PROBABILITY
AND
MATHEMATICAL STATISTICS

Vol. 1, Fasc. 2 (1980), p. 171-184

EXISTENCE OF OPTIMAL POLICIES IN STOCHASTIC DYNAMIC PROGRAMMING

BY

LAWRENCE D. BROWN AND BHARAT T. DOSHI
(NEW BRUNSWICK, NEW JERSEY)

Abstract. This paper* deals with a general discrete-time stochastic dynamic programming model. Under rather general conditions on the cost functions and the law of motion it is shown that there exists a fully optimal Borel measurable policy, that is, a policy which is optimal for future at every stage and every possible history of the process up to that stage. For the stationary dynamic programming model this implies the existence of a fully optimal stationary policy.

1. INTRODUCTION

Stochastic dynamic programming is a problem of sequential decisions under uncertainty. The basic elements of such a problem can be described as follows: We have an observable stochastic process $\{x_t: t \in T = \{1, 2, \ldots\}\}$, $x_t \in X_t$, $t \in T$. At each stage t an action a_t can be chosen from the action space A_t. This selection can be based only on the history $(x_1, a_1, x_2, \ldots, a_{t-1}, x_t)$ of the process $\{x_t, a_t: t \in T\}$ up to time t. Given the history $(x_1, a_1, x_2, \ldots, a_{t-1}, x_t)$ and the action a_t at time t, the state x_{t+1} is determined by the probability measure

$$q_t(\cdot \mid x_1, x_2, \ldots, x_t, a_1, a_2, \ldots, a_t).$$

* The first-named author, from Cornell University, was supported in part by NSF Grant MCS 75023343-A01. The second-named author, from Bell Laboratories, was supported in part by Summer Research Fellowship, Rutgers University.

This work was done while the authors were at Rutgers University.

Also a cost $l_t(x_1, x_2, \ldots, x_t, a_1, a_2, \ldots, a_t)$ is incurred at time t. For a given realization $\{x_t, a_t: t \in T\}$ the loss is

$$\sum_{t \in T} l_t(x_1, x_2, \ldots, x_t, a_1, a_2, \ldots, a_t).$$

A *policy* δ is a measurable rule of selecting actions. For a given initial state x_1 and a policy δ the *risk* $R(x_1, \delta)$ is defined to be the expectation of the loss function under δ. For a given initial measure μ and a policy δ, the risk is the expectation of $R(x_1, \delta)$ with respect to μ. An *x_1-optimal policy* is the one minimizing $R(x_1, \delta)$ over all policies. A *μ-optimal policy* is a policy which minimizes $R(\mu, \delta)$ over all policies. An *optimal policy* minimizes $R(x_1, \delta)$ over all policies for all x_1. Finally, a policy that is optimal at each stage of the process will be called *fully optimal.*

Under appropriate assumptions we show that there exists a fully optimal policy. This of course implies that for each initial measure μ there exists a μ-optimal policy, a result which was previously proved by Schäl [13] under assumptions similar to ours.

When the state and action spaces are independent of time and the one-step cost function l_t is independent of t up to a multiple α^t, the resulting problem is called *stationary dynamic programming.* Stationary dynamic programming was extensively studied by Blackwell [2], [3] and Strauch [16]. One of their main concerns was to establish conditions under which a stationary optimal policy exists. Both gave examples in which a Borel measurable stationary optimal policy does not exist. Under our assumptions on the action space and the one-step cost function this difficulty does not arise and we can show the existence of a stationary fully optimal policy.

For the stationary dynamic programming Freedman [7], Furukawa [8] and Schäl [14] also give sufficient conditions for the existence of a Borel measurable stationary optimal policy. Freedman [7] (condition (3)) and Schäl [14] (condition (W)) assume continuity of q and lower semi-continuity of l in both the state and action variables. Furukawa [8] and Schäl [14], (S), assume continuity of q and lower semi-continuity of l in the action only. Their conditions on l are similar, but less explicit than ours. Thus, in the special case of stationary dynamic programming, Schäl's Theorem 15.2 is similar to our Theorem 5.1. (It is better in the sense that his regularity condition (S2) or (S2)′ is weaker than our corresponding Assumption 2.12 when applied to the stationary case.) Under the assumptions of Freedman [7] and Schäl [14], (W), it is possible to convert the general non-stationary model into a stationary one by considering the entire history up to any stage as the state variable at that stage. Thus their results give the existence of a fully optimal Borel measurable policy in the non-stationary case under the appropriate regularity conditions on l and q. However, these conditions do

not hold in our model. Such a conversion is not possible in (S) of [14] or in [8] because conditions (S2) and (S2)′ of Schäl [14] will be violated in the resulting stationary model. In our model, Assumption 2.12 would not hold if we were to attempt such a conversion. This necessitates an explicit treatment of a non-stationary model. Such an explicit treatment of a non-stationary model may be found in Hinderer [9], Kertz and Nachman [10] and Schäl [15] among which only Kertz and Nachman [10] prove the existence of a fully optimal policy. However, their assumptions do not imply, nor are implied by ours.

Our results and proofs rely heavily on a general decision problem studied by Brown [4] and the measurable selections of extrema investigated by Brown and Purves [5].

2. CHARACTERISTICS AND FORMULATION

We now characterize a (stochastic) dynamic programming problem and specify assumptions on its various components.

Let $T_k = \{k, k+1, \ldots\}$ and $T = T_1$.

