Chapter 10

Alternative Characterizations of Markov Processes

This lecture introduces two ways of characterizing Markov processes other than through their transition probabilities.

Section 10.1 describes discrete-parameter Markov processes as transformations of sequences of IID uniform variables.

Section 10.2 describes Markov processes in terms of measure-preserving transformations (Markov operators), and shows this is equivalent to the transition-probability view.

10.1 Markov Sequences as Transduced Noise

A key theorem says that discrete-time Markov processes can be viewed as the result of applying a certain kind of filter to pure noise.

Theorem 114 *Let X be a one-sided discrete-parameter process taking values in a Borel space Ξ. X is Markov iff there are measurable functions $f_n : \Xi \times [0,1] \mapsto \Xi$ such that, for IID random variables $Z_n \sim U(0,1)$, all independent of X_1, $X_{n+1} = f_n(X_n, Z_n)$ almost surely. X is homogeneous iff $f_n = f$ for all n.*

Proof: Kallenberg, Proposition 8.6, p. 145. Notice that, in order to get the "only if" direction to work, Kallenberg invokes what we have as Proposition 26, which is where the assumptions that Ξ is a Borel space comes in. You should verify that the "if" direction does not require this assumption. □

Let us stick to the homogeneous case, and consider the function f in somewhat more detail.

In engineering or computer science, a *transducer* is an apparatus — really, a function — which takes a stream of inputs of one kind and produces a stream of outputs of another kind.

Definition 115 (Transducer) *A (deterministic) transducer is a sextuple* $\langle \Sigma, \Upsilon, \Xi, f, h, s_0 \rangle$ *where* Σ, Υ *and* Ξ *are, respectively, the state, input and output spaces,* $f : \Sigma \times \Xi \mapsto \Sigma$ *is the* state update function *or* state transition function, $h : \Sigma \times \Upsilon \mapsto \Xi$ *is the* measurement *or* observation *function, and* $s_0 \in \Sigma$ *is the* starting state. *(We shall assume both* f *and* h *are always measurable.) If* h *does not depend on its state argument, the transducer is* memoryless. *If* f *does not depend on its state argument, the transducer is* without after-effect.

It should be clear that if a memoryless transducer is presented with IID inputs, its output will be IID as well. What Theorem 114 says is that, if we have a transducer with memory (so that h depends on the state) but is without after-effect (so that f does not depend on the state), IID inputs will produce Markovian outputs, and conversely any reasonable Markov process can be represented in this way. Notice that if a transducer is without memory, we can replace it with an equivalent with a single state, and if it is without after-effect, we can identify Σ and Ξ.

Notice also that the two functions f and h determine a transition function where we use the input to update the state: $g : \Sigma \times \Upsilon \mapsto \Sigma$, where $g(s, y) = f(s, h(s, y))$. Thus, if the inputs are IID and uniformly distributed, then (Theorem 114) the successive states of the transducer are always Markovian. The question of which processes can be produced by noise-driven transducers is this intimately bound up with the question of Markovian representations. While, as mentioned, quite general stochastic processes *can* be put in this form (Knight, 1975, 1992), it is not necessarily possible to do this with a finite internal state space Σ, even when Ξ is finite. The distinction between finite and infinite Σ is crucial to theoretical computer science, and we might come back to it later, but

Two issues suggest themselves in connection with this material. One is whether, given a *two*-sided process, we can pull the same trick, and represent a Markovian X as a transformation of an IID sequence extending into the infinite past. (Remember that the theorem is for one-sided processes, and starts with an initial X_1.) This is more subtle than it seems at first glance, or even than it seemed to Norbert Wiener when he first posed the question (Wiener, 1958); for a detailed discussion, see Rosenblatt (1971), and, for recent set of applications, Wu (2005). The other question is whether the same trick can be pulled in continuous time; here much less is known.

10.2 Time-Evolution (Markov) Operators

Let's look again at the evolution of the one-dimensional distributions for a Markov process:

$$\nu_s = \nu_t \mu_{t,s} \tag{10.1}$$

$$\nu_s(B) = \int \nu_t(dx) \mu_{t,s}(x, B) \tag{10.2}$$

The transition kernels define linear operators taking probability measures on Ξ to probability measures on Ξ. This can be abstracted.

