Appendix A

Notation and Conventions

A.1 General math notation

- $:=$ for definition.
- $\mathbb{R}^N$ for N-dimensional Euclidean space. $\mathbb{R}^N_+$ restricts it to non-negative vectors.
- $\mathbb{Z}^N$ for N-dimensional integer lattice. $\mathbb{Z}^N_+$ restricts the lattice to non-negative integers.
- "sup" for supremum and "max" for maximum.
  - The supremum of a real-valued function $f(w)$ over a set W refers to the smallest upper bound of the elements in the set $\{f(w) \mid w \in W\}$. Importantly, the supremum need not be an element of that set.
  - The maximum denotes the largest element of a set and must be contained in that set.
  - As a convention, sup is used when a set is non-finite and max is reserved for finite sets. When it is shown that suprema are attained, sup is replaced by max.
  - inf and min are the analogues to sup and max, respectively, when refering to lower bounds.
- $\arg\max\{\}$ and $\arg\min\{\}$ refer to the set of arguments that maximize or minimize the expression in braces.

- For a real-valued function $f(w)$ and a set W,

$$\arg\max_{w \in W} f(w) := \{w^* \in W \mid f(w^*) \geq f(w) \text{ for all } w \in W\}.$$

- arg min is defined similarly as arg max, except with respect to minimization.
- Note that arg max and arg min may return sets consisting of multiple elements. Only if there is a unique maximizer or minimizer do they return a single element.

- $u^+ = \max\{u, 0\}$, given a scalar u.
- $\mathbf{e}$ is a vector with all components equal to one.
- $\mathbf{e}_s$ is a vector with a 1 in the s-th component and zeroes elsewhere.
- $\mathbf{I}$ is the identity matrix.
- $\mathbf{0}$ is a vector or matrix with all components equal to zero.
- $\mathbf{x}^\mathsf{T}$ denotes the transpose of the vector $\mathbf{x}$. The same operator is applied to matrices.
- $\emptyset$ is the empty set.
- $|A|$ is the number of elements in a set A.
- $I_A(x)$ is the indicator function of the set A. It equals 1 if $x \in A$ and 0 otherwise.
- $\mathcal{S}_N$ is the unit simplex for N-dimensional vectors. That is,

$$\mathcal{S}_N := \left\{ (x_1, \ldots, x_N) \;\middle|\; \sum_{i=1}^{N} x_i = 1, x_i \geq 0, i = 1, \ldots, N \right\}.$$

- $o(1)$ is a generic expression for a function $f(n)$ on the non-negative integers, that converges to zero as $n \to \infty$. It is used when one is concerned about the limiting property of a function and not its particular form. More generally a function is $o(g)$ if $f(n)/g(n) \to 0$.
- A^B denotes the set of all functions mapping elements of B to elements of A. Equivalently, it can be viewed as the Cartesian product $A \times A \times \ldots \times A$ taken $|B|$ times.
- $A \setminus B$ denotes the set of elements in set A that are not in set B.
- $\sigma(\mathbf{B})$ denotes the spectral radius (largest eigenvalue in absolute value) of the square matrix $\mathbf{B}$.

- V denotes a vector space of real-valued vectors, usually of dimension $|S|$.
- $\nabla_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta})$ denotes the gradient of a real-valued function $g(\boldsymbol{\beta})$ respect to the components of the vector $\boldsymbol{\beta}$.
- $\leftarrow$ denotes the assignment of value to a variable in an algorithm. $x \leftarrow 0$ means "x gets the value of 0" or equivalently that "x equals 0".

A.2 Notation specific to MDPs

- $p^{\pi}(\cdot)$ is the probability of a sequence of states and actions under π
- $P^{\pi}[\cdot|\cdot]$ is the conditional probability of an event under policy π.
- $p_d(j|s)$ is the one-step conditional probability of a transition to state j beginning in state s and using Markovian decision rule d.
- $\mathbf{P}_d$ is the transition probability matrix corresponding to Markovian decision rule d.
- $p_d^n(j|s)$ is the n-step conditional probability of being in state j in n steps, beginning in state s and using Markovian decision rule d. It is the (s, j)-th component of the matrix $\mathbf{P}_d^n$.
- $\rho(s)$ is an initial state distribution for a Markov chain.
- $w_d(a|s)$ is the probability that randomized decision rule d chooses action a in state s. This notation may include a superscript n if the probability depends on the epoch n (e.g., in a finite horizon model).
- $\mathbf{r}_d$ is the vector of (expected) rewards corresponding to Markovian decision rule d.
- $E^{\pi}[\cdot]$ is the unconditional expectation of a random variable under policy π.
- $E^{\pi}[\cdot|\cdot]$ is the conditional expectation of a random variable under policy π.
- $\mathbf{v}$ is a $|S|$-dimensional real-valued vector representing or approximating a value function. Value functions are usually distinguished as policy value functions or optimal value functions.
- d^{∞} denotes the stationary policy that uses decision rule d at every decision epoch.
- $d_{\mathbf{v}}$ denotes a $\mathbf{v}$-greedy decision rule. A *greedy* decision rule maximizes $L_d\mathbf{v}$ (or similar expression) over some set of decision rules. A *greedy* policy is any stationary policy composed of greedy decision rules.

- "c-max" is the component-wise maximum. Applied to a set of vectors, each component is maximized independently of the others, meaning that the vector that achieves the maximum may be different for each component.

- $\arg \text{c-max}\{\}$ returns the set of arguments that achieve the c-max. In the MDP case, c-max is applied to the set of deterministic decision rules, corresponding to the Cartesian product of all actions in all states. Hence, $\arg \text{c-max}$ will return one of the decision rules in that set.

- L is the Bellman operator. Its form varies by optimality criterion.

- L_d is the operator that applies decision rule d for one period. This varies by optimality criterion.

- B is the operator on a value function $\mathbf{v}$ defined as $L\mathbf{v} - \mathbf{v}$.

- $x(s, a)$ are dual variables in the linear programming formulation of an MDP.

- $\underline{\mathbf{v}}$ is the minimum of all components of $\mathbf{v}$.

- $\overline{\mathbf{v}}$ is the maximum of all components of $\mathbf{v}$.

- $v(s; \boldsymbol{\beta})$ denotes an approximate value function in state s parameterized by weight vector $\boldsymbol{\beta}$.

- $q(s, a; \boldsymbol{\beta})$ denotes an approximate state-action value function when the state-action pair equals (s, a), parameterized by the weight vector $\boldsymbol{\beta}$.

- Δ is an absorbing state (or set of absorbing states) in a Markov chain.

- N_Δ is a random variable representing the time to reach absorbing state(s) Δ in an expected total reward model.

A.3 Conventions

- Random variables are denoted by capital letters and their values by lower case letters.

- Vectors and matrices are bolded, while their components are not bolded. A vector written as $\mathbf{x} = (x_1, \ldots, x_n)$ should be considered a column vector, unless it is transposed explicitly. Thus, $\mathbf{x}^\mathsf{T}$ is a row vector.

- Subscripts and superscripts are defined based on the context.

  - Subscripts denote epochs for value functions, state-action value functions, rewards and transition probabilities.

  - Superscripts denote epochs for realizations of states, actions and histories.
  - Subscripts denote different states, actions or histories, within the sets of states, actions or histories, respectively.
  - Superscripts on a value function denote a specific policy (π) or an optimal policy ($*$)

- Hat denotes approximation, typically applied to a value function or parameter vector.
- Square brackets are reserved for probability and expectation.

A.4 Abbreviations

- MDP: Markov decision process
- POMDP: Partially observable Markov decision process
- HR: History-dependent and randomized
- HD: History-dependent and deterministic
- MR: Markovian and randomized
- MD: Markovian and deterministic
- MSE: Mean squared error
- RMSE: Root mean squared error (the square root of the MSE)
- TD: Temporal differencing
- PI: Policy iteration
- MPI: Modified policy iteration
- VI: Value Iteration
- LP: Linear Program

Appendix B

Markov Chains

Markov chain theory underlies Markov decision process models, especially under the average reward and expected total reward criteria. Markov chains have a long history. We especially like the seminal book Kemeny and Snell [1960] for its transparent approach and innovative applications. It provides the basis for the vanishing discount approach developed by Blackwell [1962] to analyze average reward models. Chapter 4 of Gallagher [1996] also provides an insightful discussion of the use of eigenvalues and eigenvectors to explore limiting properties of powers of transition probability matrices and of the challenges faced when analyzing countable state Markov chains.

Markov chains are now widely applied and provide the basis for Google's PageRank algorithm, speech recognition software and many reinforcement learning applications.

B.1 What is a finite Markov chain?

Let S denote a finite set of states and let $\mathcal{X} = (X_n : n = 0, 1, 2, \ldots)$ denote a sequence of random variables with values in $S = \{s_1, \ldots, s_m\}$[1]. Then $\mathcal{X}$ is a *Markov chain* if

$$P(X_n = s^n | X_{n-1} = s^{n-1}, X_{n-2} = s^{n-2}, \ldots, X_0 = s^0) = P(X_n = s^n | X_{n-1} = s^{n-1}) \tag{B.1}$$

for all $n = 0, 1, \ldots$ and $s^k \in S$ for $k = 0, 1, \ldots$.

Equation (B.1) is known as the *Markov property*, which can be stated succinctly as *"the future conditional on the present is independent of the past"*.

A Markov chain $\mathcal{X}$ is *stationary* or *homogeneous* whenever $P(X_n = s^n | X_{n-1} = s^{n-1})$ is independent of n. In this case, define for $k = 0, 1, \ldots$, the *(one-step) transition probability*

$$p(j|s) := P(X_k = j | X_{k-1} = s) \tag{B.2}$$

and the *n-step transition probability* by

$$p^n(j|s) := P(X_{n+k} = j | X_k = s). \tag{B.3}$$

[1]Recall that subscripts are used to denote different states and superscripts to denote the state visited at a decision epoch.

We emphasize that this quantity does not equal $(p(j|s))^n$.

Let $\mathbf{P}$ denote the $|S| \times |S|$ matrix with (s, j)-th component $p(j|s)$. Using the law of total probability n-times shows that this matrix provides provides a convenient way of computing the n-step transition probabilities. It follows that (s, j)-th component of the matrix product $\mathbf{P}^n$ equals $p^n(j|s)$. Also, $\mathbf{P}^0 = \mathbf{I}$.

For $s \in S$ define the initial distribution $\rho^0(s) := \rho(s) := P(X_0 = s)$ and the unconditional distribution $\rho^n(s) := P(X_n = s)$. Then again by the law of total probability, the unconditional distribution

$$\rho^n(s) = \sum_{j \in S} \rho(j) p^n(s|j),$$

which in matrix-vector notation can be written as $\boldsymbol{\rho}^n = \boldsymbol{\rho}\mathbf{P}^n$.

