Title	Constrained Markov Decision Processes : The Average Case(Mathematical Structure of Optimization Theory)
Author(s)	Kurano, Masami; Huang, Youqiang
Citation	数理解析研究所講究録 (1994), 864: 1-8
Issue Date	1994-04
URL	http://hdl.handle.net/2433/83914
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

数理解析研究所講究録
第 864 巻 1994 年 1-8

Constrained Markov Decision Processes: The Average Case

蔵野　正美 (Masami Kurano)　黄　佑強 (Youqiang Huang)
千葉大学教育学部 (Chiba University)

1. Preface

A Markov decision process (MDP) with multiple constraints is applicable in many fields. Recently, many researchers pay attention to it ([2],[10] and their references). The methods of analysis about solving it are by using Lagrange multiplier theory ([4],[5]) or by changing it to linear programming (LP) to apply LP theory ([11],[12],[13]).

In this report we consider the average reward criterion of MDP with sample path constraints by the above two ideas under state and action sets being compact. we prove the existence theorem of a constrained optimal pair. Also, by introducing the concept of state-wise mixed policy, we give a characterization of it.

2 . Formulation

In this report mentioned Borel sets are Borel subsets of a complete separable metric space. For a Borel set X, B_{X} denotes the Borel subsets of X. C(X) denotes the set of all bounded continuous functions on X. A Markov decision process with multiple constraints is a controlled dynamic system defined by following objects: S, $\{\mathrm{A}(x), x \in \mathrm{S}\}$, Q, r, c_i $(i = 1, \cdots, k)$, where S is any Borel set representing the state space of some system and for each $x \in \mathrm{S}$, the admissible action space $\mathrm{A}(x)$ is a non-empty subset of some Borel set A such that $\{(x, a) : x \in \mathrm{S}, a \in \mathrm{A}(x)\}$ is an element of $B_{\mathrm{S}} \times B_{\mathrm{A}}$, the immediate reward function r is a real-valued Borel measurable function on S $\times$ A, the immediate cost functions c_i (i=1,$\cdots$,k) are real-valued cost functions on S $\times$ A , $\mathrm{Q}(\cdot|x, a)$ is the law of motion , which is taken to be stochastic kernel on $B_{\mathrm{S}} \times \mathrm{S} \times \mathrm{A}$; i.e, for each $(x, a) \in \mathrm{S} \times \mathrm{A}$, $\mathrm{Q}(\cdot|x, a)$ is a probability measure on B_{S}; and for each $\mathrm{D} \in B_{\mathrm{S}}$, $\mathrm{Q}(\mathrm{D} \,|\cdot\,)$ is a Borel measurable function on S $\times$ A .

Throughout this report, the following assumptions will be remain operative:

(i). S and $\{\mathrm{A}(x), x \in \mathrm{S}\}$ are compacts;

(ii). r is non-negative bounded continuous;

(iii). c_i is non-negative bounded continuous $(i = 1, \cdots, k)$;

(iv). whenever $x_n \to x, a_n \to a$, $\mathrm{Q}(\cdot|x_n, a_n)$ converges weekly to $\mathrm{Q}(\cdot|x, a)$.

The sample space is the product space $\Omega = (\mathrm{S} \times \mathrm{A})^\infty$ 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 ≥ 0).

A policy is a sequence π=$(\pi_0, \pi_1, \cdots)$ such that, for each t ≥ 0, π_t is a stochastic kernel on $B_\mathrm{A} \times \mathrm{S} \times (\mathrm{A} \times \mathrm{S})^\mathrm{t}$ with $\pi_t(\mathrm{A}(x_t)|x_0, a_0, \cdots, a_{t-1}, x_t)$=1 for all $(x_0, a_0, \cdots, a_{t-1}, x_t) \in \mathrm{S} \times (\mathrm{A} \times \mathrm{S})^\mathrm{t}$.

Let Π denote the class of policies.

T(A | S) is the set of all stochastic kernels Φ on $B_\mathrm{A} \times$ S with Φ $(\mathrm{A}(x)|x)$=1 for all $x \in$ S.

