

## Chapter 7

# Further Topics in Probability

There are two important topics in probability that we haven't yet managed to cover. One is stochastic process theory and the other is simulation. Let's go over some of the key results.

### 7.1 Stochastic Processes

A **stochastic process** on  $\mathbb{R}^K$  is a sequence of random vectors  $\{\mathbf{x}_t\}_{t \geq 0}$  defined on a common probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The definition has relatively little content, but the fact that all these random vectors live on the same probability space should not be overlooked. It means that probabilities such as  $\mathbb{P}\{\mathbf{x}_t \geq 0 \text{ for all } t\}$  or  $\mathbb{P}\{\mathbf{x}_t \leq -10 \text{ for some } t\}$  are well defined. Putting the elements of the process on a common probability space pins down the joint distribution across  $t$ .

To do any kind of analysis with stochastic processes, we need to put some structure on the joint distribution, or, equivalently, on the random elements. In practice, this is almost always done recursively. Examples are given below. For the moment, however, let's look at stochastic processes in an abstract setting and state some definitions.

One kind of stochastic process is an IID sequence. Such sequences are ideal to work with in a statistical setting. For example, the law of large numbers tells us that in an IID environment, finite expectations can always be approximated with sample averages. If we want to know how good the approximation is, we can look to the central limit theorem.

Once we move out of the IID world, however, these helpful properties can break down. Without restrictions on our stochastic process, we can say very little about estimation. The first task, then is to describe some properties that make our stochastic process resemble an IID process in *some* way.

### 7.1.1 Stationarity and Ergodicity

As a first step, let's define a type of process that preserves one of the properties of IID processes (the "ID" part). A stochastic process  $\{\mathbf{x}_t\}_{t \geq 0}$  is called **stationary** if it has the property that the distribution of any subset of these random vectors is unaffected by shifting them forward in time. That is,

$$\mathcal{L}(\mathbf{x}_{t_1}, \dots, \mathbf{x}_{t_k}) = \mathcal{L}(\mathbf{x}_{t_1+m}, \dots, \mathbf{x}_{t_k+m}) \quad (7.1)$$

for any  $m \in \mathbb{N}$  and any sequence of integers  $t_1, \dots, t_k$ .

Clearly, any stationary sequence is identically distributed. In other words, for any stationary sequence, the sequence of marginals  $\{\mathcal{L}(\mathbf{x}_t)\}$  is a constant sequence.

**Example 7.1.1** An IID process  $\{\mathbf{x}_t\}$  is stationary, since independence implies that the laws on both sides of (7.1) are just  $k$  products of the marginal  $\mathcal{L}(\mathbf{x}_1)$ .

**Example 7.1.2** For a simple example of a stationary process, let  $x_1$  be a draw from some arbitrary law  $P$  on  $\mathbb{R}$  and let  $x_{t+1} = x_t$  for all  $t$ . For any collection of integers  $t_1, \dots, t_k$  and any Borel sets  $B_1, \dots, B_k$ , we have

$$\mathbb{P}\{x_{t_1} \in B_1, \dots, x_{t_k} \in B_k\} = \mathbb{P}\{x_1 \in \cap_{i=1}^k B_k\} = P(\cap_{i=1}^k B_k)$$

If we replace each  $x_{t_i}$  with  $x_{t_i+m}$ , we get the same number. Hence  $\{x_t\}$  is stationary.

**Example 7.1.3** A **random walk** on  $\mathbb{R}$  is a stochastic process  $\{x_t\}$  where  $x_t = \sum_{j=1}^t w_j$  for some IID zero-mean process  $\{w_j\}$ . Let  $\sigma^2 := \text{var } w_t$ . If  $\sigma > 0$ , then  $\{x_t\}$  is not stationary, since  $\text{var } x_t = t\sigma^2$ . In particular,  $\mathcal{L}(x_t)$  depends on  $t$ .

Stationarity alone isn't strong enough to give us properties like the LLN and CLT.

**Example 7.1.4** For the process  $\{x_t\}$  in example 7.1.2, the LLN fails whenever  $P$  is nondegenerate. Indeed  $\bar{x}_T := \frac{1}{T} \sum_{t=1}^T x_t = x_1$ , and hence  $\mathcal{L}(\bar{x}_T) = \mathcal{L}(x_1) = P$  for all  $T$ . In particular,  $\bar{x}_T$  does not converge in probability to any constant.

The problem with the process in example 7.1.4 is excessive dependence. New realizations provide insufficient (in fact no) new information. On the flip side, dependent processes can satisfy versions of the LLN if dependence is not too excessive (e.g., if  $x_t$  and  $x_{t+k}$  are "almost independent" for  $k$  sufficiently large.) These are the so-called ergodic processes.

The formal definition of ergodicity varies across different literatures. Here we define a stationary stochastic process  $\{\mathbf{x}_t\}$  on  $\mathbb{R}^K$  to be **ergodic** if the LLN holds; that is, if

$$\frac{1}{T} \sum_{t=1}^T h(\mathbf{x}_t) \xrightarrow{P} \mu_h := \mathbb{E}h(\mathbf{x}_t) \quad \text{as } T \rightarrow \infty \quad (7.2)$$

for any  $\mathcal{B}$ -measurable  $h: \mathbb{R}^N \rightarrow \mathbb{R}$  such that  $\mathbb{E}|h(\mathbf{x}_t)| < \infty$ . (Since  $\{\mathbf{x}_t\}$  is stationary, this expectation does not depend on  $t$ .) One of the main tasks of this chapter will be deriving sufficient conditions for ergodicity in the kinds of processes economists often work with.

Incidentally, a casual definition of ergodicity one often hears is that “cross sectional and time series averages coincide.” What does that mean in the present context?

To answer the question, imagine that  $\mathbf{x}_t = x_t$  is a binary variable with  $x_1 = 1$  if a given individual is currently employed and zero otherwise. In this case,  $\frac{1}{T} \sum_{t=1}^T x_t$  is the fraction of time spent employed over the period  $1, \dots, T$ . With  $h(x) = x$ , the right-hand side of (7.2) is  $\mathbb{E} x_t = \mathbb{P}\{x_t = 1\}$ , or the stationary probability of employment. If we observe time  $t$  employment outcomes  $x_t^1, \dots, x_t^N$  of individuals  $1, \dots, N$  and, moreover, the individuals all follow the same model and are sufficient independent, the cross sectional average  $\frac{1}{N} \sum_{n=1}^N x_t^n$  will also be close to  $\mathbb{P}\{x_t = 1\}$ . Hence, with ergodicity and large samples, cross-sectional and time series averages (almost) coincide.

### 7.1.2 Stochastic Recursive Sequences

In this section let’s turn to a more concrete view of stochastic processes. Afterward we’ll circle back and consider issues like ergodicity again.

Figure 7.1 shows the evolution of US GDP per capita since the early postwar era. The top panel is per capita GDP itself, while the second panel is the growth rate. How could we go about modeling such time series in a parsimonious way?

No reasonable model can be deterministic, since shocks play an important role in year-on-year change in GDP, and an IID process is too simple. We can, however, get close to both of these time series with one of the most common nontrivial stochastic processes: the scalar Gaussian AR(1) model

$$x_{t+1} = b + ax_t + cw_{t+1} \quad \text{with } \{w_t\} \stackrel{\text{iid}}{\sim} N(0, 1) \text{ and } x_0 \text{ given} \quad (7.3)$$

Here  $a, b$ , and  $c$  are parameters, while  $x_t$  is called the **state variable**. Equation (7.3) is an example of a **stochastic difference equation**. The process  $\{x_t\}$  it defines is called a stochastic process or **stochastic recursive sequence**. A realization of the process is called a **time series**.

The dynamics of  $\{x_t\}$  depends on the parameters. We can get some sense of this from simulation. Six individual time series  $\{x_t\}$  are shown in figure 7.2, each generated using a different value of  $a$ . The values of  $b$  and  $c$  are both held fixed at 1, while  $x_0 = 0$ . R code for generating the figure is given in listing 3.

As suggested by the figure, the simulated time paths are sensitive to the value of  $a$ . If  $a$  is outside the interval  $(-1, 1)$ , the series tend to diverge. If  $|a| < 1$ , then the

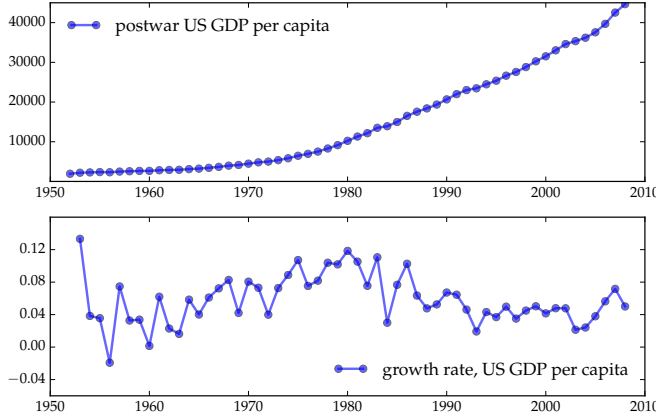


**Figure 7.1** Postwar US GDP per capita. Source: Penn World Tables

opposite is true. For example, if you look at the time series for  $a = 0.9$  in figure 7.2, you will see that, after an initial burn in period where the series is affected by the initial condition  $x_0$ , the process settles down to random motion within a band (between about 5 and 15 in this case).

If the scalar Gaussian AR(1) model in (7.3) is unable to capture important aspects of a time series that we wish to model, we can generalize it in several different directions. For example, we can remove the assumption that the shocks are normally distributed, in which case it is simply called the scalar AR(1) model. This might be desirable, if, say, the state process has heavy tails.

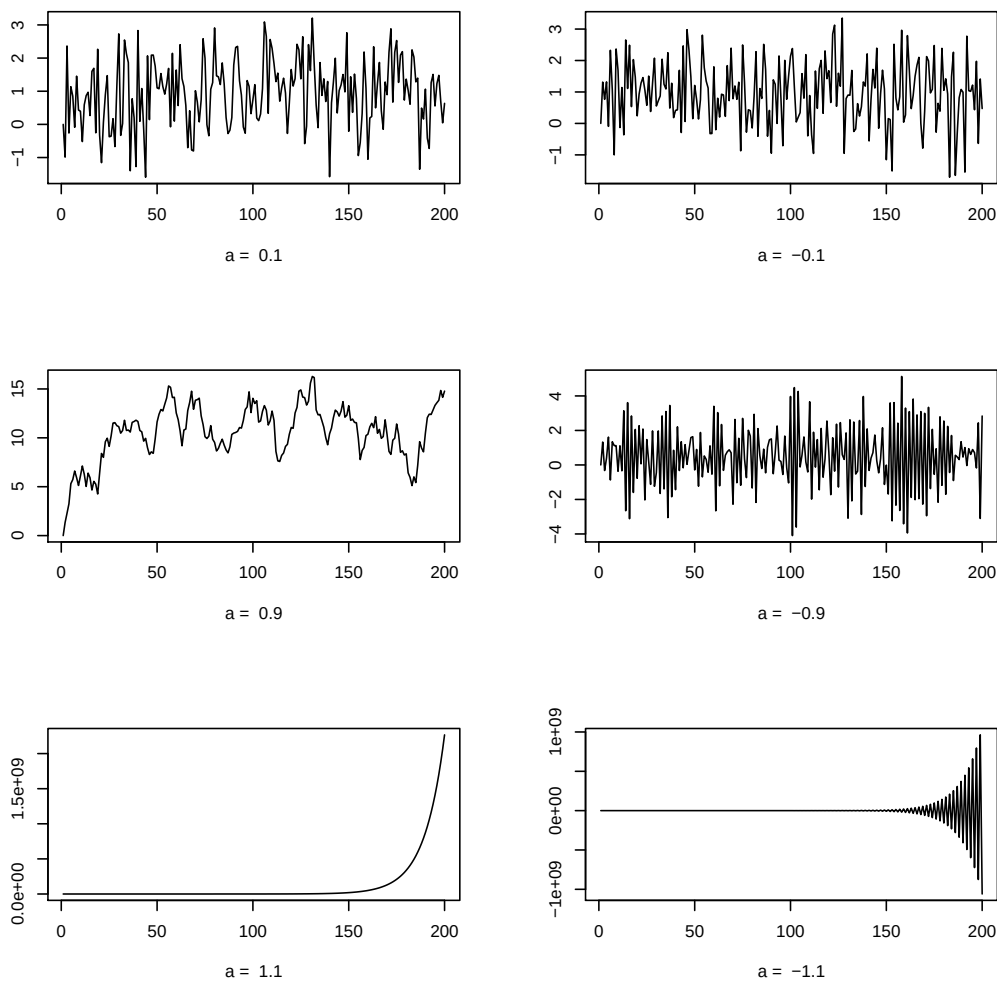
We can also generalize the scalar AR(1) model to  $\mathbb{R}^K$ , yielding the vector AR(1) model, or VAR(1) model

$$\mathbf{x}_{t+1} = \mathbf{b} + \mathbf{A}\mathbf{x}_t + \mathbf{C}\mathbf{w}_{t+1} \quad (7.4)$$

Here  $\{\mathbf{w}_t\}$  is assumed to be IID and to satisfy  $\mathbb{E}\mathbf{w}_t = \mathbf{0}$  and  $\mathbb{E}[\mathbf{w}_t\mathbf{w}_t'] = \mathbf{I}$ . If  $\mathbf{w}_t$  is multivariate normal, then (7.4) is called the Gaussian VAR(1). The vector  $\mathbf{x}_t$  is called the **state vector**. A huge variety of models can be represented as a VAR(1) process or contain VAR processes as subcomponents (e.g., linear state space models).

