R. Israel Ortega-Gutiérrez; Raúl Montes-de-Oca; Enrique Lemus-Rodríguez

Uniqueness of optimal policies as a generic property of discounted Markov decision processes: Ekeland's variational principle approach

Kybernetika, Vol. 52 (2016), No. 1, 66–75

Persistent URL: http://dml.cz/dmlcz/144863

KYBERNETIKA — VOLUME *52* (2016), NUMBER 1, PAGES 66–75

UNIQUENESS OF OPTIMAL POLICIES AS A GENERIC PROPERTY OF DISCOUNTED MARKOV DECISION PROCESSES: EKELAND'S VARIATIONAL PRINCIPLE APPROACH

R. Israel Ortega-Gutiérrez, Raúl Montes-de-Oca and Enrique Lemus-Rodríguez

Many examples in optimization, ranging from Linear Programming to Markov Decision Processes (MDPs), present more than one optimal solution. The study of this non-uniqueness is of great mathematical interest. In this paper the authors show that in a specific family of discounted MDPs, non-uniqueness is a "fragile" property through Ekeland's Principle for each problem with at least two optimal policies; a perturbed model is produced with a unique optimal policy. This result not only supersedes previous papers on the subject, but it also renews the interest in the corresponding questions of well-posedness, genericity and structural stability of MDPs.

Keywords: discounted Markov decision processes, dynamic programming, unique optimal policy, non-uniqueness of optimal policies, Ekeland's variational principle

Classification: 90C40, 93E20

1. INTRODUCTION

Existence and Uniqueness of solutions have always been main concerns of applied mathematics. Nevertheless, many examples in optimization display both existence and non-uniqueness. This non-uniqueness may be intrinsic to the system or may be *structurally non-stable*, i. e., a slightly perturbed version of the original system may present uniqueness. The authors of this paper are concerned with the question whether this is a *generic* property, i. e., if for the specific family of discounted Markov Decision Processes (MDPs) dealt with (for terminology and notation see [7]), it is always the case that a system presenting non-uniqueness can be perturbed into an arbitrarily close (in some sense to be specified later) system with a unique optimal policy. Previous results in this direction were obtained (see [10]) under convexity restrictions. Using Ekeland's celebrated Variational Principle (see [6], Section 2.1, p. 6 in [3], and Proposition 1.43, p. 31 in [11]), it is possible to obtain interesting new results without the aforementioned convexity restrictions that render article [10] obsolete.

DOI: 10.14736/kyb-2016-1-0066

The Ekeland's variational principle (or just "variational principle") is a theorem discovered by Ivar Ekeland which asserts that there exist nearly optimal solutions to some optimization problems. Roughly speaking, this principle states that, for any extended-value lower semicontinuous function which is bounded below, one can add a small perturbation to make it attain a minimum. Ekeland's variational principle can be used when the lower level set of a minimization problem is not compact, which means that the classical analysis theorems cannot be applied. Ekeland's principle relies on the completeness of the metric space and it also characterizes the completeness. Moreover, the variational principle provides a powerful tool in modern variational analysis. Its applications cover numerous areas including optimization, Banach space geometry, nonsmooth analysis, economics, control theory and game theory, to name a few (see [3] and [6]).

It is worth mentioning that the present result holds under fairly mild and general conditions (see Condition 2.2 below), satisfied by large families of MDPs (see Remark 2.5 below).

It should be stressed that the main result of this paper in fact is not a direct corollary of Ekeland's Variational Principle. The proof, instead, follows closely the pattern of Ekeland's original proof, that, as Ekeland himself asserts, is adapted from [2] (see also Remark 2.1.4, p. 9 in [3]). In particular, this step by step process is needed to establish the measurability of the optimal policy, which does not follow automatically from the classical result.

It is important to mention that Tanaka et al (in [12]) present a version of Ekeland Theorem for MDPs, related to ε-optimal policies without discussing uniqueness.

First the MDPs context is presented. Next, a result related to the existence of a unique maximal action is provided, then the main result and its proof are given. Finally, some remarks and comments are included.

2. DISCOUNTED MARKOV DECISION PROCESSES

First the preliminary Markov Decision Process context is briefly recalled (see [7]).

