Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER Computers and Mathematics with Applications 51 (2006) 279–284
An International Journal
computers & mathematics with applications
www.elsevier.com/locate/camwa

Discounted Markov Decision Processes with Utility Constraints

YOSHINOBU KADOTA
Faculty of Education, Wakayama University
Wakayama 640-8510, Japan
yoshi-k@math.edu.wakayama-u.ac.jp

MASAMI KURANO
Faculty of Education, Chiba University
Chiba 263-8522, Japan
kurano@faculty.chiba-u.jp

MASAMI YASUDA
Department of Math & Informatics, Faculty of Science
1-33 Yayoi-cho, Inage-ku, Chiba University
Chiba 263-8522, Japan
yasuda@math.s.chiba-u.ac.jp

Abstract—We consider utility-constrained Markov decision processes. The expected utility of the total discounted reward is maximized subject to multiple expected utility constraints. By introducing a corresponding Lagrange function, a saddle-point theorem of the utility constrained optimization is derived. The existence of a constrained optimal policy is characterized by optimal action sets specified with a parametric utility. © 2006 Elsevier Ltd. All rights reserved.

Keywords—Markov decision processes, Utility constraints, Discount criterion, Lagrange technique, Saddle-point, Constrained optimal policy.

1. INTRODUCTION AND PROBLEM FORMULATION

Utility-constrained Markov decision processes (MDPs) arise in the case where the decision maker wants to maximize the total reward under more than one utility function. The typical case is, for example, that in the group decision problem with different utility functions each player wants to maximize the reward under his own specified utility function. In such a case, we want to maximize the one type of expected utility of the reward while keeping other types of expected utilities higher than some given bounds.

In this paper, we consider general utility-constrained MDPs in which the expected utility of the total discounted rewards is maximized subject to multiple expected utility constraints and the objective is to show that the Lagrange approach to general utility-constrained MDPs is successfully done. In fact, by introducing a corresponding Lagrange function, a saddle-point theorem is given, by which the existence of a constrained optimal policy is proved. And a

The authors show grateful thanks to the anonymous referee who gave useful comments and suggestions on the earlier draft.

doi:10.1016/j.camwa.2005.11.013

constrained optimal policy is characterized by optimal action sets specified with a parametric utility.

However, we do not specify the kind of utility function; it is expected to enlarge the practical application of MDPs. As far as we are aware, it appears that little work has been done on the Lagrange method to general utility-constrained MDPs. The method of analysis for general utility functions is closely related to [1,2], in which discounted MDPs have been studied with general utility function and whose results are applied to characterize a constrained optimal policy. Recently, Kurano *et al.* [3] derived a saddle-point theorem for constrained MDPs with average reward criteria. For the utility treatment for MDPs and constrained MDPs, refer to [1,2,4–7] and their references.

In the remainder of this section, we define the utility-constrained problem to be examined and a constrained optimal policy. First we consider standard Markov decision processes (MDPs), specified by

$$(S, \{A(i)\}_{i\in S}, q, r)\,,$$

where $S = \{1, 2, \dots\}$ denotes the set of the states of the processes, $A(i)$ is the set of actions available at each state $i \in S$, taken to be a Borel subset of some Polish space A. The matrix $q = (q_{ij}(a))$ is a transition probability satisfying that $\sum_{j\in S} q_{ij}(a) = 1$ for all $i \in S$ and $a \in A(i)$, and $r(i, a, j)$ is an immediate reward function defined on $\{(i, a, j) \mid i \in S,\ a \in A(i),\ j \in S\}$.

Throughout this paper, the following assumption will remain operative.

ASSUMPTION 1.

