Probab. Th. Rel. Fields 78, 583–602 (1988)

Probability Theory and Related Fields

© Springer-Verlag 1988

A Convex Analytic Approach to Markov Decision Processes

Vivek S. Borkar[*,**]

Systems Research Center, University of Maryland, College Park, MD 20742, USA

Summary. This paper develops a new framework for the study of Markov decision processes in which the control problem is viewed as an optimization problem on the set of canonically induced measures on the trajectory space of the joint state and control process. This set is shown to be compact convex. One then associates with each of the usual cost criteria (infinite horizon discounted cost, finite horizon, control up to an exit time) a naturally defined occupation measure such that the cost is an integral of some function with respect to this measure. These measures are shown to form a compact convex set whose extreme points are characterized. Classical results about existence of optimal strategies are recovered from this and several applications to multicriteria and constrained optimization problems are briefly indicated.

1. Introduction

The study of Markov decision processes on a countable state space (equivalently, controlled Markov chains) usually proceeds from the dynamic programming heuristic [7]. The aim of this paper is to provide an alternative framework. The control problem is viewed here as an optimization problem on the set of canonically induced probability measures on the trajectory space by the joint state and control process. This set is shown to be compact and convex. Next one associates with each of the usual cost criteria (infinite horizon discounted cost control, finite horizon control, control up to an exit time) a naturally defined concept of an occupation measure so that the cost is the integral of some function with respect to this measure. The set of these occupation measures is then shown to be compact convex and its extreme points are characterized. This way one recovers all the classical existence theorems for optimal strategies from a different vantage point, uncovering in the process much structure that is not transparent in the conventional approaches. The latter has important implications in multiobjective and constrained control problems as will be argued in the final section of this paper.

[*] Research supported by NSF Grant CDR-85-00108

[**] *Permanent address*: Tata Institute of Fundamental Research, Bangalore Center, P.O. Box 1234, I.I. Sc. Campus, Bangalore 560012, India

The notation we use is that of [3], [4]. Let $X_n, n=1,2,\ldots$, be a controlled Markov chain on state space $S=[1,2\ldots]$ with transition matrix $P_u=[[p(i,j,u_i)]], i,j\in S$, indexed by the control vector $u=[u_1,u_2,\ldots]$. Here, $u_i\in D(i), i\in S$, for some prescribed compact metric spaces $D(i)$. The functions $p(i,j,\cdot)$ are assumed to be continous. By replacing each $D(i)$ by $\Pi D(k)$ and $p(i,j,\cdot)$ by its composition with the projection $\Pi D(k)\rightarrow D(i)$, one may assume that all $D(i)$'s are replicas of the same compact metric space D. We do so and let L denote the countable product of copies of D.

For any Polish space Y, denote by $M(Y)$ the space of probability measures on Y with the topology of weak convergence. For $n=1,2,\ldots,\infty$, Y^n will denote the n-times product of Y with itself.

A control strategy (CS) is a sequence $\{\xi_n\}$, $\xi_n=[\xi_n(1),\xi_n(2),\ldots]$ of L-valued random variables such that for $i\in S, n\geqq 1$,

$$P(X_{n+1}=i/X_m,\xi_m,m\leqq n)=p(X_n,i,\xi_n(X_n)) \tag{1.1}$$

We say that the controlled chain $\{X_n\}$ is governed by the CS $\{\xi_n\}$ wherever (1.1) holds. If for each n, ξ_n is independent of $X_m, m\leqq n$, and $\xi_m, m<n$, we call $\{\xi_n\}$ a Markov randomized strategy (MRS). If in addition $\xi_n, n\geqq 1$, are identically distributed, call it a stationary randomized strategy (SRS). An MRS for which the law of each ξ_n is a Dirac measure will be called a Markov strategy (MS). Similarly, an SRS for which the law of each ξ_n is a Dirac measure will be called a stationary strategy (SS). The motivation for this nomenclature is self-evident.

If the common law for $\xi_n, n\geqq 1$, of an SRS $\{\xi_n\}$ is $\Phi\in M(L)$, we denote it by $\gamma[\Phi]$. In view of (1.1), it is clear that as long as we are interested only in the law of the $S\times D$-valued process $\{(X_n,\xi_n(X_n))\}$, $n\geqq 1$, under an SRS $\gamma[\Phi]$, we may assume Φ to be a product measure on L. Let $\Phi_i, i\in S$, denote the image of Φ under the projection from L onto its i-th factor space. (Thus $\Phi=\Pi\Phi_i$ in view of the preceding comment). Under $\gamma[\Phi]$, $\{X_n\}$ will be a Markov chain with stationary transition probabilities given by the transition matrix $P[\Phi]=[[\int p(i,j,u)\,\Phi_i(du)]]$. If $\gamma[\Phi]$ is an SS with $\Phi=$ the Dirac measure at $\xi\in L$, denote it by $\gamma\{\xi\}$ and the corresponding transition matrix by $P\{\xi\}=P_\xi$.

Throughout this paper, we assume that the chain has a single communicating class under all SRS. This is a convenient assumption to have in the background, but can be relaxed to a varying extent for much of what follows, being completely unnecessary in some cases.

Let h: $S\rightarrow R^+$, k: $S\times D\rightarrow R^+$, l: $\mathbb{N}\times S\times D\rightarrow R^+$, be continuous functions. The various cost criteria one typically seeks to minimize over all CS are the following:

$$\text{(C1)}\qquad E\left[\sum_{n=1}^{\infty}\beta^n k(X_n,\xi_n(X_n))\right], \tag{1.2}$$

This is the 'discounted cost control problem'.

$$\text{(C2)}\qquad E\left[\sum_{n=1}^{N-1} l(n,X_n,\xi_n(X_n))+h(X_N)\right],\quad 1\leqq N<\infty \tag{1.3}$$

This is the 'finite horizon control problem'.

$$\text{(C3)} \qquad E\left[\sum_{n=1}^{\tau-1} k(X_n, \xi_n(X_n)) + h(X_\tau)\right], \tag{1.4}$$

where for some prescribed finite subset A of S,

$$\tau = \min\{n \geqq 1 \,|\, X_n \notin A\} \ (=\infty \text{ if the set on the right is empty}).$$

This is called the 'control up to a first exit time'.

$$\text{(C4)} \qquad \limsup_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} k(X_m, \xi_m(X_m)). \tag{1.5}$$

This is the 'long run average cost control problem'.

Of course, one assumes that the cost functional under consideration is finite under some CS, the problem being vacuous otherwise. In this paper, we shall concern ourselves with (C1)-(C3) only. (C4) has been adequately treated in the companion paper [4] (see also [3].).

The organization of this paper is as follows: Sect. II establishes the compactness and convexity of the laws of $\{(X_n, \xi_n(X_n))\}$ as subsets of $M((S \times D)^\infty)$ under various classes of CS. Sect. III–V study the control problems corresponding to (C1)-(C3) in that order. They follow a standard pattern. First one associates an appropriate notion of an 'occupation measure' for the joint state and control process. The attainable set of these measures under all CS is then shown to remain the same if one restricts attention to SRS in the first and the third case and MRS in the second. Furthermore, this set is shown to be compact convex using the results of Sect. II and its extreme points are shown to correspond to SS in the first and the third case and MS in the second. The choice of these measures is such that the corresponding cost can be written as an integral with respect to these measures. In view of the foregoing, this leads to the appropriate existence results for an optimal SS (or MS as the case may be) in each set-up. (Recall that a CS is optimal if the corresponding cost is the minimum cost over all CS.) Sect. VI discusses several potential applications of the foregoing to problems arising in multicriteria and constrained optimization of Markov decision processes.