Finally, for a real valued function[2] $r(\cdot)$ on S define the conditional expectation

$$E_s[r(X_n)] := E[r(X_n)|X_0 = s] = \sum_{j \in S} r(j) p^n(j|s),$$

which in matrix-vector notation can be written $E[r(X_n)|X_0 = s] = \mathbf{P}^n\mathbf{r}(s)$. Note that the expressions $\mathbf{P}^n$ and $\mathbf{P}^n\mathbf{r}$ will appear throughout this book when using matrices and vectors to analyze stationary policies because of the following important fact.

> A stationary policy d^∞ generates a Markov chain with transition probability matrix $\mathbf{P}_d$ and reward vector $\mathbf{r}_d$.

Consequently, given a stationary policy, each of the models in Chapter 3 provides an application of a Markov chain.

B.2 Classifying states

The limiting behavior of the Markov chain depends on relationships between its states. State j is said to be *accessible* from state s, written $s \rightsquigarrow j$, if $p^n(j|s) > 0$ for some $n \geq 0$. Otherwise j is *inaccessible* from S. If $s \rightsquigarrow j$ and $j \rightsquigarrow s$ then states j and s are said to *communicate*, which is written as $s \leftrightsquigarrow j$.

State $s \in S$ is said to be:

- *recurrent* if the time to return to state s is finite with probability one. This occurs if state s is accessible from all states that are accessible from s. That is, if $s \rightsquigarrow j$, then $j \rightsquigarrow s$.
- *absorbing* whenever $p(s|s) = 1$. This means that once the chain visits state s, it remains there forever.

[2] A Markov chain together with a reward function is often referred to as a *Markov reward process*.

- *transient* if the time to return to state s is finite with probability *less than* one, or equivalently. if there exists a state j for which $s \rightsquigarrow j$ but s is not accessible from j. That is after a transition to j, $p^n(s|j) = 0$ for $n = 0, 1, \ldots$.

- *periodic with period* m if the greatest common divisor of $\{n = 0, 1, \ldots \,|\, p^n(s|s) > 0\}$ is m.

- *aperiodic* if 1 is the only common divisor of $\{n = 0, 1, \ldots \,|\, p^n(s|s) > 0\}$ is 1.

Note that if s and j communicate, they will be classified in the same way. That is if $s \leftrightsquigarrow j$ and s is recurrent or periodic with period m, than j is recurrent or periodic with period m. Moreover starting in a recurrent state s, the expected number of returns to state s is infinite, while if s is transient, the expected number of returns to s is finite.

Example B.1. As a simple illustration of a periodic Markov chain, consider one with transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ p & 0 & 1-p \\ 0 & 1 & 0 \end{bmatrix}. \tag{B.4}$$

where $S = \{s_1, s_2, s_3\}$ and $0 < p < 1$. Since

$$\mathbf{P}^2 = \begin{bmatrix} p & 0 & 1-p \\ 0 & 1 & 0 \\ p & 0 & 1-p \end{bmatrix} \tag{B.5}$$

it follows that $p^2(s_i|s_i) > 0$ for $i = 1, 2, 3$. For $n \geq 1$, $\mathbf{P}^{2n+1} = \mathbf{P}$ and $\mathbf{P}^{2(n+1)} = \mathbf{P}^2$ so that it follows that each state is periodic with period 2.

Observe that when $p = 0$ or $p = 1$ the chain behaves differently.

Periodicity represents an "edge case" that complicates many analyses and necessitates averaging to ensure convergence (see Section B.4).

B.3 Classes and class structure

Call a subset C of S *closed* if no state outside of C is accessible from any state in C. Moreover C is *irreducible* if no proper subset of C is closed. An irreducible closed set C that consists of a single element is said to be *absorbing*.

A Markov chain can be partitioned into closed irreducible subsets of states $C_1, \ldots, C_M$, in which all states in each C_i are recurrent, and a set of transient states T. If the Markov chain starts at a state in some C_k, it remains in C_k forever, however if it starts in T it eventually leaves it and ends up in some C_k. Obviously when S is finite, the Markov chain contains at least one closed class.

Classification depends on the arrangement of 0 entries in $\mathbf{P}$. For example, consider the transition probability matrix with $S = \{s_1, s_2, s_3, s_4, s_5\}$,

$$\mathbf{P} = \begin{bmatrix} a & b & 0 & 0 & 0 \\ c & d & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & e & 0 & 0 & f \\ g & 0 & h & u & v \end{bmatrix},$$

where lower case letters denote non-zero probabilities with row sums equal to one. Consequently $C_1 = \{s_1, s_2\}$, $C_2 = \{s_3\}$ and $T = \{s_4, s_5\}$. Moreover s_3 is absorbing. Note that starting in state s_4, the system can either jump to C_1 in one step or remain in T for a few steps and then jump to either C_1 or C_2. Observe also the zeroes in rows 1-3 of columns 4 and 5.

A matrix partitioned as above is said to be in *canonical form*. Any transition matrix can be converted to the canonical form[3]

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & \dots & & \mathbf{P}_M & \mathbf{0} \\ \mathbf{Q}_1 & \mathbf{Q}_2 & \dots & & \mathbf{Q}_M & \mathbf{Q}_{M+1} \end{bmatrix},$$

where $\mathbf{P}_k$ corresponds transitions between states in C_k and Q_{M+1} to transitions between states in T.

B.3.1 Chain structure

Many results in this book, especially in Chapters 6 and 7, depend on the structure of Markov chains corresponding to stationary policies. A Markov chain on S is said to be:

- *regular*[4] if S is a single closed aperiodic class.
- *recurrent* or *ergodic* if it consists of a single closed class. A recurrent chain may be either periodic or aperiodic. When it is aperiodic, it is regular.
- *unichain* if it consists of a single closed class and a non-empty set T of transient states, and

[3] The Fox-Landi algorithm does this, see Section A.3 in Puterman [1994].

[4] Kemeny and Snell [1960] refer to a Markov chain as *regular* if $\mathbf{P}^N$ has all positive entries for some N. Clearly that is equivalent to our notion.

- *multi-chain* if it consists of two or more closed classes and possibly some transient states.

The analysis in the text will focus primarily on models in which all stationary policies correspond to Markov chains that are either recurrent or unichain. In the Markov decision process context, they will be referred to as recurrent or unichain models.

B.4 Limiting behavior

From a Markov decision process perspective, classification of the limiting behavior of $\mathbf{P}$ is extremely important. A sequence of matrices $\mathbf{A}_n$ converges to a matrix $\mathbf{A}$ if for each s and j in S, its components $a_n(s,j)$ converges to the components of $\mathbf{A}$, $a(s,j)$[5]. This is sometimes referred to as component-wise convergence[6].

The following important result regarding limiting behavior is stated without proof.

[5]When the matrices represent transition probability matrices, these terms will be written equivalently as $p_n(j|s)$ and $p(j|s)$, respectively.

[6]This is equivalent to convergence in norm when S is finite.

Theorem B.1. Let $\mathbf{P}$ be a transition probability matrix of a Markov chain on a finite state space S.

1. Then the limit
$$\mathbf{P}^* := \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{P}^{n-1} \tag{B.6}$$
exists.

2. $\mathbf{P}^*$ is a transition probability matrix and satisfies
$$\mathbf{P}^*\mathbf{P} = \mathbf{P}\mathbf{P}^* = \mathbf{P}^*\mathbf{P}^* = \mathbf{P}^*. \tag{B.7}$$

3. When $\mathbf{P}$ is recurrent,
$$\mathbf{P}^* = \mathbf{e}\mathbf{q}^\mathsf{T} \tag{B.8}$$
where $\mathbf{q}$ satisfies $\mathbf{q}^\mathsf{T} = \mathbf{q}^\mathsf{T}\mathbf{P}$ subject to $\mathbf{q}^\mathsf{T}\mathbf{e} = 1$[a]. Moreover each entry of $\mathbf{q}$ is strictly positive.

4. If $s \in T$, then $p^*(s|j) = 0$ for all $j \in S$.

5. When $\mathbf{P}$ is regular,
$$\mathbf{P}^* = \lim_{N\to\infty} \mathbf{P}^n. \tag{B.9}$$

6. When $\mathbf{P}$ is regular, the convergence in (B.9) is exponentially fast. That is for each j and s in S, there exists constants $k > 0$ and $0 \leq a < 1$ for which
$$|p^n(j|s) - p^*(j|s)| < ka^n. \tag{B.10}$$

[a] Recall that $\mathbf{e}$ denotes a column vector of onnes so that $\mathbf{q}^\mathsf{T}\mathbf{e}$ is a scalar and $\mathbf{e}\mathbf{q}^\mathsf{T}$ is a $|S| \times |S|$ matrix with all rows equal to $\mathbf{q}$.

Some comments on the significance of the above result follow:

1. The representation of $\mathbf{P}^*$ in (B.6) is particularly relevant in the Markov decision process setting. In this case $\sum_{n=0}^{N-1} \mathbf{P}^n\mathbf{r}$ denotes the expected total reward over N decision epochs so that $\mathbf{P}^*\mathbf{r}$ equals the *limiting average expected reward.*

2. As a consequence of (B.9), in a regular chain, $\mathbf{P}^*\mathbf{r}$ can be interpreted as the *steady state* reward. Note that in a periodic chain such as in (B.5), this limit does not exist. However the limit in (B.6) always exists (see Example B.2).

3. Part 3 provides an approach for computing $\mathbf{P}^*$ for a recurrent chain, that is by solving the system of equations
$$q(s) = \sum_{j\in S} q(j)p(s|j)$$

subject to $\sum_{j \in S} q(j) = 1$. In this case

$$\mathbf{P}^* = \begin{bmatrix} q(s_1) & q(s_2) & \dots & q(s_M) \\ . & . & & . \\ . & . & & . \\ q(s_1) & q(s_2) & \dots & q(s_M) \end{bmatrix}$$

The vector $\mathbf{q}$ is called the *stationary distribution* of the Markov chain.

4. The approach described in the previous comment applies to computing the stationary distribution and the components of $\mathbf{P}^*$ for any closed class.

5. Part 4 says that the long run average time spent in a transient state is zero. When the Markov chain is unichain, this means that for all $s \in S$

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

where $p^*(j|s)$, denotes the entries of the matrix $\mathbf{P}^*$.

6. Puterman [1994] p.593-594 describes an approach for computing $p^*(j|s)$ for $s \in T$ and j in any closed class.

7. Part 6 follows from Gallagher [1996], which shows that for a regular chain the second largest eigenvalue of $\mathbf{P}$ is strictly less than 1 and equals a.

