Chapter 4

Iteration Algorithms in Markov Decision Processes with State-Action-Dependent Discount Factors and Unbounded Costs

Fernando Luque-Vásquez and
J. Adolfo Minjárez-Sosa

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65044

Abstract

This chapter concerns discrete time Markov decision processes under a discounted optimality criterion with state-action-dependent discount factors, possibly unbounded costs, and noncompact admissible action sets. Under mild conditions, we show the existence of stationary optimal policies and we introduce the value iteration and the policy iteration algorithms to approximate the value function.

Keywords: discounted optimality, non-constant discount factor, value iteration, policy iteration, Markov decision processes

AMS 2010 subject classifications: 93E10, 90C40

1. Introduction

In this chapter we study Markov decision processes (MDPs) with Borel state and action spaces under a discounted criterion with state-action–dependent discount factors, possibly unbounded costs and noncompact admissible action sets. That is, we consider discount factors of the form

$$\alpha(x_n, a_n), \tag{1}$$

where x_n and a_n are the state and the action at time n, respectively, playing the following role during the evolution of the system. At the initial state x_0, the controller chooses an action a_0 and a cost $c(x_0,a_0)$ is incurred. Then the system moves to a new state x_1 according to a transition law. Once the system is in state x_1 the controller selects an action a_1 and incurs a discounted cost $\alpha(x_0,a_0)c(x_1,a_1)$. Next the system moves to a state x_2 and the process is repeated. In general, for the stage $n \geq 1$, the controller incurs the discounted cost

$$\prod_{k=0}^{n-1}\alpha(x_k,a_k)c(x_n,a_n), \tag{2}$$

and our objective is to show the existence of stationary optimal control policies under the corresponding performance index, as well as to introduce approximation algorithms, namely, value iteration and policy iteration.

In the scenario of assuming a constant discount factor, the discounted optimality criterion in stochastic decision problems is the best understood of all performance indices, and it is widely accepted in several application problems (see, e.g., [1–3] and reference there in). However, such assumption might be strong or unrealistic in some economic and financial models. Indeed, in these problems the discount factors are typically functions of the interest rates, which in turn depend on the amount of currency and the decision-makers actions. Hence, we have state-action–dependent discount factors, and it is indeed these kinds of situations we are dealing with.

MDPs with non constant discount factors have been studied under different approaches (see, e.g., [4–8]). In particular, our work is a sequel to [8] where is studied the control problem with state-dependent discount factor. In addition, randomized discounted criteria have been analyzed in [9–12] where the discount factor is modeled as a stochastic process independent of the state-action pairs.

Specifically, in this chapter we study control models with state-action-dependent discount factors, focusing mainly on introducing approximation algorithms for the optimal value function (value iteration and policy iteration). Furthermore, an important feature in this work is that there is no compactness assumption on the sets of admissible actions neither continuity conditions on the cost, which, in most of the papers on MDPs, are needed to show the existence of measurable selectors and continuity or semi-continuity of the minima function. Indeed, in contrast to the previously cited references, in this work, we assume that the cost and discount factor functions satisfy the $\mathcal{K}$-inf-compactness condition introduced in [13]. Then, we use a generalization of Berge's Theorem, given in [13], to prove the existence of measurable selectors. To the best of our knowledge there are no works dealing with MDPs in the context presented in this chapter.

The remainder of the chapter is organized as follows. Section 2 contains the description of the Markov decision model and the optimality criterion. In Section 3 we introduce the assumptions

http://dx.doi.org/10.5772/65044

on the model and we prove the convergence of the value iteration algorithm (Theorem 3.5). In Section 4 we define the policy iteration algorithm and the convergence is stated in Theorem 4.1.

Notation. Throughout the paper we shall use the following notation. Given a *Borel space* S—that is, a Borel subset of a complete separable metric space — $\mathcal{B}(S)$ denotes the Borel σ-algebra and "measurability" always means measurability with respect to $\mathcal{B}(S)$. Given two Borel spaces S and S', a *stochastic kernel* $\varphi(\cdot \mid \cdot)$ on S given S' is a function such that $\varphi(\cdot \mid s')$ is a probability measure on S for each $s' \in S'$, and $\varphi(B \mid \cdot)$ is a measurable function on S' for each $B \in \mathcal{B}(S)$. Moreover, $\mathbb{N}$ ($\mathbb{N}_0$) denotes the positive (nonnegative) integers numbers. Finally, $L(S)$ stands for the class of lower semicontinuous functions on S bounded below and $L_+(S)$ denotes the subclass of nonnegative functions in $L(S)$.

