Hindawi
Discrete Dynamics in Nature and Society
Volume 2022, Article ID 1080946, 11 pages
https://doi.org/10.1155/2022/1080946

Research Article

Asymptotic Optimality and Rates of Convergence of Quantized Stationary Policies in Continuous-Time Markov Decision Processes

Xiao Wu and Yanqiu Tang

School of Mathematics and Statistics, Zhaoqing University, Zhaoqing 526061, China

Correspondence should be addressed to Xiao Wu; jxwuxiao@126.com

Received 15 October 2021; Accepted 25 July 2022; Published 9 September 2022

Academic Editor: Luca Pancioni

This paper is concerned with the asymptotic optimality of quantized stationary policies for continuous-time Markov decision processes (CTMDPs) in Polish spaces with state-dependent discount factors, where the transition rates and reward rates are allowed to be unbounded. Using the dynamic programming approach, we first establish the discounted optimal equation and the existence of its solutions. Then, we obtain the existence of optimal deterministic stationary policies under suitable conditions by more concise proofs. Furthermore, we discretize and incentivize the action space and construct a sequence of quantizer policies, which is the approximation of the optimal stationary policies of the CTMDPs, and get the approximation result and the rates of convergence on the expected discounted rewards of the quantized stationary policies. Also, we give an iteration algorithm on the approximate optimal policies. Finally, we give an example to illustrate the asymptotic optimality.

1. Introduction

This paper deals with the infinite horizon discounted continuous-time Markov decision processes (CTMDPs), as well as studies the asymptotic optimality of quantized stationary policies of CTMDPs, and gives the convergence rate results. The discount factors are state-dependent, and the transition rates and reward rates are allowed to be unbounded.

It is well-known that the discounted CTMDPs have been widely studied as an important class of stochastic control problems. Generally speaking, according to the various forms of discount factors, the infinite horizon discounted CTMDPs can be classified into the following three groups: (i) MDPs with a fixed constant discount factor α, see, for instance, Doshi [1], Dynkin and Yushkevich [2], Feinberg [3], Guo [4, 5], Guo and Song [6], Guo and Hernndez-Lerma [7], Hernndez-Lerma and Lasserre [8, 9], Puterman [10], and the references therein; (ii) MDPs with varying (state-dependent or state-action dependent) discount factors, for instance, see Feinberg and Shwartz [11], Gonzlez-Hernndez et al. [12], Wu and Guo [13], Wu and Zhang [14], and the references therein; (iii) MDPs whose the discount factor is a function of the history, see Hinderer [15], for example. This paper will study the infinite horizon discounted CTMDPs in the case of the group.

For the discounted criterion of MDPs, there are many works on the existence of solutions to the discounted optimality equation and of discounted optimal stationary policies, see, for instance, [1, 4, 6, 7, 16] for the CTMDPs and [8–10, 13–15] for the discrete-time Markov decision processes (DTMDPs). These references, however, are on the discounted MDPs with a constant discount factor or the discounted DTMDPs with varying discount factors. Recently, the discounted CDMDPs with state-dependent discount factors are studied in [16], in which the authors established the discounted reward optimality equation (DROE) and obtained the existence of discounted optimal stationary policies. However, in [16], the discussion is restricted to the class of all randomized stationary policies (i.e., the policies are time-independent). Following these ideas,

still within the discounted continuous-time MDPs, models with Polish spaces are studied in this paper. We will extend some results in [16] to the case of all randomized Markov policies and obtain the existence of discounted optimal stationary policies by more concise proof.

Although the existence of the optimal policies is proved, it is difficult to compute an optimal policy even in the stationary policies class for nonfinite Polish (i.e., complete and separable metric) state and action spaces. Furthermore, in applications to networked control, the transmission of such control actions to an actuator is not realistic when there is an information transmission constraint (imposed by the presence of a communication channel) between a plant, a controller, or an actuator. Thus, from a practical point of view, it is important to study the approximation of optimal stationary policies. Several approaches have been developed in the literature to solve this problem for finite or countable state spaces, see [17–20]. Lately, for infinite Borel state and action spaces, [21, 22] give the asymptotic optimality of quantized stationary policies in stochastic control for DTMDPs. Inspired by these, in this paper, we are concerned with the asymptotic optimality of quantized stationary policies in CTMDPs with Polish spaces. To the best of our knowledge, the corresponding asymptotic optimality for CTMDPs with varying (state-dependent) discount factors has not been studied.

Therefore, this paper contains the following three main contributions:

(a) For the CDMDPs with state-dependent discount factors, we extend some results in [16] to the case of all randomized Markov policies, and the proof of the existence of discounted optimal stationary policies is simplified under mild conditions and gives an algorithm to get ε-optimal policies.

(b) We obtain that the deterministic stationary quantizer policies are able to approximate the optimal deterministic stationary policies under mild technical conditions and thus show that one can search for approximate optimal policies within the class of quantized control policies.

(c) For the asymptotic optimality, we give the corresponding convergence rates results.

This paper is organized as follows. In section 2, we introduce the models of CDMDPs with the expected discounted reward criterion and state the discounted optimality problem. In section 3, under suitable conditions, we prove the main result on the existence of the solutions to the discounted optimal equation (DOE) and the existence of optimal stationary policies. In section 4, we give an iteration algorithm on the ε-optimal policies. In section 5, we establish conditions under which quantized control policies are asymptotically optimal and give the corresponding convergence rate results and the rates of convergence on the expected discounted rewards of the quantized stationary policies. Finally, we illustrate the asymptotic optimality by an example in Section 6.

2. The Markov Decision Processes and Discounted Optimal Problem

Consider the model of continuous-time Markov decision processes $\mathscr{M}$ as follows:

$$\mathscr{M} \coloneqq \{S,\ (A(x) \subseteq A,\ x \in S),\ q(\bullet|x,a),\ \alpha(x), r(x,a)\}, \quad (1)$$

where S is the state space, $A(x)$ are sets of admissible actions, and A is a compact action space. S and A are assumed to be Polish spaces (i.e., complete and separable metric spaces) with Borel σ-field $\mathscr{B}(S)$ and $\mathscr{B}(A)$, respectively. $A(x)$ and $K \coloneqq \{(x,a)|x \in S, a \in A(x)\}$ are Borel subsets of A and $S \times A$, respectively. $q(\bullet|x,a)$ denotes the function of transition rates, which satisfy the following properties:

(P1) $q(\bullet|x,a)$ is a signed measure on $\mathscr{B}(S)$ for each fixed $(x,a) \in K$, and $q(B|\bullet,\bullet)$ is a Borel-measurable function on K for each fixed $B \in \mathscr{B}(S)$

(P2) $0 \le q(B|x,a) < \infty$ for all $(x,a) \in K$ and $x \notin B \in \mathscr{B}(S)$

(P3) $q(\bullet|x,a)$ is conservative, that is, $q(S|x,a) = 0$ for all $(x,a) \in K$, and then, $0 \le -q(\{x\}|x,a) < \infty$

(P4) the model in (1) is supposed to be stable, that is, for each $x \in S$, it holds that

$$q^*(x) \coloneqq \sup_{a \in A(x)} \{-q(\{x\}|x,a)\} < \infty. \quad (2)$$

The discount factors $\alpha(x)$ are the nonnegative measurable functions on S. Finally, the reward rate function $r(x,a)$ is assumed to be Borel-measurable on K. Note that, $r(x,a)$ is allowed to be unbounded from both above and below, and it can be regarded as a cost rate rather than a reward rate only.

