Probability in the Engineering and Informational Sciences, **21**, 2007, 635–657. Printed in the U.S.A.
DOI: 10.1017/S026996480700037X

SEMI-MARKOV DECISION PROCESSES

NONSTANDARD CRITERIA

M. BAYKAL-GÜRSOY
Department of Industrial and Systems Engineering
Rutgers University, Piscataway, NJ
E-mail: gursoy@rci.rutgers.edu

K. GÜRSOY
Department of Management Science
Kean University
Union, NJ

Considered are semi-Markov decision processes (SMDPs) with finite state and action spaces. We study two criteria: the expected average reward per unit time subject to a sample path constraint on the average cost per unit time and the expected time-average variability. Under a certain condition, for communicating SMDPs, we construct (randomized) stationary policies that are ϵ-optimal for each criterion; the policy is optimal for the first criterion under the unichain assumption and the policy is optimal and pure for a specific variability function in the second criterion. For general multichain SMDPs, by using a state space decomposition approach, similar results are obtained.

1. INTRODUCTION

We consider semi-Markov decision processes (SMDPs) with finite state and action spaces. Let R_s be the reward function at time s. R_s can be an impulse function corresponding to the reward earned immediately at a transition epoch or it can be a step function between transition epochs corresponding to the rate of reward. The great majority of the literature in this area is concerned with finding a policy $\boldsymbol{u}$ that maximizes

$$\phi_1(\boldsymbol{u}) \triangleq \liminf_{t\to\infty} \frac{1}{t} E_{\boldsymbol{u}} \left[\int_0^t R_s \, ds \right]. \tag{1}$$

 0269-9648/07 $25.00

ϕ_1 denotes the *average expected reward* [10, 11, 20, 22, 27, 35, 37]. The following alternative to ϕ_1 is given by Jewell [23], Ross [30, 31], and Mine and Osaki [28] as

$$\phi_2(\boldsymbol{u}) \triangleq \liminf_{n\to\infty} \frac{E_{\boldsymbol{u}}\left[\sum_{m=1}^{n} R_m\right]}{E_{\boldsymbol{u}}[T_n]}, \tag{2}$$

where R_m denotes the reward earned between the $(m-1)$st and the (m)th epochs and T_m denotes the (m)th transition time. The performance measure ϕ_2 is also used by other researchers (see e.g., [6, 15–17, 21, 22, 29]). In [18], ϕ_2 is referred to as the *ratio-average reward*. A sufficient condition for these two definitions to coincide under stationary policies requires that every stationary policy generates a semi-Markov chain with only one irreducible class [28, 30].

Although ϕ_1 is clearly the more appealing criterion, it is easier to write the optimality equations when establishing the existence of an optimal pure policy under criterion ϕ_2 [34, 35, 39]. On the other hand, for finite-state and finite-action SMDPs there exists an optimal pure policy under ϕ_1 [11, 34, 39], whereas such an optimal policy might not exist under ϕ_2 in a general multichain SMDP [24].

Even though there is considerable research on the nonstandard criteria for average reward Markov decision processes (MDPs), the same cannot be claimed for the average reward SMDPs. A variance-type objective function for the discrete time MDPs has been studied (see e.g., [3, 7, 19, 38]). Constraints have been introduced for the average reward MDPs (see e.g., [1, 4, 13, 14, 25, 32, 33, 37]). For the average reward SMDPs, only the constrained problem has been investigated [5, 16, 18]. Beutler and Ross [5, 6] considered the ratio-average reward with a constraint under a condition stronger than the unichain condition. In [18], Feinberg examined the problem of maximizing both ϕ_1 and ϕ_2 subject to a number of constraints. Under the condition that the initial distribution is fixed, he showed that for both criteria, there exist optimal mixed stationary policies when an associated linear program (LP) is feasible. The mixed stationary policies are defined as policies with an initial one-step randomization applied to a set of pure policies. Obviously, such a policy is not stationary. Feinberg provided a linear programming algorithm for the unichain SMDP under both criteria. However, there is a need for an efficient algorithm that would locate an optimal or ϵ-optimal stationary policy for the communicating and multichain SMDPs under ϕ_1 with constraints.

In this article we study the following criterion:

$$\psi(\boldsymbol{u}) \triangleq E_{\boldsymbol{u}}\left[\liminf_{t\to\infty} \frac{1}{t}\int_0^t R_s\, ds\right] \tag{3}$$

subject to the sample path constraint

$$P_{\boldsymbol{u}}\left\{\limsup_{t\to\infty} \frac{1}{t}\int_0^t C_s\, ds \le \alpha\right\} = 1, \tag{4}$$

where C_s denotes the cost function at time s. Fatou's lemma immediately implies that $\psi(\boldsymbol{u}) \leq \phi_1(\boldsymbol{u})$ holds for all policies. We however, prove that for a large class of policies, the two rewards are equal. We show that an ϵ-optimal randomized stationary policy can be obtained for the general SMDP, whereas such a policy might not exist for the expectation problem.

We also consider the problem of locating a policy that maximizes over all policies, $\boldsymbol{u}$, the following expected average reward:

$$\nu(\boldsymbol{u}) \triangleq E_{\boldsymbol{u}}\left[\liminf_{t\to\infty} \frac{1}{t}\int_0^t h\left(R_s, \frac{1}{t}\int_0^t R_q\, dq\right) ds\right], \tag{5}$$

where $h(\cdot,\cdot)$ is a function of the current reward at time s and the average reward over an interval that includes time s. Throughout, we assume that $h(\cdot,\cdot)$ is a continuous function. We will refer to $\nu(\boldsymbol{u})$ as the *expected time-average variability*. We show that an ϵ-optimal stationary policy can be obtained for the general SMDP. If $h(x, y) = x - \lambda(x - y)^2$, then the optimal policy is a pure policy. Note that, in this case, maximizing $\nu(\boldsymbol{u})$ corresponds to maximizing the expected average reward penalized by the expected average variability.

This article is organized as follows. In Section 2 we introduce the notation. In Section 3 we present our preliminary results, which will be used in the proceeding sections and summarize the known facts about the decomposition and sample-path theory. In Section 4, mathematical programs that will be utilized are constructed and the upper bounds for the expected average reward and the expected variability are established. Communicating SMDPs are investigated in Section 5, and it is shown that there exist ϵ-optimal stationary policies for both criteria. Multichain SMDPs are considered in Section 6, an intermediate problem is introduced, and the algorithm to locate the ϵ-optimal stationary policies is given. Finally, we conclude in Section 7 with a brief discussion on the sample-path problem with multiple constraints.

2. NOTATIONS

Denote $\{X_m, m \geq 0\}$ for the state process, which takes values in a finite state space $\mathcal{S}$. At each epoch m, the decision-maker chooses an action A_m from the finite action space $\mathcal{A}$. The sojourn time between the $(m-1)$st and the (m)th epochs is a random variable and denoted by Υ_m. The underlying sample space $\Omega = \{\mathcal{S} \times \mathcal{A} \times (0, \infty)\}^\infty$ consists of all possible realizations of states, actions, and the transition times. Throughout, the sample space will be equipped with the σ-algebra generated by the random variables $\{X_m, A_m, \Upsilon_{m+1}; m \geq 0\}$. Denote P_{xay}, $x \in \mathcal{S}$, $a \in \mathcal{A}$, $y \in \mathcal{S}$, for the law of motion of the process; that is, for all policies $\boldsymbol{u}$ and all epochs m,

$$P_{\boldsymbol{u}}\{X_{m+1} = y | X_0, A_0, \ldots, X_m = x, A_m = a\} = P_{xay}.$$

Also conditioned on the event that the next state is y, Υ_{m+1} has the distribution function $F_{xay}(\cdot)$; that is,

$$P_{\boldsymbol{u}}\{\Upsilon_{m+1} \leq t | X_0, A_0, \Upsilon_1, \ldots, X_m = x, A_m = a, X_{m+1} = y\} = F_{xay}(t).$$

Assume that $F_{xay}(0) < 1$.