Example B.2. This continues Example B.1. Since $\mathbf{P}$ is not regular, (B.9) does not hold. Using the result in part 3 of Theorem B.1 shows that:

$$\mathbf{P}^* = \begin{bmatrix} p/2 & 1/2 & (1-p)/2 \\ p/2 & 1/2 & (1-p)/2 \\ p/2 & 1/2 & (1-p)/2 \end{bmatrix}. \tag{B.11}$$

Observe that when $0 < p < 1$, all rows of $\mathbf{P}^*$ are equal since $\mathbf{P}$ consists of a single closed class. Moreover $\mathbf{P}^* = (\mathbf{P} + \mathbf{P}^2)/2$ consistent with its representation in part 1 and the observation that for $n \geq 1$, $\mathbf{P}^{2n+1} = \mathbf{P}$ and $\mathbf{P}^{2(n+1)} = \mathbf{P}^2$.

Note that when $p = 1$, states s_1 and s_2 form a recurrent class of period 2 and s_3 is transient. Similarly when $p = 0$, states s_2 and s_3 form a recurrent class of period 2 and s_1 is transient.

B.5 An important lemma

This section provides an important result regarding convergence of geometric series of matrices that is fundamental for Chapters 5-7. It holds in considerable generality and can be proved under a range of hypotheses, but the following finite dimensional version suffices for this book. Note the proof generalizes that used to derive the scalar identity $\sum_{n=0}^{\infty} a^n = (1-a)^{-1}$ when $|a| < 1$. The proof follows Kemeny and Snell [1960].

Lemma B.1. Let $\mathbf{Q}$ denote an $|S| \times |S|$ real-valued matrix for which $\mathbf{Q}^N \to \mathbf{0}$ as $N \to \infty$. Then the inverse of $\mathbf{I} - \mathbf{Q}$ exists and satisfies

$$\sum_{n=0}^{\infty} \mathbf{Q}^n = (\mathbf{I} - \mathbf{Q})^{-1}. \tag{B.12}$$

Proof. Let

$$\mathbf{S}_N := \sum_{n=0}^{N} \mathbf{Q}^n$$

Then it is easy to see that

$$(\mathbf{I} - \mathbf{Q})\mathbf{S}_{N-1} = \mathbf{I} - \mathbf{Q}^N.$$

Letting $N \to \infty$ and applying the hypothesis $\mathbf{Q}^N \to \mathbf{0}$ yields

$$(\mathbf{I} - \mathbf{Q}) \sum_{n=0}^{\infty} \mathbf{Q}^n = \mathbf{I},$$

from which the result follows. □

Since $\mathbf{Q}^n \to \mathbf{0}$ is equivalent[7] to the condition that the spectral radius (largest eigenvalue in absolute value) of $\mathbf{Q}$, $\sigma(\mathbf{Q})$, is strictly less than 1, it follows that:

Lemma B.2. The inverse of $\mathbf{I} - \mathbf{Q}$ exists and satisfies

$$\sum_{n=0}^{\infty} \mathbf{Q}^n = (\mathbf{I} - \mathbf{Q})^{-1} \tag{B.13}$$

if and only if $\sigma(\mathbf{Q}) < 1$.

Note that $\sigma(\mathbf{Q}) < 1$ even in some cases where some row sums of $\mathbf{Q}$ are greater than or equal to 1. Since $\sigma(\mathbf{Q}) \leq ||\mathbf{Q}||$ a sufficient condition for (B.12) to hold is that $||\mathbf{Q}|| < 1$.

[7] This is proved in Faddeeva [1959] and other numerical linear algebra texts.

Lemma B.1 is applied in the book as follows:

1. In discounted models, $\mathbf{Q} = \lambda \mathbf{P}$ with $0 \leq \lambda < 1$.

2. In transient and stochastic shortest path expected total reward models, $\mathbf{Q}$ equals the sub-matrix of $\mathbf{P}$ restricted to its transient states.

3. In average reward models, $\mathbf{Q} = \mathbf{P} - \mathbf{P}^*$.

The following generalization, stated without proof, will be useful when $\mathbf{P}$ is periodic. The summation below and in (B.6) are referred to as *Cesaro summation*[8].

Lemma B.3. Let $\mathbf{Q}$ denote an $M \times M$ matrix for which $\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{Q}^{N-1} \to \mathbf{0}$. Then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \left(\sum_{n=1}^{k} \mathbf{Q}^n \right) = (\mathbf{I} - \mathbf{Q})^{-1}. \tag{B.14}$$

B.6 The deviation matrix

The *deviation matrix*

$$\mathbf{H} := \sum_{n=0}^{\infty} (\mathbf{P}^n - \mathbf{P}^*)$$

is fundamental in the analysis of average reward models. It has the following closed form representation:

Proposition B.1. Let $\mathbf{H}$ be defined as above. Then if $\mathbf{P}^n \to \mathbf{P}^*$,

$$\mathbf{H} = (\mathbf{I} - (\mathbf{P} - \mathbf{P}^*))^{-1} - \mathbf{P}^* = (\mathbf{I} - (\mathbf{P} - \mathbf{P}^*))^{-1}(\mathbf{I} - \mathbf{P}^*). \tag{B.15}$$

Proof. Using equalities in (B.7), it is easy to establish directly or by induction that for $n \geq 1$, $\mathbf{P}^n - \mathbf{P}^* = (\mathbf{P} - \mathbf{P}^*)^n$.

[8] The *Cesaro summation* of the series $x_n, n = 0, 1, 2, \ldots$ equals

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

It often exists when $\lim_{n\to\infty} x_n$ does not. For example the non-convergent series $1, 0, 1, 0, \ldots$ has a *Cesaro limit* equal to $1/2$. In the Markov chain context, (component-wise) Cesaro summation is used to obtain limits in periodic chains.

Hence

$$\mathbf{H} = \sum_{n=0}^{\infty}(\mathbf{P}^n - \mathbf{P}^*) = (\mathbf{I} - \mathbf{P}^*) + \sum_{n=1}^{\infty}(\mathbf{P}^n - \mathbf{P}^*)$$
$$= \mathbf{I} + \sum_{n=1}^{\infty}(\mathbf{P} - \mathbf{P}^*)^n - \mathbf{P}^* = \sum_{n=0}^{\infty}(\mathbf{P} - \mathbf{P}^*)^n - \mathbf{P}^*.$$

Since $\mathbf{P}^n \to \mathbf{P}^*$, $(\mathbf{P}-\mathbf{P}^*)^n$ converges to zero so that the result follows from Lemma B.1. The equivalence in (B.15) follows from applying (B.7)[9]. □

Note that when a Markov chain is periodic or some recurrent class is periodic then the representation for $\mathbf{H}$ in (B.15) is still valid but the summation is in the Cesaro-sense. This means that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \sum_{i=0}^{k} \left(\mathbf{P}^k - \mathbf{P}^*\right) = \left(\mathbf{I} - (\mathbf{P} - \mathbf{P}^*)\right)^{-1} - \mathbf{P}^*.$$

Note that in both cases, $\mathbf{H}$ has the following useful and easy to prove properties:

$$(\mathbf{I} - \mathbf{P})\mathbf{H} = \mathbf{H}(\mathbf{I} - \mathbf{P}) = \mathbf{I} - \mathbf{P}^* \tag{B.16}$$
$$\mathbf{H}\mathbf{P}^* = \mathbf{P}^*\mathbf{H} = \mathbf{0}. \tag{B.17}$$

B.7 Structure of $\mathbf{P}^*$ and $\mathbf{H}$

Proofs of convergence of average reward value iteration and policy iteration in Chapter 7 exploit the following properties of $\mathbf{P}^*$ and $\mathbf{H}$ in unichain Markov chains.

Let R denote the set of recurrent states and T denote the set of transient states of $\mathbf{P}$. Then $\mathbf{P}$ can be written as the partitioned matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{RR} & \mathbf{P}_{RT} \\ \mathbf{P}_{TR} & \mathbf{P}_{TT} \end{bmatrix}, \tag{B.18}$$

where the sub-matrix $\mathbf{P}_{RR}$ corresponds to transitions between recurrent states, $\mathbf{P}_{TR}$ corresponds to transitions from transient to recurrent states, $\mathbf{P}_{RT}$ corresponds to transitions from recurrent states to transient states and $\mathbf{P}_{TT}$ corresponds to transitions between transient states. Note that all entries of $\mathbf{P}_{RT}$ equal zero since by definition such transitions cannot occur. The dimensions of these sub-matrices are *conformable*, that is they depend on the number of recurrent and transient states. The matrices $\mathbf{I}$ and $\mathbf{0}$ will be conformable with the partition.

For ease of reference parts 3 and 4 of Theorem B.1 can be restated as follows.

[9]See Appendix A in Puterman [1994] or Blackwell [1962] for two very different proofs of this result.

Lemma B.4. Let $\mathbf{P}$ be unichain. Then

$$\mathbf{P}^* = \begin{bmatrix}\mathbf{Q} & \mathbf{0}\end{bmatrix}, \tag{B.19}$$

where $\mathbf{Q}$ denotes an $|S| \times |R|$ matrix with each row equal to $\mathbf{q}$ as defined in part 3 of Theorem B.1 and $\mathbf{0}$ denotes an $|S| \times |T|$ matrix of zeroes.

The following proposition describes an important property of $\mathbf{P}_{TT}$ that provides the basis for analysis of transient models in Chapter 6.

Proposition B.2. The matrix $\mathbf{I} - \mathbf{P}_{TT}$ is invertible and satisfies

$$(\mathbf{I} - \mathbf{P}_{TT})^{-1} = \sum_{n=0}^{\infty} \mathbf{P}_{TT}^n. \tag{B.20}$$

Proof. By the definition of transience, for each $s \in T$, a state in R is accessible from s in a finite number of transitions. That is there exists a $j \in R$ for which $p^n(j|s) > 0$ for some n. Since there are finitely many states (in T), there exists a $k \geq 1$ for which each row sum of $\mathbf{P}_{TT}^k$ is strictly less than one. Hence $\mathbf{P}_{TT}^n \to \mathbf{0}$ and the result follows from Lemma B.1. □

An immediate consequence of (B.20) is that the (s, j)-th component of the matrix $(\mathbf{I} - \mathbf{P}_{TT})^{-1}$ equals the expected total number of times the chain is in state $j \in T$ starting from $s \in T$[10].

The matrix $\mathbf{H}$ can be represented in partitioned form

$$\mathbf{H} = \begin{bmatrix}\mathbf{H}_{RR} & \mathbf{H}_{RT} \\ \mathbf{H}_{TR} & \mathbf{H}_{TT}\end{bmatrix}. \tag{B.21}$$

The following lemma describes useful properties of the sub-matrices of $\mathbf{H}$ in unichain models.

