Chapter 7
Average Optimality for Unbounded Rewards

Average optimality has been studied in Chaps. 3 and 5 for finite and nonnegative models, respectively. However, these results are not applicable to the more general case where the reward function $r(i, a)$ has neither upper nor lower bounds. The average optimality for this general case will be studied in this chapter.

As in Chap. 6, *this chapter deals with the control model* (2.1).

7.1 Introduction

In Chap. 3, we have established the existence of an expected average reward (AR) optimal stationary policy for *finite* MDPs. The finiteness of the state and the action spaces is crucial, because, as in Puterman (1994) [129] or Sennott (1999) [141], we can provide an example for which no AR optimal policy exists when either the state space or the action space is infinite. Thus, to guarantee the existence of AR optimal policies for models with infinitely many states or actions, some conditions need to be imposed on the models. In short, the question dealt with in this chapter is: for non-finite MDPs, what conditions ensure the existence of an AR optimal stationary policy?

The first thing to do is to guarantee that the expected AR is indeed well defined. To this end, in Sect. 7.2, we introduce the concept of uniform exponential ergodicity under which the expected AR of a stationary policy is a constant independent of the initial state. In Sect. 7.3, we first give conditions for the existence of solutions to the *average-reward optimality equation* (AROE) (7.4) below. It is then shown that a stationary policy is AR optimal if and only if it attains the maximum in the AROE. Section 7.4 presents a policy iteration algorithm for the computation of the optimal AR function (i.e., the AR optimal value) and an AR optimal policy; in fact, this algorithm gives a solution to the AROE. Finally, in Sect. 7.5, we introduce some examples, and we conclude in Sect. 7.6 with some comments on the different approaches to obtain AR optimal policies.

X. Guo, O. Hernández-Lerma, *Continuous-Time Markov Decision Processes,*
Stochastic Modelling and Applied Probability 62,
DOI 10.1007/978-3-642-02547-1_7,

7.2 Exponential Ergodicity Conditions

Since the reward function $r(i,a)$ may be unbounded, the expected average reward $\bar{V}(i,\pi)$ of a policy $\pi \in \Pi$, defined in (2.21), may be infinite. To guarantee the finiteness of $\bar{V}(i,\pi)$, in the spirit of Lemma 6.3, we will impose the following "drift condition." For the simplicity of statements in this chapter, we will suppose that the state space S is the set of all nonnegative integers, that is, $S := \{0, 1, \ldots\}$.

Assumption 7.1

(a) (Drift condition) There exist a nondecreasing function $w \geq 1$ on S and constants $c_1 > 0$ and $b_1 \geq 0$ such that $\sum_{j \in S} w(j)q(j|i,a) \leq -c_1 w(i) + b_1 \delta_{i0}$ for all $(i,a) \in K$.
(b) $q^*(i) \leq L_0 w(i)$ for all $i \in S$, with $L_0 > 0$ and $q^*(i)$ as Assumption 2.2(b).
(c) $|r(i,a)| \leq M w(i)$ for all $(i,a) \in K$, with some $M > 0$.
(d) Assumption 6.8 holds with w there replaced by the nondecreasing function in (a) above.

The drift condition in Assumption 7.1(a) is also known as a Lyapunov or Foster–Lyapunov condition, and it is a key part of standard ergodicity hypothesis; see Chen (2000) [27], Lund et al. (1996) [113] and Meyn and Tweedie (1993) [116], among others.

Obviously, Assumption 7.1 implies Assumptions 2.2, 6.4, and 6.8. Hence, under Assumption 7.1, the transition function $p_\pi(s,i,t,j)$ is regular (see (2.8)) for every $\pi \in \Pi$.

As a consequence of (2.21)–(2.22) and Lemma 6.3, we have the following.

Lemma 7.2 *Suppose that Assumption 7.1 holds. Then the expected average reward is uniformly bounded, i.e.,*

$$\left|\bar{V}(i,\pi)\right| \leq M b_1 / c_1$$

for all $i \in S$ and $\pi \in \Pi$.

The next result will be used to show the existence of an expected AR optimal policy.

Proposition 7.3 *Suppose that Assumption 7.1 holds, and consider $\pi \in \Pi$, $u \in B_w(S)$, and a real number g. Then the following hold.*

(a) *If for all $i \in S$ and a.e. $t \geq 0$,*

$$g \geq r(i,\pi_t) + \sum_{j \in S} u(j) q(j|i,\pi_t),$$

then $g \geq \bar{V}(i,\pi)$ for all $i \in S$.

(b) *If for all* $i \in S$ *and a.e.* $t \geq 0$,

$$g \leq r(i, \pi_t) + \sum_{j \in S} u(j) q(j|i, \pi_t),$$

then $g \leq \bar{V}(i, \pi)$ *for all* $i \in S$.

Proof (a) As in the proof of (6.19), under the condition in (a), we have

$$gT \geq \int_0^T E_i^\pi r\big(x(t), \pi_t\big)\, dt + E_i^\pi u\big(x(T)\big) - u(i). \tag{7.1}$$

On the other hand, from Lemma 6.3 and Assumption 7.1(a) it is easily deduced that

$$\lim_{T \to \infty} \frac{1}{T} E_i^\pi u\big(x(T)\big) = 0 \quad \text{for all } \pi \in \Pi \text{ and } i \in S.$$

This fact, together with (2.22) and (7.1), gives (a).

(b) The proof of (b) is similar. □

Now we focus on the issue of the existence of AR optimal stationary policies (recall Definition 2.8). To do so, in addition to Assumption 7.1, we impose the following irreducibility condition.

Assumption 7.4 For each $f \in F$, the corresponding Markov process $\{x(t)\}$ with transition function $p_f(i, t, j)$ is irreducible, which means that, for any two states $i \neq j$, there exists a set of distinct states $i = i_1, \ldots, i_m$ such that

$$q(i_2|i_1, f) \cdots q(j|i_m, f) > 0.$$

Under Assumptions 7.1(a) and 7.4, for each $f \in F$, Propositions C.11 and C.12 yield that the Markov chain $\{x(t)\}$ has a unique invariant probability measure, denoted by μ_f, which satisfies that $\mu_f(j) = \lim_{t \to \infty} p_f(i, t, j)$ (independent of $i \in S$) for all $j \in S$. Thus, by Assumption 7.1(a) and Lemma 6.3(i), we have

$$\mu_f(w) := \sum_{j \in S} w(j) \mu_f(j) \leq \frac{b_1}{c_1},$$

which shows that the μ_f-expectation of w (i.e., $\mu_f(w)$) is finite. Therefore, for all $f \in F$ and $u \in B_w(S)$, the inequality $|u(i)| \leq \|u\|_w w(i)$ for all $i \in S$ gives that the expectation

$$\mu_f(u) := \sum_{i \in S} u(i) \mu_f(i) \tag{7.2}$$

exists and is finite.

Assumption 7.5 The control model (2.1) is uniformly w-exponentially ergodic, which means the following: there exist constants $\delta > 0$ and $L_2 > 0$ such that (using

the notation in (7.2))

$$\sup_{f\in F}\left|E_i^f u\big(x(t)\big) - \mu_f(u)\right| \le L_2 e^{-\delta t}\|u\|_w w(i)$$

for all $i \in S$, $u \in B_w(S)$, and $t \ge 0$.