A Markov Control Model $(X, A, \{A(x) \mid x \in X\}, Q, c)$ is considered, where X is the State Space, A is the Action Set, Q denotes the Transition Law and c denotes the cost per state per action. A measurable non-empty set $A(x) \subset A$ whose elements are all admissible actions available when the system is in state x corresponds to each $x \in X$. Define $\mathbb{K} := \{(x, a) \mid x \in X, a \in A(x)\}$. Finally, c is a non-negative measurable function defined on $\mathbb{K}$. In this paper we will assume that (X, δ) and (A, φ) are both Borel spaces. Here, δ and φ denote the corresponding metrics in X and A, respectively.

Let Π be the set of all (deterministic and randomized, history-dependent) admissible policies. In particular, a *stationary policy* is defined as a measurable function $f : X \to A$ such that $f(x) \in A(x)$, for all $x \in X$. The set of all stationary policies will be denoted by $\mathbb{F}$. For each $\pi \in \Pi$ and an initial state $x \in X$, let

$$V(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right]$$

be the *total expected discounted cost* when the policy π is applied, given the initial state x. The constant $\alpha \in (0, 1)$ is the *discount factor* and it is fixed. The sequence of consecutive states and corresponding actions will be denoted by $\{x_t\}$ and $\{a_t\}$, and E_x^π denotes the expected value.

Definition 2.1. The policy π^* is *optimal* if $V(\pi^*, x) = v^*(x)$ for all $x \in X$, where $v^*(x) = \inf_{\pi \in \Pi} V(\pi, x)$, $x \in X$. v^* is called the *optimal value function*.

Condition 2.2. (a) c is lower semicontinuous (l.s.c.) and inf-compact (i. e., for each $x \in X$ and $r \in \mathbb{R}$, the set $\{a \in A(x) \mid c(x, a) \leq r\}$ is compact).

(b) The transition law Q is strongly continuous, i. e.,

$$\theta(x, a) = \int_X u(y) Q(\mathrm{d}y|x, a),$$

is continuous for $(x, a) \in \mathbb{K}$ and bounded below on $\mathbb{K}$, for each measurable and bounded function u on X.

(c) There exists a policy $\pi \in \Pi$ such that $V(\pi, x) < \infty$, for each $x \in X$.

Lemma 2.3. (Hernández-Lerma and Lasserre [7], Theorem 4.2.3) Suppose that Condition 2.2 is fulfilled. Then

(a) The optimal value function v^* is a solution of

$$v^*(x) = \min_{a \in A(x)} \left[c(x, a) + \alpha \int_X v^*(y) Q(\,\mathrm{d}y|\, x, a) \right] \tag{1}$$

for each $x \in X$ and, if μ is another solution of the equation, then $\mu(\cdot) \geq v^*(\cdot)$.

(b) There exists a selector $f^* \in \mathbb{F}$ such that in (1) the minimum is attained, i. e., for each $x \in X$,

$$v^*(x) = c(x, f^*(x)) + \alpha \int_X v^*(y) Q(\,\mathrm{d}y|\, x, f^*(x)) \tag{2}$$

and f^* is optimal.

Definition 2.4. Define for each $(x, a) \in \mathbb{K}$,

$$G(x, a) := c(x, a) + \alpha \int_X v^*(y) Q(\mathrm{d}y|x, a).$$

Remark 2.5. Examples that satisfy Condition 2.2 are finite models (i. e., MDPs for which both X and A are finite sets) with a non-negative cost function, as well as MDPs which satisfy Condition 2.2(c), and whose Markov control models are constituted by a suitable combination of conditions (a) and (b) below.

(a) For all $x \in X$, $A(x)$ is either a compact set or a closed set, or $\mathbb{R}$. And the transition law Q is induced either by a certain dynamics with additive-noise or a dynamics of Lindley's random walk type (see Examples 4.1, 4.5, and 4.8 in [4], Examples 4.2 and 4.15 in [5], and Examples 5.1 and 5.2 in [9]). It is important to note that MDPs with this kind of dynamics include many important specific models ranging from the linear model, controlled inventory systems, controlled dams systems, etc. (see also [1] and [7]).

(b) Condition 2.2 (a) holds, for c which is not necessarily strictly convex, but is non-negative.

Let $(X, A, \{A(x) \mid x \in X\}, Q, c)$ be a given Markov Control Model. Let's denote the Markov decision process for this Markov control model as M with corresponding optimal value function v^*. M will be referred to as the original process.