2. Markov decision processes

Markov control model. Let

$$\mathcal{M} := \left(X, A, \{A(x) \subset A \mid x \in X\}, Q, \alpha, c\right) \tag{3}$$

be a discrete-time Markov control model with state-action-dependent discount factors satisfying the following conditions. The state space X and the action or control space A are Borel spaces. For each state $x \in X$, $A(x)$ is a nonempty Borel subset of A denoting the set of admissible controls when the system is in state x. We denote by $\mathbb{K}$ the graph of the multifunction $x \mapsto A(x)$, that is,

$$\mathbb{K} = \{(x,a) : x \in X, a \in A(x)\} \tag{4}$$

which is assumed to be a Borel subset of the Cartesian product of X and A. The transition law $Q(\cdot \mid \cdot)$ is a stochastic kernel on X given $\mathbb{K}$. Finally, $\alpha : \mathbb{K} \to (0,1)$ and $c : \mathbb{K} \to (0,\infty)$ are measurable functions representing the discount factor and the cost-per-stage, respectively, when the system is in state $x \in X$ and the action $a \in A(x)$ is selected.

The model $\mathcal{M}$ represents a controlled stochastic system and has the following interpretation. Suppose that at time $n \in \mathbb{N}_0$ the system is in the state $x_n = x \in X$. Then, possibly taking into account the history of the system, the controller selects an action $a_n = a \in A(x)$, and a discount factor $\alpha(x, a)$ is imposed. As a consequence of this the following happens:

1. A cost $c(x,a)$ is incurred;

2. The system visits a new state $x_{n+1} = x' \in X$ according to the transition law

$$Q(B \mid x,a) := \Pr\left[x_{n+1} \in B \mid x_n = x, a_n = a\right], \quad B \in \mathcal{B}(X). \tag{5}$$

Once the transition to state x' occurs, the process is repeated.

Typically, in many applications, the evolution of the system is determined by stochastic difference equations of the form

$$x_{n+1} = F(x_n, a_n, \xi_n), \, n \in \mathbb{N}_0, \tag{6}$$

where $\{\xi_n\}$ is a sequence of independent and identically distributed random variables with values in some Borel space S, independent of the initial state x_0, and $F: X \times A \times S \to X$ is a given measurable function. In this case, if θ denotes the common distribution of ξ_n, that is

$$\theta(D) := P\left[\xi_n \in D\right], \; D \in \mathcal{B}(S), \, n \in \mathbb{N}_0, \tag{7}$$

then the transition kernel can be written as

$$\begin{aligned} Q(B \mid x,a) &= \Pr\left[F(x_n, a_n, \xi_n) \in B \mid x_n = x, a_n = a\right] \\ &= \theta\left\{s \in S : F(x,a,s) \in B\right\} \\ &= \int_S 1_B\left[F(x,a,s)\right]\theta(ds), \quad B \in \mathcal{B}(X), (x,a) \in \mathbb{K}, \end{aligned} \tag{8}$$

where $1_B(\cdot)$ represents the indicator function of the set B.

Control policies. The actions applied by the controller are chosen by mean of rules known as control policies defined as follows. Let $\mathbb{H}_0 := X$ and $\mathbb{H}_n := \mathbb{K}^n \times X, n \geq 1$ be the spaces of admissible histories up to time n. A generic element of $\mathbb{H}_n$ is written as $h_n = (x_0, a_0, \ldots, x_{n-1}, a_{n-1}, x_n)$.

Definition 2.1 *A control policy (randomized, history-dependent) is a sequence $\pi = \{\pi_n\}$ of stochastic kernels π_n on A given $\mathbb{H}_n$ such that $\pi_n(A(x_n) \mid h_n) = 1$, for all $h_n \in \mathbb{H}_n, n \in \mathbb{N}_0$.*

We denote by Π the set of all control policies.