The process $\{S_t, B_t : t \geq 0\}$, where S_t is the state of the process at time t and B_t is the action taken at time t, is referred to as the SMDP. Let $T_n = \sum_{m=1}^{n} \Upsilon_m$. For $t \in [T_m, T_{m+1})$, clearly

$$S_t = X_m, \quad B_t = A_m.$$

A policy is called *stationary* if the decision rule at each epoch depends only on the present state of the process; denote f_{xa} for the probability of choosing action a when in state x. A stationary policy is said to be *pure* if for each $x \in \mathcal{S}$, there is only one action $a \in \mathcal{A}$ such that $\boldsymbol{f}_{xa} = 1$. Let U, F, and G denote the set of all policies, stationary policies, and pure policies, respectively.

Under a stationary policy $\boldsymbol{f}$, the state process $\{S_t : t \geq 0\}$ is a semi-Markov process, and the process $\{X_m : m \in \mathcal{N}\}$ is the embedded Markov chain with transition probabilities

$$P_{xy}(\boldsymbol{f}) = \sum_{a \in \mathcal{A}} P_{xay} f_{xa}.$$

Clearly, the process $\{S_t, B_t : t \geq 0\}$ is also a semi-Markov process under a stationary policy $\boldsymbol{f}$ with the embedded Markov chain $\{X_m, A_m : m \in \mathcal{N}\}$.

Under a stationary policy $\boldsymbol{f}$, state x is *recurrent* if and only if x is recurrent in the embedded Markov chain; similarly, x is *transient* if and only if x is transient for the embedded Markov chain. A SMDP is said to be *unichain* if the embedded Markov chain for each pure policy is unichain [i.e., if the transition matrix $P(\boldsymbol{g})$ has at most one recurrent class plus (a perhaps empty) set of transient states for all pure policies g]. Similarly, a SMDP is said to be *communicating* if $P(\boldsymbol{f})$ is irreducible for all stationary policies that satisfy $\boldsymbol{f}_{xa} > 0$, for all $x \in \mathcal{S}$, a $\in \mathcal{A}$.

Let $\tau(x, a)$ define the expected sojourn time,

$$\begin{aligned}
\tau(x,a) &\triangleq E_{\boldsymbol{u}}[\Upsilon_m | X_{m-1} = x, A_{m-1} = a] \\
&= \int_0^{\infty} \sum_{y \in \mathcal{S}} P_{\boldsymbol{u}}\{X_m = y, \Upsilon_m > t | X_{m-1} = x, A_{m-1} = a\}\, dt \\
&= \int_0^{\infty} \left[1 - \sum_{y \in \mathcal{S}} P_{xay} F_{xay}(t)\right] dt.
\end{aligned}$$

Let $W_t(x, a)$ denote the random variables representing the state-action intensities,

$$W_t(x,a) \triangleq \frac{1}{t}\int_0^t \mathbf{1}\{(S_s, B_s) = (x,a)\}\, ds,$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Let U_0 denote the class of all policies $\boldsymbol{u}$ such that $\{W_t(x, a);\ t \geq 0\}$ converges $P_{\boldsymbol{u}}$-almost surely ($P_{\boldsymbol{u}}$-a.s.) for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$. Thus, for $\boldsymbol{u} \in U_0$, there exist random variables $\{W(x, a)\}$ such that

$$\lim_{t\to\infty} W_t(x,a) = W(x,a),$$

$P_{\boldsymbol{u}}$-a.s. for all x and a. Let U_1 be the class of all policies $\boldsymbol{u}$ such that the expected state-action intensities $\{E_{\boldsymbol{u}}[W_t(x, a)];\ t \geq 0\}$ converge for all x and a. For $\boldsymbol{u} \in U_1$, denote

$$w_{\boldsymbol{u}}(x,a) = \lim_{t\to\infty} E_{\boldsymbol{u}}[W_t(x,a)].$$

From Lebesgue's Dominated Convergence Theorem, $U_0 \in U_1$.

A well-known result from renewal theory (see Çinlar [9]) is that if $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous semi-Markov process and if the embedded Markov chain is unichain, then the proportion of time spent in state y; that is,

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t \mathbf{1}\{Y_s = y\}\, ds$$

exists almost surely. Since under a stationary policy $\boldsymbol{f}$ the process $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous semi-Markov process, if the embedded Markov decision process is unichain, then the limit of $W_t(x, a)$ as t goes to infinity exists and the proportion of time spent in state x when action a is applied is given as

$$W(x,a) = \lim_{t\to\infty} W_t(x,a) = \frac{\tau(x,a)\, Z(x,a)}{\sum_{x,a} \tau(x,a)\, Z(x,a)},$$

$P_{\boldsymbol{f}}$-a.s. for all x and a, where $Z(x, a)$ denotes the associated state-action frequencies. Let $\{z_{\boldsymbol{f}}(x, a);\ x \in \mathcal{S},\ a \in \mathcal{A}\}$ denote the expected state-action frequencies; that is,

$$z_{\boldsymbol{f}}(x,a) = \lim_{n\to\infty} E_f \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}\{X_{m-1} = x, A_{m-1} = a\} = \pi_x(\boldsymbol{f})\, f_{xa},$$

where $\pi_x(\boldsymbol{f})$ is the steady-state distribution of the embedded Markov chain $P(\boldsymbol{f})$.

The long-run average number of transitions into state x when action a is applied per unit time is

$$v_{\boldsymbol{f}}(x,a) = \frac{\pi_x(\boldsymbol{f})\, f_{xa}}{\sum_{x,a} \tau(x,a)\, \pi_x(\boldsymbol{f})\, f_{xa}} = \frac{z_{\boldsymbol{f}}(x,a)}{\sum_{x,a} \tau(x,a) z_{\boldsymbol{f}}(x,a)}. \tag{6}$$

This gives $w_{\boldsymbol{f}}(x, a) = \tau(x, a)\ v_{\boldsymbol{f}}(x, a)$.

The decision-maker earns an immediate reward $R(X_m, A_m)$ and a reward with rate $r(X_m, A_m)$ until the $(m+1)$st epoch. Thus,

$$R_{m+1} = R(X_m, A_m) + r(X_m, A_m)\Upsilon_{m+1}$$

is the reward earned during the $(m+1)$st transition [5, 36]. Similarly, there is an immediate cost $C(X_m, A_m)$ and a cost with rate $c(X_m, A_m)$ with

$$C_{m+1} = C(X_m, A_m) + c(X_m, A_m)\Upsilon_{m+1}.$$

Hence, at any epoch if the process is in state $x \in \mathcal{S}$ and action $a \in \mathcal{A}$ is chosen, then the reward earned during this epoch is represented by $\bar{r}(x, a) \triangleq R(x, a) + r(x, a)\tau(x, a)$. Similarly, the cost during this epoch is represented by $\bar{c}(x, a) \triangleq C(x, a) + c(x, a)\tau(x, a)$.

We conclude this section with a fact that will be used in the subsequent theorems. It follows directly from the law of large numbers for martingale differences (see, e.g., Loeve [26]):

For all policies $\boldsymbol{u} \in U$, *if* $\sum_{m=1}^{\infty} [(\operatorname{var} \Upsilon_m)/m^2] < \infty$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} [d(X_{m-1}, A_{m-1})\Upsilon_m - d(X_{m-1}, A_{m-1})\tau(X_{m-1}, A_{m-1})] = 0 \qquad \textbf{(7)}$$

holds $P_{\boldsymbol{u}}$*-a.s. with* $d(\cdot,\cdot)$ *as an arbitrary bounded function on* $\mathcal{S} \times \mathcal{A}$.

Thus, we need the following assumption on the sojourn times.

ASSUMPTION 1: *For all policies* $\boldsymbol{u} \in U$,

$$\sum_{m=1}^{\infty} \frac{\operatorname{var} \Upsilon_m}{m^2} < \infty.$$

This condition is essentially equivalent to the assumption that $E_u[{\Upsilon_m}^2 | X_{m-1} = x, A_{m-1} = a] < \infty$ for all $x \in \mathcal{S}$ and $a \in \mathcal{A}$.

3. PRELIMINARY RESULTS

In this section we establish some facts that will be used later in the analysis. Proposition 1 shows that the expected average reward (average cost) can be written in terms of the expected state-action frequencies $\{z_{\boldsymbol{u}}(x, a)\}$ ($\{Z(x, a)\}$).

