Markov Decision Processes Discrete Stochastic Dynamic Programming
Edited by MARTIN L. PUTERMAN

APPENDIX A

Markov Chains

This appendix summarizes results from Markov chain theory that are particularly relevant for the analysis of MDPs, especially those with average reward optimality criterion. The material in Secs. A.1–A.2 and A.4 is quite standard; we refer the reader to Chung (1960), Kemeny and Snell (1960), Karlin (1969), and Çinlar (1975) for additional details. Section A.3 provides the Fox and Landi (1968) algorithm for classifying states of a Markov chain. Sections A.5 and A.6 use matrix decomposition theory to derive some fundamental results. References for those sections include Campbell and Meyer (1979), Berman and Plemmons (1979), Senata (1981), Kemeny (1981), Lamond (1986), and Lamond and Puterman (1989). In response to frequent questions of friends and students, the appendix concludes with a brief biography of A. A. Markov. Biographical sources include Ondar (1981) and bibliographic notes in Çinlar (1975).

A.1 BASIC DEFINITIONS

Let $\{X_n, n = 0, 1, 2, \dots\}$ be a sequence of random variables which assume values in a discrete (finite or countable) state space S. We say that $\{X_n, n = 0, 1, 2, \dots\}$ is a *Markov chain* if

$$P\{X_n = j_n | X_{n-1} = j_{n-1}, X_{n-2} = j_{n-2}, \dots, X_0 = j_0\} = P\{X_n = j_n | X_{n-1} = j_{n-1}\}$$

for $n \geq 1$ and $j_k \in S$, $0 \leq k \leq n$. If $P\{X_n = j | X_{n-1} = s\}$ does not depend on n, we call the Markov chain *stationary* or *time homogeneous*. In this case we write $p(j|s) \equiv P\{X_n = j | X_{n-1} = s\}$ and refer to it as a *transition probability*. We call the matrix P with (s, j)th component $p(j|s)$, the *transition probability matrix* or *transition matrix*. In a stationary Markov chain, we denote the *m-step transition probability* by $p^m(j|s) = P\{X_{n+m} = j | X_n = s\}$. As a consequence of repeated application of the law of total probabilities, $p^m(j|s)$ is an element of the matrix P^m, the mth power of the matrix P.

For a real-valued function $g(\cdot)$ on S we denote the expected value of $g(X_n)$ by $E_s\{g(X_n)\} \equiv \sum_{j \in S} g(j)p^n(j|s)$ when $X_0 = s$. When g is the indicator of $A \subset S$, we let $P_s\{X_n \in A\} \equiv E_s\{I_{\{A\}}(X_n)\} = \sum_{j \in A} p^n(j|s)$.

A.2 CLASSIFICATION OF STATES

With each $s \in S$, associate the random variables ν_s and τ_s, which represent the number of visits and time of the first visit (first return if the chain starts in s), to state s. We classify states on the basis of $P_s\{\tau_s < \infty\}$ and $E_s\{\tau_s\}$ as follows:

	$P_s\{\tau_s < \infty\} < 1$	$P_s\{\tau_s < \infty\} = 1$
$E_s\{\tau_s\} < \infty$	not possible	positive recurrent
$E_s\{\tau_s\} = \infty$	transient	null recurrent

Often we do not distinguish positive and null recurrent states, and simply refer to them as recurrent. Sometimes we refer to recurrent states as *ergodic*. Note that s is recurrent if and only if

$$E_s\{\nu_s\} = \sum_{n=0}^{\infty} p^n(s|s) = \infty,$$

and transient if and only if $E_s\{\nu_s\} < \infty$.

We say that state $j \in S$ is *accessible* from state s ($s \to j$) if $p^n(j|s) > 0$ for some $n \geq 0$; otherwise we say that j is *inaccessible* from s. Note that $p^0(j|s) = 1$ if $s = j$ and 0 if $s \neq j$. State j *communicates* with state s if $s \to j$ and $j \to s$. Call a subset C of S a *closed set* if no state outside of C is accessible from any state in C. We call a closed set C *irreducible* if no proper subset of C is closed. Every recurrent state j is a member of some irreducible subset or class of S. Irreducible closed sets consisting of a single state are said to be *absorbing*. Closed irreducible sets may be regarded as distinct Markov chains.

