Title	Constrained Markov Decision Processes With Compact State And Action Sspaces : The Average Case (Dynamic Decision Systems under Uncertain Environments)
Author(s)	Kurano, Masami; Nakagami, Jun-ichi; Huang, Youqiang
Citation	数理解析研究所講究録 (1998), 1048: 1-12
Issue Date	1998-05
URL	http://hdl.handle.net/2433/62191
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
1048巻 1998年 1-12

Constrained Markov Decision Processes With Compact State And Action Spaces: The Average Case

千葉大・教育　蔵野 正美　(Masami KURANO)
千葉大・理　中神 潤一　(Jun-ichi NAKAGAMI)
システム開発ＩＧ　黄 佑強　(Youqiang HUANG)

Abstract

Constrained Markov decision processes with compact state and action spaces are studied under long-run average reward or cost criteria. And introducing a corresponding Lagrange function, a saddle-point theorem is given, by which the existence of a constrained optimal pair of initial state distribution and policy is shown. Also, under the hypothesis of Doeblin, the functional characterization of a constrained optimal policy is obtained.

Keywords: constrained Markov decision processes; average criteria; compact state and action spaces; Lagrange technique; saddle-point; constraied optimal pair.

AMS subject classifications. 93E20; 90C40

1 Introduction And Notation

The linear programming (LP) formulation of unconstrained or constrained Markov decision processes (MDPs) has been studied by many authors (for example, [2,5,7,9,13,14] and their references). However, most of the related papers are concerned with the case of denumerable (mainly finite) states. As for the general state space, the study of the LP formulation of unconstrained MDPs has been done by Hernández-Lerma and Lasserre [11] and Hernández-Lerma and Hernández-Hernández [12], in which it has been shown that there is no duality gap between the corresponding LP and its dual LP and that strong duality condition holds using the basic facts of LP in general vector space [1].

Kurano [15] has considered the case in which the state space is compact and Doeblin's condition [10] holds and proved the existence of an optimal stationary policy for unconstrained MDPs with average cost criterion by extracting a randomized stationary policy from a limit point of empirical processes.

In this paper we consider constrained MDPs in the framework similar to [15]. Associated with constrained MDPs with average criteria, a corresponding linear program (P) and its dual (P*) are formulated by the approach of Anderson and Nash [1]. Introducing the Lagrange function we will give a saddle-point theorem for constrained MDPs by use of the absence of a duality gap between (P) and (P*) under compactness assumptions. The saddle-point statement is applied to prove the existence of a constrained optimal pair of initial state distribution and policy and obtain its functional characterization under the hypohesis of Doeblin.

In the remainder of this section we shall establish the notation that will be used throughout the paper and define the problem to be examined. Also, a constrained optimal pair of state and policy and constrained optimal policy are defined.

In Section 2 the saddle-point statements for constrained MDPs are given under a regularity condition, whose results are applied to obtain the characterization of a constrained optimal policy in Section 3.

A Borel set is a Borel subset of a complete separable metric space. For a Borel set X, $\mathcal{B}_X$ denotes the σ-algebra of Borel subsets of X. A Markov decision process with multiple constraints is a controlled dynamic system defined by the following objects: $S, \{A(x), x \in S\}, Q, r, c_l$ $(l = 1, \cdots, k)$, where S is any Borel set representing the state space of some system and for each $x \in S$, the admissable action space $A(x)$ is a non-empty subset of some Borel set A such that $\{(x, a) : x \in S, a \in A(x)\}$ is an element of $\mathcal{B}_S \times \mathcal{B}_A$, the immediate reward function r and cost functions c_l $(l = 1, 2, \cdots, k)$ are real-valued function on $S \times A$, $Q(\cdot|x, a)$ is the law of motion , which is taken to be a stochastic kernel on $\mathcal{B}_S \times S \times A$; *i.e.*, for each $(x, a) \in S \times A$, $Q(\cdot|x, a)$ is a probability measure on $\mathcal{B}_S$; and for each $D \in \mathcal{B}_S$, $Q(D|\cdot)$ is a Borel measurable function on $S \times A$.