Given the vast extent of the existing literature on this subject, it is impossible to give a decent summary of it in the short span of this introduction. We shall content ourselves with referring to the excellent texts [2], [7] as general pointers in that direction.

2. Compactness and Convexity of Attainable Laws

Let $A_C, A_{MR}, A_{SR}, A_M, A_S$ denote the sets of attainable laws of $[(X_1, \xi_1(X_1)), (X_2, \xi_2(X_2)), \ldots]$ viewed as subsets of $M((S \times D)^\infty)$ as the control strategy varies over all CS, all MRS, all SRS, all MS and all SS respectively, the initial law being held fixed. For simplicity, we take the latter to be the point mass concentrated on $1 \in S$. In this section, we show that the above sets are compact and A_C is convex.

For $n \geqq 1$ and any CS $\{\xi_m\}$, denote by $P^n(\{\xi_m\}, \cdot) \in M(S)$ the law of X_n under $\{\xi_m\}$, i.e., $P^n(\{\xi_m\}, j) = P(X_n = j)$, $j \varepsilon S$, when $\{X_m\}$ is governed by $\{\xi_m\}$.

Lemma 2.1. *For each $n \geqq 1$, the set $P^n(\{\xi_m\}, \cdot)$ as $\{\xi_m\}$ varies over all CS is tight in $M(S)$.*

Proof. We proceed by induction. The claim is trivial for $n = 1$. Suppose it holds for some $n \geqq 1$. Let $\varepsilon > 0$. Pick $N \geqq 1$ such that

$$P(1 \leqq X_i \leqq N, 1 \leqq i \leqq n) > 1 - \varepsilon/2$$

for all CS. This is possible by the induction hypothesis. For each $i \in S$ and $u_n \to u$ in D, we have $p(i, j, u_n) \to p(i, j, u)$ for all $j \in S$ and hence by Scheffe's theorem ([1], pp. 224), $p(i, \cdot, u_n) \to p(i, \cdot, u)$ in total variation and hence in $M(S)$. Thus $p(i, \cdot, u)$, $u \in D$, is tight in $M(S)$. Pick $N'(i)$ such that

$$\inf_u \sum_{j \leqq N'(i)} p(i, j, u) \geqq 1 - \varepsilon/2^i$$

Let $\bar{N} = \max\{N'(1), N'(2), \ldots, N'(N), N\}$. Then a straightforward computation using (1.1) shows that

$$P(1 \leqq X_i \leqq \bar{N}, 1 \leqq i \leqq n + 1) > 1 - \varepsilon$$

under all CS. The claim follows by induction. QED

Theorem 2.1. *A_C is compact in $M((S \times D)^\infty)$.*

Proof. Let $\{X_n^m\}, m \geqq 1$, be a sequence of controlled Markov chains governed by CS $\{\xi_n^m\}, m \geqq 1$, respectively, with $X_1^m = 1$ for all m. By the above lemma and compactness of L, the laws of $(X_n^m, \xi_n^m), m \geqq 1$, are tight in $M(S \times L)$ for each fixed n. Hence for $\xi^m = [\xi_1^m, \xi_2^m, \ldots]$, $X^m = [X_1^m, X_2^m, \ldots]$, the laws of (ξ^m, X^m), $m \geqq 1$, are tight in $M(L^\infty \times S^\infty)$ and therefore converge along a subsequence to the law of some $L^\infty \times S^\infty$-valued random variable (ξ^∞, X^∞). Restrict attention to this subsequence and denote it by $\{m\}$ again by abuse of notation. By Skorohod's theorem ([1], pp. 29), we may assume that $(\xi^m, X^m), m = 1, 2, \ldots, \infty$, are defined on a common probability space and $(\xi^m, X^m) \to (\xi^\infty, X^\infty)$ a.s. in $L^\infty \times S^\infty$. Let $\xi^\infty = [\xi_1^\infty, \xi_2^\infty, \ldots]$, $X^\infty = [X_1^\infty, X_2^\infty, \ldots]$. Let $n \geqq 1$. Let $f\colon S \to R$ be a function of finite support and $g\colon (S \times L)^n \to R$ a bounded continuous function. Then the function $F\colon (S \times L)^{n+1} \to R$ defined by

$$F((x_1, u_1), \ldots, (x_{n+1}, u_{n+1})) = (f(x_{n+1}) - \sum_{j \in S} p(x_n, j, u_n(x_n)) f(j))$$
$$g((x_1, u_1), \ldots, (x_n, u_n)) \tag{2.1}$$

is seen to be bounded continuous. For $m = 1, 2, \ldots,$

$$E[(f(X_{n+1}^m) - \sum_{j \in S} p(X_n^m, j, \xi_n^m(X_n)) f(j))\, g((X_1^m, \xi_1^m), \ldots, (X_n^m, \xi_n^m))] = 0. \tag{2.2}$$

Letting $m \to \infty$, the continuity of (2.1) implies that (2.2) holds for $m = \infty$. A standard monotone class argument establishes (1.1) with $\{X_n^\infty\}$, $\{\xi_n^\infty\}$ replacing $\{X_n\}$, $\{\xi_n\}$. The claim follows. QED

Corollary 2.1. *A_{MR}, A_{SR}, A_M, A_S are compact in $M((S \times D)^\infty)$.*

Proof. In the above proof, note that if for $m = 1, 2, \ldots, \xi_n^m$ is independent of $\xi_i^m, i < n, X_i^m, i \leqq n$, then ξ_n^∞ will be independent of $\xi_i^\infty, i < n, X_i^\infty, i \leqq n$, for $n = 1, 2, \ldots$. This is because independence is preserved under convergence in law. Similarly, if $\xi_1^m, \xi_2^m, \ldots$, are identically distributed for $m = 1, 2, \ldots$, then $\xi_1^\infty, \xi_2^\infty, \ldots$, will be identically distributed. The statement for A_{MR}, A_{SR} follows. The statement for A_M, A_S follows from the further observation that a limit of Dirac measures in $M(L)$ is again a Dirac measure. QED

Remark. Note that the above proofs in fact give the stronger claims that the attainable laws of $[(X_1, \xi_1), (X_2, \xi_2), \ldots]$ as measures in $M((S \times L)^\infty)$ are compact as the CS varies over all CS, MRS, SRS, MS or SS respectively. However, the above weaker version suffices for our purposes.

Corollary 2.2. *The law of $[X_1, X_2, \ldots]$ under an SRS $\gamma[\Phi]$ (resp., an SS $\gamma\{\xi\}$) is a continuous function of Φ (resp., ξ) when viewed as a map from the subset of $M(L)$ consisting of product measures on L to $M(S^\infty)$ (resp. L to $M(S^\infty)$).*

This is immediate in view of the foregoing. We shall establish one other corollary in anticipation of its later use in connection with the cost functional (C3). Let $A \subset S$ be a prescribed finite set and $\xi^m, X^m, m = 1, 2, \ldots, \infty$, be as above. Define $\tau^m = \min\{n \geqq 1 \mid X_n^m \notin A\}$, with $\tau^m = \infty$ when $X_n \in A$ for all n.

Corollary 2.3. $\tau^m \to \tau^\infty$ a.s. in $[1, 2, \ldots, \infty]$.

