MATHEMATICS OF OPERATIONS RESEARCH
Vol. 17, No 3, August 1992
Printed in U S A

VARIABILITY SENSITIVE MARKOV DECISION PROCESSES*[†]

MELIKE BAYKAL-GÜRSOY AND KEITH W. ROSS

Considered are time-average Markov Decision Processes (MDPs) with finite state and action spaces. Two definitions of variability are introduced, namely, the expected time-average variability and time-average expected variability. The two criteria are in general different, although they can both be employed to penalize for variance in the stream of rewards. For communicating MDPs, we construct a (randomized) stationary policy that is ϵ-optimal for both criteria; the policy is optimal and pure for a specific variability function. For general multichain MDPs, a state space decomposition leads to a similar result for the expected time-average variability. We also consider the problem of the decision maker choosing the initial state along with the policy.

1. Introduction. Considered are time-average Markov Decision Processes (MDPs) with finite state and action spaces and with a fixed and given initial state. The great majority of the literature in this area is concerned with finding a policy $\mathbf{u}$ that maximizes

$$\phi(\mathbf{u}) = \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}[R_m],$$

where R_m is the reward obtained at epoch m. It is well known that there exists an optimal pure (i.e., stationary and deterministic) policy for this criterion. Moreover, policy improvement, value iteration, and linear programming algorithms are available to locate such an optimal pure policy (e.g., see [9], [11]).

Recently there has been interest in studying criteria that take into account the variance in the stream of rewards. For instance, Filar *et al.* [8] consider the problem of maximizing $\phi(\mathbf{u}) - \lambda\,\mathrm{var}(\mathbf{u})$ over policies $\mathbf{u} \in U_1$, where $\lambda > 0$,

$$\mathrm{var}(\mathbf{u}) = \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}\Big[(R_m - \phi(\mathbf{u}))^2\Big],$$

and U_1 is the class of all policies whose expected state-action frequencies converge (see Derman [7] or §2). The quantity var(**u**) captures the notion of variability in the following sense: first, the probabilistic variance of the reward at epoch m is taken with respect to the time-average expected reward $\phi(\mathbf{u})$; then the time average of the probabilistic variance is obtained.

The problem of maximizing $\phi(\mathbf{u}) - \lambda\,\mathrm{var}(\mathbf{u})$ over policies in U_1 was addressed in a more general context by Derman (see p. 94 of [7]). It follows from Derman that there exists a pure policy that maximizes $\phi(\mathbf{u}) - \lambda\,\mathrm{var}(\mathbf{u})$ over all policies $\mathbf{u} \in U_1$. However,

*Received October 20, 1987; revised October 8, 1990.
AMS 1980 subject classification. Primary· 90C40, 90C39.
IAOR 1973 subject classification. Main: Programming: Markov Decision
OR/MS Index 1978 subject classification Primary: 118 Dynamic Programming/Markov/Finite State Space.
Key words. Markov decision processes, finite state space, expecteded time average variability, time average expected variability
[†]Supported partially through NSF Grant NCR-8707620.

0364-765X/92/1703/0558/$01.25

Derman's existence result does not point to an algorithm which would locate an optimal policy. To this end, Filar *et al.* [8] give a mathematical program with linear constraints and quadratic objective function. They then give a condition, inspired by a result of Hordijk and Kallenberg [10], so that a stationary policy obtained from the optimal solution of the mathematical program is optimal for the original problem. However, the condition does not in general hold true, so that their algorithm may give rise to a strictly suboptimal stationary policy.

Sobel [16] considers the related problem of finding Pareto-optimal policies in the sense of a high $\phi(\mathbf{u})$ and low $\operatorname{var}(\mathbf{u})$. He optimizes over stationary policies and focuses his study on the unichain case. A parametric LP algorithm and a policy improvement algorithm are given, and both produce Pareto-optimal pure policies. See also [3], [4], [13] for other recent studies on variance sensitive MDPs.

In this paper we consider the following two criteria:

$$\nu(\mathbf{u}) = E_{\mathbf{u}}\left[\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} h\left(R_m, \frac{1}{n}\sum_{l=1}^{n} R_l\right)\right] \quad \text{and}$$

$$\kappa(\mathbf{u}) = \liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}\left[h\left(R_m, \phi_n(\mathbf{u})\right)\right], \quad \text{where}$$

$$\phi_n(\mathbf{u}) = \frac{1}{n}\sum_{l=1}^{n} E_{\mathbf{u}}[R_l].$$

The *variability function* $h(\cdot,\cdot)$ compares at each epoch m the current reward with the average reward over an interval that includes m. Throughout we assume that $h(\cdot,\cdot)$ is a continuous function. We shall refer to $\nu(\mathbf{u})$ as the *expected time-average variability* and to $\kappa(\mathbf{u})$ as the *time-average expected variability*. Loosely speaking, $\nu(\mathbf{u})$ places emphasis on the time-average variability and $\kappa(\mathbf{u})$ places emphasis on the probabilistic variability.

We will show that if $h(x, y) = x - \lambda(x - y)^2$, then the time-average expected variability $\kappa(\mathbf{u})$ satisfies $\kappa(\mathbf{u}) = \phi(\mathbf{u}) - \lambda \operatorname{var}(\mathbf{u})$ for all $\mathbf{u} \in U_1$. Thus for this choice of $h(\cdot,\cdot)$ the problem of maximizing $\kappa(\mathbf{u})$ over $\mathbf{u} \in U_1$ is equivalent to the problem considered by Filar *et al.* [8] and closely related to the problem considered by Sobel [16].

In §3 we show that the two criteria, $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$, are in general different. In particular, we show that $\nu(\mathbf{f}) \neq \kappa(\mathbf{f})$ can occur for a stationary policy $\mathbf{f}$ and that the optimal policy for $\nu(\mathbf{u})$ may be suboptimal for $\kappa(\mathbf{u})$, and vice versa. However, in the *unichain* case we show that $\nu(\mathbf{f}) = \kappa(\mathbf{f})$ for all stationary policies $\mathbf{f}$. This result along with our optimization results imply that the two criteria have the same optimal stationary policy for unichain MDPs.

In §4 we consider communicating MDPs. We first give an example which illustrates that there does not in general exist an optimal stationary policy for $\nu(\mathbf{u})$ or for $\kappa(\mathbf{u})$. We then construct a stationary policy that is ϵ-optimal for both $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$. The ϵ-optimal stationary policy can be directly obtained from the solution of a mathematical program with linear constraints and nonlinear objective function. Furthermore, for the case $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$ we construct a pure policy which is optimal for both $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$. This optimal pure policy can be directly obtained from the solutions of a parametric linear program.