2.1. State space. For each $t \in T$, the *state space* X_t *at stage* t is assumed to be a Borel subset of a complete separable metric space. Let X_t be endowed with the usual topology and let $\mathscr{B}_t$ denote the Borel σ-algebra on X_t.

2.2. Sample space. The *sample space* is

$$X = \mathop{\mathbf{X}}_{t \in T} X_t.$$

Let

$$X_{(t)} = \mathop{\mathbf{X}}_{s \leqslant t} X_s, \quad \mathscr{B}_{(t)} = \mathop{\mathbf{X}}_{s \leqslant t} \mathscr{B}_t \quad \text{and} \quad \mathscr{B} = \mathop{\mathbf{X}}_{t \in T} \mathscr{B}_t = \mathscr{B}_{(\infty)}.$$

2.3. Action space. The set of available actions at stage t is an *action space* A_t, which is assumed to be a compact subset of a complete separable metric space. Let τ_t and $\mathscr{A}_t$ denote the usual topology and the Borel σ-algebra on A_t, respectively. Let

$$A = \mathop{\mathbf{X}}_{t \in T} A_t$$

with the product σ-algebra $\mathscr{A}$. For $t \in T$, let

$$A_{(t)} = \mathop{\mathbf{X}}_{s \leqslant t} A_s \quad \text{and} \quad \mathscr{A}_{(t)} = \mathop{\mathbf{X}}_{s \leqslant t} \mathscr{A}_s.$$

We will use π_t and $\pi_{(t)}$ to denote the natural projection maps on X, A, etc.

2.4. Law of motion. The *law of motion* is a sequence $q = \{q_t\}$ such that q_t is a measurable conditional distribution (Markov kernel) on $\mathscr{B}_{t+1}$ given $\mathscr{B}_{(t)} \times \mathscr{A}_{(t)}$. Formally,

(i) $q_t(\cdot | x, a)$ is a version of the conditional probability distribution on $\mathscr{B}_{t+1}$ given $\mathscr{B}_{(t)} \times \mathscr{A}_{(t)}$.

(ii) $q_t(B | \cdot, \cdot)$ is $(\mathscr{B}_{(t)} \times \mathscr{A}_{(t)})$-measurable for all $B \in \mathscr{B}_{t+1}$.

The interpretation is that, given $x_{(t)} = \pi_{(t)}(x)$ and $a_{(t)} = \pi_{(t)}(a)$, x_{t+1} is selected according to the measure $q_t(\cdot | x, a)$. (Note that $q_t(\cdot | x, a)$ depends on x and a only through $x_{(t)}$ and $a_{(t)}$. We will adopt similar notational conventions later in the paper.)

For any probability measure μ on the initial state x_1, the law of motion q will then define a conditional distribution F_μ on $(X, \mathscr{B})$ given $\mathscr{A}$ according to the natural formula

$$F_\mu\Big(\Big(\bigtimes_{t=1}^{n} B_t\Big) \times \Big(\bigtimes_{t=n+1}^{\infty} X_t\Big) \Big| a\Big)$$

$$= \int_{B_1} \cdots \int_{B_{n-1}} \int_{B_n} q_{n-1}(dx_n | x, a)\, q_{n-2}(dx_{n-1} | x, a) \cdots q_1(dx_2 | x, a)\, \mu(dx_1).$$

Let μ_{x_1} denote the distribution degenerate at $x_1 \in X_1$. We will write $F_{x_1} = F_{\mu_{x_1}}$.

2.5. Policies. A (*sequential*) *policy* is a measurable non-anticipative rule of selecting actions. Formally, $\delta = \{\delta_t\}$ is a *policy* if

(i) $\delta_t(\cdot | x, a)$ is a probability measure on $\mathscr{A}_t$ for all $x \in X$ and $a \in A$, $t \geqslant 2$;

(ii) $\delta_t(C | \cdot, \cdot)$ is $(\mathscr{B}_{(t)} \times \mathscr{A}_{(t-1)})$-measurable for each $C \in \mathscr{A}_t$, $t \geqslant 2$;

(iii) $\delta_1(\cdot | x)$ is a probability measure on $\mathscr{A}_1$ for all $x \in X$;

(iv) $\delta_1(C | \cdot)$ is $\mathscr{B}_1$-measurable for all $C \in \mathscr{A}_1$.

The interpretation is that, if $x_1, a_1, \ldots, a_{t-1}, x_t$ is the observed history up to stage t, then a_t is chosen according to the measure $\delta_t(\cdot | x, a)$. Let $\mathscr{D} = \{\delta\}$ denote the set of all policies.

2.6. Observed process. The preceding suffices to guarantee the existence of a well-defined stochastic process of the observed states and action when the initial measure is μ and a policy δ is used. This process is defined on $X \times A$ with the corresponding σ-algebra. The probability of a cylinder set

$$(B_1, C_1) \times (B_2, C_2) \times \cdots \times (B_n, C_n) \times (X_{n+1}, A_{n+1}) \times (X_{n+2}, A_{n+2}) \cdots$$

is given by

$$\Delta_{\mu,\delta}\Big(\bigtimes_{t=1}^{n} (B_t, C_t) \times \bigtimes_{t>n} (X_t, A_t)\Big)$$

$$= \int_{B_1} \int_{C_1} \cdots \int_{B_n} \int_{C_n} \delta_n(da_n | x, a)\, q_{n-1}(dx_n | x, a) \cdots \delta_1(da_1 | x)\, \mu(dx_1).$$