It is worth noting that, under our current assumptions, the gain of a deterministic stationary policy $f \in F$ is constant, i.e., $\bar{V}(i, f)$ does not depend on the initial state i and equals the μ_f-expectation of $r(f)$. More precisely, let

$$g(f) := \sum_{j\in S} r(j, f)\mu_f(j).$$

Then, by (2.22),

$$V_T(i, f) = Tg(f) + E_i^f \int_0^T \big[r\big(x(t), f\big) - g(f)\big]\,dt.$$

Hence, multiplying by $1/T$ and letting $T \to \infty$, from Assumption 7.5 we obtain

$$\bar{V}(i, f) = g(f) = \sum_{j\in S} r(j, f)\mu_f(j) \quad \forall i \in S. \tag{7.3}$$

Among Assumptions 7.1, 7.4, and 7.5 made so far on the control model, Assumption 7.5 seems to be the most difficult to verify in practice. Hence, we next propose sufficient conditions for uniform w-exponential ergodicity.

Proposition 7.6 *In addition to Assumptions* 7.1(a) *and* 7.4, *suppose that*, *for each fixed* $f \in F$,

(i) (*Stochastic monotonicity condition*) $\sum_{j\ge k} q(j|i, f) \le \sum_{j\ge k} q(j|i+1, f)$ *for all* $i, k \in S$ *such that* $k \ne i+1$.
(ii) *For each* $j > i > 0$, *there exist nonzero distinct states* $i_1, i_2, \ldots, i_m \ge j$ *such that*

$$q(i_1|i, f)\cdots q(i_m|i_{m-1}, f) > 0.$$

Then Assumption 7.5 *holds with* $\delta := c_1$ *and* $L_2 := 2(1 + \frac{b_1}{c_1})$, *where* c_1 *and* b_1 *are the constants in Assumption* 7.1.

This proposition obviously follows from Proposition C.17.

Condition (i) in Proposition 7.6 is a variant of the "monotonicity conditions" in Anderson (1991) [4, p. 249]. Condition (ii) requires that, for any two states $j > i > 0$, the process $\{x(t)\}$ can travel with positive probability from the state i to the set $\{j, j+1, \ldots\}$ without passing through the state $0 \in S$.

Other sufficient conditions for uniform w-exponential ergodicity are given by Guo, Hernández-Lerma, and Prieto-Rumeau (2006) [65] and Prieto-Rumeau and Hernández-Lerma (2006) [126], for instance.

7.3 The Existence of AR Optimal Policies

We begin with introducing the *average-reward optimality equation* (AROE) (7.4) below.

A pair $(g^*, u) \in \mathbb{R} \times B_w(S)$ is said to be a solution to the AROE if

$$g^* = \sup_{a \in A(i)} \left\{ r(i,a) + \sum_{j \in S} u(j) q(j|i,a) \right\} \quad \forall i \in S. \tag{7.4}$$

Under some assumptions, the supremum in (7.4) can be attained for every $i \in S$. In such a case, we say that $f \in F$ attains the maximum in the AROE (7.4), that is,

$$g^* = r(i,f) + \sum_{j \in S} u(j) q(j|i,f) \quad \forall i \in S. \tag{7.5}$$

Lemma 7.7 *Suppose that Assumptions 7.1, 7.4, and 7.5 are satisfied. Consider an arbitrary fixed state $i_0 \in S$. Then, for all $f \in F$ and discount factors $\alpha > 0$, the relative differences of the discounted-reward function $V_\alpha(f)$, namely,*

$$u_\alpha^f(i) := V_\alpha(i,f) - V_\alpha(i_0,f) \quad \text{for } i \in S, \tag{7.6}$$

are uniformly w-bounded in $\alpha > 0$ and $f \in F$. More precisely, we have

$$\|u_\alpha^f\|_w \le \frac{L_2 M}{\delta}\big[1 + w(i_0)\big] \quad \forall \alpha > 0 \text{ and } f \in F.$$

Proof Choose any $\alpha > 0$ and $f \in F$. By Assumption 7.1(c), $|r(i,f)| \le M w(i)$ for all $i \in S$. Recalling the notation in (7.3), from Assumptions 7.4 and 7.5 we have

$$\big|E_i^f r\big(x(t), f\big) - g(f)\big| \le L_2 M e^{-\delta t} w(i) \quad \forall i \in S. \tag{7.7}$$

Thus, for each $i \in S$, by (7.7) and (2.19) we have

$$\begin{aligned}
|u_\alpha^f(i)| &= \left| E_i^f\left[\int_0^\infty e^{-\alpha t} r\big(x(t), f\big)\, dt\right] - E_{i_0}^f\left[\int_0^\infty e^{-\alpha t} r\big(x(t), f\big)\, dt\right]\right| \\
&\le L_2 M \int_0^\infty e^{-(\alpha+\delta)t}\big(w(i) + w(i_0)\big)\, dt \\
&= \frac{L_2 M}{\alpha+\delta}\big(w(i) + w(i_0)\big) \\
&\le \frac{L_2 M}{\delta}\big(1 + w(i_0)\big) w(i),
\end{aligned}$$

which completes the proof. □

Now we present the main result of this section.

Theorem 7.8 *Suppose that Assumptions 7.1, 7.4, and 7.5 hold. Then*:

(a) *There exists a solution* $(g^*, \bar{u}) \in \mathbb{R} \times B_w(S)$ *to AROE* (7.4). *Moreover, the constant* g^* *coincides with the optimal average reward function* $\bar{V}^*$, *i.e.*,

$$g^* = \bar{V}^*(i) \quad \forall i \in S,$$

and $\bar{u}$ *is unique up to additive constants*.

(b) *A deterministic stationary policy is AR optimal if and only if it attains the maximum in AROE* (7.4).

Proof The proof proceeds in several steps. First, in (7.6) we take f as f^*_α, the α-discounted reward optimal stationary policy in Theorem 6.10; hence $V_\alpha(i, f^*_\alpha) = V^*_\alpha(i)$, the optimal discounted reward function, and instead of the function in (7.6), we now take $u_\alpha(i) := u_\alpha^{f^*_\alpha}(i)$. Then following the vanishing discount approach already used in Sects. 5.4 and 5.5, we will show the existence of a solution to the AROE.

Lemma 7.7 and Proposition A.7 give the existence of a sequence $\{\alpha_n\}$ of discount factors such that $\alpha_n \downarrow 0$, a constant g^*, and a function $\bar{u} \in B_w(S)$ such that

$$\lim_{n\to\infty} \alpha_n V^*_{\alpha_n}(i_0) = g^* \quad \text{and} \quad \lim_{n\to\infty} u_{\alpha_n}(i) = \bar{u}(i) \tag{7.8}$$

for all $i \in S$. (Observe the analogy between (7.8) and (5.12).) On the other hand, for all $n \geq 1$ and $i \in S$, by Theorem 6.10(b) we have (using the notation in (6.11) and (7.6))

$$\frac{\alpha_n V^*_{\alpha_n}(i_0)}{m(i)} + \frac{\alpha_n u_{\alpha_n}(i)}{m(i)} + u_{\alpha_n}(i) \geq \frac{r(i,a)}{m(i)} + \sum_{j\in S} u_{\alpha_n}(j)\left[\frac{q(j|i,a)}{m(i)} + \delta_{ij}\right]$$

for all $(i, a) \in K$, which, together with (7.8), gives

$$\frac{g^*}{m(i)} + \bar{u}(i) \geq \frac{r(i,a)}{m(i)} + \sum_{j\in S} \bar{u}(j)\left[\frac{q(j|i,a)}{m(i)} + \delta_{ij}\right] \quad \forall (i,a) \in K.$$