In §5 we return to general multichain MDPs. We make use of the sample path and decomposition techniques of Ross and Varadarajan [14] to develop an algorithm which constructs an ϵ-optimal stationary policy for the expected time-average variability criterion $\nu(\mathbf{u})$. In the case $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$ this policy becomes an optimal pure policy for $\nu(\mathbf{u})$. However, these optimal and ϵ-optimal policies may be strictly suboptimal for the time-average expected variability criterion $\kappa(\mathbf{u})$.

In §6, we consider the problem of choosing both the policy and the initial state, which was first posed and solved by Sobel [16] for variance sensitive MDPs. We show that the sample path and decomposition techniques lead to an alternative solution to this problem. In particular, if $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$ then $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$ have the same optimal pure policy.

We conclude in §7 with a list of problems that remain open for time average MDPs with variability sensitive criteria.

2. Preliminaries.

Denote $\{X_n, n = 1, 2, \dots\}$ for the state process, which takes values in a finite state space $\mathscr{S}$. At each epoch n the decision maker chooses an action A_n from the finite action space $\mathscr{A}$. The underlying sample space $\Omega = \{\mathscr{S} \times \mathscr{A}\}^\infty$ consists of all possible realizations of states and actions. Throughout the sample space will be equipped with the σ-algebra generated by the random variables $(X_1, A_1, X_2, A_2, \dots)$. The initial state is assumed to be fixed and given. Denote p_{xay}, $x \in \mathscr{S}$, $a \in \mathscr{A}$, $y \in \mathscr{S}$, for the law of motion of the MDP, i.e., for all policies $\mathbf{u}$ and all epochs n

$$P_{\mathbf{u}}(X_{n+1} = y | X_1, A_1, \dots, X_{n-1}, A_{n-1}, X_n = x, A_n = a) = p_{xay}.$$

A policy $\mathbf{f}$ is said to be a *stationary* if the choice of action depends only on the current state of the process; denote $f(x, a)$ for the probability of choosing action a when in state x. A stationary policy is said to be *pure* if for each $x \in \mathscr{S}$ there is one action $a \in \mathscr{A}$ such that $f(x, a) = 1$. Let U, F, G denote the classes of all policies, stationary policies, and pure policies, respectively; clearly, $G \subset F \subset U$.

Under any stationary policy $\mathbf{f}$, the state process $\{X_n, n = 1, 2, \dots\}$ is a Markov chain with transition matrix $\mathbf{P(f)}$ whose components are given by

$$P_{xy}(\mathbf{f}) = \sum_{a \in \mathscr{A}} p_{xay} f(x, a).$$

A transition matrix $\mathbf{P(f)}$ is said to be *unichain* if it has at most one recurrent class plus (a perhaps empty) set of transient states. An MDP is *unichain* if $\mathbf{P(g)}$ is unichain for all pure policies $\mathbf{g}$. An MDP is *communicating* if $\mathbf{P(f)}$ is irreducible for all stationary policies that satisfy $f(x, a) > 0$, $x \in \mathscr{S}$, $a \in \mathscr{A}$ (see [1] or [14]).

For each $x \in \mathscr{S}$ and $a \in \mathscr{A}$ define the random variables denoting the state-action frequencies through epoch n as

$$Z_n(x, a) = \frac{1}{n} \sum_{m=1}^{n} 1(X_m = x, A_m = a),$$

where $1(\cdot)$ denotes the indicator function. Let U_0 denote the class of all policies $\mathbf{u}$ such that $\{Z_n(x, a), n = 1, 2, \dots\}$ converges $P_{\mathbf{u}}$-almost surely for all $x \in \mathscr{S}$ and $a \in \mathscr{A}$. Thus, if $\mathbf{u} \in U_0$, then there exist random variables $\{Z(x, a)\}$ such that

$$\lim_{n \to \infty} Z_n(x, a) = Z(x, a)$$

$P_{\mathbf{u}}$-almost surely for all $x \in \mathscr{S}$, $a \in \mathscr{A}$. Let U_1 be the class of all policies $\mathbf{u}$ such that the expected state-action frequencies $\{E_{\mathbf{u}}[Z_n(x,a)],\ n = 1,2,\dots\}$ converge for all x and a. For $\mathbf{u} \in U_1$ denote

$$z_{\mathbf{u}}(x,a) = \lim_{n\to\infty} E_{\mathbf{u}}[Z_n(x,a)].$$

From Lebesgue's Dominated Convergence Theorem we have $U_0 \subset U_1$. Since under any stationary policy $\{(X_n, A_n),\ n = 1,2,\dots\}$ is a homogeneous Markov chain, it follows (e.g., see Çinlar [5]) that the limit of $Z_n(x,a)$ exists $P_{\mathbf{u}}$-almost surely for all x and a. Hence, $F \subset U_0$, so that $G \subset F \subset U_0 \subset U_1 \subset U$.

Denote $r(x,a)$ for the reward obtained when the state is x and action a is chosen. Thus the reward obtained at epoch n is $R_n = r(X_n, A_n)$. Recall the definition of the expected time-average variability $\nu(\mathbf{u})$ and the time-average expected variability $\kappa(\mathbf{u})$ given in the Introduction. Let

$$\nu = \sup_{\mathbf{u}\in U} \nu(\mathbf{u}), \quad \kappa = \sup_{\mathbf{u}\in U} \kappa(\mathbf{u}).$$

A policy $\mathbf{u}$ is optimal for $\nu(\cdot)$ if $\nu(\mathbf{u}) = \nu$. For a fixed $\epsilon > 0$, a policy $\mathbf{u}$ is ϵ-optimal for $\nu(\cdot)$ if $\nu(\mathbf{u}) > \nu - \epsilon$. Optimality and ϵ-optimality for $\kappa(\cdot)$ are defined in an analogous fashion.

3. Notions of variability.

3. Notions of variability. We first consider the time-average expected variability $\kappa(\mathbf{u})$. The following proposition shows that $\kappa(\mathbf{u})$ can be conveniently expressed in terms of the long-run expected state-action frequencies $\{z_{\mathbf{u}}(x,a)\}$.

