The Optimal Reward Operator

in

Negative Dynamic Programming

by

A. Maitra and W. Sudderth*
University of Minnesota

Technical Report No. 555
November 1990

* Research supported by National Science Foundation Grant DMS - 8911548.

Abstract

We consider the negative dynamic programming model of Strauch [10] and prove that the optimal reward function can be obtained by a transfinite iteration of the optimal reward operator. A departure from all previous treatments of this model is that we allow nonmeasurable policies. We prove that a player loses nothing by restricting himself to measurable policies, if the returns from nonmeasurable policies are evaluated by lower integrals.

AMS 1980 subject classification: Primary 90C39, 90C40, 93E20; Secondary 04A15, 28A05.

Key words and phrases: negative dynamic programming, stochastic control, inductive definability, analytic sets.

1. Introduction

The aim of this article is to study the optimal reward operator for the negative dynamic programming model introduced by Strauch [10]. Our paper parallels the work of Blackwell et al [2], who made a study of the optimal reward operator for the positive dynamic programming model. They proved that for an n-day horizon problem the optimal reward over measurable policies equals the optimal reward over all policies, measurable or nonmeasurable. Since, in the positive case, the optimal reward for the infinite horizon problem is the limit of the optimal rewards for the n-day problems, it follows that nonmeasurable policies do not give the player any advantage in the infinite horizon problem. In the negative case, however, the adequacy of measurable policies for finite horizon problems, which can be established simply by imitating the methods of [2], is of no help in establishing the analogous result for the infinite horizon problem. A new idea is needed. The idea is to obtain the infinite horizon optimal reward function by iterating the optimal reward operator a transfinite number of times. The measurability problems that arise in implementing this idea are handled by using a result from the theory of inductive definability. Methods from the theory of inductive definability have been used in the theory of gambling (see, for example, [3] and [5]), but, as far as we know, this is the first use of such methods in dynamic programming.

A negative dynamic programming problem is specified by the objects S, A, q, r. The state space S is a nonempty Borel subset of a Polish space. The set A is an analytic subset of $S \times X$, where X is a Polish space, with each vertical section A(s) of A nonempty. The set A(s) is to be regarded as the set of actions available at the state $s \in S$. The law of motion q is a Borel measurable transition function from A to S, that is, for each fixed $(s,a) \in A$, $q(\cdot|s,a)$ is a probability measure on the Borel subsets of S, and, for each fixed Borel subset B of S, $q(B|\cdot,\cdot)$ is a Borel measurable function on A. Finally, the daily return r is an upper analytic function on $A \times S$ into $[-\infty,0]$, that is, for every real number c, the set $\{r \geq c\}$ is an analytic subset of $A \times S$.

This is the way the game is played. When the system is in state s and we take action $a \in A(s)$, we move to a new state s' selected according to $q(\cdot|s,a)$ and receive a return of r(s,a,s'). The process is then repeated from the new state s'. We wish to maximize the total expected return over the infinite future.

A policy π is a sequence $\pi_1,\pi_2,\ldots$, where π_n is a function on $A \times A \times \cdots \times A \times S$ (n factors) to the set $\mathcal{P}(X)$ of probability measures on the Borel subsets of X such that for each partial history $h = (s_1,a_1,\ldots,s_{n-1},a_{n-1},s_n) \in A \times A \times \cdots \times A \times S$ (n factors),

$\pi_n(h)(A(s_n)) = \pi_n(A(s_n)|h) = 1$. A policy π is <u>measurable</u> if, for every $n \geq 1$, π_n is analytically measurable, that is, for every Borel subset M of $\mathcal{P}(X)$, $\pi_n^{-1}(M)$ belongs to the σ-field generated by the analytic subsets of $A \times A \times \ldots \times A \times S$ (n factors), where the Borel σ-field on $\mathcal{P}(X)$ is generated by the weak topology on $\mathcal{P}(X)$ (see [7]). A measurable policy π is <u>Markov</u> if, for every $n \geq 1$, there is an analytically measurable selector f_n: $S \rightarrow X$ for the set A such that for every partial history $h \in A \times A \times \ldots \times A \times S$, $\pi_n(\cdot|h) = \delta(f_n(s_n))$, where $\delta(a)$ is the degenerate probability measure concentrated on a. Recall that f: $S \rightarrow X$ is an analytically measurable selector for A if $(s,f(s)) \in A$ for every $s \in S$, and $f^{-1}(B)$ belongs to the σ-field generated by the analytic subsets of S for every Borel subset B of S.

