Chapter 2

Basic Markov Chain Theory

To repeat what we said in the Chapter 1, a *Markov chain* is a discrete-time stochastic process X_1, X_2, ... taking values in an arbitrary state space that has the Markov property and stationary transition probabilities:

- the conditional distribution of X_n given X_1, ..., X_{n-1} is the same as the conditional distribution of X_n given X_{n-1} only, and
- the conditional distribution of X_n given X_{n-1} does not depend on n.

The conditional distribution of X_n given X_{n-1} specifies the *transition probabilities* of the chain. In order to completely specify the probability law of the chain, we need also specify the *initial distribution*, the distribution of X_1.

2.1 Transition Probabilities

2.1.1 Discrete State Space

For a discrete state space S, the transition probabilities are specified by defining a matrix

$$P(x, y) = \Pr(X_n = y | X_{n-1} = x), \qquad x, y \in S \tag{2.1}$$

that gives the probability of moving from the point x at time $n-1$ to the point y at time n. Because of the assumption of stationary transition probabilities, the transition probability matrix $P(x, y)$ does not depend on the time n.

Some readers may object that we have not defined a "matrix." A *matrix* (I can hear such readers saying) is a rectangular array P of numbers p_{ij}, $i = 1$, ..., m, $j = 1$, ..., n, called the *entries* of P. Where is P? Well, enumerate the points in the state space $S = \{x_1, \ldots, x_d\}$, then

$$p_{ij} = \Pr\{X_n = x_j | X_{n-1} = x_i\}, \qquad i = 1, \ldots, d, \ j = 1, \ldots d.$$

I hope I can convince you this view of "matrix" is the Wrong Thing. There are two reasons.

First, the enumeration of the state space does no work. It is an irrelevancy that just makes for messier notation. The mathematically elegant definition of a matrix does not require that the index sets be $\{1, \ldots, m\}$ and $\{1, \ldots, n\}$ for some integers m and n. Any two finite sets will do as well. In this view, a *matrix* is a function on the Cartesian product of two finite sets. And in this view, the function P defined by (2.1), which is a function on $S \times S$, *is* a matrix.

Following the usual notation of set theory, the space of all real-valued functions on a set A is written $\mathbb{R}^A$. This is, of course, a d-dimensional vector space when A has d points. Those who prefer to write $\mathbb{R}^d$ instead of $\mathbb{R}^A$ may do so, but the notation $\mathbb{R}^A$ is more elegant and corresponds to our notion of A being the index set rather than $\{1, \ldots, d\}$. So our matrices P being functions on $S \times S$ are elements of the d^2-dimensional vector space $\mathbb{R}^{S \times S}$.

The second reason is that P is a conditional probability mass function. In most contexts, (2.1) would be written $p(y|x)$. For a variety of reasons, partly the influence of the matrix analogy, we write $P(x, y)$ instead of $p(y|x)$ in Markov chain theory. This is a bit confusing at first, but one gets used to it. It would be much harder to see the connection if we were to write p_{ij} instead of $P(x, y)$.

Thus, in general, we define a *transition probability matrix* to be a real-valued function P on $S \times S$ satisfying

$$P(x, y) \geq 0, \qquad x, y \in S \tag{2.2a}$$

and

$$\sum_{y \in S} P(x, y) = 1. \tag{2.2b}$$

The state space S must be countable for the definition to make sense. When S is not finite, we have an infinite matrix. Any matrix that satisfies (2.2a) and (2.2b) is said to be *Markov* or *stochastic*.

Example 2.1. Random Walk with Reflecting Boundaries.
Consider the symmetric random walk on the integers 1, ..., d with "reflecting boundaries." This means that at each step the chain moves one unit up or down with equal probabilities, $\frac{1}{2}$ each way, except at the end points. At 1, the lower end, the chain still moves up to 2 with probability $\frac{1}{2}$, but cannot move down, there being no points below to move to. Here when it wants to go down, which is does with probability $\frac{1}{2}$, it bounces off an imaginary reflecting barrier back to where it was. The behavior at the upper end is analogous. This gives a transition matrix

$$\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} & 0 & 0 & \ldots & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & \ldots & 0 & 0 & 0 \\
0 & \frac{1}{2} & 0 & \frac{1}{2} & \ldots & 0 & 0 & 0 \\
0 & 0 & \frac{1}{2} & 0 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{2} & 0 \\
0 & 0 & 0 & 0 & \ldots & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & \ldots & 0 & \frac{1}{2} & \frac{1}{2}
\end{pmatrix} \tag{2.3}$$

We could instead use functional notation

$$P(x, y) = \begin{cases} 1/2, & |x - y| = 1 \text{ or } x = y = 1 \text{ or } x = y = d \\ 0, & \text{otherwise} \end{cases}$$

Either works. We will use whichever is most convenient.

2.1.2 General State Space

For a general state space S the transition probabilities are specified by defining a *kernel*

$$P(x, B) = \Pr\{X_n \in B | X_{n-1} = x\}, \qquad x \in S, \; B \text{ a measurable set in } S,$$

satisfying

- for each fixed x the function $B \mapsto P(x, B)$ is a probability measure, and
- for each fixed B the function $x \mapsto P(x, B)$ is a measurable function.

In other words, the kernel is a *regular* conditional probability (Breiman 1968, Section 4.3).

Lest the reader worry that this definition signals an impending blizzard of measure theory, let me assure you that it does not. A little bit of measure theory is unavoidable in treating this subject, if only because the major reference works on Markov chains, such as Meyn and Tweedie (1993), are written at that level. But in practice measure theory is entirely dispensable in MCMC, because the computer has no sets of measure zero or other measure-theoretic paraphernalia. So if a Markov chain really exhibits measure-theoretic pathology, it can't be a good model for what the computer is doing.

In any case, we haven't hit serious measure theory yet. The main reason for introducing kernels here is purely notational. It makes unnecessary a lot of useless discussion of special cases. It allows us to write expressions like

$$E\{g(X_n) | X_{n-1} = x\} = \int P(x, dy) g(y) \tag{2.4}$$

using one notation for all cases. Avoiding measure-theoretic notation leads to excruciating contortions.

Sometimes the distribution of X_n given X_{n-1} is a continuous distribution on $\mathbb{R}^d$ with density $f(y|x)$. Then the kernel is defined by

$$P(x, B) = \int_B f(y|x) \, dy$$

and (2.4) becomes

$$E\{g(X_n) | X_{n-1} = x\} = \int g(y) f(y|x) \, dy.$$

Readers who like boldface for "vectors" can supply the appropriate boldface. Since both x and y here are elements of $\mathbb{R}^d$, every variable is boldfaced. I don't like the "vectors are boldface" convention. It is just one more bit of distinguishing trivial special cases that makes it much harder to see what is common to all cases.