Let $\mathbb{F}$ be the set of measurable selectors, that is, $\mathbb{F}$ is the set of measurable function $f: X \to A$ such that $f(x) \in A(x)$ for all $x \in X$.

Definition 2.2 *A control policy $\pi = \{\pi_n\}$ is said to be:*

a. *deterministic if there exists a sequence $\{g_n\}$ of measurable functions $g_n: \mathbb{H}_n \to A$ such that*

$$\pi_n(C \mid h_n) = 1_C\left[g_n(h_n)\right], \ \forall h_n \in \mathbb{H}_n, n \in \mathbb{N}_0, C \in \mathcal{B}(A); \tag{9}$$

b. *a Markov control policy if there exists a sequence* $\{f_n\}$ *of functions* $f_n \in \mathbb{F}$ *such that*

$$\pi_n(C \mid h_n) = 1_C\left[f_n(x_n)\right], \ \forall h_n \in \mathbb{H}_n, n \in \mathbb{N}_0, C \in \mathcal{B}(A). \tag{10}$$

In addition

c. *A Markov control policy is stationary if there exists* $f \in \mathbb{F}$ *such that* $f_n = f$ *for all* $n \in \mathbb{N}_0$.

If necessary, see for example [1–3, 14–16] for further information on those policies.

Observe that a Markov policy π is identified with the sequence $\{f_n\}$, and we denote $\pi = \{f_n\}$. In this case, the control applied at time n is $a_n = f_n(x_n) \in A(x_n)$. In particular, a stationary policy is identified with the function $f \in \mathbb{F}$, and following a standard convention we denote by $\mathbb{F}$ the set of all stationary control policies.

To ease the notation, for each $f \in \mathbb{F}$ and $x \in X$, we write

$$\begin{aligned} c(x, f) \quad &: \quad = c(x, f(x)), \\ Q(\cdot \mid x, f) \quad &: \quad = Q(\cdot \mid x, f(x)), \end{aligned} \tag{11}$$

and

$$\alpha(x, f) := \alpha(x, f(x)). \tag{12}$$

The underlying probability space. Let $(\Omega, \mathcal{F})$ be the canonical measurable space consisting of the sample space $\Omega = \mathbb{K}^{\infty} := \mathbb{K} \times \mathbb{K} \times \cdots$ and its product σ −algebra $\mathcal{F}$. Then, under standard arguments (see, e.g., [1, 14]) for each $\pi \in \Pi$ and initial state $x \in X$, there exists a probability measure P_x^{π} on $(\Omega,\mathcal{F})$ such that, for all $h_n \in \mathbb{H}_n$, $a_n \in A(x_n)$, $n \in \mathbb{N}_0$, $C \in \mathcal{B}(A)$, and $B \in \mathcal{B}(X)$,

$$\begin{aligned} P_x^{\pi}\left[x_0 = x\right] &= 1; \\ P_x^{\pi}\left[a_n \in C \mid h_n\right] &= \pi_n(C \mid h_n); \end{aligned} \tag{13}$$

and the Markov-like property is satisfied

$$P_x^{\pi}\left[x_{n+1} \in B \mid h_n, a_n\right] = Q(B \mid x_n, a_n). \tag{14}$$

The stochastic process $\left(\Omega, \mathcal{F}, P_x^{\pi}, \{x_n\}\right)$ is called Markov decision process.

Optimality criterion. We assume that the costs are discounted in a multiplicative discounted rate. That is, a cost C incurred at stage n is equivalent to a cost $C\Gamma_n$ at time 0, where

$$\Gamma_n := \prod_{k=0}^{n-1} \alpha(x_k, a_k) \text{ if } n \geq 1, \text{ and } \Gamma_0 = 1. \tag{15}$$

In this sense, when using a policy $\pi \in \Pi$, given the initial state $x_0 = x$, we define the total expected discounted cost (with state-action–dependent discount factors) as

$$V(\pi, x) := E_x^{\pi}\left[\sum_{n=0}^{\infty} \Gamma_n c(x_n, a_n)\right], \tag{16}$$

where E_x^{π} denotes the expectation operator with respect to the probability measure P_x^{π} induced by the policy π, given $x_0 = x$.

