TOP (2015) 23:743–772
DOI 10.1007/s11750-015-0360-5

ORIGINAL PAPER

# Markov control models with unknown random state–action-dependent discount factors

**J. Adolfo Minjárez-Sosa**

Received: 1 April 2014 / Accepted: 4 February 2015 / Published online: 13 February 2015


**Abstract** The paper deals with a class of discounted discrete-time Markov control models with non-constant discount factors of the form $\tilde{\alpha}(x_n, a_n, \xi_{n+1})$, where $x_n$, $a_n$, and $\xi_{n+1}$ are the state, the action, and a random disturbance at time $n$, respectively, taking values in Borel spaces. Assuming that the one-stage cost is possibly unbounded and that the distributions of $\xi_n$ are unknown, we study the corresponding optimal control problem under two settings. Firstly we assume that the random disturbance process $\{\xi_n\}$ is formed by observable independent and identically distributed random variables, and then we introduce an estimation and control procedure to construct strategies. Instead, in the second one, $\{\xi_n\}$ is assumed to be non-observable whose distributions may change from stage to stage, and in this case the problem is studied as a minimax control problem in which the controller has an opponent selecting the distribution of the corresponding random disturbance at each stage.

**Keywords** Discounted optimality · Non-constant discount factors · Estimation and control procedures · Minimax control systems

**Mathematics Subject Classification** 90C40 · 90C47 · 93E10 · 93E20

Work supported by Consejo Nacional de Ciencia y Tecnología (CONACYT) under Grant CB2010/154612.

J. A. Minjárez-Sosa (✉)
Departamento de Matemáticas, Universidad de Sonora, Rosales s/n,
Col. Centro, 83000 Hermosillo, Sonora, Mexico
e-mail: aminjare@gauss.mat.uson.mx; minj091200@gmail.com

## 1 Introduction

We consider discounted discrete-time Markov control models with non-constant discount factors of the form

$$\tilde{\alpha}(x_n, a_n, \xi_{n+1}), \tag{1}$$

where $x_n$ and $a_n$ are the state and the action at time $n$, respectively, and $\{\xi_n\}$ is a sequence of random variables representing a random disturbance at each time $n$. The discount factors (1) play the following role during the evolution of the system. At the initial state $x_0$, the controller chooses an action $a_0$. Then a cost $c(x_0, a_0)$ is incurred, and the system moves to a new state $x_1$ according to a transition law, and the random disturbance $\xi_1$ comes in. Once the system is in state $x_1$ the controller selects an action $a_1$ and incurs a discounted cost $\tilde{\alpha}(x_0, a_0, \xi_1)c(x_1, a_1)$. Next the system moves to a state $x_2$ and the process is repeated. That is, on the record of the states–actions and random disturbances, the controller incurs, for the stage $n \geq 1$, the discounted cost

$$\tilde{\alpha}(x_0, a_0, \xi_1)\tilde{\alpha}(x_1, a_1, \xi_2) \cdots \tilde{\alpha}(x_{n-1}, a_{n-1}, \xi_n)c(x_n, a_n). \tag{2}$$

Thus, the costs are discounted in a multiplicative discount rate, and therefore, assuming $c(\cdot, \cdot)$ possibly unbounded, our objective is to study the optimality under the performance index defined by the accumulation throughout the evolution of the system of these costs.

The main motivation in considering non-constant discount factors comes from the point of view of the applications. Indeed, in economic and financial models, the discount factors are typically functions of the interest rates, which in turn are uncertain. Such uncertainty may be due to the amount of currency, and/or the decision-makers' actions, and furthermore, to external random noises whose distributions really are unknown. Therefore, in these cases, we have random state–action-dependent discount factors which, in this paper, are supposed to be determined by a function as in (1). Then, assuming that the distributions of the random variables $\xi_n$ are unknown, we study the corresponding control problem under two settings.

Firstly, we assume that the random disturbance process $\{\xi_n\}$ is formed by observable, independent and identically distributed (i.i.d.), and independent of the state–action pairs, random variables. The common (unknown) distribution, denoted by $\theta$, is estimated from historical observations of $\xi_n$ using the empirical distribution. Then, we combine this estimation scheme with a suitable minimization process to construct asymptotically optimal strategies. It is worth remarking that the hypotheses of observability as well as the non-dependence on the state–action pairs of the random disturbance process are crucial to implement such statistical estimation and control procedure, which is also known as *certainty equivalence principle* (see, e.g., Hernández-Lerma 1989; Mandl 1974). However, there are situations, perhaps most of them in economic and financial models, where (i) the random variable $\xi_n$ really represents a random noise which is impossible to observe; or (ii) $\{\xi_n\}$ is a stochastic process which is difficult to handle inside a controlled system. The last situation might occur, for instance, when $\{\xi_n\}$ represents the interest rate. Indeed, in dynamical financial systems (see, e.g., Brigo and Mercurio 2007; Heath et al. 1992; Vasicek 1977) generally

the evolution of the interest rate is modeled by a stochastic difference equation (differential equation if the system is analyzed in continuous time), then when inserting such equation in a control system through the discount factor, as it would be our case, the study of the resulting optimal control problem becomes difficult. Considering the situations (i) and (ii), our second setting consists in supposing that $\{\xi_n\}$ is a sequence of independent and possibly non-observable random variables whose distributions may change from stage to stage. The only information possessing the controller is that at each stage, the corresponding distribution belongs to an appropriate set of probability measures $\Theta$. In this case, the optimal control problem is studied as a minimax control problem known as *game against nature*. Indeed, we suppose that the controller has an opponent, namely, the "nature", which, at each stage, is selecting a distribution from the set $\Theta$ for the corresponding random disturbance. Hence, the controller is interested in selecting actions directed to *minimize* the *maximum* discounted cost—with random state–action-dependent discount factors—generated on the set $\Theta$. Thus, the second objective is to show the existence of minimax strategies.

Observe that under the minimax approach it is possible to study the optimal control problem in general scenarios. These include the cases when the random disturbances are observable or unobservable, with constant or non-constant distribution throughout the evolution of the system. Moreover, modeling a control problem as a minimax system simplifies the mathematical analysis since it avoids dealing directly with the disturbance process, as is the case of point (ii). Although these facts constitute certain advantages, it is important to keep in mind that, under this formulation, we are obtaining minimax strategies instead of optimal strategies, which is the price we must pay. In particular, if $\xi_n$ are i.i.d. and observable random variables, as the conditions in the estimation and control scheme, the minimax procedure works if we assume that the common and unknown distribution $\theta$ belongs to a set $\Theta$. Clearly, in this case, we obtain minimax strategies instead of asymptotically optimal strategies.

Our general approach to analyze these two problems is based on the following. We introduce a minimax value iteration algorithm which converges geometrically to the minimax value function. Such value function is characterized as the unique solution of a minimax equation, and then, by imposing appropriate conditions, we prove the existence of minimax strategies. In addition, taking into account that an optimal control problem is a particular case of a minimax problem, we apply the minimax results to study the estimation and control problem, for which we first prove that the Markov strategies are sufficient.

Among the performance indices to study a stochastic optimal control problem, the discounted criterion with *constant and non-random* discount factor is the best understood. It has been widely studied under different approaches: dynamic programming (see Bertsekas 1987; Hernández-Lerma 1989; Hernández-Lerma and Lasserre 1996, 1999; Puterman 1994 and references therein); convex analysis (Altman 1999; Borkar 1998; Piunovskiy 1997); linear programming (Altman 1999; Hernández-Lerma and Lasserre 1996, 1999; Hernández-Lerma and González-Hernández 2000; Piunovskiy 1997); Lagrange multipliers (López-Martínez and Hernández-Lerma 2003); adaptive procedures (Gordienko and Minjárez-Sosa 1998; Hilgert and Minjárez-Sosa 2001, 2006); minimax systems (González-Trejo et al. 2003; Iyengar 2005; Jaskiewicz and Nowak 2011). However, although infrequently, there have been important works deal-

ing with the problem with non-constant discount factors under several settings. For instance, in Feinberg and Shwartz (1994) is studied the problem assuming $K < \infty$ different discount factors $\alpha_1, \alpha_2, \ldots, \alpha_K$ which are independent of the state–action pairs (see also Carmon and Shwartz 2009; Feinberg and Shwartz 1995, 1999). In fact, in Carmon and Shwartz (2009) is presented an extension to the case $K = \infty$. In addition, performance indices with multiplicative discount rates as (2) have been treated in Hinderer (1979) and Schäl (1975). Specifically, in Hinderer (1979), the discount factor is defined as a function of the state–action history of the system, while in Schäl (1975) it is state–action-dependent. In both papers is assumed bounded one-stage costs from below. Recently, in Wei and Guo (2011), some results in Schäl (1975) were extended to the unbounded cost case considering state-dependent discount factors. On the other hand, randomized discounted criteria have been analyzed in González-Hernández et al. (2007, 2008, 2009, 2013, 2014) addressing several issues: existence of optimal strategies, adaptive control, approximation algorithms, and problems with constraints. In these cases, the discount factor is modeled as a stochastic process, independent of the state–action pairs, which is defined in terms of a suitable discrete-time Markov process.

According to the description of the literature, our work presents an alternative form to study discounted optimal problems with non-constant discount factors. That is, to the best of our knowledge, discounted criteria with random state–action-dependent discount factors, and moreover with unknown disturbance distribution, have not been studied.

The organization of the paper is as follows. In Sect. 2, we present the control models we are concerned with. Next, Sect. 3 contains the optimality criteria, and then the minimax and the estimation and control problems are introduced. General assumptions as well as some preliminary results are stated in Sect. 4. The minimax and the asymptotically optimal strategies are constructed in Sects. 5 and 6, respectively. Finally, in Sect. 7 are given some examples to illustrate our results.

## 2 The control models

In this section, we present the control models that will be analyzed in the paper. We first introduce the Markov model corresponding to the estimation and control problem, and next we describe the minimax control model to study the case of non-observable random disturbance. We will use the following notation.

**Notation** *Given a Borel space $Z$—that is, a Borel subset of a complete separable metric space—$\mathcal{B}(Z)$ denotes the Borel $\sigma$-algebra and " measurability" always means measurability with respect to $\mathcal{B}(Z)$. The class of all probability measures on $Z$ is denoted by $\mathbb{P}(Z)$. Given two Borel spaces $Z$ and $Z'$, a stochastic kernel $\varphi(\cdot|\cdot)$ on $Z$ given $Z'$ is a function such that $\varphi(\cdot|z')$ is in $\mathbb{P}(Z)$ for each $z' \in Z'$, and $\varphi(B|\cdot)$ is a measurable function on $Z'$ for each $B \in \mathcal{B}(Z)$. Moreover, $\mathbb{R}_+$ stands for the nonnegative real numbers' subset and $\mathbb{N}(\mathbb{N}_0$, resp.) denotes the positive (nonnegative, resp.) integers' subset. The class $\mathbb{P}(Z)$ is endowed with the topology of weak convergence. That is, a sequence $\{\mu_n\}$ in $\mathbb{P}(Z)$ converges weakly to $\mu$ ($\mu_n \to \mu$) if*

$$\int_Z u \mathrm{d}\mu_n \to \int_Z u \mathrm{d}\mu$$

*for all bounded and continuous function $u$. In this case, we have that if $Z$ is a Borel space, then so is $\mathbb{P}(Z)$.*

### 2.1 Markov control model

We consider the control model with random state–action-dependent discount factors

$$\mathcal{M} = (\mathbb{X}, \mathbb{A}, S, Q, \tilde{\alpha}, c) \tag{3}$$

satisfying the following conditions. The *state space* $\mathbb{X}$, the *action space* $\mathbb{A}$, and the discount factor disturbance space $S$ are Borel spaces. To each $x \in \mathbb{X}$, we associate a nonempty measurable subset $A(x)$ of $\mathbb{A}$ denoting the set of *admissible controls* (*or actions*) when the system is in state $x$. The set

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

of *admissible state–action pairs* is assumed to be a Borel subset of the Cartesian product of $\mathbb{X}$ and $\mathbb{A}$. The *transition law* $Q(\cdot \mid \cdot)$ is a stochastic kernel on $\mathbb{X}$ given $\mathbb{K}_A$, and $\tilde{\alpha} : \mathbb{K}_A \times S \to (0, 1)$ is a function as in (1) representing the discount factors, where $\{\xi_n\}$ is a sequence of observable i.i.d. random variables on a probability space $(\Omega, \mathcal{F}, P)$ with values in $S$ and *unknown* distribution $\theta \in \mathbb{P}(S)$. Finally, the *cost-per-stage* $c$ is a measurable real-valued function on $\mathbb{K}_A$, possibly unbounded.