Thus,

$$g^* \geq \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} \bar{u}(j) q(j|i,a) \right\} \quad \forall i \in S. \tag{7.9}$$

To prove that $(g^*, \bar{u})$ is a solution to the AROE (7.4), it remains to show the reverse inequality in (7.9). As a consequence of Theorem 6.10, for each $n \geq 1$, there exists $f_n \in F$ such that

$$\frac{\alpha_n V^*_{\alpha_n}(i_0)}{m(i)} + \frac{\alpha_n u_{\alpha_n}(i)}{m(i)} + u_{\alpha_n}(i) = \frac{r(i,f_n)}{m(i)} + \sum_{j\in S} u_{\alpha_n}(j)\left[\frac{q(j|i,f_n)}{m(i)} + \delta_{ij}\right] \tag{7.10}$$

for all $i \in S$. By Assumption 7.1(d), F is compact, and thus we may suppose that there exists a policy $f' \in F$ such that

$$\lim_{n\to\infty} f_n(i) = f'(i) \quad \forall i \in S.$$

Letting $n \to \infty$ in (7.10) and applying Proposition A.4, we obtain

$$\frac{g^*}{m(i)} + \bar{u}(i) = \frac{r(i, f')}{m(i)} + \sum_{j\in S} \bar{u}(j)\left[\frac{q(j|i, f')}{m(i)} + \delta_{ij}\right] \quad \forall i \in S,$$

which can be rewritten as

$$\begin{aligned} g^* &= r(i, f') + \sum_{j\in S} \bar{u}(j) q(j|i, f') \\ &\leq \sup_{a\in A(i)} \left\{ r(i, a) + \sum_{j\in S} \bar{u}(j) q(j|i, a) \right\} \quad \forall i \in S. \end{aligned} \tag{7.11}$$

Hence, from (7.9) and (7.11) it follows that $(g^*, \bar{u})$ is a solution of the AROE.

Next, we are going to prove that $g^* = \bar{V}^*(i)$ for every $i \in S$.

Pick an arbitrary policy $\pi \in \Pi$. It follows from the AROE (7.4), together with (2.5) and (2.20), that

$$g^* \geq r(i, \pi_t) + \sum_{j\in S} \bar{u}(j) q(j|i, \pi_t) \quad \forall i \in S \text{ and } t \geq 0. \tag{7.12}$$

Then, by Proposition 7.3 we have $g^* \geq \bar{V}(i, \pi)$, and so, since $\pi \in \Pi$ is arbitrary, $g^* \geq \bar{V}^*(i)$ for all $i \in S$.

Observe now that our assumptions ensure the existence of a policy $f^* \in F$ attaining the maximum in the AROE, that is,

$$g^* = r\big(i, f^*\big) + \sum_{j\in S} \bar{u}(j) q\big(j|i, f^*\big) \quad \forall i \in S.$$

Therefore, Proposition 7.3 gives $g^* = \bar{V}(i, f^*)$ for all $i \in S$. As a consequence, $g^* = \bar{V}^*(i)$ for every $i \in S$, and, moreover, f^* is AR optimal.

Finally, note that, by (7.3),

$$\bar{V}(i, f) = \sum_{j\in S} r(j, f)\mu_f(j) = g(f) \quad \forall f \in F \text{ and } i \in S. \tag{7.13}$$

Our next step in the proof of Theorem 7.8 is to show that a necessary and sufficient condition for a deterministic stationary policy to be AR optimal is that it attains the maximum in the AROE.

In fact, we have already proved the sufficiency part. We will prove the necessity by contradiction. Thus, suppose that $f^* \in F$ is an AR optimal policy that does not

attain the maximum in the AROE (7.4). Then there exist $i' \in S$ and a constant $\beta > 0$ (depending on i' and f^*) such that

$$g^* \geq r\big(i, f^*\big) + \beta\delta_{i'i} + \sum_{j\in S} \bar{u}(j)q\big(j|i, f^*\big) \quad \forall i \in S. \tag{7.14}$$

By the irreducibility in Assumption 7.4, the invariant probability measure μ_{f^*} of $p_{f^*}(i, t, j)$ is supported on all of S, meaning that $\mu_{f^*}(j) > 0$ for every $j \in S$. Then, as in the proof of (7.13), from (7.14) and Proposition 7.3 we have

$$g^* \geq g\big(f^*\big) + \beta\mu_{f^*}(i') > g\big(f^*\big), \tag{7.15}$$

which contradicts the fact that f^* is AR optimal.

Therefore, to complete the proof of this theorem, it only remains to show that the function $\bar{u}$ in the AROE is unique up to additive constants since the constant in the AROE equals the optimal average-reward function. To this end, for each $f \in F$, let us define an operator Q^f on $B_w(S)$ as

$$\big(Q^f h\big)(i) := \sum_{j\in S} h(j)q(j|i, f) \quad \text{for } h \in B_w(S) \text{ and } i \in S. \tag{7.16}$$

Suppose now that $(g^*, \bar{u})$ and (g^*, u') are two solutions to the AROE and that $f^* \in F$ is AR optimal. Then (by part (b)) f^* attains the maximum in the *two* AROEs. Hence, subtracting the two AROEs gives

$$\sum_{j\in S}\big(\bar{u}(j) - u'(j)\big)q\big(j|i, f^*\big) = \big(Q^{f^*}(\bar{u} - u')\big)(i) = 0 \quad \forall i \in S. \tag{7.17}$$

Then, as in the proof of (7.1), by (7.17) and (6.19) (with f^* in lieu of π) we can derive that

$$E_i^{f^*}\big[\bar{u}\big(x(t)\big) - u'\big(x(t)\big)\big] = \bar{u}(i) - u'(i) \quad \forall i \in S.$$

Letting $t \to +\infty$ in this equality, it follows from Assumption 7.5 that

$$\mu_{f^*}(\bar{u} - u') = \bar{u}(i) - u'(i) \quad \forall i \in S,$$

showing that the functions $\bar{u}$ and u' differ by the constant $\mu_{f^*}(\bar{u} - u')$. □

From Proposition 7.3 and the proof of Theorem 7.8 we obtain the following.

Corollary 7.9 *Suppose that Assumptions 7.1, 7.4, and 7.5 hold, and let* $f \in F$, $g \in \mathbb{R}$, *and* $u \in B_w(S)$. *Then the following facts hold*:

(a) *If*

$$g \geq r(i, f) + \sum_{j\in S} u(j)q(j|i, f) \quad \forall i \in S,$$

then $g \geq \bar{V}(i, f) = \sum_{j\in S} r(j, f)\mu_f(j)$ *for all* $i \in S$.

(b) *If*

$$g = r(i, f) + \sum_{j \in S} u(j) q(j|i, f) \quad \forall i \in S,$$

then $g = \bar{V}(i, f) = \sum_{j \in S} r(j, f) \mu_f(j)$ *for all* $i \in S$.

Remark 7.10 The AROE (7.4) is obviously equivalent to

$$\frac{g^*}{m(i)} + \bar{u}(i) = \sup_{a \in A(i)} \left\{ \frac{r(i, a)}{m(i)} + \sum_{j \in S} \bar{u}(j) p(j|i, a) \right\} \quad \forall i \in S$$

(recall the notation in (6.12)), which is different from the discrete-time AROE (see, e.g., Hernández-Lerma and Lasserre (1996) [73], Puterman (1994) [129], and Sennott (1999) [141]) because of the denominator $m(i)$. This difference can also be explained as in Remark 6.7.