We can partition the set of recurrent states (provided some are present) into disjoint closed irreducible sets C_k, $k = 1, 2, \ldots, m$ with m finite when S is finite and possibly infinite when S is countable. We can write S as $S = C_1 \cup C_2 \cup \cdots \cup C_m \cup T$, where T denotes the set of transient states that do not belong to any closed set.

After relabeling states if necessary, we can express any transition matrix P as

$$P = \begin{bmatrix} P_1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & P_2 & 0 & \cdot & \cdot & 0 \\ \cdot & & \cdot & & & \\ \cdot & & & \cdot & & \\ 0 & & & & P_m & 0 \\ Q_1 & Q_2 & \cdot & \cdot & Q_m & Q_{m+1} \end{bmatrix}, \tag{A.1}$$

where P_i corresponds to transitions between states in C_i, Q_i to transitions from states in T to states in C_i, and Q_{m+1} to transitions between states in T. Note that Q_i may be a matrix of zeros for some values of i. We refer to this representation as the *canonical form* of P. The algorithm in the next section provides an efficient method for classifying states and transforming a transition matrix to canonical form.

We use the expression *chain structure* to refer to the extent of the decomposition of the Markov chain into classes. We call a Markov chain *irreducible* if it consists of a single closed class. For *finite* S we use the expression *unichain* to refer to chains consisting of one closed irreducible set and a (possibly empty) set of transient states. Otherwise we say that a Markov chain is *multichain*.

Both recurrence and transience are class properties. This means that, in any closed irreducible class, all states are either transient, positive recurrent, or null recurrent. We note the following important results for finite-state Markov chains.

Theorem A.1. Suppose S is finite.

a. Then any recurrent state is positive recurrent.

b. There exists at least one positive recurrent class.

Consequently, in a finite irreducible chain, all states are positive recurrent. Note that countable-state Markov chains may have more than one closed irreducible set of transient states; for example, when $p(s+2|s) = 1$ for $s = 0, 1, 2, \ldots$, the odd and even integers each form closed irreducible sets of transient states.

We refer to the greatest common divisor of all n for which $p^n(s|s) > 0$ as the *period* of s. Whenever $p(s|s) > 0$, s has period 1. We refer to such a state as *aperiodic*. Periodicity is a class property; all states in a closed irreducible class have the same period. If it exceeds 1, we call the class *periodic*; otherwise we refer to it as *aperiodic*. Similarly we call an irreducible chain periodic or aperiodic depending on the periodicity of its states. For example, the Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{A.2}$$

is irreducible, and each state has period 2.

A.3 CLASSIFYING THE STATES OF A FINITE MARKOV CHAIN

This section gives the Fox and Landi (1968) labeling algorithm for determining the closed irreducible classes and transient states of a finite Markov chain. We use this algorithm for computing the limiting matrix of a Markov chain (Section A.4) and for determining the class structure of a Markov decision process. We follow Fox and Landi; however, we state the algorithm more formally.

The algorithm cleverly exploits the following properties of a transition probability matrix.

a. State s is absorbing if and only if $p(s|s) > 0$ and $p(j|s) = 0$ or all $j \neq s$.

b. If state s is absorbing, and $p(s|k) > 0$, then state k is transient.

c. If state s is transient, and $p(s|k) > 0$, then state k is transient.

d. If state i communicates with j, and j communicates with k, then i communicates with k.

e. If i communicates with j, then i is transient if j is transient, and i is recurrent if j is recurrent.

Note that the above properties only depend on the pattern of 0 and positive entries, so that if only a decomposition is required, the matrix P may be replaced by an incidence matrix B with $b(j|i) = 1$ if $p(j|i) > 0$, and 0 otherwise.

Assume that $S = \{1, 2, \ldots, N\}$. In the algorithm, $L(i)$ denotes the label assigned to state i; recurrent states receive a label R, transient states T, and unlabeled states O. The set U denotes unlabeled states, and $S(i)$ denotes states which have been identified as communicating with i. We use the set W to indicate the indices of rows and columns in an aggregate matrix constructed in step 5. A verbal description of the algorithm follows it formal statement.

