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.*