Similarly, if μ is a probability measure on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$, $k \in T$, then define $\Delta^k_{\mu,\delta}$ by

$$\Delta^k_{\mu,\delta}\Big(\mathop{\mathbf{X}}_{t=1}^{n}(B_t, C_t) \times \mathop{\mathbf{X}}_{t>n}(X_t, A_t)\Big)$$
$$= \int_{B_1}\int_{C_1}\cdots\int_{B_n}\int_{C_n} \delta_n(da_n \mid x, a)\, q_{n-1}(dx_n \mid x, a) \cdots \delta_k(da_k \mid x, a)\, \mu\big(d(x_{(k)}, a_{(k-1)})\big).$$

Note that $\Delta^1_{\mu,\delta} = \Delta_{\mu,\delta}$ and also that $\Delta^k_{\mu,\delta}$ depends on δ only through $\{\delta_t\colon t \geqslant k\}$.

2.7. Costs. The *cost incurred at stage* t is $l_t(x_{(t)}, a_{(t)})$ if the observed history is $(x_1, a_1, \ldots, a_{t-1}, x_t)$ and a_t is chosen at stage t.

2.8. Assumption. For $t \in T$

(i) $l_t(\cdot, \cdot)$ is $(\mathscr{B}_{(t)} \times \mathscr{A}_{(t)})$-measurable;

(ii) $l_t(\cdot, \cdot)$ is non-negative extended real valued;

(iii) $l_t(x_{(t)}, \cdot)$ is lower semi-continuous for each $x_{(t)} \in X_{(t)}$.

In most of what follows the non-negativity of $l_t(\cdot, \cdot)$ can be replaced by the weaker assumption that $l_t(\cdot, \cdot)$ is bounded from below or by the even weaker assumption that $\{l_t(\cdot, \cdot)\}$ is uniformly integrable.

2.9. Loss function. The *loss function* $L\colon X \times A \to \boldsymbol{R}^+$ is defined by

$$L(x, a) = \sum_{t=1}^{\infty} l_t(x_{(t)}, a_{(t)}) \quad (x \in X,\ a \in A).$$

Also define $L^n_k\colon X \times A \to \boldsymbol{R}^+$ by

$$L^n_k(x, a) = \sum_{t=k}^{n} l_t(x_{(t)}, a_{(t)}) \quad (1 \leqslant k \leqslant n \leqslant \infty,\ k < \infty).$$

Let $L_k = L^\infty_k$ and $L^n = L^n_1$ $(1 \leqslant k \leqslant n \leqslant \infty)$.

2.10. Risk function. Let $1 \leqslant k \leqslant n \leqslant \infty$ as above. For a measure μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$ and a policy $\delta \in \mathscr{D}$, the *conditional risk* R^n_k is defined to be the expectation of the loss function $L^n_k(\cdot, \cdot)$ computed under the measure $\Delta^k_{\mu,\delta}$:

$$R^n_k(\mu, \delta) = \mathrm{E}_{\Delta^k_{\mu,\delta}}\big(L^n_k(x, a)\big) = \int_{X \times A} L^n_k(x, a)\, \Delta^k_{\mu,\delta}(dx, da).$$

Let $R_k = R^\infty_k$, $R^n = R^n_1$ and $R = R_1$. Then $R(\mu, \cdot)$ is called the *risk for the initial measure* μ (on $\mathscr{B}_1$). In particular, if μ_{x_1} is the probability measure degenerate at $x_1 \in X_1$, then let $R(x_1, \delta) = R(\mu_{x_1}, \delta)$.

Given μ on $\mathscr{B}_1$, define $\Delta_{\mu,\delta|k}$ to be the projection of $\Delta_{\mu,\delta}$ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$. It will be of special interest to consider $R_k(\Delta_{\mu,\delta|k}, \delta)$, for this can be interpreted as the conditional risk from stage k (inclusive) onward, given that the initial measure was μ, the decision rules $\delta_1, \delta_2, \ldots, \delta_{k-1}$ have already been used, and that the rules $\delta_k, \delta_{k+1}, \ldots$ will be used from stage k onward.

2.11. Optimal policies. Let $k \in T$ and let μ be a probability measure on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$. If

$$R_k(\mu, \delta') = \inf_{\delta \in \mathscr{D}} R_k(\mu, \delta),$$

then δ' is called *μ-optimal at stage k*. If $\mu = \mu_{(x_{(k)}, a_{(k-1)})}$, then δ' is called *$(x_{(k)}, a_{(k-1)})$-optimal at stage k*. A policy δ is called *optimal at stage k* if it is optimal for $(x_{(k)}, a_{(k-1)})$ at stage k for all $(x_{(k)}, a_{(k-1)}) \in X_{(k)} \times A_{(k-1)}$. A policy is called *fully optimal* if it is optimal at stage k for all $k \in T$.

For $(x_{(k)}, a_{(k-1)}) \in X_{(k)} \times A_{(k-1)}$ let

$$R_k(x_{(k)};\ a_{(k-1)}) = \inf_{\delta \in \mathscr{D}} R_k(\mu_{(x_{(k)}, a_{(k-1)})}, \delta).$$

Define $R_k^n(\cdot;\cdot)$ similarly relative to the loss function L_k^n.