The definitions of the randomized Markov policy $\coloneqq (\pi_t, t \ge 0)$, randomized stationary policy φ, and (deterministic) stationary policy f are given by [8] [Definitions 2.2.3 and 2.3.2]. The sets of all randomized Markov policies, randomized stationary policies, and (deterministic) stationary policies are denoted by Π, Φ, and F, respectively. It is clear that $F \subset \Phi \subset \Pi$, and for each $\pi = (\pi_t, t \ge 0) \in \Pi$, $x \in S$, and $B \in \mathscr{B}(S)$, we define the associated functions of transition rates $q_\pi(\bullet|x,\pi_t)$ and reward rates $r_\pi(x,\pi_t)$ by

$$\begin{aligned} q_\pi(B|x,\pi_t) &\coloneqq \int_{A(x)} q(B|x,a)\pi_t(\mathrm{d}a|x), r_\pi(x,\pi_t) \\ &\coloneqq \int_{A(x)} r(x,a)\pi_t(\mathrm{d}a|x). \end{aligned} \quad (3)$$

In general, we also write as $q(B|x,\pi_t)$ and $r(x,\pi_t)$, respectively. Furthermore, for each $\varphi \in \Phi$, we define the functions of transition rates and reward rates by

$$\begin{aligned} q(B|x,\varphi) &:= \int_{A(x)} q(B|x,a)\pi_t(\mathrm{d}a|x), r_\pi(x,\varphi) \\ &:= \int_{A(x)} r(x,a)\varphi(\mathrm{d}a|x). \end{aligned} \tag{4}$$

In particular, we write them as $q(B|x,f)$ and $r(x,f)$, respectively, when $\varphi = f \in F$, that is, $q(B|x,f) = q(B|x,f(x))$ and $r(x,f) = r(x,f(x))$. Also, for each $\in S$, we denote

$$q(x,f) := -q(\{x\}|x,f(x)). \tag{5}$$

For any fixed policy $\pi = (\pi_t, t \geq 0) \in \Pi$, $q(\bullet|x,\pi_t)$ is also called an infinitesimal generator (see Doshi [1]). As is well known, any transition function $\tilde{p}_\pi(s,x,t,B)$ depending on π such that

$$\lim_{\tau \longrightarrow 0^+} \frac{\tilde{p}_\pi(t,x,t+\tau,B) - \delta_x(B)}{\tau} = q(B|x,\pi_t). \tag{6}$$

for all $x \in S$ and $B \in \mathscr{B}(S)$ is called a Q-process with transition rates $q(\bullet|x,\pi_t)$, where $\delta_x(B)$ is the Dirac measure at $x \in S$. By Guo [4], there exists a minimal Q-process $p_\pi^{\min}(s,x,t,B)$ with transition rates $q(\cdot|x,\pi_t)$, but such a Q-process might not be regular, that is, there may exist $p_\pi^{\min}(s,x,t,S) < 1$ for some $x \in S$ and $t \geq s \geq 0$. To ensure the regularity of the Q-process, we propose the following "drift conditions."

Assumption 1. There exists a measurable function $w_1 \geq 1$ on S, and constants $c_1 \neq 0, b_1 \geq 0$ and $M_q > 0$ such that

(a) for all $(x,a) \in K$, $\int_S w_1(y)q(\mathrm{d}y|x,a) \leq c_1 w_1(x) + b_1$

(b) For each $x \in S$, $q^*(x) \leq M_q w_1(x)$

Remark 1.

(a) The function w in Assumption 1 (a) is used to guarantee the finiteness of the optimal value function as below, and by [4] [Remark 2.2(b)], it is an extension of the "drift condition" in Lund et al. [23] for a time-homogeneous Q-process. Moreover, Assumption 1 (b) is used to guarantee the regularity of the Q-process, and it is not required when the transition rates are bounded (i.e., $\sup_{x\in S} q^*(x) < \infty$).

(b) Under Assumption 1, it holds that $p_\pi^{\min}(s,x,t,S) \equiv 1$ by Guo [4] [Theorem 3.2]. Then, the Q-process with transition rates $q(\bullet|x,\pi_t)$ is regular and unique. Hence, we write $p_\pi^{\min}(s,x,t,B)$ simply as $p_\pi(s,x,t,B)$. Since it is time-homogeneous, we discuss the case that the initial time is $s = 0$, and then, we write $p_\pi(0,x,t,B)$ simply as $p_\pi(x,t,B)$.

As it is well known (e.g., see Doshi [1] and Guo [5]), for each $\pi = (\pi_t, t \geq 0) \in \Pi$ and the initial state $x \in S$, there exists a unique probability space $(\Omega, \mathscr{B}(\Omega), P_x^\pi)$, where the probability measure P_x^π is completely determined by $p_\pi(x,t,B)$ (see Guo [6], Section 2.3]), and a state and action process $\{x(t), a(t), t \geq 0\}$ with the transition probability function $p_\pi(x,t,B)$ such that (see Guo [5], Lemma 2.1])

$$\begin{aligned} P_x^\pi(x(t) \in B) &= p_\pi(x,t,B), P_x^\pi(a(t) \in \Gamma | x(t) = j) \\ &= \pi_t(\Gamma|j), \quad \forall \Gamma \in \mathscr{B}(A). \end{aligned} \tag{7}$$

The expectation operator corresponding to P_x^π can be denoted by E_x^π. Moreover, for each $x \in S, \pi \in \Pi$ and $t \geq 0$, the expected reward is given by

$$E_x^\pi r(x(t),\pi_t) := \int_S r(y,\pi_t)p_\pi(x,t,\mathrm{d}y). \tag{8}$$

Now, we state the discounted optimality problem. For each $\pi \in \Pi$ and $x \in S$, the expected discounted reward criterion is defined as

$$J(x,\pi) := E_x^\pi \left[\int_0^\infty e^{-\int_0^t \alpha(x(s))\mathrm{d}s} r(x(t),\pi_t)\mathrm{d}t \right], \tag{9}$$

and the corresponding optimal value function is given by

$$J^*(x) := \sup_{\pi\in\Pi} J(x,\pi). \tag{10}$$

Also, a policy $\pi^* \in \Pi$ is called optimal policy if $J(x,\pi^*) \geq J(x,\pi)$ for all $x \in S$ and $\pi \in \Pi$. Our main aim in Section 3 is to give conditions for the existence of optimal deterministic stationary policies.

