CDC00-INV1901

Markov Decision Processes with Constrained Stopping Times

M. Horiguchi[†], M. Kurano[‡] and M. Yasuda[§]

[†]Graduate School of Science and Technology, [‡] Faculty of Education, [§]Faculty of Science, Chiba University, Yayoi-chi, Inage-ku, Chiba 263-8522, Japan

Key words: Stopped Markov decision, constrained stopping time, Lagrange multiplier, OLA stopping time

Extended Abstract

A constrained optimal stopping problem seems to be originated by Nachman(1980) and Kennedy(1982), in which a Lagrangean approach is used to reduce the problem to an unconstrained stopping problem of a conventional type and then the constrained optimal stopping time is characterized. Markov decision process with a constraint has been studied by many authors (cf.Altman(1994,1999), Beutler and Ross(1985), Frid(1972) Hordijk and Kallenberg(1984), Sennott(1991,1993)). For the case of one constraint, Beutler and Ross has formed a Lagrange form from the average expected reward and, by analysing the corresponding parametric dynamic programming equation has shown that there exists an optimal constrained stationary policy requiring randomization between two actions in at most one state under some ergodic conditions. This Lagrange approach was generalized to the countable state case by Sennott.

On the other hand, a combined model of the Markov decision process and stopping problem, called a stopped decision process, has been considered by Furukawa and Iwamoto(1970) where the existence of an optimal pair of policy and stopping time associated with some optimality criterions are discussed. Hordijk(1974) has considered this model from a standpoint of potential theory. Also, the general utility-treatment for stopped decision processes has been studied by Kadota, Kurano and Yasuda(1996,1998).

In this paper, the optimization problem for a stopped decision process is considered over stopping times τ constrained so that $\mathbb{E}\,\tau \leqq \alpha$ for some fixed $\alpha > 0$. We introduce the concept of a randomized stationary stopping time which is a mixed extension of the entry time of a stopping region and prove the existence of an optimal constrained pair of stationary policy and stopping time utilizing a Lagrange multiplier approach. The proof is done by applying a Lagrange multiplier method developed by Frid(1972), Beutler and Ross(1985) and Sennott(1999). Also, using the idea of the one-step look ahead (OLA) policy an optimal constrained pair is sought concretely. As an example, the constrained Markov deteriorating system is considered.

In the reminder of the abstract, we shall establish the notation that will be used throughout the paper and define a problem which is to find an optimal constrained pair of policy and stopping time referring to Hordijk(1974). A dynamic system, at times $t = 0, 1, 2, \ldots$, is observed to be in one of a possible number of states. Let S be the countable state space, denoted by $S = \{0, 1, 2, \ldots\}$. We denote by $\mathcal{P}(S)$ the set of all probability vectors on S, i.e.,

$$\mathcal{P}(S) = \left\{ p = (p_0, p_1, p_2, \ldots) \,\middle|\, \sum_{i \in S} p_i \leqq 1, p_i \geqq 0 (i \in S) \right\}.$$

We allow for breaking down or disappearing of the system with positive probability, so $\sum_{i \in S} p_i \leqq 1$. For each $i \in S$, $\mathcal{P}(i)$ is a subset of $\mathcal{P}(S)$, which is assumed to be given. If at time t the system observed is state i and the decision maker takes $p(i, \cdot) \in \mathcal{P}(i)$, then the system moves to a new state $j \in S$ selected according to the probability distribution $p(i, \cdot)$. This decision process is then repeated from the new state j.

Let $\mathcal{P}$ be the set of all stochastic matrices where i-th row vector $p(i, \cdot) \in \mathcal{P}(i)$. A notion of convergence on $\mathcal{P}$ is given as follows: a sequence $P_n = (p_n(i, j); i, j \in S) \in \mathcal{P}$ converges to $P = (p(i, j); i, j \in S) \in \mathcal{P}$ if $p_n(i, j) \to p(i, j) (n \to \infty)$ for each $i, j \in S$. In this case, we write $\lim_{n \to \infty} P_n = P$. Also, $\mathcal{P}$ with this topology is a metric space (cf. Hordijk(1974)).An element of $\mathcal{P}$ is called a transition matrix. The policy R for controlling the system is a sequence of transition matrices, $P_0, P_1, \cdots \in \mathcal{P}$, denoted by $R = (P_0, P_1, \ldots)$, where P_t gives the transition probability at time $t (t \geqq 0)$. Here we confine ourselves to memoryless or Markov policies, which is shown to be sufficient to our optimization problem (cf. Theorem 13.2 in Hordijk(1974)).We denote by $\mathcal{R}$ the set of all policies. If the policy takes at all times the same transition matrix, i.e., $P^\infty := (P, P, \ldots), P \in \mathcal{P}$, it is called a stationary policy and induces a stationary Markov chain. When in state i transition $p(i, \cdot)$ is taken from $\mathcal{P}(i)$ an immediate cost depending on i and $p(i, \cdot)$ is incurred. Let $c_P(i)$ be the immediate cost when taking $p(i, \cdot)$ (the i-th row of matrix P) in state i and write c_P for the vector with i-th component $c_P(i)$. Note that for $P, Q \in \mathcal{P}$ with $p(i, \cdot) = q(i, \cdot), c_P(i) = c_Q(i)$.