Proof. Outside a set of zero probability, $X_n^m \to X_n^\infty$ a.s. for each n. Since these are discrete valued, $X_n^m = X_n^\infty$ from some n onwards depending on the sample point. The claim follows quite easily from this. QED

Theorem 2.2. *A_C is convex.*

Proof. Let $\{X_n\}, \{Y_n\}$ be controlled Markov chains governed by the CS $\{\xi_n\}, \{\phi_n\}$ respectively with $X_1 = Y_1 = 1$. For $n \geqq 1$, let $Q_n^1, Q_n^2 \in M(S^n \times D^{n-1})$ (with $S^1 \times D^0 = S$ by convention) denote the laws of $[X_1, \ldots, X_n, \xi_1(X_1), \ldots, \xi_{n-1}(X_{n-1})]$, $[Y_1, \ldots, Y_n, \phi_1(Y_1), \ldots, \phi_{n-1}(Y_{n-1})]$ respectively ($[X_1]$, $[Y_1]$ resp. when $n = 1$). Let $\alpha_1, \alpha_2 \in [0, 1]$ with $\alpha_1 + \alpha_2 = 1$. Let $Q_n = \alpha_1 Q_n^1 + \alpha_2 Q_n^2, n \geqq 1$. We shall show that for each n, Q_n is the law of $[Z_1, \ldots Z_n, \psi_1(Z_1), \ldots, \psi_{n-1}(Z_{n-1})]$ for some controlled Markov chain $\{Z_n\}$ governed by a CS $\{\psi_n\}$. This will imply the statement of the theorem by virtue of the Kolmogorov extension theorem. We proceed by induction. The claim is trivial for $n = 1$. Suppose it is true for some $n \geqq 1$. Let $s = [x_1, \ldots, x_n, y_1, \ldots, y_{n-1}] \in S^n \times D^{n-1} (= [x_1]$ for $n = 1)$ and $\bar{s} = [x_1, \ldots, x_{n+1}, y_1, \ldots y_n] \in S^{n+1} \times D^n$ denote typical elements of the respective spaces (to be used as variables of integration). For $i = 1, 2$, let $s \to \eta^i(s, \cdot)$: $S^n \times D^{n-1} \to M(D)$ denote any one representative of the regular conditional law of $\xi_n(X_n)$ (resp., $\phi_n(Y_n)$) given $[X_1, \ldots, X_n, \xi_1(X_1), \ldots, \xi_{n-1}(X_{n-1})]$ (resp., $[Y_1, \ldots, Y_n, \phi_1(Y_1), \ldots, \phi_{n-1}(Y_{n-1})]$), defined a.s. uniquely with respect to the law of the latter, with obvious modifications for $n = 1$. For simplicity, take $n \geqq 2$

in the following, the necessary modifications for $n=1$ being obvious. Let $A\subset S^n$, $A'\subset S$, $B\subset D^{n-1}$, $B'\subset D$ be measurable sets. Then

$$\begin{aligned} Q_{n+1}(A\times A'\times B\times B') &= \sum_{i=1}^{2} \alpha_i\, Q^i_{n+1}(A\times A'\times B\times B') \\ &= \sum_{i=1}^{2} \alpha_i \sum_{x_{n+1}\subset A'} \int_{A\times B} Q^i_n(ds) \int_{B'} p(x_n, x_{n+1}, y_n)\eta^i(s, dy_n). \end{aligned}$$

Define a measure $\bar{Q}\in M(S^n\times D^n)$ by

$$\bar{Q}(A\times B\times B') = \sum_{i=1}^{2} \alpha_i \int_{A\times B} Q^i_n(ds) \int_{B'} \eta_i(s, dy_n)$$

Note that the image of $\bar{Q}$ under the projection $S^n\times D^n\to S^n\times D^{n-1}$ is given by

$$\sum_{i=1}^{2} \alpha_i\, Q^i_n = Q_n.$$

Thus $\bar{Q}$ can be disintegrated as

$$\bar{Q}(ds, dy_n) = Q_n(ds)\eta(s, dy_n)$$

where $s\to\eta(s,\cdot)$: $S^n\times D^{n-1}\to M(D)$ is any representative of the appropriate regular conditional law, defined Q_n-a.s. By induction hypothesis, Q_n is the law of $[Z_1, \ldots, Z_n, \psi_1(Z_1), \ldots, \psi_{n-1}(Z_{n-1})]$ for some controlled Markov chain $Z_i, i\leqq n$, governed by a CS $\psi_i, \mathrm{i}\leqq n-1$. By enlarging the underlying probability space of these processes if necessary (e.g. by attaching to it a copy of D), construct on it a D-valued random variable $\psi_n(Z_n)$ such that the regular conditional law of $\psi_n(Z_n)$ given $Z_1, \ldots, Z_n, \psi_1(Z_1), \ldots, \psi_{n-1}(Z_{n-1})$ is $\eta([Z_1, \ldots, Z_n, \psi_1(Z_1), \ldots, \psi_{n-1}(Z_{n-1})], \cdot)$. By a further enlargement of this probability space (e.g. by attaching to it a copy of S), construct on it an S-valued random variable Z_{n+1} such that the regular conditional law of Z_{n+1} given $Z_1, \ldots, Z_n, \psi_1(Z_1), \ldots, \psi_n(Z_n)$ is $p(Z_n, \cdot, \psi_n(Z_n))$. By construction, the law of $[Z_1, \ldots, Z_{n+1}, \psi_1(Z_1), \ldots, \psi_n(Z_n)]$ is Q_{n+1}, completing the induction step. The claim follows. QED

Remarks. We have not bothered here about the components of ψ_n other than $\psi_n(Z_n)$, $n\geqq 1$. These can be easily accommodated e.g. by using the trick we used in the beginning of the preceding section to replace individual $D(i)$'s by a common space D.

3. The Discounted Cost Control Problem

Let $\{X_n\}$ be a controlled Markov chain governed by a CS $\{\xi_n\}$. Associate with it the discounted occupation measure $\nu\in M(S\times D)$ by

$$\int f\, d\nu = (\beta^{-1}-1)\, E\Big[\sum_n \beta^n f(X_n, \xi_n(X_n))\Big]$$

for all bounded continuous f: $S \times D \to R$, β being as in (C1). Let B, B_{SR}, B_S denote respectively the sets of attainable v's as the control strategy varies over all CS, SRS and SS, the initial law being held fixed. The main result of this section is that $B = B_{SR}$ and is a compact convex set with its extreme points lying in B_S, which itself is compact.

Let $v' \in M(S)$ be the image of v under the projection $S \times D \to S$ and disintegrate v as

$$\int f dv = \sum_{i \in S} v'(\{i\}) \int_D f(i, u)\, \Phi_i(du),$$

f as above, where the map $i \to \Phi_i$: $S \to M(D)$ is any representative of the appropriate regular conditional law. The suggestive notation is intentional: We associate with v an SRS $\gamma[\Phi]$ where Φ is the product measure $\Pi_i \Phi_i$.

