MATHEMATICS OF OPERATIONS RESEARCH
Vol. 14, No. 1, February 1989
Printed in U.S.A.

VARIANCE-PENALIZED MARKOV DECISION PROCESSES*†

JERZY A. FILAR,‡ L. C. M. KALLENBERG§ AND HUEY-MIIN LEE**

We consider a Markov decision process with both the expected limiting average, and the discounted total return criteria, appropriately modified to include a penalty for the variability in the stream of rewards. In both cases we formulate appropriate nonlinear programs in the space of state-action frequencies (averaged, or discounted) whose optimal solutions are shown to be related to the optimal policies in the corresponding "variance-penalized MDP". The analysis of one of the discounted cases is facilitated by the introduction of a "Cartesian product of two independent MDPs".

1. Variance-penalized MPDs—Introduction. For a major portion of work in the area of infinite horizon Markov Decision Processes (MDPs, for short), standard objective function criteria have been the expected limiting average reward per unit time, or the expected discounted total reward over the time horizon.

Nonetheless it is conceivable that the use of expected total reward (averaged or discounted) may yield "optimal" policies which are unacceptable to a "risk-sensitive" decision-maker. Such a decision-maker may well prefer to use a criterion which incorporates a penalty for the "variability" induced by a given policy. The issue of how to construct such criteria in a manner that will be both conceptually meaningful, and mathematically tractable appears to be still open. Recent researches indicate that a number of reasonable formulations lead to formidable mathematical difficulties (see for instance, Sobel [15], Chung [3], Bouakiz [1], and White [18]).

The approach proposed in this paper is reminiscent of the earlier work of Markowitz [14] on portfolio analysis, where mean and variance of returns are used to formulate the problem. Furthermore, despite the inherent differences between the averaged reward and the discounted models we propose a conceptually unified approach towards the analysis of the corresponding variance-penalized models; namely via formulating these problems as appropriate nonlinear programs in the space of state-action frequencies (averaged, or discounted).

In particular, in §2 we demonstrate that our version of the variance-penalized average reward model always possesses a deterministic optimal policy which can be obtained from an appropriate optimal solution of a certain quadratic program. The derivation of these results depends essentially on the techniques for analyzing the average reward MDPs which were recently developed by Hordijk and Kallenberg [8] and [9] (versions of some of these were also discovered independently by White [18]). It should be mentioned at this point that Sobel [16] has recently studied a similar

*Received May 31, 1986; revised September 30, 1987.
AMS 1980 subject classification. Primary: 90C47.
IAOR 1973 subject classification. Main: Programming: Markov Decision.
OR/MS Index 1978 subject classification. Primary: 117 Dynamic programming/Markov.
Key words. Markov decision processes, role of variance, non-linear programs.
†This work was supported in part by the AFOSR and the NSF under the grants ECS-8503440 and ECS-8704954.
‡University of Maryland Baltimore County.
§University of Leiden.
**The Johns Hopkins University.

0364-765X/89/1401/0147$01.25

problem. Indeed, in the so-called unichain case Sobel's formulation is equivalent to ours; in the multichain case, however, the two approaches differ. Furthermore, we assume an arbitrary initial state distribution and work in the class "C_1" of policies (see p. 4) whereas Sobel places the more restrictive assumptions that the decision-maker can choose the initial state and works in the smaller class of stationary policies.

In §3 we begin the discussion of the discounted Markov decision problem which, surprisingly perhaps, is more difficult to analyze than the average award model in our context of the ensuing variance-penalized problem. The underlying issue of what is meant by a "variance due to a policy" is challenging in this case, as some rather natural measures turn out to be difficult to analyze. We introduce three alternative such measures and their corresponding interpretations.

In §4 we demonstrate that for one of the above measures; the "discount normalized variance" the most complete analysis analogous to that of §2 can be easily given. In §5 we analyze our variance-penalized problem by using the "stagewise variance" measure of variability. While this case is more difficult mathematically than that of §4, its analysis is made possible by a technical device of a "product MDP", which in a special sense is a Cartesian product of two "independent" MDPs. We demonstrate that even though some standard properties of ordinary MDPs fail, our variance-penalized problem retains a number of desirable mathematical properties. In particular, we demonstrate the existence of optimal Markov policies, and the equivalence with a "constrained" ordinary MDP. Also, we show that in the class of stationary policies the problem is equivalent to solving a nonlinear program. For the measure of variability that we call the "variance of the present value" we do not propose a method of analysis. This measure was discussed by Sobel [15] and Bouakiz [1]; however, the analysis of these authors suggests that the corresponding variance-penalized problem is not tractable by standard techniques.

We conclude this Introduction by mentioning that, in principle at least, the risk-sensitive MDPs could be studied in terms of appropriate utility functions over the possible histories of the process. Indeed a number of approaches in this general spirit have been proposed in the literature (see for instance Jacquette [10], [11], and Howard and Matheson [9]). However, from the point of view of potential Operations Research applications the difficulty of actually constructing suitable risk-incorporating utility functions appears formidable.

2. Variance-penalized average reward MDP. A discrete *Markovian decision process* Γ is observed at discrete time points $t = 1, 2, \ldots$. The state space is denoted by $E = \{1, 2, \ldots, N\}$. With each state $i \in E$, we associate a finite action set $A(i)$. At any time point t the system is in one of the states and an action has to be chosen by the decision maker. If the system is in state i and action $a \in A(i)$ is chosen, then an immediate reward r_{ia} is earned and the process moves to a state $j \in E$ with transition probability $p_{iaj} \geq 0$ and $\sum_{j=1}^{N} p_{iaj} = 1$.

A *decision rule* π^t at time t is a function which assigns a probability to the event that action a is taken at time t. In general, π^t may depend on all realized states up to and including time t and on all realized actions up to time t. A *policy* π is a sequence of decision rules: $\pi = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$. For a *Markov policy* we require that the decision rule at time t depends only on the state at time t. A policy π is called *stationary* is all its decision rules are identical, that is, $\pi^t = \pi^1$ for all t; then we can let π_{ia} denote the probability of choosing action a in state i. A *deterministic policy* is a stationary policy with nonrandomized decision rules. Let C, $C(M)$, $C(S)$ and $C(D)$ be the sets of all policies, the Markov policies, the stationary policies, and the deterministic policies respectively.

Let X_t be the state at time t and Y_t be the action at time t and $P_\pi(X_t = j, Y_t = a|X_1 = i)$ be the conditional probability that at time t the state is j and the action taken is a, given that the initial state is i and the decision maker uses a policy π. For any policy π and initial state i, we define the *average expected reward* over the infinite horizon by

$$\Phi_i(\pi) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j,a} P_\pi(X_t = j, Y_t = a|X_1 = i) r_{ja}.$$

A policy π^* is *average optimal* in Γ if $\Phi_i(\pi^*) = \max_C \Phi_i(\pi)$ for all $i \in E$. It is well known that there exists an average optimal deterministic policy (e.g., see [4, p. 25]).