We denote by $\mathbf{F}_{\text{ao}}$ the family of AR optimal deterministic stationary policies, and by $\mathbf{F}_{\text{ca}}$ the set of *canonical policies* defined as the policies in F attaining the maximum in the AROE (7.4). Theorem 7.8(b) above shows that, in fact,

$$\mathbf{F}_{\text{ao}} = \mathbf{F}_{\text{ca}}.$$

If we drop the irreducibility hypothesis in Assumption 7.4, then $\mathbf{F}_{\text{ca}} \subseteq \mathbf{F}_{\text{ao}}$ still holds, but the reverse relationship, $\mathbf{F}_{\text{ao}} \subseteq \mathbf{F}_{\text{ca}}$, may fail. That is, we may have an AR optimal policy that is not canonical, which is in fact a situation that we already encountered in Example 5.2 and Proposition 5.11.

7.4 The Policy Iteration Algorithm

In this section, we give a policy iteration algorithm to obtain an AR optimal policy. *Throughout this section, we suppose that Assumptions* 7.1, 7.4, *and* 7.5 *are all satisfied.*

The Bias of a Stationary Policy Let $f \in F$. We say that a pair $(g, h) \in \mathbb{R} \times B_w(S)$ is a solution to the *Poisson equation* for $f \in F$ if

$$g = r(i, f) + \sum_{j \in S} h(j) q(j|i, f) \quad \forall i \in S.$$

Now recalling (7.13), the expected average reward (or gain) of f is

$$\bar{V}(i, f) = \mu_f\big(r(\cdot, f)\big) = g(f) \quad \forall i \in S.$$

We define the *bias* (or "potential"—see Remark 3.2) of f as

$$h_f(i) := \int_0^\infty \left[E_i^f r\big(x(t), f\big) - g(f)\right] dt \quad \text{for } i \in S. \tag{7.18}$$

By (7.7), h_f is finite and in $B_w(S)$. Moreover, (7.7) yields that the bias is uniformly bounded in the w-norm because

$$\sup_{f \in F} \|h_f\|_w \le L_2 M/\delta. \tag{7.19}$$

Proposition 7.11 *For every $f \in F$, the solutions to the Poisson equation for f are of the form*

$$\big(g(f), h_f + z\big) \quad \textit{with } z \textit{ any real number.}$$

Moreover, $(g(f), h_f)$ is the unique solution to the Poisson equation

$$g(f) = r(i, f) + \sum_{j \in S} h_f(j) q(j|i, f) \quad \forall i \in S \tag{7.20}$$

for which $\mu_f(h_f) = 0$.

Proof First of all, we will prove that $(g(f), h_f)$ is indeed a solution to the Poisson equation for f. Our assumptions (in particular, Assumptions 7.1 and 7.5) allow us to interchange the sums and integrals in the following equations:

$$\begin{aligned}
\sum_{j \in S} h_f(j) q(j|i, f) &= \sum_{j \in S} \left[\int_0^\infty \big(E_j^f r\big(x(t), f\big) - g(f)\big)\, dt \right] q(j|i, f) \quad \big[\text{by (7.18)}\big] \\
&= \int_0^\infty \sum_{j \in S} E_j^f r\big(x(t), f\big) q(j|i, f)\, dt \\
&\qquad \left[\text{since } \sum_{j \in S} q(j|i, f) = 0 \text{ by (2.3)}\right] \\
&= \sum_{j \in S} \int_0^\infty r(j, f) \frac{d}{dt} p_f(i, t, j)\, dt \quad \big[\text{by (2.6)}\big] \\
&= \sum_{j \in S} \Big[r(j, f) \lim_{t \to +\infty} p_f(i, t, j) \Big] - r(i, f) \\
&= \sum_{j \in S} r(j, f) \mu_f(j) - r(i, f) = g(f) - r(i, f),
\end{aligned}$$

as we wanted to prove.

Suppose now that (g, h) is a solution to the Poisson equation for f, that is, $(g, h) \in \mathbb{R} \times B_w(S)$, and

$$g = r(i, f) + \sum_{j \in S} h(j) q(j|i, f) \quad \forall i \in S.$$

Therefore, by Corollary 7.9(b), we have $g = g(f)$ because $\mu_f(r(\cdot, f)) = g(f)$.

Suppose now that $(g(f), h)$ and $(g(f), h')$ are two solutions to the Poisson equation for $f \in F$. Subtracting the corresponding Poisson equations, we get (cf. (7.17))

$$\sum_{j \in S} \big(h(j) - h'(j)\big) q(j|i, f) = 0 \quad \forall i \in S.$$

Then, as in the proof of Theorem 7.8(a), we can prove that h and h' are equal up to an additive constant.

To complete the proof, it remains to show that $\mu_f(h_f) = 0$, but this is obtained by taking the μ_f-expectation in (7.18). □

Remark 7.12 Given $f \in F$, we can determine the gain and the bias of f by solving the following system of linear equations.

First, determine the i.p.m. (invariant probability measure) μ_f as the unique non-negative solution (by Proposition C.12) to

$$\begin{cases} \sum_{i \in S} q(j|i, f) \mu_f(i) = 0 \quad \forall j \in S, \\ \sum_{j \in S} \mu_f(j) = 1. \end{cases}$$

Then, as a consequence of Proposition 7.11, the gain $g(f) = \sum_{j \in S} r(j, f) \times \mu_f(j) \in \mathbb{R}$ and the bias $h_f \in B_w(S)$ of f form the unique solution to the system of linear equations

$$\begin{cases} g = r(i, f) + \sum_{j \in S} h(j) q(j|i, f) \quad \text{for } i \in S, \\ \sum_{i \in S} h(i) \mu_f(i) = 0. \end{cases}$$

Policy Iteration Algorithm 7.1 This algorithm is a standard tool to analyze MDPs. It works as follows:

Step I. Choose an arbitrary policy $f \in F$.
Step II. Determine the gain $g(f)$ and the bias h_f of f as in Remark 7.12.
Step III. Define a policy $f' \in F$ in the following way: for each $i \in S$, if

$$r(i, f) + \sum_{j \in S} h_f(j) q(j|i, f) = \max_{a \in A(i)} \left\{ r(i, a) + \sum_{j \in S} h_f(j) q(j|i, a) \right\}, \tag{7.21}$$

let $f'(i) := f(i)$; otherwise (i.e., if in (7.21) we have a strict inequality), choose $f'(i) \in A(i)$ such that

$$r(i, f') + \sum_{j \in S} h_f(j) q(j|i, f') = \max_{a \in A(i)} \left\{ r(i, a) + \sum_{j \in S} h_f(j) q(j|i, a) \right\}. \tag{7.22}$$

Step IV. If $f' = f$ (or, equivalently, if (7.21) holds for every $i \in S$), then f is an AR optimal policy, and the algorithm stops. Otherwise, replace f with f' and return to Step II.

Let $f_0 \in F$ be the initial policy in the policy iteration algorithm (see Step I), and let $\{f_n\}$ be the sequence of stationary policies obtained by the repeated application of the algorithm.

If $f_n = f_{n+1}$ for some n, then it follows from Proposition 7.11 that the pair $(g(f_n), h_{f_n})$ is a solution to the AROE, and thus, by Theorem 7.8, f_n is AR optimal.