Lemma 3.1. *v remains unchanged if we use $\gamma[\Phi]$ instead of $\{\xi_n\}$ as the control strategy, the initial law being held fixed.*

Proof. Let $\{X'_n\}$ be a controlled Markov chain governed by $\gamma[\Phi]$ (corresponding to, say, $\{\xi'_n\}$) with the same initial law as $\{X_n\}$. Let f: $S \times D \to R$ be bounded continuous and define g: $s \to R$ by

$$g(i) = E[\sum_n \beta^n f(X'_n, \xi'_n(X'_n))/X'_1 = i], \qquad i \in S.$$

Then g is bounded and satisfies

$$g(i) = \beta \int f(i, u)\, \Phi_i(du) + \beta \sum_{j \in S} g(j) \int p(i, j, u)\, \Phi_i(du). \tag{3.1}$$

Define

$$Z_1 = g(X_1)$$

$$Z_n = \sum_{m=1}^{n-1} \beta^m f(X_m, \xi_m(X_m)) + \beta^{n-1} g(X_n), \qquad n \geqq 2$$

$$W_n = Z_{n+1} - Z_n$$

$$= \beta^n f(X_n, \xi_n(X_n)) + \beta^n g(X_{n+1}) - \beta^{n-1} g(X_n), \qquad n \geqq 1.$$

Then

$$E\left[\sum_{m=1}^{n} \beta^m f(X_m, \xi_m(X_m))\right] - E[g(X_1)] = E\left[\sum_{m=1}^{n} W_m\right] - \beta^n E[g(X_{n+1})] \tag{3.2}$$

Letting $\mathscr{F}_n = \sigma(X_m, \xi_m, m \leqq n)$, $n \geqq 1$, the sequence

$$\sum_{m=1}^{n} (W_m - E[W_m/\mathscr{F}_m])$$

is a zero mean $\{\mathscr{F}_{n+1}\}$-martingale with bounded increments. Thus

$$\begin{aligned} E\left[\sum_{m=1}^{n} W_m\right] &= E\left[\sum_{m=1}^{n} E[W_m/\mathscr{F}_m]\right] \\ &= E\left[\sum_{m=1}^{n} \beta^m (f(X_m, \xi_m(X_m)) + \sum_{j \in S} p(X_m, j, \xi_m(X_m))\, g(j) - \beta^{-1} g(X_m))\right] \end{aligned} \tag{3.3}$$

Substitute (3.3) in (3.2) and let $n \to \infty$. By the dominated convergence theorem, we get

$$\begin{aligned} & E\left[\sum_{m=1}^{\infty} \beta^m f(X_m, \xi_m(X_m))\right] - E[g(X_1)] \\ &= E\left[\sum_{m=1}^{\infty} \beta^m (f(X_m, \xi_m(X_m)) + \sum_{j \in S} p(X_m, j, \xi_m(X_m))\, g(j) - \beta^{-1} g(X_m))\right] \\ &= E\left[\sum_{m=1}^{\infty} \beta_m (\int f(X_m, u)\, \Phi_{X_m}(du) + \sum_{j \in S} g(j) \int p(X_m, j, u)\, \Phi_{X_m}(du) - \beta^{-1} g(X_m))\right] \\ &\quad \text{(by our construction of } \Phi) \\ &= 0 \end{aligned}$$

by virtue of (3.1). Thus

$$\begin{aligned} E\left[\sum_{m=1}^{\infty} \beta^m f(X_m, \xi_m(X_m))\right] &= E(g(X_1)] \\ &= E[g(X'_1)] \\ &= E\left[\sum_{m=1}^{\infty} \beta^m f(X'_n, \xi'_n(X'_n))\right] \end{aligned}$$

The claim follows. QED

Theorem 3.1. *$\mathscr{B}_{SR} = \mathscr{B}$ and is compact convex.*

Proof. Follows immediately from Lemma 3.1 and Theorems 2.1, 2.2. QED

Theorem 3.2. *$\mathscr{B}_S$ is compact and the extreme points of $\mathscr{B}_{SR}$ lie in $\mathscr{B}_S$.*

Proof. The first claim follows from Corollary 2.1. Let $\gamma[\Phi]$ be an SRS with $\Phi=\Pi_i\Phi_i$ such that for some $k\in S$, the measure $\int p(k,\cdot,u)\,\Phi_k(du)\in M(S)$ is not an extreme point of the convex set $\{\int p(k,\cdot,u)\,\mu(du),\ \mu\in M(D)\}\subset M(S)$. This means that there exist $\alpha\in(0,1)$, $u_1, u_2\in M(D)$ such that

$$\Phi_k=\alpha\mu_1+(1-\alpha)\mu_2.$$
$$\int p(k,j,u)\mu_1(du)\neq\int p(k,j,u)\mu_2(du)\qquad\text{for some } j\in S.$$

By relabelling S if necessary, take $k=1$. Define $\varphi,\psi\in M(L)$ by $\varphi=\mu_1\times\Pi_{i\geqq 2}\Phi_i$, $\psi=\mu_2\times\Pi_{i\geqq 2}\Phi_i$. Let $\eta\in M(S)$, identified with an infinite row vector $\eta=[\eta(\{1\}),\eta(\{2\}),\ldots]$. Let $P^n[\Phi], n\geqq 1$, denote the n-times matrix product of $P[\Phi]$ with itself. Take $m\geqq 1$ such that the first element of $\eta P^{m-1}[\Phi]$ $(=P(X_m=1)$ where $\{X_n\}$ is a controlled Markov chain governed by $\gamma[\Phi]$ with initial law $\eta)$ is strictly positive. By our assumption of a single communicating class, such an m exists. Let v_1, v_2, v_3 denote respectively the discounted occupation measures associated with the controlled Markov chain with initial law η and governed by (i) the SRS $\gamma[\Phi]$, (ii) the MRS $\{\xi_n\}$ where the law of ξ_n is Φ except for $n=m$, when it is φ, (iii) the same with ψ replacing φ. We shall show that

$$v_1=\alpha v_2+(1-\alpha)v_3 \tag{3.4}$$

Let $f\colon S\times D\to R$ be bounded continuous and define $f_\Phi\colon S\to R$ by

$$f_\Phi(i)=\int f(i,u)\,\Phi_i(du),\qquad i\in S.$$

Define $f_\varphi, f_\psi\colon S\to R$ analogously. Identify f_Φ with the infinite column vector $f_\Phi=[f_\Phi(1), f_\Phi(2),\ldots]^T$ and similarly for f_φ, f_ψ. Then

$$(\beta^{-1}-1)^{-1}\int f\,dv_1=\sum_{n-1}^{\infty}\beta^n\eta P^{n-1}[\Phi]f_\Phi$$
$$=\sum_{n=1}^{m-1}\beta^n\eta P^{n-1}[\Phi]f_\Phi+\beta^m\eta P^{m-1}[\Phi]f_\Phi+\sum_{n=m+1}^{\infty}\beta^n\eta P^{n-1}[\Phi]f_\Phi \tag{3.5}$$

with analogous expressions for $(\beta^{-1}-1)^{-1}\int f\,dv_i$, $i=2,3$. But

$$\eta P^{m-1}[\Phi]f_\Phi=\alpha\eta P^{m-1}[\Phi]f_\varphi+(1-\alpha)\eta P^{m-1}[\Phi]f_\psi \tag{3.6}$$

and for $n \geqq m$,

$$\eta P^n[\Phi] f_\Phi = \alpha \eta P^{m-1}[\Phi] P[\varphi] P^{n-m}[\Phi] f_\Phi$$
$$+ (1-\alpha) \eta P^{m-1}[\Phi] P[\psi] P^{n-m}[\Phi] f_\Phi. \tag{3.7}$$

