Chapter 9

Markov Processes

This chapter begins our study of Markov processes.

Section 9.1 is mainly "ideological": it formally defines the Markov property for one-parameter processes, and explains why it is a natural generalization of both complete determinism and complete statistical independence.

Section 9.2 introduces the description of Markov processes in terms of their transition probabilities and proves the existence of such processes.

Section 9.3 deals with the question of when being Markovian relative to one filtration implies being Markov relative to another.

9.1 The Correct Line on the Markov Property

The Markov property is the independence of the future from the past, given the present. Let us be more formal.

Definition 102 (Markov Property) *A one-parameter process X is a* Markov process with respect to a filtration $\{\mathcal{F}\}_t$ *when X_t is adapted to the filtration, and, for any $s > t$, X_s is independent of $\mathcal{F}_t$ given X_t, $X_s \perp\!\!\!\perp \mathcal{F}_t | X_t$. If no filtration is mentioned, it may be assumed to be the natural one generated by X. If X is also conditionally stationary, then it is a* time-homogeneous *(or just* homogeneous*) Markov process.*

Lemma 103 (The Markov Property Extends to the Whole Future) *Let X_t^+ stand for the collection of X_u, $u > t$. If X is Markov, then $X_t^+ \perp\!\!\!\perp \mathcal{F}_t | X_t$.*

Proof: See Exercise 9.1. □

There are two routes to the Markov property. One is the path followed by Markov himself, of desiring to weaken the assumption of strict statistical independence between variables to mere conditional independence. In fact, Markov

specifically wanted to show that independence was *not* a necessary condition for the law of large numbers to hold, because his arch-enemy claimed that it was, and used that as grounds for believing in free will and Christianity.[1] It turns out that all the key limit theorems of probability — the weak and strong laws of large numbers, the central limit theorem, etc. — work perfectly well for Markov processes, as well as for IID variables.

The other route to the Markov property begins with completely deterministic systems in physics and dynamics. The *state* of a deterministic dynamical system is some variable which fixes the value of all present and future observables. As a consequence, the present state determines the state at all future times. However, strictly deterministic systems are rather thin on the ground, so a natural generalization is to say that the present state determines the *distribution* of future states. This is precisely the Markov property.

Remarkably enough, it is possible to represent any one-parameter stochastic process X as a noisy function of a Markov process Z. The shift operators give a trivial way of doing this, where the Z process is not just homogeneous but actually fully deterministic. An equally trivial, but slightly more probabilistic, approach is to set $Z_t = X_t^-$, the complete past up to and including time t. (This is not necessarily homogeneous.) It turns out that, subject to mild topological conditions on the space X lives in, there is a unique *non-trivial* representation where $Z_t = \epsilon(X_t^-)$ for some function ϵ, Z_t is a homogeneous Markov process, and $X_u \perp\!\!\!\perp \sigma(\{X_t, t \leq u\}) | Z_t$. (See Knight (1975, 1992); Shalizi and Crutchfield (2001).) We may explore such *predictive Markovian representations* at the end of the course, if time permits.

9.2 Transition Probability Kernels

The most obvious way to specify a Markov process is to say what its transition probabilities are. That is, we want to know $\mathbb{P}(X_s \in B | X_t = x)$ for every $s > t$, $x \in \Xi$, and $B \in \mathcal{X}$. Probability kernels (Definition 30) were invented to let us do just this. We have already seen how to compose such kernels; we also need to know how to take their product.

Definition 104 (Product of Probability Kernels) *Let μ and ν be two probability kernels from Ξ to Ξ. Then their product $\mu\nu$ is a kernel from Ξ to Ξ, defined by*

$$(\mu\nu)(x, B) \equiv \int \mu(x, dy)\nu(y, B) \tag{9.1}$$

$$= (\mu \otimes \nu)(x, \Xi \times B) \tag{9.2}$$

[1] I am not making this up. See Basharin *et al.* (2004) for a nice discussion of the origin of Markov chains and of Markov's original, highly elegant, work on them. There is a translation of Markov's original paper in an appendix to Howard (1971), and I dare say other places as well.