The optimal control problem associated to the control model $\mathcal{M}$, is then to find an optimal policy $\pi^* \in \Pi$ such that $V(\pi^*, x) = V(x)$ for all $x \in X$, where

$$V(x) := \inf_{\pi \in \Pi} V(\pi, x) \tag{17}$$

is the optimal value function (see [10]).

3. The value iteration algorithm

In this section we give conditions on the model that imply: (i) the convergence of the value iteration algorithm; (ii) the value function V is a solution of the corresponding optimality equation; and (iii) the existence of stationary optimal policies. In order to guarantee that $V(x)$ is finite for each initial state x we suppose the following.

Assumption 3.1. *There exists $\pi_0 \in \Pi$ such that for all $x \in X, V(\pi_0, x) < \infty$.*

At the end of Section 4 we give sufficient conditions for Assumption 3.1. We also require continuity and (inf-) compactness conditions to ensure the existence of "measurable minimizers." The following definition was introduced in [13].

Definition 3.2. *A function $u: \mathbb{K} \to \mathbb{R}$ is said to be $\mathcal{K}$-inf-compact on $\mathbb{K}$ if for each compact subset K of X and $r \in \mathbb{R}$, the set*

http://dx.doi.org/10.5772/65044

$$\{(x,a) \in Gr_K(A) : u(x,a) \leq r\} \tag{18}$$

is a compact subset of $X \times A$, *where* $Gr_K(A) := \{(x,a) : x \in K, a \in A(x)\}$.

Assumption 3.3. *(a) The one-stage cost* c *and the discount factor* α *are* $\mathcal{K}$*-inf-compact functions on* $\mathbb{K}$. *In addition,* c *is nonnegative*.

(b) The transition law Q is weakly continuous, that is, the mapping

$$(x,a) \to \int_X u(y)Q(dy \mid x,a) \tag{19}$$

is continuous for each bounded and continuous function on X.

For each measurable function u on X,$x \in X$, and $f \in \mathbb{F}$, we define the operators

$$Tu(x) := \inf_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \int_X u(y)Q(dy \mid x,a) \right\} \tag{20}$$

and

$$T_f u(x) := c(x,f) + \alpha(x,f) \int_X u(y)Q(dy \mid x,f). \tag{21}$$

A consequence of Assumption 3.3 is the following.

Lemma 3.4. *Let* u *be a function in* $L_+(X)$. *If Assumption 3.3 holds then the function* $v : \mathbb{K} \to \mathbb{R}$ *defined by*

$$v(x,a) := c(x,a) + \alpha(x,a) \int_X u(y)Q(dy \mid x,a) \tag{22}$$

is $\mathcal{K}$*-inf-compact on* $\mathbb{K}$

Proof. First note that by the $\mathcal{K}$-inf-compactness hypothesis $c(\cdot,\cdot)$ and $\alpha(\cdot,\cdot)$ are l.s.c on $Gr_K(A)$ for each compact subset K of X. Then, since α and u are nonnegative functions, from Assumption 3.3 we have that $v(\cdot,\cdot)$ is l.s.c on $Gr_K(A)$. Thus, for each $r \in \mathbb{R}$, the set

$$\{(x,a) \in Gr_K(A) : v(x,a) \leq r\} \tag{23}$$

is a closed subset of the compact set $\{(x,a) \in Gr_K(A): c(x,a) \leq r\}$. Then, v is $\mathcal{K}$-inf-compact on $\mathbb{K}$.

Observe that the operator T is monotone in the sense that if $v \geq u$ then $Tv \geq Tu$. In addition, from Assumption 3.3 and ([13], Theorem 3.3), we have that T maps $L_+(X)$ into itself. Furthermore, there exists $\tilde{f} \in \mathbb{F}$ such that

$$Tu(x) = T_{\tilde{f}}u(x), \quad x \in X. \tag{24}$$

To state our first result we define the sequence $\{v_n\} \subset L_+(X)$ of value iteration functions as:

$$\begin{aligned} v_0 &\equiv 0; \\ v_n(x) &= Tv_{n-1}(x), \quad x \in X. \end{aligned} \tag{25}$$

Since T is monotone, note that $\{v_n\}$ is a nondecreasing sequence.

