MATHEMATICS OF OPERATIONS RESEARCH
Vol. 29, No. 3, August 2004, pp. 492–524
ISSN 0364-765X | EISSN 1526-5471 | 04 | 2903 | 0492

DOI 10.1287/moor.1040.0089

Continuous Time Discounted Jump Markov Decision Processes: A Discrete-Event Approach

Eugene A. Feinberg
Department of Applied Mathematics and Statistics, SUNY at Stony Brook,
Stony Brook, New York 11794-3600, eugene.feinberg@sunysb.edu

This paper introduces and develops a new approach to the theory of continuous time jump Markov decision processes (CTJMDP). This approach reduces discounted CTJMDPs to discounted semi-Markov decision processes (SMDPs) and eventually to discrete-time Markov decision processes (MDPs). The reduction is based on the equivalence of strategies that change actions between jumps and the randomized strategies that change actions only at jump epochs. This holds both for one-criterion problems and for multiple-objective problems with constraints. In particular, this paper introduces the theory for multiple-objective problems with expected total discounted rewards and constraints. If a problem is feasible, there exist three types of optimal policies: (i) nonrandomized switching stationary policies, (ii) randomized stationary policies for the CTJMDP, and (iii) randomized stationary policies for the corresponding SMDP with exponentially distributed sojourn times, and these policies can be implemented as randomized strategies in the CTJMDP.

Key words: continuous time jump Markov decision process; semi-Markov decision process
MSC2000 subject classification: Primary: 90C40
OR/MS subject classification: Primary: Dynamic programming/optimal control/Markov
History: Received December 16, 1998; revised February 1, 2003, August 31, 2003, and November 26, 2003.

1. Introduction. We introduce a new approach to discounted continuous-time jump Markov decision processes (CTJMDPs). This approach is based on the fact that strategies that change actions between jumps yield the same performance as strategies that select actions only at jump epochs. The latter strategies are the strategies in the corresponding semi-Markov decision processes (SMDPs). Because a discounted SMDP can be reduced to a discounted discrete-time Markov decision process (MDP), our approach reduces discounted CTJMDPs to discounted MDPs. This reduction takes place both for one-criterion problems and for problems with multiple criteria and constraints.

In addition, we prove in this paper that solutions for constrained discounted CTJMDPs can be found in a more natural form than solutions for constrained discounted MDPs. For constrained discounted MDPs, nonrandomized optimal strategies may not exist. However, for infinite-horizon problems there exist optimal randomized stationary policies for which the number of additional randomization procedures is limited by the number of constraints; see, e.g., Altman (1999, p. 34) or Feinberg and Shwartz (1996). In this paper, we show that optimal solutions for constrained discounted CTJMDPs can be found among the nonrandomized strategies that switch decisions between jumps, and the number of switching points is limited by the number of constraints.

The classical approach to the analysis of infinite-horizon CTJMDP (Miller 1968, 1968a and Yushkevich 1977, 1980a) is to consider a finite-horizon problem on an interval $[0, T]$, derive optimality equations in a form of differential equations, and then take $T \to \infty$. Instead of fixing the horizon $[0, T]$, we fix the maximum number of jumps N and consider the problem when the process stops at the Nth jump epoch. For infinite-horizon problems, we consider $N = \infty$. Our approach is based on Feinberg (1994), where the horizon was the interval between time 0 and the first jump epoch. For any strategy that changes actions over time, Corollary 1 in Feinberg (1994) provides a probability distribution to select constant actions in such a way that the following two characteristics remain unchanged: (i) the expected rewards until the first jump, and (ii) the probability distribution of an action at the first jump epoch. This paper extends this result to the discounted problems with the total number of jumps N being any positive integer including $N = \infty$ and this extension is accomplished by using occupation measures; see Theorem 4.5.

At least three reductions of CTJMDPs to MDPs have been developed in the literature. Lippman (1975) introduced uniformization to CTJMDPs; see also Cassandras (1993), Puterman (1994), and Bertsekas (1995). By considering fictitious jumps, uniformization relates the performance of (nonrandomized) stationary policies in CTJMDPs to the performance of the corresponding policies in MDPs. Kakumanu (1977) and Serfozo (1979) observed that optimality equations for one-criterion infinite-horizon CTJMDPs coincide with the optimality equations for the corresponding MDPs. Therefore, algorithms for discrete time one-criterion problems also find optimal policies for the corresponding continuous-time problems. Yushkevich (1980) considered a reduction of CTJMDPs to MDPs with control actions equal to functions that select actions in CTJMDPs between jump epochs.

Our reduction is different from the above reductions. First, unlike uniformization, it is applicable to all strategies in CTJMDPs. Second, it establishes the correspondence between occupation measures, which is a deeper result than the correspondence between optimality equations. In particular, our reduction is applicable to problems with multiple criteria and constraints. Third, compared to the reduction in Yushkevich (1980), the corresponding MDP has the same state and action sets as the original CTJMDP, while the action sets of the MDP considered by Yushkevich (1980) were much more complicated objects: functions on $[0, \infty)$ with the values in the sets of actions.