Intuitively, all the product does is say that the probability of starting at the point x and landing in the set B is equal the probability of first going to y and then ending in B, integrated over all intermediate points y. (Strictly speaking, there is an abuse of notation in Eq. 9.2, since the second kernel in a composition $\otimes$ should be defined over a product space, here $\Xi \times \Xi$. So suppose we have such a kernel ν', only $\nu'((x,y),B) = \nu(y,B)$.) Finally, observe that if $\mu(x,\cdot) = \delta_x$, the delta function at x, then $(\mu\nu)(x,B) = \nu(x,B)$, and similarly that $(\nu\mu)(x,B) = \nu(x,B)$.

Definition 105 (Transition Semi-Group) *For every $(t,s) \in T \times T$, $s \geq t$, let $\mu_{t,s}$ be a probability kernel from Ξ to Ξ. These probability kernels form a* transition semi-group *when*

1. *For all t, $\mu_{t,t}(x,\cdot) = \delta_x$.*
2. *For any $t \leq s \leq u \in T$, $\mu_{t,u} = \mu_{t,s}\mu_{s,u}$.*

A transition semi-group for which $\forall t \leq s \in T$, $\mu_{t,s} = \mu_{0,s-t} \equiv \mu_{s-t}$ is homogeneous.

As with the shift semi-group, this is really a monoid (because $\mu_{t,t}$ acts as the identity).

The major theorem is the existence of Markov processes with specified transition kernels.

Theorem 106 (Existence of Markov Process with Given Transition Kernels) *Let $\mu_{t,s}$ be a transition semi-group and ν_t a collection of distributions on a Borel space Ξ. If*

$$\nu_s = \nu_t \mu_{t,s} \tag{9.3}$$

then there exists a Markov process X such that $\forall t$,

$$\mathcal{L}(X_t) = \nu_t \tag{9.4}$$

and $\forall t_1 \leq t_2 \leq \ldots \leq t_n$,

$$\mathcal{L}(X_{t_1}, X_{t_2} \ldots X_{t_n}) = \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \ldots \otimes \mu_{t_{n-1},t_n} \tag{9.5}$$

Conversely, if X is a Markov process with values in Ξ, then there exist distributions ν_t and a transition kernel semi-group $\mu_{t,s}$ such that Equations 9.4 and 9.3 hold, and

$$\mathbb{P}(X_s \in B | \mathcal{F}_t) = \mu_{t,s} \ a.s. \tag{9.6}$$

Proof: (*From transition kernels to a Markov process.*) For any finite set of times $J = \{t_1, \ldots t_n\}$ (in ascending order), define a distribution on Ξ_J as

$$\nu_J \equiv \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \ldots \otimes \mu_{t_{n-1},t_n} \tag{9.7}$$

It is easily checked, using point (2) in the definition of a transition kernel semi-group (Definition 105), that the ν_J form a projective family of distributions. Thus, by the Kolmogorov Extension Theorem (Theorem 29), there exists a stochastic process whose finite-dimensional distributions are the ν_J. Now pick a J of size n, and two sets, $B \in \mathcal{X}^{n-1}$ and $C \in \mathcal{X}$.

$$\begin{aligned} \mathbb{P}(X_J \in B \times C) &= \nu_J(B \times C) && (9.8)\\ &= \mathbf{E}\left[\mathbf{1}_{B\times C}(X_J)\right] && (9.9)\\ &= \mathbf{E}\left[\mathbf{1}_B(X_{J\setminus t_n})\mu_{t_{n-1},t_n}(X_{t_{n-1}}, C)\right] && (9.10)\end{aligned}$$

Set $\{\mathcal{F}\}_t$ to be the natural filtration, $\sigma(\{X_u, u \leq s\})$. If $A \in \mathcal{F}_s$ for some $s \leq t$, then by the usual generating class arguments we have

$$\begin{aligned} \mathbb{P}\left(X_t \in C, X_s^- \in A\right) &= \mathbf{E}\left[\mathbf{1}_A \mu_{s,t}(X_s, C)\right] && (9.11)\\ \mathbb{P}(X_t \in C|\mathcal{F}_s) &= \mu_{s,t}(X_s, C) && (9.12)\end{aligned}$$

i.e., $X_t \perp\!\!\!\perp \mathcal{F}_s | X_s$, as was to be shown.