An alternative generalization of the scalar AR(1) model is the scalar AR( $p$ ) model, where the next state  $x_{t+1}$  is a linear function not just of the current state  $x_t$ , but of the last  $p$  previous states. For example, the AR(2) process has dynamics

$$x_{t+1} = b + ax_t + \gamma x_{t-1} + w_{t+1}$$

**Figure 7.2** Dynamics of the linear AR(1) model

---

**Listing 3** Code for figure 7.2 (R)

---

```
# Generates an AR(1) time series starting from x = init
ar1ts <- function(init, n, b, a) {
  x <- numeric(n)
  x[1] <- init
  w <- rnorm(n-1)
  for (t in 1:(n-1)) {
    x[t+1] <- b + a * x[t] + w[t]
  }
  return(x)
}

avec <- c(0.1, -0.1, 0.9, -0.9, 1.1, -1.1)
N <- 200
par(mfrow=c(3,2)) # Arrangement of figures
for (a in avec) {
  plot(ar1ts(0, N, 1, a), type="l", xlab=paste("a = ", a), ylab="")
}
```

---

Although  $x_{t+1}$  is a function of two lagged states,  $x_t$  and  $x_{t-1}$ , we can reformulate it as a first-order model. To begin, we define an additional state variable  $y_t$  via  $y_t = x_{t-1}$ . The dynamics can then be expressed as

$$\begin{aligned}x_{t+1} &= b + ax_t + \gamma y_t + w_{t+1} \\ y_{t+1} &= x_t\end{aligned}$$

We can write this in matrix form as

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = b \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} a & \gamma \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} w_{t+1}$$

This is a special case of the VAR(1) model in (7.4). This example shows how higher order processes can be reduced to first-order processes by increasing the number of state variables. For this reason most theoretical discussion below takes place in the context of first-order models.

### 7.1.2.1 Nonlinear Models

Linear models are simple but that isn't always an advantage. Sometimes they are unable to capture the kinds of dynamics we observe in data. For example, the bursts

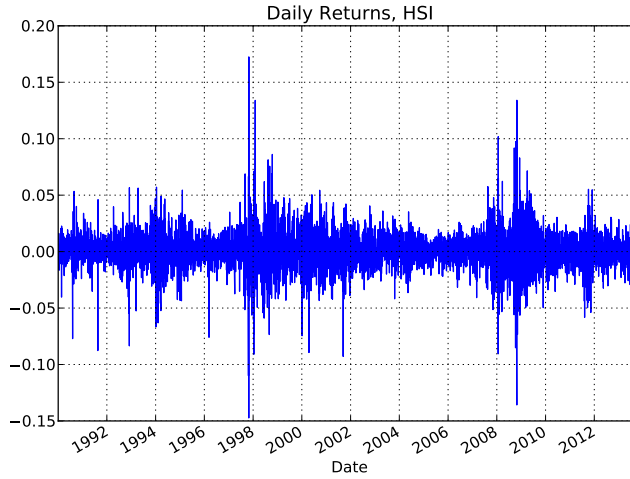


**Figure 7.3** Daily returns on the Hang Seng Index. Source: Yahoo! Finance

of volatility in the graph of daily returns on the Hang Seng Index from 1990 to 2014 shown in figure 7.3 are difficult to replicate without introducing nonlinearities into our model. It's also true that many theoretical modeling exercises produce models that are nonlinear.

One well-known nonlinear model is the  $p$ th order autoregressive conditional heteroskedasticity model (ARCH( $p$ ) model). The model arose from the need to attach a time varying volatility component  $\sigma_t$  to asset return models. Suppose, for example, that  $x_t = \sigma_t w_t$  where  $\sigma_{t+1}^2 = \alpha_0 + \alpha_1 x_t^2$ . Combining these equations gives the ARCH(1) model

$$x_{t+1} = (\alpha_0 + \alpha_1 x_t^2)^{1/2} w_{t+1} \quad \text{with} \quad \{w_t\} \stackrel{\text{iid}}{\sim} N(0, 1) \quad (7.5)$$

and  $\alpha_0 > 0, \alpha_1 \geq 0$ . Still better fits to asset returns data can be obtained via Generalized ARCH models, the simplest of which is the GARCH(1,1) process

$$\begin{aligned} x_t &= \sigma_t w_t \\ \sigma_{t+1}^2 &= \alpha_0 + \alpha_1 x_t^2 + \alpha_2 \sigma_t^2 \end{aligned}$$

Now next period volatility depends on its own lagged state as well as  $x_t$ .

Another popular nonlinear model is the **smooth transition autoregression** (STAR) model

$$x_{t+1} = g(x_t) + w_{t+1} \quad (7.6)$$

where  $g$  is of the form

$$g(s) := (b_0 + a_0s)(1 - \tau(s)) + (b_1 + a_1s)\tau(s)$$

Here  $\tau: \mathbb{R} \rightarrow [0, 1]$  is increasing and satisfies  $\lim_{s \rightarrow -\infty} \tau(s) = 0$  and  $\lim_{s \rightarrow \infty} \tau(s) = 1$ . For  $s$  near  $-\infty$ ,  $g(s)$  is similar to  $b_0 + a_0s$ . For  $s$  near  $+\infty$ ,  $g(s)$  is close to  $b_1 + a_1s$ . Thus the dynamics transition between two different linear models, with the smoothness of the transition depending on the shape of  $\tau$ .

## 7.2 Markov Processes

All of the examples in §7.1.2 are special cases of a general class of process called Markov processes. Markov processes are very rewarding to study because the concepts are broad enough to include a vast array of different stochastic processes, while at the same time having enough structure to support elegant and powerful theory. Here we cover the foundations, from definitions and basic properties to stability and asymptotics.

### 7.2.1 The Markov Assumption

Let  $S$  be a subset of  $\mathbb{R}^K$  and let  $\mathcal{B}(S)$  be the Borel subsets of  $S$ . The fundamental primitive of a discrete time Markov process on  $S$  is a **stochastic kernel** or **transition probability function**, which is a function  $Q: S \times \mathcal{B}(S) \rightarrow [0, 1]$  such that

- (i)  $Q(\mathbf{s}, \cdot)$  is a probability measure over  $\mathcal{B}(S)$  for all  $\mathbf{s} \in S$  and
- (ii)  $g(\mathbf{s}) := Q(\mathbf{s}, B)$  is  $\mathcal{B}$ -measurable for each  $B \in \mathcal{B}(S)$ .

$S$  is called the **state space** of the model. Part (ii) is a mild regularity condition that ensures certain integrals are well defined. Part (i) is the major defining characteristic, yielding a distribution supported on  $S$  for every point  $\mathbf{s}$  in  $S$ . We think of  $Q$  and some initial distribution  $P_0$  as generating a corresponding Markov process  $\{\mathbf{x}_t\}$  taking values in  $S$  via

- 1: draw  $\mathbf{x}_0$  from  $P_0$
- 2: **for**  $t$  in  $0, 1, 2, \dots$  **do**
- 3:     draw  $\mathbf{x}_{t+1}$  from the distribution  $Q(\mathbf{x}_t, \cdot)$
- 4: **end for**

The value  $Q(\mathbf{s}, B)$  is then the probability that  $\mathbf{x}_{t+1}$  lands in  $B$  given  $\mathbf{x}_t = \mathbf{s}$ . The sequence  $\{\mathbf{x}_t\}$  so generated is also called a **sample path** for  $Q$ .

There is a measure-theoretic procedure for building this process  $\{\mathbf{x}_t\}$  more formally but the construction is usually unnecessary because most Markov processes are

introduced via a stochastic difference equation, which specifies both the process  $\{\mathbf{x}_t\}$  and, implicitly, the associated stochastic kernel  $Q$ . The canonical stochastic difference equation for a **first-order Markov process** takes the form

$$\mathbf{x}_{t+1} = G(\mathbf{x}_t, \mathbf{w}_{t+1}) \quad \text{with} \quad \mathcal{L}(\mathbf{x}_0) = P_0 \quad (7.7)$$

Here  $\{\mathbf{w}_t\}_{t \geq 1}$  is an IID sequence of  $\mathbb{R}^M$ -valued shocks with common distribution  $\Psi$  and  $G$  is a given  $\mathcal{B}$ -measurable function mapping the current state  $\mathbf{x}_t \in S$  and shock  $\mathbf{w}_{t+1} \in \mathbb{R}^M$  into the new state  $\mathbf{x}_{t+1} \in S$ . The initial condition  $\mathbf{x}_0$  and the shocks  $\{\mathbf{w}_t\}_{t \geq 1}$  are also assumed to be independent of each other. The corresponding stochastic kernel is

$$Q(\mathbf{s}, B) = \mathbb{P}\{G(\mathbf{s}, \mathbf{w}_{t+1}) \in B\} \quad (\mathbf{s} \in S, B \in \mathcal{B}(S))$$

The choice of date  $t + 1$  is irrelevant here because of the IID assumption, and is used only to tie in with (7.7). Recalling that  $\Psi$  is the distribution of  $\mathbf{w}_{t+1}$ , the stochastic kernel can also be written as

$$Q(\mathbf{s}, B) = \Psi \left\{ \mathbf{w} \in \mathbb{R}^M : G(\mathbf{s}, \mathbf{w}) \in B \right\}$$

This expression makes it clear that  $Q$  is defined by the two primitives  $G$  and  $\Psi$ .

**Example 7.2.1** For the ARCH(1) model of (7.5), the stochastic kernel is

$$Q(s, B) = \mathbb{P}\{(\alpha_0 + \alpha_1 s^2)^{1/2} w_{t+1} \in B\}$$

when  $w_{t+1}$  is standard normal.

Returning to the general process (7.7), repeated substitution gives

$$\begin{aligned} \mathbf{x}_1 &= G(\mathbf{x}_0, \mathbf{w}_1) \\ \mathbf{x}_2 &= G(G(\mathbf{x}_0, \mathbf{w}_1), \mathbf{w}_2) \\ \mathbf{x}_3 &= G(G(G(\mathbf{x}_0, \mathbf{w}_1), \mathbf{w}_2), \mathbf{w}_3) \end{aligned}$$

and so on. Continuing in this fashion, we see that the state vector  $\mathbf{x}_t$  can be written as a function of  $\mathbf{x}_0$  and the shocks  $\mathbf{w}_1, \dots, \mathbf{w}_t$  for any  $t$ . In other words, for each  $t$ , there exists a function  $H_t$  such that

$$\mathbf{x}_t = H_t(\mathbf{x}_0, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_t) \quad (7.8)$$

**Example 7.2.2** For the scalar linear AR(1) process (7.3) we can write the right-hand

side of (7.8) explicitly as

$$x_t = b \sum_{k=0}^{t-1} a^k + \sum_{k=0}^{t-1} a^k c w_{t-k} + a^t x_0 \quad (7.9)$$

This is called the **moving average representation** of  $x_t$ . The proof is an exercise.