Throughout the paper, the probability space $(\Omega, \mathcal{F}, P)$ is fixed and a.s. means almost surely with respect to $P$.

In this context, since $\theta$ is unknown and the random disturbance process $\{\xi_n\}$ is observable, before choosing the action $a_n$ at stage $n \in \mathbb{N}$, the controller uses the empirical distribution to get an estimate $\hat{\theta}_n$ of $\theta$. That is, $\left\{\hat{\theta}_n\right\} \subset \mathbb{P}(S)$ is obtained by the process:

$$\hat{\theta}_n(B) := \frac{1}{n} \sum_{i=1}^{n} 1_B(\xi_i), \quad \text{for all} \quad n \in \mathbb{N} \text{ and } B \in \mathcal{B}(S), \tag{5}$$

where $1_B(\cdot)$ denotes the indicator function of the set $B \in \mathcal{B}(S)$. Next, he/she combines this with the history of the system to select a control $a = a_n(\hat{\theta}_n) \in A(x_n)$. Then, a discounted cost as in (2) is incurred, and the system moves to a new state $x_{n+1} = x'$ according to the transition law

$$Q(D|x_n, a_n) := \Pr\left[x_{n+1} \in D | x_n, a_n\right], \quad D \in \mathcal{B}(\mathbb{X}). \tag{6}$$

Once the transition to state $x_{n+1}$ occurs, the process is repeated. The costs are accumulated throughout the evolution of the system in an infinite horizon using a discounted cost criterion with random state–action-dependent discount factors defined below.

### 2.2 Minimax control model

Now we are interested in the situation where the random variable $\xi_n$ represents a random noise which is impossible to observe and, furthermore, its distribution may change from stage to stage. In opposite to the Markov model (3), in this case, the controller cannot estimate, by means of statistical methods, the unknown distribution, and under this scenario we model the control problem as a minimax system. That is, we assume that the controller has an opponent which selects the distribution $\theta_n$ for $\xi_n$ at each time $n \in \mathbb{N}$. Specifically, we consider a minimax control model of the form

$$\mathcal{M}_{\max}^{\min} = (\mathbb{X}, \mathbb{A}, \Theta, \mathbb{K}_A, \mathbb{K}, Q, \tilde{\alpha}, c), \tag{7}$$

where $\mathbb{X}$, $\mathbb{A}$, $Q$, $\tilde{\alpha}$, $c$, and $\mathbb{K}_A$ are as in (3) and (4), and $\Theta \subset \mathbb{P}(S)$ is a Borel subset of probability measures on $S$, which represents the opponent action space. The set $\mathbb{K} \in \mathcal{B}(\mathbb{X} \times \mathbb{A} \times \Theta)$ is the constraint set for the opponent. Hence, we suppose that $\{\xi_n\}$ is a sequence of independent and possibly non-observable random variables on $(\Omega, \mathcal{F}, P)$ taking values on $S$, with corresponding distribution $\theta_n \in \Theta$. That is,

$$\theta_n(B) := P\left[\xi_n \in B\right], \quad n \in \mathbb{N}, \quad B \in \mathcal{B}(S).$$

The model (7) represents a controlled stochastic systems which can be seen as a *game against the nature* whose evolution is as follows. At time $n \in \mathbb{N}$, the system is in state $x_n \in \mathbb{X}$, the controller chooses an action $a_n \in A(x_n)$ and the opponent, the "nature", picks a distribution $\theta_n \in \Theta$ for the random disturbance $\xi_n$. Then the controller incurs a discounted cost

$$\alpha(x_0, a_0, \theta_1)\alpha(x_1, a_1, \theta_2) \cdots \alpha(x_{n-1}, a_{n-1}, \theta_n)c(x_n, a_n), \tag{8}$$

where $\alpha : \mathbb{K} \to (0, 1)$ is the mean discount factor function

$$\alpha(x, a, \theta) := \int_S \tilde{\alpha}(x, a, s)\theta(\mathrm{d}s), \quad (x, a, \theta) \in \mathbb{K}. \tag{9}$$

Next, the process moves to a new state according to the transition law $Q$ and the process is repeated. Thus, the goal of the controller is to minimize the maximum cost incurred by the nature. The corresponding minimax control problem will be defined below in a precise form.

## 3 Optimality criteria

As will be stated below, some properties on the Markov control model (3) as the optimality equation, can be deduced from the results on minimax control model (7) by letting $\Theta = \{\theta\}$, where $\theta$ is the common but unknown distribution of the process $\{\xi_n\}$. Taking into account this fact, and for a clear presentation, we first define the minimax criterion, and hereupon we introduce the performance index corresponding to the model (3).

### 3.1 Minimax control problem

Let $\mathbb{H}_0 := \mathbb{X}$, $\mathbb{H}'_0 := \mathbb{K}_A$, and for $n \in \mathbb{N}$ let $\mathbb{H}_n := \mathbb{K}^n \times \mathbb{X}$ and $\mathbb{H}'_n := \mathbb{K}^n \times \mathbb{K}_A$. Generic elements of $\mathbb{H}_n$ and $\mathbb{H}'_n$ take the form $h_n = (x_0, a_0, \theta_1, \ldots, x_{n-1}, a_{n-1}, \theta_n, x_n)$ and $h'_n = (h_n, a_n)$, respectively. A strategy for the controller is a sequence $\pi = \{\pi_n\}$ of stochastic kernels on $\mathbb{A}$ given $\mathbb{H}_n$ such that $\pi_n(A(x_n)|h_n) = 1$ for all $h_n \in \mathbb{H}_n$ and $n \in \mathbb{N}_0$. If there exists a sequence $\varphi = \{\varphi_n\}$ of stochastic kernels on $\mathbb{A}$ given $\mathbb{X}$ such that $\pi_n(\cdot|h_n) = \varphi_n(\cdot|x_n)$ then $\pi$ is called a Markov strategy. We denote by $\Pi_{\mathbb{A}}$ the set of all strategies for the controller and by $\Pi^M_{\mathbb{A}} \subset \Pi_{\mathbb{A}}$ the subset of Markov strategies. A strategy $\varphi \in \Pi^M_{\mathbb{A}}$ is deterministic if there exists a sequence $\{f_n\}$ of functions in the set

$$\mathbb{F}_{\mathbb{A}} := \{f : \mathbb{X} \to \mathbb{A} | f \quad \text{is measurable and} \quad f(x) \in A(x)\ \forall x \in \mathbb{X}\}$$

such that $\varphi_n(\cdot|x_n)$ is concentrated at $f_n(x_n)$ for each $n \in \mathbb{N}_0$. If $f_n = f \in \mathbb{F}_{\mathbb{A}}$ then $\varphi$ is said to be a deterministic stationary strategy for the controller. If necessary, see for instance Hernández-Lerma and Lasserre (1996) for further information on those strategies. Following a standard convention, we denote by $\mathbb{F}_{\mathbb{A}} \subset \Pi^M_{\mathbb{A}}$ the set of deterministic stationary strategies for the controller, and we denote $\pi \in \mathbb{F}_{\mathbb{A}}$ by $f$.

The strategies for the opponent are defined similarly. That is, a strategy for the opponent is a sequence $\pi' = \{\pi'_n\}$ of stochastic kernels on $\Theta$ given $\mathbb{H}'_n$ such that $\pi'_n\left(\Theta|h'_n\right) = 1$ for all $h'_n \in \mathbb{H}'_n$ and $n \in \mathbb{N}_0$. We denote by $\Pi_\Theta$ the set of all strategies for the opponent, and by $\mathbb{F}_\Theta \subset \Pi_\Theta$ the set of all deterministic stationary strategies. We identify a deterministic stationary strategy $\pi' \in \mathbb{F}_\Theta$ with some measurable function $g : \mathbb{X} \times \mathbb{A} \to \Theta$ such that $\pi'_n\left(\cdot|h'_n\right)$ is concentrated in $g(x_n, a_n) \in \Theta$ for all $h'_n \in \mathbb{H}'_n$ and $n \in \mathbb{N}_0$.

To ease the notation, for each $f \in \mathbb{F}_{\mathbb{A}}$, we write

$$c(x, f) := c(x, f(x)) \text{ and } \alpha(x, f, \theta) := \alpha(x, f(x), \theta), \quad \theta \in \Theta, \quad x \in \mathbb{X}.$$

According to (8), we are assuming that a cost $C$ incurred at stage $n$ is equivalent to a cost $\Gamma_n C$ at time 0, where

$$\Gamma_n = \prod_{k=0}^{n-1} \alpha(x_k, a_k, \theta_{k+1}) \text{ if } n \in \mathbb{N}, \tag{10}$$

and $\Gamma_0 = 1$. Hence, for each pair of strategies $(\pi, \pi') \in \Pi_{\mathbb{A}} \times \Pi_\Theta$ and initial state $x \in \mathbb{X}$, we define the total expected discounted cost—with random state–action-dependent discount factors—as

$$V(x, \pi, \pi') := E_x^{\pi\pi'}\left[\sum_{n=0}^{\infty} \Gamma_n c(x_n, a_n)\right], \tag{11}$$

where $E_x^{\pi\pi'}$ denotes the expectation operator with respect to the probability measure $P_x^{\pi\pi'}$ induced by $(\pi, \pi') \in \Pi_{\mathbb{A}} \times \Pi_{\Theta}$, given $x_0 = x$ [for the construction of $P_x^{\pi\pi'}$ see, for instance, Dynkin and Yushkevich (1979)].

Thus, the minimax control problem to the controller is to find a strategy $\pi^* \in \Pi_{\mathbb{A}}$ such that

$$V^*(x) := \inf_{\pi \in \Pi_{\mathbb{A}}} \sup_{\pi' \in \Pi_{\Theta}} V(x, \pi, \pi') = \sup_{\pi' \in \Pi_{\Theta}} V(x, \pi^*, \pi'), \quad x \in \mathbb{X}. \tag{12}$$

In this case, the strategy $\pi^*$ is said to be minimax, whereas $V^*$ is the minimax value function.