For any Borel set X, we denote by $C(X)$ the set of all bounded continuous functions on X.

Throughout this paper, the following assumptions will remain operative:

(i). S and $K := \{(x, a)|x \in S, a \in A(x)\}$ are compact;

(ii). $r \in C(S \times A)$ and $c_l \in C(S \times A)$ $(l = 1, \cdots, k)$;

(iii). wherever $x_n \to x$, $a_n \to a$, $Q(\cdot|x_n, a_n)$ converges weakly to $Q(\cdot|x, a)$.

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projections X_t, Δ_t on the tth factors S, A describe the state and action of the tth time of the process $(t \geq 0)$.

A policy is a sequence $\pi = (\pi_0, \pi_1, \cdots)$ such that, for each $t \geq 0$, π_t is a stochastic kernel on $\mathcal{B}_A \times S \times (A \times S)^t$ with $\pi_t(A(x_t)|x_0, a_0, \cdots, a_{t-1}, x_t) = 1$ for all $(x_0, a_0, \cdots, a_{t-1}, x_t) \in S \times (A \times S)^t$. Let Π denote the class of policies.

We denote by $P(A|S)$ the set of all stochastic kernels Φ on $\mathcal{B}_A \times S$ with $\Phi(A(x)|x) = 1$ for all $x \in S$.

A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a randomized stationary policy if there is a $\Phi \in P(A|S)$ such that $\pi_t(\cdot|x_0, a_0, \cdots, x_t) = \Phi(\cdot|x_t)$ for all $(x_0, a_0, \cdots, x_t) \in S \times (A \times S)^t$ and $t \geq 0$.

Denote the corresponding policy simply by Φ.

We denote by $B(S \to A)$ the set of all Borel measurable functions $f : S \to A$ with $f(x) \in A(x)$ for all $x \in S$.

A randomized stationary policy Φ is called stationary if there exists an $f \in B(S \to A)$ satisfying $\Phi(\{f(x)\}|x) = 1$ for all $x \in S$. Such a policy will be denoted by f.

The set of all randomized stationary and stationary policies will be denoted by Π_{RS} and Π_S respectively.

Note that Π_{RS} and Π_S are $P(A|S)$ and $B(S \to A)$ respectively.

Let $H_t = (X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t)$. We assume that for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$, $Prob(\Delta_t \in D_1|H_t) = \pi_t(D_1|H_t)$ and $Prob(X_{t+1} \in D_2|H_{t-1}, \Delta_{t-1}, X_t = x, \Delta_t = a) = Q(D_2|x, a)$ for each $D_1 \in \mathcal{B}_A$ and $D_2 \in \mathcal{B}_S$ $(t \geq 0)$.

For any Borel set X, let us denote by $P(X)$ the set of all probabitity measures on X.

For each $\pi \in \Pi$ and initial state distribution $\nu \in P(S)$, we can define the probability measure P_π^ν on Ω in an obvious way.

The average expected reward and costs from a policy $\pi \in \Pi$ with an initial distribution

$\nu \in P(S)$ are defined respectively by

$$J(\nu,\pi) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\pi^\nu[r(X_t,\Delta_t)], \tag{1.1}$$

$$I_l(\nu,\pi) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\pi^\nu[c_l(X_t,\Delta_t)], \ (l=1,\cdots,k). \tag{1.2}$$

where E_π^ν is the expectation with respect to P_π^ν.

For any given vector $\alpha = (\alpha_1,\alpha_2,\cdots,\alpha_k)$, let

$$U = \{(\nu,\pi) \in P(S)\times\Pi | I_l(\nu,\pi) \le \alpha_l, \ l=1,\cdots,k\}. \tag{1.3}$$

In this paper, assuming that $U \neq \emptyset$, we consider the following problem.

$$\begin{array}{lll} \text{Problem(A)}: & \text{maximize} & J(\nu,\pi), \\ & \text{subject to} & (\nu,\pi) \in U. \end{array}$$

The optimal solution of Problem(A), (ν^*,π^*), if it exists, is called a constrained optimal pair.

