SEMI-MARKOV DECISION PROCESSES

MELIKE BAYKAL-GÜRSOY
Department of Industrial and Systems Engineering, RUTCOR, CAIT, Rutgers University, New Brunswick, New Jersey

INTRODUCTION

Semi-Markov decision processes (SMDPs) are used in modeling stochastic control problems arising in Markovian dynamic systems, where the sojourn time in each state is a general continuous random variable. They are powerful, natural tools for the optimization of queues [1–7], production scheduling [8–11], and reliability/maintenance [12].

For example, in a machine replacement problem with deteriorating performance over time, a decision maker, after observing the current state of the machine, decides whether to continue its usage, or initiate a maintenance (preventive or corrective) repair, or replace the machine. There are a reward or cost structure associated with the states and decisions, and an information pattern available to the decision maker. This decision depends on a performance measure over the planning horizon which is either finite or infinite, such as total expected discounted or long run average expected reward/cost with or without external constraints, and variance penalized average reward.

SMDPs are based on semi-Markov processes (SMPs) [13] (see also ***Semi-Markov Processes*** SMPs), that include renewal processes (see also ***Definition and Examples of Renewal Processes***) and continuous-time Markov chains (CTMCs) (see also ***Definition and Examples of Continuous-Time Markov Chains***) as special cases. In an SMP similar to Markov chains (DTMCs) (see also ***Definition and Examples of DTMCs***), state changes occur according to the Markov property, that is, states in the future do not depend on the states in the past given the present. However, the sojourn time in a state is a continuous random variable with distribution depending on that state and the next state, a Markov chain is a SMP in which the sojourn times are discrete (geometric) random variables independent of the next state; a CTMC is a SMP with exponentially distributed sojourn times; and a renewal process is a SMP with a single state. SMDPs , first introduced by Jewell [14] and De Cani [15], are also called as *Markov renewal programs* [16–19].

This article is organized as follows: the next section introduces basic definitions and notations. Various performance criteria are presented in the section titled "Performance Measures" and their solution methodologies are described in the sections titled "Discounted Reward Criterion," "Average Reward Criterion," and "Expected Time-Average Reward and Variability."

BASIC DEFINITIONS

We consider time-homogeneous, finite state, and finite action SMDPs, and give references for the more general cases. Let $\{X_m, m \geq 0\}$ denote the state process, which takes values in a finite state space $\mathcal{S}$. We also use $\{X_m, m \in \mathcal{N}\}$ to denote the state process with $\mathcal{N}$ representing the set of nonnegative integers. At each epoch m, the decision maker chooses an action A_m from a 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 $\{X_m, A_m, \Upsilon_{m+1}; m \geq 0\}$. The initial state is assumed to be fixed and given. Note that we will suppress the dependence on the initial state unless given otherwise. Denote

Wiley Encyclopedia of Operations Research and Management Science, edited by James J. Cochran

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

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

Also conditioned on the event that the next state is y, Υ_{m+1} has the distribution function $F_{xay}(.)$, 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 $T_n = \sum_{m=1}^{n} \Upsilon_m$, that is, denote the time of nth transition. For $t \in [T_m, T_{m+1})$, clearly

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

Policy Types

A *decision rule* $\boldsymbol{u}^m$ at epoch m is a vector consisting of probabilities assigned to each available action. A decision rule may depend on all of the previous states, actions, transition times, and the present state. Let u_a^m denote the ath component of $\boldsymbol{u}^m$. Thus, it is the conditional probability of choosing action a at the m-th epoch, that is,

$$P_{\boldsymbol{u}}\{A_m = a|X_0 = x_0, A_0 = a_0, \Upsilon_1 = \tau_1, \ldots, X_m = x\} = u_a^m(x_0, a_0, \tau_1, \ldots, x).$$

A *policy* is an infinite sequence of decision rules $\boldsymbol{u} = \{\boldsymbol{u}^0, \boldsymbol{u}^1, \boldsymbol{u}^2, \ldots\}$.

Policy $\boldsymbol{u}$ is called *Markov policy* if $\boldsymbol{u}^m$ at epoch m depends only on the current state not the past history, that is,

$$\boldsymbol{u}^m(x) = \boldsymbol{u}_a^m(x_0, a_0, \tau_1, \ldots, x).$$

A policy is called *stationary* if the decision rule at each epoch is the same and it depends only on the present state of the process, $\boldsymbol{u} = \{\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}, \ldots\}$; 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 $f_{xa} = 1$. Let U, M, F, and G denote the set of all policies, Markov policies, stationary policies, and pure policies, respectively. Clearly, $G \subset F \subset M \subset U$.