Let $\beta = [\beta_1, \beta_2, \ldots, \beta_N]$ be a given initial distribution, that is, β_i is the probability that $X_1 = i$, where $\beta_i \geq 0$ for all $i \in E$ and $\sum_i \beta_i = 1$. For any policy π, define $\Phi(\beta, \pi)$ by

$$\Phi(\beta, \pi) = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i} \beta_i \sum_{j,a} P_\pi(X_t = j, Y_t = a|X_1 = i) r_{ja}. \tag{2.1}$$

We shall say that π^0 is *β-optimal* in Γ if $\Phi(\beta, \pi^0) = \max_C \Phi(\beta, \pi)$. Clearly, an average optimal policy must be β-optimal, but not conversely. If β is strictly positive then a β-optimal policy is also an average optimal policy.

For any policy $\pi \in C$ and any positive integer T, we denote the *average expected state-action frequencies* in the first T periods by $x^T(\pi)$ according to the definition

$$x_{ja}^T(\pi) = \frac{1}{T} \sum_{t=1}^{T} \sum_{i} \beta_i P_\pi(X_t = j, Y_t = a|X_1 = i), \tag{2.2}$$

for all $a \in A(j)$, $j \in E$. Let $X(\pi)$ denote the set of vector-limit-points of the sequence $\{x^T(\pi), T = 1, 2, \ldots\}$. Now, define $C_1 = \{\pi \in C | X(\pi)$ is singleton$\}$. It is well known (e.g., see Kallenberg [12, p. 135]) that $C(S) \subset C_1$, and hence $X(\pi)$ is nonempty. If we denote by $x(\pi)$ the unique element of $X(\pi)$ for any $\pi \in C_1$, then from the definition of $\Phi(\beta, \pi)$ we have

$$\Phi(\beta, \pi) = \sum_{j,a} x_{ja}(\pi) r_{ja}. \tag{2.3}$$

Let $L(C) = \{x(\pi) \in X(\pi) | \pi \in C\}$ and let the sets $L(C(M))$, $L(C_1)$, $L(C(S))$ and $L(C(D))$ be defined analogously.

The linear program usually associated with the average reward MDP is the following (see [4], [7], [12])

$$\max \sum_{i,a} r_{ia} x_{ia}$$

$$\text{subject to} \quad \sum_{i,a} (\delta_{ij} - p_{iaj}) x_{ia} = 0, \qquad j \in E, \tag{2.4}$$

$$\sum_{a} x_{ja} + \sum_{i,a} (\delta_{ij} - p_{iaj}) y_{ia} = \beta_j, \qquad j \in E,$$

$$x_{ia}, y_{ia} \geq 0; \qquad i \in E, a \in A(i).$$

Let F denote the set of all feasible solutions of the above linear program, and for $(x, y) \in F$ define[1] $S_x = \{i \in E | \Sigma_a x_{ia} > 0\}$ and $S_y = \{i \in E | \Sigma_a x_{ia} = 0 \text{ and } \Sigma_a y_{ia} > 0\}$. In addition, let $X = \{x |$ there exists y such that $(x, y) \in F\}$, Ext(F) and Ext(X) denote the set of extreme points of F and X respectively.

We shall use the following important results.

THEOREM 2.1. (*Derman* [4], *Hordijk and Kallenberg* [8]).

$$\overline{L(C(D))} = \overline{L(C(S))} = L(C(M)) = L(C_1) = L(C) = X,$$

where $\bar{A}$ *denotes the closed convex hull of a set* A.

With a given $(x, y) \in F$ we can associate a policy $\pi \in C(S)$ defined by:

$$\pi_{ia}(x, y) = \begin{cases} x_{ia}/x_i & \text{if } i \in S_x,\ a \in A(i), \\ y_{ia}/y_i & \text{if } i \in S_y,\ a \in A(i), \\ \text{arbitrary} & i \notin S_x \cup S_y, \end{cases} \tag{2.5}$$

where $x_i = \Sigma_a x_{ia}$ and $y_i = \Sigma_a y_{ia}$.

Conversely, with every $\pi \in C(S)$ we can associate a point $(x(\pi), y(\pi)) \in F$ defined by

$$x_{ia}(\pi) = \left[\beta^T P^*(\pi)\right]_i \pi_{ia}, \qquad i \in E,\ a \in A(i), \tag{2.6}$$

$$y_{ia}(\pi) = \left[\beta^T D(\pi) + \gamma^T P^*(\pi)\right]_i \pi_{ia}, \qquad i \in E,\ a \in A(i) \tag{2.7}$$

where $P^*(\pi)$ and $D(\pi)$ are the stationary and the deviation matrices respectively, of the Markov Chain induced by π, and γ is an appropriately chosen vector. Recall that $P(\pi)$ is the stochastic matrix induced by π (i.e., $\pi = \Sigma_a p_{iaj}\pi_{ia}$),

$$P^*(\pi) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} [P(\pi)]^k \quad \text{and} \quad D(\pi) = [I - P(\pi) + P^*(\pi)]^{-1} - P^*(\pi).$$

For detailed definition of this transformation and the discussion of its relationship with the transformation (2.5) we refer the reader to Hordijk and Kallenberg [7].

In an attempt to define a measure of variability in the average reward process a number of criteria could be considered. For instance, we might define "the variance due to a policy" $\pi \in C_1$ by $V^1(\pi) = V_\pi(\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} R_t)$, where $V_\pi(\cdot)$ denotes the variance of a random variable with respect to the probability distribution induced by the policy π, and the random variable R_t denotes the reward at the tth stage. Alternatively, we could have defined for $\pi \in C_1$

$$V^2(\pi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} V_\pi(R_t).$$

The example below shows that both of these criteria fail to differentiate between

[1]Here $x(y)$ is treated as a vector whose components are $x_{ia}(y_{ia})$, arranged in the "natural" fashion. We shall not always differentiate between row and column vectors, this identification should be made in a manner consistent in any given equation.

streams of rewards with unequal fluctuations.

$i = 1$	$i = 2$	$i = 3$
$\begin{pmatrix} 0 \\ 100 \end{pmatrix} \begin{matrix} \to (0,1,0) \\ \to (0,0,1) \end{matrix}$	$(0) \to (1,0,0)$	$(-100) \to (1,0,0)$

The notation $(*) \to (\cdot, \cdot, \cdot)$ gives the reward and probability transitions when a given action is chosen. For instance, the choice of action 2 in state 1 yields the reward of 100, and results in a transition to state 3 with probability one. Here there are only two deterministic policies: π^1 which chooses action 1 in state 1, and π^2 which chooses action 2. Both π^1 and π^2 are average optimal; with $\Phi(\beta, \pi^1) = \Phi(\beta, \pi^2) = 0$ for all β. Further

$$0 = V^1(\pi^1) = V^1(\pi^2) = V^2(\pi^1) = V^2(\pi^2),$$

and yet π^1 and π^2 induce the streams of rewards $\{0, 0, 0, 0, \dots\}$, and $\{100, -100, 100, -100, \dots\}$ respectively. Thus both $V^1(\cdot)$ and $V^2(\cdot)$ fail to detect the variability of the second stream.

