Statistics, Probability and Game Theory
IMS Lecture Notes – Monograph Series (1996) Volume 30

ON MEASURABILITY AND REPRESENTATION OF STRATEGIC MEASURES IN MARKOV DECISION PROCESSES

EUGENE A. FEINBERG
State University of New York at Stony Brook

Abstract.

This paper deals with a discrete time Markov Decision Process with Borel state and action spaces. We show that the set of all strategic measures generated by randomized stationary policies is Borel. Combined with known results, this fact implies measurability of the sets of strategic measures generated by stationary, Markov, and randomized Markov policies. We consider applications of these measurability results to two groups of problems: (i) measurability of value functions for various classes of policies and (ii) integral representation of strategic measures for randomized Markov and arbitrary randomized policies through strategic measures for corresponding nonrandomized policies.

1. Introduction. The foundations of dynamic programming for problems with uncountable state spaces were built by David Blackwell (1965, 1965a) and his student Ralph Strauch (1966). For the past thirty years, these pioneering results have been developed in various directions such as: (i) problems with more general measurability conditions (Blackwell, Orkin, Freedman 1974, Freedman 1974, Bertsekas and Shreve 1978, Schäl and Sudderth 1987, (ii) problems with more general summation assumptions than positive, negative, and discounted dynamic programming problems (Hinderer 1970, Dynkin and Yushkevich 1979, Schäl 1983, Feinberg 1982, 1982a, 1992, Schäl and Sudderth 1987), (iii) dynamic programming on compact sets (Schäl 1975, Balder 1989). The previous sentence represents just a small part of research directions and publications stimulated by research of David Blackwell on dynamic programming.

One of the remarkable discoveries in these pioneering papers by Blackwell (1965, 1965a) and Strauch (1966) was that value functions may not be measurable in a standard Borel sense, but they are measurable in a more general sense, namely they are universally measurable. More precisely, if the objective is to maximize the expected total rewards, the value function is upper semianalytic and therefore it is universally measurable. This discovery established connections between dynamic programming and the theory of analytic sets (Lusin 1927, Kuratowski 1966), an area of pure mathematics developed in the first part of twentieth century, and stimulated additional research in the fields of topology, set theory, and analysis, including new

developments related to selection theorems; Wagner (1977). It also allowed Blackwell, Strauch, and future researchers to write optimality operators and optimality equations for dynamic programming problems with uncountable state spaces and to analyze these equations.

Dynamic programming models are a particular case of Markov Decision Processes (MDPs) when the criterion is an expected total reward; see Puterman (1994) and references therein for various models of MDPs and related problems. All natural criteria, including expected total rewards, expected rewards per unit time, the Dubins-Savage criterion, and their measurable combinations, belong to the class of measurable criteria introduced in Feinberg (1982a). The measurability property of value functions described in the previous paragraph holds for any measurable criterion in an MDP with Borel state and action spaces; Feinberg (1982a). In fact, it follows directly from the Borel measurability of the set of all strategic measures; Dynkin and Yushkevich (1979), sections 3.5, 3.6, and 5.5. We recall that any initial distribution and any policy define a probability measure on the set of trajectories. This measure is called strategic.

Dubins and Savage (1965) introduced gambling models which are close relatives of MDPs; see the papers by Blackwell (1976) and Schäl (1989) on the relationship between gambling models and MDPs. The first fundamental contributions to the theory of Borel gambling models were by Strauch (1967) and Sudderth (1969). In particular, Sudderth (1969, Theorem 1 on p. 403) proved the measurability of the set of strategic measures for Borel gambling problems. Theorems 1, 2, and 4(b) in Blackwell (1976) imply the similar result for MDPs where it means the measurability of the set of all strategic measures generated by nonrandomized policies.

In this paper, we show that the set of all strategic measures generated by randomized stationary policies is measurable. This result and measurability of the sets of all strategic measures (Dynkin and Yushkevich 1979) and of all strategic measures generated by nonrandomized policies (Sudderth 1969, Blackwell 1976) imply measurability of the sets of strategic measures generated by the following classes of policies: (nonrandomized) stationary, Markov, and randomized Markov. These results imply that value functions of these classes of policies are upper semianalytic and therefore they are universally measurable. These results also allow us to write representation of randomized Markov and general randomized policies through nonrandomized ones in a simpler form than has been known before; see Gikhman and Skorokhod (1979), Feinberg (1982, 1982a), and Kadelka (1983).

This paper is organized in the following way. Section 2 introduces major definitions. Section 3 describes the measurability of various sets of strategic measures. Sections 4 and 5 deal with two different applications of the results of Section 3. We describe the results related to measurability of value functions in Section 4. Section 5 deals with the representation of randomized strategic measures through nonrandomized ones.

2. Definitions. We consider a standard discrete time MDP with Borel state and action spaces with the following elements:

(i) The *state space* X is a standard Borel space (i.e. a nonempty Borel subset of some Polish space endowed with the σ–field of Borel subsets of X).

(ii) The *action space* A is also a standard Borel space.

(iii) The mapping D which assigns to each $x \in X$ the set of *available actions* $D(x)$ which is a nonempty measurable subset of A. It is assumed that the set

$$graph\ D = \{(x, a); x \in X, a \in D(x)\}$$

is a measurable subset of $X \times A$ and contains the graph of a measurable map of X into A. (Throughout the paper "measurable" means "Borel measurable").

(iv) The *law of motion* (or *transition probabilities*) p which is a measurable stochastic kernel on X given $X \times A$; that is $p(\cdot|x, a)$ is a probability measure on the σ–field of Borel subsets of X and $p(B|\cdot, \cdot)$ is a measurable function on $X \times A$ for each $B \subseteq X$.

(v) The reward criterion w which will be defined later.