The Chain Decomposition Algorithm

1. *Initialization.* Set $S(i) = \{i\}$, $L(i) = O$ for $i = 1, 2, \ldots, N$, $U = S$, and $W = S$.

2. *Preliminary identification.*

a. For each $i \in U$, if $p(i|i) > 0$ and $p(j|i) = 0$ for all $j \neq i$, set $L(i) = R$, and replace U by $U/\{i\}$.

b. If $U = \varnothing$, go to step 6; otherwise, for each j for which $L(j) = R$, if $p(j|i) > 0$ for any $i \in U$, set $L(i) = T$ and replace U by $U/\{i\}$.

3. *Stopping.* If $U = \varnothing$, go to step 6; otherwise, set $r = 0$ and go to step 4.

4. *Path formation.*

a. Select an $i \in U$, set $i_r = i$.

b. Choose a state $j \neq i_r$ for which $p(j|i_r) > 0$; set $i_{r+1} = j$.

i. If $L(i_{r+1}) = T$, set $L(i) = T$ for $i \in S(i_0) \cup \cdots \cup S(i_{r+1})$, replace U by $U/(S(i_0) \cup \cdots \cup S(i_r))$, and go to step 3.

ii. If $L(i_{r+1}) \neq T$, and $i_{r+1} = i_k$ for some k, $0 \leq k \leq r$, go to step 5. Otherwise replace r by $r + 1$, and go to step 4(b).

5. *Path aggregation.*

a. Replace $p(j|i_k)$ by $p(j|i_k) + p(j|i_{k+1}) + \cdots + p(j|i_r)$ for all $j \in W$.

b. Replace $p(i_k|i)$ by $p(i_k|i) + p(i_{k+1}|i) + \cdots + p(i_r|i)$ for all $i \in W$.

c. Replace $S(i_k)$ by $S(i_k) \cup \cdots \cup S(i_r)$.

d. Replace W by $W/\{i_{k+1}, \ldots, i_r\}$. (i.e. Delete rows and columns $i_{k+1}, \ldots, i_r$ from the matrix).

e. If $p(i_k|i_k) > 0$ and $p(j|i_k) = 0$ for all $j \in W$, do the following.
   i. Set $L(i) = R$ for all $i \in S(i_k)$, and replace U by $U/S(i_k)$.
   ii. If $k > 0$, set $L(i) = T$ for all $i \in S(i_0) \cup \cdots \cup S(i_{k-1})$, and replace U by $U/(S(i_0) \cup \cdots \cup S(i_{k-1}))$.
   iii. For $j \in U$, if $p(i_h|j) > 0$ for some h, $0 \le h \le k$, set $L(i) = T$ for all $i \in S(j)$, and replace U by $U/S(j)$.
   iv. Go to Step 3.

f. Set $r = k$ and go to step 4.

6. *Classification.* For each $i \in W$ for which $L(i) = R$, $S(i)$ is a closed irreducible class. All other states are transient.

We may paraphrase steps 4 and 5 of the algorithm as follows. *Pick an unclassified state and begin a path starting at it. Extend the path until it either identifies a transient state or it cycles; that is, it repeats a previous state in the chain. In the former case, classify all states on the path as transient; otherwise, combine all states in the cycle and determine whether they are recurrent.* Note that steps 5(a)–5(d) are a rather formal way of saying "replace row and columns i_k by the sum of entries in rows and columns $i_k, i_{k+1}, \ldots, i_r$, and delete these rows and columns from the matrix." The above algorithm works with the full matrix but ignores the "deleted" columns.

If instead of using the transition probability matrix, we use the incidence matrix for computation, we replace steps 5(a) and 5(b) by a Boolean "or" operation; that is, if $u = 0$ and $v = 0$, then $u + v = 0$, otherwise $u + v = 1$. Using the incidence matrix allows for efficient storage of large matrices and faster calculation.