Hence, to analyze the convergence of the policy iteration algorithm, we will consider the case

$$f_n \neq f_{n+1} \quad \text{for every } n \geq 0. \tag{7.23}$$

Define, for $n \geq 1$ and $i \in S$,

$$\varepsilon(f_n, i) := r(i, f_n) + \sum_{j \in S} h_{f_{n-1}}(j) q(j|i, f_n) - \left[r(i, f_{n-1}) + \sum_{j \in S} h_{f_{n-1}}(j) q(j|i, f_{n-1}) \right],$$

which by Proposition 7.11 can be expressed as

$$\varepsilon(f_n, i) = r(i, f_n) + \sum_{j \in S} h_{f_{n-1}}(j) q(j|i, f_n) - g(f_{n-1}). \tag{7.24}$$

Observe (by Step III above) that $\varepsilon(f_n, i) = 0$ if $f_n(i) = f_{n-1}(i)$, whereas $\varepsilon(f_n, i) > 0$ if $f_n(i) \neq f_{n-1}(i)$. Hence, $\varepsilon(f_n, i)$ can be interpreted as the "improvement" of the nth iteration of the algorithm.

Lemma 7.13 *Suppose that (7.23) is satisfied. Then the following statements hold.*

(a) *The sequence $\{g(f_n)\}$ is strictly increasing and it has a finite limit.*
(b) *For every $i \in S$, $\varepsilon(f_n, i) \to 0$ as $n \to \infty$.*

Proof (a) As in the proof of (7.3), from (7.24) we have

$$\sum_{i \in S} \varepsilon(f_n, i) \mu_{f_n}(i) = g(f_n) - g(f_{n-1}) \quad \forall n \geq 1. \tag{7.25}$$

On the other hand, the hypothesis (7.23) implies that, for every $n \geq 1$, there exists some $i \in S$ with $\varepsilon(f_n, i) > 0$, and, besides, by the irreducibility condition in Assumption 7.4, $\mu_{f_n}(i) > 0$ for every $i \in S$. Hence, $g(f_n) > g(f_{n-1})$, as we wanted to prove. Moreover, as a consequence of Lemma 7.2, the sequence $\{g(f_n)\}$ is bounded above and, therefore, converges to some finite limit.

(b) Let $\mu(i) := \inf_{f \in F} \mu_f(i)$ for all $i \in S$. We will show that $\mu(i) > 0$ for all $i \in S$. To this end, first *fix* an arbitrary state $i \in S$. Since F is compact, there exist a sequence $\{f_m\}$ and f (depending on i) in F such that

$$\mu(i) = \lim_{m \to \infty} \mu_{f_m}(i) \quad \text{and} \quad \lim_{m \to \infty} f_m(j) = f(j) \quad \forall j \in S. \tag{7.26}$$

On the other hand, we can prove that $\lim_{m \to \infty} p_{f_m}(i, t, j) = p_f(i, t, j)$ for all $i, j \in S$ and $t \geq 0$. To do so, let

$$\gamma_{ij}^{\alpha}(h) := \int_0^{\infty} e^{-\alpha t} p_h(i, t, j)\, dt \quad (\text{for each } h \in F, i, j \in S, \alpha > 0)$$

be the Laplace transform of $p_h(i, t, j)$. Then, we have

$$\gamma_{ij}^{\alpha}(h) \geq 0, \quad \alpha \sum_{j \in S} \gamma_{ij}^{\alpha}(h) = 1 \quad \forall h \in F, i, j \in S \text{ and } \alpha > 0,$$

and it follows from Theorem 6.9(c) that, for all $m \geq 1$ and $j, k \in S$,

$$\lim_{m \to \infty} q(k|j, f_m) = q(k|j, f), \qquad \alpha \gamma_{ij}^{\alpha}(f_m) = \delta_{ij} + \sum_{k \in S} \gamma_{ik}^{\alpha}(f_m) q(k|j, f_m). \tag{7.27}$$

We now fix $i, j \in S$ and $\alpha > 0$, and then choose an arbitrary subsequence $\{\gamma_{ij}^{\alpha}(f_{m_k})\}$ of $\{\gamma_{ij}^{\alpha}(f_m)\}$ converging to a nonnegative number v_{ij}^{α}. Since S is denumerable, there exists a subsequence $\{l\}$ of $\{m_k\}$ such that $\{\gamma_{i'j'}^{\alpha}(f_l)\}$ converges to a nonnegative number $v_{i'j'}^{\alpha}$ for each $i', j' \in S$ (as $l \to \infty$). Hence, by (7.26)–(7.27) and Proposition A.4 we have

$$\alpha v_{ij}^{\alpha} = \delta_{ij} + \sum_{k \in S} v_{ik}^{\alpha} q(k|j, f). \tag{7.28}$$

By the uniqueness of solution to (7.28) in Theorem 6.9(c) we have $v_{ij}^{\alpha} = \gamma_{ij}^{\alpha}(f)$. Then, as the above subsequence $\{\gamma_{ij}^{\alpha}(f_{m_k})\}$ was arbitrarily chosen and all such subsequences have the same limit $\gamma_{ij}^{\alpha}(f)$, we obtain $\lim_{m \to \infty} \gamma_{ij}^{\alpha}(f_m) = \gamma_{ij}^{\alpha}(f)$. This means that $\gamma_{ij}^{\alpha}(f)$ is continuous on F, and so is $p_f(i, t, j)$ in $f \in F$. Thus, replacing f and the function $r(j, f)$ in (7.7) with f_n and the function δ_{ij}, respectively, and then letting $m \to \infty$, by (7.26) we obtain

$$\left| p_f(i, t, i) - \mu(i) \right| \leq L_2 e^{-\delta t} w(i).$$

Consequently, $\mu(i) = \lim_{t \to \infty} p_f(i, t, i) = \mu_f(i) > 0$. Therefore, since $i \in S$ was arbitrary, $\mu(i) := \inf_{f \in F} \mu_f(i)$ is positive for all $i \in S$. Hence, recalling that

$\varepsilon(f_n, i) \geq 0$ for all $n \geq 1$ and $i \in S$, by (7.25) we see that, for any $i \in S$,

$$0 \leq \varepsilon(f_n, i)\mu(i) \leq \varepsilon(f_n, i)\mu_{f_n}(i) \leq g(f_n) - g(f_{n-1}).$$

This implies (since $\mu(i) > 0$ and recalling that $\{g(f_n)\}$ is converging) that $\lim_{n\to\infty} \varepsilon(f_n, i) = 0$. □

Proposition 7.14 *Let $\{f_n\}$ be the sequence obtained from the policy iteration algorithm 7.1. Then $g(f_n)$ converges to g^*, the optimal average reward function of the continuous-time MDP, and any limit point $f \in F$ of the sequence $\{f_n\}$ is an AR optimal stationary policy.*

Proof First of all, without loss of generality we suppose that the sequence $\{f_n\}$ satisfying (7.23) has limit points since F is compact.