### 3.2 The estimation and control problem

Since the disturbance process $\{\xi_n\}$ is a sequence of observable i.i.d. random variables, the actions or controls applied at time $n \in \mathbb{N}_0$ by the controller are selected on the knowledge of the observed history $h_n = (x_0, a_0, \xi_1, \ldots, x_{n-1}, a_{n-1}, \xi_n, x_n)$. That is, we consider $\mathbb{H}_0 := \mathbb{X}$ and $\mathbb{H}_n := (\mathbb{K}_A \times S)^n \times \mathbb{X}$ for $n \in \mathbb{N}$ , as the spaces of admissible histories up to time $n$. Then the strategies for the controller under this context are defined similarly as the minimax criterion. For notational convenience, we will keep denoting by $\Pi_{\mathbb{A}}$ and $\mathbb{F}_{\mathbb{A}}$ the sets of all strategies and stationary strategies for the controller, respectively.

Now, taking into account (2), when using a strategy $\pi \in \Pi_{\mathbb{A}}$, given the initial state $x_0 = x \in \mathbb{X}$, we define the total expected discounted cost—with random state–action-dependent discount factors—as

$$V(x, \pi) := E_x^{\pi} \left[ \sum_{n=0}^{\infty} \tilde{\Gamma}_n c(x_n, a_n) \right], \tag{13}$$

where

$$\tilde{\Gamma}_n = \prod_{k=0}^{n-1} \tilde{\alpha}(x_k, a_k, \xi_{k+1}) \ \text{ if } \ n \in \mathbb{N}, \ \text{ and } \ \tilde{\Gamma}_0 = 1.$$

Then, the optimal control problem associated with the control model (3) is to find an optimal strategy $\pi^* \in \Pi_{\mathbb{A}}$ such that $V(x, \pi^*) = V(x)$ for all $x \in \mathbb{X}$, where

$$V(x) := \inf_{\pi \in \Pi_{\mathbb{A}}} V(x, \pi), \quad x \in \mathbb{X}, \tag{14}$$

is the optimal value function.

However, as is proved in Lemma 15 in Sect. 6, the Markov strategies are sufficient to solve the optimal control problem. That is, for each $\pi \in \Pi_{\mathbb{A}}$ there exists $\varphi \in \Pi_{\mathbb{A}}^M$ such that

$$V(x, \pi) = V(x, \varphi), \quad x \in \mathbb{X}. \tag{15}$$

Therefore, our problem is to find a strategy $\varphi^* \in \Pi_{\mathbb{A}}^M$ such that

$$V(x, \varphi^*) = \inf_{\varphi \in \Pi_{\mathbb{A}}^M} V(x, \varphi) =: V(x), \quad x \in \mathbb{X}. \tag{16}$$

Furthermore (see Lemma 16, Sect. 6), on the class of Markov strategies $\Pi_{\mathbb{A}}^M$ we can write the performance index (13) in terms of the unknown distribution $\theta$ of the disturbance process $\{\xi_n\}$ as (see Lemma 16 , below)

$$V(x, \varphi) := E_x^{\varphi}\left[\sum_{n=0}^{\infty} \Gamma_n c(x_n, a_n)\right], \quad x \in \mathbb{X}, \ \varphi \in \Pi_{\mathbb{A}}^M, \tag{17}$$

where $\Gamma_n$ is as in (10). Observe that in this case

$$\Gamma_n = \prod_{k=0}^{n-1} \alpha(x_k, a_k, \theta), \ \text{ if } \ n \in \mathbb{N}.$$

The optimal control problem is studied by combining the empirical estimation process (5) of the distribution $\theta$ with minimization procedures. However, as the performance index (13) depends strongly on the controls selected at the first stages, precisely when the information about the unknown distribution is poor, we cannot ensure, in general, the existence of optimal strategies. Hence, the optimality of strategies constructed under this context will be studied in the following asymptotic sense.

**Definition 1** A strategy $\pi \in \Pi_A$ is said to be asymptotically optimal for the control model $\mathcal{M}$ if, for all $x \in \mathbb{X}$,

$$\left| V^{(n)}(x, \pi) - E_x^{\pi}\left[V(x_n)\right] \right| \to 0 \ \text{ as } \ n \to \infty,$$

where

$$V^{(n)}(x, \pi) := E_x^{\pi}\left[\sum_{t=n}^{\infty} \Gamma_{n,t} c(x_t, a_t)\right] \tag{18}$$

is the total expected discounted cost—with random state–action-dependent discount factors—from stage $n$ onward, and

$$\Gamma_{n,t} = \prod_{k=n}^{t-1} \alpha(x_k, a_k, \theta), \quad \text{for } t > n \text{ and } \Gamma_{n,n} = 1. \tag{19}$$

The notion of asymptotic optimality was introduced by Schäl (1987) to study a problem of estimation and control in dynamic programming with constant random discount factors (see also Gordienko and Minjárez-Sosa 1998; Hilgert and Minjárez-Sosa 2001).

## 4 Assumptions and preliminary results

We shall require the following general boundedness and continuity assumptions on the control models. Observe that Assumption 2(a) allows an unbounded cost function $c(x, a)$ provided that it is majorized by a "bounding" function $W$. Under these assumptions, we next establish some preliminary facts which will be used to prove our main results.

**Assumption 2** (a) The cost function $c(x, a)$ is lower semi-continuous (l.s.c.) on $\mathbb{K}_A$. Moreover, there exist a continuous function $W : \mathbb{X} \to [1, \infty)$ and a positive constant $c_0$ such that

$$|c(x, a)| \leq c_0 W(x) \qquad \forall (x, a) \in \mathbb{K}_A. \tag{20}$$

(b) The transition law $Q$ is weakly continuous, that is, for each continuous and bounded function $u : \mathbb{X} \to \mathbb{R}$, the function

$$(x, a) \longmapsto \int_{\mathbb{X}} u(y)\, Q(\mathrm{d}y \mid x, a) \tag{21}$$

is continuous on $\mathbb{K}_A$.

(c) The function

$$(x, a) \longmapsto \int_{\mathbb{X}} W(y)\, Q(\mathrm{d}y \mid x, a)$$

is continuous on $\mathbb{K}_A$.

(d) The multifunction $x \to A(x)$ is upper semi-continuous (u.s.c.), and the set $A(x)$ is compact for each $x \in \mathbb{X}$.

(e) The function $\tilde{\alpha}(x, a, s)$ is continuous on $\mathbb{K}_A \times S$, and

$$\alpha^* := \sup_{(x,a,s) \in \mathbb{K}_A \times S} \tilde{\alpha}(x, a, s) < 1. \tag{22}$$

(f) There exists a positive constant $b$ such that

$$1 \leq b < (\alpha^*)^{-1},$$

and for all $(x, a) \in \mathbb{K}_A$

$$\int_{\mathbb{X}} W(y) Q(\mathrm{d}y \mid x, a) \leq b W(x). \tag{23}$$

We denote by $\mathbb{B}_W$ the Banach space of all measurable functions $u : \mathbb{X} \to \mathbb{R}$ with the $W$-norm

$$\|u\|_W := \sup_{x \in \mathbb{X}} \frac{|u(x)|}{W(x)} < \infty, \tag{24}$$

and by $\mathbb{L}_W$ the subspace of l.s.c. functions in $\mathbb{B}_W$. We will repeatedly use the following inequalities. For any $u \in \mathbb{B}_W$,

$$|u(x)| \leq \|u\|_W \, W(x) \tag{25}$$

and

$$\int_{\mathbb{X}} u(y) \, Q(\mathrm{d}y \mid x, a) \leq b \, \|u\|_W \, W(x), \tag{26}$$

for all $(x, a) \in \mathbb{K}_A$. The inequality (25) is a consequence of the definition (24), whereas (26) follows from (23) and (25).

*Remark 3* (a) As is well known (see Hernández-Lerma and Lasserre 1996), the Assumption 2(b) can be substituted by the following equivalent condition: For each l.s.c. and bounded below function $u : \mathbb{X} \to \mathbb{R}$, the function in (21) is l.s.c. on $\mathbb{K}_A$.

(b) Under Assumption 2(e), the monotone convergence theorem yields that the function $\alpha : \mathbb{K} \to (0, 1)$ defined in (9) is continuous. Indeed, let $\{(x_n, a_n, \theta_n)\} \in \mathbb{K}$ be a sequence converging to $(x, a, \theta) \in \mathbb{K}$ and denote $\tilde{\alpha}_*(s) := \liminf_{n\to\infty} \tilde{\alpha}(x_n, a_n, s)$ and $\tilde{\alpha}_k(s) := \inf_{j\geq k} \tilde{\alpha}(x_j, a_j, s)$. Observe that $\tilde{\alpha}_k(s) \nearrow \tilde{\alpha}_*(s)$, as $k \to \infty$. Then, for all $n \geq k$, $\tilde{\alpha}(x_n, a_n, \cdot) \geq \tilde{\alpha}_k(\cdot)$. Hence, since $\theta_n \to \theta$ weakly,

$$\liminf_{n\to\infty} \int_S \tilde{\alpha}(x_n, a_n, s)\theta_n(\mathrm{d}s) \geq \liminf_{n\to\infty} \int_S \tilde{\alpha}_k(s)\theta_n(\mathrm{d}s) = \int_S \tilde{\alpha}_k(s)\theta(\mathrm{d}s).$$

Now, letting $k \to \infty$, by Assumption 2(e) and the monotone convergence theorem we obtain

$$\liminf_{n\to\infty} \int_S \tilde{\alpha}(x_n, a_n, s)\theta_n(\mathrm{d}s) \geq \int_S \tilde{\alpha}_*(s)\theta(\mathrm{d}s) = \int_S \tilde{\alpha}(x, a, s)\theta(\mathrm{d}s). \tag{27}$$

Similarly we can prove the inequality

$$\limsup_{n\to\infty} \int_S \tilde{\alpha}(x_n, a_n, s)\theta_n(\mathrm{d}s) \leq \int_S \tilde{\alpha}(x, a, s)\theta(\mathrm{d}s),$$

which combined with (27) implies that $\alpha : \mathbb{K} \to (0, 1)$ is a continuous function.

(c) From Assumption 2(f), for all $x \in \mathbb{X}$, $n \in \mathbb{N}_0$, and $(\pi, \pi') \in \Pi_{\mathbb{A}} \times \Pi_\Theta$,

$$E_x^{\pi\pi'}\left[W(x_{n+1})\right] \leq b E_x^{\pi\pi'}\left[W(x_n)\right].$$

Iterating this inequality we obtain

$$E_x^{\pi\pi'}\left[W(x_{n+1})\right] \leq b^{n+1} W(x), \quad x \in \mathbb{X}. \tag{28}$$

Therefore, from Assumptions 2(a), (e), and (f), for each $x \in \mathbb{X}$, and $(\pi, \pi') \in \Pi_{\mathbb{A}} \times \Pi_{\Theta}$,

$$\left|V\left(x,\pi,\pi'\right)\right| \leq E_x^{\pi\pi'}\left[\sum_{n=0}^{\infty}\left(\alpha^*\right)^n |c(x_n,a_n)|\right]$$
$$\leq c_0 W(x)\sum_{n=0}^{\infty}\left(b\alpha^*\right)^n = \frac{c_0 W(x)}{1-b\alpha^*}.$$

This implies that $V^* \in \mathbb{B}_W$. In fact we have,

$$\left\|V^*\right\|_W \leq \frac{c_0}{1-b\alpha^*}. \tag{29}$$

We next introduce the following family of operators. For each function $u$ on $\mathbb{X}$ and $(x,a,\theta) \in \mathbb{K}$ we define:

$$T_{(a,\theta)}u(x) := c(x,a) + \alpha(x,a,\theta)\int_{\mathbb{X}} u(y)Q(\mathrm{d}y \mid x,a), \tag{30}$$

$$\hat{T}_a u(x) := \sup_{\theta\in\Theta} T_{(a,\theta)}u(x), \tag{31}$$

and

$$Tu(x) := \inf_{a\in A(x)}\sup_{\theta\in\Theta} T_{(a,\theta)}u(x) = \inf_{a\in A(x)}\hat{T}_a u(x). \tag{32}$$

We conclude this section summarizing some useful properties of the operators (30)–(32).

