数理解析研究所講究録
第 780 巻 1992 年 46-50

On an ε-optimal policy in dynamic programming with a finite horizon

新潟大学理学部 田中謙輔 (Kensuke Tanaka)

1 Abstract

In this paper, a dynamic programming with a finite horizon is investigated. In many cases, the concept of an optimal policy is introduced and, then, the existence of the optimal policy is shown. But, in order to show the existence of the optimal policy, we need to assume the stronger condition such that an action space is compact. Here, excluding the compact property of the action space and applying Ekeland's theorem, we shall study the characterization of an ε-optimal policy on the dynamic programming.

2 Formulation of dynamic programming with a finite horizon

A dynamic decision model is specified by a set of seven elements

$$(S, A, F, q, r, N, v), \tag{1}$$

where

1. $S = \{1, 2, 3, \cdots\}$ is a countable set, the set of states of the decision system.
2. A is a Polish (i.e., complete, separable, metric) space, namely, the action space.
3. F is a multifunction which assigns to each state $i \in S$ a non-empty feasible set of actions $F(i) \subset A$.
4. q is a transition probability on S given each $i \in S$ and $a \in F(i)$,i.e., $q(j \mid i, a)$ is a probability of a state j for each $i \in S$ and $a \in F(i)$ and a Borel measurable function with respect to $a \in F(i)$ for each state $i, j \in S$.
5. $r(i, a)$ is a real-valued function, $S \times A \to R$, the one-step loss function.
6. N is a positive integer, a terminal stage number.
7. v is a real-valued function on S, $S \to R$, a terminal loss function on the stage N.

In the specification, we should note that the feasible set of actions $F(i)$ depends on a state $i \in S$ and $q(\cdot \mid i, a)$ is independent of the times.

Then, a policy is defined as a finite sequence $(f_1, f_2, \cdots, f_{N-1})$, each element f_n of which is a decision function from S into A such that $f_n(i) \in F(i)$ for each $i \in S$. Thus, each decision function indicates an action to use for each $i \in S$. We assume that we use only such policies on the decision system. Throughout this paper, the class of all policies is denoted by π . Thus, the dynamic decision system is interpreted as follows. If a policy $\pi = (f_1, f_2, f_3, \cdots, f_{N-1})$ is employed, at the decision time $n, n = 1, 2, 3, \cdots, N-1$, we observe the state of the decision system and classify it to a possible state $x_n \in S$. So, we choose an action $f_n(x_n) \in F(x_n)$ by the decision function f_n. As a result of the choice $a_n, a_n = f_n(x_n)$, we will incur a loss $r(x_n, a_n)$. Then, the decision system moves to a new state $x_{n+1} \in S$ according to transition probability $q(\cdot \mid x_n, a_n)$. After that, the process of the dynamic decision system is developed from x_{n+1} up to the terminal stage N and, at the terminal stage N, we will incur a loss $v(x_N)$. So, given an initial distribution $p(\cdot)$ on S and any policy π together with transition probability q, we define a probability p_n^π on the state space S at each time n (see K.Hinderer [7, p.80] in detail). Thus, the expected loss at each time n is given by

$$E_\pi[r(x_n, a_n)] = \sum_{i \in S} r(i, f_n(i)) p_n^\pi(i). \tag{2}$$

So, if a policy $\pi = (f_1, f_2, \cdots, f_{N-1})$ is employed, the total expected loss is given by

$$I(\pi) = \sum_{n=1}^{N-1} E_\pi[r(x_n, a_n)] + E_\pi[v(x_N)] \tag{3}$$

$$= \sum_{n=1}^{N-1} \sum_{i \in S} r(i, f_n(i)) p_n^\pi(i) + \sum_{i \in S} v(i) p_N^\pi(i). \tag{4}$$

Then, we consider a basic minimization problem (P) for the dynamic decision system:

(P) minimize $I(\pi)$ subject to π.

3 The existence of an ε-optimal policy in the dynamic system

In order to show that there exists at least an ε-optimal policy, let $M(S)$ be the set of all real-valued functions, bounded from below, on S. We impose some assumptions on F, q, r, and v as follows.

(A1) F is a closed-valued multifunction from S into A, that is, $F(i)$ is a closed subset in A for each $i \in S$.

