Under appropriate conditions, the Langevin diffusion has a stationary distribution with densityt h_π and is reversible. Assume that the proposal state Y_{k+1} in the Metropolis-Hastings algorithm corresponds to the Euler discretization of the Langevin SDE for some step size h:

$$Y_{k+1} = X_k + \frac{\gamma}{2}\nabla \log h_\pi(X_k) + \sqrt{\gamma} Z_{k+1} \,, \qquad Z_{k+1} \sim \mathrm{N}(0,\mathrm{I}) \,.$$

Such algorithms are known as Langevin Metropolis-Hastings algorithms. The gradient can be approximated numerically via finite differences and does not require knowledge of the normalizing constant of the target distribution h_π.

4.1.2 Data augmentation

Throughout this section, $(\mathsf{X}, \mathscr{X})$ and $(\mathsf{Y}, \mathscr{Y})$ are Polish spaces equipped with their Borel σ-fields. Again, we wish to simulate from a probability measure π defined on $(\mathsf{X}, \mathscr{X})$ using a sequence $\{X_k,\, k \in \mathbb{N}\}$ of X-valued random variables. Data augmentation algorithms consist in writing the target distribution π as the marginal of the distribution π^* on the product space $(\mathsf{X}\times\mathsf{Y}, \mathscr{X}\otimes\mathscr{Y})$ defined by $\pi^* = \pi \otimes R$ where R is a kernel on $\mathsf{X}\times\mathscr{Y}$. By **??**, there exists also a kernel S on $\mathsf{Y}\times\mathscr{X}$ and a probability measure $\tilde{\pi}$ on $(\mathsf{Y}, \mathscr{Y})$ such that $\pi^*(C) = \iint \mathbb{1}_C(x,y)\tilde{\pi}(\mathrm{d}y)S(y,\mathrm{d}x)$ for $C \in \mathscr{X}\otimes\mathscr{Y}$. In other words, if (X,Y) is a pair of random variables with distribution π^*, then $R(x,\cdot)$ is the distribution of Y conditionally on $X = x$ and $S(y,\cdot)$ is the distribution of X conditionally on $Y = y$. The bivariate distribution π^* can then be expressed as follows

$$\pi^*(\mathrm{d}x\mathrm{d}y) = \pi(\mathrm{d}x)R(x,\mathrm{d}y) = S(y,\mathrm{d}x)\tilde{\pi}(\mathrm{d}y) \,. \tag{4.11}$$

A data augmentation algorithm consists in running a Markov Chain $\{(X_k,Y_k),\, k\in\mathbb{N}\}$ with invariant probability π^* and to use $n^{-1}\sum_{k=0}^{n-1} f(X_k)$ as an approximation of $\pi(f)$. A significant difference between this general approach and a Metropolis-Hastings algorithm associated to the target distribution π is that $\{X_k,\, k\in\mathbb{N}\}$ is no longer constrained to be a Markov chain. The transition from (X_k,Y_k) to (X_{k+1},Y_{k+1}) is decomposed into two successive steps: Y_{k+1} is first drawn given (X_k,Y_k) and then X_{k+1} is drawn given (X_k,Y_{k+1}). Intuitively, Y_{k+1} can be used as an auxiliary variable, which directs the moves of X_k toward interesting regions with respect to the target distribution.

When sampling from R and S is feasible, a classical choice consists in following the two successive steps: given (X_k,Y_k),

(i) sample Y_{k+1} from $R(X_k,\cdot)$,
(ii) sample X_{k+1} from $S(Y_{k+1},\cdot)$.

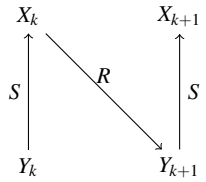

Fig. 4.1 In this example, sampling from R and S is feasible.

It turns out that $\{X_k,\, k\in\mathbb{N}\}$ is a Markov chain with Markov kernel RS and π is reversible with respect to RS.

Lemma 4.7 *The distribution π is reversible with respect to the kernel RS.*

PROOF. By **??**, we must prove that the measure $\pi\otimes RS$ on X^2 is symmetric. For $A,B\in\mathscr{X}$, applying (4.11), we have

$$\begin{aligned}
&\pi\otimes RS(A\times B)\\
&\quad=\int_{\mathsf{X}\times\mathsf{Y}}\pi(\mathrm{d}x)R(x,\mathrm{d}y)\mathbb{1}_A(x)S(y,B)=\int_{\mathsf{X}\times\mathsf{Y}}\mathbb{1}_A(x)S(y,B)\pi^*(\mathrm{d}x\mathrm{d}y)\\
&\quad=\int_{\mathsf{X}\times\mathsf{Y}}\mathbb{1}_A(x)S(y,B)S(y,\mathrm{d}x)\tilde{\pi}(\mathrm{d}y)=\int_{\mathsf{Y}}S(y,A)S(y,B)\tilde{\pi}(\mathrm{d}y)\,.
\end{aligned}$$

This proves that $\pi\otimes RS$ is symmetric. ■

Assume now that sampling from R or S is infeasible. In this case, we consider two instrumental kernels Q on $(\mathsf{X}\times\mathsf{Y})\times\mathscr{Y}$ and T on $(\mathsf{X}\times\mathsf{Y})\times\mathscr{X}$ which will be used to propose successive candidates for Y_{k+1} and X_{k+1}. For simplicity, assume that $R(x,\mathrm{d}y')$ and $Q(x,y;\mathrm{d}y')$ (resp. $S(y',\mathrm{d}x')$ and $T(x,y';\mathrm{d}x')$) are dominated by the same measure and call r and q (resp. s and t) the associated transition densities. We assume that r and s are known up to a normalizing constant. Define the Markov chain $\{(X_k,Y_k),\, k\in\mathbb{N}\}$ as follows. Given $(X_k,Y_k)=(x,y)$,

(DA1) draw a candidate $\tilde{Y}_{k+1}$ according to the distribution $Q(x,y;\cdot)$ and accept $Y_{k+1}=\tilde{Y}_{k+1}$ with probability $\alpha(x,y,\tilde{Y}_{k+1})$ defined by

$$\alpha(x,y,y')=\frac{r(x,y')q(x,y';y)}{r(x,y)q(x,y;y')}\wedge 1\,;$$

otherwise, set $Y_{k+1}=Y_k$; the Markov kernel on $\mathsf{X}\times\mathsf{Y}\times\mathscr{Y}$ associated to this transition is denoted by K_1;

(DA2) draw then a candidate $\tilde{X}_{k+1}$ according to the distribution $T(x,Y_{k+1};\cdot)$ and accept $X_{k+1}=\tilde{X}_{k+1}$ with probability $\beta(x,Y_{k+1},\tilde{X}_{k+1})$ defined by

$$\beta(x,y,x')=\frac{s(y,x')t(x',y;x)}{s(y,x)t(x,y;x')}\wedge 1\,;$$

otherwise, set $X_{k+1} = X_k$; the Markov kernel on $\mathsf{X} \times \mathsf{Y} \times \mathscr{X}$ associated to this transition is denoted by K_2.

For $i = 1, 2$, let K_i^* be the kernels associated to K_1 and K_2 as follows: for $x \in \mathsf{X}$, $y \in \mathsf{Y}$, $A \in \mathscr{X}$ and $B \in \mathscr{Y}$,

$$K_1^*(x, y; A \times B) = \mathbb{1}_A(x) K_1(x, y; B) \,. \tag{4.12}$$

$$K_2^*(x, y; A \times B) = \mathbb{1}_B(y) K_2(x, y; A) \,. \tag{4.13}$$

Then, the kernel of the chain $\{(X_n, Y_n),\, n \in \mathbb{N}\}$ is $K = K_1^* K_2^*$. The process $\{X_n,\, n \in \mathbb{N}\}$ is in general not a Markov chain since the distribution of X_{k+1} conditionally on (X_k, Y_k) depends on (X_k, Y_k) and on X_k only, except in some special cases. Obviously, this construction includes the previous one where sampling from R and S was feasible. Indeed, if $Q(x, y; \cdot) = R(x, \cdot)$ and $T(x, y; \cdot) = S(x, \cdot)$, then the acceptance probabilities α and β defined above simplify to one, the candidates are always accepted and we are back to the previous algorithm.

Proposition 4.8 *The extended target distribution π^* is reversible with respect to the kernels K_1^* and K_2^* and invariant with respect to K.*

PROOF. The reversibility of π^* with respect to K_1^* and K_2^* implies its invariance and consequently its invariance with respect to the product $K = K_1^* K_2^*$. Let us prove the reversibility of π^* with respect to K_1^*. For each $x \in \mathsf{X}$, the kernel $K_1(x, \cdot; \cdot)$ on $\mathsf{Y} \times \mathscr{Y}$ is the kernel of a Metropolis-Hastings algorithm with target density $r(x, \cdot)$, proposal kernel density $q(x, \cdot; \cdot)$ and acceptation probability $\alpha(x, \cdot, \cdot)$. By Proposition 4.3, this implies that the distribution $R(x, \cdot)$ is reversible with respect to the kernel $K_1(x, \cdot; \cdot)$. Applying the definition (4.12) of K_1^* and $\pi^* = \pi \otimes R$ yields, for $A, C \in \mathscr{X}$ and $B, D \in \mathscr{Y}$,

$$\begin{aligned}
\pi^* \otimes K_1^*(A \times B \times C \times D) &= \iint_{A \times B} \pi(\mathrm{d}x \mathrm{d}y) K_1^*(x, y; C \times D) \\
&= \iint_{A \times B} \pi(\mathrm{d}x) R(x, \mathrm{d}y) \mathbb{1}_C(x) K_1(x, y, D) \\
&= \int_{A \cap C} \pi(\mathrm{d}x) [R(x, \cdot) \otimes K_1(x, \cdot; \cdot)](B \times D) \,.
\end{aligned}$$

We have seen that for each $x \in \mathsf{X}$, the measure $R(x, \cdot) \otimes K_1(x, \cdot; \cdot)$ is symmetric, thus $\pi^* \otimes K^*$ is also symmetric. The reversibility of π^* with respect to K_2^* is proved similarly. ■

Example 4.9 (The slice sampler). Set $\mathsf{X} = \mathbb{R}^d$ and $\mathscr{X} = \mathscr{B}(\mathsf{X})$. Let μ be a σ-finite measure on $(\mathsf{X}, \mathscr{X})$ and let h be the density with respect to μ of the target distribution. We assume that for all $x \in \mathsf{X}$,

$$h(x) = C \prod_{i=0}^{k} f_i(x) \,,$$

where C is a constant (which is not necessarily known) and $f_i : \mathbb{R}^d \to \mathbb{R}_+$ are nonnegative measurable functions. For $y = (y_1, \ldots, y_k) \in \mathbb{R}_+^k$, define

$$L(y) = \left\{ x \in \mathbb{R}^d \,:\, f_i(x) \geq y_i \,,\; i = 1, \ldots, k \right\} \,.$$

The f_0-slice-sampler algorithm proceeds as follows:

- given X_n, draw independently a k-tuple $Y_{n+1} = (Y_{n+1,1}, \ldots, Y_{n+1,k})$ of independent random variables such that $Y_{n+1,i} \sim \mathrm{Unif}(0, f_i(X_n))$, $i = 1, \ldots, k$.
- sample X_{n+1} from the distribution with density proportional to $f_0 \mathbb{1}_{L(Y_{n+1})}$.

Set $\mathsf{Y} = \mathbb{R}_+^k$ and for $(x, y) \in \mathsf{X} \times \mathsf{Y}$,

$$h^*(x, y) = C f_0(x) \mathbb{1}_{L(y)}(x) = h(x) \prod_{i=1}^{k} \frac{\mathbb{1}_{[0, f_i(x)]}(y_i)}{f_i(x)} \,.$$

Let π^* be the probability measure with density h^* with respect to Lebesgue's measure on $\mathsf{X}\times\mathsf{Y}$. Then $\int_{\mathsf{Y}} h^*(x,y)\mathrm{d}y = h(x)$ i.e. π is the first marginal of π^*. Let R be the kernel on $\mathsf{X}\times\mathscr{Y}$ with kernel denisty r defined by

$$r(x,y) = \frac{h^*(x,y)}{h(x)}\mathbb{1}\{h(x)>0\} \ .$$

Then $\pi^* = \pi\otimes R$. Define the distribution $\tilde{\pi} = \pi R$, its density $\tilde{h}(y) = \int_{\mathsf{X}} h^*(u,y)\mathrm{d}u$ and the kernel S on $\mathsf{Y}\times\mathscr{X}$ with density s by

$$s(y,x) = \frac{h^*(x,y)}{\tilde{h}(y)}\mathbb{1}\left\{\tilde{h}(y)>0\right\} \ .$$

If (X,Y) is a vector with distribution π^*, then $S(y,\cdot)$ is the conditional distribution of X given $Y=y$ and the Markov kernel of the chain $\{X_n,\ n\in\mathbb{N}\}$ is RS and Lemma 4.7 can be applied to prove that π is reversible, hence invariant, with respect to RS.

4.1.3 Two-stage Gibbs sampler

The Gibbs sampler is a simple method which decomposes a complex multidimensional distribution into a collection of smaller dimensional ones. Let $(\mathsf{X},\mathscr{X})$ and $(\mathsf{Y},\mathscr{Y})$ be complete separable metric spaces endowed with their Borel σ-fields. To construct the Markov chain $\{(X_n,Y_n),\ n\in\mathbb{N}\}$ with π^* as an invariant probability, we proceed exactly as in data-augmentation algorithms. Assume that π^* may be written as

$$\pi^*(\mathrm{d}x\mathrm{d}y) = \pi(\mathrm{d}x)R(x,\mathrm{d}y) = \tilde{\pi}(\mathrm{d}y)S(y,\mathrm{d}x) \tag{4.14}$$

where π and $\tilde{\pi}$ are probability measures on X and Y respectively and R and S are kernels on $\mathsf{X}\times\mathscr{Y}$ and $\mathsf{Y}\times\mathscr{X}$ respectively.

The deterministic updating (two-stage) Gibbs (DUGS) sampler

When sampling from R and S is feasible, the DUGS sampler proceeds as follows: given (X_k,Y_k),

(DUGS1) sample Y_{k+1} from $R(X_k,\cdot)$,
(DUGS2) sample X_{k+1} from $S(Y_{k+1},\cdot)$.

For both the Data Augmentation algorithms and the two-stage Gibbs sampler we consider a distribution π^* on the product space $\mathsf{X}\times\mathsf{Y}$. In the former case, the distribution of interest is a marginal distribution of π^* and in the latter case the target distribution is π^* itself.

We may associate to each update (DUGS1)-(DUGS2) of the algorithm a transition kernel on $(\mathsf{X}\times\mathsf{Y})\times(\mathscr{X}\otimes\mathscr{Y})$ defined for $(x,y)\in\mathsf{X}\times\mathsf{Y}$ and $A\times B\in\mathscr{X}\otimes\mathscr{Y}$ by

$$R^*(x,y;A\times B) = \mathbb{1}_A(x)R(x,B) \ , \tag{4.15}$$
$$S^*(x,y;A\times B) = \mathbb{1}_B(y)S(y,A) \ . \tag{4.16}$$

The transition kernel of the DUGS is then given by

$$P_{\mathrm{DUGS}} = R^*S^* \ . \tag{4.17}$$

Note that for $A\times B\in\mathscr{X}\otimes\mathscr{Y}$,

$$P_{\mathrm{DUGS}}(x,y;A\times B) = \iint_{\mathsf{X}\times\mathsf{Y}} R^*(x,y;\mathrm{d}x'\mathrm{d}y')S^*(x',y';A\times B)$$
$$= \iint_{\mathsf{X}\times\mathsf{Y}} R(x,\mathrm{d}y')\mathbb{1}_B(y')S(y',A)$$
$$= \int_B R(x,\mathrm{d}y')S(y',A) = R\otimes S(x,B\times A)\,. \tag{4.18}$$

As a consequence of Proposition 4.8, we obtain the invariance of π^*.

Lemma 4.10 *The distribution π^* is reversible with respect to the kernels R^* and S^* and invariant with respect to P_{DUGS}.*

The Random Scan Gibbs sampler (RSGS)

At each iteration, the RSGS algorithm consists in updating one component chosen at random. It proceeds as follows: given (X_k,Y_k),

(RSGS1) sample a Bernoulli random variable B_{k+1} with probability of success $1/2$.
(RSGS2) If $B_{k+1}=0$, then sample Y_{k+1} from $R(X_k,\cdot)$ else sample X_{k+1} from $S(Y_{k+1},\cdot)$.

The transition kernel of the RSGS algorithm can be written

$$P_{\mathrm{RSGS}} = \frac{1}{2}R^* + \frac{1}{2}S^*\,. \tag{4.19}$$

Lemma 4.10 implies that P_{RSGS} is reversible with respect to π^* and therefore π^* is invariant for P_{RSGS}.

If sampling from R or S is infeasible, the Gibbs transitions can be replaced by a Metropolis-Hastings algorithm on each component as in the case of the DUGS algorithm. The agorithm is then called the Two-Stage Metropolis-within-Gibbs algorithm.

Example 4.11 (Scalar Normal-Inverse Gamma). In a statistical problem one may be presented with a set of independent observations $\mathbf{y}=\{y_1,\,\ldots,\,y_n\}$, assumed to be normally distributed, but with unknown mean μ and variance τ^{-1} (τ is often referred to as the precision). The Bayesian approach to this problem is to assume that μ and τ are themselves random variables, with a given prior distribution. For example, we might assume that

$$\mu\sim \mathrm{N}(\theta_0,\,\phi_0^{-1})\,,\quad \tau\sim\Gamma(a_0,\,b_0)\,, \tag{4.20}$$

i.e. μ is normally distributed with mean θ_0 and variance ϕ_0^{-1} and τ has a Gamma distribution with parameters a_0 and b_0.

The parameters θ_0, ϕ_0, a_0 and b_0 are assumed to be known. The posterior density h of (μ,τ) defined as the conditional density given the observations, is then given, using the Bayes formula, by

$$h(u,t)\propto \exp(-\phi_0(u-\theta_0)^2/2)\exp\left(-t\sum_{i=1}^n (y_i-u)^2/2\right)t^{a_0-1+n/2}\exp(-b_0t)\,.$$

Conditioning on the observations introduces a dependence between μ and τ. Nevertheless, the conditional laws of μ given τ and τ given μ have a simple form. Write $\bar{y}=n^{-1}\sum_{i=1}^n y_i$ and $S^2=n^{-1}\sum_{i=1}^n (y_i-\bar{y})^2$,

$$\theta_n(t) = (\phi_0\theta_0+nt\bar{y})/(\phi_0+nt)\,,\quad \phi_n(t)=\phi_0+nt\,,$$
$$a_n = a_0+n/2\,,\quad b_n(u)=b_0+nS^2/2+n(\bar{y}-u)^2/2\,.$$

Then,

$$\mathscr{L}(\mu|\tau)=\mathrm{N}(\theta_n(\tau),\,\phi_n^{-1}(\tau))\,,\qquad \mathscr{L}(\tau|\mu)=\Gamma(a_n,\,b_n(\mu))\,.$$

The Gibbs sampler provides a simple approach to define a Markov chain whose invariant probability has the density h. First we simulate μ_0 and τ_0 independently with distribution as in (4.20). At the k-th stage,

given $(\mu_{k-1},\ \tau_{k-1})$, we first simulate $N_k \sim \mathrm{N}(0,1)$ and $G_k \sim \Gamma(a_n,1)$ and we set

$$\begin{aligned}\mu_k &= \theta_n(\tau_{k-1}) + \phi_n^{-1/2}(\tau_{k-1})N_k \\ \tau_k^{-1} &= b_n(\mu_k)G_k\ .\end{aligned}$$

In the simple case where $\theta_0 = 0$ and $\phi_0 = 0$ which corresponds to a flat prior for μ (an improper distribution with a constant density on $\mathbb{R}$), the above equation can be rewritten as

$$\begin{aligned}\mu_k &= \bar{y} + (n\tau_{k-1})^{-1/2}N_k \\ \tau_k^{-1} &= \left(b_0 + nS^2/2 + n(\bar{y}-\mu_k)^2/2\right)G_k\ .\end{aligned}$$

Thus, $\{\tau_k^{-1},\ k\in\mathbb{N}\}$ and $\{(\mu_k-\bar{y})^2,\ k\in\mathbb{N}\}$ are Markov chains which follow the random coefficient autoregressions

$$\begin{aligned}\tau_k^{-1} &= \frac{N_k^2 G_k}{2}\tau_{k-1}^{-1} + \left(b_0 + \frac{nS^2}{2}\right)G_k\ , \\ (\mu_k-\bar{y})^2 &= \frac{N_k^2 G_{k-1}}{2}(\mu_{k-1}-\bar{y})^2 + \left(b_0 + \frac{nS^2}{2}\right)G_{k-1}\ .\end{aligned}$$

4.1.4 Hit-and-run algorithm

Let K be a bounded subset of $\mathbb{R}^d$ with non-empty interior. Let $\rho : K \to [0,\infty)$ be a (not necessarily normalized) density, i.e. a non-negative Lebesgue-integrable function. We define the probability measure π_ρ with density ρ by

$$\pi_\rho(A) = \frac{\int_A \rho(x)\mathrm{d}x}{\int_K \rho(x)\mathrm{d}x} \tag{4.21}$$

for all measurable sets $A \subset K$. For example, if $\rho(x) \equiv 1$ then π is simply the uniform distribution on K. The hit-and-run Markov kernel, presented below, can be used to sample approximately from π_ρ. The hit-and-run algorithm consists of two steps. Starting from $x \in K$, we first choose a random direction $\theta \in \mathrm{S}_{d-1}$, the unit sphere in $\mathbb{R}^d$ according to a uniform distribution on S_{d-1}. We then choose the next state of the Markov chain with respect to the density ρ restricted to the chord determined by the current state x and and the direction $\theta \in \mathrm{S}_{d-1}$: for any function $f \in \mathbb{F}_+(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$,

$$H_\theta f(x) = \frac{1}{\ell_\rho(x,\theta)}\int_{s=-\infty}^{\infty} \mathbb{1}_K(x+s\theta)f(x+s\theta)\rho(x+s\theta)\mathrm{d}s\ , \tag{4.22}$$

where $\ell_\rho(x,\theta)$ is the normalizing constant defined as

$$\ell_\rho(x,\ \theta) = \int_{-\infty}^{\infty} \mathbb{1}_K(x+s\theta)\rho(x+s\theta)\mathrm{d}s\ . \tag{4.23}$$

The Markov kernel H that corresponds to the hit-and-run algorithm is therefore defined by, for all $f \in \mathbb{F}_+(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$,

$$Hf(x) = \int_{\mathrm{S}_{d-1}} H_\theta f(x)\sigma_{d-1}(\mathrm{d}\theta)\ , \tag{4.24}$$

where σ_{d-1} is the uniform distribution on S_{d-1}.