**Lemma 4** *If Assumption 2 holds, then:*

(a) $\left|\hat{T}_a u(x)\right| \leq \hat{M}W(x)$ *for all* $(x,a) \in \mathbb{K}_A$, $u \in \mathbb{B}_W$, *and some* $\hat{M} < \infty$.
(b) *The mapping* $(x,a,\theta) \to T_{(a,\theta)}u(x)$ *is l.s.c. on* $\mathbb{K}$ *for all* $u \in \mathbb{L}_W$.
(c) *The operator* $T$ *is a contraction on* $\mathbb{B}_W$ *with modulus* $b\alpha^* < 1$.
*In addition, if* $\Theta$ *is a compact set, then*
(d) $T$ *maps* $\mathbb{L}_W$ *into itself.*
(e) *For each* $u \in \mathbb{L}_W$, *there exists* $f^* \in \mathbb{F}_{\mathbb{A}}$ *such that*

$$Tu(x) = \sup_{\theta\in\Theta} T_{(f^*,\theta)}u(x), \quad x \in \mathbb{X}.$$

*Proof* (a) This part follows from (20), (22), (25), and (26) by taking $\hat{M} := c_0 + \alpha^* b\,\|u\|_W$.
(b) Observe that if $u \in \mathbb{L}_W$ then, from (24) and Assumption 2(a), the function $v(x) := u(x) + \|u\|_W W(x)$ is nonnegative and l.s.c. Thus, Assumption 2(b) (see Remark 3(a)) implies that the mapping $(x,a) \to \int_{\mathbb{X}} v(y)Q(\mathrm{d}y \mid x,a)$ is l.s.c. on $\mathbb{K}_A$, which, together with Assumption 2(c) implies that the mapping

$(x,a) \to \int_{\mathbb{X}} u(y)Q(\mathrm{d}y \mid x,a)$ is l.s.c. on $\mathbb{K}_A$. Then, Assumptions 2(a), (e) yield the lower semi-continuity of $T_{(a,\theta)}u(x)$ on $\mathbb{K}$.

(c) Let $u, u' \in \mathbb{B}_W$. Then, from definition (32) of the operator $T$, we have, for each $x \in \mathbb{X}$,

$$\begin{aligned}\left|Tu(x) - Tu'(x)\right| &\leq \sup_{a\in A(x)}\sup_{\theta\in\Theta}\alpha(x,a,\theta)\int_{\mathbb{X}}\left|u(y) - u'(y)\right| Q(\mathrm{d}y \mid x,a)\\ &\leq \alpha^*\left\|u-u'\right\|_W\int_{\mathbb{X}} W(y)Q(\mathrm{d}y \mid x,a)\\ &\leq \alpha^* b\left\|u-u'\right\|_W W(x),\end{aligned}$$

where the last two inequalities follows from (25) and (26 ). Therefore,

$$\left\|Tu - Tu'\right\|_W \leq \alpha^* b\left\|u-u'\right\|_W.$$

(d) Let $\{(x_m,a_m)\} \subset \mathbb{K}_A$ be a sequence such that $(x_m,a_m) \to (x,a) \in \mathbb{K}_A$, and $\theta \in \Theta$ be arbitrary. From the compactness of the set $\Theta$ there exists a sequence $\{\theta_m\}$ such that $\theta_m \to \theta$. Then, from the part (b) of the lemma,

$$\begin{aligned}\liminf_{m\to\infty}\hat{T}_{a_m}u(x_m) &= \liminf_{m\to\infty}\sup_{\theta\in\Theta}T_{(a_m,\theta)}u(x_m)\\ &\geq \liminf_{m\to\infty}T_{(a_m,\theta_m)}u(x_m)\\ &\geq T_{(a,\theta)}u(x).\end{aligned}$$

Since $\theta$ is arbitrary, we have

$$\liminf_{m\to\infty}\hat{T}_{a_m}u(x_m) \geq \hat{T}_a u(x),$$

which implies that $\hat{T}_a u(x)$ is l.s.c. on $\mathbb{K}_A$. Therefore, from the part (a), the function $\hat{T}_a u(x) + \hat{M}W(x)$ is nonnegative and l.s.c. on $\mathbb{K}_A$. Thus, from Assumption 2(d) and due to a well-known result in Schäl (1975) [see also Proposition D.5 in Hernández-Lerma and Lasserre (1996) and Rieder (1978)], we have that

$$\inf_{a\in A(x)}\left\{\hat{T}_a u(x) + \hat{M}W(x)\right\} = Tu(x) + \hat{M}W(x) \tag{33}$$

is a l.s.c. function on $\mathbb{X}$, which, in turn implies the lower semi-continuity of $Tu(x)$ on $\mathbb{X}$. Finally, because $\|Tu\|_W < \infty$ [see part (a) of the Lemma and (32)] we have that $T$ maps $\mathbb{L}_W$ into itself.

(e) Since $\hat{T}_a u(x) + \hat{M}W(x)$ is a nonnegative and l.s.c. function, from Assumption 2(d) and by applying standard arguments on the existence of minimizers [see, for instance, Rieder (1978); Schäl (1975)] there exists $f^* \in \mathbb{F}_A$ such that

$$\inf_{a\in A(x)}\left\{\hat{T}_a u(x) + \hat{M}W(x)\right\} = \hat{T}_{f^*}u(x) + \hat{M}W(x),$$

which, together with (33) proves the part (e).

□

## 5 Minimax strategies

In this section, we present the results corresponding to the minimax criterion. First, we define the minimax value iteration function and prove the geometric convergence to the minimax value function. Then we show the existence of minimax strategies

It is easy to prove that $\mathbb{L}_W \subset \mathbb{B}_W$ is a closed set, which implies that it is a complete subset of $\mathbb{B}_W$. Then, from Lemma 4(c) and the Banach's Fixed Point Theorem, there exists a unique function $\tilde{u} \in \mathbb{L}_W$ such that for all $x \in \mathbb{X}$,

$$\tilde{u}(x) = T\tilde{u}(x) \tag{34}$$

and

$$\left\|T^n u - \tilde{u}\right\|_W \leq \left(b\alpha^*\right)^n \|u - \tilde{u}\|_W \quad \forall u \in \mathbb{L}_W,\ n \in \mathbb{N}_0. \tag{35}$$

Now we define the sequence of minimax value iteration function $\{v_n\}$ in $\mathbb{L}_W$ as

$$\begin{aligned} v_0 &= 0, \\ v_n(x) &= T v_{n-1}(x) = T^n v_0, \quad n \in \mathbb{N}, x \in \mathbb{X}. \end{aligned} \tag{36}$$

Observe that from (35) and (36), taking $u = v_0$ we get

$$\|v_n - \tilde{u}\|_W \leq \left(b\alpha^*\right)^n \|\tilde{u}\|_W \quad \forall n \in \mathbb{N}_0. \tag{37}$$

We state our minimax result as follows.

**Theorem 5** *If Assumption 2 holds and $\Theta$ is a compact set, then:*

(a) *The minimax value function (12) is the unique solution in $\mathbb{L}_W$ satisfying*

$$V^*(x) = TV^*(x), \quad x \in \mathbb{X}. \tag{38}$$

(b) *For each $n \in \mathbb{N}$,*

$$\left\|v_n - V^*\right\|_W \leq c_0 \frac{(b\alpha^*)^n}{1 - b\alpha^*}.$$

(c) *There exists $f^* \in \mathbb{F}_{\mathbb{A}}$ such that*

$$V^*(x) = \sup_{\theta \in \Theta} T_{(f^*,\theta)} V^*(x) = \hat{T}_{f^*} V^*(x), \quad x \in \mathbb{X}, \tag{39}$$

*and moreover, $f^*$ is a minimax strategy for the controller, that is,*

$$V^*(x) = \sup_{\pi' \in \Pi_\Theta} V(x, f^*, \pi').$$

*Proof* From (34) and (37), parts (a) and (b) will be proved if we show that $\tilde{u} = V^*$. To this end, let $f \in \mathbb{F}_{\mathbb{A}}$ be a selector such that

$$\tilde{u}(x) = \sup_{\theta \in \Theta} T_{(f,\theta)} \tilde{u}(x), \quad x \in \mathbb{X},$$

which exists because of Lemma 4(e). Then, from (30)

$$\tilde{u}(x) \geq c(x, f) + \alpha(x, f, \theta) \int_{\mathbb{X}} \tilde{u}(y) Q(\mathrm{d}y \mid x, f) \quad \forall x \in \mathbb{X},\ \theta \in \Theta. \tag{40}$$

Now, for an arbitrary strategy $\pi' \in \Pi_\Theta$ for the opponent, iteration of the inequality (40) yields

$$\begin{aligned} \tilde{u}(x) &\geq E_x^{f\pi'} \left[ c(x_0, f) + \sum_{n=1}^{m-1} \prod_{k=0}^{n-1} \alpha(x_k, f, \theta_{k+1}) c(x_n, f) \right] \\ &\quad + E_x^{f\pi'} \left[ \prod_{k=0}^{m-1} \alpha(x_k, f, \theta_{k+1}) \tilde{u}(x_m) \right] \\ &= E_x^{f\pi'} \left[ \sum_{n=1}^{m-1} \Gamma_n c(x_n, f) \right] + E_x^{f\pi'} \left[ \Gamma_m \tilde{u}(x_m) \right]. \end{aligned} \tag{41}$$

Combining (22), (25), and (28), we have

$$E_x^{f\pi'} \left[ \Gamma_m \tilde{u}(x_m) \right] \leq \left( b\alpha^* \right)^m \|\tilde{u}\|_W W(x), \quad x \in \mathbb{X}.$$

Hence, letting $m \to \infty$ in (41), from (11) we get

$$\tilde{u}(x) \geq V\left(x, f, \pi'\right) \quad \forall x \in \mathbb{X},\ \pi' \in \Pi_\Theta. \tag{42}$$

As $\pi' \in \Pi_\Theta$ is arbitrary, (42) and (12) yield

$$\tilde{u}(x) \geq V^*(x) \quad \forall x \in \mathbb{X}. \tag{43}$$

On the other hand, since $\alpha$ is a continuous function in $\theta$ (see Remark 3(b)), from (34) and the compactness of $\Theta$, for each $(x, a) \in \mathbb{K}_{\mathbb{A}}$, there exists $g : \mathbb{K}_{\mathbb{A}} \to \Theta$, such that $g(x, a) \in \Theta$ satisfies

$$\begin{aligned} \tilde{u}(x) &= \inf_{a \in A(x)} T_{(a,g)} \tilde{u}(x) = \inf_{a \in A(x)} \left\{ c(x, a) + \alpha(x, a, g) \int_{\mathbb{X}} \tilde{u}(y) Q(\mathrm{d}y \mid x, a) \right\} \\ &\leq c(x, a) + \alpha(x, a, g) \int_{\mathbb{X}} \tilde{u}(y) Q(\mathrm{d}y \mid x, a) \quad \forall x \in \mathbb{X},\ a \in A(x). \end{aligned} \tag{44}$$

Similarly as in (41) and (42), for a strategy $\pi \in \Pi_{\mathbb{A}}$, iteration of (44) yields

$$\tilde{u}(x) \leq V(x, \pi, g) \quad \forall x \in \mathbb{X}. \tag{45}$$

Thus, since

$$V(x, \pi, g) \leq \sup_{g \in \mathbb{F}_\Theta} V(x, \pi, g) \leq \sup_{\pi' \in \Pi_\Theta} V\left(x, \pi, \pi'\right) \quad \forall x \in \mathbb{X},$$

and, in addition, $\pi \in \Pi_{\mathbb{A}}$ is arbitrary, (45) implies

$$\tilde{u}(x) \leq \inf_{\pi \in \Pi_{\mathbb{A}}} \sup_{\pi' \in \Pi_\Theta} V\left(x, \pi, \pi'\right) = V^*(x) \quad \forall x \in \mathbb{X}. \tag{46}$$

Therefore, combining (43) and (46) we get $\tilde{u} = V^*$ which proves the parts (a) and (b) of the theorem.