Theorem 3.5. *Suppose that Assumptions 3.1 and 3.3 hold. Then*

a. $v_n \nearrow V$.

b. *V is the minimal solution in $L_+(X)$ of the Optimality Equation, i.e.,*

$$V(x) = TV(x) = \inf_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \int_X V(y) Q(dy \mid x,a) \right\}. \tag{26}$$

c. *There exists a stationary policy $f^* \in \mathbb{F}$ such that, for all $x \in X, V(x) = T_{f^*}V(x)$, that is*

$$V(x) = c(x, f^*) + \alpha(x, f^*) \int_X u(y) Q(dy \mid x, f^*), \tag{27}$$

and f^* is an optimal policy.

Proof. Since $\{v_n\}$ is nondecreasing, there exists $v \in L_+(X)$ such that $v_n \nearrow v$. Hence, from Monotone Convergence Theorem, ([13], Lemmas 2.2, 2.3), and ([1], Lemma 4.2.4), we obtain, for each $x \in X$, $v_n(x) = Tv_{n-1}(x) \to Tv(x)$, as $n \to \infty$, which, in turn implies

$$Tv = v. \tag{28}$$

http://dx.doi.org/10.5772/65044

Therefore, to get (a)-(b) we need to prove that $v = V$. To this end, observe that for all $x \in X$ and $\pi \in \Pi$

$$v_n(x) \le \int_A c(x,a)\pi(da \mid x) + \int_A \alpha(x,a)\int_X v_{n-1}(x_1)Q(dx_1 \mid x,a)\pi(da \mid x). \tag{29}$$

Then, iterating (29) we obtain

$$v_n(x) \le V_n(\pi,x), \ \ n \in \mathbb{N}, \tag{30}$$

where

$$V_n(\pi,x) = E_x^{\pi}\left[\sum_{t=0}^{n-1}\Gamma_t c(x_t,a_t)\right], \tag{31}$$

is the n −stage discounted cost V_n. Then, letting $n \to \infty$ we get $v(x) \le V(\pi,x)$, for all $\pi \in \Pi$ and $x \in X$. Thus,

$$v(x) \le V(x), \ \ x \in X. \tag{32}$$

On the other hand, from (28) and (24), let $f \in \mathbb{F}$ such that $v(x) = T_f v(x)$, $x \in X$. Iterating this equation, we have (see (31))

$$\begin{aligned} v(x) &= E_x^f\left[c(x,f) + \sum_{t=1}^{n-1}\prod_{k=0}^{t-1}\alpha(x_k,f)c(x_t,f)\right] \\ &\quad + E_x^f\left[\prod_{k=0}^{n-1}\alpha(x_k,f)v(x_n)\right] \\ &\ge V_n(f,x). \end{aligned} \tag{33}$$

Hence, letting $n \to \infty$,

$$v(x) \ge V(f,x) \ge V(x), \ \ x \in X. \tag{34}$$

Combining (32) and (34) we get $v = V$.

Now, let $u \in L_+(X)$ be an arbitrary solution of the optimality equation, that is, $Tu = u$. Then, applying the arguments in the proof of (34) with u instead of v we conclude that $u \ge V$. That is, V is minimal in $L_+(X)$.

Part (c) follows from (b) and ([13], Theorem 3.3). Indeed, there exists a stationary policy $f^* \in \mathbb{F}$ such that $V(x) = T_{f^*}V(x), x \in X$. Then, iteration of this equation yields $V(x) = V(f^*,x)$, which implies that f^* is optimal.

4. Policy iteration algorithm

In Theorem 3.5 is established an approximation algorithm for the value function V by means of the sequence of the value iteration functions $\{v_n\}$. In this case the sequence $\{v_n\}$ increase to V and it is defined recursively. Now we present the well-known policy iteration algorithm which provides a decreasing approximation to V in the set of the control policies.