PROPOSITION 1: *Assume that the SMDP is unichain. For any policy* $\boldsymbol{u} \in F$, *the expected average reward and the average cost are given respectively as*

$$\psi(\boldsymbol{u}) = \frac{\sum_{x,a} \bar{r}(x, a) z_{\boldsymbol{u}}(x, a)}{\sum_{x',a'} \tau(x', a') z_{\boldsymbol{u}}(x', a')} = \sum_{x,a} \bar{r}(x, a) v_{\boldsymbol{u}}(x, a) \qquad \textbf{(8)}$$

and

$$\limsup_{t\to\infty}\frac{1}{t}\int_0^t C_s\,ds=\frac{\sum_{x,a}\bar{c}(x,a)Z(x,a)}{\sum_{x',a'}\tau(x',a')Z(x',a')}, \tag{9}$$

$P_{\boldsymbol{u}}$-*a.s.*

PROOF: Fix a policy $\boldsymbol{u}\in F$. Equation (8) is written as

$$\begin{aligned}
\psi(\boldsymbol{u}) &= E_{\boldsymbol{u}}\left[\liminf_{t\to\infty}\frac{1}{t}\int_0^t R_s\,ds\right]\\
&= E_{\boldsymbol{u}}\left[\liminf_{t\to\infty}\frac{1}{t}\left[\sum_{m=0}^{n(t)}R(X_m,A_m)+\sum_{m=0}^{n(t)-1}r(X_m,A_m)\Upsilon_{m+1}+(t-t_{n(t)})r(X_{n(t)},A_{n(t)})\right]\right]\\
&= E_{\boldsymbol{u}}\left[\liminf_{t\to\infty}\frac{1}{t}\sum_{x,a}\left[R(x,a)\sum_{m=0}^{n(t)}\mathbf{1}\{X_m=x,A_m=a\}\right.\right.\\
&\qquad\left.\left.+\,r(x,a)\tau(x,a)\sum_{m=0}^{n(t)-1}\mathbf{1}\{X_m=x,A_m=a\}\right]\right]\\
&= \sum_{x,a}R(x,a)v_{\boldsymbol{u}}(x,a)+\sum_{x,a}r(x,a)\tau(x,a)v_{\boldsymbol{u}}(x,a),
\end{aligned}$$

where $n(t)\triangleq\max\{m:T_m\le t\}$ denotes the number of transitions up to time t. Note that as t goes to infinity, so does $n(t)$. Thus, the last term in the second equality goes to zero as t goes to infinity. Equation (9) is similarly obtained. ■

Now, consider the expected time-average variability $\nu(\boldsymbol{u})$. The following proposition shows that the time-average variability can also be expressed in terms of the long-run state-action frequencies $\{Z(x,a)\}$. Let Ψ_t denote the time average reward random variable (r.v.) up to time t:

$$\Psi_t\triangleq\frac{1}{t}\int_0^t R_s\,ds$$

and

$$\Psi\triangleq\sum_{x,a}\bar{r}(x,a)V(x,a).$$

PROPOSITION 2: *For all* $\boldsymbol{u} \in U_0$,

$$\liminf_{t\to\infty} \frac{1}{t}\int_0^t h\left(R_s, \frac{1}{t}\int_0^t R_q\, dq\right) ds$$
$$= \left(\sum_{x,a} h\left[\bar{r}(x,a), \frac{\sum_{y,b} \bar{r}(y,b)Z(y,b)}{\sum_{x',a'} \tau(x',a')Z(x',a')}\right] Z(x,a)\right)$$
$$\times \left[\sum_{x',a'} \tau(x',a')Z(x',a')\right]^{-1},$$

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

$$v(\boldsymbol{u}) = \psi(\boldsymbol{u}) - \lambda \lim_{t\to\infty} \frac{1}{t}\int_0^t E_{\boldsymbol{u}}\left[R_s - \frac{1}{t}\int_0^t R_q\, dq\right]^2 ds.$$

PROOF: Fix a policy $\boldsymbol{u} \in U_0$. Similar to Proposition 1, it is straightforward to establish that

$$\liminf_{t\to\infty} \frac{1}{t}\int_0^t h(R_s, \Psi)\, ds = \sum_{x,a} h[\bar{r}(x,a), \Psi]V(x,a),$$

$P_{\boldsymbol{u}}$-a.s. The rest of the proof follows from Proposition 1 in [3]. ■

The proof of the following proposition that defines ψ and v for the multichain case, is straightforward.

PROPOSITION 3: *Let* $\boldsymbol{f}$ *be a stationary policy and let* $\mathcal{R}_1, \ldots, \mathcal{R}_q$ *be the recurrent classes associated with* $P(\boldsymbol{f})$. *Denote* $(\pi_x^i(\boldsymbol{f}); x \in \mathcal{R}_i)$ *for the equilibrium probability vector associated with class* i, $i = 1, \ldots, q$. *Further, denote*

$$\psi_i(\boldsymbol{f}) = \frac{\sum_{x,a} \bar{r}(x,a)\pi_x^i(\boldsymbol{f})f_{xa}}{\sum_{y,b} \tau(y,b)\pi_y^i(\boldsymbol{f})f_{yb}}. \tag{10}$$

Then

$$\psi(\boldsymbol{f}) = \sum_{i=1}^{q} P_{\boldsymbol{f}}\{X_n \in \mathcal{R}_i \ \ a.s.\}\psi_i(\boldsymbol{f}) \tag{11}$$

and

$$\nu(\boldsymbol{f}) = \sum_{i=1}^{q} P_{\boldsymbol{f}}\{X_n \in \mathcal{R}_i \ \ a.s.\} \sum_{x,a} \frac{h[\bar{r}(x,a), \psi_i(\boldsymbol{f})]\, \pi_x^i(\boldsymbol{f}) f_{xa}}{\sum_{y,b} \tau(y,b) \pi_y^i(\boldsymbol{f}) f_{yb}}. \quad \textbf{(12)}$$

If SMDP is unichain, then

$$\nu(f) = \sum_{x,a} \frac{h[\bar{r}(x,a), \psi(\boldsymbol{f})]\, \pi_x(\boldsymbol{f}) f_{xa}}{\sum_{y,b} \tau(y,b) \pi_y(\boldsymbol{f}) f_{yb}}. \quad \textbf{(13)}$$

Note that Schäl [34] showed in Lemma 2.7 that for finite-state finite-action multichain SMDPs under a pure policy, ϕ_1 is equivalently given by Eqs. (10) and (11), which define the expected average reward ψ.

Decomposition and Sample Path Theory

The following notation will be used in the subsequent sections. A set $\mathcal{C} \subseteq \mathcal{S}$ is said to be a *strongly communicating* class if (1) $\mathcal{C}$ is a recurrent class for some stationary policy, (2) $\mathcal{C}$ is not a proper subset of some $\mathcal{C}'$ for which (1) holds true. Let $\{\mathcal{C}_1, \ldots, \mathcal{C}_I\}$ be the collection of all strongly communicating classes. Let $\mathcal{T}$ be the (possibly empty) set of states that are transient under all stationary policies. It is shown in [33] that $\{\mathcal{C}_1, \ldots, \mathcal{C}_I, \mathcal{T}\}$ forms a partition of the state space $\mathcal{S}$. The decomposition ideas was first introduced by Bather [2]. For each $i = 1, \ldots, I$, denote the for each $x \in \mathcal{C}_i$ the set

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

The following result is also proved in [33].

PROPOSITION 4: *For all policies* $\boldsymbol{u}$,

$$\sum_{i=1}^{I} P_{\boldsymbol{u}}\{X_n \in \mathcal{C}_i \ \ a.s.\} = 1 \quad \textbf{(14)}$$

and

$$P_{\boldsymbol{u}}\{A_n \in \mathcal{F}_{X_n} \ \ a.s.\} = 1. \quad \textbf{(15)}$$

For each $i = 1, \ldots, I$, define a new SMDP, called SMDP-i, as follows: The state space is $\mathcal{C}_i$; for each $x \in \mathcal{C}_i$, the set of available actions is given by the state-dependent action spaces $\mathcal{F}_x$; the law of motion P_{xay}, the conditional sojourn time distribution $F_{xay}(\cdot)$, the reward function $\bar{r}(x, a)$, and the cost function $\bar{c}(x, a)$ are the same as earlier but restricted to $\mathcal{C}_i$ and $\mathcal{F}_x$ for $x \in \mathcal{C}_i$. Now, each SMDP-i is communicating for all $i = 1, \ldots, I$. For each SMDP-i, let $\nu_i(\boldsymbol{u})$ be the expected average variability under policy $\boldsymbol{u}$.