PROPOSITION 1. *For all* $\mathbf{u} \in U_1$,

$$\phi(\mathbf{u}) = \sum_{x,a} r(x,a) z_{\mathbf{u}}(x,a) \quad \textit{and} \tag{1}$$

$$\kappa(\mathbf{u}) = \sum_{x,a} h[r(x,a), \phi(\mathbf{u})] z_{\mathbf{u}}(x,a). \tag{2}$$

Moreover, if $h(x,y) = x - \lambda(x-y)^2$ *then for all* $\mathbf{u} \in U_1$,

$$\kappa(\mathbf{u}) = \phi(\mathbf{u}) - \lambda \operatorname{var}(\mathbf{u}).$$

PROOF. Fix a policy $\mathbf{u} \in U_1$. It is straightforward to establish (1) and that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}[h(R_m, \phi(\mathbf{u}))] = \sum_{x,a} h[r(x,a), \phi(\mathbf{u})] z_{\mathbf{u}}(x,a). \tag{3}$$

Let $\Lambda = \{r(x,a) \colon x \in \mathscr{S},\ a \in \mathscr{A}\}$. We have

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} E_{\mathbf{u}} \left| \sum_{m=1}^{n} h(R_m, \phi_n(\mathbf{u})) - \sum_{m=1}^{n} h(R_m, \phi(\mathbf{u})) \right| \\ &\leqslant \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \max_{r\in\Lambda} |h(r, \phi_n(\mathbf{u})) - h(r, \phi(\mathbf{u}))| \\ &= \lim_{n\to\infty} \max_{r\in\Lambda} |h(r, \phi_n(\mathbf{u})) - h(r, \phi(\mathbf{u}))| = 0, \end{aligned} \tag{4}$$

where the last equality follows from the continuity of $h(\cdot, \cdot)$. Combining (3) and (4) gives (2). The last statement follows directly from (2). □

It follows from Proposition 1 that if $h(x, y) = x - \lambda(x - y)^2$, then the problem of maximizing $\kappa(\mathbf{u})$ over $\mathbf{u} \in U_1$ is equivalent to the problem considered by Filar *et al.* Other interesting choices for $h(\cdot, \cdot)$ include

$$h(x, y) = x - \sum_{k=1}^{K} \lambda_k |x - y|^k,$$

and continuous approximations of

$$h(x, y) = \begin{cases} x, & x \geqslant y - \alpha, \\ x - \lambda & \text{otherwise}, \end{cases}$$

where $\lambda > 0$, $\alpha > 0$. Note that the first variability function takes into account higher moments, whereas the second variability function attempts to make the average expected reward $\phi(\mathbf{u})$ high while keeping the current reward R_m above $\phi(\mathbf{u}) - \alpha$ with high probability.

Now consider the expected time-average variability criterion $\nu(\mathbf{u})$. The following proposition shows that the time-average variability can be conveniently expressed in terms of the long-run state-action frequencies $\{Z(x, a)$: $x \in \mathscr{S}$, $a \in \mathscr{A}\}$. Its proof is similar to that of Proposition 1 and will be omitted. Let

$$\bar{R}_n = \frac{1}{n} \sum_{l=1}^{n} R_l.$$

PROPOSITION 2. *For all* $\mathbf{u} \in U_0$ *we have*

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} h\left(R_m, \bar{R}_n\right) = \sum_{x, a} h\left[r(x, a), \sum_{y, b} r(y, b) Z(y, b)\right] Z(x, a)$$

$P_{\mathbf{u}}$*-almost surely. If* $h(x, y) = x - \lambda(x - y)^2$, *then for* $\mathbf{u} \in U_0$ *we have*

$$\nu(\mathbf{u}) = \phi(\mathbf{u}) - \lambda \lim_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}\left[\left(R_m - \bar{R}_n\right)^2\right].$$

We now compare the expected time-average variability $\nu(\mathbf{u})$ with the time-average expected variability $\kappa(\mathbf{u})$. We first show that the two criteria can be quite different.

EXAMPLE 1. Consider an MDP with state space $\{0, 1, 2\}$. Let the initial state be 0 and let the states 1 and 2 be absorbing. Let there be two possible actions to choose from when in state 0: action a, under which the process moves to state 1 with probability 1; action b, under which the process moves to states 1 and 2 with probabilities 0.1 and 0.9, respectively. Let the (single-stage) rewards for states 1 and 2 be equal to 0 and 10, respectively. Let $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$.

There are exactly 2 pure policies for this MDP: $\mathbf{g}_a$ which chooses action a when in state 0; $\mathbf{g}_b$ which chooses action b when in state 0. It is easily seen that $\nu(\mathbf{g}_a) = \kappa(\mathbf{g}_a) = 0 - \lambda \cdot 0 = 0$. But $\nu(\mathbf{g}_b) = 9 - \lambda \cdot 0 = 9$ and $\kappa(\mathbf{g}_b) = 9 - \lambda \cdot 9$. Hence, $\nu(\mathbf{g}_b) \neq \kappa(\mathbf{g}_b)$.

Now suppose that $\lambda > 1$. Then for the expected time-average variability criterion, $\nu(\mathbf{u})$, the optimal policy is $\mathbf{g}_b$, which produces a constant stream of 10s with probability 0.9 and produces a constant stream of 0s with probability 0.1. For the time-average expected variability criterion, $\kappa(\mathbf{u})$, the optimal policy is $\mathbf{g}_a$, which produces a constant stream of 0s with probability 1. Hence, the two criteria lead to different optimal policies. □

Propositions 3 and 4 shed additional insight on the relationship between the criteria $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$. The proof of Proposition 4 is straightforward and omitted.

PROPOSITION 3. *Suppose* $h(x, y) = x - \lambda(x - y)^2$ *for some* $\lambda > 0$ (*which corresponds to the criterion of Filar et al.* [8] *and is closely related to the criterion of Sobel* [16]). *Then* $\nu(\mathbf{u}) \geqslant \kappa(\mathbf{u})$ *for all* $\mathbf{u} \in U_0$.