3. The Existence of Optimal Stationary Policies

In this section, the existence and uniqueness of the discounted optimal equation (DOE) are shown, and the existence of the optimal policies is given for the CTMDPs M defined in (1).

Note that, for any given measurable function $w \geq 1$ on S, a function v on S is called $w-$ bounded if the $w-$ weighted norm $\|v\|_w := \sup_{x\in S}|v(x)/w(x)|$ is finite. Such a function w is called a weight function. It is clear that $\mathscr{B}_w(S) := \{v: \|v\|_w < \infty\}$ is a Banach space for all real-valued measurable functions v on S. To guarantee the finiteness of the optimal value function, we need the following assumptions.

Assumption 2. Let w_1 and c_1 be as in Assumption 1. For each $x \in S$, suppose that the following conditions hold:

(a) $A(x)$ is a compact set

(b) The function $r(x,a)$ is continuous on $a \in A(x)$, and for each $(x,a) \in K$, there exists a constant $M_1 > 0$ such that $|r(x,a)| \leq M_1 w_1(x)$

(c) The discount factor $\alpha(x)$ is continuous on S, and there is a constant $\alpha_0 > c_1$ such that $\alpha(x) \geq \alpha_0 > c_1$

(d) For any bounded measurable function $u(x)$ on S, the functions $\int_S u(y)q(\mathrm{d}y|x,a)$ and $\int_S w_1(y)q(dy\ |x,a)$ are continuous on $a \in A(x)$

(e) There exists a nonnegative measurable function $w_2(x)$ on S, and constants $c_2 > 0, b_2 \geq 0$ and $M_2 > 0$ such that $q^*(x)w_1(x) \leq M_2 w_2(x)$ and $\int_S u(y)q(\mathrm{d}y|x,a) \leq c_2 w_2(x) + b_2$ for all $x \in S$ and $a \in A(x)$

For each $x \in S$, let $m(x)$ be any positive measurable function on S such that $m(x) \geq q^*(x) \geq 0$, and

$$P(B|x,a) := \frac{q(B|x,a)}{m(x)} + \delta_x(B), \quad \forall B \in \mathscr{B}(S), (x,a) \in K. \tag{11}$$

where $\delta_x(B)$ is Dirac measure (i.e., it is equal to 1 if $x \in B$ and 0 otherwise). It is clear that $P(\cdot|x,a)$ is a probability measure on S for each $(x,a) \in K$. For any $u \in \mathscr{B}_{w_1}(S)$, define an operator T on $\mathscr{B}_{w_1}(S)$ as

$$Tu(x) := \sup_{a \in A(x)} \left\{ \frac{r(x,a)}{\alpha(x) + m(x)} + \frac{m(x)}{\alpha(x) + m(x)} \int_S u(y)P(\mathrm{d}y|x,a) \right\}, \quad \forall x \in S. \tag{12}$$

And, define a recursive sequence $\{u_n, n \geq 0\}$ as

$$u_0(x) := -\frac{M_1 b_1}{\alpha_0(\alpha_0 - c_1)} - \frac{M_1 w_1(x)}{\alpha_0 - c_1}, \quad u_n(x) := Tu_{n-1}(x). \tag{13}$$

Now, we give the discounted optimal equation (DOE).

Theorem 1. *Under Assumptions 1 and 2 (b)-(c), the following assertions hold.*

(a) $|J(x,\pi)| \leq b_1 M_1/\alpha_0(\alpha_0 - c_1) + M_1/\alpha_0 - c_1 w_1(x)$ *for all* $x \in S$ *and* $\pi \in \Pi$, *and* $J(\cdot,\pi) \in \mathscr{B}_{w_1}(S)$

(b) Let $u^* := lim_{n \longrightarrow \infty} u_n$, *then we have* $u^* \in \mathscr{B}_{w_1}(S)$, *and it is the solution of the following discounted optimal equation (DOE):*

$$\alpha(x)u(x) = \sup_{a \in A(x)} \left\{ r(x,a) + \int_S u(y)q(\mathrm{d}y|x,a) \right\}, \quad \forall x \in S. \tag{14}$$

Proof. (a) By the assumptions, we have

$$\begin{aligned} |J(x,\pi)| &= \left| E_x^\pi \left[\int_0^\infty e^{-\int_0^t \alpha(x(s))\mathrm{d}s} r(x(t),\pi_t)\mathrm{d}t \right] \right| \leq \int_0^\infty \left| E_x^\pi e^{-\alpha_0 t} r(x(t),\pi_t) \right| \mathrm{d}t, \\ &\leq M_1 \int_0^\infty e^{-\alpha_0 t} E_x^\pi [w_1(x(t))]\mathrm{d}t \leq M_1 \int_0^\infty e^{-\alpha_0 t} \left[e^{c_1 t} w_1(x) + \frac{b_1}{c_1}\left(e^{c_1 t} - 1\right) \right] \mathrm{d}t, \\ &= \frac{b_1 M_1}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1}{\alpha_0 - c_1} w_1(x), \end{aligned} \tag{15}$$

where the last inequality holds by Guo [4] [Theorem 3.2(b)]. Then, $J(x,\pi) \in \mathscr{B}_{w_1}(S)$ for each $x \in S$, and part (a) holds.

(b) First, we obtain $\{u_n\}$ is monotone and nondecreasing by a similar calculation as in Ye and Guo [16] [Equation (15)]. Furthermore, it is clear that the operator T is monotone and nondecreasing. Then, we have $\{u_n\}$ is monotone and nondecreasing, which yields that $u^* := \lim_{n \longrightarrow \infty} u_n \geq u_n$ for all $n \geq 0$.