Substituting (3.6), (3.7) in (3.5), (3.4) follows. Next we show that $v_2 \neq v_3$. It suffices to show that $v'_2 \neq v'_3$ where for $i=2,3$, v'_i is the image of v under the projection $S \times D \to S$. Note that the space l^∞ of bounded $f\colon S \to R$ (with supremum norm) separates points of $M(S)$ in the sense that $\eta_1 = \eta_2$ in $M(S)$ if and only if $\int f d\eta_1 = \int f d\eta_2$ for all $f \in l^\infty$. Note also that $\sum_{n=0}^{\infty} \beta^n P^n[\Phi]$ as a map from l^∞ to l^∞ is invertible with inverse $I - \beta P[\Phi]$, I being the infinite identity matrix. In view of these and the explicit expansions for $\int f dv'_i$, $i=2,3$, along the lines of (3.5), it suffices to check that

$$\eta P^{m-1}[\Phi] P[\varphi] f \neq \eta P^{m-1}[\Phi] P[\psi] f$$

for some $f \in l^\infty$, written as a column vector $[f(1), f(2), \ldots]^T$. Since φ, ψ coincide in all their factors except the first and since the first component of $\eta P^{m-1}[\Phi]$ is strictly positive by our choice of m, this reduces to

$$\sum_{j \in S} f(j) \int p(1,j,u) \mu_1(du) \neq \sum_{j \in S} f(j) \int p(1,j,u) \mu_2(du)$$

for some $f \in l^\infty$, which is certainly true. Thus $v_2 \neq v_3$ and hence v_1 cannot be an extreme point of $\mathscr{B} = \mathscr{B}_{SR}$. It follows that if the discounted occupation measure for some SRS $\gamma[\Phi]$ with $\Phi = \Pi_i \Phi_i$ is an extreme point of $\mathscr{B}_{SR}$, then for each $i \in S$, the measure $\int p(i,\cdot,u) \Phi_i(du) \in M(S)$ is an extreme point of the convex set $\{\int p(i,\cdot,u)\mu(du), \mu \in m(D)\} \subset M(S)$. Since the set of extreme points of the latter set is contained in $\{p(i,\cdot,u), u \in D\}$, it follows that $P[\Phi] = P\{\xi\}$ for some $\xi \in L$. QED

Note that we have in fact the stronger claim that the extreme points of $\mathscr{B} = \mathscr{B}_{SR}$ correspond to $\gamma\{\xi\}, \xi = [\xi(1), \xi(2), \ldots]$, for which $p(i,\cdot,\xi(i))$ is an extreme point of $\{p(i,\cdot,u), u \in D\}$ for each $i \in S$. This is a necessary condition. It is not clear whether it is also sufficient, i.e., whether all $\gamma\{\xi\}$ satisfying the above extreme point property for transition probabilities lead to v which are extreme points of $\mathscr{B}$.

Theorem 3.3. *The control problem with* $(C1)$ *as the cost criterion has an optimal* SS *which is optimal for any initial law.*

Proof. (1.2) is of the form

$$(\beta^{-1}-1)^{-1}\int k\,dv=(\beta^{-1}-1)^{-1}\sup_{N}\int(k\wedge N)\,dv$$

and is thus a lower semicontinuous function of v. By Theorem 3.1, an optimal SRS $\gamma[\Phi]$ exists. Let $\bar{v}$ be the discounted occupation measure corresponding to $\gamma[\phi]$ and $\mathscr{B}_e$ the set of extreme points of $\mathscr{B}$. Then by Choquet's theorem [5],

$$\int k\,d\bar{v}=\int_{\mathscr{B}_e} d\mu(v)\left(\int k\,dv\right)$$

for some $\mu\in M(\mathscr{B}_e)$. Since $\int k\,dv\geqq\int k\,d\bar{v}$ for all $v\in\mathscr{B}$, it follws that $\int k\,d\bar{v}=\int k\,dv$ for some $v\in\mathscr{B}_e$. The claim now follows from Theorem 3.2 when the initial law is fixed. Let $\gamma\{\xi\}$ be an optimal SS for the initial law η. Suppose it is not optimal for some other initial law. Then it is easy to see that for some $i\in S$ and $\xi'\in L$, the chain starting at $X_1=i$ and governed by $\gamma\{\xi'\}$ has a strictly lower cost than the one governed by $\gamma\{\xi\}$ with the same initial law. Let $m\geqq 1$ be such that the i-th component of $\eta P^{m-1}\{\xi\}$ is strictly positive. Let $\{X_n\}$ be a controlled Markov chain with initial law η and governed by the CS $\{\xi_n\}$ defined by $\xi_n=$the Dirac measure at $\xi I\{n<m\}+\xi I\{n\geqq m, X_m\neq i\}+\xi' I\{n\geqq m, X_m=i\}$, $n\geqq 1$. Let $\{X'_n\}$ be a controlled Markov chain governed by $\gamma\{\xi\}$ and with initial law η. Then

$$E\left[\sum_{n=1}^{\infty}\beta^n k(X_n,\xi_n(X_n))\right]=E\left[\sum_{n=1}^{m-1}\beta^n k(X_n,\xi(X_n))\right]$$
$$+E\left[I\{X_m\neq i\}\left(\sum_{n=m}^{\infty}\beta^n k(X_n,\xi(X_n))\right)\right]+E\left[I\{X_m=i\}\right.$$
$$\left.\left(\sum_{n=m}^{\infty}\beta^n k(X_n,\xi'(X_n))\right)\right]$$

The first two terms on the right are unchanged if we replace $\{X_n\}$ by $\{X'_n\}$ and the third becomes strictly larger. Thus $\{\xi_n\}$ gives a strictly lower cost than $\gamma\{\xi\}$, a contradiction. The claim follows. QED

Remarks. It is not hard to recover the dynamic programming equations from the above.

4. The Finite Horizon Control Problem

In this and the next section, the development closely parallels that of the preceding section. To emphasize this analogy and economize on notation, we duplicate much of the notation of the preceding section here and again in the next section. We shall also omit details of the arguments used when they are routine modifications of those employed earlier.

Let $N \geqq 1$ be as in (C2). Let $B = \{1, 2, \ldots, N\}$. Let $\{X_n\}$ be a controlled Markov chain governed by the CS $\{\xi_n\}$. Define the 'finite horizon occupation measure' $v \in M(B \times S \times D)$ by

$$\int f \, dv = \frac{1}{N} \sum_{m=1}^{N} E[f(m, X_m, \xi_m(X_m))] \tag{4.1}$$

for bounded continuous f: $B \times S \times D \to R$. Let $\mathscr{B}, \mathscr{B}_{MR}, \mathscr{B}_M$ denote respectively the sets of attainable v's as the control strategy varies over all CS, all MRS, all MS, with the initial law held fixed. Let v' denote the image of v under the projection $B \times D \times S \to B \times D$ and disintegrate v as

$$\int f \, dv = \sum_{m=1}^{N} \sum_{i \in S} v'(\{(m, i)\}) \int f(m, i, u) \, \psi_{m,i}(du)$$

for bounded continuous f: $B \times S \times D \to R$, where the map $(m, i) \to \psi_{m,i}$: $B \times S \to M(D)$ is any representative of the regular conditional law (defined v'-a.s.). Let $\psi_m = \Pi_i \psi_{m,i} \in M(L)$ for $m \geqq 1$. Let $\{X'_n\}$ be a controlled Markov chain governed by an MRS $\{\xi'_n\}$ where the law of ξ'_n is ψ_n for each n, with the same initial law as $\{X_n\}$.