Under a stationary policy $\boldsymbol{f}$, the state process $\{S_t : t \geq 0\}$ is a SMP, while 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 SMP under a stationary policy $\boldsymbol{f}$ with the embedded Markov chain $\{X_m, A_m : m \in \mathcal{N}\}$.

Chain Structure

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* (*multichain*) if the embedded Markov chain for each pure policy is unichain (multichain), that is, if the transition matrix $P(\boldsymbol{g})$ has at most one (more than one) recurrent class plus (a perhaps empty) set of transient states for all pure policies $\boldsymbol{g}$. It is called *irreducible* if $P(\boldsymbol{g})$ is irreducible under all pure policies $\boldsymbol{g}$. Similarly, a SMDP is said to be *communicating* if $P(\boldsymbol{f})$ is irreducible for all stationary policies that satisfy $f_{xa} > 0$, for all $x \in \mathcal{S},\ a \in \mathcal{A}$.

Let $\tau(x, a)$ define the expected sojourn time given that the state is x and the action a is chosen just before a transition, that is,

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

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

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

where $\mathbf{1}\{.\}$ 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. Thus, for $\boldsymbol{u} \in U_0$, there exist random variables $\{W(x,a)\}$ such that

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

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

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

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

A well-known result from renewal theory [13] is that if $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous SMP, and if the embedded Markov chain $\{X_m, m \in \mathcal{N}\}$ is unichain then, the proportion of time spent in state y, that is,

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

exists. Since under a stationary policy $\boldsymbol{f}$, the process $\{Y_t = (S_t, B_t) : t \geq 0\}$ is a homogeneous SMP, 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

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

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,

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

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,

$$\begin{aligned} v_{\boldsymbol{f}}(x,a) &= \frac{\pi_x(\boldsymbol{f})f_{xa}}{\sum_{x,a}\tau(x,a)\pi_x(\boldsymbol{f})f_{xa}} \\ &= \frac{z_{\boldsymbol{f}}(x,a)}{\sum_{x,a}\tau(x,a)z_{\boldsymbol{f}}(x,a)}. \end{aligned} \tag{1}$$

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

Reward Structure

Let R_t be the reward function at time t. R_t can be an impulse function corresponding to the reward earned immediately at a transition epoch and/or it can be a step function between transition epochs corresponding to the rate of reward as described below. The decision maker earns an immediate reward $R(X_m, A_m)$ and a reward with rate $r(X_m, A_m)$ until the $(m+1)$-th epoch, that is,

$$R_t = \begin{cases} R(X_m, A_m), & \text{if } t = T_m, \\ r(X_m, A_m), & \text{if } t \in [T_m, T_{m+1}). \end{cases} \tag{2}$$

Thus,

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

is the reward earned during the $(m+1)$-th transition [20–22].

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 $\bar{c}(x,a) \triangleq C(x,a) + c(x,a)\tau(x,a)$.

Example 1. Consider the *machine replacement problem* mentioned in the section titled "Introduction" with states (1) machine is in good condition, (2) machine has some minor problems, (3) machine is down and needs to be replaced. Time to failure of the machine follows a Weibull distribution with scale parameter equal to 8000 h and shape parameter equal to 4. The failure is minor with probability 0.95 and major requiring replacement of the machine with probability 0.05. Life time of a machine with minor problems follows a Weibull distribution with scale parameter equal to 20,000 h and shape parameter equal to 4. However, a machine with minor problems could be maintenance repaired to make it as good as new. Maintenance repair takes Weibull distributed amount of time with scale parameter equal to 8 h and shape parameter equal to 0.5. On the other hand, the machine replacement time is normally distributed with mean 72 h and variance of 8 h. Running a fully working machine earns \$100/h, and a machine with minor problem earns \$75/h profit. It costs \$40/h to repair and \$10,000 to replace a machine. Note that there is no control action available in states 1 and 3. In state 1 the decision maker needs to “wait” and in state 3 s/he needs to order a new machine. Let us denote this action as action 1. In state 2, there are two possible actions to choose: “wait” action denoted as action 1, and “initiate repair” denoted as action 2. Parameters of this model are:

$$P_{113} = 0.95, \quad P_{113} = 0.05, \quad P_{213} = 1, \quad P_{221} = 1, \quad P_{311} = 1,$$

$$\tau(1,1) = 7252, \quad \tau(2,1) = 1813, \quad \tau(2,2) = 48, \quad \tau(3,1) = 72,$$
$$\bar{r}(1,1) = 725,200, \quad \bar{r}(2,1) = 135,975, \quad \bar{r}(2,2) = -1920, \quad \bar{r}(3,1) = -10,000.$$

The last two reward values correspond to the incurred costs under repair and replacement, respectively.