Often the distribution of X_n given X_{n-1} is more complicated. A common situation in MCMC is that the distribution is continuous except for an atom at x. The chain stays at x with probability $r(x)$ and moves with probability $1 - r(x)$, and when it moves the distribution is given by a density $f(y|x)$. Then (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = r(x)g(x) + [1 - r(x)] \int g(y)f(y|x)\,dy.$$

The definition of the kernel in this case is something of a mess

$$P(x, B) = \begin{cases} r(x) + [1 - r(x)] \int_B f(y|x)\,dy, & x \in B \\ [1 - r(x)] \int_B f(y|x)\,dy, & \text{otherwise} \end{cases} \tag{2.5}$$

This can be simplified by introducing the *identity kernel* (yet more measure-theoretic notation) defined by

$$I(x, B) = \begin{cases} 1, & x \in B \\ 0, & x \notin B \end{cases} \tag{2.6}$$

which allows us to rewrite (2.5) as

$$P(x, B) = r(x)I(x, B) + [1 - r(x)] \int_B f(y|x)\,dy.$$

We will see why the identity kernel has that name a bit later.

Another very common case in MCMC has the distribution of X_n given X_{n-1} changing only one component of the state vector, say the i-th. The Gibbs update discussed in Chapter 1 is an example. The distribution of the i-th component has a density $f(y|x)$, but now x is an element of $\mathbb{R}^d$ and y is an element of $\mathbb{R}$ (not $\mathbb{R}^d$). Then (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = \int g(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d)f(y|x)\,dy.$$

The notation for the kernel is even uglier unless we use "probability is a special case of expectation." To obtain the kernel just take the special case where g is the indicator function of the set B.

The virtue of the measure-theoretic notation (2.4) is that it allows us to refer to all of these special cases and many more without getting bogged down in a lot of details that are irrelevant to the point under discussion. I have often wondered why this measure-theoretic notation isn't introduced in lower

level courses. It would avoid tedious repetition, where first we woof about the discrete case, then the continuous case, even rarely the mixed case, thus obscuring what is common to all the cases. One can use the notation without knowing anything about measure-theoretic probability. Just take (2.4) as the definition of the notation. If you understand what expectations mean in the model at hand, then you can write out what the notation means in each case, as we have done above. Regardless of whether you think this would be a good idea in lower level courses, or not, I hope you are convinced that the notation is necessary in dealing with Markov chains. One would never see the forest for the trees without it.

2.1.3 Existence of Infinite Random Sequences

Transition probabilities do not by themselves define the probability law of the Markov chain, though they do define the law conditional on the initial position, that is, given the value of X_1. In order to specify the unconditional law of the Markov chain we need to specify the *initial distribution* of the chain, which is the marginal distribution of X_1.

If λ is the initial distribution and P is the transition kernel and g_1, ..., g_n are any real-valued functions, then

$$\begin{aligned} &E\{g_1(X_1) \ldots g_n(X_n)\} \\ &\qquad = \int \cdots \int \lambda(dx_1) P(x_1, dx_2) \cdots P(x_{n-1}, dx_n) g_1(x_1) \cdots g_n(x_n) \end{aligned}$$

provided the expectation exists. This determines the joint probability distribution of X_1, ..., X_n for any n. Just take the special case where the g_i are indicator functions.

Let Q_n denote the probability distribution of X_1, ..., X_n, a measure on the cartesian product S^n, where S is the state space. The Q_n are called the *finite-dimensional* distributions of the infinite random sequence X_1, X_2, ... The finite-dimensional distributions satisfy the obvious consistency property: $Q_n(A) = Q_{n+1}(A \times S)$. It is a theorem of measure-theoretic probability (Fristedt and Gray 1997, Theorem 3 of Chapter 22 and Definition 10 of Chapter 21) that for any consistent sequence of finite-dimensional distributions, there exists a unique probability measure Q_∞ for the infinite sequence such that Q_∞ agrees with the finite-dimensional distributions, that is, if A is a measurable set in S^n and

$$B = \{\, (x_1, x_2, \ldots) \in S^\infty : (x_1, \ldots, x_n) \in A \,\},$$

then $Q_n(A) = Q_\infty(B)$.

We will only rarely refer explicitly or even implicitly to Q_∞. One place where it cannot be avoided is the strong law of large numbers, which says that the set of infinite sequences $(X_1, X_2, \ldots)$ having the property that $\overline{X}_n \to \mu$ has probability one, the probability here referring to Q_∞, since it refers to probabilities on the space of infinite sequences. But mostly we deal only with

finite-dimensional distributions. The CLT, for example, is a statement about finite-dimensional distributions only.

Anyway, this issue of Q_∞ has nothing to do particularly with Markov chains. It is needed for the SLLN in the i. i. d. case too. If you are not bothered by the SLLN for i. i. d. random sequences, then the SLLN for Markov chains should not bother you either. The measure-theoretic technicalities are exactly the same in both cases.

2.2 Transition Probabilities as Operators

When the state space is finite, we have seen that the transition probabilities form a matrix, an $d \times d$ matrix if the state space has d points. From linear algebra, the reader should be familiar with the notion that a matrix represents a linear operator. This is true for Markov transition matrices as well. Actually, we will see it represents two different linear operators.

In the general state space case, transition probabilities also represent linear operators. In this case the vector spaces on which they operate are infinite-dimensional. We do not assume the reader should be familiar with these notions and so develop what we need of this theory to work with Markov chains.

2.2.1 Finite State Space

Right Multiplication

When the state space S is finite (2.4) becomes

$$E\{g(X_n)|X_{n-1} = x\} = \sum_{y \in S} P(x, y)g(y).$$

Although the notation is unusual, the right hand side corresponds to the matrix multiplication of the matrix P on the right by the "column vector" g. Using this notation we write the function defined by the right hand side as Pg. Hence we have

$$Pg(x) = E\{g(X_n)|X_{n-1} = x\}.$$

If we were fussy, we might write the left hand side as $(Pg)(x)$, but the extra parentheses are unnecessary, since the other interpretation of $Pg(x)$, that P operates on the real number $g(x)$, is undefined.

As mentioned above, the vector space of all real-valued functions on S is denoted $\mathbb{R}^S$. The operation of right multiplication defined above takes a function g in $\mathbb{R}^S$ to another function Pg in $\mathbb{R}^S$. This map $R_P : g \mapsto Pg$ is a linear operator on $\mathbb{R}^S$ represented by the matrix P. When we are fussy, we distinguish between the matrix P and the linear operator R_P it represents, as is common in introductory linear algebra books (Lang 1987, Chapter IV). But none of the Markov chain literature bothers with this distinction. So we will bother with making this distinction only for a little while. Later we will just write P instead of R_P as all the experts do, relying on context to make it clear whether P means

a matrix or a linear operator. We don't want the reader to think that making a clear distinction between the matrix P and the linear operator R_P is essential. Holding fast to that notational idiosyncrasy will just make it hard for you to read the literature.