If π is a measurable policy, then, for each initial state s, π, together with the law of motion q, induces a probability measure $e_\pi(\cdot|s)$ on the Borel subsets of $Z = X \times S \times X \times \ldots$ such that $e_\pi(A(s) \times A \times A \times \ldots|s) = 1$ ([1], Proposition 7.45). The total expected reward $I(\pi)(s_1)$ from π, when we start in s_1, is given by the formula

$$I(\pi)(s_1) = \int_{Z(s_1)} \sum_{n=1}^{\infty} r(s_n,a_n,s_{n+1})\, e_\pi(dz|s_1)$$

where $Z(s_1) = A(s_1) \times A \times A \times \ldots$ and $z = (a_1,s_2,a_2,\ldots) \in Z(s_1)$. Suppose, next, that π is a policy which is not necessarily measurable. We wish to define $I(\pi)$. Set

$$r_n(h) = \sum_{i=1}^{n} r(s_i,a_i,s_{i+1})$$

where the partial history $h = (s_1,a_1,\ldots,a_{n-1},s_n,a_n,s_{n+1})$. Let $\rho_n(\pi,s_1)$ be the 2n-fold iterated lower integral of $r_n(h)$ with respect to the 2n probability measures

$$q(ds_{n+1}|s_n,a_n)\, \pi_n(da_n|s_1,a_1,\ldots,s_n) \ldots q(ds_2|s_1,a_1)\, \pi_1(da_1|s_1).$$

Finally, set

$$I(\pi)(s_1) = \lim_n \rho_n(\pi,s_1).$$

Using the monotone convergence theorem, one checks easily that the two definitions of $I(\pi)$, when π is measurable, are in fact equivalent.

Blackwell et al [2] used the upper integral in defining the total expected reward from a nonmeasurable policy. In the negative dynamic programming problem, however, use of the upper integral turns out to be too generous an evaluation of the total expected reward and, then, nonmeasurable policies have a distinct advantage over measurable ones. This is shown by the following example, which is based on an unpublished example of S. Ramakrishnan in the theory of gambling.

Example. Let $S = \{-1\} \cup (0,1) \cup \{0,1,2,...\}$; $A(s) = \{0\}$, $s = -1,0,1,...$; $A(s) = \{1,2,...\}$, $s \in (0,1)$. Let $q(\cdot|-1,0) = \lambda$, where λ is Lebesgue measure on $(0,1)$; $q(\cdot|s,n) = \delta(n)$, $s \in (0,1)$, $n = 1,2,...$; $q(\cdot|n,0) = \delta(n-1)$, $n = 1,2,...$; $q(\cdot|0,0) = \delta(0)$. Let $r(1,0,0) = -1$ and let $r = 0$, elsewhere. It is straightforward to verify that $I(\pi)(-1) = -1$ for every measurable policy π.

Suppose now that in evaluating the expected reward from a nonmeasurable policy we had used a convex combination of the lower and upper integrals. To fix ideas, suppose that we had defined $\rho_n(\pi,s)$ as follows: For $n = 1$, let

$$\rho_1(\pi,s) = \alpha \int^* \int^* r(s,a,s')\, q(ds'|s,a)\, \pi_1(s)(da)$$

$$+ (1-\alpha) \int_* \int_* r(s,a,s')\, q(ds'|s,a)\, \pi_1(s)(da),$$

where α is fixed and $0 \le \alpha \le 1$. For $n > 1$, define ρ_n analogously.

Consider the following nonmeasurable policy for the example above. Fix a partition $\{B_n, n \ge 1\}$ of $(0,1)$ such that $\lambda_*(\bigcup_{i=1}^{n} B_i = 0)$ for every $n \ge 1$, λ_* being inner Lebesgue measure. The existence of such a partition is proved in [1,p.279]. Let π be the policy that chooses action n when the current state $s \in B_n$. Then

$$I(\pi)(-1) = \lim_{n\to\infty} \rho_n(\pi,-1)$$

$$= \lim_{n\to\infty} -(1-\alpha)\, \lambda^*\Big(\bigcup_{i=1}^{n} B_i\Big)$$

$$= -(1-\alpha).$$

Hence, for any choice of $\alpha > 0$, the policy π would fare strictly better at $s = -1$ than any measurable policy by the fixed amount α. Indeed, $\alpha = 0$ is the only choice for which nonmeasurable policies do not provide an advantage over measurable policies. This is true in general and is the import of Theorem 1.1.