PERFORMANCE MEASURES

We will focus on the optimality criteria over the infinite horizon, since some general results could be obtained for these models. We will first consider finding a policy $\boldsymbol{u}$ that will maximize the *total discounted reward* defined as

$$\phi_\alpha(\boldsymbol{u}) \triangleq E_{\boldsymbol{u}}\left[\int_0^\infty e^{-\alpha s} R_s \, \mathrm{d}s\right], \tag{3}$$

where α represents the discount factor [14,23–26]. Discounted reward optimality criterion is easier to analyze and understand than the average reward criterion, because the results for these models hold regardless of the chain structure of the embedded Markov chain. In fact, the existence of this integral is immediate under finite rewards. In addition, discounting lands itself naturally in economic problems in which the present value of future earnings is discounted as a function of the interest rate. Another interpretation of these models implies the importance of the initial decisions.

The great majority of the literature, on the other hand, is concerned with the *long run average expected reward* criterion with

$$\phi_1(\boldsymbol{u}) \triangleq \liminf_{t\to\infty} \frac{1}{t} E_{\mathbf{u}}\left[\int_0^t R_s \, \mathrm{d}s\right], \tag{4}$$

ϕ_1 denoting the *long run average expected reward* [14,17,22,27–29]. The following alternative to ϕ_1 is given by Jewell [30], Ross [31,32], and Mine and Osaki [18] as

$$\phi_2(\boldsymbol{u}) \triangleq \liminf_{n\to\infty} \frac{E_{\mathbf{u}}\left[\sum_{m=1}^{n} R_m\right]}{E_{\mathbf{u}}[T_n]}, \tag{5}$$

referred to as the *ratio-average reward* [33]. The performance measure ϕ_2 is also used by other researchers [7,14,34–38].

Let

$$\phi_\alpha^* = \sup_{\boldsymbol{u}\in U} \phi_\alpha(\boldsymbol{u}), \quad \phi_1^* = \sup_{\boldsymbol{u}\in U} \phi_1(\boldsymbol{u}), \quad \phi_2^* = \sup_{\boldsymbol{u}\in U} \phi_2(\boldsymbol{u}).$$

A policy $\boldsymbol{u}$ is optimal for $\phi_\alpha(\cdot)$ if $\phi_\alpha(\boldsymbol{u}) = \phi_\alpha^*$. For a fixed $\epsilon > 0$, a policy $\boldsymbol{u}$ is ϵ-optimal for $\phi_\alpha(\cdot)$ if $\phi_\alpha(\boldsymbol{u}) > \phi_\alpha^* - \epsilon$. Optimality and ϵ-optimal for $\phi_1(\cdot)$, $\phi_2(\cdot)$ and the other performance measures we consider in this article are defined analogously.

The following *expected time-average reward* criterion has been considered recently by Baykal-Gürsoy and Gürsoy [39–41]

$$\psi(\boldsymbol{u}) \triangleq E_{\mathbf{u}}\left[\liminf_{t\to\infty} \frac{1}{t}\int_0^t R_s\,\mathrm{d}s\right] \tag{6}$$

subject to the sample path constraint,

$$P_{\mathbf{u}}\left\{\limsup_{t\to\infty} \frac{1}{t}\int_0^t C_s\,\mathrm{d}s \le \gamma\right\} = 1. \tag{7}$$

This constraint requires that the long-run average costs on almost all sample paths should be bounded by γ.

More generally, they investigate the following *expected time-average variability*

$$\nu(\boldsymbol{u}) \triangleq E_{\mathbf{u}}\left[\liminf_{t\to\infty} \frac{1}{t}\int_0^t h\big(R_s, \frac{1}{t}\int_0^t R_q\,\mathrm{d}q\big)\,\mathrm{d}s\right], \tag{8}$$

where $h(.,.)$ is a continuous function of the current reward at time s and the average reward over an interval that includes time s. By letting $\nu^* = \sup_{\boldsymbol{u}\in U}\nu(\boldsymbol{u})$, the optimality and ϵ-optimality for $\nu(.)$ are analogously defined.