For each $\nu \in P(S)$, letting

$$J(\nu) = \sup\{J(\nu,\pi) | (\nu,\pi) \in U\},$$

a policy π^* is a constrained optimal policy if $J(\nu) = J(\nu,\pi^*)$ for all $\nu \in P(S)$ with $(\nu,\pi^*) \in U$.

For any $\mu \in P(K)$, let $\mu_S \in P(S)$ denote the marginal of μ on S; *i.e.*,

$$\mu_S(D) := \mu(D\times A), \ D \in \mathcal{B}_S.$$

Let $O(K) := \{\mu \in P(K) | \mu_S(\cdot) = \int Q(\cdot|x,a)\mu(d(x,a))\}$. Note that an element belonging to $O(K)$ is called an ergodic occupation measure in [6].

We shall use the following.

Lemma 1.1 (*cf.* [15]). For any $(\nu,\pi) \in P(S)\times\Pi$ and $g \in C(K)$, there exists a $\mu \in O(K)$ such that

$$\int g(x,a)\mu(d(x,a)) \ge \limsup_{T\to\infty} \frac{1}{T} E_\pi^\nu[\sum_{t=0}^{T-1} g(X_t,\Delta_t)]. \tag{1.4}$$

Lemma 1.1 asserts that for any given reward $g \in C(K)$ any pair $(\nu,\pi) \in P(S)\times\Pi$ can be replaced by an ergodic occupation measure $\mu \in O(K)$, the expected reward from which is at least as large as that from (ν,π).

2 Saddle-point Theorem For Constrained MDPs

In this section we define the Lagrangian associated with Problem(A), by which the saddle-point statement is given. And using the absence of a duality gap between a corresponding linear program (P) and its dual (P*) the saddle-point theorem is proved.

First we define the Lagrangian, L, that corresponds to Problem(A) as follows:

$$L((\nu,\pi),\lambda) = J(\nu,\pi) + \sum_{l=1}^{k} \lambda_l(\alpha_l - I_l(\nu,\pi)), \tag{2.1}$$

for any $(\nu,\pi) \in P(S) \times \Pi$ and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \in R_+^k$, where R_+^k is the positive orthant of a k-dimensional Euclidian space.

Henceforth $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \in R_+^k$ will be written simply by $\lambda \geq 0$.

For the Lagrangian approach in general non-linear programming problems we refer to ([3,16]).

The following saddle-point statement can be made, whose proof is similar to ([16], p.221, Theorem 2) and omitted.

Theorem 2.1 (*cf.* [16]). Suppose that there exists a $(\nu^*,\pi^*) \in P(S) \times \Pi$ and $\lambda^* \geq 0$ such that $L((\nu,\pi),\lambda)$ possess a saddle-point at (ν^*,π^*), λ^*; *i.e.*,

$$L((\nu,\pi),\lambda^*) \leq L((\nu^*,\pi^*),\lambda^*) \leq L((\nu^*,\pi^*),\lambda), \tag{2.2}$$

for all $(\nu,\pi) \in P(S) \times \Pi$ and $\lambda \geq 0$. Then, (ν^*,π^*) solves Problem(A) and is a constrained optimal pair.

From the above theorem, it will be of value to obtain sufficient conditions for the existence of a saddle-point of the Lagrangian L.

To this end let us introduce several notations, the linear program (P) and its dual (P*) corresponding to Problem(A).

For a Borel set X, let $B(X)$ be the set of all bounded Borel measurable functions on X.

For any $u \in B(X)$ and $\eta \in P(X)$, we denote the integral as follows:

$$\langle u, \eta \rangle = \int u d\eta.$$

For any $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \in R_+^k$, let

$$r(x,a|\lambda) := r(x,a) + \sum_{l=1}^{k} \lambda_l(\alpha_l - c_l(x,a)). \tag{2.3}$$

Here, consider the following linear programs (P) and (P*).

$$\begin{aligned}(\mathrm{P}):\ &\text{maximize} && \langle r,\mu\rangle,\\ &\text{subject to} && \mu\in O(K),\\ & && \langle c_l,\mu\rangle\le\alpha_l\ \ (l=1,\cdots,k).\end{aligned}\tag{2.4}$$

$$\begin{aligned}(\mathrm{P}^*):\ &\text{minimize} && \rho\\ &\text{subject to} && \rho+h(x)\ge r(x,a|\lambda)+\int h(x')Q(dx'|x,a)\\ & && \lambda\ge 0,\ h\in B(S).\end{aligned}\tag{2.5}$$

If the program (P) is consistent(resp., solvable), then its value is denoted by sup(P)(resp., max(P)); the value of the dual program (P^*) is written by inf(P*)(resp., min(P*)).