Lemma 4.12 *For all $\theta \in \mathrm{S}_{d-1}$, the Markov kernel H_θ is reversible with respect to π_ρ defined in* (4.21). *Furthermore, H is also reversible with respect to π_ρ.*

PROOF. Let $c = \int_K \rho(x)\mathrm{d}x$ and $A, B \in \mathscr{B}(K)$. By elementary computations, we have

$$\begin{aligned}
\int_A H_\theta(x,\, B)\pi_\rho(\mathrm{d}x) &= \int_A \int_{-\infty}^{\infty} \frac{\mathbb{1}_B(x+s\theta)\rho(x+s\theta)\mathrm{d}s}{\ell_\rho(x,\theta)} \frac{\rho(x)\mathrm{d}x}{c} \\
&= \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \frac{\mathbb{1}_A(x)\mathbb{1}_B(x+s\theta)\rho(x+s\theta)\rho(x)}{\ell_\rho(x,\theta)} \frac{\mathrm{d}s\mathrm{d}x}{c} \\
&= \int_{\mathbb{R}^d} \int_{-\infty}^{\infty} \frac{\mathbb{1}_A(y-s\theta)\mathbb{1}_B(y)\rho(y)\rho(y-s\theta)}{\ell_\rho(y-s\theta,\theta)} \frac{\mathrm{d}s\mathrm{d}y}{c} \\
&= \int_B \int_{-\infty}^{\infty} \frac{\mathbb{1}_A(y-s\theta)\rho(y-s\theta)}{\ell_\rho(y-s\theta,\theta)} \mathrm{d}s\pi_\rho(\mathrm{d}y) \\
&= \int_B \int_{-\infty}^{\infty} \frac{\mathbb{1}_A(y-s\theta)\rho(y-s\theta)}{\ell_\rho(y,\theta)} \mathrm{d}s\pi_\rho(\mathrm{d}y) \\
&= \int_B \int_{-\infty}^{\infty} \frac{\mathbb{1}_A(y+t\theta)\rho(y+t\theta)}{\ell_\rho(y,\theta)} \mathrm{d}t\pi_\rho(\mathrm{d}y) \\
&= \int_B H_\theta(x,\, A)\pi_\rho(\mathrm{d}x)\,.
\end{aligned}$$

The reversibility of H_θ is proved. The reversibility of H_ρ follows from the reversibility of H_θ: for any $A, B \in \mathscr{B}(\mathbb{R}^d)$, we have

$$\begin{aligned}
\int_A H(x,B)\pi_\rho(\mathrm{d}x) &= \int_{\mathrm{S}_{d-1}} \int_A H_\theta(x,B)\pi_\rho(\mathrm{d}x)\sigma_{d-1}(\mathrm{d}\theta) \\
&= \int_{\mathrm{S}_{d-1}} \int_B H_\theta(x,A)\pi_\rho(\mathrm{d}x)\sigma_{d-1}(\mathrm{d}\theta) \\
&= \int_B H(x,A)\pi_\rho(\mathrm{d}x)\,.
\end{aligned}$$

■

4.1.5 Gibbs sampling

When the distribution of interest is multivariate, it may be the case that for each particular variable, its conditional distribution given all remaining variables has a simple form. In this case, a natural algorithm is the *Gibbs sampler*, which is now described. Its name somehow inappropriately stems from its use for the simulation of Gibbs Markov random fields by **?**.

Assume that $\mathsf{X} = \mathsf{X}_1 \times \cdots \times \mathsf{X}_m$ is a product space equipped with the product σ-algebra $\mathscr{X} = \mathscr{X}_1 \otimes \cdots \otimes \mathscr{X}_m$. Let λ_k, $k \in \{1,\ldots,m\}$, be σ-finite measures on $(\mathsf{X}_k, \mathscr{X}_k)$, and let $\lambda = \lambda_1 \otimes \lambda_2 \otimes \cdots \otimes \lambda_m$ be the product measure. Suppose we are given a joint distribution with probability density function π with respect to the product measure λ. For simplicity, we assume that π is everywhere positive. An element $x \in \mathsf{X}$ may be decomposed into m components $x = (x^{[1]},\ldots,x^{[m]})$, where $x^{[k]} \in \mathsf{X}_k$. If k is an index in $\{1,\ldots,m\}$, we shall denote by $x^{[k]}$ the kth component of x and by $x^{[-k]} = \{x^{[l]}\}_{l\neq k}$ the collection of remaining components. We further denote by $\pi_k(\cdot \mid x^{[-k]})$ the conditional probability density function, defined as

$$\pi_k(x^{[k]} \mid x^{[-k]}) = \frac{\pi(x^{[1]},x^{[2]},\ldots,x^{[m]})}{\int \pi(x^{[1]},x^{[2]}\ldots,x^{[m]})\lambda_k(\mathrm{d}x^{[k]})}\,.$$

We assume that sampling from this conditional distribution is feasible (for $k = 1,\ldots,m$). Note that $x^{[k]}$ is not necessarily scalar but may be itself vector-valued.

The *deterministic scan Gibbs* sampler is an MCMC algorithm which, starting from an initial arbitrary state X_0, updates the current state $X_i = (X_i^{[1]},\ldots,X_i^{[m]})$ to a new state $X^{[i+1]}$ as follows.

For $k = 1,2,\ldots,m$: Simulate $X_{i+1}^{[k]}$ from $\pi_k(\cdot \mid X_{i+1}^{[1]},\ldots,X_{i+1}^{[k-1]},X_i^{[k+1]},\ldots,X_i^{[m]})$.

In other words, in the kth round of the cycle needed to simulate X_{i+1}, the kth component is updated by simulation from its conditional distribution given all other components, which remain fixed. This new value then supersedes the old one and is used in the subsequent simulation steps. A complete cycle of m condi-

tional simulations is usually referred to as a *sweep* of the algorithm. We denote by K_k the corresponding Markov kernel:

$$K_k(x, A_1 \times \cdots \times A_m) = K_k(x^{[1]}, \ldots, x^{[m]}; A) = \int \cdots \int \prod_{j \neq k} \delta_{x^{[j]}}(\mathrm{d}x'^{[j]}) \pi_k(x'^{[k]} \mid x^{[-k]}) \, \mathbb{1}_A(x') \, \lambda_k(\mathrm{d}x'^{[k]}) \,. \quad (4.25)$$

Proposition 4.13 (Reversibility of individual Gibbs steps) *Each of the m individual kernels K_k, $k \in \{1, \ldots, m\}$ is reversible with respect to π and thus admits π as a stationary probability density function.*

PROOF. See Exercise 8.26. ■

Each step corresponds to a very special type of Metropolis-Hastings move where the acceptance probability is equal to 1, due to choice of π_k as the proposal distribution. However, Proposition 4.13 does not suffice to establish the convergence of the Gibbs sampler. Only the combination of the m moves in the complete cycle has a chance of producing a chain with the ability to visit the whole space X from any starting point.

Example 4.14 (Scalar normal–inverse gamma) *In a statistical problem one may be presented with a set of independent observations $\mathbf{y} := \{y_1, \ldots, y_n\}$, which we assume to be normally distributed, but with unknown mean μ and variance τ^{-1} (τ is often referred to as the* precision*). The Bayesian approach to this problem is to assume that μ and τ are themselves random variables, with a given prior distribution. For example, we might assume that*

$$\mu \sim \mathrm{N}(\mu_0, \tau_0^{-1}) \,, \quad \tau \sim \Gamma(a_0, b_0) \,,$$

i.e., μ is normal with mean μ_0 and variance τ_0^{-1} and τ has gamma distribution with parameters a_0 and b_0. [1]

The parameters μ_0, τ_0, a_0 and b_0 are assumed to be known. Then the prior density for (μ, τ) is given by

$$\pi(\mu, \tau) \propto \exp\{-\tau_0(\mu - \mu_0)^2/2\} \tau^{a_0 - 1} \exp\{-b_0 \tau\} \,.$$

The posterior density for (μ, τ) defined as the conditional density given the observations, is then given, using the Bayes formula, by

$$\pi(\mu, \tau \mid \mathbf{y}) \propto \pi(\mu, \tau) f(\mathbf{y} \mid \mu, \tau) \propto \exp(-\tau_0(\mu - \mu_0)^2/2) \times \exp\left(-\tau \sum_{i=1}^{n} (y_i - \mu)^2/2\right) \tau^{a_0 - 1 + n/2} \exp(-b_0 \tau) \,.$$

Conditioning with respect to the observations has introduced a dependence between μ and τ. Nevertheless, the full conditional distributions *still have a simple form*

$$\pi(\mu \mid \mathbf{y}, \tau) = \mathrm{N}(m_n(\tau), t_n^{-1}(\tau)) \,,$$
$$\pi(\tau \mid \mathbf{y}, \mu) = \Gamma(a_n, b_n(\mu)) \,,$$

where, denoting $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$

$$m_n(\tau) := (\tau_0 \mu_0 + n\tau \bar{y}) / (\tau_0 + n\tau) \,, \quad t_n(\tau) := \tau_0 + n\tau \,,$$
$$a_n := a_0 + n/2 \,, \quad b_n(\mu) := b_0 + 1/2 \sum_{i=1}^{n} (y_i - \mu)^2 \,.$$

[1] Z has an inverse gamma distribution if $1/Z$ has a gamma distribution; general properties can be found, for example, in (**?**, Section 8.5)

The Gibbs sampler *provides a particularly simple approach to sample from* $\pi(\mu,\ \tau \mid \mathbf{y})$. *We set*

$$\mathsf{X} = \mathsf{X}_1 \times \mathsf{X}_2 = \mathbb{R} \times [0,\ \infty)\ .$$

We wish to simulate $X = (\mu,\ \tau)$ *with density* $\pi(\mu,\ \tau \mid \mathbf{y})$. *First we simulate* X_0, *say from the product form density* $\pi(\mu,\ \tau)$. *At the* t*-th stage, given* $X_{t-1} = (\mu_{t-1},\ \tau_{t-1})$, *we first simulate* $N_t \sim \mathrm{N}(0,1)$ *and* $G_t \sim \Gamma(a_n, 1)$ *and we put*

$$\begin{aligned}\mu_t &= m_n(\tau_{t-1}) + t_n^{-1/2}(\tau_{t-1})N_t \\ \tau_t &= b_n(\mu_t)G_t\ ,\end{aligned}$$

and $X_t = (\mu_t,\ \tau_t)$. *In the simple case where* $\mu_0 = 0$ *and* $\tau_0 = 0$, *which corresponds to a* flat prior *for* μ *(an improper distribution with a "constant" density on* $\mathbb{R}$*), the above equation can be rewritten as*

$$\begin{aligned}\mu_t &= \bar{y} + (n\tau_{t-1})^{-1/2}N_t \\ \tau_t &= \left(b_0 + 1/2\sum_{i=1}^{n}(y_i - \mu_t)^2\right)G_t\ .\end{aligned}$$

By alternating and iterating these two updates, we see that τ_t^{-1} *is a Markov chain that iterates following a random coefficient autoregression (see* **??***),*

$$\tau_t^{-1} = A_t\tau_{t-1}^{-1} + B_t\ , \tag{4.26}$$

where $A_t = N_t^2/(2G_t)$, *and* $B_t = (b_0 + nS^2/2)/G_t)$. *It is easily checked that* $(\mu_t - \bar{y})^2$ *is also a random coefficient autoregressive process. The Gibbs sampler can be implemented in* R *using the script provided for this example.*

4.2 Ordering the asymptotic variances

Define $\mathsf{L}^2(\pi) = \{f : \mathsf{X} \to \mathbb{R} :\ f$ is $\mathscr{X}/\mathscr{B}(\mathbb{R})$-measurable$\}$ and $\mathsf{L}_0^2(\pi) = \left\{f \in \mathsf{L}^2(\pi) :\ \pi(f) = 0\right\}$. Moreover, for all $f, g \in \mathsf{L}^2(\pi)$, write

$$\langle f, g\rangle = \pi(fg)\ .$$

Definition 4.15. Let P_0 and P_1 be Markov transition kernels on $(\mathsf{X}, \mathscr{X})$ with invariant probability π. We say that P_1 dominates P_0 on the off-diagonal, writen $P_0 \le P_1$, if for all $A \in \mathscr{X}$, and π-a.e. all x in X,

$$P_0(x, A \setminus \{x\}) \le P_1(x, A \setminus \{x\})\ .$$

Definition 4.16. Let P_0 and P_1 be Markov transition kernels on $(\mathsf{X}, \mathscr{X})$ with invariant probability π. We say that P_1 dominates P_0 in the covariance ordering, writen $P_0 \preceq P_1$, if for all $f \in \mathsf{L}^2(\pi)$,

$$\langle f, P_1 f\rangle \le \langle f, P_0 f\rangle\ .$$

Lemma 4.17 *Let* P_0 *and* P_1 *be Markov transition kernels on* $(\mathsf{X}, \mathscr{X})$ *with invariant probability* π. *Assume moreover that* $P_0 \le P_1$. *Then,* $P_0 \preceq P_1$.

PROOF. For all $x \in \mathsf{X}$ and $A \in \mathscr{X}$, define

$$P(x,A) = \delta_x(A) + P_1(x,A) - P_0(x,A) .$$

P is a nonnegative kernel since for all $x \in \mathsf{X}$ and $A \in \mathscr{X}$,

$$\begin{aligned} P(x,\{x\}) &= 1 + P_1(x,\{x\}) - P_0(x,\{x\}) \geq 0 , \\ P(x,A\setminus\{x\}) &= P_1(x,A\setminus\{x\}) - P_0(x,A\setminus\{x\}) \geq 0 . \end{aligned}$$

Combining with $P(x,\mathsf{X}) = 1$, this implies that P is a Markov kernel. We now show that for all $f \in \mathsf{L}^2(\pi)$

$$\langle f, P_0 f\rangle - \langle f, P_1 f\rangle = \iint \pi(\mathrm{d}x) P(x,\mathrm{d}y) \left(f(x) - f(y)\right)^2 /2$$

Indeed,

$$\begin{aligned} \langle f, P_0 f\rangle - \langle f, P_1 f\rangle &= \iint \pi(\mathrm{d}x) f(x) (P_0(x,\mathrm{d}y) - P_1(x,\mathrm{d}y)) f(y) \\ &= \iint \pi(\mathrm{d}x) f(x) (\delta_x(\mathrm{d}y) - P(x,\mathrm{d}y)) f(y) \\ &= \int \pi(\mathrm{d}x) f^2(x) - \iint \pi(\mathrm{d}x) P(x,\mathrm{d}y) f(x) f(y) \\ &= \iint \pi(\mathrm{d}x) P(x,\mathrm{d}y) \left[\frac{f^2(x) - f^2(y)}{2} + f(x) f(y)\right] \end{aligned}$$

where the last inequality follows from the fact that P is clearly π-invariant. Finally,

$$\langle f, P_0 f\rangle - \langle f, P_1 f\rangle = \iint \pi(\mathrm{d}x) P(x,\mathrm{d}y) \left(f(x) - f(y)\right)^2 /2 \geq 0 .$$

And thus, $P_0 \preceq P_1$. ■

Proposition 4.18 *Let P_0 and P_1 be Markov kernels on $\mathsf{X}\times\mathscr{X}$ and $\pi \in \mathbb{M}_1(\mathscr{X})$. Assume that π is reversible with respect to P_0 and P_1 and that for all $i \in \{0,1\}$, $\pi(|f|^2) + 2\lim_{n\to\infty} \sum_{k=1}^n \pi(|f|P_i^k|f|) < \infty$. Assume that $P_0 \preceq P_1$. Then, for any $f \in \mathsf{L}^2_0(\pi)$,*

$$v_1(f,P_1) \leq v_0(f,P_0) ,$$

where for $i \in \{0,1\}$

$$v_i(f,P_i) = \pi(f^2) + 2\lim_{n\to\infty} \sum_{k=1}^n \pi(f P_i^k f) .$$

PROOF. For all $\alpha \in (0,1)$, denote $P_\alpha = (1-\alpha)P_0 + \alpha P_1$. For $\lambda \in (0,1)$, define

$$w_\lambda(\alpha) = \sum_{k=0}^\infty \lambda^k \left\langle f, P_\alpha^k f\right\rangle ,$$

We first show that for all $\alpha \in (0,1)$,

$$\frac{\mathrm{d}w_\lambda(\alpha)}{\mathrm{d}\alpha} = \sum_{k=0}^\infty \lambda^k \sum_{i=1}^k \left\langle f, P_\alpha^{i-1}(P_1 - P_0) P_\alpha^{k-i} f\right\rangle . \tag{4.27}$$

Indeed, for all $1 \leq \ell \leq k$ and all $\alpha_1, \ldots, \alpha_k$,

$$\begin{aligned} \left\langle f, P_{\alpha_1} \ldots P_{\alpha_k} f\right\rangle &= (1-\alpha_\ell) \left\langle f, P_{\alpha_1} \ldots P_{\alpha_{\ell-1}} P_0 P_{\alpha_{\ell+1}} \ldots P_{\alpha_k} f\right\rangle \\ &\quad + \alpha_\ell \left\langle f, P_{\alpha_1} \ldots P_{\alpha_{\ell-1}} P_1 P_{\alpha_{\ell+1}} \ldots P_{\alpha_k} f\right\rangle , \end{aligned}$$

so that

$$\frac{\partial}{\partial \alpha_\ell} \left\langle f, P_{\alpha_1} \ldots P_{\alpha_k} f\right\rangle = \left\langle f, P_{\alpha_1} \ldots P_{\alpha_{\ell-1}} (P_1 - P_0) P_{\alpha_{\ell+1}} \ldots P_{\alpha_k} f\right\rangle ,$$

and thus, we obtain by differentiating $\alpha \mapsto w_\lambda(\alpha)$,

$$\frac{\mathrm{d}w_\lambda(\alpha)}{\mathrm{d}\alpha} = \sum_{k=0}^\infty \lambda^k \sum_{i=1}^k \left\langle f, P_\alpha^{i-1}(P_1 - P_0) P_\alpha^{k-i} f\right\rangle .$$

This shows (4.27). Using now that π is reversible for the kernel P_α,

$$\begin{aligned}\frac{\mathrm{d}w_\lambda(\alpha)}{\mathrm{d}\alpha} &= \sum_{i=1}^{\infty}\sum_{k\geq i}^{\infty}\lambda^k\left\langle P_\alpha^{i-1}f,(P_1-P_0)P_\alpha^{k-i}f\right\rangle \\ &= \lambda\left\langle\sum_{\ell=0}^{\infty}\lambda^\ell P_\alpha^\ell f,(P_1-P_0)\sum_{\ell=0}^{\infty}\lambda^\ell P_\alpha^\ell f\right\rangle \leq 0\,,\end{aligned}$$

which completes the proof. ■

Chapter 5

Ergodic theory for Markov chains

Contents

This chapter is concerned with the asymptotic behaviour of sample averages of stationary ergodic Markov chains. For this purpose, it is convenient to link the Markov chain to a certain dynamical system. The Law of Large Numbers for Markov chains is then obtained as a consequence of the classical Birkhoff theorem. It turns out that under appropriate assumptions, this approach still holds true for functions that actually depend on the whole trajectory, such as $n^{-1}\sum_{k=0}^{n-1} f(\{X_{k+\ell},\, \ell \in \mathbb{N}\})$ or $n^{-1}\sum_{k=0}^{n-1} f(\{X_{k-\ell},\, \ell \in \mathbb{N}\})$. A key result of this chapter is Theorem 5.21 which shows that the existence of a unique invariant probability measure implies the ergodicity of the associated dynamical system, which in turn allows to apply the Birkhoff ergodic theorem. Still, the price to pay for using the dynamical system theory is the stationarity assumption. Typically in this chapter, the Law of Large Numbers will be proved $\mathbb{P}_\pi$ − a.s. (where π is the unique invariant probability measure for P) and will be extended to other initial distributions. Sufficient conditions are given in this chapter but a more thorough treatment for other initial distributions requires notions that will be introduced in later chapters.

5.1 Dynamical systems

We first briefly introduce some basic definitions and properties of dynamical systems that will be useful when applying them to Markov chains.