The *history spaces* are defined as $H_n = X \times (X \times A)^n$, $n = 0, 1, 2, \ldots, \infty$. On each set H_n a Borel σ-field generated as a product of Borel σ–fields on X and A is denoted. As usual, a general (randomized) policy $\pi = \{\pi_n\}$ is defined as a sequence of transition probabilities from H_n to A such that $\pi_n(D(x_n)|x_0a_0 \ldots x_n) = 1$ for each $x_0a_0 \ldots x_n \in H_n$, $n = 0, 1, 2, \ldots$. A nonrandomized policy $\phi = \{\phi_n\}$ is defined as a sequence of measurable functions from H_n to A such that $\phi_n(x_0a_0 \ldots x_n) \in D(x_n)$ for each $x_0a_0 \ldots x_n \in H_n$, $n = 0, 1, 2, \ldots$. A randomized Markov policy is defined as a sequence of transition probabilities from X to A such that $\pi_n(D(x)|x) = 1$ for each $x \in X$, $n = 0, 1, 2, \ldots$. A Markov policy is defined as a sequence $\phi = \{\phi_n\}$ of measurable functions from X to A such that $\phi_n(x) \in D(x)$ for each $x \in X$, $n = 0, 1, 2, \ldots$. If decisions depend just on current states, a policy is called *randomized stationary*. A randomized stationary policy is defined by a transition probability π from X to A such that $\pi(D(x)|x) = 1$, $x \in X$. And a stationary policy is a measurable function ϕ from X to A such that $\phi(x) \in D(x)$.

We denote by $R\Pi$, Π, RM, M, RS, and by S the set of all, nonrandomized, randomized Markov, Markov, randomized stationary, and stationary policies respectively. Obviously, $RS \subseteq RM \subseteq R\Pi$, $S \subseteq M \subseteq \Pi$, and $F \subseteq RF$, where $F = S, M$, or Π.

For any standard Borel space E we use the following notations:

$\mathcal{B}(E)$ — the σ–field of Borel subsets of E;

$\mathcal{P}(E)$ — the set of probability measures on $(E, \mathcal{B}(E))$;

$\mathcal{M}(E)$ — the minimal σ–field on $\mathcal{P}(E)$ such that for any $A' \in \mathcal{B}(E)$ the function $\mu \to \mu(A')$ is measurable. If E is a standard Borel space then

$(\mathcal{P}(E), \mathcal{M}(E))$ is a standard Borel space too; f.i. see Dynkin and Yushkevich (1979), Appendix 5.

By the Ionescu Tulcea theorem (Neveu 1965), each policy π and initial distribution $\mu \in \mathcal{P}(X)$ define a probability measure $\mathbf{P}_\mu^\pi$ on the space $(H_\infty, \mathcal{B}(H_\infty))$,

$$\mathbf{P}_\mu^\pi(dx_0 da_0 dx_1 \dots dx_n da_n \dots) = \mu(dx_0) \prod_{i=0}^{\infty} (\pi_i(da_i|x_0 a_0 x_1 a_1 \dots x_i) p(dx_{i+1}|x_i, a_i)). \tag{2.1}$$

We denote by $\mathbf{E}_\mu^\pi$ the mathematical expectation with respect to this measure. If $\mu(x) = 1$ for some $x \in X$ we will write $\mathbf{P}_x^\pi$ and $\mathbf{E}_x^\pi$ instead of $\mathbf{P}_\mu^\pi$ and $\mathbf{E}_\mu^\pi$.

For $\Delta \subseteq R\Pi$ we define the set $L_\Delta = \{\mathbf{P}_\mu^\pi; \mu \in \mathcal{P}(X), \pi \in \Delta\}$ of strategic measures generated by Δ. Then $L = L_{R\Pi}$ is the set of all strategic measures. If one considers a σ–field $\mathcal{L}$ on L induced by σ–field $\mathcal{M}(H_\infty)$ then $(L, \mathcal{L})$ is a Borel space (Dynkin and Yushkevich 1979, Section 5.5), i.e. $L \in \mathcal{M}(H_\infty)$. It is also known that $L_\Pi \in \mathcal{M}(H_\infty)$; see Blackwell (1976) and, for gambling problems, Sudderth (1969).

We consider a general situation when a criterion w is an arbitrary function on L. In other words, a numerical function $w : L \to [-\infty; \infty]$ is called *a criterion.* We define $w^\pi(x) = w(\mathbf{P}_x^\pi)$.

Let Δ be a subset of $R\Pi$. We define a value of Δ by

$$v_\Delta(x) = \sup_{\pi \in \Delta} w^\pi(x).$$

We also define $s(x) = v_S(x)$, $s_R(x) = v_{RS}(x)$, and $v(x) = v_{R\Pi}(x)$.

Following Feinberg (1982a) we say that a criterion w is called *measurable* if the function $w(P)$ is measurable on L. As was observed in Feinberg (1982a), if w is a measurable criterion, then $v(x)$ is upper semianalytic and therefore it is universally measurable function on X.

In fact every natural criterion for a Borel MDP is measurable. In conclusion of this section, we provide examples of measurable criteria.

Let g be a measurable function on H_∞. Then any expected utility criterion, defined by

$$w^\pi(\mu) = \mathbf{E}_\mu^\pi g(x_0 a_0 x_1 a_1 \dots)$$

is measurable. In order to provide the correctness of integration, we agree everywhere in the paper that $(-\infty) + (+\infty) = -\infty$. In particular, one can consider measurable functions r, R, and u on $X \times A$ and define

$$g(x_0 a_0 x_1 a_1 \dots) = \limsup_{n\to\infty} \sum_{n=0}^{\infty} r(x_n, a_n) + \limsup_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} R(x_n, a_n) + \limsup_{n\to\infty} u(x_n, a_n). \tag{2.2}$$

When $R = u = 0$ we have a dynamic programming problem in a broader sense than it is usually considered in the literature, when $r = u = 0$ we have a version of an average reward criterion, and when $r = R = 0$ we have the Dubins-Savage criterion for gambling problems. If lim sup is replaced with lim inf in some of three summands in (2.2), we get different versions of these criteria.