Left Multiplication

A probability distribution on S is also determines a vector in $\mathbb{R}^S$. In this case the vector is the probability mass function $\lambda(x)$. If X_{n-1} has the distribution λ, then the distribution of X_n is given by

$$\Pr(X_n = y) = \sum_{x \in S} \lambda(x) P(x, y). \tag{2.7}$$

Again we can recognize a matrix multiplication, this time of the matrix P on the left by the "row vector" λ. Using this notation we write the probability distribution defined by the right hand side as λP. and hence have

$$\lambda P(y) = \Pr(X_n = y),$$

when X_{n-1} has the distribution λ. Again if we were fussy, we might write the left hand side as $(\lambda P)(y)$, but again the extra parentheses are unnecessary, since the other interpretation of $\lambda P(y)$, that $P(y)$ operates on λ, is undefined because $P(y)$ is undefined.

Equation (2.7) makes sense when λ is an arbitrary element of $\mathbb{R}^S$, in which case we say it represents a *signed measure* rather than a probability measure. Thus the matrix P also represents another linear operator on $\mathbb{R}^S$, the operator $L_P : \lambda \mapsto \lambda P$. Note that L_P and R_P are not the same operator, because P is not a symmetric matrix, so right and left multiplication produce different results.

When we are not being pedantic, we will usually write P instead of L_P or R_P. So how do we tell these two operators apart? In most contexts only one of the two is being used, so there is no problem. In contexts where both are in use, the notational distinction between Pf and λP helps distinguish them.

Invariant Distributions

Recall from Section 1.5 that a probability distribution π is an *invariant* distribution for a specified transition probability matrix P if the Markov chain that results from using π as the initial distribution is stationary. (An invariant distribution is also called a *stationary* or an *equilibrium* distribution.) Because the transition probabilities are assumed stationary, as we always do, it is enough to check that $X_{n-1} \sim \pi$ implies $X_n \sim \pi$. But we have just learned that $X_{n-1} \sim \lambda$ implies $X_n \sim \lambda P$. Hence we can use our new notation to write the characterization of invariant distributions very simply: a probability distribution π is invariant for a transition probability matrix P if and only if $\pi = \pi P$.

Recall from Section 1.7 that the "first task in MCMC" is to find a Markov update mechanism that preserves a specified distribution. Now we can state

that in notation. We are given a distribution π. The "first task" is to find one transition probability matrix P such that $\pi = \pi P$. Often, we want to find several such matrices or kernels, intending to combine them by composition or mixing.

Matrix Multiplication (Composition of Operators)

The distribution of X_{n+2} given X_n is given by

$$\Pr(X_{n+2} = z | X_n = x) = \sum_{y \in S} P(x, y) P(y, z).$$

Now we recognize a matrix multiplication. The right hand side is the (x, z) entry of the matrix P^2, which we write $P^2(x, z)$. Carrying the process further we see that

$$\Pr(X_{n+k} = z | X_n = x) = P^k(x, z),$$

where $P^k(x, z)$ denotes the (x, z) entry of the matrix P^k.

We can use these operations together. $P^k g$ is the conditional expectation of $g(X_{n+k})$ given X_n, and λP^k is the marginal distribution of X_{n+k} when X_n has marginal distribution λ.

We also want to use this operation when the transition probability matrices are different. Say $P(x, y)$ and $Q(x, y)$ are two transition probability matrices, their product is defined in the obvious way

$$(PQ)(x, z) = \sum_{y \in S} P(x, y) Q(y, z).$$

We met this object in Chapter 1 under the name of the composition of P and Q, which we wrote as PQ, anticipating that it would turn out to be a matrix multiplication. The reason for calling it "composition" is that it is functional composition when we think of P and Q as linear operators. Obviously, $(PQ)g = P(Qg)$. This translates to

$$R_{PQ} = R_P \circ R_Q \tag{2.8a}$$

when we use the notation R_P for the linear operator $f \mapsto Pf$. It translates to

$$L_{PQ} = L_Q \circ L_P \tag{2.8b}$$

when we use the notation L_P for the linear operator $\lambda \mapsto \lambda P$. In both cases matrix multiplication represents functional composition, but note that P and Q appear in opposite orders on the right hand sides of (2.8a) and (2.8b), the reason being the difference between right and left multiplication.

Convex Combinations of Matrices (Mixing)

Besides multiplication of matrices, linear algebra also defines the operations of matrix addition and multiplication of a matrix by a scalar. Neither of these

operations turns a Markov matrix into a Markov matrix, because matrix addition loses property (2.2b) and multiplication by a negative scalar loses property (2.2a).

If we use both operations together, we can get an operation that preserves Markovness. Transition probability matrices are elements of the vector space $\mathbb{R}^{S\times S}$, a d^2-dimensional vector space if the state space S has d elements. Addition of matrices is just vector addition in this vector space. Multiplication of a matrix by a scalar is just scalar multiplication in this vector space. If P_1, ..., P_k are elements of any vector space, and a_1, ..., a_k are scalars, then

$$P = a_1 P_1 + \cdots + a_k P_k \tag{2.9}$$

is called a *linear combination* of the P_i. If the a_i also satisfy $\sum_i a_i = 1$, a linear combination is called an *affine combination*. If the a_i also satisfy $a_i \geq 0$ for each i, an affine combination is called a *convex combination*.

For Markov matrices P_1, ..., P_k,

- if P in (2.9) is Markov, then linear combination is affine,
- conversely, if the linear combination is convex, then P is Markov.

(Exercise 2.2).

Convex combinations correspond exactly to the operation of mixing of update mechanisms (also called "random scan") described in Section 1.7. if there are k update mechanisms, the i-th mechanism described by transition probability matrix P_i, and we choose to execute the i-the mechanism with probability a_i, then the transition probability matrix for the combined update mechanism is given by (2.9). In order to be probabilities the a_i must be nonnegative and sum to one, which is exactly the same as the requirement for (2.9) to be a convex combination. We would have called this notion "convex combination" rather than "mixture," but that seemed too long for everyday use.

2.2.2 General State Space

Now we turn to general state spaces, and kernels replace matrices. The objects on which the kernels operate on the left and right now are very different, a function on the state space (an object for right multiplication) is not at all like a measure on the state space (and object for left multiplication).

Signed Measures

In the discrete case we wanted to talk about measures that were not probability measures. We need a similar notion for general state spaces. A real-valued measure on a measurable space[1] $(S, \mathcal{B})$ is a function $\mu : \mathcal{B} \to \mathbb{R}$ that is countably additive.

[1] A *measurable space* is a pair $(S, \mathcal{B})$ consisting of a set S, in this case the state space, and a σ-field of subsets of S. The elements of $\mathcal{B}$ are called the *measurable sets* or, when we are talking about probabilities, *events*. So $\mathcal{B}$ is just the set of all possible events.

Although not part of the definition, it is a theorem of real analysis that μ is actually a bounded function (Rudin 1987, Theorem 6.4), that is, there are constants a and b such that $a \le \mu(B) \le b$ for all $B \in \mathcal{B}$. If $\mu(B) \ge 0$ for all measurable sets B, then we say μ is a *positive* measure. The general case, in which $\mu(B)$ takes values of both signs, is sometimes called a real *signed* measure, although strictly speaking the "signed" is redundant.

