Chapter 3

FINITE-STATE MARKOV CHAINS

3.1 Introduction

The counting processes $\{N(t);\, t > 0\}$ described in Section 2.1.1 have the property that $N(t)$ *changes* at discrete instants of time, but is *defined* for all real $t > 0$. The Markov chains to be discussed in this chapter are stochastic processes *defined* only at integer values of time, $n = 0, 1, \ldots$. At each integer time $n \geq 0$, there is an integer-valued random variable (rv) X_n, called the *state* at time n, and the process is the family of rv's $\{X_n;\, n \geq 0\}$. We refer to these processes as *integer-time processes*. An integer-time process $\{X_n;\, n \geq 0\}$ can also be viewed as a process $\{X(t);\, t \geq 0\}$ defined for all real t by taking $X(t) = X_n$ for $n \leq t < n + 1$, but since changes occur only at integer times, it is usually simpler to view the process only at those integer times.

In general, for Markov chains, the set of possible values for each rv X_n is a countable set $\mathcal{S}$. If $\mathcal{S}$ is countably infinite, it is usually taken to be $\mathcal{S} = \{0, 1, 2, \ldots\}$, whereas if $\mathcal{S}$ is finite, it is usually taken to be $\mathcal{S} = \{1, \ldots, \mathsf{M}\}$. In this chapter (except for Theorems 3.2.2 and 3.2.3), we restrict attention to the case in which $\mathcal{S}$ is finite, *i.e.*, processes whose sample functions are sequences of integers, each between 1 and M. There is no special significance to using integer labels for states, and no compelling reason to include 0 for the countably infinite case and not for the finite case. For the countably infinite case, the most common applications come from queueing theory, where the state often represents the number of waiting customers, which might be zero. For the finite case, we often use vectors and matrices, where positive integer labels simplify the notation. In some examples, it will be more convenient to use more illustrative labels for states.

Definition 3.1.1. *A Markov chain is an integer-time process, $\{X_n, n \geq 0\}$ for which the sample values for each rv X_n, $n \geq 1$, lie in a countable set $\mathcal{S}$ and depend on the past only through the most recent rv X_{n-1}. More specifically, for all positive integers n, and for all $i, j, k, \ldots, m$ in $\mathcal{S}$,*

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i, X_{n-2}{=}k, \ldots, X_0{=}m\} = \Pr\{X_n{=}j \mid X_{n-1}{=}i\}\,. \tag{3.1}$$

Furthermore, $\Pr\{X_n{=}j \mid X_{n-1}{=}i\}$ *depends only on* i *and* j *(not* n*) and is denoted by*

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i\} = P_{ij}. \tag{3.2}$$

The initial state X_0 *has an arbitrary probability distribution. A finite-state Markov chain is a Markov chain in which* $\mathcal{S}$ *is finite.*

Equations such as (3.1) are often easier to read if they are abbreviated as

$$\Pr\{X_n \mid X_{n-1}, X_{n-2}, \ldots, X_0\} = \Pr\{X_n \mid X_{n-1}\}.$$

This abbreviation means that equality holds for all sample values of each of the rv's. *i.e.*, it means the same thing as (3.1).

The rv X_n is called the state of the chain at time n. The possible values for the state at time n, namely $\{1, \ldots, \mathsf{M}\}$ or $\{0, 1, \ldots\}$ are also generally called states, usually without too much confusion. Thus P_{ij} is the probability of going to state j given that the previous state is i; the new state, given the previous state, is independent of all earlier states. The use of the word *state* here conforms to the usual idea of the state of a system — the state at a given time summarizes everything about the past that is relevant to the future.

Definition 3.1.1 is used by some people as the definition of a *homogeneous Markov chain*. For them, Markov chains include more general cases where the transition probabilities can vary with n. Thus they replace (3.1) and (3.2) by

$$\Pr\{X_n{=}j \mid X_{n-1}{=}i, X_{n-2}{=}k, \ldots, X_0{=}m\} = \Pr\{X_n{=}j \mid X_{n-1}{=}i\} = P_{ij}(n). \tag{3.3}$$

We will call a process that obeys (3.3), with a dependence on n, a *non-homogeneous Markov chain*. We will discuss only the homogeneous case, with no dependence on n, and thus restrict the definition to that case. Not much of general interest can be said about non-homogeneous chains.[1]

An initial probability distribution for X_0, combined with the transition probabilities $\{P_{ij}\}$ (or $\{P_{ij}(n)\}$ for the non-homogeneous case), define the probabilities for all events in the Markov chain.

Markov chains can be used to model an enormous variety of physical phenomena and can be used to approximate many other kinds of stochastic processes such as the following example:

Example 3.1.1. Consider an integer process $\{Z_n;\ n \geq 0\}$ where the Z_n are finite integer-valued rv's as in a Markov chain, but each Z_n depends probabilistically on the previous k rv's, $Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k}$. In other words, using abbreviated notation,

$$\Pr\{Z_n \mid Z_{n-1}, Z_{n-2}, \ldots, Z_0\} = \Pr\{Z_n \mid Z_{n-1}, \ldots Z_{n-k}\}. \tag{3.4}$$

[1] On the other hand, we frequently find situations where a small set of rv's, say W, X, Y, Z satisfy the *Markov condition* that $\Pr\{Z \mid Y, X, W\} = \Pr\{Z \mid Y\}$ and $\Pr\{Y \mid X, W\} = \Pr\{Y \mid X\}$ but where the conditional distributions $\Pr\{Z \mid Y\}$ and $\Pr\{Y \mid X\}$ are unrelated. In other words, *Markov chains* imply homogeneity here, whereas the *Markov condition* does not.

We now show how to view the condition on the right side of (3.4), *i.e.*, $(Z_{n-1}, Z_{n-2}, \ldots, Z_{n-k})$ as the state of the process at time $n-1$. We can rewrite (3.4) as

$$\Pr\{Z_n, Z_{n-1}, \ldots, Z_{n-k+1} \mid Z_{n-1}, \ldots, Z_0\} = \Pr\{Z_n, \ldots, Z_{n-k+1} \mid Z_{n-1}, \ldots Z_{n-k}\},$$

since, for each side of the equation, any given set of values for $Z_{n-1}, \ldots, Z_{n-k+1}$ on the right side of the conditioning sign specifies those values on the left side. Thus if we define $X_{n-1} = (Z_{n-1}, \ldots, Z_{n-k})$ for each n, this simplifies to

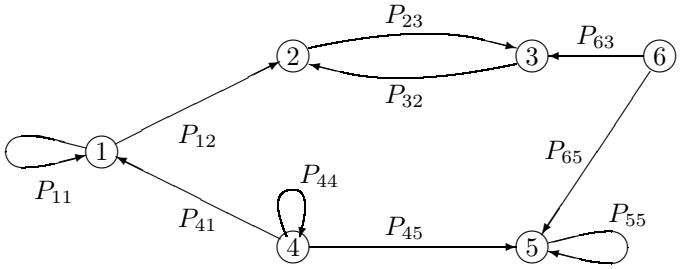
$$\Pr\{X_n \mid X_{n-1}, \ldots, X_{k-1}\} = \Pr\{X_n \mid X_{n-1}\}.$$

We see that by expanding the state space to include k-tuples of the rv's Z_n, we have converted the k dependence over time to a unit dependence over time, *i.e.*, a Markov process is defined using the expanded state space.

Note that in this new Markov chain, the initial state is $X_{k-1} = (Z_{k-1}, \ldots, Z_0)$, so one might want to shift the time axis to start with X_0.

Markov chains are often described by a directed graph (see Figure 3.1 a). In this graphical representation, there is one node for each state and a directed arc for each non-zero transition probability. If$P_{ij} = 0$, then the arc from node i to node j is omitted, so the difference between zero and non-zero transition probabilities stands out clearly in the graph. The classification of states, as discussed in Section 3.2, is determined by the set of transitions with non-zero probabilities, and thus the graphical representation is ideal for that topic.

A finite-state Markov chain is also often described by a matrix $[P]$ (see Figure 3.1 b). If the chain has M states, then $[P]$ is an M by M matrix with elements P_{ij}. The matrix representation is ideally suited for studying algebraic and computational issues.



$$[P] = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{16} \\ P_{21} & P_{22} & \cdots & P_{26} \\ \vdots & \vdots & \vdots\vdots & \vdots \\ P_{61} & P_{62} & \cdots & P_{66} \end{bmatrix}$$

a) Graphical **b)** Matrix

Figure 3.1: Graphical and Matrix Representation of a 6 state Markov Chain; a directed arc from i to j is included in the graph if and only if (iff) $P_{ij} > 0$.

3.2 Classification of states

This section, except where indicated otherwise, applies to Markov chains with both finite and countable state spaces. We start with several definitions.

Definition 3.2.1. *An (n-step) walk is an ordered string of nodes, $(i_0, i_1, \dots i_n)$, $n \geq 1$, in which there is a directed arc from i_{m-1} to i_m for each m, $1 \leq m \leq n$. A path is a walk in which no nodes are repeated. A cycle is a walk in which the first and last nodes are the same and no other node is repeated.*

Note that a walk can start and end on the same node, whereas a path cannot. Also the number of steps in a walk can be arbitrarily large, whereas a path can have at most $\mathsf{M} - 1$ steps and a cycle at most M steps for a finite-state Markov chain with $|\mathcal{S}| = \mathsf{M}$.

Definition 3.2.2. *A state j is* accessible *from i (abbreviated as $i \to j$) if there is a walk in the graph from i to j.*

For example, in Figure 3.1(a), there is a walk from node 1 to node 3 (passing through node 2), so state 3 is accessible from 1. There is no walk from node 5 to 3, so state 3 is not accessible from 5. State 2 is accessible from itself, but state 6 is not accessible from itself. To see the probabilistic meaning of accessibility, suppose that a walk $i_0, i_1, \dots i_n$ exists from node i_0 to i_n. Then, conditional on $X_0 = i_0$, there is a positive probability, $P_{i_0 i_1}$, that $X_1 = i_1$, and consequently (since $P_{i_1 i_2} > 0$), there is a positive probability that $X_2 = i_2$. Continuing this argument, there is a positive probability that $X_n = i_n$, so that $\Pr\{X_n{=}i_n \mid X_0{=}i_0\} > 0$. Similarly, if $\Pr\{X_n{=}i_n \mid X_0{=}i_0\} > 0$, then an n-step walk from i_0 to i_n must exist. Summarizing, $i \to j$ if and only if (iff) $\Pr\{X_n{=}j \mid X_0{=}i\} > 0$ for some $n \geq 1$. We denote $\Pr\{X_n{=}j \mid X_0{=}i\}$ by P_{ij}^n. Thus, for $n \geq 1$, $P_{ij}^n > 0$ if and only if the graph has an n step walk from i to j (perhaps visiting the same node more than once). For the example in Figure 3.1(a), $P_{13}^2 = P_{12}P_{23} > 0$. On the other hand, $P_{53}^n = 0$ for all $n \geq 1$. An important relation that we use often in what follows is that if there is an n-step walk from state i to j and an m-step walk from state j to k, then there is a walk of $m + n$ steps from i to k. Thus

$$P_{ij}^n > 0 \text{ and } \mathrm{P_{jk}^m} > 0 \quad \text{imply} \quad \mathrm{P_{ik}^{n+m}} > 0. \tag{3.5}$$

This also shows that

$$i \to j \text{ and } \mathrm{j} \to \mathrm{k} \quad \text{imply} \quad \mathrm{i} \to \mathrm{k}. \tag{3.6}$$

Definition 3.2.3. *Two distinct states i and j* communicate *(abbreviated $i \leftrightarrow j$) if i is accessible from j and j is accessible from i.*

An important fact about communicating states is that if $i \leftrightarrow j$ and $m \leftrightarrow j$ then $i \leftrightarrow m$. To see this, note that $i \leftrightarrow j$ and $m \leftrightarrow j$ imply that $i \to j$ and $j \to m$, so that $i \to m$. Similarly, $m \to i$, so $i \leftrightarrow m$.

Definition 3.2.4. *A* class *$\mathcal{C}$ of states is a non-empty set of states such that each $i \in \mathcal{C}$ communicates with every other state $j \in \mathcal{C}$ and communicates with no $j \notin \mathcal{C}$.*

For the example of Figure 3.1(a), $\{2, 3\}$ is one class of states, $\{1\}$, $\{4\}$, $\{5\}$, and $\{6\}$ are the other classes. Note that state 6 does not communicate with any other state, and is not even accessible from itself, but the set consisting of $\{6\}$ alone is still a class. The entire set of states in a given Markov chain is partitioned into one or more disjoint classes in this way.

Definition 3.2.5. *For finite-state Markov chains, a* recurrent state *is a state i that is accessible from all states that are accessible from i (i is recurrent if $i \to j$ implies that $j \to i$). A* transient state *is a state that is not recurrent.*

Recurrent and transient states for Markov chains with a countably-infinite state space will be defined in Chapter 5.

According to the definition, a state i in a finite-state Markov chain is recurrent if there is no possibility of going to a state j from which there can be no return. As we shall see later, if a Markov chain ever enters a recurrent state, it returns to that state eventually with probability 1, and thus keeps returning infinitely often (in fact, this property serves as the definition of recurrence for Markov chains without the finite-state restriction). A state i is transient if there is some j that is accessible from i but from which there is no possible return. Each time the system returns to i, there is a possibility of going to j; eventually this possibility will occur with no further returns to i.

Theorem 3.2.1. *For finite-state Markov chains, either all states in a class are transient or all are recurrent.*[2]

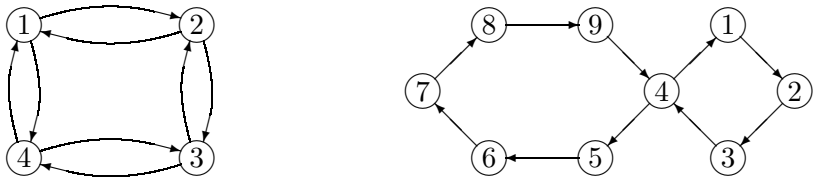
Proof: Assume that state i is transient (i.e., for some j, $i \to j$ but $j \not\to i$) and suppose that i and m are in the same class (i.e., $i \leftrightarrow m$). Then $m \to i$ and $i \to j$, so $m \to j$. Now if $j \to m$, then the walk from j to m could be extended to i; this is a contradiction, and therefore there is no walk from j to m, and m is transient. Since we have just shown that all nodes in a class are transient if any are, it follows that the states in a class are either all recurrent or all transient. □

For the example of Figure 3.1(a), $\{2, 3\}$ and $\{5\}$ are recurrent classes and the other classes are transient. In terms of the graph of a Markov chain, a class is transient if there are any directed arcs going from a node in the class to a node outside the class. Every finite-state Markov chain must have at least one recurrent class of states (see Exercise 3.2), and can have arbitrarily many additional classes of recurrent states and transient states.

States can also be classified according to their periods (see Figure 3.2). For $X_0 = 2$ in Figure 3.2(a), X_n must be 2 or 4 for n even and 1 or 3 for n odd. On the other hand, if X_0 is 1 or 3, then X_n is 2 or 4 for n odd and 1 or 3 for n even. Thus the effect of the starting state never dies out. Figure 3.2(b) illustrates another example in which the memory of the starting state never dies out. The states in both of these Markov chains are said to be periodic with period 2. Another example of periodic states are states 2 and 3 in Figure 3.1(a).

Definition 3.2.6. *The* period *of a state i, denoted $d(i)$, is the greatest common divisor (gcd) of those values of n for which $P_{ii}^n > 0$. If the period is 1, the state is aperiodic, and if the period is 2 or more, the state is periodic.*

[2] As shown in Chapter 5, this theorem is also true for Markov chains with a countably infinite state space, but the proof given here is inadequate. Also recurrent classes with a countably infinite state space are further classified into either *positive-recurrent* or *null-recurrent*, a distinction that does not appear in the finite-state case.



(a) (b)

Figure 3.2: Periodic Markov Chains

For example, in Figure 3.2(a), $P_{11}^n > 0$ for $n = 2, 4, 6, \ldots$. Thus $d(1)$, the period of state 1, is two. Similarly, $d(i) = 2$ for the other states in Figure 3.2(a). For Figure 3.2(b), we have $P_{11}^n > 0$ for $n = 4, 8, 10, 12, \ldots$; thus $d(1) = 2$, and it can be seen that $d(i) = 2$ for all the states. These examples suggest the following theorem.

Theorem 3.2.2. *For any Markov chain (with either a finite or countably infinite number of states), all states in the same class have the same period.*

Proof: Let i and j be any distinct pair of states in a class $\mathcal{C}$. Then $i \leftrightarrow j$ and there is some r such that $P_{ij}^r > 0$ and some s such that $P_{ji}^s > 0$. Since there is a walk of length $r + s$ going from i to j and back to i, $r + s$ must be divisible by $d(i)$. Let t be any integer such that $P_{jj}^t > 0$. Since there is a walk of length $r + t + s$ from i o j, then back to j, and then to i, $r + t + s$ is divisible by $d(i)$, and thus t is divisible by $d(i)$. Since this is true for any t such that $P_{jj}^t > 0$, $d(j)$ is divisible by $d(i)$. Reversing the roles of i and j, $d(i)$ is divisible by $d(j)$, so $d(i) = d(j)$. □

Since the states in a class $\mathcal{C}$ all have the same period and are either all recurrent or all transient, we refer to $\mathcal{C}$ itself as having the period of its states and as being recurrent or transient. Similarly if a Markov chain has a single class of states, we refer to the chain as having the corresponding period.

Theorem 3.2.3. *If a recurrent class $\mathcal{C}$ in a finite-state Markov chain has period d, then the states in $\mathcal{C}$ can be partitioned into d subsets, $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_d$, in such a way that all transitions from $\mathcal{S}_1$ go to $\mathcal{S}_2$, all from $\mathcal{S}_2$ go to $\mathcal{S}_3$, and so forth up to $\mathcal{S}_{m-1}$ to $\mathcal{S}_m$. Finally, all transitions from $\mathcal{S}_m$ go to $\mathcal{S}_1$.*

Proof: See Figure 3.3 for an illustration of the theorem. For a given state in $\mathcal{C}$, say state 1, define the sets $\mathcal{S}_1, \ldots, \mathcal{S}_d$ by

$$\mathcal{S}_m = \{j : P_{1j}^{nd+m} > 0 \text{ for some } n \geq 0\}; \quad 1 \leq m \leq d. \tag{3.7}$$

For each $j \in \mathcal{C}$, we first show that there is one and only one value of m such that $j \in \mathcal{S}_m$. Since $1 \leftrightarrow j$, there is some r for which $P_{1j}^r > 0$ and some s for which $P_{j1}^s > 0$. Thus there is a walk from 1 to 1 (through j) of length $r + s$, so $r + s$ is divisible by d. For the given r,

let m, $1 \leq m \leq d$, satisfy $r = m + nd$, where n is an integer. From (3.7), $j \in \mathcal{S}_m$. Now let r' be any other integer such that $P_{1j}^{r'} > 0$. Then $r' + s$ is also divisible by d, so that $r' - r$ is divisible by d. Thus $r' = m + n'd$ for some integer n' and that same m. Since r' is *any* integer such that $P_{1j}^{r'} > 0$, j is in $\mathcal{S}_m$ for only that one value of m. Since j is arbitrary, this shows that the sets $\mathcal{S}_m$ are disjoint and partition $\mathcal{C}$.

Finally, suppose $j \in \mathcal{S}_m$ and $P_{jk} > 0$. Given a walk of length $r = nd + m$ from state 1 to j, there is a walk of length $nd + m + 1$ from state 1 to k. It follows that if $m < d$, then $k \in \mathcal{S}_{m+1}$ and if $m = d$, then $k \in \mathcal{S}_1$, completing the proof. □

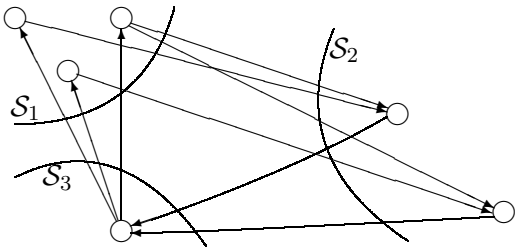


Figure 3.3: Structure of a periodic Markov chain with $d = 3$. Note that transitions only go from one subset $\mathcal{S}_m$ to the next subset $\mathcal{S}_{m+1}$ (or from $\mathcal{S}_d$ to $\mathcal{S}_1$).

We have seen that each class of states (for a finite-state chain) can be classified both in terms of its period and in terms of whether or not it is recurrent. The most important case is that in which a class is both recurrent and aperiodic.

Definition 3.2.7. *For a finite-state Markov chain, an ergodic class of states is a class that is both recurrent and aperiodic*[3]*. A Markov chain consisting entirely of one ergodic class is called an ergodic chain.*

We shall see later that these chains have the desirable property that P_{ij}^n becomes independent of the starting state i as $n \to \infty$. The next theorem establishes the first part of this by showing that $P_{ij}^n > 0$ for all i and j when n is sufficiently large. A guided proof is given in Exercise 3.5.

Theorem 3.2.4. *For an ergodic M state Markov chain, $P_{ij}^m > 0$ for all i, j, and all $m \geq (\mathsf{M}-1)^2 + 1$.*

Figure 3.4 illustrates a situation where the bound $(\mathsf{M}-1)^2 + 1$ is met with equality. Note that there is one cycle of length $\mathsf{M} - 1$ and the single node not on this cycle, node 1, is the unique starting node at which the bound is met with equality.

[3]For Markov chains with a countably infinite state space, ergodic means that the states are positive-recurrent and aperiodic (see Chapter 5, Section 5.1).