We are thus led to the notion of a *long-run variance of a policy* $\pi \in C_1$ and propose to consider it is a reasonable measure of variability. This notion is based on the observation that it is the long-run frequency of occurrence of state-action pairs (i, a) which determine the variability in the long-run average rewards. Further, since for a policy $\pi \in C_1$ the probability vector $x(\pi)$ is the (unique) limit of $x^T(\pi)$, it can be regarded as the *long-run probability distribution* on the "outcomes" (i, a) induced by π. Thus the *long-run variance* of $\pi \in C_1$ will be defined by

$$V(\pi) = \sum_i \sum_a \left[r_{ia} - \Phi(\beta, \pi)\right]^2 x_{ia}(\pi),$$

where r_{ia} are the rewards of the original MDP-process. Note that for $\pi \in C_1$ the above definition is equivalent to

$$V(\pi) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} E_\pi\left[\left(R_t - \Phi(\beta, \pi)\right)^2\right],$$

where $E_\pi(\cdot)$ is the expectation operator induced by the policy π. We will consider the optimization problem

$$\max_{\pi \in C_1} \left[\Phi(\beta, \pi) - \lambda V(\pi)\right], \qquad \lambda \in [0, \infty). \tag{2.8}$$

The objective function of (2.8) represents the long-run average gain penalized by a multiple of the long-run variance induced by the same policy π of C_1. It turns out that the problem (2.8) is closely related to the following convex quadratic program (with $\lambda \in [0, \infty)$):

$$\max \left[\sum_i \sum_a r_{ia} x_{ia} - \lambda \sum_i \sum_a r_{ia}^2 x_{ia} + \lambda \left[\sum_i \sum_a r_{ia} x_{ia}\right]^2\right]$$

subject to:

$$x \in X. \tag{2.9}$$

THEOREM 2.2. (i) *Let* $\lambda \in [0, \infty)$ *and* $\pi^* \in C_1$ *be optimal for* (2.8), *then* $x(\pi^*)$ *is optimal for* (2.9), *and the maxima in* (2.8) *and* (2.9) *are equal.*

(ii) *Let* x^* *be an optimal solution of* (2.9) *and* $\pi^* \in C_1$ *be such that* $x^* = x(\pi^*)$, *then* π^* *is optimal for* (2.8).

(iii) *Let* $\lambda > 0$ *and* $\pi^* \in C_1$ *be optimal for* (2.8), *then there does not exist* $\pi \in C_1$ *such that componentwise* $(\Phi(\beta, \pi), -V(\pi)) \geq (\Phi(\beta, \pi^*), -V(\pi^*))$ *with a strict inequality holding for at least one of the two components. That is,* π^* *is efficient or Pareto-Optimal.*

PROOF. From the definitions of $\Phi(\beta, \pi)$, $V(\pi)$ and C_1 and since $X = L(C_1)$ we have from (2.3) that for any $\pi \in C_1$

$$\Phi(\beta, \pi) - \lambda V(\pi) \tag{2.10}$$

$$= \sum_i \sum_a r_{ia} x_{ia}(\pi) - \lambda \sum_i \sum_a \left[r_{ia} - \sum_i \sum_a r_{ia} x_{ia}(\pi) \right]^2 x_{ia}(\pi)$$

$$= \sum_i \sum_a r_{ia} x_{ia}(\pi) - \lambda \sum_i \sum_a r_{ia}^2 x_{ia}(\pi) + \lambda \left[\sum_i \sum_a r_{ia} x_{ia}(\pi) \right]^2.$$

Since X is a bounded polyhedron which is a projection of F onto its x-coordinates, the maximum in (2.9) is achieved and by (2.10) is the same as the maximum of (2.8). The rest of the claims in (i) and (ii) now follow easily from the identity $X = L(C_1)$.

For part (iii) suppose π^* is not efficient, then there exists a policy $\pi \in C_1$ such that $[\Phi(\beta, \pi), -V(\pi)] \geq [\Phi(\beta, \pi^*), -V(\pi^*)]$ and $[\Phi(\beta, \pi), -V(\pi)] \neq [\Phi(\beta, \pi^*), -V(\pi^*)]$. This implies that $\Phi(\beta, \pi) + \lambda(-V(\pi)) > \Phi(\beta, \pi^*) + \lambda(-V(\pi^*))$ for $\lambda > 0$. This contradicts the optimality of π^* in (2.8). ■

COROLLARY 2.1. *There exists a deterministic optimal policy* π^* *for* (2.8). *If* $\lambda > 0$, *then* π^* *is efficient for the two objectives* $\Phi(\beta, \pi)$ *and* $-V(\pi)$, $\pi \in C_1$.

PROOF. Consider the problem of maximizing the convex objective function of (2.9) over the bounded polyhedron X. Since there exists $x^* \in \text{Ext}(X)$ at which this maximum is achieved we have from Hordijk and Kallenberg [8] (see p. 279) that there exists a policy $\pi^* \in C(D)$ such that $x(\pi^*) = x^*$. Thus by parts (ii) and (iii) of Theorem 2.2, π^* is a desired deterministic policy. ■

REMARK 2.1. The preceding theorem and corollary lead to the following still open problems. Is there an efficient way of computing an optimal extreme solution of (2.9) (perhaps, by an adaptation of the algorithms of either [2] or [5])? If x_0 is such an optimal solution how can we construct an optimal policy for the original problem (2.8)? If $x^0 \in L(C(S))$ and if $x^0 \neq x(\pi^0)$ we can still find $\pi^* \in C(S)$ such that $x^0 = x(\pi^*)$ by solving a linear system of equations given on p. 147 of [12]. However, it is conceivable that $x^0 \in L(C) \setminus L(C(S))$. Of course, in the latter case a policy $\hat{\pi} \in C_1$ optimal for (2.8) could be found by the algorithm described on p. 282 of [8] which, however, appears to be computationally prohibitive. On the other hand, the latter problem vanishes in the important "unichain" case since now $L(C) = L(C(S))$ and $x^0 = x(\pi^0)$ (see Theorems 6 and 7 in [8], or [4, p. 95]), and the optimality of π^0 in (2.8) follows from (2.10).

3. Measures of variability in discounted MDP's. We now consider a finite state/action Markovian Decision Process over the infinite time horizon and with a

constant discount factor. For any policy π and initial state i, we define the *expected discounted reward*, over the infinite horizon by

$$\Phi_i^{\alpha}(\pi) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_j \sum_a P_{\pi}(X_t = j, Y_t = a | X_1 = i) r_{ja} \tag{3.1}$$

where $\alpha \in (0,1)$ is the *discount factor*. The process Γ with the reward criterion (3.1) will be called the *discounted Markovian decision process* or DMDP, for short.