To define the algorithm, first observe that from the Markov property (14) and applying properties of conditional expectation, for any stationary policy $f \in \mathbb{F}$ and $x \in X$, the corresponding cost $V(f,x)$ satisfies

$$\begin{aligned} V(f,x) &= c(x,f) + \alpha(x,f)E_x^f\left[\sum_{t=1}^{\infty}\prod_{k=0}^{t-1}\alpha(x_k,f)c(x_t,f)\right] \\ &= c(x,f) + \alpha(x,f)\int_X E^f\left[c(x_1,f) + \sum_{t=2}^{\infty}\prod_{k=0}^{t-1}\alpha(x_k,f)c(x_t,f) \mid x_1 = y\right]Q(dy \mid x,f) \\ &= c(x,f) + \alpha(x,f)\int_X V(f,y)Q(dy \mid x,f) = T_f V(f,x), \quad x \in X. \end{aligned} \tag{35}$$

Let $f_0 \in \mathbb{F}$ be a stationary policy with a finite valued cost $w_0(\cdot) := V(f_0,\cdot) \in L_+(X)$. Then, from (35),

$$\begin{aligned} w_0(x) &= c(x,f_0) + \alpha(x,f_0)\int_X w_0(y)Q(dy \mid x,f_0) \\ &= T_{f_0}w_0(x), \quad x \in X. \end{aligned} \tag{36}$$

Now, let $f_1 \in \mathbb{F}$ be such that

$$Tw_0(x) = T_{f_1}w_0(x), \tag{37}$$

and define $w_1(\cdot) = V(f_1,\cdot)$.

In general, we define a sequence $\{w_n\}$ in $L_+(X)$ as follows. Given $f_n \in \mathbb{F}$, compute $w_n(\cdot) := V(f_n,\cdot) \in L_+(X)$ Next, let $f_{n+1} \in \mathbb{F}$ be such that

http://dx.doi.org/10.5772/65044

$$T_{f_{n+1}} w_n(x) = T w_n(x), \quad x \in X, \tag{38}$$

that is,

$$\begin{aligned} T_{f_{n+1}} w_n(x) &= c(x, f_{n+1}) + \alpha(x, f_{n+1}) \int_X w_n(y) Q(dy|x, f_{n+1}) \\ &= \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \int_X w_n(y) Q(dy|x,a) \right\} \\ &= T w_n(x), \quad x \in X. \end{aligned} \tag{39}$$

Then we define $w_{n+1}(\cdot) = V(f_{n+1}, \cdot)$

Theorem 4.1. *Under Assumptions 3.1 and 3.3, there exists a measurable nonnegative function* $w \geq V$ *such that* $w_n \searrow w$, *and* $Tw = w$. *Moreover, if w satisfies*

$$\lim_{n \to \infty} E_x^{\pi} [\Gamma_n w(x_n)] = 0 \quad \forall \pi \in \Pi, \ x \in X, \tag{40}$$

then $w = V$.

To prove the Theorem 4.1 we need the following result.

Lemma 4.2. *Under Assumption 3.3, if* $u: X \to \mathbb{R}$ *is a measurable function such that* Tu *is well defined,* $u \leq Tu$, *and*

$$\lim_{n \to \infty} E_x^{\pi} \left[\Gamma_n u(x_n) \right] = 0 \quad \forall \pi \in \Pi, x \in X, \tag{41}$$

then $u \leq V$.

Proof. From the Markov property (14), for each $\pi \in \Pi$ and $x \in X$,

$$E_x^{\pi} \left[\Gamma_{n+1} u(x_{n+1}) \mid h_n, a_n \right] = \Gamma_{n+1} \int_X u(y) Q(dy \mid x_n, a_n) \tag{42}$$

$$= \Gamma_n \left[c(x_n, a_n) + \alpha(x_n, a_n) \int_X u(y) Q(dy \mid x_n, a_n) - c(x_n, a_n) \right] \tag{43}$$

$$\geq \Gamma_n \inf_{a \in A(x_n)} \left[c(x_n, a) + \alpha(x_n, a) \int_X u(y) Q(dy \mid x_n, a) \right] - \Gamma_n c(x_n, a_n) \tag{44}$$

$$= \Gamma_n T u(x_n) - \Gamma_n c(x_n, a_n) \geq \Gamma_n u(x_n) - \Gamma_n c(x_n, a_n), \tag{45}$$

which, in turn implies

$$\Gamma_n c(x_n, a_n) \geq E_x^{\pi}\left[\Gamma_n u(x_n) - \Gamma_{n+1} u(x_{n+1}) \mid h_n, a_n\right]. \tag{46}$$

Therefore, for all $k \in \mathbb{N}$ (see (31)),

$$V_k(\pi, x) = E_x^{\pi} \sum_{n=0}^{k-1} \Gamma_n c(x_n, a_n) \geq u(x) - E_x^{\pi}\left[\Gamma_k u(x_k)\right]. \tag{47}$$

Finally, letting $k \to \infty$, (41) yields $V(\pi,x) \geq u(x)$, and since π is arbitrary we obtain $V(x) \geq u(x)$.