Another theorem (Rudin 1987, Theorem 6.14) says that there exists a partition[2] of the state space into two measurable sets A_1 and A_2 such that

$$\begin{aligned} \mu(B) &\le 0, \qquad B \subset A_1 \\ \mu(B) &\ge 0, \qquad B \subset A_2 \end{aligned}$$

This is called the *Hahn decomposition* of the state space S. Then the measures μ^+ and μ^- defined by

$$\begin{aligned} \mu^-(B) &= -\mu(B \cap A_1), && B \in \mathcal{B} \\ \mu^+(B) &= \mu(B \cap A_2), && B \in \mathcal{B} \end{aligned}$$

are both positive measures on S and they are mutually singular. Note that $\mu = \mu^+ - \mu^-$, which is called the *Jordan decomposition* of μ. It is entirely analogous to the decomposition $f = f^+ - f^-$ of a function into its positive and negative parts. The measure $|\mu| = \mu^+ + \mu^-$ is called the *total variation* of μ. And $\|\mu\| = |\mu|(S)$ is called the *total variation norm* of μ.

Let $\mathcal{M}(S)$ denote the set of all real signed measures on S. From the Jordan decomposition, we see that every element of $\mathcal{M}(S)$ is a difference of positive finite measures, hence a linear combination of probability measures. Thus $\mathcal{M}(S)$ is the vector space spanned by the probability measures. Hence it is the proper replacement for $\mathbb{R}^S$ in our discussion of left multiplication in the discrete case.

Norms and Operator Norm

For any vector space V, a function $x \mapsto \|x\|$ from V to $[0, \infty)$ is called a *norm* on V if it satisfies the following axioms (Rudin 1987, p. 95)

(a) $\|x + y\| \le \|x\| + \|y\|$ for all $x, y \in V$,

(b) $\|ax\| = |a| \cdot \|x\|$ for all $a \in \mathbb{R}$ and $x \in V$, and

(c) $\|x\| = 0$ implies $x = 0$.

Axiom (a) is called the *triangle inequality*. The pair $(V, \|\cdot\|)$ is called a *normed vector space* or a *normed linear space*.

Total variation norm makes $\mathcal{M}(S)$ a normed vector space. We do need to verify that total variation norm does satisfy the axioms for a norm (Exercise 2.3).

Denote the set of all linear operators on a vector space V by $L(V)$. Then $L(V)$ is itself a vector space if we define vector addition by

$$(S + T)(x) = S(x) + T(x), \qquad S, T \in L(V),\ x \in V \tag{2.10a}$$

[2]Partition means $A_1 \cap A_2 = \varnothing$ and $A_1 \cup A_2 = S$

and scalar multiplication by

$$(aT)(x) = aT(x), \qquad a \in \mathbb{R},\ T \in L(V),\ x \in V. \tag{2.10b}$$

These definitions are the obvious ones, arrived at almost without thinking. How else would you define the sum of two functions S and T except as the sum (2.10a)?

When V is normed, there is a natural corresponding norm for $L(V)$ defined by

$$\|T\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Tx\|}{\|x\|} \tag{2.11}$$

Or, more precisely, we should say that (2.11) defines a norm for the subset of $L(V)$ consisting of T such that (2.11) is finite. We denote that subset $B(V)$, and call its elements the *bounded operators* on $L(V)$. The bounded operators are the well behaved ones.

A normed linear space is also a metric space, the metric being defined by $d(x, y) = \|x - y\|$. Hence we can discuss topological notions like continuity and convergence of sequences. A sequence $\{x_n\}$ in V converges to a point x if $\|x_n - x\| \to 0$. An operator $T \in L(V)$ is continuous at a point x if $Tx_n \to Tx$ (meaning $\|Tx_n - Tx\| \to 0$) for every sequence $\{x_n\}$ converging to x. Since $Tx_n - Tx = T(x_n - x)$ by linearity, a linear operator T is continuous at x if and only if it is continuous at zero. Thus linear operators are either everywhere continuous or nowhere continuous. A linear operator T is continuous if and only if it is bounded (Rudin 1991, Theorem 1.32). Thus the unbounded operators are nowhere continuous, a fairly obnoxious property. If V is finite-dimensional, then every operator in $L(V)$ is bounded (Halmos 1958, p. 177). But if V is infinite-dimensional, there are lots of unbounded operators.

Let's check that operator norm satisfies the norm axioms. Essentially it satisfies the axioms because vector norm does. For the triangle inequality

$$\begin{aligned}
\|S + T\| &= \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx + Tx\|}{\|x\|} \\
&\leq \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx\| + \|Tx\|}{\|x\|} \\
&\leq \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Sx\|}{\|x\|} + \sup_{\substack{y \in V \\ y \neq 0}} \frac{\|Ty\|}{\|y\|} \\
&= \|S\| + \|T\|
\end{aligned}$$

The first inequality is the triangle inequality for the vector norm. The second inequality is subadditivity of the supremum operation. For any functions f and g on any set S

$$f(x) + g(x) \leq f(x) + \sup_{y \in S} g(y),$$

so taking the sup over x gives

$$\sup_{x \in S} [f(x) + g(x)] \le \sup_{x \in S} f(x) + \sup_{y \in S} g(y).$$

For axiom (b),

$$\|aT\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|aTx\|}{\|x\|} = \sup_{\substack{x \in V \\ x \neq 0}} \frac{|a| \cdot \|Tx\|}{\|x\|} = a\|T\|.$$

Finally, for axiom (c), $\|T\| = 0$ only if $\|Tx\| = 0$ for all $x \in V$, but axiom (c) for vector norm implies $\|Tx\| = 0$ if and only if $Tx = 0$. Thus $\|T\| = 0$ implies that T is the operator that maps every x to 0. And this operator is indeed the zero of the vector space $L(V)$, because then

$$(S + T)(x) = S(x) + T(x) = S(x) + 0 = S(x), \qquad x \in V$$

so $S + T = S$ for all $S \in L(V)$, and this is the property that makes T the zero of the vector space $L(V)$.

Operator norm satisfies two important inequalities. The first

$$\|Tx\| \le \|T\| \cdot \|x\| \tag{2.12}$$

follows immediately from the definition (2.11).

The second involves the notion of operator "multiplication," which is defined as composition of functions: ST is shorthand for $S \circ T$. As we saw above, this agrees with our usual notation in the finite-dimensional case: matrix multiplication corresponds to functional composition of the corresponding operators. With this notion of multiplication $B(V)$ becomes an *operator algebra*. A *vector algebra*, also called *linear algebra*, is a vector space in which a multiplication is defined. The reason the subject "linear algebra" is so called is because matrices form a vector algebra.

The second important inequality is

$$\|ST\| \le \|S\| \cdot \|T\|. \tag{2.13}$$

I call (2.13) the *Banach algebra inequality* because it is one of the defining properties of a Banach algebra. Since we will have no need of Banach algebras in this course, it is a really horrible name. Maybe we should call it the *mumble mumble* inequality. Whatever we call it, the proof is a trivial consequence of operator "multiplication" actually being functional composition.

$$\|ST\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|S(Tx)\|}{\|x\|} \le \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|S\| \cdot \|Tx\|}{\|x\|} = \|S\| \cdot \|T\|$$

where the inequality is just (2.12).

