Kybernetika

Xiao Wu; Yanqiu Tang

Constrained optimality problem of Markov decision processes with Borel spaces and varying discount factors

Kybernetika, Vol. 57 (2021), No. 2, 295–311

Persistent URL: http://dml.cz/dmlcz/149040

Terms of use:

© Institute of Information Theory and Automation AS CR, 2021

Institute of Mathematics of the Czech Academy of Sciences provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these *Terms of use*.

This document has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library* http://dml.cz

KYBERNETIKA — VOLUME 57 (2021), NUMBER 2, PAGES 295–311

CONSTRAINED OPTIMALITY PROBLEM OF MARKOV DECISION PROCESSES WITH BOREL SPACES AND VARYING DISCOUNT FACTORS

Xiao Wu and Yanqiu Tang

This paper focuses on the constrained optimality of discrete-time Markov decision processes (DTMDPs) with state-dependent discount factors, Borel state and compact Borel action spaces, and possibly unbounded costs. By means of the properties of so-called occupation measures of policies and the technique of transforming the original constrained optimality problem of DTMDPs into a convex program one, we prove the existence of an optimal randomized stationary policies under reasonable conditions.

Keywords: constrained optimality problem, discrete-time Markov decision processes, Borel state and action spaces, varying discount factors, unbounded costs

Classification: 90C40, 60J27

1. INTRODUCTION

This paper attempts to study the constrained optimality problem of discrete-time Markov decision processes (DTMDPs) with varying (state-dependent) discount factors, Borel state and compact Borel action spaces, and possibly unbounded costs functions from above and below.

As is well known, more and more researchers focus on the discounted DTMDPs, such as [1, 2, 10, 11], which share a common feature: the discount factor is a constant. For a general case of varying discount factors, [6] studies the non-stationary DTMDPs with the time- and state-dependent discount factors and proves the existence of an optimal Markov policies under suitable conditions, [19] deals with unconstrained DTMDPs with state-dependent discount factors and unbounded rewards/costs and also proves the existence of optimal stationary policies. Recently, [20] considers the discounted DTMDPs in Borel state and compact action spaces with state-dependent discount factors and unbounded rewards, and shows the existence of optimal policies. Also, there are many works on continuous-time MDPs (CTMDPs), such as [8, 13], etc.

On the other hand, constrained DTMDPs have also been discussed widely. One of the pioneering papers is [4], in which the author briefly mentions the constrained case in countable state spaces. More studies on the constrained discounted DTMDPs

DOI: 10.14736/kyb-2021-2-0295

arise in [1, 2] with countable state spaces and a constant discount factor. Furthermore, the constrained discounted DTMDPs with a Borel state space and a constant discount factor have been studied with the development of the convex analytic approach in [5, 9, 15, 16, 17]. Recently, [14] shows the existence of a constrained optimal policy for the constrained DTMDPs with state-action dependent discount factors, denumerable state and Borel action spaces. In addition, [3] and [7] research the convergence of the optimal values of constrained DTMDPs and CDMDPs with denumerable states and a constant discount factor, respectively. These convergence results are crucial for the approximation of optimal value or optimal policies. It is worth noting that [21] develops the convex analytic approach to the constrained DTMDPs in Borel state and action spaces with varying discount factors and bounded costs from below and proves the existence of a randomized stationary optimal policy under slightly stronger conditions. However, in [21], the varying discount factors are assumed to be continuous. To the best of our knowledge, the constrained optimality problem of DTMDPs with Borel state space, measurable discount factors, and unbounded costs from above and below, has not been studied yet.

Inspired by the ideas in [3, 7, 20, 21], we study the constrained optimality problem of DTMDPs in Borel state and action spaces with possibly unbounded costs and the state-dependent discount factors are assumed to be measurable. By introducing so-called occupation measures of policies and studying the characterization of the occupation measures, we prove the solvability of the constrained optimality problem of DTMDPs is equivalent to the solvability of a convex program one, and then, establish the existence of an optimal randomized stationary policy under reasonable conditions weaker than those in [3, 21].

The organization of this paper is as follows. First, in Section 2, we introduce the constrained DTMDPs with state-dependent discount factors, and state the constrained optimality problem of the DTMDPs. After that, we study the occupation measures and some corresponding properties in Subsection 3.1, and prove the existence of the optimal randomized stationary policy for the constrained DTMDPs by means of the technique of transforming the original optimality problem into a convex program in Subsection 3.2. In Section 4, we give an application example of a controlled cash-balance system and verify the existence of an optimal stationary policy. Finally, we finish this article with a conclusion in Section 5.

2. THE MODEL OF CONSTRAINED DTMDPS

Consider the constrained DTMDPs with varying discount factors as below:

$$\mathcal{M}:=\left\{S, A, A(x), Q(dy|x,a), \alpha(x), c^0(x,a), \left(c^l(x,a), d^l, 1\leq l\leq q\right), \gamma\right\}, \tag{1}$$

where S and A are state and action spaces, which are assumed to be Borel spaces with Borel σ-fields $\mathcal{B}(S)$ and Borel σ-fields $\mathcal{B}(A)$ respectively, $A(x)$ is the set of admissible actions at state $x \in S$. Let $K := \left\{(x,a)|x \in S, a \in A(x)\right\}$ be the set of all feasible state-action pairs, and suppose that K is a measurable Borel subset of $S \times A$. The transition law $Q(dy|x,a)$ is the one-step (homogeneous) transition probability on S given K, and the discount factor $\alpha(x)$ is a measurable function from S to $[0,1)$. In addition, $c^0(x,a)$ and $c^l(x,a)$ $(1 \leq l \leq q)$ denote the objective cost and constrained cost functions

respectively, which are assumed to be measurable on K. Finally, The real numbers d^l $(1 \le l \le q)$ denote the constraints, and γ denotes the initial distribution on S.

Let H_m be the family of admissible histories up to time m for each $m = 0, 1, \ldots$, that is, $H_0 := S$ and $H_m := K^m \times S$ for $m = 1, 2, \ldots$, and the control policies are given as follows:

Definition 2.1. A randomized history-dependent policy is a sequence $\pi = \{\pi_m,\ m = 0, 1, \ldots\}$ of stochastic kernels π_m on A given H_m such that

$$\pi_m(A(x_m)|h_m) = 1 \quad \forall h_m := (x_0, a_0, x_1, a_1, \ldots, x_m) \in H_m,\ m = 0, 1, \ldots.$$

Definition 2.2. A randomized history-dependent $\pi = \{\varphi_m,\ m = 0, 1, \ldots\}$ is said to be (randomized) stationary if $\varphi_m(\mathrm{d}a|x_0, a_0, x_1, a_1, \ldots, x_{m-1}, a_{m-1}, x_m)$ are independent of m and all histories $(x_0, a_0, x_1, a_1, \ldots, x_{m-1}, a_{m-1})$, so that π is of the form $\pi = \{\varphi, \varphi, \varphi, \ldots\}$ with φ being a stochastic kernel on A given S such that $\varphi_m(A(x)|x) \equiv 1$ on S. In this case, the policy π is denoted by φ. All randomized history-dependent and stationary policies are denoted by Π and Φ, respectively.