(*From the Markov property to the transition kernels.*) From the Markov property, for any measurable set $C \in \mathcal{X}$, $\mathbb{P}(X_t \in C|\mathcal{F}_s)$ is a function of X_s alone. So define the kernel $\mu_{s,t}$ by $\mu_{s,t}(x, C) = \mathbb{P}(X_t \in C|X_s = x)$, with a possible measure-0 exceptional set from (ultimately) the Radon-Nikodym theorem. (The fact that Ξ is Borel guarantees the existence of a regular version of this conditional probability.) We get the semi-group property for these kernels thus: pick any three times $t \leq s \leq u$, and a measurable set $C \subseteq \Xi$. Then

$$\begin{aligned} \mu_{t,u}(X_t, C) &= \mathbb{P}(X_u \in C|\mathcal{F}_t) && (9.13)\\ &= \mathbb{P}(X_u \in C, X_s \in \Xi|\mathcal{F}_t) && (9.14)\\ &= (\mu_{t,s} \otimes \mu_{s,u})(X_t, \Xi \times C) && (9.15)\\ &= (\mu_{t,s}\mu_{s,u})(X_t, C) && (9.16)\end{aligned}$$

The argument to get Eq. 9.3 is similar. □

Note: For one-sided discrete-parameter processes, we could use the Ionescu-Tulcea Extension Theorem 33 to go from a transition kernel semi-group to a Markov process, even if Ξ is not a Borel space.

Definition 107 (Invariant Distribution) *Let X be a homogeneous Markov process with transition kernels μ_t. A distribution ν on Ξ is* invariant *when, $\forall t$, $\nu = \nu\mu_t$, i.e.,*

$$\begin{aligned} (\nu\mu_t)(B) &\equiv \int \nu(dx)\mu_t(x, B) && (9.17)\\ &= \nu(B) && (9.18)\end{aligned}$$

ν is also called an equilibrium distribution.

The term "equilibrium" comes from statistical physics, where however its meaning is a bit more strict, in that "detailed balance" must also be satisified: for any two sets $A, B \in \mathcal{X}$,

$$\int \nu(dx)\mathbf{1}_A \mu_t(x, B) = \int \nu(dx)\mathbf{1}_B \mu_t(x, A) \tag{9.19}$$

i.e., the flow of probability from A to B must equal the flow in the opposite direction. Much confusion has resulted from neglecting the distinction between equilibrium in the strict sense of detailed balance and equilibrium in the weaker sense of invariance.

Theorem 108 (Stationarity and Invariance for Homogeneous Markov Processes) *Suppose X is homogeneous, and $\mathcal{L}(X_t) = \nu$, where ν is an invariant distribution. Then the process X_t^+ is stationary.*

Proof: Exercise 9.4. □

9.3 The Markov Property Under Multiple Filtrations

Definition 102 specifies what it is for a process to be Markovian relative to a given filtration $\{\mathcal{F}\}_t$. The question arises of when knowing that X Markov with respect to one filtration $\{\mathcal{F}\}_t$ will allow us to deduce that it is Markov with respect to another, say $\{\mathcal{G}\}_t$.

To begin with, let's introduce a little notation.

Definition 109 (Natural Filtration) *The* natural filtration *for a stochastic process X is $\left\{\mathcal{F}^X\right\}_t \equiv \sigma(\{X_u,\ u \leq t\})$. Every process X is adapted to its natural filtration.*

Definition 110 (Comparison of Filtrations) *A filtration $\{\mathcal{G}\}_t$ is* finer than *or* more refined than *or* a refinement of *$\{\mathcal{F}\}_t$, $\{\mathcal{F}\}_t \prec \{\mathcal{G}\}_t$, if, for all t, $\mathcal{F}_t \subseteq \mathcal{G}_t$, and at least sometimes the inequality is strict. $\{\mathcal{F}\}_t$ is* coarser *or* less fine *than $\{\mathcal{G}\}_t$. If $\{\mathcal{F}\}_t \prec \{\mathcal{G}\}_t$ or $\{\mathcal{F}\}_t = \{\mathcal{G}\}_t$, we write $\{\mathcal{F}\}_t \preceq \{\mathcal{G}\}_t$.*