Next, we show that $u^* \in \mathscr{B}_{w_1}(S)$. Note that $w_1(x) \geq 1$ by Assumption 1, which yields that

$$|u_0(x)| \leq \frac{M_1 b_1}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1 w_1(x)}{\alpha_0 - c_1} \leq \frac{M_1(b_1 + \alpha_0)}{\alpha_0(\alpha_0 - c_1)} w_1(x). \tag{16}$$

Then, by induction argument, for all $n \geq 1$, we have

$$\begin{aligned} |u_n(x)| &\leq \sup_{a \in A(x)} \left\{ \frac{M_1 w_1(x)}{\alpha_0 + m(x)} + \frac{m(x)}{\alpha_0 + m(x)} \int_S \left(\frac{b_1 M_1}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1}{\alpha_0 - c_1} w_1(y) \right) P(\mathrm{d}y|x,a) \right\}, \\ &= \frac{M_1 b_1}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1 w_1(x)}{\alpha_0 - c_1} \leq \frac{M_1(b_1 + \alpha_0)}{\alpha_0(\alpha_0 - c_1)} w_1(x). \end{aligned} \tag{17}$$

Thus,
$|u^*| = |\lim_{n\longrightarrow\infty} u_n| \le M_1(b_1 + \alpha_0)/\alpha_0(\alpha_0 - c_1)w_1(x)$, that is, $u^* \in \mathcal{B}_{w_1}(S)$.

Last, we show $Tu^* = u^*$. By the monotonicity of T and u_n, we have $Tu^* \ge Tu_n = u_{n+1}$ for all $n \ge 0$, and so $Tu^* \ge u^*$.

On the other hand, by the definition of the operator T, we have

$$u_{n+1}(x) \ge \frac{r(x,a)}{\alpha(x)+m(x)} + \frac{m(x)}{\alpha(x)+m(x)} \int_S u_n(y) P(\mathrm{d}y|x,a). \tag{18}$$

Then, letting $n \longrightarrow \infty$, by Hern $a\prime$ Hernndez-Lerma and Lasserre [9], [Lemma 8.3.7], we obtain

$$u^*(x) \ge \frac{r(x,a)}{\alpha(x)+m(x)} + \frac{m(x)}{\alpha(x)+m(x)} \int_S u^*(y) P(\mathrm{d}y|x,a), \tag{19}$$

which follows that $u^* \ge Tu^*$. Thus, we have $Tu^* = u^*$, that is, u^* is the solution of DOE in (14). □

Remark 2. Theorem 1 is not only the generalization of the control model with a constant discount factor in Guo [4] [Theorem 3.3(a)-(b)] but also the model in Ye and Guo [16] whose policies are restricted within the family Φ of all randomized stationary policies.

The following Lemma 1 is a direct consequence of [16] [Theorem 3.2].

Lemma 1. *Under Assumptions 1 and 2, for each $x \in S$ and $\varphi \in \Phi$, the expected discounted reward criterion $J(x,\varphi)$ is the unique solution of the following equation:*

$$\alpha(x)u(x) = r(x,\varphi) + \int_S u(y) q(\mathrm{d}y|x,\varphi). \tag{20}$$

Lemma 2. *Under Assumptions 1 and 2, for each $x \in S$, $\varphi \in \Phi$, and $u \in \mathcal{B}_{w_1}(S)$, the following assertions hold.*

(a) if

$$\alpha(x)u(x) \ge r(x,\varphi) + \int_S u(y) q(\mathrm{d}y|x,\varphi), \tag{21}$$

then, we have $u(x) \ge J(x,\varphi)$.

(b) if

$$\alpha(x)u(x) \le r(x,\varphi) + \int_S u(y) q(\mathrm{d}y|x,\varphi), \tag{22}$$

then we have $u(x) \le J(x,\varphi)$.

Proof. By (21), there exists a nonnegative measurable function $v(x)$ on S such that

$$\alpha(x)u(x) = r(x,\varphi) + v(x) + \int_S u(y) q(\mathrm{d}y|x,\varphi). \tag{23}$$

Now, let $\overline{r}(x,a) = r(x,a) + v(x)$, and we get the new Markov decision processes:

$$\overline{M} := \{S, (A(x) \subseteq A, x \in S), q(\cdot|x,a), \alpha(x), \overline{r}(x,a)\}, \tag{24}$$

in which only the reward rate function is different from the model in (1). Moreover, for each $x \in S$, $\varphi \in \Phi$, the expected discounted reward criterion is given by

$$\overline{J}(x,\varphi) := J(x,\varphi) + \int_0^\infty E_x^\varphi \left[e^{-\int_0^t \alpha(x(s))\mathrm{d}s} v(x(t))\mathrm{d}t \right] \ge 0. \tag{25}$$

By Lemma 1, we have $u(x) = \overline{J}(x,\varphi) \ge J(x,\varphi)$, which gives part (a).

Similarly, we can prove (b). □

Remark 3. Lemma 2 is the generalization of Ye and Guo [16] [Lemma 6.3].

Theorem 2. *Under Assumptions 1 and 2, for each $x \in S$, the optimal value function $J^*(x)$ is the solution of DOE in (12), and there exists a (deterministic) stationary policy $f^* \in F$ such that*

$$\alpha(x)J^*(x) \ge r(x,f^*) + \int_S J^*(y) q(\mathrm{d}y|x,f^*). \tag{26}$$

Proof. By Theorem 1(b), for each $x \in S$ and $\pi \in \Pi$, we have

$$\alpha(x)u^*(x) \ge r(x,\pi_t) + \int_S u^*(y) q(\mathrm{d}y|x,\pi_t), \tag{27}$$

which together with Lemma 2(a) yields that $u^*(x) \ge J(x,\pi)$, and then, $u^*(x) \ge J^*(x)$. Note that $\int_S u^*(y) q(\mathrm{d}y|x,a)$ is upper semicontinuous on $a \in A(x)$; then, by [9] [Lemma 8.3.8], we can obtain that there exists a policy $f^* \in F$ such that

$$\alpha(x)u^*(x) = r(x,f^*(x)) + \int_S u^*(y) q(\mathrm{d}y|x,f^*(x)), \tag{28}$$

for all $x \in S$. Thus, by Lemma 1, we have $u^*(x) = J(x,f^*)$. □

Remark 4.

(a) Theorem 2 shows that the optimal value function is a solution to the DOE ((21)]) and ensures the existence of an optimal (deterministic) stationary policy.

(b) By the construction of the new Markov decision processes, the proof of Theorem 2 is more concise than in [16] Theorem 3.3.

4. An Iteration Algorithm for ε-Optimal Policies

In this section, we provide an iteration algorithm for ε-optimal policies.