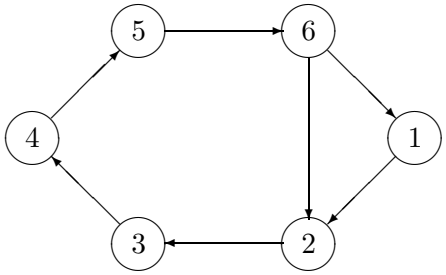


Figure 3.4: An ergodic chain with $\mathsf{M} = 6$ states in which $P_{ij}^m > 0$ for all $m > (\mathsf{M}-1)^2$ and all i, j but $P_{11}^{(\mathsf{M}-1)^2} = 0$ The figure also illustrates that an M state Markov chain must have a cycle with $\mathsf{M}-1$ or fewer nodes. To see this, note that an ergodic chain must have cycles, since each node must have a walk to itself, and subcycles of repeated nodes can be omitted from that walk, converting it into a cycle. Such a cycle might have M nodes, but a chain with only an M node cycle would be periodic. Thus some nodes must be on smaller cycles, such as the cycle of length 5 in the figure.

3.3 The matrix representation

The matrix $[P]$ of transition probabilities of a Markov chain is called a stochastic matrix; that is, a *stochastic matrix* is a square matrix of nonnegative terms in which the elements in each row sum to 1. We first consider the n step transition probabilities P_{ij}^n in terms of [P]. The probability, starting in state i, of going to state j in two steps is the sum over k of the probability of going first to k and then to j. Using the Markov condition in (3.1),

$$P_{ij}^2 = \sum_{k=1}^{\mathsf{M}} P_{ik} P_{kj}.$$

It can be seen that this is just the ij term of the product of the matrix $[P]$ with itself; denoting $[P][P]$ as $[P^2]$, this means that P_{ij}^2 is the (i,j) element of the matrix $[P^2]$. Similarly, P_{ij}^n is the ij element of the nth power of the matrix $[P]$. Since $[P^{m+n}] = [P^m][P^n]$, this means that

$$P_{ij}^{m+n} = \sum_{k=1}^{\mathsf{M}} P_{ik}^m \, P_{kj}^n. \tag{3.8}$$

This is known as the *Chapman-Kolmogorov* equation. An efficient approach to compute $[P^n]$ (and thus P_{ij}^n) for large n, is to multiply $[P]^2$ by $[P]^2$, then $[P]^4$ by $[P]^4$ and so forth. Then $[P], [P^2], [P^4], \ldots$ can be multiplied as needed to get $[P^n]$.

3.3.1 Steady state and $[P^n]$ for large n

The matrix $[P^n]$ (i.e., the matrix of transition probabilities raised to the nth power) is important for a number of reasons. The i, j element of this matrix is $P_{ij}^n = \Pr\{X_n{=}j \mid X_0{=}i\}$. If memory of the past dies out with increasing n, then we would expect the dependence of P_{ij}^n on both n and i to disappear as $n \to \infty$. This means, first, that $[P^n]$ should converge to a limit as $n \to \infty$, and, second, that for each column j, the elements in that column, $P_{1j}^n, P_{2j}^n, \ldots, P_{\mathsf{M}j}^n$ should all tend toward the same value, say π_j, as $n \to \infty$. If this type of convergence occurs, (and we later determine the circumstances under which it occurs), then $P_{ij}^n \to \pi_j$ and each row of the limiting matrix will be $(\pi_1, \ldots, \pi_{\mathsf{M}})$, *i.e.*, each row is the same as each other row.

If we now look at the equation $P_{ij}^{n+1} = \sum_k P_{ik}^n P_{kj}$, and assume the above type of convergence as $n \to \infty$, then the limiting equation becomes $\pi_j = \sum \pi_k P_{kj}$. In vector form, this equation is $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$. We will do this more carefully later, but what it says is that if P_{ij}^n approaches a limit denoted π_j as $n \to \infty$, then $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathsf{M}})$ satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$. If nothing else, it is easier to solve the linear equations $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ than to multiply $[P]$ by itself an infinite number of times.

Definition 3.3.1. *A steady-state vector (or a steady-state distribution) for an M state Markov chain with transition matrix $[P]$ is a row vector $\boldsymbol{\pi}$ that satisfies*

$$\boldsymbol{\pi} = \boldsymbol{\pi}[P] \; ; \quad \text{where } \sum_i \pi_i = 1 \text{ and } \pi_i \geq 0 \, , \, 1 \leq i \leq \mathsf{M}. \tag{3.9}$$

If $\boldsymbol{\pi}$ satisfies (3.9), then the last half of the equation says that it must be a probability vector. If $\boldsymbol{\pi}$ is taken as the initial PMF of the chain at time 0, then that PMF is maintained forever. That is, post-multiplyng both sides of (3.9) by $[P]$, we get $\boldsymbol{\pi}[P] = \boldsymbol{\pi}[P^2]$, and iterating this, $\boldsymbol{\pi} = \boldsymbol{\pi}[P^2] = \boldsymbol{\pi}[P^3] = \cdots$ for all n.

It is important to recognize that we have shown that if $[P^n]$ converges to a matrix all of whose rows are $\boldsymbol{\pi}$, then $\boldsymbol{\pi}$ is a steady-state vector, *i.e.*, it satisfies (3.9). However, finding a $\boldsymbol{\pi}$ that satisfies (3.9) does not imply that $[P^n]$ converges as $n \to \infty$. For the example of Figure 3.1, it can be seen that if we choose $\pi_2 = \pi_3 = 1/2$ with $\pi_i = 0$ otherwise, then $\boldsymbol{\pi}$ is a steady-state vector. Reasoning more physically, we see that if the chain is in either state 2 or 3, it simply oscillates between those states for all time. If it starts at time 0 being in states 2 or 3 with equal probability, then it persists forever being in states 2 or 3 with equal probability. Although this choice of $\boldsymbol{\pi}$ satisfies the definition in (3.9) and also is a steady-state distribution in the sense of not changing over time, it is not a very satisfying form of steady state, and almost seems to be concealing the fact that we are dealing with a simple oscillation between states.

This example raises one of a number of questions that should be answered concerning steady-state distributions and the convergence of $[P^n]$:

1. Under what conditions does $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ have a probability vector solution?
2. Under what conditions does $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$ have a unique probability vector solution?

3. Under what conditions does each row of $[P^n]$ converge to a probability vector solution of $\boldsymbol{\pi} = \boldsymbol{\pi}[P]$?

We first give the answers to these questions for finite-state Markov chains and then derive them. First, (3.9) *always* has a solution (although this is not necessarily true for infinite-state chains). The answers to the second and third questions are simplified if we use the following definition:

Definition 3.3.2. *A unichain is a finite-state Markov chain that contains a single recurrent class plus, perhaps, some transient states. An ergodic unichain is a unichain for which the recurrent class is ergodic.*

A unichain, as we shall see, is the natural generalization of a recurrent chain to allow for some initial transient behavior without disturbing the long term aymptotic behavior of the underlying recurrent chain.

The answer to the second question above is that the solution to (3.9) is unique if and only if [P] is the transition matrix of a *unichain*. If there are c recurrent classes, then (3.9) has c linearly independent solutions, each nonzero only over the elements of the corresponding recurrent class. For the third question, each row of $[P^n]$ converges to the unique solution of (3.9) if and only if [P] is the transition matrix of an *ergodic unichain*. If there are multiple recurrent classes, and each one is aperiodic, then $[P^n]$ still converges, but to a matrix with non-identical rows. If the Markov chain has one or more periodic recurrent classes, then $[P^n]$ does not converge.

We first look at these answers from the standpoint of the transition matrices of finite-state Markov chains, and then proceed in Chapter 5 to look at the more general problem of Markov chains with a countably infinite number of states. There we use renewal theory to answer these same questions (and to discover the differences that occur for infinite-state Markov chains).

The matrix approach is useful computationally and also has the advantage of telling us something about rates of convergence. The approach using renewal theory is very simple (given an understanding of renewal processes), but is more abstract.

In answering the above questions (plus a few more) for finite-state Markov chains, it is simplest to first consider the third question,[4] *i.e.*, the convergence of $[P^n]$. The simplest approach to this, for each column j of $[P^n]$, is to study the difference between the largest and smallest element of that column and how this difference changes with n. The following almost trivial lemma starts this study, and is valid for all finite-state Markov chains.

Lemma 3.3.1. *Let $[P]$ be the transition matrix of a finite-state Markov chain and let $[P^n]$ be the nth power of $[P]$* i.e.*, the matrix of nth order transition probabilities, P_{ij}^n. Then for each state j and each integer $n \geq 1$*

$$\max_i P_{ij}^{n+1} \leq \max_\ell P_{\ell j}^n \qquad \min_i P_{ij}^{n+1} \geq \min_\ell P_{\ell j}^n. \tag{3.10}$$

[4] One might naively try to show that a steady-state vector exists by first noting that each row of P sums to 1. The column vector $\boldsymbol{e} = (1, 1, \ldots, 1)^\mathsf{T}$ then satisfies the eigenvector equation $\boldsymbol{e} = [P]\boldsymbol{e}$. Thus there must also be a left eigenvector satisfying $\boldsymbol{\pi}[P] = \boldsymbol{\pi}$. The problem here is showing that $\boldsymbol{\pi}$ is real and non-negative.

Discussion The lemma says that for each column j, the largest of the elementsis non-increasing with n and the smallest of the elements is non-decreasing with n. The elements in a column that form the maximum and minimum can change with n, but the range covered by those elements is nested in n, either shrinking or staying the same as $n \to \infty$.

Proof: For each i, j, n, we use the Chapman-Kolmogorov equation, (3.8), followed by the fact that $P_{kj}^n \leq \max_\ell P_{\ell j}^n$, to see that

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \leq \sum_k P_{ik} \max_\ell P_{\ell j}^n = \max_\ell P_{\ell j}^n. \tag{3.11}$$

Since this holds for all states i, and thus for the maximizing i, the first half of (3.10) follows. The second half of (3.10) is the same, with minima replacing maxima, *i.e.*,

$$P_{ij}^{n+1} = \sum_k P_{ik} P_{kj}^n \geq \sum_k P_{ik} \min_\ell P_{\ell j}^n = \min_\ell P_{\ell j}^n. \tag{3.12}$$

□

For some Markov chains, the maximizing elements in each column decrease with n and reach a limit equal to the increasing sequence of minimizing elements. For these chains, $[P^n]$ converges to a matrix where each column is constant, *i.e.*, the type of convergence discussed above. For others, the maximizing elements converge at some value strictly above the limit of the minimizing elements, Then $[P^n]$ does not converge to a matrix where each column is constant, and might not converge at all since the location of the maximizing and minimizing elements in each column can vary with n.

The following three subsections establish the above kind of convergence (and a number of subsidiary results) for three cases of increasing complexity. The first assumes that $P_{ij} > 0$ for all i, j. This is denoted as $[P] > 0$ and is not of great interest in its own right, but provides a needed step for the other cases. The second case is where the Markov chain is ergodic, and the third is where the Markov chain is an ergodic unichain.

3.3.2 Steady state assuming $[P] > 0$

Lemma 3.3.2. *Let the transition matrix of a finite-state Markov chain satisfy* $[P] > 0$ *(*i.e.*, $P_{ij} > 0$ for all i, j), and let $\alpha = \min_{i,j} P_{ij}$. Then for all states j and all $n \geq 1$:*

$$\max_i P_{ij}^{n+1} - \min_i P_{ij}^{n+1} \leq \left(\max_\ell P_{\ell j}^n - \min_\ell P_{\ell j}^n \right) (1 - 2\alpha). \tag{3.13}$$

$$\left(\max_\ell P_{\ell j}^n - \min_\ell P_{\ell j}^n \right) \leq (1 - 2\alpha)^n. \tag{3.14}$$

$$\lim_{n\to\infty} \max_\ell P_{\ell j}^n = \lim_{n\to\infty} \min_\ell P_{\ell j}^n > 0. \tag{3.15}$$

Discussion: Since $P_{ij} > 0$ for all i, j, we must have $\alpha > 0$. Thus the theorem says that for each j, the elements P_{ij}^n in column j of $[P^n]$ approach equality over both i and n as

$n \to \infty$, *i.e.*, the state at time n becomes independent of the state at time 0 as $n \to \infty$. The approach is exponential in n.

Proof: We first slightly tighten the inequality in (3.11). For a given j and n, let $\ell_{\min}$ be a value of ℓ that minimizes $P^n_{\ell j}$. Then

$$\begin{aligned} P^{n+1}_{ij} &= \sum_k P_{ik} P^n_{kj} \\ &\leq \sum_{k \neq \ell_{\min}} P_{ik} \max_\ell P^n_{\ell j} \;+\; P_{i\ell_{\min}} \min_\ell P^n_{\ell j} \\ &= \max_\ell P^n_{\ell j} - P_{i\ell_{\min}} \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right) \\ &\leq \max_\ell P^n_{\ell j} - \alpha \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right), \end{aligned}$$

where in the third step, we added and subtracted $P_{i\ell_{\min}} \min_\ell P^n_{\ell j}$ to the right hand side, and in the fourth step, we used $\alpha \leq P_{i\ell_{\min}}$ in conjuction with the fact that the term in parentheses must be nonnegative.

Repeating the same argument with the roles of max and min reversed,

$$P^{n+1}_{ij} \geq \min_\ell P^n_{\ell j} + \alpha \left(\max_\ell P^n_{\ell j} - \min_\ell P^n_{\ell j} \right).$$

The upper bound above applies to $\max_i P^{n+1}_{ij}$ and the lower bound to $\min_i P^{n+1}_{ij}$. Thus, subtracting the lower bound from the upper bound, we get (3.13).

Finally, note that

$$\min_\ell P_{\ell j} \geq \alpha > 0 \qquad \max_\ell P_{\ell j} \leq 1 - \alpha.$$

Thus $\max_\ell P_{\ell j} - \min_\ell P_{\ell j} \leq 1 - 2\alpha$. Using this as the base for iterating (3.13) over n, we get (3.14). This, in conjuction with (3.10), shows not only that the limits in (3.10) exist and are positive and equal, but that the limits are approached exponentially in n. □

3.3.3 Ergodic Markov chains

Lemma 3.3.2 extends quite easily to arbitrary ergodic finite-state Markov chains. The key to this comes from Theorem 3.2.4, which shows that if $[P]$ is the matrix for an M state ergodic Markov chain, then the matrix $[P^h]$ is positive for any $h \geq (\mathsf{M}-1)^2 - 1$. Thus, choosing $h = (\mathsf{M}-1)^2 - 1$, we can apply Lemma 3.3.2 to $[P^h] > 0$. For each integer $\nu \geq 1$,

$$\max_i P^{h(\nu+1)}_{ij} - \min_i P^{h(\nu+1)}_{ij} \leq \left(\max_m P^{h\nu}_{mj} - \min_m P^{h\nu}_{mj} \right) (1 - 2\beta) \tag{3.16}$$

$$\left(\max_m P^{h\nu}_{mj} - \min_m P^{h\nu}_{mj} \right) \leq (1 - 2\beta)^\nu$$

$$\lim_{\nu \to \infty} \max_m P^{h\nu}_{mj} = \lim_{\nu \to \infty} \min_m P^{h\nu}_{mj} > 0, \tag{3.17}$$

where $\beta = \min_{i,j} P_{ij}^h$. Lemma 3.3.1 states that $\max_i P_{ij}^{n+1}$ is nondecreasing in n, so that the limit on the left in (3.17) can be replaced with a limit in n. Similarly, the limit on the right can be replaced with a limit on n, getting

$$\left(\max_m P_{mj}^n - \min_m P_{mj}^n\right) \leq (1-2\beta)^{\lfloor n/h \rfloor} \tag{3.18}$$

$$\lim_{n\to\infty} \max_m P_{mj}^n = \lim_{n\to\infty} \min_m P_{mj}^n > 0. \tag{3.19}$$

Now define $\boldsymbol{\pi} > 0$ by

$$\pi_j = \lim_{n\to\infty} \max_m P_{mj}^n = \lim_{n\to\infty} \min_m P_{mj}^n > 0. \tag{3.20}$$

Since π_j lies between the minimum and maximum P_{ij}^n for each n,

$$\left|P_{ij}^n - \pi_j\right| \leq (1-2\beta)^{\lfloor n/h \rfloor}. \tag{3.21}$$

In the limit, then,

$$\lim_{n\to\infty} P_{ij}^n = \pi_j \quad \text{for each } i,j. \tag{3.22}$$

This says that the matrix $[P^n]$ has a limit as $n \to \infty$ and the i,j term of that matrix is π_j for all i,j. In other words, each row of this limiting matrix is the same and is the vector $\boldsymbol{\pi}$. This is represented most compactly by

$$\lim_{n\to\infty} [P^n] = \boldsymbol{e}\boldsymbol{\pi} \qquad \text{where } \boldsymbol{e} = (1,1,\ldots,1)^{\mathsf{T}}. \tag{3.23}$$

The following theorem[5] summarizes these results and adds one small additional result.

Theorem 3.3.1. *Let $[P]$ be the matrix of an ergodic finite-state Markov chain. Then there is a unique steady-state vector $\boldsymbol{\pi}$, that vector is positive and satisfies (3.22) and (3.23). The convergence in n is exponential, satisfying (3.18).*

Proof: We need to show that $\boldsymbol{\pi}$ as defined in (3.20) is the unique steady-state vector. Let $\boldsymbol{\mu}$ be any steady state vector, *i.e.*, any probability vector solution to $\boldsymbol{\mu}[P] = \boldsymbol{\mu}$. Then $\boldsymbol{\mu}$ must satisfy $\boldsymbol{\mu} = \boldsymbol{\mu}[P^n]$ for all $n > 1$. Going to the limit,

$$\boldsymbol{\mu} = \boldsymbol{\mu} \lim_{n\to\infty} [P^n] = \boldsymbol{\mu}\boldsymbol{e}\boldsymbol{\pi} = \boldsymbol{\pi}.$$

Thus $\boldsymbol{\pi}$ is a steady state vector and is unique □

3.3.4 Ergodic Unichains

Understanding how P_{ij}^n approaches a limit as $n \to \infty$ for ergodic unichains is a straightforward extension of the results in Section 3.3.3, but the details require a little care. Let $\mathcal{T}$ denote the set of transient states (which might contain several transient classes), and

$$[P] = \left[\begin{array}{c|c} [P_{\mathcal{T}}] & [P_{\mathcal{TR}}] \\ \hline [0] & [P_{\mathcal{R}}] \end{array}\right] \qquad \text{where} \qquad [P_{\mathcal{T}}] = \begin{bmatrix} P_{11} & \cdots & P_{1t} \\ \cdots & \cdots & \cdots \\ P_{t1} & \cdots & P_{tt} \end{bmatrix}$$

$$[P_{\mathcal{TR}}] = \begin{bmatrix} P_{1,t+1} & \cdots & P_{1,t+r} \\ \cdots & \cdots & \cdots \\ P_{t,t+1} & \cdots & P_{t,t+r} \end{bmatrix} \qquad [P_{\mathcal{R}}] = \begin{bmatrix} P_{t+1,t+1} & \cdots & P_{t+r,t+1} \\ \cdots & \cdots & \cdots \\ P_{t+r,t+1} & \cdots & P_{t+r,t+r} \end{bmatrix}$$

Figure 3.5: The transition matrix of a unichain. The block of zeroes in the lower left corresponds to the absence of transitions from recurrent to transient states.

assume the states of $\mathcal{T}$ are numbered $1, 2, \ldots, t$. Let $\mathcal{R}$ denote the recurrent class, assumed to be numbered $t{+}1, \ldots, t{+}r$ (see Figure 3.5).

If i and j are both recurrent states, then there is no possibility of leaving the recurrent class in going from i to j. Assuming this class to be ergodic, the transition matrix $[P_{\mathcal{R}}]$ as shown in Figure 3.5 has been analyzed in Section 3.3.3.

If the initial state is a transient state, then eventually the recurrent class is entered, and eventually after that, the distribution approaches steady state within the recurrent class. This suggests (and we next show) that there is a steady-state vector $\boldsymbol{\pi}$ for $[P]$ itself such that $\pi_j = 0$ for $j \in \mathcal{T}$ and π_j is as given in Section 3.3.3 for each $j \in \mathcal{R}$.

Initially we will show that P_{ij}^n converges to 0 for $i, j \in \mathcal{T}$. The exact nature of how and when the recurrent class is entered starting in a transient state is an interesting problem in its own right, and is discussed more later. For now, a crude bound will suffice.

For each transient state, there must a walk to some recurrent state, and since there are only t transient states, there must be some such path of length at most t. Each such path has positive probability, and thus for each $i \in \mathcal{T}$, $\sum_{j\in\mathcal{R}} P_{ij}^t > 0$. It follows that for each $i \in \mathcal{T}$, $\sum_{j\in\mathcal{T}} P_{ij}^t < 1$. Let $\gamma < 1$ be the maximum of these probabilities over $i \in \mathcal{T}$, *i.e.*,

$$\gamma = \max_{i\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{ij}^t < 1.$$

Lemma 3.3.3. *Let $[P]$ be a unichain with a set $\mathcal{T}$ of t transient states. Then*

$$\max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^n \leq \gamma^{\lfloor n/t \rfloor}. \tag{3.24}$$

[5]This is essentially the Frobenius theorem for non-negative irreducible matrices, specialized to Markov chains. A non-negative matrix $[P]$ is *irreducible* if its graph (containing an edge from node i to j if $P_{ij} > 0$) is the graph of a recurrent Markov chain. There is no constraint that each row of $[P]$ sums to 1. The proof of the Frobenius theorem requires some fairly intricate analysis and seems to be far more complex than the simple proof here for Markov chains. A proof of the general Frobenius theorem can be found in [10].