(A2) The loss function r is a real-valued function, $S \times A \to R$, bounded from below on $S \times A$, and, for each $i \in S$, $r(i, a)$ is a lower semi-continuous (l.s.c.) with respect to

$a \in F(i)$, that is, for any convergent sequence of actions $\{a_k\}, k = 1, 2, \cdots$, in $F(i)$ such that $\rho(a_k, a) \to 0$ as $k \to \infty$,

$$\liminf_{k\to\infty} r(i, a_k) \geq r(i, a) \qquad \text{for each } i \in S,$$

where ρ is the metric on the action space A.

(A3) For any $u \in M(S)$ and $i \in S$,

$$\sum_{j\in S} u(j)q(j \mid i, a)$$

is a l.s.c. function with respect to $a \in F(i)$.

(A4) $\quad v \in M(S)$.

Now, let D denote the set of all feasible decision functions $f : S \to A$, such that $f(i) \in F(i)$ for each $i \in S$. We define, for each $f \in D$, an operator $T(f)$ on $M(S)$ as follows: for each $u \in M(S)$ and $i \in S$,

$$T(f)u(i) = r(i, f(i)) + \sum_{j\in S} u(j)q(j \mid i, f(i)). \tag{5}$$

Then, if a policy $\pi = (f_1, f_2, f_3, \cdots, f_{N-1}), f_n \in D, n = 1, 2. \cdots, N-1$, is employed, the total expected loss $I(\pi)(i)$ with an initial state $i \in S$ can be rewritten as

$$I(\pi)(i) = T(f_1)T(f_2)\cdots T(f_{N-1})v(i). \tag{6}$$

So, in order to show the existence of an ε-optimal policy, first we use the following lemma.

Lemma 3.1 Let U be a complete metric space and $G : U \to R$, l.s.c. function, bounded from below. Then, for any $\varepsilon > 0$, there is some point $u^* \in U$ such that

$$G(u^*) \leq \inf_{u\in U} G(u) + \varepsilon$$

and, for all $u \in U$,

$$G(u) \geq G(u^*) - \varepsilon d(u^*, u),$$

where d is the metric on U.

The proof of the lemma is easily derived by Ekeland's theorem (see [7] and [7] in detail).

Lemma 3.2 For any $\varepsilon > 0$, $u \in M(S)$ and $i \in S$, there is some action $a^* \in F(i)$ such that

$$T(a^*)u(i) \leq \inf_{a\in F(i)} T(a)u(i) + \varepsilon \tag{7}$$

and, for all actions $a \in F(i)$,

$$T(a)u(i) \geq T(a^*)u(i) - \varepsilon\rho(a^*, a). \tag{8}$$

Proof. Since, from (A1),$F(i)$ is a closed set, $F(i)$ is complete for each state i. Further, from (A2) and (A3), it follows that, for each $i \in S$, $T(a)u(i)$ is l.s.c. on $F(i)$ and bounded from below on $S \times A$. So, using Lemma 3.1, we can easily show the result of the lemma.

Theorem 3.1 For any $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{N-1}), \varepsilon_n > 0, n = 1, 2, \cdots, N-1$, and any policy $\pi = (f_1, f_2, \cdots, f_{N-1}) \in \Pi$, there exists an ε-optimal policy $\pi = (f_1^*, f_2^*, \cdots, f_{N-1}^*)$ such that

$$I(\pi) \geq I(\pi^*) - \sum_{n=1}^{N-1} \varepsilon_n E_\pi[\rho(f_n^*(x_n), f_n(x_n))] \tag{9}$$

$$= I(\pi^*) - \sum_{n=1}^{N-1} \varepsilon_n \sum_{j \in S} [\rho(f_n^*(j), f_n(j)) p_n^\pi(j) \tag{10}$$

Proof. For any policy $\pi = (f_1, f_2, \cdots, f_{N-1})$ and the initial distribution p, it follows from (6) that

$$I(\pi) = \sum_{i \in S} T(f_1)T(f_2), \cdots, T(f_{N-1})v(i)p(i). \tag{11}$$

So, from (11), $I(\pi)$ is successively constructed by the operator T. From Lemma 3.2, it follows that there exists a decision function $f_{N-1}^* \in D$ such that, for each $i \in S$,

$$T(f_{N-1}^*)v(i) \leq \inf_f T(f)v(i) + \varepsilon_{N-1}$$

and, for all $f \in D$,

$$T(f)v(i) \geq T(f_{N-1}^*)v(i) - \varepsilon_{N-1}\rho(f_{N-1}^*(i), f(i)). \tag{12}$$