If sup(P) = inf(P*), it is said that there is no duality gap, whereas if they are solvable and max(P) = min(P*) we say that the strong duality condition holds ([1]).

Throughout this paper we assume that both programs (P) and (P*) are consistent.

For any $\mu\in P(K)$ and $\lambda\ge 0$, let

$$L(\mu,\lambda)=\int r(x,a|\lambda)\mu(d(x,a)).\tag{2.6}$$

It will become obvious that the above definition is compatible with the Lagrangian defined in (2.1).

The following lemma, the proof of which is obtained by applying Lemma 1.1., will play a crucial role in the sequel.

Lemma 2.1. For any $\lambda\ge 0$ and $(\nu,\pi)\in P(S)\times\Pi$, there exists $\mu\in O(K)$ such that

$$L(\mu,\lambda)\ge L((\nu,\pi),\lambda).\tag{2.7}$$

Proof. Applying Lemma 1.1 with $g(x,a)=r(x,a|\lambda)$, there exists $\mu\in O(K)$ such that

$$L(\mu,\pi)\ge\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}E_\pi^\nu[r(X_t,\Delta_t|\lambda)].\tag{2.8}$$

Thus, from $\lambda\ge 0$ and (2.1)-(2.2), we get

$$\begin{aligned}L(\mu,\lambda)&\ge\liminf_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}E_\pi^\nu[r(X_t,\Delta_t|\lambda)]\\ &\ge J(\nu,\pi)+\sum_{l=1}^{k}\lambda_l(\alpha_l-I_l(\nu,\pi))\\ &=L((\nu,\pi),\lambda),\end{aligned}$$

which completes the proof.□

Using Lemma 2.1 we can derive the following.

Corollary 2.1. (i) For each $\lambda \geq 0$,

$$\sup_{(\nu,\pi)\in P(S)\times\Pi} L((\nu,\pi),\lambda) = \sup_{\mu\in O(K)} L(\mu,\lambda).$$

$$\text{(ii)} \quad \sup_{(\nu,\pi)\in P(S)\times\Pi} \inf_{\lambda\geq 0} L((\nu,\pi),\lambda) \geq \sup_{\mu\in O(K)} \inf_{\lambda\geq 0} L(\mu,\lambda).$$

Proof. Let $\mu \in O(K)$. Then we decompose the probability measure μ into $\mu_S \in P(S)$ and $\Phi \in P(A|S)$ such that

$$\mu(D_1 \times D_2) = \int_{D_1} \Phi(D_2|x)\mu_S(dx) \quad \text{for any } D_1 \in \mathcal{B}_S \text{ and } D_2 \in \mathcal{B}_A,$$

(for example, [4]).

Letting $Q(\cdot|x,\Phi) = \int Q(\cdot|x,a)\Phi(da|x)$, it follows that

$$\mu_S(\cdot) = \int Q(\cdot|x,\Phi)\mu_S(dx),$$

which means that μ_S is a stationary absolute probability measure for the Markov process induced by $\{Q(\cdot|x,\Phi)\}$.

Hence, $L(\mu,\lambda) = L((\mu_S,\Phi),\lambda)$, which shows (i) together with Lemma 2.1. Also, (ii) follows similarly. □

Lemma 2.2. (i) $\inf(\mathrm{P}^*) \geq \inf_{\lambda\geq 0} \sup_{\mu\in O(K)} L(\mu,\lambda)$.

$$\text{(ii)} \quad \sup_{\mu\in O(K)} \inf_{\lambda\geq 0} L(\mu,\lambda) \geq \sup(\mathrm{P}).$$

Proof. Let (ρ,h,λ) be any feasible solution for (P^*) and $\mu \in O(K)$.