Left Multiplication

If λ is a probability measure on the state space, and X_{n-1} has distribution λ, then the distribution of X_n is given by

$$\lambda P(A) = \int \lambda(dx) P(x, A). \tag{2.14}$$

This is no longer a matrix multiplication, but it does define a linear operator, because integration is a linear operation. Using the Jordan decomposition, we see that (2.14) makes sense for any $\lambda \in \mathcal{M}(S)$. Hence (2.14) defines a linear operator on $\mathcal{M}(S)$.

The next question to answer is whether it is a well-behaved operator, that is, whether it is bounded. In fact, it is. For any Markov kernel P, let L_P denote the linear operator on $\mathcal{M}(S)$ defined by $\lambda \mapsto \lambda P$. Then $\|L_P\| = 1$ (Exercise 2.5).

As was the case for discrete state spaces, a probability measure π is invariant for a transition probability kernel if and only if $\pi = \pi P$. This is an integral equation

$$\pi(B) = \int \pi(dx) P(x, B), \qquad B \in \mathcal{B}$$

but we do not usually attempt to find a P that satisfies this equation by direct means. Usually we exploit some trick (if this is mysterious, it will all become clear in the next chapter).

Function Spaces

Before we can define the analog to right matrix multiplication, we must decide what space the linear operator $f \mapsto Pf$ is to act upon. There are a number of possibilities. The ones we will consider are the so-called $L^p(\pi)$ spaces, where $1 \le p \le \infty$ and π is a probability measure.

The $L^p(\pi)$ norm of a real-valued measurable function f on the probability space $(S, \mathcal{B}, \pi)$ is defined by

$$\|f\|_p = \left(\int |f(x)|^p \pi(dx) \right)^{1/p}$$

when $1 \le p < \infty$. The vector space $L^p(\pi)$ is the set of all measurable functions f on $(S, \mathcal{B})$ such that $\|f\|_p < \infty$. It is easy to see that the $L^p(\pi)$ norm satisfies axiom (b) for norms. That it satisfies axiom (a) is a well-known inequality called *Minkowski's inequality* (Rudin 1987, Theorem 3.5). It is also easy to see that the $L^p(\pi)$ norm fails to satisfy axiom (c), since $\|f\|_p = 0$ only implies $\pi\{|f(X)| > 0\} = 0$. If S is not discrete, there must be nonempty sets of probability zero, and any function f that is zero except on a set of probability zero has $\|f\|_p = 0$.

In order to make $L^p(\pi)$ a normed vector space, we need to work around this problem by redefining equality in $L^p(\pi)$ to mean equal except on a set of probability zero. Then axiom (c) is satisfied too, and $L^p(\pi)$ is a legitimate normed vector space.

We also redefine what we mean by inequalities as well. The statement $f \le g$ only means $f(x) \le g(x)$ except on a set of probability zero, and similarly for the other inequality relations. The space $L^\infty(\pi)$ consists of the bounded elements of $L^p(\pi)$, that is $|f| \le c$ for some real number c. Following the conventions for L^p spaces, this only means $|f(x)| \le c$ except on a set of probability zero. The $L^\infty(\pi)$ norm is the smallest c that will work

$$\|f\|_\infty = \inf\{\, c > 0 : \pi\{|f(X)| > c\} = 0 \,\}$$

This is also now easily seen to satisfy the axioms for norms, axiom (c) holding because we consider $f = 0$ if it is zero except on a set of probability zero. Thus all the $L^p(\pi)$ spaces for $1 \le p \le \infty$ are normed vector spaces[3].

An useful fact about $L^p(\pi)$ spaces is that $1 \le p \le q \le \infty$ implies $L^p(\pi) \supset L^q(\pi)$ (Exercise 2.12). (Warning: this uses the fact that π is a bounded measure. It is not true otherwise. However, we will be interested only in the case where π is a probability measure.)

Right Multiplication

We are finally ready to define "multiplication" of a kernel on the right by a function. If f is any nonnegative measurable function on $(S, \mathcal{B})$,

$$Pf(x) = \int P(x, dy) f(y) \tag{2.15}$$

is well-defined, though possibly $+\infty$. So we have no trouble defining "right multiplication" for nonnegative functions.

General functions are a bit more tricky. The issue is whether we can even define Pf for f that are both positive and negative. The trouble is that we want f to be integrable with respect to an infinite collection of probability measures, $P(x, \cdot)$, $x \in S$.

It turns out that we get everything we need, if π is an invariant probability measure for a transition probability kernel P and we use integrability with respect to π as our criterion. For $f \in L^1(\pi)$, define

$$g(x) = \int P(x, dy) |f(y)|.$$

Then

$$\begin{aligned} \int \pi(dx) g(x) &= \iint \pi(dx) P(x, dy) |f(y)| \\ &= \int \pi(dx) |f(y)| \\ &= \|f\|_1 \end{aligned} \tag{2.16}$$

[3]Actually they are Banach spaces, a *Banach space* being a complete normed vector space, where *complete* means every Cauchy sequence converges. But that will not play any role in the theory used in this course.

because $\pi = \pi P$. The interchange of the order of integration going from line 2 to line 3 is the conditional Fubini theorem (Fristedt and Gray 1997, Theorem 2 of Chapter 22). Hence the set

$$B = \{\, x \in S : g(x) < \infty \,\}.$$

satisfies $\pi(B^c) = 0$, because if g were infinite on a set of positive probability, the integral (2.16) would be infinite. This means we can define $Pf(x)$ by (2.15) for $x \in B$ and arbitrarily (say $Pf(x) = 0$) for $x \in B^c$ and have a function well defined in the $L^p(\pi)$ sense. Since $L^p(\pi) \subset L^1(\pi)$ for any $p > 1$, this makes the map $f \mapsto Pf$ well-defined on $L^p(\pi)$ for $1 \le p \le \infty$.

Now we want to show that the linear transformation $R_P : f \mapsto Pf$ actually maps $L^p(\pi)$ into $L^p(\pi)$. For $x \in B$ and $1 \le p < \infty$, Jensen's inequality gives

$$\begin{aligned} |Pf(x)|^p &= \left| \int P(x, dy) f(y) \right|^p \\ &\le \int P(x, dy) |f(y)|^p \end{aligned}$$

When we integrate both sides with respect to π, the fact that the left hand side is not defined for $x \in B^c$ does not matter because $\pi(B^c) = 0$. Hence

$$\begin{aligned} \|Pf\|_p^p &= \int \pi(dx) |Pf(x)|^p \\ &\le \iint \pi(dx) P(x, dy) |f(y)|^p \\ &= \int \pi(dy) |f(y)|^p \\ &= \|f\|_p^p \end{aligned}$$

Again $\pi = \pi P$ and the conditional Fubini theorem were used in going from line 2 to line 3.