A policy π^* is *optimal* in DMDP if for every $i \in E$

$$\Phi_i^{\alpha}(\pi^*) = \max_C \Phi_i^{\alpha}(\pi). \tag{3.2}$$

It is well known that there exists an optimal policy $\pi^* \in C(D)$ (e.g., see Derman [4, p. 23]).

Again, we shall assume that the *initial distribution* of Γ is the given vector: $\beta = (\beta_1, \beta_2, \ldots, \beta_N)^T$. Let

$$\Phi^{\alpha}(\beta, \pi) = \sum_{t=1}^{\infty} \alpha^{t-1} \sum_{i=1}^{N} \beta_i \sum_j \sum_a P_{\pi}(X_t = j, Y_t = a | X_1 = i) r_{ja}, \tag{3.3}$$

be defined for every $\pi \in C$.

Surprisingly, perhaps, the question of how to measure the variability in the stream of rewards appears to be more difficult in the discounted process, than in the average reward process. Below we again define three alternative measures of a "variance due to a policy π" and briefly discuss some of their respective advantages and disadvantages.

Recall that the reward at the tth stage of the process was denoted by R_t, and that for a given initial distribution β and a policy $\pi \in C$, the expectation and the variance: $E_{\pi}(R_t)$ and $V_{\pi}(R_t)$ are well defined for every t. We now, introduce the following:

(i) Variance of the "Present Value" due to a policy π. This measure was introduced by Sobel [15] and is defined by

$$V^1(\pi) = V_{\pi}\left[\sum_{t=1}^{\infty} \alpha^{t-1} R_t\right].$$

(ii) The "stagewise-variance" due to a policy π. This measure was introduced in Filar and Lee [6], and is defined by

$$V^2(\pi) = \sum_{t=1}^{\infty} \alpha^{t-1} V_{\pi}(R_t).$$

(iii) The "discount normalized variance" due to a policy π. This is defined by

$$V^3(\pi) = \sum_{t=1}^{\infty} \alpha^{t-1} E_{\pi}\left[\left(R_t - (1-\alpha)\Phi^{\alpha}(\beta, \pi)\right)^2\right]. \tag{3.4}$$

Note that in the discounted model it is the variability in the early stages of the process that "carry more weight" and hence $V^1(\pi)$ and $V^2(\pi)$ are now meaningful, as well as $V^3(\pi)$. In the context of this paper, the relevance of each one of the above three measures should, perhaps, be evaluated on the basis of:

(a) the meaning in the underlying physical decision model, and

(b) whether it possesses desirable mathematical properties that ensure that the corresponding variance-penalized optimization problem is tractable.

In line with the above comments we can state that the measure $V^1(\pi)$ is clearly meaningful since it is precisely the variance of the random variable (often referred to as the present value) whose expectation is maximized in the classical problem (3.2). Unfortunately, Sobel [15] has demonstrated that this measure lacks the monotonicity property of dynamic programming, and it is not clear how the related variance-penalized problem could be analyzed even with the help of certain analytic properties demonstrated in Bouakiz [1].

The stagewise-variance $V^2(\pi)$ would, of course, coincide with $V^1(\pi)$ if the random variables R_t were independent. However, in a nontrivial Markov Decision Process such independence is usually absent. Nonetheless, in the context of maximizing the variance-penalized objective function

$$W^2(\pi, \lambda, \beta) = \sum_{t=1}^{\infty} \alpha^{t-1}\left[E_\pi(R_t) - \lambda V_\pi(R_t)\right] = \Phi^\alpha(\beta, \pi) - \lambda V^2(\pi)$$

this measure can be given the following appealing interpretation: suppose that a *risk-sensitive decision maker* perceives his reward at stage t to be (while he uses a policy π)

$$\hat{R}_t = R_t - \lambda f(R_t - E_\pi(R_t)),$$

for some penalty function f of the deviations from the mean. Assuming that f can be approximated by the first two terms of its Taylor's series, with some coefficients a_0, a_1, and $a_2 > 0$, we have

$$\hat{R}_t \approx R_t - \lambda a_0 - \lambda a_1(R_t - E_\pi(R_t)) - \lambda a_2(R_t - E_\pi(R_t))^2.$$

Thus,

$$E_\pi(\hat{R}_t) \approx E_\pi(R_t) - \lambda a_0 - \lambda a_2 V_\pi(R_t).$$

Hence, the "usual" maximization of the discounted sum of $E_\pi(\hat{R}_t)$'s is equivalent to

$$(\dagger) \qquad \max_C \sum_{t=1}^{\infty} \alpha^{t-1}\left[E_\pi(R_t) - \lambda V_\pi(R_t)\right].$$

REMARK 3.1. It may appear at first that the problem (†) is equivalent to a dynamic program. The example given below shows otherwise.

$i = 1$	$i = 2$
$(1) \to (\frac{1}{2}, \frac{1}{2})$	$(0) \to (0, 1)$

In the above, there is only one action in each of the two states, and the choice of the single action in state 1 results in a reward of 1 and the transitions to states 1 and 2 with probabilities $\frac{1}{2}$ and $\frac{1}{2}$. Let $\lambda = 1$, $\beta_1^T = (1, 0)$ and $\beta_2^T = (0, 1)$, and π^0 denote the unique stationary policy. If (†) were a dynamic program, we would expect the recursive equation

$$W^2(\pi^0, 1, \beta_1) = \left[E_{\pi^0}(R_1) - V_{\pi^0}(R_1)\right] + \alpha\left[\tfrac{1}{2}W^2(\pi^0, 1, \beta_1) + \tfrac{1}{2}W^2(\pi^0, 1, \beta_2)\right]$$

to hold. However, a straightforward calculation shows that the left side of the above equation is $4/(4 - \alpha)$, while the right side is $(4 + \alpha)/(4 - \alpha/2)$ (where α is the discount factor).

Despite the above, we demonstrate in §5 that an analysis of problem (†) is possible and that it possesses desirable asymptotic properties, if viewed as a limit of finite-stage problems.

Finally, we turn our attention to the discount normalized variance $V^3(\pi)$. It will be seen in §4 that the variance-penalized optimization problem

$$(*) \qquad \max_{\pi} \left[\Phi^{\alpha}(\beta, \pi) - \lambda V^3(\pi)\right]$$

lends itself to the straightforward mathematical analysis analogous to that used in §2 for the average reward process and that for $\pi \in C_1$, in the limit as $\alpha \to 1$, the results are consistent with those of §2. Furthermore $V^3(\pi)$ is a stagewise variance in the same sense as $V^2(\pi)$ except that at every stage t, instead of fluctuations about $E_{\pi}(R_t)$ it aggregates the squared fluctuations about $(1-\alpha)\Phi^{\alpha}(\beta, \pi)$. However, for $\pi \in C_1$ and α sufficiently near 1, $(1-\alpha)\Phi^{\alpha}(\beta, \pi)$ is approximately equal to $\Phi(\beta, \pi)$, the long-run expected average reward corresponding to the same policy. Thus the quantity $(1-\alpha)\Phi^{\alpha}(\beta, \pi)$ can be regarded as a measure of "per-stage return" due to a policy π used in the discounted process. Consequently, the general term of $V^3(\pi)$, $E_{\pi}\{[R_t - (1-\alpha)\Phi^{\alpha}(\beta, \pi)]^2\}$ represents what might be called "discount normalized variance" of the tth stage reward R_t, thereby motivating the use of $V^3(\pi)$.