Integrating both sides of (2.5) with respect to μ, we get $\rho \geq L(\mu,\lambda)$, which implies (i).

For (ii), let μ be a feasible solution for (P). Then, since $\lambda \geq 0$, $L(\mu,\lambda) \geq \langle r,\mu\rangle$, which shows (ii). □

Applying general LP results ([1], Theorem 3.10 and Theorem 3.22), the following lemma can be obtained by checking the closedness of (P^*) from the compactness of K. The proof is tedious and omitted.

Lemma 2.3. There is no duality gap between (P) and (P^*) and (P) is solvable, *i.e.*,

$$\inf(\mathrm{P}^*) = \max(\mathrm{P}). \tag{2.9}$$

Observing Corollary 2.1 and Lemmas 2.2-2.3, the following minimax theorem holds.

Theorem 2.2. $\inf_{\lambda\geq 0}\sup_{(\nu,\pi)\in P(S)\times\Pi} L((\nu,\pi),\lambda) = \sup_{(\nu,\pi)\in P(S)\times\Pi}\inf_{\lambda\geq 0} L((\nu,\pi),\lambda).$ (2.10)

Henceforth the common value of (2.10) will be denoted by L^*.

Corollary 2.2. $\inf_{\lambda\geq 0}\sup_{\mu\in O(K)} L(\mu,\lambda) = \sup_{\mu\in O(K)}\inf_{\lambda\geq 0} L(\mu,\lambda) = L^*.$

Here we define a convex function on R_+^k by

$$L_1(\lambda) := \sup_{(\nu,\pi)\in P(S)\times\Pi} L((\nu,\pi),\lambda). \tag{2.11}$$

In order to prove the existence of a saddle-point, we need the following condition.

Condition A (Slater condition). There exists a $(\bar{\nu},\bar{\Phi})\in P(S)\times\Pi$ such that

$$I_l(\bar{\nu},\bar{\Phi}) < \alpha_l, \text{ for all } l\ (1\leq l\leq k).$$

Since $L((\bar{\nu},\bar{\Phi}),\lambda)\to\infty$ as $\|\lambda\|\to\infty$ under Condition A, the following lemma obviously holds, where $\|\cdot\|$ is a norm on R_+^k.

Lemma 2.4. Under Condition A, $L_1(\cdot)$ is bounded from below and $L_1(\lambda)\to\infty$ as $\|\lambda\|\to\infty$.

In view of Lemma 2.4, under Condition A there exists $\lambda^*\geq 0$ such that

$$\inf_{\lambda\geq 0} L_1(\lambda) = L(\lambda^*).$$

For this λ^*, Theorem 2.2 shows that

$$L((\nu,\pi),\lambda^*)\leq L^*, \text{ for all } (\nu,\pi)\in P(S)\times\Pi. \tag{2.12}$$

Now, let $L_2(\mu) = \inf_{\lambda\geq 0} L(\mu,\lambda)$ for each $\mu\in O(K)$.

Since $L(\mu,\lambda)$ is continuous in $(\mu,\lambda)\in O(K)\times R_+^k$, $L_2(\cdot)$ is upper semicontinuous on $O(K)$.

Thus, from the compactness of $O(K)$ ([17]), there exists $\mu^*\in O(K)$ such that

$$\sup_{\mu\in O(K)} L_2(\mu) = L_2(\mu^*).$$

Decomposing μ^* into $\mu^* = \nu^*\times\Phi^*$ with $\nu^* = \mu_S^*$ and $\Phi^*\in P(A|S)$, it follows from Corollary 2.2 that

$$L^*\leq L((\nu^*,\Phi^*),\lambda), \quad \text{for all } \lambda\geq 0. \tag{2.13}$$

Let $U_{RS} = \{(\nu,\pi)\in U|\pi\in\Pi_{RS}\}$.

From (2.12) and (2.13), we get the following main theorem.