DISCOUNTED REWARD CRITERION

Discounted reward can be rewritten as:

$$\begin{aligned}\phi_\alpha(\boldsymbol{u}) &= E_{\boldsymbol{u}}\left[\sum_{m=0}^{\infty} \mathrm{e}^{-\alpha T_m}\left(R(X_m, A_m) + \frac{r(X_m, A_m)}{\alpha}(1-\mathrm{e}^{-\alpha\Upsilon_m})\right)\right]\\ &= \sum_{m=0}^{\infty}\sum_{x,a}\int_0^\infty \mathrm{e}^{-\alpha t}\left[R(x,a) + \frac{r(x,a)}{\alpha}\times\left(1-\sum_y P_{xay}\int_0^\infty \mathrm{e}^{-\alpha\tau}\,\mathrm{d}F_{xay}(\tau)\right)\right]\\ &\quad P_{\boldsymbol{u}}\{X_m = x,\ A_m = a, T_m \le t\}.\end{aligned}$$

The terms inside the second integral could be recognized as the Laplace transform of the density function $f_{xay}(\cdot)$ and will be denoted as $\tilde{f}_{xay}(\alpha)$.

The optimal discounted reward vector is represented by ϕ_α^{*x} for each initial state x, and it can be shown that it satisfies the optimality equation for all $x \in \mathcal{S}$:

$$\begin{aligned}\phi_\alpha^x &= \max_a\left\{\left[R(x,a) + \frac{r(x,a)}{\alpha}\left(1-\sum_j P_{xaj}\tilde{f}_{xaj}\right)\right] + \sum_y P_{xay}\tilde{f}_{xay}(\alpha)\phi_\alpha^y\right\}\\ &= \max_a\left\{r^\alpha(x,a) + \sum_y P_{xay}^\alpha\phi_\alpha^y\right\}.\end{aligned} \tag{9}$$

Second equality is obtained from the first by denoting the terms inside the square bracket as $r^\alpha(x,a)$ and writing $P_{xay}\tilde{f}_{xay}(\alpha)$ as P_{xay}^α. Note that the second equality is similar to the one obtained for the Markov Decision Processes (MDPs) (see also ***Total Expected Discounted Reward MDPs: Existence of Optimal Policies***). Thus, discounted SMDPs can be reduced to discounted MDPs by using these transformations. Since $P_{xay}^\alpha < 1$, the right hand side of the optimality Equation (9) is a contraction mapping and the next theorem is immediate.

Theorem 1. *For SMDPs under the discounted reward criterion:*

- *There exists a unique solution to the optimality equation (9) and it is equal to* ϕ_α^*.
- *There exists an optimal pure policy* $\boldsymbol{g}^*$ *given by,* $\phi_\alpha(\boldsymbol{g}^*) = \phi_\alpha^* = (I - P_{xay}^\alpha)^{-1} r^\alpha(\boldsymbol{g}^*)$ *where* $r^\alpha(\boldsymbol{g}^*$ *denotes the single-period discounted reward earned under policy* $\boldsymbol{g}^*$.

This optimal pure policy could be obtained using the policy iteration (see also ***Total Expected Discounted Reward MDPs: Policy Iteration Algorithm***), value iteration (see also ***Total Expected Discounted Reward MDPs: Value Iteration Algorithm***), or linear programming (LP) algorithms [5–7,38]. The LP algorithm is discussed next. Consider the following LP with given numbers $\beta_x > 0$ for $x \in \mathcal{S}$.

$$\max \sum_{x\in\mathcal{S},a\in\mathcal{A}} r^{\alpha}(x,a)z(x,a)$$
$$s.t. \sum_{x\in\mathcal{S},a\in\mathcal{A}} (\delta_{xy} - P^{\alpha}_{xay})z(x,a) = \beta_y, \quad y \in \mathcal{S}$$
$$z(x,a) \geq 0, \quad x \in \mathcal{S},\ a \in \mathcal{A}.$$