Proof: For each integer multiple νt of t and each $i \in \mathcal{T}$,

$$\sum_{j\in\mathcal{T}} P_{ij}^{(\nu+1)t} = \sum_{k\in\mathcal{T}} P_{ik}^{t} \sum_{j\in\mathcal{T}} P_{kj}^{\nu t} \le \sum_{k\in\mathcal{T}} P_{ik}^{t} \max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t} \le \gamma \max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t}.$$

Recognizing that this applies to all $i \in \mathcal{T}$, and thus to the maximum over i, we can iterate this equation, getting

$$\max_{\ell\in\mathcal{T}} \sum_{j\in\mathcal{T}} P_{\ell j}^{\nu t} \le \gamma^{\nu}.$$

Since this maximum is nonincreasing in n, (3.24) follows. □

We now proceed to the case where the initial state is $i \in \mathcal{T}$ and the final state is $j \in \mathcal{R}$. Let $m = \lfloor n/2 \rfloor$. For each $i \in \mathcal{T}$ and $j \in \mathcal{R}$, the Chapman-Kolmogorov equation, says that

$$P_{ij}^{n} = \sum_{k\in\mathcal{T}} P_{ik}^{m} P_{kj}^{n-m} + \sum_{k\in\mathcal{R}} P_{ik}^{m} P_{kj}^{n-m}.$$

Let π_j be the steady-state probability of state $j \in \mathcal{R}$ in the recurrent Markov chain with states $\mathcal{R}$, *i.e.*, $\pi_j = \lim_{n\to\infty} P_{kj}^{n}$. Then for each $i \in \mathcal{T}$,

$$\begin{aligned}
\left|P_{ij}^{n} - \pi_j\right| &= \left|\sum_{k\in\mathcal{T}} P_{ik}^{m}\left(P_{kj}^{n-m} - \pi_j\right) + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left(P_{kj}^{n-m} - \pi_j\right)\right| \\
&\le \sum_{k\in\mathcal{T}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| \\
&\le \sum_{k\in\mathcal{T}} P_{ik}^{m} + \sum_{k\in\mathcal{R}} P_{ik}^{m}\left|P_{kj}^{n-m} - \pi_j\right| \qquad (3.25) \\
&\le \gamma^{\lfloor m/t \rfloor} + (1-2\beta)^{\lfloor (n-m)/h \rfloor}, \qquad (3.26)
\end{aligned}$$

where the first step upper bounded the absolute value of a sum by the sum of the absolute values. In the last step, (3.24) was used in the first half of (3.25) and (3.21) (with $h = (r-1)^2 - 1$ and $\beta = \min_{i,j\in\mathcal{R}} P_{ij}^{h} > 0$) was used in the second half.

This is summarized in the following theorem.

Theorem 3.3.2. *Let $[P]$ be the matrix of an ergodic finite-state unichain. Then $\lim_{n\to\infty}[P^n] = \boldsymbol{e\pi}$ where $\boldsymbol{e} = (1, 1, \ldots, 1)^{\mathsf{T}}$ and $\boldsymbol{\pi}$ is the steady-state vector of the recurrent class of states, expanded by 0's for each transient state of the unichain. The convergence is exponential in n for all i, j.*

3.3.5 Arbitrary finite-state Markov chains

The asymptotic behavior of $[P^n]$ as $n \to \infty$ for arbitrary finite-state Markov chains can mostly be deduced from the ergodic unichain case by simple extensions and common sense.

First consider the case of $m > 1$ aperiodic classes plus a set of transient states. If the initial state is in the κth of the recurrent classes, say $\mathcal{R}^{\kappa}$ then the chain remains in $\mathcal{R}^{\kappa}$ and there

is a unique finite-state vector $\boldsymbol{\pi}^\kappa$ that is non-zero only in $\mathcal{R}^\kappa$ that can be found by viewing class κ in isolation.

If the initial state i is transient, then, for each $\mathcal{R}^\kappa$, there is a certain probability that $\mathcal{R}^\kappa$ is eventually reached, and once it is reached there is no exit, so the steady state over that recurrent class is approached. The question of finding the probability of entering each recurrent class from a given transient class will be discussed in the next section.

Next consider a recurrent Markov chain that is periodic with period d. The dth order transition probability matrix, $[P^d]$, is then constrained by the fact that $P^d_{ij} = 0$ for all j not in the same periodic subset as i. In other words, $[P^d]$ is the matrix of a chain with d recurrent classes. We will obtain greater facility in working with this in the next section when eigenvalues and eigenvectors are discussed.

3.4 The eigenvalues and eigenvectors of stochastic matrices

For ergodic unichains, the previous section showed that the dependence of a state on the distant past disappears with increasing n, *i.e.*, $P^n_{ij} \to \pi_j$. In this section we look more carefully at the eigenvalues and eigenvectors of $[P]$ to sharpen our understanding of how fast $[P^n]$ converges for ergodic unichains and what happens for other finite-state Markov chains.

Definition 3.4.1. *The row*[6] *vector $\boldsymbol{\pi}$ is a left eigenvector of $[P]$ of eigenvalue λ if $\boldsymbol{\pi} \neq \boldsymbol{0}$ and $\boldsymbol{\pi}[P] = \lambda\boldsymbol{\pi}$,* i.e., $\sum_i \pi_i P_{ij} = \lambda\pi_j$ *for all j. The column vector $\boldsymbol{\nu}$ is a right eigenvector of eigenvalue λ if $\boldsymbol{\nu} \neq \boldsymbol{0}$ and $[P]\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$,* i.e., $\sum_j P_{ij}\nu_j = \lambda\nu_i$ *for all i.*

We showed that a stochastic matrix always has an eigenvalue $\lambda = 1$, and that for an ergodic unichain, there is a unique steady-state vector $\boldsymbol{\pi}$ that is a left eigenvector with $\lambda = 1$ and (within a scale factor) a unique right eigenvector $\boldsymbol{e} = (1, \ldots, 1)^\mathsf{T}$. In this section we look at the other eigenvalues and eigenvectors and also look at Markov chains other than ergodic unichains. We start by limiting the number of states to $\mathsf{M} = 2$. This provides insight without requiring much linear algebra. After that, the general case with arbitrary $\mathsf{M} < \infty$ is analyzed.

3.4.1 Eigenvalues and eigenvectors for $\mathsf{M} = 2$ states

The eigenvalues and eigenvectors can be found by elementary (but slightly tedious) algebra. The left and right eigenvector equations can be written out as

$$\begin{aligned} \pi_1 P_{11} + \pi_2 P_{21} &= \lambda\pi_1 \\ \pi_1 P_{12} + \pi_2 P_{22} &= \lambda\pi_2 \end{aligned} \quad \text{(left)} \qquad \begin{aligned} P_{11}\nu_1 + P_{12}\nu_2 &= \lambda\nu_1 \\ P_{21}\nu_1 + P_{22}\nu_2 &= \lambda\nu_2 \end{aligned} \quad \text{(right).} \tag{3.27}$$

[6] Students of linear algebra usually work primarily with right eigenvectors (and in abstract linear algebra often ignore matrices and concrete M-tuples altogether). Here a more concrete view is desirable because of the direct connection of $[P^n]$ with transition probabilities. Also, although left eigenvectors could be converted to right eigenvectors by taking the transpose of $[P]$, this would be awkward when Markov chains with rewards are considered and both row and column vectors play important roles.

Each set of equations have a non-zero solution if and only if the matrix $[P-\lambda I]$, where $[I]$ is the identity matrix, is singular (i.e., there must be a non-zero $\boldsymbol{\nu}$ for which $[P-\lambda I]\boldsymbol{\nu}=\boldsymbol{0}$). Thus λ must be such that the determinant of $[P-\lambda I]$, namely $(P_{11}-\lambda)(P_{22}-\lambda)-P_{12}P_{21}$, is equal to 0. Solving this quadratic equation in λ, we find that λ has two solutions,

$$\lambda_1=1 \qquad \lambda_2=1-P_{12}-P_{21}.$$

Assuming initially that P_{12} and P_{21} are not both 0, the solution for the left and right eigenvectors, $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\nu}^{(1)}$, of λ_1 and $\boldsymbol{\pi}^{(2)}$ and $\boldsymbol{\nu}^{(2)}$ of λ_2, are given by

$$\begin{array}{llll} \pi_1^{(1)}=\frac{P_{21}}{P_{12}+P_{21}} & \pi_2^{(1)}=\frac{P_{12}}{P_{12}+P_{21}} & \nu_1^{(1)}=1 & \nu_2^{(1)}=1 \\ \pi_1^{(2)}=1 & \pi_2^{(2)}=-1 & \nu_1^{(2)}=\frac{P_{12}}{P_{12}+P_{21}} & \nu_2^{(2)}=\frac{-P_{21}}{P_{12}+P_{21}} \end{array}.$$

These solutions contain arbitrary normalization factors. That for $\boldsymbol{\pi}^{(1)}=(\pi_1^{(1)},\ \pi_2^{(1)})$ has been chosen so that $\boldsymbol{\pi}^{(1)}$ is a steady-state vector (*i.e.*, the components sum to 1). The solutions have also been normalized so that $\boldsymbol{\pi}_i\boldsymbol{\nu}_i=1$ for $i=1,2$. Now define

$$[\Lambda]=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix} \qquad \text{and} \qquad [U]=\begin{bmatrix}\nu_1^{(1)} & \nu_1^{(2)}\\ \nu_2^{(1)} & \nu_2^{(2)}\end{bmatrix},$$

i.e., $[U]$ is a matrix whose columns are the eigenvectors $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$. Then the two right eigenvector equations in (3.27) can be combined compactly as $[P][U]=[U][\Lambda]$. It turns out (for the given normalization of the eigenvectors) that the inverse of $[U]$ is just the matrix whose rows are the left eigenvectors of $[P]$ (this can be verified by noting that $\boldsymbol{\pi}_1\boldsymbol{\nu}_2=\boldsymbol{\pi}_2\boldsymbol{\nu}_1=0$. We then see that $[P]=[U][\Lambda][U]^{-1}$ and consequently $[P^n]=[U][\Lambda]^n[U]^{-1}$. Multiplying this out, we get

$$[P^n]=\begin{bmatrix}\pi_1+\pi_2\lambda_2^n & \pi_2-\pi_2\lambda_2^n\\ \pi_1-\pi_1\lambda_2^n & \pi_2+\pi_1\lambda_2^n\end{bmatrix}, \tag{3.28}$$

where $\boldsymbol{\pi}=(\pi_1,\pi_2)$ is the steady state vector $\boldsymbol{\pi}^{(1)}$. Recalling that $\lambda_2=1-P_{12}-P_{21}$, we see that $|\lambda_2|\le 1$. There are 2 trivial cases where $|\lambda_2|=1$. In the first, $P_{12}=P_{21}=0$, so that $[P]$ is just the identity matrix. The Markov chain then has 2 recurrent states and stays forever where it starts. In the other trivial case, $P_{12}=P_{21}=1$. Then $\lambda_2=-1$ so that $[P^n]$ alternates between the identity matrix for n even and $[P]$ for n odd. In all other cases, $|\lambda_2|<1$ and $[P^n]$ approaches the steady state matrix $\lim_{n\to\infty}[P^n]=\boldsymbol{e\pi}$.

What we have learned from this is the exact way in which $[P^n]$ approaches $\boldsymbol{e\pi}$. Each term in $[P^n]$ approaches the steady state value exponentially in n as λ_2^n. Thus, in place of the upper bound in (3.21), we have an exact expression, which in this case is simpler than the bound. As we see shortly, this result is representative of the general case, but the simplicity is lost.

3.4.2 Eigenvalues and eigenvectors for $\mathsf{M} > 2$ states

For the general case of a stochastic matrix, we start with the fact that the set of eigenvalues is given by the set of (possibly complex) values of λ that satisfy the determinant equation $\det[P - \lambda I] = 0$. Since $\det[P - \lambda I]$ is a polynomial of degree M in λ, this equation has M roots (*i.e.*, M eigenvalues), not all of which need be distinct.[7]

Case with M distinct eigenvalues: We start with the simplest case in which the M eigenvalues, say $\lambda_1, \ldots, \lambda_{\mathsf{M}}$, are all distinct. The matrix $[P - \lambda_i I]$ is singular for each i, so there must be a right eigenvector $\boldsymbol{\nu}^{(i)}$ and a left eigenvector $\boldsymbol{\pi}^{(i)}$ for each eigenvalue λ_i. The right eigenvectors span M dimensional space and thus the matrix U with columns $(\boldsymbol{\nu}^{(1)}, \ldots, \boldsymbol{\nu}^{(\mathsf{M})})$ is nonsingular. The left eigenvectors, if normalized to satisfy $\boldsymbol{\pi}^{(i)}\boldsymbol{\nu}^{(i)} = 1$ for each i, then turn out to be the rows of $[U^{-1}]$ (see Exercise 3.11). As in the two state case, we can then express $[P^n]$ as

$$[P^n] = [U^{-1})[\Lambda^n][U], \tag{3.29}$$

where Λ is the diagonal matrix with terms $\lambda_1, \ldots, \lambda_{\mathsf{M}}$.

If Λ is broken into the sum of M diagonal matrices,[8] each with only a single nonzero element, then (see Exercise 3.11) $[P^n]$ can be expressed as

$$[P^n] = \sum_{i=1}^{\mathsf{M}} \lambda_i^n \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i)}. \tag{3.30}$$

Note that this is the same form as (3.28), where in (3.28), the eigenvalue $\lambda_1 = 1$ simply appears as the value 1. We have seen that there is always one eigenvalue that is 1, with an accompanying steady-state vector $\boldsymbol{\pi}$ as a left eigenvector and the unit vector $\boldsymbol{e} = (1, \ldots, 1)^{\mathsf{T}}$ as a right eigenvector. The other eigenvalues and eigenvectors can be complex, but it is almost self evident from the fact that $[P^n]$ is a stochastic matrix that $\left|\lambda_i\right| \leq 1$. A simple guided proof of this is given in Exercise 3.12.

We have seen that $\lim_{n\to\infty}[P^n] = \boldsymbol{e}\boldsymbol{\pi}$ for ergodic unichains. This implies that all terms except $i = 1$ in (3.30) die out with n, which further implies that $\left|\lambda_i\right| < 1$ for all eigenvalues except $\lambda = 1$. In this case, we see that the rate at which $[P^n]$ approaches steady state is given by the second largest eigenvalue in magnitude, *i.e.*, $\max_{i:|\lambda_i|<1} |\lambda_i|$.

If a recurrent chain is periodic with period d, it turns out that there are d eigenvalues of magnitude 1, and these are uniformly spaced around the unit circle in the complex plane. Exercise 3.19 contains a guided proof of this.

Case with repeated eigenvalues and M linearly independent eigenvectors: If some of the M eigenvalues of $[P]$ are not distinct, the question arises as to how many linearly independent left (or right) eigenvectors exist for an eigenvalue λ_i of multiplicity m, *i.e.*, a λ_i that is an mth order root of $\det[P - \lambda I]$. Perhaps the ugliest part of linear algebra is the fact that an eigenvalue of multiplicity m need not have m linearly independent eigenvectors.

[7] Readers with little exposure to linear algebra can either accept the linear algebra results in this section (without a great deal of lost insight) or can find them in Strang [19] or many other linear algebra texts.

[8] If 0 is one of the M eigenvalues, then only $\mathsf{M} - 1$ such matrices are required.

An example of a very simple Markov chain with $\mathsf{M} = 3$ but only two linearly independent eigenvectors is given in Exercise 3.14. These eigenvectors do not span M-space, and thus the expansion in (3.30) cannot be used.

Before looking at this ugly case, we look at the case where the right eigenvectors, say, span the space, *i.e.*, where each distinct eigenvalue has a number of linearly independent eigenvectors equal to its multiplicity. We can again form a matrix $[U]$ whose columns are the M linearly independent right eigenvectors, and again $[U^{-1}]$ is a matrix whose rows are the corresponding left eigenvectors of $[P]$. We then get (3.30) again. Thus, so long as the eigenvectors span the space, the asymptotic expression for the limiting transition probabilities can be found in the same way.

The most important situation where these repeated eigenvalues make a major difference is for Markov chains with $\kappa > 1$ recurrent classes. In this case, κ is the multiplicity of the eigenvalue 1. It is easy to see that there are κ different steady-state vectors. The steady-state vector for recurrent class ℓ, $1 \leq \ell \leq \kappa$, is strictly positive for each state of the ℓth recurrent class and is zero for all other states.

The eigenvalues for $[P]$ in this case can be found by finding the eigenvalues separately for each recurrent class. If class j contains r_j states, then r_j of the eigenvalues (counting repetitions) of $[P]$ are the eigenvalues of the r_j by r_j matrix for the states in that recurrent class. Thus the rate of convergence of $[P^n]$ within that submatrix is determined by the second largest eigenvalue (in magnitude) in that class.

What this means is that this general theory using eigenvalues says exactly what common sense says: if there are κ recurrent classes, look at each one separately, since they have nothing to do with each other. This also lets us see that for any recurrent class that is aperiodic, all the other eigenvalues for that class are strictly less than 1 in magnitude.

The situation is less obvious if there are κ recurrent classes plus a set of t transient states. All but t of the eigenvalues (counting repetitions) are associated with the recurrent classes, and the remaining t eigenvalues are the eigenvalues of the t by t matrix, say $[P_t]$, between the transient states. Each of these t eigenvalues are strictly less than 1 (as seen in Section 3.3.4) and neither these eigenvalues nor their eigenvectors depend on the transition probabilities from the transient to recurrent states. The left eigenvectors for the recurrent classes also do not depend on these transient to recurrent states. The right eigenvector for $\lambda = 1$ for each recurrent class $\mathcal{R}_\ell$ is very interesting however. It's value is 1 for each state in $\mathcal{R}_\ell$, is 0 for each state in the other recurrent classes, and is equal to $\lim_{n\to\infty} \Pr\{X_n \in \mathcal{R}_\ell \mid X_0 = i\}$ for each transient state i (see Exercise 3.13).

The Jordan form case: As mentioned before, there are cases in which one or more eigenvalues of $[P]$ are repeated (as roots of $\det[P - \lambda I]$) but where the number of linearly independent right eigenvectors for a given eigenvalue is less than the multiplicity of that eigenvalue. In this case, there are not enough eigenvectors to span the space, so there is no M by M matrix whose columns are linearly independent eigenvectors. Thus $[P]$ can not be expressed as $[U^{-1}][\Lambda][U]$ where Λ is the diagonal matrix of the eigenvalues, repeated according to their multiplicity.

The Jordan form is the cure for this unfortunate situation. The Jordan form for a given

$[P]$ is the following modification of the diagonal matrix of eigenvalues: we start with the diagonal matrix of eigenvalues, with the repeated eigenvalues as neighboring elements. Then for each missing eigenvector for a given eigenvalue, a 1 is placed immediately to the right and above a neighboring pair of appearances of that eigenvalue, as seen by example[9] below:

$$[J] = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}.$$

There is a theorem in linear algebra that says that an invertible matrix $[U]$ exists and a Jordan form exists such that $[P] = [U^{-1}][J][U]$. The major value to us of this result is that it makes it relatively easy to calculate $[J]^n$ for large n (see Exercise 3.15). This exercise also shows that for all stochastic matrices, each eigenvalue of magnitude 1 has precisely one associated eigenvector. This is usually expressed by the statement that all the eigenvalues of magnitude 1 are *simple*, meaning that their multiplicity equals their number of linearly independent eigenvectors. Finally the exercise shows that $[P^n]$ for an aperiodic recurrent chain converges as a polynomial[10] in n times λ_s^n where λ_s is the eigenvalue of largest magnitude less than 1.

The most important results of this section on eigenvalues and eigenvectors can be summarized in the following theorem.

Theorem 3.4.1. *The transition matrix of a finite state unichain has a single eigenvalue $\lambda = 1$ with an accompanying left eigenvector $\boldsymbol{\pi}$ satisfying (3.9) and a left eigenvector $\boldsymbol{e} = (1, 1, \ldots, 1)^\intercal$. The other eigenvalues λ_i all satisfy $|\lambda_i| \leq 1$. The inequality is strict unless the unichain is periodic, say with period d, and then there are d eigenvalues of magnitude 1 spaced equally around the unit circle. If the unichain is ergodic, then $[P^n]$ converges to steady state $\boldsymbol{e\pi}$ with an error in each term bounded by a fixed polynomial in n times $|\lambda_s|^n$, where λ_s is the eignevalue of largest magnitude less than 1.*

Arbitrary Markov chains can be split into their recurrent classes, and this theorem can be applied separately to each class.

3.5 Markov chains with rewards

Suppose that each state i in a Markov chain is associated with a reward, r_i. As the Markov chain proceeds from state to state, there is an associated sequence of rewards that are not independent, but are related by the statistics of the Markov chain. The concept of a reward in each state[11] is quite graphic for modeling corporate profits or portfolio performance, and

[9] See Strang [19], for example, for a more complete description of how to construct a Jordan form

[10] This polynomial is equal to 1 if these eigenvalues are simple.

[11] Occasionally it is more natural to associate rewards with transitions rather than states. If r_{ij} denotes a reward associated with a transition from i to j and P_{ij} denotes the corresponding transition probability, then defining $r_i = \sum_j P_{ij} r_{ij}$ essentially simplifies these transition rewards to rewards over the initial state for the transition. These transition rewards are ignored here, since the details add complexity to a topic that is complex enough for a first treatment.

is also useful for studying queueing delay, the time until some given state is entered, and many other phenomena. The reward r_i associated with a state could equally well be veiwed as a cost or any given real-valued function of the state.

In Section 3.6, we study dynamic programming and Markov decision theory. These topics include a "decision maker," "policy maker," or "control" that modify both the transition probabilities and the rewards at each trial of the 'Markov chain.' The decision maker attempts to maximize the expected reward, but is typically faced with compromising between immediate reward and the longer-term reward arising from the choice of transition probabilities that lead to 'high reward' states. This is a much more challenging problem than the current study of Markov chains with rewards, but a thorough understanding of the current problem provides the machinery to understand Markov decision theory also.

The steady-state expected reward per unit time, assuming a single recurrent class of states, is defined to be the *gain*, expressed as $g = \sum_i \pi_i r_i$ where π_i is the steady-state probability of being in state i.

3.5.1 Examples of Markov chains with rewards

The following examples demonstrate that it is important to understand the transient behavior of rewards as well as the long-term averages. This transient behavior will turn out to be even more important when we study Markov decision theory and dynamic programming.

Example 3.5.1 (Expected first-passage time). First-passage times, *i.e.*, the number of steps taken in going from one given state, say i, to another, say 1, are frequently of interest for Markov chains, and here we solve for the expected value of this random variable.