5.1.1 Definitions

Definition 5.1 (Dynamical system). Let $(\Omega, \mathscr{B}, \mathbb{P})$ be a probability space.

- A measurable map T from $(\Omega, \mathscr{B})$ to $(\Omega, \mathscr{B})$ is a measure-preserving transformation if for all $A \in \mathscr{B}$,
$$\mathbb{P}(\mathrm{T}^{-1}(A)) = \mathbb{P}(A)\,.$$
The probability $\mathbb{P}$ is then said to be invariant under the transformation T and $(\Omega, \mathscr{B}, \mathbb{P}, \mathrm{T})$ is said to be a dynamical system.
- The application T is said to be an invertible measure-preserving transformation if it is measure-preserving, invertible and its inverse T^{-1} is measurable.

If the transformation T is measure preserving and invertible, then T^{-1} is also measure-preserving since, for all $A \in \mathscr{B}$,

$$\mathbb{P}((\mathrm{T}^{-1})^{-1}(A)) = \mathbb{P}(\mathrm{T}(A)) = \mathbb{P}(\mathrm{T}^{-1}\{\mathrm{T}(A)\}) = \mathbb{P}(A)\,.$$

Note also that if T is measure-preserving, then for all integer $n \in \mathbb{N}$ and $A \in \mathscr{B}$,

$$\mathbb{P}(\mathrm{T}^{-n}(A)) = \mathbb{P}(A)\,.$$

Let $(\mathsf{X}, \mathscr{X})$ be a measurable space. Denote by $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$ the associated canonical space and by $\{X_n,\, n \in \mathbb{N}\}$ the coordinate process. The shift operator θ (see **??**) is defined, for $\omega = (\omega_k)_{k\in\mathbb{N}} \in \mathsf{X}^{\mathbb{N}}$, by

$$\theta(\omega_0, \omega_1, \dots) = (\omega_1, \omega_2, \dots)\,.$$

Note that $X_k \circ \theta = X_{k+1}$ for all $k \geq 0$. By **??**, θ is a measurable map from $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$ to $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$, but it is not invertible. Recall that $\{X_n,\, n \in \mathbb{N}\}$ is stationary if the distribution of $(X_k, \dots, X_{k+n})$ is independent of k for all $n \in \mathbb{N}$. The next Lemma shows the connection between the stationarity of the coordinate process and the invariance of $\mathbb{P}$ under the shift operator θ.

Lemma 5.2 *A probability measure $\mathbb{P}$ on $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$ is invariant under the shift operator θ if and only if the coordinate process is stationary under $\mathbb{P}$.*

PROOF. It suffices to note (**??**) that $\mathbb{P}$ is measure preserving if and only if for all $n \geq 1$ and all $f \in \mathbb{F}_b(\mathsf{X}^n, \mathscr{X}^{\otimes n})$, $\mathbb{E}[f(X_0, \ldots, X_{n-1})] = \mathbb{E}[f(X_1, \ldots, X_n)]$. ■

Example 5.3 (One-sided Markov shift). Let P be a kernel on $(\mathsf{X}, \mathscr{X})$ which admits an invariant probability π on $(\mathsf{X}, \mathscr{X})$. By **??**, there exists a unique probability measure $\mathbb{P}_\pi$ on the canonical space such that the coordinate process is a Markov chain with kernel P and initial distribution π. Then, by **??**, the canonical chain is a stationary process. Lemma 5.2 then shows that $\mathbb{P}_\pi$ is invariant under θ, i.e. $\mathbb{P}_\pi \circ \theta^{-1} = \mathbb{P}_\pi$.

5.1.2 Invariant events

Definition 5.4 (Invariant random variable, invariant event). Let T be a measurable map from $(\Omega, \mathscr{B})$ to $(\Omega, \mathscr{B})$.

- A $\bar{\mathbb{R}}$-valued random variable Y on $(\Omega, \mathscr{B})$ is invariant for T if $Y \circ \mathrm{T} = Y$.
- An event A is invariant for T if $A = \mathrm{T}^{-1}(A)$ or equivalently if its indicator function $\mathbb{1}_A$ is invariant for T.

Proposition 5.5 *Let T be a measurable map from $(\Omega, \mathscr{B})$ to $(\Omega, \mathscr{B})$.*

(i) The collection $\mathscr{I}$ of invariant sets for is a sub-σ-field of $\mathscr{B}$.
(ii) Let Y be a $\bar{\mathbb{R}}$-valued random variable. Y is invariant if and only if Y is $\mathscr{I}$-measurable.

PROOF. The proof of (i) is elementary and omitted. Consider now (ii). If $Y \circ \mathrm{T} = Y$, then for all $B \in \mathscr{B}(\bar{\mathbb{R}})$,

$$\mathrm{T}^{-1}(Y^{-1}(B)) = (Y \circ \mathrm{T})^{-1}(B) = Y^{-1}(B)\,.$$

Thus $Y^{-1}(B) \in \mathscr{I}$ and Y is $\mathscr{I}$-measurable.

Conversely, if Y is $\mathscr{I}$-measurable, define $A_{k,n} = \{\frac{k}{n} \leq Y < \frac{k+1}{n}\} \in \mathscr{I}$, $n \geq 1$, $k \in \mathbb{Z}$. Then, with the convention $\infty \times 0 = 0$, Y is the pointwise limit of the sequence $\{Y_n,\ n \in \mathbb{N}^*\}$ defined by

$$Y_n = \sum_{k \in \mathbb{Z}} \frac{k}{n} \mathbb{1}_{A_{k,n}} + \infty \mathbb{1}_{\{Y = +\infty\}} - \infty \mathbb{1}_{\{Y = -\infty\}}\,.$$

Since Y is $\mathscr{I}$-measurable, the sets $A_{k,n}$, $\{Y = -\infty\}$ and $\{Y = +\infty\}$ belong to $\mathscr{I}$ hence the functions Y_n are invariant for all n and

$$Y \circ T = (\lim_{n\to\infty} Y_n) \circ T = \lim_{n\to\infty} (Y_n \circ T) = \lim_{n\to\infty} Y_n = Y\,.$$

■

The most important examples of invariant random variables which we will be considered in the sequel are defined as limits.

Lemma 5.6 *Let $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$ be the canonical space, $\{X_n,\ n \in \mathbb{N}\}$ the coordinate process and θ the shift operator. Then $\mathscr{I} \subset \cap_{k\geq 0} \sigma(X_\ell,\ \ell \geq k)$. Moreover, for any $f \in \mathbb{F}\left(\bar{\mathbb{R}}, \mathscr{B}(\bar{\mathbb{R}})\right)$, $\limsup_{n\to\infty} f(X_n)$, $\liminf_{n\to\infty} f(X_n)$, $\limsup_{n\to\infty} n^{-1}(f(X_0) + \cdots + f(X_{n-1}))$ and $\liminf_{n\to\infty} n^{-1}(f(X_0) + \cdots + f(X_{n-1}))$ are invariant random variables.*

PROOF. Set $\mathscr{G}_k = \sigma(X_\ell, \ell \geq k)$ and $\mathscr{G}_\infty = \cap_{k\geq 0} \mathscr{G}_k$. Let A be an invariant set. Then, $A \in \mathscr{G}_0 = \mathscr{X}^{\otimes\mathbb{N}}$. Since we have the implication: if $A \in \mathscr{G}_k$, then $A = \theta^{-1}(A) \in \mathscr{G}_{k+1}$, we obtain by induction that $A \in \mathscr{G}_k$ for all k and thus $A \in \cap_{k\geq 0} \mathscr{G}_k$.

The remaining statements of the lemma are straightforward. ■

Let now $(\Omega, \mathscr{B}, \mathbb{P}, \mathrm{T})$ be a dynamical system, that is, T is measure preserving for $\mathbb{P}$. A $\bar{\mathbb{R}}$-valued random variable Y defined on Ω is said to be $\mathbb{P}-$a.s. invariant (for T) if $Y \circ \mathrm{T} = Y$, $\mathbb{P}-$a.s. Similarly, an event $A \in \mathscr{B}$ is $\mathbb{P}-$a.s. invariant (for T) if its indicator function $\mathbb{1}_A$ is $\mathbb{P}-$a.s. invariant.

Lemma 5.7 *If Y is $\mathbb{P}-$a.s. invariant, then there exists an invariant random variable Z, such that $Y = Z$ $\mathbb{P}-$a.s. In particular, if $A \in \mathscr{B}$ is $\mathbb{P}-$a.s. invariant, there exists $B \in \mathscr{I}$ such that $\mathbb{1}_A = \mathbb{1}_B$ $\mathbb{P}-$a.s.*

PROOF. The random variable $Z = \limsup_{n\to\infty} Y \circ \mathrm{T}^n$ is invariant. Since Y is $\mathbb{P}-$a.s. invariant, $Y = Y \circ \mathrm{T}$ $\mathbb{P}-$a.s., hence $Y = Y \circ \mathrm{T}^n$ $\mathbb{P}-$a.s. for all $n \geq 1$. This yields $Y = Z$ $\mathbb{P}-$a.s. If $Y = \mathbb{1}_A$, then there exists an invariant random variable Z such that $\mathbb{1}_A = Z$ $\mathbb{P}-$a.s.. The set $B = \{Z = 1\}$ is therefore invariant and $\mathbb{1}_A = \mathbb{1}_B$ $\mathbb{P}-$a.s.. ■

It is easy to check that the family of $\mathbb{P}-$a.s. invariant sets for T is a σ-algebra $\mathscr{I}_{\mathbb{P}}$. Lemma 5.7 shows that $\mathscr{I}_{\mathbb{P}}$ is the $\mathbb{P}$-completion of the invariant σ-algebra $\mathscr{I}$ (the σ-algebra generated by $\mathscr{I}$ and the family of sets which are $\mathbb{P}$-negligible).

Denote by T^n the transformation T iterated n-times and by convention, we let T^0 be the identity function. The behavior of time averages is given by the following fundamental result.

Theorem 5.8 (Birkhoff's ergodic theorem). *Let $(\Omega, \mathscr{B}, \mathbb{P}, \mathrm{T})$ be a dynamical system and Y be a random variable such that $\mathbb{E}[|Y|] < \infty$. Then,*

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} Y \circ \mathrm{T}^k = \mathbb{E}[Y \,|\, \mathscr{I}] \qquad \mathbb{P}-\text{a.s.} \tag{5.1}$$

Moreover, the convergence also holds in $\mathsf{L}^1(\mathbb{P})$.

The proof is based on the following lemma.

Lemma 5.9 *Let Z be a random variable such that $\mathbb{E}[|Z|] < \infty$. If $\mathbb{E}[Z \,|\, \mathscr{I}] > 0$ $\mathbb{P}-$a.s., then,*

$$\liminf_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} Z \circ \mathrm{T}^k \geq 0 \qquad \mathbb{P}-\text{a.s.}$$

PROOF. For all $n \in \mathbb{N}^*$, write $S_n = \sum_{k=0}^{n-1} Z \circ \mathrm{T}^k$. Note that for all $n \geq 1$, $\mathbb{E}[|S_n|] \leq n\mathbb{E}[|Z|] < \infty$. Denote $L_n = \inf\{S_k : 1 \leq k \leq n\}$ and $A = \{\inf_{n\in\mathbb{N}^*} L_n = -\infty\}$. Since $|Z| < \infty$ $\mathbb{P}-$a.s., $\{\inf_{n\geq 1} S_n = -\infty\} = \{\inf_{n\geq 1} S_n \circ \mathrm{T} = -\infty\}$ $\mathbb{P}-$a.s., the set A is $\mathbb{P}-$a.s. invariant. Since $L_{n-1} \geq L_n$,

$$\begin{aligned} L_n &= Z + \inf\{S_k - Z : 1 \leq k \leq n\} \\ &= Z + \inf(0, L_{n-1} \circ \mathrm{T}) \geq Z + \inf(0, L_n \circ \mathrm{T})\,. \end{aligned} \tag{5.2}$$

Since $\mathbb{E}[|S_k|] < \infty$ for all $k \in \mathbb{N}$, for all $n \geq 1$, $\mathbb{E}[|L_n|] \leq \sum_{k=0}^{n-1} \mathbb{E}[|S_k|] < \infty$. and (5.2) implies that $Z \leq L_n + (L_n \circ \mathrm{T})^- = L_n + L_n^- \circ \mathrm{T}$ $\mathbb{P}-$a.s. Then, using $\mathbb{1}_A = \mathbb{1}_A \circ \mathrm{T}$ $\mathbb{P}-$a.s., we get

$$\begin{aligned} \mathbb{E}[\mathbb{1}_A Z] &\leq \mathbb{E}[\mathbb{1}_A L_n] + \mathbb{E}\left[\mathbb{1}_A L_n^- \circ \mathrm{T}\right] = \mathbb{E}[\mathbb{1}_A L_n] + \mathbb{E}\left[\mathbb{1}_A \circ \mathrm{T}\; L_n^- \circ \mathrm{T}\right] \\ &\leq \mathbb{E}[\mathbb{1}_A L_n] + \mathbb{E}\left[\mathbb{1}_A L_n^-\right] = \mathbb{E}\left[\mathbb{1}_A L_n^+\right]\,. \end{aligned} \tag{5.3}$$

Since $L_n^+ \leq Z^+$ with $\mathbb{E}[Z^+] < \infty$ and $\lim_{n\to\infty} \mathbb{1}_A L_n^+ = 0$ $\mathbb{P}-$a.s., Lebesgue's dominated convergence theorem shows that $\mathbb{E}[\lim_{n\to\infty} \mathbb{1}_A L_n^+] = 0$. Therefore, since $0 \leq \mathbb{E}[\mathbb{1}_A \mathbb{E}[Z \,|\, \mathscr{I}]] = \mathbb{E}[\mathbb{1}_A Z]$, we finally get using (5.3)

$$\mathbb{E}[\mathbb{1}_A \mathbb{E}[Z \,|\, \mathscr{I}]] = \mathbb{E}[\mathbb{1}_A Z] \leq \mathbb{E}\left[\lim_{n\to\infty} \mathbb{1}_A L_n^+\right] = 0\,.$$

By assumption, $\mathbb{E}[Z \,|\, \mathscr{I}] > 0$ $\mathbb{P}-$a.s., the previous inequality shows $\mathbb{P}(A) = 0$. We conclude that $\liminf_{n\to\infty} n^{-1} S_n \geq 0$ $\mathbb{P}-$a.s.

■

PROOF. [of Theorem 5.8] Let $\varepsilon > 0$ and set $Z = Y - \mathbb{E}[Y \,|\, \mathscr{I}] + \varepsilon$. Note that $\mathbb{E}[|Z|] \leq 2\mathbb{E}[|Y|] + \varepsilon$ showing that $\mathbb{E}[Z \,|\, \mathscr{I}]$

is well-defined and, by construction, $\mathbb{E}[Z|\mathscr{I}] > 0$. Using that $\mathbb{E}[Y|\mathscr{I}]$ being $\mathscr{I}$-measurable, it is invariant according to Proposition 5.5 and Lemma 5.9 implies that

$$\liminf_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k \geq \mathbb{E}[Y|\mathscr{I}] - \varepsilon \qquad \mathbb{P}-\text{a.s.}$$

Replacing Y by $-Y$, we finally obtain

$$-\varepsilon + \limsup_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k \leq \mathbb{E}[Y|\mathscr{I}] \leq \liminf_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k + \varepsilon \qquad \mathbb{P}-\text{a.s.}$$

This shows (5.1) since $\varepsilon > 0$ is arbitrary.

We now turn to the $\mathsf{L}^1(\mathbb{P})$ convergence. Denote by $M_n(Y) = n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k$. If Y bounded, then Lebesgue's dominated convergence theorem shows that the convergence in (5.1) also holds in $\mathsf{L}^1(\mathbb{P})$. For a bounded random variable $\bar{Y}$, consider the decomposition

$$|M_n(Y) - \mathbb{E}[Y|\mathscr{I}]| \leq |M_n(Y) - M_n(\bar{Y})| + |M_n(\bar{Y}) - \mathbb{E}[\bar{Y}|\mathscr{I}]| + \mathbb{E}[|\bar{Y}-Y||\mathscr{I}] \ .$$

Let us denote by $\|U\|_1 = \mathbb{E}[|U|]$. Note that $\|\mathbb{E}[|\bar{Y}-Y||\mathscr{I}]\|_1 \leq \|\bar{Y}-Y\|_1$ and

$$\||M_n(Y) - M_n(\bar{Y})\|_1 \leq n^{-1}\sum_{k=0}^{n-1} \left\|(Y-\bar{Y})\circ \mathrm{T}^k\right\|_1 = \|Y-\bar{Y}\|_1 \ ,$$

where we have used that the transformation T is measure preserving and thus $\left\|(Y-\bar{Y})\circ \mathrm{T}^k\right\|_1 = \|Y-\bar{Y}\|_1$ for all $k\in\mathbb{N}$. Therefore,

$$\limsup_{n\to\infty} \|M_n(Y) - \mathbb{E}[Y|\mathscr{I}]\|_1 \leq 2\,\|\bar{Y}-Y\|_1 \ .$$

The proof is complete since bounded random variables are dense in $\mathsf{L}^1(\mathbb{P})$. ■

The most interesting case of application of Theorem 5.8 is when the σ-field $\mathscr{I}$ is trivial, in which case the conditional expectation $\mathbb{E}[Y|\mathscr{I}]$ can be replaced by $\mathbb{E}[Y]$ in (5.1).

Definition 5.10 (Ergodic dynamical system). A dynamical system $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$ is ergodic if the invariant σ-field $\mathscr{I}$ is trivial for $\mathbb{P}$, i.e. for all $A\in\mathscr{I}$, $\mathbb{P}(A)\in\{0,1\}$.

Corollary 5.11 *Let $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$ be an ergodic dynamical system and Y be a $\bar{\mathbb{R}}$-valued random variable such that $\mathbb{E}[|Y|] < \infty$. Then,*

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k = \mathbb{E}[Y] \qquad \mathbb{P}-\text{a.s.} \tag{5.4}$$

5.1.2.1 Dynamical systems associated to one-sided and two-sided sequences

In the context of Markov chains on a measurable space $(\mathsf{X},\mathscr{X})$, the dynamical systems will be associated to either one-sided sequences $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}})$, or two-sided sequences $(\mathsf{X}^{\mathbb{Z}},\mathscr{X}^{\otimes\mathbb{Z}})$. Denote by $\bar{\theta}:\mathsf{X}^{\mathbb{Z}}\to\mathsf{X}^{\mathbb{Z}}$ the shift operator on $\mathsf{X}^{\mathbb{Z}}$: for any two-sided sequence $\omega=(\omega_n)_{n\in\mathbb{Z}}\in\mathsf{X}^{\mathbb{Z}}$,

$$[\bar{\theta}(\omega)]_n = \omega_{n+1} \ , \qquad \text{for all } n\in\mathbb{Z} \ . \tag{5.5}$$

Moreover, set $\bar{\theta}_1 = \bar{\theta}$ and for all $n>1$, $\bar{\theta}_n = \bar{\theta}_{n-1}\circ\bar{\theta}$. Let $\bar{\mathbb{P}}$ be a probability measure on $(\mathsf{X}^{\mathbb{Z}},\mathscr{X}^{\otimes\mathbb{Z}})$ and denote by $\bar{\mathbb{E}}$ the associated expectation operator. Let Π be the measurable map from $(\mathsf{X}^{\mathbb{Z}},\mathscr{X}^{\otimes\mathbb{Z}})$ to $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}})$ defined by

$$\Pi(\omega) = (\omega_n)_{n\in\mathbb{N}} \ , \qquad \text{for all } \omega=(\omega_n)_{n\in\mathbb{Z}}\in\mathsf{X}^{\mathbb{Z}} \ . \tag{5.6}$$

We denote by $\mathbb{P} = \bar{\mathbb{P}}\circ\Pi^{-1}$ the probability induced on $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}})$ by the probability $\bar{\mathbb{P}}$ and the map Π. Let θ be the shift operator on $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}})$ defined in **??** and note that $\theta\circ\Pi = \Pi\circ\bar{\theta}$.

Lemma 5.12 *If* $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}}, \bar{\mathbb{P}}, \bar{\theta})$ *is a dynamical system, then* $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}, \theta)$ *is also a dynamical system.*

PROOF. By assumption, $\bar{\mathbb{P}} \circ \bar{\theta}^{-1} = \bar{\mathbb{P}}$. Combining with $\theta \circ \Pi = \Pi \circ \bar{\theta}$, this implies for all $A \in \mathscr{X}^{\otimes\mathbb{N}}$,

$$\begin{aligned}\mathbb{P} \circ \theta^{-1}(A) &= (\bar{\mathbb{P}} \circ \Pi^{-1}) \circ \theta^{-1}(A) = \bar{\mathbb{P}} \circ (\theta \circ \Pi)^{-1}(A) = \bar{\mathbb{P}} \circ (\Pi \circ \bar{\theta})^{-1}(A) \\ &= (\bar{\mathbb{P}} \circ \bar{\theta}^{-1}) \circ \Pi^{-1}(A) = \bar{\mathbb{P}} \circ \Pi^{-1}(A) = \mathbb{P}(A) \,.\end{aligned}$$

■

Proposition 5.13 *Assume that* $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}}, \bar{\mathbb{P}}, \bar{\theta})$ *is a dynamical system. If the dynamical system* $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}, \theta)$ *is ergodic, then* $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}}, \bar{\mathbb{P}}, \bar{\theta})$ *is ergodic.*

PROOF.