Often there is no neat expression for the function  $H_t$ . Nonetheless, (7.8) clarifies the fact that (7.7) pins down each  $\mathbf{x}_t$  as a well-defined random vector, depending on the initial condition and the shocks up until date  $t$ . Since  $\mathbf{x}_t$  is a function of  $\mathbf{x}_0, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_t$ , and these random vectors are all, by the IID assumption, independent of  $\mathbf{w}_{t+k}$  for any  $k \geq 1$ , fact 5.1.10 on page 137 tells us that the current state  $\mathbf{x}_t$  and future shocks are independent. Let's record this:

**Fact 7.2.1** For the process (7.7), the pair  $\mathbf{x}_t$  and  $\mathbf{w}_{t+k}$  are independent for all  $k \geq 1$ .

Although it's rarely made explicit, behind all the random vectors  $\{\mathbf{x}_t\}$  lies a single probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . The idea is that an element  $\omega$  of  $\Omega$  is selected by "nature" at the start of time. This determines the initial condition  $\mathbf{x}_0$  and the shocks  $\mathbf{w}_t$  as

$$\mathbf{x}_0(\omega), \mathbf{w}_1(\omega), \mathbf{w}_2(\omega), \mathbf{w}_3(\omega), \dots$$

From these, each state vector  $\mathbf{x}_t$  is determined via

$$\mathbf{x}_t(\omega) = H_t(\mathbf{x}_0(\omega), \mathbf{w}_1(\omega), \mathbf{w}_2(\omega), \dots, \mathbf{w}_t(\omega))$$

where  $H_t$  is the function in (7.8).

### 7.2.1.1 The Density Case

If  $Q(\mathbf{s}, \cdot)$  is absolutely continuous for all  $\mathbf{s}$ , then we can define the corresponding **stochastic density kernel** or the **transition density** via

$$q(\mathbf{s}, \cdot) := \text{the density of } Q(\mathbf{s}, \cdot) \text{ for all } \mathbf{s} \in S \quad (7.10)$$

In other words,  $q(\mathbf{s}, \cdot)$  is the conditional density of  $\mathbf{x}_{t+1}$  given  $\mathbf{x}_t = \mathbf{s}$ . Heuristically,  $q(\mathbf{s}, \mathbf{s}') d\mathbf{s}'$  is the probability of transitioning from  $\mathbf{s}$  to  $\mathbf{s}'$  in one step.

**Example 7.2.3** Recall the ARCH(1) process, with stochastic kernel  $Q$  as given in example 7.2.1. The transition density  $q(s, \cdot)$  is the density of  $y = (\alpha_0 + \alpha_1 s^2)^{1/2} w_{t+1}$  when  $\mathcal{L}(w_{t+1}) = N(0, 1)$ . Hence

$$q(s, s') = \frac{1}{\sqrt{2\pi\sigma_s^2}} \exp \left\{ -\frac{(s')^2}{2\sigma_s^2} \right\} \quad \text{where} \quad \sigma_s^2 := \alpha_0 + \alpha_1 s^2 \quad (7.11)$$

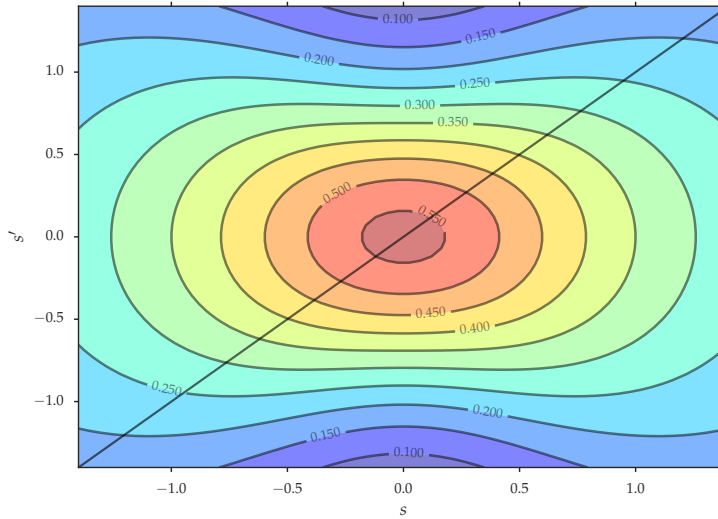


**Figure 7.4** Transition density  $q(s, s')$  for the ARCH kernel (7.11)

A visualization of the ARCH kernel (7.11) is given in figure 7.4 when  $\alpha_0 = 0.5$  and  $\alpha_1 = 0.8$ . The diagonal line is the 45 degree line. Each vertical fiber of the graph corresponds to one density  $s' \mapsto q(s, s')$ . For large  $s$  (to the right on the horizontal axis), the corresponding density puts most mass below the 45 degree line, indicating a tendency to decrease. Conversely, when  $-s$  is large, the process tends to increase.

**Example 7.2.4** Consider a general model with additive shock, so that

$$\mathbf{x}_{t+1} = G(\mathbf{x}_t, \mathbf{w}_{t+1}) = g(\mathbf{x}_t) + \mathbf{w}_{t+1} \quad \text{with} \quad \{\mathbf{w}_t\} \stackrel{\text{iid}}{\sim} \psi \quad (7.12)$$

for some function  $g$  and density  $\psi$ . In this case the transition density has the form

$$q(\mathbf{s}, \mathbf{s}') = \psi(\mathbf{s}' - g(\mathbf{s})) \quad (7.13)$$

In the scalar case, (7.13) follows from example 4.2.16 on page 112, since  $q(s, \cdot)$  is the density of  $y = g(s) + w_{t+1}$  with  $\mathcal{L}(w_{t+1}) = \psi$ . This gives  $q(s, s') = \psi(s' - g(s))$ . The multivariate proof uses a similar argument.

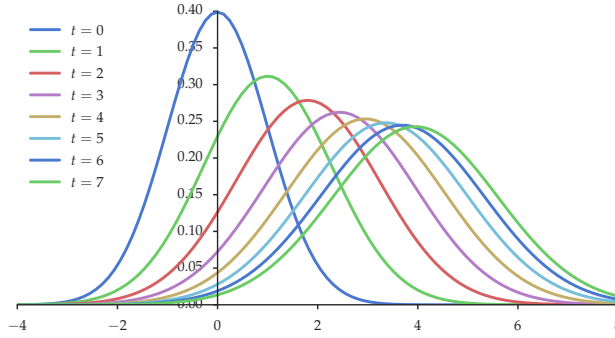


**Figure 7.5** Sequence of marginal densities for the Gaussian AR(1) process

## 7.2.2 Marginal and Joint Distributions

Let's note some features of the marginal and joint distributions of a Markov process. Consider our generic process  $\mathbf{x}_{t+1} = G(\mathbf{x}_t, \mathbf{w}_{t+1})$  from (7.7), where  $\mathcal{L}(\mathbf{x}_0) = P_0$  and  $\{\mathbf{w}_t\}$  is IID on  $\mathbb{R}^M$  with common distribution  $\Psi$ . The **marginal distribution of the state vector** is just  $\mathcal{L}(\mathbf{x}_t)$ . Since  $\mathbf{x}_t$  is a well-defined random vector (see in particular (7.8)),  $\mathcal{L}(\mathbf{x}_t)$  also well-defined, and we denote it in what follows by  $P_t$ . That is,

$$P_t(B) = \mathbb{P}\{\mathbf{x}_t \in B\} \quad (B \in \mathcal{B}(S))$$

**Example 7.2.5** Consider the Gaussian AR(1) process in (7.3). Suppose that  $\mathcal{L}(x_0) = N(\mu_0, \sigma_0^2)$ , where  $\mu_0$  and  $\sigma_0$  are given constants, and that  $x_0$  is independent of the shock process  $\{w_t\}$ . Combining the moving average representation of  $x_t$  in (7.9) with fact 5.1.14 on page 137, we have

$$P_t = N(\mu_t, \sigma_t^2) \quad \text{where} \quad \mu_t := b \sum_{k=0}^{t-1} a^k + a^t \mu_0 \quad \text{and} \quad \sigma_t^2 := \sum_{k=0}^{t-1} a^{2k} c^2 + a^{2t} \sigma_0^2$$

Figure 7.5 shows a visualization of this sequence when  $\mu_0 = 0, \sigma_0 = b = c = 1$  and  $a = 0.8$ . In fact we don't actually need to recompute the sums in the definition of  $\mu_t$  and  $\sigma_t^2$  at each step because these sequences satisfy the recursions

$$\mu_{t+1} = b + a\mu_t \quad \text{and} \quad \sigma_{t+1}^2 = a^2 \sigma_t^2 + c^2 \quad (7.14)$$

You can verify this by taking expectations and variances on both side of (7.3). For code replicating the figure, see [johnstachurski.net/emet.html](http://johnstachurski.net/emet.html).

For a general Markov process, tracking evolution of the first two moments is usu-

ally more complicated, and even then fails to provide full information on the distributions. We do, however, have the following general result:

**Fact 7.2.2** The marginal distributions of a Markov process with stochastic kernel  $Q$  obey the recursion

$$P_{t+1}(B) = \int Q(\mathbf{s}, B) P_t(d\mathbf{s}) \quad (B \in \mathcal{B}(S), t \geq 0) \quad (7.15)$$

Exercise 7.6.4 asks you to prove this. Fact 7.2.2 is a version of the law of total probability (see page 142). To calculate the probability of reaching  $B$  next period, we sum the probability of that event given outcome  $\mathbf{s}$  today, weighted by the probability that  $\mathbf{s}$  occurs today. Here's a parallel result for the density case.

**Fact 7.2.3** If  $Q(\mathbf{s}, \cdot)$  is absolutely continuous for all  $\mathbf{s} \in S$ , then  $P_t$  is absolutely continuous for all  $t \geq 1$ . Letting  $p_t$  be the corresponding density, the sequence  $\{p_t\}$  satisfies

$$p_{t+1}(\mathbf{s}') = \int q(\mathbf{s}, \mathbf{s}') p_t(\mathbf{s}) d\mathbf{s} \quad (B \in \mathcal{B}(S), t \geq 1) \quad (7.16)$$

where  $q$  is the transition density corresponding to  $Q$ .

Let's now look at joint distributions. We'll cover these only in the density case, which is the most important case in econometrics.

**Fact 7.2.4** Let  $\{\mathbf{x}_t\}$  be a Markov process with transition density  $q$  and initial density  $p_0$ . Under the conditions of fact 7.2.3, the joint density  $p_T$  of  $\mathbf{x}_0, \dots, \mathbf{x}_T$  is

$$p_T(\mathbf{s}_0, \dots, \mathbf{s}_T) = p_0(\mathbf{s}_0) \prod_{t=0}^{T-1} q(\mathbf{s}_t, \mathbf{s}_{t+1}) \quad (7.17)$$

This result can be obtained in a few different ways. One is to repeatedly apply (5.27) on page 142 to factor the joint density into

$$p_T(\mathbf{s}_0, \dots, \mathbf{s}_T) = p_0(\mathbf{s}_0) \prod_{t=0}^{T-1} p(\mathbf{s}_{t+1} | \mathbf{s}_t, \dots, \mathbf{s}_0)$$

This is the expression for the joint density of a general stochastic process, where the current state depends on lagged states of all orders. With the first-order Markov property, an expression like  $p(\mathbf{s}_{t+1} | \mathbf{s}_t, \dots, \mathbf{s}_0)$  reduces to  $p(\mathbf{s}_{t+1} | \mathbf{s}_t)$ , which is just  $q(\mathbf{s}_t, \mathbf{s}_{t+1})$ .

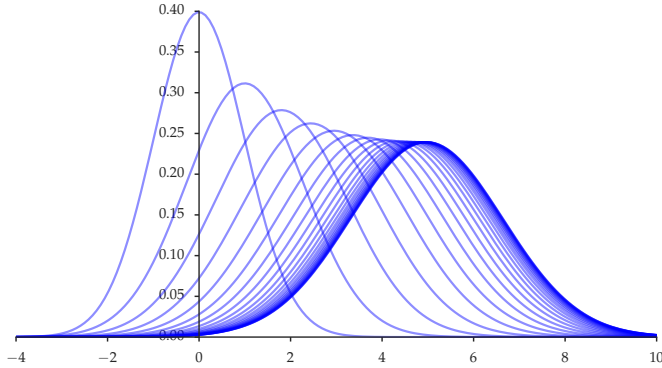


**Figure 7.6** Convergence of marginal densities, Gaussian AR(1) case

### 7.2.3 Stationarity of Markov Processes

Consider the notion of stationarity introduced in §7.1.1. Not every Markov process is stationary. As stated above, a necessary condition is that the marginals are constant, and in figure 7.5 we saw an example of how the marginals  $\{P_t\}$  can change over time. Figure 7.6 shows the same sequence but now longer, with identical parameters. Probability mass moves from left to right.