The above algorithm requires $O(|S|^2)$ comparisons. Since there are S^2 pairs of indices, this algorithm is extremely efficient.

A.4 THE LIMITING MATRIX

Results in this section apply to both finite- and countable-state Markov chains. Let $\{A_n: n \ge 0\}$ be a sequence of matrices. We write $\lim_{n\to\infty} A_n = A$ if $\lim_{n\to\infty} A_n(j|s) = A(j|s)$ for each (s, j) in $S \times S$. When analyzing periodic Markov chains, limits of this form do not exist and instead we consider the Cesaro limit which we define as follows. We say that A is the *Cesaro limit* (of order one) of $\{A_n: n \ge 0\}$ if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} A_n = A,$$

and write

$$C - \lim_{N\to\infty} A_N = A$$

to distinguish this as a Cesaro limit. Sometimes, the ordinary limit is called a Cesaro limit of order zero.

Define the *limiting matrix* P^* by

$$P^* = C - \lim_{N\to\infty} P^N. \tag{A.3}$$

In component notation, where $p^*(j|s)$ denotes the $(j|s)$th element of P^*, this means that, for each s and j,

$$p^*(j|s) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} p^{n-1}(j|s)$$

where, as above, p^{n-1} denotes a component of P^{n-1} and $p^0(j|s)$ is a component of a an $S \times S$ identity matrix. When P is aperiodic, $\lim_{N\to\infty} P^N$ exists and equals P^*.

We define P^* in terms of the Cesaro limit to account for the nonconvergence of powers of transition matrices of periodic chains. For example, with P given by (A.2), $P^{2n} = I$ and $P^{2n+1} = P$, so that $\lim_{N\to\infty} P^N$ does not exist but

$$C - \lim_{N\to\infty} P^N = P^* = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}.$$

Doob (1953) and Chung (1960, p. 33) prove that the limit in (A.3) exists. Doob's proof assumes finite S and uses matrix methods, while Chung's uses probabilistic argument. In the next section, we provide an alternative derivation for finite-state Markov chains using matrix decomposition theory.

The limiting matrix P^* satisfies the following equalities

$$PP^* = P^*P = P^*P^* = P^*. \tag{A.4}$$

Since $(P^*)^2 = P^*$, $(I - P^*)^2 = (I - P^*)$, and $P^*(I - P^*) = 0$, P^* and $(I - P^*)$ are orthogonal projection matrices. Note further that, for finite chains, P^* is a stochastic matrix (has row sums equal to 1), while in countable-state chains, some or all row sums might be less than 1, for example, when $p(s + 1|s) = 1$ for $s = 0, 1, \ldots,$ and $p(j|s) = 0$ otherwise, then P^* is a zero matrix.

Since

$$E_s\{\nu_j\} = \sum_{n=1}^{\infty} p^{n-1}(j|s), \tag{A.5}$$

we may interpret $p^*(j|s)$ as the long-run fraction of time that the system occupies state j starting in state s. In aperiodic chains, $\lim_{N\to\infty} p^N(j|s)$ exists, in which case we may interpret $p^*(j|s)$ as the steady-state probability that the chain is in state j when it starts in state s.

We now show how to compute P^*. We begin with the following key result which holds for both finite and countable state chains. In it, q is a column vector and q^T denotes its transpose.

Theorem A.2. Suppose P is the transition matrix of a positive recurrent irreducible chain. Then the system of equations $q^T = q^TP$ subject to $\sum_{j\in S} q(j) = 1$, has a unique positive solution.

We call this solution the *stationary distribution* of P. Since $P^*P = P^*$, when P is recurrent and irreducible, P^* has identical rows and we can write

$$P^* = eq^T, \tag{A.6}$$

where q is the stationary distribution of P, e denotes a column vector of ones, and e^T its transpose.