The sample space is the product space $\Omega = S^\infty$ such that the projection X_n on the n-th factor S describes the state at time n. For each $R \in \mathcal{R}$ and initial state $i \in S$, we can define the measure $\mathbb{P}_{i,R}$ in an obvious way. The expectation of the cost at time n is denoted by $\mathbb{E}_{i,R}\, c(X_n)$, where $\mathbb{E}_{i,R}$ is the expectation with respect to $\mathbb{P}_{i,R}$. The vector whose i-th component is $\mathbb{E}_{i,R}\, c(X_n)$ is written by $\mathbb{E}_R\, c(X_n)$. For stationary policy P^∞, it is written by $\mathbb{E}_P\, c(X_n)$ instead of $\mathbb{E}_{P^\infty}\, c(X_n)$. It is easily seen that, for $R = (P_0, P_1, \ldots)$,

$$\mathbb{E}_R\, c(X_n) = P_0 P_1 \cdots P_{n-1} c_{P_n} \quad (n \geqq 1).$$

We denote by $\sigma(X_0, X_1, \ldots, X_n)$ the σ-field induced by $\{X_0, X_1, \ldots, X_n\}$. Let $\mathcal{F}_n = \sigma(X_0, X_1, \ldots, X_n), (n \geqq 0)$ and $\mathcal{F}_\infty$ be the smallest σ-field containing each $\mathcal{F}_n, n \geqq 0$. Let $N = \{0, 1, 2, \ldots\} \cup \{\infty\}$. We call a map $\tau : \Omega \to N$ a stopping time with respect to the filtration $\mathcal{F} = \{\mathcal{F}_n, n \in N\}$ if $\{\tau = n\} \in \mathcal{F}_n$ for all $n \in N$. The class of stopping time with respect to $\mathcal{F}$ will be denoted by $C(\mathcal{F})$. In order to solve our problem described in the sequel, we need introduce randomized stopping time (cf. Chow, Robbins and Siegmund(1976)).To this purpose, enlarging Ω to $\overline{\Omega} := \Omega \times [0,1]$, we can embed $(\Omega, \mathcal{F}_\infty, \rho)$ to $(\overline{\Omega}, \mathcal{F}_\infty \times \mathbb{B}_1, \rho \times \mu)$, where $\mathbb{B}_1$ is Borel subsets of $[0,1]$ and μ a Lebesgue measure on $\mathbb{B}_1$. For a filtration $\mathcal{G} = \{\mathcal{G}_n, n \in N\}$ with $\mathcal{G}_n = \mathcal{F}_n \times \mathbb{B}_1$ we can assume without loss of generality that for each $n \in N$

$$\mathcal{F}_n \subset \mathcal{G}_n \tag{1.1}$$

and

$$\text{Prob.}(A|\mathcal{F}_n) = \text{Prob.}(A|\mathcal{G}_n) \text{ for } A \in \mathcal{F}_\infty. \tag{1.2}$$

We call a map $\tau : \overline{\Omega} \to N$ a (randomized) stopping time with respect to $\mathcal{G}$ if $\{\tau = n\} \in \mathcal{G}_n$ for each $n \in N$. The class of stopping times with respect to $\mathcal{G}$ will be denoted by $C(\mathcal{G})$. Obviously $C(\mathcal{G}) \supset C(\mathcal{F})$. A $\tau \in C(\mathcal{G})$ is called randomized stationary if for each $n \geqq 0$ Prob. $\left(\tau = n \middle| X_0, X_1, \ldots, X_{n-1}, X_n = i, \tau \geqq n\right)$ is independent of n and depending only on $i \in S$. In such a case, we can define the set $\{\delta(i), i \in S\}$ by

$$\delta(i) = \text{Prob.}\left(\tau = n \middle| X_0, X_1, \ldots, X_{n-1}, X_n = i, \tau \geqq n\right). \tag{1.3}$$

Then obviously

$$0 \leqq \delta(i) \leqq 1 \text{ for each } i \in S. \tag{1.4}$$

Conversely, for any set $\{\delta(i), i \in S\}$ satisfying (1.4), we can define a randomized stationary stopping time τ through (1.3). Such a stopping time is said to be determined by $\{\delta(i)\}$. When $\delta(i) = 0$ or $\delta(i) = 1$ for all $i \in S$, the corresponding stopping time is called simply stationary, which is a entry time of $\Gamma := \{i \in S | \delta(i) = 1\}$, denoted by τ_Γ. Given a reward function r in S, the general optimization problem for stopped decision process is to find a policy $R \in \mathcal{R}$ and stopping time $\tau \in C(\mathcal{G})$ such that

$$J_{R,\tau}(i) = \mathbb{E}_{i,R}\left(\sum_{n=0}^{\tau-1} c(X_n) + r(X_\tau)\right) \tag{1.5}$$

is maximized. Note that $\tau = \infty$ with positive probability is admissible with zero reward. Let $\alpha > 0$ be given arbitrarily. Constrained optimal pairs are defined with respect to a given initial state. So without loss of generality we may assume initial state "0". Let

$$\Delta(\mathcal{G}) = \left\{(R,\tau) \in \mathcal{R} \times C(\mathcal{G}) \middle| \mathbb{E}_{0,R}(\tau) \leqq \alpha \text{ and } \mathbb{E}_R(r(X_\tau)) < \infty\right\}.$$

In this paper, we will consider the constrained optimization problem:

$$\text{maximize } J_{R,\tau}(0), \text{ subject to } (R,\tau) \in \Delta(\mathcal{G}). \tag{1.6}$$

The constrained pair $(R^*, \tau^*) \in \Delta(\mathcal{G})$ is called optimal in state $0 \in S$ if

$$J_{R^*,\tau^*}(0) \geqq J_{R,\tau}(0) \tag{1.7}$$

for all $(R,\tau) \in \Delta(\mathcal{G})$. In Section 2, we survey some results on constrained problems referring Hordijk(1974) which are used in the sequel sections. Also, new results are given concerning with the OLA stopping time. In Section 3, the Lagrange multiplier is introduced to the constrained problem and the parametrized version of stopped decision process is analysed, whose results are utilized for solving the constrained optimization problem in Section 4.