Lemma 111 (The Natural Filtration Is the Coarsest One to Which a Process Is Adapted) *If X is adapted to $\{\mathcal{G}\}_t$, then $\left\{\mathcal{F}^X\right\}_t \preceq \{\mathcal{G}\}_t$.*

Proof: For each t, X_t is $\mathcal{G}_t$ measurable. But $\mathcal{F}_t^X$ is, by construction, the smallest σ-algebra with respect to which X_t is measurable, so, for every t, $\mathcal{F}_t^X \subseteq \mathcal{G}_t$, and the result follows. □

Theorem 112 (Markovianity Is Preserved Under Coarsening) *If X is Markovian with respect to $\{\mathcal{G}\}_t$, then it is Markovian with respect to any coarser filtration to which it is adapted, and in particular with respect to its natural filtration.*

PROOF: Use the smoothing property of conditional expectations: For any two σ-fields $\mathcal{H} \subset \mathcal{K}$ and random variable Y, $\mathbf{E}[Y|\mathcal{H}] = \mathbf{E}[\mathbf{E}[Y|\mathcal{K}]|\mathcal{H}]$ a.s. So, if $\{\mathcal{F}\}_t$ is coarser than $\{\mathcal{G}\}_t$, and X is Markovian with respect to the latter, for any function $f \in L_1$ and time $s > t$,

$$\begin{aligned} \mathbf{E}[f(X_s)|\mathcal{F}_t] &= \mathbf{E}[\mathbf{E}[f(X_s)|\mathcal{G}_t]|\mathcal{F}_t] \ a.s. & (9.20)\\ &= \mathbf{E}[\mathbf{E}[f(X_s)|X_t]|\mathcal{F}_t] & (9.21)\\ &= \mathbf{E}[f(X_s)|X_t] & (9.22) \end{aligned}$$

The next-to-last line uses the fact that $X_s \perp\!\!\!\perp \mathcal{G}_t | X_t$, because X is Markovian with respect to $\{\mathcal{G}\}_t$, and this in turn implies that conditioning X_s, or any function thereof, on $\mathcal{G}_t$ is equivalent to conditioning on X_t alone. (Recall that X_t is $\mathcal{G}_t$-measurable.) The last line uses the facts that (i) $\mathbf{E}[f(X_s)|X_t]$ is a function X_t, (ii) X is adapted to $\{\mathcal{F}\}_t$, so X_t is $\mathcal{F}_t$-measurable, and (iii) if Y is $\mathcal{F}$-measurable, then $\mathbf{E}[Y|\mathcal{F}] = Y$. Since this holds for all $f \in L_1$, it holds in particular for $\mathbf{1}_A$, where A is any measurable set, and this established the conditional independence which constitutes the Markov property. Since (Lemma 111) the natural filtration is the coarsest filtration to which X is adapted, the remainder of the theorem follows. □

The converse is false, as the following example shows.

Example 113 (The Logistic Map Shows That Markovianity Is Not Preserved Under Refinement) *We revert to the symbolic dynamics of the logistic map, Examples 39 and 40. Let S_1 be distributed on the unit interval with density $1/\pi\sqrt{s(1-s)}$, and let $S_n = 4S_{n-1}(1-S_{n-1})$. Finally, let $X_n = \mathbf{1}_{[0.5,1.0]}(S_n)$. It can be shown that the X_n are a Markov process with respect to their natural filtration; in fact, with respect to that filtration, they are independent and identically distributed Bernoulli variables with probability of success $1/2$. However, $\mathbb{P}\left(X_{n+1}|\mathcal{F}_n^S, X_n\right) \neq \mathbb{P}(X_{n+1}|X_n)$, since X_{n+1} is a* deterministic *function of S_n. But, clearly, $\mathcal{F}_n^X \subset \mathcal{F}_n^S$ for each n, so $\left\{\mathcal{F}^X\right\}_t \prec \left\{\mathcal{F}^S\right\}_t$.*