While  $\{P_t\}$  is not typically constant, for the particular sequence shown in figure 7.6, the difference between successive distributions is diminishing over time. We can see this analytically as well by recalling that  $P_t = N(\mu_t, \sigma_t^2)$ , where  $\mu_t$  and  $\sigma_t^2$  are as given in (7.14). Provided that  $|a| < 1$ , these sequences converge, with

$$\mu_t \rightarrow \mu_\infty := \frac{b}{1-a} \quad \text{and} \quad \sigma_t^2 \rightarrow \sigma_\infty^2 := \frac{c^2}{1-a^2}$$

Let  $P_\infty$  denote the distribution associated with these limits. That is,

$$P_\infty = N(\mu_\infty, \sigma_\infty^2) = N\left[\frac{b}{1-a}, \frac{c^2}{1-a^2}\right] \quad (7.18)$$

What happens if we actually attain this limiting distribution, or if we take our initial draw  $x_0$  from this distribution? On an intuitive level, it seems that the distribution sequence will be constant, remaining at  $P_\infty$  over time. Indeed this is the case, as exercise 7.6.1 asks you to confirm.

The limiting distribution is an example of what is called a stationary distribution. More generally, a **stationary distribution** for a Markov process on  $S$  with stochastic

kernel  $Q$  is any distribution  $P_\infty$  on  $S$  satisfying

$$P_\infty(B) = \int Q(\mathbf{s}, B) P_\infty(d\mathbf{s}) \quad \text{for all } B \in \mathcal{B}(S)$$

Comparing with the recursion  $P_{t+1}(B) = \int Q(\mathbf{s}, B) P_t(d\mathbf{s})$  from fact 7.2.2, we see that the definition of stationarity of  $P_\infty$  means precisely that

$$\mathcal{L}(\mathbf{x}_t) = P_\infty \implies \mathcal{L}(\mathbf{x}_{t+1}) = P_\infty$$

From this we see that any process starting at  $P_\infty$  will be identically distributed. In fact it is stationary, in the sense that (7.1) holds. Let's record this:

**Fact 7.2.5** Let  $\{\mathbf{x}_t\}$  be a Markov process with stationary distribution  $P_\infty$ . If  $\mathcal{L}(\mathbf{x}_0) = P_\infty$ , then  $\{\mathbf{x}_t\}$  is a stationary stochastic process, with  $\mathcal{L}(\mathbf{x}_t) = P_\infty$  for all  $t$ .

**Example 7.2.6** Consider the scalar Gaussian linear AR(1) process from (7.3). If  $|a| < 1$  and  $\mathcal{L}(x_0) = P_\infty$ , where  $P_\infty$  is as in (7.18), then  $\{x_t\}$  is stationary.

Moments of the stationary distribution are called the **stationary moments** of the Markov process. Other names for the stationary distribution include the "invariant" or "unconditional" distribution, and the moments are then called the invariant or unconditional moments.

**Fact 7.2.6** If  $Q(\mathbf{s}, \cdot)$  is absolutely continuous for all  $\mathbf{s} \in S$  with transition density  $q(\mathbf{s}, \cdot)$ , then every stationary distribution  $P_\infty$  is absolutely continuous, and its density  $p_\infty$  satisfies

$$p_\infty(\mathbf{s}') = \int q(\mathbf{s}, \mathbf{s}') p_\infty(\mathbf{s}) d\mathbf{s} \quad \text{for all } \mathbf{s}' \in S \quad (7.19)$$

Such a density is called a **stationary density**.

### 7.2.3.1 Existence

Not every Markov process has a stationary distribution. For example, in the linear scalar AR(1) model  $x_{t+1} = b + ax_t + cw_{t+1}$ , we know that the marginal variance evolves according to  $\sigma_{t+1}^2 = a^2\sigma_t^2 + c^2$ . If  $|a| \geq 1$ , then this sequence diverges. Because the variance is not constant, the distribution sequence cannot be constant either.

Let's state one sufficient condition for existence. The condition we give slightly specializes a famous result known as the Krylov–Bogolyubov theorem. It uses the notion of a **coercive** function, which is any nonnegative function  $V$  on the state space  $S$  such that

$$C_\gamma := \{\mathbf{s} \in S : V(\mathbf{s}) \leq \gamma\}$$

is a compact subset of  $S$  for all  $\gamma \in \mathbb{R}$ . Intuitively, coercive functions are functions that go to  $+\infty$  as we move “away from the center” of the state space.

**Example 7.2.7** If  $S = \mathbb{R}^K$  and  $V(\mathbf{s}) = \|\mathbf{s}\|$ , then  $V$  is coercive on  $S$ . Indeed, for this function, each  $C_\gamma$  is a closed sphere centered on the origin. Closed and bounded subsets of  $\mathbb{R}^K$  are compact.

**Example 7.2.8** If  $S = \mathbb{R}$  and  $V(s) = s^2$ , then  $V$  is coercive on  $S$ . To see this, just observe that  $C_\gamma = [-\sqrt{\gamma}, \sqrt{\gamma}]$ . This is a closed, bounded subset of  $S$ .

Now we can state the following:

**Theorem 7.2.1** Let  $\mathbf{x}_{t+1} = G(\mathbf{x}_t, \mathbf{w}_{t+1})$  be the process on  $S$  defined in (7.7). If

- (i)  $\mathbf{s} \mapsto G(\mathbf{s}, \mathbf{w})$  is continuous for each fixed  $\mathbf{w} \in \mathbb{R}^M$ , and
- (ii) there exists a coercive function  $V$  on  $S$  and constants  $\lambda$  and  $L$  such that  $0 \leq \lambda < 1$  and

$$\mathbb{E} V[G(\mathbf{s}, \mathbf{w}_{t+1})] \leq \lambda V(\mathbf{s}) + L \quad (\mathbf{s} \in S) \quad (7.20)$$

then the Markov process has at least one stationary distribution.

See §11.2 of Stachurski (2009) for proofs and references. Condition (7.20) is an example of a **drift condition**. One way to understand it is to write it as

$$\frac{\mathbb{E} V[G(\mathbf{s}, \mathbf{w}_{t+1})]}{V(\mathbf{s})} \leq \lambda + \frac{L}{V(\mathbf{s})}$$

As we move “away from the center” of the state space,  $V(\mathbf{s})$  becomes large, and the term on the right is eventually  $< 1$ . Looking at the left hand side, this tells us that the value attached to the state by  $V$  is expected to decrease, which in turn suggests movement away from the “edges” of the state space and back towards the center.

**Example 7.2.9** Consider again the ARCH(1) process  $x_{t+1} = (\alpha_0 + \alpha_1 x_t^2)^{1/2} w_{t+1}$  from (7.5), where, as usual  $\alpha_0 > 0$  and  $\alpha_1 \geq 0$ . Let  $S = \mathbb{R}$ . If  $\alpha_1 < 1$ , then this process has a stationary distribution on  $S$ . To see this, observe that

$$G(s, w) = (\alpha_0 + \alpha_1 s^2)^{1/2} w \quad (s, w \in \mathbb{R})$$

is continuous in  $s$  for each fixed  $w$ . Moreover  $V(s) = s^2$  is coercive on  $S$  by example 7.2.8, and

$$\mathbb{E} G(s, w_{t+1})^2 = (\alpha_0 + \alpha_1 s^2) \mathbb{E} w_{t+1}^2 = \alpha_0 + \alpha_1 s^2 \quad (7.21)$$

Hence (7.20) holds with equality when  $V(s) := s^2$ ,  $\lambda := \alpha_1$  and  $L := \alpha_0$ .

### 7.2.4 Asymptotics of Markov Processes

When do Markov processes exhibit nice properties such as ergodicity? To gain some intuition, let's first consider the process  $x_{t+1} = ax_t + w_{t+1}$ , where  $\{w_t\}$  is IID  $N(0, 1)$  and  $|a| < 1$ . We consider a stationary process, where each  $x_t$  has marginal

$$P_\infty := N\left(0, (1 - a^2)^{-1}\right) \quad (7.22)$$

A law of large numbers for this process takes the form

$$\bar{x}_T \xrightarrow{P} \int s P_\infty(ds) = 0 \quad \text{as } T \rightarrow \infty \quad (7.23)$$

Fact 6.1.2 on page 162 tells us that for (7.23) to hold it suffices that

$$\mathbb{E}\bar{x}_T \rightarrow 0 \quad \text{and} \quad \text{var } \bar{x}_T \rightarrow 0 \quad \text{as } T \rightarrow \infty \quad (7.24)$$

The first part is obvious, since  $\mathbb{E}\bar{x}_T = \frac{1}{T} \sum_{t=1}^T \mathbb{E}x_t$  and  $\mathbb{E}x_t = 0$ . For the second part, since  $\mathbb{E}\bar{x}_T = 0$ ,

$$\text{var} \left[ \frac{1}{T} \sum_{t=1}^T x_t \right] = \mathbb{E} \left[ \frac{1}{T} \sum_{t=1}^T x_t \right]^2 = \frac{1}{T^2} \sum_{n=1}^T \sum_{m=1}^T \mathbb{E}[x_n x_m]$$

Using  $\mathbb{E}[x_n x_m] = \mathbb{E}[x_m x_n]$  we obtain the bound

$$\text{var } \bar{x}_T \leq \frac{2}{T^2} \sum_{1 \leq n \leq m \leq T} \mathbb{E}[x_n x_m] = \frac{2}{T^2} \sum_{j=0}^{T-1} \sum_{t=1}^{T-j} \mathbb{E}[x_{t+j} x_t] \quad (7.25)$$

Thus, to obtain  $\text{var } \bar{x}_T \rightarrow 0$  as  $T \rightarrow \infty$ , we require the covariance terms  $\mathbb{E}[x_{t+j} x_t]$  to converge to zero in  $j$  sufficiently quickly. That they do in this particular instance is the subject of exercise 7.6.3.

To summarize, if correlations die out sufficiently quickly, then we can regain the LLN. The same is true for the CLT when second moment conditions hold. We now state a much more general result for sample paths of Markov processes. The proofs are omitted but the essential ideas are the same.

**Theorem 7.2.2** *Let  $\{x_t\}$  be a Markov process on  $S$  with stochastic kernel  $Q$ . Suppose that  $Q(\mathbf{s}, \cdot)$  is absolutely continuous for all  $\mathbf{s} \in S$  and let  $q(\mathbf{s}, \cdot)$  be the corresponding transition density. If*

(a)  *$q$  is strictly positive and continuous on  $S \times S$ , and*

(b) there exists a coercive function  $V$  on  $S$  and constants  $\lambda$  and  $L$  such that  $0 \leq \lambda < 1$  and

$$\int V(\mathbf{s}')q(\mathbf{s}, \mathbf{s}') d\mathbf{s}' \leq \lambda V(\mathbf{s}) + L \quad (\mathbf{s} \in S)$$

then the following statements are true:

- (i)  $Q$  has a unique stationary distribution  $P_\infty$  with density  $p_\infty$ .
- (ii)  $P_t \xrightarrow{w} P_\infty$  as  $t \rightarrow \infty$  for all initial  $P_0$ .
- (iii) If  $h: S \rightarrow \mathbb{R}$  is  $\mathcal{B}$ -measurable and  $\int |h(\mathbf{s})|p_\infty(\mathbf{s}) d\mathbf{s} < \infty$ , then

$$\frac{1}{T} \sum_{t=1}^T h(\mathbf{x}_t) \xrightarrow{p} \int h(\mathbf{s})p_\infty(\mathbf{s}) d\mathbf{s} \quad \text{as } T \rightarrow \infty \quad (7.26)$$

(iv) If, in addition,  $h^2 \leq V$ , then there exists a  $\sigma_h^2 \geq 0$  such that

$$\sqrt{T} \left\{ \frac{1}{T} \sum_{t=1}^T h(\mathbf{x}_t) - \int h(\mathbf{s})p_\infty(\mathbf{s}) d\mathbf{s} \right\} \xrightarrow{d} N(0, \sigma_h^2) \quad \text{as } T \rightarrow \infty \quad (7.27)$$

Conditions (a) and (b) are not dissimilar to the conditions for existence in theorem 7.2.1. Condition (b) is the drift condition in (7.20) specialized to the density setting. The major difference is that  $q$  is required to be strictly positive. This gives enough mixing to generate uniqueness and the sample path properties. Theorem 7.2.2 can be proved from results in Meyn and Tweedie (2009).<sup>1</sup>

Theorem 7.2.2 is but one of a vast range of results on asymptotic properties of Markov processes. Discussion of alternatives is given in §7.5. In fact the conditions are rather strict and can certainly be weakened at the expense of a more complex theorem statement. Nonetheless, they suffice for many applications of interest.