Lemma 4.1. *v defined by (4.1) remains unchanged if $\{X_n\}, \{\xi_n\}$ are replaced by $\{X'_n\}, \{\xi'_n\}$ respectively.*

Proof. Let f: $B \times S \times D \to R$ be bounded continuous. Define g: $B \times S \to R$ by

$$g(m, i) = E\left[\sum_{n=m}^{N} f(n, X'_n, \xi'_n(X'_n)) / X'_m = i\right], \qquad 1 \leqq m \leqq N.$$

Also, let $g(N+1, i) = 0$, $i \in S$. Then g is bounded and satisfies

$$g(m, i) = \int f(m, i, u) \psi_{m,i}(du) + \sum_{j \in S} g(m+1, j) \int p(i, j, u) \psi_{m,i}(du) \tag{4.2}$$

for $1 \leqq m \leqq N$, $i \in S$. Define

$$\begin{aligned} Z_1 &= g(1, X_1) \\ Z_n &= \sum_{m=1}^{n-1} f(m, X_m, \xi_m(X_m)) + g(n, X_n), \qquad 2 \leqq n \leqq N+1 \\ W_n &= Z_{n+1} - Z_n \\ &= f(n, X_n, \xi_n(X_n)) + g(n+1, X_{n+1}) - g(n, X_n), \qquad 1 \leqq n \leqq N \end{aligned}$$

Then

$$E\left[\sum_{n=1}^{N} f(n, X_n, \xi_n(X_n))\right] - E[g(1, X_1)] = E\left[\sum_{n=1}^{N} W_n\right]$$
$$= E\left[\sum_{n=1}^{N} E[W_n/\mathscr{F}_n]\right]$$

with $\{\mathscr{F}_n\}$ as before. This equals

$$E\left[\sum_{m=1}^{N} (f(m, X_m, \xi_m(X_m)) + \sum_{j\in S} p(X_m, j, \xi_m(X_m)) g(m+1, j) - g(m, X_m))\right]$$
$$= E\left[\sum_{m=1}^{N} (\int f(m, X_m, u)\psi_{m,X_m}(du) + \sum_{j\in S} g(m+1, j)\right.$$
$$\left. \cdot \int p(X_m, j, u)\psi_{m,X_m}(du) - g(m, X_m))\right]$$
$$\text{(by the definition of } \psi_{m,i})$$
$$= 0$$

by (4.2). Thus

$$E\left[\sum_{n=1}^{N} f(n, X_n, \xi_n(X_n))\right] = E(g(1, X_1)]$$
$$= E[g(1, X'_1)]$$
$$= E\left[\sum_{n=1}^{N} f(n, X'_n, \xi'_n(X'_n))\right].$$

The claim follows. QED

Theorem 4.1. *For a fixed initial law, $\mathscr{B} = \mathscr{B}_{MR}$ and is compact convex with its extreme points in $\mathscr{B}_M$ which itself is compact.*

Proof. This follows along the same lines as Theorems 3.1, 3.2 in view of the preceding lemma. The only significant change required is in the part of the proof of Theorem 3.2 where one proves $v_2 \neq v_3$. Instead of considering bounded $f\colon S \to R$ as there, consider now bounded $f\colon B \times S \to R$ and then further restrict them to those which vanish outside $\{m+1\} \times S$ for m as in that proof. The rest is easy. QED

Theorem 4.2. *The control problem with (C2) as the cost criterion has an optimal Markov strategy with the property that for any $m \in B$, the restriction of this strategy* *to the time interval* $\{n \mid m \leqq n \leqq N\}$ *is optimal for the control problem with the cost criterion*

$$E\left[\sum_{n=m}^{N-1} l(n, X_n, \xi_n(X_n)) + h(X_N)\right]$$

with arbitrary initial data.

Proof. Define $f\colon B \times S \times D \to R$ by

$$f(m, i, u) = l(m, i, u), \qquad 1 \leqq m < N$$
$$f(N, i, u) = h(i).$$

Then (1.3) equals $N \int f \, dv$. Now argue along the lines of Theorem 3.3. QED

Corollary 4.1. *For* $\{X_n\}, \{X'_n\}$ *as in Lemma 4.1, the laws of* X_n, X'_n *agree for each* n, $1 \leqq n \leqq N$.

Proof. Take $f(m, i, u) = \bar{f}(i)$ for $m = n$, 0 otherwise, $\bar{f}$ being an arbitrary bounded map $S \to R$. Then $E[\bar{f}(X_n)] = E[\bar{f}(X'_n)]$ by Lemma 4.1 and the claim follows. QED

Repeating this argument on successive time intervals $\{jN+1, \ldots, (j+1)N\}$, $j = 1, 2, \ldots$, it follows that for any controlled Markov chain governed by an arbitrary CS, there exists another controlled Markov chain governed by an MRS having the same one dimensional marginals.

5. Control Up to an Exit Time

Let A, τ be as in (C3). (See (1.4)). Before establishing the analogs of the results of Sect. III, IV for (C3), we shall first establish certain uniform moment bounds on τ. Note that without any loss of generality we may assume the initials law to be supported in A.

Lemma 5.1. *There exists an* $N \geqq 1$ *and* $\alpha \in (0, 1)$ *such that*

$$\sup P(\tau > N) < \alpha$$

where the supremum is over all CS and all initial laws supported in A.

Proof. Suppose not. Then there exists a sequence of controlled Markov chains $\{X_n^m\}$, $m = 1, 2, \ldots$, governed by CS $\{\xi_n^m\}$, $m = 1, 2, \ldots$, resp. with initial laws supported in A and satisfying: If $\tau^m = \min\{n \geqq 1 \mid X_n^m \notin A\}$ ($= \infty$ if $X_n^m \in A \forall n$), then

$$P(\tau^m > m) > 1 - \frac{1}{m}, \qquad m = 1, 2, \ldots$$

By dropping to a subsequence if necessary and invoking Skorohod's theorem as in the proof of Theorem 2.1, we may assume that these chains are defined on a common probability space and there exists a controlled Markov chain

$\{X_n^\infty\}$ governed by a CS $\{\xi_n^\infty\}$ with initial law supported in A such that $[X_1^m, X_2^m, \ldots, \xi_1^m, \xi_2^m, \ldots] \to [X_1^\infty, X_2^\infty, \ldots, \xi_1^\infty, \xi_2^\infty, \ldots]$ a.s. Since

$$\mathbb{P}(\tau^m > j) = E\left[\prod_{i=1}^{j} I\{X_i^m \in A\}\right], \qquad m, j = 1, 2, \ldots,$$

a straightforward limiting argument shows that $\tau^\infty = \min\{n \geqq 1 \mid X_n^\infty \notin A\}$ $(= \infty$ if $X_n^\infty \in A \forall n)$ satisfies

$$\mathbb{P}(\tau^\infty > m) > 1 - \frac{1}{m}, \qquad m = 1, 2, \ldots$$

This implies that $\tau^\infty = \infty$ a.s., i.e., $X_n^\infty \in A$ for all n, a.s. This is possible only if there exists a nonempty subset G of A such that for $i \in G, j \notin G$,

$$\inf_u p(i, j, u) = 0.$$

Given our hypotheses on $p(i, j, \cdot)$ and D, this infimum is a minimum attained at some point in D. But this means that one can construct an SS $\gamma\{\xi\}$ under which a chain starting in G never leaves G, contradicting our assumption of a single communicating class. The claim follows. QED

Corollary 5.1. *For* $n = 1, 2, \ldots,$

$$\sup E[(\tau)^n] < \infty,$$

where the supremum is overall CS *and all initial laws supported in* A.