In fact, for the operator T on $\mathcal{B}_{w_1}(S)$ in Section 3, with $m(x) = q^* + 1$, it holds that

Step 1. (Initialization). Choose any $\varepsilon > 0$, let $w_1 \geq \max\{1, 2b_1/\alpha_0 - c_1\}$ in Assumption 1, and for each $x \in S$, let
$u_0(x) := -M_1 b_1/\alpha_0(\alpha_0 - c_1) - M_1 w_1(x)/\alpha_0 - c_1$
Step 2. (Iteration). For each $x \in S$, let
$u_{n+1}(x) := \max_{a \in A(x)} \left\{ r(x,a)/\alpha(x) + q^* + 1 + q^* + 1/\alpha(x) + q^* + 1 \int_S u_n(y) P(\mathrm{d}y|x,a) \right\}$
where $P(\mathrm{d}y|x,a) := q(\mathrm{d}y|x,a)/q^* + 1 + \delta_x(\{y\})$..
Step 3. (Approximation value). If
$u_{n+1} - u_{nw_1} \leq \varepsilon(1-\lambda)/2\lambda$
where $\lambda := \alpha_0 + c_1/2 + q^* + 1/\alpha_0 + q^* + 1 \leq 1$, go on step 4, otherwise increment n by 1 and return to step 2.
Step 4. (ε-optimal policy). For each $x \in S$, choose
$f_\varepsilon(x) \in \operatorname{argmax}_{a \in A(x)} \left\{ r(x,a)/\alpha(x) + q^* + 1 + q^* + 1/\alpha(x) + q^* + 1 \int_S u_{n+1}(y) P(\mathrm{d}y|x,a) \right\}$,
and f_ε is ε-optimal policy.

Algorithm 1: (An iteration algorithm).

$$\|Tu - Tv\|_{w_1} \leq \lambda \|u - v\|_{w_1}, \text{ for all, } \quad u, v \in \mathscr{B}_{w_1}(S). \tag{29}$$

Then, by Algorithm 1, we have

$$\begin{aligned} \left|u_{n+1} - u^*\right| &\leq \left|u_{n+1} - \mathrm{T}u_{n+1}\right| + \left|Tu_{n+1} - u^*\right| \\ &= \left|Tu_n - \mathrm{T}u_{n+1}\right| + \left|Tu_{n+1} - Tu^*\right|, \\ &\leq \lambda \left\|u_{n+1} - u_n\right\|_{w_1} w_1(x) + \lambda \left\|u_{n+1} - u^*\right\|_{w_1} w_1(x), \end{aligned} \tag{30}$$

which yields that

$$\left\|u_{n+1} - u^*\right\|_{w_1} \leq \frac{\lambda}{1-\lambda} \left\|u_{n+1} - u_n\right\|_{w_1}. \tag{31}$$

By a similar argument, we have

$$\left\|J(g, f_\varepsilon) - u_{n+1}\right\|_{w_1} \leq \frac{\lambda}{1-\lambda} \left\|u_{n+1} - u_n\right\|_{w_1}, \tag{32}$$

and then,

$$\left\|J(g, f_\varepsilon) - u^*\right\|_{w_1} \leq \frac{2\lambda}{1-\lambda} \left\|u_{n+1} - u_n\right\|_{w_1} \leq \varepsilon. \tag{33}$$

5. Asymptotic Optimality of Quantized Stationary Policies

5.1. Approximation of Deterministic Stationary Policies. In Section 3, we give the existence of the deterministic stationary policies for the CTMDPs in (1) under suitable conditions. However, in practice, sometimes, the action space cannot satisfy the continuity conditions in theoretical research. Thus, in this section, we will discretize and incentivize the action space, so that we can construct a sequence of policies, namely "quantizer policies," which is the approximation of the deterministic stationary policies of the CTMDPs in (1).

To this end, we first give the definitions of quantizers and deterministic stationary quantizer policies.

Definition 1. A measurable function $f: S \longrightarrow A$ is called a quantizer from S to A, if $f(S) := \{f(x) \in A: x \in S\}$ is finite. Let $\mathscr{F}$ denote the set of all quantizers from S to A.

Definition 2. A policy is called a deterministic stationary quantizer policy, if there exists a constant sequence $\pi = \{\pi_n, n \geq 0\}$ of stochastic kernels on A given S such that $\pi_n(\cdot|x) = \delta_{f(x)}(\cdot)$ for all n for some $f \in \mathscr{F}$, where $\delta_{f(x)}(\cdot)$ is Dirac measure as in (11).

For any finite set $\Lambda \subset A$, let $\mathscr{F}(\Lambda)$ denotes the set of all quantizers having range Λ, and let $S\mathscr{F}(\Lambda)$ denotes the set of all deterministic stationary quantizer policies induced by $\mathscr{F}(\Lambda)$.

Denote the metric on A as d_A, and then, the action space A is totally bounded by its compactness. For any fixed integer $k \geq 1$, there exists a finite point set $\{a_i\}_{i=1}^{n_k}$ such that for all $a \in A$,

$$\min_{1 \leq i \leq n_k} \mathrm{d}_A(a, a_i) \leq \frac{1}{k}, \tag{34}$$

where $\{a_i\}_{i=1}^{n_k}$ is called the $1/k-$ net in A. From this, for any a deterministic stationary policy $f \in F$, we can construct a sequence of quantizer policies to approximate to f by the following methods.

Lemma 3 (The construction of quantizer policies). *Let $\Lambda_k := \{a_i\}_{i=1}^{n_k}$ is the $1/k-$ net in A, for each $x \in S$ and deterministic stationary policy $f \in F$, we define*

$$f_k(x) := \operatorname{argmin}_{a \in \Lambda_k} \mathrm{d}_A(f(x), a). \tag{35}$$

Then, $\{f_k\}_{k \geq 1}$ is a deterministic stationary quantizer policy sequence, and f_k converges uniformly to f as $k \longrightarrow \infty$.

Proof. Lemma 3 holds obviously by [21] [Section 3].

We also call $\{f_k\}_{k \geq 1}$ as the quantized approximations of f. Next, we show their expected discounted rewards also satisfy the approximation. For this purpose, we need the following conditions as follows. □

Assumption 3. Let c_1 be as in Assumption 1, for each $x \in S$, suppose that $q(\{x\}|x,a)$ and $q(B|x,a)$ are setwise continuous in $a \in A(x)$ for each $B \in \mathscr{B}(S)$ and $x \notin B$, that is, if $a_k \longrightarrow a$, then $q(\cdot|x,a_n) \longrightarrow q(\cdot|x,a)$ setwise.

Lemma 4. *Suppose that Assumptions 2 and 3 hold. Let $f \in F$ be a deterministic stationary policy of the control model in (1)*

and $\{f_k\}_{k\geq 1}$ *be the quantized approximations of* f *as in Lemma 3, then for each* $x \in S$, *the strategic measures* $\{P_x^{f_k}\}$ *induced by the quantized approximations* f_k *of* f *converge to* $\{P_x^{f_k}\}$ *in the weak topology. Therefore,* $E_x^{f_k} r(x(t), a(t))$ *converges to* $E_x^f r(x(t), a(t))$.