2.12. Assumption. (a) The law of motion $q = \{q_t\}$ is *dominated*. That is, there exists a law of motion $v = \{v_t\}$ which is independent of $a \in A$ and such that $q_t \ll v_t$. Let

$$\varphi_t(\cdot \mid x, a) = \frac{dq_t(\cdot \mid x, a)}{dv_t(\cdot \mid x)} \qquad (a \in A,\ t \in T).$$

We also assume that $\varphi_t(\cdot \mid \cdot, \cdot)$ is $(\mathscr{B}_{t+1} \times \mathscr{B}_{(t)} \times \mathscr{A}_{(t)})$-measurable.

(b) Assume also that for each $t \in T$, $x \in X$, the map

$$A \to L_1\left(X_{t+1}, \mathscr{B}_{t+1}, v_t(\cdot \mid x)\right)$$

defined by $a \to \varphi_t(\cdot \mid x, a)$ is continuous. Note that this map depends on a only through $a_{(t)} = \pi_{(t)}(a)$.

Let μ denote a probability measure on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$ and let $\hat{\mu}$ denote its projection (marginal measure) on $\mathscr{B}_{(k)}$. Let V_μ^k denote the measure on $\mathscr{B}$ generated by $\hat{\mu}$ through stage k followed by v. Let $\varphi_\mu^k(\cdot \mid a)$ denote the conditional measure on $\mathscr{B}$ defined by

$$\varphi_\mu^k\Big(B_{(k)} \times \mathop{\mathbf{X}}_{t=k+1}^{n} B_t \times \mathop{\mathbf{X}}_{t=n+1}^{\infty} X_t \mid a\Big)$$
$$= \int_{A_{(k-1)}} \int_{B_{(k)}} \int_{B_{k+1}} \cdots \int_{B_n} q_{n-1}(dx_n \mid x, a) \cdots q_k(dx_{k+1} \mid x, a)\, \mu\big(d(x_{(k)}, a_{(k-1)})\big).$$

Note that $\varphi_\mu^k(\cdot \mid a)$ depends on a only through $a_k, a_{k+1}, \ldots$ Assumption 2.12 (a) guarantees that $\varphi_\mu^k(\cdot \mid a) \ll V_\mu^k(\cdot)$ for all $a \in A$, $k \in T$, μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$. (It also guarantees that the projection of $\Delta_{\mu,\delta}^k$ on $\mathscr{B}$ is dominated by V_μ^k.) Let

$$f_\mu^k(\cdot \mid a) = \frac{d\varphi_\mu^k(\cdot \mid a)}{dV_\mu^k(\cdot)}.$$

Then Assumption 2.12 (b) implies that for each $k \in T$ the map

$$N_k(\mu)\colon A \to L_1(X, \mathscr{B}, V_\mu^k)$$

defined by $a \to f_\mu^k(\cdot \mid a)$ is continuous for each $k \in T$ and μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$.

Note that $\varphi_\mu^1 = F_\mu$, and so, in particular, the map

$$N_1(\mu)\colon a \to \frac{dF_\mu}{dV_\mu^1} = f_\mu^1$$

is continuous for each initial measure μ on $\mathscr{B}_1$.

The assumption above is a significant restriction on the family q. However, the necessity of some assumption similar to 2.12 (b) is demonstrated in Example 3.4.

2.13. Topologizing the policies. Corresponding to each policy $\delta = \{\delta_t\} \in \mathscr{D}$ there is a conditional measure on A, given $\mathscr{B}$, defined by

$$\delta(C_1 \times C_2 \times \ldots \times C_t \times A_{t+1}, \ldots \mid x) = \int_{C_1} \int_{C_2} \cdots \int_{C_{t-1}} \delta_t(C_t \mid x, a)\, \delta_{t-1}(da_{t-1} \mid x, a) \ldots \delta_1(da_1 \mid x).$$

For each $k \in T$ and μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$ define the topology $\tau_k(\mu)$ on $\mathscr{D}$ as the weak topology generated by the functions

$$\delta \to \int \delta(da \mid x)\, f(x)\, c(a)\, V_\mu^k(dx)$$

for $f \in L_1(X, \prod_{t \in T_k} \mathscr{B}_t, V_\mu^k)$ and $c \in C(\prod_{t \in T_k} A_t)$, where $C(\prod_{t \in T_k} A_t)$ denotes the set of continuous functions on $\prod_{t \in T_k} A_t$. Note that $\tau_k(\mu)$ really depends only on the projection measure $\hat{\mu}$ on $\mathscr{B}_{(k)}$. Let $\tau(\mu) = \tau_1(\mu)$ and $\tau(x_1) = \tau(\mu_{x_1})$, $x_1 \in X_1$.

These topologies are not Hausdorff. For an intensive study of $\mathscr{D}$ with these topologies it is useful to define the space $\mathscr{D}_k(\mu)$ of equivalence classes of $\mathscr{D}$ with the topology $\tau_k(\mu)$. This is done in [4]. It may help the reader to visualize $\mathscr{D}$, $\tau(\mu)$, etc., in this manner but we will not explicitly need this terminology in the sequel.

The following important results may be directly deduced from Theorems 3.6 and 3.14 of [4].

2.14. THEOREM. *For each $k \in T$, μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$,*

(a) *the spaces $\mathscr{D}$, $\tau_k(\mu)$ are compact,*

(b) *the maps $R_k(\mu, \cdot)\colon \mathscr{D} \to [0, \infty]$ are lower semi-continuous relative to $\tau_k(\mu)$.*

It is implicit in Theorem 2.14 (b) that $R_k(\mu, \delta)$ depends only on the $\tau_k(\mu)$ equivalence class of $\delta \in \mathscr{D}$. For $k = 1$ this latter fact follows from the expression

$$R(\mu, \delta) = \int L(x, a) f_\mu^1(x|a)\, \delta(da|x)\, V_\mu(dx)$$

and from Theorem 3.10 in [4]. This expression is derived in [4], Proposition 2.2. Similar expressions are valid for $R_k(\mu, \delta)$ but are not explicitly needed in the sequel. (The above result is comparable to results in [13].)