Proof. Let $\{X_n\}$ be a controlled Markov chain governed by a CS $\{\xi_n\}$ with $X_1 \in A$ a.s. and let τ be its first exit time from A. Then for N, α as in Lemma 5.1,

$$\begin{aligned}
P(\tau > nN) &= E\left[\prod_{m=1}^{nN} I\{X_m \in A\}\right] \\
&= E\left[E\left[\prod_{m=(n-1)N+1}^{nN} I\{X_m \in A\} / \mathscr{F}_{(n-1)N}\right] \prod_{i=1}^{(n-1)N} I\{X_m \in A\}\right] \\
&= E[P(\tau > nN / \mathscr{F}_{(n-1)N}) I\{\tau > (n-1)N\}] \\
&\leqq \alpha P(\tau > (n-1)N)
\end{aligned}$$

where $\{\mathscr{F}_n\}$ are defined as before. Iterating,

$$P(\tau > nN) \leqq \alpha^n.$$

The claim follows easily from this. QED

In particular, $E[\tau]<\infty$. Thus given a controlled Markov chain $\{X_n\}$ governed by a CS $\{\xi_n\}$ with $X_1\in A$ a.s., we can associate with it the 'occupation measure up to the first exit from A', denoted $\nu\in M(A\times D)$, by

$$\int f\,d\nu=E\left[\sum_{n=1}^{\tau-1} f(X_n,\xi_n(X_n))\right]/(E[\tau]-1) \tag{5.1}$$

for bounded continuous $f\colon A\times D\to R$. Let $\nu'\in M(A)$ be the image of ν under the projection $A\times D\to A$ and disintegrate ν as

$$\int f\,d\nu=\sum_{i\in A}\nu'(i)\int f(i,u)\psi_i(du),$$

f being as above, where the map $i\to\psi_i\colon A\to M(D)$ is any representative of the regular conditional law (defined ν'-a.s.). Let $\Phi=\Pi_i\,\Phi_i\in M(L)$ with $\Phi_i=\psi_i$ for $i\in A$, arbitrary otherwise. Let $\{X'_n\}$ be a controlled Markov chain with the same initial law as $\{X_n\}$ but governed by the SRS $\gamma[\Phi]$, with $\{\xi'_n\}$ denoting the actual control sequence. Let $\tau'=\min\{n\geq 1\,|\,X'_n\notin A\}$ $(=\infty$ if $X'_n\in A$ for all $n)$.

Let $\bar M(A\times D)$ denote the space of finite nonnegative measures on $A\times D$ with the coarsest topology that makes the maps $\mu\in\bar M(A\times D)\to\int f\,d\mu\in R$ continuous for continuous $f\colon A\times D\to R$. For ν as in (5.2), define $\bar\nu\in\bar M(A\times D)$ by

$$\int f\,d\bar\nu=(E[\tau]-1)\int f\,d\nu \tag{5.2}$$

$$=E\left[\sum_{n=1}^{\tau-1} f(X_n,\xi_n(X_n))\right], \tag{5.3}$$

f being as before.

Lemma 5.2. *$\bar\nu$ defined by (5.3) is unchanged if $\{X_n\},\{\xi_n\},\tau$ are replaced by $\{X'_n\},\{\xi'_n\},\tau'$ respectively.*

Proof. Let $f\colon S\times D\to R$ be bounded continuous. Define $g\colon S\to R$ by

$$g(i)=E\left[\sum_{n=1}^{\tau'-1} f(X'_n,\xi'_n(X'_n))/X'_1=i\right],\qquad i\in A$$
$$=0,\qquad \text{otherwise}$$

Then for $i\in A$, $g(i)$ satisfies

$$g(i)=\int f(i,u)\Phi_i(du)+\sum_{j\in S} g(j)\int p(i,j,u)\Phi_i(du) \tag{5.4}$$

Define

$$Z_1=g(X_1)$$
$$Z_n=\sum_{m=1}^{n-1} f(X_m,\xi_m(X_m))+g(X_n)$$
$$W_n=Z_{n+1}-Z_n$$
$$=f(X_n,\xi_n(X_n))+g(X_{n+1})-g(X_n).$$

Then

$$E\left[\sum_{m=1}^{\tau-1} f(X_m, \xi_m(X_m))\right] - E[g(X_1)] = E\left[\sum_{m=1}^{\tau-1} W_m\right]$$

$$= E\left[\sum_{m=1}^{\tau-1} E[W_m/\mathscr{F}_m]\right]$$

by a straightforward application of the optional sampling theorem, since $\sum_{m=1}^{n} (W_m - E[W_m/\mathscr{F}_m])$, $n=1, 2, \ldots$, is an $\{\mathscr{F}_{n+1}\}$-martingale. The right hand side equals

$$E\left[\sum_{m=1}^{\tau-1} (f(X_m, \xi_m(X_m)) + \sum_{j\in S} g(j)p(X_m, j, \xi_m(X_m)) - g(X_m))\right]$$

$$E\left[\sum_{m=1}^{\tau-1} (\int f(X_m, u)\Phi_{X_m}(du) + \sum_{j\in S} g(j)\int p(X_m, j, u)\Phi_{X_m}(du) - g(X_m))\right]$$

(by the definition of Φ)

$$=0$$

by (5.4). Thus

$$E\left[\sum_{m=1}^{\tau-1} f(X_m, \xi_m(X_m))\right] = E[g(X_1)]$$

$$= E[g(X_1')]$$

$$= E\left[\sum_{m=1}^{\tau'-1} f(X_m', \xi_m'(X_m'))\right]$$

The claim follows. QED

Let $\mathscr{B}, \mathscr{B}_{SR}, \mathscr{B}_S$ denote the sets of attainable $\bar{v}$ as the control strategy varies over all CS, all SRS and all SS respectively, with the initial law being held fixed at some $\eta \in M(A)$.

Theorem 5.1. *$\mathscr{B} = \mathscr{B}_{SR}$ and is compact convex with its extreme points lying in $\mathscr{B}_S$ which itself is compact.*

Proof. Lemma 5.2 implies that $\mathscr{B} = \mathscr{B}_{SR}$. Convexity of $\mathscr{B}$ is immediate from Theorem 2.2. Let $\{X_n^m\}, \{\xi_n^m\}$, $m=1, 2, \ldots, \infty$, be as in the proof of Theorem 2.1 with the initial law now set equal to η. Define τ^m, $m=1, 2, \ldots, \infty$, correspondingly as in Corollary 2.3. Let f: $A \times D \to R$ be continuous. By Corollary 2.3,

$$\sum_{n=1}^{\tau^m - 1} f(X_n^m, \xi_n^m(X_n^m)) \to \sum_{n=1}^{\tau^\infty - 1} f(X_n^\infty, \xi_n^\infty(X_n^\infty)) \qquad \text{a.s.}$$

By Corollary (5.1), $\{\tau^m, m \geqq 1\}$ are uniformly integrable. Thus we can take expectations in the above to conclude that

$$E\left[\sum_{n=1}^{\tau^m - 1} f(X_n^m, \xi_n^m(X_n^m))\right] \to E\left[\sum_{n=1}^{\tau^\infty - 1} f(X_n^\infty, \xi_n^\infty(X_n^\infty))\right].$$