We say that the General Convergence Condition holds if

$$\mathbf{E}_x^\pi \sum_{n=0}^{\infty} r^+(x_n, a_n) < \infty \tag{2.3}$$

for all $x \in X$ and for all $\pi \in \Pi$. If the General Convergence Condition holds then the criterion

$$w^\pi(x) = \mathbf{E}_x^\pi \sum_{i=0}^{\infty} r(x_i, a_i)$$

is well-defined for any initial state x and any policy π. Problems satisfying the General Convergence Condition are more general than positive programming (r is nonnegative and (2.3) holds for $r^+ = r$, Blackwell 1965a), negative programming (r is nonpositive, Strauch 1966), and discounted programming (r is bounded and the system moves at each step to an absorbent state with a given fixed positive probability and the one-step reward in this absorbent state is 0, Blackwell 1965).

For discounted dynamic programming problems, one can also write

$$g(x_0 a_0 x_1 a_1 \ldots) = \sum_{n=0}^{\infty} \beta^n r(x_n, a_n),$$

where $\beta \in [0, 1[$. In a more general situation,

$$g(x_0 a_0 x_1 a_1 \ldots) = \sum_{k=1}^{K} \sum_{n=0}^{\infty} \beta_k^n r_k(x_n, a_n),$$

where K is a positive integer, r_k are bounded above measurable functions on $X \times A$, and $\beta_k \in [0, 1[$, we get weighted discounted criteria; see Feinberg and Shwartz (1994).

Examples of measurable criteria, which are not expected utility criteria, are an average reward per unit time criterion

$$w^\pi(\mu) = \liminf_{N \to \infty} \frac{1}{N} \mathbf{E}_\mu^\pi \sum_{n=0}^{N-1} r(x_n, a_n)$$

and its lim sup version; see Derman (1970), Feinberg and Park (1994) and references therein. We also notice that a measurable function of a several measurable criteria is a measurable criterion. It is true, for example,

for a sum of a finite number of measurable criteria and other operations such as maximum and minimum; see examples of particular criteria in Feinberg (1982a), Krass, Filar, and Sinha (1992), Filar and Vrieze (1992), and Fernández-Gaucherand, Ghosh, and Marcus (1994).

3. Measurability of Sets of Strategic Measures. The goal of this section is to show measurability of the sets L_{SR}, L_S, L_{MR}, and L_M in addition to the known results that the sets L and L_Π are measurable. Our central result is that the set L_{RS} is measurable. Measurability of L_S, L_{MR}, and L_M follows from this result and from measurability of L_Π.

For any$P \in L$ we define a probability measure $Q(\cdot|P)$ on $(X \times A, \mathcal{B}(X \times A))$ and a probability measure $\nu(\cdot|P)$ on $(X, \mathcal{B}(X))$:

$$Q(E|P) = \sum_{n=0}^{\infty} 2^{-(n+1)} P\{(x_n, a_n) \in E\},$$
$$\nu(E'|P) = Q(E' \times A|P),$$

where $E \in \mathcal{B}(X \times A)$, $E' \in \mathcal{B}(X)$.

Since the map $P \to P\{(x_n, a_n) \in E\}$ is measurable for each $n = 0, 1, \ldots$ and $E \in \mathcal{B}(X \times A)$ then Q is a measurable map from $(L, \mathcal{B}(L))$ to $(\mathcal{P}(X \times A), \mathcal{M}(X \times A))$. Hence ν is a measurable map from $(L, \mathcal{B}(L))$ to $(\mathcal{P}(X), \mathcal{M}(X))$. By Proposition 7.27 in Bertsekas and Shreve (1978), there exists a measurable map $q(\cdot|P, x)$: from $(L \times X, \mathcal{B}(L \times X))$ to $(\mathcal{P}(A), \mathcal{M}(A))$ such that

$$Q(E \times D|P) = \int_E q(D|P, x)\nu(dx|P) \tag{3.1}$$

for any $E \in \mathcal{B}(X)$ and any $D \in \mathcal{B}(A)$. We fix some measurable map q : $(L \times X, \mathcal{B}(L \times X)) \to (\mathcal{P}(A), \mathcal{M}(A))$ satisfying (3.1).

For any measure $P = \mathbf{P}_\mu^\pi \in L$ we define a measure $F(P)$ on $(H_\infty, \mathcal{B}(H_\infty))$

$$F(P)(dx_0 da_0 dx_1 da_1 \ldots) = P(dx_0) \prod_{i=0}^{\infty} (q(a_i|P, x_i) p(dx_{i+1}|x_i a_i)). \tag{3.2}$$

Note that $P \to P(\cdot)$ is a measurable map from $(L, \mathcal{B}(L))$ to $(\mathcal{P}(X), \mathcal{M}(X))$, and $q(\cdot|P, x)$ is a measurable map from $\mathrm{v}(L \times X, \mathcal{B}(L \times X))$, and $p(\cdot|x, a)$ is a measurable map from $(X \times A, \mathcal{B}(X \times A))$ to $(X, \mathcal{B}(X))$. By the Ionescu Tulcea theorem (Neveu 1965), $F(P)$ is a measurable map from $(L, \mathcal{B}(L))$ to $(\mathcal{P}(H_\infty), \mathcal{M}(H_\infty))$.

Lemma 3.1. $L_{RS} = \{P \in L : P = F(P)\}$.

Proof. First we show that $P = F(P)$ for all $P \in L_{RS}$. Fix $P \in L_{RS}$. Then $P = \mathbf{P}_\mu^\sigma$ for some $\mu \in \mathcal{P}(X)$, $\sigma \in RS$. Consider marginal distributions

$P_n(X', A') = P\{x_n \in X', a_n \in A'\}$, $P_n^1(X') = P\{x_n \in X', a_n \in A\}$. Then $P_n(dxda) = P_n^1(dx)\sigma(da|x)$ and

$$Q(dxda|P) = \sum_{n=0}^{\infty} 2^{-(n+1)} P_n(dxda)$$
$$= \sum_{n=0}^{\infty} 2^{-(n+1)} P_n^1(dx)\sigma(da|x) = \nu(dx|P)\sigma(da|x).$$

So for any $A' \in \mathcal{B}(A)$

$$q(A'|P, x) = \sigma(A'|x) \qquad (\nu(\cdot|P) - \text{a.s.}), \tag{3.3}$$

i.e. $q(A'|P, x_n) = \sigma(A'|x_n)$ $\quad (P - \text{a.s.})$ for any $n = 0, 1, 2, \ldots$.