Finally, the existence of $f^* \in \mathbb{F}_{\mathbb{A}}$ satisfying (39) follows from (38) and Lemma 4(e). Moreover, similarly as in (42), we have that for an arbitrary strategy $\pi' \in \Pi_\Theta$,

$$V^*(x) \geq V\left(x, f^*, \pi'\right) \quad \forall x \in \mathbb{X},$$

which implies that

$$V^*(x) = \sup_{\pi' \in \Pi_\Theta} V\left(x, f^*, \pi'\right) \quad \forall x \in \mathbb{X}.$$

□

## 6 Asymptotically optimal strategies

We consider the control model (3). In this case we are supposing that $\{\xi_n\}$ is a sequence of observable i.i.d. random variables with unknown distribution $\theta \in \mathbb{P}(S)$, and our objective is to study the optimal control problem (16) which, taking $\Theta = \{\theta\}$, can be seen as a particular case of the minimax control problem (12).

Considering this fact and (17) (see (11)), the operator $T$ defined in (32) takes the form

$$\begin{aligned} Tu(x) &= \inf_{a \in A(x)} T_{(a,\theta)} u(x) \\ &= \inf_{a \in A(x)} \left\{ c(x, a) + \alpha(x, a, \theta) \int_{\mathbb{X}} u(y) Q(\mathrm{d}y \mid x, a) \right\}, \end{aligned}$$

for each function $u$ on $\mathbb{X}$. Hence, we have the following consequences of Theorem 5.

**Proposition 6** *Suppose that Assumption 2 holds. Then:*

(a) *The value function* (14) (*see* (16)) *satisfies*

$$V(x) = TV(x), \quad x \in \mathbb{X}, \tag{47}$$

*and moreover* (*see* (29)),

$$\|V\|_W \leq \frac{c_0}{1 - b\alpha^*}. \tag{48}$$

(b) *There exists* $f^* \in \mathbb{F}_{\mathbb{A}}$ *such that*

$$V(x) = T_{(f^*,\theta)}V(x) = c\left(tx, f^*\right) + \alpha\left(x, f^*, \theta\right) \int_{\mathbb{X}} V(y)Q(\mathrm{d}y \mid x, f^*), \quad x \in \mathbb{X}.$$

Since $\theta$ is unknown, the solution given for the Proposition 6 is not accessible for the controller. Under these circumstances, using the empirical distribution $\hat{\theta}_n$ to estimate $\theta$ (see (5)), our objective is to show the existence of asymptotically optimal strategies.

### 6.1 Empirical estimation

It is well known that $\hat{\theta}_n$ converges weakly to $\theta$ a.s. Hence, for each $(x, a) \in \mathbb{K}_A$

$$\int_S \tilde{\alpha}(x, a, s)\hat{\theta}_n(\mathrm{d}s) \to \int_S \tilde{\alpha}(x, a, s)\theta(\mathrm{d}s) \ \text{ a.s}, \quad \text{as } n \to \infty.$$

That is, as $n \to \infty$,

$$\alpha\left(x, a, \hat{\theta}_n\right) \to \alpha(x, a, \theta) \quad \text{a.s.}$$

However, this convergence is not sufficient for our objective. Specifically we require uniform convergence on the set $\mathbb{K}_A$. Then, to state the suitable estimation process we need to impose the following assumption.

**Assumption 7** The family of functions

$$\mathcal{A} := \{\tilde{\alpha}(x, a, \cdot) : (x, a) \in \mathbb{K}_A\}$$

is equicontinuous on $S$.

Then, as a consequence of Theorem 6.4 in Ranga Rao (1962) we have the following result.

**Lemma 8** *Under Assumption 7, as* $n \to \infty$,

$$\sup_{(x,a)\in\mathbb{K}_A} \left|\alpha(x, a, \hat{\theta}_n) \to \alpha(x, a, \theta)\right| \to 0 \quad \text{a.s.} \tag{49}$$

*Remark 9* An obvious sufficient condition for Assumption 7 is that the disturbance set $S$ is countable with the discrete topology.

The uniform convergence (49) is used to obtain a (non-stationary) value iteration algorithm to approximate the value function (16), which will be a key point to construct asymptotically optimal strategies in the next subsection.

Let $\{V_n\}$ be a sequence of functions defined as

$$\begin{aligned} V_0 &\equiv 0; \\ V_n(x) &= \inf_{a\in A(x)} T_{(a,\hat{\theta}_n)} V_{n-1}(x). \end{aligned} \tag{50}$$

A straightforward calculation shows that (see (48))

$$\|V_n\|_W \leq \frac{c_0}{1-b\alpha^*} \quad \text{a.s., } \forall n \in \mathbb{N}. \tag{51}$$

That is, $V_n \in \mathbb{B}_W$ a.s., for all $n \in \mathbb{N}$.

**Proposition 10** *Under Assumptions 2 and 7,*

$$\|V - V_n\|_W \to 0 \text{ a.s.,} \quad \text{as} \quad n \to \infty.$$

*Proof* From (47) and (50), by adding and subtracting the term $\alpha(x,a,\hat{\theta}_n)\int_{\mathbb{X}} V(y)\, Q(\mathrm{d}y \mid x,a)$, we have, for each $x \in \mathbb{X}$ and $n \in \mathbb{N}$,

$$\begin{aligned} |V(x) - V_n(x)| &\leq \sup_{a\in A(x)} \left\{ \left|\alpha(x,a,\theta) - \alpha\left(x,a,\hat{\theta}_n\right)\right| \int_{\mathbb{X}} |V(y)|\, Q(\mathrm{d}y|x,a) \right. \\ &\quad \left. + \alpha(x,a,\hat{\theta}_n) \int_{\mathbb{X}} |V(y) - V_{n-1}(y)|\, Q(\mathrm{d}y|x,a) \right\} \\ &\leq b\,\|V\|_W\, W(x) \sup_{a\in A(x)} \left|\alpha(x,a,\theta) - \alpha\left(x,a,\hat{\theta}_n\right)\right| \\ &\quad + \alpha^* b\,\|V - V_{n-1}\|_W\, W(x), \end{aligned}$$

where the last inequality comes from (22)–(26). Therefore, from (48),

$$\|V - V_n\|_W \leq \frac{bc_0}{1-\alpha^* b} \sup_{(x,a)\in\mathbb{K}} \left|\alpha(x,a,\theta) - \alpha\left(x,a,\hat{\theta}_n\right)\right| + \alpha^* b\,\|V - V_{n-1}\|_W. \tag{52}$$

Now, let $l := \limsup_{n\to\infty} \|V - V_n\|_W < \infty$ (see (48) and (51)). Taking lim sup on both sides of (52), from Lemma 8 we get $l < \alpha^* b l$. Then, observing that $0 < \alpha^* b < 1$ (see Assumption 2(f)) we obtain $l = 0$ which proves the proposition. □

## 6.2 Asymptotically optimal strategies

Consider the value iteration function $V_n$ defined in (50). Observe that under Assumption 2, applying standard arguments on the existence of minimizers (see Proposition 6), for each $n \in \mathbb{N}$, there exists $f_n \in \mathbb{F}_A$ such that

$$V_n(x) = T_{\left(f_n,\hat{\theta}_n\right)} V_{n-1}(x), \quad x \in \mathbb{X}. \tag{53}$$

Now, let $\hat{\pi} = \{\hat{\pi}_n\} \in \Pi_{\mathbb{A}}^M$ be the strategy determined by the sequence $\{f_n\}$, that is, $\hat{\pi}_n(\cdot|h_n)$ is concentrated on $f_n(x_n)$ for all $h_n \in \mathbb{H}_n$ and $n \in \mathbb{N}$, with $\hat{\pi}_0$ any fixed action. Then, our objective is to show that $\hat{\pi}$ is an asymptotically optimal strategy.

Before to state the result, we need to impose the following technical requirement.

**Assumption 11** There exist positive constants $d_0 < \infty$, $\beta_0 < 1$, and $p > 1$ such that for all $(x, a) \in \mathbb{K}_A$,

$$\int_{\mathbb{X}} W^p(y) Q(\mathrm{d}y|x, a) \leq \beta_0 W^p(x) + d_0. \tag{54}$$

*Remark 12* (a) Applying Jensen's inequality to (54) we have, for all $(x, a) \in \mathbb{K}_A$,

$$\int_{\mathbb{X}} W(y) Q(\mathrm{d}y|x, a) \leq \beta' W(x) + d', \tag{55}$$

where $\beta' = \beta_0^{1/p}$ and $d' = d_0^{1/p}$. Moreover, as a consequence of both inequalities (54) and (55) we have

$$\sup_{n\in\mathbb{N}_0} E_x^\pi \left[W^p(x_n)\right] < \infty \tag{56}$$

and

$$\sup_{n\in\mathbb{N}_0} E_x^\pi \left[W(x_n)\right] < \infty. \tag{57}$$

Indeed, first note that from (54)

$$E_x^\pi \left[W^p(x_n)\right] \leq \beta_0 E_x^\pi [W^p(x_{n-1})] \Big] + d_0, \quad n \in \mathbb{N}.$$

Then, iterating this inequality and using the fact $\beta_0 < 1$ we obtain

$$E_x^\pi \left[W^p(x_n)\right] \leq \beta_0^n W^p(x) + \left(1 + \beta_0 + \cdots + \beta_0^{n-1}\right) d_0 \leq W^p(x) + d_0/(1-\beta_0),$$

which, in turns implies (56). Similarly, (57) follows from (55).

(b) In addition, since $W(\cdot) \geq 1$, observe that if $\left(\beta_0^{1/p} + d_0^{1/p}\right) \alpha^* < 1$, the relation (55) implies Assumption 2(f).

**Theorem 13** *Under Assumptions 2, 7, and 11 the strategy $\hat{\pi}$ is asymptotically optimal.*

The proof of Theorem 13 is based on the following characterization of asymptotic optimality which is an adaptation of the Theorem 4.6.2 in Hernández-Lerma and Lasserre (1996).