Throughout the paper, original processes M that satisfy Condition 2.2 are considered. Condition 2.2 for a process M will not be mentioned in each Lemma or Theorem in this paper, but it is supposed to hold. Moreover, given a model M, let f^* denote the corresponding optimal policy whose existence is ensured in Lemma 2.3.

Let ε be a positive number that will be assumed fixed in the rest of the paper. Now, let's define for the following MDP, denoted by M_ε (from the original process M): $(X, A, \{A(x) \mid x \in X\}, Q, c^*)$, where $c^*(x, a) = c(x, a) + \varepsilon\varphi(a, f^*(x))$, $x \in X$, $a \in A(x)$ and f^* appears in Lemma 2.3, where c is the original cost function from M. Note that both MDPs, M and M_ε, are equally excepted for the cost function; moreover, the set of $\mathbb{F}$ stationary policies is the same for both models (in fact, the set Π of all admissible policies is also the same for both models). M_ε will be referred to as the perturbed process.

For M_ε, let $W(\pi, x)$ be the total expected discounted cost when the policy π is applied, given the initial state x, and let w^* denote the corresponding optimal value function.

Remark 2.6. Theorem 4.2 will establish the existence of a **unique** optimal stationary policy for the system M_ε even if the original process M has non-unique optimal policies.

3. EXISTENCE OF A UNIQUE MAXIMAL STATIONARY POLICY

Define for each $x \in X$, an order relationship on the action set $A(x)$, in the following manner, for a_1, $a_2 \in A(x)$,

$$a_1 \prec a_2 \text{ if and only if } G(x, a_2) - G(x, a_1) + \varepsilon\varphi(a_1, a_2) \leq 0.$$

This clearly establishes a partial order, i. e., a reflexive, antisymmetric and transitive relationship.

For $x \in X$, a control $d \in A(x)$ is said to be *maximal* for $A(x)$ if $a \prec d$ for all $a \in A(x)$.

Lemma 3.1. For each $\eta \in \mathbb{F}$, there exists a unique stationary policy $g^* \in \mathbb{F}$ such that, for each $x \in X$, $\eta(x) \prec g^*(x)$ and $g^*(x)$ is maximal for $A(x)$.

P r o o f . Let $\eta \in \mathbb{F}$. Let a sequence of stationary policies $\{f_k\}_{k=1,2,\ldots}$ be defined inductively as follows.
Take $f_1 \equiv \eta$. Define, for each $x \in X$,

$$S_1(x) = \{a \in A(x) \mid f_1(x) \prec a\}. \tag{3}$$

Since $S_1(x) \neq \emptyset$, for each $x \in X$ (observe that $f_1(x) \in S_1(x)$), f_2 is chosen as it is shown below. By definition of $S_1(x)$, $x \in X$, it follows that for each $g \in S_1(x)$:

$$\begin{aligned} 0 \leq \varepsilon\varphi(g, f_1(x)) &\leq G(x, f_1(x)) - G(x, g) &(4)\\ &\leq G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x, a). &(5) \end{aligned}$$

Consider that $x \in X$ is arbitrary. Observe that Conditions 2.2 (a) and 2.2 (b) imply that $G(x,\cdot)$ is l.s.c. Hence, as the distance function is continuous, it follows that the set $S_1(x)$ is closed. Moreover, it is easy to prove that $S_1(x) \subseteq \{a \in A(x) \mid c(x,a) \leqslant G(x, f_1(x))\}$. Hence, $S_1(x)$ is a compact set (see Condition 2.2 (a)).

Note that, as $G(x,\cdot)$ is l.s.c. on $S_1(x) \subseteq A(x)$ and $S_1(x)$ is compact, for every $x \in X$, by Proposition D.5, p. 182 in [7], it follows that there exists $f_2 \in \mathbb{F}$ such that, for all $x \in X$, $f_2(x) \in S_1(x)$, and

$$\inf_{a \in S_1(x)} G(x,a) = G(x, f_2(x)).$$

Now, take $\frac{1}{2}\varepsilon\varphi(f_1(x), f_2(x)) \geqslant 0$, for each $x \in X$. Then, the following inequality evidently holds for each $x \in X$,

$$G(x, f_2(x)) \quad \leq \quad \inf_{a \in S_1(x)} G(x,a) + \frac{1}{2}\varepsilon\varphi(f_1(x), f_2(x)). \tag{6}$$