To prove $F(P) = P$ it is sufficient to prove that

$$F(P)(dx_0 da_0 \ldots dx_n) = P(dx_0 da_1 \ldots dx_n) \tag{3.4}$$

for each $n = 0, 1, 2, \ldots$. For $n = 0$ we have from (3.2) that $F(P)(dx_0) = P(dx_0)$. Let (3.4) be fulfilled for some $n = 0, 1, \ldots$. Then

$$P(dx_0 da_0 \ldots dx_n da_n) = P(dx_0 da_0 \ldots dx_n)\sigma(da_n|x_n) = \tag{3.5}$$

$$F(P)(dx_0 da_0 \ldots dx_n) \cdot q(da_n|P, x_n) = F(P)(dx_0 da_0 \ldots dx_n da_n)$$

(the first and the last equations follow from (2.1) and (3.2); the second equation follows from the induction hypothesis and (3.3)). And

$$P(dx_0 da_0 \ldots dx_{n+1}) = P(dx_0 da_0 \ldots da_n)p(dx_{n+1}|a_n) =$$

$$F(P)(dx_0 da_0 \ldots da_n) \cdot p(dx_{n+1}|a_n) = F(P)(dx_0 da_0 \ldots dx_{n+1})$$

(the first and last equations follow from (2.1) and (3.2); the second equation follows from (3.5)). So (3.4) is proved for any n.

Now we prove that if $P = F(P)$ then $P \in L_{RS}$. Let $P = F(P)$ and $P = \mathbf{P}_\mu^\pi$. Then by (2.1) and by (3.2)

$$\pi_n(da_n|x_0 a_0 x_1 a_1 \ldots x_n) = q(da_n|P, x_n) \qquad (P - \text{a.s.})$$

for any $n = 0, 1, 2, \ldots$.

For $n = 0, 1, 2, \ldots$ we consider the sets

$$H_n(P) = \{h_n \in H_n;\ \pi_n(A'|h_n) \neq q(A'|P, x_n),\ h_n = x_0 a_0 x_1 \ldots x_n\}.$$

Then $P(H_n(P)) = 0$ for any $n = 0, 1, 2, \ldots$.

We fix some $\sigma' \in RS$ and define

$$\sigma(\cdot|h_n) = \begin{cases} q(\cdot|P, x_n), & \text{if } h_n \notin H_n(P), \\ \sigma'(\cdot|x_n), & \text{if } h_n \in H_n(P), \end{cases}$$

where $n = 0, 1, 2, \ldots$, $h_n = x_0 a_0 x_1 \ldots x_n$.

Then $\pi_n(da_n|x_0 a_0 x_1 \ldots x_n) = \sigma(\cdot|x_n)$ $(P-\text{a.s.})$ for any $n = 0, 1, 2, \ldots$. Consequently, $P = \mathbf{P}_\mu^\pi = \mathbf{P}_\mu^\sigma$, where $\sigma \in RS$. ∎

Theorem 3.2. (i) $L \in \mathcal{M}(H_\infty)$;
(ii) $L_\Pi \in \mathcal{M}(H_\infty)$;
(iii) $L_{RM} \in \mathcal{M}(H_\infty)$;
(iv) $L_M \in \mathcal{M}(H_\infty)$;
(v) $L_{RS} \in \mathcal{M}(H_\infty)$;
(vi) $L_S \in \mathcal{M}(H_\infty)$.

Proof. (i) See sections 3.5, 3.6, and 5.5 in Dynkin and Yushkevich (1979). (ii) This fact was proved by Blackwell (1976, Theorems 1, 2, and 4(b)). Similar results were established for Borel gambling problems by Sudderth (1969) and for analytic gambling problems by Dellacherie (1985). We remark that, since the derived model defined in Blackwell (1976) has the state space $A \times X$ and its transition probabilities do not depend on the first component a, Blackwell (1976) proved in fact that $\mathcal{P}(A) \times L_\Pi \in \mathcal{M}(A \times H_\infty)$. This fact implies (ii). (v) By Lemma 3.1 $L_{RS} = \{P \in L : P = F(P)\}$. Let $I(P) = P$. Since $L \in \mathcal{M}(H_\infty)$ and I and F are (Borel) measurable maps, then L_{RS} is measurable. (iii) We expand the state space X to $X \times \mathbf{N}$, where $\mathbf{N} = \{0, 1, \ldots\}$. This is a standard construction which transforms the sets of Markov and randomized Markov policies respectively into the sets of stationary and randomized stationary policies in a new model; see Feinberg and Sonin (1985) or Feinberg and Shwartz (1994). Let $\tilde{H}_\infty = (X \times \mathbf{N} \times A)^\infty$ be the set of trajectories in a new model. We slightly abuse the notations and write $\tilde{H}_\infty = X^\infty \times A^\infty \times \mathbf{N}^\infty = H_\infty \times \mathbf{N}$. Let $\tilde{L}_{RS}$ be the set of strategic measures in the new model. By (iii) $\tilde{L}_{RS} = \mathcal{M}(\tilde{H}_\infty) = \mathcal{M}(H_\infty) \times \mathcal{M}(\mathbf{N}^\infty)$. Then $\tilde{L}_{RS} = L_{RM} \times \{\delta(0), \delta(1), \ldots, \}$, where $\delta(i)$ is a probability distribution on $\mathbf{N}$ concentrated at i. Therefore (iii) is proved. (iv, vi) $L_M = L_{RM} \cap L_\Pi \in \mathcal{M}(H_\infty)$ and $L_S = L_{RS} \cap L_\Pi \in \mathcal{M}(H_\infty)$. ∎