Definition 116 (Markov Operator on Measures) *Take any measurable space $\Xi, \mathcal{X}$. A* Markov operator on measures *is an operator M which takes finite measures on this space to other finite measures on this space such that*

1. *M is linear, i.e., for any $a_1, a_2 \in [0,1]$ and any two measures μ_1, μ_2,*

$$M(a_1\mu_1 + a_2\mu_2) = a_1 M\mu_1 + a_2 M\mu_2 \tag{10.3}$$

2. *M is norm-preserving, i.e., $M\mu(\Xi) = \mu(\Xi)$.*

(In particular P must take probability measures to probability measures.)

Definition 117 (Markov Operator on Densities) *Take any probability space $\Xi, \mathcal{X}, \mu$, and let L_1 be as usual the class of all μ-integrable generalized functions on Ξ. A linear operator $P : L_1 \mapsto L_1$ is a* Markov operator on densities *when:*

1. *If $f \geq 0$ (a.e. μ), $Pf \geq 0$ (a.e. μ).*
2. *If $f \geq 0$ (a.e. μ), $\|Pf\| = \|f\|$.*

By "a Markov operator" I will often mean a Markov operator on densities, with the reference measure μ being some suitable uniform distribution on Ξ. However, the general theory applies to operators on measures.

Lemma 118 (Markov Operators on Measures Induce Those on Densities) *Let M be a Markov operator on measures. If M takes measures absolutely continuous with respect to μ to measures absolutely continuous with respect to μ, i.e., it preserves domination by μ, then it induces an almost-unique Markov operator P on densities with respect to μ.*

Proof: Let f be a function which is in $L_1(\mu)$ and is non-negative (μ-a.e.). If $f \geq 0$ μ a.e., the set function $\nu_f(A) = \int_A f(x)d\mu$ is a finite measure which is absolutely continuous with respect to μ. (Why is it finite?) By hypothesis, then, $M\nu_f$ is another finite measure which is absolutely continuous with respect to μ, and $\nu_f(\Xi) = M\nu_f(\Xi)$. Hence, by the Radon-Nikodym theorem, there is an $L_1(\mu)$ function, call it Pf, such that $M\nu_f(A) = \int_A Pf(x)d\mu$. ("Almost unique"

refers to the possibility of replacing Pf with another version of $dM\nu_f/d\mu$.) In particular, $Pf(x) \geq 0$ for μ-almost-all x, and so $\|Pf\| = \int_\Xi |Pf(x)|d\mu = M\nu_f(\Xi) = \nu_f(\Xi) = \int_\Xi |f(x)|d\mu = \|f\|$. Finally, the linearity of the operator on densities follows directly from the linearity of the operator on measures and the linearity of integration. If f is sometimes negative, apply the reasoning above to f^+ and f^-, its positive and negative parts, and then use linearity again.□

Recall from Definition 30 that, for an arbitrary kernel κ, $\kappa f(x)$ is defined as $\int f(y)\kappa(x, dy)$. Applied to our transition kernels, this suggests another kind of operator.

Definition 119 (Transition Operators) *Take any measurable space $\Xi, \mathcal{X}$, and let $B(\Xi)$ be the class of bounded measurable functions. An operator $K : B(\Xi) \mapsto B(\Xi)$ is a* transition operator *when:*

1. *K is linear*
2. *If $f \geq 0$ (a.e. μ), $Kf \geq 0$ (a.e. μ)*
3. *$K1_\Xi = 1_\Xi$.*
4. *If $f_n \downarrow 0$, then $Kf_n \downarrow 0$.*

Definition 120 (L_∞ Transition Operators) *For a probability space $\Xi, \mathcal{X}, \mu$, an L_∞-transition operator is an operator on $L_\infty(\mu)$ satisfying points (1)–(4) of Definition 119.*

Note that every function in $B(\Xi)$ is in $L_\infty(\mu)$ for each μ, so the definition of an L_∞ transition operator is actually stricter than that of a plain transition operator. This is unlike the case with Markov operators, where the L_1 version is weaker than the unrestricted version.

Lemma 121 (Kernels and Operators) *Every probability kernel κ from Ξ to Ξ induces a Markov operator M on measures,*

$$M\nu = \nu\kappa \tag{10.4}$$

and every Markov operator M on measures induces a probability kernel κ,

$$\kappa(x, B) = M\delta_x(B) \tag{10.5}$$

Similarly, every transition probability kernel induces a transition operator K on functions,

$$Kf(x) = \kappa f(x) \tag{10.6}$$

and every transition operator K induces a transition probability kernel,

$$\kappa(x, B) = K\mathbf{1}_B(x) \tag{10.7}$$

PROOF: Exercise 10.1. □

Now we need some basic concepts from functional analysis; see, e.g., Kolmogorov and Fomin (1975) for background.