(i) *For each $i \in S$, $A(i)$ is a closed set of a compact metric space A.*
(ii) *For each $i, j \in S$, both $q_{ij}(\cdot)$ and $r(i, \cdot, j)$ are continuous on $A(i)$.*
(iii) *The function r is uniformly bounded, i.e., $|r(i, a, j)| \le M$ for all $i, j \in S$, $a \in A(i)$, and some $M > 0$.*

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projection X_t, Δ_t on the t^{th} factors S, A describe the state and the action of t-time of the process ($t \ge 0$). A policy $\pi = (\pi_0, \pi_1, \dots)$ is a sequence of conditional probabilities π_t such that $\pi_t(A(i_t) \mid i_0, a_0, \dots, i_t) = 1$ for all histories $(i_0, a_0, \dots, i_t) \in (S \times A)^t \times S$. The set of policies is denoted by Π. Let $H_t = (X_0, \Delta_0, \dots, \Delta_{t-1}, X_t)$ for $t \ge 0$.

ASSUMPTION 2. *We assume that*

(i) $\text{Prob}(X_{t+1} = j \mid H_{t-1}, \Delta_{t-1}, X_t = i,\ \Delta_t = a) = q_{ij}(a)$,
(ii) $\text{Prob}(\Delta_{t+1} \in D \mid H_t) = \pi_t(D \mid H_t)$

for all $t \ge 0$, $i, j \in S$, $a \in A(i)$, any Borel subset $D \in A$, and for any given $\pi = (\pi_0, \pi_1, \dots) \in \Pi$.

Let $\mathcal{P}(X)$ be denoted by the set of all probability measures on any Borel measurable set X. Then, any initial probability measure $\nu \in \mathcal{P}(S)$ and policy $\pi \in \Pi$ determine the probability measure $P_\pi^\nu \in \mathcal{P}(\Omega)$ in a usual way.

For the state-action process $\{X_t, \Delta_t;\ t = 0, 1, 2, \dots\}$, its discounted present value is defined by

$$\mathcal{B} := \sum_{t=0}^{\infty} \beta^t r(X_t, \Delta_t, X_{t+1}), \tag{1.1}$$

where β ($0 < \beta < 1$) is a discount factor. Then, for each $\nu \in P(S)$ and $\pi \in \Pi$, $\mathcal{B}$ is a random variable from the probability space (Ω, P_π^ν) into the interval $[-M/(1-\beta), M/(1-\beta)]$.

ASSUMPTION 3. *Let g, h_i ($1 \le i \le k$) be any real-valued functions on the set of real numbers* R *satisfying that*

(i) *g is upper semicontinuous;*
(ii) *each h_i ($1 \le i \le k$) is lower semicontinuous.*

For any given threshold vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in \mathbf{R}^k$ and any initial probability measure $\nu \in \mathcal{P}(S)$, let

$$\mathcal{V}(\nu, \alpha) := \{\pi \in \Pi \mid E_\pi^\nu(h_i(\mathcal{B})) \leq \alpha_i, \text{ for all } i(1 \leq i \leq k)\},$$

where E_π^ν is the expectation with respect to P_π^ν. Interpreting g, h_i $(1 \leq i \leq k)$ as given utility functions, we will consider the following utility-constrained optimization problem:

Problem A: *maximize* $E_\pi^\nu(g(\mathcal{B}))$ *subject to* $\pi \in \mathcal{V}(\nu, \alpha)$.

The optimal solution $\pi^* \in \mathcal{V}(\nu, \alpha)$ of Problem A, if it exists, is called a ν-constrained optimal policy, or simply a constrained optimal policy.

Note that Problem A includes, for example, the constrained moment problem (cf. [8]): for the i^{th} moment of $\mathcal{B}$ with a sign $(-1)^i$,

- maximize $E_\pi^\nu(\mathcal{B})$ subject to $(-1)^i E_\pi^\nu(\mathcal{B}^i) \leq \alpha_i$ $(2 \leq i \leq k+1)$,

and the constrained threshold probability problem (cf. [9,10]):

- maximize $P_\pi^\nu(\mathcal{B} \geq a)$ subject to $P_\pi^\nu(\mathcal{B} \leq b) \leq \alpha$ for some $b < a$.

We shall use the following result in the sequel.