Remark 3.3. Our proof of $L_{RS} \in \mathcal{M}(H_\infty)$ is based on Lemma 3.1. Ashok Maitra and Bill Sudderth pointed out to the author an alternative proof of this fact which is based on Lemma 2.2 in Maitra, Purves, and Sudderth (1990), according to which for any $P \in \mathcal{P}(H_\infty)$ it is possible to fix versions $P[x_0]$, $P[x_0 a_0]$, $P[x_0 a_0 x_1], \ldots$ of the conditional distributions of a_0 given x_1, x_1 given $(x_0 a_0)$, a_1 given $(x_0 a_0 x_1)$, ..., respectively, that are jointly measurable in P and in conditioning variables. Then L_{RS} is a collection of all P such that

$$P[P[x_0](D(x_0)) = 1,\ P[x_0 a_0] = p(\cdot|x_0, a_0),\ P[x_0 a_0 x_1](D(x_1)) = 1, \ldots] = 1$$

and $P[P[x_0a_0 \dots x_n] = P[x_n],\ n = 1, 2 \dots] = 1$. The measurability of L_{RS} can be established by using Corollary 1 on p. 403 in Sudderth (1969).

4. Measurability of Value Functions. A function $g: X \times [-\infty, \infty]$ is called upper semianalytic if the set $\{x \in X : g(x) > c\}$ is analytic for each c. If all these sets are universally measurable, the function is called universally measurable; see Bertsekas and Shreve (1978) and Dynkin and Yushkevich (1979) for details.

It is well-known that v is upper semianalytic for dynamic programming problems; Strauch (1966). This follows from Theorem 3.2 (i); see Dynkin and Yushkevich (1979). The similar proof holds for an arbitrary measurable criterion; Feinberg (1982a). In this section, we show that Theorem 3.1 implies that v_Π, v_{RM}, v_M, s_R, and s are upper semianalytic functions. For gambling problems, the result for v_Π was established by Sudderth (1969).

Lemma 4.1 *Let $E \in \mathcal{M}(H_\infty)$ and for each $x \in X$ there exist a policy π such that $\mathbf{P}_x^\pi \in E$. Then the function $g(x) = \sup_{\{\pi:\ \mathbf{P}_x^\pi \in E\}} v(P_x^\pi)$ is upper semianalytic.*

Proof. Consider the sets $L(x) = \{P \in L : P\{x_0 = x\} = 1\}$, where $x \in X$, $L^0 = \bigcup_{x \in X} L(x)$, $E(x) = L(x) \cap E$, and $E^0 = L^0 \cap E$. By Dynkin and Yushkevich 1979, Sections 3.5, 3.6, and 5.5, the sets $L(x), x \in X$, and L^0 are measurable. Therefore the sets $E(x), x \in X$, and E^0 are measurable too.

Consider a map $k : L^0 \to X$, $k(P) = x$ if $P \in L^0(x)$. By Dynkin and Yushkevich 1979, Section 5.5, k is a measurable map from $(L_0, \mathcal{B}(L_0))$ onto $(X, \mathcal{B}(X))$. We consider a map $l : E^0 \to X$ which is equal to k when the argument is from E^0, $l(P) = k(P)$ for $P \in E^0$. Since $E^0 \in \mathcal{B}(L^0)$, the map l is measurable. Since

$$g(x) = \sup\{w(P); P \in l^{-1}(x)\}, \qquad x \in X,$$

the function $g(x)$ is upper semianalytic; see Theorem B from Chapter 3 in Dynkin and Yushkevich (1979) or Proposition 7.47 in Bertsekas and Shreve (1979). ∎

We remark that the proof of Lemma 4.1 is similar to the proof that v is upper semianalytic in Dynkin and Yushkevich (1979). Theorem 3.2 and Lemma 4.1 imply the following result.

Theorem 4.2. *If w is a measurable criterion than each of the value functions v, v_Π, v_{RM}, v_M, s_R, and s is upper semianalytic and therefore universally measurable.*

Until the end of this section, consider a dynamic programming problem (or an MDP with the expected total rewards). For a universally measurable function g on X, we define an optimality operator

$$Tg(x) = \sup_{a \in D(x)} T^a g(x),$$

where for $x \in X$ and $a \in D(x)$

$$T^a g(x) = r(x,a) + \int_X g(y)p(dy|x,a).$$

In view of Theorem 4.2, Tg is defined for $g \in \{v, v_\Pi, v_{RM}, v_M, s_R, s\}$.

If the General Convergence Condition holds then the optimality equation $v = Tv$ holds; see f.i. Dynkin and Yushkevich (1979). Under this condition $s = s_R$ (Feinberg 1992) and $v_M = v$ under even weaker conditions (Feinberg 1982, 1982a). However, it is possible that $v \neq s$ for negative dynamic programming (Strauch 1966) which is a particular case of dynamic programming problems satisfying the General Convergence Condition. Under the General Convergence Condition, Feinberg and Sonin (1983) proved that $s = Ts$ when the state space X is countable. It is easy to see that $Ts \geq s$ when X is Borel. Indeed, $w^\phi(x) = T^{\phi(x)}w^\phi(x)$ for any stationary policy ϕ, where $\phi(x)$ is an action that stationary policy ϕ prescribes at state x. Therefore $Ts(x) = \sup_{a\in D(x)} T^a s(x) \geq \sup_{\phi\in S} T^{\phi(x)}w^\phi(x) = \sup_{\pi\in S} w^\phi(x) = s(x)$. If X is Borel, the validity of $s = Ts$ is an open question, because it is not clear why $s \geq Ts$. For the countable state space case, the proof in Feinberg and Sonin (1983) used the existence of uniformly nearly optimal stationary policies proved in that paper. The example by Blackwell and Ramakrishnan (1988) demonstrates that this fact does not hold for Borel state problems even for universally measurable policies. The question whether there exist (a.s.) uniformly nearly optimal policies within the class of stationary policies is open for Borel state dynamic programming problems satisfying the General Convergence Conditions. In Schäl and Sudderth (1987) the existence of stationary uniformly (a.s) uniformly nearly optimal policies was proved for some classes of Borel models for which $s = v$.