By (5) and (6) it is obtained that, for each $x \in X$,

$$\begin{aligned} G(x, f_2(x)) \quad &\leq \quad \inf_{a \in S_1(x)} G(x,a) + \frac{1}{2}\left[G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x,a)\right] \\ &= \quad \frac{1}{2}\left[G(x, f_1(x)) + \inf_{a \in S_1(x)} G(x,a)\right], \end{aligned}$$

hence, it results that, for each $x \in X$,

$$2G(x, f_2(x)) - G(x, f_1(x)) \quad \leq \quad \inf_{a \in S_1(x)} G(x,a). \tag{7}$$

So, take for each $x \in X$,

$$S_2(x) = \{a \in A(x) \,|f_2(x) \prec a\}. \tag{8}$$

Again, consider that $x \in X$ is arbitrary. Now, let $f_k \in \mathbb{F}$ be given for a positive integer $k > 2$. Suppose that for each $x \in X$, $f_k(x) \in S_{k-1}(x)$ and $f_k(x)$ minimizes $G(x,a)$ over all $a \in S_{k-1}(x)$. Define $S_k(x)$ as:

$$S_k(x) = \{a \in A(x) \,|f_k(x) \prec a\}. \tag{9}$$

Then $f_{k+1} \in \mathbb{F}$, satisfying that $f_{k+1}(x) \in S_k(x)$ for $x \in X$ is chosen in a way similar to the one in which $f_2 \in \mathbb{F}$ was obtained from f_1. Besides, the following inequality also holds for each $x \in X$:

$$2G(x, f_{k+1}(x)) - G(x, f_k(x)) \quad \leq \quad \inf_{a \in S_k(x)} G(x,a), \tag{10}$$

and

$$S_{k+1}(x) = \{a \in A(x) \,|f_{k+1}(x) \prec a\}. \tag{11}$$

With this procedure, the sequence $\{f_k\}_{k=1,2,\ldots}$ is constructed. Note that, for each k and $x \in X$, $S_k(x)$ is compact.

Now, once more consider that $x \in X$ is arbitrary. Observe that, for each k, $f_{k+1}(x) \in S_k(x)$, satisfies

$$G(x, f_{k+1}(x)) \leq \inf_{a \in S_k(x)} G(x,a) + \frac{1}{2}\left[G(x, f_k(x)) - \inf_{a \in S_k(x)} G(x,a)\right]. \tag{12}$$

In fact, the last inequality is a consequence of (10). Now, by construction, for each $k \geq 1$, $S_{k+1}(x) \subset S_k(x)$ (recall that $\prec$ is transitive), then

$$\inf_{a \in S_k(x)} G(x,a) \leq \inf_{a \in S_{k+1}(x)} G(x,a);$$

this implies that

$$G(x, f_{k+1}(x)) - \inf_{a \in S_{k+1}(x)} G(x,a) \leq G(x, f_{k+1}(x)) - \inf_{a \in S_k(x)} G(x,a), \tag{13}$$

for each $x \in X$.

From inequalities (12) and (13) (recall that $f_{k+1}(x) \in S_{k+1}(x)$, for each $x \in X$) , it results that, for each $x \in X$,

$$\begin{aligned}\left|G(x, f_{k+1}(x)) - \inf_{a \in S_{k+1}(x)} G(x,a)\right| &\leq \frac{1}{2}\left|G(x, f_k(x)) - \inf_{a \in S_k(x)} G(x,a)\right| \\ &\leq \frac{1}{2^k}\left|G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x,a)\right|.\end{aligned}$$

From the last inequality, it is obtained that, for each $x \in X$ and $l \in S_{k+1}(x)$ (observe that $f_{k+1}(x) \prec l$):

$$\begin{aligned}|G(x, f_{k+1}(x)) - G(x,l)| &= G(x, f_{k+1}(x)) - G(x,l) \\ &\leq G(x, f_{k+1}(x)) - \inf_{a \in S_{k+1}(x)} G(x,a) \\ &\leq \frac{1}{2^k}\left|G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x,a)\right|.\end{aligned}$$

This, from definition of $S_{k+1}(x)$, implies that, for each $x \in X$,

$$\varepsilon\varphi(f_{k+1}(x), l) \leq \frac{1}{2^k}\left|G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x,a)\right|.$$

So, using the triangular inequality, it follows that

$$\varepsilon\varphi(\beta, l) \leq \frac{2}{2^k}\left|G(x, f_1(x)) - \inf_{a \in S_1(x)} G(x,a)\right|,$$

for all $x \in X$, $\beta, l \in S_{k+1}(x)$, and $k \geq 1$.