Let $\boldsymbol{z}^*$ be an optimum solution of the above LP. Clearly, any extreme point of this LP has $|\mathcal{S}|$ number of basic variables, where $|\cdot|$ denotes the number of elements in a given set. Thus, $z^*(x,a)$ is positive only for one action a. The optimum pure policy $\boldsymbol{g}^*$ is then obtained by assigning $\boldsymbol{g}^*_x$ in such a way that $z^*(x,\boldsymbol{g}^*_x) > 0$. Constraints defined in a similar fashion,

$$E_{\boldsymbol{u}}\left[\int_0^{\infty} e^{-\alpha s} C_s \,\mathrm{d}s\right] < \gamma,$$

could be included into the LP as

$$\sum_{x\in\mathcal{S},a\in\mathcal{A}} c^{\alpha}(x,a)z(x,a) < \gamma,$$

with $c^{\alpha}(x,a) = C(x,a) + \frac{c(x,a)}{\alpha}(1 - \sum_y P_{xay}\tilde{f}_{xay})$. Since every new constraint will increase the number of basic variables, the optimum policy will no longer be pure but randomized stationary [24].

For the countable state case, we need the assumption,

Assumption 1. There exists $\delta > 0$ and $\varepsilon > 0$, such that

$$F_{xay}(\delta) \leq 1 - \varepsilon \quad \text{for all } x \text{ and } y \in \mathcal{S} \text{ and a } \in \mathcal{A},$$

together with $|r^{\alpha}(x,a)| \leq M < \infty$ to ensure the existence of an optimal pure policy. Additional conditions are required for SMDPs with Borel state and action spaces, and unbounded rewards [21,24,38].

AVERAGE REWARD CRITERION

Average or *ratio-average* expected reward criterion is applied to systems in which the system dynamics is not slow enough to warrant discounting. This criterion is more difficult to analyze since the existence of the optimal stationary policy depends on the chain structure. Under the condition that the SMDP is irreducible, $\phi_1(\boldsymbol{f}) = \phi_2(\boldsymbol{f})$ for every stationary policy $\boldsymbol{f}$ [18,31]. However, this may not hold even for unichain SMDPs [33]. While 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 [22,29,42]. On the other hand, for finite state and finite action SMDPs, there exists an optimal pure policy under ϕ_1 [27,29,42], while such an optimal policy may not exist under ϕ_2 in a general multichain SMDP [43]. Jianyong and Xiaobo [43] investigate average reward SMDPs focusing on ϕ_2 and using a data-transformation method [19]. They show that the optimal pure policy exists in some special cases such as the unichain case and the weakly communicating case.

The optimal pure policy for the average expected reward criterion in multichain SMDPs is obtained from the optimal solution of the following LP [25] under the assumption on the sojourn times.

$$max \sum_{x\in\mathcal{S},a\in\mathcal{A}} \overline{r}(x,a)v(x,a)$$
$$s.t. \sum_{x\in\mathcal{S},a\in\mathcal{A}} (\delta_{xy} - P_{xay})v(x,a) = 0, \quad y \in \mathcal{S}$$
$$\sum_{a\in\mathcal{A}} \tau(y,a)v(y,a) + \sum_{x\in\mathcal{S},a\in\mathcal{A}} (\delta_{xy} - P_{xay})t(x,a) = \beta_y, \quad y \in \mathcal{S}$$
$$v(x,a) \geq 0,\ t(x,a) \geq 0 \quad x \in \mathcal{S},\ a \in \mathcal{A},$$

where $\beta_x > 0$ for $x \in \mathcal{S}$ and $\sum_y \beta_y = 1$. The optimum average expected reward for each initial state is obtained from the dual of this LP.

In the unichain case, the average reward remains constant regardless of the initial state, and the LP reduces to

$$\begin{aligned}\phi_1^* = max & \sum_{x\in\mathcal{S},a\in\mathcal{A}} \bar{r}(x,a)v(x,a) \\ s.t. & \sum_{x\in\mathcal{S},a\in\mathcal{A}} (\delta_{xy} - P_{xay})v(x,a) = 0, \quad y \in \mathcal{S} \\ & \sum_{x\in\mathcal{S},a\in\mathcal{A}} \tau(y,a)v(y,a) = 1 \\ & v(x,a) \geq 0, \quad x \in \mathcal{S},\ a \in \mathcal{A},\end{aligned}$$

with the optimum solution denoted as $\boldsymbol{v}^*$. The optimum pure policy $\boldsymbol{g}^*$ is then obtained by assigning $\boldsymbol{g}_x^*$ in such a way that $v^*(x,\boldsymbol{g}_x^*) > 0$. The optimality equations are given for each state x by

$$\zeta_x = \max_a \left\{ \bar{r}(x,a) - g\tau(x,a) + \sum_y P_{xay}\zeta_y \right\}.$$

The solution to these equations, $\{\zeta^*, g^*\}$ provides the optimum average expected reward, $\phi_1^* = g^*$.