5. Representation of Strategic Measures. In this section, we give a new formulation of the following results: (i) any strategic measure can be represented as a mixture (integral convex combination) of strategic measures from L_Π; (i) any strategic measure from L_{RM} can be represented as a mixture of strategic measures from L_M.

Each nonrandomized policy is defined by a measurable map ϕ from $H = \cup_{n=0}^\infty (X \times A)^n \times X$ to A such that $\phi(x_0a_0 \dots x_n) \in D(x_n)$ for each $(x_0a_0 \dots x_n) \in H$. Let $(\Omega, \mathcal{B}(\Omega))$ be a Borel space. We consider a measurable map ϕ of (Ω, H) to A such that for a given ω each map $\phi(\omega, \cdot)$ defines a nonrandomized policy which we denote by $\phi[\omega]$.

Similarly, each Markov policy is defined by a measurable map ϕ from $X \times \mathbf{N}$ to A such that $\phi(x_n, n) \in D(x_n)$. We consider a measurable map ϕ of $(\Omega, X \times \mathbf{N})$ to A such that $\phi(\omega, x, n) \in D(x_n)$ for all $\omega \in \Omega,$, $x \in X$, and $n \in \mathbf{N}$. If we fix some ω, the map $\phi(\omega, \cdot, \cdot)$ defines a Markov policy which we denote by $\phi[\omega]$. By the Ionescu Tulcea theorem (Neveu 1965), given an

initial distribution μ, $\omega \to \mathbf{P}_\mu^{\phi[\omega]}$ are measurable maps from Ω to $\mathcal{M}(H_\infty)$ in the both cases of Markov and arbitrary policies.

Let E be a Borel set and ν be a probability measure on $(\mathcal{P}(E), \mathcal{M}(E))$. We write $\eta = \int\limits_{\mathcal{P}(E)} p\nu(dp)$ if $\eta(C) = \int\limits_{\mathcal{P}(E)} p(C)\nu(dp)$ for each $C \in \mathcal{B}(E)$. Also, let m be a probability measure on a Borel set Ω and let l be a measurable map of Ω to $(\mathcal{P}(E), \mathcal{M}(E))$. Then we write $\eta = \int\limits_{\Omega} l(\omega)m(d\omega)$ if $\eta(C) = \int\limits_{\Omega} l(\omega)(C)m(d\omega)$ for any measurable subset C of E.

Theorem 5.1. (Feinberg 1982, Theorem 1). *Let an initial distribution μ be fixed. There exists a Borel space Ω and a probability measure m on $(\Omega, \mathcal{B}(\Omega))$ with the following properties:*

(i) *For any policy π there exists a measurable map $\phi : (\Omega \times H) \to A$ such that $\phi(\omega, x_0, a_0, \dots, x_n) \in D(x_n)$ for all $(\omega, x_0 a_0 \dots x_n) \in (\Omega \times H)$ and*

$$\mathbf{P}_\mu^\pi = \int\limits_{\Omega} \mathbf{P}_\mu^{\phi[\omega]} m(d\omega); \tag{5.1}$$

(ii) *For any randomized Markov policy π there exists a measurable map $\phi : (\Omega \times X \times \mathbf{N}) \to A$ such that $\phi(\omega, x, n) \in D(x)$ for all $(\omega, x, n) \in (\Omega \times X \times \mathbf{N})$ and (5.1) holds.*

The method of using an auxiliary space $(\Omega, \mathcal{B}(\Omega), m)$ was introduced by Aumann (1964) for games. A version of Theorem 5.1 can be found in Section 1.2 of Gikhman and Skorokhod (1979). Feinberg (1982) used Theorem 5.1(ii) to prove that, given an initial distribution, for any policy there exists a Markov policy with the same or better expected total rewards. A question on the existence of such a policy was formulated by Strauch (1966) for positive programming. Feinberg (1982a) studied applications of Theorem 5.1 to various criteria. Kadelka (1983) announced a result similar to Theorem 5.1. Feinberg (1991) described a sufficient condition (Strong Non-Repeating Condition) that a result similar to Theorem 5.1 holds for a class of policies. In view of Feinberg (1991), M and Π are particular classes of policies for which that condition holds.

For a countable state problem, the measure can be introduced directly on the sets of Markov and nonrandomized policies by using Kolmogorov's theorem. In this case there is no need to consider an auxiliary space. For countable state MDPs, Krylov (1965) proved the result similar to Theorem 5.1(i) and Feinberg (1986) proved such result for arbitrary classes strategies satisfying the so-called Non-Repeating Condition. Hill and Pestien (1987) applied a statement similar to Theorem 5.1 to a countable gambling problem and Sonin (1991) applied it to a finite state gambling problem.

Theorem 3.2 (ii, iv) allows us to formulate a more natural version of Theorem 5.1, in which we do not have to introduce an auxiliary space Ω.

Theorem 5.2. *Let an initial distribution μ be fixed. For any policy π there exists a probability measure ν on L_Π such that*

$$\mathbf{P}_\mu^\pi = \int_{L_\Pi} P\nu(dP). \tag{5.2}$$

If π is a Markov policy then ν can be chosen in a way that $\nu(L_M) = 1$.