The compactness of $\mathscr{B}$ follows. That of $\mathscr{B}_S$ follows by the same additional observations as in the proof of Corollary 2.1. The proof that the extreme points of $\mathscr{B}_{SR}$ lie in $\mathscr{B}_S$ follows along the lines of the proof of Theorem 3.2 with some important modifications, which are given next. Let $P_1[\Phi]$ denote the matrix whose (i,j)-th element equals that of $P[\Phi]$ when $i, j \in \bar{A}$ and is zero otherwise, where $\bar{A} \subset A$ is the set $\{i \in A |$ the i-th component of $\eta P_1^m[\Phi]$ is >0 for some $m \geqq 1\}$ $(=\{i \in A | P(X_{m \wedge \tau} = i) > 0$ for some $m \geqq 1\}$ where $\{X_m\}$ is the chain governed by $\gamma[\Phi]$ with initial law η). Clearly, η is supported in $\bar{A}$. Note that the transition probabilities $p(j,\cdot,\cdot)$ for $j \notin \bar{A}$ may be changed arbitrarily without affecting v. In particular, they can be set equal to those corresponding to some SS. Thus we only need repeat the argument of Theorem 3.2 for $\bar{A}$, $P_1[\Phi]$ replacing S, $P[\Phi]$ respectively with $\beta = 1$ (which is okay because $P_1[\Phi]$ is a strictly substochastic matrix). The details are omitted. QED

Theorem 5.2. *For the control problem with cost criterion (C3), an optimal stationary strategy exists which is optimal for arbitrary initial data.*

Proof. Define f: $A \times D \to R$ by

$$f(i,u) = k(i,u) + \sum_{j \in S} p(i,j,u)\,h(j) - h(i) \tag{5.5}$$

Then

$$\sum_{m=1}^{n} [k(X_m, \xi_m(X_m)) + h(X_{m+1}) - h(X_m) - f(X_m, \xi_m(X_m))],$$

for $n = 1, 2, \ldots$, is an $\{\mathscr{F}_{n+1}\}$-martingale with zero mean and a simple application of the optional sampling theorem shows that (1.4) equals $\int f\,d\bar{v} - E[h(X_1)]$. The rest of the proof follows along the lines of that of Theorem 3.3 with a few minor modifications. QED

For $\{X_n\}$, $\{X_n'\}$, τ, τ' as in the proof of Lemma 5.2 and f as in (5.5) with k identically equal to zero, we have

$$E\left[\sum_{m=1}^{\tau-1} f(X_m, \xi_m(X_m))\right] = E\left[\sum_{m=1}^{\tau'-1} f(X_m', \xi_m'(X_m'))\right].$$

It is easy to see that this leads to

$$E[h(X_\tau)] = E[h(X_{\tau'}')]$$

implying that X_τ, $X_{\tau'}'$, have the same law. (Compare this with Corollary 4.1).

6. Applications

In this section, we shall briefly indicate some situations, several of them arising from multiobjective or constrained optimization problems, where the foregoing theory offers some immediate insight. We shall confine ourselves to the discounted cost set-up. The analogs thereof for the other two (or even mixed) situations will be self-evident. Let v denote the discounted occupation measure for some $\beta \in (0, 1)$, defined as in Sect. III.

Let k_i: $S \times D \to R$, $1 \leqq i \leqq n$, be bounded continuous and F: $R^n \to R$ continuous. Suppose we want to minimize $F(\int k d v, \ldots, \int k_n d v)$ over all CS. By Theorem 3.1, an optimal SRS exists. If we are able to show the extreme point property for this SRS by some means, an optimal SS will also exist.

A typical situation is $F(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$. A related criterion is $\|v - \mu\|$ where $\mu \in M(S \times D)$ is prescribed and $\|\cdot\|$ denotes the total variation norm (i.e., we want the occupation measure to approximate a prescribed distribution as closely as possible.). This can be rewritten as

$$\sup(\int f d\mu - \int f d v)$$

where the supremum is over all continuous f: $S \times D \to R$ satisfying $\sup_{i,u} |f(i, u)| = 1$. Since this is a lower semicontinuous function of v, an optimal SRS exists.

Another analogous situation arises as follows: Suppose several optimal SRS exist for the cost criterion (C1) (with, say, bounded k for sake of simplicity). One may want to pick from among those the SRS that minimizes the 'variance' of the cost given as

$$\int k^2 d v - (\int k d v)^2 \tag{6.1}$$

This fits the above framework with $n = 2$, $k_1 = k^2$, $k_2 = k$ and $F(x, y) = x - y^2$. The set of v corresponding to optimal SRS is easily seen to be compact. Thus an SRS that further minimizes (6.1) exists.

A somewhat different situation arises when for $\{k_i\}$ as above, one wants to minimize $\int k_1 \, d v$ with the constraints $\int k_i \, d v \in A_i$, $2 \leqq i \leqq n$, for some prescribed closed subsets $\{A_i\}$ of R. Again the existence of an optimal SRS follows from Theorem 3.1.

Suppose instead that we have a vector cost $[\int k_1 \, d v, \ldots, \int k_n \, d v]$ and we know its value for a finite collection of CS. Then any value lying in the closed convex hull of these will also be attainable for some SRS by virtue of Theorem 3.2. Such considerations may be useful in implementational schemes which start with a few educated guesses and use some recursive adaptation to zero in on the desired strategy. For a work in this spirit (albeit with a different cost criterion viz. (C4)), see [6] where a similar situation arises from a constrained optimization problem.

When the only optimal SRS available is not an SS, one may still want to approximate it by a convex combination of finitely many SS in a suitable sense for implementational ease. For example, one may want to approximate the v corresponding to the above SRS by $\sum_{i=1}^{n} \alpha_i v_i$ where $\alpha_i \in [0, 1]$, $1 \leqq i \leqq n$, with $\sum_{i=1}^{n} \alpha_i = 1$ and v_i are the discounted occupation measures corresponding to some SS $\gamma\{\xi_i\}$, $1 \leqq i \leqq n$, respectively. One may then either pick SS $\gamma\{\xi_i\}$ with probability α_i based on some random experiment performed beforehand or interlace the $\gamma\{\xi_i\}$'s along the time axis in a suitable manner ('time-multiplexing') to obtain the desired result. (Again, see [6] for a representative situation in connection with (C4).) Theorem 3.2 makes such approximations possible.

Theorem 3.1 also guarantees an optimal CS for general cost criteria of the type

$$E[F([X_1, X_2, \ldots, \xi_1(X_1), \xi_2(X_2), \ldots])] \tag{6.2}$$

for a lower semicontinuous F: $S^\infty \times D^\infty \to R$ such that (6.2) is finite for at least one CS.

Finally, the author would like to mention that the principal raison d'etre for this work is the hope that the techniques of convex analysis can be made to have a direct and fruitful bearing on the difficult problems of Markov decision processes such as multicriteria and constrained optimization. The results here are only a small beginning in this direction.

References

1. Billingsley, P.: Convergence of probability measures. New York: Wiley 1968
2. Bertsekas, D.P.: Dynamic Programming and stochastic control. New York: Academic 1976
3. Borkar, V.S.: On minimum cost per unit time control of Markov chains, SIAM J. Control Optimization **22**, 965–978 (1984)
4. Borkar, V.S.: Control of Markov chains with long-run average cost criterion. In: Fleming, W., Lions, P.L. (eds.) Stochastic differential systems, stochastic control theory and applications, IMA vol. 10, pp. 57–77. Berlin Heidelberg New York: Springer 1988
5. Phelps, R.: Lectures on Choquet's theorem. New York: Nostrand 1966
6. Makowski, A., Schwartz, A.: Implementation issues for Markov decision processes. In: Fleming, W., Lions, P.L. (eds.). Stochastic differential systems, stochastic control theory and applications, IMA vol. 10, pp. 323–337. Berlin Heidelberg New York: Springer 1988
7. Ross, S.: Introduction to stochastic dynamic programming. New York: Academic 1984

Received March 15, 1987; received in revised form January 8, 1988