The chain structure of the Markov chain determines the form of P^*. For P in canonical form (A.1),

$$P^* = \begin{bmatrix} P_1^* & 0 & 0 & \cdot & \cdot & 0 \\ 0 & P_2^* & 0 & \cdot & \cdot & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ 0 & & & & P_m^* & 0 \\ Q_1^* & Q_2^* & \cdot & \cdot & Q_m^* & 0 \end{bmatrix}, \tag{A.7}$$

where P_i^* is the limiting matrix of P_i. For each i, we find P_i^* by solving $q_i P_i = q_i$ subject to $q_i^T e_i = 1$, where e_i is a column vector of ones with $|C_i|$ components. Then, following (A.5), we set $P_i^* = q_i e_i^T$.

We determine $Q_1^*, \ldots, Q_m^*$ as follows. Assume P is in canonical form and write it as

$$P = \begin{bmatrix} U & 0 \\ V & W \end{bmatrix},$$

where U corresponds to transitions between recurrent states, W to transitions between transient states, and V to transitions from transient to recurrent states.

We use the following results [Çinlar (1975, p. 144–146) and Berman and Plemmons (1979, p. 223)] which follow from (A.5) and the definition of transience.

Proposition A.3. When S is finite

a. the spectral radius of W, $\sigma(W) < 1$, and

b. $(I - W)^{-1}$ exists and satisfies

$$(I - W)^{-1} = \sum_{n=0}^{\infty} W^n.$$

Theorem A.4. Let P be in canonical form. Then

$$Q_i^* = (I - Q_{m+1})^{-1} Q_i P_i^*.$$

Proof. Write P as

$$P = \begin{bmatrix} U & 0 \\ V & W \end{bmatrix},$$

and the corresponding limiting matrix as

$$P^* = \begin{bmatrix} U^* & 0 \\ V^* & 0 \end{bmatrix}.$$

Since $PP^* = P^*$,

$$VU^* + WV^* = V^*,$$

so that

$$V^* = (I - W)^{-1}VU^*,$$

where Proposition A.3 ensures the existence of $(I - W)^{-1}$. From (A.1) and (A.7), it follows that

$$Q_i^* = (I - Q_{m+1})^{-1}Q_iP_i^*. \qquad \square$$

Observe that for unichain P with closed irreducible recurrent class C_1 and transient states T, P^* has equal rows and

$$p^*(j|s) = \begin{cases} q(j) & j \in C_1 \\ 0 & j \in T, \end{cases}$$

where q satisfies $qP_1 = q$ subject to $\sum_{j \in C_1} q(j) = 1$.

In summary, to compute P^*, do the following:

1. Use the Fox-Landi algorithm to transform P to canonical form.
2. For each P_i, solve $q_iP_i = q_i$ subject to $q_i^Te_i = 1$, and set $P_i^* = q_ie_i^T$.
3. Set $Q_i^* = (I - Q_{m+1})^{-1}Q_iP_i^*$.

A.5 MATRIX DECOMPOSITION, THE DRAZIN INVERSE, AND THE DEVIATION MATRIX

The results in this section rely on decomposition of the transition matrix and are often referred to as "the modern theory of Markov chains." We follow Lamond and Puterman (1989), who draw on Campbell and Meyer (1979). Berman and Plemmons (1979) also cover some of this material and state

> "...virtually everything that one would want to know about the chain can be determined by investigating a certain group inverse (the Drazin Inverse) and a limiting matrix... their introduction into the theory provides practical advantages over the more classical techniques and serves to unify the theory to a certain extent."

It appears that the results have only been established for finite-state Markov chains, therefore, *we assume finite S in this section*. An important research contribution would be to generalize these results to nonfinite S. The following theorem is fundamental. We denote the spectral radius of P by $\sigma(P)$ (see Sec. C.2 of Appendix C). Note that in the following results, our approach to matrix decomposition differs from that used to establish (A.1).

Theorem A.5. Suppose S is finite and P has m recurrent classes.

a. Then 1 is an eigenvalue of P with (algebraic and geometric) multiplicity m and with m linearly independent eigenvectors.

b. There exists a nonsingular matrix W for which

$$P = W^{-1}\begin{bmatrix} Q & 0 \\ 0 & I \end{bmatrix} W, \tag{A.8}$$

where I is an $m \times m$ identity matrix and Q is an $(|S| - m) \times (|S| - m)$ matrix with the following properties.

1. 1 is not an eigenvalue of Q.