LEMMA 1.1. *(See [11].) For any* $\nu \in \mathcal{P}(S)$, $\tilde{\varphi}(\nu) := \{P_\pi^\nu \in \mathcal{P}(\Omega) \mid \pi \in \Pi\}$ *is convex and compact in the weak topology.*

In Section 2, the saddle-point statement for Problem A is given, whose results are applied to obtain the existence of a constrained optimal policy. The characterization of a constrained optimal policy is given and the exponential case is discussed in Section 3.

2. SADDLE-POINT THEOREM FOR UTILITY-CONSTRAINED MDPS

In this section, we prove the saddle-point theorem for the Lagrangian associated with Problem A. For any initial probability measure $\nu \in \mathcal{P}(S)$, we define the Lagrangian, L^ν, that corresponds to Problem A as follows:

$$L^\nu(\pi, \lambda) := E_\pi^\nu(g(\mathcal{B})) + \sum_{i=1}^{k} \lambda_i \left(\alpha_i - E_\pi^\nu(h_i(\mathcal{B}))\right) \tag{2.1}$$

for any $\pi \in \Pi$ and $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k) \in \mathbf{R}_+^k := \mathbf{R}^k \cap \{\lambda_i \geq 0 \ (1 \leq i \leq k)\}$. Without any confusion, $\lambda \in \mathbf{R}_+^k$ will be written simply by $\lambda \geq 0$.

The following statement on saddle-points can be proved similarly to that of Luenberger [12, p. 221, Theorem 2] and so omitted.

THEOREM 2.1. *(Cf. [12].) Suppose that there exists* $\pi^* \in \Pi$ *and* $\lambda^* \geq 0$ *such that* $L^\nu(\cdot, \cdot)$ *with* $\nu \in \mathcal{P}(S)$ *possesses a saddle-point at* π^*, λ^*, *i.e.,*

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

for all $\pi \in \Pi$ *and* $\lambda \geq 0$. *Then,* π^* *solves Problem A and is a* ν*-constrained optimal policy.*

The above theorem motivates us to obtain sufficient conditions for the existence of a saddle-point of the Lagrangian L^ν. To this purpose, it is convenient to rewrite the expected utility using the distribution function of the present value.

Let, for each $\nu \in \mathcal{P}(S)$ and $\pi \in \Pi$,

$$F_\pi^\nu(x) := P_\pi^\nu(\mathcal{B} \leq x), \tag{2.3}$$

$$\Phi(\nu) := \{F_\pi^\nu(\cdot) \mid \pi \in \Pi\}. \tag{2.4}$$

Now, with some abuse of notation, we define

$$L^{\nu}(F,\lambda) := \int g_{\lambda}(x)\, dF(x) \tag{2.5}$$

for any $F \in \Phi(\nu)$ and $\lambda \geq 0$, where

$$g_{\lambda}(x) := g(x) + \sum_{i=1}^{k} \lambda_i(\alpha_i - h_i(x)). \tag{2.6}$$

Then, the Lagrangian L^{ν} defined in (2.1) is obviously rewritten by $L^{\nu}(\pi,\lambda) = L^{\nu}(F,\lambda)$ with $F = F^{\nu}_{\pi}$. Thus, we have the following corollary.

COROLLARY 2.1. *Let $\pi^* \in \Pi$ and $\lambda^* \geq 0$. Then, $L^{\nu}(\cdot,\cdot)$ with $\nu \in \mathcal{P}(S)$ possesses a saddle-point at π^*, λ^* if and only if the following relation holds with $F^* = F^{\nu}_{\pi^*}$.*

$$L^{\nu}(F,\lambda^*) \leq L^{\nu}(F^*,\lambda^*) \leq L^{\nu}(F^*,\lambda), \tag{2.7}$$

for all $F \in \Phi(\nu)$ and $\lambda \geq 0$. Then, π^* solves Problem A and is a ν-constrained optimal policy.