Theorem 2.3. Under Condition A, the Lagrangian $L(\cdot,\cdot)$ has a saddle-point with a randomized stationary policy; *i.e.*, there exists $\lambda^* \geq 0$ and $(\nu^*, \Phi^*) \in P(S) \times \Pi_{RS}$ such that

$$L((\nu,\pi),\lambda^*) \leq L((\nu^*,\Phi^*),\lambda^*) = L^* \leq L((\nu^*,\pi),\lambda), \tag{2.14}$$

for all $(\nu,\pi) \in P(S) \times \Pi$ and $\lambda \geq 0$.

Then, from Theorem 2.1 and 2.3 the following corollary follows.

Corollary 2.3. Under Assumption A, there exists a constrained optimal pair in U_{RS}.

3 Characterization of Φ^*

In this section we use the hypothesis of Doeblin [10] and give the functional characterization of Φ^*.

Denoting by MDPs(λ) unconstrained MDPs with $r(\cdot,\cdot|\lambda)$ as an immediate reward function, we define the average expected reward in MDPs(λ) as

$$\phi_\lambda(\nu,\pi) := \liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} E_\nu^\pi[r(X_t,\Delta_t|\lambda)],\ (\nu,\pi) \in P(S) \times \Pi. \tag{3.1}$$

We say that $(\bar{\nu},\bar{\pi}) \in P(S) \times \Pi$ is an optimal pair for MDPs(λ) if $\phi_\lambda(\bar{\nu},\bar{\pi}) \geq \phi_\lambda(\nu,\pi)$ for all $(\nu,\pi) \in P(S) \times \Pi$.

The following lemma can be proved easily (*cf.* [3]).

Lemma 3.1. Let $(\bar{\nu},\bar{\pi}) \in P(S) \times \Pi$ and $\bar{\lambda} = (\bar{\lambda_1}, \bar{\lambda_2}, \cdots, \bar{\lambda_k}) \in R_+^k$. Then, the Lagrangian L has a saddle-point at $(\bar{\nu},\bar{\pi}), \bar{\lambda}$ iff the following holds:

(i). $(\bar{\nu},\bar{\pi})$ is a optimal pair for MDPs $(\bar{\lambda})$,
(ii). $I_l(\bar{\nu},\bar{\pi}) \leq \alpha_l$ for all l $(1 \leq l \leq k)$,
(iii). $\sum_{l=1}^k \bar{\lambda_l}(\alpha_l - I_l(\bar{\nu},\bar{\pi})) = 0$.

For any $\Phi \in \Pi_{RS}$, the t-step transition probabilities are defined by

$$Q^{(1)}(\cdot|x,\Phi) = \int Q(\cdot|x,a)\Phi(da|x),$$
$$Q^{(t+1)}(\cdot|x,\Phi) = \int Q^{(t)}(\cdot|x_1,\Phi)Q^{(1)}(dx_1|x,\Phi)\ \ (t \geq 1).$$

Henceforth we assume that the following hypothesis holds.

Hypothesis (Doeblin [10]). There is a finite measure γ of sets $D \in \bar{B}_S$ with $\gamma(S) > 0$, an integer m and a positive ϵ, such that

$$Q^{(m)}(D|x,\Phi) \leq 1-\epsilon \text{ if } \gamma(D) \leq \epsilon, \text{ for all } \Phi \in \Pi_{RS} \text{ and } x \in S.$$

Here we need the following condition.

Condition B. The following B1-B2 holds:

B1. The finite measure γ in the hypothesis of Doeblin is non-atomic, that is, $\gamma(U_\epsilon(x)) > 0$ for any $\epsilon > 0$ and $x \in S$,
where $U_\epsilon(x) = \{y \in S | d(x,y) < \epsilon\}$ and d is a metric on S.

B2. For any $f \in \Pi_S, \nu_f$ and γ are absolutely continuous each other, where ν_f is a stationary absolute probability measure for the Markov process induced by $\{Q(\cdot|x, f(x))\}$.

For (ν^*, Φ^*) and $\lambda^* \geq 0$ in Theorem 2.3, considering unconstrained MDPs(λ^*), from Lemma 3.1 we can obtain the following functional characterization of Φ^*, whose proof is obtained by observing that of ([15], Theorem 2.2) and left to the reader.

