数理解析研究所講究録
第 829 巻 1993 年 107-116

On an ε-optimal Policy in Dynamic Programming with a Discount Factor

新潟大学理学部 田中謙輔 (Kensuke Tanaka)

1. 序 論

この発表では、或 $D.P$ システムが与えられている時に最適政策は存在しないかも知れないので、任意の指定された政策 f_0 が任意の$\varepsilon > 0$ に対して $\varepsilon-$最適になるように $D.P$ システムの損失関数 F を修正したい。これに関連すると思われる内容は、少し古くなるが inverse optimal problem と呼ばれる $D.P$ 問題であろう（参考文献 [1],[2],[3],[4] ）。このような問題に対して、いろいろな接近方法が考えられるかも知れませんが、ここでは一つの接近方法として *epigraph* の概念より$\varepsilon-$劣微分の性質を用いる方法を展開したい、即ち $f^* \in \partial_\varepsilon F(f_0)$ に対して、

$$F(f) - F(f_0) \geq < f^*, f - f_0 > -\varepsilon \quad \forall f.$$

この時に、損失関数を

$$G(f) = F(f) - < f^*, f >$$

に修正すれば f_0 は修正 $D.P$ システムの ε-最適政策となるだろう。

ここで、上の ε-最適政策 f_0 のもとで以下の Ekeland's theorem が適用可能となり、この指定された政策 f_0 の近くにある、或政策がこの修正 $D.P$ システムの最適政策になるように更に修正出来る事になるだろう。

Ekeland's theorem

Let (X, d) be a complete metric space, and $G : X \longrightarrow R\cup\{+\infty\}$, a l.s.c. function, $\not\equiv +\infty$, bounded from below. Let $\varepsilon > 0$, and a point $f_0 \in X$ such that

$$G(f_0) \leq \inf_{f \in X} G(f) + \varepsilon.$$

Then there exists some point $f_* \in X$ such that

$$G(f_*) \leq G(f_0)$$

$$d(f_*, f_0) \leq 1$$

$$\forall f \neq f_* \quad G(f) > G(f_*) - \varepsilon d(f_*, f).$$

2. Formulation of Markov decision problem

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

$$(S, A, F, q, r, \beta), \tag{2.1}$$

where

(i) S is a non-empty Borel subset of a Polish(i.e., complete, separable, metric) space with the Borel $\sigma-$field $\beta(S)$, the set of states of the decision system.

(ii) A is a Polish space with the Borel $\sigma-$field $\beta(A)$, namely, the action space.

(iii) F is a multifunction which assigns to each state $s \in S$ a non-empty permissible set of actions $F(s) \subset A$. We assume that $GrF = \{(s,a)|a \in F(s), s \in S\}$ is a Borel subset in $S \times A$ with the Borel $\sigma-$field $\beta(S) \times \beta(A)$.

(iv) q is a transition probability measure $q(\cdot|s,a)$ on the Borel subsets of S given each $(s,a) \in GrF$, i.e., $q(B|s,a)$ is a probability of a Borel subset $B \in \beta(S)$ for each $(s,a) \in GrF$ and a Borel measurable function of $(s,a) \in GrF$ for each Borel subset B. The law of motion of the decision system is given by q.

(v) $r(s,a)$ is a real-valued Borel measurable function, $GrF \longrightarrow R$, the one-step loss function.

(vi) β is a discount factor, $0 < \beta < 1$.

In the specification, we should note that the permissible set of actions $F(s)$ depends on a state $s \in S$ and $q(\cdot|s,a)$ is independent of the time.

Then, a policy π is defined as a sequence of infinite decision functions $(f_1, f_2, \cdots, f_t, \cdots)$, each function f_t of which is a Borel measurable selection for the constraint multifunction F, i.e., f_t is a Borel measurable mapping from S into A such that $f_t(s) \in F(s)$ for each state $s \in S$. Thus, such a decision function indicates an action to use for each state $s \in S$. We assume that we use only such policies, which are called Markov policies, on the decision system. Especially, if any decision function f_t in a Markov policy π is independent of the stage number and dependent only on the present state, that is, $f_t = f$ for all t, this policy π is said to be stationary and is written as f instead of π. The class of all Markov policies is denoted by Π in the paper.