Lemma B.5. Let $\mathbf{P}$ be unichain. Then

- All entries of $\mathbf{H}_{RT}$ equal zero,
- $\mathbf{H}_{TT} = (\mathbf{I} - \mathbf{P}_{TT})^{-1}$, and
- $\mathbf{H}_{TT} \geq \mathbf{I}$.

[10] See Chapter III in Kemeny and Snell [1960].

Proof. Writing $\mathbf{Z} := \mathbf{I} - (\mathbf{P} - \mathbf{P}^*)$ in partitioned form gives

$$\mathbf{Z} = \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{W} & \mathbf{I} - \mathbf{P}_{TT} \end{bmatrix},$$

where $\mathbf{I}$ denotes a $|T| \times |T|$ identity matrix and $\mathbf{0}$ denotes an $|R| \times |T|$ matrix of zeroes and $\mathbf{U}$ and $\mathbf{W}$ are specific matrices that will not be further examined. Standard formulae for inverting a partitioned matrix establish that

$$\mathbf{Z}^{-1} = \begin{bmatrix} \mathbf{X} & \mathbf{0} \\ \mathbf{Y} & (\mathbf{I} - \mathbf{P}_{TT})^{-1} \end{bmatrix},$$

where $\mathbf{X}$ and $\mathbf{Y}$ are specific matrices that are not of interest. From Lemma B.4

$$\mathbf{I} - \mathbf{P}^* = \begin{bmatrix} \mathbf{I} - \mathbf{P}^*_{RR} & \mathbf{0}_{RT} \\ -\mathbf{P}^*_{TR} & \mathbf{I}_{TT} \end{bmatrix},$$

where subscripts suggest the sub-matrix dimensions. Since $\mathbf{H} = \mathbf{Z}^{-1}(\mathbf{I} - \mathbf{P}^*)$, parts 1 and 2 follow. Part c follows from Proposition B.2 and the fact that the entries of $\mathbf{P}_{TT}$ are non-negative. $\square$

B.8 Some examples

Example B.3. A two-state Markov chain.
Suppose $S = \{s_1, s_2\}$ and

$$\mathbf{P} = \begin{bmatrix} a & 1-a \\ 1-b & b \end{bmatrix}.$$

Limiting properties of P depend on the values of a and b as follows:

1. If $0 < a < 1$ and $0 < b < 1$, the Markov chain is regular and $\mathbf{P}^n \to \mathbf{P}^*$ where
$$\mathbf{P}^* = \begin{bmatrix} \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \\ \frac{1-b}{2-a-b} & \frac{1-a}{2-a-b} \end{bmatrix}.$$

2. If $a = 1$ and $0 < b < 1$, then the Markov chain is unichain, s_1 is absorbing, $\mathbf{P}^n \to \mathbf{P}^*$ and
$$\mathbf{P}^* = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix},$$
as noted in Lemma B.4.

3. When $a = b = 1$ the Markov chain is multi-chain and $\mathbf{P}^n = \mathbf{P}^*$ for all $n = 0, 1, \ldots$.

4. Suppose $a = b = 0$, the chain consists of single closed periodic class (with period 2), $\mathbf{P}^n$ does not converge to a limiting matrix, but from part 3 of Theorem B.1,
$$\mathbf{P}^* = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{P}^n = \begin{bmatrix} 1/2 & 1/2 \\ \\ 1/2 & 1/2 \end{bmatrix}.$$
This means that the average time in each state is 1/2, which is obvious from the structure of $\mathbf{P}$.

It is left as an exercise to compute $\mathbf{H}$ for each of these cases.

Example B.4. A multi-state Markov chain.
This example analyzes a slightly more complicated Markov chain in which $S = \{s_1, \ldots, s_6\}$ and
$$\mathbf{P} = \begin{bmatrix} a & 1-a & 0 & 0 & 0 & 0 \\ 1-b & b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ c & d & e & f & g & h \end{bmatrix}.$$

Assume $0 < a < 1$ and $0 < b < 1$. This multi-chain matrix is in canonical form and corresponds to closed classes $C_1 = \{s_1, s_2\}$, $C_2 = \{s_3, s_4, s_5\}$ and transient class $T = \{s_6\}$. Because C_2 is periodic, $\mathbf{P}^n$ does not converge, but $\mathbf{P}^*$ can be defined using (B.6).

Letting $w = (1-b)/(2-a-b)$, gives:
$$\mathbf{P}^* = \begin{bmatrix} w & 1-w & 0 & 0 & 0 & 0 \\ w & 1-w & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 \\ \\ (\frac{c+d}{1-h})w & (\frac{c+d}{1-h})(1-w) & (\frac{e+f+g}{1-h})\frac{1}{3} & (\frac{e+f+g}{1-h})\frac{1}{3} & (\frac{e+f+g}{1-h})\frac{1}{3} & 0 \end{bmatrix}.$$

A general formula for deriving the last row appears on p. 594 of Puterman [1994] however in this example, it can be argued that starting in state s_6, the chain ends in C_1 with probability $(c+d)/(1-h)$ and ends in C_2 with probability $(e+f+g)/(1-h)$. Then the probability of being in a particular state in steady-state is the probability of ending in that state's class times the probability of being in that state given it is in that class in steady-state.

B.9 Eigenvalues and eigenvectors of a transition matrix*

Eigenvectors and eigenvalues provide insight into the limiting behavior of Markov chains. This section uses some basic linear algebra results[11].

The following key result will shed much insight into the rate at which a Markov chain converges to its limit.

Theorem B.2. Suppose the $m \times m$ matrix $\mathbf{P}$ has m independent eigenvectors. Then

$$\mathbf{P} = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^{-1}, \tag{B.22}$$

where $\mathbf{W}$ is a matrix with columns equal to the right eigenvectors of $\mathbf{P}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix with entries equal to eigenvalues of $\mathbf{P}$.
Moreover

$$\mathbf{P}^n = \mathbf{W}\boldsymbol{\Lambda}^n\mathbf{W}^{-1}. \tag{B.23}$$

The following example illustrates this result and its consequences using the two-state example.

Example B.5. Let

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix}.$$

Note that this matrix corresponds to the decision rule $d(s_1) = a_{1,1}$ and $d(s_2) = a_{2,2}$ in Example 2.2. Note that the Markov chain corresponding to $\mathbf{P}$ is regular.

It is easy to see that[a]

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1.0 & 0 \\ 0 & 0.4 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1.0 & 0.2 \\ 1.0 & -0.4 \end{bmatrix} \quad \text{and} \quad \mathbf{W}^{-1} = \begin{bmatrix} 2/3 & 1/3 \\ 5/3 & -5/3 \end{bmatrix}.$$

Expanding the right hand side of (B.22) gives

$$\mathbf{P} = 1 \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix} + 0.4 \begin{bmatrix} 1/3 & -1/3 \\ 2/3 & -2/3 \end{bmatrix}. \tag{B.24}$$

It follows from (B.23) that for $n \geq 1$

$$\mathbf{P}^n = 1^n \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix} + 0.4^n \begin{bmatrix} 1/3 & -1/3 \\ 2/3 & -2/3 \end{bmatrix}. \tag{B.25}$$

[11] For example see Strang [2023] and his brilliant online lectures.

Therefore

$$\mathbf{P}^n \to \mathbf{P}^* = \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix} \tag{B.26}$$

and the convergence is at rate 0.4^n as noted in Theorem B.1.

[a] Recall that the sum of the eigenvalues equals the trace of a matrix and the product of eigenvalues equals its determinant.

Observe that in this example the eigenvalues of $\mathbf{P}$ are 1 and 0.4. This is an example of the following important result often referred to as the *Perron-Frobenius theorem.*

Theorem B.3. Let $\mathbf{P}$ be a transition probability matrix. Then

1. 1 is an eigenvalue of $\mathbf{P}$,
2. all eigenvalues of $\mathbf{P}$ are less than or equal to 1 in absolute value, and
3. the eigenvector corresponding to eigenvalue 1 has non-negative components.

Observe also that the stationary distribution $(2/3, 1/3)$ arises naturally in this example as the left eigenvector of $\mathbf{P}$ corresponding to the eigenvalue 1. Note that the right eigenvector of 1 equals $(1, 1)$. Note also that when $\mathbf{P}$ is regular, all eigenvalues other than 1 are strictly less than 1.

The following example analyzes the periodic model in Example B.3.

Example B.6. Let

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

Then

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{W}^{-1} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{bmatrix}.$$

Since $\mathbf{\Lambda}^n$ does not converge, $\lim_{n\to\infty} \mathbf{P}^n$ does not exist. But

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{\Lambda}^n = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

so that from (B.6)

$$\mathbf{P}^* = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{P}^n = \mathbf{W} \left[\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbf{\Lambda}^n \right] \mathbf{W}^{-1} = \begin{bmatrix} 1/2 & 1/2 \\ \\ 1/2 & 1/2 \end{bmatrix}.$$

The following result (stated without proof) is important for the analysis of models in Chapter 7.

Theorem B.4. If the Markov chain corresponding to the transition matrix $\mathbf{P}$ has k closed classes, then the eigenvalue 1 has multiplicity k and k independent eigenvectors.

A consequence of this theorem is that when a Markov chain is unichain or recurrent, the eigenvalue 1 has multiplicity 1 and (right) eigenvector $\mathbf{e} = (1, \ldots, 1)$.

B.10 Absorbing chains

Chapter 6 analyzes Markov decision process models in which Markov chains corresponding to stationary policies can be transformed into a model with $|S - 1|$ transient states and 1 absorbing state. That is $\mathbf{P}$ has the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0}^\mathsf{T} & 1 \end{bmatrix}, \tag{B.27}$$

where at least one component of the $(|S| - 1) \times 1$ matrix $\mathbf{R}$ is positive. Thus by the above the theorem, 1 is an eigenvalue of $\mathbf{P}$ of multiplicity 1. Moreover the following important result follows directly from Lemma B.2.

Theorem B.5. Suppose $\mathbf{P}$ can be partitioned as in (B.27). Then the spectral radius of $\mathbf{Q}$ is strictly less than 1, $\mathbf{Q}^n \to \mathbf{0}$ and

$$(\mathbf{I} - \mathbf{Q})^{-1} = \sum_{n=1}^{\infty} \mathbf{Q}^{n-1}. \tag{B.28}$$

The following example illustrates this result.