The case $p = \infty$ is even simpler, for $x \in B$

$$\begin{aligned} |Pf(x)| &= \left| \int P(x, dy) f(y) \right| \\ &\le \int P(x, dy) |f(y)| \\ &\le \|f\|_\infty \int P(x, dy) \\ &= \|f\|_\infty \end{aligned}$$

Integrating with respect to π gives $\|Pf\|_\infty \le \|f\|_\infty$.

Thus we see that for $1 \le p \le \infty$ the linear transformation $R_P : f \mapsto Pf$ maps $L^p(\pi)$ into $L^p(\pi)$ and the corresponding operator norm satisfies

$$\|R_P\|_p = \sup_{\substack{f \in L^p(\pi) \\ f \neq 0}} \frac{\|R_P f\|_p}{\|f\|_p} \le 1. \tag{2.17}$$

In fact $\|R_P\|_p = 1$ because for $f \equiv 1$,

$$Pf(x) = \int P(x, dy) = 1 = f(x)$$

so $\|Pf\|_p = \|f\|_p$ for constant functions and the supremum in (2.17) is actually equal to one.

This has been an important section, so we summarize our results. If f is a measurable function from the state space to $[0, \infty]$, then $Pf(x)$ is well defined, though it may have the value $+\infty$. Since the set of functions on which this operation is defined is not a vector space, we cannot call P a linear operator here, but this notion is useful in various places in the theory of Markov chains.

If a kernel P has an invariant distribution π and $f \in L^p(\pi)$ for some $p \geq 1$, then Pf is a well defined element of $L^p(\pi)$. The linear operator $R_P : f \mapsto Pf$ is a bounded operator on $L^p(\pi)$ having operator norm equal to one.

General Kernels

In discrete state spaces, we wanted to discuss matrices that were not necessarily Markov. We need the analogous definitions for kernels. If $(S, \mathcal{B})$ is a measurable space, then a map K from $S \times \mathcal{B}$ to $\mathbb{R}$ is a *kernel* if

- for each fixed x the function $B \mapsto K(x, B)$ is a real signed measure, and
- for each fixed B the function $x \mapsto K(x, B)$ is a measurable function.

Multiplication of Kernels

The operation on kernels that is analogous to matrix multiplication is defined by

$$(K_1 K_2)(x, A) = \int K_1(x, dy) K_2(y, A).$$

Kernel multiplication is associative,

$$(K_1 K_2) K_3 = K_1 (K_2 K_3) \tag{2.18}$$

for any kernels K_1, K_2, and K_3, by the conditional Fubini theorem (Fristedt and Gray 1997, Theorem 2 of Chapter 22).

Kernel multiplication is not, in general, commutative: $K_1 K_2 = K_2 K_1$ may be false.

All of the results for composition and mixing of transition operators that we described in the discrete case carry over unchanged to the general case. In particular, multiplication of kernels corresponds to composition of operators (also called "fixed scan") in just the same way as we saw in (2.8a) and (2.8b). And a convex combination of Markov operators again produces a Markov operator and still corresponds to the operation of choosing an update mechanism at random and applying it (also called "random scan").

The Identity Kernel

The identity element any of the kernel operations is indeed the identity kernel defined back in (2.6). The identity kernel has connections with other notations widely used in probability. For fixed x, the measure $I(x, \cdot)$ is the probability measure concentrated at x, sometimes written δ_x, sometimes called the *Dirac measure*. For fixed A, the function $I(\cdot, A)$ is the indicator of the set A, more commonly written 1_A.

The identity kernel is the identity for kernel multiplication because

$$(IK)(x, A) = \int I(x, dy)K(y, A) = \int \delta_x(dy)K(y, A) = K(x, A),$$

and

$$(KI)(x, A) = \int K(x, dy)I(y, A) = \int K(x, dy)1_A(y) = \int_A K(x, dy) = K(x, A).$$

For this reason, we define $K^0 = I$ for any kernel K. Then the so-called Chapman-Kolmogorov equation

$$K^n = K^m K^{n-m}$$

holds whenever $0 \le m \le n$ as a direct consequence of the associative law (2.18).

The identity kernel is the identity for left multiplication of a kernel by a signed measure because

$$(\lambda I)(A) = \int \lambda(dx)I(x, A) = \int \lambda(dx)1_A(x) = \int_A \lambda(dx) = \lambda(A)$$

It is the identity for right multiplication of a kernel by a function because

$$(If)(x) = \int I(x, dy)f(y) = \int \delta_x(dy)f(y) = f(x).$$

Needless to say, the operators $L_P : \lambda \mapsto \lambda P$ and $R_P : f \mapsto Pf$ are the identity operators on the relevant vector spaces when P is the identity kernel.

The identity kernel is Markov, because, as we have seen $I(x, \cdot)$ is a probability measure, δ_x, for each x. If $X_n \sim \delta_x$, then $X_{n+1} \sim \delta_x$, because $\delta_x I = \delta_x$. Hence the chain never moves. Thus the identity kernel is the transition probability for the "maximally uninteresting chain" described in Example 1.4.

2.2.3 Hilbert Space Theory

Inner Product Spaces

An *inner product* on a complex vector space V is a map from $V \times V$ to $\mathbb{C}$, the value for the ordered pair of vectors x and y being written (x, y), that satisfies the following axioms (Halmos 1958, p. 121)

(a) $(x, y) = \overline{(y, x)}$,

(b) $(ax + by, z) = a(x, z) + b(y, z)$, for $a, b \in \mathbb{C}$,

(c) $(x, x) \geq 0$, and

(d) $(x, x) = 0$ implies $x = 0$.

where the overline in (a) denotes complex conjugation. An *inner product space* is a vector space equipped with an inner product.

For the most part, we will only be interested in real inner product spaces, in which case the complex conjugation in (a) does nothing and the scalars in (b) must be real. Since in applications we have no complex numbers, why should the theory involve them? The answer is eigenvalues and eigenvectors. Transition probability matrices are nonsymmetric and hence may have complex eigenvalues even though all their entries are real. So we will not be able to avoid mentioning complex inner product spaces. However, we will see they play a very minor role in Markov chain theory.

An inner product space is also a normed vector space with the norm defined by $\|x\| = \sqrt{(x, x)}$. It is easily verified that the norm axioms are implied by the inner product axioms (Exercise 2.6), the only bit of the proof that is nontrivial being the triangle inequality, which follows directly from

$$|(x, y)| \leq \|x\| \cdot \|y\|,$$

which is known to statisticians as the *Cauchy-Schwarz* inequality. It, of course, is proved exactly the same way as one proves that correlations are between -1 and 1.

Hilbert Spaces

A *Hilbert space* is a complete inner product space, where *complete* means every Cauchy sequence converges, a sequence $\{x_n\}$ being *Cauchy* if $\|x_m - x_n\| \to 0$ as $\min(m, n) \to \infty$. We will not develop any of the consequences of this definition, since they are well beyond the level of real analysis taken by most statistics graduate students, but we will steal a few results here and there from Hilbert space theory, explaining what they mean but blithely ignoring proofs.