Now, the dynamic decision system is interpreted as follows. If a policy $\pi = (f_1, f_2, f_3, \cdots, f_t, \cdots)$ is employed, at the successive decision time $t, t = 1, 2, 3, \cdots$, we observe the state of the decision system and classify it to a possible state $s_t \in S$. So, we choose an action $a_t \in F(s_t), a_t = f_t(s_t)$, by the decision function f_t. As a result of the state s_t and the choice a_t at the time t, we will incur a loss $r(s_t, a_t)$. Then, the decision system moves to a new state $s_{t+1} \in S$ according to transition probability $q(\cdot|s_t, a_t)$. After that, the process of the dynamic decision system is analogously developed from s_{t+1}. So, given an initial state $s_1 = x$ on S, any policy π together with the transition probability q, gives a probability measure p_t^π on the state space S at each time t in the decision system.

Thus, the expected loss at each time t is given by

$$E_\pi[r(s_t, a_t)|s_1 = x] = \int_S r(s, f_t(s)) p_t^\pi(ds|s_1 = x). \tag{2.2}$$

So, if a policy $\pi = (f_1, f_2, \cdots, f_t, \cdots)$ is employed under the discount factor β, the total expected loss is given by

$$\begin{aligned} I(\pi)(x) &= \sum_{t=1}^{\infty} \beta^{t-1} E_\pi[r(s_t, a_t)|s_1 = x] \\ &= \sum_{t=1}^{\infty} \beta^{t-1} \int_S r(s, f_t(s)) p_t^\pi(ds|s_1 = x). \end{aligned} \tag{2.3}$$

Then, assuming that $\inf_{\pi \in \Pi} I(\pi)(x) > -\infty$ under an initial state $x \in S$, we consider a basic minimization problem (P) for the dynamic decision system:

$$\text{(P)} \qquad \text{minimize } I(\pi)(x) \quad \text{subject to } \Pi.$$

In this problem (P), if there exists a policy $\overline{\pi}$ such that, for $\varepsilon \geq 0$

$$I(\pi)(x) \geq I(\overline{\pi})(x) - \varepsilon \quad \text{for all } \pi \in \Pi,$$

the policy $\overline{\pi}$ is said to be an $\varepsilon-$optimal one. Especially, if $\varepsilon = 0$, $\overline{\pi}$ is said to be an optimal policy.

3. An $\varepsilon-$optimality of a given stationary policy in the modified dynamic system

In order to show that some specified stationary policy becomes an optimal one in the Markov decision system with modified loss functions, let $M(S)$ be the set of all real-valued Borel measurable and bounded functions on S. Further, let $V(S)$ be the set of all extended real-valued Borel measurable functions on S, each of which is function from S into $R \cup \{\infty\}$. We impose some assumptions on F, q, and r as follows.

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

(A2) The loss function r is a real-valued Borel measurable, bounded function, $GrF \to R$ and, for each $s \in S$, $r(s, a)$ is convex, and lower semi-continuous (l.s.c.) with respect to $a \in F(s)$,

(A3) For any $u \in V(S)$ and $s \in S$,

$$\int_S u(y)q(dy|s,a)$$

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

$$\inf_{a \in F(s)} \int_S u(y)q(dy|s,a) > -\infty.$$

REMARK 3.1 *If stochastic Markov policies are only used in the decision system, it will be reasonable that the integral in* $(A3)$ *is a convex and continuous function.*