LEMMA 2.1. *For any $\nu \in \mathcal{P}(S)$, it holds that*

(i) *$\Phi(\nu)$ is convex and compact in the week topology;*
(ii) *$L^{\nu}(\cdot,\lambda)$ is concave and upper semicontinuous for each $\lambda \geq 0$;*
(iii) *$L^{\nu}(F,\cdot)$ is convex and continuous for each $F \in \Phi(\nu)$.*

PROOF. Noting that the present value $\mathcal{B}$ is a continuous map from Ω to $[-M/(1-\beta), M/(1-\beta)]$,

(i) follows from Lemma 1.1. Since $g_{\lambda}(\cdot)$ is upper semicontinuous,
(ii) follows from (2.5), also,
(iii) clearly holds. ∎

From Lemma 2.1, we observe that Fan's minimax theorem (cf. [13]) is applicable to obtain the following.

LEMMA 2.2. *It holds that, for any $\nu \in \mathcal{P}(S)$,*

$$\inf_{\lambda \geq 0} \max_{F \in \Phi(\nu)} L^{\nu}(F,\lambda) = \max_{F \in \Phi(\nu)} \inf_{\lambda \geq 0} L^{\nu}(F,\lambda). \tag{2.8}$$

Henceforth, the common value of (2.8) will be denoted by L^*. In order to prove the existence of a saddle-point with (2.7), we need the following condition.

SLATER CONDITION. *There exists a $\bar{\pi} \in \Pi$ such that*

$$E^{\nu}_{\bar{\pi}}(h_i(\mathcal{B})) < \alpha_i, \qquad \text{for all } i,\ 1 \leq i \leq k. \tag{2.9}$$

Since $L^{\nu}(\bar{F},\lambda) \longrightarrow \infty$ as $\|\lambda\| \longrightarrow \infty$ with $\bar{F} = F^{\nu}_{\bar{\pi}}$ under condition (2.9), the convex function $\max_{F \in \Phi(\nu)} L^{\nu}(F,\lambda)$ is bounded from below, so that there exists $\lambda^* \geq 0$ such that

$$L^{\nu}(F,\lambda^*) \leq L^*, \qquad \text{for all } F \in \Phi(\nu) \tag{2.10}$$

by (2.8). On the other hand, by Lemma 2.2, there exists $F^* \in \Phi(\nu)$ with

$$L^{\nu}(F^*,\lambda) \geq L^*, \qquad \text{for all } \lambda \geq 0. \tag{2.11}$$

Thus, applying Corollary 2.1, (2.10) and (2.11) lead the following main theorem.

THEOREM 2.2. *Under condition (2.9), the Lagrangian $L^{\nu}(\cdot,\cdot)$ with the initial probability measure $\nu \in \mathcal{P}(S)$ has a saddle-point, i.e., there exists $\pi^* \in \Pi$ and $\lambda^* \geq 0$ satisfying (2.2).*

Also, from Theorem 2.1 and 2.2, the following corollary holds.

COROLLARY 2.2. *Under condition (2.9), there exists a constrained optimal policy.*

3. CHARACTERIZATION OF THE CONSTRAINED OPTIMAL POLICY

In this section, by applying the results in [1], a constrained optimal policy is characterized by optimal action sets.

Let $\nu \in \mathcal{P}(S)$. Then, for each $\lambda \geq 0$, $\pi^* \in \Pi$ is called g_λ-optimal if

$$E_{\pi^*}^{\nu}(g_\lambda(\mathcal{B})) \geq E_{\pi}^{\nu}(g_\lambda(\mathcal{B})), \qquad \text{for all } \pi \in \Pi,$$

where g_λ is given in (2.6).

The following lemma can be easily proved (cf. [14]).