Since the first-passage time is independent of the transitions after the first entry to state 1, we can modify the chain to convert the final state, say state 1, into a trapping state (a *trapping state* i is a state from which there is no exit, *i.e.*, for which $P_{ii} = 1$). That is, we modify P_{11} to 1 and P_{1j} to 0 for all $j \neq 1$. We leave P_{ij} unchanged for all $i \neq 1$ and all j (see Figure 3.6). This modification of the chain will not change the probability of any sequence of states up to the point that state 1 is first entered.

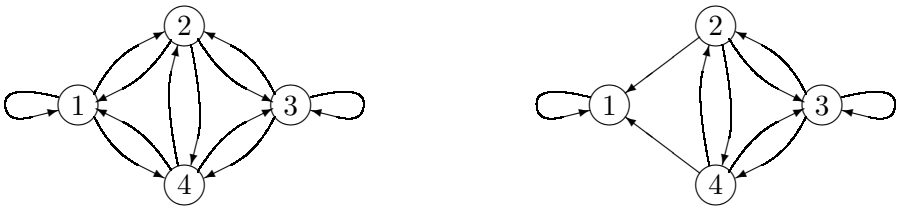


Figure 3.6: The conversion of a recurrent Markov chain with $\mathsf{M} = 4$ into a chain for which state 1 is a trapping state, *i.e.*, the outgoing arcs from node 1 have been removed.

Let v_i be the expected number of steps to first reach state 1 starting in state $i \neq 1$. This number of steps includes the first step plus the expected number of remaining steps to reach state 1 starting from whatever state is entered next (if state 1 is the next state entered, this

remaining number is 0). Thus, for the chain in Figure 3.6, we have the equations

$$\begin{aligned} v_2 &= 1 + P_{23}v_3 + P_{24}v_4. \\ v_3 &= 1 + P_{32}v_2 + P_{33}v_3 + P_{34}v_4. \\ v_4 &= 1 + P_{42}v_2 + P_{43}v_3. \end{aligned}$$

For an arbitrary chain of M states where 1 is a trapping state and all other states are transient, this set of equations becomes

$$v_i = 1 + \sum_{j \neq 1} P_{ij} v_j; \quad i \neq 1. \tag{3.31}$$

If we define $r_i = 1$ for $i \neq 1$ and $r_i = 0$ for $i = 1$, then r_i is a unit reward for not yet entering the trapping state, and v_i is the expected aggregate reward before entering the trapping state. Thus by taking $r_1 = 0$, the reward ceases upon entering the trapping state, and v_i is the expected transient reward, *i.e.*, the expected first-passage time from state i to state 1. Note that in this example, rewards occur only in transient states. Since transient states have zero steady-state probabilities, the steady-state gain per unit time, $g = \sum_i \pi_i r_i$, is 0.

If we define $v_1 = 0$, then (3.31), along with $v_1 = 0$, has the vector form

$$\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}; \quad v_1 = 0. \tag{3.32}$$

For a Markov chain with M states, (3.31) is a set of $\mathsf{M} - 1$ equations in the $\mathsf{M} - 1$ variables v_2 to v_{M}. The equation $\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}$ is a set of M linear equations, of which the first is the vacuous equation $v_1 = 0 + v_1$, and, with $v_1 = 0$, the last $\mathsf{M} - 1$ correspond to (3.31). It is not hard to show that (3.32) has a unique solution for $\boldsymbol{v}$ under the condition that states 2 to M are all transient and 1 is a trapping state, but we prove this later, in Theorem 3.5.1, under more general circumstances.

Example 3.5.2. Assume that a Markov chain has M states, $\{0, 1, \ldots, \mathsf{M} - 1\}$, and that the state represents the number of customers in an integer-time queueing system. Suppose we wish to find the expected sum of the customer waiting times, starting with i customers in the system at some given time t and ending at the first instant when the system becomes idle. That is, for each of the i customers in the system at time t, the waiting time is counted from t until that customer exits the system. For each new customer entering before the system next becomes idle, the waiting time is counted from entry to exit.

When we discuss Little's theorem in Section 4.5.4, it will be seen that this sum of waiting times is equal to the sum over τ of the state X_τ at time τ, taken from $\tau = t$ to the first subsequent time the system is empty.

As in the previous example, we modify the Markov chain to make state 0 a trapping state and assume the other states are then all transient. We take $r_i = i$ as the "reward" in state i, and v_i as the expected aggregate reward until the trapping state is entered. Using the same reasoning as in the previous example, v_i is equal to the immediate reward $r_i = i$ plus the expected aggregate reward from whatever state is entered next. Thus $v_i = r_i + \sum_{j \geq 1} P_{ij} v_j$. With $v_0 = 0$, this is $\boldsymbol{v} = \boldsymbol{r} + [P]\boldsymbol{v}$. This has a unique solution for $\boldsymbol{v}$, as will be shown later in Theorem 3.5.1. This same analysis is valid for any choice of reward r_i for each transient state i; the reward in the trapping state must be 0 so as to keep the expected aggregate reward finite.

In the above examples, the Markov chain is converted into a trapping state with zero gain, and thus the expected reward is a transient phenomena with no reward after entering the trapping state. We now look at the more general case of a unichain. In this more general case, there can be some gain per unit time, along with some transient expected reward depending on the initial state. We first look at the aggregate gain over a finite number of time units, thus providing a clean way of going to the limit.

Example 3.5.3. The example in Figure 3.7 provides some intuitive appreciation for the general problem. Note that the chain tends to persist in whatever state it is in. Thus if the chain starts in state 2, not only is an immediate reward of 1 achieved, but there is a high probability of additional unit rewards on many successive transitions. Thus the aggregate value of starting in state 2 is considerably more than the immediate reward of 1. On the other hand, we see from symmetry that the gain per unit time, over a long time period, must be one half.

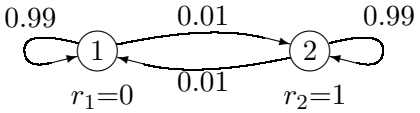


Figure 3.7: Markov chain with rewards and nonzero steady-state gain.

3.5.2 The expected aggregate reward over multiple transitions

Returning to the general case, let X_m be the state at time m and let $R_m = R(X_m)$ be the reward at that m, *i.e.*, if the sample value of X_m is i, then r_i is the sample value of R_m. Conditional on $X_m = i$, the aggregate expected reward $v_i(n)$ over n trials from X_m to X_{m+n-1} is

$$\begin{aligned} v_i(n) &= \mathsf{E}\left[R(X_m) + R(X_{m+1}) + \cdots + R(X_{m+n-1}) \mid X_m = i\right] \\ &= r_i + \sum_j P_{ij} r_j + \cdots + \sum_j P_{ij}^{n-1} r_j. \end{aligned}$$

This expression does not depend on the starting time m because of the homogeneity of the Markov chain. Since it gives the expected reward for each initial state i, it can be combined into the following vector expression $\boldsymbol{v}(n) = (v_1(n), v_2(n), \ldots, v_{\mathsf{M}}(n))^{\mathsf{T}}$,

$$\boldsymbol{v}(n) \;=\; \boldsymbol{r} + [P]\boldsymbol{r} + \cdots + [P^{n-1}]\boldsymbol{r} \;=\; \sum_{h=0}^{n-1} [P^h]\boldsymbol{r}, \tag{3.33}$$

where $\boldsymbol{r} = (r_1, \ldots, r_{\mathsf{M}})^{\mathsf{T}}$ and P^0 is the identity matrix. Now assume that the Markov chain is an ergodic unichain. Then $\lim_{n\to\infty}[P]^n = \boldsymbol{e\pi}$ and $\lim_{n\to\infty}[P]^n\boldsymbol{r} = \boldsymbol{e\pi r} = g\boldsymbol{e}$ where $g = \boldsymbol{\pi r}$ is the steady-state reward per unit time. If $g \neq 0$, then $\boldsymbol{v}(n)$ changes by approximately $g\boldsymbol{e}$ for each unit increase in n, so $\boldsymbol{v}(n)$ does not have a limit as $n \to \infty$. As shown below, however, $\boldsymbol{v}(n) - ng\boldsymbol{e}$ does have a limit, given by

$$\lim_{n\to\infty} (\boldsymbol{v}(n) - ng\boldsymbol{e}) = \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^h - \boldsymbol{e\pi}]\boldsymbol{r}. \tag{3.34}$$

To see that this limit exists, note from (3.26) that $\epsilon > 0$ can be chosen small enough that $P_{ij}^n - \pi_j = o(\exp(-n\epsilon))$ for all states i, j and all $n \geq 1$. Thus $\sum_{h=n}^{\infty}(P_{ij}^h - \pi_j) = o(\exp(-n\epsilon))$ also. This shows that the limits on each side of (3.34) must exist for an ergodic unichain.

The limit in (3.34) is a vector over the states of the Markov chain. This vector gives the asymptotic relative expected advantage of starting the chain in one state relative to another. This is an important quantity in both the next section and the remainder of this one. It is called the relative-gain vector and denoted by $\boldsymbol{w}$,

$$\boldsymbol{w} = \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^h - \boldsymbol{e}\boldsymbol{\pi}]\boldsymbol{r} \tag{3.35}$$

$$= \lim_{n\to\infty} (\boldsymbol{v}(n) - ng\boldsymbol{e}) \,. \tag{3.36}$$

Note from (3.36) that if $g > 0$, then $ng\boldsymbol{e}$ increases linearly with n and and $\boldsymbol{v}(n)$ must asymptotically increase linearly with n. Thus the relative-gain vector $\boldsymbol{w}$ becomes small relative to both $ng\boldsymbol{e}$ and $\boldsymbol{v}(n)$ for large n. As we will see, $\boldsymbol{w}$ is still important, particularly in the next section on Markov decisions.

We can get some feel for $\boldsymbol{w}$ and how $v_i(n) - n\pi_i$ converges to w_i from Example 3.5.3 (as described in Figure 3.7). Since this chain has only two states, $[P^n]$ and $v_i(n)$ can be calculated easily from (3.28). The result is tabulated in Figure 3.8, and it is seen numerically that $\boldsymbol{w} = (-25, +25)^{\mathsf{T}}$. The rather significant advantage of starting in state 2 rather than 1, however, requires hundreds of transitions before the gain is fully apparent.

n	$\boldsymbol{\pi}\boldsymbol{v}(n)$	$v_1(n)$	$v_2(n)$
1	0.5	0	1
2	1	0.01	1.99
4	2	0.0592	3.9408
10	5	0.4268	9.5732
40	20	6.1425	33.8575
100	50	28.3155	71.6845
400	200	175.007	224.9923

Figure 3.8: The expected aggregate reward, as a function of starting state and stage, for the example of figure 3.7. Note that $\boldsymbol{w} = (-25, +25)^{\mathsf{T}}$, but the convergence is quite slow.

This example also shows that it is somewhat inconvenient to calculate $\boldsymbol{w}$ from (3.35), and this inconvenience grows rapidly with the number of states. Fortunately, as shown in the following theorem, $\boldsymbol{w}$ can also be calculated simply by solving a set of linear equations.

Theorem 3.5.1. *Let $[P]$ be the transition matrix for an ergodic unichain. Then the relative-gain vector $\boldsymbol{w}$ given in (3.35) satisfies the following linear vector equation.*

$$\boldsymbol{w} + g\boldsymbol{e} = [P]\boldsymbol{w} + \boldsymbol{r} \qquad \textit{and} \quad \boldsymbol{\pi}\boldsymbol{w} = 0. \tag{3.37}$$

Furthermore (3.37) has a unique solution if $[P]$ is the transition matrix for a unichain (either ergodic or periodic).

Discussion: For an ergodic unichain, the interpretation of $\boldsymbol{w}$ as an asymptotic relative gain comes from (3.35) and (3.36). For a periodic unichain, (3.37) still has a unique solution, but (3.35) no longer converges, so the solution to (3.37) no longer has a clean interpretation as an asymptotic limit of relative gain. This solution is still called a relative-gain vector, and can be interpreted as an asymptotic relative gain over a period, but the important thing is that this equation has a unique solution for arbitrary unichains.

Definition 3.5.1. *The relative-gain vector $\boldsymbol{w}$ of a unichain is the unique vector that satisfies (3.37).*

Proof: Premultiplying both sides of (3.35) by $[P]$,

$$\begin{aligned}
[P]\boldsymbol{w} &= \lim_{n\to\infty} \sum_{h=0}^{n-1} [P^{h+1} - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \lim_{n\to\infty} \sum_{h=1}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \lim_{n\to\infty} \left(\sum_{h=0}^{n} [P^{h} - \boldsymbol{e\pi}]\boldsymbol{r} \right) - [P^0 - \boldsymbol{e\pi}]\boldsymbol{r} \\
&= \boldsymbol{w} - [P^0]\boldsymbol{r} + \boldsymbol{e\pi r} = \boldsymbol{w} - \boldsymbol{r} + g\boldsymbol{e}.
\end{aligned}$$

Rearranging terms, we get (3.37). For a unichain, the eigenvalue 1 of $[P]$ has multiplicity 1, and the existence and uniqueness of the solution to (3.37) is then a simple result in linear algebra (see Exercise 3.23). □

The above manipulations conceal the intuitive nature of (3.37). To see the intuition, consider the first-passage-time example again. Since all states are transient except state 1, $\pi_1 = 1$. Since $r_1 = 0$, we see that the steady-state gain is $g = 0$. Also, in the more general model of the theorem, $v_i(n)$ is the expected reward over n transitions starting in state i, which for the first-passage-time example is the expected number of transient states visited up to the nth transition. In other words, the quantity v_i in the first-passage-time example is $\lim_{n\to\infty} v_i(n)$. This means that the $\boldsymbol{v}$ in (3.32) is the same as $\boldsymbol{w}$ here, and it is seen that the formulas are the same with g set to 0 in (3.37).

The reason that the derivation of aggregate reward was so simple for first-passage time is that there was no steady-state gain in that example, and thus no need to separate the gain per transition g from the relative gain $\boldsymbol{w}$ between starting states.

One way to apply the intuition of the $g = 0$ case to the general case is as follows: given a reward vector $\boldsymbol{r}$, find the steady-state gain $g = \boldsymbol{\pi r}$, and then define a modified reward vector $\boldsymbol{r'} = \boldsymbol{r} - g\boldsymbol{e}$. Changing the reward vector from $\boldsymbol{r}$ to $\boldsymbol{r'}$ in this way does not change $\boldsymbol{w}$, but the modified limiting aggregate gain, say $\boldsymbol{v'}(n)$ then has a limit, which is in fact $\boldsymbol{w}$. The intuitive derivation used in (3.32) again gives us $\boldsymbol{w} = [P]\boldsymbol{w} + \boldsymbol{r'}$. This is equivalent to (3.37) since $\boldsymbol{r'} = \boldsymbol{r} - g\boldsymbol{e}$.

There are many generalizations of the first-passage-time example in which the reward in each recurrent state of a unichain is 0. Thus reward is accumulated only until a recurrent

state is entered. The following corollary provides a monotonicity result about the relative-gain vector for these circumstances that might seem obvious[12]. Thus we simply state it and give a guided proof in Exercise 3.25.

Corollary 3.5.1. *Let $[P]$ be the transition matrix of a unichain with the recurrent class $\mathcal{R}$. Let $\boldsymbol{r} \geq 0$ be a reward vector for $[P]$ with $r_i = 0$ for $i \in \mathcal{R}$. Then the relative-gain vector $\boldsymbol{w}$ satisfies $\boldsymbol{w} \geq 0$ with $w_i = 0$ for $i \in \mathcal{R}$ and $w_i > 0$ for $r_i > 0$. Furthermore, if $\boldsymbol{r}'$ and $\boldsymbol{r}''$ are different reward vectors for $[P]$ and $\boldsymbol{r}' \geq \boldsymbol{r}''$ with $r'_i = r''_i$ for $i \in \mathcal{R}$, then $\boldsymbol{w}' \geq \boldsymbol{w}''$ with $w'_i = w''_i$ for $i \in \mathcal{R}$ and $w'_i > w''_i$ for $r'_i > r''_i$.*

3.5.3 The expected aggregate reward with an additional final reward

Frequently when a reward is aggregated over n transitions of a Markov chain, it is appropriate to assign some added reward, say u_i, as a function of the final state i. For example, it might be particularly advantageous to end in some particular state. Also, if we wish to view the aggregate reward over $n + \ell$ transitions as the reward over the first n transitions plus that over the following ℓ transitions, we can model the expected reward over the final ℓ transitions as a final reward at the end of the first n transitions. Note that this final expected reward depends only on the state at the end of the first n transitions.

As before, let $R(X_{m+h})$ be the reward at time $m + h$ for $0 \leq h \leq n - 1$ and $U(X_{m+n})$ be the final reward at time $m + n$, where $U(X) = u_i$ for $X = i$. Let $v_i(n, \boldsymbol{u})$ be the expected reward from time m to $m + n$, using the reward $\boldsymbol{r}$ from time m to $m + n - 1$ and using the final reward $\boldsymbol{u}$ at time $m+n$. The expected reward is then the following simple modification of (3.33):

$$\boldsymbol{v}(n, \boldsymbol{u}) \;=\; \boldsymbol{r} + [P]\boldsymbol{r} + \cdots + [P^{n-1}]\boldsymbol{r} + [P^n]\boldsymbol{u} \;=\; \sum_{h=0}^{n-1} [P^h]\boldsymbol{r} \;+\; [P^n]\boldsymbol{u}. \tag{3.38}$$

This simplifies considerably if $\boldsymbol{u}$ is taken to be the relative-gain vector $\boldsymbol{w}$.

Theorem 3.5.2. *Let $[P]$ be the transition matrix of a unichain and let $\boldsymbol{w}$ be the corresponding relative-gain vector. Then for each $n \geq 1$,*

$$\boldsymbol{v}(n, \boldsymbol{w}) = ng\boldsymbol{e} + \boldsymbol{w}. \tag{3.39}$$

Also, for an arbitrary final reward vector $\boldsymbol{u}$,

$$\boldsymbol{v}(n, \boldsymbol{u}) = ng\boldsymbol{e} + \boldsymbol{w} + [P^n](\boldsymbol{u} - \boldsymbol{w}). \tag{3.40}$$

Discussion: An important special case of (3.40) arises from setting the final reward $\boldsymbol{u}$ to 0, thus yielding the following expression for $\boldsymbol{v}(n)$:

$$\boldsymbol{v}(n) = ng\boldsymbol{e} + \boldsymbol{w} - [P^n]\boldsymbol{w}. \tag{3.41}$$

[12] An obvious counterexample if we omit the condition $r_i = 0$ for $i \in \mathcal{R}$ is given by Figure 3.7 where $\boldsymbol{r} = (0, 1)^{\mathsf{T}}$ and $\boldsymbol{w} = (-25, 25)^{\mathsf{T}}$.

For an ergodic unichain, $\lim_{n\to\infty}[P^n] = \boldsymbol{e}\boldsymbol{\pi}$. Since $\boldsymbol{\pi}\boldsymbol{w} = 0$ by definition of $\boldsymbol{w}$, the limit of (3.41) as $n \to \infty$ is

$$\lim_{n\to\infty}(\boldsymbol{v}(n) - ng\boldsymbol{e}) = \boldsymbol{w},$$

which agrees with (3.36). The advantage of (3.41) over (3.36) is that it provides an explicit expression for $\boldsymbol{v}(n)$ for each n and also that it continues to hold for a periodic unichain.

Proof: For $n = 1$, we see from (3.38) that

$$\boldsymbol{v}(1, \boldsymbol{w}) \;=\; \boldsymbol{r} + [P]\boldsymbol{w} \;=\; g\boldsymbol{e} + \boldsymbol{w},$$

so the theorem is satisfied for $n = 1$. For $n > 1$,

$$\begin{aligned}
\boldsymbol{v}(n, \boldsymbol{w}) &= \sum_{h=0}^{n-1}[P^h]\boldsymbol{r} \;+\; [P^n]\boldsymbol{w} \\
&= \sum_{h=0}^{n-2}[P^h]\boldsymbol{r} \;+\; [P^{n-1}]\,(\boldsymbol{r} + [P]\boldsymbol{w}) \\
&= \sum_{h=0}^{n-2}[P^h]\boldsymbol{r} \;+\; [P^{n-1}]\,(g\boldsymbol{e} + \boldsymbol{w}) \\
&= \boldsymbol{v}(n{-}1, \boldsymbol{w}) + g\boldsymbol{e}.
\end{aligned}$$

Using induction, this implies (3.39).

To establish (3.40), note from (3.38) that

$$\boldsymbol{v}(n, \boldsymbol{u}) - \boldsymbol{v}(n, \boldsymbol{w}) \;=\; [P^n](\boldsymbol{u} - \boldsymbol{w}).$$

Then (3.40) follows by using (3.39) for the value of $\boldsymbol{v}(n, \boldsymbol{w})$. □

3.6 Markov decision theory and dynamic programming

In the previous section, we analyzed the behavior of a Markov chain with rewards. In this section, we consider a much more elaborate structure in which a decision maker can choose among various possible rewards and transition probabilities. In place of the reward r_i and the transition probabilities $\{P_{ij}; 1 \le j \le \mathsf{M}\}$ associated with a given state i, there is a choice between some number K_i of different rewards, say $r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(K_i)}$ and a corresponding choice between K_i different sets of transition probabilities, say $\{P_{ij}^{(1)}; 1 \le j \le \mathsf{M}\}, \{P_{ij}^{(2)}, 1 \le j \le \mathsf{M}\}, \ldots \{P_{ij}^{(K_i)}; 1 \le j \le \mathsf{M}\}$. At each time m, a decision maker, given $X_m = i$, selects one of the K_i possible choices for state i. Note that if decision k is chosen in state i, then the reward is $r_i^{(k)}$ and the transition probabilities from i are $\{P_{ij}^{(k)}; 1 \le j \le \mathsf{M}\}$; it is not permissable to choose $r_i^{(k)}$ for one k and $\{P_{ij}^{(k)}; 1 \le j \le \mathsf{M}\}$ for another k. We also assume that if decision k is selected at time m, the probability of entering state j at time $m + 1$ is $P_{ij}^{(k)}$, independent of earlier states and decisions.