4. The discount normalized variance-penalized MDP. In this section we shall consider the variance-penalized optimization problem

$$(4.1) \qquad \max_{\pi} \left[\Phi^{\alpha}(\beta, \pi) - \lambda V^3(\pi)\right],$$

where λ is a nonnegative penalty parameter. Towards this end we shall need the following notation and known results.

For any policy $\pi \in C$ we shall define the "discounted, expected, state-action frequencies" by

$$(4.2) \qquad x_{ja}(\pi) = \sum_{i=1}^{N} \beta_i \sum_{t=1}^{\infty} \alpha^{t-1} P_{\pi}(X_t = j, Y_t = a | X_1 = i)$$

where $(j, a) \in E \times A(j)$. Recalling that every stationary policy $\pi \in C(S)$ defines a Markov chain determined by the transition matrix $P(\pi) = (p_{ij}(\pi))_{i,j=1}^{N}$, with $p_{ij}(\pi) = \sum_{a \in A(i)} p_{iaj}\pi_{ia}$, it is easy to check that for all $\pi \in C(S)$ and $(j, a) \in E \times A(j)$

$$(4.3) \qquad x_{ja}(\pi) = \left[\beta^T (I - \alpha P(\pi))^{-1}\right]_j \pi_{ja}$$

where $[\cdot]_j$ denotes the jth entry of a vector. We shall denote by $x(\pi)$ the vector $\{x_{ja}(\pi) | (j, a) \in E \times A(j)\}$ (note that $x(\pi)$ can also be regarded as a vector function of π).

Let $K(C) = \{x(\pi) | \pi \in C\}$ and let the sets $K(C(M))$, $K(C(S))$, $K(C(D))$ be defined analogously. There is an interesting correspondence between these sets and the feasible region of the standard linear programming formulation of Γ. It is summarized in the following theorem, that can be established by an argument analogous to the one used in Theorems 3.4.8 and 3.4.9 in Kallenberg [12], and which will be important in the sequel.

THEOREM 4.1. *Let Γ be a discounted MDP, and let β, $K(C)$ and $K(C(S))$ be as above. Consider the linear program: maximize $\sum_{i=1}^{N} \sum_{a \in A(i)} r_{ia} x_{ia}$ over the set*

$$X(\beta) = \left\{x | \sum_i \sum_a (\delta_{ij} - \alpha p_{iaj}) x_{ia} = \beta_j; \quad j \in E, x_{ia} \geq 0; (i, a) \in E \times A(i)\right\},$$

then

(i) $K(C) = K(C(M)) = K(C(S)) = \overline{K(C(D))} = X(\beta)$, *where* $\bar{A}$ *denotes the closed convex hull of a set* A. *Further*, $\Phi^{\alpha}(\beta, \pi) = \sum_j \sum_a x_{ja}(\pi) r_{ja}$,

(ii) $\max_C \Phi^{\alpha}(\beta, \pi) = \max_{X(\beta)} \sum_i \sum_a r_{ia} x_{ia}$,

(iii) *If* π^* *maximizes the left side of* (ii), *then* $x(\pi^*)$ *maximizes the right side of* (ii).

(iv) *If* x^* *maximizes the right side of* (ii), *then* $\pi(x^*) \in C(S)$ *maximizes the left side of* (ii), *where* $\pi(x) \in C(S)$ *is defined for all* $x \in X(\beta)$ *by*

$$\pi_{ia}(x) = \begin{cases} x_{ia}/x_i & \text{if } x_i = \sum_a x_{ia} > 0,\ a \in A(i), \\ \text{arbitrary}, & \text{if } x_i = 0,\ a \in A(i). \end{cases}$$

(v) *If* $x \in X(\beta)$, *and* $\pi(x) \in C(S)$ *is defined as in* (iv), *then* $x = x(\pi(x))$, *where* $x(\pi(x))$ *is defined by* (4.3).

Now, from the definition of $V^3(\pi)$ we see with the help of (4.2) that

(4.4)

$$\begin{aligned} V^3(\pi) &= \sum_{t=1}^{\infty} \alpha^{t-1} \sum_i \sum_a E_{\pi}\left\{ \left[R_t - (1-\alpha)\Phi^{\alpha}(\beta, \pi) \right]^2 | X_t = i, Y_t = a \right\} \\ &\quad \times P_{\pi}(X_t = i, Y_t = a) \\ &= \sum_{t=1}^{\infty} \alpha^{t-1} \sum_i \sum_a \left[r_{ia} - (1-\alpha)\Phi^{\alpha}(\beta, \pi) \right]^2 \sum_{k=1}^{N} P_{\pi}(X_t = i, Y_i = a | X_1 = k) \beta_k \\ &= \sum_i \sum_a \left[r_{ia} - (1-\alpha)\Phi^{\alpha}(\beta, \pi) \right]^2 x_{ia}(\pi). \end{aligned}$$

In view of (4.4) we will be able to show that, in a manner analogous to that of §2, the variance-penalized optimization problem (4.1) is closely related to the following convex quadratic program

$$\max \left\{ \sum_i \sum_a r_{ia} x_{ia} - \lambda \sum_i \sum_a r_{ia}^2 x_{ia} + \lambda(1-\alpha) \left(\sum_i \sum_a r_{ia} x_{ia} \right)^2 \right\}$$

subject to:

$$x \in X(\beta). \tag{4.5}$$

THEOREM 4.2. (i) *Let* $\lambda \in [0, \infty)$ *and* $\pi^* \in C$ *be optimal for* (4.1), *then* $x(\pi^*)$ *is optimal for* (4.5), *and the maxima in* (4.1) *and* (4.5) *are equal.*

(ii) *Let* x^* *be an optimal solution of* (4.5), *then* $\pi(x^*)$ (*defined as in Theorem* 4.1(iv)) *is optimal for* (4.1).

(iii) *Let* $\lambda > 0$, *and* π^* *be optimal for* (4.1), *then* π^* *is Pareto optimal for the two objectives* $\Phi^{\alpha}(\beta, \pi)$ *and* $[-V^3(\pi)]$.