Next, we define the optimal reward operator T. For any function $w: S \to [-\infty,0]$, define $Tw: S \to [-\infty,0]$ by

$$(Tw)(s) = \sup_{a \in A(s)} [\int_*(r(s,a,s') + w(s'))\, q(ds'|s,a)],$$

where $\int_*$ is the lower integral. Define by transfinite induction functions Q_ξ, $\xi < \omega_1$, on S into $[-\infty,0]$ as follows:

$$Q_0 = T\underset{\sim}{O}$$

and, for $\xi > 0$,

$$Q_\xi = T(\inf_{\eta<\xi} Q_\eta),$$

where $\underset{\sim}{O}$ is the function which is identically zero on S. Finally, set

$$Q = \inf_{\xi<\omega_1} Q_\xi.$$

We can now state the main results of the paper.

Theorem 1.1.

(a) Q is upper analytic, $TQ = Q$ and Q is the largest function w such that $Tw \geq w$.

(b) Q is the optimal reward function, that is,

$$Q(s) = \sup \{I(\pi)(s): \pi \text{ any policy}\}$$

for every $s \in S$.

(c) For every $\varepsilon > 0$, there is an ε-optimal Markov policy, that is, there is a Markov policy π such that $I(\pi) \geq Q - \varepsilon$.

As was mentioned earlier, the negative dynamic programming model was introduced by Strauch [10] and studied by him, by Bertsekas and Shreve in their

monograph [1] and by Schäl [8,9]. None of these authors considered nonmeasurable policies. Denote by v the optimal reward function over measurable policies. Strauch [10] proved that v is upper analytic and that $Tv = v$. Bertsekas and Shreve proved that v is the largest upper analytic function w from S to $[-\infty,0]$ such that $Tw \geq w$ ([1], Proposition 9.10) and also that, given $\varepsilon > 0$, there is a Markov policy π such that $I(\pi) \geq v - \varepsilon$ ([1], Proposition 9.19). The thrust of Schäl's work was in a different direction. He found sufficient conditions under which the functions $Q_{n-1} = T^n\underset{\sim}{Q}$ converge to v and also conditions ensuring the existence of a stationary optimal policy. Bertsekas and Shreve also dealt with these problems (see [1, section 9.6]).

Our paper is organized as follows. Section 2 is devoted to the statements of auxiliary results that will be used in this article. The optimal reward operator is studied in section 3. The proof of Theorem 1.1 is completed in section 4. In section 5, we give a sufficient condition which ensures that $Q_\xi = Q$ for some countable ordinal ξ and conclude with an example for which $Q_\xi \neq Q$ for every $\xi < \omega_1$.

2. Measure and set-theoretic preliminaries

In the sequel, a number of results from measure theory and set theory will be used. This section contains the statements of these results. In particular, it includes a brief exposition of that part of the theory of inductive definability, which we use in this paper.

Let $(\Omega,\mathcal{Q},\mu)$ be a probability space and let $w: \Omega \to [-\infty,0]$. We define the lower integral $\int_* w d\mu$ by the formula:

$$\int_* w d\mu = \sup \int u d\mu,$$

where the supremum is taken over all $\mathcal{Q}$-measurable functions $u: \Omega \to [-\infty,0]$ such that $u \leq w$.

(2.1) Suppose $w: \Omega \to [-\infty,0]$, $u: \Omega \to [-\infty,0]$ and assume u is measurable with respect to the completion of $\mathcal{Q}$ with respect to μ. Then

$$\int_*(u+w)\, d\mu = \int u d\mu + \int_* w d\mu$$

(2.2) Let $w: \Omega \to [-\infty,0]$, $G = \{(\omega,c) \in \Omega \times [-\infty,0]: w(\omega) < c\}$ and let λ be Lebesgue measure on $[-\infty,0]$. Then

$$\int_* w d\mu = -(\mu \times \lambda)^*(G),$$

where $(\mu \times \lambda)^*$ denotes the outer measure induced by the product measure $\mu \times \lambda$.

(2.3) Suppose $w, w_n: \Omega \to [-\infty, 0]$, $n \geq 1$, and suppose $w_n \downarrow w$. Then

$$\int_* w d\mu = \lim_{n \to \infty} \int_* w_n d\mu.$$

The proofs of (2.1) and (2.3) can be found in [1, Appendix A], while (2.2) is easily proved by using the results there. The Bertsekas-Shreve definition of the outer integral for nonpositive functions agrees with our definition of the lower integral above.

Let Y, Y' be analytic subsets of Polish spaces.

(2.4) Suppose $w: Y \to [-\infty, 0]$. Then w is upper analytic iff $\{(y,c) \in Y \times [-\infty, 0]: w(y) \geq c\}$ is an analytic subset of $Y \times [-\infty, 0]$.

(2.5) Suppose $g: Y' \to Y$ is Borel measurable and $w: Y \to [-\infty, 0]$ is upper analytic (universally measurable). Then $w \circ g$ is upper analytic (universally measurable).