Figure 3.9 shows an example of this situation in which the decision maker can choose between two possible decisions in state 2 ($K_2 = 2$), and has no freedom of choice in state 1 ($K_1 = 1$). This figure illustrates the familiar tradeoff between instant gratification (alternative 2) and long term gratification (alternative 1).

Figure 3.9: A Markov decision problem with two alternatives in state 2.

The set of rules used by the decision maker in selecting an alternative at each time is called a *policy*. We want to consider the expected aggregate reward over n steps of the "Markov chain" as a function of the policy used by the decision maker. If for each state i, the policy uses the same decision, say k_i, at each occurrence of i, then that policy corresponds to a homogeneous Markov chain with transition probabilities $P_{ij}^{(k_i)}$. We denote the matrix of these transition probabilities as $[P^{\boldsymbol{k}}]$, where $\boldsymbol{k} = (k_1, \ldots, k_{\mathsf{M}})$. Such a policy, i.e., mapping each state i into a fixed decision k_i, independent of time and past, is called a stationary policy. The aggregate gain for any such stationary policy was found in the previous section. Since both rewards and transition probabilities depend only on the state and the corresponding decision, and not on time, one feels intuitively that stationary policies make a certain amount of sense over a long period of time. On the other hand, if we look at the example of Figure 3.9, it is clear that decision 2 is the best choice in state 2 at the nth of n trials, but it is less obvious what to do at earlier trials.

In what follows, we first derive the optimal policy for maximizing expected aggregate reward over an arbitrary number n of trials, say at times m to $m + n - 1$. We shall see that the decision at time $m + h$, $0 \leq h < n$, for the optimal policy can in fact depend on h and n (but not m). It turns out to simplify matters considerably if we include a final reward $\{u_i; 1 \leq i \leq \mathsf{M}\}$ at time $m + n$. This final reward $\boldsymbol{u}$ is considered as a fixed vector, to be chosen as appropriate, rather than as part of the choice of policy.

This optimized strategy, as a function of the number of steps n and the final reward $\boldsymbol{u}$, is called an *optimal dynamic policy* for that $\boldsymbol{u}$. This policy is found from the dynamic programming algorithm, which, as we shall see, is conceptually very simple. We then go on to find the relationship between optimal dynamic policies and optimal stationary policies. We shall find that, under fairly general conditions, each has the same long-term gain per trial.

3.6.1 Dynamic programming algorithm

As in our development of Markov chains with rewards, we consider the expected aggregate reward over n time periods, say m to $m + n - 1$, with a final reward at time $m + n$. First consider the optimal decision with $n = 1$. Given $X_m = i$, a decision k is made with immediate reward $r_i^{(k)}$. With probability $P_{ij}^{(k)}$ the next state X_{m+1} is state j and the final

reward is then u_j. The expected aggregate reward over times m and $m+1$, maximized over the decision k, is then

$$v_i^*(1,\boldsymbol{u}) = \max_k\{r_i^{(k)} + \sum_j P_{ij}^{(k)} u_j\}. \tag{3.42}$$

Being explicit about the maximizing decision k', (3.42) becomes

$$v_i^*(1,\boldsymbol{u}) = r_i^{(k')} + \sum_j P_{ij}^{(k')} u_j \qquad \text{for } k' \text{ such that}$$

$$r_i^{(k')} + \sum_j P_{ij}^{(k')} u_j = \max_k\{r_i^{(k)} + \sum_j P_{ij}^{(k)} u_j\}. \tag{3.43}$$

Note that a decision is made only at time m, but that there are two rewards, one at time m and the other, the final reward, at time $m+1$. We use the notation $v_i^*(n,\boldsymbol{u})$ to represent the maximum expected aggregate reward from times m to $m+n$ starting at $X_m = i$. Decisions (with the reward vector $\boldsymbol{r}$) are made at the n times m to $m+n-1$, and this is followed by a final reward vector $\boldsymbol{u}$ (without any decision) at time $m+n$. It often simplifies notation to define the vector of maximal expected aggregate rewards

$$\boldsymbol{v}^*(n,\boldsymbol{u}) = (v_1^*(n,\boldsymbol{u}), v_2^*(n,\boldsymbol{u}), \ldots, v_{\mathsf{M}}^*(1,\boldsymbol{u}))^{\mathsf{T}}.$$

With this notation, (3.42) and (3.43) become

$$\boldsymbol{v}^*(1,\boldsymbol{u}) = \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{u}\} \quad \text{where } \boldsymbol{k} = (k_1,\ldots,k_{\mathsf{M}})^{\mathsf{T}},\ \boldsymbol{r}^{\boldsymbol{k}} = (r_1^{k_1},\ldots,r_{\mathsf{M}}^{k_{\mathsf{M}}})^{\mathsf{T}}. \tag{3.44}$$

$$\boldsymbol{v}^*(1,\boldsymbol{u}) = \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{u} \qquad \text{where } \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{u} = \max_{\boldsymbol{k}} \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{u}. \tag{3.45}$$

Now consider $v_i^*(2,\boldsymbol{u})$, *i.e.*, the maximal expected aggregate reward starting at $X_m = i$ with decisions made at times m and $m+1$ and a final reward at time $m+2$. The key to dynamic programming is that an optimal decision at time $m+1$ can be selected based only on the state j at time $m+1$; this decision (given $X_{m+1} = j$) is optimal independent of the decision at time m. That is, whatever decision is made at time m, the maximal expected reward at times $m+1$ and $m+2$, given $X_{m+1} = j$, is $\max_k \left(r_j^{(k)} + \sum_\ell P_{j\ell}^{(k)} u_\ell\right)$. Note that this maximum is $v_j^*(1,\boldsymbol{u})$, as found in (3.42).

Using this optimized decision at time $m+1$, it is seen that if $X_m = i$ and decision k is made at time m, then the sum of expected rewards at times $m+1$ and $m+2$ is $\sum_j P_{ij}^{(k)} v_j^*(1,\boldsymbol{u})$. Adding the expected reward at time m and maximizing over decisions at time m,

$$v_i^*(2,\boldsymbol{u}) = \max_k \left(r_i^{(k)} + \sum\nolimits_j P_{ij}^{(k)} v_j^*(1,\boldsymbol{u})\right). \tag{3.46}$$

In other words, the maximum aggregate gain over times m to $m+2$ (using the final reward $\boldsymbol{u}$ at $m+2$) is the maximum over choices at time m of the sum of the reward at m plus the

maximum aggregate expected reward for $m+1$ and $m+2$. The simple expression of (3.46) results from the fact that the maximization over the choice at time $m+1$ depends on the state at $m+1$ but, given that state, is independent of the policy chosen at time m.

This same argument can be used for all larger numbers of trials. To find the maximum expected aggregate reward from time m to $m+n$, we first find the maximum expected aggregate reward from $m+1$ to $m+n$, conditional on $X_{m+1}=j$ for each state j. This is the same as the maximum expected aggregate reward from time m to $m+n-1$, which is $v_j^*(n-1,\boldsymbol{u})$. This gives us the general expression for $n\geq 2$,

$$v_i^*(n,\boldsymbol{u})=\max_k\left(r_i^{(k)}+\sum\nolimits_j P_{ij}^{(k)}v_j^*(n-1,\boldsymbol{u})\right). \tag{3.47}$$

We can also write this in vector form as

$$\boldsymbol{v}^*(n,\boldsymbol{u})=\max_{\boldsymbol{k}}\left(\boldsymbol{r}^{\boldsymbol{k}}+[P^{\boldsymbol{k}}]\boldsymbol{v}^*(n-1,\boldsymbol{u})\right). \tag{3.48}$$

Here $\boldsymbol{k}$ is a set (or vector) of decisions, $\boldsymbol{k}=(k_1,k_2,\ldots,k_{\mathsf{M}})^{\mathsf{T}}$, where k_i is the decision for state i. $[P^{\boldsymbol{k}}]$ denotes a matrix whose (i,j) element is $P_{ij}^{(k_i)}$, and $\boldsymbol{r}^{\boldsymbol{k}}$ denotes a vector whose ith element is $r_i^{(k_i)}$. The maximization over $\boldsymbol{k}$ in (3.48) is really M separate and independent maximizations, one for each state, i.e., (3.48) is simply a vector form of (3.47). Another frequently useful way to rewrite (3.48) is as follows:

$$\boldsymbol{v}^*(n,\boldsymbol{u})=\boldsymbol{r}^{\boldsymbol{k}'}+[P^{\boldsymbol{k}'}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})\quad\text{for } \boldsymbol{k}' \text{ such that}$$

$$\boldsymbol{r}^{\boldsymbol{k}'}+[P^{\boldsymbol{k}'}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})=\max_{\boldsymbol{k}}\left(\boldsymbol{r}^{\boldsymbol{k}}+[P^{\boldsymbol{k}}]\boldsymbol{v}^*(n{-}1,\boldsymbol{u})\right). \tag{3.49}$$

If $\boldsymbol{k}'$ satisfies (3.49), then $\boldsymbol{k}'$ is an optimal decision at an arbitrary time m given, first, that the objective is to maximize the aggregate gain from time m to $m+n$, second, that optimal decisions for this objective are to be made at times $m+1$ to $m+n-1$, and, third, that $\boldsymbol{u}$ is the final reward vector at $m+n$. In the same way, $\boldsymbol{v}^*(n,\boldsymbol{u})$ is the maximum expected reward over this finite sequence of n decisions from m to $m+n-1$ with the final reward $\boldsymbol{u}$ at $m+n$.

Note that (3.47), (3.48), and (3.49) are valid with no restrictions (such as recurrent or aperiodic states) on the possible transition probabilities $[P^{\boldsymbol{k}}]$. These equations are also valid in principle if the size of the state space is infinite. However, the optimization for each n can then depend on an infinite number of optimizations at $n-1$, which is often infeasible.

The *dynamic programming algorithm* is just the calculation of (3.47), (3.48), or (3.49), performed iteratively for $n=1,2,3,\ldots$. The development of this algorithm, as a systematic tool for solving this class of problems, is due to Bellman [Bel57]. Note that the algorithm is independent of the starting time m; the parameter n, usually referred to as stage n, is the number of decisions over which the aggregate gain is being optimized. This algorithm yields the optimal dynamic policy for any fixed final reward vector $\boldsymbol{u}$ and any given number of trials. Along with the calculation of $\boldsymbol{v}^*(n,\boldsymbol{u})$ for each n, the algorithm also yields the optimal decision at each stage (under the assumption that the optimal policy is to be used for each lower numbered stage, *i.e.*, for each later trial of the process).

The surprising simplicity of the algorithm is due to the Markov property. That is, $v_i^*(n, \boldsymbol{u})$ is the aggregate present and future reward conditional on the present state. Since it is conditioned on the present state, it is independent of the past (i.e., how the process arrived at state i from previous transitions and choices).

Although dynamic programming is computationally straightforward and convenient[13], the asymptotic behavior of $\boldsymbol{v}^*(n, \boldsymbol{u})$ as $n \to \infty$ is not evident from the algorithm. After working out some simple examples, we look at the general question of asymptotic behavior.

Example 3.6.1. Consider Figure 3.9, repeated below, with the final rewards $u_2 = u_1 = 0$.

Since $r_1 = 0$ and $u_1 = u_2 = 0$, the aggregate gain in state 1 at stage 1 is

$$v_1^*(1, \boldsymbol{u}) = r_1 + \sum_j P_{1j} u_j = 0.$$

Similarly, since policy 1 has an immediate reward $r_2^{(1)} = 1$ in state 2, and policy 2 has an immediate reward $r_2^{(2)} = 50$,

$$v_2^*(1, \boldsymbol{u}) = \max\left\{ \left[r_2^{(1)} + \sum_j P_{2j}^{(1)} u_j\right], \quad \left[r_2^{(2)} + \sum_j P_{2j}^{(2)} u_j\right] \right\} = \max\{1,\, 50\} = 50.$$

We can now go on to stage 2, using the results above for $v_j^*(1, \boldsymbol{u})$. From (3.46),

$$\begin{aligned} v_1^*(2, \boldsymbol{u}) &= r_1 + P_{11} v_1^*(1, \boldsymbol{u}) + P_{12} v_2^*(1, \boldsymbol{u}) = P_{12} v_2^*(1, \boldsymbol{u}) = 0.5 \\ v_2^*(2, \boldsymbol{u}) &= \max\left\{ \left[r_2^{(1)} + \sum_j P_{2j}^{(1)} v_j^*(1, \boldsymbol{u})\right], \quad \left[r_2^{(2)} + P_{21}^{(2)} v_1^*(1, \boldsymbol{u})\right] \right\} \\ &= \max\left\{ [1 + P_{22}^{(1)} v_2^*(1, \boldsymbol{u})],\, 50 \right\} = \max\{50.5,\, 50\} = 50.5. \end{aligned}$$

Thus for two trials, decision 1 is optimal in state 2 for the first trial (stage 2), and decision 2 is optimal in state 2 for the second trial (stage 1). What is happening is that the choice of decision 2 at stage 1 has made it very profitable to be in state 2 at stage 1. Thus if the chain is in state 2 at stage 2, it is preferable to choose decision 1 (i.e., the small unit gain) at stage 2 with the corresponding high probability of remaining in state 2 at stage 1. Continuing this computation for larger n, one finds that $v_1^*(n, \boldsymbol{u}) = n/2$ and $v_2^*(n, \boldsymbol{u}) = 50 + n/2$. The optimum dynamic policy (for $\boldsymbol{u} = 0$) is decision 2 for stage 1 (*i.e.*, for the last decision to be made) and decision 1 for all stages $n > 1$ (i.e., for all decisions before the last).

This example also illustrates that the maximization of expected gain is not necessarily what is most desirable in all applications. For example, risk-averse people might well prefer decision 2 at the next to final decision (stage 2). This guarantees a reward of 50, rather than taking a small chance of losing that reward.

[13]Unfortunately, many dynamic programming problems of interest have enormous numbers of states and possible choices of decision (the so called curse of dimensionality), and thus, even though the equations are simple, the computational requirements might be beyond the range of practical feasibility.

Example 3.6.2 (Shortest Path Problems). The problem of finding the shortest paths between nodes in a directed graph arises in many situations, from routing in communication networks to calculating the time to complete complex tasks. The problem is quite similar to the expected first-passage time of example 3.5.1. In that problem, arcs in a directed graph were selected according to a probability distribution, whereas here decisions must be made about which arcs to take. Although this is not a probabilistic problem, the decisions can be posed as choosing a given arc with probability one, thus viewing the problem as a special case of dynamic programming.

Consider finding the shortest path from each node in a directed graph to some particular node, say node 1 (see Figure 3.10). Each arc (except the special arc $(1,1)$) has a positive *link length* associated with it that might reflect physical distance or an arbitrary type of cost. The special arc (1, 1) has 0 link length. The length of a path is the sum of the lengths of the arcs on that path. In terms of dynamic programming, a policy is a choice of arc out of each node (state). Here we want to minimize cost (i.e., path length) rather than maximizing reward, so we simply replace the maximum in the dynamic programming algorithm with a minimum (or, if one wishes, all costs can be replaced with negative rewards).

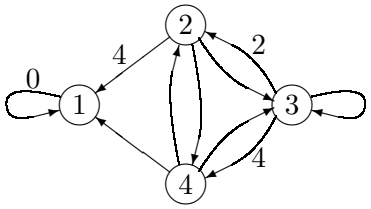


Figure 3.10: A shortest path problem. The arcs are marked with their lengths. Unmarked arcs have unit length.

We start the dynamic programming algorithm with a final cost vector that is 0 for node 1 and infinite for all other nodes. In stage 1, the minimal cost decision for node (state) 2 is arc $(2,1)$ with a cost equal to 4. The minimal cost decision for node 4 is $(4,1)$ with unit cost. The cost from node 3 (at stage 1) is infinite whichever decision is made. The stage 1 costs are then

$$v_1^*(1,\boldsymbol{u}) = 0, \quad v_2^*(1,\boldsymbol{u}) = 4, \quad v_3^*(1,\boldsymbol{u}) = \infty, \quad v_4^*(1,\boldsymbol{u}) = 1.$$

In stage 2, the cost $v_3^*(2,\boldsymbol{u})$, for example, is

$$v_3^*(2,\boldsymbol{u}) = \min\Big[2 + v_2^*(1,\boldsymbol{u}), \quad 4 + v_4^*(1,\boldsymbol{u})\Big] = 5.$$

The set of costs at stage 2 are

$$v_1^*(2,\boldsymbol{u}) = 0, \quad v_2^*(2,\boldsymbol{u}) = 2, \quad v_3^*(2,\boldsymbol{u}) = 5, \quad v_4^*(2,\boldsymbol{u}) = 1.$$

The decision at stage 2 is for node 2 to go to 4, node 3 to 4, and 4 to 1. At stage 3, node 3 switches to node 2, reducing its path length to 4, and nodes 2 and 4 are unchanged. Further iterations yield no change, and the resulting policy is also the optimal stationary policy.

The above results at each stage n can be interpreted as the shortest paths constrained to at most n hops. As n is increased, this constraint is successively relaxed, reaching the true shortest paths in less than M stages.

It can be seen without too much difficulty that these final aggregate costs (path lengths) also result no matter what final cost vector $\boldsymbol{u}$ (with $u_1 = 0$) is used. This is a useful feature for many types of networks where link lengths change very slowly with time and a shortest path algorithm is desired that can track the corresponding changes in the shortest paths.

3.6.2 Optimal stationary policies

In Example 3.6.1, we saw that there was a final transient (at stage 1) in which decision 1 was taken, and in all other stages, decision 2 was taken. Thus, the optimal dynamic policy consisted of a long-term stationary policy, followed by a transient period (for a single stage in this case) over which a different policy was used. It turns out that this final transient can be avoided by choosing an appropriate final reward vector $\boldsymbol{u}$ for the dynamic programming algorithm. If one has very good intuition, one would guess that the appropriate choice of final reward $\boldsymbol{u}$ is the relative-gain vector $\boldsymbol{w}$ associated with the long-term optimal policy.

It seems reasonable to expect this same type of behavior for typical but more complex Markov decision problems. In order to understand this, we start by considering an arbitrary stationary policy $\boldsymbol{k}' = (k'_1, \ldots, k'_{\mathsf{M}})$ and denote the transition matrix of the associated Markov chain as $[P^{\boldsymbol{k}'}]$. We assume that he associated Markov chain is a unichain, or, abbrevating terminology, that $\boldsymbol{k}'$ is a unichain. Let $\boldsymbol{w}'$ be the unique relative-gain vector for $\boldsymbol{k}'$. We then find some necessary conditions for $\boldsymbol{k}'$ to be the optimal dynamic policy at each stage using $\boldsymbol{w}'$ as the final reward vector.

First, from (3.45) $\boldsymbol{k}'$ is an optimal dynamic decision (with the final reward vector $\boldsymbol{w}'$ for $[P^{\boldsymbol{k}'}]$) at stage 1 if

$$\boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}' = \max_{\boldsymbol{k}} \{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'\}. \tag{3.50}$$

Note that this is more than a simple statement that $\boldsymbol{k}'$ can be found by maximizing $\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'$ over $\boldsymbol{k}$. It also involves the fact that $\boldsymbol{w}'$ is the relative-gain vector for $\boldsymbol{k}'$, so there is no immediately obvious way to find a $\boldsymbol{k}'$ that satisfies (3.50), and no a priori assurance that this equation even has a solution. The following theorem, however, says that this is the only condition required to ensure that $\boldsymbol{k}'$ is the optimal dynamic policy at every stage (again using $\boldsymbol{w}'$ as the final reward vector).

Theorem 3.6.1. *Assume that (3.50) is satisfied for some policy $\boldsymbol{k}'$ where the Markov chain for $\boldsymbol{k}'$ is a unichain and $\boldsymbol{w}'$ is the relative-gain vector of $\boldsymbol{k}'$. Then the optimal dynamic policy, using $\boldsymbol{w}'$ as the final reward vector, is the stationary policy $\boldsymbol{k}'$. Furthermore the optimal gain at each stage n is given by*

$$\boldsymbol{v}^*(n, \boldsymbol{w}') = \boldsymbol{w}' + ng'\boldsymbol{e}, \tag{3.51}$$

where $g' = \boldsymbol{\pi}'\boldsymbol{r}^{\boldsymbol{k}'}$ and $\boldsymbol{\pi}'$ is the steady-state vector for $\boldsymbol{k}'$.

Proof: We have seen from (3.45) that $\boldsymbol{k}'$ is an optimal dynamic decision at stage 1. Also, since $\boldsymbol{w}'$ is the relative-gain vector for $\boldsymbol{k}'$, Theorem 3.5.2 asserts that if decision $\boldsymbol{k}'$ is used at each stage, then the aggregate gain satisfies $\boldsymbol{v}(n, \boldsymbol{w}') = ng'\boldsymbol{e} + \boldsymbol{w}'$. Since $\boldsymbol{k}'$ is optimal at stage 1, it follows that (3.51) is satisfied for $n = 1$.