Let A be an invariant set for the dynamical system $(\mathsf{X}^{\mathbb{Z}}, \mathscr{X}^{\otimes\mathbb{Z}}, \bar{\mathbb{P}}, \bar{\theta})$, that is $\mathbb{1}_A = \mathbb{1}_A \circ \bar{\theta}$. We will show that $\bar{\mathbb{P}}(A) = 0$ or 1.

Note first that $\mathscr{X}^{\otimes\mathbb{Z}} = \sigma(\mathscr{F}^+_{-k}, k \in \mathbb{N})$ where $\mathscr{F}^+_\ell = \sigma(X_i, \ell \le i < \infty)$ and $\{X_i,\ i \in \mathbb{Z}\}$ are the coordinate process on $\mathsf{X}^{\mathbb{Z}}$. This allows to apply the approximation **??** showing that for all $\varepsilon > 0$, there exists $k_\varepsilon \in \mathbb{N}^*$ and a $\mathscr{F}^+_{-k_\varepsilon}$-measurable random variable Z_ε such that $\bar{\mathbb{E}}[|Z_\varepsilon|] < \infty$ and $\bar{\mathbb{E}}[|\mathbb{1}_A - Z_\varepsilon|] \le \varepsilon$. Set $Y_\varepsilon = Z_\varepsilon \circ \bar{\theta}_{k_\varepsilon}$. By construction, Y_ε is $\mathscr{F}^+_0$-measurable. Using that A is an invariant set, we obtain

$$\bar{\mathbb{E}}[|\mathbb{1}_A - Y_\varepsilon|] = \bar{\mathbb{E}}[|\mathbb{1}_A \circ \bar{\theta}_k - Z_\varepsilon \circ \bar{\theta}_k|] = \bar{\mathbb{E}}[|\mathbb{1}_A - Z_\varepsilon|] \le \varepsilon \,.$$

Since ε is arbitrary, there exists a $\mathscr{F}^+_0$-measurable random variable Y satisfying $\bar{\mathbb{E}}[|Y|] < \infty$ and $\mathbb{1}_A = Y$, $\bar{\mathbb{P}}$ − a.s. Since $1 = \bar{\mathbb{P}}(\mathbb{1}_A = Y) \le \bar{\mathbb{P}}(Y \in \{0,1\}) \le 1$ there exists $B \in \mathscr{F}^+_0$ such that

$$\mathbb{1}_B = Y = \mathbb{1}_A \,, \quad \bar{\mathbb{P}} - \text{a.s.} \tag{5.7}$$

Eq. (5.7) and the invariance of A then shows that

$$\bar{\mathbb{P}}(\mathbb{1}_B \circ \bar{\theta} = \mathbb{1}_A \circ \bar{\theta} = \mathbb{1}_A = \mathbb{1}_B) = 1 \,.$$

Now, note that $\mathscr{F}^+_0 = \sigma(\Pi)$, the σ-algebra generated by Π, where the canonical projection $\Pi : \mathsf{X}^{\mathbb{Z}} \to \Omega$ is defined in (5.6). Then, since $B \in \mathscr{F}^+_0$, there exists $C \in \mathscr{F}^+$ such that $B = \Pi^{-1}(C)$ and thus,

$$\begin{aligned}1 &= \bar{\mathbb{P}}(\mathbb{1}_B = \mathbb{1}_B \circ \bar{\theta}) \\ &= \bar{\mathbb{P}}(\mathbb{1}_C \circ \Pi = \mathbb{1}_C \circ \Pi \circ \bar{\theta}) \\ &\overset{(i)}{=} \bar{\mathbb{P}}(\mathbb{1}_C \circ \Pi = \mathbb{1}_C \circ \theta \circ \Pi) = \bar{\mathbb{P}} \circ \Pi^{-1}(\mathbb{1}_C = \mathbb{1}_C \circ \theta) \overset{(ii)}{=} \mathbb{P}(\mathbb{1}_C = \mathbb{1}_C \circ \theta) \,,\end{aligned}$$

where $\overset{(i)}{=}$ follows from $\Pi \circ \bar{\theta} = \theta \circ \Pi$ and $\overset{(ii)}{=}$ from $\mathbb{P} = \bar{\mathbb{P}} \circ \Pi^{-1}$. The dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}, \theta)$ being ergodic, it implies that $\mathbb{P}(C) = 0$ or 1 which concludes the proof since

$$\mathbb{P}(C) = \bar{\mathbb{P}} \circ \Pi^{-1}(C) = \bar{\mathbb{P}}(B) = \bar{\mathbb{P}}(A) \,.$$

■

Proposition 5.13 allows to study only the ergodicity of dynamical systems on one-sided sequences (instead of two-sided sequences) and then to use Birkhoff's ergodic theorem either to functions depending on the future $n^{-1}\sum_{k=0}^{n-1} f(\{X_{k+\ell},\ \ell \in \mathbb{N}\})$ or even on the whole past $n^{-1}\sum_{k=0}^{n-1} f(\{X_{k-\ell},\ \ell \in \mathbb{N}\})$. From now on, we only consider the ergodicity of dynamical systems associated to one-sided sequences.

5.2 Markov chains ergodicity

We specialize the results of the previous section in the context of Markov chains. Here and subsequently, we consider a Markov kernel P on a measurable space $(\mathsf{X}, \mathscr{X})$ and the coordinate process $\{X_k,\ k \in \mathbb{N}\}$ on the

canonical space $(\Omega, \mathscr{F}) = (\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}})$, endowed with the family of probability measures $\mathbb{P}_\xi$, $\xi \in \mathbb{M}_1(\mathscr{X})$ under which the coordinate process is a Markov chain with kernel P and initial distribution ξ.

As a consequence of Birkhoff's ergodic theorem and of Corollary 5.11, we obtain the ergodic theorem for Markov chains.

Theorem 5.14. *Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$. Assume that P admits an invariant probability measure π and that the associated dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is ergodic. Then, for all random variables $Y \in \mathsf{L}^1(\mathbb{P}_\pi)$,*

$$\lim_{n\to\infty} n^{-1} \sum_{k=0}^{n-1} Y \circ \theta_k = \mathbb{E}_\pi[Y] \qquad \mathbb{P}_\pi - \text{a.s.}$$

Moreover, the convergence also holds in $\mathsf{L}^1(\mathbb{P}_\pi)$.

The condition $Y \in \mathsf{L}^1(\mathbb{P}_\pi)$ may be relaxed to $\mathbb{E}_\pi[Y^+] < \infty$ as shown by Exercise 5.33.

Definition 5.15 (Absorbing set). A set $B \in \mathscr{X}$ is called absorbing if $P(x,B) = 1$ for all $x \in B$.

This definition subsumes that the empty set is absorbing. Of course the interesting absorbing sets are non-empty.

Proposition 5.16 *Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$ admitting an invariant probability measure π. If $B \in \mathscr{X}$ is an absorbing set, then $\pi_B = \pi(B \cap \cdot)$ is an invariant finite measure. Moreover, if the invariant probability measure is unique, then $\pi(B) \in \{0,1\}$.*

PROOF. Let B be an absorbing set. Using that $\pi_B \le \pi$, $\pi P = \pi$ and B is absorbing, we get that for all $C \in \mathscr{X}$,

$$\pi_B P(C) = \pi_B P(C \cap B) + \pi_B P(C \cap B^c) \le \pi P(C \cap B) + \pi_B P(B^c) = \pi(C \cap B) = \pi_B(C)\,.$$

Replacing C by C^c and noting that $\pi_B P(\mathsf{X}) = \pi_B(\mathsf{X}) < \infty$ show that π_B is an invariant finite measure. To complete the proof, assume that P has a unique invariant probability measure. If $\pi(B) > 0$ then, $\pi_B/\pi(B)$ is an invariant probability measure and is therefore equal to π. Since $\pi_B(B^c) = 0$, we get $\pi(B^c) = 0$. Thus, $\pi(B) \in \{0,1\}$. ■

We now relate harmonic functions (defined in **??**) with invariant random variables for the shift transformation θ.

Proposition 5.17 *Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$.*

(i) Let Y be a bounded invariant random variable for the shift transformation θ. Then the function $h_Y : x \mapsto h_Y(x) = \mathbb{E}_x[Y]$ is a bounded harmonic function.

(ii) Let h be a bounded harmonic function and define $Y = \limsup_{n\to\infty} h(X_n)$. Then Y is an invariant random variable for θ and for any $\xi \in \mathbb{M}_1(\mathscr{X})$, the sequence $\{h(X_n),\, n \in \mathbb{N}\}$ converges to Y $\mathbb{P}_\xi$ − a.s. and in $\mathsf{L}^1(\mathbb{P}_\xi)$. Moreover, $h(x) = \mathbb{E}_x[Y]$ for all $x \in \mathsf{X}$.

(iii) Let π be an invariant probability measure and $Y \in \mathsf{L}^1(\mathbb{P}_\pi)$ be an invariant random variable for θ. Then, $\mathbb{E}_x[|Y|] < \infty$ π − a.e., the function $x \mapsto \mathbb{E}_x[Y]$ is π-integrable and $Y = \mathbb{E}_{X_0}[Y]$ $\mathbb{P}_\pi$ − a.s.

PROOF.

(i) Assume that $Y : \mathsf{X}^{\mathbb{N}} \to \mathbb{R}$ is a bounded invariant random variable, i.e. $Y \circ \theta = Y$. By the Markov property, for any $x \in \mathsf{X}$,

$$Ph_Y(x) = \mathbb{E}_x[h_Y(X_1)] = \mathbb{E}_x[\mathbb{E}_{X_1}[Y]] = \mathbb{E}_x[Y \circ \theta_1] = \mathbb{E}_x[Y] = h_Y(x) \,,$$

showing that h_Y is harmonic.

(ii) Let h be a bounded harmonic function: $Ph(x) = h(x)$ for all $x \in \mathsf{X}$. Then, $\{(h(X_n), \mathscr{F}_n),\ n \in \mathbb{N}\}$ is a bounded $\mathbb{P}_\xi$-martingale, for any initial distribution $\xi \in \mathbb{M}_1(\mathscr{X})$. By Doob's martingale convergence theorem, the sequence $\{h(X_n), n \in \mathbb{N}\}$ converges $\mathbb{P}_\xi$ – a.s. and in $\mathsf{L}^1(\mathbb{P}_\xi)$ to a limit. Hence, we get

$$Y = \lim_{n\to\infty} h(X_n)\ \mathbb{P}_\xi \text{ – a.s. and } \mathbb{E}_\xi[Y] = \lim_{n\to\infty} \mathbb{E}_\xi[h(X_n)]. \tag{5.8}$$

The function h being harmonic, we have $h(x) = P^n h(x) = \mathbb{E}_x[h(X_n)]$ for all $x \in \mathsf{X}$ and $n \in \mathbb{N}$. Applying (5.8) with $\xi = \delta_x$

$$\mathbb{E}_x[Y] = \lim_{n\to\infty} \mathbb{E}_x[h(X_n)] = h(x) \,.$$

(iii) Since $Y \in \mathsf{L}^1(\mathbb{P}_\pi)$, $\mathbb{E}_\pi[|Y|] = \int_{\mathsf{X}} \pi(\mathrm{d}x)\mathbb{E}_x[|Y|]$ showing that $\mathbb{E}_x[|Y|] < \infty$ π-a.e. and that the function $x \mapsto \mathbb{E}_x[Y]$ is integrable with respect to π. By the Markov property and the invariance of Y, we get

$$\mathbb{E}_{X_k}[Y] = \mathbb{E}_\pi[Y \circ \theta_k | \mathscr{F}_k] = \mathbb{E}_\pi[Y | \mathscr{F}_k] \qquad \mathbb{P}_\pi \text{ – a.s.}$$

Therefore, $\{(\mathbb{E}_{X_k}[Y], \mathscr{F}_k), k \in \mathbb{N}\}$ is a uniformly integrable $\mathbb{P}_\pi$-martingale. By **??**,

$$\lim_{k\to\infty} \mathbb{E}_{X_k}[Y] = \lim_{k\to\infty} \mathbb{E}_\pi[Y | \mathscr{F}_k] = \mathbb{E}_\pi[Y | \mathscr{F}] = Y \qquad \mathbb{P}_\pi \text{ – a.s.} \tag{5.9}$$

and in $\mathsf{L}^1(\mathbb{P}_\pi)$. Moreover, applying successively that the translation operator θ is measure preserving for $\mathbb{P}_\pi$ and $Y = Y \circ \theta_k$, we obtain for any $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_\pi[|Y - \mathbb{E}_{X_0}[Y]|] &= \mathbb{E}_\pi[|Y - \mathbb{E}_{X_0}[Y]| \circ \theta_k] \\ &= \mathbb{E}_\pi[|Y \circ \theta_k - \mathbb{E}_{X_k}[Y]|] = \mathbb{E}_\pi[|Y - \mathbb{E}_{X_k}[Y]|] \,.\end{aligned}$$

Taking the limit as k goes to infinity, (5.9) yields

$$\mathbb{E}_\pi[|Y - \mathbb{E}_{X_0}[Y]|] = \lim_{k\to\infty} \mathbb{E}_\pi[|Y - \mathbb{E}_{X_k}[Y]|] = 0 \,.$$

■

Remark 5.18. Proposition 5.17 shows that the map $Y \mapsto h_Y$, where $h_Y(x) = \mathbb{E}_x[Y]$, $x \in \mathsf{X}$ defines a one-to-one correspondence between the bounded harmonic functions and the bounded invariant random variables. If Y is a bounded invariant random variable, then $h_Y : x \mapsto h_Y(x) = \mathbb{E}_x[Y]$ is a bounded harmonic function. If h is a bounded harmonic function, then $h(x) = \mathbb{E}_x[Y]$ where $Y = \limsup_{n\to\infty} h(X_n)$ is an invariant random variable (hence $h = h_Y$). ▲

Corollary 5.19 *Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$. The following statements are equivalent.*

(i) The bounded harmonic functions are constant.
(ii) The invariant σ-field $\mathscr{I}$ is trivial up to an equivalence, i.e. for all $A \in \mathscr{I}$, we get for all $\xi \in \mathbb{M}_1(\mathscr{X})$, $\mathbb{P}_\xi(A) = 0$ or $\mathbb{P}_\xi(A) = 1$.

PROOF. (i) $\Rightarrow$ (ii): Let A be an invariant set. Then $h_A : x \mapsto h_A(x) = \mathbb{P}_x(A)$ is a harmonic function by Proposition 5.17 which is constant under (i), i.e. $h_A(x) = c$ for all $x \in \mathsf{X}$. By the Markov property, we get that, for all $\xi \in \mathbb{M}_1(\mathscr{X})$, $h_A(X_n) = \mathbb{E}_{X_n}[\mathbb{1}_A] = \mathbb{E}_\xi[\,\mathbb{1}_A \circ \theta_n \,|\, \mathscr{F}_n]$ $\mathbb{P}_\xi$ – a.s. Since $A \in \mathscr{I}$, it holds that $\mathbb{E}_\xi[\,\mathbb{1}_A \circ \theta_n \,|\, \mathscr{F}_n] = \mathbb{E}_\xi[\,\mathbb{1}_A \,|\, \mathscr{F}_n]$ $\mathbb{P}_\xi$ – a.s. and **??** shows that $\mathbb{E}_\xi[\,\mathbb{1}_A \,|\, \mathscr{F}_n] \overset{\mathbb{P}_\xi\text{-a.s.}}{\longrightarrow} \mathbb{1}_A$ $\mathbb{P}_\xi$ – a.s. which implies that $c \in \{0, 1\}$.

(ii) $\Rightarrow$ (i): Let h be a bounded harmonic function and $\xi \in \mathbb{M}_1(\mathscr{X})$. The random variable $Y = \limsup_{n\to\infty} h(X_n)$ is invariant and under (ii), there exists a constant $c < \infty$ (possibly depending on ξ) such that $\limsup_{n\to\infty} h(X_n) = c$ $\mathbb{P}_\xi$ – a.s. By Proposition 5.17 $\mathbb{E}_x[Y] = c = h(x)$ for all $x \in \mathsf{X}$. Therefore, h is constant which shows (i). ■

If we wish to obtain the Law of Large Numbers for a particular Markov chain by applying Theorem 5.14, we have to check the ergodicity assumption. It is therefore convenient to have sufficient conditions ensuring ergodicity.

We now give a sufficient condition for $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ to be ergodic expressed in terms of absorbing sets.

Lemma 5.20 *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting an invariant probability measure π. If for all absorbing sets $B\in\mathscr{X}$, $\pi(B)\in\{0,1\}$, then the dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is ergodic.*

PROOF. Let $A\in\mathscr{I}$ and define $h(x)=\mathbb{E}_x[\mathbb{1}_A]$ and $B=\{x\in\mathsf{X}: h(x)=1\}$. By Proposition 5.17-(i), h is a nonnegative harmonic function bounded by 1. For any $x\in B$ we have $\mathbb{E}_x[h(X_1)]=Ph(x)=h(x)=1$, which implies $\mathbb{P}_x(h(X_1)=1)=1$. Therefore for any $x\in B$, we get $\mathbb{P}_x(X_1\in B)=\mathbb{P}_x(h(X_1)=1)=1$. Therefore B is absorbing and hence, under the stated assumption, we have $\pi(B)\in\{0,1\}$.

By Proposition 5.17 (iii), we know that $\mathbb{P}_\pi(\mathbb{E}_{X_0}[\mathbb{1}_A]=\mathbb{1}_A)=1$ which implies that $\mathbb{P}_\pi(h(X_0)\in\{0,1\})=1$. This yields

$$\begin{aligned}\mathbb{P}_\pi(A)=\mathbb{E}_\pi[\mathbb{E}_{X_0}[\mathbb{1}_A]]&=\int_{\mathsf{X}}\pi(\mathrm{d}x)h(x)\\&=\int_{\mathsf{X}}\pi(\mathrm{d}x)\mathbb{1}\{h(x)=1\}=\int_{\mathsf{X}}\pi(\mathrm{d}x)\mathbb{1}_B(x)=\pi(B)\,.\end{aligned}$$

Thus $\mathbb{P}_\pi(A)\in\{0,1\}$ and $\mathscr{I}$ is trivial for $\mathbb{P}_\pi$. ■

It turns out that the sufficient condition in Lemma 5.20 is also a necessary condition. Before showing the necessary part, we first draw an easy and useful consequence of Lemma 5.20.

Theorem 5.21. *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting a unique invariant probability measure π. The dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is ergodic.*

PROOF. By Proposition 5.16, since P has unique invariant probability π, for every absorbing set B, $\pi(B)\in\{0,1\}$. We conclude by Lemma 5.20. ■

The uniqueness of the invariant probability measure is a sufficient but not a necessary condition for ergodicity as illustrated in Exercise 5.35. Comparing with Lemma 5.20, the following Lemma goes one step further. When the dynamical system is not ergodic, the state space X contains not only one but at least two disjoints absorbing sets which are not trivial with respect to π.

Lemma 5.22 *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting an invariant probability measure π. If the dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is not ergodic, then, there exist two disjoint absorbing sets B and B' in $\mathscr{X}$ such that $\pi(B)=1-\pi(B')\in(0,1)$ and $\pi_B(\cdot)=\pi(B\cap\cdot)/\pi(B)$ and $\pi_{B'}(\cdot)=\pi(B'\cap\cdot)/\pi(B')$ are invariant probability measures.*

PROOF. Since the dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is not ergodic, there exists $A\in\mathscr{I}$ such that $\mathbb{P}_\pi(A)=\alpha\in(0,1)$. As $\mathscr{I}$ is a σ-field, we also have $A^c\in\mathscr{I}$. Define $B=\{x\in\mathsf{X}:\mathbb{E}_x[\mathbb{1}_A]=1\}$ and $B'=\{x\in\mathsf{X}:\mathbb{E}_x[\mathbb{1}_{A^c}]=1\}$. As noted in the proof of Lemma 5.20 (see also Exercise 5.31), the sets B and B' are absorbing and

$$\pi(B)=\mathbb{P}_\pi(A)=1-\mathbb{P}_\pi(A^c)=1-\pi(B')\in(0,1)\,.$$

By Proposition 5.16, π_B and $\pi_{B'}$ are invariant probability measures. ■

Without ergodicity assumption, the generalized version of Birkhoff's ergodic theorem in Theorem 5.8 shows that the normalized partial sums still converges but the limit is a random variable which is not necessarily almost surely constant (see also an illustration in Exercise 5.35). In the context of Markov chains this limit turns out to be a function of X_0. More precisely we have the following theorem.