Example B.7. Let $S = \{s_1, s_2, s_3\}$ and

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.6 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In this case $\mathbf{Q} = \begin{bmatrix} 0.5 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$ and $\mathbf{R} = \begin{bmatrix} 0.2 \\ 0 \end{bmatrix}$. The eigenvalues of $\mathbf{P}$ are $1, 0.9$ and 0.2 so that as a consequence of (B.23),

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

Moreover as stated in Theorem B.5 the eigenvalues of $\mathbf{Q}$ are 0.9 and 0.2 implying that $\mathbf{Q}^n \to \mathbf{0}$, so the $(\mathbf{I} - \mathbf{P})^{-1}$ exists. Direct computation shows that

$$(\mathbf{I} - \mathbf{Q})^{-1} = \begin{bmatrix} 5.00 & 3.75 \\ 5.00 & 6.25 \end{bmatrix}.$$

Note that the (s, j)-th component of $(\mathbf{I} - \mathbf{Q})^{-1}$ can be interpreted as the (finite) expected number of visits to state j. Hence, starting in s_1, the expected number of visits to s_1 equals 5 and the expected number of visits to s_2 equal 3.75 so on average starting in state s_1 the absorbing state will be reached after 8.75 transitions.

The following example provides another illustration of the consequences of Theorem B.4.

Example B.8. Consider the Markov chain with states $\{s_1, s_2, s_3, s_4\}$ and transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0.4 & 0.6 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.2 & 0.2 & 0.4 & 0.2 \end{bmatrix}.$$

Then the Markov chain corresponding to this matrix has 2 closed classes $C_1 = \{s_1, s_2\}$ and $C_2 = \{s_3\}$ and a transient state s_4. Hence Theorem B.4 implies that the eigenvalue 1 has multiplicity 2.
Direct computation shows that $\mathbf{P}$ has 4 linearly independent eigenvectors so as

a result of Theorem B.2 it can be expressed as $\mathbf{P} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^{-1}$ where[a]

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1 & 0 & 0.2 & 0 \\ 1 & 0 & -0.4 & 0 \\ 0 & 1 & 0 & 0 \\ 0.5 & 0.5 & 0.2 & 1 \end{bmatrix},$$

$$\text{and } \mathbf{W}^{-1} = \begin{bmatrix} 0.667 & 0.333 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1.667 & 1.667 & 0 & 0 \\ 0 & -0.5 & -0.375 & 1 \end{bmatrix}.$$

Letting $\mathbf{L} = \lim_{n\to\infty} \mathbf{\Lambda}^n$ yields:

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

Hence $\mathbf{P}^* = \mathbf{W}\mathbf{L}\mathbf{W}^{-1}$ is given by:

$$\mathbf{P}^* = \begin{bmatrix} 0.667 & 0.333 & 0 & 0 \\ 0.667 & 0.333 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.333 & 0.1667 & 0.5 & 0 \end{bmatrix}.$$

Note that starting in state s_4, the chain ends in C_1 or C_2 with probability 0.5 and then follows the limiting distribution for that class.

[a] To find the eigenvectors of $\mathbf{P}$ first find the eigenvectors for each closed class and adjust values in transient states to ensure that $\mathbf{P}\mathbf{v} = \lambda\mathbf{v}$ for all states. See Example B.5.

B.11 Countable state chains*

Although not used in this book, a brief discussion of some distinctions that arise when analyzing countable state Markov chains is included. Countable state chains arise naturally in queuing models in which the state represents the number of jobs in the queue. In most places in the book this issue has been avoided by truncating the state space.

This brief section points out some of the challenges that arise when analyzing countable state Markov chains as the following simple example shows.

Example B.9. This example shows that when S is countable, the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{P}^n \tag{B.29}$$

exists but that its limit $\mathbf{P}^*$ need not be a transition probability matrix. Let $S = \{1, 2, \dots\}$ and consider a Markov chain with transition probabilities $p(s+1|s) = 1$, $p(j|s) = 0$ for $j \neq s+1$. That is at each transition the Markov chain moves to the right with probability one.

Clearly all states are transient, aperiodic and $\lim_{n\to\infty} \mathbf{P}^n$ exists so that the limit in (B.29) also exists. But the limiting matrix $\mathbf{P}^*$ with elements $p^*(j|s) = 0$ is not a transition probability matrix since all the mass "went off to infinity". Hence if there exists a non-zero reward $\mathbf{r}$, the long run average reward *cannot* be represented by $\mathbf{P}^*\mathbf{r}$.

Such a model is not typical of Markov decision process applications in queuing in which good policies (such as control limit policies in admission control models) result in a Markov chain that can be partitioned into a finite set of recurrent states and infinite set of transient states.

Note further that there are some distinctions with respect to the concepts of recurrence and transience, which are defined in terms of random variables describing transition behavior. Define ν_s to be the number of times a Markov chain visits state s. If s is transient $E_s[\nu_s] < \infty$ and if s is recurrent $E_s[\nu_s] = \infty$. However, recurrence can be further refined.

Let τ_s denote the time of the *first visit time to state s*[12]. For a recurrent state s $P(\tau_s < \infty | X_0 = s) = 1$ and if s is transient, $P(\tau_s < \infty | X_0 = s) < 1$. That is, starting from a recurrent state, the Markov chain returns to it in a finite time with probability one. Starting from a transient state, however, it does not necessarily return. This leads to a refinement on the concept of recurrent: a state s is *positive recurrent* if $E[\tau_s | X_0 = s] < \infty$ and *null recurrent* if $E[\tau_s | X_0 = s] = \infty$. Null recurrence is a curious phenomenon. It refers to a situation in which the chain returns to its starting state with certainty but the expected time to do so is infinite.

The important consequence of this is that in a transient or null recurrent class, the limiting probabilities are 0, while in a positive recurrent class, the limiting probabilities are non-zero. The following example adopted from Sennott [1999] (p.293-295) illustrates this distinction.

Example B.10. Consider a single server queue with batch arrivals and deterministic service rate of one job per period. Assume further that jobs are admitted

[12] If the chain starts in s, τ_s denotes the first return time.

only when the server is idle, which occurs when the queue is empty.
To model this let $S = \{0, 1, 2, ..\}$, $p(j|0) = q_j$ for $j \geq 1$, $p(s-1|s) = 1$ for $s \geq 1$ and $p(j|s) = 0$ otherwise. Thus each time the queue is empty, the system jumps to state j with probability q_j. It is left as an exercise to show that:

- All states are recurrent.
- If $\sum_{j=1}^{\infty} jq_j < \infty$, all states are positive recurrent.
- If $\sum_{j=1}^{\infty} jq_j = \infty$, all states are null recurrent.

Thus if the batch sizes are too large, the Markov chain is null recurrent.

Appendix C

Linear Programming

A linear program is an optimization problem that comprises an *objective function* and *constraints*, which are restricted to linear functions of the *decision variables*. Let $\mathbf{c} \in \mathbb{R}^n$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Consider the following linear program.

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{n} c_j x_j \\ \text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \geq b_i, \quad \forall i = 1, \ldots, m. \end{aligned} \tag{C.1}$$

It has a linear objective function and constraints in the form of linear inequalities with respect to the decision variables $x_1, \ldots, x_n$. The *parameters* $c_j, j = 1, \ldots, n, a_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$, and $b_i, i = 1, \ldots, m$ are fixed and typically derived from data.

The above formulation can be written compactly in vector form:

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^\mathsf{T} \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \geq \mathbf{b}. \end{aligned} \tag{C.2}$$

The i-th row of the $\mathbf{A}$ matrix will be written as a row vector $\mathbf{a}_i^\mathsf{T}$. The j-th column of $\mathbf{A}$ will be written $\mathbf{A}_j$.

Note that many equivalent forms of an LP can be written in the above manner. For example, if the objective is to maximize instead of minimize, simply replace $\mathbf{c}$ with $-\mathbf{c}$. Similarly, for "less than or equal to" inequalities, multiply $\mathbf{A}$ and $\mathbf{b}$ by -1. Equality constraints can also be represented in the above form by noting that for a given i, $\mathbf{a}_i^\mathsf{T}\mathbf{x} = b_i$ can be enforced with two constraints: $\mathbf{a}_i^\mathsf{T}\mathbf{x} \geq b_i$ and $-\mathbf{a}_i^\mathsf{T}\mathbf{x} \geq -b_i$. Sign constraints on the variables (e.g., $x_1 \geq 0$ or $x_2 \leq 0$) can be enforced with appropriate choices of $\mathbf{A}$ and $\mathbf{b}$.

A *standard form* linear program is written

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^\mathsf{T} \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0}, \end{aligned} \tag{C.3}$$

where the constraints comprise equality constraints involving $\mathbf{A}$ and $\mathbf{b}$, and non-negativity constraints on the decision variables. Any linear program can be transformed to one in standard form as follows. An unconstrained variable x_j can be replaced by $x_j^+ - x_j^-$, with constraints $x_j^+ \geq 0$ and $x_j^- \geq 0$. An inequality $\mathbf{a}_i^\mathsf{T}\mathbf{x} \geq b_i$ can be written as $\mathbf{a}_i^\mathsf{T}\mathbf{x} - s_i = b_i$, where $s_i \geq 0$. If the inequality is a "less than or equal to", add s_i instead of subtracting it.

A vector $\mathbf{x}$ that satisfies all constraints is a *feasible solution*. The set of vectors that satisfy the constraints is referred to as the *feasible region*. The feasible region of a linear program is a *polyhedron* by definition, since a polyhedron is defined as the intersection of a finite set of linear inequalities. Among the feasible solutions, the one with the lowest value of $\mathbf{c}^\mathsf{T}\mathbf{x}$, if it exists, is an *optimal solution*, denoted $\mathbf{x}^*$. If a linear program is feasible, but for every real number r there exists a feasible solution $\mathbf{x}$ such that $\mathbf{c}^\mathsf{T}\mathbf{x} < r$, then the problem is *unbounded* or has *unbounded objective function value*. Such an LP may be referred to as having an optimal value of $-\infty$. The following theorem establishes when an LP will have an optimal solution.

Theorem C.1. Every feasible LP with bounded objective function has an optimal solution.

If the feasible region of an LP is non-empty and bounded, then the objective function value cannot go to $-\infty$. Hence, the LP will have an optimal solution.

Corollary C.1. An LP with a non-empty bounded feasible region has an optimal solution.

Consider an LP with the set of constraints $\mathbf{a}_i^\mathsf{T}\mathbf{x} \geq b_i, i = 1, \ldots, m$. If a vector $\mathbf{x}$ satisfies a subset of the constraints at equality $\mathbf{a}_i^\mathsf{T}\mathbf{x} = b_i, i \in I \subseteq \{1, \ldots, m\}$, and if there are n coefficient vectors in this subset $\{\mathbf{a}_i | i \in I\}$ that are linearly independent, then $\mathbf{x}$ is a *basic feasible solution*. Figure 5.14 from Chapter 5 illustrates a polyhedron. Basic feasible solutions are the "corner points" of the polyhedron. Basic feasible solutions are important because under mild conditions, optimal solutions to an LP, if they exist, can be found there.