One important fact about Hilbert space theory is the existence of the adjoint of an operator, which is analogous to the transpose of a matrix. If T is a bounded operator on a Hilbert space H. Then there is a unique bounded operator T^* on H that satisfies

$$(x, Ty) = (T^*x, y), \qquad x, y \in H$$

(Rudin 1991, Section 12.9). T^* is called the *adjoint* of T. If $T^* = T$, then T is said to be *self-adjoint*.

To see the connection between adjoints and transposes, equip the vector space $\mathbb{R}^S$ for some finite set S with the usual inner product

$$(f, g) = \sum_{x \in S} f(x)g(x). \tag{2.19}$$

A linear operator on $\mathbb{R}^S$ is represented by a matrix $M(x, y)$, the linear operator being $T_M : f \mapsto Mf$ (the same as the right multiplication we studied in Section 2.1.1 but with M not necessarily a transition probability matrix). Then

$$(f, T_M g) = \sum_{x \in S} \sum_{y \in S} f(x) M(x, y) g(y)$$

and

$$(T_M^* f, g) = \sum_{x \in S} \sum_{y \in S} g(x) M^*(x, y) f(y)$$

where M^* is the matrix that represents T_M^*. Clearly, M and M^* are transposes of each other.

For Markov chain theory, there are only two important Hilbert spaces. The first we have already met: $L^2(\pi)$ is a Hilbert space when the inner product is defined by

$$(f, g) = \int f(x) \overline{g(x)} \pi(dx). \tag{2.20}$$

That this defines an inner product (with the usual proviso that equality means only equality with probability one) is obvious. The completeness comes from the fact that every $L^p(\pi)$ is a complete metric space (Rudin 1987, Theorem 3.11). Usually we consider $L^p(\pi)$ a real Hilbert space, in which case the complex conjugate in (2.20) does nothing.

The reason why $L^2(\pi)$ is so important is that (2.20) is $\operatorname{Cov}\{f(X), g(X)\}$ in the special case when both variables have mean zero. In order to cater to this special case of interest to statisticians, we introduce the subspace of $L^2(\pi)$ that consists of mean-zero functions

$$L_0^2(\pi) = \left\{ f \in L^2(\pi) : \int f(x) \pi(dx) = 0 \right\}$$

Another characterization of $L_0^2(\pi)$ uses the notion of orthogonality. Vectors x and y in a Hilbert space are *orthogonal* if $(x, y) = 0$. If 1 represents the constant function equal to 1 almost surely, then we can also write

$$L_0^2(\pi) = \left\{ f \in L^2(\pi) : (f, 1) = 0 \right\}$$

Thus $L_0^2(\pi)$ is the subspace of $L^2(\pi)$ orthogonal to the constant functions. Since the linear function $f \mapsto (f, 1)$ is continuous, $L_0^2(\pi)$ is a topologically closed subspace of $L^2(\pi)$ and hence is also a Hilbert space.

Warning: The characterization of the adjoint as the transpose is incorrect for $L^2(\pi)$ even in the finite state space case. The reason is that (2.19) is not the inner product on $L^2(\pi)$. The inner product is defined by (2.20). The same formula applies to finite state spaces as for general state spaces (general includes finite). Exercise 2.9 derives the correct formula for the adjoint.

In the preceding section, we saw that the operator norm for the linear operator $f \mapsto Pf$ is exactly equal to one, no matter which $L^p(\pi)$ we have the operator act on. The Hilbert space $L^2(\pi)$ is no exception, but $L_0^2(\pi)$ is different. Reducing the domain of the operator cannot increase the norm, but may decrease it, the supremum in (2.17) being over a smaller set. The proof that the norm is exactly one no longer applies, because it used the fact that $Pf = f$ for constant functions f, and those functions are no longer in the domain. Thus when we consider $R_P : f \mapsto Pf$ an operator on $L_0^2(\pi)$ we have $\|R_P\|_2 \leq 1$ with strict inequality now a possibility.

2.2.4 Time-Reversed Markov Chains

The measure-theoretic construction of infinite sequences of random variables discussed in Section 2.1.3, says that specification of the probability distribution of an infinite sequence is equivalent to specifying a consistent set of finite-dimensional distributions. This allows us to specify a stationary Markov chain as a doubly infinite sequence $\ldots$, X_{-2}, X_{-1}, X_0, X_1, X_2, $\ldots$. Specifying the distribution of the doubly infinite sequence is the same as specifying the joint distribution of X_n, X_{n+1}, $\ldots$, X_{n+k} for any $k > 0$. Stationarity implies that this joint distribution does not depend on n.

Two questions naturally arise about the time-reversed sequence. First, is it Markov? Second, what is its kernel? That the time-reversed sequence has the Markov property is a trivial consequence of conditional independence being a symmetric property, that is, the following three statements are equivalent.

- The future is independent of the past given the present.
- The past is independent of the future given the present.
- The past and future are independent given the present.

If this isn't mathy enough for you, here are some equations. What is to be shown is that

$$\begin{aligned} E\{f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n\} & \\ = E\{f(X_{n+1}, X_{n+2}, \ldots) | X_n\} E\{g(X_{n-1}, X_{n-2}, \ldots) | X_n\} & \end{aligned} \tag{2.21}$$

for any functions f and g such that both sides are well defined. This says the σ-field generated by X_{n+1}, X_{n+2}, $\ldots$ (the future) and the σ-field generated by X_{n-1}, X_{n-2}, $\ldots$ (the past) are conditionally independent given the σ-field generated by X_n (the present) (Fristedt and Gray 1997, Definition 23 of Chapter 21).

The proof is

$$\begin{aligned}
&E\{f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n\} \\
&\quad = E\{E[f(X_{n+1}, X_{n+2}, \ldots) g(X_{n-1}, X_{n-2}, \ldots) | X_n, X_{n-1}, X_{n-2}, \ldots] | X_n\} \\
&\quad = E\{g(X_{n-1}, X_{n-2}, \ldots) E[f(X_{n+1}, X_{n+2}, \ldots) | X_n, X_{n-1}, X_{n-2}, \ldots] | X_n\} \\
&\qquad = E\{g(X_{n-1}, X_{n-2}, \ldots) E[f(X_{n+1}, X_{n+2}, \ldots) | X_n] | X_n\} \\
&\qquad\quad = E\{f(X_{n+1}, X_{n+2}, \ldots) | X_n\} E\{g(X_{n-1}, X_{n-2}, \ldots) | X_n\}
\end{aligned}$$

The equality between lines 3 and 4 is the Markov property of the original chain running forwards in time. The other equalities are standard properties of conditional expectation. The equalities between lines 2 and 3 and between lines 4 and 5 are the property that functions of the conditioning variables can be taken outside a conditional expectation (Fristedt and Gray 1997, Problem 27 of Chapter 23). The equality between lines 1 and 2 is the general iterated conditional expectation formula (Fristedt and Gray 1997, Proposition 6 of Chapter 23).