3. POLICIES OPTIMAL FOR μ

3.1. THEOREM. *For each $k \in T$ and each μ on $\mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)}$ there exists an optimal policy for μ at stage k. In particular, for each $x_1 \in X_1$ there is an x_1-optimal policy at stage* 1.

Proof. The theorem follows directly from Theorem 2.14 since any lower semi-continuous function on a compact set achieves its infimum.

This theorem leaves many questions unanswered. For example, if μ is a given initial measure on $\mathscr{B}_1$ and δ is μ-optimal at stage 1, then is δ an x_1-optimal policy at stage 1 for μ-almost every x_1? We have been able to adapt methods developed in Eaton [6] to answer this question (and other similar ones) in the affirmative when the measures $\{V^1_{\mu_{x_1}} : x_1 \in X_1\}$ form a dominated family. However, the optimality results proved in the next section by a different method are stronger than the results we can prove using Eaton's methods, and so we will omit the detailed argument.

The following technical result will later be useful:

3.2. LEMMA. $R(x_1) = \lim_{n \to \infty} R^n(x_1)$.

Proof. $R^n(x_1, \delta)$ is the risk in the problem with the loss function

$$L^n(x, a) = \sum_{t=1}^{n} l_t(x_{(t)}, a_{(t)}).$$

By Theorem 3.1 there exists an optimal policy δ_n for each such problem. Since $\mathscr{D}, \tau(x_1)$ is compact, there exists a "convergent" subnet $\delta_{n'} \to \delta$. (Actually δ is determined only up to its $\tau(x_1)$ equivalence class and $\delta_{n'} \to \delta$ only in this sense.) By Theorem 2.14 (b) we have

$$\liminf_{\{n'\}} R^m(x_1, \delta_{n'}) \geqslant R^m(x_1, \delta) \quad \text{for each } m \in T.$$

Also

$$R(x_1) \geqslant R^n(x_1, \delta_n) \geqslant R^m(x_1, \delta_n) \quad (n \geqslant m).$$

Hence

$$R(x_1) \geqslant \lim_{m \to \infty} \limsup_{\{n'\}} R^m(x_1, \delta_{n'}) \geqslant \lim_{m \to \infty} R^m(x_1, \delta) = R(x_1, \delta) \geqslant R(x_1).$$

This proves the lemma.

3.3. Remark. The assumption that A be compact may be relaxed. For example, it suffices for A to be locally compact with

$$\lim_{a \to \infty} L(x, a) = \infty$$

and for certain other technical conditions to be satisfied. Details may be deduced from Section 4 of [4] together with the above theorems. A similar remark is also valid for later theorems in this paper.

The forthcoming example demonstrates the important role of Assumption 2.12 (b) concerning the continuity of the maps $a \to \varphi_t(\cdot|x, a)$. The example demonstrates that Theorem 3.1 is false under the weaker assumption that for fixed μ the map

$$a \to \frac{dF_\mu(\cdot|a)}{dV_\mu^1(\cdot)}$$

is continuous as a map from A to $L_1(X, \mathscr{B}, V_\mu^1)$ with the weak topology.

3.4. Example. Let S denote the trivial space consisting of one point. Let

$$X_1 = S, \quad A_1 = \left\{\frac{1}{m} : m = 1, 2, \ldots\right\} \cup \{0\},$$

$$X_2 = [0, 1], \quad A_2 = \{0, 1\} = X_3, \quad A_k = S = X_{k+1}, \ k = 3, 4, \ldots,$$

and let

$$l_3(x_{(3)}, a_{(3)}) = \begin{cases} 0 & \text{if } a_2 = x_3, \\ 1 & \text{if } a_2 \neq x_3, \end{cases} \qquad l_k \equiv 0 \quad \text{for } k \neq 3.$$

In standard statistical terminology, x_3 is the unknown parameter and the problem is one of testing the simple hypothesis, $x_3 = 0$, versus the simple alternative, $x_3 = 1$, based on the observation of $x_2 \in X_2$. The prior distribution of x_3 and the distribution of x_2, given x_3, are determined by the choice of a_1 as specified in the sequel.

For $y \in [0, 1]$, let $b_k(y)$ denote the k-th digit in the binary expansion of y. (Adopt any convention to define $b_k(y)$ uniquely when y has a terminating binary expansion.) Let λ denote the Lebesgue measure on $[0, 1]$ and let

$$\frac{dq_1(x_2|x, a)}{d\lambda} = \begin{cases} (2/3 - a_1) \cdot 2b_{a_1^{-1}}(x_2) + 1/3 + a_1 & \text{if } a_1 > 0, \\ 1 & \text{if } a_1 = 0, \end{cases}$$

$$q_2(\{0\}|x, a) = \begin{cases} \dfrac{2(2/3 - a_1)}{5/3 - a_1} & \text{if } b_{a_1^{-1}}(x_2) = 1, \ a_1 > 0, \\ 0 & \text{if } b_{a_1^{-1}}(x_2) = 0, \ a_1 > 0, \\ 2/5 & \text{if } a_1 = 0. \end{cases}$$