The constant  $\sigma_h^2$  in (7.27) is given by the expression

$$\sigma_h^2 := \text{var } h(\mathbf{x}_0) + 2 \sum_{t=1}^{\infty} \text{cov}[h(\mathbf{x}_0), h(\mathbf{x}_t)] \quad (7.28)$$

with the understanding that  $\{\mathbf{x}_t\}$  is a stationary process. If  $\sigma_h^2 = 0$ , then the limiting distribution is degenerate.

---

1. The stated results are a consequence of theorem 17.0.1 in Meyn and Tweedie (2009), which in turn requires that the process is  $V$ -uniformly ergodic. In our case, the latter property follows from aperiodicity and irreducibility, which are immediate consequences of strict positivity of  $q$  on  $S \times S$ , and from the drift condition (b). In the language of Meyn and Tweedie (2009), the function  $V$  is “unbounded off petite sets.” This holds for any coercive function  $V$  when  $q$  is positive and continuous, since compact sets are then petite.

**Fact 7.2.7** Let  $\mathbf{x}_{t+1} = g(\mathbf{x}_t) + \mathbf{w}_{t+1}$  be the additive shock process on  $\mathbb{R}^K$  from (7.12). If  $\mathbb{E}\|\mathbf{w}_t\|$  is finite,  $g$  is continuous, the density  $\psi$  of  $\mathbf{w}_t$  is continuous and positive everywhere on  $\mathbb{R}^K$ , and there exist constants  $\lambda$  and  $L$  such that  $0 \leq \lambda < 1$  and

$$\|g(\mathbf{s})\| \leq \lambda \|\mathbf{s}\| + L \quad \text{for all } \mathbf{s} \in \mathbb{R}^K \quad (7.29)$$

then the conditions of theorem 7.2.2 are satisfied.

To see why, recall from (7.13) that the transition density satisfies  $q(\mathbf{s}, \mathbf{s}') = \psi(\mathbf{s}' - g(\mathbf{s}))$ . Since  $\psi$  is assumed continuous and positive on  $\mathbb{R}^K$ , the transition density has these properties on  $\mathbb{R}^K \times \mathbb{R}^K$ . Hence condition (a) in theorem 7.2.2 is satisfied. Condition (b) is also satisfied because, taking  $\lambda$  and  $L$  as in (7.29), we have

$$\int \|\mathbf{s}'\| q(\mathbf{s}, \mathbf{s}') d\mathbf{s}' = \mathbb{E} \|g(\mathbf{s}) + \mathbf{w}_{t+1}\| \leq \lambda \|\mathbf{s}\| + L + \int \|\mathbf{w}\| \psi(\mathbf{w}) d\mathbf{w}$$

**Example 7.2.10** Let  $\{x_t\}$  be a scalar AR(1) process  $x_{t+1} = ax_t + b + cw_{t+1}$ , where  $\{w_t\}$  is IID and standard normal. If  $|a| < 1$  and  $c > 0$ , then the conditions of fact 7.2.7 are satisfied. The density of  $cw_t$  is  $N(0, c^2)$ , which is continuous and strictly positive on  $\mathbb{R}$ . To see that (7.29) holds, rewrite the model as  $x_{t+1} = g(x_t) + cw_{t+1}$ , where  $g(s) := as + b$ . By the triangle inequality,  $|g(s)| \leq \lambda|s| + L$ , where  $\lambda := |a|$  and  $L := |b|$ .

**Example 7.2.11** Here's a variation on the scalar "threshold autoregressive" model:

$$x_{t+1} = a|x_t| + (1 - a^2)^{1/2}w_{t+1} \quad \text{with } |a| < 1 \quad \text{and} \quad \{w_t\} \stackrel{\text{IID}}{\sim} N(0, 1)$$

The conditions of fact 7.2.7 are satisfied. To see this, we can rewrite the model as

$$x_{t+1} = g(x_t) + v_{t+1}, \quad g(s) = a|s| \quad \text{and} \quad \{v_t\} \stackrel{\text{IID}}{\sim} N(0, 1 - a^2)$$

Clearly (7.29) is satisfied with  $\lambda = |a|$  and  $L = 0$ . Moreover  $\mathcal{L}(v_t) = N(0, 1 - a^2)$ , which has finite mean and a bounded density everywhere positive on  $\mathbb{R}$ .

For most Markov processes the stationary density has no known closed-form. However, in the case of example 7.2.11, the stationary density is known to have the form  $p_\infty(s) = 2\phi(s)\Phi(qs)$ , where  $q := a(1 - a^2)^{-1/2}$ ,  $\phi$  is the standard normal density and  $\Phi$  is the standard normal CDF.

**Example 7.2.12** Consider the STAR model from (7.6), where the function  $g$  is given by  $g(s) := (b_0 + a_0s)(1 - \tau(s)) + (b_1 + a_1s)\tau(s)$  for some  $\tau: \mathbb{R} \rightarrow [0, 1]$ . Applying the triangle inequality and using  $0 \leq \tau(s) \leq 1$ , we have

$$|g(s)| \leq |b_0| + |b_1| + \max\{|a_0|, |a_1|\}|s|$$

Letting  $L := |b_0| + |b_1|$  and  $\lambda := \max\{|a_0|, |a_1|\}$ , we then have  $|g(s)| \leq \lambda|s| + L$ . If both  $|a_0|$  and  $|a_1|$  are strictly less than one, then condition (7.29) is satisfied.

**Example 7.2.13** Recall the ARCH process  $x_{t+1} = (\alpha_0 + \alpha_1 x_t^2)^{1/2} w_{t+1}$  from (7.5), where  $\{w_t\}$  is IID and standard normal. In example 7.2.9 we saw that this model has a stationary distribution on  $S = \mathbb{R}$  whenever  $\alpha_0 > 0$  and  $0 \leq \alpha_1 < 1$ . In particular, we showed in (7.21) that it satisfies drift with respect to the coercive function  $V(s) = s^2$ . This gives condition (b) of theorem 7.2.2. Condition (a) also holds because the transition density is continuous and everywhere positive (see page 186).

### 7.2.5 The Linear Case

Recall the VAR process in (7.4), which, to repeat, takes the form

$$\mathbf{x}_{t+1} = \mathbf{b} + \mathbf{A}\mathbf{x}_t + \mathbf{C}\mathbf{w}_{t+1} \quad (7.30)$$

where  $\{\mathbf{w}_t\}$  is assumed to be IID and to satisfy  $\mathbb{E}\mathbf{w}_t = \mathbf{0}$  and  $\mathbb{E}[\mathbf{w}_t\mathbf{w}_t'] = \mathbf{I}$ . Regarding the asymptotics of this process, if the shock vector has a continuous and positive density, then we can apply fact 7.2.7 on page 195 to obtain asymptotic stability and LLN and CLT results whenever the matrix norm of  $\mathbf{A}$  satisfies  $\|\mathbf{A}\| < 1$ . To see this, let  $g(\mathbf{s}) = \mathbf{b} + \mathbf{A}\mathbf{s}$ . By the norm triangle inequality and the definition of the matrix norm, we have

$$\|g(\mathbf{s})\| = \|\mathbf{b} + \mathbf{A}\mathbf{s}\| \leq \|\mathbf{b}\| + \|\mathbf{A}\mathbf{s}\| \leq \|\mathbf{b}\| + \|\mathbf{A}\|\|\mathbf{s}\|$$

Hence (7.29) holds.

As mentioned above, the stability and sample path conditions considered above are relatively strict. In the linear case, we can certainly do better. Here's one such result:

**Fact 7.2.8** For the VAR process (7.30), the conclusions of theorem 7.2.2 hold with  $V = \|\cdot\|$  whenever  $\rho(\mathbf{A}) < 1$ .

Here  $\rho(\mathbf{A})$  is the spectral radius of  $\mathbf{A}$ , as defined on page 69. Under this condition, a unique stationary distribution exists. Taking expectations across (7.30), we see that its mean  $\boldsymbol{\mu}_\infty$  must satisfy  $\boldsymbol{\mu}_\infty = \mathbf{b} + \mathbf{A}\boldsymbol{\mu}_\infty$ . Applying the Neumann series lemma (page 69), this equation has the unique solution

$$\boldsymbol{\mu}_\infty = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \sum_{i=0}^{\infty} \mathbf{A}^i \mathbf{b}$$

Let  $\boldsymbol{\Sigma}_\infty$  be the asymptotic variance (i.e., the variance–covariance matrix of the station-

ary distribution). From (7.30) it must satisfy

$$\Sigma_\infty = \mathbf{A}\Sigma_\infty\mathbf{A}^\top + \mathbf{C}\mathbf{C}^\top$$

This is an example of a **discrete time Lyapunov equation**. Under the condition  $\rho(\mathbf{A}) < 1$  it can be solved by iteration or by using existing numerical solvers. Here's an example in Julia that uses a solver from the QuantEcon package:

```
using QuantEcon
```

```
A = [0.8 -0.2;
      -0.1 0.7]
C = [0.5 0.4;
      0.4 0.6]
```

```
solve_discrete_lyapunov(A, C * C')
```

## 7.3 Martingales

Martingales generalize the notion of a random walk, where increments to the current state are draws from a white noise process. The concept of martingales has been central to many advances in probability since the mid twentieth century. Econometricians also like martingales because they neatly capture the idea of “unpredictable” change.

### 7.3.1 Definitions

To define martingales, we need the notion of a filtration, which is an increasing sequence of information sets. Recall from §5.2.4 that an information set is just a set of random variables. Let  $\{\mathcal{F}_t\}$  be a sequence of information sets. That is,  $\mathcal{F}_t$  is an information set for each  $t$ . The sequence  $\{\mathcal{F}_t\}$  is called a **filtration** if, in addition,  $\mathcal{F}_t \subset \mathcal{F}_{t+1}$  for all  $t$ . Intuitively, expanding information sets reflect the idea that more information is revealed over time.

**Example 7.3.1** Let  $\{\mathbf{x}_t\}$  be a sequence of random vectors, and let

$$\mathcal{F}_0 := \emptyset \quad \text{and} \quad \mathcal{F}_t := \{\mathbf{x}_1, \dots, \mathbf{x}_t\} \quad \text{for all } t \geq 1 \quad (7.31)$$

The sequence  $\{\mathcal{F}_t\}$  is a filtration. It is called the **filtration generated by**  $\{\mathbf{x}_t\}$ .

Now let  $\{m_t\}$  be a scalar stochastic process and let  $\{\mathcal{F}_t\}$  be a filtration. We say that  $\{m_t\}$  is **adapted** to  $\{\mathcal{F}_t\}$  if  $m_t$  is  $\mathcal{F}_t$ -measurable for every  $t$ . (For the definition of measurability, see §5.2.4.)

The intuitive meaning is this:  $\mathcal{F}_t$  represents the variables we know as of time  $t$ . If  $\{m_t\}$  is adapted to  $\mathcal{F}_t$ , then we can compute  $m_t$  at time  $t$  as well.

**Example 7.3.2** Let  $\{x_t\}$  be any stochastic process, let  $\{\mathcal{F}_t\}$  be the filtration generated by  $\{x_t\}$  and let  $\bar{x}_t := \frac{1}{t} \sum_{j=1}^t x_j$  be the sample mean up until date  $t$ . Then  $\{\bar{x}_t\}$  is adapted to  $\{\mathcal{F}_t\}$ .

**Example 7.3.3** Let  $\{w_t\}$  be IID standard normal, let  $\{\mathcal{F}_t\}$  be the filtration generated by  $\{w_t\}$  and let  $x_t := \sup_{j \geq t} w_j$ . This process is not adapted to  $\{\mathcal{F}_t\}$  because the value  $x_t$  cannot be determined from  $\{w_1, \dots, w_t\}$ .

**Fact 7.3.1** If  $\{m_t\}$  is adapted to  $\{\mathcal{F}_t\}$ , then  $\mathbb{E}[m_t | \mathcal{F}_{t+j}] = m_t$  for any  $j \geq 0$ .

The proof goes like this: By adaptedness,  $m_t$  is  $\mathcal{F}_t$ -measurable. From the definition of a filtration we have  $\mathcal{F}_t \subset \mathcal{F}_{t+j}$ . Applying fact 5.2.10 on page 150, it follows that  $m_t$  is  $\mathcal{F}_{t+j}$ -measurable. Fact 7.3.1 now follows from fact 5.2.13 on page 151.