2. $\sigma(Q) \leq 1$ and if all recurrent subchains of P are aperiodic, $\sigma(Q) < 1$.

3. $(I - Q)^{-1}$ exists.

4. $\sigma(I - Q) = \sigma(I - P)$.

c. There exists a unique matrix P^* which satisfies (A.4) and $P_i^* > 0$ for $i = 1, 2, \ldots, m$. It may be represented by

$$P^* = W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W. \tag{A.9}$$

Note in Theorem A.5(b) that even when $\sigma(Q) = 1$, the fact that 1 is not an eigenvalue of Q implies that 0 is not an eigenvalue of $I - Q$, so that $(I - Q)^{-1}$ exists. If $\sigma(Q) = 1$, $\lim_{n \to \infty} Q^n$ need not exist, so that

$$(I - Q)^{-1} = C - \lim_{N \to \infty} \sum_{n=0}^{N-1} Q^n.$$

We use Theorem A.5 to establish the following result.

Theorem A.6. Suppose S is finite. Then the limit in (A.3) exists and equals P^*.

Proof. From (A.8),

$$P^n = W^{-1}\begin{bmatrix} Q^n & 0 \\ 0 & I \end{bmatrix} W,$$

so that

$$\frac{1}{N}\sum_{n=0}^{N-1} P^n = W^{-1}\begin{bmatrix} \frac{1}{N}\sum_{n=0}^{N-1} Q^n & 0 \\ 0 & I \end{bmatrix} W.$$

The nonsingularity of $I - Q$ implies

$$\sum_{n=0}^{N-1} Q^n = (I - Q^N)(I - Q)^{-1}. \tag{A.10}$$

Because $\sigma(Q) \leq 1$, Q^N is bounded, so that (A.10) implies $\Sigma_{n=0}^{N-1} Q^n$ is bounded and

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} Q^n = 0.$$

Therefore, from Theorem A.5(c), we conclude that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=0}^{N-1} P^n = W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W = P^*. \qquad \square$$

In Markov decision process theory we frequently solve systems of the form

$$(I - P)v = r.$$

Since $(I - P)$ is singular, this system does not have a unique solution. We use the matrix representation in (A.8) to distinguish a solution which plays a key role in the theory of average reward Markov decision processes. From (A.8),

$$I - P = W^{-1}\begin{bmatrix} I - Q & 0 \\ 0 & 0 \end{bmatrix} W, \tag{A.11}$$

where 0 in the lower-right-hand corner denotes an $m \times m$ matrix of zeros. As a consequence of Theorem A.1(b), $m \geq 1$, so that $(I - P)^{-1}$ does not exist. Instead we seek a *generalized inverse* of $(I - P)$.

Given a matrix B with representation

$$B = W^{-1}\begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix} W,$$

in which C is nonsingular, we define the matrix $B^\#$ by

$$B^\# = W^{-1}\begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix} W. \tag{A.12}$$

It is easy to see that $B^{\#}$ satisfies

$$B^{\#}BB^{\#} = B^{\#}, \qquad BB^{\#} = B^{\#}B, \qquad \text{and} \qquad BB^{\#}B = B. \tag{A.13}$$

We call a matrix which satisfies (A.13) a *Drazin inverse* or *group inverse* of B. It is a particular generalized inverse of B. We now derive a representation for $(I - P)^{\#}$ and show how the matrix decomposition theory may be used to provide direct proofs of many key matrix identities. The following theorem defines the matrices Z_P and H_P.

Theorem A.7.

a. The matrix $(I - P + P^*)$ is nonsingular, with inverse denoted by Z_P.

b. The Drazin inverse of $(I - P)$, denoted by H_P, satisfies

$$(I - P)^{\#} = (I - P + P^*)^{-1}(I - P^*) \equiv H_P. \tag{A.14}$$

c.

$$H_P = C - \lim_{N\to\infty} \sum_{n=0}^{N-1} (P^N - P^*), \tag{A.15}$$

where the ordinary limit exists whenever P is aperiodic.