(2.6) If $w_1, w_2: Y \to [-\infty, 0]$ are upper analytic, then so is $w_1 + w_2$.

(2.7) Suppose $w: Y \to [-\infty, 0]$ is upper analytic (universally measurable). Equip $\mathcal{P}(Y)$, the set of probability measures on the Borel σ-field of Y, with the usual weak topology. Then the function $\mu \to \int w d\mu$ is an upper analytic (universally measurable) function from $\mathcal{P}(Y)$ into $[-\infty, 0]$.

(2.8) Let Ω be a Borel subset of a Polish space. Suppose C is a coanalytic subset of $\Omega \times [-\infty, 0]$. If λ is Lebesgue measure on $[-\infty, 0]$, then the function $\omega \to -\lambda(C_\omega)$ is an upper analytic function from Ω to $[-\infty, 0]$.

The proofs of (2.4)-(2.8) may be found in [1,Chapter 7] or in [2].

Finally, we state the result from the theory of inductive definability. Let Y be an infinite set and Φ a mapping from the power-set of Y to the power-set of Y. Say that Φ is a monotone operator if, whenever $E_1 \subseteq E_2 \subseteq Y$, then $\Phi(E_1) \subseteq \Phi(E_2)$. Define the iterates of Φ by transfinite induction as follows:

$$\Phi^{\xi} = \Phi(\cup_{\eta<\xi}\Phi^{\eta}),$$

where ξ is any ordinal. So, in particular, $\Phi^0 = \Phi(\phi)$. It is easy to verify that Φ^{∞}, the least fixed point of Φ, is given by $\cup\{\Phi^{\eta}: \eta<\kappa\}$, where κ is the least cardinal greater than the cardinality of Y.

Suppose that Y is a Borel subset of a Polish space and Φ is a monotone operator on Y. We say Φ respects coanalytic sets if, whenever Ω is a Polish space and C is a coanalytic subset of $\Omega \times Y$, then the set $C^* = \{(\omega,y) \in \Omega \times Y: y \in \Phi(C_{\omega})\}$ is also coanalytic. (Here $C_{\omega} = \{y \in Y: (\omega,y) \in C\}$.)

(2.9) Let Φ be a monotone operator on a Borel subset Y of a Polish space and suppose that Φ respects coanalytic sets. Then
(a) Φ^{∞} is a coanalytic subset of Y,
(b) $\Phi^{\infty} = \cup_{\xi<\omega_1}\Phi^{\xi}$, where ω_1 is the first uncountable ordinal.

(2.9) is a special case of a very general result of a Moschovakis [6, 7C.8, p.414]. Zinsmeister [11] gives a nice exposition of Moschovakis's theorem.

3. The optimal reward operator

We are now in a position to establish the basic properties of the function Q, introduced in section 1.

Define an operator Φ on the power-set of $S \times [-\infty,0]$ to the power set of $S \times [-\infty,0]$ as follows:

$$(3.1)\quad \Phi(E) = \{(s,c) \in S \times [-\infty,0]: \sup_{a\in A(s)} [\int r(s,a,s')\, q(ds'|s,a) - (q(\cdot|s,a) \times \lambda)^*(E)] < c\}.$$

It is straightforward to verify that Φ is monotone. The next lemma shows that Φ respects coanalytic sets.

Lemma 3.1. Let Ω be a Polish space and let C be a coanalytic subset of $\Omega \times S \times [-\infty,0]$. Then

$$C^* = \{(\omega,s,c) \in \Omega \times S \times [-\infty,0]: (s,c) \in \Phi(C_\omega)\}$$

is a coanalytic subset of $\Omega \times S \times [-\infty,0]$. In particular, if E is a coanalytic subset of $S \times [-\infty,0]$, then so is $\Phi(E)$.

Proof. Observe that $(\omega,s,c) \in C^*$ if and only if

$$(3.2) \quad (c>-\infty) \;\&\; (\exists n\geq 1)(\forall a\in A(s))[\{\int r(s,a,s')q(ds'|s,a) - \int\lambda(C_{\omega,s'})q(ds'|s,a)\} \leq c-1/n].$$

For $(\omega,s,a) \in \Omega \times A$, set

$$\varphi_1(\omega,s,a) = \int r(s,a,s')\, q(ds'|s,a)$$

and

$$\varphi_2(\omega,s,a) = -\int\lambda(C_{\omega,s'})\, q(ds'|s,a).$$

It follows from (2.7), (2.8) and (2.5) that φ_1 and φ_2 are upper analytic functions on $\Omega \times A$. So, by (2.6), $\varphi_1 + \varphi_2$ is also upper analytic. Consequently, the subset of $\Omega \times S \times [-\infty,0]$ defined by the condition within [ ] in (3.2) is coanalytic. It now follows by using the well known closure properties of the point class of coanalytic sets that C^* is coanalytic.