LEMMA 3.1. *Let $\bar{\pi} \in \Pi$ and $\bar{\lambda} = (\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_k) \in \mathbf{R}_+^k$. Then, for any $\nu \in \mathcal{P}(S)$, the Lagrangian $L^{\nu}(\cdot,\cdot)$ given in (2.1) has a saddle-point at $\bar{\pi}$, $\bar{\lambda}$ iff the following holds:*

(i) *$\bar{\pi}$ is $g_{\bar{\lambda}}$-optimal;*
(ii) *$\bar{\pi} \in \mathcal{V}(\nu, \alpha)$;*
(iii) *$\sum_{i=1}^{k} \bar{\lambda}_i(\alpha_i - E_{\bar{\pi}}^{\nu}(h_i(\mathcal{B}))) = 0$.*

To characterize $g_{\bar{\lambda}}$-optimality in Lemma 3.1(i), let

$$U_t\{g_\lambda\}(s,i,a,j) := \max_{F \in \Phi(j)} \int g_\lambda\left(s + \beta^t r(i,a,j) + \beta^{t+1} x\right) F(dx), \tag{3.1}$$

for $t \geq 0$, $s \in [-M/(1-\beta), M/(1-\beta)]$, and $i, j \in S$, where if $\nu \in \mathcal{P}(S)$ is degenerate at $\{j\}$, ν is simply denoted by j and $\Phi(\nu)$ by $\Phi(j)$. Since $g_\lambda(\cdot)$ is upper semicontinuous and $\Phi(j)$ is compact in the week topology, the maximum in (3.1) is attained. Here, for each $\lambda \geq 0$, we define the sequence $\{A_t^\lambda\}_{t=0}^{\infty}$ by

$$A_t^\lambda(s,i) := \arg\max_{a \in A(i)} \sum_{j \in S} q_{ij}(a) U_t\{g_\lambda\}(s,i,a,j), \tag{3.2}$$

for $s \in [-M/(1-\beta), M/(1-\beta)]$ and $i \in S$. Then, we have the following.

THEOREM 3.1. *For any $\nu \in \mathcal{P}(S)$, a policy $\pi^* \in \mathcal{V}(\nu, \alpha)$ is a constrained optimal policy iff there exists $\lambda^* \geq 0$ such that*

(i) *$P_{\pi^*}^{\nu}(\Delta_t \in A_t^{\lambda^*}(\mathcal{B}_{t-1}, X_t)) = 1$ where $\mathcal{B}_t = \sum_{s=0}^{t-1} \beta^s r(X_s, \Delta_s, X_{s+1})$ $(t \geq 1)$;*
(ii) *$\sum_{i=1}^{k} \lambda_i^*(\alpha_i - E_{\pi^*}^{\nu}(h_i(\mathcal{B}))) = 0$.*

PROOF. Applying the results of Theorem 3.3 in [1], it can be shown that π^* is g_{λ^*}-optimal iff the above (i) holds. So, Theorem 3.1 follows from Lemma 3.1. ∎

Consider the exponential utility case with $k = 1$, i.e., $g(x) = h_{\lambda_1}(x)$ and $h_1(x) = h_{\lambda_2}(x)$ $(\lambda_1, \lambda_2 \neq 0)$, where $h_\delta(\cdot)$ is a utility function with constant risk sensitivity δ, as follows:

$$h_\delta(x) := \begin{cases} \operatorname{sign}(\delta) e^{\delta x}, & \delta \neq 0, \\ x, & \delta = 0. \end{cases}$$

In this case, $g_\lambda(x)$ in (2.6) is given as $g_\lambda(x) = g(x) + \lambda(\alpha - h_1(x))$ with a Lagrange multiplier λ.

For each $\lambda \geq 0$ and $i \in S$, $t \geq 0$, $-\infty < x < \infty$, let

$$P_t^\lambda(i,s) = \sup_{F \in \Phi(i)} \int \left\{ \operatorname{sign}(\lambda_1) e^{\lambda_1 s + \beta^t \lambda_1 x} - \lambda \operatorname{sign}(\lambda_2) e^{\lambda_2 s + \beta^t \lambda_2 x} \right\} dF(x). \tag{3.3}$$

Then, the following recursive equation holds:

$$P_t^\lambda(i,s) = \max_{a \in A(i)} \sum q_{ij}(a) P_{t+1}^\lambda\left(j, s + \beta^t r(i,a,j)\right). \tag{3.4}$$