4. OPTIMIZATION RESULTS

In the constrained problem, say $T^{(1)}$, we seek to maximize the expected average reward $\psi(\boldsymbol{u})$ [Eq. (3)] over the policies that satisfy the sample-path constraint [Eq. (4)]. Let U_f denote the class of feasible policies. The *optimal constrained average reward* is given as

$$\psi^* = \sup_{\boldsymbol{u} \in U_f} \psi(\boldsymbol{u}).$$

A policy $\boldsymbol{u}^* \in U_f$ is said to *constrained average optimal* if $\psi(\boldsymbol{u}^*) = \psi^*$. A policy $\boldsymbol{u} \in U_f$ is said to be ϵ-*average optimal* if $\psi(\boldsymbol{u}) > \psi^* - \epsilon$. The second problem, $T^{(2)}$, maximizes the expected time-average variability [Eq. (5)]. Let

$$\nu^* = \sup_{\boldsymbol{u} \in U} \nu(\boldsymbol{u}).$$

A policy $\boldsymbol{u}^*$ is optimal for $\nu(\cdot)$ if $\nu(\boldsymbol{u}^*) = \nu^*$. An ϵ-*optimal policy* for $\nu(\cdot)$ is defined as a policy $\boldsymbol{u}$ such that $\nu(\boldsymbol{u}) > \nu^* - \epsilon$.

Note that by choosing α to be sufficiently large, the unconstrained problem can be viewed as a special case of the constrained optimization problem. Also, by choosing $h^{(1)}(x, y) = x$, we have $\nu(\boldsymbol{u}) = \psi(\boldsymbol{u})$. Thus, in the next section we will present the general problem of maximizing $\nu^{(j)}(\boldsymbol{u})$ subject to the sample-path constraint (4), where $j = 1$ corresponds to the constrained problem with $\nu^{(1)}(\boldsymbol{u}) = \psi(\boldsymbol{u})$ and $j = 2$ corresponds to the expected average variability with $\nu^{(2)}(\boldsymbol{u}) = \nu(\boldsymbol{u})$ and $\alpha^{(2)} = \infty$.

For each $j = 1, 2$ and $i = 1, \ldots, I$, consider the following fractional program with decision variables $z(x, a)$, $x \in \mathcal{C}_i$, $a \in \mathcal{F}_x$. Let $\delta_{xy} = 1$ if $x = y$ and $\delta_{xy} = 0$ otherwise.

Program $T_i^{(j)}$

$$t_i^{(j)} = \max \left\{ \frac{\displaystyle\sum_{x \in \mathcal{C}_i} \sum_{a \in \mathcal{F}_x} h^{(j)} \left[\bar{r}(x, a), \frac{\displaystyle\sum_{y \in \mathcal{C}_i, b \in \mathcal{F}_y} \bar{r}(y, b)\, z(y, b)}{\displaystyle\sum_{x' \in \mathcal{C}_i, a' \in \mathcal{F}_{x'}} \tau(x', a')\, z(x', a')} \right] z(x, a)}{\displaystyle\sum_{x' \in \mathcal{C}_i, a' \in \mathcal{F}_{x'}} \tau(x', a')\, z(x', a')} \right\} \tag{16}$$

$$\text{s.t.} \sum_{x \in \mathcal{C}_i, a \in \mathcal{F}_x} (\delta_{xy} - P_{xay})\, z(x, a) = 0, \qquad y \in \mathcal{C}_i \tag{17}$$

$$\sum_{x \in \mathcal{C}_i, a \in \mathcal{F}_x} z(x, a) = 1, \tag{18}$$

$$\frac{\sum_{x\in\mathcal{C}_i,a\in\mathcal{F}_x}\bar{c}(x,a)z(x,a)}{\sum_{x'\in\mathcal{C}_i,a'\in\mathcal{F}_{x'}}\tau(x',a')z(x',a')}\le\alpha^{(j)}, \tag{19}$$

$$z(x,a)\ge 0 \quad x\in\mathcal{C}_i,\ a\in\mathcal{F}_x. \tag{20}$$

For each $\eta\ge 0$, we will also need to refer to the following fractional program with decision variables $z(x, a)$, for all $x\in\mathcal{S}$, $a\in\mathcal{A}$.

Program $Q_\eta^{(j)}$

$$q_\eta^{(j)}=\max\left\{\frac{\sum_{x\in\mathcal{S},a\in\mathcal{A}}h^{(j)}\left[\bar{r}(x,a),\dfrac{\sum_{y\in\mathcal{S},b\in\mathcal{A}}\bar{r}(y,b)\,z(y,b)}{\sum_{x',a'}\tau(x',a')\,z(x',a')}\right]z(x,a)}{\sum_{x',a'}\tau(x',a')\,z(x',a')}\right\} \tag{21}$$

$$\text{s.t.}\sum_{x\in\mathcal{S},a\in\mathcal{A}}(\delta_{xy}-P_{xay})z(x,a)=0,\qquad y\in\mathcal{S}, \tag{22}$$

$$\sum_{x\in\mathcal{S},a\in\mathcal{A}}z(x,a)=1, \tag{23}$$

$$\frac{\sum_{x\in\mathcal{S},a\in\mathcal{A}}\bar{c}(x,a)z(x,a)}{\sum_{x'\in\mathcal{S},a'\in\mathcal{A}}\tau(x',a')z(x',a')}\le\alpha^{(j)}, \tag{24}$$

$$z(x,a)\ge\eta,\quad x\in\mathcal{S},\ a\in\mathcal{A}. \tag{25}$$

We will refer to the feasible regions of Program $T_i^{(j)}$ and Program $Q_\eta^{(j)}$ simply as $T_i^{(j)}$ and $Q_\eta^{(j)}$, respectively. Note that the objective functions for both sets of mathematical programs are continuous functions over polytopes. As long as the cost constraint is satisfied for some $\{z(x, a)\}$, then $T_i^{(1)}$ for $i = 1, \ldots, I$ and $Q_0^{(1)}$ are nonempty. Note that $T_i^{(2)}$ for $i = 1, \ldots, I$ and $Q_0^{(2)}$ are always nonempty. For a given solution $\{z(x, a)\}$, we will write

$$z(x)=\sum_a z(x,a).$$

First, we consider the constrained problem, $T^{(1)}$ given by Eqs. (3) and (4). Thus, use $h^{(1)}(x, y) = x$ in Eqs. (16) and (21). The following lemmas provide bounds on $\phi(\boldsymbol{u})$ and $\nu(\boldsymbol{u})$. The proof of Lemma 2 is similar to the proof of Lemma 1, thus only an outline of the proof will be given.

LEMMA 1: *If U_f is nonempty, then for all $i = 1, \ldots, I$, $T_i^{(1)}$ is nonempty, and for $\boldsymbol{u} \in U_f$,*

$$P_{\boldsymbol{u}}\left\{\liminf_{t\to\infty} \frac{1}{t}\int_0^t R_s\, ds \le t_i^{(1)} | X_n \in \mathcal{C}_i \ \ a.s.\right\} = 1 \tag{26}$$

and, consequently

$$\psi(\boldsymbol{u}) \le \sum_{i=1}^{I} t_i^{(1)} P_{\boldsymbol{u}}\{X_n \in \mathcal{C}_i \ \ a.s.\}. \tag{27}$$

PROOF: Fix a policy $\boldsymbol{u} \in U_f$. Let Γ be the set of all sample paths $\omega = (x_0, a_0, \tau_1, x_1, a_1, \tau_2, \ldots)$ that satisfy the following:

(i) $a_n \in \mathcal{F}_{x_n}$, $\forall\, n \ge N$ for some positive integer N
(ii) $\sum_{x\in\mathcal{S}} \sum_{a\in\mathcal{A}} P_{xay}\, Z(x, a) = \sum_{a\in\mathcal{A}} Z(y, a)$, $\forall\, y \in \mathcal{S}$
(iii) $\limsup_{t\to\infty}(1/t)\int_0^t C_s\, ds \le \alpha^{(1)}$.