Suppose that $(\Omega, \mathcal{F})$ is the measurable space, where Ω is the canonical sample space and $\mathcal{F}$ is the corresponding product $\sigma-$algebra. For any initial distribution γ on S and $\pi = \{\pi_m\} \in \Pi$, by the well-known Tulcea's theorem in [10, p.178], there exist a unique probability P_γ^π on $(\Omega, \mathcal{F})$, and a state-action process $\{i_m, a_m, m = 0, 1, \ldots\}$ defined on this space such that, for each $\Gamma \in \mathcal{B}(S), C \in \mathcal{B}(A)$ and $m \ge 0$,

$$\begin{aligned}
&P_\gamma^\pi(x_0 \in \Gamma) = \mu(\Gamma),\\
&P_\gamma^\pi(a_m \in C|h_m) = \pi_n(C|h_m),\\
&P_\gamma^\pi(x_{m+1} \in \Gamma|h_m, a_m) = Q(\Gamma|x_m, a_m),
\end{aligned}$$

see, e. g., [10, p.16] for the construction of the probability measure P_γ^π. If γ is concentrated at some state x, then we write P_γ^π as P_x^π. E_γ^π is the expectation operation corresponding to P_γ^π, and E_γ^π is denoted by E_x^π when γ is concentrated at some state x.

Definition 2.3. For each $\pi \in \Pi$, $0 \le l \le q$ and initial distribution γ, the discounted criteria is defined by

$$V^l(\pi) \ := \ E_\gamma^\pi \left[\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \alpha(x_k) c^l(x_m, a_m) \right], \tag{2}$$

where, $V^l(\pi)$ will be replaced by $V^l(x, \pi)$ if γ is concentrated at some state x.

Remark 2.4. (a) In (2) and what follows, for any sequence $\{y_j, j = 0, 1, \ldots, \}$ we use the convention that

$$\prod_{j=m}^{k} y_j := 1 \quad \text{and} \quad \sum_{j=m}^{k} y_j := 0 \quad \text{if } k < m.$$

(b) If $B = \emptyset$ and $\alpha(\cdot) \equiv \alpha$ is a constant in $(0, 1)$, then $V^l(\pi)$ in (2) becomes the standard infinite-horizon α-discounted cost as in [3, 10, 11].

Let $U := \{\pi \in \Pi \big| V^l(\pi) \le d^l, 1 \le l \le q\}$ be the set of all *feasible policies*, which is assumed to be nonempty. Then, the optimal value of $\mathcal{M}$ is given by

$$V^{0*}(\pi) = \inf_{\pi \in U} V^0(\pi).$$

Thus, the corresponding *constrained optimality problem* (COP) of the DTMDP in (1) is:

$$\text{COP}: \qquad \text{minimize } V^0(\pi) \text{ over } \pi \in U.$$

Moreover, we say that π^* is an *optimal policy* for COP if $\pi^* \in U$ minimizes $V^0(\pi)$ over $\pi \in U$, that is, $V^0(\pi^*) = \inf_{\pi \in U} V^0(\pi)$.

Note that, since $c^0(x, a)$ is allowed to be unbounded from above and below, it can be regarded as rewards rather than costs only, and thus the corresponding COP is to maximize $V^0(\pi)$ over $\pi \in U$.

Now, we introduce some notations and terminology. We say that $\omega : X \to [1, \infty)$ is a *strictly unbounded* function on a Borel space X if there exists a nondecreasing sequence $\{\Gamma_m\}$ of compact sets such that $\Gamma_m \uparrow X$ and $\lim_{m\to\infty} \inf_{x \in \Gamma_m^c} \omega(x) = \infty$. Let $B_\omega(S)$ be the Banach space of real-valued measurable functions u on S with the finite norm $\|u\|_\omega := \sup_{x \in S} \frac{|u(x)|}{\omega(x)}$.

Next, we state the hypotheses on the control model $\mathcal{M}$ in (1).

Assumption 2.5. (a) There exists a constant $\alpha \in (0, 1)$ such that $\sup_{x \in S} \alpha(x) \le \alpha$;

(b) There exist nonnegative constants β $(0 < \beta < \frac{1}{\alpha})$ and L, and a strictly unbounded function $\omega \ge 1$ on S, such that

$$\int_S \omega(x)\gamma(\mathrm{d}x) < \infty, \qquad \sup_{a \in A(x)} |c^l(x, a)| \le L\omega(x) \ (l = 0, 1, \ldots, q)$$

and

$$\sup_{a \in A(x)} \int_S Q(\mathrm{d}y|x, a)\omega(y) \le \beta\omega(x) \qquad \forall x \in S;$$

(c) For each $x \in S$, the set $A(x)$ is compact;

(d) For each $0 \le l \le q$, $x \in S$, and $\Gamma \in \mathcal{B}(S)$, the cost function $c^l(x, \cdot)$ and transition law $Q(\Gamma|x, \cdot)$ are continuous on $A(x)$;

(e) The function $\int_S Q(\mathrm{d}y|x, \cdot)\omega(y)$ is continuous on $A(x)$, for each $x \in S$.

Remark 2.6. (a) Assumption 2.5(a) holds obviously for the cases $\alpha(i) \equiv \alpha \in (0, 1)$, and for the general state spaces case, some application examples are given in [19, Example 4.1] and [20, Example 6.1].

(b) Assumption 2.5(c) implies that, for each $x \in S$, the space $\mathbb{P}(A(x))$ with the topology of weak convergence is compact, where $\mathbb{P}(A(x))$ is the set of all probability measures on $\mathcal{B}(A(x))$. Then, by Tychonoff theorem, $\Phi = \times_{x \in S} \mathbb{P}(A(x))$ is also compact.

(c) Assumption 2.5(b) extends the conditions of [3]Assumption 3.3, and Assumption 2.5(c)-(e) are the continuity-compactness conditions commonly assumed to proving the solvability of optimization problems in [7, 10, 11, 19, 20].

3. MAIN STATEMENTS

In this section, we introduce the concept of an occupation measure and prove the existence of the optimal stationary policies for the constrained MDPs in (1).

3.1. Properties of occupation measures

Definition 3.1. The occupation measure $\mu^\pi(\mathrm{d}x \times \mathrm{d}a)$ of a policy $\pi \in \Pi$ is a measure on $\mathcal{B}(S \times A)$ defined by

$$\mu^\pi(\Gamma \times C) := E_\gamma^\pi \left[\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \alpha(x_k) I_{\{x_m \in \Gamma, a_m \in C\}} \right], \tag{3}$$

for each $\Gamma \in \mathcal{B}(S)$ and $C \in \mathcal{B}(A)$, where $I_{\{\cdot\}}$ is the indicator function. The space of occupation measures is denoted by $\mathcal{D}$, and the marginal measure of $\mu(\mathrm{d}x \times \mathrm{d}a)$ on S by $\hat{\mu}(\mathrm{d}x) := \mu(\mathrm{d}x \times A)$.