Proof of Theorem 4.1. According to Lemma 4.2, it is sufficient to show the existence of a function $w \geq V$ such that $w_n \searrow w$ and $Tw = w$. To this end, from (36)–(38),

$$\begin{aligned} w_0(x) &\geq \min_{a \in A(x)} \left\{ c(x,a) + \alpha(x,a) \int_X w_0(y) Q(dy|x,a) \right\} = T_{f_1} w_0(x) \\ &= c(x, f_1) + \alpha(x, f_1) \int_X w_0(y) Q(dy|x, f_1). \end{aligned} \tag{48}$$

Iterating this inequality, a straightforward calculation as in (34) shows that

$$w_0(x) \geq V(f_1, x) = w_1(x), \quad x \in X. \tag{49}$$

In general, similar arguments yield

$$w_n \geq T w_n \geq w_{n+1}, \quad n \in \mathbb{N}. \tag{50}$$

Therefore, there exists a nonnegative measurable function w such that $w_n \searrow w$. In addition, since $w_n \geq V \ \ \forall n \in \mathbb{N}_0, \ \forall n \in \mathbb{N}_0$, we have $w \geq V$. Next, letting $n \to \infty$ in (47) and applying ([17], Lemma 3.3), we obtain $w \geq Tw \geq w$, which implies $w = Tw$.

4.1. Sufficient conditions for Assumption 3.1 and (40)

An obvious sufficient condition for Assumption 3.1 and (40) is the following:

C1 (a) There exists $\bar{\alpha} \in (0,1)$ such that for all $(x,a) \in \mathbb{K}$, $\alpha(x,a) < \bar{\alpha}$.

(b) For some constant m, $0 \leq c(x,a) \leq m$ for all $(x,a) \in \mathbb{K}$.

Indeed, under condition C1, $V(\pi,x) \leq m/(1-\bar{\alpha})$ for all $x \in X$ and $\pi \in \Pi$, and $\{w_n\}$ is a bounded sequence which in turn implies (since $w_n \searrow w$ the boundedness of the function w. This fact clearly yields (40).

http://dx.doi.org/10.5772/65044

Other less obvious sufficient conditions are the following (see, e.g., [15, 16, 2]).

C2 (a) Condition C1 (a).

(b) There exist a measurable function $W: X \to (1,\infty)$ and constants $M > 0$, $\beta \in (1, 1/\bar{\alpha})$, such that for all $(x,a) \in \mathbb{K}$,

$$\sup_{A(x)} c(x,a) \le MW(x) \tag{51}$$

and

$$\int_X W(y) Q(dy \mid x,a) \le \beta W(x). \tag{52}$$

First note that by condition C2 and the Markov property (14), for any policy $\pi \in \Pi$ and initial state $x_0 = x \in X$,

$$E_x^\pi \left[W(x_{n+1}) \mid h_n, a_n \right] = \int_X W(y) Q(dy \mid x_n, a_n) \le \beta W(x_n), \ \ \forall n \in \mathbb{N}_0. \tag{53}$$

Then, using properties of conditional expectation,

$$E_x^\pi \left[W(x_{n+1}) \right] \le \beta E_x^\pi \left[W(x_n) \right], \ \ \forall n \in \mathbb{N}_0. \tag{54}$$

Iterating inequality (51) we get

$$E_x^\pi \left[W(x_n) \right] \le \beta^n W(x), \ \ \forall n \in \mathbb{N}_0. \tag{55}$$

Therefore, by condition C2, for any policy $\pi \in \Pi$ and $x \in X$,

$$\begin{aligned} V(\pi, x) &\le E_x^\pi \sum_{n=0}^{\infty} \bar{\alpha}^n c(x_n, a_n) \le \sum_{n=0}^{\infty} M \bar{\alpha}^n E_x^\pi W(x_n) \\ &\le \frac{M}{1 - \bar{\alpha}\beta} W(x). \end{aligned} \tag{56}$$

Thus, Assumption 3.1 holds.