By (7.19), there exists a subsequence $\{f_m\}$ of f_n such that h_{f_m} converges pointwise to some $h \in B_w(S)$. Therefore, we have

$$\begin{aligned} g(f_m) \to g \quad &\big[\text{Lemma 7.13(a)}\big], \qquad f_m \to f, \quad \text{and} \\ h_{f_m} \to h \quad &[\text{pointwise}]. \end{aligned} \tag{7.29}$$

Now, by Proposition 7.11 and the definition of the improvement term $\varepsilon(f_{m+1}, i)$ in (7.24), we have

$$\begin{aligned} g(f_m) &= r(i, f_m) + \sum_{j\in S} h_{f_m}(j) q(j|i, f_m) \\ &= \max_{a\in A(i)} \left\{ r(i, a) + \sum_{j\in S} h_{f_m}(j) q(j|i, a) \right\} - \varepsilon(f_{m+1}, i) \end{aligned} \tag{7.30}$$

for all $i \in S$. As in the proof of Theorem 7.8(a), letting $m \to \infty$ in (7.30), by (7.29) and Lemma 7.13(b) we have

$$\begin{aligned} g &= r(i, f) + \sum_{j\in S} h(j) q(j|i, f) \\ &= \max_{a\in A(i)} \left\{ r(i, a) + \sum_{j\in S} h(j) q(j|i, a) \right\} \end{aligned} \tag{7.31}$$

for all $i \in S$. This shows (by Theorem 7.8(b)) that f is AR optimal and also that g equals the optimal AR function. □

We summarize our results in the following theorem.

Theorem 7.15 *Suppose that Assumptions 7.1, 7.4, and 7.5 hold, and let $f_1 \in F$ be an arbitrary initial policy for the policy iteration algorithm 7.1. Let $\{f_n\} \subseteq F$ be the sequence of policies obtained by the policy iteration algorithm 7.1. Then one of the following results hold.*

(a) *Either*
   (i) *the algorithm converges in a finite number of iterations to an AR optimal policy, or*
   (ii) *as* $n \to \infty$, *the sequence* $\{g(f_n)\}$ *converges to the optimal AR function value* g^*, *and any limit point of* $\{f_n\}$ *is an AR optimal stationary policy.*

(b) *There exists a subsequence* $\{f_m\} \subset \{f_n\}$ *for which* (7.29) *holds, and, in addition, the limiting triplet* $(g, f, h) \in \mathbb{R} \times F \times B_w(S)$ *satisfies* (7.31), *so that* (g, h) *satisfies the AROE, and* f *is a canonical policy.*

As a consequence of Theorems 7.15 and 7.8(b), if there exists a unique AR optimal stationary policy $f^* \in F$, then the whole sequence $\{f_n\}$ converges pointwise to f^*.

7.5 Examples

This section presents some applications of the main results in this chapter.

Example 7.1 (Average-optimal control of a birth-and-death system) Consider a controlled birth-and-death system in which the state variable denotes the population size at any time $t \geq 0$. The birth rate is assumed to be a *fixed* constant $\lambda > 0$, but the death rates μ are assumed to be controlled by a decision-maker; hence, we interpret a death rate μ as an *action* a (i.e., $\mu =: a$). When the system's state is $i \in S := \{0, 1, \ldots\}$, the decision-maker takes an action a from a given set $A(i) \equiv [\mu_1, \mu_2]$ with $\mu_2 > \mu_1 > 0$, which increases or decreases the death rates given by (7.33)–(7.34) below. This action incurs a cost $c(i, a)$. In addition, suppose that there is a benefit represented by $p > 0$ for each unit of time, and then the decision-maker gets a reward pi for each unit of time during which the system remains in state i.

We now formulate this system as a continuous-time MDP. The corresponding transition rates $q(j|i, a)$ are given as follows: for each $a \in [\mu_1, \mu_2]$,

$$q(1|0, a) = -q(0|0, a) := \lambda, \quad \text{and} \quad q(j|0, a) = 0 \quad \text{for } j \geq 2, \tag{7.32}$$

$$q(0|1, a) := a, \qquad q(1|1, a) = -a - \lambda, \qquad q(2|1, a) := \lambda, \qquad q(j|1, a) = 0 \tag{7.33}$$

for all $j \geq 3$. For all $i \geq 2$ and $a \in A(i) = [\mu_1, \mu_2]$,

$$q(j|i, a) := \begin{cases} p_1 ai & \text{if } j = i - 2, \\ p_2 ai & \text{if } j = i - 1, \\ -(a + \lambda)i & \text{if } j = i, \\ \lambda i & \text{if } j = i + 1, \\ 0 & \text{otherwise,} \end{cases} \tag{7.34}$$

where $p_1 \geq 0$ and $p_2 \geq 0$ are fixed constants such that $p_1 + p_2 = 1$.

By the model's description we see that the reward rates $r(i,a)$ are of the form

$$r(i,a) := pi - c(i,a) \quad \text{for } (i,a) \in K := \big\{(i,a) : i \in S, a \in A(i)\big\}. \tag{7.35}$$

We wish to find conditions that ensure the existence of an AR optimal stationary policy. To do this, we consider the following assumptions:

D$_1$. $\mu_1 - \lambda > 0$.
D$_2$. $p_1 \le \frac{\mu_1}{2\mu_2}$ with p_1 as in (7.34). (This condition trivially holds when $p_1 = 0$.)
D$_3$. The function $c(i,a)$ is continuous in $a \in A(i) = [\mu_1, \mu_2]$ for each fixed $i \in S$, and $\sup_{a\in A(i)} |c(i,a)| < \tilde{M}(i+1)$ for all $i \in S$, for some constant $\tilde{M} \ge 0$.

Under these conditions, we obtain the following.

Proposition 7.16 *Under conditions* **D$_1$**, **D$_2$**, *and* **D$_3$**, *the above controlled birth-and-death system satisfies Assumptions* 7.1, 7.4, *and* 7.5. *Therefore* (*by Theorem* 7.8), *there exists an AR optimal stationary policy.*

Proof We shall first verify Assumption 7.1. Let $c_1 := \frac{1}{2}(\mu_1 - \lambda) > 0$ (by **D$_1$**), and $w(i) := i + 1$ for all $i \in S$. Then, from (7.32) and (7.33) we have

$$\sum_{j\in S} w(j)q(j|0,a) = \lambda \le -c_1 w(0) + \mu_1 + \lambda \quad \forall a \in A(i); \tag{7.36}$$

$$\sum_{j\in S} w(j)q(j|1,a) = -(a-\lambda) \le -c_1 w(1) \quad \forall a \in A(i). \tag{7.37}$$

Moreover, for all $i \ge 2$ and $a \in [\mu_1, \mu_2]$, from (7.34) we have

$$\begin{aligned}\sum_{j\in S} w(j)q(j|i,a) &= -(a + ap_1 - \lambda)i \\ &\le -\frac{2}{3}(a + ap_1 - \lambda)w(i) \le -c_1 w(i).\end{aligned} \tag{7.38}$$

By inequalities (7.36)–(7.38) we have, for all $i \in S$ and $a \in A(i)$,

$$\begin{aligned}\sum_{j\in S} w(j)q(j|i,a) &\le -c_1 w(i) + (\mu_1 + \lambda)\delta_{i0} \\ &\le -c_1 w(i) + \mu_1 + \lambda,\end{aligned} \tag{7.39}$$

which verifies Assumption 7.1(a). On the other hand, by (7.32)–(7.34), we have $q^*(i) \le (\mu_2 + \lambda)(i+1) = (\mu_2 + \lambda)w(i)$, and so Assumption 7.1(b) follows. By (7.35) and **D$_3$** we have $|r(i,a)| \le (p + \tilde{M})w(i)$ for all $i \in S$, which implies Assumption 7.1(c). We now verify Assumption 7.1(d). By (7.32)–(7.34) and **D$_3$** we see that Assumptions 6.8(a) and 6.8(b) hold. To verify Assumption 6.8(c), let

$$w'(i) := (i+1)(i+2) \quad \text{for each } i \in S.$$

Then by (7.32)–(7.34) we have

$$q^*(i)w(i) \le (\mu_2 + \lambda)w'(i) \quad \forall i \in S, \quad \text{and}$$
$$\sum_{j \in S} w'(j)q(j|i,a) \le 6\lambda w'(i) \quad \forall a \in [\mu_1, \mu_2] \text{ and } i \in S,$$

which imply Assumption 6.8(c) with $M' := (\mu_2 + \lambda)$, $c' := 6\lambda$, $b' := 0$. Thus, Assumption 7.1(d) is verified. Hence, Assumption 7.1 holds.