Remark 3.2. Since $\mu \in \mathcal{D}$ is concentrated on K, by Definition 3.1, we have

$$V^l(\pi) = \int_{S \times A} c^l(x,a) \mu^\pi(\mathrm{d}x \times \mathrm{d}a) = \int_K c^l(x,a) \mu^\pi(\mathrm{d}x \times \mathrm{d}a) \quad \forall \pi \in \Pi, 0 \le l \le q.$$

For any policy $\pi \in \Pi$, $x \in S$ and bounded measurable function $g(x,a)$ on K, let

$$V(x,\pi,g) := E_x^\pi \left[\sum_{m=0}^{\infty} \prod_{k=0}^{m-1} \alpha(x_k) g(x_m, a_m) \right].$$

Obviously, $V(x,\pi,g)$ is bounded measurable function in $x \in S$. Moreover, we have the following result.

Lemma 3.3. (a) Suppose that Assumption 2.5(a) holds. Then, for any bounded measurable function $g(x,a)$ on K and $\varphi \in \Phi$, $V(x,\varphi,g)$ is the unique bounded solution to the following equation

$$u(x) = g(x,\varphi) + \alpha(x) \int_S u(y) Q(\mathrm{d}y|x,\varphi) \quad \forall x \in S. \tag{4}$$

(b) Suppose that Assumption 2.5(a)-(b) hold, the result of part (a) holds for any $g \in B_\omega(S)$.

Proof. The proof is exactly the same as [20, Theorem 3.1] with $V(x,f)$ in [20] is replaced by $V(x,\varphi,g)$. □

Theorem 3.4. Suppose that Assumption 2.5 is satisfied. Then, the following assertions hold.

(a) For each $\varphi \in \Phi$, it follows that

$$\mu^{\varphi}(\Gamma \times C) = \int_{\Gamma} \hat{\mu}^{\varphi}(\mathrm{d}x)\varphi(C|x) \quad \forall \Gamma \in \mathcal{B}(S), C \in \mathcal{B}(A).$$

(b) For each $\pi \in \Pi$, there exists a policy $\tilde{\varphi} \in \Phi$ such that

$$\mu^{\pi}(\Gamma \times C) = \int_{\Gamma} \hat{\mu}^{\pi}(\mathrm{d}x)\tilde{\varphi}(C|x) = \mu^{\tilde{\varphi}}(\Gamma \times C) \quad \forall \Gamma \in \mathcal{B}(S), C \in \mathcal{B}(A).$$

(c) A measure $\mu(\mathrm{d}x \times \mathrm{d}a)$ on $S \times A$ is an occupation measure if and only if

$$\mu(S \times A) \le \frac{1}{1-\alpha} \tag{5}$$

and

$$\hat{\mu}(\mathrm{d}x) = \gamma(\mathrm{d}x) + \int_{S \times A} \alpha(y) Q(\mathrm{d}x|y,a)\mu(\mathrm{d}y \times \mathrm{d}a). \tag{6}$$

Proof. (a) For any $\varphi \in \Phi$, $\Gamma \in \mathcal{B}(S)$ and $C \in \mathcal{B}(A)$, we have

$$\begin{aligned}\mu^{\varphi}(\Gamma \times C) &= E_{\gamma}^{\varphi}\Big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)I_{\{x_m \in \Gamma, a_m \in C\}}\Big] \\ &= E_{\gamma}^{\varphi}\Big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)I_{\{x_m \in \Gamma\}}\varphi(C|x_m)\Big] = \int_{\Gamma}\hat{\mu}^{\varphi}(\mathrm{d}x)\varphi(C|x).\end{aligned}$$

(b) For any fixed $\pi \in \Pi$, by [10]Proposition D.8, there exists $\tilde{\varphi} \in \Phi$ such that

$$\mu^{\pi}(\Gamma \times C) = \int_{\Gamma} \tilde{\varphi}(C|x)\hat{\mu}^{\pi}(\mathrm{d}x) \quad \forall \Gamma \in \mathcal{B}(S), C \in \mathcal{B}(A). \tag{7}$$

Next, we show that, for any $\Gamma \in \mathcal{B}(S)$ and $C \in \mathcal{B}(A)$, $\mu^{\pi}(\Gamma \times C) = \mu^{\tilde{\varphi}}(\Gamma \times C)$.

Fix a bounded measurable function $g(x,a)$ on K, and, for each $j = 1, 2, \ldots$, define

$$W_j = \prod_{k=0}^{j-2}\alpha(x_k)g(x_{j-1}, a_{j-1}) + \prod_{k=0}^{j-1}\alpha(x_k)V(x_j, \tilde{\varphi}, g) - \prod_{k=0}^{j-2}\alpha(x_k)V(x_{j-1}, \tilde{\varphi}, g).$$

Then, we can get

$$\begin{aligned}&E_{\gamma}^{\pi}\Big[\sum_{m=1}^{j} W_m\Big] = E_{\gamma}^{\pi}\Big[\sum_{m=1}^{j} E_{\gamma}^{\pi}[W_m \big| h_{m-1}, a_{m-1}]\Big] \\ &= E_{\gamma}^{\pi}\Big[\sum_{m=1}^{j}\prod_{k=0}^{m-2}\alpha(x_k)\big[g(x_{m-1}, a_{m-1}) \\ &\quad + \alpha(x_{m-1})\int_{S} V(y, \tilde{\varphi}, g)Q(\mathrm{d}y|x_{m-1}, a_{m-1}) - V(x_{m-1}, \tilde{\varphi}, g)\big]\Big].\end{aligned}$$

Since

$$\sum_{m=1}^{j}\prod_{k=0}^{m-2}\alpha(x_k)g(x_{m-1},a_{m-1})=\sum_{m=1}^{j}W_m+V(x_0,\tilde{\varphi},g)-\prod_{k=0}^{j-1}\alpha(x_k)V(x_j,\tilde{\varphi},g),$$

it follows that

$$\begin{aligned}
&E_\gamma^\pi\Big[\sum_{m=1}^{j}\prod_{k=0}^{m-2}\alpha(x_k)g(x_{m-1},a_{m-1})\Big]\\
=\;&E_\gamma^\pi\Big[\sum_{m=1}^{j}\prod_{k=0}^{m-2}\alpha(x_k)\big[g(x_{m-1},a_{m-1})\\
&+\alpha(x_{m-1})\int_S V(y,\tilde{\varphi},g)Q(\mathrm{d}y|x_{m-1},a_{m-1})-V(x_{m-1},\tilde{\varphi},g)\big]\Big]\\
&+\int_S V(y,\tilde{\varphi},g)\gamma(\mathrm{d}y)-E_\gamma^\pi\big[\prod_{k=0}^{j-1}\alpha(x_k)V(x_j,\tilde{\varphi},g)\big].
\end{aligned}$$

Let $j\to\infty$, then the above equality may be written

$$\begin{aligned}
&E_\gamma^\pi\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_i)g(x_{m-1},a_{m-1})\Big]\\
=\;&\int_S V(y,\tilde{\varphi},g)\gamma(\mathrm{d}y)+E_\gamma^\pi\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)\big[g(x_{m-1},a_{m-1})\\
&+\alpha(x_{m-1})\int_S V(y,\tilde{\varphi},g)Q(\mathrm{d}y|x_{m-1},a_{m-1})-V(x_{m-1},\tilde{\varphi},g)\big]\Big].
\end{aligned}\tag{8}$$