A policy π=$(\pi_0, \pi_1, \cdots)$ is a randomize stationary policy if there is a $\Phi \in$ T(A | S) such that $\pi_t(\cdot|x_0, a_0, \cdots, x_t) = \Phi(\cdot|x_t)$ for all $(x_0, a_0, \cdots, x_t) \in \mathrm{S} \times (\mathrm{A} \times \mathrm{S})^\mathrm{t}$ and t ≥ 0. Let Φ^∞ denote the corresponding policy.

For any D $\in B_\mathrm{S}$, B(D $\rightarrow$ A) denotes the set of all Borel measurable functions u: D $\rightarrow$ A with $u(x) \in \mathrm{A}(x)$ for all $x \in$ D.

A randomize stationary policy Φ^∞ is called stationary if there is an f $\in$ B(S $\rightarrow$ A) such that $\Phi(f(x)|x)$=1 for all $x \in$ S . Such a policy will be written by f^∞ .

Π_RS and Π_S are respectively the sets of all randomize stationary and stationary policies.

Let $\mathrm{H_t} = (\mathrm{X}_0, \Delta_0, \cdots, \Delta, \mathrm{X_t})$. It is assumed that, for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$, $\mathrm{Prof}(\Delta_t \in \mathrm{D}_1|\mathrm{H_t}) = \pi_t(\mathrm{D}_1|\mathrm{H_t})$ and $\mathrm{Prof}(\mathrm{X_{t+1}} \in \mathrm{D}_2|\ \mathrm{H_{t-1}},\ \Delta_{t-1},\ \mathrm{X_t} = x,\ \Delta_t = a) = \mathrm{Q}(\mathrm{D}_2|x, a)$, for every $\mathrm{D}_1 \in B_\mathrm{A}$ and $\mathrm{D}_2 \in B_\mathrm{S}$, and t=0,1,2,$\cdots$.

For any Borel set X , P(X) is the set of all probability measures on X. Then, for each $\pi \in \Pi$ and initial state distribution $\nu \in$ P(S) , P_π^ν is probability measure on Ω, which can be defined in an abvious way, and E_π^ν is the expectation with respect to P_π^ν.

We define measurable funstions on Ω as follows:

$$\text{(2.3)} \qquad \begin{aligned} \tilde{\mathrm{R}}_\mathrm{T} &:= \frac{1}{\mathrm{T}} \sum_{t=0}^{\mathrm{T}-1} r(\mathrm{X_t}, \Delta_\mathrm{t}) \quad (\mathrm{T} \geq 1), \\ \tilde{\mathrm{R}} &:= \liminf_{\mathrm{T} \to \infty} \tilde{\mathrm{R}}_\mathrm{T}. \end{aligned}$$

$$\text{(2.4)} \qquad \begin{aligned} \tilde{\mathrm{C}}_\mathrm{T}^i &:= \frac{1}{\mathrm{T}} \sum_{t=0}^{\mathrm{T}-1} c_i(\mathrm{X_t}, \Delta_\mathrm{t}) \quad (\mathrm{T} \geq 1), \\ \tilde{\mathrm{C}}^i &:= \limsup_{\mathrm{T} \to \infty} \tilde{\mathrm{C}}_\mathrm{T}^i \quad (i = 1, \cdots, k). \end{aligned}$$

For any $(\nu, \pi) \in P(S) \times \Pi$,

$$\text{(2.5)} \qquad \mathrm{R}(\nu, \pi) := \mathrm{ess} \cdot \inf \tilde{\mathrm{R}} (= \sup\{a|\mathrm{P}_\nu^\pi(\tilde{\mathrm{R}} \geq a) = 1\})$$

$$\text{(2.6)} \qquad \mathrm{C}_i(\nu, \pi) := \mathrm{ess} \cdot \sup \tilde{\mathrm{C}}^i (= \inf\{a|\mathrm{P}_\nu^\pi(\tilde{\mathrm{C}}^i \leq a) = 1\}) \quad (i = 1, \cdots, k,)$$

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)$ is given.