**Lemma 14** *Under Assumption 2, a policy $\pi \in \Pi_A$ is asymptotically optimal for the control model $\mathcal{M}$ if for all $x \in \mathbb{X}$,*

$$\lim_{n\to\infty} E_x^\pi \left[\Phi(x_n, a_n)\right] = 0,$$

*where*

$$\Phi(x,a) := c(x,a) + \alpha(x,a,\theta) \int_{\mathbb{X}} V(y)\, Q(\mathrm{d}y|x,a) - V(x), \quad (x,a) \in \mathbb{K}_A.$$

*Proof* First observe the following facts. For each $x \in \mathbb{X}$, $\pi \in \Pi_A$, and $n \in \mathbb{N}$, from Assumptions 2(e), (f), and relations (28) and (48),

$$E_x^\pi \left[\Gamma_{n,m} V(x_m)\right] \le \frac{(\alpha^*)^{m-n} b^m c_0}{1 - b\alpha^*} W(x).$$

Hence, because $\alpha^* b \in (0,1)$,

$$\lim_{m\to\infty} E_x^\pi \left[\Gamma_{n,m} V(x_m)\right] = 0. \tag{58}$$

In addition, since $\Phi$ is nonnegative (see (47)), for each $x \in \mathbb{X}$ and $\pi \in \Pi_A$,

$$\lim_{t\to\infty} E_x^\pi \left[\Phi(x_t, a_t)\right] = 0 \text{ implies } \lim_{n\to\infty} \sum_{t=n}^{\infty} E_x^\pi \left[\Gamma_{n,t} \Phi(x_t, a_t)\right] = 0. \tag{59}$$

Now, for each $x \in \mathbb{X}$, $\pi \in \Pi_A$, and $t \in \mathbb{N}$,

$$\Phi(x_t, a_t) = E_x^\pi \left[c(x_t, a_t) + \alpha(x_t, a_t, \theta) V(x_{t+1}) - V(x_t) | h_t, a_t\right], \tag{60}$$

where $h_t$ is the history of the system up to time $t$. Then, from (18), (19), (60), and applying properties of conditional expectation, for each $n \ge t$, $x \in \mathbb{X}$, and $\pi \in \Pi_A$,

$$\begin{aligned}
&\sum_{t=n}^{\infty} E_x^\pi \left[\Gamma_{n,t} \Phi(x_t, a_t)\right] \\
&= \sum_{t=n}^{\infty} E_x^\pi \left[\Gamma_{n,t} E_x^\pi \left[c(x_t, a_t) + \alpha(x_t, a_t, \theta) V(x_{t+1}) - V(x_t) | h_t, a_t\right]\right] \\
&= \sum_{t=n}^{\infty} \left\{E_x^\pi \left[\Gamma_{n,t} c(x_t, a_t)\right] + E_x^\pi \left[\Gamma_{n,t+1} V(x_{t+1}) - \Gamma_{n,t} V(x_t)\right]\right\} \\
&= V^{(n)}(x,\pi) - E_x^\pi \left[V(x_n)\right] + \lim_{m\to\infty} E_x^\pi \left[\Gamma_{n,m} V(x_m)\right] \\
&= V^{(n)}(x,\pi) - E_x^\pi \left[V(x_n)\right],
\end{aligned} \tag{61}$$

where the last equality follows from (58). Finally, (61), Definition 1 and (59) yield the desired result. □

*Proof of Theorem 13* According to Lemma 14, to prove the theorem it is sufficient to show

$$\lim_{n\to\infty} E_x^{\hat{\pi}}\left[\Phi\left(x_n, a_n\right)\right] = 0.$$

To this end, for each $n \in \mathbb{N}$, we define the function $\Phi_n : \mathbb{K}_A \to \mathbb{R}$ as

$$\Phi_n(x,a) := c(x,a) + \alpha(x,a,\hat{\theta}_n)\int_{\mathbb{X}} V_{n-1}(y)\,Q(\mathrm{d}y|x,a) - V_n(x).$$

Thus, from (53)

$$\Phi_n(x, f_n) = 0, \quad \forall n \in \mathbb{N},\ x \in \mathbb{X}. \tag{62}$$

Now, if $\{(x_n, a_n)\}$ is the sequence of state–action pairs corresponding to the application of the strategy $\hat{\pi}$, observe that from (62)

$$\begin{aligned}\Phi\left(x_n, a_n\right) &\le \left|\Phi\left(x_n, a_n\right) - \Phi_n\left(x_n, a_n\right)\right| \\ &\le \sup_{a\in A(x_n)} \left|\Phi(x_n,a) - \Phi_n(x_n,a)\right| \\ &\le W(x_n)\mathcal{L}_n \ \text{ a.s.},\end{aligned}$$

where

$$\mathcal{L}_n := \sup_{(x,a)\in\mathbb{K}} \frac{\left|\Phi(x,a) - \Phi_n(x,a)\right|}{W(x)}.$$

Hence, the remainder of the proof consists in proving

$$\lim_{n\to\infty} E_x^{\hat{\pi}}\left(W(x_n)\mathcal{L}_n\right) = 0. \tag{63}$$

By adding and subtracting the term $\alpha\left(x, a, \hat{\theta}_n\right)\int_{\mathbb{X}} V\left(y\right) Q(\mathrm{d}y|x,a)$, using (25), (26), and (48), it is easy to see that, for each $n \in \mathbb{N}$,

$$\begin{aligned}\mathcal{L}_n \le \frac{bc_0}{1-b\alpha^*} &\sup_{(x,a)\in\mathbb{K}} \left|\alpha(x,a,\theta) - \alpha(x,a,\hat{\theta}_n)\right| \\ &+ b\alpha^*\left\|V - V_{n-1}\right\|_W + \left\|V - V_n\right\|_W.\end{aligned}$$

Thus, from Lemma 8 and Proposition 10

$$\mathcal{L}_n \to 0 \ \text{ a.s., as } n \to \infty.$$

Note that from the property (69) below we also have that

$$\mathcal{L}_n \to 0 \quad P_x^{\hat{\pi}} - \text{a.s., as } n \to \infty. \tag{64}$$

Furthermore, observe that $\sup_{n\in\mathbb{N}} \mathcal{L}_n \leq L_1 < \infty$ for some constant $L_1$, and from (64) we have the convergence in probability

$$\mathcal{L}_n \xrightarrow{P_x^{\hat{\pi}}} 0 \quad \text{as} \quad n \to \infty. \tag{65}$$

Also, from (56)

$$\sup_n E_x^{\hat{\pi}} \left(W(x_n)\mathcal{L}_n\right)^p \leq L_1^p \sup_n E_x^{\hat{\pi}} \left(W^p(x_n)\right) < \infty.$$

This implies (see, for instance, Lemma 7.6.9 in Ash 1972)) that $\{W(x_n)\mathcal{L}_n\}$ is $P_x^{\hat{\pi}}$-uniformly integrable.

On the other hand, for arbitrary positive numbers $m_1$ and $m_2$, we have,

$$P_x^{\hat{\pi}} \left(W(x_n)\mathcal{L}_n > m_1\right) \leq P_x^{\hat{\pi}} \left(\mathcal{L}_n > \frac{m_1}{m_2}\right) + P_x^{\hat{\pi}} \left(W(x_n) > m_2\right),$$

which implies, from Chebyshev's inequality, that

$$P_x^{\hat{\pi}} \left(W(x_n)\mathcal{L}_n > m_1\right) \leq P_x^{\hat{\pi}} \left(\mathcal{L}_n > \frac{m_1}{m_2}\right) + \frac{E_x^{\hat{\pi}} \left(W(x_n)\right)}{m_2}.$$

This relation, together with (65), yields the convergence in probability

$$W(x_n)\mathcal{L}_n \xrightarrow{P_x^{\hat{\pi}}} 0 \quad \text{as } n \to \infty.$$

Therefore, since $\{W(x_n)\mathcal{L}_n\}$ is $P_x^{\hat{\pi}}$ -uniformly integrable, we obtain (63) which implies the asymptotic optimality of the strategy $\hat{\pi}$. □

### 6.3 Sufficiency of Markov strategies

We conclude proving the relations (15) and (17). To this end, we will use the following well-known properties of the probability measure $P_x^{\pi}$ (see, e.g., Hernández-Lerma and Lasserre 1996, 1999). For each $\pi = \{\pi_n\} \in \Pi_{\mathbb{A}}$ and $x \in \mathbb{X}$,

$$P_x^{\pi} \left[x_0 \in X\right] = \delta_x(X), \quad X \in \mathcal{B}(\mathbb{X}); \tag{66}$$

$$P_x^{\pi} \left[a_n \in A | h_n\right] = \pi_n \left(A | h_n\right), \quad A \in \mathcal{B}(\mathbb{A}); \tag{67}$$

$$P_x^{\pi} \left[x_{n+1} \in X | h_n, a_n, \xi_{n+1}\right] = Q \left(X | x_n, a_n\right), \quad X \in \mathcal{B}(\mathbb{X}); \tag{68}$$

$$P_x^{\pi} \left[\xi_{n+1} \in B | h_n, a_n\right] = \theta(B), \quad B \in \mathcal{B}(S), \tag{69}$$

where $\delta_x$ is the Dirac measure concentrated at $x$ and $h_n = (x_0, a_0, \xi_1, \ldots, x_{n-1}, a_{n-1}, \xi_n, x_n) \in \mathbb{H}_n := (\mathbb{K}_A \times S)^n \times \mathbb{X}$ for $n \in \mathbb{N}$.

**Lemma 15** *For each $\pi \in \Pi_{\mathbb{A}}$ there exists $\varphi \in \Pi_{\mathbb{A}}^{M}$ such that*

$$V(x,\pi) = V(x,\varphi), \quad x \in \mathbb{X}.$$

*Proof* For each $\pi \in \Pi_{\mathbb{A}}$, $x \in \mathbb{X}$, and $n \in \mathbb{N}_0$, we define the finite measures $M_{x,n}^{\pi}$ on $\mathbb{X} \times \mathbb{A}$ and $m_{x,n}^{\pi}$ on $\mathbb{X}$ as

$$M_{x,n}^{\pi}(K) := E_x^{\pi} \tilde{\Gamma}_n 1_{\{(x_n,a_n)\in K\}}, \quad K \in \mathcal{B}(\mathbb{X} \times \mathbb{A}) \tag{70}$$

and

$$m_{x,n}^{\pi}(X) := E_x^{\pi} \tilde{\Gamma}_n 1_{\{x_n \in X\}}, \quad X \in \mathcal{B}(\mathbb{X}). \tag{71}$$

Observe that $m_{x,n}^{\pi}$ is the marginal of $M_{x,n}^{\pi}$ on $\mathbb{X}$. Then, by Corollary 7.27.2 in Bertsekas and Shreve (1978), there exists a stochastic kernel $\varphi_n$ on $\mathbb{A}$ given $\mathbb{X}$ such that, for $X \in \mathcal{B}(\mathbb{X})$ and $A \in \mathcal{B}(\mathbb{A})$,

$$M_{x,n}^{\pi}(X \times A) = \int_X \varphi_n(A|y) m_{x,n}^{\pi}(\mathrm{d}y) = \int_X \int_A \varphi_n(\mathrm{d}a|y) m_{x,n}^{\pi}(\mathrm{d}y). \tag{72}$$

Since $M_{x,n}^{\pi}$ is concentrated on $\mathbb{K}_A$, we can select versions of $\varphi_n$, $n \in \mathbb{N}_0$, such that $\varphi_n(A(y)|y) = 1$, for $y \in \mathbb{X}$. Thus, $\varphi := \{\varphi_n\} \in \Pi_{\mathbb{A}}^{M}$. Therefore, to prove the lemma, it is enough to prove the equality

$$M_{x,n}^{\pi} = M_{x,n}^{\varphi}, \quad x \in \mathbb{X}, \ n \in \mathbb{N}_0. \tag{73}$$

Indeed, first note that from (70) and applying standard arguments on integration theory as linearity and the monotone convergence theorem, we can obtain

$$E_x^{\pi} \tilde{\Gamma}_n g(x_n, a_n) = \int_{\mathbb{X}\times\mathbb{A}} g(y,a) M_{x,n}^{\pi}(\mathrm{d}(y,a)), \tag{74}$$

for any measurable function $g : \mathbb{X} \times \mathbb{A} \to \Re$. Then, from (73) and (74) with $g = c$, we get

$$E_x^{\pi} \tilde{\Gamma}_n c(x_n, a_n) = \int_{\mathbb{X}\times\mathbb{A}} c(y,a) M_{x,n}^{\varphi}(\mathrm{d}(y,a)) = E_x^{\varphi} \tilde{\Gamma}_n c(x_n, a_n),$$

which, from (13) proves the lemma.