A sequence of integrable random variables  $\{m_t\}$  adapted to a filtration  $\{\mathcal{F}_t\}$  is called a **martingale** with respect to  $\{\mathcal{F}_t\}$  if

$$\mathbb{E}[m_{t+1} | \mathcal{F}_t] = m_t \quad \text{for all } t$$

The expression  $\mathbb{E}[m_{t+1} | \mathcal{F}_t]$  is sometimes simplified to  $\mathbb{E}_t[m_{t+1}]$  when the filtration is understood.

**Example 7.3.4** (Random Walk) Let  $\{\eta_t\}$  be an IID sequence of zero-mean random variables and let  $m_t := \sum_{j=1}^t \eta_j$ . For example,  $\eta_t$  might be the payoff on the  $t$ th round of a fair game and  $m_t$  is the wealth of a gambler after the  $t$ th round.<sup>2</sup> We can also express  $\{m_t\}$  recursively via  $m_{t+1} = m_t + \eta_{t+1}$  with  $m_0 = 0$ . The process  $\{m_t\}$  is a martingale with respect to the filtration  $\{\mathcal{F}_t\}$  generated by  $\{\eta_t\}$ . That  $\{m_t\}$  is adapted to  $\{\mathcal{F}_t\}$  follows immediately from the definitions. Moreover

$$\mathbb{E}[m_{t+1} | \mathcal{F}_t] = \mathbb{E}[m_t + \eta_{t+1} | \mathcal{F}_t] = m_t + \mathbb{E}[\eta_{t+1} | \mathcal{F}_t]$$

where the last equality follows from linearity and fact 7.3.1. Since  $\{\eta_t\}$  is IID, we know that  $\eta_{t+1}$  is independent of the variables in  $\mathcal{F}_t$ , and hence  $\mathbb{E}[\eta_{t+1} | \mathcal{F}_t] = \mathbb{E}[\eta_{t+1}] = 0$ . Therefore  $\{m_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$  as claimed.

**Example 7.3.5** A well-known example of a martingale from economic theory is the consumption process in Hall (1978). To understand the idea, consider the Euler equation

$$u'(c_t) = \mathbb{E}_t \left[ \frac{1 + r_{t+1}}{1 + \rho} u'(c_{t+1}) \right]$$

2. We are assuming that wealth starts at zero and may take arbitrarily large negative values without the gambler getting knee-capped by the mafia.

where  $u'$  is the derivative of a utility function  $u$ ,  $r_t$  is an interest rate and  $q$  is a discount factor. The expectation  $\mathbb{E}_t[\cdot]$  can be thought of as  $\mathbb{E}[\cdot | \mathcal{F}_t]$ , where  $\mathcal{F}_t$  contains all variables observable at time  $t$ . Specializing to the case  $r_{t+1} = q$  and  $u(c) = c - ac^2/2$ , the Euler equation reduces to

$$c_t = \mathbb{E}_t[c_{t+1}] =: \mathbb{E}[c_{t+1} | \mathcal{F}_t]$$

Thus consumption is a martingale with respect to  $\{\mathcal{F}_t\}$ .

### 7.3.1.1 Martingale Difference Sequences

A stochastic process  $\{d_t\}$  is called a **martingale difference sequence** (or **MDS**) with respect to some filtration  $\{\mathcal{F}_t\}$  if

$$\mathbb{E}[d_{t+1} | \mathcal{F}_t] = 0 \quad \text{for all } t$$

An MDS is so named because if  $\{m_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ , then  $d_t = m_t - m_{t-1}$  is an MDS with respect to  $\{\mathcal{F}_t\}$ . See exercise 7.6.7.

**Fact 7.3.2** If  $\{d_t\}$  is an MDS, then  $\mathbb{E}d_t = 0$  for all  $t$ .

Indeed, by the law of iterated expectations,  $\mathbb{E}d_t = \mathbb{E}[\mathbb{E}[d_t | \mathcal{F}_{t-1}]] = \mathbb{E}0 = 0$ .

### 7.3.2 Martingale Difference LLN and CLT

Next we consider asymptotics for martingale difference sequences. Such sequences are good candidates for the LLN and CLT. To see this, recall that integrable IID sequences always satisfy the LLN. With finite second moment they also satisfy the CLT. Suppose now that  $\{m_t\}$  is a stationary MDS with respect to some filtration  $\{\mathcal{F}_t\}$ . We don't have independence in this case but we do have the following:

**Fact 7.3.3** If  $\{m_t\}$  is an MDS, then  $\text{cov}[m_{t+j}, m_t] = 0$  for all  $j \geq 1$ .

To see this, fix  $t \geq 0$  and  $j \geq 1$ . We have

$$\text{cov}[m_{t+j}, m_t] = \mathbb{E}[m_{t+j}m_t] = \mathbb{E}[\mathbb{E}[m_{t+j}m_t | \mathcal{F}_{t+j-1}]]$$

Since  $t + j - 1 \geq t$  and  $\{\mathcal{F}_t\}$  is a filtration, we know that  $m_t$  is  $\mathcal{F}_{t+j-1}$ -measurable. Hence

$$\mathbb{E}[\mathbb{E}[m_{t+j}m_t | \mathcal{F}_{t+j-1}]] = \mathbb{E}[m_t \mathbb{E}[m_{t+j} | \mathcal{F}_{t+j-1}]] = \mathbb{E}[m_t \cdot 0] = 0$$

The lack of correlation in fact 7.3.3 is a key part of the next result.

**Theorem 7.3.1** *If  $\{m_t\}$  is a stationary MDS with respect to some filtration  $\{\mathcal{F}_t\}$ , then*

$$\frac{1}{T} \sum_{t=1}^T m_t \xrightarrow{p} 0 \quad \text{as } T \rightarrow \infty \quad (7.32)$$

*If, in addition,  $\gamma^2 := \mathbb{E}[m_t^2] < \infty$  and  $\frac{1}{T} \sum_{t=1}^T \mathbb{E}[m_t^2 \mid \mathcal{F}_{t-1}] \xrightarrow{p} \gamma^2$  as  $T \rightarrow \infty$ , then*

$$T^{-1/2} \sum_{t=1}^T m_t \xrightarrow{d} N(0, \gamma^2) \quad \text{as } T \rightarrow \infty \quad (7.33)$$

The proof of (7.32) is identical to that of the classical LLN in theorem 6.1.1. Theorem 7.3.1 can be established from a martingale CLT proved in Durrett (2010).

## 7.4 Simulation

Simulation is invaluable as a tool for exploring statistical concepts and solving applied problems. For example, the current popularity of Bayesian methods for econometrics and machine learning has been driven to a large extent by clever simulation algorithms that have proved successful at solving hard estimation problems.

Modern scientific programming environments all have good routines for simulating independent observations from a range of common named distributions. We've seen examples of this in the displayed computer code. But it's easy to run into situations where the distribution we are working with is nonstandard, or has no closed-form solution. We then need to roll our own routines for producing random variates. Let's review two of the most important algorithms.

### 7.4.1 Inverse Transforms

Suppose that we wish to draw from a distribution on  $\mathbb{R}$  with CDF  $F$ . Assuming that we can generate uniform random variables, we can do so via the **inverse transform** method:

- 1: draw  $u$  from the uniform distribution on  $[0, 1]$
- 2: return  $F^{-1}(u)$ , where  $F^{-1}$  is the quantile function of  $F$  (see (4.35))

We wish to show that this technique is valid, or, equivalently, that

$$\mathcal{L}(u) = U[0, 1] \implies \mathcal{L}[F^{-1}(u)] = F \quad (7.34)$$

In the proof we'll assume that  $F$  is strictly increasing. (This isn't necessary, but the proof of the general case provides no additional intuition.) Let  $\mathcal{L}(u) = U[0, 1]$  and let

$z := F^{-1}(u)$ . To see that  $\mathcal{L}(z) = F$ , observe that, by strict monotonicity,  $F(a) \leq F(b)$  if and only if  $a \leq b$ . As a result, for any  $s \in \mathbb{R}$ ,

$$\mathbb{P}\{z \leq s\} = \mathbb{P}\{F^{-1}(u) \leq s\} = \mathbb{P}\{u \leq F(s)\}$$

The defining property of a  $U[0, 1]$  random variable  $u$  is that  $\mathbb{P}\{u \leq t\} = t$  whenever  $0 \leq t \leq 1$ . Using this fact in the preceding calculation gives  $\mathbb{P}\{z \leq s\} = F(s)$ , proving (7.34).

Here's one extension for the basic inverse transform method. Suppose that we want to simulate from a joint distribution on  $\mathbb{R}^N$  that is defined in terms of a copula (see §5.1.5). In particular, suppose that  $F_n$  is a continuous CDF on  $\mathbb{R}$  for  $n = 1, \dots, N$ , that  $C$  is a copula on  $\mathbb{R}^N$  and that the joint distribution of interest is

$$F(s_1, \dots, s_N) = C(F_1(s_1), \dots, F_N(s_N))$$

(This is just (5.22) from page 138.) If we can simulate from  $C$ , then we can simulate from  $F$ . The method is

- 1: draw  $u_1, \dots, u_N$  from  $C$
- 2: return  $(x_1, \dots, x_N) := (F_1^{-1}(u_1), \dots, F_N^{-1}(u_N))$

Exercise 7.6.13 asks you to prove that  $(x_1, \dots, x_N)$  is a draw from  $F$ .

## 7.4.2 Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is a way of simulating from a given density  $\pi$  on  $S \subset \mathbb{R}^K$ . The idea is to construct a stochastic kernel  $P$  on  $S$  such that

- (i)  $\pi$  is a stationary distribution for  $P$ , and
- (ii)  $P$  is sufficiently ergodic that its sample path averages converge to expectations under  $\pi$

See (7.26) on page 194 for a clearer statement of (ii). If we specialize to indicator functions we can approximate probabilities under  $\pi$  too.

There are two major strands of MCMC algorithms: Metropolis–Hastings and the Gibbs sampler. Here we'll cover the former. While boldface symbols are omitted in what follows, the ideas are valid in any dimension. In the section on Bayesian estimation in §8.3.4, I will explain why Metropolis–Hastings fits so beautifully with Bayesian updating.

The **Metropolis–Hastings algorithm** starts with a Markov process in the form of a stochastic density kernel  $q = q(s, s')$  called the **proposal density**. Draws from the proposal density are called **proposals**. Each time we draw a proposal, we either (i) accept it and move to that new state, or (ii) reject it and stay put. The probability of

accepting is structured so that the chain tends to stay in regions where  $\pi$  puts most probability mass. As a consequence, for a sequence  $\{x_t\}$  generated by this process, we get

$$\text{fraction of time spent in } B = \frac{1}{T} \sum_{t=1}^T \mathbb{1}\{x_t \in B\} \approx \pi(B) \quad \text{for large } T$$

This is a version of (7.26) with  $h = \mathbb{1}_B$ .

The ideas are subtle, so let's build up in steps. Consider first a general accept-reject style process of the kind discussed above, where the probability of acceptance is determined by an arbitrary function  $\alpha = \alpha(x_t, y)$  of the current state  $x_t$  and proposal  $y$ . The function takes values in  $[0, 1]$ . The algorithm for generating  $x_{t+1}$  from  $x_t$  given  $\alpha$  and proposal density  $q$  is given in algorithm 1.

---

**Algorithm 1** A Markov process with proposal density  $q$  and acceptance function  $\alpha$

---

```

1: draw  $y$  from  $q(x_t, \cdot)$ 
2: draw  $u$  independently from the uniform  $[0, 1]$  distribution
3: if  $u \leq \alpha(x_t, y)$  then ▷ with probability  $\alpha(x_t, y)$ 
4:   set  $x_{t+1} = y$ 
5: else ▷ with probability  $1 - \alpha(x_t, y)$ 
6:   set  $x_{t+1} = x_t$ 
7: end if

```

---

The stochastic kernel  $P$  associated with  $\{x_t\}$  in algorithm 1 has the form

$$P(s, B) = \int_B p(s, s') ds' + (1 - \lambda(s)) \mathbb{1}\{s \in B\} \quad (7.35)$$

for all  $s \in S$  and  $B \in \mathcal{B}(S)$ , where

$$p(s, s') := q(s, s')\alpha(s, s') \quad \text{and} \quad \lambda(s) := \int p(s, s') ds'$$

To show this, we can break the probabilities up as follows. Fix any Borel set  $B$  and let  $x_t$  be given at  $s$ . Let  $x_{t+1}$  be drawn according to algorithm 1 when  $x_t = s$ . By additivity of  $\mathbb{P}$  we can write  $\mathbb{P}\{x_{t+1} \in B\}$  as

$$\mathbb{P}\{x_{t+1} \in B\} \cap \{u \leq \alpha(s, y)\} + \mathbb{P}\{x_{t+1} \in B\} \cap \{u > \alpha(s, y)\} \quad (7.36)$$

Our aim is to show that the two terms in (7.36) match the two terms on the right-hand side of (7.35).