On the other hand, if $L_+^W(X)$ denotes the subclass of all functions u in $L_+(X)$ such that

$$\|u\|_W := \sup_{x \in X} \frac{h(x)}{W(x)} < \infty, \tag{57}$$

then, because $w_k(\cdot) = V(f_{k+1}, \cdot)$, from (53) and condition C2, we have that $w_k \in L_+^W(X)$ for all $k = 1,2,...$ and

$$\lim_{n \to \infty} E_x^\pi \left[\Gamma_n w_k(x_n) \right] = 0 \quad \forall \pi \in \Pi, x \in X. \tag{58}$$

Since $w \leq w_k$, (40) follows from (55).

Acknowledgements

Work supported partially by Consejo Nacional de Ciencia y Tecnologa (CONACYT) under grant CB2015/254306.

Author details

Fernando Luque-Vásquez and J. Adolfo Minjárez-Sosa*

*Address all correspondence to: aminjare@gauss.mat.uson.mx

Department of Mathematics, University of Sonora, Hermosillo, Sonora, México

References

[1] O. Hernández-Lerma, J.B. Lasserre, Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer-Verlag, New York, NY, 1996.

[2] O. Hernández-Lerma, J.B. Lasserre, Further Topics on Discrete-Time Markov Control Processes. Springer-Verlag, New York, NY, 1999.

[3] M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York, NY, 1994.

[4] Y. Carmon, A. Shwartz, Markov decision processes with exponentially representable discounting. Oper. Res. Lett. 37 (2009), 51–55.

http://dx.doi.org/10.5772/65044

[5] E.A. Feinberg, A. Shwartz, Constrained dynamic programming with two discount factors: applications and an algorithm. IEEE Trans. Autom. Control, 44 (1999), 628–631.

[6] K. Hinderer, Foundations of non-stationary dynamical programming with discrete time parameter, Lecture Notes Oper. Res. 33 Springer, New York, NY, 1970.

[7] M. Schäl, Conditions for optimality in dynamic programming and for the limit of n-stages optimal policies to be optimal, Z. Wahr. Verw. Geb. 32 (1975), 179–196.

[8] Q. Wei, X. Guo, Markov decision processes with state-dependent discount factors and unbounded rewards/costs. Oper. Res. Lett. 39 (2011), 369–374.

[9] J. González-Hernández, R.R. López-Martnez, R. Pérez-Hernández, Markov control processes with randomized discounted cost in Borel space. Math. Meth. Oper. Res., 65 (2007), 27–44.

[10] J. González-Hernández, R.R. López-Martnez, J.A. Minjárez-Sosa, Adaptive policies for stochastic systems under a randomized discounted criterion. Bol. Soc. Mat. Mex., 14 (2008), 149–163.

[11] J. González-Hernández, R.R. López-Martnez, J.A. Minjárez-Sosa, Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion. Kybernetika, 45 (2009), 737–754.

[12] J. González-Hernández, R.R. López-Martnez, J.A. Minjárez-Sosa, J.R. Gabriel-Arguelles, Constrained Markov control processes with randomized discounted cost criteria: occupation measures and extremal points. Risk Decis. Anal., 4 (2013), 163–176.

[13] E.A. Feinberg, P.O. Kasyanov, N.V. Zadoianchuck, Berge's theorem for non compact image sets, J. Math. Anal. Appl. 397 (2013), 255–259.

[14] E.B. Dynkin, A.A. Yushkevich, Controlled Markov Processes. Springer-Verlag, New York, NY, 1979.

[15] E.I. Gordienko, O. Hernández-Lerma, Average cost Markov control processes with weighted norms: existence of canonical policies. Appl. Math. (Warsaw), 23 (1995), 199–218.

[16] E.I. Gordienko, O. Hernández-Lerma, Average cost Markov control processes with weighted norms: value iterations. Appl. Math. (Warsaw), 23 (1995), 219–237.

[17] O. Hernández-Lerma, W. Runggaldier, Monotone approximations for convex stochastic control problems. J. Math. Syst., Est. Control, 4 (1994), 99–140.

IntechOpen