Now, let D denote the set of all permissible decision functions $f : S \to A$, in which each f is Borel measurable selection and $f(s) \in F(s)$ for each $s \in S$. In view of a selection theorem [8], $D \neq \emptyset$ if F is lower measurable set-valued function, that is, for every open set O in the action space A, the set $\{s \in S | F(s) \cap O \neq \emptyset\} \in \beta(S)$. We define, for each $f \in D$, an operator $T(f)$ on $V(S)$ as follows: for each $u \in V(S)$ and $s \in S$,

$$T(f)u(s) = r(s, f(s)) + \beta \int_S u(y)q(dy|s, f(s)). \tag{3.1}$$

Further, we define an operator T_0 on $V(S)$ by

$$T_0 u(s) = \inf_{f \in D} T(f)u(s). \tag{3.2}$$

Evidently, from (A3), $T_0 u(s) \in V(S)$, whenever $u \in V(S)$. If the domain of the operator T is limited to $M(S)$, the operator T_0 is a contraction operator on Banach space $M(S)$ with supnorm. So, T_0 has a unique fixed point u^* in $M(S)$, that is, $u^* = T_0 u^*$. It is well known that $u^*(x)$ is an optimal value for the problem (P), that is,

$$u^*(x) = \inf_{\pi \in \Pi} I(\pi)(x). \tag{3.3}$$

See E.B.Dynkin and A.A.Yushkevich [5] in detail.

Then, if a policy $\pi = (f_1, f_2, f_3, \cdots, f_t, \cdots)$, $f_t \in D, t = 1, 2. \cdots$, is employed, for any time k, the total expected loss $I(\pi)(x)$ with the initial state $x \in S$ can be rewritten as

$$\begin{aligned} I(\pi)(x) &= \sum_{t=1}^{\infty} \beta^{t-1} E_\pi[r(s_t, a_t)|s_1 = x] \\ &= \sum_{t=1}^{k} \beta^{t-1} E_\pi[r(s_t, a_t)|s_1 = x] + \beta^k E_\pi[I(\pi^{k+1})(s_{k+1})|s_1 = x] \\ &= T(f_1)T(f_2)\cdots T(f_k)I(\pi^{k+1})(x), \end{aligned} \tag{3.4}$$

where π^{k+1} denotes a policy constructed by a sequence of decision functions after the time $t = k + 1$ in the policy π, i.e., $\pi^{k+1} = (f_{k+1}, f_{k+2}, f_{k+3}, \cdots)$.

Now, we need the notations in the convex analysis to prove the main theorem in this paper. So, firstly we define the extended function $G(\cdot)u(s)$ of $T(\cdot)u(s)$ on $F(s) \subset A$ as follows:

$$G(a)u(s) = \begin{cases} T(a)u(s) & \text{if } a \in F(s) \\ \infty & \text{if } a \notin F(s) \text{ and } a \in A. \end{cases}$$

Secondly, we define the epigraph of $G(a)u(s)$ as follows:

$$\text{epi}G(\cdot)u(s) = \{(a, r) | r \geq G(a)u(s), a \in A\}.$$

Then, the following lemma in the convex analysis plays an important role.

LEMMA 3.1 *Let $B$ be a Banach space and any function $g : B \longrightarrow R \cup \{\infty\}$, $\not\equiv \infty$, convex, and l.s.c.. Then, for any $\varepsilon > 0$, the $\varepsilon-$subdifferential of $g$ at $b_0 \in dom(g)(dom(g)$ is the set where $g$ is finite), $\partial_\varepsilon g(b_0)$ is a nonempty, convex and weak*-closed subset in B^*, where B^* denotes the dual space of $B$ and*

$$\partial_\varepsilon g(b_0) = \{b^* \in B^* | g(b) \geq g(b_0) + < b^*, b - b_0 > -\varepsilon \text{ for all } b \in B\}. \tag{3.5}$$

This lemma is proved by using the properties of epigraph of $g$, epi$g$, in [11].

REMARK 3.2 *If $b_0 \in int(dom(g))$, for any $\varepsilon > 0$, $\partial_\varepsilon g(b_0)$ is a nonempty, convex, weak* compact and locally bounded.*

REMARK 3.3 *In [10], the subdifferential $\partial g(b_0)$ of $g$ for $\varepsilon = 0$ is discussed in detail.*