PROOF. Employing Propositions 1 and 2 it is straightforward to show that for all $\mathbf{u} \in U_0$

$$(5) \qquad \kappa(\mathbf{u}) = \phi(\mathbf{u}) - \lambda \sum_{x,a} r^2(x,a) z_{\mathbf{u}}(x,a) + \lambda\phi^2(\mathbf{u}),$$

and that

$$(6) \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} h(R_m, \bar{R}_n)$$

$$= \sum_{x,a} r(x,a)Z(x,a) - \lambda \sum_{x,a} r^2(x,a)Z(x,a) + \lambda\left[\sum_{x,a} r(x,a)Z(x,a)\right]^2$$

$P_{\mathbf{u}}$-almost surely. Taking the expectation of (6), employing Jensen's inequality, and comparing with (5) gives $\nu(\mathbf{u}) \geqslant \kappa(\mathbf{u})$. □

PROPOSITION 4. *Let* $\mathbf{f}$ *be a stationary policy and let* $\mathscr{R}_1, \ldots, \mathscr{R}_q$ *be the recurrent classes associated with* $\mathbf{P}(\mathbf{f})$. *Denote* $(\pi_i^x(\mathbf{f})\colon x \in \mathscr{R}_i)$ *for the equilibrium probability vector associated with class* i, $i = 1, \ldots, q$. *Further denote*

$$\psi_i = \sum_{x,a} r(x,a)\pi_i^x(\mathbf{f})f(x,a).$$

Then

$$\kappa(\mathbf{f}) = \sum_{i=1}^{q} P_{\mathbf{f}}(X_n \in \mathscr{R}_i \ a.a.) \sum_{x,a} h\left[r(x,a), \sum_{j=1}^{q} P_{\mathbf{f}}(X_n \in \mathscr{R}_j \ a.a.)\psi_j\right]\pi_i^x(\mathbf{f})f(x,a),$$

$$\nu(\mathbf{f}) = \sum_{i=1}^{q} P_{\mathbf{f}}(X_n \in \mathscr{R}_i \ a.a.) \sum_{x,a} h[r(x,a), \psi_i]\pi_i^x(\mathbf{f})f(x,a),$$

where a.a. abbreviates almost always. If the MDP is unichain (*in which case* $q = 1$ *and we remove the subscript* i), *then*

$$\nu(\mathbf{f}) = \kappa(\mathbf{f}) = \sum_{x,a} h[r(x,a), \phi(\mathbf{f})]\pi^x(\mathbf{f})f(x,a).$$

3.1. *Decomposition and sample path theory*. At this juncture it is convenient to collect some results that will be needed in the subsequent sections.

A set $\mathscr{C} \subseteq \mathscr{S}$ is said to be a *strongly communicating class* if: (i) $\mathscr{C}$ is a recurrent class for some stationary policy; (ii) $\mathscr{C}$ is not a proper subset of some $\mathscr{C}'$ for which (i) holds true. Let $\{\mathscr{C}_1, \ldots, \mathscr{C}_I\}$ be the collection of all strongly communicating classes. Let $\mathscr{T}$ be the (possibly empty) set of states that are transient under all stationary policies. It is shown in [14] that $\{\mathscr{C}_1, \ldots, \mathscr{C}_I, \mathscr{T}\}$ forms a partition of the state space $\mathscr{S}$ (see also Bather [2] for a related decomposition). For each $i = 1, \ldots, I$, denote for each $x \in \mathscr{C}_i$ the set

$$\mathscr{F}_x = \{a \in \mathscr{A}: p_{xay} = 0 \text{ for all } y \notin \mathscr{C}_i\}.$$

The following result is also proved in [14].

PROPOSITION 5. *For all policies* $\mathbf{u}$ *we have*

$$\sum_{i=1}^{I} P_{\mathbf{u}}(X_n \in \mathscr{C}_i \text{ a.a.}) = 1 \quad \text{and} \tag{7}$$

$$P_{\mathbf{u}}(A_n \in \mathscr{F}_{X_n} \text{ a.a.}) = 1. \tag{8}$$

For each $i = 1, \ldots, I$, define a new MDP, called MDP-i, as follows: the state space is $\mathscr{C}_i$; for each $x \in \mathscr{C}_i$, the set of available actions is given by the state dependent action spaces $\mathscr{F}_x$; the law of motion p_{xay} and reward function $r(x, a)$ are the same as for the original MDP but restricted to $\mathscr{C}_i$ and to the state dependent action spaces $\mathscr{F}_x$, $x \in \mathscr{C}_i$. It is straightforward to show that MDP-i is a communicating MDP for all $i = 1, \ldots, I$. For MDP-i, let $\nu_i(\mathbf{u})$ be the expected time-average variability under policy $\mathbf{u}$.

For each $i = 1, \ldots, I$, consider the following mathematical program with decision variables $z(x, a)$, $a \in \mathscr{F}_x$, $x \in \mathscr{C}_i$. Let $\delta_{xy} := 1$ if $x = y$ and $\delta_{xy} = 0$ otherwise.

Program T_i.

$$t_i = \max \sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{F}_x} h\left[r(x,a), \sum_{y \in \mathscr{C}_i} \sum_{b \in \mathscr{F}_y} r(y,b) z(y,b)\right] z(x,a)$$

s.t.

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{F}_x} (\delta_{xy} - p_{xay}) z(x,a) = 0, \quad y \in C_i,$$

$$\sum_{x \in \mathscr{C}_i} \sum_{a \in \mathscr{F}_x} z(x,a) = 1,$$

$$z(x,a) \geqslant 0, \qquad a \in \mathscr{F}_x, \quad x \in \mathscr{C}_i.$$

For each $\eta \geqslant 0$, we will also need to refer to the following mathematical program with decision variables $z(x, a)$, $x \in \mathscr{S}$, $a \in \mathscr{A}$.

Program Q^η.

$$q^\eta = \max \sum_{x \in \mathscr{S}} \sum_{a \in \mathscr{A}} h\left[r(x,a), \sum_{y \in \mathscr{S}} \sum_{b \in \mathscr{A}} r(y,b) z(y,b)\right] z(x,a)$$

s.t.

$$\sum_{x \in \mathscr{S}} \sum_{a \in \mathscr{A}} (\delta_{xy} - p_{xay}) z(x,a) = 0, \qquad y \in \mathscr{S},$$

$$\sum_{x \in \mathscr{S}} \sum_{a \in \mathscr{A}} z(x,a) = 1,$$

$$z(x,a) \geqslant \eta, \qquad x \in \mathscr{S}, \quad a \in \mathscr{A}.$$

We will refer to the feasible regions of Program T_i and Program Q^η simply as T_i and Q^η, respectively. Note that the objective functions for both sets of the mathematical programs are continuous functions over polytopes. Also note that T_i, $i = 1, \dots, I$, and that Q^0 are nonempty.