Let

$$\mathrm{U}_\alpha := \{(\nu,\pi) \in \mathrm{P(S)} \times \Pi | \mathrm{C}^i(\nu,\pi) \le \alpha_i,\ i=1,\cdots,k\}$$
$$\mathrm{U}_\alpha^{\mathrm{RS}} := \{(\nu,\Phi^\infty) \in \mathrm{U}_\alpha | (\nu,\Phi^\infty) \in \mathrm{P(S)} \times \Pi_{\mathrm{RS}}\}$$

In this report we mainly consider the following problem.

$$\text{Problem(A)}: \quad \text{Maximum} \quad \mathrm{R}(\nu,\pi)$$
$$\text{subject to} \quad (\nu,\pi) \in \mathrm{U}_\alpha$$

$(\nu^*,\pi^*) \in \mathrm{U}_\alpha$ will be called a constrained optimal pair if $\mathrm{R}(\nu^*,\pi^*) \ge \mathrm{R}(\nu,\pi)$ for all $(\nu,\pi) \in \mathrm{U}_\alpha$.

For any $\epsilon > 0$,

$(\nu^*,\pi^*) \in \mathrm{U}_\alpha$ is called a constrained ϵ-optimal pair if $\mathrm{R}(\nu^*,\pi^*) \ge \mathrm{R}(\nu,\pi) - \epsilon$ for all $(\nu,\pi) \in \mathrm{U}_\alpha$.

In Section 3 we shall prove that a constrained optimal pair exists in $\mathrm{U}_\alpha^{\mathrm{RS}}$ and in Section 4 give characterization of a constrained optimal pair.

3. Existence of optimal pair and related linear programmings

In this section, we transform Problem (A) given in the preceding section to LP equivalently and prove the existence of optimal pair by using compactness.

Let $\{x_i\}$ be dense in S and define $g_{ij} \in \mathrm{C(S)}$ for i , j=1,2, $\cdots$, by

$$g_{ij}(x) = 2(1 - jd(x,x_i)) \vee 0, \tag{3.1}$$

where d is the metric defined in S and $x \vee y = \max\{x,y\}$.

Let

$$\mathrm{G} = \{g_{ij} : i,j = 1,2,\cdots\}.$$

Then G is separating, i.e, whenever $\mathrm{P}_1,\ \mathrm{P}_2 \in \mathrm{P(S)}$ and

$$\int g d\mathrm{P}_1 = \int g d\mathrm{P}_2$$

for all $g \in \mathrm{G}$, we have $\mathrm{P}_1 = \mathrm{P}_2$ ([7]).

For $\mu \in \mathrm{P(S \times A)}, \mathrm{h} \in \mathrm{C(S \times A)}$ we denote the integral as follows:

$$(h,\mu) := \int h(x,a)\mu(d(x,a)). \tag{3.2}$$

We can verify the following lemma by referring to the proof of Lemma 2.1 in ([8]).

Lemma 3.1 . For any positive ε , and any $(\nu,\pi) \in U_\alpha$, there is $\mu \in \mathrm{P(S \times A)}$ such that:

(i). $(r,\mu) \geq R(\nu,\pi) - \epsilon$
(ii). $(c_i,\mu) \leq \alpha_i + \epsilon \quad (i = 1,2,\cdots,k)$
(iii). $\int g(x)\mu(d(x,a)) = \int \mu(d(x,a)) \int g(x')Q(dx'|x,a)$ for all $g \in$ G.

Brief proof.

By the definition of $\mathrm{R}(\nu,\pi), \mathrm{C}^i(\nu,\pi)$ and stability theorem [9] we can get the following:

$$\mathrm{P}^{\nu}_{\pi}(\tilde{\mathrm{R}} \geq \mathrm{R}(\nu,\pi) - \epsilon) = 1 \tag{3.3}$$

$$\mathrm{P}^{\nu}_{\pi}(\tilde{\mathrm{C}}_i \leq \alpha_i + \epsilon) = 1 \quad (i = 1,\cdots,k) \tag{3.4}$$

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T-1}\{g(\mathrm{X}_t) - E[g(\mathrm{X}_t)|B_{t-1}]\}}{T} = 0 \quad \text{for all } g \in \mathrm{G}, \mathrm{P}^{\nu}_{\pi} - \text{almost surely} \tag{3.5}$$

From (3.3) to (3.5) there exists a sample path $\omega \in \Omega$ such that $\tilde{\mathrm{R}}(\omega) \geq \mathrm{R}(\nu,\pi) - \epsilon$, $\tilde{\mathrm{C}}^i(\omega) \leq \alpha_i + \epsilon$ $(i = 1,\cdots,k)$ and (3.5) hold.