LEMMA 3.2 *Suppose that $G(\cdot)u(s)$ is finite at $f_0 \in D$ and $\partial_\varepsilon G(f_0)u(\cdot)$ is lower measurable on $S$. Then, if A^* is separable, for any $u \in V(S)$ and $s \in S$, there exists a Borel measurable function $f^* : S \longrightarrow A^*$ such that for all $f \in D$*

$$T(f)u(s) \geq T(f_0)u(s) + < f^*(s), f(s) - f_0(s) > -\varepsilon, \tag{3.6}$$

that is,

$$T(f)u(s) - < f^*(s), f(s) > \geq T(f_0)u(s) - < f^*(s), f_0(s) > -\varepsilon, \tag{3.7}$$

Proof. From (A1), (A2), (A3), and conditions of the lemma, for each $s \in S$, the extended function $G(\cdot)u(s) : A \longrightarrow R \cup \{\infty\}$, is convex, and l.s.c. at $f_0(s) \in F(s) \subset A$. Then, since $G(f_0)u(s) < \infty$, it follows from Lemma 3.1 that $\partial_\varepsilon G(f_0)u(s)$ is a nonempty, convex and weak*-closed subset of A^* for each $s \in S$. So, we get for all $a^* \in \partial_\varepsilon G(f_0)u(s)$

$$T(f)u(s) \geq T(f_0)u(s) + < a^*, f(s) - f_0(s) > -\varepsilon \text{ for all } f \in D \tag{3.8}$$

Then, A^* is a Polish space since it is assumed that A^* is separable. Further, we assume that $\partial_\varepsilon G(f_0)u(s)$ is lower measurable on S. In view of a selection theorem in [8], there is a Borel measurable selection $f^*(s) \in \partial_\varepsilon G(f_0)u(s)$. Thus, we constract a Borel measurable function $f^* : S \rightarrow A^*$ satisfying (3.6) for the given policy $f_0 \in D$. Thus, the proof of the lemma is completed.

Then, in order to show that a given stationary policy, f_0, is an ε−optimal one for a dynamic decision model with the loss functions modified by using the measurable function $f^*_\cdot : S \rightarrow A^*$ as follows:

$$(S, A, F, q, r- < f^*_\cdot, \cdot >, \beta), \tag{3.9}$$

we introduce, for each decision function $f \in D$, a modified operator $T(f^*, f)$ of $T(f)$ on $V(S)$ as follows: for each $u \in V(S)$ and $s \in S$

$$T(f^*, f)u(s) = T(f)u(s) - < f^*(s), f(s) >, \tag{3.10}$$

that is,

$$T(f^*, f)u(s) = r(s, f(s)) - < f^*(s), f(s) > +\beta \int_S u(y)q(dy|s, f(s)). \tag{3.11}$$

Remark 3.4 *Since $G(\cdot)u(s)$ is convex, and l.s.c., $G(\cdot)u(s)$ is locally Lipschitzian at $f_0(s) \in int(domG(\cdot)u(s))$. So, there exists $M(s) > 0$ and a neighborhood U of $f_0(s)$ such that*

$$|G(f)u(s) - G(f_0)u(s)| \leq M(s) \parallel f(s) - f_0(s) \parallel, \tag{3.12}$$

whenever $f(s) \in U$, where $\parallel \cdot \parallel$ denotes the metric on A. Since $f_0(s) \in U$ and $f^*(s) \in \partial_\varepsilon G(f_0)u(s)$, we have, for all $f(s) \in U$

$$< f^*(s), f(s) - f_0(s) > \leq M(s) \parallel f(s) - f_0(s) \parallel + \varepsilon, \tag{3.13}$$

which shows that $\parallel f^*(s) \parallel_* \leq M(s) + \varepsilon$, where $\parallel \cdot \parallel_*$ denotes the norm on A^*.

Thus, using Lemma 3.2, we can prove the main theorem.

THEOREM 3.1 *Let everything be as in Lemma* 3.2 *and suppose that, for a specified stationary policy* f_0 *with* $\sup_{s\in S} \| f_0(s) \| = \| f_0 \| < \infty$, *there exists a sequence of infinite Borel measurable functions* $\pi^* = (f_1^*, f_2^*, \cdots, f_t^*, \cdots), f_t^* : S \to A^*$, *and* $\sup_{s\in S} \| f_t^*(s) \|_* \leq M$, *that is, bounded on* S.