Obviously, Assumption 7.4 follows from the description of the model.

Finally, we verify Assumption 7.5. Since $0 \le p_1 \le \frac{\mu_1}{2\mu_2}$, by (7.32)–(7.34) we have, for each fixed $f \in F$,

$$\sum_{j \ge k} q\big(j|i, f(i)\big) \le \sum_{j \ge k} q\big(j|i+1, f(i+1)\big) \quad \forall i, k \in S \text{ such that } k \ne i+1,$$

which, together with Proposition C.16, implies that the corresponding Markov process $x(t)$ is stochastically ordered. Thus, Assumption 7.5 follows from (7.34), (7.39), and Proposition 7.6. □

Example 7.2 (Average-optimal control of upwardly skip-free processes) We may recall from Example 1.3 that the upwardly skip-free processes, also known as birth-and-death processes with *catastrophes*, belong to the category of *population processes* (see Anderson (1991) [4], Chap. 9, p. 292) with the state space $S := \{0, 1, 2, \ldots\}$. Here we are interested in the AR optimal control problem for such processes with catastrophes of *two* sizes, and so the transition rates are of the form

$$q(j|i,a) := \begin{cases} \lambda i + a_1 & \text{if } j = i+1, \\ -(\lambda i + \mu i + d(i,a_2) + a_1) & \text{if } j = i, \\ \mu i + d(i,a_2)\gamma_i^1 & \text{if } j = i-1, \\ d(i,a_2)\gamma_i^2 & \text{if } j = i-2, \\ 0 & \text{otherwise,} \end{cases} \tag{7.40}$$

where $i \ge 2$, $a := (a_1, a_2)$, and the birth rate $\lambda > 0$ and death rate $\mu > 0$ are fixed. Moreover, the *immigration* rates $a_1 \ge 0$, and the $d(i, a_2)$ are nonnegative numbers representing the rates at which the "catastrophes" occur and which are assumed to be controlled by decisions a_2 in some compact set $B(i)$ when the process is in state $i \ge 2$. The numbers γ_i^1 and γ_i^2 are nonnegative and such that $\gamma_i^1 + \gamma_i^2 = 1$ for all $i \ge 2$, with γ_i^k denoting the probability that the process makes a transition from i to $i - k$ $(k = 1, 2)$, given that a catastrophe occurs when the process is in state $i \ge 2$. When the state i is 0 or 1, it is natural to let $q(1|0,a) = a_1, q(0|0,a) =$

$-a_1, q(j|0,a) = 0$ for $j \geq 2$, and

$$q(j|1,a) = \begin{cases} \lambda + a_1 & \text{if } j = 2, \\ -\lambda - \mu - a_1 - d(1,a_2) & \text{if } j = 1, \\ \mu + d(1,a_2) & \text{if } j = 0, \\ 0 & \text{otherwise,} \end{cases}$$

with $d(1,a_2)$ having the similar meaning as $d(i,a_2)$ above. On the other hand, we suppose that the immigration rates a_1 can also be controlled, and so we interpret $a := (a_1, a_2)$ as an action. Thus, we consider the admissible action sets $A(0) := [0,b]$ and $A(i)$ that are finite subsets of $[0,b] \times B(i)$ for $i \geq 1$, for some constant $b > 0$. In addition, we suppose that the damage caused by a catastrophe is represented by $p > 0$ for each unit of time and that it incurs a cost at rate $c(i,a_2)$ to take decision $a_2 \in B(i)$ at state $i \geq 1$. Let $c(0,\cdot) \equiv 0$. Also, we assume that the benefits obtained by the transitions to $i-1$ and $i-2$ from i (≥ 2) are given by positive constants q_1 and q_2, respectively, and the benefit earned by each $a_1 \in [0,b]$ is denoted by $\tilde{r}(a_1)$. Hence, the reward function is of the form

$$r(i,a) := \tilde{r}(a_1) - c(i,a_2) - pd(i,a_2) + q_1\gamma_i^1 d(i,a_2) + q_2\gamma_i^2 d(i,a_2)$$

for all $a = (a_1,a_2) \in A(i)$, where $\gamma_0^1 = \gamma_0^2 := 0, \gamma_1^1 := 1, \gamma_1^2 := 0, d(0,a_2) := 0$.

As in Example 7.1, it can be verified that, under the following conditions $\mathbf{E_k}$ $(k = 1,2,3)$, the above controlled upwardly skip-free process satisfies Assumptions 7.1, 7.4, and 7.5.

$\mathbf{E_1}$. $\mu - \lambda > 0$; $\gamma_{i+1}^2 \leq \inf_{\{a_2 \in B(i)\}} \frac{d(i,a_2)+\mu i}{d(i+1,a_2)}$ for all $i \geq 1$.

$\mathbf{E_2}$. $b \leq \lambda - \mu + \inf_{\{i \geq 1, a_2 \in B(i)\}} \{d(i,a_2) + \gamma_i^2 d(i,a_2)\}$.

$\mathbf{E_3}$. For each $i \in S$, the functions $\tilde{r}(a_1)$, $d(i,a_2)$, and $c(i,a_2)$ are continuous in $(a_1,a_2) \in A(i)$, and

$$\sup_{a_2 \in B(i)} \left|d(i,a_2)\right| \leq L_1'(i+1), \qquad \sup_{a_2 \in B(i)} \left|c(i,a_2)\right| < L_2'(i+1)$$

for some constants $L_1' > 0$ and $L_2' > 0$.

In particular, these conditions $\mathbf{E_k}$ $(k = 1,2,3)$ hold when $\lambda < \mu \leq b + \lambda$, $\tilde{r}(a_1) := \tau a_1$, $d(i,a_2) := 2a_2 i$, $\gamma_i^2 \leq \frac{1}{2} + \frac{\mu}{4\beta}$, and $B(i) := [b,\beta]$ for all $i \geq 1$, for some constants $\tau > 0$ and $\beta > b$.

Therefore (by Theorem 7.8), we have the following fact.

Proposition 7.17 *Under conditions* $\mathbf{E_1}$, $\mathbf{E_2}$, *and* $\mathbf{E_3}$, *the above controlled upwardly skip-free process satisfies Assumptions* 7.1, 7.4, *and* 7.5. *Therefore (by Theorem* 7.8*), there exists an AR optimal stationary policy.*

We next provide two examples about queueing systems which also satisfy Assumptions 7.1, 7.4, and 7.5. Therefore (by Theorem 7.8), there exists an AR optimal stationary policy for each of the two examples.