Therefore, for each $x \in X$, $\mathrm{Diam}(S_k(x)) \to 0$ (where $\mathrm{Diam}(D)$ is the supremum of the distances between any pair of elements of D, for D is a subset of a metric space),

when $k \to \infty$; as, for each $x \in X$, $S_1(x)$ is complete, the sets $S_k(x)$ have for each $k \geq 1$ a unique common point, denote it by $g^*(x)$, that is,

$$\{g^*(x)\} = \bigcap_{k=1}^{\infty} S_k(x).$$

Now, by definition of $S_k(x) \subseteq A(x)$, it is also obtained that $f_k(x) \prec g^*(x)$ for all k and $x \in X$. In particular, for $k = 1$, it follows that $f_1(x) \prec g^*(x)$. Suppose that $g^*(x)$ is not maximal for $A(x)$. Then there exists $z(x) \in A(x)$ such that $g^*(x) \prec z(x)$. By transitivity, $f_k(x) \prec z(x)$ for any k. Consequently, $z(x) \in \{g^*(x)\}$, i. e., $z(x) = g^*(x)$. Therefore, $g^*(x)$ is in fact the unique maximal element for $A(x)$.

Since x is arbitrary, it is possible to define a function $g^* : X \to A$ such that $g^*(x) \in A(x)$ is maximal and unique, for each $x \in X$.

Now it will be established that g^* is a stationary policy. Fix $x \in X$. As $f_k(x) \prec f_{k+1}(x)$, then:

$$0 \leq \varepsilon\varphi(f_k(x), f_{k+1}(x)) \leq G(x, f_k(x)) - G(x, f_{k+1}(x)), \tag{14}$$

for $k = 1, 2, \ldots$. Adding, through $m > k$:

$$\begin{aligned} \varepsilon\varphi(f_k(x), f_m(x)) &\leq \varepsilon \sum_{j=k}^{m-1} \varphi(f_j(x), f_{j+1}(x)) && (15) \\ &\leq \sum_{j=k}^{m-1} \{G(x, f_j(x)) - G(x, f_{j+1}(x))\} \\ &= G(x, f_k(x)) - G(x, f_m(x)). && (16) \end{aligned}$$

By (14) it is obtained that $\{G(x, f_k(x))\}_{k=1,2,\ldots}$ is decreasing, and bounded below, and then $\{G(x, f_k(x))\}_{k=1,2,\ldots}$ converges. This establishes that the right term of equation (16) converges to 0 when $k, m \to \infty$. Consequently $\{f_k(x)\}_{k=1,2,\ldots}$ is a Cauchy sequence in $S_1(x) \subset A(x)$. The set $S_1(x)$ is complete, and hence the control sequence $\{f_k(x)\}_{k=1,2,\ldots}$ converges to $g^*(x)$ in $A(x)$. Hence, since x is arbitrary, it results that $\{f_k(x)\}_{k=1,2,\ldots}$ converges to $g^*(x)$ in $A(x)$, for each $x \in X$. As $f_k \in \mathbb{F}$, that is, each f_k is measurable, it results that g^* is also measurable, i. e., $g^* \in \mathbb{F}$. □

4. UNIQUENESS OF THE OPTIMAL POLICY IN THE PERTURBED PROCESS

Lemma 4.1. For an optimal policy f^*, given by Lemma 2.3, for the MDP M, there exists $g^* \in \mathbb{F}$ (given by Lemma 3.1 taking $\eta \equiv f^*$), such that, the following properties are satisfied:

(i) $\varphi(f^*(x), g^*(x)) = 0$, then $f^*(x) = g^*(x)$ and $G(x, g^*(x)) = G(x, f^*(x))$, for all $x \in X$;

(ii)

$$G(x, g^*(x)) < G(x, f(x)) + \varepsilon\varphi(f(x), g^*(x)), \tag{17}$$

for each $f \in \mathbb{F}$ with $f(x) \neq g^*(x)$, for $x \in X$. Moreover, g^* is the **unique** optimal policy for the perturbed process M_ε.