Then, we have for any policy $\pi \in \Pi$,

$$I^*(\pi^*, \pi)(x) \geq I^*(\pi^*, f_0)(x) - \varepsilon,$$

where, for a policy $\pi = (f_1, f_2, \cdots, f_t, \cdots)$,

$$\begin{aligned} I^*(\pi^*, \pi)(x) &= \sum_{t=1}^{\infty} \beta^{t-1} E_\pi[r(s_t, f_t(s_t)) - < f_t^*(s_t), f_t(s_t) > |s_1 = x] \\ &= \sum_{t=1}^{\infty} \beta^{t-1} \int_S [r(s, f_t(s)) - < f_t^*(s), f_t(s) >] p_t^\pi(ds|s_1 = x). \end{aligned}$$

Proof. For any policy $\pi = (f_1, f_2, \cdots, f_t, \cdots)$, the initial state $x \in S$ and the optimal value $u^*(x)$(see (3.3)), we define a notation as follows

$$V^k(\pi^*, \pi)(x) = T(f_1^*, f_1)T(f_2^*, f_2) \cdots T(f_k^*, f_k)u^*(x), \tag{3.14}$$

Since, from (A2), the loss function r is bounded, i.e., $|r| \leq$ N for some positive number N, on GrF, we have $|u^*(s)| \leq N/(1-\beta)$ for any state $s \in S$. Further, from the condition of the theorem, we have for each $s \in S$ and $t = 1, 2, \cdots$,

$$\sup_{s\in S} | < f_t^*(s), f_0(s) > | \leq \sup_{s\in S} \| f_t^*(s) \|_* \| f_0(s) \| \leq M \| f_0 \| . \tag{3.15}$$

So, it follows that for each $s \in S$

$$|I^*(\pi^{*(k+1)}, f_0)(s)| \leq \frac{N + M\|f_0\|}{1-\beta}, \tag{3.16}$$

where $\pi^{*(k+1)}$ denotes a sequence of infinite functions after the time $t = k+1$ in π^*, i.e., $\pi^{*(k+1)} = (f_{k+1}^*, f_{k+2}^*, f_{k+3}^*, \cdots)$. Thus, we need to show the result of the theorem for $I^*(\pi^*, \pi)(\cdot) \in V(S)$. Then, if $I^*(\pi^*, \pi)(x) = \infty$, the result of the theorem is obvious. So, it is sufficient to show that the result holds only when $I^*(\pi^*, \pi)(x) < \infty$. If $I^*(\pi^*, \pi)(x) < \infty$, for sufficiently small $\eta > 0$, from (3.4), it follows that there exists a sufficiently large integer $k > 0$ such that for any state $x \in S$,

$$|I^*(\pi^*, \pi)(x) - V^k(\pi^*, \pi)(x)| < \eta.$$

Thus, to prove the theorem, it is sufficient to show the result of the theorem for $V^k(\pi^*; \pi)(\cdot) \in V(S), \neq \infty$. From (3.14), $V^k(\pi^*, \pi)(x)$ is successively constructed by the modified operators $T(f^*, f)$. So, from Lemma 3.2 and the conditions of the theorem, it follows that, for f_0, there exists a function $f_k^* : S \to A^*$, such that, for each $s \in S$ and $f \in D$,

$$T(f_k^*, f)u^*(x) \geq T(f_k^*, f_0)u^*(x) - \varepsilon. \tag{3.17}$$

So, applying Lemma 3.2 to $T(f_k^*, f_k)u^*(\cdot) \in V(S), \neq \infty$ with the k th decision function $f_k$in π instead of f in (3.17), we obtain a function $f_{k-1} \in \partial_\varepsilon G(f_0)T(f_k^*, f_k)u^*(s)$ such that, for each $x \in S$ and $f \in D$

$$T(f_{k-1}^*, f)T(f_k^*, f_k)u^*(x) \geq T(f_{k-1}^*, f_0)T(f_k^*, f_k)u^*(x) - \varepsilon. \tag{3.18}$$