The choice of $\{\nu_t\}$ with $\nu_1 = \lambda$ and $\nu_2 = \nu^*$, where $\nu^*(\{0\}) = \nu^*(\{1\}) = 1/2$, yields

$$V_\mu^1 = \tau \times \lambda \times \nu^* \times \tau \times \tau \times \tau \times \ldots,$$

where τ denotes the unique probability measure on S. Then it can easily be checked that the map $a \to dF_\mu(\cdot|a)/dV_\mu^1(\cdot)$ is continuous as a map from A to $L_1(X, \mathscr{B}, V_\mu^1)$ with the *weak topology*. It is also true that the map $a \to q_1(\cdot|x, a)$ has a similar continuity property, but note that the map $a \to q_2(\cdot|x, a)$ does not.

Consider the sequence of policies $\{\delta^{(m)} : m = 1, 2, \ldots\}$ defined by

$$\delta_1^{(m)}(\{m^{-1}\}|x) = 1,$$

$$\delta_2^{(m)}(\{0\}|a) = \begin{cases} 1 & \text{if } b_m(x_2) = 1, \\ 0 & \text{if } b_m(x_2) = 0. \end{cases}$$

Then a simple computation yields $R(\mu, \delta^{(m)}) = 1/6 + 1/2m$. Hence $R(\mu) \leqslant 1/6$, but it can be readily seen that there is no policy which satisfies $R(\mu, \delta) \leqslant 1/6$. So there is no optimal policy.

4. EXISTENCE OF FULLY OPTIMAL POLICIES

Now we state the main theorem of the paper.

4.1. THEOREM. *Let the assumptions of Section 2 be satisfied*(1). *Then there exists a fully optimal policy.*

Proof. For $1 \leqslant k \leqslant n < \infty$ define $S_k^n(\cdot;\cdot,\cdot)$ inductively (backwards on k) by

$$S_n^n(a_n; x_{(n)}, a_{(n-1)}) = l_n(x_{(n)}, (a_{(n-1)}, a_n)),$$

$$(1)\quad S_{k-1}^n(a_{k-1}; x_{(k-1)}, a_{(k-2)}) = l_{k-1}(x_{(k-1)}, (a_{(k-2)}, a_{k-1})) + \int \{ \inf_{a_k \in A_k} S_k^n(a_k; (x_{(k-1)}, x_k), (a_{(k-2)}, a_{k-1}))\} \varphi_{k-1}(x_k|x, a)\, \nu_{k-1}(dx_k|x),$$

where $\int$ denotes the lower integral. Write $S_1^n(a_1; x_1)$ instead of $S_1^n(a_1; x_{(1)}, a_{(0)})$.

The following claims will now be proved for $1 \leqslant k \leqslant n < \infty$:

(A) $S_k^n(\cdot;\cdot,\cdot)$ *is* $(\mathscr{A}_k \times \mathscr{B}_{(k)} \times \mathscr{A}_{(k-1)})$*-measurable.*

(B) $S_k^n(\cdot; x_{(k)}, a_{(k-1)})$ *is lower semi-continuous on* A_k *for all* $x_{(k)}, a_{(k-1)}$.

(C) *There exists a measurable map* $h_{k,n} : X_{(k)} \times A_{(k-1)} \to A_k$ *such that*

$$(2)\quad \inf_{a_k \in A_k} S_k^n(a_k; x_{(k)}, a_{(k-1)}) = S_k^n(h_{k,n}(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}) = R_k^n(x_{(k)}; a_{(k-1)}).$$

(1) For the proof of this theorem the most important are the assumptions stated in Sections 2.3, 2.8, 2.12 and 3.3.

For $k = n$, claims (A) and (B) follow from Assumption 2.8. In view of the assumptions on A_t stated in Section 2.3, the existence of $h_{n,n}$ satisfying the first part of (2) follows from the selection theorem of Brown and Purves [5]. It is clear that

$$S_n^n\big(h_{n,n}(x_{(n)}, a_{(n-1)}); x_{(n)}, a_{(n-1)}\big) \leqslant R_n^n(x_{(n)}; a_{(n-1)}).$$

On the other hand,

$$R_n^n(\mu_{x_{(n)},a_{(n-1)}}, \delta) = S_n^n\big(h_{n,n}(x_{(n)}, a_{(n-1)}); x_{(n)}, a_{(n-1)}\big)$$

for the policy δ with

$$\delta_n(\cdot|x_{(n)}, a_{(n-1)}) = \mu_{h_{n,n}(x_{(n)},a_{(n-1)})}.$$

Hence the second equality of (2) is also satisfied for $k = n$.

Let $2 \leqslant m \leqslant n$. Suppose that (A), (B) and (C) are satisfied for $m \leqslant k \leqslant n$. Then

$$(3)\quad S_{m-1}^n(a_{m-1}; x_{(m-1)}, a_{(m-2)}) = l_{m-1}\big(x_{(m-1)}, (a_{(m-2)}, a_{m-1})\big) + \\ + \int R_m^n\big((x_{(m-1)}, x_m); (a_{(m-2)}, a_{m-1})\big)\, \varphi_{m-1}(x_m|x, a)\, \nu_{m-1}(dx_m|x).$$

By (A) and (C), $R_m^n(\cdot;\cdot)$ is $(\mathscr{B}_{(m)} \times \mathscr{A}_{(m-1)})$-measurable. Hence S_{m-1}^n satisfies (A) by standard Fubini-type theorems (see, e.g., [12], p. 74).