PROOF. The proof is logically analogous to that of Theorem 2.1, with Theorem 4.1 providing nearly all the necessary tools. ■

COROLLARY 4.1. *There exists a deterministic optimal policy* π^* *for* (4.1). *If* $\lambda > 0$, *then* π^* *is Pareto optimal for the two objectives* $\Phi^{\alpha}(\beta, \pi)$ *and* $[-V^3(\pi)]$.

PROOF. Analogous to that of Corollary 2.3, except that in the discounted case all extreme points of $X(\beta)$ correspond to policies in $C(D)$, thereby simplifying the argument. ■

REMARK 4.1. Part (v) of Theorem 4.1 demonstrates that the computational difficulty mentioned in the Remark 2.1 for the undiscounted case does not arise in the discounted model. Again, we refer the reader to [2] and [5] for algorithms available for tackling the problem (4.5).

5. The stagewise variance-penalized MDP. In this section we shall discuss the variance-penalized optimization problem

$$\max_{\pi} \left[\Phi^{\alpha}(\beta, \pi) - \lambda V^2(\pi)\right], \tag{5.1}$$

where λ is a nonnegative penalty parameter, and $V^2(\pi)$ is the stagewise variance due to a policy π, that was introduced in §3.

Let us now denote the given discounted Markov Decision Process by $\Gamma = \langle E, A, r, p\rangle$ where $A = \bigcup_{i=1}^{N} A(i)$, $r = \{r_{ia} | i \in E, a \in A(i)\}$ and $p = \{p_{iaj} | (i, a, j) \in E \times A(i) \times E\}$.

The analysis of the above "variance-penalized" problem (5.1) is facilitated by the following technical device. We shall "pretend" that every policy $\pi \in C$ is being applied independently and simultaneously in two processes: Γ and Γ^c, where Γ^c is an identical copy of Γ and X_t^c, Y_t^c and R_t^c will denote the state, the action, and the reward at t in Γ^c, respectively, when the policy $\pi \in C$ is being used. Also the initial state distribution β will be assumed to be the same in Γ and Γ^c.

From above we know that for every t, R_t, and R_t^c are independent, identically distributed random variables (as long as the same $\pi \in C$ is used in *both* Γ and Γ^c), and hence that for every $t = 1, 2, \ldots$.

$$E_{\pi}(R_t) - \lambda V_{\pi}(R_t) = \tfrac{1}{2}E_{\pi}(R_t + R_t^c) - \tfrac{1}{2}\lambda E_{\pi}(R_t - R_t^c)^2 = \tfrac{1}{2}E_{\pi}\left[\psi(R_t, R_t^c, \lambda)\right]$$

where $\psi(R_t, R_t^c, \lambda) = (R_t + R_t^c) - \lambda(R_t - R_t^c)^2$. Now, the original problem (5.1) can be rewritten as

$$\max_{C} \left\{\frac{1}{2}\sum_{t=1}^{\infty} \alpha^{t-1} E_{\pi}\left[\psi(R_t, R_t^c, \lambda)\right]\right\}. \tag{5.2}$$

Since the objective of (5.2) resembles the ordinary discounted gain objective of a Markovian Decision Process we are led to consider the following *product DMDP* $\hat{\Gamma} = \langle \hat{E}, \hat{A}, \hat{r}, \hat{p}\rangle$: $\hat{E}$ = the state space = $E \times E$, $\hat{A}(i, j)$ = the action space in state $(i, j) = A(i) \times A(j)$, $\hat{r}_{ij, aa'} = \psi(r_{ia}, r_{ja'}, \lambda) = (r_{ia} + r_{ja'}) - \lambda(r_{ia} - r_{ja'})^2$ for all $[ij, aa'] \in \hat{E} \times \hat{A}(i, j)$, and $\hat{p}_{kl, aa', ij} = p_{kai}p_{la'j}$ for all $[kl, aa', ij] \in \hat{E} \times \hat{A}(k, l) \times \hat{E}$. Further, let $\hat{X}_t$, $\hat{Y}_t$ and $\hat{R}_t$ denote the state, the action, and the reward at time t, respectively, in the process $\hat{\Gamma}$. The initial distribution in $\hat{\Gamma}$ will be the N^2-dimensional probability vector $\hat{\beta}$ with $\hat{\beta}_{ij} = \beta_i\beta_j$, for all $(ij) \in \hat{E}$. Of course, $\hat{C}$, $\hat{C}(M)$ and $\hat{C}(S)$ will denote the sets of all, Markov, and stationary policies in $\hat{\Gamma}$, and $\Phi_{ij}^{\alpha}(\hat{\pi})$ will denote the expected discounted reward from using $\hat{\pi}$ given initial state ij in $\hat{\Gamma}$.

REMARK 5.1. Recall that a policy $\pi = (\pi^1, \pi^2, \ldots, \pi^t, \ldots)$ is a sequence of decision rules, where π^t is a function of the history $H_t = (X_1, Y_1, \ldots, X_t)$ and the action set $A(X_t)$ which assigns a probability to the event that action $a \in A(X_t)$ is taken at time t. For every realization $h_t = (i_1, a_1, \ldots, i_t)$ of H_t and action $a \in A(i_t)$ we have that $0 \le \pi^t(h_t, a) \le 1$ and $\sum_{a \in A(i_t)} \pi^t(h_t, a) = 1$. If the decision-maker uses policy π simultaneously in Γ and Γ^c, and computes his "combined reward" from the two

processes according to $\psi(R_t, R_t^c, \lambda)$, he is behaving precisely as if he were using the policy $\hat{\pi} = (\hat{\pi}^1, \hat{\pi}^2, \ldots, \hat{\pi}^t, \ldots) = \pi \times \pi$ defined as follows:

$$\hat{\pi}^t\big(\hat{h}_t, (a, a')\big) = \pi^t(h_t, a)\pi^t(h_t^c, a')$$

where $\hat{h}_t = ((i_1, j_1), (a_1, a_1'), \ldots, (i_t, j_t))$, $h_t = (i_1, a_1, \ldots, i_t)$ and $h_t^c = (j_1, a_1', \ldots, j_t)$.

We shall be particularly interested in the subset $\hat{C}^2(M)$ of Markov policies of $\hat{\Gamma}$ defined by $\hat{C}^2(M) = \{\hat{\pi} = \pi \times \pi | \pi \in C(M)$ and $\hat{\pi}$ as in Remark 5.1$\}$, and let $\hat{C}^2(S)$ be defined analogously. The following lemmas are straightforward and are proved in Lee [13].

LEMMA 5.1. *Let* $\hat{\pi} = \pi \times \pi \in \hat{C}^2(M)$ *then*

$$P_{\hat{\pi}}\big(\hat{X}_t = (i, j)\big) = P_\pi(X_t = i)P_\pi(X_t^c = j),$$

for every $t = 1, 2, \ldots$.