Proof. Combining (A.8) and (A.9) yields

$$I - P + P^* = W^{-1}\begin{bmatrix} I - Q & 0 \\ 0 & I \end{bmatrix} W.$$

Theorem A.5(b) implies that the matrix on the right-hand side is nonsingular, so that (a) follows.

To derive (b), note that the definition of the Drazin inverse implies

$$\begin{aligned}(I - P)^{\#} &= W^{-1}\begin{bmatrix} (I - Q)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W \\ &= W^{-1}\begin{bmatrix} (I - Q)^{-1} & 0 \\ 0 & I \end{bmatrix} W - W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W \\ &= (I - P + P^*)^{-1} - P^* = Z_P - P^*.\end{aligned}$$

Expression (A.14) follows by noting that $P^* = Z_P P^*$.

The proof of (c) is slightly more involved, and we refer the reader to Lamond and Puterman (1989) for details. It relies on the identity

$$(P - P^*)^n = P^n - P^*, \tag{A.16}$$

which holds for $n \geq 1$. □

We refer to H_P as the *deviation matrix* and Z_P as the *fundamental matrix*. These play a key role in average reward MDP theory and computation. The following useful identities may be derived by similar methods to the proof of Theorem A.7(b) or directly from the power series expansion for $[I - (P - P^*)]^{-1}$:

$$(I - P)H_P = H_P(I - P) = I - P^*, \tag{A.17}$$

$$H_P P^* = P^* H_P = 0, \tag{A.18}$$

$$H_P = Z_P - P^*, \tag{A.19}$$

$$Z_P P^* = P^*, \tag{A.20}$$

$$P^* = I - (I - P)(I - P)^{\#}. \tag{A.21}$$

For aperiodic chains, Theorem A.7(c) implies that

$$H_P = \lim_{N\to\infty} \left[\sum_{k=0}^{N-1} P^k - NP^* \right],$$

so that we may interpret $H_P(j|s)$ as the difference between the expected number of visits to state j starting in s, and the expected number of visits to j starting in s for a chain with transition probability matrix P^*. Note that the Cesaro limit in (A.15) is equivalent to

$$H_P = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \sum_{i=0}^{k} (P^k - P^*). \tag{A.22}$$

The following example illustrates some of these calculations and the need for the Cesaro limit when defining H_P.

Example A.1. Suppose

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Decompose P as in (A.8) as

$$P = W^{-1} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} W,$$

where

$$W = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad W^{-1} = \begin{bmatrix} 1 & \frac{1}{2} \\ -1 & \frac{1}{2} \end{bmatrix}.$$

In the notation above $Q = -1$ and $m = 1$. Note that

$$P^* = W^{-1} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} W = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

and

$$I - P = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = W^{-1}\begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}W,$$

so that

$$H_P = (I - P)^{\#} = W^{-1}\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{bmatrix}W = \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

Since $P^{2n} - P^* = I - P^*$ and $P^{2n+1} = P - P^*$,

$$\lim_{N\to\infty} \sum_{n=0}^{N-1} (P^N - P^*)$$

does not exist in the ordinary sense and we require a Cesaro limit to define H_P as a limit of partial sums of matrix powers.

A.6 THE LAURENT SERIES EXPANSION OF THE RESOLVENT

The Laurent series expansion provides a powerful tool for analyzing undiscounted Markov decision processes. For $\rho > 0$, define the *resolvent* of $P - I$, (Sec. C.2) denoted R^ρ, by

$$R^\rho \equiv (\rho I + [I - P])^{-1}. \tag{A.23}$$

Letting $\lambda = (1 + \rho)^{-1}$, we see that

$$(I - \lambda P) = (1 + \rho)^{-1}(\rho I + [I - P]).$$

When $0 \le \lambda < 1$, $\sigma(\lambda P) < 1$, so that $(I - \lambda P)^{-1}$ exists. Consequently the above identity implies that if $\rho > 0$, R^ρ exists. In our analysis of MDPs in Chap. 10, we will often appeal to the following easily derived identities:

$$(I - \lambda P)^{-1} = (1 + \rho)R^\rho$$

and

$$R^\rho = \lambda(I - \lambda P)^{-1}.$$

We use the matrix decomposition methods of the previous section to derive a series expansion for R_ρ.