Proposition 5.23 *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting an invariant probability measure π and let $Y\in\mathsf{L}^1(\mathbb{P}_\pi)$. Let Z be a version of the conditional expectation $\mathbb{E}_\pi[Y\,|\,\mathscr{I}]$, i.e. $Z=\mathbb{E}_\pi[Y\,|\,\mathscr{I}]$ $\mathbb{P}_\pi-$a.s. Then, there exists a set $S\in\mathscr{X}$, such that $\pi(S)=1$ and for each $x\in S$,*

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\theta_k = \mathbb{E}_x[Z] \qquad \mathbb{P}_x - \text{a.s.} \tag{5.10}$$

PROOF. Define $\phi(x) = \mathbb{E}_x[Z]$. It follows from Proposition 5.17 (iii) that $\mathbb{E}_\pi[Y\,|\,\mathscr{I}] = \phi(X_0)$, $\mathbb{P}_\pi$ − a.s. Hence, Theorem 5.8 yields

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} Y\circ\theta_k = \mathbb{E}_\pi[Y|\,\mathscr{I}] = \phi(X_0) \qquad \mathbb{P}_\pi - \text{a.s.}$$

Set $A = \left\{\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} Y\circ\theta_k = \phi(X_0)\right\}$. The previous relation implies $\mathbb{P}_\pi(A) = 1$, i.e. $\int \pi(\mathrm{d}x)\mathbb{P}_x(A) = 1$. Since $\mathbb{P}_x(A) \le 1$ for all $x\in\mathsf{X}$, this implies that $\mathbb{P}_x(A) = 1$ for π-almost all $x\in\mathsf{X}$. Setting $S = \{x\in\mathsf{X}:\ \mathbb{P}_x(A) = 1\}$ concludes the proof. ■

In Proposition 5.23, the limit $\mathbb{E}_x[Z]$ in (5.10) is expressed in terms of Z, a version of the conditional expectation $\mathbb{E}_\pi[Y\,|\,\mathscr{I}]$. If we choose another version of $\mathbb{E}_\pi[Y\,|\,\mathscr{I}]$, say Z', under $\mathbb{P}_\pi$, then obviously, $Z = Z'$ $\mathbb{P}_\pi$ − a.s. but we do not necessarily have $\mathbb{E}_x[Z'] = \mathbb{E}_x[Z]$ $\mathbb{P}_x$ − a.s. since without additional assumption, $\mathbb{P}_x$ is not necessarily dominated by $\mathbb{P}_\pi$. The situation is different when the dynamical system is ergodic since the limit is then $\mathbb{P}_\pi$ − a.s. constant.

Theorem 5.24 (Birkhoff's Theorem for Markov chains). *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ and assume that P admits an invariant probability measure π such that $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,\theta)$ is ergodic. Let $Y\in\mathsf{L}^1(\mathbb{P}_\pi)$. Then, for π-almost all $x\in\mathsf{X}$,*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} Y\circ\theta_k = \mathbb{E}_\pi[Y] \qquad \mathbb{P}_x - \text{a.s.}$$

PROOF. Since π is ergodic, the invariant σ-field $\mathscr{I}$ is trivial for $\mathbb{P}_\pi$. This implies $\mathbb{E}_\pi[Y\,|\,\mathscr{I}] = \mathbb{E}_\pi[Y]$ $\mathbb{P}_\pi$ − a.s. ■

Theorem 5.25. *Let P a Markov kernel on $\mathsf{X}\times\mathscr{X}$. If π_1 and π_2 are distinct invariant probability measures such that $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\pi_1},\theta)$ and $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\pi_2},\theta)$ are ergodic, then π_1 and π_2 are mutually singular.*

PROOF. Note first that, if $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,\theta)$ is ergodic and $f\in\mathbb{F}_b(\mathsf{X})$, then applying Theorem 5.24 to the random variable $Y = f(X_0)$ and the dominated convergence theorem, we obtain that there exists a set $S\in\mathscr{X}$, such that $\pi(S) = 1$ and for all $x\in S$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} P^k f(x) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}_x[f(X_k)] = \pi(f)\,.$$

Now assume that π_1 and π_2 are different invariant probability measures such that $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\pi_1},\theta)$ and $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\pi_2},\theta)$ are ergodic. Let $C\in\mathscr{X}$ such that $\pi_1(C)\neq\pi_2(C)$ and set, for $i = 1,2$,

$$S_i = \left\{x\in\mathsf{X}:\ \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} P^k\mathbb{1}_C(x) = \pi_i(C)\right\},$$

We have $S_1\cap S_2 = \emptyset$, $\pi_1(S_1) = 1$ and $\pi_2(S_2) = 1$, which means that π_1 and π_2 are mutually singular. ■

We have now all the tools for getting a necessary and sufficient condition for the dynamical system to be ergodic.

Theorem 5.26. *Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting an invariant probability measure π. The dynamical system $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,\theta)$ is ergodic if and only if for all absorbing sets $B\in\mathscr{X}$, $\pi(B)\in\{0,1\}$.*

PROOF. The sufficient condition follows from Lemma 5.20. We now consider an ergodic dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ and we let B be an absorbing set. Assume first that $\pi(B) > 0$. Then, by Proposition 5.16, $\bar{\pi}_B(\cdot) = \pi(B \cap \cdot)/\pi(B)$ is an invariant probability measure. Moreover, note that for all $A \in \mathscr{X}^{\otimes\mathbb{N}}$,

$$\mathbb{P}_{\bar{\pi}_B}(A) = \int \pi(\mathrm{d}x)\mathbb{P}_x(A)\frac{\mathbb{1}_B(x)}{\pi(B)} \leq \frac{\mathbb{P}_\pi(A)}{\pi(B)} .$$

Combining with the ergodicity of the dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ we deduce that any invariant set $A \in \mathscr{I}$ satisfies either $0 = \mathbb{P}_\pi(A) = \mathbb{P}_{\bar{\pi}_B}(A)$ or $0 = \mathbb{P}_\pi(A^c) = \mathbb{P}_{\bar{\pi}_B}(A^c)$. The dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_{\bar{\pi}_B}, \theta)$ is therefore ergodic and by Theorem 5.25, $\bar{\pi}_B = \pi$ since they are not mutually singular. This implies that $\pi(B^c) = \bar{\pi}_B(B^c) = 0$. Finally, $\pi(B) \in \{0,1\}$, which concludes the proof. ■

In Proposition 5.23 and Theorem 5.24, the law of large numbers is obtained under $\mathbb{P}_x$ for all x belonging to a set S such that $\pi(S) = 1$ and which may depend on the random variable Y under consideration. This is unsatisfactory since it does not tell if the LLN holds for a given $x \in \mathsf{X}$ or more generally for a given initial distribution ξ. We now give a criterion to obtain the LLN when the chain does not start from stationarity. Recall that the total variation distance between two probability measures $\mu, \nu \in \mathbb{M}_1(\mathscr{X})$ is defined by

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{h \in \mathbb{F}_b(\mathsf{X}), |h|_\infty \leq 1} |\mu(h) - \nu(h)| .$$

More details and basic properties on the total variation distance are given in **??**.

Proposition 5.27 *Let P be a Markov kernel on $\mathsf{X} \times \mathscr{X}$ admitting an invariant probability measure π. If the dynamical system $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is ergodic and if $\xi \in \mathbb{M}_1(\mathscr{X})$ is such that $\lim_{n\to\infty} \left\|n^{-1}\sum_{k=1}^n \xi P^k - \pi\right\|_{\mathrm{TV}} = 0$ then for all $Y \in \mathsf{L}^1(\mathbb{P}_\pi)$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} Y \circ \theta_k = \mathbb{E}_\pi[Y] \qquad \mathbb{P}_\xi - \text{a.s.}$$

PROOF. Set $A = \{\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} Y \circ \theta_k = \mathbb{E}_\pi[Y]\}$. Since $(\mathsf{X}^{\mathbb{N}}, \mathscr{X}^{\otimes\mathbb{N}}, \mathbb{P}_\pi, \theta)$ is ergodic, we already know that $\mathbb{P}_\pi(A) = 1$. We show that $\mathbb{P}_\xi(A) = 1$. Define the function h by $h(x) = \mathbb{E}_x[\mathbb{1}_A]$. Since $A \in \mathscr{I}$, Proposition 5.17 implies that h is harmonic. Then, for all $n \in \mathbb{N}$, $n^{-1}\sum_{k=1}^n \xi P^k h = \xi(h)$. Moreover, noting that $h \leq 1$, we have by assumption, $\lim_{n\to\infty} n^{-1}\sum_{k=1}^n \xi P^k(h) = \pi(h)$. Thus, $\xi(h) = \pi(h)$ and

$$\mathbb{P}_\xi(A) = \int_{\mathsf{X}} \xi(\mathrm{d}x)\mathbb{E}_x[\mathbb{1}_A] = \xi(h) = \pi(h) = \mathbb{P}_\pi(A) = 1 .$$

■

The condition $\lim_n \|\xi P^n - \pi\|_{\mathrm{TV}} = 0$ is not mandatory for having a Law of Large Numbers under $\mathbb{P}_\xi$. In some situations, one can get the same result without any straightforward information in the decrease of $\|\xi P^n - \pi\|_{\mathrm{TV}}$ toward 0. This is the case for example for Metropolis-Hastings kernels as illustrated in Exercise 5.36. Another illustration can be found in Exercise 5.41 where the Law of Large Numbers is extended to different initial distributions in the case where (X, d) is a complete separable metric space.

5.3 Exercises

Exercise 5.28. Let $(\Omega, \mathscr{B}, \mathbb{P}, \mathrm{T})$ be a dynamical system. Show that $\mathscr{I} \neq \cap_{k\geq 0}\sigma(X_l, l > k)$.

Exercise 5.29. Let $(\Omega, \mathscr{B}, \mathbb{P})$ be a probability space and $\theta : \Omega \to \Omega$ be a measurable transformation. Let $\mathscr{B}_0$ a family of sets, stable under finite intersection and generating $\mathscr{B}$. If for all $B \in \mathscr{B}_0$, $\mathbb{P}[\theta^{-1}(B)] = \mathbb{P}(B)$, then T is measure-preserving.

Exercise 5.30. Let $(\Omega, \mathscr{B}, \mathbb{P}, \mathrm{T})$ be a dynamical system. Let Y be a $\bar{\mathbb{R}}$-valued random variable such that $\mathbb{E}[Y^+] < \infty$. Show that for all $k \geq 0$,

$$\mathbb{E}\left[Y\circ \mathrm{T}^k \,\middle|\, \mathscr{I}\right] = \mathbb{E}\left[Y \,|\, \mathscr{I}\right] \qquad \mathbb{P}-\text{a.s.}$$

Exercise 5.31. Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$ and let $(\Omega,\mathscr{B})$ be the canonical space. For $A\in\mathscr{I}$, define $B=\{x\in\mathsf{X}:\ \mathbb{P}_x(A)=1\}$.

1. Show that B is absorbing.
2. Let π be an invariant probability. Show that $\pi(A)=\mathbb{P}_\pi(B)$.

Exercise 5.32. The following exercise provides the converse of Theorem 5.14. Let P be a Markov kernel on $\mathsf{X}\times\mathscr{X}$. Let π be a probability measure, $\pi\in\mathbb{M}_1(\mathscr{X})$. Assume that for all $f\in\mathbb{F}_b(\mathsf{X})$, we get

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k) = \pi(f) \qquad \mathbb{P}_\pi-\text{a.s.}$$

1. Show that π is invariant.
2. Let $A\in\mathscr{I}$ and set $B=\{x\in\mathsf{X}:\ \mathbb{P}_x(A)=1\}$. Show that

$$\mathbb{1}_A = \mathbb{P}_{X_0}(A) = \mathbb{1}_B(X_0) \qquad \mathbb{P}_\pi-\text{a.s.}$$

   and that for all $k\in\mathbb{N}$, $\mathbb{1}_A=\mathbb{1}_B(X_0)=\cdots=\mathbb{1}_B(X_k)$ $\mathbb{P}_\pi-$a.s..
3. Show that the dynamical system $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,\theta)$ is ergodic.

Exercise 5.33. In this exercise, we prove various extensions of Birkhoff's ergodic theorem. Let $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$ be a dynamical system. In the first two questions, we assume that the dynamical system is ergodic.

1. Let Y be nonnegative random variable such that $\mathbb{E}[Y]=\infty$. Show that

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\mathrm{T}^k = \infty \qquad \mathbb{P}-\text{a.s.}$$

   [Hint: use Corollary 5.11 with $Y_M=Y\wedge M$ and let M tends to infinity.]
2. Let Y be a random variable such that $\mathbb{E}[Y^+]<\infty$. Show that

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\mathrm{T}^k = \mathbb{E}[Y] \qquad \mathbb{P}-\text{a.s.}$$

3. In what follows, we do not assume any ergodicity of the dynamical system $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$. Let Y be a nonnegative random variable. Set

$$A=\{\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\mathrm{T}^k = \mathbb{E}\left[Y\,|\,\mathscr{I}\right]\}$$

   Let $M>0$. Using Theorem 5.8 with $Y\,\mathbb{1}\{\mathbb{E}\left[Y\,|\,\mathscr{I}\right]\le M\}$, show that on $\{\mathbb{E}\left[Y\,|\,\mathscr{I}\right]\le M\}$

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\mathrm{T}^k = \mathbb{E}\left[Y\,|\,\mathscr{I}\right] \qquad \mathbb{P}-\text{a.s.}$$

   Deduce that $\mathbb{P}(A^c\cap\{\mathbb{E}\left[Y\,|\,\mathscr{I}\right]<\infty\})=0$. Moreover, show that for all $M>0$,

$$\liminf_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\mathrm{T}^k \ge \mathbb{E}\left[Y\wedge M\,|\,\mathscr{I}\right] \qquad \mathbb{P}-\text{a.s.}$$

   Deduce that $\mathbb{P}(A^c\cap\{\mathbb{E}\left[Y\,|\,\mathscr{I}\right]=\infty\})=0$.
4. Let Y be a random variable such that $\mathbb{E}[Y^+]<\infty$, show that

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ \mathrm{T}^k = \mathbb{E}\left[Y \,|\, \mathscr{I}\right] \qquad \mathbb{P}-\text{a.s.}$$

[Hint: write $Y = Y^+ - Y^-$ and use the previous question with Y^-]

Exercise 5.34. Let P be the kernel on $\mathsf{X}\times\mathscr{X}$ defined by for all $(x,A)\in\mathsf{X}\times\mathscr{X}$, $P(x,A)=\delta_x(A)$. Find all the probability measures $\pi\in\mathbb{M}_1(\mathscr{X})$ such that $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_\pi,\theta)$ is ergodic.

Exercise 5.35. Let μ_0 (resp. μ_1) be a probability measure on $\mathbb{R}^+$ (resp. $\mathbb{R}^-\setminus\{0\}$). Let P be a Markov kernel on $\mathbb{R}\times\mathscr{B}(\mathbb{R})$ defined by $P(x,\cdot)=\mu_0$ if $x\geq 0$ and $P(x,\cdot)=\mu_1$ otherwise. Set for all $\alpha\in(0,1)$, $\mu_\alpha=(1-\alpha)\mu_0+\alpha\mu_1$.

1. Show that for all $\alpha\in(0,1)$, μ_α is an invariant probability measure but the dynamical system $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\mu_\alpha},\theta)$ is not ergodic.
2. Show that for $\alpha\in\{0,1\}$, $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\mu_\alpha},\theta)$ is ergodic.
3. Let $f\in\mathsf{L}^1(\mu_0)\cap\mathsf{L}^1(\mu_1)$. Find a function ϕ such that for all $x\in\mathbb{R}$,

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} f(X_k) = \phi(X_0) \qquad \mathbb{P}_{\mu_\alpha}-\text{a.s.}$$

For all $\alpha\in(0,1)$, find a version of $\mathbb{E}_{\mu_\alpha}[f|\mathscr{I}]$.

Exercise 5.36. In this exercise, we will find a sufficient condition for a Metropolis-Hastings kernel to satisfy the law of large numbers starting from any initial distribution. We make use of the notation of Section 4.1.1. Let π be target distribution on a measurable space $(\mathsf{X},\mathscr{X})$ and assume that π has a positive density h with respect to a measure $\mu\in\mathbb{M}_+(\mathscr{X})$. Let Q be a proposal kernel on $\mathsf{X}\times\mathscr{X}$ and assume that Q has a positive kernel density $y\mapsto q(x,y)$ with respect to μ. The Metropolis-Hasting kernel P is then defined by (4.4).

1. Show that
$$P(x,A)\geq\int_A \alpha(x,y)q(x,y)\mu(\mathrm{d}y)\,.$$
Deduce that π is the unique invariant probability of P.
2. Let $A\in\mathscr{I}$ and assume that $\mathbb{P}_\pi(A)=0$. Set $\phi(x)=\mathbb{P}_x(A)$. For all $x\in\mathsf{X}$, show that
$$\phi(x)=P\phi(x)=\int\frac{\alpha(x,y)q(x,y)}{h(y)}\pi(\mathrm{d}y)\phi(y)+\bar{\alpha}(x)\phi(x)\,.$$
and deduce that $\mathbb{P}_x(A)=0$.
3. Let $\xi\in\mathbb{M}_1(\mathscr{X})$. Deduce from the previous question that for all random variables $Y\in\mathsf{L}^1(\mathbb{P}_\pi)$,

$$\lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} Y\circ\theta_k = \mathbb{E}_\pi[Y] \qquad \mathbb{P}_\xi-\text{a.s.}$$

Exercise 5.37. Let P a Markov kernel on $\mathsf{X}\times\mathscr{X}$ admitting an invariant probability measure π_1 such that $(\mathsf{X}^{\mathbb{N}},\mathscr{X}^{\otimes\mathbb{N}},\mathbb{P}_{\pi_1},\theta)$ is ergodic. Let μ be another invariant probability measure.

1. Show that $\pi_1\wedge\mu$ is an invariant finite measure.
2. If $\pi_1\wedge\mu(\mathsf{X})\neq 0$, show, using Theorem 5.25, that $\pi_1\wedge\mu/\pi_1\wedge\mu(\mathsf{X})=\pi_1$.
3. Show that there exists an invariant probability measure π_2 satisfying: π_1 and π_2 are mutually singular and there exists $\alpha\in[0,1]$ such that $\mu=\alpha\pi_1+(1-\alpha)\pi_2$.

Exercise 5.38. Let $\{X_n,\, n\in\mathbb{Z}\}$ be a canonical stationary Markov chain on $(\mathsf{X}^{\mathbb{Z}},\mathscr{X}^{\otimes\mathbb{Z}})$. We set

$$\overline{\mathscr{F}^0_{-\infty}}=\overline{\sigma(X_k,k\leq 0)}^P,\quad \overline{\mathscr{F}^\infty_0}=\overline{\sigma(X_k,k\geq 0)}^P.$$

We consider an invariant bounded random variable Y.

(i) Show that Y is $\overline{\mathscr{F}_0^\infty}$ and $\overline{\mathscr{F}_{-\infty}^0}$ measurable.
(ii) Deduce from the previous question that $Y = \mathbb{E}[Y\,|X_0]\ \ \mathbb{P}-$a.s.
(iii) Show that the previous identity holds true for all $\mathbb{P}$-integrable or positive invariant random variable Y.

The following exercises deal with subbadditive sequences. A sequence of random variables $\{Y_n,\ n\in\mathbb{N}^\star\}$ is said to be subadditive for the dynamical system $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$ if for all $(n,p)\in\mathbb{N}^\star$, $Y_{n+p}\le Y_n + Y_p\circ\mathrm{T}^n$. The sequence is said to be additive if for all $(n,p)\in\mathbb{N}^\star$, $Y_{n+p} = Y_n + Y_p\circ\mathrm{T}^n$.

Exercise 5.39 (Fekete Lemma). Consider $\{a_n,\ n\in\mathbb{N}^\star\}$, a sequence in $[-\infty,\infty)$ such that, for all $(m,n)\in\mathbb{N}^\star\times\mathbb{N}^\star$, $a_{n+m}\le a_n+a_m$. Then,

$$\lim_{n\to\infty}\frac{a_n}{n} = \inf_{m\in\mathbb{N}^\star}\frac{a_m}{m}\ ;$$

in other words, the sequence $\{n^{-1}a_n, n\in\mathbb{N}^\star\}$ either converges to its lower bounds or diverges to $-\infty$.

Exercise 5.40. Let $(\Omega,\mathscr{B},\mathbb{P},\mathrm{T})$ be a dynamical system and $\{Y_n,\ n\in\mathbb{N}^\star\}$ be a subadditive sequence of functions such that $\mathbb{E}\left[Y_1^+\right]<\infty$. Show that for any $n\in\mathbb{N}^\star$, $\mathbb{E}[Y_n^+]\le n\mathbb{E}\left[Y_1^+\right]<\infty$ and

$$\lim_{n\to\infty} n^{-1}\mathbb{E}[Y_n] = \inf_{n\in\mathbb{N}^\star} n^{-1}\mathbb{E}[Y_n]\ , \tag{5.11}$$

$$\lim_{n\to\infty} n^{-1}\mathbb{E}[Y_n\,|\mathscr{I}] = \inf_{n\in\mathbb{N}^\star} n^{-1}\mathbb{E}[Y_n\,|\mathscr{I}]\ ,\quad \mathbb{P}-\text{a.s.} \tag{5.12}$$

where $\mathscr{I}$ is the invariant σ-field [Hint: use Exercise 5.30].

Exercise 5.41. Let P be a Markov kernel on a complete separable metric space (X,d) which admits a unique invariant probability measure π. We assume that there exists a probability space $(\Omega,\mathscr{F},\mathbb{P})$ and a stochastic process $\{(X_n,X_n'),\ n\in\mathbb{N}\}$ such that $\{X_n,\ n\in\mathbb{N}\}$ and $\{X_n',\ n\in\mathbb{N}\}$ are Markov chains with kernel P and initial distribution $\xi\in\mathbb{M}_1(\mathscr{X})$ and π, respectively. Assume that $\mathrm{d}(X_n,X_n')\xrightarrow{\mathbb{P}\text{-a.s.}}0$.

We recall the Parthasaraty's theorem (see (**?**, Theorem 6.6)): Then there exists a countable set H of bounded continuous functions such that for all $\{\mu,\mu_n,n\ge1\}\subset\mathbb{M}_1(\mathsf{X})$, the following assertions are equivalent:

(i) μ_n converges weakly to μ.
(ii) For all $h\in H$, $\lim_{n\to\infty}\mu_n(h)=\mu(h)$.