Using the fact that intersections become products when we switch to indicators

(fact 4.1.4, page 90), we can applying the law of iterated expectations to obtain

$$\begin{aligned}
 \mathbb{P}\{x_{t+1} \in B\} \cap \{u \leq \alpha(s, y)\} &= \mathbb{P}\{y \in B\} \cap \{u \leq \alpha(s, y)\} \\
 &= \mathbb{E} [\mathbb{1}\{y \in B\} \mathbb{1}\{u \leq \alpha(s, y)\}] \\
 &= \mathbb{E} \{ \mathbb{E} [\mathbb{1}\{y \in B\} \mathbb{1}\{u \leq \alpha(s, y)\} \mid y] \} \\
 &= \mathbb{E} \{ \mathbb{1}\{y \in B\} \mathbb{E} [\mathbb{1}\{u \leq \alpha(s, y)\} \mid y] \}
 \end{aligned}$$

The term  $\mathbb{E} [\mathbb{1}\{u \leq \alpha(s, y)\} \mid y]$  is just  $\mathbb{P}\{u \leq \alpha(s, y)\}$  with  $y$  given, which, since  $u$  is uniform, evaluates to  $\alpha(s, y)$ . Hence the final expression in the last display evaluates to

$$\begin{aligned}
 \int \mathbb{1}\{s' \in B\} \alpha(s, s') q(s, s') ds' &= \int_B \alpha(s, s') q(s, s') ds' \\
 \therefore \mathbb{P}\{x_{t+1} \in B\} \cap \{u \leq \alpha(s, y)\} &= \int_B p(s, s') ds' \tag{7.37}
 \end{aligned}$$

Equation (7.37) says that the first term in (7.36) is equal to the first term on the right-hand side of (7.35). For the second terms, note that  $x_{t+1} = x_t = s$  when  $u > \alpha(s, y)$ , so

$$\begin{aligned}
 \mathbb{P}\{x_{t+1} \in B\} \cap \{u > \alpha(s, y)\} &= \mathbb{E} [\mathbb{1}\{s \in B\} \mathbb{1}\{u > \alpha(s, y)\}] \\
 &= \mathbb{1}\{s \in B\} \mathbb{E} [\mathbb{1}\{u > \alpha(s, y)\}] \\
 &= \mathbb{1}\{s \in B\} [1 - \mathbb{P}\{u \leq \alpha(s, y)\}]
 \end{aligned}$$

To match this expression with the second term in (7.35), we just need to show that  $\mathbb{P}\{u \leq \alpha(s, y)\} = \int p(s, s') ds'$ . But this is immediate from (7.37) if we take  $B = S$ .

So far we have verified that the kernel  $P$  in (7.35) is the stochastic kernel corresponding to algorithm 1. Now let's double back and design the selection function  $\alpha$  so that the stationary distribution of  $P$  is the target density  $\pi$ . In the Metropolis–Hastings algorithm, this is done by setting

$$\alpha(s, s') := \min \left\{ \frac{\pi(s') q(s', s)}{\pi(s) q(s, s')}, 1 \right\} \tag{7.38}$$

If  $\pi(s) q(s, s') = 0$ , then  $\alpha(s, s') := 1$ . The min is in place so  $\alpha(s, s')$  can be interpreted as a probability. We don't actually need to impose it in algorithm 1, since for  $u \in [0, 1]$  we have  $u \leq \min\{a, 1\}$  iff  $u \leq a$ . The important element is the ratio inside the min, which is large when the new proposal  $s'$  has high probability under  $\pi$  relative to the current location  $s$  and small when the reverse is true. This keeps the chain in regions where  $\pi$  has high mass, and leads to our next result.

**Theorem 7.4.1** *Let  $P$  be the stochastic kernel in (7.35). If  $\alpha$  is as defined (7.38), then  $\pi$  is a*

stationary distribution for  $P$ .

*Proof.* The proof turns on the fact that for this choice of  $\alpha$  we have  $p(s, s')\pi(s) = p(s', s)\pi(s')$  for all  $s, s'$  (see ex. 7.6.14). From that observation and the definition of  $P$  we obtain

$$\begin{aligned}
 \int P(s, B)\pi(s) \, ds &= \int \left\{ \int_B p(s, s') \, ds' \right\} \pi(s) \, ds + \int (1 - \lambda(s)) \mathbb{1}\{s \in B\} \pi(s) \, ds \\
 &= \int_B \left\{ \int p(s, s') \pi(s) \, ds \right\} ds' + \int_B (1 - \lambda(s)) \pi(s) \, ds \\
 &= \int_B \left\{ \int p(s', s) \pi(s') \, ds \right\} ds' + \int_B (1 - \lambda(s)) \pi(s) \, ds \\
 &= \int_B \lambda(s') \pi(s') \, ds' + \int_B (1 - \lambda(s)) \pi(s) \, ds \\
 &= \int_B \pi(s) \, ds
 \end{aligned}$$

Thus  $\pi$  is a stationary distribution of  $P$ , as was to be shown.  $\square$

It remains to establish that  $P$  has the kind of stability properties discussed in theorem 7.2.2 on page 193 (and in particular the LLN result in (7.26) of that theorem). This isn't universally true, and requires at least some conditions on the primitives  $q$  and  $\pi$ . One well-known result is given in corollary 3 of Tierney (1994) for the case where  $q$  and  $\pi$  are continuous and strictly positive on  $S$  and  $S$  itself is compact. In essence, the continuity and positivity corresponds to condition (a) of theorem 7.2.2, while the compactness of  $S$  is a strict version of the drift condition in (b) of the same theorem.<sup>3</sup> For further results and references, see Roberts and Rosenthal (2004).

One disadvantage of MCMC methods is that the variates they generate are not in general independent. Sample averages from dependent samples are typically less informative than IID samples. (See example 7.1.4 for the extreme case.) It is desirable, then, that dependence between  $x_t$  and  $x_{t+k}$  dies out quickly as  $k$  increases. Whether it does or not is determined by the choice of  $q$ . Extensive discussion of such issues can be found in Brooks et al. (2011).

A function implementing the Metropolis–Hastings algorithm in Julia with a random walk proposal density is given in listing 4. The proposal density takes the form  $q(s, s') = \psi(s' - s)$ , where  $\psi$  is a given density. This is the system (7.12)–(7.13) with  $g$  set to the identity. If  $\psi$  is symmetric, then  $q(s, s') = q(s', s)$ , and  $\alpha(s, s')$  simplifies to  $\min\{\pi(s')/\pi(s), 1\}$  if  $\pi(s) > 0$  and 1 otherwise. This is the case shown in listing 4,

---

3. When  $S$  is compact, every constant nonnegative function on  $S$  is coercive, and if  $V$  is constant then the drift condition in (7.20) is trivial.

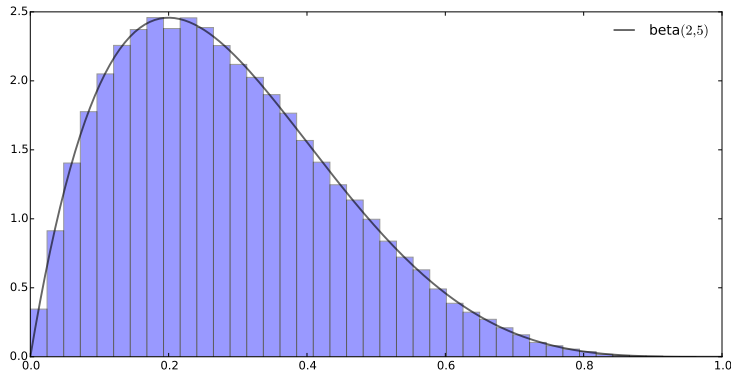


**Figure 7.7** Observations from the random walk Metropolis–Hastings algorithm

with  $\psi$  set to the  $N(0, \sigma^2)$  distribution. Look up Julia’s ternary operator if you find it hard to follow the code.

---

**Listing 4** The random walk Metropolis–Hastings algorithm (Julia)

---

```
function rw_metropolis(pi_density, T, init=0, sigma=1)
    xvec = Array{Float64, T}()
    xvec[1] = init
    for t in 1:(T - 1)
        x = xvec[t]
        y = x + sigma * randn()
        alpha = pi_density(x) > 0 ? pi_density(y) / pi_density(x) : 1
        xvec[t+1] = rand() < alpha ? y : x
    end
    return xvec
end
```

---

Figure 7.7 shows a histogram of  $10^6$  draws generated from this algorithm when  $\pi$  is the beta distribution with parameters (2, 5). The black line is the target density. This is just a toy example, since routines for sampling from beta distributions are readily available. You can experiment with the code at [johnstachurski.net/emet.html](http://johnstachurski.net/emet.html).

## 7.5 Further Reading

The modern theory of Markov processes owes much to a brilliant mathematician called Wolfgang Doeblin. His short life and shining contributions to probability theory are discussed in Bru and Yor (2002).

An excellent overview of Markov processes and their stability can be found in Meyn and Tweedie (2009). For more on Markov processes from a density perspective see Lasota and Mackey (1994) or chapter 8 of Stachurski (2009). For a nice survey of sample path properties with a focus on the central limit theorem, see Jones (2004).

Diaconis (2009) provides a fascinating overview of MCMC and its impact on different fields. Richey (2010) is a more general survey. Good references on the theoretical side include Roberts and Rosenthal (2004), Tierney (1994), and Geweke (1996). See Kroese and Chan (2014) for a more practical discussion with a focus on applications.

## 7.6 Exercises

**Ex. 7.6.1** Show that the process in example 7.2.6 on page 191 is identically distributed.

**Ex. 7.6.2** Let  $x_{t+1} = ax_t + w_{t+1}$ , where  $\{w_t\}$  is IID standard normal and  $\mathcal{L}(x_0) = \mathcal{N}(0, 1/(1 - a^2))$ . Show that  $\text{cov}[x_t, x_{t+j}] = a^j(1 - a^2)^{-1}$  for all  $j \geq 0$ .

**Ex. 7.6.3** Show that the right-hand side of (7.25) converges to zero in  $T$ .

**Ex. 7.6.4** Prove fact 7.2.2 on page 189.<sup>4</sup>

**Ex. 7.6.5** Consider the additive shock process in fact 7.2.7. Assume the conditions stated in the fact and let  $\mathbb{E}\|x_0\|$  be finite. Show that  $\mathbb{E}\|x_t\|$  is bounded in  $t$ .

**Ex. 7.6.6** Consider again the scalar Markov sequence  $x_{t+1} = ax_t + w_{t+1}$ . Assume that  $\{w_t\}$  is IID, having Student's  $t$ -distribution with 2 degrees of freedom, and that  $|a| < 1$ . Show that the conditions of fact 7.2.7 are satisfied.

**Ex. 7.6.7** Let  $\{\mathcal{F}_t\}$  be a filtration. Show that if  $\{m_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ , then  $d_t = m_t - m_{t-1}$  is a martingale difference sequence with respect to  $\{\mathcal{F}_t\}$ .

**Ex. 7.6.8** Let  $\{\mathcal{F}_t\}$  be a filtration, and let  $\{m_t\}$  be a martingale with respect to  $\{\mathcal{F}_t\}$ . Let  $\alpha_t := m_t + 1$ , let  $\kappa_t := 2m_t$ , and let  $\gamma_t := m_t^2$ . Which of these is a martingale with respect to  $\{\mathcal{F}_t\}$ ?

**Ex. 7.6.9** Let  $\{\eta_t\}$  be an IID sequence of scalar random variables with  $\mathbb{E}\eta_1 = 0$  and  $\text{var}\eta_1 = \sigma^2 > 0$ . Let  $\{\mathcal{F}_t\}$  be the filtration defined by  $\mathcal{F}_t := \{\eta_1, \dots, \eta_t\}$ , and let  $z_t := t \cdot \eta_t$  for each  $t$ .