LEMMA 5.2. *Let* $\hat{\pi} = \pi \times \pi \in \hat{C}^2(M)$, *and let* $\hat{\Phi}^\alpha(\hat{\beta}, \hat{\pi})$ *be the discounted reward, induced by* $\hat{\pi}$ *in* $\hat{\Gamma}$ *(see* (3.3)*). Then the objective function* $W^2(\pi, \lambda, \beta)$ *of* (5.1) *satisfies* $W^2(\pi, \lambda, \beta) = \frac{1}{2}\hat{\Phi}^\alpha(\hat{\beta}, \hat{\pi})$.

THEOREM 5.1.

$$\max_C W^2(\pi, \lambda\beta) = \max_{C(M)} W^2(\pi, \lambda, \beta) = \max_{\hat{C}^2(M)} \tfrac{1}{2}\hat{\Phi}^\alpha(\hat{\beta}, \hat{\pi}).$$

PROOF. Since

$$\begin{aligned} W^2(\pi, \lambda, \beta) &= \sum_{t=1}^{\infty} \alpha^{t-1}\big[E_\pi(R_t) - \lambda V_\pi(R_t)\big] \\ &= \sum_{t=1}^{\infty} \alpha^{t-1}\big[E_\pi(R_t) - \lambda E_\pi\big(R_t - E_\pi(R_t)\big)^2\big] \end{aligned}$$

is a continuous function of π, and C is compact (see Derman [4, p. 20]); there exists a policy $\pi^* \in C$ such that $W^2(\pi^*, \lambda, \beta) = \max_C W^2(\pi, \lambda, \beta)$. For policy π^*, there exists a policy $\pi_0 \in C(M)$ satisfying (see Derman [4, p. 91])

$$P_{\pi^*}(X_t = j, Y_t = a) = P_{\pi_0}(X_t = j, Y_t = a),$$

for all $t = 1, 2, \ldots$ and $j \in E$, $a \in A(j)$. Thus $W^2(\pi^*, \lambda, \beta) = W^2(\pi_0, \lambda, \beta)$. From the above and Lemma 5.2, the theorem is proven. ∎

In view of the above the variance-penalized problem (5.1) has been transformed to the problem of maximizing the "usual" discounted expected gain of the product MDP $\hat{\Gamma}$, but over the "product" space of Markov policies $\hat{C}^2(\hat{M})$.

REMARK 5.2. While the first equality in the statement of Theorem 5.1 establishes the existence of an optimal Markov policy for our variance-penalized problem (5.1), an example given by Lee [13] (see pp. 69–70) shows that "natural" restriction to the class of stationary strategies involves a loss of optimality in (5.1). That is, it shows that the strict inequality

$$\max_C W^2(\pi, \lambda, \beta) > \max_{C(S)} W^2(\pi, \lambda, \beta) \tag{5.3}$$

is in general possible.

The inequality (5.3) suggests that, perhaps, the natural problem to consider first is the *stationary variance-penalized problem*, namely

$$\max_{C(S)} \left\{ W^2(\pi, \lambda, \beta) = \sum_{t=1}^{\infty} \alpha^{t-1} \left[E_\pi(R_t) - \lambda V_\pi(R_t) \right] \right\}. \tag{5.4}$$

This can be reformulated via Lemma 5.2 as a problem concerning $\hat{\Gamma}$, namely

$$\max_{\hat{C}^2(S)} \tfrac{1}{2}\hat{\Phi}^\alpha(\hat{\beta}, \hat{\pi}). \tag{5.5}$$

If the maximization in (5.5) were over all of $\hat{C}(S)$, the solution could be readily obtained by linear programming (e.g., see Theorem 3.4.9 in [12]). Unfortunately, the set $\hat{C}^2(S)$ of those stationary policies of $\hat{\Gamma}$ which are "products" of stationary policies of Γ has a somewhat complicated structure. Nonetheless, it will be seen that an optimal solution of (5.5) can be obtained from a mathematical program. We shall need additional notation.

We now formulate a mathematical program which will later be seen to be related to (5.4). Let $L(\hat{x}) = \frac{1}{2}\sum_{(i,j)\in\hat{E}} \sum_{(a,a')\in\hat{A}(i,j)} \hat{r}_{ij,aa'}\hat{x}_{ij,aa'}$, where $\hat{x}$ can be regarded as a vector whose entries $\hat{x}_{ij,aa'}$, range over $\hat{E} \times \hat{A}(i,j)$. Let $\hat{x}_{ij} = \sum_{(a,a')\in\hat{A}(i,j)} \hat{x}_{ij,aa'}$ for all $(i,j) \in \hat{E}$ and formulate the mathematical program

$$\text{maximize } L(\hat{x}) \tag{5.6}$$

subject to:

(a) $\hat{x} \in \hat{X}(\beta)$,

(b) $x \in X(\beta)$,

(c) $x_i x_j \hat{x}_{ijaa'} = \hat{x}_{ij} x_{ia} x_{ja'}$; $(ij, aa') \in \hat{E} \times \hat{A}(i,j)$.

In the above x and $X(\beta)$ are defined in Theorem 4.1 for the process Γ, and $\hat{x}$ and $\hat{X}(\hat{\beta})$ are the analogous entities for the process $\hat{\Gamma}$. Note that the objective for (5.6) is linear, as are all the constraints in (a) and (b), while the constraints in (c) are trilinear in the variables x_{ia} and $\hat{x}_{ij,aa'}$.

THEOREM 5.2. *Let $\xi^* = (x^*, \hat{x}^*)$ be an optimal solution of* (5.6) *and define a policy $\pi^* \in C(S)$ as $\pi_{ia}^* = x_{ia}^*/x_i^*$, if $x_i^* = \sum_{a\in A(i)} x_{ia}^* > 0$, and arbitrarily otherwise. Then π^* is an optimal solution of* (5.4).

PROOF. We prove this result only for the case where β, the initial state distribution of the original process Γ, is strictly positive in every entry. The proof of the general case is quite cumbersome and can be found in Filar and Lee [6] and Lee [13].

It should first be noted that since the feasible region of (5.6) is compact, an optimal solution $\xi^* = (x^*, \hat{x}^*)$ exists. Note that since ξ^* is feasible for (5.6) we can now define the policy $\tilde{\pi} \in \hat{C}(S)$ by (recall that $\beta_i > 0$ for all i) $\tilde{\pi}_{ij,aa'} = \hat{x}_{ij,aa'}^*/\hat{x}_{ij}^*$, for all $(i,j) \in \hat{E}$ and $(a,a') \in \hat{A}(i,j)$. Note that $\tilde{\pi}$ satisfies $\tilde{\pi}_{ij,aa'} = (x_{ia}^*/x_i^*)(x_{ja'}^*/x_j^*) = \pi_{ia}^*\pi_{ja'}^*$. Thus by definition $\tilde{\pi} = \pi^* \times \pi^* := \hat{\pi}^* \in \hat{C}^2(S)$. Also by Theorem 4.1 parts (i) and (v), and the above we have that $\hat{x}^* = \hat{x}(\hat{\pi}^*)$ and hence $L(\hat{x}^*) = L(\hat{x}(\hat{\pi}^*)) = \frac{1}{2}\hat{\Phi}^\alpha(\hat{\beta}, \hat{\pi}^*)$. Now by Lemma 5.2

$$W^2(\pi^*, \lambda, \beta) = L(\hat{x}^*). \tag{5.7}$$

However, with an arbitrary $\pi \in C(S)$ and $\hat{\pi} = \pi \times \pi \in \hat{C}(S)$, we can associate