Example 7.3 (Average-optimal control of a pair of $M/M/1$ queues in tandem) Consider a tandem queueing system consisting of two servers in series. Customers arrive as a Poisson stream with *unit* rate to the first queue where they are served with mean service time a_1^{-1}. After service is completed at the first queue, each customer immediately departs and joins the second queue, where the mean service time is a_2^{-1}. After service is completed at the second queue, the customers leave the system. The state space is $S := \{0, 1, 2, \ldots\}^2$. We interpret a given pair $(a_1, a_2) =: a$ of mean service times as an action taken from the action set $A(i_1, i_2) \equiv [\mu_1, \mu_1^*] \times [\mu_2, \mu_2^*]$, with positive constants $\mu_1^* > \mu_1,\ \mu_2^* > \mu_2$. Let

$$w(i_1, i_2) := \sigma_1^{i_1 - 1} + \sigma_2^{i_1 + i_2 - 1} + \gamma \sigma_1^{-\beta_1 (i_1 - 1)} \sigma_2^{-\beta_2 (i_1 + i_2 - 1)},$$

where $\sigma_1 = 1.06,\ \sigma_2 = 1.03,\ \gamma = 0.4,\ \beta_1 = 1.5,\ \beta_2 = 0.3,\ \mu_1 \geq 3$, and $\mu_2 \geq 2$. Suppose that $r(i_1, i_2, a)$ is *bounded* in all (i_1, i_2, a) and *continuous* in $a \in A(i_1, i_2)$ for each fixed $(i_1, i_2) \in S$. Then by a direct calculation and Proposition 7.6, we see that Assumption 7.1 is satisfied with $c_1 := 0.002$ and a constant $b_1 > 0$. Moreover, under these parameter values which are found by the computer program Mathematica, a straightforward calculation yields Assumptions 7.1, 7.4, and 7.5. (If necessary, see Lund, Meyn, and Tweedie (1996) [113] for details.) Therefore, there exists an AR optimal stationary policy for our tandem queueing system.

Example 7.4 (Average-optimal control of $M/M/N/0$ queue systems) Here the state space is $S := \{0, 1, \ldots, N\}$ for some integer $N \geq 1$. Suppose that the arrival rate λ is fixed but the service rates can be controlled. Therefore, we interpret the service rates a as actions which may depend on the current states $i \in S$. We denote by $A(i)$ the action set at state $i \in S$. When the system is empty, we may suppose that $A(0) := \{0\}$. For each $i \geq 1$, let $A(i) := [\mu_1, \mu_2]$ with constants $\mu_2 > \mu_1 > 0$. The transition rates are given as follows:

$$q(0|0, 0) = -\lambda = -q(1|0, 0), \qquad q(j|0, 0) = 0 \quad \text{for } 2 \leq j \leq N;$$
$$q(N|N, a) = -Na = -q(N - 1|N, a), \qquad q(j|N, a) = 0$$

for all $0 \leq j \leq N - 2$ and $a \in A(N)$; moreover, for all $1 \leq i \leq N - 1$ and $a \in A(i)$,

$$q(j|i, a) := \begin{cases} \lambda & \text{if } j = i + 1, \\ -(\lambda + ai) & \text{if } j = i, \\ ai & \text{if } j = i - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, suppose that $\mu_1 > \lambda$, and that the given reward function $r(i, a)$ is continuous in $a \in A(i)$ for all $i \in S$. Then, as in the Example 7.1, we can see that this controlled $M/M/N/0$ system satisfies Assumptions 7.1, 7.4, and 7.5.

Remark 7.18 In the verification of Assumptions 7.1, 7.4, and 7.5 for our four examples, a key step is the verification of Assumption 7.5 by means of Proposition 7.6.

This is due to the advantageous feature of Proposition 7.6 of being expressed in terms of the primitive data of the model; this allows a direct verification of conditions (a) and (b) in the proposition. We should also note that these conditions have to be *uniform* with respect to the actions. In fact, this uniformity is crucial to show that the exponential convergence rate δ and the constant L_2 in Assumption 7.5 are *independent* of all the stationary policies. To conclude, we mention that other examples and approaches that yield Assumption 7.5 can be seen in Down, Meyn, and Tweedie (1995) [37], Lund, Meyn, and Tweedie (1996) [113], and Tweedie (1981) [149], for instance.

7.6 Notes

As we already mentioned in Sect. 7.1, an AR optimal policy may not exist for a *non*-finite MDP. Thus, the main aim of the study on the average-reward optimality criterion is to find conditions ensuring the existence of AR optimal policies, and many conditions have indeed been proposed; see, for instance, Hou and Guo (1998) [84] and Kakumanu (1972) [93] for bounded transition rates and rewards; Haviv and Puterman (1998) [68], Lewis and Puterman (2001) [107], Puterman (1994) [129], Sennott (1999) [141], Serfozo (1981) [143], and Yushkevich and Feinberg (1979) [165] for bounded transition rates but unbounded rewards; Bather (1976) [9], Guo and Liu (2001), Hou and Guo (1998) [84], and Song (1987) [145] for unbounded transition rates but bounded rewards; and Guo and Zhu (2002b) [63], and Guo and Hernández-Lerma (2003c) [55] for unbounded transition rates and unbounded rewards. For the case of a Polish state space (that is, where S is a complete separable metric space), the reader is referred to Doshi (1976) [36], Guo and Rieder (2006) [59], and Hernández-Lerma (1994) [69]. We also note that when the transition rates are bounded, some results for continuous-time MDPs can be obtained from those for discrete-time MDPs by using the uniformization technique; see, e.g., Haviv and Puterman (1998) [68], Lembersky (1974) [106], Lewis and Puterman (2001) [107], Puterman (1994) [129], or Veinott (1969) [151].

This chapter concerns MDPs with *unbounded transition rates*, *unbounded reward functions*, and a *denumerable state space*, as in Guo and Liu (2001) [58], Guo and Zhu (2002b) [63], and Guo and Hernández-Lerma (2003c) [55]. The main results in this chapter are from Guo and Hernández-Lerma (2003c) [55].

On the other hand, the existing approaches used to showing the existence of AR optimal policies include *Kolmogorov's forward equation approach* by Guo and Liu (2001) [58], Guo and Zhu (2002a, 2002b) [62, 63], Kakumanu (1971) [92], and Miller (1968) [117], for instance; the *uniformization technique* by Lewis and Puterman (2001) [107], Puterman (1994) [129], and Sennott (1999) [141]; the *extended generator approach* by Guo and Cao (2005) [52] and Guo and Hernández-Lerma (2003) [53, 54, 56]; the *average-cost minimum nonnegative solution approach* (that is, the optimality inequality approach associated to the vanishing discount method) provided in Sect. 5.4; and the *convex analytic approach* in Piunovskiy (2004) [120].

Now, in view of Theorem 7.15 (in Sect. 7.4), we can add another approach, namely the policy iteration algorithm.

Each of these approaches has its own advantages. Roughly speaking, Kolmogorov's forward equation and the extended generator approaches, as well as the policy iteration algorithm, can deal with the case that the reward function may have neither upper nor lower bounds, but of course other conditions are also required. The uniformization technique (see Remark 6.1) is applicable only when the transition rates are bounded. The minimum nonnegative solution approach can show that an AR optimal policy may exist even when the AROE approach fails; however, it cannot deal with the case of unbounded from below rewards. Finally, the convex analytic approach can also be used to study multi-criteria and multi-constrained problems, but it is not very common because it mainly deals with the problem of *existence* of optimal policies; it is not obvious at all that it can be used for computational or approximation purposes, such as, for instance, the policy iteration algorithm in Sect. 7.4.

http://www.springer.com/978-3-642-02546-4

Continuous-Time Markov Decision Processes
Theory and Applications
Guo, X.; Hernández-Lerma, O.
2009, XVIII, 234 p., Hardcover
ISBN: 978-3-642-02546-4