Theorem 3.1. Suppose that Assumption A and B hold. Then, for (ν^*, Φ^*) and λ^* in Theorem 2.3 there exist $v \in B(S)$, $v_l \in B(S)$ $(1 \leq l \leq k)$ such that for all $x \in S$, the following (i)-(iii) holds:

(i). $v(x) + L^* = \int \Phi^*(da|x)\{r(x,a|\lambda^*) + \int v(x')Q(dx'|x,a)\}$, (3.2)

(ii). $u_l\{\Phi^*\}(x) \geq 0$, (3.3)

(iii). $\sum_{l=1}^k \lambda_l^* u_l\{\Phi^*\}(x) = 0$, (3.4)

where, for $\Phi \in \Pi_{RS}$, $1 \leq l \leq k$,

$$u_l\{\Phi\}(x) = \int \Phi(da|x)\{\alpha_l - c_l(x,a) + \int v_l(x')Q(dx'|x,a)\} - v_l(x).$$

Corollary 3.1. Under Assumption A and B, Φ^* in Theorem 2.3 is a constrained optimal policy.

Proof. From Theorem 3.1 (i), for any $x \in S$, (x, Φ^*) is a optimal pair for MDPs(λ^*), where the initial distribution degenerate at a point x is denoted by x.

Also, (ii) and (iii) in Theorem 3.1 implies (ii) and (iii) in Lemma 3.1 with (x, Φ^*) so that (x, Φ^*) give a saddle-point of the Lagrangian L, which is as required. □

For further working, we need the following condition.

Condition C. The following C1-C2 holds:

C1. For any $g \in B(S)$, $\int g(x')Q(dx'|x,a)$ is continuous in $(x,a) \in K$.

C2. The set-valued function $A(\cdot)$ is lower semicontinuous, *i.e.*, for any sequence $\{x_n\}$ and $x \in S$ with $x_n \to x$ as $n \to \infty$ and any $a \in A(x)$, there exists a sequence $\{a_n\}$ with $a_n \in A(x_n)$ and $a_n \to a$ as $n \to \infty$.

For $v \in B(S)$ in Theorem 3.1, let

$$U(x,a) := r(x,a|\lambda^*) + \int v(x')Q(dx'|x,a), \text{ for } (x,a) \in K.$$

Under Condition C, $U(x,a)$ is continuous in $(x,a) \in K$ and $U^*(x) = \max_{a \in A(x)} U(x,a)$ is so too.

Thus, if we put $K_0 := \{(x,a) \in K | U^*(x) = U(x,a)\}$, K_0 is closed.

Also, (3.2) implies $v(x) + L^* \leq U^*(x)$ for all $x \in S$.

Lemma 3.2. Under Assumption A, B,and C, we have:

(i). $v(x) + L^* = U^*(x), \quad \gamma - a.s.,$

(ii). $\Phi^*(K_0(x)|x) = 1, \quad \gamma - a.s.,$

(iii). For any $\Phi \in \Pi_{RS}$ with $\Phi(K_0(x)|x) = 1$ for all $x \in S$ is optimal in MDPs(λ^*), that is, $\phi_{\lambda^*}(x,\Phi) \geq \phi_{\lambda^*}(x,\pi)$ for all $\pi \in \Pi$ and $x \in S$, where $K_0(x) = \{a \in A(x) | (x,a) \in K_0\}$.

Proof. Suppose that (i) does not hold, then, there exists a $D \in B_S$ such that $\gamma(D) > 0$ and $U(x) + L^* < U^*(x)$, for all $x \in D$.

By the measurable selection theorem (for example, [4,8]) there exists a $f \in B(S \to A)$ with $U^*(x) = U(x, f(x))$ for all $x \in S$.

For this stationary policy f, we have

$$\nu(x) + L^* \leq U(x,f), \text{ for all } x \in S$$

and

$$\nu(x) + L^* < U(x,f), \text{ for all } x \in D.$$

By the usual discussion (for example, [9,15]), it follows from B that

$$L^* < \phi_{\lambda^*}(\nu_f, f),$$

which is a contradiction.