Combining Eq. (14) with Eq. (7) where $d(\cdot,\cdot) = 1$ and $\Upsilon_m = \mathbf{1}\{X_m = y\}$ and the fact that $\boldsymbol{u}$ is feasible yields

$$P_{\boldsymbol{u}}(\Gamma) = 1.$$

Let $(x_0, a_0, \tau_1, x_1, a_1, \tau_2, \ldots) \in \{X_n \in \mathcal{C}_i \ a.s.\} \cap \Gamma$ and define

$$Z_n(x, a) \triangleq \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}\{X_{m-1} = x, A_{m-1} = a\}.$$

Since $0 \le Z_n(x,\ a) \le 1$, by the standard compactness argument there exists a subsequence $\{N_k(\omega)\}$ along which $\{Z_n(x, a;\ \omega)\}$ converges to some $Z'(x, a;\ \omega)$ on $\Phi = \{X_n \in \mathcal{C}_i \text{ a.s.}\} \cap \Gamma$; that is,

$$\lim_{k\to\infty} Z_{N_k}(x, a) = Z'(x, a). \tag{28}$$

By definition, it follows that

$$Z'(x, a) = 0 \quad \text{whenever } x \notin \mathcal{C}_i \text{ or } a \notin \mathcal{F}_x$$

on the set Φ. Thus, on Φ,

$$\sum_{x\in\mathcal{C}_i} \sum_{a\in\mathcal{F}_x} P_{xay} Z'(x, a) = \sum_{a\in\mathcal{F}_y} Z'(y, a), \quad \forall\, y \in \mathcal{C}_i,$$

and

$$\sum_{x\in\mathcal{C}_i} \sum_{a\in\mathcal{F}_x} Z'(x, a) = 1, \qquad Z'(x, a) \ge 0, \quad \forall\, x \in \mathcal{C}_i, a \in \mathcal{F}_x.$$

Observe that for any bounded function $d(\cdot,\cdot)$ on Φ,

$$\left|\frac{1}{N_k}\sum_{m=1}^{N_k} d(X_{m-1},A_{m-1})\Upsilon_m - \sum_{x,a} d(x,a)\tau(x,a)Z'(x,a)\right|$$
$$\leq \left|\frac{1}{N_k}\sum_{m=1}^{N_k}[d(X_{m-1},A_{m-1})\Upsilon_m - d(X_{m-1},A_{m-1})\,\tau(X_{m-1},A_{m-1})]\right|$$
$$+\left|\frac{1}{N_k}\sum_{m=1}^{N_k} d(X_{m-1},A_{m-1})\,\tau(X_{m-1},A_{m-1}) - \sum_{x,a} d(x,a)\tau(x,a)Z'(x,a)\right|,$$

which combined with Eq. (7) and Eq. (28) gives

$$\lim_{k\to\infty}\frac{1}{N_k}\sum_{m=1}^{N_k} d(X_{m-1},A_{m-1})\Upsilon_m = \sum_{x,a} d(x,a)\tau(x,a)Z'(x,a).$$

From this equation the following holds:

$$\lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{m=1}^{N_k} C_m = \frac{\lim_{k\to\infty}\frac{1}{N_k}\sum_{m=1}^{N_k}[C(X_{m-1},A_{m-1}) + c(X_{m-1},A_{m-1})\Upsilon_m]}{\frac{1}{N_k}\sum_{m=1}^{N_k}\Upsilon_m}$$
$$= \frac{\sum_{x,a}\bar{c}(x,a)Z'(x,a)}{\sum_{x,a}\tau(x,a)Z'(x,a)}.$$

Also, on Φ,

$$\alpha^{(1)} \geq \limsup_{t\to\infty}\frac{1}{t}\int_0^t C_s\,ds$$
$$\geq \lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{m=1}^{N_k} C_m = \frac{\sum_{x,a}\bar{c}(x,a)Z'(x,a)}{\sum_{x,a}\tau(x,a)Z'(x,a)}$$

Thus, $Z'(x,a)$ is in the feasible set implying that $T_i^{(1)}$ is nonempty. Hence, on Φ,

$$\frac{\sum_{x,a}\bar{r}(x,a)Z'(x,a)}{\sum_{x,a}\tau(x,a)Z'(x,a)} \leq t_i^{(1)}.$$

In a similar manner,

$$\liminf_{t\to\infty}\frac{1}{t}\int_0^t R_s\,ds \le \lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{m=1}^{N_k} R_m = \frac{\sum_{x,a}\bar{r}(x,a)Z'(x,a)}{\sum_{x,a}\tau(x,a)Z'(x,a)},$$

which gives the desired result. Combining Eq. (26) with Proposition 4 gives Eq. (27).

Next, we consider the expected time-average variability criterion. ■

LEMMA 2: *For all $i = 1, \ldots, I$ and for all policies $\boldsymbol{u}$, we have*

$$P_{\boldsymbol{u}}\left\{\liminf_{t\to\infty}\frac{1}{t}\int_0^t h\left(R_s,\frac{1}{t}\int_0^t R_q\,dq\right)ds \le t_i^{(2)} | X_n \in \mathcal{C}_i \text{ a.s.}\right\} = 1 \qquad \textbf{(29)}$$

and, consequently,

$$v(\boldsymbol{u}) \le \sum_{i=1}^{I} t_i^{(2)} P_{\boldsymbol{u}}\{X_n \in \mathcal{C}_i \text{ a.s.}\}. \qquad \textbf{(30)}$$

PROOF: The proof is similar to the proof of Lemma 1. We only need to note that

$$\begin{aligned}
&\liminf_{t\to\infty}\frac{1}{t}\int_0^t h\left(R_s,\frac{1}{t}\int_0^t R_q\,dq\right)ds \\
&\quad\le \lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{m=1}^{N_k} h\left(R_m,\frac{1}{T_{N_k}}\sum_{l=1}^{N_k} R_l\right) \\
&\quad= \lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{m=1}^{N_k} h\left(R_m,\lim_{k\to\infty}\frac{1}{T_{N_k}}\sum_{l=1}^{N_k} R_l\right) \\
&\quad= \frac{\displaystyle\sum_{x\in\mathcal{C}_i}\sum_{a\in\mathcal{F}_x} h\left[\bar{r}(x,a),\frac{\displaystyle\sum_{y\in\mathcal{C}_i}\sum_{b\in\mathcal{F}_y}\bar{r}(y,b)Z'(y,b)}{\displaystyle\sum_{x'\in\mathcal{C}_i}\sum_{a'\in\mathcal{F}_{x'}}\tau(x',a')Z'(x',a')}\right]Z'(x,a)}{\displaystyle\sum_{x'\in\mathcal{C}_i}\sum_{a'\in\mathcal{F}_{x'}}\tau(x',a')Z'(x',a')},
\end{aligned}$$

on Φ. ■

5. THE COMMUNICATING CASE

We assume that the SMDP is communicating. This implies that there is only one strongly communicating class and that $\mathcal{S} = \mathcal{C}_1$. The analysis of this section draws on results and observations from [32].

In this section we will show that, in general, there does not exist an optimal stationary policy for both criteria. Instead, we show that an ϵ-optimal stationary policy can be constructed. First, we consider the constrained problem: Eqs. (3) and (4). Let $h^{(1)}(x, y) = x$ in Eqs. (16) and (21).

PROPOSITION 5: *Fix $\eta \geq 0$ and let $\{ z^\eta(x, a)\}$ be an optimal extreme point for $Q_\eta^{(1)}$. Define a policy $\boldsymbol{f}^\eta$ by the transformation*

$$\boldsymbol{f}^\eta = \begin{cases} \dfrac{z^\eta(x,a)}{z^\eta(x)} & \text{if } z^\eta(x) > 0 \\ \text{uniformly over the actions} & \text{otherwise.} \end{cases} \tag{31}$$

Then

$$\sum_x z^\eta(x) P_{xy}(\boldsymbol{f}^\eta) = z^\eta(y), \tag{32}$$

$$\sum_x z^\eta(x) = 1. \tag{33}$$

If $P(\boldsymbol{f}^\eta)$ is unichain, then $\boldsymbol{f}^\eta \in U_f$ and $P_{f^\eta}\{\liminf_{t\to\infty}(1/t)\int_0^t R_s\, ds = q_\eta^{(1)}\} = 1$. In particular, if $P(\boldsymbol{f}^0)$ is unichain, then $\boldsymbol{f}^0$ is an optimal stationary policy for the constrained problem.