Definition 122 (Functional) *A* functional *is a map $g : V \mapsto \mathbb{R}$, that is, a real-valued function of a function. A functional g is*

- linear *when $g(af_1 + bf_2) = ag(f_1) + bg(f_2)$;*
- continuous *when $f_n \to f$ implies $g(f_n) \to f$;*
- bounded by M *when $|g(f)| \leq M$ for all $f \in V$;*
- bounded *when it is bounded by M for some M;*
- non-negative *when $g(f) \geq 0$ for all f;*

etc.

Definition 123 (Conjugate or Adjoint Space) *The* conjugate space *or* adjoint space *of a vector space V is the space $V^\dagger$ of its continuous linear functionals. For $f \in V$ and $g \in V^\dagger$, $\langle f, g \rangle$ denotes $g(f)$. This is sometimes called the* inner product.

Proposition 124 (Conjugate Spaces are Vector Spaces) *For every V, $V^\dagger$ is also a vector space.*

Proposition 125 (Inner Product is Bilinear) *For any $a, b, c, d \in \mathbb{R}$, any $f_1, f_2 \in V$ and any $g_1, g_2 \in V^\dagger$,*

$$\langle af_1 + bf_2, cg_1 + dg_2 \rangle = ac\langle f_1, g_1 \rangle + ad\langle f_1, g_2 \rangle + bc\langle f_2, g_1 \rangle + bd\langle f_2, g_2 \rangle \quad (10.8)$$

PROOF: Follows from the fact that $V^\dagger$ is a vector space, and each g_i is a linear operator. □

You are already familiar with an example of a conjugate space.

Example 126 (Vectors in $\mathbb{R}^n$) *The vector space $\mathbb{R}^n$ is self-conjugate. If $g(\vec{x})$ is a continuous linear function of $\vec{x}$, then $g(\vec{x}) = \sum_{i=1}^n y_i x_i$ for some real constants y_i, which means $g(\vec{x}) = \vec{y} \cdot \vec{x}$.*

Here is the simplest example where the conjugate space is not equal to the original space.

Example 127 (Row and Column Vectors) *The space of row vectors is conjugate to the space of column vectors, since every continuous linear functional of a column vector x takes the form of $y^{\mathrm{T}}x$ for some other column vector y.*

Example 128 (L_p spaces) *The function spaces $L_p(\mu)$ and $L_q(\mu)$ are conjugate to each other, when $1/p + 1/q = 1$, and the inner product is defined through*

$$\langle f, g \rangle \equiv \int f g d\mu \tag{10.9}$$

In particular, L_1 and L_∞ are conjugates.

Example 129 (Measures and Functions) *The space of $C_b(\Xi)$ of bounded, continuous functions on Ξ and the spaces $\mathcal{M}(\Xi, \mathcal{X})$ of finite measures on Ξ are conjugates, with inner product*

$$\langle \mu, f \rangle = \int f d\mu \tag{10.10}$$

Definition 130 (Adjoint Operator) *For conjugate spaces V and $V^\dagger$, the* adjoint operator*, $O^\dagger$, to an operator O on V is an operator on $V^\dagger$ such that*

$$\langle Of, g \rangle = \langle f, O^\dagger g \rangle \tag{10.11}$$

for all $f \in V, g \in V^\dagger$.

Proposition 131 (Adjoint of a Linear Operator) *If O is a continuous linear operator on V, then its adjoint $O^\dagger$ exists and is linear.*

Lemma 132 (Markov Operators on Densities and L_∞ Transition Operators) *Every Markov operator P on densities induces an L_∞ transition operator U on essentially-bounded functions, and vice versa.*

PROOF: Exercise 10.2. □

Clearly, if κ is part of a transition kernel semi-group, then the collection of induced Markov operators and transition operators also form semi-groups.

Theorem 133 (Transition operator semi-groups and Markov processes) *Let X be a Markov process with transition kernels $\mu_{t,s}$, and let $K_{t,s}$ be the corresponding semi-group of transition operators. Then for any $f \in B(\Xi)$,*

$$\mathbf{E}\left[f(X_s)|\mathcal{F}_t\right] = (K_{t,s}f)(X_t) \tag{10.12}$$

Conversely, let X be any stochastic process, and let $K_{t,s}$ be a semi-group of transition operators such that Equation 10.12 is valid (a.s.). Then X is a Markov process.