We now use induction on n, with $n = 1$ as a basis, to verify (3.51) and the optimality of this same $\boldsymbol{k}'$ at each stage n. Thus, assume that (3.51) is satisfied for n. Then, from (3.48),

$$\begin{aligned}
\boldsymbol{v}^*(n+1, \boldsymbol{w}') &= \max_{\boldsymbol{k}} \{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{v}^*(n, \boldsymbol{w}')\} && (3.52)\\
&= \max_{\boldsymbol{k}} \Big\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\{\boldsymbol{w}' + ng'\boldsymbol{e}\}\Big\} && (3.53)\\
&= ng'\boldsymbol{e} + \max_{\boldsymbol{k}} \{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}'\} && (3.54)\\
&= ng'\boldsymbol{e} + \boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}'\} && (3.55)\\
&= (n+1)g'\boldsymbol{e} + \boldsymbol{w}'. && (3.56)
\end{aligned}$$

Eqn (3.53) follows from the inductive hypothesis of (3.51), (3.54) follows because $[P^{\boldsymbol{k}}]\boldsymbol{e} = \boldsymbol{e}$ for all $\boldsymbol{k}$, (3.55) follows from (3.50), and (3.56) follows from the definition of $\boldsymbol{w}'$ as the relative-gain vector for $\boldsymbol{k}'$. This verifies (3.51) for $n + 1$. Also, since $\boldsymbol{k}'$ maximizes (3.54), it also maximizes (3.52), showing that $\boldsymbol{k}'$ is the optimal dynamic decision at stage $n + 1$. This completes the inductive step. □

Since our major interest in stationary policies is to help understand the relationship between the optimal dynamic policy and stationary policies, we define an optimal stationary policy as follows:

Definition 3.6.1. *A unichain stationary policy $\boldsymbol{k}'$ is optimal if the optimal dynamic policy with $\boldsymbol{w}'$ as the final reward uses $\boldsymbol{k}'$ at each stage.*

This definition side-steps several important issues. First, we might be interested in dynamic programming for some other final reward vector. Is it possible that dynamic programming performs much better in some sense with a different final reward vector. Is it possible that there is another stationary policy, especially one with a larger gain per stage? We answer these questions later and find that stationary policies that are optimal according to the definition do have maximal gain per stage compared with dynamic policies with arbitrary final reward vectors.

From Theorem 3.6.1, we see that if there is a policy $\boldsymbol{k}'$ which is a unichain with relative-gain vector $\boldsymbol{w}'$, and if that $\boldsymbol{k}'$ is a solution to (3.50), then $\boldsymbol{k}'$ is an optimal stationary policy.

It is easy to imagine Markov decision models for which each policy corresponds to a Markov chain with multiple recurrent classes. There are many special cases of such situations, and their detailed study is inappropriate in an introductory treatment. The essential problem with such models is that it is possible to get into various sets of states from which there is no exit, no matter what decisions are used. These sets might have different gains, so that there is no meaningful overall gain per stage. We avoid these situations by a modeling assumption called *inherent reachability*, which assumes, for each pair (i, j) of states, that there is some decision vector $\boldsymbol{k}$ containing a path from i to j.

The concept of inherent reachability is a little tricky, since it does not say the same $\boldsymbol{k}$ can be used for all pairs of states (*i.e.*, that there is some $\boldsymbol{k}$ for which the Markov chain is recurrent). As shown in Exercise 3.31, however, inherent reachability does imply that for any state j, there is a $\boldsymbol{k}$ for which j is accessible from all other states. As we have seen a number of times, this implies that the Markov chain for $\boldsymbol{k}$ is a unichain in which j is a recurrent state.

Any desired model can be modified to satisfy inherent reachability by creating some new decisions with very large negative rewards; these allow for such paths but very much discourage them. This will allow us to construct optimal unichain policies, but also to use the appearance of these large negative rewards to signal that there was something questionable in the original model.

3.6.3 Policy improvement and the seach for optimal stationary policies

The general idea of policy improvement is to start with an arbitrary unichain stationary policy $\boldsymbol{k'}$ with a relative gain vector $\boldsymbol{w'}$ (as given by (3.37)). We assume inherent reachability throughout this section, so such unichains must exist. We then check whether (3.50), is satisfied, and if so, we know from Theorem 3.6.1 that $\boldsymbol{k'}$ is an optimal stationary policy. If not, we find another stationary policy $\boldsymbol{k}$ that is 'better' than $\boldsymbol{k'}$ in a sense to be described later. Unfortunately, the 'better' policy that we find might not be a unichain, so it will also be necessary to convert this new policy into an equally 'good' unichain policy. This is where the assumption of of inherent reachability is needed. The algorithm then iteratively finds better and better unichain stationary policies, until eventually one of them satisfies (3.50) and is thus optimal.

We now state the policy-improvement algorithm for inherently reachable Markov decision problems. This algorithm is a generalization of Howard's policy-improvement algorithm, [How60].

Policy-improvement Algorithm

1. Choose an arbitrary unichain policy $\boldsymbol{k'}$
2. For policy $\boldsymbol{k'}$, calculate $\boldsymbol{w'}$ and g' from $\boldsymbol{w'} + g'\boldsymbol{e} = \boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'}$ and $\boldsymbol{\pi' w'} = 0$
3. If $\boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} = \max_{\boldsymbol{k}}\{\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}\}$, then stop; $\boldsymbol{k'}$ is optimal.
4. Otherwise, choose ℓ and k_ℓ so that $r_\ell^{(k'_\ell)} + \sum_j P_{\ell j}^{(k'_\ell)} w'_j < r_\ell^{(k_\ell)} + \sum_j P_{\ell j}^{(k_\ell)} w'_j$. For $i \neq \ell$, let $k_i = k'_i$.
5. If $\boldsymbol{k} = (k_1, \ldots k_{\mathsf{M}})$ is not a unichain, then let $\mathcal{R}$ be the recurrent class in $\boldsymbol{k}$ that contains state ℓ, and let $\tilde{\boldsymbol{k}}$ be a unichain policy for which $\tilde{k}_i = k_i$ for each $i \in \mathcal{R}$. Alternatively, if $\boldsymbol{k}$ is already a unichain, let $\tilde{\boldsymbol{k}} = \boldsymbol{k}$.
6. Update $\boldsymbol{k'}$ to the value of $\tilde{\boldsymbol{k}}$ and return to step 2.

If the stopping test in step 3 fails, there must be an ℓ and k_ℓ for which $r_\ell^{(k'_\ell)} + \sum_j P_{\ell j}^{(k'_\ell)} w'_j < r_\ell^{(k_\ell)} + \sum_j P_{\ell j}^{(k_\ell)} w'_j$. Thus step 4 can always be executed if the algorithm does not stop in

step 3, and since the decision is changed only for the single state ℓ, the resulting policy $\boldsymbol{k}$ satisfies

$$\boldsymbol{r}^{\boldsymbol{k}'} + [p^{\boldsymbol{k}'}]\boldsymbol{w}' \leq \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}' \qquad \text{with strict inequality for component } \ell. \tag{3.57}$$

The next three lemmas consider the different cases for the state ℓ whose decision is changed in step 4 of the algorithm. Taken together, they show that each iteration of the algorithm either increases the gain per stage or keeps the gain per stage constant while increasing the relative gain vector. After proving these lemmas, we return to show that the algorithm must converge and explain the sense in which the resulting stationary algorithm is optimal.

For each of the lemmas, let $\boldsymbol{k}'$ be the decision vector in step 1 of a given iteration of the policy improvement algorithm and assume that the Markov chain for $\boldsymbol{k}'$ is a unichain. Let g', $\boldsymbol{w}'$, and $\mathcal{R}'$ respectively be the gain per stage, the relative gain vector, and the recurrent set of states for $\boldsymbol{k}'$. Assume that the stopping condition in step 3 is not satisfied and that ℓ denotes the state whose decision is changed. Let k_ℓ be the new decision in step 4 and let $\boldsymbol{k}$ be the new decision vector.

Lemma 3.6.1. *Assume that $\ell \in \mathcal{R}'$. Then the Markov chain for $\boldsymbol{k}$ is a unichain and ℓ is recurrent in $\boldsymbol{k}$. The gain per stage g for $\boldsymbol{k}$ satisfies $g > g'$.*

Proof: The Markov chain for $\boldsymbol{k}$ is the same as that for $\boldsymbol{k}'$ except for the transitions out of state ℓ. Thus every path into ℓ in $\boldsymbol{k}'$ is still a path into ℓ in $\boldsymbol{k}$. Since ℓ is recurrent in the unichain $\boldsymbol{k}'$, it is accessible from all states in $\boldsymbol{k}'$ and thus in $\boldsymbol{k}$. It follows (see Exercise 3.3) that ℓ is recurrent in $\boldsymbol{k}$ and $\boldsymbol{k}$ is a unichain. Since $\boldsymbol{r}^{\boldsymbol{k}'} + [P^{\boldsymbol{k}'}]\boldsymbol{w}' = \boldsymbol{w}' + g'\boldsymbol{e}$ (see (3.37)), we can rewrite (3.57) as

$$\boldsymbol{w}' + g'\boldsymbol{e} \leq \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w}' \qquad \text{with strict inequality for component } \ell. \tag{3.58}$$

Premultiplying both sides of (3.58) by the steady-state vector $\boldsymbol{\pi}$ of the Markov chain $\boldsymbol{k}$ and using the fact that ℓ is recurent and thus $\pi_\ell > 0$,

$$\boldsymbol{\pi}\boldsymbol{w}' + g' < \boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}} + \boldsymbol{\pi}[P^{\boldsymbol{k}}]\boldsymbol{w}'.$$

Since $\boldsymbol{\pi}[P^{\boldsymbol{k}}] = \boldsymbol{\pi}$, this simplifies to

$$g' < \boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}}. \tag{3.59}$$

The gain per stage g for $\boldsymbol{k}$ is $\boldsymbol{\pi}\boldsymbol{r}^{\boldsymbol{k}}$, so we have $g' < g$. □

Lemma 3.6.2. *Assume that $\ell \notin \mathcal{R}'$ (*i.e.*, ℓ is transient in $\boldsymbol{k}'$) and that the states of $\mathcal{R}'$ are not accessible from ℓ in $\boldsymbol{k}$. Then $\boldsymbol{k}$ is not a unichain and ℓ is recurrent in $\boldsymbol{k}$. A decision vector $\tilde{\boldsymbol{k}}$ exists that is a unichain for which $\tilde{k}_i = k_i$ for $i \in \mathcal{R}$, and its gain per stage $\tilde{g}$ satisfies $\tilde{g} > g$.*

Proof: Since $\ell \notin \mathcal{R}'$, the transition probabilities from the states of $\mathcal{R}'$ are unchanged in going from $\boldsymbol{k}'$ to $\boldsymbol{k}$. Thus the set of states accessible from $\mathcal{R}'$ remains unchanged, and $\mathcal{R}'$ is a recurrent set of $\boldsymbol{k}$. Since $\mathcal{R}'$ is not accessible from ℓ, there must be another recurrent

set, $\mathcal{R}$, in $\boldsymbol{k}$, and thus $\boldsymbol{k}$ is not a unichain. The states accessible from $\mathcal{R}$ no longer include $\mathcal{R}'$, and since ℓ is the only state whose transition probabilities have changed, all states in $\mathcal{R}$ have paths to ℓ in $\boldsymbol{k}$. It follows that $\ell \in \mathcal{R}$.

Now let $\boldsymbol{\pi}$ be the steady-state vector for $\mathcal{R}$ in the Markov chain for $\boldsymbol{k}$. Since $\pi_\ell > 0$, (3.58) and (3.59) are still valid for this situation. Let $\tilde{\boldsymbol{k}}$ be a decision vector for which $\tilde{k}_i = k_i$ for each $i \in \mathcal{R}$. Using inherent reachability, we can also choose $\tilde{k}_i$ for each $i \notin \mathcal{R}$ so that ℓ is reachable from i (see Exercise 3.31). Thus $\tilde{\boldsymbol{k}}$ is a unichain with the recurrent class $\mathcal{R}$. Since $\tilde{\boldsymbol{k}}$ has the same transition probabilities and rewards in $\mathcal{R}$ as $\boldsymbol{k}$, we see that $\tilde{g} = \pi \boldsymbol{r}^{\boldsymbol{k}}$ and thus $\tilde{g} > g'$. □

The final lemma now includes all cases not in Lemmas 3.6.1 and 3.6.2

Lemma 3.6.3. *Assume that $\ell \notin \mathcal{R}'$ and that $\mathcal{R}'$ is accessible from ℓ in $\boldsymbol{k}$. Then $\boldsymbol{k}$ is a unichain with the same recurrent set $\mathcal{R}'$ as $\boldsymbol{k}'$. The gain per stage g is equal to g' and the relative-gain vector $\boldsymbol{w}$ of $\boldsymbol{k}$ satisfies*

$$\boldsymbol{w}' \leq \boldsymbol{w} \qquad \text{with } w'_\ell < w_\ell \text{ and } w'_i = w_i \text{ for } i \in \mathcal{R}'. \tag{3.60}$$

Proof: Since $\boldsymbol{k}'$ is a unichain, $\boldsymbol{k}'$ contains a path from each state to $\mathcal{R}'$. If such a path does not go through state ℓ, then $\boldsymbol{k}$ also contains that path. If such a path does go through ℓ, then that path can be replaced in $\boldsymbol{k}$ by the same path to ℓ followed by a path in $\boldsymbol{k}$ from ℓ to $\mathcal{R}'$. Thus $\mathcal{R}'$ is accessible from all states in $\boldsymbol{k}$. Since the states accessible from $\mathcal{R}'$ are unchanged from $\boldsymbol{k}'$ to $\boldsymbol{k}$, $\boldsymbol{k}$ is still a unichain with the recurrent set $\mathcal{R}'$ and state ℓ is still transient.

If we write out the defining equation (3.37) for $\boldsymbol{w}'$ component by component, we get

$$w'_i + g' = r_i^{k'_i} + \sum_j P_{ij}^{k'_i} w'_j. \tag{3.61}$$

Consider the set of these equations for which $i \in \mathcal{R}'$. Since $P_{ij}^{k'_i} = 0$ for all transient j in $\boldsymbol{k}'$, these are the same relative-gain equations as for the Markov chain restricted to $\mathcal{R}'$. Therefore $\boldsymbol{w}'$ is uniquely defined for $i \in \mathcal{R}'_i$ by this restricted set of equations. These equations are not changed in going from $\boldsymbol{k}'$ to $\boldsymbol{k}$, so it follows that $w_i = w'_i$ for $i \in \mathcal{R}'$. We have also seen that the steady-state vector $\boldsymbol{\pi}'$ is determined solely by the transition probabilities in the recurrent class, so $\boldsymbol{\pi}'$ is unchanged from $\boldsymbol{k}'$ to $\boldsymbol{k}$, and $g = g'$.

Finally, consider the difference between the relative-gain equations for $\boldsymbol{k}'$ in 3.61 and those for $\boldsymbol{k}$. Since $g' = g$,

$$w_i - w'_i = r_i^{k_i} - r_i^{k'_i} + \sum_j \left(P_{ij}^{k_i} w_j - P_{ij}^{k'_i} w'_j \right). \tag{3.62}$$

For all $i \neq \ell$, this simplifies to

$$w_i - w'_i = \sum_j P_{ij}^{k_i} (w_j - w'_j). \tag{3.63}$$

For $i = \ell$, (3.62) can be rewritten as

$$w_\ell - w'_\ell = \sum_j P_{\ell j}^{k_\ell}(w_j - w'_j) + \left[r_\ell^{k_\ell} - r_\ell^{k'_\ell} + \sum_j \left(P_{\ell j}^{k_\ell} w'_j - P_{\ell j}^{k'_\ell} w'_j \right) \right]. \tag{3.64}$$

The quantity in brackets must be positive because of step 4 of the algorithm, and we denote it as $\hat{r}_\ell - \hat{r}'_\ell$. If we also define $\hat{r}_i = \hat{r}'_i$ for $i \neq \ell$, then we can apply the last part of Corollary 3.5.1 (using $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{r}}'$ as reward vectors) to conclude that $\boldsymbol{w} \geq \boldsymbol{w}'$ with $w_\ell > w'_\ell$. □

We now see that each iteration of the algorithm either increases the gain per stage or holds the gain per stage the same and increases the relative-gain vector $\boldsymbol{w}$. Thus the sequence of policies found by the algorithm can never repeat. Since there are a finite number of stationary policies, the algorithm must eventually terminate at step 3. This means that the optimal dynamic policy using the final reward vector $\boldsymbol{w}'$ for the terminating decision vector $\boldsymbol{k}'$ must in fact be the stationary policy $\boldsymbol{k}'$.

The question now arises whether the optimal dynamic policy using some other final reward vector can be substantially better than that using $\boldsymbol{w}'$. The answer is quite simple and is developed in Exercise 3.30. It is shown there that if $\boldsymbol{u}$ and $\boldsymbol{u}'$ are arbitrary final reward vectors used on the dynamic programming algorithm, then $v^*(n, \boldsymbol{u})$ and $v^*(n, \boldsymbol{u}')$ are related by

$$v^*(n, \boldsymbol{u}) \leq v^*(n, \boldsymbol{u}') + \alpha \boldsymbol{e},$$

where $\alpha = \max_i(u_i - u'_i)$. Using $\boldsymbol{w}'$ for $\boldsymbol{u}'$, it is seen that the gain per stage of dynamic programming, with any final reward vector, is at most the gain g' of the stationary policy at the termination of the policy-improvement algorithm.

The above results are summarized in the following theorem.

Theorem 3.6.2. *For any inherently reachable finite-state Markov decision problem, the policy-improvement algorithm terminates with a stationary policy $\boldsymbol{k}'$ that is the same as the solution to the dynamic programming algorithm using $\boldsymbol{w}'$ as the final reward vector. The gain per stage g' of this stationary policy maximizes the gain per stage over all stationary policies and over all final-reward vectors for the dynamic programming algorithm.*

One remaining issue is the question whether the relative-gain vector found by the policy-improvement algorithm is in any sense optimal. The example in Figure 3.11 illustrates two different solutions terminating the policy-improvement algorithm. They each have the same gain (as guaranteed by Theorem 3.6.2) but their relative-gain vectors are not ordered.

In many applications such as variations on the shortest path problem, the interesting issue is what happens before the recurrent class is entered, and there is often only one recurrent class and one set of decisions within that class of interest. The following corollary shows that in this case, the relative-gain vector for the stationary policy that terminates the algorithm is maximal not only among the policies visited by the algorithm but among all policies with the same recurrent class and the same decisions within that class. The proof is almost the same as that of Lemma 3.6.3 and is carried out in Exercise 3.33.

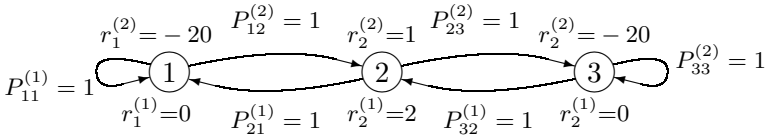


Figure 3.11: A Markov decision problem in which there are two unichain decision vectors (one left-going, and the other right-going). For each, (3.50) is satisfied and the gain per stage is 0. The dynamic programming algorithm (with no final reward) is stationary but has two recurrent classes, one of which is $\{3\}$, using decision 2 and the other of which is $\{1,2\}$, using decision 1 in each state.

Corollary 3.6.1. *Assume the policy improvement algorithm terminates with the recurrent class $\mathcal{R}'$, the decision vector $\boldsymbol{k}'$, and the relative-gain vector $\boldsymbol{w}'$. Then for any stationary policy that has the recurrent class $\mathcal{R}'$ and a decision vector $\boldsymbol{k}$ satisfying $k_i = k_i'$ for all $i \in \mathcal{R}'$, the relative gain vector $\boldsymbol{w}$ satisfies $\boldsymbol{w} \leq \boldsymbol{w}'$.*

3.7 Summary

This chapter has developed the basic results about finite-state Markov chains. It was shown that the states of any finite-state chain can be partitioned into classes, where each class is either transient or recurrent, and each class is periodic or aperiodic. If a recurrent class is periodic of period d, then the states in that class can be partitioned into d subsets where each subset has transitions only into the next subset.

The transition probabilities in the Markov chain can be represented as a matrix $[P]$, and the n-step transition probabilities are given by the matrix product $[P^n]$. If the chain is ergodic, *i.e.*, one aperiodic recurrent class, then the limit of the n-step transition probabilities become independent of the initial state, *i.e.*, $\lim_{n\to\infty} P_{ij}^n = \pi_j$ where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathsf{M}})$ is called the steady-state probability. Thus the limiting value of $[P^n]$ is an M by M matrix whose rows are all the same, *i.e.*, the limiting matrix is the product $\boldsymbol{e\pi}$. The steady state probabilities are uniquely specified by $\sum_j \pi_i P_{ij} = \pi_j$ and $\sum_i \pi_i = 1$. That unique solution must satisfy $\pi_i > 0$ for all i. The same result holds (see Theorem 3.3.2) for aperidodic unichains with the exception that $\pi_i = 0$ for all transient states.

The eigenvalues and eigenvectors of $[P]$ are useful in many ways, but in particular provide precise results about how P_{ij}^n approaches π_j with increasing n. An eigenvalue equal to 1 always exists, and its multiplicity is equal to the number of recurrent classes. For each recurrent class, there is a left eigenvector $\boldsymbol{\pi}$ of eigenvalue 1. It is the steady-state vector for the given recurrent class. If a recurrent class is periodic with period d, then there are d corresponding eigenvalues of magnitude 1 uniformly spaced around the unit circle. The left eigenvector corresponding to each is nonzero only on that periodic recurrent class.

All other eigenvalues of $[P]$ are less than 1 in magnitude. If the eigenvectors of the entire set of eigenvalues span M dimensional space, then $[P^n]$ can be represented by (3.30) which shows explicitly how steady state is approached for aperiodic recurrent classes of states. If the eigenvectors do not span M-space, then (3.30) can be replaced by a Jordan form.

For an arbitrary finite-state Markov chain, if the initial state is transient, then the Markov chain will eventually enter a recurrent state, and the probability that this takes more than n steps approaches zero geometrically in n; Exercise 3.18 shows how to find the probability that each recurrent class is entered. Given an entry into a particular recurrent class, the results about recurrent chains can be used to analyze the behavior within that class.

The results about Markov chains were extended to Markov chains with rewards. The use of reward functions (or cost functions) provides a systematic way to approach a large class of problems ranging from first-passage times to dynamic programming. For unichains, the key result here is Theorem 3.5.2, which provides both an exact expression and an asymptotic expression for the expected aggregate reward over n stages. Markov chains with rewards and multiple recurrent classes are best handled by considering the individual recurrent classes separately.

Finally, the results on Markov chains with rewards were used to understand Markov decision theory. The Bellman dynamic programming algorithm was developed, and the policy improvement algorithm was discussed and analyzed. Theorem 3.6.2 demonstrated the relationship between the optimal dynamic policy and the optimal stationary policy. This section provided only an introduction to dynamic programming and omitted all discussion of discounting (in which future gain is considered worth less than present gain because of interest rates). The development was also restricted to finite-state spaces.