PROOF: It is straightforward to show equations (32) and (33). If $P(\boldsymbol{f}^\eta)$ is unichain, there is a unique probability vector $\pi(f^\eta)$ associated with $P(\boldsymbol{f}^\eta)$. Hence, $\pi_x(\boldsymbol{f}^\eta) = z^\eta(x)$, giving P_{f^η}-almost surely

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t C_s\, ds = \frac{\sum_{x,a} \bar{c}(x,a)\pi_x(\boldsymbol{f}^\eta) f_{xa}^\eta}{\sum_{x,a} \tau(x,a)\pi_x(\boldsymbol{f}^\eta) f_{xa}^\eta}$$

$$= \frac{\sum_{x,a} \bar{c}(x,a) z^\eta(x,a)}{\sum_{x,a} \tau(x,a) z^\eta(x,a)} \leq \alpha^{(1)}.$$

In a similar manner, we have P_{f^η}-a.s.

$$\limsup_{t\to\infty}\frac{1}{t}\int_0^t R_s\ ds=\frac{\sum_{x,a}\bar{r}(x,a)z^\eta(x,a)}{\sum_{x,a}\tau(x,a)z^\eta(x,a)}=q_\eta^{(1)} \qquad \blacksquare$$

Only the outline of the proof of the following theorem will be given since it follows the proofs of Propositions 5–7 in [32].

THEOREM 1: *Suppose that the SMDP is communicating: Then U_f is nonempty if and only if $Q_0^{(1)}$ is nonempty. If $Q_0^{(1)}$ is nonempty, then for each $\epsilon>0$, there exists an ϵ-optimal stationary policy for the constrained problem.*

PROOF: Proposition 5 proves the (only if) part. To prove the (if) part assume that $\{z^0(x,a)\}$ is an optimal extreme point of $Q_0^{(1)}$. Let f^0 be the policy obtained via transformation (31). It follows from Eq. (32) that the set of states where $z^0(x)>0$ is a closed set, and by Lemma 2 of [32], all states outside of this closed set are transient. This closed set can be composed of the union of m recurrent classes $R_1,\ldots,R_m$ associated with $P(f^0)$. For each recurrent class, we can define

$$d_k=\frac{\sum_{x\in R_k}\sum_a \bar{c}(x,a)z^0(x,a)}{\sum_{x\in R_k}\sum_a \tau(x,a)z^0(x,a)}.$$

The value d_k has the interpretation of being the average cost per unit time, given that the process has entered R_k. Let $l=\arg\min_{1\le k\le m} d_k$. Then since $\{z^0(x,a)\}$ is feasible for $Q_0^{(1)}$, we have $d_l \le \alpha^{(1)}$. Since d_k can be greater than $\alpha^{(1)}$ for some k, f^0 does not necessarily belong to U_f. However, we can define a stationary policy $\tilde{f}$ that is equal to f^0 in R_l, and outside R_l it takes every available action with equal probability. Clearly, since the SMDP is communicating, R_l is the only recurrent class associated with $P(\tilde{f})$ and $\tilde{f}$ is in U_f. Thus, U_f is nonempty.

For the second part of the theorem, we assume that $Q_0^{(1)}$ is nonempty. Using the machinery developed in [32], whenever there exists a policy that strictly meets the constraint, one can construct a feasible stationary policy that chooses every action with positive probability and gives rise to an irreducible Markov chain. Otherwise, the stationary policy f^0 given by the transformation (31) gives rise to a unichain $P(f^0)$; thus, f^0 is the optimal policy.

Thus, we assume that there exists a policy that strictly meets the constraint. In this case, there exists an $\zeta>0$ such that for each η that satisfies $0<\eta<\zeta$, there is a feasible solution for $Q_\eta^{(1)}$, and $P(f^\eta)$ is irreducible for f^η obtained via transformation (31). From Proposition 5, we have $f^\eta \in U_f$ and $P_{f^\eta}\{\liminf_{t\to\infty}(1/t)\int_0^t R_s\,ds=q_\eta^{(1)}=1\}$. To prove that $\lim_{\eta\to 0} q_\eta^{(1)}=q_0^{(1)}$, we can transform the fractional program into a linear

program using transformation (6) for v_{f^η} [8, 12]. Then the desired continuity holds by the piecewise linearity and convexity of the objective function with respect to the right-hand-side value of $\eta / \sum_{x,\,a} \tau(x, a)$. ■

Next, we present the mathematical programs obtained via transformation (6) explicitly, in terms of the decision variables $v(x, a)$,

Program $LT_i^{(j)}$

$$t_i^{(j)} = \max \left\{ \sum_{x \in \mathcal{C}_i} \sum_{a \in \mathcal{F}_x} h^{(j)} \left[\bar{r}(x, a), \sum_{y \in \mathcal{C}_i, b \in \mathcal{F}_y} \bar{r}(y, b) v(y, b) \right] v(x, a) \right\}$$

$$\text{s.t.} \sum_{x \in \mathcal{C}_i, a \in \mathcal{F}_x} (\delta_{xy} - P_{xay}) v(x, a) = 0, \qquad y \in \mathcal{C}_i$$

$$\sum_{x \in \mathcal{C}_i, a \in \mathcal{F}_x} \tau(x, a) v(x, a) = 1,$$

$$\sum_{x \in \mathcal{C}_i, a \in \mathcal{F}_x} \bar{c}(x, a) v(x, a) \leq \alpha^{(j)},$$

$$v(x, a) \geq 0, x \in \mathcal{C}_i, a \in \mathcal{F}_x.$$

For each $\eta \geq 0$, we also define the following program with decision variables $v(x, a)$, $x \in \mathcal{S}$, $a \in \mathcal{A}$.

Program $LQ_\eta^{(j)}$

$$q_\eta^{(j)} = \max \left\{ \sum_{x \in \mathcal{S}, a \in \mathcal{A}} h^{(j)} \left[\bar{r}(x, a), \sum_{y \in \mathcal{S}, b \in \mathcal{A}} \bar{r}(y, b) v(y, b) \right] v(x, a) \right\}$$

$$\text{s.t.} \sum_{x \in \mathcal{S}, a \in \mathcal{A}} (\delta_{xy} - P_{xay}) v(x, a) = 0, \quad y \in \mathcal{S},$$

$$\sum_{x \in \mathcal{S}, a \in \mathcal{A}} \tau(x, a) v(x, a) = 1,$$

$$\sum_{x \in \mathcal{S}, a \in \mathcal{A}} \bar{c}(x, a) v(x, a) \leq \alpha^{(j)},$$

$$v(x, a) \geq \eta, \quad x \in \mathcal{S}, \quad a \in \mathcal{A}.$$

Now, we can present the following procedure to locate the optimal or near optimal policies for the constrained problem.

Step 1: Solve the LP $LQ_0^{(1)}$ by the simplex method. If $LQ_0^{(1)}$ is not feasible, then there does not exist a policy that meets the sample path constraint, stop; otherwise go to Step 2.
Step 2: Let $\{ v^0(x,a) \}$ be an optimal extreme point for the LP $LQ_0^{(1)}$ and let $\boldsymbol{f}^0$ be the corresponding stationary policy obtained via transformation (31). If $P(\boldsymbol{f}^0)$ is unichain, then $\boldsymbol{f}^0$ is an optimal stationary policy, stop; otherwise go to Step 3.
Step 3: Solve the parametric LP $LQ_\eta^{(1)}$, $\eta \geq 0$ over some interval $[0, \delta]$ beginning with $\eta = 0$. Then employ the transformation (31) to obtain an ϵ-optimal stationary policy for ϵ as small as desired.

For the second criterion, we consider that the right-hand-side value of the cost constraint is equal to infinity; that is, $\alpha^{(2)} = \infty$ and the objective function is equal to $\nu(\boldsymbol{u})$. We have the following lemma, which easily follows from the invariance of the steady-state distribution.