For this $\omega \in \Omega$ we discuss the following empirical probability measure $\mu_{\mathrm{T}} \in \mathrm{P}(\mathrm{S} \times \mathrm{A})$.

$$\mu_{\mathrm{T}}(\mathrm{D}) = \frac{\sum_{\mathrm{t}=0}^{\mathrm{T}-1} \mathrm{I_D}(\mathrm{X_t},\Delta_{\mathrm{t}})}{\mathrm{T}} \quad (\mathrm{T} \geq 0), \quad \text{for all } \mathrm{D} \in B_{\mathrm{S}\times\mathrm{A}},$$

where $\mathrm{I_D}$ is the indicator function of D.

From weak compactness of $\mathrm{P}(\mathrm{S} \times \mathrm{A})$ there exists a sequence $\{\mu_{\mathrm{T_j}}\}$ such that (3.7) to (3.10) hold.

$$\lim_{j\to\infty}[\int g(x)\mu_{\mathrm{T}_j}(d(x,a)) - \int \mu_{\mathrm{T}_j-1}(d(x,a)) \int g(x')\mathrm{Q}(dx'|x,a)] = 0 \quad \text{for all } g \in \mathrm{G} \tag{3.7}$$

$$\lim_{j\to\infty}(r,\mu_{\mathrm{T}_j}) \geq R(\nu,\pi) - \epsilon \tag{3.8}$$

$$\lim_{j\to\infty}(c_i,\mu_{\mathrm{T}_j}) \leq \alpha_i + \epsilon \quad (i = 1,2,\cdots,k) \tag{3.9}$$

$$\mu_{\mathrm{T_j}} \to \mu \in \mathrm{P}(\mathrm{S} \times \mathrm{A}), \text{ in the weak topology} \tag{3.10}$$

It is obvious that μ in (3.10) satisfies (i),(ii),(iii) of lemma 3.1.

We now consider infinite linear programming as follows:

LP* Maximum (r,μ)
subject to (i) $(c_i,\mu) \leq \alpha_i \quad (i = 1,\cdots,k)$
(ii) $\int g(x)\mu(d(x,a)) = \int \mu(d(x,a)) \int g(x')\mathrm{Q}(dx'|x,a)$
(iii) $\mu \in \mathrm{P}(\mathrm{S} \times \mathrm{A})$

$P_F(S \times A)$ is the set of all μ which satisfy the LP* conditions (i),(ii),(iii).

In order to prove that Problem (A) is equivalent to LP*, we need the following assumption 1, which remains operative thereafter.

Assumption 1.

For any $\Phi^\infty \in \Pi_{RS}$, the Markov chain induced by $Q(\cdot|x,\Phi)$ satisfies the Doeblin condition and is one-ergodic,where $Q(\cdot|x,\Phi) = \int Q(\cdot|x,a)\Phi(da|x)$.

Under assumption 1 and lemma 3.1 we can verify the following theorem 3.1 by using weak compactness of $P(S \times A)$ and ergodic theorem ([6]).

Theorem 3.1. For any $(\nu,\pi) \in U_\alpha$, there exists a $\mu \in P_F(S \times A)$ such that for the decomposition $\mu = \nu_s \times \Phi, \nu_s \in P(S), \Phi \in P(A|S)$, it holds:

(i). $(\nu_s, \Phi^\infty) \in U_\alpha$;

(ii). $R(\nu_s, \Phi^\infty) \geq R(\nu,\pi)$;

(iii). $R(\nu_s, \Phi^\infty) = (r, \mu)$.

From the above result, the following follows.

Corollary 3.1. Problem (A) is equivalent to LP*

Since it is shown by compactness that LP* has optimal solution (see [1]), the following holds from corollary 3.1.

Corollary 3.2. Optimal pair exists in U_α^{RS}.