Proof. The proof is similar to that of [21], Proposition 3.1, and by Assumption 3 and the definition of the strategic measures P_x^f as in [6], [Section 2.3] or [5], [Section II], we can get Lemma 4 holds.

Now, we give the approximation result on the expected discounted rewards of the deterministic stationary quantizer policies. □

Theorem 3. *Suppose that Assumptions 1–3 hold. Let* $f \in F$ *be a deterministic stationary policy of the control model in (1), and* $\{f_k\}_{k\geq 1}$ *be the quantized approximations of* f *as in Lemma 3, then for each* $x \in S$, *we have*

$$\lim_{k\longrightarrow\infty} J(x, f_k) = J(x, f). \tag{36}$$

Proof. By the definition of the expected discounted reward criterion, we can get

$$\begin{aligned} \left|J(x, f_k) - J(x, f)\right| &= \int_0^\infty e^{-\int_0^t \alpha(x(s))\mathrm{d}s} \left|\left[E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right]\mathrm{d}t\right|, \\ &\leq \int_0^T e^{-\alpha_0 t} \left|\left[E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right]\mathrm{d}t\right|, \\ &+ \int_0^T e^{-\alpha_0 t} \left|\left[E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right]\mathrm{d}t\right|. \end{aligned} \tag{37}$$

Note that, by Lemmas 3 and 4, we have

$$\begin{aligned} &\left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right|, \\ &\leq \left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^{f_k} r(x(t), f(x(t)))\right|, \\ &+ \left|E_x^{f_k} r(x(t), f(x(t))) - E_x^f r(x(t), f(x(t)))\right|, \\ &\longrightarrow 0, \end{aligned} \tag{38}$$

which yields that

$$\begin{aligned} &\int_0^T e^{-\alpha_0 t} \left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right| \\ &\qquad \mathrm{d}t \longrightarrow 0 \,(k \longrightarrow \infty). \end{aligned} \tag{39}$$

On the other hand, we have

$$\begin{aligned} &\int_T^\infty e^{-\alpha_0 t} \left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right|\mathrm{d}t, \\ &\leq \int_T^\infty e^{-\alpha_0 t} \left(\left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right|\right)\mathrm{d}t, \\ &\leq M_1 \int_T^\infty e^{-\alpha_0 t} \left[E_x^{f_k} w_1(x(t)) + E_x^f w_1(x(t))\right]\mathrm{d}t, \\ &\leq 2M_1 \int_T^\infty e^{-\alpha_0 t} \left[e^{c_1 t} w_1(x) + \frac{b_1}{c_1}\left(e^{c_1 t} - 1\right)\right]\mathrm{d}t, \\ &= 2M_1 \left(\frac{w_1(x)}{\alpha_0 - c_1} e^{-(\alpha_0 - c_1)T} + \frac{b_1}{c_1(\alpha_0 - c_1)} e^{-(\alpha_0 - c_1)T} - \frac{b_1}{\alpha_0 c_1} e^{-\alpha_0 T}\right), \end{aligned} \tag{40}$$

where the last inequality holds by [4] [Theorem 3.2(b)]. Then, we have

$$\int_T^\infty e^{-\alpha_0 t} \left|E_x^{f_k} r(x(t), f_k(x(t))) - E_x^f r(x(t), f(x(t)))\right|\mathrm{d}t \longrightarrow 0, \tag{41}$$

as $T \longrightarrow \infty$. By (40), we can get

$$\lim_{k\longrightarrow\infty} J(x, f_k) = J(x, f). \tag{42}$$

□

5.2. Rates of Convergence

Definition 3. Let $\|\bullet\|_{TV}$ denote the total variation distance between measures P_1 and P_2 on the probability space $(\Omega, \mathfrak{F})$, which satisfies

$$\|P_1 - P_{2TV}\| 2 \sup_{D\in\mathfrak{F}(\Omega)} |P_1(D) - P_2(D)|. \tag{43}$$

Assumption 4. For each $x \in S$, suppose that the model (1) satisfies the following conditions:

(a) A is a compact subset of $\mathbb{R}^{\mathrm{d}}$ for some $\mathrm{d} \geq 1$

(b) For all $(x, a_1), (x, a_2) \in K$, there exists a constant $K_1 > 0$ such that

$$|r(x, a_1) - r(x, a_2)| \leq K_1 \mathrm{d}_A(a_1, a_2). \tag{44}$$

(c) For all $(x, a_1), (x, a_2) \in K$ and the function of transition rates $q(\cdot|x, a)$, there exists a constant $K_2 > 0$ such that

$$\|q(\cdot|x, a_1) - q(\cdot|x, a_2)\|_{TV} \leq K_2 \mathrm{d}_A(a_1, a_2). \tag{45}$$

By Lemma 3 and Assumption 4, the following Lemma holds.

Lemma 5. *For any measurable function $f\colon S \longrightarrow A$, we can construct a sequence of quantizers $\{f_k\}_{k\geq 1}$ from S to A, and there exists some constant $K_3 > 0$ such that*

$$\sup_{x\in S} \mathrm{d}_A(f(x), f_k(x)) \leq K_3 (1/k)^{1/\mathrm{d}}. \tag{46}$$

Now, we give the convergence rates result.

Theorem 4. *Suppose that Assumptions 1–4 hold. Let $f \in F$ be a deterministic stationary policy of the control model in (1)*

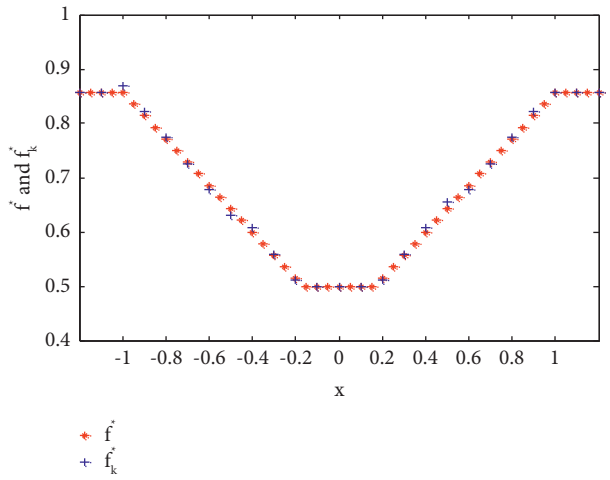

FIGURE 1: The optimal policy $f^*(x)$ and $f_k^*(x)$ as $k = 15$.