$\xi(\pi) = (x(\pi), \hat{x}(\hat{\pi}))$ defined by (see (4.3))

$$x_{ja'}(\pi) = \left\{ \beta^T \left[I - \alpha P(\pi) \right]^{-1} \right\}_j \pi_{ja'}, \qquad (j, a') \in E \times A(j), \quad \text{and}$$

$$\hat{x}_{ij,\,aa'}(\hat{\pi}) = \left\{ \hat{\beta}^T \left[I - \alpha \hat{P}(\hat{\pi}) \right]^{-1} \right\}_{ij} \hat{\pi}_{ij,\,aa'}, \qquad (ij, aa') \in \hat{E} \times \hat{A}(i, j).$$

where I is the $N \times N$, and $N^2 \times N^2$ identity matrix respectively. By Theorem 4.1(i), $x(\pi) \in X(\beta)$ and $\hat{x}(\hat{\pi}) \in \hat{X}(\hat{\beta})$. Further, since $\hat{\pi}_{ij,\,aa'} = \pi_{ia}\pi_{ja'}$, we easily obtain from the above

$$\hat{x}_{ij}(\hat{\pi}) x_{ia}(\pi) x_{ja'}(\pi) = \hat{x}_{ij,\,aa'}(\hat{\pi}) x_i(\pi) x_j(\pi),$$

where $\hat{x}_{ij}(\hat{\pi}) = \sum_{(a,\,a') \in \hat{A}(i,\,j)} \hat{x}_{ij,\,aa'}(\hat{\pi})$; $x_i(\pi) = \sum_{a \in A(i)} x_{ia}(\pi)$, etc. as usual. Thus $\xi(\pi)$ is feasible for (5.6).

Now, an argument analogous to that used to derive (5.7) shows that $W^2(\pi, \lambda, \beta) = L(x(\hat{\pi}))$, so that the optimality of ξ^* in (5.6) yields

$$W^2(\pi^*, \lambda, \beta) \geq W^2(\pi, \lambda, \beta), \tag{5.8}$$

for all $\pi \in C(S)$. ■

REMARK 5.3. It is worth mentioning that the original problem $\max_\pi W^2(\pi, \lambda, \beta)$ can be regarded as a limiting case of finite horizon problems in the following sense. Let

$$W_T^2(\pi, \lambda, \beta) = \sum_{t=1}^{T} \alpha^{t-1} \left[E_\pi(R_t) - \lambda V_\pi(R_t) \right],$$

where the dependence on β of the right side of the above equation is suppressed. Also, let π_T^* and π^* denote optimal policies that maximize $W_T^2(\pi, \lambda, \beta)$ and $W^2(\pi, \lambda, \beta)$ respectively, recalling that the existence of π^* follows from Theorem 5.2 and the existence of π_T^* can be established similarly. It is shown in [6] and [13] that

$$\lim_{T \to \infty} W^2(\pi_T^*, \lambda, \beta) = W^2(\pi^*, \lambda, \beta),$$

and that the limit of every convergent subsequence of $\{\pi_T^*\}_{T=1}^{\infty}$ is optimal for (5.1). Thus the infinite horizon optimization problem (5.1) is consistent with the corresponding finite horizon problems.

Acknowledgements. We are indebted to C. G. Bird for the motivation of one of the objective criteria in the discounted model, to A. J. Goldman for streamlining the analysis of §5, and to M. J. Sobel for a number of helpful discussions. This paper supersedes reference [6] by two of the present authors.

References

[1] Bouakiz, M. (1985). Risk Sensitivity in Stochastic Optimization with Applications. Ph.D. Thesis, Georgia Institute of Technology, Atlanta.

[2] Carbot, A. V. and Francis, R. L. (1970). Solving Nonconvex Quadratic Minimization Problems by Ranking the Extreme Points. *Oper. Res.* **18** 82–86.

[3] Chung, K.-J. (1985). Some Topics in Risk-Sensitive Stochastic Dynamic Models. Ph.D. Thesis, Georgia Institute of Technology, Atlanta.

[4] Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

[5] Falk, J. and Hoffman, K. L. R. (1976). A Successive Underestimation Method for Concave Minimization Problems. *Math. Oper. Res.* **1** 251–259.

[6] Filar, J. A. and Lee, H.-M. (1986). Variance-penalized Markov Decision Processes. Technical Report #463 Department of Mathematical Sciences, The Johns Hopkins University, Baltimore, MD.
[7] Hordijk, A. and Kallenberg, L. C. M. (1979). Linear Programming and Markov Decision Chains. *Management Sci.* **25** 352–362.
[8] _____and_____. (1984). Constrained Undiscounted Dynamic Programming. *Math. Oper. Res.* **9** 276–289.
[9] Howard, R. S. and Matheson, J. E. (1972). Risk-Sensitive Markov Decision Processes. *Management Sci.* **18** 356–369.
[10] Jacquette, S. C. (1973). Markov Decision Processes with a New Optimality Criterion: Discrete Time. *Ann. Statist.* **1** 496–505.
[11] _____. (1976). A Utility Criterion for Markov Decision Process. *Management Sci.* **23** 43–49.
[12] Kallenberg, L. C. M. (1983). Linear Programming and Finite Markovian Control Problems. Mathematisch Centrum Tract #148, Amsterdam.
[13] Lee, H.-M. (1985). Gain/Variability Tradeoffs in Markovian Decision Processes and Related Problems. Ph.D. Thesis, The Johns Hopkins University, Baltimore, MD.
[14] Markowitz, H. (1959). *Portfolio Selection*. Wiley, New York.
[15] Sobel, M. J. (1982). The Variance of Discounted MDP's. *J. Appl. Probab.* **19** 794–802.
[16] _____. (1984). Mean-Variance Tradeoffs in an Undiscounted MDP. Unpublished manuscript, Georgia Institute of Technology, Atlanta.
[17] _____. (1985). Maximal Mean-Variance Ratio in an Undiscounted MDP. *Oper. Res. Lett.* **4** 157–159.
[18] White, D. J. (1984). Probabilistic Constraints and Variance in Markov Decision Processes. *Notes in Decision Theory* **149**, University of Manchester, Manchester, United Kingdom.

FILAR: DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND, BALTIMORE COUNTY CAMPUS, CATONSVILLE, MARYLAND 21228
KALLENBERG: UNIVERSITY OF LEIDEN, LEIDEN, THE NETHERLANDS
LEE: THE JOHNS HOPKINS UNIVERSITY, BALTIMORE, MARYLAND