We then proceed to prove (73) by induction. First observe that from ( 66)

$$m_{x,0}^{\pi}(X) := E_x^{\pi} 1_{\{x_0 \in X\}} = \delta_x(X) = m_{x,0}^{\varphi}(X), \quad X \in \mathcal{B}(\mathbb{X}).$$

Then, from (70) and (72),

$$\begin{aligned} M_{x,0}^{\pi}(X \times A) &:= E_x^{\pi} 1_{\{(x_0,a_0)\in X\times A\}} = \int_X \varphi_0(A|y) m_{x,0}^{\pi}(\mathrm{d}y) \\ &= \int_X \varphi_0(A|y) m_{x,0}^{\varphi}(\mathrm{d}y) = M_{x,0}^{\varphi}(X \times A), \quad X \times A \in \mathcal{B}(\mathbb{X} \times \mathbb{A}). \end{aligned}$$

Now we assume that (73) holds for some $n \in \mathbb{N}_0$. Then, using properties of conditional expectation, from (68) and (69),

$$\begin{aligned} m^{\pi}_{x,n+1}(X) &= E^{\pi}_x \tilde{\Gamma}_{n+1} 1_{\{x_{n+1}\in X\}} = E^{\pi}_x \left[ E^{\pi}_x \left[ \tilde{\Gamma}_n \tilde{\alpha}(x_n, a_n, \xi_{n+1}) 1_{\{x_{n+1}\in X\}} | h_n, a_n \right] \right] \\ &= E^{\pi}_x \left[ \tilde{\Gamma}_n E^{\pi}_x \left[ \tilde{\alpha}(x_n, a_n, \xi_{n+1}) 1_{\{x_{n+1}\in X\}} | h_n, a_n \right] \right] \\ &= E^{\pi}_x \left[ \tilde{\Gamma}_n Q\left(X | x_n, a_n\right) \int_S \tilde{\alpha}(x_n, a_n, s)\theta(\mathrm{d}s) \right]. \end{aligned}$$

Then, taking $g(y, a) = Q\left(X|y, a\right) \int_S \tilde{\alpha}(y, a, s)\theta(\mathrm{d}s)$, from (74) we have, for each $\pi \in \Pi_{\mathbb{A}}$,

$$\begin{aligned} m^{\pi}_{x,n+1}(X) &= \int_{\mathbb{X}\times\mathbb{A}} g(y, a) M^{\pi}_{x,n}(\mathrm{d}(y, a)) \\ &= \int_{\mathbb{X}\times\mathbb{A}} g(y, a) M^{\varphi}_{x,n}(\mathrm{d}(y, a)). \end{aligned} \tag{75}$$

In particular, letting $\pi = \varphi$, we obtain

$$m^{\varphi}_{x,n+1}(X) = \int_{\mathbb{X}\times\mathbb{A}} g(y, a) M^{\varphi}_{x,n}(\mathrm{d}(y, a)),$$

which, together with (75), yields

$$m^{\pi}_{x,n+1}(X) = m^{\varphi}_{x,n+1}(X). \tag{76}$$

Now we use this fact to prove (73). From (72) and (76 ),

$$\begin{aligned} M^{\pi}_{x,n+1}(X \times A) &= \int_X \int_A \varphi_{n+1}(\mathrm{d}a|y) m^{\pi}_{x,n+1}(\mathrm{d}y) \\ &= \int_X \int_A \varphi_{n+1}(\mathrm{d}a|y) m^{\varphi}_{x,n+1}(\mathrm{d}y). \end{aligned} \tag{77}$$

On the other hand, observe that, similar to (74), we have that for each $\pi \in \Pi_{\mathbb{A}}$ and $n \in \mathbb{N}_0$,

$$E^{\pi}_x \tilde{\Gamma}_n h(x_n) = \int_{\mathbb{X}} h(y) m^{\pi}_{x,n}(\mathrm{d}y), \tag{78}$$

for any measurable function $h : \mathbb{X} \to \Re$. Then, from ( 70) and (67),

$$\begin{aligned} M^{\varphi}_{x,n+1}(X \times A) &= E^{\varphi}_x \tilde{\Gamma}_{n+1} 1_{\{x_{n+1}\in X, a_{n+1}\in A\}} \\ &= E^{\varphi}_x \left[ E^{\varphi}_x \left[ \tilde{\Gamma}_{n+1} 1_{\{x_{n+1}\in X, a_{n+1}\in A\}} | h_{n+1} \right] \right] \end{aligned}$$

$$= E_x^{\varphi}\left[\tilde{\Gamma}_{n+1}1_{\{x_{n+1}\in X\}}E_x^{\varphi}\left[1_{\{a_{n+1}\in A\}}|h_{n+1}\right]\right]$$
$$= E_x^{\varphi}\left[\tilde{\Gamma}_{n+1}1_{\{x_{n+1}\in X\}}\int_A \varphi_{n+1}(da_{n+1}|x_{n+1})\right].$$

Taking $h(y) = 1_{\{y\in X\}}\int_A \varphi_{n+1}(\mathrm{d}a|y)$, from (78) we get

$$M^{\varphi}_{x,n+1}(X\times A) = \int_{\mathbb{X}} 1_{\{y\in X\}}\int_A \varphi_{n+1}(\mathrm{d}a|y)m^{\pi}_{x,n+1}(\mathrm{d}y)$$
$$= \int_X\int_A \varphi_{n+1}(\mathrm{d}a|y)m^{\pi}_{x,n+1}(\mathrm{d}y). \tag{79}$$

Thus, (77) and (79) yield (73), which proves the lemma. □

**Lemma 16** *For each $x\in\mathbb{X}$ and $\varphi\in\Pi^M_{\mathbb{A}}$, the relation (17) holds, that is*

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

*where $\Gamma_n = \prod_{k=0}^{n-1}\alpha(x_k,a_k,\theta)$, if $n\in\mathbb{N}$ and $\Gamma_0 = 1$.*

*Proof* The relation (17) is consequence of the properties of the probability measure $P_x^{\varphi}$ (66)–(69) and Lemma 15. Indeed, for each $x\in\mathbb{X}$ and $\varphi\in\Pi^M_{\mathbb{A}}$,

$$\begin{aligned}
E_x^{\varphi}\tilde{\alpha}(x_0,a_0,\xi_1)c(x_1,a_1) &= \iiiint_{\mathbb{A}\times S\times\mathbb{X}\times\mathbb{A}} \tilde{\alpha}\left(x_0,a_0',\xi_1'\right)c\left(x_1',a_1'\right)\\
&\quad\times\varphi_1\left(\mathrm{d}a_1'|x_1'\right)Q\left(\mathrm{d}x_1'|x_0,a_0'\right)\theta\left(\mathrm{d}\xi_1'\right)\varphi_0\left(\mathrm{d}a_0'|x_0\right)\\
&= \int_{\mathbb{A}}\int_S \tilde{\alpha}\left(x_0,a_0',\xi_1'\right)\theta\left(\mathrm{d}\xi_1'\right)\int_{\mathbb{X}}\int_{\mathbb{A}} c\left(x_1',a_1'\right)\varphi_1\left(\mathrm{d}a_1'|x_1'\right)\\
&\quad\times Q\left(\mathrm{d}x_1'|x_0,a_0'\right)\varphi_0\left(\mathrm{d}a_0'|x_0\right)\\
&= \int_{\mathbb{A}}\int_{\mathbb{X}}\int_{\mathbb{A}} \alpha\left(x_0,a_0',\theta\right)c\left(x_1',a_1'\right)\varphi_1\left(\mathrm{d}a_1'|x_1'\right)\\
&\quad\times Q\left(\mathrm{d}x_1'|x_0,a_0'\right)\varphi_0\left(\mathrm{d}a_0'|x_0\right)\\
&= E_x^{\varphi}\alpha\left(x_0,a_0,\theta\right)c(x_1,a_1). \quad \text{(see (9))}
\end{aligned}$$

Applying similar arguments, it is easy to see that, for $n = 2,3,\ldots,$

$$E_x^{\varphi}\tilde{\Gamma}_n c(x_n,a_n) = E_x^{\varphi}\Gamma_n c(x_n,a_n),$$

which implies (17). □

## 7 Example

Besides the problems with constant discount factor, our theory, additionally, covers some particular cases with non-constant discount factor, for instance, state-dependent discount factor systems, and problems with random discount factor but independent of the state–action process, see Wei and Guo (2011) and González-Hernández et al. (2008, 2013, 2014). In these works are introduced application examples corresponding to each particular case. Now we present further examples which satisfy our hypotheses.

### 7.1 A cash-balance model

We consider a simple discrete-time cash-balance model introduced in Hordjik and Yushkevich (1999) (see also Wei and Guo 2011). The problem consists in to control the level of a firm's cash balance to meet its demand for cash at minimum total discounted cost.

We define the following variables:

$x_n$ is the cash balance at time $n$;
$a_n$ is the withdrawal of size $-a_n$ (if $a_n < 0$) of money in cash, or a supply in amount $a_n$ (if $a_a > 0$), at time $n$;
$w_n$ is the demand for cash during the stage $n$. A positive demand means cash outflow and a negative demand means a cash inflow.

Then, the cash-balance process $\{x_n\}$ evolves on the state space $\mathbb{X} = \mathbb{R}$ according to the recursive equation

$$x_{n+1} = x_n + a_n + w_n, \quad n = 0, 1, \ldots \tag{80}$$

We assume that $\{w_n\}$ is a sequence of i.i.d. random variables with standard normal distribution

$$\rho(w) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right), \quad w \in \mathbb{R};$$

and furthermore the set of admissible actions when the cash balance is $x \in \mathbb{R}$ is $A(x) = [-|x|, |x|]$. The one-stage cost function is an arbitrary l.s.c. on

$$\mathbb{K}_A := \{(x, a) : x \in \mathbb{R}, a \in [-|x|, |x|]\}$$

such that

$$|c(x, a)| \leq c_0(x^2 + 1),$$

for some positive constant $c_0$.

Under these conditions, clearly Assumptions 2(a) and (d) are satisfied with

$$W(x) := x^2 + 1, \quad x \in \mathbb{R}. \tag{81}$$

To verify Assumptions 2(b) and (c), let us, first, note that the transition law (6) takes the form

$$Q(D|x,a) = \int_{\mathbb{R}} 1_D[x+a+w]\rho(w)\mathrm{d}w$$
$$= \int_{\mathbb{R}} 1_D[x+a+w]\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right)\mathrm{d}w, \quad D \in \mathcal{B}(\mathbb{R}). \tag{82}$$

Hence, from the continuity of the density $\rho$ as well as of the function $(x,a,w) \to x+a+w$, it is easy to prove that the functions

$$(x,a) \longmapsto \int_{\mathbb{R}} u[x+a+w]\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right)\mathrm{d}w$$

and

$$(x,a) \longmapsto \int_{\mathbb{R}} W[x+a+w]\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right)\mathrm{d}w$$

are continuous on $\mathbb{K}_A$, for each continuous and bounded function $u : \mathbb{R} \to \mathbb{R}$ [see, e.g., Examples C.6 and C.8, Appendix C in Hernández-Lerma and Lasserre (1996)], which yield Assumptions 2(b) and (c).