P r o o f. Take $\eta \equiv f^* \equiv f_1$ in Lemma 3.1, and let g^* be defined as in Lemma 3.1.

(i) Taking $k = 1$ in (16), it is obtained for each $m = 2, \ldots$, and $x \in X$, that

$$\varepsilon\varphi(f^*(x), f_m(x)) \leq G(x, f^*(x)) - G(x, f_m(x)). \tag{18}$$

As $G(x, \cdot)$ is l.s.c. in $A(x)$ for each $x \in X$, from (18), and as $f_m(x) \to g^*(x)$ for each $x \in X$, it follows that

$$\begin{aligned} G(x, g^*(x)) &\leq \liminf_{m\to\infty} G(x, f_m(x)) \\ &\leq \liminf_{m\to\infty} \{G(x, f^*(x)) - \varepsilon\varphi(f^*(x), f_m(x))\} \\ &= G(x, f^*(x)) - \varepsilon\varphi(f^*(x), g^*(x)), \end{aligned}$$

for each $x \in X$. So

$$G(x, g^*(x)) + \varepsilon\varphi(f^*(x), g^*(x)) \leq G(x, f^*(x))$$

and, since f^* is optimal for M,

$$\varepsilon\varphi(f^*(x), g^*(x)) \leq G(x, f^*(x)) - G(x, g^*(x)) \leq G(x, f^*(x)) - \inf_{a\in A(x)} G(x, a) = 0,$$

for each $x \in X$. Hence, $f^*(\cdot) = g^*(\cdot)$.

(ii) From (i) it results that for each $x \in X$, $W(g^*, x) = V(f^*, x) = v^*(x)$. And it is easy to verify that for each $\pi \in \Pi$ and $x \in X$, $W(\pi, x) \geq V(\pi, x) \geq v^*(x)$ (observe that $c^*(x, a) \geq c(x, a)$, for all $x \in X$ and $a \in A(x)$). Combining these equalities and inequalities, it follows that, for each $\pi \in \Pi$ and $x \in X$, $W(\pi, x) \geq W(g^*, x)$, hence $w^*(\cdot) = v^*(\cdot) = W(g^*, \cdot)$ and g^* is optimal for M_ε. Now, take $f \in \mathbb{F}$ and $x \in X$, such that $f(x) \neq g^*(x)$. Then, using that $\varphi(g^*(x), f(x)) > 0$ and the maximality of g^*, it results that

$$\begin{aligned} G(x, g^*(x)) &\leq G(x, f(x)) - \varepsilon\varphi(g^*(x), f(x)) \\ &\leq G(x, f(x)) \\ &< G(x, f(x)) + \varepsilon\varphi(g^*(x), f(x)), \end{aligned}$$

hence, inequality (17) follows. Finally, let h be a stationary policy such that h is optimal for M_ε and $h \neq g^*$. Take $x \in X$, such that $h(x) \neq g^*(x)$. From (17) and using that $w^*(\cdot) = v^*(\cdot) = W(h, \cdot)$ and $f^*(\cdot) = g^*(\cdot)$, it is obtained that

$$\begin{aligned} w^*(x) &= G(x, g^*(x)) \\ &< G(x, h(x)) + \varepsilon\varphi(g^*(x), h(x)) \\ &= c(x, h(x)) + \varepsilon\varphi(f^*(x), h(x)) + \alpha \int_X v^*(y) Q(\mathrm{d}y|x, h(x)) \\ &= c(x, h(x)) + \varepsilon\varphi(f^*(x), h(x)) + \alpha \int_X W(h, y) Q(\mathrm{d}y|x, h(x)). \end{aligned}$$

So,

$$\begin{aligned} w^*(x) &< c^*(x, h(x)) + \alpha \int_X W(h, y) Q(\mathrm{d}y|x, h(x)) \\ &= W(h, x), \end{aligned}$$

where the last equality follows from (4.2.15), p. 51 in [7]. Therefore, $w^*(x) < W(h, x)$, i. e., $h(x)$ is not optimal, which is a contradiction for the optimality of h. This ends the proof of part (ii) of Lemma 4.1.