For a review of vectors, matrices, and linear algebra, see any introductory text on linear algebra such as Strang [19]. For further reading on Markov decision theory and dynamic programming, see Bertsekas, [3]. Bellman [1] is of historic interest and quite readable.

3.8 Exercises

Exercise 3.1. Let $[P]$ be the transition matrix for a finite state Markov chain and let state i be recurrent. Prove that i is aperiodic if $P_{ii} > 0$.

Exercise 3.2. Show that every Markov chain with $\mathsf{M} < \infty$ states contains at least one recurrent set of states. Explaining each of the following statements is sufficient.

a) If state i_1 is transient, then there is some other state i_2 such that $i_1 \to i_2$ and $i_2 \not\to i_1$.

b) If the i_2 of part a) is also transient, there is an i_3 such that $i_2 \to i_3$, $i_3 \not\to i_2$, and consequently $i_1 \to i_3$, $i_3 \not\to i_1$.

c) Continuing inductively, if i_k is also transient, there is an i_{k+1} such that $i_j \to i_{k+1}$ and $i_{k+1} \not\to i_j$ for $1 \leq j \leq k$.

d) Show that for some $k \leq \mathsf{M}$, k is not transient, *i.e.*, it is recurrent, so a recurrent class exists.

Exercise 3.3. Consider a finite-state Markov chain in which some given state, say state 1, is accessible from every other state. Show that the chain has at most one recurrent class

$\mathcal{R}$ of states and state $1 \in \mathcal{R}$. (Note that, combined with Exercise 3.2, there is exactly one recurrent class and the chain is then a unichain.)

Exercise 3.4. Show how to generalize the graph in Figure 3.4 to an arbitrary number of states $\mathsf{M} \geq 3$ with one cycle of M nodes and one of $\mathsf{M}-1$ nodes. For $\mathsf{M}=4$, let node 1 be the node not in the cycle of $\mathsf{M}-1$ nodes. List the set of states accessible from node 1 in n steps for each $n \leq 12$ and show that the bound in Theorem 3.2.4 is met with equality. Explain why the same result holds for all larger M.

Exercise 3.5. (Proof of Theorem 3.2.4)

a) Show that an ergodic Markov chain with M states must contain a cycle with $\tau < \mathsf{M}$ states. Hint: Use ergodicity to show that the smallest cycle cannot contain M states.

b) Let ℓ be a fixed state on this cycle of length τ. Let $\mathcal{T}(m)$ be the set of states accessible from ℓ in m steps. Show that for each $m \geq 1$, $\mathcal{T}(m) \subseteq \mathcal{T}(m+\tau)$. Hint: For any given state $j \in \mathcal{T}(m)$, show how to construct a walk of $m+\tau$ steps from ℓ to j from the assumed walk of m steps.

c) Define $\mathcal{T}(0)$ to be the singleton set $\{i\}$ and show that

$$\mathcal{T}(0) \subseteq \mathcal{T}(\tau) \subseteq \mathcal{T}(2\tau) \subseteq \cdots \subseteq \mathcal{T}(n\tau) \subseteq \cdots .$$

d) Show that if one of the inclusions above is satisfied with equality, then all subsequent inclusions are satisifed with equality. Show from this that at most the first $M-1$ inclusions can be satisfied with strict inequality and that $\mathcal{T}(n\tau) = \mathcal{T}((M-1)\tau)$ for all $n \geq M-1$.

e) Show that all states are included in $\mathcal{T}((M-1)\tau)$.

f) Show that $P_{ij}^{(\mathsf{M}-1)^2+1} > 0$ for all i, j.

Exercise 3.6. Consider a Markov chain with one ergodic class of m states, say $\{1, 2, \ldots, m\}$ and $\mathsf{M}-m$ other states that are all transient. Show that $P_{ij}^n > 0$ for all $j \leq m$ and $n \geq (m-1)^2 + 1 + \mathsf{M} - m$.

Exercise 3.7. a) Let τ be the number of states in the smallest cycle of an arbitrary ergodic Markov chain of $\mathsf{M} \geq 3$ states. Show that $P_{ij}^n > 0$ for all $n \geq (\mathsf{M}-2)\tau + \mathsf{M}$. Hint: Look at the proof of Theorem 3.2.4 in Exercise 3.5.

b) For $\tau = 1$, draw the graph of an ergodic Markov chain (generalized for arbitrary $\mathsf{M} \geq 3$) for which there is an i, j for which $P_{ij}^n = 0$ for $n = 2\mathsf{M}-3$. Hint: Look at Figure 3.4.

c) For arbitrary $\tau < \mathsf{M}-1$, draw the graph of an ergodic Markov chain (generalized for arbitrary M) for which there is an i, j for which $P_{ij}^n = 0$ for $n = (\mathsf{M}-2)\tau + \mathsf{M} - 1$.

Exercise 3.8. A transition probability matrix $[P]$ is said to be doubly stochastic if

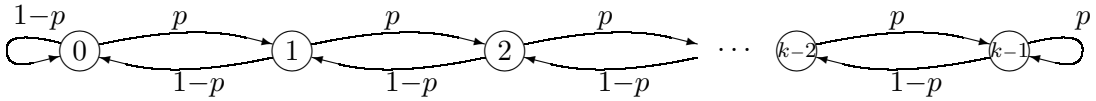
$$\sum_j P_{ij} = 1 \quad \text{for all } i; \quad \sum_i P_{ij} = 1 \quad \text{for all } j.$$

That is, the row sum and the column sum each equal 1. If a doubly stochastic chain has M states and is ergodic (i.e., has a single class of states and is aperiodic), calculate its steady-state probabilities.

Exercise 3.9. **a)** Find the steady-state probabilities $\pi_0, \ldots, \pi_{k-1}$ for the Markov chain below. Express your answer in terms of the ratio $\rho = p/q$. Pay particular attention to the special case $\rho = 1$.

b) Sketch $\pi_0, \ldots, \pi_{k-1}$. Give one sketch for $\rho = 1/2$, one for $\rho = 1$, and one for $\rho = 2$.

c) Find the limit of π_0 as k approaches ∞; give separate answers for $\rho < 1$, $\rho = 1$, and $\rho > 1$. Find limiting values of π_{k-1} for the same cases.



Exercise 3.10. **a)** Find the steady-state probabilities for each of the Markov chains in Figure 3.2. Assume that all clockwise probabilities in the first graph are the same, say p, and assume that $P_{4,5} = P_{4,1}$ in the second graph.

b) Find the matrices $[P]^2$ for the same chains. Draw the graphs for the Markov chains represented by $[P]^2$, i.e., the graph of two step transitions for the original chains. Find the steady-state probabilities for these two step chains. Explain why your steady-state probabilities are not unique.

c) Find $\lim_{n\to\infty}[P^{2n}]$ for each of the chains.

Exercise 3.11. **a)** Assume that $\boldsymbol{\nu}^{(i)}$ is a right eigenvector and $\boldsymbol{\pi}^{(j)}$ is a left eigenvector of an M by M stochastic matrix $[P]$ where $\lambda_i \neq \lambda_j$. Show that $\boldsymbol{\pi}^{(j)}\boldsymbol{\nu}^{(i)} = 0$. Hint: Consider two ways of finding $\boldsymbol{\pi}^{(j)}[P]\boldsymbol{\nu}^{(i)}$.

b) Assume that $[P]$ has M distinct eigenvalues. The right eigenvectors of $[P]$ then span M space (see section 5.2 of Strang, [19]), so the matrix $[U]$ with those eigenvectors as columns is nonsingular. Show that U^{-1} is a matix whose rows are the M left eigenvectors of $[P]$. Hint: use part a).

c) For each i, let $[\Lambda^{(i)}]$ be a diagonal matrix with a single nonzero element, $[\Lambda^{(i)}_{ii}] = \lambda_i$. Assume that $\boldsymbol{\pi}_i\boldsymbol{\nu}_k = 0$. Show that

$$\boldsymbol{\nu}^{(j)}[\Lambda^{(i)}]\boldsymbol{\pi}^{(k)} = \lambda_i\delta_{ik}\delta_{jk},$$

where δ_{ik} is 1 if $i = k$ and 0 otherwise. Hint visualize straightforward vector/matrix multiplication.

d) Verify (3.30).

Exercise 3.12. **a)** Let λ_k be an eigenvalue of a stochastic matrix $[P]$ and let $\boldsymbol{\pi}^{(k)}$ be an eigenvector for λ_k. Show that for each component $\pi^{(k)}_j$ of $\boldsymbol{\pi}^{(k)}$ and each n that

$$\lambda_k^n\,\pi_j^{(k)} = \sum_i \pi_i^{(k)} P_{ij}^n.$$

b) By taking magnitudes of each side, show that

$$\left|\lambda_k\right|^n \left|\pi_j^{(k)}\right| \leq \mathsf{M}.$$

c) Show that $\left|\lambda_k\right| \leq 1$.

Exercise 3.13. Consider a finite state Markov chain with matrix $[P]$ which has κ aperiodic recurrent classes, $\mathcal{R}_1, \ldots, \mathcal{R}_\kappa$ and a set $\mathcal{T}$ of transient states. For any given recurrent class ℓ, consider a vector $\boldsymbol{\nu}$ such that $\nu_i = 1$ for each $i \in \mathcal{R}_\ell$, $\nu_i = \lim_{n\to\infty} \Pr\{X_n \in \mathcal{R}_\ell | X_0 = i\}$ for each $i \in \mathcal{T}$, and $\nu_i = 0$ otherwise. Show that $\boldsymbol{\nu}$ is a right eigenvector of $[P]$ with eigenvalue 1. Hint: Redraw Figure 3.5 for multiple recurrent classes and first show that $\boldsymbol{\nu}$ is an eigenvector of $[P^n]$ in the limit.

Exercise 3.14. Answer the following questions for the following stochastic matrix $[P]$

$$[P] = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix}.$$

a) Find $[P]^n$ in closed form for arbitrary $n > 1$.

b) Find all distinct eigenvalues and the multiplicity of each distinct eigenvalue for $[P]$.

c) Find a right eigenvector for each distinct eigenvalue, and show that the eigenvalue of multiplicity 2 does not have 2 linearly independent eigenvectors.

d) Use (c) to show that there is no diagonal matrix $[\Lambda]$ and no invertible matrix $[U]$ for which $[P][U] = [U][\Lambda]$.

e) Rederive the result of part d) using the result of a) rather than c).

Exercise 3.15. a) Let $[J_i]$ be a 3 by 3 block of a Jordan form, *i.e.*,

$$[J_i] = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix}.$$

Show that the nth power of $[J_i]$ is given by

$$[J_i^n] = \begin{bmatrix} \lambda_i^n & n\lambda_1^{n-1} & \binom{n}{2}\lambda_i^{n-1} \\ 0 & \lambda_i^n & n\lambda_1^{n-1} \\ 0 & 0 & \lambda_i^n \end{bmatrix}.$$

Hint: Perhaps the easiest way is to calculate $[J_i^2]$ and $[J_i^3]$ and then use iteration.

b) Generalize a) to a k by k block of a Jordan form. Note that the nth power of an entire Jordan form is composed of these blocks along the diagonal of the matrix.

c) Let $[P]$ be a stochastic matrix represented by a Jordan form $[J]$ as $[P] = U^{-1}[J][U]$ and consider $[U][P][U^{-1}] = [J]$. Show that any repeated eigenvalue of $[P]$ (*i.e.*, any eigenvalue represented by a Jordan block of 2 by 2 or more) must be strictly less than 1. Hint: Upper bound the elements of $[U][P^n][U^{-1}]$ by taking the magnitude of the elements of $[U]$ and $[U^{-1}]$ and upper bounding each element of a stochastic matrix by 1.

d) Let λ_s be the eigenvalue of largest magnitude less than 1. Assume that the Jordan blocks for λ_s are at most of size k. Show that each ergodic class of $[P]$ converges at least as fast as $n^k \lambda_s^k$.

Exercise 3.16. **a)** Let λ be an eigenvalue of a matrix $[A]$, and let $\boldsymbol{\nu}$ and $\boldsymbol{\pi}$ be right and left eigenvectors respectively of λ, normalized so that $\boldsymbol{\pi\nu} = 1$. Show that

$$[[A] - \lambda \boldsymbol{\nu\pi}]^2 = [A^2] - \lambda^2 \boldsymbol{\nu\pi}.$$

b) Show that $[[A^n] - \lambda^n \boldsymbol{\nu\pi}][[A] - \lambda \boldsymbol{\nu\pi}] = [A^{n+1}] - \lambda^{n+1} \boldsymbol{\nu\pi}$.

c) Use induction to show that $[[A] - \lambda \boldsymbol{\nu\pi}]^n = [A^n] - \lambda^n \boldsymbol{\nu\pi}$.

Exercise 3.17. Let $[P]$ be the transition matrix for an aperiodic Markov unichain with the states numbered as in Figure 3.5.

a) Show that $[P^n]$ can be partitioned as

$$[P^n] = \begin{bmatrix} [P_{\mathcal{T}}{}^n] & [P_x^n] \\ 0 & [P_{\mathcal{R}}^n] \end{bmatrix}.$$

That is, the blocks on the diagonal are simply products of the corresponding blocks of $[P]$, and the upper right block is whatever it turns out to be.

b) Let q_i be the probability that the chain will be in a recurrent state after t transitions, starting from state i, i.e., $q_i = \sum_{t<j\leq t+r} P_{ij}^t$. Show that $q_i > 0$ for all transient i.

c) Let q be the minimum q_i over all transient i and show that $P_{ij}^{nt} \leq (1-q)^n$ for all transient i, j (i.e., show that $[P_{\mathcal{T}}]^n$ approaches the all zero matrix $[0]$ with increasing n).

d) Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_{\mathcal{T}}, \boldsymbol{\pi}_{\mathcal{R}})$ be a left eigenvector of $[P]$ of eigenvalue 1. Show that $\boldsymbol{\pi}_{\mathcal{T}} = \boldsymbol{0}$ and show that $\boldsymbol{\pi}_{\mathcal{R}}$ must be positive and be a left eigenvector of $[P_{\mathcal{R}}]$. Thus show that $\boldsymbol{\pi}$ exists and is unique (within a scale factor).

e) Show that $\boldsymbol{e}$ is the unique right eigenvector of $[P]$ of eigenvalue 1 (within a scale factor).

Exercise 3.18. Generalize Exercise 3.17 to the case of a Markov chain $[P]$ with m recurrent classes and one or more transient classes. In particular,

a) Show that $[P]$ has exactly κ linearly independent left eigenvectors, $\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)}, \ldots, \boldsymbol{\pi}^{(\kappa)}$ of eigenvalue 1, and that the mth can be taken as a probability vector that is positive on the mth recurrent class and zero elsewhere.

b) Show that $[P]$ has exactly κ linearly independent right eigenvectors, $\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)}, \ldots, \boldsymbol{\nu}^{(\kappa)}$ of eigenvalue 1, and that the mth can be taken as a vector with $\nu_i^{(m)}$ equal to the probability that recurrent class m will ever be entered starting from state i.

c) Show that

$$\lim_{n\to\infty}[P^n] = \sum_m \boldsymbol{\nu}^{(m)}\boldsymbol{\pi}^{(m)}.$$

Exercise 3.19. Suppose a recurrent Markov chain has period d and let $\mathcal{S}_m$, $1 \le m \le d$, be the mth subset in the sense of Theorem 3.2.3. Assume the states are numbered so that the first s_1 states are the states of $\mathcal{S}_1$, the next s_2 are those of $\mathcal{S}_2$, and so forth. Thus the matrix $[P]$ for the chain has the block form given by

$$[P] = \begin{bmatrix} 0 & [P_1] & \ddots & \ddots & 0 \\ 0 & 0 & [P_2] & \ddots & \ddots \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & \ddots & \ddots & [P_{d-1}] \\ [P_d] & 0 & \ddots & \ddots & 0 \end{bmatrix},$$

where $[P_m]$ has dimension s_m by s_{m+1} for $1 \le m \le d$, where $d+1$ is interpreted as 1 throughout. In what follows it is usually more convenient to express $[P_m]$ as an M by M matrix $[P'_m]$ whose entries are 0 except for the rows of $\mathcal{S}_m$ and the columns of $\mathcal{S}_{m+1}$, where the entries are equal to those of $[P_m]$. In this view, $[P] = \sum_{m=1}^{d}[P'_m]$.

a) Show that $[P^d]$ has the form

$$[P^d] = \begin{bmatrix} [Q_1] & 0 & \ddots & 0 \\ 0 & [Q_2] & \ddots & \ddots \\ 0 & 0 & \ddots & [Q_d] \end{bmatrix},$$

where $[Q_m] = [P_m][P_{m+1}]\ldots[P_d][P_1]\ldots[P_{m-1}]$. Expressing $[Q_m]$ as an M by M matrix $[Q'_m]$ whose entries are 0 except for the rows and columns of $\mathcal{S}_m$ where the entries are equal to those of $[Q_m]$, this becomes $[P^d] = \sum_{m=1}^{d} Q'_m$.

b) Show that $[Q_m]$ is the matrix of an ergodic Markov chain, so that with the eigenvectors $\hat{\boldsymbol{\pi}}_m, \hat{\boldsymbol{\nu}}_m$ as defined in Exercise 3.18, $\lim_{n\to\infty}[P^{nd}] = \sum_{m=1}^{d} \hat{\boldsymbol{\nu}}^{(m)}\hat{\boldsymbol{\pi}}^{(m)}$.

c) Show that $\hat{\boldsymbol{\pi}}^{(m)}[P'_m] = \hat{\boldsymbol{\pi}}^{(m+1)}$. Note that $\hat{\boldsymbol{\pi}}^{(m)}$ is an M-tuple that is nonzero only on the components of $\mathcal{S}_m$.

d) Let $\phi = \frac{2\boldsymbol{\pi}\sqrt{-1}}{d}$ and let $\boldsymbol{\pi}^{(k)} = \sum_{m=1}^{d} \hat{\boldsymbol{\pi}}^{(m)} e^{mk\phi}$. Show that $\boldsymbol{\pi}^{(k)}$ is a left eigenvector of $[P]$ of eigenvalue $e^{-\phi k}$.

Exercise 3.20. (continuation of Exercise 3.19). **a)** Show that, with the eigenvectors defined in Exercises 3.19,

$$\lim_{n\to\infty}[P^{nd}][P] = \sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i+1)},$$

where, as before, $d+1$ is taken to be 1.

b) Show that, for $1 \le j < d$,

$$\lim_{n\to\infty}[P^{nd}][P^j] = \sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\boldsymbol{\pi}^{(i+j)}.$$

c) Show that

$$\lim_{n\to\infty}[P^{nd}]\left\{I + [P] + \ldots, [P^{d-1}]\right\} = \left(\sum_{i=1}^{d} \boldsymbol{\nu}^{(i)}\right)\left(\sum_{i=1}^{d} \boldsymbol{\pi}^{(i+j)}\right).$$

d) Show that

$$\lim_{n\to\infty}\frac{1}{d}\left([P^n] + [P^{n+1}] + \cdots + [P^{n+d-1}]\right) = \boldsymbol{e}\boldsymbol{\pi},$$

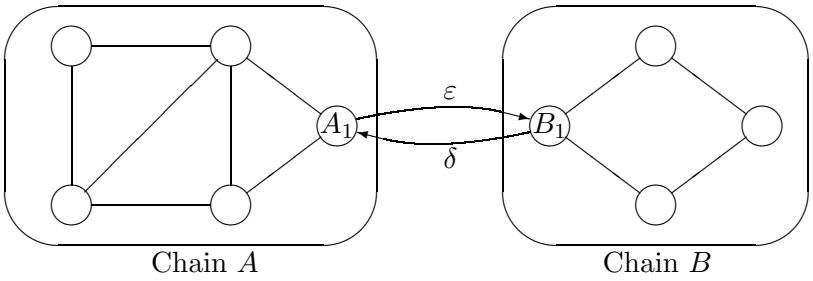
where $\boldsymbol{\pi}$ is the steady-state probability vector for $[P]$. Hint: Show that $\boldsymbol{e} = \sum_m \boldsymbol{\nu}^{(m)}$ and $\boldsymbol{\pi} = (1/d)\sum_m \boldsymbol{\pi}^{(m)}$.

e) Show that the above result is also valid for periodic unichains.

Exercise 3.21. Suppose A and B are each ergodic Markov chains with transition probabilities $\{P_{A_i,A_j}\}$ and $\{P_{B_i,B_j}\}$ respectively. Denote the steady-state probabilities of A and B by $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$ respectively. The chains are now connected and modified as shown below. In particular, states A_1 and B_1 are now connected and the new transition probabilities P' for the combined chain are given by

$$\begin{aligned} P'_{A_1,B_1} &= \varepsilon, \quad P'_{A_1,A_j} = (1-\varepsilon)P_{A_1,A_j} \qquad \text{for all } A_j \\ P'_{B_1,A_1} &= \delta, \quad P'_{B_1,B_j} = (1-\delta)P_{B_1,B_j} \qquad \text{for all } B_j. \end{aligned}$$

All other transition probabilities remain the same. Think intuitively of ε and δ as being small, but do not make any approximations in what follows. Give your answers to the following questions as functions of ε, δ, $\{\pi_{A_i}\}$ and $\{\pi_{B_i}\}$.



a) Assume that $\epsilon > 0$, $\delta = 0$ (i.e., that A is a set of transient states in the combined chain). Starting in state A_1, find the conditional expected time to return to A_1 given that the first transition is to some state in chain A.

b) Assume that $\epsilon > 0$, $\delta = 0$. Find $T_{A,B}$, the expected time to first reach state B_1 starting from state A_1. Your answer should be a function of ϵ and the original steady state probabilities $\{\pi_{A_i}\}$ in chain A.

c) Assume $\varepsilon > 0$, $\delta > 0$, find $T_{B,A}$, the expected time to first reach state A_1, starting in state B_1. Your answer should depend only on δ and $\{\pi_{B_i}\}$.

d) Assume $\varepsilon > 0$ and $\delta > 0$. Find $P'(A)$, the steady-state probability that the combined chain is in one of the states $\{A_j\}$ of the original chain A.

e) Assume $\varepsilon > 0$, $\delta = 0$. For each state $A_j \neq A_1$ in A, find v_{A_j}, the expected number of visits to state A_j, starting in state A_1, before reaching state B_1. Your answer should depend only on ε and $\{\pi_{A_i}\}$.

f) Assume $\varepsilon > 0$, $\delta > 0$. For each state A_j in A, find π'_{A_j}, the steady-state probability of being in state A_j in the combined chain. Hint: Be careful in your treatment of state A_1.