and $\{f_k\}_{k\geq 1}$ be the quantized approximations of f as in Lemma 3, then for each $x \in S$, it holds that

$$|J(x, f) - J(x, f_k)| \leq \frac{K_3 (1/k)^{1/\mathrm{d}}}{\alpha_0} \left\{K_1 + \frac{b_1 M_1 K_2}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1 K_2}{\alpha_0 - c_1} w_1(x)\right\}. \tag{47}$$

Proof. By Lemma 1, we can get

$$\begin{aligned} \alpha(x) J(x, f) &= r(x, f) + \int_S q(\mathrm{d}y|x, f), \\ \alpha(x) J(x, f_k) &= r(x, f_k) + \int_S J(y, f_k) q(\mathrm{d}y|x, f). \end{aligned} \tag{48}$$

Then, by Lemma 5 and Theorem 1, we have

$$\begin{aligned} \alpha_0 |J(x, f) - J(x, f_k)| &\leq |\alpha(x)(J(x, f) - J(x, f_k))|, \\ &\leq |r(x, f) - r(x, f_k)| + \left|\int_S q(\mathrm{d}y|x, f) - \int_S J(y, f_k) q(\mathrm{d}y|x, f_k)\right|, \\ &\leq K_1 \mathrm{d}_A(f(x), f_k(x)) + \sup_{x\in S, \pi\in\Pi} \leq K_1 K_3 (1/k)^{1/\mathrm{d}} + \left\{\frac{b_1 M_1}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1}{\alpha_0 - c_1} w_1(x)\right\} K_2 \mathrm{d}_A(f(x), f_k(x)), \\ &= K_3 (1/k)^{1/\mathrm{d}} \left\{K_1 + \frac{b_1 M_1 K_2}{\alpha_0(\alpha_0 - c_1)} + \frac{M_1 K_2}{\alpha_0 - c_1} w_1(x)\right\}, \end{aligned} \tag{49}$$

which yields that

$$\left|J(x,f)-J(x,f_k)\right| \le \frac{K_3(1/k)^{1/d}}{\alpha_0}\left\{K_1+\frac{b_1M_1K_2}{\alpha_0(\alpha_0-c_1)}+\frac{M_1K_2}{\alpha_0-c_1}w_1(x)\right\}. \tag{50}$$

□

6. An Example

In this section, we give an example to illustrate our main results.

Consider a control problem of hypertension. As it is well known, we can describe the blood pressure with Gaussian distribution, and thus, the quantity of blood pressure may take values in $S := (-\infty, \infty)$. When the current amount of blood pressure is at $x \in S$, a controlled amount a is given by $a \in A = A(x) := [L_1, L_2]$ for each $x \in S$ with $L_1 > 0$. The rate of change of blood pressure is given as follows:

$$q(B|x,a) := a(|x|+1)\int_{B-\{x\}} \frac{1}{\sqrt{2\pi}} e^{-(y-x)^2/2}\,dy - a(|x|+1)\delta_x(B), \tag{51}$$

for $(x,a) \in K\{(x,a)|x \in S, a \in A(x)\}$ and $B \in \mathcal{B}(S)$, where π is the circumference ratio, and $\delta_x(\cdot)$ is the Dirac measure at x. It is clear that $q(\cdot|x,a)$ is a transition rate function. We denote by $r(x,a) := n_0(x^2 + n_1|x| - n_2a^2)$ the cost of taking control a when the current amount of blood pressure is at $x \in S$, and regard a as an action. The discount factor is defined by $\alpha(x) := 1/|x| + 1 + \beta$ for $x \in S$ with a constant $\beta > L_1/2$. Suppose that the constants $n_k\,(k = 0, 1, 2)$ satisfy that

(i) $n_0 > 0, n_1 > 1 + 1/\beta$ and $n_2 = \rho^2/4(\beta\rho - 1)$, where $\rho := n_1(\beta+1) - (\beta+2)$

(ii) $\rho L_2 > \rho L_1 > 2(\beta\rho - 1)$

Let $w_1(x) = x^2 + 1$, $w_2(x) = x^4 + 1$ and $q^*(x) = L_1(|x|+1)$, and then, by Steps 1–4 of the iteration algorithm in Section 4, the approximate optimal value is

$$J^*(x) = \rho n_0 x^2 + (2\rho - n_1 + 1)n_0|x| + (\rho - n_1 + 1)n_0, \tag{52}$$

and the optimal stationary policy is

$$f^*(x) := \begin{cases} L_1, & |x| < U, \\ \dfrac{2(\beta\rho-1)}{\rho}(|x|+1), & U \le |x| \le V, \\ L_2, & |x| > V, \end{cases} \tag{53}$$

where $U := \rho L_1/2(\beta\rho - 1) - 1$, $V := \rho L_2/2(\beta\rho - 1) - 1$.

Now, we can construct a sequence of quantizer policies of f^* as follows:

$$f_k^*(x) := \begin{cases} L_1, & |x| < U, \\ \dfrac{2(\beta\rho-1)}{\rho}\left(U + \dfrac{(2j+1)}{2}k_0 + 1\right), & k_0 j \le |x| - U < k_0(j+1), \quad j = 0, 1, \ldots, k, \\ L_2, & |x| > V, \end{cases} \tag{54}$$

where $k_0 := V - U/k$.

Now, we compute f^*, f_k^*, and J^* by assigning values to parameters β, L_1, L_2, n_1, and n_2 as follows:

$$\beta = \frac{1}{2}, L_1 = \frac{1}{2}, L_2 = \frac{6}{7}, n_0 = 1, n_1 = 4, \tag{55}$$

then, the optimal value is $J^*(x) = 7/2x^2 + 4|x| + 1/2$, and the optimal stationary policy is

$$f^*(x) := \begin{cases} \dfrac{1}{2}, & |x| < \dfrac{1}{6}, \\ \dfrac{3}{7}(|x|+1), & \dfrac{1}{6} \le |x| \le 1, \\ \dfrac{6}{7}, & |x| > 1. \end{cases} \tag{56}$$

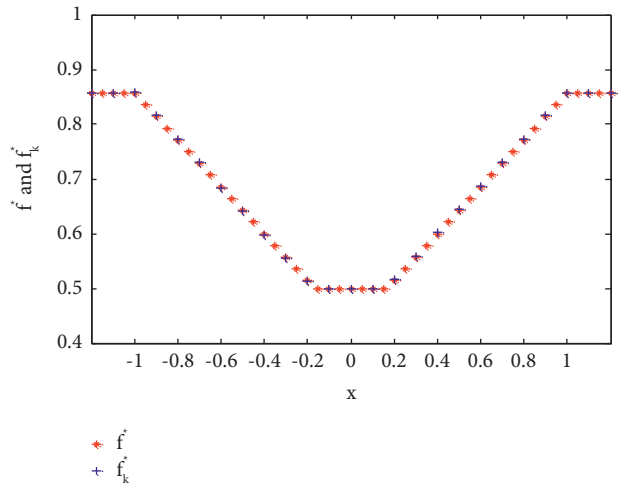

FIGURE 2: The optimal policy $f^*(x)$ and $f_k^*(x)$ as $k = 100$.