Now, we denote $f(x,a):=g(x,a)+\alpha(x)\int_S V(y,\tilde{\varphi},g)Q(\mathrm{d}y|x,a)$. It is clear that $f(x,a)$ is bounded and measurable on K, and then,

$$\begin{aligned}
&E_\gamma^\pi\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)f(x_{m-1},a_{m-1})\Big]=E_\gamma^\pi\Big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)f(x_m,a_m)\Big]\\
=\;&\int_{S\times A}f(x,a)\mu^\pi(\mathrm{d}x\times\mathrm{d}a)=\int_S\int_A f(x,a)\tilde{\varphi}(\mathrm{d}a|x)\hat{\mu}^\pi(\mathrm{d}x)\\
=\;&E_\gamma^\pi\big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)\int_A f(x_{m-1},a)\tilde{\varphi}(\mathrm{d}a|x_{m-1})\big],
\end{aligned}\tag{9}$$

where the second to the last equality is due to (7) and the last equality is a result of the

definition of $\hat{\mu}^{\pi}(\mathrm{d}x)$. By (8) and (9), we have

$$\begin{aligned}
&\int_{S\times A} g(x,a)\mu^{\pi}(\mathrm{d}x\times\mathrm{d}a)\\
=&\int_S V(y,\tilde{\varphi},g)\gamma(\mathrm{d}y)+E_{\gamma}^{\pi}\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)f(x_{m-1},a_{m-1})\Big]\\
&-E_{\gamma}^{\pi}\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)V(x_{m-1},\tilde{\varphi},g)\Big]\\
=&\,E_{\gamma}^{\pi}\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-2}\alpha(x_k)\int_A\big[f(x_{m-1},a)-V(x_{m-1},\tilde{\varphi},g)\big]\tilde{\varphi}(\mathrm{d}a|x_{m-1})\Big]\\
&+\int_S V(y,\tilde{\varphi},g)\gamma(\mathrm{d}y)\\
=&\int_S V(y,\tilde{\varphi},g)\gamma(\mathrm{d}y)=E_{\gamma}^{\tilde{\varphi}}\Big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)g(x_m,a_m)\Big]=\int_{S\times A}g(x,a)\mu^{\tilde{\varphi}}(\mathrm{d}x\times\mathrm{d}a),
\end{aligned}$$

where the third equality is due to Lemma 3.3. Note that the bounded measurable function $g(x,a)$ is arbitrarily fixed, then

$$\mu^{\pi}(\Gamma\times C)=\mu^{\tilde{\varphi}}(\Gamma\times C)\quad\ \forall\Gamma\in\mathcal{B}(S),C\in\mathcal{B}(A).$$

So, part (b) holds.

(c) First, we prove the 'only if' part. By part (b), for each occupation measure μ, there exists a stationary policy $\varphi\in\Phi$ such that $\mu(\mathrm{d}x\times\mathrm{d}a)=\mu^{\varphi}(\mathrm{d}x\times\mathrm{d}a)$. Thus, we assume that $\mu=\mu^{\varphi}$.

For any bounded measurable function $g(x)$ on S, we have

$$\begin{aligned}
&\int_S g(x)\hat{\mu}^{\varphi}(\mathrm{d}x)=E_{\gamma}^{\varphi}\big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)g(x_m)\big]\\
=&\int_S g(x)\gamma(\mathrm{d}x)+E_{\gamma}^{\varphi}\Big[E_{\gamma}^{\varphi}\big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)g(x_m)\big|h_{m-1},a_{m-1}\big]\Big]\\
=&\,E_{\gamma}^{\varphi}\Big[\sum_{m=1}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)\int_S g(y)Q(\mathrm{d}y|x_{m-1},a_{m-1})\Big]+\int_S g(x)\gamma(\mathrm{d}x)\\
=&\int_S g(x)\gamma(\mathrm{d}x)+\int_{S\times A}\alpha(x)\int_S g(y)Q(\mathrm{d}y|x,a)\mu^{\varphi}(\mathrm{d}x\times\mathrm{d}a),
\end{aligned}$$

which implies the 'only if' part.

Second, we prove the 'if' part. Suppose that $\mu(\mathrm{d}x\times\mathrm{d}a)$ is an arbitrarily fixed measure on $\mathcal{B}(S\times A)$ concentrated on K such that (5) and (6) hold. By [10, Proposition D.8], there exists a policy $\varphi\in\Phi$ such that $\mu(\mathrm{d}x\times\mathrm{d}a)=\hat{\mu}(\mathrm{d}x)\varphi(\mathrm{d}a|x)$. Next, we show that $\mu(\mathrm{d}x\times\mathrm{d}a)=\mu^{\varphi}(\mathrm{d}x\times\mathrm{d}a)$, i. e., $\mu(\mathrm{d}x\times\mathrm{d}a)$ is an occupation measure.

For arbitrarily fixed bounded measurable function $g(x,a)$ on K, by Lemma 3.3, we have

$$\int_{S\times A} g(x,a)\mu(\mathrm{d}x\times\mathrm{d}a)=\int_{S\times A} g(x,a)\hat{\mu}(\mathrm{d}x)\varphi(\mathrm{d}a|x)=\int_S g(x,\varphi)\hat{\mu}(\mathrm{d}x)$$
$$=\int_S V(x,\varphi,g)\hat{\mu}(\mathrm{d}x)-\int_{S\times A}\alpha(x)\int_S V(y,\varphi,g)Q(dy|x,a)\mu(\mathrm{d}x\times\mathrm{d}a),$$

which together with (6) yields, for the bounded measurable function $V(x,\varphi,g)$,

$$\int_S V(x,\varphi,g)\gamma(\mathrm{d}x)$$
$$=\int_S V(x,\varphi,g)\hat{\mu}(\mathrm{d}x)-\int_{S\times A}\alpha(x)\int_S V(y,\varphi,g)Q(dy|x,a)\mu(\mathrm{d}x\times\mathrm{d}a),$$

and then

$$\int_{S\times A} g(x,a)\mu(\mathrm{d}x\times\mathrm{d}a)=\int_S V(x,\varphi,g)\gamma(\mathrm{d}x)$$
$$=E_\gamma^\varphi\Big[\sum_{m=0}^{\infty}\prod_{k=0}^{m-1}\alpha(x_k)g(y)Q(dy|x_m,a_m)\Big]=\int_{S\times A} g(x,a)\mu^\varphi(\mathrm{d}x\times\mathrm{d}a).$$

Since $g(x,a)$ is arbitrarily fixed, it follows that $\mu(\mathrm{d}x\times\mathrm{d}a)=\mu^\varphi(\mathrm{d}x\times\mathrm{d}a)$.

Hence, part (c) holds. □

Under Assumption 2.5, it is obvious that $\mathcal{D}$ is convex by Theorem 3.4(c), and by Theorem 3.4(b), here is the following results.

Corollary 3.5. (a) If a finite measure μ on $S\times A$ satisfies (5) and (6), then there exists a stationary policy $\varphi\in\Phi$ such that $\mu=\mu^\varphi$, and φ can be obtained from the following decomposition: $\mu(\mathrm{d}x\times\mathrm{d}a)=\hat{\mu}(\mathrm{d}x)\varphi(\mathrm{d}a|x)$, which will be written as $\mu=\hat{\mu}\circ\varphi$.