---

4. Hint: Try converting probabilities into expectations of indicators and using the law of iterated expectations.

- (i) Is  $\{z_t\}$  IID? Why or why not?  
 (ii) Is  $\{z_t\}$  a martingale difference sequence with respect to  $\{\mathcal{F}_t\}$ ? Why or why not?

**Ex. 7.6.10** Let  $\{\eta_t\}$  be an IID sequence of scalar random variables with  $\mathbb{P}\{\eta_1 = 1\} = \mathbb{P}\{\eta_1 = -1\} = 0.5$  and let  $\{\mathcal{F}_t\}$  be the filtration defined by  $\mathcal{F}_t := \{\eta_1, \dots, \eta_t\}$ . Let

$$m_t := \sum_{j=1}^t \eta_j \quad \text{and} \quad \kappa_t := m_t^2 - t$$

Show that  $\{\kappa_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ .

**Ex. 7.6.11** Consider the scalar sequence  $x_{t+1} = ax_t + w_{t+1}$ , where  $\{w_t\} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$  and  $x_0 = 0$ . Let  $\mathcal{F}_t := \{w_1, \dots, w_t\}$ . Give conditions on  $a$  such that

- (i)  $\{x_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$  and  
 (ii)  $\{x_t\}$  is a martingale difference sequence with respect to  $\{\mathcal{F}_t\}$ .

**Ex. 7.6.12** Let  $\{\mathcal{F}_t\}$  be a filtration. A sequence of random variables  $\{\gamma_t\}$  is called **previsible** with respect to  $\{\mathcal{F}_t\}$  if  $\gamma_t$  is  $\mathcal{F}_{t-1}$ -measurable for every  $t$ . Show that if  $\{x_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ , and  $\{\gamma_t\}$  is previsible with respect to  $\{\mathcal{F}_t\}$ , then the process  $\{y_t\}$  defined by

$$y_t := \sum_{i=1}^t \gamma_i (x_i - x_{i-1})$$

is a martingale with respect to  $\{\mathcal{F}_t\}$ .

**Ex. 7.6.13** Let  $F(s_1, \dots, s_N) = C(F_1(s_1), \dots, F_N(s_N))$ , where  $C$  is a copula and each  $F_n$  is a continuous CDF on  $\mathbb{R}$ . Let  $\mathcal{L}(u_1, \dots, u_N) = C$  and let

$$(x_1, \dots, x_N) := (F_1^{-1}(u_1), \dots, F_N^{-1}(u_N))$$

Prove that  $\mathcal{L}(x_1, \dots, x_N) = F$ .

**Ex. 7.6.14** Show that if  $a \geq 0$  and  $b$  and  $c$  are any real numbers, then  $a \min\{b, c\} = \min\{ab, ac\}$ . Next let  $\alpha$  be as defined in (7.38) and let  $p(s, s') := q(s, s')\alpha(s, s')$ . Show that  $\pi(s)p(s, s') = p(s', s)\pi(s')$  for all  $s, s'$ .

## 7.6.1 Solutions to Selected Exercises

**Solution to Ex. 7.6.1.** Here's a proof for the case  $b = 0$  and  $c = 1$ . For the induction step, suppose that  $\mathcal{L}(x_t) = P_\infty$ , where  $P_\infty$  is as defined in (7.18) on page 190. With

$b = 0$  and  $c = 1$ , this becomes  $P_\infty = N(0, 1/(1 - a^2))$ . Since  $x_{t+1} = ax_t + w_{t+1}$  and linear combinations of independent normals are normal, we know that  $x_{t+1}$  is normal. Thus it remains only to show that  $\mathbb{E}x_{t+1} = 0$  and  $\text{var } x_{t+1} = 1/(1 - a^2)$ . The first claim is true because, by linearity of expectations,  $\mathbb{E}x_{t+1} = \mathbb{E}[ax_t + w_{t+1}] = a\mathbb{E}[x_t] + \mathbb{E}[w] = 0$ . The second claim also holds because

$$\text{var } x_{t+1} = \text{var}[ax_t + w_{t+1}] = a^2 \text{var}[x_t] + \text{var}[w_{t+1}] + 2a \text{cov}[x_t, w_{t+1}]$$

Since  $x_t$  and  $w_{t+1}$  are independent the final term is zero, and we get

$$\text{var } x_{t+1} = a^2 \frac{1}{1 - a^2} + 1 = \frac{1}{1 - a^2} \quad \square$$

**Solution to Ex. 7.6.2.** We aim to show that  $\text{cov}[x_t, x_{t+j}] = a^j/(1 - a^2)$ . Fix  $j \geq 1$  and iterate with  $x_{t+1} = ax_t + w_{t+1}$  to get  $x_{t+j} = a^j x_t + \sum_{k=1}^j a^{j-k} w_{t+k}$ . Using this expression and independence of  $x_t$  and  $w_{t+k}$  gives

$$\mathbb{E}[x_{t+j}x_t] = \mathbb{E}\left[a^j x_t^2 + \sum_{k=1}^j a^{j-k} w_{t+k} x_t\right] = a^j \mathbb{E}[x_t^2] + \sum_{k=1}^j a^{j-k} \mathbb{E}[w_{t+k}] \mathbb{E}[x_t]$$

Since  $\mathbb{E}[w_{t+k}] = 0$ , this gives  $\text{cov}[x_{t+j}, x_t] = \mathbb{E}[x_{t+j}x_t] = a^j \mathbb{E}[x_t^2]$ . The process  $\{x_t\}$  is identically distributed (fact 7.2.5) with common variance  $1/(1 - a^2)$ . The result follows.  $\square$

**Solution to Ex. 7.6.3.** As shown in exercise 7.6.2, we have  $\mathbb{E}[x_t x_{t+j}] = a^j/(1 - a^2)$ . Hence

$$\frac{2}{T^2} \sum_{j=0}^{T-1} \sum_{t=1}^{T-j} \frac{a^j}{1 - a^2} = \frac{2}{T^2} \sum_{j=0}^{T-1} (T - j) \frac{a^j}{1 - a^2} \leq \frac{2}{T} \frac{1}{(1 - a)(1 - a^2)}$$

The term on the right converges to 0 as  $T \rightarrow \infty$ .  $\square$

**Solution to Ex. 7.6.4.** Fix  $B \in \mathcal{B}(S)$  and let  $P_t$  be the  $t$ th marginal. By the law of iterated expectations, we have

$$\mathbb{P}\{\mathbf{x}_{t+1} \in B\} = \mathbb{E}[\mathbb{1}\{\mathbf{x}_{t+1} \in B\}] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{\mathbf{x}_{t+1} \in B\} \mid \mathbf{x}_t]] = \mathbb{E}[Q(\mathbf{x}_t, B)]$$

In other words,  $P_{t+1}(B) = \int Q(\mathbf{s}, B) P_t(d\mathbf{s})$ , as was to be shown.  $\square$

**Solution to Ex. 7.6.7.** To prove that  $\{d_t\}$  is adapted to  $\{\mathcal{F}_t\}$ , we need to show that  $d_t$  is a function of variables in  $\mathcal{F}_t$ . To see this, observe that  $m_t$  is a function of variables in  $\mathcal{F}_t$  and  $m_{t-1}$  is a function of variables in  $\mathcal{F}_{t-1}$ . By the definition of filtrations, every

variable in  $\mathcal{F}_{t-1}$  is also in  $\mathcal{F}_t$ . Hence  $m_{t-1}$  is also a function of variables in  $\mathcal{F}_t$ . Since both  $m_t$  and  $m_{t-1}$  are functions of variables in  $\mathcal{F}_t$ , so is  $d_t = m_t - m_{t-1}$ .

Next we show that  $\mathbb{E}[d_{t+1} | \mathcal{F}_t] = 0$ . Since  $\mathbb{E}[m_{t+1} | \mathcal{F}_t] = m_t$  and  $\mathbb{E}[m_t | \mathcal{F}_t] = m_t$  (by fact 7.3.1), we have

$$\mathbb{E}[d_{t+1} | \mathcal{F}_t] = \mathbb{E}[m_{t+1} | \mathcal{F}_t] - \mathbb{E}[m_t | \mathcal{F}_t] = m_t - m_t = 0 \quad \square$$

**Solution to Ex. 7.6.8.** First,  $\{\alpha_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ . To see that  $\{\alpha_t\}$  is adapted to  $\{\mathcal{F}_t\}$ , note that  $\{m_t\}$  is adapted to  $\{\mathcal{F}_t\}$  by assumption. Therefore  $m_t$  is a function of variables in  $\mathcal{F}_t$  and hence  $\alpha_t = m_t + 1$  is a function of variables in  $\mathcal{F}_t$ . Moreover, using the fact that  $\{m_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ , we have

$$\mathbb{E}[\alpha_{t+1} | \mathcal{F}_t] = \mathbb{E}[m_{t+1} + 1 | \mathcal{F}_t] = \mathbb{E}[m_{t+1} | \mathcal{F}_t] + 1 = m_t + 1 = \alpha_t$$

Second,  $\{\kappa_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ . The proof is similar and hence omitted.

Third,  $\{\gamma_t\}$  is not generally a martingale with respect to  $\{\mathcal{F}_t\}$ . For example, let  $\{m_t\}$  be the random walk martingale in example 7.3.4. The process  $\{\gamma_t\}$  given by  $\gamma_t = m_t^2$  is not a martingale whenever  $\sigma^2 := \mathbb{E}[\eta_1^2]$  is strictly positive. To see this, observe that

$$\gamma_{t+1} = m_{t+1}^2 = \left( \sum_{j=1}^t \eta_j + \eta_{t+1} \right)^2 = (m_t + \eta_{t+1})^2 = m_t^2 + 2m_t\eta_{t+1} + \eta_{t+1}^2$$

$$\therefore \mathbb{E}[\gamma_{t+1} | \mathcal{F}_t] = \mathbb{E}[m_t^2 | \mathcal{F}_t] + 2\mathbb{E}[m_t\eta_{t+1} | \mathcal{F}_t] + \mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t]$$

Since

$$\mathbb{E}[m_t\eta_{t+1} | \mathcal{F}_t] = m_t\mathbb{E}[\eta_{t+1} | \mathcal{F}_t] = m_t\mathbb{E}[\eta_{t+1}] = 0$$

and  $\mathbb{E}[\eta_{t+1}^2 | \mathcal{F}_t] = \mathbb{E}[\eta_{t+1}^2] = \sigma^2$  we now have

$$\mathbb{E}[\gamma_{t+1} | \mathcal{F}_t] = \mathbb{E}[m_t^2 | \mathcal{F}_t] + \sigma^2 = m_t^2 + \sigma^2 = \gamma_t + \sigma^2 > \gamma_t$$

Hence  $\{\gamma_t\}$  is not a martingale with respect to  $\{\mathcal{F}_t\}$ .  $\square$

**Solution to Ex. 7.6.11.** It is clear from (7.8) on page 185 that  $\{x_t\}$  is adapted to the filtration for all values of  $a$ . Regarding part (i), if  $a = 1$ , then  $\{x_t\}$  is the random walk in example 7.3.4. Hence  $\{x_t\}$  is a martingale with respect to  $\{\mathcal{F}_t\}$ . Regarding part (ii), if  $a = 0$ , then  $x_t = w_t$ , and

$$\mathbb{E}[x_{t+1} | \mathcal{F}_t] = \mathbb{E}[w_{t+1} | \mathcal{F}_t] = \mathbb{E}[w_{t+1}] = 0$$

Hence  $\{x_t\}$  is a martingale difference sequence with respect to  $\{\mathcal{F}_t\}$ .  $\square$

**Solution to Ex. 7.6.13.** We aim to show that  $\mathcal{L}(x_1, \dots, x_N) = F$ . We have

$$\mathbb{P} \cap_n \{x_n \leq s_n\} = \mathbb{P} \cap_n \{F_n^{-1}(u_n) \leq s_n\} = \mathbb{P} \cap_n \{u_n \leq F_n(s_n)\}$$

Since  $C$  is the CDF of  $u_1, \dots, u_N$ , the right-hand side is equal to  $C(F_1(s_1), \dots, F_N(s_N))$ . This is  $F(s_1, \dots, s_N)$  by the definition of  $F$ .  $\square$

**Solution to Ex. 7.6.14.** The first claim is immediate, since  $a \geq 0$  implies that  $f(x) = ax$  is an order preserving transform. The second claim is just algebra. You've made it this far through the book, so I guess you don't need my help with that.  $\square$