Theorem C.2. If a linear program has an optimal solution and its feasible region contains at least one basic feasible solution, then at least one basic feasible solution is an optimal solution.

The following characterizes when the feasible region of a linear program will have an extreme point.

Theorem C.3. A polyhedron has at least one extreme point if and only if it does not contain a line.

The well-known Simplex Algorithm progressively searches basic feasible solutions with improving objective function value until it finds an optimal basic feasible solution or determines that the optimal value is $-\infty$.

Note that there are LPs that have optimal solutions, but no corner points. Consider the feasible region $\{(x_1, x_2) \mid x_2 \geq 0\}$, which is a half space of $\mathbb{R}^2$. If the objective is to minimize x_2, then all points on the line $(x_1, 0), x_1 \in (-\infty, \infty)$ are optimal but none are corner points.

To guarantee that a polyhedron has a corner point, it is necessary and sufficient that it does not contain a line. That is, there must not be a vector $\mathbf{x}$ and nonzero vector $\mathbf{d}$ such that $\mathbf{x} + \alpha\mathbf{d}$ remains in the polyhedron for all scalars (negative and positive) α. Note that a linear program in standard form does not contain a line, since the feasible region is a subset of $\{\mathbf{x} \mid \mathbf{x} \geq \mathbf{0}\}$. Hence, if a linear program in standard form has an optimal solution, there must be one at a basic feasible solution.

Theorem C.4. A polyhedron has at least one extreme point if and only if it does not contain a line.

C.1 Duality

Given any linear program, there exists a *dual* linear program. The former or original linear program is referred to as the *primal*. The following pair of linear programs are dual to each other.

$$
\begin{array}{rlrl}
\text{minimize} & \mathbf{c}^\mathsf{T}\mathbf{x} & \text{maximize} & \mathbf{b}^\mathsf{T}\mathbf{y} \\
\text{subject to} & \mathbf{a}_i^\mathsf{T}\mathbf{x} \geq b_i, \quad i \in I_1, & \text{subject to} & y_i \geq 0, \quad i \in I_1, \\
& \mathbf{a}_i^\mathsf{T}\mathbf{x} \leq b_i, \quad i \in I_2, & & y_i \leq 0, \quad i \in I_2, \\
& \mathbf{a}_i^\mathsf{T}\mathbf{x} = b_i, \quad i \in I_3, & & y_i \text{ free}, \quad i \in I_3, \\
& x_j \geq 0, \quad j \in J_1, & & \mathbf{A}_j^\mathsf{T}\mathbf{y} \leq c_j, \quad i \in J_1, \\
& x_j \leq 0, \quad j \in J_2, & & \mathbf{A}_j^\mathsf{T}\mathbf{y} \geq c_j, \quad i \in J_2, \\
& x_j \text{ free}, \quad j \in J_3, & & \mathbf{A}_j^\mathsf{T}\mathbf{y} = c_j, \quad i \in J_3.
\end{array}
$$

Let the minimization problem be the primal. Then the dual problem is a maximization problem. A dual variable is associated with every non-sign constraint in the primal. Whether the dual variable is subject to a sign constraint or is free depends on whether the corresponding primal constraint is an inequality or equality, respectively. Similarly, every primal variable is associated with a non-sign constraint in the dual, and whether the primal variable is sign-constrained or is free determines whether the dual constraint is an inequality or equality, respectively. Notice that the constraints involving $\mathbf{A}$ in the dual use the transpose of the $\mathbf{A}$ matrix. That is, the constraints in the primal are written based on the rows, $\mathbf{a}_i^\mathsf{T}$, of $\mathbf{A}$. However, the constraints in the dual are written based on the columns, $\mathbf{A}_j$, of $\mathbf{A}$. Correspondingly, the parameters $\mathbf{c}$

and $\mathbf{b}$ have switched places. The objective coefficients $\mathbf{c}$ in the primal have become the right hand side parameters in the constraints in the dual, and the right hand side parameters $\mathbf{b}$ in the primal have become the objective coefficients in the dual.

First, it is straightforward to show that the value of the dual (maximization) objective is a lower bound on the value of the primal (minimization) objective. This result is known as *Weak Duality*.

Theorem C.5. Let $\mathbf{x}$ be a feasible solution to the primal and $\mathbf{y}$ be a feasible solution to the dual. Then $\mathbf{b}^\mathsf{T}\mathbf{y} \leq \mathbf{c}^\mathsf{T}\mathbf{x}$.

An immediate consequence of this result is that if there is a primal solution and dual solution that have the same objective function value, they are optimal solutions for their respective problems.

Corollary C.2. Let $\mathbf{x}$ be a feasible solution to the primal and $\mathbf{y}$ be a feasible solution to the dual. If $\mathbf{b}^\mathsf{T}\mathbf{y} = \mathbf{c}^\mathsf{T}\mathbf{x}$, then $\mathbf{x}$ and $\mathbf{y}$ are optimal solutions to the primal and dual, respectively.

A fundamental theorem of linear programming is the theorem of *Strong Duality*.

Theorem C.6. Let $\mathbf{x}^*$ be an optimal solution for an LP. Then its dual also has an optimal solution, $\mathbf{y}^*$, and their optimal values are equal.

Finally, the *Complementary Slackness* conditions provide another set of necessary and sufficient conditions for feasible solutions to the primal and dual to be optimal.

Theorem C.7. Let $\mathbf{x}$ be a feasible solution to the primal and $\mathbf{y}$ be a feasible solution to the dual. They are optimal if and only if

$$
\begin{aligned}
y_i(\mathbf{a}_i^\mathsf{T}\mathbf{x} - b_i) &= 0, \quad \forall i \\
(c_j - \mathbf{A}_j^\mathsf{T}\mathbf{y})x_j &= 0, \quad \forall j.
\end{aligned}
$$

Bibliography

Optimal decision procedures for finite Markov chains, I], author= J. A. Bather, journal=Adv. Appl. Prob., volume=5, pages=328-339, year=1973.

J. Abounadi, D. Bertsekas, and V.S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.*, 40(3):681–698, 2001.

O. Alagoz, L. M. Maillart, A. J. Schaefer, and M. S. Roberts. Determining the acceptance of cadaveric livers using an implicit model of the waiting list. *Oper. Res.*, 55: 24–36, 2007.

American Cancer Society. *Cancer Facts and Figures 2024*. American Cancer Society, Atlanta, GA, 2024.

Aristotle. *Nicomachean Ethics*. Oxford University Press, Oxford, 1925. Book II, chapter 1 (1103a31–32). Translated by W. D. Ross.

K. J. Arrow, D. Blackwell, and M. A. Girshick. Bayes and minimax solutions of sequential decision problems. *Econometrica*, 17:213–244, 1949.

K. J. Arrow, T. Harris, and J. Marschak. Optimal inventory policy. *Econometrica*, 19: 250–272, 1951.

K. J. Arrow, S. Karlin, and H. E. Scarf. *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Stanford, CA, 1958.

T. Ayer, O. Alagoz, and N. K. Stout. Or forum–A POMDP approach to personalize mammography screening decisions. *Oper. Res.*, 60:1019–1034, 2012.

L. C. Baird. Residual algorithms: Reinforcement learning with function approximation. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 30–37, San Francisco, CA, USA, 1995. Morgan Kaufmann.

A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Trans. Syst. Man. Cybern.*, SMC-13 (5):834–846, 1983.

A. G. Barto, R. S. Sutton, and C. J. C. H. Watkins. Sequential decision problems and neural networks. In D.S. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2, pages 686–693. 1989.

J. A. Bather. Optimal decision procedures for finite Markov chains, II. *Adv. Appl. Prob.*, 5:521–540, 1973.

J. A. Bather. Dynamic programming. In W. Lederman and S. Vajda, editors, *Analysis, Vol. 4., Handbook of Applicable Analysis*. John Wiley and Sons, NY, 1982.

R. Bellman. Some directions of research in dynamic programming. *Unternehmensforschung*, 7:97–102, 1963.

R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

M. Bennett. *A brief history of intelligence*. William Collins, 2023.

A. Berman and R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice Hall, 1987.

D. P. Bertsekas. Approximate policy iteration: a survey and some new methods. *J. Control Theory Appl.*, 9:310–335, 2011.

D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volume 2, 4th Edition*. Athena Scientific, 2012.

D. P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

D. P. Bertsekas. *Lessons from AlphaZero for Optimal, Model Predictive and Adaptive Control*. Athena Scientific, 2022.

D. P. Bertsekas and J. N. Tsitsiklis. An analysis of stochastic shortest path problems. *Math. Oper. Res.*, 16:580–595, 1991.

D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

D. Bertsimas and R. Shioda. Restaurant revenue management. *Oper. Res.*, 51:472–486, 2003.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Belmont, Massachusetts, 1997.

T. Bi and R. D'Andrea. Sample-efficient learning to solve a real-world labyrinth game using data-augmented model-based reinforcement learning, 2023. URL https://arxiv.org/abs/2312.09906.

D. Blackwell. Discrete dynamic programming. *Ann. Math. Stat.*, 33:719–726, 1962.

D. Blackwell. Discounted dynamic programming. *Ann. Math. Stat.*, 36:226–235, 1965.

D. Blackwell. Positive dynamic programming. In *Proceedings of the 5th Berkeley Symposium on Probability and Statistics, Volume 1*, pages 415–418, 1967.

J. E. Blum. Approximation methods which converge with probability one. *Ann. Math. Stat.*, 25:382–386, 1954.

M. Broadie. *Every Shot Counts.* Avery, 2014.

W. J. Bryan. America's Mission. Speech, Virginia Democratic Association Banquet, Washington, D.C., 1899. Delivered on February 22, 1899.

L. Buşoniu, D. Ernst, B. De Schutter, and R. Babuška. Online least-squares policy iteration for reinforcement learning control. In *Proceedings of the 2010 American Control Conference*, pages 486–491, 2010.

L. Buşoniu, A. Lazaric, M. Ghavamzadeh, M. Rémi, and R. Babuška. Least-squares methods for policy iteration. In M. Wiering and M. van Otterio, editors, *Reinforcement Learning: State of the Art*, pages 75–109. Springer Berlin, 2012.

P. G. Canbolat and U. G. Rothblum. (Approximate) iterated successive approximations algorithm for sequential decision processes. *Ann. Oper. Res.*, 208:309–320, 2013.