Take $C = \Omega \times E$ to get the final assertion of the lemma.

If $w: S \to [-\infty,0]$, set $E(w) = \{(s,c) \in S \times [-\infty,0]: w(s) < c\}$. So if $w = \underset{\sim}{O}$, then $E(w) = \phi$.

Lemma 3.2. If $w: S \to [-\infty,0]$, then $E(Tw) = \Phi(E(w))$.

Proof. $E(Tw) = \{(s,c) \in S \times [-\infty,0]: (Tw)(s) < c\}$

$$= \{(s,c) \in S \times [-\infty,0]: \sup_{a\in A(s)} [\int_*(r(s,a,s') + w(s'))\, q(ds'|s,a)] < c\}$$

$$= \{(s,c) \in S \times [-\infty,0]: \sup_{a\in A(s)} [\int r(s,a,s')\, q(ds'|s,a) + \int_* w(s')\, q(ds'|s,a)] < c\}$$

$$= \{(s,c) \in S \times [-\infty,0]: \sup_{a \in A(s)} [\int r(s,a,s')\, q(ds'|s,a) - (q(\cdot|s,a) \times \lambda)^*(E(w))] < c\}$$

$$= \Phi(E(w)),$$

where the third equality is justified by using (2.1) and the fourth by using (2.2).

Lemma 3.3. (a) $\Phi^\xi = E(Q_\xi)$, $\xi < \omega_1$
(b) For every $\xi < \omega_1$, Q_ξ is upper analytic.

Proof. $\Phi^0 = \Phi(\phi)$

$$= \Phi(E(\underset{\sim}{Q}))$$

$$= E(T\underset{\sim}{Q})$$

$$= E(Q_0),$$

where the third equality is by virtue of Lemma 3.2. Suppose next that (a) is true for all $\eta < \xi$ and ξ is an ordinal greater than 0. Then

$$\Phi^\xi = \Phi(\cup_{\eta<\xi} \Phi^\eta)$$

$$= \Phi(\cup_{\eta<\xi} E(Q_\eta))$$

$$= \Phi(E(\inf_{\eta<\xi} Q_\eta))$$

$$= E(T(\inf_{\eta<\xi} Q_\eta))$$

$$= E(Q_\xi),$$

where the second equality uses the inductive hypothesis and the fourth is by virtue of Lemma 3.2. This establishes (a). To prove (b), use induction on ξ and Lemma 3.1 to see that Φ^ξ is a coanalytic subset of $S \times [-\infty,0]$ for every $\xi < \omega_1$. (b) now follows from (a) and (2.4).

Theorem 3.4. The function Q is upper analytic and the largest fixed point of the operator T. Indeed, if w: $S \to [-\infty,0]$ and $Tw \geq w$, then $Q \geq w$.

Proof. $\Phi^{\infty} = \cup_{\xi<\omega_1} \Phi^{\xi}$

$= \cup_{\xi<\omega_1} E(Q_{\xi})$

$= E(\inf_{\xi<\omega_1} Q_{\xi})$

$= E(Q)$,

where the first equality is by virtue of (2.9 (b)) and the second is by Lemma 3.3 (a). By (2.9) (a)), Φ^{∞} is coanalytic, so Q is upper analytic by (2.4). On the other hand, by Lemma 3.2,

$$\begin{aligned} E(TQ) &= \Phi(E(Q)) \\ &= \Phi(\Phi^{\infty}) \\ &= \Phi^{\infty}, \end{aligned}$$

so that $TQ = Q$. Finally, let w: $S \to [-\infty,0]$ and suppose $Tw \geq w$. Then, by Lemma 3.2,

$$\begin{aligned} E(w) &\supseteq E(Tw) \\ &= \Phi(E(w)). \end{aligned}$$

Since Φ is monotone, it follows by induction on ξ that $\Phi^{\xi} \subseteq E(w)$ for every $\xi < \omega_1$. So

$$\Phi^{\infty} = \cup_{\xi<\omega_1} \Phi^{\xi} \subseteq E(w).$$

Hence $E(Q) \subseteq E(w)$, so that $Q \geq w$. This completes the proof.

4. The optimal reward function

We will now identify the function Q with the optimal reward function for the negative dynamic programming problem. Set

(4.1) $\quad v^*(s) = \sup I(\pi)(s), \; s \in S,$

where the supremum is taken over all policies, measurable or nonmeasurable.