The following lemma provides bounds on $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$. Only an outline of the proof will be given since it is similar to the proof of Proposition 2 of [15].

LEMMA 1. (i) *For all $i = 1, \dots, I$ and for all policies* $\mathbf{u}$ *we have*

$$P_{\mathbf{u}}\left(\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} h(R_m, \bar{R}_n) \leqslant t_i | X_n \in \mathscr{C}_i \text{ a.a.}\right) = 1; \tag{9}$$

consequently,

$$\nu(\mathbf{u}) \leqslant \sum_{i=1}^{I} t_i P_{\mathbf{u}}(X_n \in \mathscr{C}_i \text{ a.a.}) \tag{10}$$

for all policies $\mathbf{u}$. (ii) $\nu(\mathbf{u}) \leqslant q^0$ *for all policies* $\mathbf{u}$; (iii) $\kappa(\mathbf{u}) \leqslant q^0$ *for all policies* $\mathbf{u}$.

PROOF. Fix a policy $\mathbf{u}$. Since $0 \leqslant Z_n(x, a) \leqslant 1$, it follows from standard compactness arguments that for each $\omega \in \Omega$ there exists a subsequence $\{N_k(\omega), k = 1, 2, \dots\}$ along which $\{Z_n(x, a; \omega), n = 1, 2, \dots\}$ converges to some $W(x, a; \omega)$ for all $x \in \mathscr{S}$, $a \in \mathscr{A}$.With the aid of (8) it can then be shown that $W(x, a)$, $x \in \mathscr{C}_i$, $a \in \mathscr{F}_x$ is a feasible solution to Program T_i on the set $\Phi = \{X_n \in \mathscr{C}_i \text{ a.a.}\} - \Gamma$, where Γ is a set of $P_{\mathbf{u}}$-measure zero. Thus,

$$\sum_{x\in\mathscr{C}_i}\sum_{a\in\mathscr{F}_x} h\left[r(x, a), \sum_{y\in\mathscr{C}_i}\sum_{b\in\mathscr{F}_y} r(y, b)W(y, b)\right]W(x, a) \leqslant t_i \tag{11}$$

on Φ. We also have

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} h(R_m, \bar{R}_n) \\ &\quad\leqslant \lim_{k\to\infty} \frac{1}{N_k} \sum_{m=1}^{N_k} h(R_m, \bar{R}_{N_k}) \\ &\quad= \lim_{k\to\infty} \frac{1}{N_k} \sum_{m=1}^{N_k} h\left(R_m, \lim_{k\to\infty} \bar{R}_{N_k}\right) \\ &\quad= \sum_{x\in\mathscr{C}_i}\sum_{a\in\mathscr{F}_x} h\left[r(x, a), \sum_{y\in\mathscr{C}_i}\sum_{b\in\mathscr{F}_y} r(y, b)W(y, b)\right]W(x, a) \end{aligned} \tag{12}$$

on Φ. Combining (11) and (12) gives (9), and combining (9) with Proposition 5 gives (10). A similar argument leads to

$$P_{\mathbf{u}}\left(\liminf_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} h(R_m, \bar{R}_n) \leqslant q^0\right) = 1,$$

from which (ii) follows. The proof for (iii) is also similar to (i) except we use a limit point of $\{\mathrm{E}_{\mathbf{u}}[Z_n(x,a)]: n = 1,2,\dots\}$ in place of $W(x,a)$. □

4. The communicating case.

Throughout this section we assume that the MDP is communicating. This implies that there is only one strongly communicating class and that $\mathscr{S} = \mathscr{C}_1$. The analysis of this section draws on results and observations from [15].

The following example illustrates some of the subtleties that can arise in the communicating case.

EXAMPLE 2. Consider the MDP with state space $\{0,1,2\}$ depicted in Figure 1. In the middle state 1 there is only one action available: it brings the state process to state 0 with probability 0.5 and to state 2 with probability 0.5. States 0 and 2 each have two actions: action 1 keeps the process in the current state with probability 1; action 2 moves the process to the middle state 1 with probability 1. Let $r(x,a) = x$ for $x = 0,1,2$ and let the initial state be state 0. Let the variability function be given by $h(x,y) = (x-y)^2$ so that we are *maximizing* the variance. It is not difficult to see that $\nu(\mathbf{f}) = \kappa(\mathbf{f}) < 1$ for all stationary policies $\mathbf{f}$, and that $\sup_{\mathbf{f} \in F} \nu(\mathbf{f}) = 1$. Thus there does not in general exist an optimal stationary policy for either $\nu(\mathbf{u})$ or $\kappa(\mathbf{u})$. □

We can, however, obtain several positive results. To this end define for each $\mathbf{z}$ belonging to Q^0:

$$g(\mathbf{z}) = \sum_{x \in \mathscr{S}} \sum_{a \in \mathscr{A}} h\left[r(x,a), \sum_{y \in \mathscr{S}} \sum_{b \in \mathscr{A}} r(y,b) z(y,b)\right] z(x,a) \quad \text{and}$$

$$I_{\mathbf{z}} = \{x \in \mathscr{S}: z(x,a) > 0 \text{ for some } a \in \mathscr{A}\}.$$

For each $\mathbf{z}$ belonging to Q^0 construct a stationary policy $\mathbf{f}$ as follows: If $x \in I_{\mathbf{z}}$, let

$$f(x,a) = \frac{z(x,a)}{\sum_{a \in \mathscr{A}} z(x,a)};$$

if $x \notin I_{\mathbf{z}}$ choose actions in a deterministic way so that the state process eventually enters $I_{\mathbf{z}}$ (this can be done since the MDP is communicating).

LEMMA 2. *Let* $\mathbf{z}$ *be a feasible solution for Program* Q^0 *and let* $\mathbf{f}$ *be defined as above. If* $\mathbf{P}(\mathbf{f})$ *is unichain, then* $\nu(\mathbf{f}) = \kappa(\mathbf{f}) = g(\mathbf{z})$.