Then, since $T(\cdot, \cdot)$ is a monotone operator on $V(S)$, combining (3.17) with (3.18) and using the $(k - 1)$ th decision function f_{k-1}in π instead of f in (3.18), we obtain for the policy $\pi = (f_1, f_2, \cdots, f_t, \cdots)$,

$$\begin{aligned} T(f_{k-1}^*, f_{k-1})T(f_k^*, f_k)u^*(x) &\geq T(f_{k-1}^*, f_0)T(f_k^*, f_k)u^*(x) - \varepsilon \\ &\geq T(f_{k-1}^*, f_0)T(f_k^*, f_0)u^*(x) - \beta\varepsilon - \varepsilon. \end{aligned} \tag{3.19}$$

Further, applying Lemma 3.2 to (3.19) repeatedly, we arrive at

$$\begin{aligned} &T(f_1^*, f_1)T(f_2^*, f_2)\cdots T(f_{k-1}^*, f_{k-1})T(f_k^*, f_k)u^*(x) - \sum_{i=1}^{k-1} \beta^i \varepsilon \\ &\geq T(f_1^*, f_0)T(f_2^*, f_0)\cdots T(f_{k-1}^*, f_0)T(f_k^*, f_0)u^*(x) - \sum_{i=1}^{k} \beta^i \varepsilon. \end{aligned} \tag{3.20}$$

Thus, from (3.20), we get for sufficiently large k

$$V^k(\pi^*, \pi)(x) \geq V^k(\pi^*, f_0)(x) - \frac{\varepsilon}{1 - \beta}. \tag{3.21}$$

So, taking $\varepsilon(1 - \beta)^{-1}$ as ε and k as ∞ in (3.21), the proof of the theorem is completed.

THEOREM 3.2 *Let everything be as in Lemma 3.2 and assume that, for each $u \in M(S)$, the zero vector, θ^* belongs to $\partial_\varepsilon G(f_0)u(s)$ for all $s \in S$, that is, $\theta^* \in \partial_\varepsilon G(f_0)u(s)$ for all $u \in M(S)$ and $s \in S$.*

Then, for any policy π, we have

$$I(\pi)(x) \geq I(f_0)(x) - \varepsilon.$$

Proof. Since $\theta^* \in \partial_\varepsilon G(f_0)u(s)$ for all $u \in M(S)$ and $s \in S$, we can choose $f_t^* = \theta^*, t = 1, 2, \cdots$, as each function f_t^* of π^* in Theorem 3.1. So, for any policy π, $I^*(\pi^*, \pi)(x)$ is equal to $I(\pi)(x)$. Thus, the proof is completed.

参考文献

[1] R.Bellman (1970), *Dynamic programming and inverse optimal problems in Mathematical economics* , J. Math. Anal. Appl. 29, 424-428.

[2] R.Bellman, H.Kagiwada and R.Kalaba (1967),*Dynamic programming and an inverse problem in transport theory* , Computing, 2, 5-16.

[3] R.Bellman and R.Kalaba (1963), *An inverse problem in dynamic programming and automatic control* , J. Math. Anal. Appl. 7, 322-325.

[4] R.Bellman and J.M.Richardson (1961), *A note on an inverse problem in mathematical Physics* , Quart. Appl. Math. 19, 269-271.

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

[6] I.Ekeland (1979), *Nonconvex minimization problems*, Bull. Amer. Math. Soc. 3, No.1, 443-474.

[7] I.Ekeland (1974), *On the variational principle*, J. Math. Anal. Appl. 47, 324-353.

[8] K.Kuratowaski and Ryll-Nardzewski (1965), *A general theorem in selectors*, Bull. Acad. Polon. Sci. 13, 379-403.

[9] D.G.Luenberger (1968), Optimization by Vector Space Methods, John Wiley and Sons, Inc., New York.

[10] R.R.Phelps (1989), Convex Functions, Monotone Operators and Differentiability, *Lecture notes in Mathematics* 1364, Springer-Verlag, Berlin.

[11] J.-B.Hiriart-Urruty (1980), $\mathcal{E}$-subdifferential Calculus, Universitê de Clermont.

[12] 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.