(b) For each policy $\pi\in\Pi$, there exists a stationary policy $\varphi\in\Phi$ such that $V^l(\pi)=V^l(\varphi)$ for all $l=0,1,\ldots,q$. Furthermore, if there exists an optimal policy $\pi\in\Pi$ for COP, then there will be an optimal stationary policy $\varphi\in\Phi$ for COP.

Let $\mathcal{M}(K)$ be the set of finite measures on K and $\tau(\mathcal{M}(K))$ the weak topology on it generated by the set of bounded continuous functions on K. Then, by [18], $(\mathcal{M}(K),\tau(\mathcal{M}(K)))$ is metrizable. Thus, $(\mathcal{M}(K),\tau(\mathcal{M}(K)))$ is regarded as a metric space below, so is $(\mathcal{D},\tau(\mathcal{D}))$.

Lemma 3.6. (a) Let $\{\varphi_m\}$ be a sequence in Φ such that $\varphi_m(\cdot|x)\to\varphi(\cdot|x)$ for each $x\in S$, and $\{\nu_m\}$ is a finite measure sequence in $\mathcal{M}(S)$ such that $\nu_m\to\nu$ weakly. If $\mu_m:=\nu_m\circ\varphi_m$ and $\mu:=\nu\circ\varphi$, then $\mu_m\to\mu$ weakly.

(b) If $\{\mu_m\}\subset\mathcal{D}$ such that $\mu_m\to\mu$ weakly, then there exist a subsequence $\{\mu_{m_k}\}$ of $\{\mu_m\}$, a corresponding sequence $\{\varphi_{m_k}\}\subset\Phi$ and a stationary policy $\varphi\in\Phi$, such that $\varphi_{m_k}\to\varphi$ weakly and $\mu_{m_k}=\hat{\mu}_{m_k}\circ\varphi_{m_k}$ converges weakly to $\mu=\hat{\mu}\circ\varphi$.

P r o o f . Since the proof of part (a) is similar as part (b), next, we only give the proof of part (b).

For the sequence $\{\mu_m\} \subset \mathcal{D}$, by [10, Proposition D.8] there exists a sequence $\{\varphi_m\} \subset \Phi$ such that $\mu_m(\mathrm{d}x \times \mathrm{d}a) = \hat{\mu}_m(\mathrm{d}x)\varphi_m(\mathrm{d}a|x)$.

For each $x \in S$, by the compactness of $\mathbb{P}(A(x))$ there exist a subsequence $\{\varphi_{m_k}(\cdot|x)\}$ of $\{\varphi_m(\cdot|x)\}$ and a policy $\varphi \in \Phi$ such that $\varphi_{m_k}(\cdot|x) \to \varphi(\cdot|x)$ weakly. Pick an arbitrary $h \in C_b(K)$, where $C_b(K)$ is the set of the bounded continuous functions on K, and then, for each $x \in S$, we have

$$h(x, \varphi_{m_k}) = \int_{A(x)} h(x,a)\varphi_{m_k}(\mathrm{d}a|x) \to \int_{A(x)} h(x,a)\varphi(\mathrm{d}a|x) = h(x,\varphi). \tag{10}$$

On the other hand, for any $\tilde{\mu} \in \mathcal{D}$, there exists $\tilde{\varphi} \in \Phi$ such that

$$\begin{aligned}
\hat{\tilde{\mu}}(\mathrm{d}x) &= \hat{\tilde{\mu}}^{\tilde{\varphi}}(\mathrm{d}x) = E_\gamma^{\tilde{\varphi}}\big[\sum_{j=0}^{\infty}\prod_{k=0}^{j-1}\alpha(x_k)I_{\{x_j\in dx\}}\big] \\
&= \gamma(\mathrm{d}x) + E_\gamma^{\tilde{\varphi}}\Big[E_\gamma^{\tilde{\varphi}}\big[\sum_{j=1}^{\infty}\prod_{k=0}^{j-1}\alpha(x_k)I_{\{x_j\in \mathrm{d}x\}}\big]\big|h_{j-1}, a_{j-1}\Big] \\
&= \gamma(\mathrm{d}x) + [E_\gamma^{\tilde{\varphi}}\big[\sum_{j=1}^{\infty}\prod_{k=0}^{j-1}\alpha(x_k)Q(\mathrm{d}x|x_{j-1}, a_{j-1})\big] \\
&= \gamma(\mathrm{d}x) + \int_S \alpha(y)Q(\mathrm{d}x|y,\tilde{\varphi})\hat{\tilde{\mu}}^{\tilde{\varphi}}(\mathrm{d}y).
\end{aligned} \tag{11}$$

By iteration, we can get

$$\hat{\tilde{\mu}}^{\varphi}(\mathrm{d}x) \le \gamma(\mathrm{d}x) + \sum_{j=1}^{N-1}\alpha^j\int_S Q^j(\mathrm{d}x|y,\tilde{\varphi})\gamma(\mathrm{d}y) + \alpha^N\int_S Q^N(\mathrm{d}x|y,\tilde{\varphi})\hat{\tilde{\mu}}^{\tilde{\varphi}}(\mathrm{d}y),$$

letting $N \to \infty$ in the above inequality, and obtain

$$\hat{\tilde{\mu}}^{\tilde{\varphi}}(\mathrm{d}) \le \gamma(\mathrm{d}x) + \sum_{j=1}^{\infty}\alpha^j\int_S Q^j(\mathrm{d}x|y,\tilde{\varphi})\gamma(\mathrm{d}y). \tag{12}$$

Now, let

$$v_0(\cdot) := \gamma(\cdot) + \sum_{j=1}^{\infty}\alpha^j\int_S Q^j(\cdot|y,\varphi)\gamma(\mathrm{d}y), \tag{13}$$

then, it is clear that v_0 is a finite measure, and for sufficiently large k, we have $\hat{\mu}_{m_k} \le v_0$ by (12) and $Q^j(\cdot|y,\varphi_{m_k}) \to Q^j(\cdot|y,\varphi)$. Since $|h(x,\varphi)| \le \|h\| < \infty$, it holds that

$$\int_S h^+(x,\varphi)v_0(\mathrm{d}x) \le \int_S \|h\|v_0(\mathrm{d}x) < \infty.$$

Note that $\hat{\mu}_{m_k} \to \hat{\mu}$ weakly, by [12, Theorem 2.1b], we have

$$\int_S h^+(x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) \to \int_S h^+(x,\varphi)\hat{\mu}(\mathrm{d}x),$$

and also $\int_S h^-(x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) \to \int_S h^-(x,\varphi)\hat{\mu}(\mathrm{d}x)$, which yields that

$$\int_S h(x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) \to \int_S h(x,\varphi)\hat{\mu}(\mathrm{d}x). \tag{14}$$