Theorem 4.1. For every $\xi < \omega_1$, $v^* \le Q_\xi$.

Proof. The proof is by induction on ξ. Consider first the case when $\xi = 0$. Let π be a policy, possibly nonmeasurable. Then, for any $s \in S$,

$$I(\pi)(s) \le \rho_1(\pi,s)$$

$$= \int_*\int_* r(s,a_1,s_2)\, q(ds_2|s_1,a_1)\, \pi_1(da_1|s)$$

$$\le \int_* (\underset{\sim}{TO})(s)\, \pi_1(da_1|s)$$

$$= (\underset{\sim}{TO})(s)$$

$$= Q_0(s).$$

This proves that $v^* \le Q_0$. Suppose now that $\xi > 0$ and that $v^* \le Q_\eta$ for every $\eta < \xi$. Let π be a policy, possibly nonmeasurable. For $(s,a) \in A$, define the policy $\pi^{s,a}$ as follows:

$$\pi_n^{s,a}(s_1,a_1,\ldots,s_{n-1},a_{n-1},s_n) = \pi_{n+1}(s,a,s_1,a_1,\ldots,s_{n-1},a_{n-1},s_n)$$

for $n=1,2,\ldots$. It is easy to verify that

$$\rho_{n+1}(\pi,s) = \int_*\int_* \{r(s,a,s_2) + \rho_n(\pi^{s,a},s_2)\}\, q(ds_2|s,a)\, \pi_1(da|s)$$

for $n=1,2,\ldots,s\in S$. Hence, for any $s \in S$,

$$I(\pi)(s) = \lim_{n\to\infty} \rho_{n+1}(\pi,s)$$

$$= \lim_{n\to\infty} \int_*\int_* \{r(s,a,s_2) + \rho_n(\pi^{s,a},s_2)\}\, q(ds_2|s,a)\, \pi_1(da|s)$$

$$= \int_* \int_* \{r(s,a,s_2) + \lim_{n\to\infty} \rho_n(\pi^{s,a},s_2)\}\ q(ds_2|s,a)\ \pi_1(da|s)$$

$$= \int_* \int_* \{r(s,a,s_2) + I(\pi^{s,a})(s_2)\}\ q(ds_2|s,a)\ \pi_1(da|s)$$

$$\leq \int_* \int_* \{r(s,a,s_2) + v^*(s_2)\}\ q(ds_2|s,a)\ \pi_1(da|s)$$

$$\leq \int_* \int_* \{r(s,a,s_2) + (\inf_{\eta<\xi} Q_\eta)(s_2)\}\ q(ds_2|s,a)\ \pi_1(da|s)$$

$$\leq \int_* (T(\inf_{\eta<\xi} Q_\eta))(s)\ \pi_1(da|s)$$

$$= (T(\inf_{\eta<\xi} Q_\eta))(s)$$

$$= Q_\xi(s),$$

where the third equality uses (2.3) and the second inequality is by virtue of the inductive hypothesis. Hence, $v^* \leq Q_\xi$ and the proof is complete.

The next corollary is an immediate consequence of Theorem 4.1.

Corollary 4.2. $v^* \leq Q$.

We will now prove the reverse inequality. Indeed, we will establish an apparently stronger result. As in the Introduction, let

(4.2) $$v(s) = \sup I(\pi)(s),\ s \in S,$$

where the supremum is taken over measurable policies. For the next result, we need to explain some more notation.

With each analytically measurable selector f of the set A, associate an operator L(f) as follows:

(4.3) $$(L(f)w)(s) = \int \{r(s,f(s),s') + w(s')\}\ q(ds'|s,a),$$

where w: $S \to [-\infty,0]$ is universally measurable. It follows from (2.5) and (2.7) that L(f)w is a universally measurable function on S into $[-\infty,0]$.

Suppose that $\pi = (f_1,f_2,\ldots)$ is a Markov policy. It is easy to prove by induction on n that for any $s \in S$

$$\rho_n(\pi,s) = (L(f_1)L(f_2)\cdots L(f_n)\,\underset{\sim}{O})(s)$$

so that

$$I(\pi)(s) = \lim_{n\to\infty} (L(f_1)L(f_2)\cdots L(f_n)\,\underset{\sim}{O})(s), \tag{4.4}$$

by the monotone convergence theorem.

Theorem 4.3. For every $\varepsilon > 0$, there is a Markov policy π such that $I(\pi)(s) \geq Q(s) - \varepsilon$ for every $s \in S$. Consequently, $v \geq Q$.