Let α_i be any convergent sequence in A_{m-1}, say $\alpha_i \to \alpha$. Let $\delta^{(i)}$ be a policy satisfying

(i) $\delta_k^{(i)}(\cdot|x, a) = \mu_{h_{k,n}}(\cdot)_{(x_{(k)},a_{(k-1)})}$ $(m \leqslant k \leqslant n)$,

(ii) $\delta_{m-1}^{(i)}(\{\alpha_i\}|\cdot,\cdot) \equiv 1$.

Then, by (2) and (3),

$$(4)\qquad S_{m-1}^n(\alpha_i; x_{(m-1)}, a_{(m-2)}) = R_{m-1}^n(\mu_{(x_{(m-1)},a_{(m-2)})}, \delta^{(i)}).$$

Let δ be any $\tau_{m-1}(\mu_{(x_{(m-1)},a_{(m-2)})})$ accumulation point of the sequence $\{\delta^{(i)}\}$ (δ exists by Theorem 2.14 (a)). Note that every accumulation point of $\{\delta^{(i)}\}$ is equivalent to δ relative to this topology. By (4) and Theorem 2.14 (b),

$$\begin{aligned} S_{m-1}^n(\alpha; x_{(m-1)}, a_{(m-2)}) &= R_{m-1}^n(\mu_{(x_{(m-1)},a_{(m-2)})}, \delta) \\ &\leqslant \liminf_{i\to\infty} R_{m-1}^n(\mu_{(x_{(m-1)},a_{(m-2)})}, \delta^{(i)}) \\ &= \liminf_{i\to\infty} S_{m-1}(\alpha_i; x_{(m-1)}, a_{(m-2)}). \end{aligned}$$

Hence S_{m-1}^n satisfies (B). The existence of $h_{m-1,n}$ satisfying the first equality of (2) then follows from the selection theorem of Brown and Purves [5]. The second equality of (2) then follows from the first equality and the definitions of S_{m-1}^n and R_{m-1}^n by the same reasoning as in the case $k = n$. This completes the proof of (A), (B) and (C) for $1 \leqslant k \leqslant n < \infty$.

Clearly, S_k^n is non-decreasing in n for fixed k. Let

$$S_k = \lim_{n\to\infty} S_k^n.$$

Then S_k satisfies (A) and (B). So there exist measurable functions h_k satisfying the first equality in (2) for S_k. By Lemma 3.2,

$$\begin{aligned} R_k\big((x_{(k)}, a_{(k-1)})\big) &= \lim_{n\to\infty} R_k^n(x_{(k)}, a_{(k-1)}) \\ &= \lim_{n\to\infty} \inf_{a_k\in A_k} S_k^n(a_k; x_{(k)}, a_{(k-1)}) \\ &= \inf_{a_k\in A_k} S_k(a_k; x_{(k)}, a_{(k-1)}) \end{aligned} \tag{5}$$

since A_k is compact and $\{S_k^n(\cdot\,; x_{(k)}, a_{(k-1)})\colon n = 1, 2, \ldots\}$ is a non-decreasing sequence of lower semi-continuous functions. Thus S_k and R_k also satisfy the second equality in (2). It follows from the above and from the definition (1) that, furthermore,

$$\lim_{n\to\infty} S_k^n\big(h_{k,n}(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}\big) = S_k\big(h_k(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}\big) \tag{6}$$

since A_k is compact and $\{S_k^n(\cdot\,; x_{(k)}, a_{(k-1)})\colon n = 1, 2, \ldots\}$ is a non-decreasing sequence of lower semi-continuous functions. So $\{S_k\}$ satisfies the recursion (1) for all $k \geqslant 2$, $(x_{(k-1)}, a_{(k-2)}) \in X_{(k-1)} \times A_{(k-2)}$.

Let δ^* be the policy defined by

$$\delta_k^*\big(\{h_k(x_{(k)}, a_{(k-1)})\} \mid x, a\big) = 1 \qquad (k \in T).$$

Then, by (1),

$$R_k^n(\mu, \delta^*) = \mathrm{E}_{\Delta_{\mu,\delta^*}^k}\Big(\sum_{t=k}^{n} l_t(x_{(t)}, a_{(t)})\Big) \leqslant S_k\big(h_k(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}\big)$$

$$(1 \leqslant k \leqslant n < \infty),$$

where $\mu = \mu_{(x_{(k)}, a_{(k-1)})}$. Hence

$$R_k\big(\mu_{(x_{(k)}, a_{(k-1)})}, \delta^*\big) \leqslant S_k\big(h_k(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}\big).$$

By Lemma 3.2 and formulas (5) and (6),

$$\begin{aligned} R_k(x_{(k)}; a_{(k-1)}) &= S_k\big(h_k(x_{(k)}, a_{(k-1)}); x_{(k)}, a_{(k-1)}\big) \\ &\geqslant R_k\big(\mu_{(x_{(k)}, a_{(k-1)})}, \delta^*\big) \geqslant R_k(x_{(k)}; a_{(k-1)}). \end{aligned}$$

It follows that $R_k\big(\mu_{(x_{(k)}, a_{(k-1)})}, \delta^*\big) = R_k(x_{(k)}; a_{(k-1)})$, and hence δ^* is optimal at stage k for all $k \in T$. This proves the theorem.

Note that the fully optimal policy δ^*, derived above, is non-randomized.