PROOF. Suppose that $\mathbf{P}(\mathbf{f})$ is unichain. A standard argument gives $z(x,a) = \pi^x(\mathbf{f}) f(x,a)$ for all $x \in \mathscr{S}$, $a \in \mathscr{A}$, where $[\pi^x(\mathbf{f}): x \in \mathscr{S}]$ is the equilibrium vector

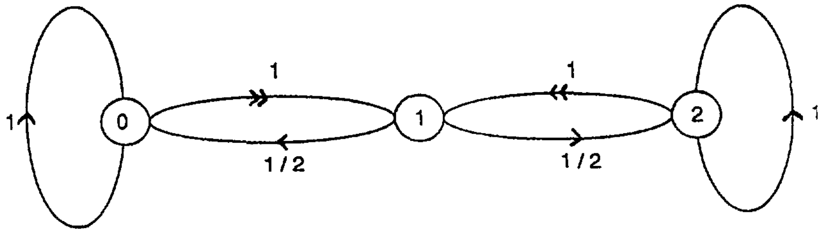

FIGURE 1. A Communicating MDP Which Does Not Have an Optimal Stationary Policy.

associated with $\mathbf{P}(\mathbf{f})$. Thus

$$\begin{aligned} g(\mathbf{z}) &= \sum_{x\in\mathscr{S}}\sum_{a\in\mathscr{A}} h\left[r(x,a),\ \sum_{y\in\mathscr{S}}\sum_{b\in\mathscr{A}} r(y,b)z(y,b)\right]z(x,a) \\ &= \sum_{x\in\mathscr{S}}\sum_{a\in\mathscr{A}} h\left[r(x,a),\ \sum_{y\in\mathscr{S}}\sum_{b\in\mathscr{A}} r(y,b)\pi^{y}(\mathbf{f})f(y,b)\right]\pi^{x}(\mathbf{f})f(x,a) \\ &= \nu(\mathbf{f}) = \kappa(\mathbf{f}), \end{aligned}$$

where the last two equalities follow from Proposition 4. □

For a communicating MDP, Q^{η} is nonempty for all $\eta \in [0,\delta]$ for some $\delta > 0$. Now for each $\eta \in [0,\delta]$, let $\mathbf{z}^{\eta}$ be an optimal solution to Program Q^{η}. If there is an optimal extreme point solution to Program Q^{0}, further require $\mathbf{z}^{0}$ to be an extreme point. For each $\eta \in [0,\delta]$, let $\mathbf{f}^{\eta}$ be defined from $\mathbf{z}^{\eta}$ according to the above transformation.

THEOREM 1. *Fix $\epsilon > 0$. If the MDP is communicating, then for $\eta > 0$ sufficiently small the stationary policy $\mathbf{f}^{\eta}$ is ϵ-optimal for $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$. If, in addition, $h(x,y) = x - \lambda(x-y)^2$ with $\lambda > 0$, then the policy $\mathbf{f}^{0}$ is a pure policy and is optimal for both $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$.*

PROOF. In order to prove the first statement we note that $\mathbf{P}(\mathbf{f}^{\eta})$ is unichain for all $\eta \in (0,\delta]$. Thus, by Lemma 2,

$$\nu(\mathbf{f}^{\eta}) = \kappa(\mathbf{f}^{\eta}) = g(\mathbf{z}^{\eta}) \tag{13}$$

for all $\eta \in (0,\delta]$. Also note that since $g(\mathbf{z})$ is continuous over Q^{0}, there is a neighborhood of $\mathbf{z}^{0}$ such that $g(\mathbf{z}) > q^{0} - \epsilon$ for all $\mathbf{z}$ in the neighborhood. It is easily seen that there exists a $\mathbf{z}'$ that belongs to this neighborhood and to Q^{γ} for some $\gamma > 0$ sufficiently small. Thus

$$g(\mathbf{z}^{\gamma}) \geqslant g(\mathbf{z}') > q^{0} - \epsilon. \tag{14}$$

Combining (13) and (14) with Lemma 1 establishes the ϵ-optimality of $\mathbf{f}^{\gamma}$ for $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$.

In order to prove the second statement, we note that for this choice of $h(\cdot,\cdot)$ the objective function $g(\mathbf{z})$ is convex over Q^{0}. Therefore, we may choose $\mathbf{z}^{0}$ as an optimal extreme point for Program Q^{0}. This property implies that $\mathbf{f}^{0}$ is a pure policy and that $\mathbf{P}(\mathbf{f}^{0})$ is unichain (e.g., see [6]). The proof is then completed by again invoking Lemmas 1 and 2. □

It follows from the above proof that if $\mathbf{P}(\mathbf{f}^{0})$ is unichain, then $\mathbf{f}^{0}$ is optimal for both $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$ (but is not necessarily pure). This condition will always be satisfied if the MDP itself is unichain. It also follows from the above proof that if $g(\mathbf{z})$ is convex over Q^{0}, then the policy $\mathbf{f}^{0}$ is pure and optimal for $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$.

From Theorem 1 we see that the two criteria, $\nu(\mathbf{u})$ and $\kappa(\mathbf{u})$, are quite similar for communicating MDPs as they share the same ϵ-optimal (or optimal) policy. But it is interesting to note that we still may have $\nu(\mathbf{g}) \neq \kappa(\mathbf{g})$ for a pure policy $\mathbf{g}$ for a communicating MDP. Indeed in Example 2, if the initial state is the middle state 1 and if $\mathbf{g}$ chooses action 1 at both of the side states 0 and 2, then $0 = \nu(\mathbf{g}) \neq \kappa(\mathbf{g}) = \frac{1}{2}$.

5. Multichain MDPs.

In this section we impose no restrictions on the law of motion p_{xay}, $x \in \mathscr{S}$, $a \in \mathscr{A}$, $y \in \mathscr{S}$. We now construct a stationary policy that is

ϵ-optimal for $\nu(\mathbf{u})$. For the case of $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, the policy becomes an optimal pure policy. Since many of the arguments are similar to those in [14], only outlines of the proofs will be given.

Recall that MDP-i is communicating. By Theorem 1 we can therefore construct for each $i = 1, \ldots, I$ an ϵ-optimal stationary policy $\mathbf{f}_i$ for MDP-i. In the case of $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, $\mathbf{f}_i$ becomes an optimal pure policy for each $i = 1, \ldots, I$. Recall that t_i is the value of Program T_i, $i = 1, \ldots, I$. We also need to consider the problem of finding a policy that maximizes the following time-average expected reward:

$$\beta(\mathbf{u}) = \liminf_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} E_{\mathbf{u}}\left[\sum_{i=1}^{I} t_i 1(X_n \in \mathscr{C}_i)\right].$$

We refer to this problem as the *intermediate MDP*. It is well known that there exists an optimal pure policy $\mathbf{g}^*$ for this problem which can be found by policy improvement, value iteration, or linear programming. Let

$$H = \{i \colon \mathscr{C}_i \text{ contains a recurrent class under } \mathbf{P}(\mathbf{g}^*)\}.$$

Modify $\mathbf{g}^*$ so that $\mathscr{C}_i$ is closed for each $i \in H$ and so that $\mathbf{g}^*$ remains optimal for the intermediate MDP (see [14]).