We will now proceed to verify Assumptions 2(e) and (f). Let us assume that the random disturbance process $\{\xi_n\}$ is a sequence of independent random variables taking values on $S := [0,1]$ with unknown distribution $\theta \in \Theta = \mathbb{P}(S)$. In addition, the discount factor function is defined as

$$\tilde{\alpha}(x,a,s) = \frac{s}{\gamma(x^2+a^2+1)}, \quad (x,a) \in \mathbb{K}_A, s \in S, \tag{83}$$

for a constant $\gamma > 4$. Hence, clearly

$$\alpha^* := \sup_{(x,a,s)\in\mathbb{K}_A\times S} \tilde{\alpha}(x,a,s) < \frac{1}{4}, \tag{84}$$

which implies that Assumption 2(e) holds.

Furthermore, from (81) and (82), observe that

$$\int_{\mathbb{R}} W(y)Q(\mathrm{d}y|x,a) = \int_{\mathbb{R}} [(x+a+w)^2+1]\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right)\mathrm{d}w$$
$$= \int_{\mathbb{R}} (x+a+w)^2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w^2}{2}\right)\mathrm{d}w + 1$$
$$= \int_{\mathbb{R}} y^2\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(y-(x+a)^2}{2}\right)\mathrm{d}y + 1$$
$$= (x+a)^2 + 2, \quad (x,a) \in \mathbb{K}_A.$$

Now, because $A(x) = [-|x|, |x|]$, it is easy to see that

$$\int_{\mathbb{R}} W(y)Q(\mathrm{d}y|x,a) \leq 4x^2 + 2 \leq 4(x+a)^2 = 4W(x).$$

Thus, from (84), Assumption 2(f) is satisfied with $b = 4$.

Therefore, from Theorem 5, there exists a minimax strategy.

### 7.2 An autoregressive control model

We consider a controlled process of the form

$$x_{n+1} = G(a_n)x_n + w_n, \quad n = 0, 1, \ldots,$$

$x_0$ given, with state space $\mathbb{X} = [0, \infty)$, and compact action set $A(x) = \mathbb{A} \subset \mathbb{R}$, $x \in \mathbb{X}$, and $G : \mathbb{A} \to (0, \lambda]$ is a given measurable function with $\lambda < 1$. The random disturbance process $\{w_n\}$ is formed by i.i.d. and nonnegative random variables with a continuous density $\rho$ and finite expectation, say $E[w_0] = \bar{w} < \infty$. Hence, the transition law is

$$Q(D|x,a) = \int_{\mathbb{R}} 1_D[G(a)x + w]\rho(w)\mathrm{d}w, \quad D \in \mathcal{B}(\mathbb{X}).$$

In addition, the one-stage cost $c$ is an arbitrary l.s.c. function on $\mathbb{K}_A = \{(x,a) : x \geq 0, x \in \mathbb{A}\}$ such that

$$|c(x,a)| \leq (x + \bar{b})^{1/p}, \quad (x,a) \in \mathbb{K}_A,$$

for some constants $\bar{b} > 1$ and $p > 1$.

Similarly as previous example, defining $W(x) := (x+\bar{b})^{1/p}$ we have that Assumptions 2(a)–(d) hold. Moreover, for all $(x,a) \in \mathbb{K}_A$,

$$\begin{aligned}\int_0^\infty W^p[G(a)x + w]\,\rho(w)\mathrm{d}w &= \int_0^\infty \left[G(a)x + w + \bar{b}\right]\rho(w)\mathrm{d}w \\ &\leq \lambda(x + \bar{b}) + \bar{b} + \bar{w},\end{aligned}$$

which implies Assumption 11 with $\beta_0 := \lambda$ and $d_0 := \bar{b} + \bar{w}$. From Remark 12(a), we also have

$$\begin{aligned}\int_0^\infty W[G(a)x + w]\,\rho(w)\mathrm{d}w &= \int_0^\infty \left[G(a)x + w + \bar{b}\right]^{1/p}\rho(w)\mathrm{d}w \\ &\leq \lambda^{1/p}W(x) + \left(\bar{b} + \bar{w}\right)^{1/p}, \quad (x,a) \in \mathbb{K}_A. \quad (85)\end{aligned}$$

We again assume that the discount factor function is as (83), but where the constant $\gamma$ satisfies

$$1 < \lambda^{1/p} + \left(\bar{b} + \bar{w}\right)^{1/p} = \beta_0^{1/p} + d_0^{1/p} < \gamma.$$

In addition, $\xi_n$, $n \geq 0$, are i.i.d. random variables taking values on $S = [0, 1]$ with unknown distribution $\theta \in \mathbb{P}(S)$. Then, under these conditions, we have

$$\alpha^* < \frac{1}{\lambda^{1/p} + \left(\bar{b} + \bar{w}\right)^{1/p}} < 1, \tag{86}$$

and, because $W(\cdot) \geq 1$, from (85) we get

$$\int_0^\infty W\left[G(a)x + w\right] \rho(w) \mathrm{d}w \leq \left(\lambda^{1/p} + \left(\bar{b} + \bar{w}\right)^{1/p}\right) W(x), \quad (x, a) \in \mathbb{K}_A. \tag{87}$$

Therefore, defining $b := (\lambda^{1/p} + \left(\bar{b} + \bar{w}\right)^{1/p})$, (86) and (87) yield Assumptions 2(e) and (f).

Finally, observe that the derivative of $\tilde{\alpha}$ with respect to $s$ satisfies

$$\tilde{\alpha}'(x, a, s) = \frac{1}{\gamma(x^2 + a^2 + 1)} < 1, \quad \forall (x, a) \in \mathbb{K}_A, s \in [0, 1].$$

This fact implies that the family of functions

$$\mathcal{A} := \{\tilde{\alpha}(x, a, \cdot) : (x, a) \in \mathbb{K}_A\}$$

is equi-Lipschitz, and therefore equicontinuous on $S$. Then Assumption 7 is satisfied.

Therefore, from Theorem 13, there exists an asymptotically optimal strategy.

**Acknowledgments** Work supported by Consejo Nacional de Ciencia y Tecnología (CONACYT, MEXICO) under Grant CB2010/154612.

## References

Altman E (1999) Constrained Markov decision processes. Chapman and Hall, London

Ash RB (1972) Real analysis and probability. Academic Press, New York

Bertsekas DP, Shreve SE (1978) Stochastic optimal control: the discrete-time case. Academic Press, New York

Bertsekas DP (1987) Dynamic programming: deterministic and stochastic models. Prentice-Hall, Englewood Cliffs

Borkar VS (1998) A convex analytic approach to Markov decision processes. Probab Theory Relat Fields 78:583–602

Brigo D, Mercurio F (2007) Interest rate models: theory and practice. Springer, New York

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

Dynkin EB, Yushkevich AA (1979) Controlled Markov processes. Springer, New York

Feinberg EA, Shwartz A (1994) Markov decision models with weighted discounted criteria. Math Oper Res 19:152–168

Feinberg EA, Shwartz A (1995) Constrained Markov decision models with weighted discounted rewards. Math Oper Res 20:302–320

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

González-Hernández J, López-Martínez RR, Pérez-Hernández R (2007) Markov control processes with randomized discounted cost in Borel space. Math Methods Oper Res 65:27–44

González-Hernández J, López-Martínez RR, Minjárez-Sosa JA (2008) Adaptive policies for stochastic systems under a randomized discounted criterion. Bol Soc Mat Mex 14:149–163
González-Hernández J, López-Martínez RR, Minjárez-Sosa JA (2009) Approximation, estimation and control of stochastic systems under a randomized discounted cost criterion. Kybernetika 45:737–754
González-Hernández J, López-Martínez RR, Minjárez-Sosa JA, Gabriel-Arguelles JR (2013) Constrained Markov control processes with randomized discounted cost criteria: occupation measures and extremal points. Risk Decis Anal 4:163–176
González-Hernández J, López-Martínez RR, Minjárez-Sosa JA, Gabriel-Arguelles JR (2014) Constrained Markov control processes with randomized discounted rate: infinite linear programming approach. Optim Control Appl Methods 35:575–591
González-Trejo TJ, Hernández-Lerma O, Hoyos-Reyes LF (2003) Minimax control of discrete-time stochastic systems. SIAM J Control Optim 41:1626–1659
Gordienko EI, Minjárez-Sosa JA (1998) Adaptive control for discrete-time Markov processes with unbounded costs: discounted criterion. Kybernetika 34:217–234
Heath D, Jarrow R, Morton A (1992) Bond pricing and the term structure of interest rates: a new methodology. Econometrica 60:77–105
Hernández-Lerma O (1989) Adaptive Markov control processes. Springer, New York
Hernández-Lerma O, Lasserre JB (1996) Discrete-time Markov control processes: basic optimality criteria. Springer, New York
Hernández-Lerma O, Lasserre JB (1999) Further topics on discrete-time Markov control processes. Springer, New York
Hernández-Lerma O, González-Hernández J (2000) Constrained Markov control processes in Borel spaces: the discounted case. Math Methods Oper Res 52:271–285
Hilgert N, Minjárez-Sosa JA (2001) Adaptive policies for time-varying stochastic systems under discounted criterion. Math Methods Oper Res 54:491–505
Hilgert N, Minjárez-Sosa JA (2006) Adaptive control of stochastic systems with unknown disturbance distribution: discounted criteria. Math Methods Oper Res 63:443–460
Hinderer K (1979) Foundations of non-stationary dynamic programming with discrete time parameter. In: Lecture Notes Oper. Res., vol 33. Springer, New York
Hordjik A, Yushkevich AA (1999) Blackwell optimality in the class of all policies in Markov decision chains with Borel state space an unbounded rewards. Math Methods Oper Res 50:421–448
Iyengar GN (2005) Robust dynamic programming. Math Oper Res 30:257–280
Jaskiewicz A, Nowak AS (2011) Stochastic games with unbounded payoffs: applications to robust control in economics. Dyn Games Appl 1:253–279
López-Martínez RR, Hernández-Lerma O (2003) The Lagrange approach to constrained Markov processes: a survey and extension of results. Morfismos 7:1–26
Mandl P (1974) Estimation and control in Markov chains. Adv Appl Probab 6:40–60
Piunovskiy AB (1997) Optimal control of random sequences in problems with constraints. Kluwer, Dordrecht
Puterman ML (1994) Markov decision processes. In: Discrete stochastic dynamic programming. Wiley, New York
Ranga Rao R (1962) Relations between weak and uniform convergence of measures with applications. Ann Math Stat 33:659–680
Rieder U (1978) Measurable selection theorems for optimization problems. Manuscripta Math 24:115–131
Schäl M (1975) Conditions for optimality and for the limit on $n$-stage optimal policies to be optimal. Z Wahrs Verw Gebiete 32:179–196
Vasicek O (1977) An equilibrium characterisation of the term structure. J Financ Econ 5:177–180
Schäl M (1987) Estimation and control in discounted stochastic dynamic programming. Stochastics 20:51–71
Wei Q, Guo X (2011) Markov decision processes with state-dependent discount factors and unbounded rewards/costs. Oper Res Lett 39:274–368