The constrained problem has been investigated for the average reward SMDPs [20,33,34]. Beutler and Ross [20,34] consider the ratio-average reward with a constraint under a condition stronger than the unichain condition. In Ref. 33, Feinberg examines 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 shows that for both criteria, there exist optimal mixed stationary policies when an associated LP is feasible. The mixed policies are defined as policies with an initial one-step randomization applied to a set of pure policies, hence they are not stationary. He provides an LP algorithm for the unichain SMDP under both criteria. Average expected reward SMDPs with Borel state and action spaces and unbounded rewards are considered by Schäl [42], Sennott [21], and Luque-Vásquez and Hernández-Lerma [44].

EXPECTED TIME-AVERAGE REWARD AND VARIABILITY

The expected time-average reward criterion is similar to the average expected reward criterion. Fatou's lemma immediately implies that $\psi(\boldsymbol{u}) \leq \phi_1(\boldsymbol{u})$ holds for all policies. Baykal-Gürsoy and Gürsoy [40] show that for a large class of policies, these two rewards are equal and an ϵ-optimal randomized stationary policy can be obtained for the general (communicating, multichain) SMDP, while such a policy may not exist for the average reward problem [40]. Multiple constraints and the more general expected time-average variability criterion are also discussed. They show that an ϵ-optimal stationary policy can be obtained for the general SMDPs. 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. A decomposition algorithm to locate the ϵ-optimal stationary policy for both problems is given in Ref. 40. This algorithm utilizes an LP of the form:

$$\begin{aligned} max & \sum_{x\in\mathcal{S},a\in\mathcal{A}} h\left[\bar{r}(x,a), \sum_{y\in\mathcal{S},b\in\mathcal{A}} \bar{r}(y,b)v(y,b)\right]v(x,a) \\ s.t. & \sum_{x\in\mathcal{S},a\in\mathcal{A}} (\delta_{xy} - P_{xay})v(x,a) = 0, \quad y \in \mathcal{S} \\ & \sum_{x\in\mathcal{S},a\in\mathcal{A}} \tau(x,a)v(x,a) = 1 \\ & \sum_{x\in\mathcal{S},a\in\mathcal{A}} \bar{c}(x,a)v(x,a) \leq \gamma \\ & v(x,a) \geq 0, \quad x \in \mathcal{S},\ a \in \mathcal{A}.\end{aligned}$$

REFERENCES

1. Hajek B. Optimal control of two interacting service stations. IEEE Trans Automatic Control 1984;29:491–499.
2. Stidham S, Weber RR Jr. A survey of Markov decision models for control of networks of queues. Queuing Syst 1993;13:291–314.
3. Sennott LI. Stochastic dynamic programming and the control of queuing systems. New York: Wiley and Sons; 1999.
4. Feinberg EA, Kella O. Optimality of D-policies for an M/G/1 queue with a removable server. Queueing Syst 2002;42:355–376.
5. Tijms HC. Stochastic modeling and analysis: a computational approach. Chichester: Wiley; 1986.
6. Tijms HC. A first course in stochastic models. Chichester, England: Wiley; 2003.