Exercise 3.22. Example 3.5.1 showed how to find the expected first passage times to a fixed state, say 1, from all other nodes. It is often desirable to include the expected first recurrence time from state 1 to return to state 1. This can be done by splitting state 1 into 2 states, first an initial state with no transitions coming into it but the original transitions going out, and second, a final trapping state with the original transitions coming in.

a) For the chain on the left side of Figure 3.6, draw the graph for the modified chain with 5 states where state 1 has been split into 2 states.

b) Suppose one has found the expected first-passage-times v_j for states $j = 2$ to 4 (or in general from 2 to M). Find an expression for v_1, the expected first recurrence time for state 1 in terms of $v_2, v_3, \ldots v_{\mathsf{M}}$ and $P_{12}, \ldots, P_{1\mathsf{M}}$.

Exercise 3.23. a) Assume throughout that $[P]$ is the transition matrix of a unichain (and thus the eigenvalue 1 has multiplicity 1). Show that a solution to the equation $[P]\boldsymbol{w} - \boldsymbol{w} = \boldsymbol{r} - g\boldsymbol{e}$ exists if and only if $\boldsymbol{r} - g\boldsymbol{e}$ lies in the column space of $[P - I]$ where I is the identity matrix.

b) Show that this column space is the set of vectors $\boldsymbol{x}$ for which $\boldsymbol{\pi x} = 0$. Then show that $\boldsymbol{r} - g\boldsymbol{e}$ lies in this column space.

c) Show that, with the extra constraint that $\boldsymbol{\pi w} = 0$, the equation $[P]\boldsymbol{w} - \boldsymbol{w} = \boldsymbol{r} - g\boldsymbol{e}$ has a unique solution.

Exercise 3.24. For the Markov chain with rewards in Figure 3.7,

a) Find the solution to (3.5.1) and find the gain g.

b) Modify Figure 3.7 by letting P_{12} be an arbitrary probability. Find g and $\boldsymbol{w}$ again and give an intuitive explanation of why P_{12} effects w_2.

Exercise 3.25. (Proof of Corollary 3.5.1) **a)** Show that the gain per stage g is 0. Hint: Show that $\boldsymbol{r}$ is zero where the steady-state vector $\boldsymbol{\pi}$ is nonzero.

b) Let $[P_{\mathcal{R}}]$ be the transition matrix for the recurrent states and let $\boldsymbol{r}_{\mathcal{R}} = 0$ be the reward vector and $\boldsymbol{w}_{\mathcal{R}}$ the relative-gain vector for $[P_{\mathcal{R}}]$. Show that $\boldsymbol{w}_{\mathcal{R}} = 0$. Hint: Use Theorem 3.5.1.

c) Show that $w_i = 0$ for all $i \in \mathcal{R}$. Hint: Compare the relative-gain equations for $[P]$ to those for $[P_{\mathcal{R}}]$.

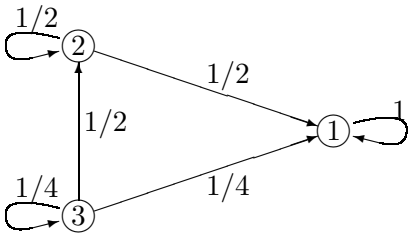
d) Show that for each $n \geq 0$, $[P^n]\boldsymbol{w} = [P^{n+1}]\boldsymbol{w} + [P^n]\boldsymbol{r}$. Hint: Start with the relative-gain equation for $[P]$.

e) Show that $\boldsymbol{w} = [P^{n+1}]\boldsymbol{w} + \sum_{m=0}^{n}[P^m]\boldsymbol{r}$. Hint: Sum the result in b).

f) Show that $\lim_{n\to\infty}[P^{n+1}]\boldsymbol{w} = 0$ and that $\lim_{n\to\infty}\sum_{m=0}^{n}[P^m]\boldsymbol{r}$ is finite, non-negative, and has positive components for $r_i > 0$. Hint: Use lemma 3.3.3.

g) Demonstrate the final result of the corollary by using the previous results on $\boldsymbol{r} = \boldsymbol{r}' - \boldsymbol{r}''$.

Exercise 3.26. Consider the Markov chain below:



a) Suppose the chain is started in state i and goes through n transitions; let $v_i(n)$ be the expected number of transitions (out of the total of n) until the chain enters the trapping state, state 1. Find an expression for $\boldsymbol{v}(n) = (v_1(n), v_2(n), v_3(n))^{\mathsf{T}}$ in terms of $\boldsymbol{v}(n-1)$ (take $v_1(n) = 0$ for all n). (Hint: view the system as a Markov reward system; what is the value of $\boldsymbol{r}$?)

b) Solve numerically for $\lim_{n\to\infty}\boldsymbol{v}(n)$. Interpret the meaning of the elements v_i in the solution of (3.32).

c) Give a direct argument why (3.32) provides the solution directly to the expected time from each state to enter the trapping state.

Exercise 3.27. a) Show that (3.48) can be rewritten in the more compact form

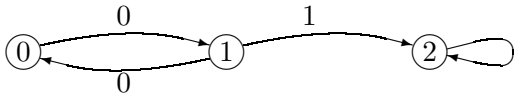
$$\boldsymbol{v}^*(n, \boldsymbol{u}) = \boldsymbol{v}^*(1, \boldsymbol{v}^*(n-1, \boldsymbol{u})).$$

b) Explain why it is also true that

$$\boldsymbol{v}^*(2n, \boldsymbol{u}) = \boldsymbol{v}^*(n, \boldsymbol{v}^*(n, \boldsymbol{u})). \tag{3.65}$$

c) One might guess that (3.65) could be used iteratively, finding $\boldsymbol{v}^*(2^{n+1}, \boldsymbol{u})$ from $\boldsymbol{v}^*(2^n, \boldsymbol{u})$. Explain why this is not possible in any straighttforward way. Hint: Think through explicitly how one might calculate $\boldsymbol{v}^*(n, \boldsymbol{v}^*(n, \boldsymbol{u}))$ from $\boldsymbol{v}^*(n, \boldsymbol{u})$.

Exercise 3.28. Consider a sequence of IID binary rv's $X_1, X_2, \ldots$. Assume that $\Pr\{X_i = 1\} = p_1$, $\Pr\{X_i = 0\} = p_0 = 1 - p_1$. A binary string $(a_1, a_2, \ldots, a_k)$ *occurs* at time n if $X_n = a_k, X_{n-1} = a_{k-1}, \ldots X_{n-k+1} = a_1$. For a given string $(a_1, a_2, \ldots, a_k)$, consider a Markov chain with $k+1$ states $\{0, 1, \ldots, k\}$. State 0 is the initial state, state k is a final trapping state where $(a_1, a_2, \ldots, a_k)$ has already occurred, and each intervening state i, $0 < i < k$, has the property that if the subsequent $k - i$ variables take on the values $a_{i+1}, a_{i+2}, \ldots, a_k$, the Markov chain will move successively from state i to $i+1$ to $i+2$ and so forth to k. For example, if $k = 2$ and $(a_1, a_2) = (0, 1)$, the corresponding chain is given by



a) For the chain above, find the mean first-passage time from state 0 to state 2.

b) For parts b) to d), let $(a_1, a_2, a_3, \ldots, a_k) = (0, 1, 1, \ldots, 1)$, i.e., zero followed by $k-1$ ones. Draw the corresponding Markov chain for $k = 4$.

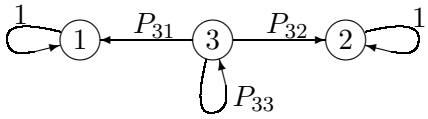
c) Let v_i, $1 \leq i \leq k$ be the expected first-passage time from state i to state k. Note that $v_k = 0$. Show that $v_0 = 1/p_0 + v_1$.

d) For each i, $1 \leq i < k$, show that $v_i = \alpha_i + v_{i+1}$ and $v_0 = \beta_i + v_{i+1}$ where α_i and β_i are each a product of powers of p_0 and p_1. Hint: use induction, or iteration, starting with $i = 1$, and establish both equalities together.

e) Let $k = 3$ and let $(a_1, a_2, a_3) = (1, 0, 1)$. Draw the corresponding Markov chain for this string. Evaluate v_0, the expected first-passage time for the string 1,0,1 to occur.

Exercise 3.29. a) Find $\lim_{n\to\infty}[P^n]$ for the Markov chain below. Hint: Think in terms of the long term transition probabilities. Recall that the edges in the graph for a Markov chain correspond to the positive transition probabilities.

b) Let $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ denote the first two rows of $\lim_{n\to\infty}[P^n]$ and let $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$ denote the first two columns of $\lim_{n\to\infty}[P^n]$. Show that $\boldsymbol{\pi}^{(1)}$ and $\boldsymbol{\pi}^{(2)}$ are independent left eigenvectors of $[P]$, and that $\boldsymbol{\nu}^{(1)}$ and $\boldsymbol{\nu}^{(2)}$ are independent right eigenvectors of $[P]$. Find the eigenvalue for each eigenvector.



c) Let $\boldsymbol{r}$ be an arbitrary reward vector and consider the equation

$$\boldsymbol{w} + g^{(1)}\boldsymbol{\nu}^{(1)} + g^{(2)}\boldsymbol{\nu}^{(2)} = \boldsymbol{r} + [P]\boldsymbol{w}. \tag{3.66}$$

Determine what values $g^{(1)}$ and $g^{(2)}$ must have in order for (3.66) to have a solution. Argue that with the additional constraints $w_1 = w_2 = 0$, (3.66) has a unique solution for $\boldsymbol{w}$ and find that $\boldsymbol{w}$.

Exercise 3.30. Let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be arbitrary final reward vectors with $\boldsymbol{u} \leq \boldsymbol{u}'$.

a) Let $\boldsymbol{k}$ be an arbitrary stationary policy and prove that $\boldsymbol{v}^{\boldsymbol{k}}(n, \boldsymbol{u}) \leq \boldsymbol{v}^{\boldsymbol{k}}(n, \boldsymbol{u}')$ for each $n \geq 1$.

b) For the optimal dynamic policy, prove that $\boldsymbol{v}^*(n, \boldsymbol{u}) \leq \boldsymbol{v}^*(n, \boldsymbol{u}')$ for each $n \geq 1$. This is known as the monotonicity theorem.

c) Now let $\boldsymbol{u}$ and $\boldsymbol{u}'$ be arbitrary. Let $\alpha = \max_i(u_i - u'_i)$. Show that

$$\boldsymbol{v}^*(n, \boldsymbol{u}) \leq \boldsymbol{v}^*(n, \boldsymbol{u}') + \alpha \boldsymbol{e}.$$

Exercise 3.31. Consider a Markov decision problem with M states in which some state, say state 1, is inherently reachable from each other state.

a) Show that there must be some other state, say state 2, and some decision, k_2, such that $P_{21}^{(k_2)} > 0$.

b) Show that there must be some other state, say state 3, and some decision, k_3, such that either $P_{31}^{(k_3)} > 0$ or $P_{32}^{(k_3)} > 0$.

c) Assume, for some i, and some set of decisions $k_2, \ldots, k_i$ that, for each j, $2 \leq j \leq i$, $P_{jl}^{(k_j)} > 0$ for some $l < j$ (i.e., that each state from 2 to j has a non-zero transition to a lower numbered state). Show that there is some state (other than 1 to i), say $i+1$ and some decision k_{i+1} such that $P_{i+1,l}^{(k_{i+1})} > 0$ for some $l \leq i$.

d) Use parts a), b), and c) to observe that there is a stationary policy $\boldsymbol{k} = k_1, \ldots, k_{\mathsf{M}}$ for which state 1 is accessible from each other state.

Exercise 3.32. George drives his car to the theater, which is at the end of a one-way street. There are parking places along the side of the street and a parking garage that costs \$5 at the theater. Each parking place is independently occupied or unoccupied with probability 1/2. If George parks n parking places away from the theater, it costs him n cents (in time and shoe leather) to walk the rest of the way. George is myopic and can only see the parking place he is currently passing. If George has not already parked by the time he reaches the nth place, he first decides whether or not he will park if the place is unoccupied, and then

observes the place and acts according to his decision. George can never go back and must park in the parking garage if he has not parked before.

a) Model the above problem as a 2 state dynamic programming problem. In the "driving" state, state 2, there are two possible decisions: park if the current place is unoccupied or drive on whether or not the current place is unoccupied.

b) Find $v_i^*(n, \boldsymbol{u})$, the *minimum* expected aggregate cost for n stages (i.e., immediately before observation of the nth parking place) starting in state $i = 1$ or 2; it is sufficient to express $v_i^*(n, \boldsymbol{u})$ in terms of $v_i^*(n-1)$. The final costs, in cents, at stage 0 should be $v_2(0) = 500$, $v_1(0) = 0$.

c) For what values of n is the optimal decision the decision to drive on?

d) What is the probability that George will park in the garage, assuming that he follows the optimal policy?

Exercise 3.33. (Proof of Corollary 3.6.1) **a)** Show that if two stationary policies $\boldsymbol{k}'$ and $\boldsymbol{k}$ have the same recurrent class $\mathcal{R}'$ and if $k_i' = k_i$ for all $i \in \mathcal{R}'$, then $w_i' = w_i$ for all $i \in \mathcal{R}'$. Hint: See the first part of the proof of Lemma 3.6.3.

b) Assume that $\boldsymbol{k}'$ satisfies 3.50 (*i.e.*, that it satisfies the termination condition of the policy improvement algorithm) and that $\boldsymbol{k}$ satisfies the conditions of part a). Show that (3.64) is satisfied for all states ℓ.

c) Show that $\boldsymbol{w} \leq \boldsymbol{w}'$. Hint: Follow the reasoning at the end of the proof of Lemma 3.6.3.

Exercise 3.34. Consider the dynamic programming problem below with two states and two possible policies, denoted $\boldsymbol{k}$ and $\boldsymbol{k}'$. The policies differ only in state 2.

a) Find the steady-state gain per stage, g and g', for stationary policies $\boldsymbol{k}$ and $\boldsymbol{k}'$. Show that $g = g'$.

b) Find the relative-gain vectors, $\boldsymbol{w}$ and $\boldsymbol{w}'$, for stationary policies $\boldsymbol{k}$ and $\boldsymbol{k}'$.

c) Suppose the final reward, at stage 0, is $u_1 = 0$, $u_2 = u$. For what range of u does the dynamic programming algorithm use decision $\boldsymbol{k}$ in state 2 at stage 1?

d) For what range of u does the dynamic programming algorithm use decision $\boldsymbol{k}$ in state 2 at stage 2? at stage n? You should find that (for this example) the dynamic programming algorithm uses the same decision at each stage n as it uses in stage 1.

e) Find the optimal gain $v_2^*(n, \boldsymbol{u})$ and $v_1^*(n, \boldsymbol{u})$ as a function of stage n assuming $u = 10$.

f) Find $\lim_{n\to\infty} v^*(n, \boldsymbol{u})$ and show how it depends on u.

Exercise 3.35. Consider a Markov decision problem in which the stationary policies $\boldsymbol{k}$ and $\boldsymbol{k'}$ each satisfy (3.50) and each correspond to ergodic Markov chains.

a) Show that if $\boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} \geq \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}$ is not satisfied with equality, then $g' > g$.

b) Show that $\boldsymbol{r}^{\boldsymbol{k'}} + [P^{\boldsymbol{k'}}]\boldsymbol{w'} = \boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'}$ (Hint: use part a).

c) Find the relationship between the relative gain vector $\boldsymbol{w}^{\boldsymbol{k}}$ for policy $\boldsymbol{k}$ and the relative-gain vector $\boldsymbol{w'}$ for policy $\boldsymbol{k'}$. (Hint: Show that $\boldsymbol{r}^{\boldsymbol{k}} + [P^{\boldsymbol{k}}]\boldsymbol{w'} = g\boldsymbol{e} + \boldsymbol{w'}$; what does this say about $\boldsymbol{w}$ and $\boldsymbol{w'}$?)

e) Suppose that policy $\boldsymbol{k}$ uses decision 1 in state 1 and policy $\boldsymbol{k'}$ uses decision 2 in state 1 (i.e., $k_1 = 1$ for policy $\boldsymbol{k}$ and $k_1 = 2$ for policy $\boldsymbol{k'}$). What is the relationship between $r_1^{(k)}, P_{11}^{(k)}, P_{12}^{(k)}, \ldots P_{1J}^{(k)}$ for k equal to 1 and 2?

f) Now suppose that policy $\boldsymbol{k}$ uses decision 1 in each state and policy $\boldsymbol{k'}$ uses decision 2 in each state. Is it possible that $r_i^{(1)} > r_i^{(2)}$ for all i? Explain carefully.

g) Now assume that $r_i^{(1)}$ is the same for all i. Does this change your answer to part f)? Explain.

Exercise 3.36. Consider a Markov decision problem with three states. Assume that each stationary policy corresponds to an ergodic Markov chain. It is known that a particular policy $\boldsymbol{k'} = (k_1, k_2, k_3) = (2, 4, 1)$ is the unique optimal stationary policy (i.e., the gain per stage in steady state is maximized by always using decision 2 in state 1, decision 4 in state 2, and decision 1 in state 3). As usual, $r_i^{(k)}$ denotes the reward in state i under decision k, and $P_{ij}^{(k)}$ denotes the probability of a transition to state j given state i and given the use of decision k in state i. Consider the effect of changing the Markov decision problem in each of the following ways (the changes in each part are to be considered in the absence of the changes in the other parts):

a) $r_1^{(1)}$is replaced by $r_1^{(1)} - 1$.

b) $r_1^{(2)}$is replaced by $r_1^{(2)} + 1$.

c) $r_1^{(k)}$ is replaced by $r_1^{(k)} + 1$ for all state 1 decisions k.

d) for all i, $r_i^{(k_i)}$ is replaced by $r^{(k_i)} + 1$ for the decision k_i of policy $\boldsymbol{k'}$.

For each of the above changes, answer the following questions; *give explanations*:

1) Is the gain per stage, g', increased, decreased, or unchanged by the given change?

2) Is it possible that another policy, $\boldsymbol{k} \neq \boldsymbol{k'}$, is optimal after the given change?

Exercise 3.37. (The Odoni Bound) Let $\boldsymbol{k'}$ be the optimal stationary policy for a Markov decision problem and let g' and $\boldsymbol{\pi'}$ be the corresponding gain and steady-state probability respectively. Let $v_i^*(n, \boldsymbol{u})$ be the optimal dynamic expected reward for starting in state i at stage n with final reward vector $\boldsymbol{u}$.

a) Show that $\min_i[v_i^*(n,\boldsymbol{u}) - v_i^*(n-1,\boldsymbol{u})] \leq g' \leq \max_i[v_i^*(n,\boldsymbol{u}) - v_i^*(n-1,\boldsymbol{u})]$; $n \geq 1$. Hint: Consider premultiplying $\boldsymbol{v}^*(n,\boldsymbol{u}) - \boldsymbol{v}^*(n-1,\boldsymbol{u})$ by $\boldsymbol{\pi}'$ or $\boldsymbol{\pi}$ where $\boldsymbol{k}$ is the optimal dynamic policy at stage n.

b) Show that the lower bound is non-decreasing in n and the upper bound is non-increasing in n and both converge to g' with increasing n.

Exercise 3.38. Consider an integer-time queueing system with a finite buffer of size 2. At the beginning of the n^{th} time interval, the queue contains at most two customers. There is a cost of one unit for each customer in queue (i.e., the cost of delaying that customer). If there is one customer in queue, that customer is served. If there are two customers, an extra server is hired at a cost of 3 units and both customers are served. Thus the total immediate cost for two customers in queue is 5, the cost for one customer is 1, and the cost for 0 customers is 0. At the end of the nth time interval, either 0, 1, or 2 new customers arrive (each with probability 1/3).

a) Assume that the system starts with $0 \leq i \leq 2$ customers in queue at time -1 (i.e., in stage 1) and terminates at time 0 (stage 0) with a final cost $\boldsymbol{u}$ of 5 units for each customer in queue (at the beginning of interval 0). Find the expected aggregate cost $v_i(1,\boldsymbol{u})$ for $0 \leq i \leq 2$.

b) Assume now that the system starts with i customers in queue at time -2 with the same final cost at time 0. Find the expected aggregate cost $v_i(2,\boldsymbol{u})$ for $0 \leq i \leq 2$.

c) For an arbitrary starting time $-n$, find the expected aggregate cost $v_i(n,\boldsymbol{u})$ for $0 \leq i \leq 2$.

d) Find the cost per stage and find the relative cost (gain) vector.

e) Now assume that there is a decision maker who can choose whether or not to hire the extra server when there are two customers in queue. If the extra server is not hired, the 3 unit fee is saved, but only one of the customers is served. If there are two arrivals in this case, assume that one is turned away at a cost of 5 units. Find the minimum dynamic aggregate expected cost $v_i^*(1)$, $0 \leq i \leq$, for stage 1 with the same final cost as before.

f) Find the minimum dynamic aggregate expected cost $v_i^*(n,\boldsymbol{u})$ for stage n, $0 \leq i \leq 2$.

g) Now assume a final cost $\boldsymbol{u}$ of one unit per customer rather than 5, and find the new minimum dynamic aggregate expected cost $v_i^*(n,\boldsymbol{u})$, $0 \leq i \leq 2$.

MIT OpenCourseWare
http://ocw.mit.edu

6.262 Discrete Stochastic Processes
Spring 2011

For information about citing these materials or our Terms of Use, visit: http://ocw.mit.edu/terms.