4. State-wise mixed stationary policies

From corollary 3.1 we can get optimal pair or ϵ-optimal pair of problem (A) by solving LP*. In this section we give characterization of solutions of LP*.

Let $\mu_A \in P(A)$. If there exists an integer $l(l \geq 1), a_i \in A(i = 1,\cdots,l)$ and $p_i(i = 1,\cdots,l)$ with $p_i \geq 0$, $\sum_{i=1}^{l} p_i = 1$ such that $\mu_A(\{a_i\}) = p_i(i = 1,\cdots,l)$, we dnote μ_A by

$$\mu_A = \begin{pmatrix} a_1, a_2, \cdots, a_l \\ p_1, p_2, \cdots, p_l \end{pmatrix}. \tag{4.1}$$

Let $\Phi \in P(A|S)$, If there exists an integer $l(l \geq 1), f_i \in B(S \to A)$ and $p_i \in B(S \to [0,1])$ $(i = 1,\cdots,l)$ with $p_i(x) \geq 0$, $\sum_{i=1}^{l} p_i(x) = 1$ for all $x \in S$. such that

$$\Phi(\cdot|x) = \begin{pmatrix} f_1(x), f_2(x), \cdots, f_l(x) \\ p_1(x), p_2(x), \cdots, p_l(x) \end{pmatrix}, \tag{4.2}$$

Φ is called an l-state-wise mixed kernel (l-s.m.k) and corresponding policy Φ^∞ is called l-state-wise mixed stationary policy. Moreover, if Φ is an l-s.m.k for some $l \geq 1$, we say Φ is a s.m.k.

For $l \geq 1$, let

$$F^l := \{\mu \in \mathrm{P}(\mathrm{S} \times \mathrm{A}) | \mu = \nu_s \times \Phi,\ \nu_s \in \mathrm{P}(\mathrm{S}),\ \Phi : l - \text{s.m.k}\}$$

$$F := \bigcup_{l=1}^{\infty} F^l$$

We note F^1 represents the set of all non-randomized stationary policies.

Now, we need the following assumption 2, which remains operative thereafter.

Assumption 2.

The set of inner points of U_α is non-empty.

Theorem 4.1. For any $\epsilon > 0$, an ϵ-optimal solution of LP* exists in F.

The proof of theorem 4.1 is given in a sequence of lemmas, some of interest.

Lemma 4.1. F is convex and if S is a finite set F^l is compact for every l ($l \geq 1$).

The proof theorem 4.1 is done by induction on l which is the number of constraints in LP*. For that, the following definitions are given.

$$\mathrm{P}_\mathrm{F}^{(m)}(\mathrm{S} \times \mathrm{A}) := \{\mu \in \mathrm{P}(\mathrm{S} \times \mathrm{A}) | (c_i, \mu) \leq \alpha_i (i = 1, \cdots, m) \text{ and } \mu \text{ satisfies (ii) of LP*}\}$$

$$\mathrm{P}_\mathrm{F}^{(0)}(\mathrm{S} \times \mathrm{A}) := \{\mu \in \mathrm{P}(\mathrm{S} \times \mathrm{A}) | \mu \text{ satisfies (ii) of LP*}\}$$

We note $\mathrm{P}_\mathrm{F}^{(m)}(\mathrm{S} \times \mathrm{A})$ is compact and convex.

For each $m (0 \leq m \leq k)$, consider the following LP problem:

$$\mathrm{LP}^{(m)} \quad \text{Maximum} \quad (r, \mu)$$
$$\text{subject to} \quad \mu \in \mathrm{P}_\mathrm{F}^{(m)}(\mathrm{S} \times \mathrm{A})$$

In case that $m = 0$, the following holds (see [8]).

Lemma 4.2. $\mathrm{LP}^{(0)}$ has optimal solution in F^1.

Lemma 4.2 shows that theorem 3.1 is true for $\mathrm{LP}^{(0)}$. Now for $\mathrm{LP}^{(m)}$ $(1 \leq m \leq k)$ it is supposed that theorem 4.1 is true. Here we use Lagrangian multiplier techniques [4].