LEMMA 3: *Let* $\mathbf{z}$ *be a feasible solution for Program* $LQ_0^{(2)}$ *and let* $\boldsymbol{f}$ *be defined as in Eq. (31). If* $P(\boldsymbol{f})$ *is unichain, then*

$$\nu(\boldsymbol{f}) = \frac{\displaystyle\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}} h^{(2)}\left[\bar{r}(x,a), \frac{\displaystyle\sum_{x,a}\bar{r}(x,a)z(x,a)}{\displaystyle\sum_{x,a}\tau(x,a)z(x,a)}\right] z(x,a)}{\displaystyle\sum_{x\in\mathcal{S}}\sum_{a\in\mathcal{A}}\tau(x,a)z(x,a)}.$$

For the communicating SMDP $Q_\eta^{(2)}$, consequently the feasible region of program $LQ_\eta^{(2)}$ is nonempty for all $\eta \in [0, \delta]$ for some $\delta > 0$. Now for each η, let v^η be an optimal solution to Program $LQ_\eta^{(2)}$. If there is an optimal extreme point solution to Program $LQ_0^{(2)}$, further require that v^0 to be an extreme point. For each $\eta \in [0,\delta]$, let $\boldsymbol{f}^\eta$ be defined from v^η according to transformation given in Eq. (31).

THEOREM 2: *Fix* $\epsilon > 0$. *If the SMDP is communicating, then for* $\eta > 0$ *sufficiently small, the stationary policy* $\boldsymbol{f}^\eta$ *is* ϵ*-optimal for* $\nu(\boldsymbol{u})$. *If, in addition,* $h^{(2)}(x,y) = x - \lambda(x-y)^2$ *with* $\lambda > 0$, *then the policy* $\boldsymbol{f}^0$ *is the optimal pure policy for the expected average variability criterion.*

PROOF: Noting that the objective function of program $LQ_\eta^{(2)}$ is continuous over the feasible region of Program $LQ_0^{(2)}$, the proof follows from the proof of Theorem 1 in [3]. ■

6. MULTICHAIN SMDPs

In this section we impose no restrictions on the law of motion P_{xay}, $x \in \mathcal{S}$, $a \in \mathcal{A}$, and $y \in \mathcal{S}$. We now construct ϵ-optimal stationary policies for the constrained problem and for the expected average variability problem. Since the arguments are similar for both criteria, we will present the combined results. The construction of the optimal policy follows closely the developments for the MDP problem in [33]; thus, we will only give the outlines of the proofs.

Recall that SMDP-i is communicating. By Theorems 1 and 2 we can construct an ϵ-optimal stationary policy $\boldsymbol{f}_i^{(j)}$ for each SMDP-i, $i = 1, \ldots, I$, and for either criterion, $j = 1, 2$. Recall that $t_i^{(j)}$ is the value of Program $LT_i^{(j)}$. We will make the following modification to $t_i^{(1)}$ in the constrained problem. Although $t_i^{(1)}$ is assigned to each communicating class i whenever program $LT_i^{(1)}$ has a feasible solution, if there does not exist any feasible policy for program $LT_i^{(1)}$, then $t_i^{(1)} = -\infty$ is assigned to discourage the process from going into class $\mathcal{C}_i$.

Consider the problem of finding a policy that maximizes the following time-average expected reward for each criterion:

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

This problem is referred as the intermediate SMDP. At this stage, the decision-maker decides which communicating class generates the maximum reward while satisfying the constraint. It is known that there exists an optimal pure policy $\boldsymbol{g}^{(j)}$ for each criterion that can be found by policy improvement, value iteration, or linear programming. Let

$$H^{(j)} = \{i : \ \mathcal{C}_i \text{ contains a recurrent class under } P(\boldsymbol{g}^{(j)})\}.$$

Modify $\boldsymbol{g}^{(j)}$ so that $\mathcal{C}_i$ is closed for each $i \in H^{(j)}$ and so that $\boldsymbol{g}^{(j)}$ remains optimal for the intermediate problem (see [33]).

We now construct stationary policy $\boldsymbol{f}^{(1)*}$ ($\boldsymbol{f}^{(2)*}$) as follows: When in state $x \in \mathcal{C}_i$, $i \in H^{(1)}$ ($H^{(2)}$), apply $\boldsymbol{f}_i^1$ ($\boldsymbol{f}_i^2$); otherwise, apply $\boldsymbol{g}^{(1)}$ ($\boldsymbol{g}^{(2)}$). The main result is as follows:

THEOREM 3: *The stationary policy* $\boldsymbol{f}^{(1)*}$ ($\boldsymbol{f}^{(2)*}$) *is* ϵ*-optimal for* $\psi(\boldsymbol{u})$ ($v(\boldsymbol{u})$).

PROOF: Employing Eq. (14) it can be shown that

$$\beta^{(j)}(\boldsymbol{u}) = \sum_{i=1}^{I} t_i^{(j)} P_{\boldsymbol{u}}\{X_n \in \mathcal{C}_i \text{ a.s.}\}$$

for all policies $\boldsymbol{u} \in U_f$ and $j = 1, 2$. Thus, from Lemma 1, we have

$$\psi(\boldsymbol{u}) \le \beta^{(1)}(\boldsymbol{g}^{(1)})$$

for all policies $\boldsymbol{u} \in U_f$. From Lemma 2, we have

$$v(\boldsymbol{u}) \leq \beta^{(2)}(\boldsymbol{g}^{(2)})$$

for all policies $\boldsymbol{u}$. From Proposition 3 and the construction of $\boldsymbol{f}^{(1)^*}$ and $\boldsymbol{f}^{(2)^*}$, we have

$$\psi(\boldsymbol{f}^{(1)^*}) = \sum_{i=1}^{I} \psi_i(\boldsymbol{f}_i^{(1)}) P_{\boldsymbol{g}^{(1)}}\{X_n \in \mathcal{C}_i \text{ a.s.}\}$$

and

$$v(\boldsymbol{f}^{(2)^*}) = \sum_{i=1}^{I} v_i(\boldsymbol{f}_i^{(2)}) P_{\boldsymbol{g}^{(2)}}\{X_n \in \mathcal{C}_i \text{ a.s.}\},$$

Combining the above equations with Theorems 1 and 2 gives the desired results. ■

In order to construct the ϵ-optimal (respectively optimal) stationary (respectively pure) policy $\boldsymbol{f}^*$ for the constained problem and for the expected variability criteria (expected time-average variability criteria when $h^{(2)}(x, y) = x - \lambda(x - y)^2$, $\lambda > 0$), we can use the following procedure.

Step 1: Determine the strongly communicating classes $\mathcal{C}_i$, $i = 1, \ldots, I$.
Step 2: For the constrained problem (respectively the expected time-average variability criterion), solve Program $LT_i^{(1)}$ and obtain policies $\boldsymbol{f}_i^{(1)}$ and optimal values $t_i^{(1)}$ (respectively $LT_i^{(2)}$, $\boldsymbol{f}_i^{(2)}$, and $t_i^{(2)}$) for $i = 1, \ldots, I$.
Step 3: For the constrained problem (respectively the expected time-average variability criterion) solve the intermediate SMDP and obtain $\boldsymbol{g}^{(1)}$ and $H^{(1)}$ (respectively $\boldsymbol{g}^{(2)}$and $H^{(2)}$). Then combine it with $\boldsymbol{f}_i^{(1)}$ ($\boldsymbol{f}_i^{(2)}$) for $i \in H^{(1)}$ ($H^{(2)}$), to get the ϵ-optimal (or optimal) policy $\boldsymbol{f}^{(1)^*}$ ($\boldsymbol{f}^{(2)^*}$).

7. CONCLUSIONS

In this article, we first considered the expected time-average reward $\psi(\boldsymbol{u})$ subject to a sample path constraint on the time-average cost. In general, there exists an ϵ-optimal stationary policy that can be obtained from the decomposition algorithm outlined in Section 6. If the SMDP is unichain, then the policy is optimal for the constrained problem. The optimal (ϵ-optimal) policy can be found for unichain (respectively communicating) SMDPs from the algorithm presented in Section 5.