Hence, by (10) and (14), we can get

$$\begin{aligned}
&\Big|\int_{S\times A} h(x,a)\mu_{m_k}(\mathrm{d}x\times d\mathrm{d}a) - \int_{S\times A} h(x,a)\mu(\mathrm{d}x\times \mathrm{d}a)\Big| \\
&= \Big|\int_S h(x,\varphi_{m_k})\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S h(x,\varphi)\hat{\mu}(\mathrm{d}x)\Big| \\
&\le \Big|\int_S h(x,\varphi_{m_k})\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S h(x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x)\Big| + \Big|\int_S h(x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S h(x,\varphi)\hat{\mu}(\mathrm{d}x)\Big| \\
&\to 0,
\end{aligned}$$

that is, $\mu_{m_k} \to \mu(=\hat{\mu}\circ\varphi)$ weakly. □

Lemma 3.7. Suppose that Assumption 2.5 holds. If $\{\mu_m\} \subset \mathcal{D}$ satisfies $\mu_m \to \mu$ weakly, then

$$\lim_{m\to\infty}\int_K c^l(x,a)\mu_m(\mathrm{d}x\times\mathrm{d}a) = \int_K c^l(x,a)\mu(\mathrm{d}x\times\mathrm{d}a) \quad l = 0,1,\ldots,q.$$

P r o o f. For fixed $l = 0,1,\ldots,q$, let $f_m := \int_K c^l(x,a)\mu_m(\mathrm{d}x\times\mathrm{d}a)$ $(m = 0,1,\ldots)$ and $\{f_{m_k}\}$ be an arbitrary subsequence of $\{f_m\}$. For the subsequence $\{\mu_{m_k}\}$ of $\{\mu_m\}$, by Lemma 3.6(b) there exist a corresponding sequence $\{\varphi_{m_k}\} \subset \Phi$ and a stationary policy $\varphi \in \Phi$ such that $\varphi_{m_k} \to \varphi$ weakly and $\mu_{m_k} = \hat{\mu}_{m_k}\circ\varphi_{m_k}$ converges weakly to $\mu = \hat{\mu}\circ\varphi$.

In addition, for the finite measure v_0 as in (13), by Assumption 2.5, we have

$$\begin{aligned}
\int_S \omega(x)v_0(\mathrm{d}x) &\le \int_S \omega(x)\gamma(\mathrm{d}x) + \sum_{m=1}^{\infty}(\alpha\beta)^m\int_S \omega(y)\gamma(\mathrm{d}y) \\
&\le \frac{1}{1-\alpha\beta}\int_S \omega(x)\gamma(\mathrm{d}x) < \infty.
\end{aligned} \tag{15}$$

Since, for each $x \in S$, $c^l(x,\cdot)$ $(0 \le l \le q)$ and $Q(\Gamma|x,\cdot)$ are continuous and bounded on $A(x)$, which yields that $c^l(x,\varphi_{m_k}) \to c^l(x,\varphi)$ by $\varphi_{m_k}(\cdot|x) \to \varphi(\cdot|x)$ weakly. And, $\hat{\mu}_{m_k}(\Gamma) \to \hat{\mu}_{m_k}(\Gamma)$ for arbitrary $\Gamma \in \mathcal{B}(S)$ by $\hat{\mu}_{m_k} \to \hat{\mu}$ weakly, $\hat{\mu}_{m_k} \le v_0$ for sufficiently large k by (11)-(12), thus, by the Generalized Dominated Convergence Theorem for Measures [12, Theorem 2.2], it holds that

$$\int_S c^l(x,\varphi_{m_k})\hat{\mu}_{m_k}(\mathrm{d}x) \to \int_S c^l(x,\varphi)\hat{\mu}(\mathrm{d}x),$$

i. e.,

$$f_{m_k} = \int_K c^l(x,a)\mu_{m_k}(\mathrm{d}x \times \mathrm{d}a) \to \int_K c^l(x,a)\mu(\mathrm{d}x \times \mathrm{d}a).$$

Since the subsequence $\{f_{m_k}\}$ of $\{f_m\}$ is arbitrary, it follows that

$$f_m = \int_K c^l(x,a)\mu_m(\mathrm{d}x \times \mathrm{d}a) \to \int_K c^l(x,a)\mu(\mathrm{d}x \times d\mathrm{d}a) \quad l = 0, 1, \ldots, q.$$

□

Definition 3.8. A family D of finite measures on K is called tight, if for arbitrary $\varepsilon > 0$, there exists a compact subset $\widetilde{\Gamma} \subseteq K$ such that $\mu(\widetilde{\Gamma}) > 1 - \varepsilon$ for each $\mu \in D$.

Lemma 3.9. Under Assumption 2.5, the family $\mathcal{D}$ of all occupation measures is compact in the weak topology.

P r o o f . Let $\hat{\mathcal{D}}$ be the space of the marginal measures of occupation measures with the weak topology on S. Then, by Assumption 2.5(b), (12) and (15), we have

$$\int_S \omega(x)\hat{\mu}^\pi(\mathrm{d}x) = \int_S \omega(x)\hat{\mu}^\varphi(\mathrm{d}x) \le \int_S \omega(x)v_0(\mathrm{d}x) < \infty.$$

Hence, by [10, Proposition E.8], $\hat{\mathcal{D}}$ is tight. By Prokhorov's Theorem [10, Theorem E.6], $\hat{\mathcal{D}}$ is relatively compact. Now, we assert that $\mathcal{D}$ is relatively compact. In fact, for any $\{\mu_m\} \subset \mathcal{D}$, there exists a sequence $\{\varphi_m\} \subset \Phi$ such that $\mu_m(\mathrm{d}x \times \mathrm{d}a) = \hat{\mu}_m(\mathrm{d}x)\varphi(\mathrm{d}a|x)$, and further, there exist a subsequence $\{\hat{\mu}_{m_k}\}$ and the corresponding subsequence $\{\varphi_{m_k}\}$ such that $\hat{\mu}_{m_k} \to \nu$ weakly in $\mathcal{M}(S)$ and $\varphi_{m_k}(\cdot|x) \to \varphi(\cdot|x)$ weakly in Φ. Therefore, by Lemma 3.6(a), $\mu_{m_k} \to \nu \circ \varphi$, i. e., $\mathcal{D}$ is relatively compact.

Next, we prove that $\mathcal{D}$ is close. Choose arbitrarily a sequence $\{\mu_m\} \subset \mathcal{D}$ such that $\mu_m \to \mu$ weakly, by Lemma 3.6(b), there exist a subsequence $\{\mu_{m_k}\}$, a sequence $\{\varphi_{m_k}\} \subset \Phi$ and a stationary policy $\varphi \in \Phi$, such that $\varphi_{m_k} \to \varphi$ weakly and $\mu_{m_k} = \hat{\mu}_{m_k} \circ \varphi_{m_k}$ converges weakly to $\mu = \hat{\mu} \circ \varphi$. By Theorem 3.4(c) one can obtain that μ_m satisfies (5) and (6) for each $m = 0, 1, \ldots$, and then μ satisfies (5). It remains to verify that (6) holds for $\hat{\mu}$.