Proof. Consider the auxiliary space $(\Omega, \mathcal{B}(\Omega), m)$ introduced in Theorem 5.1. Since the set L_Π is measurable, the measurable map $\Omega \to \mathbf{P}_\mu^{\phi[\omega]}$, considered in Theorem 5.1(i), induces a probability measure ν on L_Π. Therefore, (5.1) implies (5.2). If π is a randomized Markov policy then the proof is the same, but we follow Theorem 5.2(ii) and consider a measurable map $\omega \to \mathbf{P}_\mu^{\phi[\omega]}$ from Ω to L_M which satisfies (5.1). ∎

For $\mu \in \mathcal{P}(X)$ we denote $L(\mu) = \{P \in L : P = \mathbf{P}_\mu^\pi \text{ for some } \pi \in R\Pi\}$. This set is measurable; Dynkin and Yushkevich (1979). We remark that, for an arbitrary policy, the measure ν from (5.2) is concentrated on $L_\Pi \cap L(\mu)$. If π is randomized Markov and ν satisfies (5.2) and is concentrated on L_M then ν is concentrated on $L_M \cap L(\mu)$.

A natural question is whether a measure ν that satisfies (5.2) is unique. The following example gives a negative answer to this question.

Example 5.3. Let $X = \{0,1,2,3\}$, $A = \{1,2\}$, $D(0) = D(3) = \{1\}$, and $D(1) = D(2) = A$. Let also $p(i|0,1) = .5$, $p(3|i,j) = p(3|3,1) = 1$, $i,j = 1,2$, with other probabilities equal zero. Let $x_0 = 0$. The process always moves from 0 to either 1 or 2 with probabilities .5 and then it moves to 3 which is an absorbing state. We consider a randomized Markov policy π such that $\pi_1(i|j) = .5$, $i,j = 1,2$, and four Markov policies $\phi[ij]$ with $\phi[ij]_1(1) = i$ and $\phi[ij]_1(2) = j$. Then $\mathbf{P}_0^\pi = .5\mathbf{P}_0^{\phi[11]} + .5\mathbf{P}_0^{\phi[22]} = .5\mathbf{P}_0^{\phi[12]} + .5\mathbf{P}_0^{\phi[21]}$. ∎

Let π be a randomized Markov policy. As we see from the previous example, a measure ν that satisfies (5.2) may not be unique. According to Theorem 5.2, it is possible to select ν such that (5.2) holds and ν is concentrated on L_M. The following example shows that it is possible that there exists ν which satisfies (5.2) and is not concentrated on L_M.

Example 5.4. Let $X = \{0,1,2,3\}$, $A = \{1,2\}$, $D(0) = D(1) = D(2) = \{1\}$, and $D(3) = A$. Let also $p(i|0,1) = .5$, $p(3|i,1) = p(3|3,i) = 1$, $i = 1,2$, with other probabilities equal zero. Let $x_0 = 0$. Like in the previous example, the process always moves from 0 to either 1 or 2 with probabilities .5 and then it moves to 3 which is an absorbing state. We consider a randomized Markov policy π such that $\pi_2(i|3) = .5$, $i = 1,2$, and $\pi_n(1|3) = 1$ for $n = 3,4,\ldots$. We also consider four nonrandomized policies $\phi[ij]$, $i,j = 1,2$, such that $\phi[ij]_2(0,1,1,1,3) = i$, $\phi[ij]_2(0,1,2,1,3) = j$, and $\phi[ij]_n(x_0a_0\ldots x_n) = 1$ for $n = 3,4,\ldots$.

Then $\mathbf{P}_0^\pi = \sum_{i=1}^{2}\sum_{j=1}^{2} .25\mathbf{P}_0^{\phi[ij]}$. In this case, ν is concentrated on four points and $\nu(\mathbf{P}_0^{\phi[ij]}) = .25$. Since each policy $\phi[ij]$ is not Markov, $\nu(L_M) = 0$.

On the other side, one can consider Markov policies $\phi[i]$, $i = 1, 2$, with $\phi_2[i](3) = i$ and $\phi[i]_n(3) = 1$ for $n = 2, 3, \ldots$. In this case $\nu(L_M) = 1$, where ν is concentrated at two points with $\nu(P_0^{\phi[i]}) = .5$, $i = 1, 2$. We also have $\mathbf{P}_0^{\pi} = .5\mathbf{P}_0^{\phi[1]} + .5\mathbf{P}_0^{\phi[2]}$. The second selection of measure ν is consistent with the statement of Theorem 5.2 for Markov policies. ∎

We also remark that if π is a randomized stationary policy then it is randomized Markov. Theorem 5.2 states that (5.2) holds for some ν concentrated on L_M. However, it is possible that there is no ν for which (5.2) holds and which is concentrated on L_S; see Remark 3.1 in Feinberg (1986) or Example 2.3 in Hill and Pestien (1987).

Acknowledgement. The author expresses deep thanks to Ashok Maitra and Bill Sudderth for pointing out the results of Sudderth (1969) and Blackwell (1976) on measurability of L_Π. The first version of this paper left open the questions of measurability of the sets L_Π, L_M, and L_S and of the functions v_Π, v_M. and s. In particular, measurability of L_Π allowed us to formulate Theorem 5.2. I also thank them for Remark 3.3. This research was partially supported by National Science Foundation grant DMI-9500746.

REFERENCES

AUMANN, R. J. (1964). Mixed and behavior strategies in infinite extensive games. *Ann. Math. Studies* **53** 627–650.

BALDER, E. (1989). On compactness of the space of policies in stochastic dynamic programming. *Stochastic Process. Appl.* **32** 141–150.

BERTSEKAS, D. P. AND SHREVE, S. E. (1978). *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, New York.

BLACKWELL, D. (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

BLACKWELL, D. (1965a). Positive dynamic programming. *Proc. 5th Berkeley Symp. Math. Statist. and Probability* **1** 415–418.

BLACKWELL, D. (1976). The stochastic processes of Borel gambling and dynamic programming. *Ann. Statist.* **4** 370–374.

BLACKWELL, D., FREEDMAN, D., AND ORKIN, M. (1974). The optimal reward operator in dynamic programming. *Ann. Probab.* **2** 926–941.