By Propositions 25 and 27 of Chapter 23 in Fristedt and Gray (1997) (2.21) implies the Markov property for the time-reversed chain

$$E\{1_A(X_{n-1}) | X_n, X_{n+1}, X_{n+2}, \ldots\} = E\{1_A(X_{n-1}) | X_n\}.$$

Clearly, the time-reversed chain is also stationary, in particular, it has stationary transition probabilities. As to whether these transition probabilities are representable by a kernel, the answer is not necessarily, but usually. The issue is whether there exists a kernel P^* satisfying

$$\int_A \pi(dx) P^*(x, B) = \int_B \pi(dx) P(x, A), \qquad A, B \in \mathcal{B}, \tag{2.22}$$

(where $\mathcal{B}$ is the σ-field of the state space), that is, whether P^* exists as a *regular* conditional probability. Conditional probabilities always exist, but *regular* ones do not. The key is whether the state space is "nice" enough. If the state space is a so-called *Borel space*, then regular conditional probabilities (a. k. a. kernels) exist (Fristedt and Gray 1997, Theorem 19 of Chapter 21). Euclidean spaces $\mathbb{R}^d$ are Borel spaces, as are most (all?) other state spaces that arise in practical examples. So we may take it for granted that P^* exists. It is not, however, uniquely defined. $P^*(x, \cdot)$ can be defined arbitrarily for x in a set of π-probability zero without effecting (2.22). Thus there are many kernels P^*, all of which give the same probability law for the time-reversed chain.

Now that we have a kernel P^* for the time-reversed chain, we know that P^* and the marginal distribution π of X_n, which is invariant for both P and P^*, determine the probability distribution of the infinite sequence. We can also look at P^* as an operator. In particular, (2.22) is equivalent to

$$\int \pi(dx) P^*(x, dy) f(x) g(y) = \int \pi(dx) P(x, dy) g(x) f(y), \qquad f, g \in L^2(\pi) \tag{2.23}$$

by linearity of expectation and monotone convergence. In Hilbert space notation (2.23) is

$$(f, P^* g) = (Pf, g)$$

so now we see why the choice of P^* for the kernel of the time-reversed chain. It is the adjoint operator on $L^2(\pi)$.

2.2.5 Reversibility

A stationary Markov chain is *reversible* (also called *time-reversible*) if the doubly infinite sequence has the same probability distribution when time is reversed. We also say a kernel P is reversible with respect to π if (2.22) holds with $P^* = P$, that is,

$$\int_A \pi(dx) P(x, B) = \int_B \pi(dx) P(x, A), \qquad A, B \in \mathcal{B}. \tag{2.24}$$

Taking the case where A is the whole state space in (2.24) gives

$$\int \pi(dx) P(x, B) = \int_B \pi(dx) = \pi(B), \qquad B \in \mathcal{B},$$

which says $\pi P = \pi$. Thus (2.24) implies that π is invariant for P.

This is a very important principle.

> *If P is reversible with respect to π, then P preserves π.*

This will turn out to be our main method for accomplishing the "first task" of MCMC. Given a distribution π, how do we find Markov update mechanisms that preserve π? Answer: show they are reversible with respect to π.

If (2.24) holds, then so does (2.23) with $P^* = P$, that is,

$$\iint f(x) g(y) \pi(dx) P(x, dy) = \iint g(x) f(y) \pi(dx) P(x, dy), \qquad f, g \in L^2(\pi). \tag{2.25}$$

Hence P is self-adjoint.

> *P is reversible with respect to π, if and only if P is a self-adjoint operator on $L^2(\pi)$.*

We can rewrite (2.24) as

$$\Pr(X_n \in A \,\&\, X_{n+1} \in B) = \Pr(X_n \in B \,\&\, X_{n+1} \in A) \tag{2.26}$$

This gives yet another slogan.

> *A stationary Markov chain is reversible, if and only if X_n and X_{n+1} are exchangeable.*

For a discrete state space, transition probability matrix P and invariant distribution π, and state space S, the reversibility property is

$$\Pr(X_n = x \,\&\, X_{n+1} = y) = \Pr(X_n = y \,\&\, X_{t+1} = x),$$

or stated in terms of π and P

$$\pi(x)P(x, y) = \pi(y)P(y, x), \qquad x, y \in S, \tag{2.27}$$

a condition that is referred to as *detailed balance*. Our main tool for establishing that a particular transition probability P has a specified invariant distribution π will be verification of the detailed balance condition (2.27) and its counterparts for general state spaces. This is generally much easier than verifying $\pi P = \pi$ directly.

The analogue of (2.27) for general state spaces (2.26) involves probabilities of sets rather than points, and so does not lead to an analog of the detailed balance condition. You will sometimes see

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx)$$

called "detailed balance for general state spaces," but strictly speaking this is merely a shorthand for (2.24) or (2.25).

Exercises

2.1. Find an invariant distribution and show that it is unique for

(a) The random walk with reflecting barriers, Example 2.1.

(b) The modification of random walk with reflecting barriers, so that the first row of the transition probability matrix is $0, 1, 0, \ldots$ and the last row is modified similarly to $\ldots, 0, 1, 0$, the rest of the rows remaining as in (2.3).

2.2.

(a) Show that a linear combination of Markov transition operators is Markov if and only if the linear combination is an affine combination.

(b) Provide a counterexample that shows an affine combination of Markov transition operators that is not a convex combination but is still Markov.

2.3. Show that total variation norm satisfies the norm axioms.

2.4. Show that the map $L_P : \lambda \mapsto \lambda P$ is a linear operator on $\mathcal{M}(S)$ when P is a Markov kernel. There are two things to show, first that L_P is a linear transformation

$$L_P(a\lambda + b\mu) = aL_P(\lambda) + bL_P(\mu), \qquad a, b \in \mathbb{R},\ \lambda, \mu \in \mathcal{M}(S),$$

and second that L_P maps $\mathcal{M}(S)$ to $\mathcal{M}(S)$ (that is, λP is a countably additive set function).

2.5. Show that the map $L_P : \lambda \mapsto \lambda P$ satisfies $\|L_P\| = 1$ when P is a Markov kernel.

2.6. Show that $\|x\| = \sqrt{(x, x)}$ defines a norm, when (x, y) is an inner product. Include a proof of the Cauchy-Schwarz inequality for inner product spaces.

2.7. Show that the stationary scalar-valued AR(1) time series discussed in Examples 1.2 and 1.5 is reversible.

2.8.

(a) Show that the random walk with reflecting barriers of Example 2.1 is reversible.

(b) Show that the modified random walk of Problem 2.1 (b) is reversible.

(c) Show that the "maximally uninteresting chain" having the identity kernel as its kernel is reversible for any invariant distribution π.

2.9. Suppose P is a transition probability matrix on a finite state space S having invariant distribution π considered as a vector $\pi \in \mathbb{R}^S$. Find the formula for the adjoint of $R_P : f \to Pf$ considered as an operator on $L^2(\pi)$.

2.10. Find a Markov chain transition probability kernel that is *not* reversible.

2.11. Show that the Gibbs update described in Section 1.7 is reversible.

2.12. If π is a probability measure, show that $1 \le p \le q \le \infty$ implies $L^p(\pi) \supset L^q(\pi)$.