In fact, by using the dynamic programming method,

$$\begin{aligned} P_t^{\lambda}(i,s) &= \sup_{F\in\Phi(i)} \int \left\{ \operatorname{sign}(\lambda_1) e^{\lambda_1 s+\beta^t \lambda_1 x} - \lambda \operatorname{sign}(\lambda_2) e^{\lambda_2 s+\beta^t \lambda_2 x} \right\} dF(x) \\ &= \max_{a\in A(i)} \sum_j q_{ij}(a) \sup_{F\in\Phi(j)} \int \left\{ \operatorname{sign}(\lambda_1) e^{\lambda_1 \left(s+\beta^t r(i,a,j)\right)+\lambda_1 \beta^{t+1} x} \right. \\ &\quad \left. -\lambda \operatorname{sign}(\lambda_2) e^{\lambda_2 \left(s+\beta^t r(i,a,j)\right)+\lambda_2 \beta^{t+1} x} \right\} dF(x) \\ &= \max_{a\in A(i)} \sum_j q_{ij}(a) P_{t+1}^{\lambda}\left(j, s+\beta^t r(i,a,j)\right). \end{aligned}$$

Obviously,

$$\lim_{t\to\infty} P_t^{\lambda}(i,s) = \operatorname{sign}(\lambda_1) e^{\lambda_1 s} - \lambda \operatorname{sign}(\lambda_2) e^{\lambda_2 s}. \tag{3.5}$$

Also, $U_t\{g_\lambda\}$ in (3.4) is written as follows:

$$U_t\{g_\lambda\}(s,i,a,j) = P_{t+1}^{\lambda}\left(j, s+\beta^t r(i,a,j)\right) + \lambda\alpha. \tag{3.6}$$

We note that the efficient algorithm for obtaining a constrained optimal policy by Theorem 3.1 is not so easy. Implementing a numerical work or applying the result in the real world problem should be our future work.

REFERENCES

1. Y. Kadota, M. Kurano and M. Yasuda, Discounted Markov decision processes with general utility, In *Proceeding of APORS' 94*, pp. 330–337, World Scientific, (1995).
2. Y. Kadota, M. Kurano and M. Yasuda, On the general utility of discounted Markov decision processes, *Int. Trans. Oper. Res.* **5** (1), 27–34, (1998).
3. M. Kurano, J. Nakagami and Y. Huang, Markov decision processes with compact state and action spaces: The average case, *Optimization* **48**, 255–269, (2000).
4. K.J. Chung and M.J. Sobel, Discounted MDP's: Distribution functions and exponential utility maximization, *SIAM J. Control Optim.* **25**, 49–62, (1987).
5. E.V. Denardo and U.G. Rothblum, Optimal stopping, exponential utility and linear programming, *Math. Prog.* **16**, 228–244, (1979).
6. E. Altman, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, (1999).
7. L.I. Sennot, Constrained discounted Markov decision chains, *Probability in the Engineering and Information Sciences* **5**, 463–475, (1991).
8. S.C. Jaquette, Markov decision processes with a new optimality criterion: Discrete time, *Ann. Stat.* **1**, 496–505, (1973).
9. M. Bouakit and Y. Kebir, Target-level criterion in Markov decision processes, *J. Optim. Theory Appl.* **86**, 1–15, (1995).
10. D.J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173**, 634–646, (1993).
11. V.S. Borkar, *Topics in Controlled Markov Chains*, Longman Scientific Technical, (1991).
12. D. Luenberger, *Optimization by Vector Space Methods*, John Wiley, New York, (1969).
13. J.M. Borwein and D. Zhuang, On Fan's minimax theorem, *Math. Programming* **34**, 232–244, (1986).
14. M. Avriel, *Nonlinear Programming, Analysis and Methods*, Prentice-Hall, (1976).
15. R.S. Howard and J.E. Matheson, Risk-sensitive Markov decision processes, *Management Sci.* **8**, 356–369, (1972).