L. Carroll. *Alice's Adventures in Wonderland.* Macmillan and Co., London, 1865.

V. Carter and R. E. Machol. Technical note–Operations research on football. *Oper. Res.*, 19:541–544, 1971.

A. R. Cassandra. The POMDP Page, 2025. URL https://www.pomdp.org/. Accessed: 2025-01-30.

A. R. Cassandra, M. L. Littman, and N. L. Zhang. Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI1997)*, pages 54–61, 1997.

A. Cayley. Mathematical questions with their solutions, no. 4528. *Educational Times*, 23:18, 1875.

E. Çinlar. *Introduction to Stochastic Processes.* Prentice Hall, Englewood Cliffs, N.J., 1975.

T. C. Y. Chan and R. Singal. A Markov Decision Process-based handicap system for tennis. *Journal Quant. Anal. Sports*, 12:179–189, 2016.

T. C. Y. Chan, C. Fernandes, and M. L. Puterman. Points gained in football: Using Markov process-based value functions to assess team performance. *Oper. Res.*, 69: 877–894, 2021.

T. C. Y. Chan, C. Fernandes, and R. Walker. No more throwing darts at the wall: Developing fair handicaps for darts using a Markov Decision Process. In *Proceedings of the 18th Annual MIT Sloan Sports Analytics Conference*, 2024.

H. S. Chang, M. C. Fu, J. Hu, and S. I. Marcus. An adaptive sampling algorithm for solving Markov decision processes. *Oper. Res.*, 53(1):126–139, 2005.

R. R. Chen, T. C. E. Cheng, T. M. Choi, and Y. Wang. Novel advances in applications of the newsvendor model. *Decis. Sci.*, 47(1):8–10, 2016.

H. Cheng. *Algorithms for partially observed Markov decision processes*. PhD thesis, University of British Columbia, 1988.

Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The theory of optimal stopping*. Houghton-Mifflin, New York, 1971.

C. W. Clark. The lazy adaptable lions: A Markovian model of group foraging. *Anim. Behav.*, 35:361–368, 1987.

Confucius. *Analects 15.10*, page 178. Hackett Publishing, Indianapolis, 2003. Translated by E. Slingerland.

C. Darken, J. Chang, and J. Moody. Learning rate schedules for faster stochastic gradient search. In *Neural Networks for Signal Processing II Proceedings of the 1992 IEEE Workshop*, pages 3–12, 1992.

P. Dayan and T.J. Sejnowski. TD(λ) converges with probability 1. *Machine Laearning*, 14:295–301, 1994.

D. P. de Farias and B. Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *J. Opt. Theor. Appl.*, 105:589–608, 2000.

D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Oper. Res.*, 51:850–865, 2003.

V. de Paul. Letter 1796 to Charles Ozenne, 13 Nov. 1654. In Sr. Marie Poole et al., editors, *Correspondence, Conferences, Documents*, volume 5, page 219. New City Press, Brooklyn, NY, 1995. Translated by Sr. Marie Poole and Rev. Francis Germovnik.

E. Denardo and B. Fox. Multichain markov renewal programs. *SIAM J. Appl. Math.*, 16:468–487, 1968.

F. D'Epenoux. Sur un problème de production et de stockage dans l'aléatoire. *RAIRO*, 14:3–16, 1960.

C. Derman. *Finite State Markovian Decision Processes*. Academic Press, New York, NY, 1970.

J. Desrosiers and M. E. Lübbecke. A primer in column generation. In G. Desaulniers, J. Desrosiers, and M. M. Solomon, editors, *Column Generation*, pages 1–32. Springer, 2005.

N. Draper and H. Smith. *Applied Regression Analysis, 3rd Edition*. Wiley-Interscience, 1998.

A. Dvoretsky, J. Kiefer, and J. Wolfowitz. The inventory problem: I. case of known distributions of demand. *Econometrica*, 20:187, 1952.

F. Dyson. A Meeting with Enrico Fermi. *Nature*, 427(6972):297, 2004.

J. H. Eaton and L. A. Zadeh. Optimal pursuit strategies in discrete state probabilistic systems. *Trans. ASME, Series D*, 84:23–29, 1962.

F. Edgeworth. The mathematical theory of banking. *J. R. Stat. Soc.*, 51(1):113–127, 1888.

G. Eliot. *Middlemarch: A Study of Provincial Life*. William Blackwood and Sons, Edinburgh and London, 1874.

V. N. Faddeeva. *Computational Methods of Linear Algebra*. Dover, New York, 1959.

W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1*. John Wiley & Sons., 1982.

T. S. Ferguson. Who solved the secretary problem? *Stat. Sci.*, 4(3):282–296, 1989.

R. P. Feynman, R. B. Leighton, and M. Sands. Lecture 1: Atoms in motion. In *The Feynman Lectures on Physics*, volume I, pages 1–2. Addison–Wesley, 1963.

J. Forrest and the COIN-OR contributors. *COIN-OR CLP: Simplex Solver*. COIN-OR Foundation, 2024. URL `https://www.coin-or.org/Clp`.

M. C. Fu. Monte Carlo tree search: a tutorial. In *Proceedings of the 2018 Winter Simulation Conference*, pages 222–236, 2018.

R. G. Gallagher. *Discrete Stochastic Processes*. Springer, 1996.

G. Gallego and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Manag. Sci.*, 40:999–1020, 1994.

J. Gessford and S. Karlin. Optimal policies for hydroelectric operations. In K. Arrow, S. Karlin, and H. Scarf, editors, *Studies in the Mathematical Theory of Inventory and Production*, pages 179–200. Stanford University Press, Stanford, CA, 1958.

G. T. De Ghelinick. Les problèmes de décisions séquentielles. *Cahiers du Centre d'Études Rec. Opér.*, 2:161–179, 1960.

J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*. North-Holland, Amsterdam, 1974.

Y. Goçgun and M. L. Puterman. Dynamic scheduling with due dates and time windows: An application to chemotherapy patient appointment booking. *Health Care Manag. Sci.*, 17:60–76, 2014.

A. Gosavi. *Simulation-Based Optimization - Parametric Optimization Techniques and Reinforcement Learning*. Springer, 2015.

S. Gronauer, M. Gottwald, and K. Diepold. The sucessful ingredients of policy gradient algorithms. In *Proceedings of the Thirtieth Joint Conference on Artificial Intelligence (IIJCAI-21)*, pages 2455–2461, 2021.

I. Grondman, L. Busoniu, A. Lopes, and R. Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradient. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307, 2012.

LLC Gurobi Optimization. *Gurobi Optimizer Reference Manual*, 2024. URL `https://www.gurobi.com`.

D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025.

M. Hall. The state of goalie pulling in the NHL. `https://hockey-graphs.com/2020/05/18/the-state-of-goalie-pulling-in-the-nhl/`, May 2020. [accessed 14-September-2021].

N. A. J. Hastings. Some notes on dynamic programming and replacement. *Op. Res. Quart.*, 19:453–464, 1968.

M. Haviv and M. L. Puterman. Bias optimality in controlled queueing systems. *Jour. Appl. Prob.*, 35:136–150, 1998.

J. Herrman. The Great Amazon Flip-a-Thon. *The New York Times*, 2021. URL `https://www.nytimes.com/2021/03/17/style/amazon-brand-flippers.html`. Published online 17 Mar. 2021.

D. P. Heyman. Optimal operating policies for M/G/1 queueing systems. *Oper. Res.*, 16:362–382, 1968.

B. Hill. The glory of love. Sheet music, 1936.

F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. McGraw-Hill, 2021.

R. Howard. *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA, 1960.

T. Hubel, J. Myatt, N. Jordan, O. Dewhirst, J. McNutt, and A. Wilson. Energy cost and return in African wild dogs and cheetahs. *Nat. Commun.*, 7:11034, 2016.

G. Hübner. Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties. In J. Kožešnik, editor, *Transactions of the Seventh Prague conference on Information Theory, Statistical Decision Functions, Random Processes and the 1974 European Meeting of Statisticians*, pages 257–263, 1977.

IBM. *IBM ILOG CPLEX Optimization Studio*, 2024. URL https://www.ibm.com/products/ilog-cplex-optimization-studio.

T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Comput.*, 6(6):1185–1201, 1994.

L. C. M. Kallenberg. *Linear Programming and Finite Markovian Control Problems*. Mathematisch Centrum, Amsterdam, 1983.

S. Karlin. Stochastic models and optimal policies for selling an asset. In K. Arrow, S. Karlin, and H. Scarf, editors, *Studies in Applied Probability and Management Science*, pages 148–158. Stanford University Press, Stanford, CA, 1962.

D. Katselis. ECE586: MDPs and reinforcement learning: Lecture 10, 2019. URL http://katselis.web.engr.illinois.edu/ECE586/Lecture10.pdf.

E. J. Kelly and P. L. Kennedy. A dynamic stochastic model of mate desertion. *Ecology*, 74:351–366, 1993.

J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Van Nostrand-Reinhold, New York, 1960.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

L. Kleinrock. *Queuing Systems, Volume 1*. John Wiley & Sons, 1975.

M. J. Kochenderfer, T. A. Wheeler, and K. H. Wray. *Algorithms for Decision Making*. MIT Press, 2022.

R. Kohavi and S. Thomke. The surprising power of online experiments. *Harvard Business Review*, pages 74–82, September 2017.

V. Konda and J. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 12. MIT Press, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/0199d4b5c234ed1324952503c3e02160-Paper.pdf.

V. Krishnamurthy. *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge University Press, Cambridge, UK, 2016.

M. Kurt, B. T. Denton, A. J. Schaefer, N. D. Shah, and S. A. Smith. The structure of optimal statin initiation policies for patients with type 2 diabetes. *IIE Trans. Health. Syst. Eng.*, 1:49–65, 2011.

H. Kushner and A.J. Kleimnan. Accelerated procedures for the solution of discrete Markov control problems. *IEEE Trans. Autom. Control*, 16:147–152, 1971.

M. G. Lagoudakis and R. Parr. Least-squares policy evaluation. *J. Mach. Learn. Res.*, 4:1107–1149, 2003.

M. Lehmann. The definitive guide to policy gradients in deep reinforcement learning: Theory, algorithms and implementations, 2024. URL https://arxiv.org/abs/2401.13662.

S. Levine. CS285: Deep reinforcement learning; Lectures 4-9, 2024. URL https://rail.eecs.berkeley.edu/deeprlcourse/.

S. A. Lippman. Applying a new device in the optimization of exponential queuing systems. *Oper. Res.*, 23:687–710, 1975.

J. London. What Life Means to Me. *Cosmopolitan Magazine*, 40(5):526–530, March 1906.

W. S. Lovejoy. Computationally feasible bounds for partially observed Markov decision processes. *Oper. Res.*, 39:162–175, 1991a.