Proof. Fix $\varepsilon > 0$. By Theorem 3.4, $TQ = Q$. Hence, by a selection theorem ([2, p.936] or [1, Proposition 7.50]), for every $n \geq 1$, we can find an analytically measurable selector f_n: $S \to X$ for the set A such that

$$(L(f_n)Q)(s) \geq Q(s) - \frac{\varepsilon}{2^n} \tag{4.5}$$

for every $s \in S$. Now, using the fact that the operator L(f) of (4.3) is monotone and iterating (4.5), we get:

$$(L(f_1)L(f_2)\ldots L(f_n)Q)(s) \geq Q(s) - \left(\frac{\varepsilon}{2} + \frac{\varepsilon}{2^2} + \cdots + \frac{\varepsilon}{2^n}\right)$$

$$\geq Q(s) - \varepsilon$$

for every $s \in S$ and $n \geq 1$. But, since $Q \leq \underset{\sim}{O}$, we have for every $n \geq 1$ and $s \in S$,

$$(L(f_1)L(f_2)\cdots L(f_n)\underset{\sim}{O})(s) \geq (L(f_1)L(f_2)\cdots L(f_n)Q)(s)$$

and, consequently,

$$(L(f_1)L(f_2)...L(f_n)\underset{\sim}{O})(s) \geq Q(s) - \varepsilon.$$

Now let $n \to \infty$ and use (4.4) to see that

$$I(\pi)(s) \geq Q(s) - \varepsilon$$

for every $s \in S$. This completes the proof.

Theorem 1.1 now follows from Theorem 3.4, Corollary 4.2 and Theorem 4.3.

5. The absolutely continuous case

According to Theorem 1.1, the optimal reward function Q can be obtained by iterating the operator T ω_1 times. The question arises if this process of iteration terminates at some countable stage, that is, if there is a countable ordinal ξ such that $Q_\xi = Q$. An example will be given at the end of this section to show that, in general, it is possible that $Q_\xi \neq Q$ for every countable ordinal ξ. The next result gives a sufficient condition for the existence of a countable ordinal ξ such that $Q_\xi = Q$.

Theorem 5.1. Assume that there is a probability measure μ on the Borel σ-field of S such that $q(\cdot|s,a)$ is absolutely continuous with respect to μ for every $(s,a) \in A$. Then there is an ordinal $\xi < \omega_1$ such that $Q = Q_\xi$.

Proof. Since $Q_\xi \leq Q_\eta$ for $\eta < \xi < \omega_1$, it follows that $E(Q_\eta) \subseteq E(Q_\xi)$. Set $E_\xi = E(Q_{\xi+1}) - E(Q_\xi)$, $\xi<\omega_1$. Then the sets E_ξ are disjoint and $(\mu\times\lambda)$-measurable, where λ is Lebesgue measurable on $[-\infty,0]$. Since $\mu\times\lambda$ is a σ-finite measure, it follows from the countable chain condition that only countably many E_ξ's can have positive $(\mu\times\lambda)$-measure. So there is $\xi_0<\omega_1$ such that $(\mu\times\lambda)(E_{\xi_0}) = 0$. It follows by Fubini's theorem that $Q_{\xi_0} = Q_{\xi_0+1}$ a.s. (μ). Hence, for each $(s,a) \in A$,

$$\int\{r(s,a,s') + Q_{\xi_0}(s')\}\ q(ds'|s,a)$$

$$= \int\{r(s,a,s') + Q_{\xi_0+1}(s')\}\ q(ds'|s,a),$$

since $q(\cdot|s,a)$ is absolutely continuous with respect to μ. So, for any $s \in S$,

$$Q_{\xi_0+1}(s) = \sup_{a \in A(s)} [\int \{r(s,a,s') + Q_{\xi_0}(s')\}\ q(ds'|s,a)]$$

$$= \sup_{a \in A(s)} [\int \{r(s,a,s') + Q_{\xi_0+1}(s')\ q(ds'|s,a)]$$

$$= (TQ_{\xi_0+1})(s).$$

It follows that $Q = Q_{\xi_0+1}$. This completes the proof.

It should be apparent from the proof above that in the absolutely continuous case Theorem 1.1 can be established without recourse to the methods of inductive definability.

Strauch [10, p.880] gives an example of a negative dynamic programming problem with countable state and action spaces - and therefore satisfying the conditions of Theorem 5.1 - such that $Q_\omega = Q$, but $Q_n \neq Q$ for every natural number n. Without undue effort, one can generalize Strauch's example to establish the following: For every $\xi < \omega_1$, there is a negative dynamic programming problem with countable state and action spaces such that $Q_\xi = Q$, but $Q_\eta \neq Q$ for every $\eta < \xi$. In other words, even for countable problems, there is no uniform bound on the number of times T has to be iterated to obtain Q.