For this set H, define the event

$$A = \left\{\omega\in\Omega\ :\ \forall h\in H\ ,\ \lim_{n\to\infty} n^{-1}\sum_{k=0}^{n-1} h(X_k'(\omega)) = \pi(h)\right\}\ .$$

1. Show that $\mathbb{P}(A)=1$.
2. Deduce that there exists a set $\tilde\Omega$ such that $\mathbb{P}(\tilde\Omega)=1$ and for all bounded continuous functions h on X and all $\tilde\omega\in\tilde\Omega$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} h(X_k(\tilde\omega)) = \pi(h)\ .$$

Let V be a nonnegative and uniformly continuous such that $\pi(V)<\infty$.

3. Show that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1} V(X_k) = \pi(V)\ ,\qquad \mathbb{P}-\text{a.s.}$$

4. Show that there exists $\bar\Omega$ such that for all $\omega\in\bar\Omega$ and all continuous functions f such that $\sup_{x\in\mathsf{X}}|f(x)|/V(x)<\infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} f(X_k(\omega)) = \pi(f) \,, \qquad \mathbb{P} - \text{a.s.}$$

5.4 Bibliographical notes

Ergodic theory is a very important area of probability theory which has given rise to a great deal of work. The interested reader will find an introduction to this field in the books **?** and **?**. The application of ergodic theory to Markov chains is a very classic subject. a A detailed study of the ergodic theory of Markov chains can be found in (**?**, Chapter 4) and (**?**, Chapter 2); These books contain many references to works on this subject that began in the early 1960s.

The proof of the Birkhoff Theorem (Theorem 5.8) is borrowed from unpublished notes by B. Delyon and extended to possibly non-ergodic dynamical systems. This approach is closely related to the very short proof of the Law of Large Numbers for i.i.d. random variables written by Jacques Neveu in his unpublished lecture notes at Ecole Polytechnique. Several other proofs of the ergodic theorem are also given in **?**.

Theorem 5.21 is essentially borrowed from (**?**, Proposition 2.4.3) even if the statements of the two results are slightly different.

Chapter 6

Pseudo Marginal Monte Carlo methods and applications

Chapter 7

Hamiltonian Monte Carlo methods

Contents

Let $\pi(q) \propto \mathrm{e}^{-U(q)}$ be a target distribution. We may consider π as the marginal of the extended target

$$\Pi(q,p) \propto \exp\left\{-U(q) - p^2/2\right\} .$$

We make use of the following terminology

- $q \in \mathbb{R}^d$ is the position and $U(q)$ is the potential energy.
- $p \in \mathbb{R}^d$ is the momentum and $K(p) = p^T p/2$ is the kinetic energy.
- $H(q,p) = U(q) + K(p)$ is called the Hamiltonian.

7.1 MH with deterministic moves

If we want a deterministic move to target Π in a Metropolis Hasting algorithm, then the proposal kernel writes: $Q(x, \mathrm{d}y) = \delta_{\varphi(x)}(\mathrm{d}y)$ and the acceptance probability is

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(x,y) \wedge 1 \quad \text{where} \quad \mu(\mathrm{d}x\mathrm{d}y) = \Pi(\mathrm{d}y)\delta_{\varphi(y)}(\mathrm{d}x) , \quad \nu(\mathrm{d}x\mathrm{d}y) = \Pi(\mathrm{d}x)\delta_{\varphi(x)}(\mathrm{d}y)$$

but this is only possible if μ and ν are equivalent (ie dominated one by the other and conversely). To get that, write

$$\begin{aligned}\int_A \mu(\mathrm{d}x\mathrm{d}y) &= \int \Pi(\mathrm{d}u)\mathbb{1}_A(\underbrace{\varphi(u)}_{v}, \underbrace{u}_{\varphi^{-1}(v)}) \\ &= \int \Pi\circ\varphi^{-1}(\mathrm{d}v)\mathbb{1}_A(v,\varphi^{-1}(v)) = \int \frac{\mathrm{d}\Pi\circ\varphi^{-1}}{\mathrm{d}\Pi}(v)\Pi(\mathrm{d}v)\mathbb{1}_A(v,\varphi^{-1}(v)) = \int_A \frac{\mathrm{d}\Pi\circ\varphi^{-1}}{\mathrm{d}\Pi}(x)\nu(\mathrm{d}x\mathrm{d}y)\end{aligned}$$

only if $\varphi^{-1}(v) = \varphi(v)$ that is φ is an **involution**.

Remark 7.1 *(i) For any involution, you can get a Metropolis Hasting with a theoretical expression of the acceptance probability as*

$$\frac{\mathrm{d}\Pi \circ \varphi^{-1}}{\mathrm{d}\Pi}(x) \wedge 1$$

but the ideal HMC goes one step further since we can show that this is equal to 1. To get this, if we work on $\mathbb{R}^d$ *and if* Π *has density wrt the Lebesgue measure that we still denote* Π*, we get*

$$\frac{\mathrm{d}\Pi \circ \varphi^{-1}}{\mathrm{d}\Pi}(x) = \frac{\Pi(\varphi^{-1}(x))}{\Pi(x)} \left| \frac{\partial \varphi^{-1}(x)}{\partial x} \right|$$

where the second term $\left| \frac{\partial \varphi^{-1}(x)}{\partial x} \right|$ *is the Jacobian of the transformation* φ*. To get* 1 *in the acceptance probability, we can impose that the two terms are equal to* 1*. The first term is one if the involution stays on* the same level set *and the second is one if the involution is* volume invariant*. If for example the involution only keeps the volume then the Radon Nikodym simplifies to*

$$\frac{\Pi(\varphi^{-1}(x))}{\Pi(x)} = \frac{\Pi(\varphi(x))}{\Pi(x)} \quad \textit{since} \quad \varphi \circ \varphi = \mathrm{I}$$

(ii) The second point is that if we only use the involution, then after two steps we land up to the initial state... Not very interesting... Therefore, this deterministic transition is often combined with another move that is not deterministic.

7.2 Hamiltonian dynamics

7.2.1 Level sets

If we now let (p,q) depend on a real parameter t and we impose to stay on a level set of H, we get:

$$\frac{\mathrm{d}H(q_t,p_t)}{\mathrm{d}t} = 0 = \sum_{i=1}^{d} \frac{\partial H(q_t,p_t)}{\partial q_t^i} \frac{\mathrm{d}q_t^i}{\mathrm{d}t} + \frac{\partial H(q_t,p_t)}{\partial p_t^i} \frac{\mathrm{d}p_t^i}{\mathrm{d}t} \ .$$

This gives the idea of using the following dynamics: for all $i \in \{1,\dots,d\}$,

$$\begin{aligned} \frac{\partial H}{\partial q_t^i}(q_t,p_t) &= \frac{\partial U(q_t)}{\partial q_t^i} = -\frac{\mathrm{d}p_t^i}{\mathrm{d}t} \\ \frac{\partial H}{\partial p_t^i}(q_t,p_t) &= \frac{\partial K(p_t)}{\partial p_t^i} = p_t^i = \frac{\mathrm{d}q_t^i}{\mathrm{d}t} \end{aligned}$$

The very last equation leads to the interpretation of p_t^i as a speed since it is the derivative of the position wrt time. Note $\phi_t(q,p) = (q_t,p_t)$ the (deterministic) position and momentum at time t when (q_s,p_s) follows the Hamiltonian dynamics.

7.2.2 Involution

Denote $s(q,p) = (q,-p)$ and set $f_T = s \circ \phi_T$. We now show that f_T is an involution. Indeed, write $f_T(q,p) = (q_T,-p_T)$. To see what we obtain by applying again f_T, set $\tilde{q}_t = q_{T-t}$ and $\tilde{p}_t = -p_{T-t}$ so that $(\tilde{q}_0,\tilde{p}_0) = f_T(q,p)$. Then,

$$\frac{\mathrm{d}\tilde{q}_t^i}{\mathrm{d}t} = \frac{\mathrm{d}q_{T-t}^i}{\mathrm{d}t} = -\left.\frac{\mathrm{d}q_s^i}{\mathrm{d}s}\right|_{s=T-t} = -p_{T-t}^i = \tilde{p}_t^i$$
$$\frac{\mathrm{d}\tilde{p}_t^i}{\mathrm{d}t} = -\frac{\mathrm{d}p_{T-t}^i}{\mathrm{d}t} = \left.\frac{\mathrm{d}p_s^i}{\mathrm{d}s}\right|_{s=T-t} = -\frac{\partial U(q_{T-t})}{\partial q_{T-t}^i} == -\frac{\partial U(\tilde{q}_{T-t})}{\partial \tilde{q}_{T-t}^i}$$

Finally, the process $(\tilde{q}_t, \tilde{p}_t)$ follows the Hamiltonian dynamics so that $f_T(\tilde{q}_0, \tilde{p}_0) = (\tilde{q}_T, -\tilde{p}_T)$ and by definition this quantity is equal to (q_0, p_0). We finally obtain that f_T is an involution.

7.3 The leapfrog integrator

$$\begin{aligned} p_{k+1/2} &= p_k - (h/2)\nabla U(q_k) \\ q_{k+1} &= q_k - h p_{k+1/2} \\ p_{k+1} &= p_{k+1/2} - (h/2)\nabla U(q_{k+1}) \end{aligned}$$

Chapter 8

Exercises

Exercise 8.1. Let $\{Z_t,\, t \in \mathbb{N}^*\}$ be an i.i.d. Bernoulli sequence such that $\mathbb{P}(Z_t = 1) = p = 1 - \mathbb{P}(Z_t = 0)$, and $p \in (0,1)$. Define $\{X_t,\, t \in \mathbb{N}\}$ by $X_0 = 0$ and $X_t = X_{t-1} + Z_t = \sum_{s=1}^{t} Z_s$, for $t \geq 1$. Denote by $\mathscr{F}_t^X = \sigma(X_s,\, 0 \leq s \leq t)$ the σ-algebra generated by the random variables $\{X_0, \ldots, X_t\}$.

(a) Show that

$$\mathbb{P}(X_{t+1} = x_{t+1} \,\big|\, X_t = x_t, \ldots, X_1 = x_1) = \mathbb{P}(X_{t+1} = x_{t+1} \,\big|\, X_t = x_t) = \begin{cases} p & \text{if } x_{t+1} - x_t = 1, \\ 1-p & \text{if } x_{t+1} - x_t = 0, \\ 0 & \text{otherwise.} \end{cases}$$

(b) Consider $A = \{X_1 = 1, X_2 = 1\}$ and $B = \{X_2 = 1, X_3 = 1\}$. Show that $A = \{Z_1 = 1, Z_2 = 0\}$ and $B = \{(Z_1 = 0, Z_2 = 1, Z_3 = 0) \cup (Z_1 = 1, Z_2 = Z_3 = 0)\}$. Show that A and B are not independent.

(c) Show that A and B are conditionally independent given X_1. $\mathbb{P}(A \cap B \,\big|\, X_2 = 1) = (1-p)/2 = \mathbb{P}(A \,\big|\, X_2 = 1)\mathbb{P}(B \,\big|\, X_2 = 1)$.

Exercise 8.2. Let $\{(X_t, \mathscr{F}_t),\, t \in \mathbb{N}\}$ be a Markov chain.

(a) Show that for all $0 \leq r \leq t$ and $f \in \mathbb{F}_b(\mathsf{X})$

$$\mathbb{E}\left[f(X_{t+1}) | X_t \right] = \mathbb{E}\left[f(X_{t+1}) | \, \sigma(X_r, \ldots, X_t) \right] = \mathbb{E}\left[f(X_{t+1}) | \, \mathscr{F}_t \right] ,$$

(b) Consider the property $(\mathscr{P}_n)$: for all $Y = \prod_{k=0}^{n} g_k(X_{t+k})$ where $g_k \in \mathbb{F}_b(\mathsf{X})$, $\mathbb{E}\left[Y | \, \mathscr{F}_t \right] = \mathbb{E}\left[Y | X_t \right]$, $\mathbb{P}-$a.s. Show that $\mathscr{P}_0$ is satisfied.

(c) Assume that $(\mathscr{P}_n)$ is satisfied. Then, for any $g_k \in \mathbb{F}_b(\mathsf{X})$, prove that

$$\mathbb{E}\left[g_0(X_t) \ldots g_n(X_{t+n}) g_{n+1}(X_{t+n+1}) | \, \mathscr{F}_t \right] = \mathbb{E}\left[g_0(X_t) \ldots g_n(X_{t+n}) g_{n+1}(X_{t+n+1}) | X_t \right] ,$$

showing that $(\mathscr{P}_{n+1})$ is true.

(d) Consider the vector space

$$\mathscr{H} = \left\{ Y \in \sigma(X_s, s \geq t), \mathbb{E}\left[Y | \, \mathscr{F}_t \right] = \mathbb{E}\left[Y | X_t \right], \mathbb{P} - \text{a.s.} \right\} .$$

Let $\{Y_n,\, n \in \mathbb{N}\}$ be an increasing sequence of nonnegative random variables in $\mathscr{H}$ such that $Y = \lim_{n\to\infty} Y_n$ is bounded. Show that

$$\mathbb{E}\left[Y | \, \mathscr{F}_t \right] = \lim_{n\to\infty} \mathbb{E}\left[Y_n | \, \mathscr{F}_t \right] = \lim_{n\to\infty} \mathbb{E}\left[Y_n | X_t \right] = \mathbb{E}\left[Y | X_t \right] , \quad \mathbb{P} - \text{a.s.}$$

(e) Show that, for all $t \in \mathbb{N}$ and all bounded $\sigma(X_s, s \geq t)$-measurable random variables Y,

$$\mathbb{E}\left[Y | \, \mathscr{F}_t \right] = \mathbb{E}\left[Y | X_t \right] , \quad \mathbb{P} - \text{a.s.} \tag{8.1}$$

Exercise 8.3. Let $\{(X_t, \mathscr{F}_t),\ t \in \mathbb{N}\}$ be an adapted stochastic process such that for all $t \in \mathbb{N}$ and all bounded $\sigma(X_s, s \geq t)$-measurable random variables Y, $\mathbb{E}[Y|\mathscr{F}_t] = \mathbb{E}[Y|X_t]$, $\mathbb{P}-$a.s.. For $A \in \mathscr{F}_t$ and $B \in \sigma(X_s, s \geq t)$, denote $Z = \mathbb{1}_A$ and $Y = \mathbb{1}_B$. Show that $\mathbb{P}[A \cap B|\mathscr{F}_t] = Z\mathbb{E}[Y|X_t]$ $\mathbb{P}-$a.s.and that $\mathbb{P}[A \cap B|X_t] = \mathbb{P}[A|X_t]\mathbb{P}[B|X_t]$ $\mathbb{P}-$a.s.

Exercise 8.4. Let $\{(X_t, \mathscr{F}_t),\ t \in \mathbb{N}\}$ be an adapted stochastic process such that for all $t \in \mathbb{N}$, $A \in \mathscr{F}_t$ and $B \in \sigma(X_s, s \geq t)$, $\mathbb{P}[A \cap B|X_t] = \mathbb{P}[A|X_t]\mathbb{P}[B|X_t]$, $\mathbb{P}-$a.s.

(a) Show that $\mathbb{E}[YZ|X_t] = \mathbb{E}[Y|X_t]\mathbb{E}[Z|X_t]$ for all bounded $\sigma(X_s, s \geq t)$-measurable random variables Y and $\mathscr{F}_t$-measurable random variables Z.
(b) Show that $\{(X_t, \mathscr{F}_t),\ t \in \mathbb{N}\}$ is a Markov chain.

Exercise 8.5. Combine the following arguments to get the proof of Proposition 2.5.

(a) Assume first that f is a simple nonnegative function, i.e., $f = \sum_{i\in I} \beta_i \mathbb{1}_{B_i}$ for a finite collection of nonnegative numbers β_i and sets $B_i \in \mathscr{Y}$. Then, using property (ii) of Theorem 2.2, for all $x \in \mathsf{X}$, $Nf(x) = \sum_{i\in I} \beta_i N(x, B_i)$, the function Nf is measurable.
(b) Let now $f \in \mathbb{F}_+(\mathsf{X})$ and let $\{f_n,\ n \in \mathbb{N}\}$ be a nondecreasing sequence of measurable nonnegative simple functions such that $\lim_{n\to\infty} f_n(x) = f(x)$. Then, by the monotone convergence theorem, for all $x \in \mathsf{X}$, $Nf(x) = \lim_{n\to\infty} Nf_n(x)$.
(c) Since any nonnegative measurable function is a pointwise limit of a nondecreasing sequence of nonnegative measurable functions, and since a pointwise limit of measurable function is measurable, Nf is measurable.

Exercise 8.6. Combine the following arguments to get the proof of Proposition 2.6.

(a) It is easily seen that $\mu N(A) \geq 0$ for all $A \in \mathscr{Y}$ and $\mu N(\emptyset) = 0$ ($N(x, \emptyset) = 0$ for all $x \in \mathsf{X}$.)
(b) Let $\{A_i,\ i \in \mathbb{N}\} \subset \mathscr{X}$ be a sequence of pairwise disjoints sets. For any integer n,

$$\mu N\left(\bigcup_{i=1}^{n} A_i\right) = \int \mu(\mathrm{d}x) N\left(x, \bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{\infty} \int \mu(\mathrm{d}x) N(x, A_i) = \sum_{i=1}^{n} \mu N(A_i)\ .$$

The monotone convergence theorem therefore implies that

$$\mu N\left(\bigcup_{i=1}^{\infty} A_i\right) = \int \mu(\mathrm{d}x) N\left(x, \bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \int \mu(\mathrm{d}x) N(x, A_i) = \sum_{i=1}^{\infty} \mu N(A_i)\ .$$

which establishes the σ-additivity.

Exercise 8.7. Prove Proposition 2.7.

Exercise 8.8. Prove Proposition 2.8.

Exercise 8.9. Let $(\mathsf{X}, \mathscr{X})$, $(\mathsf{Y}, \mathscr{Y})$, and $(\mathsf{Z}, \mathscr{Z})$ be measurable spaces. Let M be a kernel on $\mathsf{X} \times \mathscr{Y}$ and N be a kernel on $\mathsf{Y} \times \mathscr{Z}$.

(a) Show that, if M and N are both finite (resp. bounded) kernels, then $M \otimes N$ is finite (resp. bounded) kernel.
(b) Show that, if M and N are both Markov kernels, then $M \otimes N$ is a Markov kernel.
(c) Show that, if $(\mathsf{U}, \mathscr{U})$ is a measurable space and P is a kernel on $\mathsf{Z} \times \mathscr{U}$, then $(M \otimes N) \otimes P = M \otimes (N \otimes P)$, the tensor product of kernels is associative.

Exercise 8.10 (Proof of Theorem 2.10: direct implication). Denote by $\mathscr{H}$ the set of measurable functions $f \in \mathbb{F}_b(\mathsf{X}^{t+1}, \mathscr{X}^{\otimes(t+1)})$ such that (2.6) is satisfied.

(a) Show that $\mathscr{H}$ is a vector space.
(b) Let $\{f_n,\ n \in \mathbb{N}\}$ be an nondecreasing sequence of nonnegative functions in $\mathscr{H}$ such that $\lim_{n\to\infty} f_n := f$ is bounded. Show that f belongs to $\mathscr{H}$.

(c) For $t \geq 1$, assume that (2.6) holds for $t-1$ and f of the form

$$f_0 \otimes \cdots \otimes f_t(x_0, \ldots, x_t) = f_0(x_0) \cdots f_t(x_t) , \quad f_s \in \mathbb{F}_b(\mathsf{X}) . \tag{8.2}$$

Show that

$$\mathbb{E}\left[\prod_{s=0}^{t} f_s(X_s)\right] = \mathbb{E}\left[\prod_{s=0}^{t-1} f_s(X_s)\mathbb{E}\left[f_t(X_t)|\mathscr{F}_{t-1}\right]\right] = \mathbb{E}\left[\prod_{s=0}^{t-1} f_s(X_s) P f_t(X_{t-1})\right]$$
$$= \nu \otimes P^{\otimes(t-1)}(f_0 \otimes \cdots \otimes f_{t-1} P f_t) = \nu \otimes P^{\otimes t}(f_0 \otimes \cdots \otimes f_t) .$$

(d) Show that $\mathscr{H}$ contains the functions of the form (8.2).
(e) Conclude.

Exercise 8.11 (Proof of Theorem 2.10: converse). For $t = 0$, (2.6) implies that ν is the law of X_0. We have to prove that, for all $t \geq 1$, $f \in \mathbb{F}_+(\mathsf{X})$ and $\mathscr{F}_{t-1}^X$-measurable random variable Y:

$$\mathbb{E}[f(X_t)Y] = \mathbb{E}[Pf(X_{t-1})Y] . \tag{8.3}$$

Denote by $\mathscr{H}$ the set of $\mathscr{F}_{t-1}^X$-measurable random variables Y satisfying (8.3).

(a) Show that $\mathscr{H}$ is a vector space.
(b) Show that if $\{Y_t,\ t \in \mathbb{N}\}$ is an increasing sequence of nonnegative random variables such that $Y = \lim_{n\to\infty} Y_n$ is bounded, then $Y \in \mathscr{H}$.
(c) Show that (8.3) holds for $Y = f_0(X_0) f_1(X_1) \ldots f_{t-1}(X_{t-1})$ where $f_s \in \mathbb{F}_b(\mathsf{X})$ and conclude.

Exercise 8.12 (Functional autoregressive process). Consider the following recursion

$$X_t = g(X_{t-1}) + \sigma(X_{t-1}) Z_t ,$$

where $\{Z_t,\ t \in \mathbb{N}\}$ is an i.i.d. sequence, $g : \mathbb{R} \to \mathbb{R}$ and $\sigma : \mathbb{R} \to \mathbb{R}_+^*$ are measurable functions.