□

Now all is set for the statement of the main result of this paper, that is clearly an immediate consequence of the two previous lemmas:

Theorem 4.2. For an MDP M, let f^* be an optimal policy (note that M does not necessarily have a unique optimal policy). Then there exists an MDP M_ε with a unique optimal policy which coincides with f^*.

5. FINAL REMARKS

The following remarks regarding the present paper are relevant:

1. This paper generalizes [10], extending its basic results for not-necessarily convex cost functions.

2. There is no clear way of obtaining the main result of this paper as a direct corollary of Ekeland's Variational Principle, but, surprisingly, its classical proof (see [6]) can be suitably adapted.

The main theorem of the present paper may be interpreted as a specific density property of the set of Markov Decision Models with unique optimal policies. Due to the fact that the very construction of the perturbed process uses explicitly an optimal policy of the original model, this approach has not been yet fully established if under some natural topology, the set of all Markov Decision Models with unique optimal policies (the class of MDPs with the UOPP —Unique Optimal Policy Property—) is open or of first category, and, hence, is *generic*. The authors are currently working on this open problem. For instance, it is of great interest to study the Well-Posedness (in Luccheti's sense — see [8]) of this family of Markov Decision Processes.

ACKNOWLEDGEMENT

This work was partially supported by CONACYT (México) and ASCR (Czech Republic) under Grant No. 171396.

(Received February 8, 2015)

REFERENCES

[1] D. P. Bertsekas: Dynamic Programming: Deterministic and Stochastic Models. Prentice-Hall, NJ 1987.

[2] E. Bishop and R. R. Phelps: The support functionals of a convex set. In: Proc. Sympos. Pure Math. Vol. VII, 1963 (V. L. Klee, ed.), Amer. Math. Soc., pp. 27–35. DOI:10.1090/pspum/007/0154092

[3] J. M. Borwein and Q. J. Zhu: Techniques of Variational Analysis. Springer, New York 2005.

[4] D. Cruz-Suárez, R. Montes-de-Oca, and F. Salem-Silva: Conditions for the uniqueness of optimal policies of discounted Markov decision processes. Math. Methods Oper. Res. *60* (2004), 415–436. DOI:10.1007/s001860400372

[5] D. Cruz-Suárez and R. Montes-de-Oca: Uniform convergence of the value iteration policies for discounted Markov decision processes. Bol. Soc. Mat. Mexicana *12* (2006), 133–152.

[6] I. Ekeland: On the variational principle. J. Math. Anal. Appl. *67* (1974), 324–353. DOI:10.1016/0022-247x(74)90025-0

[7] O. Hernández-Lerma and J. B. Lasserre: Discrete-Time Markov Control Processes: Basic Optimality Criteria. Springer-Verlag, New York 1996. DOI:10.1007/978-1-4612-0729-0

[8] R. Lucchetti: Convexity and Well-Posed Problems. CMS Books in Mathematics, Springer, New York 2006. DOI:10.1007/0-387-31082-7

[9] R. Montes-de-Oca and E. Lemus-Rodríguez: An unbounded Berge's minimum theorem with applications to discounted Markov decision processes. Kybernetika *48* (2012), 268–286.

[10] R. Montes-de-Oca, E. Lemus-Rodríguez, and F. Salem-Silva: Nonuniqueness versus uniqueness of optimal policies in convex discounted Markov decision processes. J. Appl. Math. *2013* (2013), 1–5.

[11] R. T. Rockafellar and R. J. B. Wets: Variational Analysis. Springer, New York 2004.

[12] K. Tanaka, M. Hosino, and D. Kuroiwa: On an ε-optimal policy of discrete time stochastic control processes. Bull. Inform. Cybernet. *27* (1995), 107–119.

R. Israel Ortega-Gutiérrez, Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Av. San Rafael Atlixco 186, Col. Vicentina, 09340 México, D.F. México.

e-mail: rei_israel@yahoo.com.mx

Raúl Montes-de-Oca, Departamento de Matemáticas, Universidad Autónoma Metropolitana-Iztapalapa, Av. San Rafael Atlixco 186, Col. Vicentina, 09340 México, D.F. México.

e-mail: momr@xanum.uam.mx

Enrique Lemus-Rodríguez, Escuela de Actuaría, Universidad Anáhuac México Norte, Huixquilucan, Estado de México. México.

e-mail: enrique.lemus@gmail.com