We now construct a stationary policy $\mathbf{f}^*$ as follows: when in state $x \in \mathscr{C}_i$, $i \in H$, apply $\mathbf{f}_i$; otherwise, apply $\mathbf{g}^*$. Our main result is

THEOREM 2. *The stationary policy* $\mathbf{f}^*$ *is* ϵ*-optimal for* $\nu(\mathbf{u})$. *Moreover, if* $h(x, y) = x - \lambda(x - y)^2$ *with* $\lambda > 0$, *then the policy* $\mathbf{f}^*$ *is a pure policy and it is optimal for* $\nu(\mathbf{u})$.

PROOF. Employing (7) it can be shown that

$$\beta(\mathbf{u}) = \sum_{i=1}^{I} t_i P_{\mathbf{u}}(X_n \in \mathscr{C}_i \text{ a.a.})$$

for all policies $\mathbf{u}$. Thus, from Lemma 1 we have

$$\nu(\mathbf{u}) \leqslant \beta(\mathbf{g}^*) \tag{15}$$

for all policies $\mathbf{u}$. From Proposition 4 and the construction of $\mathbf{f}^*$ we have

$$\nu(\mathbf{f}^*) = \sum_{i=1}^{I} \nu_i(\mathbf{f}_i) P_{\mathbf{g}^*}(X_n \in \mathscr{C}_i \text{ a.a.}). \tag{16}$$

Combining (15) and (16) with Theorem 1 gives the desired results. □

5.1. *Computational considerations.* In order to construct the ϵ-optimal (respectively optimal) stationary (respectively pure) policy $\mathbf{f}^*$ for the expected time-average variability criterion we can use the following recipe.

1. Determine the strongly communicating classes $\mathscr{C}_i$, $i = 1, \ldots, I$.
2. Solve Program T_i for $i = 1, \ldots, I$ and obtain policies $\mathbf{f}_i$, $i = 1, \ldots, I$ and optimal values t_i, $i = 1, \ldots, I$.
3. Solve the intermediate MDP and obtain $\mathbf{g}^*$ and H. Then combine $\mathbf{g}^*$ with $\mathbf{f}_i$, $i \in H$, to get the ϵ-optimal (or optimal) policy $\mathbf{f}^*$.

Step 1 can be done with a graph-theoretic algorithm that is outlined in [14]. Its worst case complexity is $O(|\mathscr{S}|^3|\mathscr{A}|)$.

The intermediate MDP problem in Step 3 can be solved by standard MDP algorithms. Computational savings can be obtained, however, by observing that the reward function for the intermediate MDP is *constant* over $\mathscr{C}_i$ for each $i = 1, \dots, I$. This enables one to aggregate all states in $\mathscr{C}_i$ into one state for each $i = 1, \dots, I$ and solve an MDP with $I + |\mathscr{T}|$ states with one of the standard algorithms. This *aggregated MDP* is discussed in [14].

Now consider the problem of solving Program T_i for a fixed i. In general the objective function of Program T_i can have numerous local maxima within the feasible region, in which case the problem is difficult to solve. However, Katoh and Ibaraki [12] have shown that if the objective function for this class of mathematical programs possesses a certain quasi-convexity property, then the problem of solving Program T_i becomes quite tractable. (See also Sobel [16] for a related approach; also see the references within [12].) For example, suppose that $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$. Then the objective function of Program T_i takes the form

$$g(\mathbf{z}) = f_1(\mathbf{z}) + \lambda f_2^2(\mathbf{z}), \quad \text{where}$$

$$f_1(\mathbf{z}) = \sum_{x,a} \left[r(x,a) - \lambda r^2(x,a)\right] z(x,a),$$

$$f_2(\mathbf{z}) = \sum_{x,a} r(x,a) z(x,a).$$

For $\gamma \geqslant 0$ let

$$\begin{aligned} s_\gamma(\mathbf{z}) &= f_1(\mathbf{z}) + \gamma f_2(\mathbf{z}) \\ &= \sum_{x,a} \left[r(x,a) - \lambda r^2(x,a) + \gamma r(x,a)\right] z(x,a), \end{aligned}$$

and consider the problem of maximizing $s_\gamma(\mathbf{z})$ over T_i, henceforth referred to as problem $P(\gamma)$. Since $g(\cdot)$ is a convex function of $f_1(\cdot)$ and $f_2(\cdot)$, it follows from Theorem 2 of [12] that there exists a $\gamma > 0$ such that the optimal $\mathbf{z}$ for $P(\gamma)$ is also optimal for Program T_i. Thus, we can find an optimal solution to Program T_i by solving $P(\gamma)$ for all $\gamma \geqslant 0$ and checking to see which of these optimal solutions maximizes $g(\mathbf{z})$. But for fixed γ, $P(\gamma)$ is a linear program! Moreover, $P(\gamma)$ can be readily solved for all $\gamma \geqslant 0$ with parametric linear programming.

6. Choice-of-initial-state optimality.

In many MDP applications, the decision maker not only chooses a policy $\mathbf{u}$, but also chooses the initial state $x \in \mathscr{S}$ in order to maximize the objective. In the context of risk-sensitive time-average MDPs, Sobel [16] considers this problem and gives an example from inventory theory. We now show how the decomposition and sample path theory lead to an alternative approach to solve this problem.

Denote $P_{\mathbf{u}}^x$ for the probability measure corresponding to policy $\mathbf{u}$ and initial state x. In order to emphasize the dependence on the initial state $x \in \mathscr{S}$, we write in this section $\nu^x(\mathbf{u})$ and $\kappa^x(\mathbf{u})$ for the two variability criteria. Let j be such that

$$t_j = \max\{t_i : i = 1, \dots, I\},$$

and recall the definition of the stationary (possibly pure) policy $\mathbf{f}_j$ for MDP-j. Extend the definition of $\mathbf{f}_j$ to the original MDP by defining $\mathbf{f}_j$ to be deterministic but otherwise arbitrary in $\mathscr{S} - \mathscr{C}_j$.