The quantizer policies f_k^* of f^* are

$$f_k^*(x) := \begin{cases} \dfrac{1}{2}, & |x| < \dfrac{1}{6}, \\ \dfrac{1}{2} + \dfrac{3}{7k}\left(\dfrac{5}{6}j + \dfrac{5}{12}\right), & \dfrac{5}{6k}j \le |x| - \dfrac{1}{6} < \dfrac{5}{6k}(j+1), \quad j = 0, 1, \ldots, k, \\ \dfrac{6}{7}, & |x| > 1. \end{cases} \tag{57}$$

Furthermore, the asymptotic approximation of the optimal policy $f^*(x)$ is given by Figures 1 and 2 when $k = 15$ and $k = 100$, respectively. This verifies $f_k^*(x) \longrightarrow f^*(x)$ for each $x \in S$, and by Theorem 3, it holds that $\lim_{k\longrightarrow\infty} J(x, f_k) = J(x, f)$.

7. Conclusions

In this paper, we are concerned with the asymptotic optimality of quantized stationary policies in CTMDPs with Polish spaces and varying (state-dependent) discount factors. First, we establish the discounted optimal equation (DOE) and give the existence of its solutions. Then, by a relatively simple proof, we obtain the existence of optimal deterministic stationary policies under suitable conditions in Theorem 2. Meanwhile, we generalize the relevant conclusions of Ye and Guo [16] in Theorem 1 and Lemma 2. Next, we discretize and incentivize the action space, construct a sequence of policies, namely "quantizer policies," and obtain the approximation results and the rates of convergence for the optimal policies on the CTMDPs in (1) as in Theorem 3. Finally, we give an example to illustrate the asymptotic optimality.

Data Availability

No data were used to support the findings of this study.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 11961005) and the Opening Project of Guangdong Province Key Laboratory of Computational Science at Sun Yat-Sen University (Grant No. 2021021).

References

[1] B. T. Doshi, "Continuous-time control of Markov processes on an arbitrary state space: discounted rewards," *Annals of Statistics*, vol. 4, no. 6, pp. 1219–1235, 1976.

[2] E. B. Dynkin and A. A. Yushkevich, *Controlled Markov Processes*, Springer, New York, 1979.

[3] E. A. Feinberg, "Continuous time discounted jump Markov decision processes: a discrete-event approach," *Mathematics of Operations Research*, vol. 29, no. 3, pp. 492–524, 2004.
[4] X. P. Guo, "Continuous-time Markov decision processes with discounted rewards: the case of Polish spaces," *Mathematics of Operations Research*, vol. 32, no. 1, pp. 73–87, 2007.
[5] X. P. Guo, "Constrained optimization for average cost continuous-time Markov decision processes," *IEEE Transactions on Automatic Control*, vol. 52, no. 6, pp. 1139–1143, 2007.
[6] X. P. Guo and O. Hernndez-Lerma, *Continuous-time Markov Decision Processes: Theory and Applications*, Springer, Berlin, 2009.
[7] X. P. Guo and X. Y. Song, "Discounted continuous-time constrained Markov decision processes in Polish spaces," *Annals of Applied Probability*, vol. 21, no. 5, pp. 2016–2049, 2011.
[8] O. Hernndez-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes*, Springer, New York, 1996.
[9] O. Hernndez-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.
[10] M. L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.
[11] E. A. Feinberg and A. Shwartz, "Markov decision models with weighted discounted criteria," *Mathematics of Operations Research*, vol. 19, no. 1, pp. 152–168, 1994.
[12] J. González-Hernández, R. R. López-Martínez, and J. R. Pérez-Hernández, "Markov control processes with ran- domized discounted cost," *Mathematical Methods of Operations Research*, vol. 65, no. 1, pp. 27–44, 2007.
[13] X. Wu and X. P. Guo, "First passage optimality and variance minimisation of Markov decision processes with varying discount factors," *Journal of Applied Probability*, vol. 52, no. 02, pp. 441–456, 2015.
[14] X. Wu and J. Y. Zhang, "Finite approximation of the first passage models for discrete-time Markov decision processes with varying discount factors," *Discrete Event Dynamic Systems*, vol. 26, no. 4, pp. 669–683, 2016.
[15] K. Hinderer, "Foundations of non stationary dynamic programming with discrete time parameter," *Lecture Notes in Operations Research*, vol. 33, 1970.
[16] L. E. Ye and X. P. Guo, "Continuous-time Markov decision processes with state-dependent discount factors," *Acta Applicandae Mathematica*, vol. 121, no. 1, pp. 5–27, 2012.
[17] D. Bertsekas and J. Tsitsiklis, *Neuro-Dynammic Programming*, MA: Athena Scientific, Boston, 1996.
[18] R. Cavazos-Cadena, "Finite-state approximations for denumerable state discounted Markov decision processes," *Applied Mathematics and Optimization*, vol. 14, pp. 1–26, 1986.
[19] F. Dufour and T. Prieto-Rumeau, "Finite linear programming approximations of constrained discounted Markov decision processes," *SIAM Journal on Control and Optimization*, vol. 51, no. 2, pp. 1298–1324, 2013.
[20] X. Wu and X. P. Guo, "Convergence of Markov decision processes with constraints and state-action dependent discount factors," *Science China Mathematics*, vol. 63, no. 1, pp. 167–182, 2020.
[21] N. Saldi, T. Linder, and S. Yuksel, "Asymptotic optimality and rates of convergence of quantized stationary policies in stochastic control," *IEEE Transactions on Automatic Control*, vol. 60, no. 2, pp. 553–558, 2015.
[22] N. Saldi, S. Yüksel, and T. Linder, "Near optimality of quantized policies in stochastic control under weak continuity conditions," *Journal of Mathematical Analysis and Applications*, vol. 435, no. 1, pp. 321–337, 2016.
[23] R. B. Lund, S. P. Meyn, and R. L. Tweedie, "Computable exponential convergence rates for stochastically ordered Markov processes," *Annals of Applied Probability*, vol. 6, no. 1, pp. 218–237, 1996.