Let

$$(4.3) \qquad r_\lambda := r - \lambda c_{m+1} \qquad (\lambda \geq 0)$$

We consider the following LP.

$$\begin{array}{lll} \mathrm{LP}^{**} & \text{Maximum} & (r_\lambda, \mu) \\ & \text{subject to} & \mu \in \mathrm{P}_{\mathrm{F}}^{(m)}(\mathrm{S} \times \mathrm{A}) \end{array}$$

For any $\epsilon > 0$, it follows from the assumption of induction that an ϵ-optimal solution of LP** exist in F. Consequently, an optimal solution μ^λ of LP** exists in $\bar{\mathrm{F}}$ (the closure of F).

Let

$$\begin{aligned} \mathrm{J}^\lambda &:= (r_\lambda, \mu^\lambda) = (r, \mu^\lambda) - \lambda(c_{(m+1)}, \mu^\lambda); \\ \mathrm{R}^\lambda &:= (\nu, \mu^\lambda); \\ \mathrm{K}^\lambda &:= (c_{m+1}, \mu^\lambda). \end{aligned}$$

The following two lemmas can be verified by the a way as these in [4].

Lemma 4.3. $\mathrm{J}^\lambda, \mathrm{R}^\lambda, \mathrm{K}^\lambda$ are non-decrease functions of λ.

Lemma 4.4. $\gamma := \inf\{\lambda : \mathrm{K}^\lambda \leq \alpha_{m+1}\} < \infty$.

Under the above preliminaries we know that if $\mathrm{K}^\gamma = \alpha_{m+1}$, μ^γ is the optimal solution of $\mathrm{LP}^{(m+1)}$ by usually Lagarange multiplier way and if $\mathrm{K}^\gamma \neq \alpha_{m+1}$ Theorem 4.1 is shown to be true for $\mathrm{LP}^{(m+1)}$ by using lemma 4.1.

When S is a finite set, the following results are obtained by compact property of F^l.

Corollary 4.1. If S is finite set an optimal solution of LP* exist in F^{2k}.

From corollary 3.1 the following fact can be easily obtained.

Corollary 4.2. For any $\epsilon > 0$ there are $\nu \in \mathrm{P(S)}$ and s.m.k Φ such that (ν, Φ^∞) is an ϵ-optimal pair. If S is finite set, there are $\nu \in \mathrm{P(S)}$ and 2k-s.m.k Φ such that (ν, Φ^∞) is an optimal pair.

Bibliography

[1] Anderson E.J. and Nash P. *Linear Programming in Infinite Dimensional Space*,Wiley,Chichester,1987.

[2] Arapostathis A. Borkar V.S. Fernandez-Gaucherand E. Ghosh M.K.and Marcus S.I. *Discrete time controlled Markov process with average cost criterion: a survey*,SIAM J.Control optim.vol.31,no.2,282-344,1993.

[3] Bertsekas D.P. and Shreve S.E. *Stochastic optimal control the discrete times case*,Academic Press,New York,1978.

[4] Beutler F.J.and Ross K.W. *Optimal policies for controlled Markov chair with a constraint*,J.Math. Anal.Appal.112(1985),236-256.

[5] Borkar V.S. *Topics in controlled Markov chain*,Pitman Research Nates in Math.No.240, Longman Scientific and Technical,Harlow,1991.

[6] Doob J.L. *Stochastic Processes*,John Wiley,New York,1953.

[7] Etier S.N.and Kurtz T.G. *Markov process characterization and convergence*,John Wiley,1986.

[8] Kurano M. *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), 296-307.

[9] Loeve M. *Probability theory*,Second edition Van Nostrand,Princeton,NJ,1960.

[10] Ohnishi M. *Markov* 決定過程に関する最近の話題,Proceeding of the Fourth RAMP Symposium, 1992.

[11] Ross K.W. *Randomized and path dependent policies for Markov decision processes with multiple constrains*,Open.Res.37(1989)474-477.

[12] Ross K.W.and Varadarajan R. *Markov decision processes with sample path constrains:The communicating cases*,Open.Res.37(1989),780-790.

[13] Ross K.W.and Varadarajan R. *Multiple Markov decision processes with a sample path constrain: A decomposition approach*,Math.O.R.16(1991),195-207.