Then we considered the expected time-average variability $v(\boldsymbol{u})$. In general, there exists an ϵ-optimal stationary policy that can be obtained from the decomposition algorithm outlined in Section 6. If $h(x, y) = x - \lambda(x - y)^2$ with $\lambda > 0$, then there exists an optimal pure policy that can again be obtained from the decomposition algorithm; moreover, in this case, each restricted SMDP can be solved with parameteric LP. For general $h(\cdot,\cdot)$ an optimal (ϵ-optimal) policy can be found for unichain (respectively communicating) SMDPs by solving the mathematical program $LQ_0^{(2)}$ (respectively mathematical programs $LQ_\eta^{(2)}$, $\eta \geq 0$).

Multiple Constraints

Multiple sample-path constraints could be handled by the theory presented above and in [32]; they were omitted in order to simplify the notation. Multiple sample-path constraints could be introduced as

$$P_{\boldsymbol{u}}\left\{\limsup_{n\to\infty}\frac{1}{t}\int_0^t C_s\,ds \le \alpha_k^{(1)}\right\} = 1$$

for all $k = 1, \ldots, K$. To incorporate these constraints, the programs $T_i^{(j)}$, $Q_\eta^{(j)}$, $LT_i^{(j)}$, and $LQ_\eta^{(j)}$ should be modified accordingly. One can see that all of the results in Sections 3, 4, and 6 continue to hold. However, note that except in the unichain case, for general SMDPs, the existence of a stationary policy is not implied by the nonemptiness of $Q_0^{(1)}$ when there is more than one constraint. Thus, Theorem 1 should be altered similar to [32], as below.

THEOREM 4: *Suppose that SMDP is communicating. If there exists a policy* $\boldsymbol{u}$ *and a* $\nu > \delta$ *such that*

$$P_{\boldsymbol{u}}\left\{\limsup_{t\to\infty}\frac{1}{t}\int_0^t C_s\,ds \le \alpha_k^{(1)} - \delta\right\} = 1$$

for all $k = 1, \ldots, K$, *then for any* $\epsilon > 0$, *there exists an* ϵ*-optimal stationary policy for the sample-path criterion.*

Since the modified program $LQ_0^{(1)}$ is an LP with $|\mathcal{S}| + K$ linearly independent constraints, one could see that the number of additional actions that an ϵ-optimal policy uses in communicating SMDP problems is equal to the number of constraints.

Acknowledgments

The first author's research was supported by the NSF under grant No. NCR-9110105. The first author would like to thank K.W. Ross for introducing this problem and for his valuable comments. The authors acknowledge with gratitude the insightful comments and suggestions by an anonymous referee that improved the presentation substantially.

References

1. Altman, E. (1993). Asymptotic properties of constrained Markov decision processes. *Mathematical Methods of Operations Research* 37: 151–170.
2. Bather, J. (1973). Optimal decision procedures in finite Markov chains. Part II: Communicating systems. *Advances in Applied Probability* 5: 521–552.
3. Baykal-Gürsoy, M. & Ross, K.W. (1992). Variability sensitive Markov decision processes. *Mathematics of Operations Research* 17: 558–571.
4. Beutler, F.J. & Ross, K.W. (1985). Optimal policies for controlled Markov chains with a constraint. *Journal of Mathematical Analysis and Applications* 112: 236–252.

5. Beutler, F.J. & Ross, K.W. (1986). Time-average optimal constrained semi-Markov decision processes. *Advances in Applied Probability* 18: 341–359.
6. Beutler, F.J. & Ross, K.W. (1987). Uniformization for semi-Markov decision processes under stationary policies. *Advances in Applied Probability* 24: 644–656.
7. Bouakiz, M.A. & Sobel, M.J. (1985). Nonstationary policies are optimal for risk-sensitive Markov decision processes. Technical Report, Georgia Institute of Technology.
8. Charnes, A. & Cooper, W.W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly* 9: 181–186.
9. Çinlar, E. (1975). *Introduction to stochastic processes*. Englewood Cliffs, NJ: Prentice-Hall.
10. Denardo, E.V. (1971). Markov renewal programs with small interest rate. *Annals of Mathematical Statistics* 42: 477–496.
11. Denardo, E.V. & Fox, B.L. (1968). Multichain Markov renewal programs. *SIAM Journal of Applied Mathematics* 16: 468–487.
12. Derman, C. (1962). On sequential decisions and Markov chains. *Management Science* 9: 16–24.
13. Derman, C. (1970). *Finite state Markovian decision processes*. New York: Academic Press.
14. Derman, C. & Veinott, A.F., Jr. (1972). Constrained Markov decision chains. *Management Science* 19: 389–390.
15. Federgruen, A., Hordijk, A., & Tijms, H.C. (1979). Denumerable state semi-Markov decison processes with unbounded costs, average cost criterion. *Stochastic Processes and Applications* 9: 223–235.
16. Federgruen, A. & Tijms, H.C. (1978). The optimality equation in average cost denumerable state semi-Markov decison problems, recurrency conditions and algorithms. *Journal of Applied Probability* 15: 356–373.
17. Federgruen, A., Schweitzer, P.J., & Tijms, H.C. (1983). Denumerable undiscounted semi-Markov decision processes with unbounded rewards. *Mathematics of Operations Research* 8(2): 298–313.
18. Feinberg, E.A. (1994). Constrained semi-Markov decision processes with average rewards. *Mathematical Methods of Operations Research* 39: 257–288.
19. Filar, J.A., Kallenberg, L.C.M., & Lee, H.M. (1989). Variance penalized Markov decision processes. *Mathematics of Operations Research* 14: 147–161.
20. Fox, B. (1966). Markov renewal programming by linear fractional programming. *SIAM Journal of Applied Mathematics* 16: 1418–1432.
21. Heyman, D.P. & Sobel, M.J. (1983). *Stochastic models in operations research.* Vol. II*: Stochastic optimization*. New York: McGraw-Hill.
22. Jewell, W.S. (1963). Markov renewal programming I: Formulation, finite return models. *Journal of Operations Research* 11: 938–948.
23. Jewell, W.S. (1963). Markov renewal programming II: Inifinite return models, example. *Operations Research* 11: 949–971.
24. Jianyong, L. & Xiaobo, Z. (2004). On average reward semi-Markov decision processes with a general multichain structure. *Mathematics of Operations Research* 29(2): 339–352.
25. Kallenberg, L.C.M. (1983). *Linear programming and finite Markovian control problems.* Mathematical Centre Tracts Vol. 146. Amsterdam: Elsevier.
26. Loeve, M. (1978). *Probability theory*, Vol. 2. New York: Springer-Verlag.
27. Lippman, S.A. (1971). Maximal average reward policies for semi-Markov renewal processes with arbitrary state and action spaces. *Annals of Mathematical Statistics* 42: 1717–1726.
28. Mine, H. & Osaki, S. (1970). *Markovian decision processes*. New York: Elsevier.
29. Puterman, M.L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: Wiley.
30. Ross, S.M. (1970). Average cost semi-Markov processes. *Journal of Applied Probability* 7: 649–656.
31. Ross, S. (1971). *Applied probability models with optimization applications*. San Francisco: Holden-Day.
32. Ross, K.W. & Varadarajan, R. (1989). Markov decision processes with sample path constraints: The communicating case. *Operations Research* 37: 780–790.

33. Ross, K.W. & Varadarajan, R. (1991). Multichain Markov decision processes with a sample path constraint: A decomposition approach. *Mathematics of Operations Research* 16: 195–207.
34. Schäl, M. (1992). On the second optimality equation for semi-Markov decision models. *Mathematics of Operations Research* 17(2): 470–486.
35. Schweitzer, P.J. & Federgruen, A.F. (1978). The functional equations of undiscounted Markov renewal programming. *Mathematics of Operations Research* 3: 308–321.
36. Sennott, L.I. (1989). Average cost semi-Markov decision processes and the control of queueing systems. *Probability in the Engineering and Informational Sciences* 3: 247–272.
37. Sennott, L.I. (1993). Constrained average cost Markov decision chains. *Probability in the Engineering and Informational Sciences* 7: 69–83.
38. Sobel, M.J. (1994). Mean variance tradeoffs in an undiscounted MDP. *Operations Research* 42(1): 175–183.
39. Yushkevich, A.A. (1981). On semi-Markov controlled models with an average reward criterion. *Theory of Probability and Its Applications* 26: 796–802.