W. S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Ann. Oper. Res.*, 28:47–66, 1991b.

J. MacQueen. A modified dynamic programming method for Markov decision problems. *J. Math. Anal. Appl.*, 14:38–43, 1966.

J. MacQueen. Letter to the editor – A test for suboptimal actions in Markov decision problems. *Oper. Res.*, 15:559–561, 1967.

S. Mahadevan. Average reward reinforcement learning: Foundations, algorithms and empirical results. *Mach. Learn.*, 22:159–195, 1996.

A. Makhorin. *GNU Linear Programming Kit, Version 5.0.* GNU Project, 2024. URL https://www.gnu.org/software/glpk/.

M. Mangel and C. W. Clark. Towards a unified foraging theory. *Ecology*, 67:1127–1138, 1986.

A. Manne. Linear programming and sequential decisions. *Manag. Sci.*, 6:259–267, 1960.

P. Massé. *Les reserves et la regulation se l'avenir dans la vie economique, Vol I and II.* Hermann, 1946.

M. C. McKenzie and M. D. McDonnell. Modern value based reinforcement learning: A chronological review. *IEEE Access*, 10:134704–134725, 2022.

S. Meyn. *Control Systems and Reinforcement Learning.* Cambridge University Press, 2022.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with deep reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning*, pages 330–337, 2013.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

G. E. Monahan. State of the art–A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Manag. Sci.*, 28:1–16, 1982.

D. G. Morrison. On the optimal time to pull the goalie: A Poisson model applied to a common strategy used in ice hockey. *TIMS Studies in Management Science*, 4: 67–78, 1976.

P. Naor. On the regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

I. Newton. Letter to Robert Hooke, 15 Feb. 1676. In H. W. Turnbull, editor, *The Correspondence of Isaac Newton: 1661–1675*, volume 1, page 416. Cambridge University Press, Cambridge, 1959.

M. Nielsen. *Neural Networks and Deep Learning.* Determination Press, 2015.

J. M. Norman. Dynamic programming in tennis – When to use a fast serve. *J. Oper. Res. Soc.*, 36:75–77, 1985.

NVIDIA Isaac. AI robot development platform. https://developer.nvidia.com/isaac/, 2025. Retrieved Feb. 20, 2025.

J. Patrick, M. L. Puterman, and M. Queyranne. Dynamic multi-priority patient scheduling. *Oper. Res.*, 56:1507–152, 2008.

A. Patterson, S. Neumann, M. White, and A. White. Empirical design in reinforcement learning. *J. Mach. Learn. Res.*, 25:1–63, 2024.

J. Pineau, G.. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1025–1032, 2003.

L. K. Platzman. Negative dynamic programming. *Oper. Res.*, 25:529–533, 1977.

E. Porteus. *Foundations of Stochastic Inventory Theory*. Stanford Business Books, 2002.

W. Powell. *Approximate Dynamic Programming*. J. Wiley and Sons, 2007.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons., 1994.

M. L. Puterman and S. Brumelle. On the convergence of policy iteration in stationary dynamic programming. *Math. Oper. Res.*, 4:60–69, 1979.

M. L. Puterman and M. C. Shin. Modified policy iteration algorithms for discounted Markov decision problems. *Manag. Sci.*, 24:1127–1137, 1978.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

A. Ribes, T. Terraz, B. Iooss, Y. Fournier, and B. Raffin. Large scale in transit computation of quantiles for ensemble runs. 2019. URL https://arxiv.org/abs/1905.04180.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

O. Roeder. *Seven games: a human history*. W. W. Norton, New York, 2022.

S. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1:206–215, 2019.

W. Rudin. *Principles of Mathematical Analysis, 2nd Edition*. McGraw-Hill, Inc., 1964.

A. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. and Dev.*, 3:210–229, 1959.

A. Sauré, J. Patrick, and M. L. Puterman. Optimal multi-appointment scheduling. *Eur. J. Oper. Res.*, 223:573–584, 2012.

A. Sauré, J. Patrick, and M.L. Puterman. Simulation-based approximate policy iteration with generalized logistic functions. *INFORMS J. Comput.*, 27:579–595, 2015.

H. Scarf. The optimality of (s, S)-policies in the dynamic inventory problem. In S. Karlin K. Arrow and P. Suppes, editors, *Studies in the Mathematical Theory of Inventory and Production*, pages 196–202. Stanford University Press, Stanford, CA, 1960.

J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. 2017. URL https://arxiv.org/abs/1707.06347.

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

P. J. Schweitzer. Iterative solution of the functional equations of undiscounted Markov renewal programming. *J. Math. Anal. Appl.*, 34:495–501, 1971.

P. J. Schweitzer and A. Federgruen. Asymptotic behavior of value iteration in Markov decision problems. *Math. Oper. Res.*, 2:360–381, 1977.

P. J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markov decision processes. *J. Math. Anal. Appl.*, 110:568–582, 1985.

L. Sennott. *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley and Sons, 1999.

R. Serfozo. Monotone optimal policies for Markov decision processes. *Math. Program. Stud.*, 6:202–215, 1976.

L. S. Shapley. Stochastic games. *Proc. Natl. Acad. Sci. USA*, 39:1095–1100, 1953.

S. M. Shechter, M. D. Bailey, A. J. Schaefer, and M. S. Roberts. The optimal time to initiate HIV therapy under ordered health states. *Oper. Res.*, 56:20–33, 2008.

S. M. Shechter, F. Ghassemi, Y. Gocgun, and M. L. Puterman. Technical note–Trading off quick versus slow actions in optimal search. *Oper. Res.*, 63:353–362, 2015.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

R. Simmons and S. Koenig. Probabilistic robot navigation in partially observable environments. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1080–1087, 1995.

S. Singh, T. Jaakkola, M. Littman, and C. Szepesvári. Convergence results for some reinforcement learning algorithms. In *Proceedings of the seventeenth International Conference on Machine Learning*, pages 791–798, 2000.

E. Sirot and C. Bernstein. Time sharing between host searching and food searching in parasitoids: state-dependent optimal strategies. *Behav. Ecol.*, 7:189–194, 1996.

R. D. Smallwood and E. J. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.*, 21:1071–1088, 1973.

B. C. Smith, J. F. Leimkuhler, and R. M. Darrow. Yield management at American Airlines. *Interfaces*, 22:8–31, 1992.

G. L. Smuts. Diet of lions and spotted hyenas assessed from stomach contents. *S. Afr. J. Wildl. Res.*, 9:19–25, 1979.

M. J. Sobel. Optimal average-cost policy for a queue with start-up and shut-down costs. *Oper. Res.*, 17:145–162, 1969.

E. J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Oper. Res.*, 26:282–304, 1978.

M. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *J. Artif. Intell. Res.*, 24:195–220, 2005.

D. Stengos and L. C. Thomas. The blast furnaces problem. *Eur. J. Oper. Res.*, 4: 330–336, 1980.

G. Strang. *Introduction to Linear Algebra, 6th edition.* Wellesley-Cambridge Press, 2023.

R. Strauch. Negative dynamic programming. *Ann. Math. Stat.*, 37:871–890, 1966.

H. Sugiyama. Solving deterministic problems via stochastic approximation – A simple yet powerful numerical method. *Computers Math. Applic.*, 27(9/10):129–144, 1994.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Mach. Learn.*, 3:9–44, 1988.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An introduction, second edition.* MIT Press, Cambridge, MA, 2018.

R. S. Sutton, D. MacAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 1999.

C. Szepesvari. *Algorithms for Reinforcement Learning.* Morgan and Claypool, 2010.

K. Talluri and G. van Ryzin. *The Theory and Practice of Revenue Management.* Springer US, 2004.

J. C. Y. Tang, V. Paixao, F. Carvalho, A. Silva, A. Klaus, J. Alves da Silva, and R. M. Costa. Dynamic behaviour restructuring mediates dopamine-dependent credit assignment. *Nature*, 626:583–592, 2024.

G. Tesauro. Practical issues in temporal difference learning. *Mach. Learn.*, 8:257–277, 1992.

P. S. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, PMLR*, volume 48, pages 2139–2148, 2016.

H. D. Thoreau. *Walden; or, Life in the Woods.* Ticknor and Fields, Boston, 1st edition, 1854.

J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Mach. Learn.*, 16:185–202, 1994.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximations. *IEEE Trans. Autom. Control*, 42:674–690, 1997.

J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35:1799–1808, 1999.

J. W. Tukey. The future of data analysis. *Ann. Math. Stat.*, 33(1):1–67, 1962.

J. A. E. E. van Nunen. A set of successive approximation methods for discounted Markovian decision problems. *Z. Oper. Res.*, 20:203–208, 1976.

A. F. Veinott, Jr. Discrete dynamic programming programming with sensitive discount optimality criteria. *Ann. Math. Stat.*, 40:1635–1660, 1969a.

A. F. Veinott, Jr. Dynamic programming and stochastic control course notes and homework. OR351, Stanford University, 1969b. Unpublished.

A. Wald. *Sequential Analysis.* John Wiley & Sons, New York, 1947.

A. Wald and J. Wolfowitz. Optimal character of the sequential probability ratio test. *Ann. Math. Stat.*, 19:326–339, 1948.

J. Wardle. Wordle, 2021. URL https://www.nytimes.com/games/wordle/index.html. Accessed: 2024-07-12.

A. Washburn. Still more on pulling the goalie. *Interfaces*, 21:59–64, 1991.

C. Watkins and P. Dayan. Q-learning. *Mach. Learn.*, 8:279–292, 1992.

J. Wessels. Markov programming by successive approximations with respect to weighted supremum norms. *J. Math. Anal. Appl.*, 58:326–335, 1977.

D. J. White. Dynamic programming, markov chains and the method of sucessive approximations. *J. Math. Anal. Appl.*, 6:373–376, 1963.

P. Whittle. *Optimization over time, dynamic programming and stochastic control, Volume II.* Wiley, 1983.

P. Whittle. Restless bandits: Activity allocation in a changing world. *J. Appl. Probab.*, 25:287–298, 1988.

B. Widrow, N. K. Gupta, and S. Maitra. Punish/reward: learning with a critic in adaptive threshold systems. *IEEE Trans. Syst. Man. Cybern.*, 3(5):455–465, 1973.

A. A. Wieschenberg. Making mathematics: The coffee connection. *College Teaching*, 47:102–105, 1999.

R. Williams. Some statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992.

C. Xiao, I. Lee, B. Dai, D. Schuurmans, and C. Szepesvari. The curse of passive data collection in batch reinforcement learning. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pages 8413–8438, 2022.

M. Yadin and P. Naor. On queueing systems with variable service capacities. *Naval Res. Logist. Quart.*, 14:43–53, 1967.

Index