PROOF: Exercise 10.3. □

Remark. The proof works because the expectations of all $B(\Xi)$ functions together determine a probability measure. (Recall that $\mathbb{P}(B) = \mathbf{E}[\mathbf{1}_B]$, and indicators are bounded everywhere.) If we knew of another collection of functions which also sufficed to determine a measure, then linear operators on that collection would work just as well, in the theorem, as do the transition operators, which by definition apply to all of $B(\Xi)$. In particular, it is sometimes possible to define operators only on much smaller, more restricted collections of functions, which can have technical advantages. See Ethier and Kurtz (1986, ch. 4, sec. 1) for details.

The next two lemmas are useful in establishing asymptotic results.

Lemma 134 (Markov Operators are Contractions) *For any Markov operator P and $f \in L_1$,*

$$\|Pf\| \leq \|f\| \tag{10.13}$$

PROOF (after Lasota and Mackey (1994, prop. 3.1.1, pp. 38–39)): First, notice that $(Pf(x))^+ \leq Pf^+(x)$, because

$$(Pf(x))^+ = (Pf^+ - Pf^-)^+ = \max(0, Pf^+ - Pf^-) \leq \max(0, Pf^+) = Pf^+$$

Similarly $(Pf(x))^- \leq Pf^-(x)$. Therefore $|Pf| \leq P|f|$, and then the statement follows by integration. □

Lemma 135 (Markov Operators Bring Distributions Closer) *For any Markov operator, and any $f, g \in L_1$, $\|P^n f - P^n g\|$ is non-increasing.*

PROOF: By linearity, $\|P^n f - P^n g\| = \|P^n(f-g)\|$. By the definition of P^n, $\|P^n(f-g)\| = \|PP^{n-1}(f-g)\|$. By the contraction property (Lemma 134), $\|PP^{n-1}(f-g)\| \leq \|P^{n-1}(f-g)\| = \|P^{n-1}f - P^{n-1}g\|$ (by linearity again). □

Theorem 136 (Invariant Measures Are Fixed Points) *A probability measure ν is invariant for a homogeneous Markov process iff it is a fixed point of all the Markov operators, $M_t\nu = \nu$.*

PROOF: Clear from the definitions! □

10.3 Exercises

Exercise 10.1 (Kernels and Operators) *Prove Lemma 121.* Hints*: 1. You will want to use the fact that $\mathbf{1}_A \in B(\Xi)$ for all measurable sets A. 2. In going back and forth between transition kernels and transition operators, you may find Proposition 32 helpful.*

Exercise 10.2 (L_1 and L_∞) *Prove Lemma 132.*

Exercise 10.3 (Operators and Expectations) *Prove Theorem 133. Hint: in showing that a collection of operators determines a Markov process, try using mathematical induction on the finite-dimensional distributions.*

Exercise 10.4 (Bayesian Updating as a Markov Process) *Consider a simple version of Bayesian learning, where the set of hypotheses Θ is finite, and, for each $\theta \in \Theta$, $f_\theta(x)$ is a probability density on Ξ with respect to a common dominating measure, μ say, and the Ξ-valued data $X_1, X_2, \ldots$ are all IID, both under the hypotheses and in reality. Given a prior probability vector π_0 on Θ, the posterior π_n is defined via Bayes's rule:*

$$\pi_n(\theta) = \frac{\pi_0(\theta)\prod_{i=1}^{n} f_\theta(X_i)}{\sum_{\theta\in\Theta}\pi_0(\theta)\prod_{i=1}^{n} f_\theta(X_i)}$$

1. *Prove that the random sequence $\pi_1, \pi_2, \ldots$ is adapted to $\{\mathcal{F}\}_t$ if X is adapted.*
2. *Prove that the sequence of posterior distributions is Markovian with respect to its natural filtration.*
3. *Is this still a Markov process if X is not IID? If the hypotheses θ do not model X as IID?*
4. *When, if ever, is the Markov process homogeneous? (If it is sometimes homogeneous, you may give either necessary or sufficient conditions, as you find easier.)*

Exercise 10.5 (More on Bayesian Updating) *Consider a more complicated version of Bayesian updating. Let T be one-sided, H be a Θ-valued random variable, and $\{\mathcal{G}\}_t$ be any filtration. Assume that $\pi_t = \mathcal{L}(H|\mathcal{G}_t)$ is a regular conditional probability distribution on Θ for all t and all ω. As before, π_0 is the prior. Prove that π_t is Markovian with respect to $\{\mathcal{G}\}_t$. (Hint: $\mathbf{E}[\mathbf{E}[Y|\mathcal{G}_t]|\mathcal{G}_s] = \mathbf{E}[X|\mathcal{G}_s]$ a.s., when $s \leq t$ and $Y \in L_1$ so the expectations exist.)*