In fact, by (11), we have

$$\begin{aligned}
&\Big|\hat{\mu}_{m_k} - \big(\gamma(\mathrm{d}x) + \int_{S\times A} \alpha(x)Q(dy|x,a)\mu(\mathrm{d}x \times \mathrm{d}a)\big)\Big| \\
&= \Big| \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi_{m_k})\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi)\hat{\mu}(\mathrm{d}x)\Big| \\
&\le \Big| \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi_{m_k})\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x)\Big| \\
&\quad + \Big| \int_S \alpha(x)Q(\mathrm{d}|x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) - \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi)\hat{\mu}(\mathrm{d}x)\Big|. \qquad (16)
\end{aligned}$$

For the subsequence $\{\mu_{m_k}\}$, by (12)-(13) and $Q^j(\cdot|y,\varphi_{m_k}) \to Q^j(\cdot|y,\varphi)$, we have $\hat{\mu}_{m_k} \leq v_0$ for sufficiently large k. Hence, by [12, Theorem 2.1*b*], we can obtain

$$\int_S \alpha(x)Q(\mathrm{d}y|x,\varphi)\hat{\mu}_{m_k}(\mathrm{d}x) \to \int_S \alpha(x)Q(\mathrm{d}y|x,\varphi)\hat{\mu}(\mathrm{d}x),$$

and note that $Q(dy|x,\varphi_{m_k}) \to Q(dy|x,\varphi)$ since $Q(dy|x,\cdot)$ are continuous and bounded on $A(x)$, which together with (16) gives

$$\hat{\mu}_{m_k} \to \gamma(\mathrm{d}x) + \int_{S\times A} \alpha(x)Q(\mathrm{d}y|x,a)\mu(\mathrm{d}x\times \mathrm{d}a)$$

Then, it follows that $\hat{\mu} = \gamma(\mathrm{d}x) + \int_{S\times A}\alpha(x)Q(\mathrm{d}y|x,a)\mu(\mathrm{d}x\times \mathrm{d}a)$, i. e., (6) holds for $\hat{\mu}$.

Therefore, $\mathcal{D}$ is compact in the weak topology. □

3.2. Existence of the optimal policy for constrained MDPs

Let

$$\mathbf{F} := \big\{\mu \in \mathcal{D}\big| \int_K c^l(x,a)\mu(\mathrm{d}x\times \mathrm{d}a) \leq d^l,\ 1 \leq l \leq q\big\}.$$

By Assumption 2.5, Remark 3.2, Lemma 3.9 and (2), the constrained optimality problem COP is equivalent to the following well-defined convex program:

$$\mathrm{COP}': \qquad \text{minimize } \int_K c^0(x,a)\mu(\mathrm{d}x\times \mathrm{d}a) \ \text{ over } \mu \in \mathbf{F}.$$

$\mathbf{F}$ is called the set of all feasible policies for COP′ and the optimal value of COP′ is denoted by $\inf \mathrm{COP}'$. If there exists a feasible solution $\mu \in \mathbf{F}$ to COP′ such that

$$\int_K c^0(x,a)\mu(\mathrm{d}x\times \mathrm{d}a) = \inf \mathrm{COP}',$$

then μ is called an optimal solution to COP′. By the standing assumption $U \neq \emptyset$ in Section 2, it is clear that $\mathbf{F} \neq \emptyset$.

Furthermore, for any $\mu \in \mathbf{F}$, by Assumption 2.5(b) and (12) we have

$$\begin{aligned}\int_{S\times A} c^0(x,a)\mu(\mathrm{d}x\times \mathrm{d}a) \ &\leq\ L\int_{S\times A}\omega(x)\mu(\mathrm{d}x\times \mathrm{d}a) \leq L\int_S \omega(x)\hat{\mu}(\mathrm{d}x)\\ &\leq\ L\int_S \omega(x)v_0(\mathrm{d}x) < \infty,\end{aligned}$$

and then $\inf \mathrm{COP}' < \infty$.

Lemma 3.10. Under Assumption 2.5, the set $\mathbf{F} \subset \mathcal{D}$ of feasible solutions to COP′ is compact in the weak topology.

P r o o f . Let $\{\mu_{m_k}\}$ be an arbitrary subsequence of $\{\mu_m\} \subset \mathbf{F} \subset \mathcal{D}$. By Lemma 3.9, there exists a subsequence of $\{\mu_{m_k}\}$, still denoted as $\{\mu_{m_k}\}$ without loss of generality, such that $\{\mu_{m_k}\} \to \mu$ weakly in $\mathcal{D}$. Hence, it only remains to prove that μ satisfies the constraints for COP$'$. Indeed, by Lemma 3.7, we have

$$\int_K c^l(x,a)\mu(\mathrm{d}x \times \mathrm{d}a) = \lim_{m\to\infty} \int_K c^l(x,a)\mu_{m_k}(\mathrm{d}x \times \mathrm{d}a) \le d^l. \quad l = 1, 2, \ldots, q.$$

Therefore, it holds that $\mu \in \mathbf{F}$. □

Theorem 3.11. Suppose that Assumption 2.5 holds. Then,

(a) COP$'$ is solvable.

(b) there exists an optimal stationary policy $\varphi \in \Phi$ for COP.

P r o o f . (a) Let $\{\mu_m\}$ be a minimizing sequence for COP$'$, that is, $\{\mu_m\} \subset \mathbf{F}$ such that $\{\int_K c^0(x,a)\mu_m(\mathrm{d}x \times \mathrm{d}a)\}$ is a decreasing sequence and

$$\lim_{m\to\infty} \int_K c^0(x,a)\mu_m(\mathrm{d}x \times d\mathrm{d}a) = \inf \mathrm{COP}',$$

where $\inf \mathrm{COP}'$ stands for the value of COP$'$. Since $\mathbf{F}$ is compact by Lemma 3.10, it follows that there exists a subsequence $\{\mu_{m_k}\}$ of $\{\mu_m\}$ in $\mathbf{F}$ such that $\mu_{m_k} \to \mu$ weakly in $\mathbf{F}$. Then, by Lemma 3.7, we have

$$\int_K c^0(x,a)\mu(\mathrm{d}x \times \mathrm{d}a) = \lim_{k\to\infty} \int_K c^0(x,a)\mu_{m_k}(\mathrm{d}x \times \mathrm{d}a) = \inf \mathrm{COP}'.$$

Thus, μ is an optimal solution for COP$'$.

(b) It is clear that the solvability of COP$'$ implies the solvability of COP, which together with Corollary 3.5 yields that part (b) holds. □

4. AN EXAMPLE

In this section, we give an application example to show that all the conditions imposed in this paper can be satisfied simultaneously, so that the results for the existence of the optimal policy are directly applicable.

Example 4.1. (*A Cash-balance system*) Consider the controlled cash-balance system $\mathcal{M}$, which satisfies

$$x_{n+1} = x_n + a_n + \varepsilon_n, \quad n = 0, 1, \ldots, \tag{17}$$

where the state x_n and action a_n denote the amount of cash balance, and a withdrawal of size $-a_n$ (if $a_n < 0$) of the money in cash, or a supply in the amount a_n (if $a_n > 0$) at time n, respectively. Let the state space be $S := (-\infty, +\infty)$. If the state of the system is $x \in (-\infty, +\infty)$, a decision-maker takes an action a in a given set $A(x)$, which is

assumed to be $[-|x|, |x|]$. The disturbances $\{\varepsilon_n, n \geq 0\}$ are assumed to be independent standard normal random variables, that is, the transition law $Q(\cdot|x, a)$ is given by

$$Q(\Gamma|x, a) = \int_{\Gamma} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-x-a)^2}{2}} \mathrm{d}y \qquad \forall\ (x, a) \in K,\ \Gamma \in \mathcal{B}(S).$$

When the amount of the cash balance is x in this system, it incurs the holding cost $c^0(x, a)$ and the running cost $c^1(x, a)$ during a stage. We assume that $c^l(x, a)$ $(l = 0, 1)$ are continuous on $a \in [-|x|, |x|]$ and there exists a constant $L > 0$ such that $|c^l(x, a)| \leq L(x^2 + 1)$ for all $x \in S$, $a \in [-|x|, |x|]$ and $l = 0, 1$. In addition, let γ be an initial distribution on S such that $\int_{-\infty}^{+\infty} x^2 \gamma(\mathrm{d}x) < \infty$, and assume that the discount factors $\alpha(x)$ is measurable on S and satisfies that $\sup_{x \in S} \alpha(x) < \frac{1}{4}$.