5. STATIONARY CASE

In this section we consider an important special case of the general problem. Assume that the state and action spaces, the law of motion and the cost structure are time invariant (stationary), that is,

$$X_t = X_1, \quad A_t = A_1,$$

$$q_t(B_{t+1} \mid x, a) = q_t(B_{t+1} \mid x_t, a_t) = q_{t-1}(B_t \mid x_{t-1}, a_{t-1}) \quad (t \geqslant 2)$$

if $B_{t+1} = B_t \in \mathscr{B}_1$, $x_t = x_{t-1}$ and $a_t = a_{t-1}$. Moreover, for some α, $0 < \alpha \leqslant 1$,

$$l_t(x_{(t)}, a_{(t)}) = \alpha^{t-1} l(x_t, a_t),$$

where $l(\cdot, \cdot)$ is a $(\mathscr{B}_1 \times \mathscr{A}_1)$-measurable non-negative function which is lower semi-continuous with respect to the action variable.

When $\alpha = 1$, this corresponds to the negative dynamic programming of Strauch [16], and when $\alpha < 1$ and $l(\cdot, \cdot)$ is bounded (not necessarily non-negative), this corresponds to the discounted dynamic programming of Blackwell [2]. One of their main concerns was to determine whether a stationary optimal policy necessarily exists, that is, whether there always exists a $\delta \in \mathscr{D}$ such that δ is optimal in $\mathscr{D}$ and

$$\delta_t(C \mid x, a) = \delta_t(C \mid x_t) = \delta_{t-1}(C \mid x_{t-1}) \quad (t \geqslant 2)$$

for all $C \in A_1$ and $x_t = x_{t-1} \in X_1$.

Blackwell [2] gave a counterexample showing that this is not the case. However, in his example, depending on the interpretation, either A_1 is not compact or $l(x, \cdot)$ is not lower semi-continuous. With our assumptions, which include compactness of A_1 (but see Remark 3.3), lower semi-continuity of $l(x, \cdot)$ and assumptions on q, we can show that a fully optimal stationary policy does exist.

The desired result follows almost immediately from Theorem 4.1. (It could also be proved independently of Theorem 4.1 by specializing the proof of that theorem to the stationary case.) As noted in the Introduction the following result is similar to the regularity condition (S) of Schäl [14]:

5.1. THEOREM. *Let the assumptions of Theorem* 4.1 *and also the stationarity assumptions above be satisfied. Then there exists a stationary fully optimal policy.*

Proof. Let δ^* be the fully optimal policy of Theorem 4.1. Note that δ^* need not be stationary. However, define δ^{**} by

$$\delta_k^{**}(B \mid x, a) = \delta_k^{**}(B \mid x_{k-1}) = \delta_1^*(B \mid x_{k-1})$$

for $B \in \mathscr{B}_k = \mathscr{B}_1$, $x_{k-1} \in X_1$. Then δ^{**} is stationary. It can readily be shown that $R(\cdot, \delta^{**}) = R(\cdot, \delta^*)$, thus implying that the stationary policy δ^{**} is fully optimal.

REFERENCES

[1] R. Bellman, *Dynamic programming*, Princeton University Press, 1957.
[2] D. Blackwell, *Discounted dynamic programming*, Ann. Math. Statist. 36 (1965), p. 226-235.
[3] – *Positive dynamic programming*, Proc. Fifth Berkeley Symp. Math. Stat. Prob. 1 (1967), p. 415-418.
[4] L. D. Brown, *Closure theorems for procedures in sequential-design processes*, p. 57-91 in: S. S. Gupta and D. S. Moore (editors), *Statistical decision theory and related topics, II*, Academic Press, 1977.
[5] – and R. Purves, *Measurable selections of extrema*, Ann. Statist. 1 (1973), p. 902-912.
[6] M. L. Eaton, *Complete class theorems derived from conditional complete class theorems*, ibidem 6 (1978), p. 820-827.
[7] D. A. Freedman, *The optimal reward operator in special classes of dynamic programming problems*, Ann. Probability 2 (1974), p. 942-949.
[8] N. Furukawa, *Markovian decision processes with compact action spaces*, Ann. Math. Statist. 43 (1972), p. 1612-1622.
[9] K. Hinderer, *Foundations of non-stationary dynamic programming with discrete time parameter*, Springer-Verlag, 1970.
[10] R. P. Kertz and D. C. Nachman, *Optimal plans for discrete-time non-stationary dynamic programming with general total reward function, I, II*, Tech. Reps., Georgia Institute of Technology, 1977.
[11] L. LeCam, *An extension of Wald's theory of statistical decision functions*, Ann. Math. Statist. 26 (1955), p. 69-81.
[12] J. Neveu, *Mathematical foundations of the calculus of probability*, Holden-Day, 1965.
[13] M. Schäl, *On dynamic programming: Compactness of the space of policies*, Stochastic Processes Appl. 3 (1975), p. 345-364.
[14] – *Conditions for optimality in dynamic programming and the limit of n-stage optimal policies to be optimal*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 32 (1975), p. 179-196.
[15] – *On dynamic programming and statistical decision theory*, Ann. Statist. 7 (1979), p. 432-445.
[16] R. E. Strauch, *Negative dynamic programming*, Ann. Math. Statist. 37 (1966), p. 871-890.

Department of Mathematics
Cornell University
Ithaca, New York 14853
White Hall, U.S.A.

Received on 1. 12. 1979