Finally, we give an example to show that it is possible that $Q_\xi \neq Q$ for every $\xi < \omega_1$. The example is due to Blackwell ([10], p.881). First, we need to explain some concepts and results from classical descriptive set theory.

Let $\mathcal{R}$ be the set of rationals in (0,1) with its usual ordering. Fix a system $\{W_r, r \in \mathcal{R}\}$ of Borel subsets of [0,1]. Such a system is called a Borel sieve. For each $x \in [0,1]$, let

$$M_x = \{r \in Q: x \in W_r\},$$

$$D = \{x \in [0,1]: M_x \text{ is not well-ordered}\}$$

and, for each countable ordinal ξ,

$$C_\xi = \{x \in [0,1]: \text{the order-type of } M_x = \xi\}.$$

The set D is called the set sifted through the sieve $\{W_r\}$ and the sets C_ξ are called the constituents determined by the sieve $\{W_r\}$. Here are the facts that we will need.

(a) For any Borel sieve $\{W_r\}$, the set D is analytic ([4], p.465). Conversely, every analytic subset of [0,1] is of the form D for some Borel sieve $\{W_r\}$ ([4], p.483).
(b) For any Borel sieve $\{W_r\}$, the constituents C_ξ are Borel subsets of [0,1] and $[0,1] - D = \cup_{\xi<\omega_1} C_\xi$ ([4], p.500).

Now fix a Borel sieve $\{W_r, r \in \mathcal{R}\}$ such that the sifted set D is not Borel. Fix also an enumeration $r_1, r_2, \ldots$ of the elements of $\mathcal{R}$. Consider now the negative dynamic programming problem defined as follows:

Let $S = [[0,1] \times (\mathcal{R} \cup \{0,1\})] \cup \{t\}$, where $t \notin [0,1] \times (\mathcal{R} \cup \{0,1\})$, and let $A(s) = \{1,2,3,\ldots\}$ for every $s \in S$. The law of motion is given by

$$q(\cdot|(x,r),a) = \delta(x,r_a) \quad \text{if } r_a < r \ \& \ x \in W_{r_a}$$
$$= \delta(t) \qquad \text{otherwise,}$$
$$q(\cdot|t,a) = \delta(t)$$

and the return function is given by

$$r(s,a,t) = -1 \quad \text{if } s \neq t$$
$$= 0 \quad \text{otherwise.}$$

It was proved in [10] that

$$Q(x,1) = 0 \quad \text{if } x \in D$$
$$= -1 \quad \text{if } x \in [0,1] - D.$$

On the other hand, it is not hard to verify that

$$Q_\xi(x,1) = 0 \ \text{ if } \ x \in [0,1] - \cup_{\eta\leq\xi} C_\eta$$
$$= -1 \text{ if } x \in \cup_{\eta\leq\xi} C_\eta.$$

Since the set D is not Borel, it follows from (b) that $\cup_{\eta\leq\xi} C_\eta \neq [0,1] - D$ for every $\xi < \omega_1$. So $Q_\xi \neq Q$ for every $\xi < \omega_1$.

References

1. D.P. Bertsekas & S.E. Shreve, Stochastic Optimal Control: The Discrete Time Case, Academic Press, New York (1978).

2. D. Blackwell, D. Freedman & M. Orkin, The optimal reward operator in dynamic programming, Ann. Probab. 2 (1974), 926-941.

3. L. Dubins, A. Maitra, R. Purves and W. Sudderth, Measurable, nonleavable gambling problems, Israel J. Math. 67 (1989), 257-271.

4. K. Kuratowski, Topology I, Academic Press, New York (1966).

5. A. Maitra, R. Purves & W. Sudderth, Leavable gambling problems with unbounded utilities, Trans. Amer. Math. Soc. 320 (1990), 543-567.

6. Y.N. Moschovakis, Descriptive Set Theory, North-Holland, Amsterdam (1980).

7. K.R. Parthasarathy, Probability Measures on Metric Spaces, Academic Press, New York (1980).

8. M. Schäl, Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 32 (1975), 179-196.

9. M. Schäl, An operator-theoretical treatment of negative dynamic programming, in Dynamic Programming and its Applications (ed. M.L. Puterman), Academic Press, New York (1978), 351-368.

10. R.E. Strauch, Negative dynamic programming, Ann. Math. Statist. 37 (1966), 871-890.

11. M. Zinsmeister, Les derivations analytiques, Seminaire de Probabilites XXIII, Lecture Notes in Math. 1372, Springer-Verlag, Berlin and New York (1989), 21-46.