A decision-maker wishes to minimize the expected discounted holding cost, while the expected discounted running cost is maintained bounded above by a positive constant d.

Proposition 4.2. The controlled system $\mathcal{M}$ in Example 4.1 verifies Assumption 2.5, and then there exists an optimal stationary policy.

Proof. For $x \in S$, let $\omega(x) = x^2 + 1$. Then, it is clear that $\omega(x)$ satisfies that $\omega(x) \geq 1$ on S and it is strictly unbounded. Moreover, by $\int_{-\infty}^{+\infty} x^2 \gamma(\mathrm{d}x) < \infty$ we have $\int_S \omega(x)\gamma(\mathrm{d}x) < \infty$ and

$$\int_S \omega(y) Q(\mathrm{d}y|x, a) = \int_{-\infty}^{+\infty} \omega(y) Q(\mathrm{d}y|x, a) = (x + a)^2 + 2 < 4\omega(x),$$

$$|c^l(x, a)| \leq L\omega(x) \qquad \forall\ x \in X, a \in A(x) \text{ and } l = 0, 1.$$

Thus, Assumption 2.5(b) holds. In addition, it is clear that Assumption 2.5(a) and (c)-(e) are satisfies by the [19, Proposition 4.1] and the definition of $\mathcal{M}$ in Example 4.1. □

5. CONCLUSION

This paper studies the solvability of constrained optimality problem of DTMDPs with varying discount factors, Borel states space, compact Borel action sets, and possibly unbounded cost functions. By means of the so-called occupation measure of policies and its properties, we prove the existence of an optimal randomized stationary policies (see Theorem 3.11). This result extends those of constrained DTMDPs with a constant discount factor in [3] and constrained DTMDPs with the bounded costs from below as in [21] under weaker conditions.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (Grant No. 11961005) and Characteristic Innovation Projects of General Colleges and Universities in Guangdong Province (Grant No. 2018KTSCX253). The authors also thank the Associate Editor and the referee for many valuable comments and suggestions which have improved this paper.

(Received February 23, 2020)

REFERENCES

[1] E. Altman: Denumerable constrained Markov decision processes and finite approximations. Math. Meth. Operat. Res. *19* (1994), 169–191. DOI:10.1155/S1073792894000188

[2] E. Altman: Constrained Markov decision processes. Chapman and Hall/CRC, Boca Raton 1999.

[3] J. Alvarez-Mena and O. Hernández-Lerma: Convergence of the optimal values of constrained Markov control processes. Math. Meth. Oper. Res. *55* (2002), 461–484.

[4] V. Borkar: A convex analytic approach to Markov decision processes. Probab. Theory Relat. Fields *78* (1988), 583–602.

[5] J. González-Hernández and O. Hernández-Lerma: Extreme points of sets of randomized strategies in constrained optimization and control problems. SIAM. J. Optim. *15* (2005), 1085–1104. DOI:10.1137/040605345

[6] X. P. Guo, A. Hernández-del-Valle, and O. Hernández-Lerma: First passage problems for nonstationary discrete-time stochastic control systems. Europ. J. Control *18* (2012), 528–538. DOI:10.3166/EJC.18.528-538

[7] X. P. Guo and W. Z. Zhang: Convergence of controlled models and finite-state approximation for discounted continuous-time Markov decision processes with constraints. Europ. J, Oper. Res. *238* (2014), 486–496. DOI:10.1016/j.ejor.2014.03.037

[8] X. P. Guo, X. Y. Song, and Y. Zhang: First passage criteria for continuous-time Markov decision processes with varying discount factors and history-dependent policies. IEEE Trans. Automat. Control *59* (2014), 163–174. DOI:10.1109/tac.2013.2281475

[9] O. Hernández-Lerma and J. González-Hernández: Constrained Markov Decision Processes in Borel spaces: the discounted case. Math. Meth. Operat. Res. *52* (2000), 271–285. DOI:10.1155/S1073792800000167

[10] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes. Springer-Verlag, New York 1996.

[11] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes. Springer-Verlag, New York 1999.

[12] O. Hernández-Lerma and J. B. Lasserre: Fatou's lemma and Lebesgue's convergence theorem for measures. J. Appl. Math. Stoch. Anal. *13(2)* (2000), 137–146. DOI:10.1155/s1048953300000150

[13] Y. H. Huang and X. P. Guo: First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs. Acta. Math. Appl. Sin-E. *27(2)* (2011), 177–190. DOI:10.1007/s10255-011-0061-2

[14] Y. H. Huang, Q. D. Wei, and X. P. Guo: Constrained Markov decision processes with first passage criteria. Ann. Oper. Res. *206* (2013), 197–219. DOI:10.1007/s10479-012-1292-1

[15] X. Mao and A. Piunovskiy: Strategic measures in optimal control problems for stochastic sequences. Stoch. Anal. Appl. *18* (2000), 755–776. DOI:10.1080/07362990008809696

[16] A. Piunovskiy: Optimal Control of Random Sequences in Problems with Constraints. Kluwer Academic, Dordrecht 1997.

[17] A. Piunovskiy: Controlled random sequences: the convex analytic approach and constrained problems. Russ. Math. Surv., *53* (2000), 1233–1293. DOI:10.1070/rm1998v053n06abeh000090

[18] Y. Prokhorov: Convergence of random processes and limit theorems in probability theory. Theory Probab Appl. *1* (1956), 157–214. DOI:10.1137/1101016

[19] Q. D. Wei and X. P. Guo: Markov decision processes with state-dependent discount factors and unbounded rewards/costs. Oper. Res. Lett. *39* (2011), 369–374. DOI:10.1016/j.orl.2011.06.014

[20] X. Wu and X. P. Guo: First passage optimality and variance minimization of Markov decision processes with varying discount factors. J. Appl. Probab. *52(2)* (2015), 441–456. DOI:10.1017/S0021900200012560

[21] Y. Zhang: Convex analytic approach to constrained discounted Markov decision processes with non-constant discount factors. TOP *21* (2013), 378–408. DOI:10.1007/s11750-011-0186-8

Xiao Wu, School of Mathematics and Statistics, Zhaoqing University, Zhaoqing 526061*. P. R. China.*
e-mail: jxwuxiao@126.com

Yanqiu Tang, Corresponding author. School of Mathematics and Statistics, Zhaoqing University, Zhaoqing 526061*. P. R. China.*
e-mail: tangyanqiu1985@126.com