Thus, applying the Lemma 3.2 to $T(f)T(f_{N-1}^*)v(i)$, we obtain a decision function $f_{N-2}^* \in D$ such that

$$T(f_{N-2}^*)T(f_{N-1}^*)v(i) \leq \inf_f T(f)T(f_{N-1}^*)v(i) + \varepsilon_{N-2},$$

and, for all $f \in D$,

$$T(f)T(f_{N-1}^*)v(i) \geq T(f_{N-2}^*)T(f_{N-1}^*)v(i) - \varepsilon_{N-2}\rho(f_{N-2}^*(i), f(i)). \tag{13}$$

Then, since T is a monotone operator, combining (12) with (13), we obtain for the given policy $\pi = (f_1, f_2, \cdots, f_{N-1})$,

$$\begin{aligned} T(f_{N-2})T(f_{N-1})v(i) &\geq T(f_{N-2})T(f_{N-1}^*)v(i) \\ &\quad - \varepsilon_{N-1}E_{f_{N-2}}[\rho(f_{N-1}^*(x_{N-1}), f_{N-1}(x_{N-1})) \mid x_{N-2} = i] \\ &\geq T(f_{N-2}^*)T(f_{N-1}^*)v(i) - \varepsilon_{N-2}\rho(f_{N-2}^*(i), f_{N-2}(i)) \\ &\quad - \varepsilon_{N-1}E_{f_{N-2}}[\rho(f_{N-1}^*(x_{N-1}), f_{N-1}(x_{N-1})) \mid x_{N-2} = i] \end{aligned} \tag{14}$$

Applying Lemma 3.2 to (14) repeatedly, we obtain

$$I(\pi)(i) \geq T(f_{N-2}^*)T(f_{N-1}^*) \cdots T(f_{N-1}^*)v(i) - \sum_{n=1}^{N-1} \varepsilon_n E_\pi[\rho(f_n^*(x_n), f_n(x_n)) \mid x_1 = i] \tag{15}$$

Thus, multiplying both sides of (15) by the initial distribution p, and then, summing with respect to i on S, we have

$$I(\pi) \geq I(\pi^*) - \sum_{n=1}^{N-1} \varepsilon_n E_\pi[\rho(f_n^*(x_n), f_n(x_n))].$$

which completes the proof of the theorem.

Remark. From the theorem, there exists an ε-optimal policy $\pi = (f_1^*, f_2^*, \cdots, f_{N-1}^*) \in \Pi$ such that, for any policy $\pi = (f_1, f_2, \cdots, f_{N-1})$ and any $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{N-1}), \varepsilon_n > 0, n = 1, 2, \cdots, N-1$, (10) holds. So, we define a function $H : \Pi \to R$, as follows

$$H(\pi) = I(\pi) + \sum_{n=1}^{N-1} \varepsilon_n E_\pi[\rho(f_n^*(x_n), f_n(x_n))]$$

Then, $H(\pi)$ attains the minimum at π^* and $H(\pi^*) = I(\pi^*)$.

References

[1] E.B.Dynkin and A.A.Yushkevich (1979), Controlled Markov Process, Springer-Verlag, New York.

[2] I.Ekeland (1974), *On the variational principle* , J. Math. Anal. Appl., Vol.47, 325-353.

[3] I.Ekeland (1979), *Nonconvex minimization problems* , Bull. Am. Math. Soci., Vol.3, No.1, 443-474.

[4] C.J.Himmelberg, T.Pathasarathy, and F.S.van Vleck (1976), *Optimal plans for dynamic programming problems* , Math. Oper. Res., Vol.1, 390-394.

[5] K.Hinderer (1970), Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter, Springer-Verage, Berlin.

[6] K.Tanaka and K.Yokoyama (1991), *On ε-equilibrium point in a noncooperative n-person game* , J. Math. Anal. Appl. Vol.160, No.2, 413-423.