Theorem A.8. For $0 < \rho < \sigma(I - P)$,

$$R^\rho = \rho^{-1}P^* + \sum_{n=0}^{\infty} (-\rho)^n H_P^{n+1} \tag{A.24}$$

Proof. Let Q be defined through (A.8), and let $B = I - Q$. Then

$$\rho I + I - P = W^{-1}\begin{bmatrix} \rho I + B & 0 \\ 0 & \rho I \end{bmatrix} W,$$

so that

$$\begin{aligned} R^{\rho} &= W^{-1}\begin{bmatrix} (\rho I + B)^{-1} & 0 \\ 0 & \rho^{-1} I \end{bmatrix} W \\ &= \rho^{-1} W^{-1}\begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} W + W^{-1}\begin{bmatrix} (\rho I + B)^{-1} & 0 \\ 0 & 0 \end{bmatrix} W. \end{aligned} \tag{A.25}$$

From Theorem A.5(c), the first term in (A.25) equals $\rho^{-1} P^*$. Since

$$(\rho I + B)^{-1} = \left(I + \rho B^{-1}\right)^{-1} B^{-1},$$

and whenever $\sigma(\rho B^{-1}) = [\rho/\sigma(I - Q)] < 1$ or $\rho < \sigma(I - Q) = \sigma(I - P)$,

$$(\rho I + B)^{-1} = \sum_{n=0}^{\infty} (-\rho)^n (B^{-1})^n,$$

the second term in (A.25) becomes

$$\sum_{n=0}^{\infty} (-\rho)^n W^{-1}\begin{bmatrix} (I - Q)^{-n} & 0 \\ 0 & 0 \end{bmatrix} W = \sum_{n=0}^{\infty} (-\rho)^n H_P^{n+1},$$

from which we conclude the result. □

Example A.1 (ctd.). Since $\sigma(I - Q) = 2$, it follows from Theorem A.7 that, for $\rho < 2$,

$$R_{\rho} = \rho^{-1}\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} + \sum_{n=0}^{\infty} (-\rho)^n \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}^{n+1}$$

A.7 A. A. MARKOV

The subject of this book derives its name from the Russian mathematician Andrei Andreivich Markov, who introduced the concept of a Markov chain in the early 20th century. His work in this area concerned what we now refer to as finite-state, irreducible, aperiodic Markov chains. In addition to considerable theoretical research on these models, he applied them to analyze Bernoulli and LaPlace's urn problem and the pattern of vowels and consonants in Russian novels. He did not develop Markov decision processes; this subject evolved much later.

A. A. Markov was born on July 14, 1856 in the province of Ryazan, and moved to St. Petersburg in the 1860's. In high school, he was a mediocre student but excelled in mathematics. He enrolled at the University of St. Petersburg in 1874 and was greatly influenced by the emminent mathematician P. L. Chebyshev. Markov received a gold medal for his paper "On the integration of differential equations with the aid of continued fractions," written prior to his graduation in 1878. Markov was awarded his doctoral degree in 1884 and went on to a distinguished career as Professor at The University of St. Petersburg. He became an Academician in the Russian Academy of Sciences in 1896 and died on July 20, 1922. His son, also named Andrei Andreivich Markov (1903–1979), was a emminent mathematician who made significant contributions to mathematical logic and the theory of algorithms.

In addition to his seminal work on dependent processes, which inspired Kolmogorv's development of the theory of stochastic processes, Markov studied important problems in probability and statistics. He formulated many fundamental concepts in probability theory while completing Chebyshev's proof of the central limit theorem. He was interested in laws of large numbers and developed them for sequences of dependent random variables. His justification of the method of least squares, now referred to as the Gauss-Markov Theorem, led to the modern statistical ideas of efficiency and unbiasedness. His writings in all areas emphasized his mathematical rigor.

Markov was also a tenacious social activist. He was often referred to (in rough translation) as "Andrei the irrepressible, who does not pull any punches." In 1902, he renounced all government honors when the Tsarist government expelled writer A. M. Gorky from the Russian Academy.