(a) Determine the Markov kernel of this chain.
(b) Assume that the distribution of Z_0 has a density q with respect to Lebesgue's measure on $\mathbb{R}$; show that the Markov kernel P has a density. Give the expression of this density.

Exercise 8.13 (Setwise convergence / existence of a stationary distribution). Assume that for some initial probability ξ, the sequence of probabilities $\{\xi P^t, t \in \mathbb{N}\}$ converges setwise, i.e., for any $A \in \mathscr{X}$, the sequence $\{\xi P^t(A), t \in \mathbb{N}\}$ has a limit, denoted $\gamma_\xi(A)$, i.e., $\lim_{t\to\infty} \xi P^t(A) = \gamma_\xi(A)$.

(a) Show that, $\gamma_\xi(A) = \gamma_\xi P(A)$.
(b) Assume that there exists a unique invariant probability measure. Show that the limit γ_ξ is independent of the initial distribution ξ.
(c) Conversely, assume that for any initial distribution ξ on $(\mathsf{X}, \mathscr{X})$, the sequence $\{\xi P^t, t \in \mathbb{N}\}$ has the same setwise limit. Show that there exists a unique stationary distribution.

Exercise 8.14 (Weak convergence / existence of a stationary distribution). Let (X, d) be a Polish space. We assume that for some initial measure ξ, the sequence of probability measures $\{\xi P^t, t \in \mathbb{N}\}$ converges weakly to γ_ξ. In addition, for any $f \in \mathrm{C}_b(\mathsf{X})$ (the set of bounded continuous functions), $Pf \in \mathrm{C}_b(\mathsf{X})$.

(a) Show that $\gamma_\xi = \gamma_\xi P$.
(b) Show that if the kernel P has a unique invariant distribution, then the weak limit of the sequence $\{\xi P^t, t \in \mathbb{N}\}$ does not depend on the initial distribution ξ.
(c) Conversely, if for any initial distribution ξ the sequence of probability $\{\xi P^t, t \in \mathbb{N}\}$ converges weakly to the same limiting distribution, then the invariant distribution is unique.

Exercise 8.15 (Existence and uniqueness of the stationary measure of an AR(1)). Consider a first order autoregressive process $X_t = \phi X_{t-1} + Z_t$, $t \geq 1$, where $\{Z_t,\ t \in \mathbb{N}\}$ is a zero mean i.i.d. sequence and X_0 is independent of $\{Z_t,\ t \in \mathbb{N}\}$ and is distributed according to some probability ξ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. We assume that $|\phi| < 1$.

(a) Show that $X_t = \phi^t X_0 + \sum_{i=0}^{t-1} \phi^i Z_{t-i}$
(b) Show that, for each integer t, the random variable X_t has the same distribution as $\phi^t X_0 + \sum_{i=0}^{t-1} \phi^i Z_{i+1}$.
(c) Define $\mathscr{F}_t^Z = \sigma(\{Z_s, 0 \le s \le t\})$ and

$$Y_t = \sum_{i=0}^{t-1} \phi^i Z_{i+1} = Y_{t-1} + \phi^{t-1} Z_t \ .$$

Show that $\{(Y_t, \mathscr{F}_t^Z),\ t \in \mathbb{N}\}$ is an L^1-bounded martingale.
(d) Show that the sequence $\{Y_t,\ t \in \mathbb{N}\}$ converges $\mathbb{P}-$a.s. to the integrable random variable Y_∞ . We denote by π the law of Y_∞.
(e) Show that for any function f and any initial distribution ξ, $\mathbb{E}[\xi]\,[f(X_t)] \to \pi(f)$.
(f) Using (8.14), show that π is the unique stationary distribution.

Exercise 8.16. The DAR(1) (see **??**) is given by $X_t = V_t X_{t-1} + (1-V_t)Z_t$ where $\{V\}$ is an i.i.d. sequence of binary random variables with $\mathbb{P}(V_t = 1) = \alpha$, $\{Z\}$ is an i.i.d. sequence of random variables with distribution given by π on $(\mathsf{X}, \mathscr{X})$ and $\{V\}$ and $\{Z\}$ are independent. The initial random variable X_0 is assumed to be independent of $\{V_t,\ t \in \mathbb{N}\}$ and $\{Z_t,\ t \in \mathbb{N}\}$ and is distributed according to ξ.

(a) Show that, for all $t \ge 1$ and any bounded measurable function,

$$\mathbb{E}[\xi]\,[f(X_t)] = \alpha \mathbb{E}[\xi]\,[f(X_{t-1})] + (1-\alpha)\pi(f) \ .$$

(b) Show that $\mathbb{E}[\xi]\,[f(X_{t-1})] = \pi(f)$, then $\mathbb{E}[\xi]\,[f(X_t)] = \pi(f)$ and that π is the unique stationary distribution.
(c) Assume that $\mathsf{X} = \mathbb{N}$ and that $\sum_{k=0}^{\infty} k^2 \pi(k) < \infty$. For any positive integer h, show that $\mathbb{C}\mathrm{ov}[\pi] X_h X_0 = \alpha \mathbb{C}\mathrm{ov}[\pi] X_{h-1} X_0$ and $\mathbb{C}\mathrm{ov}[\pi] X_h X_0 = \alpha^h \mathbb{V}\mathrm{ar}[\pi] X_0$.
(d) Show that the DAR(1) process has exactly the same autocovariance structure as an AR(1) process.
(e) Explain why the DAR(1) process cannot exhibit negative dependence.

Exercise 8.17. Consider the self-excited threshold autoregressive model defined by

$$X_t = \phi_0^{(j)} + \sum_{i=1}^{p_j} \phi_i^{(j)} X_{t-i} + Z_t^{(j)} \quad \text{if} \quad X_{t-d} \in \left(r_{j-1}, r_j\right] \tag{8.4}$$

where $\{Z_t^{(j)},\ t \in \mathbb{N}\} \sim_{\text{iid}} \mathrm{N}(0, \sigma_j^2)$, for $j = 1, \ldots, k$, the positive integer d is a specified delay and $r_0 = -\infty < r_1 < \cdots < r_{p-1} < r_p = +\infty$ is a partition of $\mathbb{R}$.

(a) Show that $\{X_t,\ t \in \mathbb{N}\}$ is a m-th order Markov chain.
(b) Determine the Markov kernel of this chain.

Exercise 8.18. Let $\{Z_t,\ t \in \mathbb{N}\}$ be Gaussian white noise with variance σ^2 and let $|\phi| < 1$ be a constant. Consider the process $X_0 = Z_0$, and $X_t = \phi X_{t-1} + Z_t$, for $t = 1, 2, \ldots$.

(a) Find the mean and the variance of X_t. Is $\{X_t,\ t \in \mathbb{N}\}$ stationary?
(b) Show that for all $h \in \{0, \ldots, t\}$,

$$\mathrm{Cor}(X_t, X_{t-h}) = \phi^h \left[\frac{\mathrm{Var}(X_{t-h})}{\mathrm{Var}(X_t)} \right]^{1/2} \ .$$

(c) Show that $\lim_{t\to\infty} \mathrm{Var}(X_t) = \sigma^2/(1-\phi^2)$ and that $\lim_{t\to\infty} \mathrm{Cor}(X_t, X_{t-h}) = \phi^h$, $h \ge 0$.
(d) Comment on how you could use these results to simulate n observations of a stationary Gaussian AR(1) model from simulated i.i.d. $\mathrm{N}(0,1)$ values.
(e) Now suppose $X_0 = Z_0/\sqrt{1-\phi^2}$. Is $\{X_t,\ t \in \mathbb{N}\}$ stationary?

Exercise 8.19 (Simple Markovian bilinear model). Consider the simple Markovian bilinear model

$$X_t = aX_{t-1} + bZ_t X_{t-1} + Z_t = (a + bZ_t)X_{t-1} + Z_t \ . \tag{8.5}$$

where $\{Z_t,\, t \in \mathbb{N}\}$ is an i.i.d. sequence of standard Gaussian variables and is independent of $X_0 \sim \xi$. We assume that $a^2 + b^2 < 1$.

(a) Show that, for all $t \in \mathbb{N}$,

$$X_t = \prod_{i=1}^{t}(a+bZ_i)X_0 + Z_t + \sum_{j=1}^{t-1} Z_{t-j} \prod_{i=t-j+1}^{t} (a+bZ_i)\,.$$

(b) Show that $\lim_{t\to\infty} \left(\prod_{i=1}^{t} |a+bZ_i|\right)^{1/t}$ exists $\mathbb{P}-$a.s. and that this limit is strictly less than 1.
(c) Set $Y_t = Z_0 + \sum_{j=1}^{t-1} Z_j \prod_{i=0}^{j-1}(a+bZ_i)$. Show that the random variables Y_t and $Z_t + \sum_{j=1}^{t-1} Z_{t-j} \prod_{i=t-j+1}^{t}(a+bZ_i)$ have the same distributions.
(d) Show that $\{(Y_t, \mathscr{F}_t^Z),\, t \in \mathbb{N}\}$ is an L^2-bounded martingale converging $\mathbb{P}-$a.s. and in $\mathsf{L}^2(\mathbb{P})$ to $Y_\infty = Z_0 + \sum_{j=1}^{\infty} Z_j \prod_{i=0}^{j-1}(a+bZ_i)$.
(e) Show that for any $f \in \mathrm{C}_b(\mathbb{R})$, $\lim_{t\to\infty} \mathbb{E}[\xi][][f(X_t)] = \pi(f)$, where $\pi_{a,b}$ is the distribution of Y_∞ (the subscripts (a,b) stress the dependence of the stationary distributions in a and b).
(f) Using Exercise 8.14, show that $\pi_{a,b}$ is the unique stationary distribution.
(g) Show that $\int x\pi_{a,b}(\mathrm{d}x) = 0$ and $\int x^2\pi_{a,b}(\mathrm{d}x) = 1/(1-a^2-b^2)$.

Exercise 8.20 (Simple Markovian bilinear model cont.). Consider again the simple bilinear model

$$X_t^{(a,b)} = aX_{t-1}^{(a,b)} + bZ_tX_{t-1}^{(a,b)} + Z_t = (a+bZ_t)X_{t-1}^{(a,b)} + Z_t\,, \tag{8.6}$$

where $\{Z_t,\, t \in \mathbb{N}\}$ is an i.i.d. sequence of standard Gaussian variables and is independent of $X_0 \sim \pi_{a,b}$ where $\pi_{a,b}$ is the stationary distribution. The superscript (a,b) is used to stress the dependence of the process in these coefficients.

(a) Show that

$$\mathbb{E}|X_t^{(a,b)} - X_t^{(a,0)}|^2 = \frac{b^2}{(1-a^2-b^2)(1-a^2)} \to 0\,, \quad \text{as } b \to 0\,.$$

(b) Assume that $b > 0$. Show that the cumulative distribution function $x \mapsto F_{a,b}(x)$ of $\pi_{a,b}$ is the solution of an integral equation.
(c) Show that, if $x \neq a/b$, the cumulative dist is differentiable at x.
(d) Show that $F_{a,b}$ does not have a discrete component at $x = -a/b$.
(e) Conclude that the stationary distribution $\pi_{a,b}$ has a density satisfying the following integral equation

$$f_{a,b}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{f_{a,b}(s)}{|1+bs|} \exp\left[-\frac{1}{2}\left(\frac{x-as}{1+bs}\right)^2\right] \mathrm{Leb}(\mathrm{d}s)\,.$$

(f) Show that $f_{a,b}(-1/b) > 0$ and that $\lim_{x\to -a/b} f_b(x) = \infty$.

Exercise 8.21. Consider a GARCH(p,q) model

$$X_t = \sigma_t\varepsilon_t \tag{8.7}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2 + \beta_1\sigma_{t-1}^2 + \cdots + \beta_q\sigma_{t-q}^2\,, \tag{8.8}$$

where $\{\varepsilon_t,\, t \in \mathbb{N}\}$ is an i.i.d. sequence with zero-mean and unit-variance, and the coefficients $\alpha_j,\ j \in \{0,\ldots,p\}$ and $\beta_j,\ j \in \{1,\ldots,q\}$ are nonnegative. State conditions upon which the GARCH(p,q) model admits a unique invariant stationary distribution.

Exercise 8.22 (Basic Gibbs). Suppose we wish to obtain samples from a bivariate normal distribution,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix},\, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right],$$

where $|\theta| < 1$.

(a) Determine the univariate conditionals $\mathrm{p}_1^\theta(X \mid Y)$ and $\mathrm{p}_2^\theta(Y \mid X)$; see (**??**)–(**??**).
(b) Consider the Markov chain: Pick $X^{(0)} = x_0$, and then iterate the process $X^{(0)} \mapsto Y^{(0)} \mapsto X^{(1)} \mapsto Y^{(1)} \mapsto \cdots \mapsto X^{(t)} \mapsto Y^{(t)} \mapsto \cdots$, where $Y^{(t)}$ is a sample from $\mathrm{p}_2^\theta(\cdot \mid X^{(t)})$, and $X^{(t)}$ is a sample from $\mathrm{p}_1^\theta(\cdot \mid Y^{(t-1)})$. Write the joint distribution of $(X^{(t)}, Y^{(t)})$ in terms of the starting value x_0 and the correlation θ.
(c) What is the asymptotic distribution of $(X^{(t)}, Y^{(t)})$ as $t \to \infty$?
(d) How can you use the results of part (c) to obtain a pseudo random sample of n bivariate normals using only pseudo samples from a univariate normal distribution? How does the value of θ affect the procedure?

Exercise 8.23 (Accept-reject algorithm). We wish to sample the target density π (with respect to a dominating measure λ), which is known up to a multiplicative constant. Let q be a proposal density (assumed to be easier to sample.) We assume that there exists a constant $M < \infty$ such that $\pi(x)/q(x) \leq M$ for all $x \in \mathsf{X}$. The Accept-Reject goes as follows: first, we generate $Y \sim g$ and, independently, we generate $U \sim \mathrm{Unif}([0,1])$. If $Mq(Y)U \leq \pi(Y)$, then we set $X = Y$. If the inequality is not satisfied, we then discard Y and U and start again.

(a) Show that X is distributed according to π.
(b) What is the distribution of the number of trials required per sample?
(c) What is the average number of samples needed for one simulation?
(d) Propose a method to estimate the normalizing constant of π.
(e) Compare with the Metropolis-Hastings algorithm with Independent proposal.

Exercise 8.24 (The independent sampler MCMC). We wish to sample a target distribution $\Gamma(\alpha, \beta)$ using the independent MCMC algorithm. We will use as the proposal distribution $\gamma(\lfloor \alpha \rfloor, b)$.

(a) Derive the corresponding Accept-Reject method. Explain why, when $\beta = 1$, the choice of b minimizing the average number of simulations per sample is $b = \lfloor \alpha \rfloor / \alpha$.
(b) Generate 10000 $\Gamma(5, 5/5.25)$ random variables to derive a $\Gamma(5.25, 1)$ sample.
(c) Use the same sample in the corresponding Metropolis-Hastings algorithm to generate 10000 $\Gamma(5.25, 1)$ random variables.
(d) Compare the algorithms using (i) their acceptance rates and (ii) the estimates of the mean and variance of the $\Gamma(5.25, 1)$ along with their errors.

Exercise 8.25 (Functional autoregressive). Let P be a Markov kernel on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Show that $(x,u) \mapsto G(x,u) = \inf\{x' \in \mathbb{R} : F(x,x') \geq u\}$ is a Borel function where $F : (x,x') \mapsto P(x, (\infty, x'])$ is a Borel function.

Exercise 8.26 (Reversibility of the Gibbs sampler). We use the notations introduced in Section 4.1.5.

(a) Show that for any functions $f, g \in \mathbb{F}_+(\mathsf{X})$,

$$\iint f(x) g(x') \pi(x) \, \lambda(\mathrm{d}x) \, K_k(x, \mathrm{d}x') = \int \left\{ f(x) \pi(x) \, \lambda_k(\mathrm{d}x^{[k]}) \int g(x'^{[k]}, x^{[-k]}) \pi_k(x'^{[k]} \mid x^{[-k]}) \, \lambda_k(\mathrm{d}x'^{[k]}) \right\} \lambda_{-k}(\mathrm{d}x^{[-k]}) \,,$$

where $(x'^{[k]}, x^{[-k]})$ refers to the element u of X such that $u_k = x'^{[k]}$ and $(u_{\ell \neq k}) = x^{[-k]}$
(b) Show that

$$\pi(x^{[k]}, x^{[-k]}) \pi_k(x'^{[k]} \mid x^{[-k]}) = \pi_k(x^{[k]} \mid x^{[-k]}) \pi(x'^{[k]}, x^{[-k]}) \,,$$

(c) Conclude.

Exercise 8.27 (A simple time series of counts). Consider the following iterated random function:

$$X_{k+1} = d + aX_k + b\lfloor -X_k \ln(U_{k+1}) \rfloor \,, \quad k \geq 0 \tag{8.9}$$

where $\{U_k,\ k \in \mathbb{N}^\star\}$ is a sequence of i.i.d. random variables such that $U_k \sim \mathrm{Unif}[0,1]$. We assume that X_0 is independent of $\{U_k,\ k \in \mathbb{N}^\star\}$ and that $d, a, b > 0$ and $a + b < 1$. By simplification, we assume $X_0 \geq d$, so that $\{X_k,\ k \in \mathbb{N}\}$ is a Markov chain taking values on $\mathsf{X} = [d, \infty)$.

(a) Show that for some function f, the recursion on the $\{X_k,\, k \in \mathbb{N}\}$ may be written as: for all $k \geq 0$,

$$\begin{aligned} Y_{k+1} | \mathscr{F}_k &\sim \text{Geom}(f(X_k)) \,, \\ X_{k+1} &= d + aX_k + b(Y_{k+1} - 1) \,. \end{aligned} \tag{8.10}$$

where $\mathscr{F}_k = \sigma(X_0, \ldots, X_k, Y_1, \ldots, Y_k)$.

(b) Denote $X_1(x) = d + ax + b\lfloor -x \ln(U) \rfloor$ where $U \sim \text{Unif}[0,1]$. Compute $\mathbb{E}[X_1(x)]$ explicitly.

(c) Deduce that $\varphi(x) : x \mapsto \mathbb{E}[X_1(x)]$ can be differentiated and show that $|\varphi'(x)| < 1$ for all $x \in \mathsf{X}$.

(d) By noting that $X_1(x) \geq X_1(x')$ for all $x \geq x'$, deduce that there exists $\rho < 1$ such that $\mathbb{E}[|X_1(x) - X_1(x')|] \leq \rho |x - x'|$ for all $x, x' \in \mathsf{X}$. Conclude.

Appendix A

Technical results

A.1 Limit theorems for triangular arrays

In this section, we derive limit theorems for triangular arrays of dependent random variables. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, let X be a random variable and let $\mathscr{G}$ be a a sub-σ field of $\mathscr{F}$. Let $\{M_N\}_{N\geq 0}$ be a sequence of positive integers, $\{U_{N,i}\}_{i=1}^{\infty}$ be a triangular array of random variables, and $\{\mathscr{F}_{N,i}\}_{0\leq i\leq\infty}$ be a triangular array of sub-sigma-fields of $\mathscr{F}$. Throughout this section, it is assumed that $\mathscr{F}_{N,i-1} \subseteq \mathscr{F}_{N,i}$ and for each N and $i = 1,\ldots,M_N$, $U_{N,i}$ is $\mathscr{F}_{N,i}$-measurable. We preface the proof with some technical lemmas.

Lemma A.1 *Let $\{(Z_n, \mathscr{F}_n),\ n \in \mathbb{N}\}$ be an adapted sequence of nonnegative random variables. Then, for all $\varepsilon > 0$,*

$$\xi_n = \mathbb{E}\left[\sum_{i=1}^{n} Z_i \mathbb{1}\Big(\sum_{j=1}^{i} \mathbb{E}\left[Z_j \middle| \mathscr{F}_{j-1}\right]\Big) \leq \varepsilon\right] \leq \varepsilon\,.$$

PROOF. We set $\mu_i = \sum_{j=1}^{i} \mathbb{E}\left[Z_j \middle| \mathscr{F}_{j-1}\right]$. Then,

$$\xi_n = \sum_{i=1}^{n} \mathbb{E}\left[Z_i \mathbb{1}(\{\mu_i \leq \varepsilon\})\right] = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[Z_i \middle| \mathscr{F}_{i-1}\right] \mathbb{1}(\{\mu_i \leq \varepsilon\})\right]\,.$$

Set $\tau := \max\{1 \leq i \leq n,\ \mu_i \leq \varepsilon\}$ and $\tau = 0$ on $\{\min 1 \leq i \leq n,\ \mu_i > \varepsilon\}$. On $\{\tau = 0\}$ we get $\sum_{i=1}^{n} \mathbb{E}\left[Z_i \middle| \mathscr{F}_{i-1}\right] \mathbb{1}(\{\mu_i \leq \varepsilon\}) = 0$ and, on$\{\tau > 0\}$, since $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$,

$$\sum_{i=1}^{n} \mathbb{E}\left[Z_i \middle| \mathscr{F}_{i-1}\right] \mathbb{1}(\{\mu_i \leq \varepsilon\}) = \sum_{i=1}^{\tau} \mathbb{E}\left[Z_i \middle| \mathscr{F}_{i-1}\right] \leq \varepsilon\,.$$

■

The following Lemma has been established in **?**, Lemma 3.5.