7. Heyman DP, Sobel MJ. Stochastic models in operations research-volume II. New York: McGraw-Hill Book Company; 1984.
8. Ross SM. Introduction to stochastic dynamic programming. New York: Academic Press; 1983.
9. Pinedo M. Scheduling: theory, algorithms, and systems. Englewood Cliffs: Prentice Hall; 1995.
10. Baykal-Gürsoy M, Altiok T, Danhong H. Control policies for two-stage production/inventory systems with constrained average cost criterion. Int J Prod Res 1994;32:2005–2014.
11. Baykal-Gürsoy M, Altiok T, Danhong H. Look-back policies for two-stage, pull-type production/inventory systems. Ann Oper Res 1994;48:381–400. Special Issue on Queueing Networks.
12. Hu Q, Yue W. Optimal replacement of a system according to a semi-Markov decision process in a semi-Markov environment. Optim Methods Softw 2003;18(2):181–196.
13. Çinlar E. Introduction to stochastic processes. New Jersey: Prentice Hall; 1975.
14. Jewell WS. Markov renewal programming I: Formulation, finite return models. J Oper Res 1963;11:938–948.
15. De Cani JS. A dynamic programming algorithm for embedded Markov chains when the planning horizon is at infinity. Manage Sci 1964;10(4):716–733.
16. Denardo EV. Markov renewal programs with small interest rate. Ann Math Stat 1971;42:477–496.
17. Fox B. Markov renewal programming by linear fractional programming. SIAM J Appl Math 1966;16:1418–1432.
18. Mine H, Osaki S. Markovian decision processes. New York: Elsevier; 1970.
19. Schweitzer PJ. Iterative solution of the functional equations of undiscounted Markov renewal programming. J Math Anal Appl 1971;34:495–501.
20. Beutler FJ, Ross KW. Time-average optimal constrained semi-Markov decision processes. Adv App Prob 1986;18:341–359.
21. Sennott LI. Average cost semi-Markov decision processes and the control of queuing systems. Prob Eng Info Sci 1989;3:247–272,.
22. Yushkevich AA. On semi-Markov controlled models with an average reward criterion. Theory Probab Appl 1981;26:796–802.
23. Schweitzer PJ, Puterman ML, Kindle KW. Iterative aggregation-disaggregation procedures for solving discounted semi-Markovian reward processes. Oper Res 1985;33:589–605.
24. Feinberg EA. Constrained discounted semi-Markov decision processes. In: How Z, Filar JA, Chen A, editors. Markov Processes and Controlled Markov Chains. Dordrecht, The Netherlands: Kluwer; 2002. pp. 233–244.
25. Kallenberg LCM. Linear programming and finite Markovian control problems. Volume 148, Amsterdam: Mathematical Centre Tracts; 1983.
26. Lippman SA. On dynamic programming with unbounded rewards. Manage Sci 1975;21(11):1225–1233.
27. Denardo EV, Fox BL. Multichain Markov renewal programs. SIAM J Appl Math 1968;16:468–487.
28. Lippman SA. Maximal average reward policies for semi-Markov renewal processes with arbitrary state and action spaces. Ann Math Stat 1971;42:1717–1726.
29. Schweitzer PJ, Federgruen AF. The functional equations of undiscounted Markov renewal programming. Math Oper Res 1978;3:308–321.
30. Jewell WS. Markov renewal programming II: Infinite return models, example. J Oper Res 1963;11:949–971.
31. Ross SM. Average cost semi-Markov processes. J Appl Probab 1970;7:649–656.
32. Ross SM. Applied probability models with optimization applications. San Francisco: Holden-Day; 1971.
33. Feinberg EA. Constrained semi-Markov decision processes with average rewards. Math Meth Oper Res 1994;39:257–288.
34. Beutler FJ, Ross KW. Uniformization for semi-Markov decision processes under stationary policies. Adv App Probab 1987; 24:644–656.
35. Federgruen A, Tijms HC. The optimality equation in average cost denumerable state semi-Markov decision problems, recurrency conditions and algorithms. J Appl Probab 1978;15:356–373.
36. Federgruen A, Hordijk A, Tijms HC. Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion. Stoch Proc Appl 1979;9:223–235.
37. Federgruen A, Schweitzer PJ, Tijms HC. Denumerable undiscounted semi-Markov decision processes with unbounded rewards. Math Oper Res 1983;8(2):298–313.

38. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: John Wiley and Sons; 1994.
39. Baykal-Gürsoy M. Semi-Markov decision processes: nonstandard criteria. *IEEE Regional NY/NJ Conference*, New Brunswick (NJ) 1994.
40. Baykal-Gürsoy M, Gürsoy K. Semi-Markov decision processes: Nonstandard criteria. Prob Eng Info Sci 2007;21:635–657.
41. Baykal-Gürsoy M, Ross KW. Variability sensitive Markov decision processes. Math Oper Res 1992;17:558–571.
42. Schäl M. On the second optimality equation for semi-Markov decision models. Math Oper Res 1992;17(2):470–486.
43. Jianyong L, Xiaobo Z. On average reward semi-Markov decision processes with a general multichain structure. Math Oper Res 2004;29(2):339–352.
44. Luque-Vásquez F, Hernández-Lerma O. Semi-Markov control models with average costs. Appl Math 1999;26(3):315–331.