THEOREM 3. *Suppose the decision maker can choose the initial state $x \in \mathscr{S}$ as well as the policy $\mathbf{u} \in U$. Then the stationary policy $\mathbf{f}_j$ along with any initial state $\Delta \in \mathscr{E}_j$ is ϵ-optimal for $\nu^x(\mathbf{u})$. Moreover, if $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, then $\mathbf{f}^*$ is pure and optimal for $\nu^x(\mathbf{u})$ and $\kappa^x(\mathbf{u})$.*

PROOF. From Lemma 1 we have

$$\nu^x(\mathbf{u}) \leqslant \sum_{i=1}^{I} t_i P_{\mathbf{u}}^x(X_t \in \mathscr{E}_i \text{ a.a.}) \leqslant t_j \tag{17}$$

for all policies $\mathbf{u}$ and all initial states $x \in \mathscr{S}$. But from Theorem 1 we have $\nu^{\Delta}(\mathbf{f}_j) \geqslant t_j - \epsilon$ for all $\Delta \in \mathscr{E}_j$. This establishes the first statement. We have from Theorem 1 that if $h(x, y) = x - \lambda(x - y)^2$, with $\lambda > 0$, then $\mathbf{f}_j$ is pure and $\nu^{\Delta}(\mathbf{f}_j) = \kappa^{\Delta}(\mathbf{f}_j) = t_j$ for all $\Delta \in \mathscr{E}_j$. Combining this with (17) and Proposition 3 establishes the second statement. □

Note that the choice-of-initial-state problem is easier to solve than the original problem of maximizing $\nu(\mathbf{u})$ over $\mathbf{u} \in U$ since it is not required to solve an intermediate MDP. This technique can also be employed to solve the choice-of-initial-state problem for MDPs with sample path constraints.

7. Conclusion. Let us now take stock of what we know about time-average MDPs with variability sensitive criteria. Throughout this discussion we assume that the initial state is fixed and given.

First consider the time-average expected variability $\kappa(\mathbf{u})$. In general there does not exist an ϵ-optimal stationary policy for $\kappa(\mathbf{u})$. (This has not been shown. The claim can be verified by considering Example 1 of [15] with $h(x, y) = (x - y)^2$.) But if the variability function takes the specific form $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, then there exists an optimal pure policy, albeit the only known algorithm to locate it is complete enumeration of all pure policies. If the MDP is either communicating or unichain and if $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, then there exists an optimal pure policy which can be obtained from the solution of a parametric LP. For general $h(\cdot, \cdot)$ an optimal (respectively ϵ-optimal) policy can be found for unichain (respectively communicating) MDPs by solving the mathematical program Q^0 (respectively mathematical programs Q^{η}, $\eta \geqslant 0$).

Now consider the expected time-average variability $\nu(\mathbf{u})$. In general there exists an ϵ-optimal stationary policy which can be obtained from the decomposition algorithm outlined in §5. If $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, then there exists an optimal pure policy which can again be obtained from the decomposition algorithm; moreover, in this case each restricted MDP can be solved with parametric LP. For communicating and unichain MDPs the results are the same as for $\kappa(\mathbf{u})$, and the ϵ-optimal and optimal policies are the same for the two criteria.

We end this paper with a list of problems that remain open for time-average MDPs with variability sensitive criteria.

1. In the general multichain case with $h(x, y) = x - \lambda(x - y)^2$, $\lambda > 0$, can we find an efficient algorithm to locate the optimal pure policy for $\kappa(\mathbf{u})$?
2. What variability functions, besides $h(x, y) = x - \lambda(x - y)^2$, $\lambda > 0$, lead to optimal pure policies? In particular, do the variability functions $h(\cdot, \cdot)$ discussed in §3 give rise to optimal pure policies?
3. Can the results be extended to MDPs with infinite state or action spaces?

References

[1] Bather, J. (1973). Optimal Decision Procedures in Finite Markov Chains. Part II. Communicating Systems. *Adv. Appl. Probab.* **5** 521–552

[2] ______(1973). Optimal Decision Procedures in Finite Markov Chains. Part III. General Convex Systems. *Adv. Appl. Probab.* **5** 541–553.

[3] Bouakiz, M. A. and Sobel, M. J. (1985). Nonstationary Policies are Optimal for Risk-Sensitive Markov Decision Processes. Technical Report, Georgia Institute of Technology. Atlanta, GA.

[4] Chung, K. J. (1985). Some Topics in Risk-Sensitive Stochastic Models. PhD thesis, Georgia Institute of Technology, Atlanta, GA.

[5] Çinlar, E. (1975). *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs, NJ.

[6] Denardo, E. (1970). On Linear Programming in a Markov Decision Chain. *Management Sci.* **16** 281–288.

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

[8] Filar, J. A., Kallenberg. L. C. M. and Lee, H. M. (1989). Variance Penalized Markov Decision Processes. *Math. Oper. Res.* **14** 147–161.

[9] Heyman, D. and Sobel, M. J. (1984). *Stochastic Models in Operations Research. Vol. II.* McGraw-Hill Company, New York.

[10] Hordijk, A. and Kallenberg, L. C. M. (1984). Constrained Undiscounted Dynamic Programming. *Math. Oper. Res.* **9** 276–289.

[11] Kallenberg, L. C. M. (1983). *Linear Programming and Finite Markovian Control Problems.* vol. 148. Mathematical Centre Tracts, Amsterdam.

[12] Katoh, N. and Ibaraki, T. (1987). A Parametric Characterization of an ϵ-Approximation Scheme for the Minimization of a Quasiconcave Program. *Discrete Appl. Math.* **17** 39–66.

[13] Kawai, H. (1987). A Variance Minimization Problem for Markov Decision Processes. *European J. Oper. Res* **31** 140–145.

[14] Ross, K. W. and Varadarajan, R. (1991). Multichain Markov Decision Processes with a Sample-Path Constraint: A Decomposition Approach. *Math. Oper. Res.* **16** 195–207.

[15] ______ and ______ (1989). Markov Decision Processes with Sample Path Constraints: The Communicating Case. *Oper. Res.* **37** 780–790.

[16] Sobel, M. J. (1984). Mean Variance Tradeoffs in an Undiscounted MDP. Preprint.

BAYKAL-GÜRSOV: INDUSTRIAL ENGINEERING DEPARTMENT, RUTGERS UNIVERSITY, PISCATAWAY, NEW JERSEY 08854

ROSS: DEPARTMENT OF SYSTEMS, UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PENNSYLVANIA 19104