The issue can be illustrated with graphical models (Spirtes *et al.*, 2001; Pearl, 1988). A discrete-time Markov process looks like Figure 9.1*a*. X_n blocks all the pasts from the past to the future (in the diagram, from left to right), so it produces the desired conditional independence. Now let's add another variable which actually drives the X_n (Figure 9.1*b*). If we can't measure the S_n variables, just the X_n ones, then it can still be the case that we've got the conditional independence among what we can see. But if we can see X_n as well as S_n — which is what refining the filtration amounts to — then simply conditioning on X_n does not block all the paths from the past of X to its future, and, generally speaking, we will lose the Markov property. Note that knowing S_n *does* block all paths from past to future — so this remains a *hidden* Markov model. Markovian representation theory is about finding conditions under which we can get things to look like Figure 9.1*b*, even if we can't get them to look like Figure 9.1*a*.

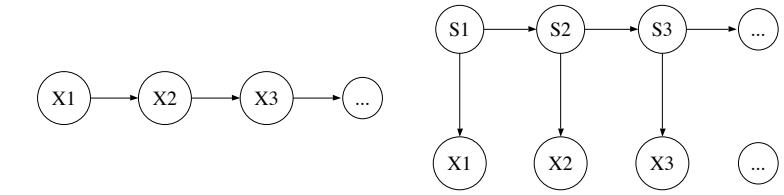

Figure 9.1: (a) Graphical model for a Markov chain. (b) Refining the filtration, say by conditioning on an additional random variable, can lead to a failure of the Markov property.

9.4 Exercises

Exercise 9.1 (Extension of the Markov Property to the Whole Future) *Prove Lemma 103.*

Exercise 9.2 (Futures of Markov Processes Are One-Sided Markov Processes) *Show that if X is a Markov process, then, for any $t \in T$, X_t^+ is a one-sided Markov process.*

Exercise 9.3 (Discrete-Time Sampling of Continuous-Time Markov Processes) *Let X be a continuous-parameter Markov process, and t_n a countable set of strictly increasing indices. Set $Y_n = X_{t_n}$. Is Y_n a Markov process? If X is homogeneous, is Y also homogeneous? Does either answer change if $t_n = nt$ for some constant interval $t > 0$?*

Exercise 9.4 (Stationarity and Invariance for Homogeneous Markov Processes) *Prove Theorem 108.*

Exercise 9.5 (Rudiments of Likelihood-Based Inference for Markov Chains) *(This exercise presumes some knowledge of sufficient statistics and exponential families from theoretical statistics.)*

Assume $T = \mathbb{N}$, Ξ is a finite set, and X a homogeneous Markov Ξ-valued Markov chain. Further assume that X_1 is constant, $= x_1$, with probability 1. (This last is not an essential restriction.) Let $p_{ij} = \mathbb{P}(X_{t+1} = j | X_t = i)$.

1. *Show that p_{ij} fixes the transition kernel, and vice versa.*
2. *Write the probability of a sequence $X_1 = x_1, X_2 = x_2, \ldots X_n = x_n$, for short $X_1^n = x_1^n$, as a function of p_{ij}.*
3. *Write the log-likelihood ℓ as a function of p_{ij} and x_1^n.*

4. *Define n_{ij} to be the number of times t such that $x_t = i$ and $x_{t+1} = j$. Similarly define n_i as the number of times t such that $x_t = i$. Write the log-likelihood as a function of p_{ij} and n_{ij}.*

5. *Show that the n_{ij} are sufficient statistics for the parameters p_{ij}.*

6. *Show that the distribution has the form of a canonical exponential family, with sufficient statistics n_{ij}, by finding the natural parameters. (*Hint*: the natural parameters are transformations of the p_{ij}.) Is the family of full rank?*

7. *Find the maximum likelihood estimators, for either the p_{ij} parameters or for the natural parameters. (*Hint*: Use Lagrange multipliers to enforce the constraint $\sum_j n_{ij} = n_i$.)*

The classic book by Billingsley (1961) remains an excellent source on statistical inference for Markov chains and related processes.