BLACKWELL, D. AND RAMAKRISHNAN, S. (1988). Stationary plans need not be uniformly adequate for leavable, Borel gambling problems. Proc. Amer. Math. Soc. **102** 1024–1027.

DELLACHERIE, C. (1985). Quelques résultats sur les maisons de jeux analytiques. *Lecture Notes in Math.* **1123** 222-229.

DERMAN, C. (1970). *Finite State Markovian Decision Processes.* Academic Press, New York.

DYNKIN, E. B. AND YUSHKEVICH, A. A. (1979). *Controlled Markov Processes.* Springer-Verlag, New York.

FEINBERG, E. A. (1982). Non-randomized Markov and semi-Markov strategies in dynamic programming. *Theory Probab. Appl.* **27** 116–126.

FEINBERG, E. A. (1982a). Controlled Markov processes with arbitrary numerical criteria. *Theory Probab. Appl.* **27** 486–503.

FEINBERG, E. A. (1986). Sufficient classes of strategies in stochastic dynamic programming. I: Decomposition of randomized strategies and imbedded models. *Theory Probab. Appl.* **31** 658–668.

FEINBERG, E. A. (1987). Sufficient classes of strategies in discrete dynamic programming. II: Locally stationary policies. *Theory Probab. Appl.* **32** 478–493.

FEINBERG, E. A. (1991). Non-randomized strategies in stochastic decision processes. *Ann. Oper. Res.* **29** 315–332.

FEINBERG, E. A. (1992). On stationary strategies in Borel dynamic programming. *Math. Oper. Res.* **17** 392–397.

FEINBERG, E. A. AND PARK, H. (1994). Finite state Markov decision models with average reward criteria. *Stochastic Process. Appl.* **49** 159–177.

FEINBERG, E. A. AND SHWARTZ, A. (1994). Markov decision models with weighted discounted criteria. *Math. Oper. Res.* **19** 152–168.

FEINBERG, E. A. AND SONIN, I. M. (1983). Stationary and Markov policies in countable state dynamic programming, *Lecture Notes in Math.* **1021** 111–129.

FEINBERG, E. A. AND SONIN, I. M. (1985). Persistently nearly optimal strategies in stochastic dynamic programming, in: *Statistics and Control of Stochastic Processes (Steklov Seminar* 1984*)* (eds. N. V. Krylov, R. S. Liptser, and A. A. Novikov), Optimization Software, New York, 69–101.

FERNÁNDEZ-GAUCHERAND, E., GHOSH, M.K., AND MARCUS, S. I. (1994). Controlled Markov processes on the infinite planning horizon: weighted and overtaking cost criteria. *ZOR-Methods and Models of Oper. Res.* **39** 131–155.

FILAR, J. A. AND VRIEZE, O. J. (1992). Weighted reward criteria in competitive Markov decision processes. *ZOR-Methods and Models of Oper. Res.* **36** 343–358.

FREEDMAN, D. (1974). The optimal reward operator in special classes of dynamic programming problems. *Ann. Probab.* **2** 942–949.

GIKHMAN, I. I. AND SKOROKHOD, A. V. (1979). *Controlled Random Processes.* Springer-Verlag, New York.

HILL, T. P. AND PESTIEN, V. C. (1987). The existence of good Markov strategies for decision processes with general payoffs. *Stochastic Process. Appl.* **24** 61–76.

HINDERER, K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter.* Springer-Verlag, Berlin.

KADELKA, D. (1983). On randomized policies and mixtures of deterministic policies in dynamic programming. *ZOR - Methods of Operations Research* **46** 67–75.

KRASS, D., FILAR, J. A., AND SINHA, S. S. (1992). A weighted Markov decision process. *Oper. Res.* **40** 1180–1187.

KRYLOV, N. V. (1965). The construction of an optimal strategy for a finite controlled chain. *Theory Probab. Appl.* **10** 45–54.

KURATOWSKI, K. (1966). *Topology I.* Academic Press, New York.

LUSIN, N. (1927). Sur les ensembles analytiques. *Fund. Math.* **10** 1–95.

MAITRA, A., PURVES, R., AND SUDDERTH, W. (1990). Leavable gambling problems with unbounded utilities. *Trans. Amer. Math. Soc.* **320** 543–567.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability.* Holden-Day, San Francisco.

PUTERMAN, M. L. (1994). *Markov Decision Processes.* Wiley, New York.

SCHÄL, M. (1975). On dynamic programming: compactness of the space of policies. *Stochastic Process. Appl.* **3** 345–364.

SCHÄL, M. (1983). Stationary policies in dynamic programming models under compactness assumptions. *Math. Oper. Res.* **8** 366–372.

SCHÄL, M. (1989). On stochastic dynamic programming: a bridge between Markov decision processes and gambling. in *Markov Processes and Control Theory* (eds. H. Langer and V. Nollau), Mathematical Research **54**, 178–216, Akademie-Verlag, Berlin.

SCHÄL, M. AND SUDDERTH, W. (1987). Stationary policies and Markov policies in Borel dynamic programming. *Prob. Th. Rel. Fields* **74** 91–111.

SONIN, I. M. (1991). On an extremal property of Markov chains and sufficiency of Markov strategies in Markov decision processes with the Dubins-Savage criterion. *Ann. Oper. Res.* **29** 417–426.

STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.

STRAUCH, R. E. (1967). Measurable gambling houses. *Trans. Amer. Math. Soc.* **126** 64–72.

SUDDERTH, W. D. (1969). On the existence of good stationary strategies. *Trans. Amer. Math. Soc.* **135** 399–414.

WAGNER, D. H. (1977). Survey of measurable selection theorems. *SIAM J. Control Optimization* **15** 859–903.

W.A. HARRIMAN SCHOOL FOR MANAGEMENT AND POLICY
STATE UNIVERSITY OF NEW YORK AT STONY BROOK
STONY BROOK, NY 11794-3775
efeinber@fac.har.sunysb.edu