Lemma A.2 *Let $\{(Z_n, \mathscr{F}_n),\ n \in \mathbb{N}\}$ be an adapted sequence of nonnegative random variables. Then, for all $\varepsilon > 0$, and $\alpha > 0$,*

$$\mathbb{P}(\max_{1\leq i\leq n} Z_i > \varepsilon) \leq \alpha + \mathbb{P}\left(\sum_{i=1}^{n} \mathbb{P}\left[Z_i > \varepsilon \middle| \mathscr{F}_{i-1}\right] > \alpha\right)\,.$$

PROOF. We set $\nu_i := \sum_{j=1}^{i} \mathbb{P}\left[Z_j > \varepsilon \middle| \mathscr{F}_{j-1}\right]$. Then,

$$\mathbb{P}(\max_{1\leq i\leq n} Z_i > \varepsilon) \leq \mathbb{P}(\max_{1\leq i\leq n} Z_i > \varepsilon,\ \nu_n \leq \alpha)\ + \mathbb{P}(\nu_n > \alpha)\,,$$

and

$$\begin{aligned}\mathbb{P}(\max_{1\leq i\leq n} Z_i > \varepsilon,\ \nu_n \leq \alpha) &\leq \sum_{i=1}^{n} \mathbb{P}(Z_i > \varepsilon,\ \nu_n \leq \alpha) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(Z_i > \varepsilon,\ \nu_i \leq \alpha) = \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{Z_i>\varepsilon\}} \mathbb{1}_{\{\nu_i\leq\alpha\}}\right] \leq \alpha\,,\end{aligned}$$

using Lemma A.1. ■

Lemma A.3 *Let $\mathscr{G}$ be a σ-field and X a random variable such that $\mathbb{E}\left[X^2 \middle| \mathscr{G}\right] < \infty$. Then, for any $\varepsilon > 0$,*

$$4\mathbb{E}\left[|X|^2 \mathbb{1}\{|X| \geq \varepsilon\} \middle| \mathscr{G}\right] \geq \mathbb{E}\left[|X - \mathbb{E}\left[X \middle| \mathscr{G}\right]|^2 \mathbb{1}\{|X - \mathbb{E}\left[X \middle| \mathscr{G}\right]| \geq 2\varepsilon\} \middle| \mathscr{G}\right]\,.$$

PROOF. Let $Y = X - \mathbb{E}\left[X \middle| \mathscr{G}\right]$. We have $\mathbb{E}\left[Y \middle| \mathscr{G}\right] = 0$. It is equivalent to show that for any $\mathscr{G}$-measurable random variable Z,

$$\mathbb{E}\left[Y^2 \mathbb{1}\{|Y| \geq 2\varepsilon\} \middle| \mathscr{G}\right] \leq 4\mathbb{E}\left[|Y + Z|^2 \mathbb{1}\{|Y + Z| \geq \varepsilon\} \middle| \mathscr{G}\right]\,.$$

On the set $\{|Z| < \varepsilon\}$,

$$\begin{aligned}\mathbb{E}\left[Y^2\mathbb{1}\{|Y|\geq 2\varepsilon\}\,\middle|\,\mathscr{G}\right] &\leq 2\mathbb{E}\left[((Y+Z)^2+Z^2)\mathbb{1}\{|Y+Z|\geq\varepsilon\}\,\middle|\,\mathscr{G}\right]\\ &\leq 2(1+Z^2/\varepsilon^2)\mathbb{E}\left[(Y+Z)^2\mathbb{1}\{|Y+Z|\geq\varepsilon\}\,\middle|\,\mathscr{G}\right]\\ &\leq 4\mathbb{E}\left[(Y+Z)^2\mathbb{1}\{|Y+Z|\geq\varepsilon\}\,\middle|\,\mathscr{G}\right]\,.\end{aligned}$$

Moreover, on the set $\{|Z|\geq\varepsilon\}$, using that $\mathbb{E}[ZY\,|\,\mathscr{G}]=Z\mathbb{E}[Y\,|\,\mathscr{G}]=0$.

$$\begin{aligned}\mathbb{E}\left[Y^2\mathbb{1}\{|Y|\geq 2\varepsilon\}\,\middle|\,\mathscr{G}\right] &\leq \mathbb{E}\left[Y^2+Z^2-\varepsilon^2\,\middle|\,\mathscr{G}\right]\\ &\leq \mathbb{E}\left[(Y+Z)^2-\varepsilon^2\,\middle|\,\mathscr{G}\right]\leq\mathbb{E}\left[(Y+Z)^2\mathbb{1}\{|Y+Z|\geq\varepsilon\}\,\middle|\,\mathscr{G}\right]\,.\end{aligned}$$

The proof is completed. ■

Theorem A.4. *Assume that* $\mathbb{E}\left[|U_{N,j}|\,\middle|\,\mathscr{F}_{N,j-1}\right]<\infty\ \mathbb{P}-$a.s. *for any N and any* $j=1,\ldots,M_N$, *and*

$$\sup_N\mathbb{P}\left(\sum_{j=1}^{M_N}\mathbb{E}\left[|U_{N,j}|\,\middle|\,\mathscr{F}_{N,j-1}\right]\geq\lambda\right)\to 0 \qquad \text{as } \lambda\to\infty \tag{A.1}$$

$$\sum_{j=1}^{M_N}\mathbb{E}\left[|U_{N,j}|\mathbb{1}\{|U_{N,j}|\geq\varepsilon\}\,\middle|\,\mathscr{F}_{N,j-1}\right]\overset{\mathbb{P}\text{-prob}}{\longrightarrow}0 \qquad \text{for any } \varepsilon>0\,. \tag{A.2}$$

Then, $\max_{1\leq i\leq M_N}\left|\sum_{j=1}^{i}U_{N,j}-\sum_{j=1}^{i}\mathbb{E}\left[U_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right|\overset{\mathbb{P}\text{-prob}}{\longrightarrow}0\,.$

PROOF. Assume first that for each N and each $i=1,\ldots,M_N$, $U_{N,i}\geq 0$, $\mathbb{P}-$a.s. By Lemma A.2, we have that for any constants ε and $\eta>0$,

$$\mathbb{P}\left[\max_{1\leq i\leq M_N}U_{N,i}\geq\varepsilon\right]\leq\eta+\mathbb{P}\left[\sum_{i=1}^{M_N}\mathbb{P}(\,U_{N,i}\geq\varepsilon\,|\,\mathscr{F}_{N,i-1})\geq\eta\right]\,.$$

From the conditional version of the Chebyshev identity,

$$\mathbb{P}\left[\max_{1\leq i\leq M_N}U_{N,i}\geq\varepsilon\right]\leq\eta+\mathbb{P}\left[\sum_{i=1}^{M_N}\mathbb{E}\left[U_{N,i}\mathbb{1}\{U_{N,i}\geq\varepsilon\}\,|\,\mathscr{F}_{N,i-1}\right]\geq\eta\varepsilon\right]\,. \tag{A.3}$$

Let ε and $\lambda>0$ and define

$$\bar{U}_{N,i}:=U_{N,i}\mathbb{1}\{U_{N,i}<\varepsilon\}\,\mathbb{1}\left\{\sum_{j=1}^{i}\mathbb{E}\left[U_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]<\lambda\right\}\,.$$

For any $\delta>0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq i\leq M_N}\left|\sum_{j=1}^{i}U_{N,j}-\sum_{j=1}^{i}\mathbb{E}\left[U_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right|\geq 2\delta\right)\\ &\quad\leq\mathbb{P}\left(\max_{1\leq i\leq M_N}\left|\sum_{j=1}^{i}\bar{U}_{N,j}-\sum_{j=1}^{i}\mathbb{E}\left[\bar{U}_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right|\geq\delta\right)\\ &\quad+\mathbb{P}\left(\max_{1\leq i\leq M_N}\left|\sum_{j=1}^{i}U_{N,j}-\bar{U}_{N,j}-\sum_{j=1}^{i}\mathbb{E}\left[U_{N,j}-\bar{U}_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right|\geq\delta\right)\,.\end{aligned}$$

The second term in the right-hand side is bounded by

$$\begin{aligned}&\mathbb{P}\left(\max_{1\leq i\leq M_N}U_{N,i}\geq\varepsilon\right)+\mathbb{P}\left(\sum_{j=1}^{M_N}\mathbb{E}\left[U_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\geq\lambda\right)+\\ &\qquad\qquad\mathbb{P}\left(\sum_{j=1}^{M_N}\mathbb{E}\left[U_{N,j}\mathbb{1}\{U_{N,j}\geq\varepsilon\}\,\middle|\,\mathscr{F}_{N,j-1}\right]\geq\delta\right)\,.\end{aligned}$$

Eqs. (A.2) and (A.3) imply that the first and last terms in the last expression converge to zero for any $\varepsilon>0$ and (A.1) implies that the second term may be arbitrarily small by choosing for λ sufficiently large. Now, by the Doob maximal inequality,

$$\mathbb{P}\left(\max_{1\le i\le M_N}\left|\sum_{j=1}^{i}\bar{U}_{N,j}-\mathbb{E}\left[\bar{U}_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right|\ge\delta\right)$$
$$\le\delta^{-2}\mathbb{E}\left[\sum_{j=1}^{M_N}\mathbb{E}\left[\left(\bar{U}_{N,j}-\mathbb{E}\left[\bar{U}_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right)^2\middle|\,\mathscr{F}_{N,0}\right]\right].$$

This last term does not exceed

$$\delta^{-2}\mathbb{E}\left[\sum_{i=1}^{M_N}\mathbb{E}\left[\bar{U}_{N,j}^2\middle|\,\mathscr{F}_{N,0}\right]\right]\le\delta^{-2}\varepsilon\mathbb{E}\left[\sum_{j=1}^{M_N}\mathbb{E}\left[\bar{U}_{N,j}\middle|\,\mathscr{F}_{N,0}\right]\right]$$
$$\le\delta^{-2}\varepsilon\mathbb{E}\left[\sum_{j=1}^{M_N}\mathbb{E}\left[\bar{U}_{N,j}\,\middle|\,\mathscr{F}_{N,j-1}\right]\right]\le\delta^{-2}\varepsilon\lambda\,.$$

Since ε is arbitrary, the proof follows for $U_{N,j}\ge 0$, $\mathbb{P}-$a.s., for each N and $j=1,\ldots,M_N$. The proof extends to an arbitrary triangular array $\{U_{N,j}\}_{i=1}^{M_N}$ by applying the preceding result to $\{U_{N,j}^+\}_{1\le j\le M_N}$ and $\{U_{N,j}^-\}_{1\le j\le M_N}$. ■

Lemma A.5 *Assume that for all N,* $\sum_{i=1}^{M_N}\mathbb{E}\left[U_{N,i}^2\,\middle|\,\mathscr{F}_{N,i-1}\right]=1$, $\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right]=0$ *for* $i=1,\ldots,M_N$, *and for all* $\varepsilon>0$,

$$\sum_{i=1}^{M_N}\mathbb{E}\left[U_{N,i}^2\mathbb{1}\{|U_{N,i}|\ge\varepsilon\}\,\middle|\,\mathscr{F}_{N,0}\right]\overset{\mathbb{P}\text{-}prob}{\longrightarrow}0\,. \tag{A.4}$$

Then, for any real u, $\mathbb{E}\left[\exp\left(\mathrm{i}u\sum_{j=1}^{M_N}U_{N,j}\right)\,\middle|\,\mathscr{F}_{N,0}\right]-\exp\left(-u^2/2\right)\overset{\mathbb{P}\text{-}prob}{\longrightarrow}0.$

PROOF. Denote $\sigma_{N,i}^2:=\mathbb{E}\left[U_{N,i}^2\,\middle|\,\mathscr{F}_{N,i-1}\right]$. Write the following decomposition (with the convention $\sum_{j=a}^{b}=0$ if $a>b$):

$$\mathrm{e}^{\mathrm{i}u\sum_{j=1}^{M_N}U_{N,j}}-\mathrm{e}^{-\frac{u^2}{2}\sum_{j=1}^{M_N}\sigma_{N,j}^2}=\sum_{l=1}^{M_N}\mathrm{e}^{\mathrm{i}u\sum_{j=1}^{l-1}U_{N,j}}\left(\mathrm{e}^{\mathrm{i}uU_{N,l}}-\mathrm{e}^{-\frac{u^2}{2}\sigma_{N,l}^2}\right)\mathrm{e}^{-\frac{u^2}{2}\sum_{j=l+1}^{M_N}\sigma_{N,j}^2}\,.$$

Since $\sum_{j=1}^{l-1}U_{N,j}$ and $\sum_{j=l+1}^{M_N}\sigma_{N,j}^2=1-\sum_{j=1}^{l}\sigma_{N,j}^2$ are $\mathscr{F}_{N,l-1}$-measurable,

$$\left|\mathbb{E}\left[\exp\left(\mathrm{i}u\sum_{j=1}^{M_N}U_{N,j}\right)-\exp\left(-(u^2/2)\sum_{j=1}^{M_N}\sigma_{N,j}^2\right)\,\middle|\,\mathscr{F}_{N,0}\right]\right|$$
$$\le\sum_{l=1}^{M_N}\mathbb{E}\left[\left|\mathbb{E}\left[\exp(\mathrm{i}uU_{N,l})\,\middle|\,\mathscr{F}_{N,l-1}\right]-\exp(-u^2\sigma_{N,l}^2/2)\right|\,\middle|\,\mathscr{F}_{N,0}\right]\,. \tag{A.5}$$

For any $\varepsilon>0$, it is easily shown that

$$\mathbb{E}\left[\sum_{l=1}^{M_N}\left|\mathbb{E}\left[\exp\left(\mathrm{i}uU_{N,l}\right)-1+\frac{1}{2}u^2\sigma_{N,l}^2\,\middle|\,\mathscr{F}_{N,l-1}\right]\right|\,\middle|\,\mathscr{F}_{N,0}\right]$$
$$\le\frac{1}{6}\varepsilon|u|^3+u^2\sum_{l=1}^{M_N}\mathbb{E}\left[U_{N,l}^2\mathbb{1}\{|U_{N,l}|\ge\varepsilon\}\,\middle|\,\mathscr{F}_{N,0}\right]\,. \tag{A.6}$$

Since $\varepsilon>0$ is arbitrary, it follows from (A.4) that the right-hand side tends in probability to 0 as $N\to\infty$. Finally, for all $\varepsilon>0$,

$$\mathbb{E}\left[\sum_{l=1}^{M_N}\left|\mathbb{E}\left[\exp\left(-u^2\sigma_{N,l}^2/2\right)-1+\frac{1}{2}u^2\sigma_{N,l}^2\,\middle|\,\mathscr{F}_{N,l-1}\right]\right|\,\middle|\,\mathscr{F}_{N,0}\right]$$
$$\le\frac{u^4}{8}\sum_{l=1}^{M_N}\mathbb{E}\left[\sigma_{N,l}^4\,\middle|\,\mathscr{F}_{N,0}\right]\le\frac{u^4}{8}\left(\varepsilon^2+\sum_{j=1}^{M_N}\mathbb{E}\left[U_{N,j}^2\mathbb{1}\{|U_{N,j}|\ge\varepsilon\}\,\middle|\,\mathscr{F}_{N,0}\right]\right)\,.$$

(A.4) shows that the right-hand side of previous equation tends in probability to 0 as $N\to\infty$. The proof follows. ■

Theorem A.6. *Assume that for each N and* $i=1,\ldots,M_N$, $\mathbb{E}\left[U_{N,i}^2\,\middle|\,\mathscr{F}_{N,i-1}\right]<\infty$ *and*

$$\sum_{i=1}^{M_N}\{\mathbb{E}\left[U_{N,i}^2\,\middle|\,\mathscr{F}_{N,i-1}\right]-(\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right])^2\}\stackrel{\mathbb{P}\text{-}prob}{\longrightarrow}\sigma^2 \qquad \textit{for some } \sigma^2>0\,, \tag{A.7}$$

$$\sum_{i=1}^{M_N}\mathbb{E}\left[U_{N,i}^2\mathbb{1}_{\{|U_{N,i}|\geq\varepsilon\}}\,\middle|\,\mathscr{F}_{N,i-1}\right]\stackrel{\mathbb{P}\text{-}prob}{\longrightarrow}0 \qquad \textit{for any } \varepsilon>0\,. \tag{A.8}$$

Then, for any real u,

$$\mathbb{E}\left[\exp\left(\mathrm{i}u\sum_{i=1}^{M_N}\{U_{N,i}-\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right]\}\right)\,\middle|\,\mathscr{F}_{N,0}\right]\stackrel{\mathbb{P}\text{-}prob}{\longrightarrow}\exp(-(u^2/2)\sigma^2)\,. \tag{A.9}$$

PROOF. We first assume that $\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right]=0$ for all $i=1,\ldots,M_N$, and $\sigma^2=1$. Define $\tau_N:=\max\left\{1\leq k\leq M_N:\sum_{j=1}^k\sigma_{N,j}^2\leq 1\right\}$, with the convention $\max\emptyset=0$. Put $\bar{U}_{N,k}=U_{N,k}$ for $k\leq\tau_N$, $\bar{U}_{N,k}=0$ for $\tau_N<k\leq M_N$ and $\bar{U}_{N,M_N+1}=\left(1-\sum_{j=1}^{\tau_N}\sigma_{N,j}^2\right)^{1/2}Y_N$, where $\{Y_N\}$ are $\mathscr{N}(0,1)$ independent and independent of $\mathscr{F}_{N,M_N}$. Put

$$\sum_{j=1}^{M_N}U_{N,j}=\sum_{j=1}^{M_N+1}\bar{U}_{N,j}-\bar{U}_{N,M_N+1}+\sum_{j=\tau_N+1}^{M_N}U_{N,j}\,. \tag{A.10}$$

We will prove that (a) $\{\bar{U}_{N,j}\}_{1\leq j\leq M_N+1}$ satisfies the assumptions of Lemma A.5, (b) $\bar{U}_{N,M_N+1}\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0$, (c) $\sum_{j=\tau_N+1}^{M_N}U_{N,j}\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0$. If $\tau_N<M_N$, then for any $\varepsilon>0$,

$$0\leq 1-\sum_{j=1}^{\tau_N}\sigma_{N,j}^2\leq\sigma_{N,\tau_N+1}^2\leq\max_{1\leq j\leq M_N}\sigma_{N,j}^2\leq\varepsilon^2+\sum_{j=1}^{M_N}\mathbb{E}\left[U_{N,j}^2\mathbb{1}\{|U_{N,j}|\geq\varepsilon\}\,\middle|\,\mathscr{F}_{N,j-1}\right]$$

Since $\varepsilon>0$ is arbitrary, it follows from (A.8) that $1-\sum_{j=1}^{\tau_N}\sigma_{N,j}^2\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0$, which implies that $\mathbb{E}\left[\bar{U}_{N,M_N+1}^2\,\middle|\,\mathscr{F}_{N,0}\right]\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0$, showing (a) and (b). It remains to prove (c). We have

$$\sum_{j=\tau_N+1}^{M_N}\sigma_{N,j}^2=\sum_{j=1}^{M_N}\sigma_{N,j}^2-1+\left(1-\sum_{j=1}^{\tau_N}\sigma_{N,j}^2\right)\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0. \tag{A.11}$$

For any $\lambda>0$,

$$\left(\mathbb{E}\left[\sum_{j=\tau_N+1}^{M_N}U_{N,j}\mathbb{1}\left\{\sum_{i=\tau_N+1}^{j}\sigma_{N,i}^2\leq\lambda\right\}\right]\right)^2=\mathbb{E}\left[\sum_{j=\tau_N+1}^{M_N}\sigma_{N,j}^2\mathbb{1}\left\{\sum_{i=\tau_N+1}^{j}\sigma_{N,i}^2\leq\lambda\right\}\right]\,.$$

The term between braces converges to 0 in probability by (A.11) and its value is bounded by λ, which shows that $\sum_{j=\tau_N+1}^{M_N}U_{N,j}\mathbb{1}\left\{\sum_{i=\tau_N+1}^{j}\sigma_{N,i}^2\leq\lambda\right\}\stackrel{\mathbb{P}\text{-prob}}{\longrightarrow}0$ Moreover,

$$\mathbb{P}\left(\sum_{j=\tau_N+1}^{M_N}U_{N,j}\mathbb{1}\left\{\sum_{i=\tau_N+1}^{j}\sigma_{N,i}^2>\lambda\right\}\neq 0\right)\leq\mathbb{P}\left(\sum_{i=\tau_N+1}^{M_N}\sigma_{N,i}^2>\lambda\right)\,,$$

which converges to 0 by (A.11). The proof is completed when $\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right]=0$. To deal with the general case, it suffices to set $\bar{U}_{N,i}=U_{N,i}-\mathbb{E}\left[U_{N,i}\,|\,\mathscr{F}_{N,i-1}\right]$ and use Lemma A.3. ■

We may now specialize these results to stationary martingale increment sequences. This result was established by **?**.

Theorem A.7. *Assume that $\{X_k,\ k\in\mathbb{N}\}$ is a strict-sense stationary, ergodic process such that $\mathbb{E}\left[X_1^2\right]$ is finite and $\mathbb{E}\left[X_k|\,\mathscr{F}_{k-1}^X\right]=0$, where $\{\mathscr{F}_k^X,\ k\in\mathbb{N}\}$ is the natural filtration. Then, $n^{-1/2}\sum_{k=1}^n X_k\stackrel{\mathscr{L}_{\mathbb{P}}}{\Longrightarrow}\mathrm{N}(0,\mathbb{E}\left[X_1^2\right])$.*