Also, from (3.2) and (i), (ii) and (iii) obviously follow. □

Theorem 3.2. Suppose that Assumption A, B and C hold. For v, $v_l (1 \leq l \leq k) \in B(S)$ in Theorem 3.1, let $\Phi \in \Pi_{RS}$ satisfy the following (i)-(iii):

(i). $\Phi(K_0(x)|x) = 1$ for all $x \in S$,

(ii). $u_l\{\Phi\}(x) \geq 0$ for all $x \in S$,

(iii). $\sum_{l=1}^k \lambda_l^* u_l\{\Phi\}(x) = 0$, for all $x \in S$, where $\lambda^* = (\lambda_1^*, \cdots, \lambda_k^*)$ is in Theorem 2.3.

Then, Φ is a constrained optimal policy.

Proof. In view of Lemma 3.2 (iii), we observe that Φ is optimal for MDPs(λ^*).

Also, (ii) and (iii) implies that $I_l(x,\Phi) \leq \alpha_l$ for all $x \in S$ and $l (1 \leq l \leq k)$ and that $\sum_{l=1}^k \lambda_l^*(\alpha_l - I_l(x,\Phi)) = 0$ for all $x \in S$. This fact implies from Lemma 3.1 that the Lagrangian L has a saddle-point at (x,Φ), λ^*, which is as required. □

References

[1] E.J.Anderson and P.Nash, *Linear programming in infinite-dimensional spaces*, Wiley, Chichester, England, 1987.

[2] A.Arapostathis, V.S.Borkar, E.Fernandez-Gaucherand, M. K.Ghosh and S.I.Marcus, *Discrete-time Controlled Markov Processes with Average Cost Criterion:A Survey*, SIAM J.Control Optim., Vol.31(1993), pp.282-344.

[3] M.Avriel, *Nonlinear programming, Analysis and Methods*, Prentice-Hall,Inc., 1976.

[4] D.P.Bertsekas and S.E.Shreve, *Stochastic optimal control-the discrete time case*, Academic press, New York, 1978.

[5] F.J.Beutler and K.W.Ross, *Optimal policies for controlled Markov chains with a constraint*, J.Math.Anal.Appl., Vol.112 (1985), pp.236-252.

[6] V.S.Borkar, *Topics in controlled Markov chains*, Pitman Research Notes in Math. No.240, Longman Scientific and Technical,Harlow 1991.

[7] V.S.Borkar, *Ergodic control of Markov chains with constraints-the general case*, SIAM J.Control Optim., Vol.32(1994), pp.176-186.

[8] L.D.Brown and R.Purve, *Measurable selection of extrema*, Ann.Statist., Vol.1(1973), pp.902-912.

[9] C.Derman, *Finite state Markovian decision processes*, Academic Press, New York, 1970.

[10] J.L.Doob, *Stochastic processes*, John Wiley, New York, 1953.

[11] O.Hernández-Lerma and J.B.Lasserre, *Linear programming and average optimality of Markov control processes on Borel spaces-Unbounded costs*, SIAM J.Control Optim., 32(1994), pp.480-500.

[12] O.Hernández-Lerma and D.Hernández-Hernández, *Discounted cost Markov decision processes on Borel spaces:The linear programming formulation*, J.Math.Anal.Appl., Vol.183(1994), pp.335-351.

[13] A.Hordijk and J.B.Lasserre, *Linear programming formulation of MDPs in countable state space:the multichain case*, ZOR-Math.Meth.O.R., Vol.40(1994), pp.91-108.

[14] L.C.M.Kallenberg, *Survey of linear programming for standard and nonstandard Markovian control problems. Part I:Theory*, ZOR-Math.Meth.O.R., 40(1994), pp.1-42.

[15] M.Kurano, *The existence of a minimum pair of state and policy for Markov decision processes under the hypothesis of Doeblin*, SIAM J.Control Optim., 27(1989), pp.296-307.

[16] D.Luenberger, *Optimization by vector space methods*, John Wiley, New York, 1969.

[17] K.R.Parthasarathy, *Probability measure on metric space*, Academic Press, New York, 1967.