Exercise 9.6 (Implementing the MLE for a Simple Markov Chain) *(This exercise continues the previous one.)*
Set $\Xi = \{0, 1\}$, $x_1 = 0$, and

$$\mathbf{p}_0 = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix}$$

1. *Write a program to simulate sample paths from this Markov chain.*

2. *Write a program to calculate the maximum likelihood estimate $\hat{\mathbf{p}}$ of the transition matrix from a sample path.*

3. *Calculate $2(\ell(\hat{\mathbf{p}}) - \ell(\mathbf{p}_0))$ for many independent sample paths of length n. What happens to the distribution as $n \to \infty$? (*Hint*: see Billingsley (1961), Theorem 2.2.)*

Exercise 9.7 (The Markov Property and Conditional Independence from the Immediate Past) *Let X be a Ξ-valued discrete-parameter random process. Suppose that, for all t, $X_{t-1} \perp\!\!\!\perp X_{t+1} | X_t$. Either prove that X is a Markov process, or provide a counter-example. You may assume that Ξ is a Borel space if you find that helpful.*

Exercise 9.8 (Higher-Order Markov Processes) *A discrete-time process X is a k^{th} order Markov process with respect to a filtration $\{\mathcal{F}\}_t$ when*

$$X_{t+1} \perp\!\!\!\perp \mathcal{F}_t | \sigma(X_t, X_{t-1}, \ldots X_{t-k+1}) \tag{9.23}$$

*for some finite integer k. For any Ξ-valued discrete-time process X, define the k-*block process *$Y^{(k)}$ as the Ξ^k-valued process where $Y_t^{(k)} = (X_t, X_{t+1}, \ldots X_{t+k-1})$.*

1. *Prove that if X is Markovian to order k, it is Markovian to any order $l > k$. (For this reason, saying that X is a k^{th} order Markov process conventionally means that k is the* smallest *order at which Eq. 9.23 holds.)*

2. *Prove that X is k^{th}-order Markovian if and only if $Y^{(k)}$ is Markovian.*

The second result shows that studying on the theory of first-order Markov processes involves no essential loss of generality. For a test of the hypothesis that X is Markovian of order k against the alternative that is Markovian of order $l > k$, see Billingsley (1961). For recent work on estimating the order of a Markov process, assuming it is Markovian to some finite order, see the elegant paper by Peres and Shields (2005).

Exercise 9.9 (AR(1) Models) *A first-order autoregressive model, or AR(1) model, is a real-valued discrete-time process defined by an evolution equation of the form*

$$X(t) = a_0 + a_1 X(t-1) + \epsilon(t)$$

where the innovations *$\epsilon(t)$ are independent and identically distributed, and independent of $X(0)$. A p^{th}-order autoregressive model, or AR(p) model, is correspondingly defined by*

$$X(t) = a_0 + \sum_{i=1}^{p} a_i X(t-i) + \epsilon(t)$$

Finally an AR(p) model in state-space form *is an $\mathbb{R}^p$-valued process defined by*

$$\vec{Y}(t) = a_0 \vec{e}_1 + \begin{bmatrix} a_1 & a_2 & \dots & a_{p-1} & a_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \vec{Y}(t-1) + \epsilon(t)\vec{e}_1$$

where $\vec{e}_1$ is the unit vector along the first coordinate axis.

1. *Prove that AR(1) models are Markov for all choices of a_0 and a_1, and all distributions of $X(0)$ and ϵ.*

2. *Give an explicit form for the transition kernel of an AR(1) in terms of the distribution of ϵ.*

3. *Are AR(p) models Markovian when $p > 1$? Prove or give a counterexample.*

4. *Prove that $\vec{Y}$ is a Markov process, without using Exercise 9.8. (*Hint*: What is the relationship between $Y_i(t)$ and $X(t-i+1)$?)*