In fact, our reduction of a CTJMDP consists of two steps. First, we reduce a CTJMDP to an SMDP. An SMDP is a generalization of an MDP to continuous time when sojourn times are random variables, while in MDPs all sojourn times are equal to 1. If the strategies are allowed to select actions only at jump epochs in a CTJMDP, this CTJMDP becomes a special case of an SMDP with all sojourn times having exponential distributions. Such SMDPs, called exponential continuous-time jump Markov decision processes (ESMDPs) in this paper, are much simpler objects than CTJMDPs. In fact, because of this simplicity, the material on CTJMDPs in almost all textbooks (Bertsekas 1995, Cassandras 1993, Sennott 1999) is limited to ESMDPs. Second, discounted SMDPs can be reduced to discounted MDPs; see Appendix A for details.

Section 2 introduces CTJMDPs. Section 3, which was written for the reader's convenience, provides informal description of models and strategies considered in this paper.

Section 4 reduces discounted CTJMDPs to discounted ESMDPs. We introduce occupation measures for CTJMDPs, and for a strategy in a CTJMDP we construct in Theorem 4.5 a randomized Markov policy in the corresponding ESMDP such that their occupation measures coincide. For a randomized Markov policy in an ESMDP, the decisions are selected only at jump epochs and depend only on the current states and on the current jump numbers. In our opinion, Theorem 4.5 is the central statement of this paper. This theorem is similar to Lemma A.2, which, for a discounted criterion, generalizes from MDPs to SMDPs the well-known result by Derman and Strauch (1966).

Section 5 establishes the converse of Theorem 4.5. Under certain conditions, Theorem 5.2 constructs an equivalent switching Markov policy for a randomized Markov policy in the corresponding ESMDP. For switching Markov policies, decisions depend on the current state, the jump number, and the time passed since the last jump. If the decisions depend only on two parameters—the current state and the time passed since the last jump—the strategy is called a switching stationary policy. Thus, §§4 and 5 establish the equivalence between the switching stationary (or Markov) policies in a CTJMDP and the randomized stationary (respectively, Markov) policies in the corresponding ESMDP.

Section 6 deals with the optimization problems. Because the theory of constrained optimization is better developed for countable MDPs than for Borel state MDPs, we restrict our attention to countable state problems in §6. We establish the existence of optimal switching stationary policies, and in addition, the number of switching points is bounded by the number of constraints. We discuss linear programs for finding such policies.

In §§2 and 4–6 we limit our consideration to only nonrandomized strategies. In fact, even in these sections we consider more general classes of strategies than is usually done in the literature on CTJMDPs, because early papers on CTJMDPs (Miller 1968, 1968a, and Kakumanu 1971) and almost all later papers considered (nonrandomized) Markov strategies as the most general class of strategies. For a Markov strategy, decisions are functions of the current state and time. Yushkevich (1977) introduced past-dependent strategies, and Kitaev (1985) gave an equivalent definition of past-dependent strategies based on the paper by Jacod (1975); see also Chapter 4 in Kitaev and Rykov (1995).

Section 7 deals with randomized strategies for CTJMDPs. We prove that switching stationary (and Markov) policies are optimal within the larger class of randomized strategies. Randomized Markov strategies for CTJMDPs were introduced by Hordijk and van der Duyn Schouten (1979) and general randomized strategies were introduced by Kitaev (1985). In §8 we prove in Theorem 8.3 the existence of optimal randomized policies for CTJMDPs such that the number of randomization procedures is not greater than the number of constraints. We also prove in Theorem 8.4 that a randomized policy for an ESMDP can be implemented as a randomized policy for the corresponding CTJMDP. Thus, for infinite-horizon constrained discounted CTJMDPs, there exist three types of optimal strategies: (i) switching stationary policies, (ii) randomized stationary policies for ESMDPs, and (iii) randomized stationary policies for CTJMDPs. The number of switching points in (i) and the number of randomization procedures in (ii) and (iii) is less than or equal to the number of constraints. In our opinion, switching policies are the most natural among (i)–(iii). Appendix A contains all necessary results for SMDPs and their reduction to MDPs.

2. Definitions. The probability structure of a CTJMDP is defined by $\{X, A, D(x), q(Y|x, a)\}$, where

(i) X is a Borel state space (a measurable space $(X, \mathscr{X})$ for which there exists a one-to-one measurable correspondence onto a Borel subset of a separable complete metric space. The σ-field $\mathscr{X}$ is called Borel; see Dynkin and Yushkevich 1979, Appendix 1, for details).

(ii) A is a Borel action space endowed with the Borel σ-field $\mathscr{A}$.

(iii) $D(x)$ are Borel sets of actions available at $x \in X$. It is assumed that

$$\text{graph } D = \{(x, a)\colon x \in X, a \in D(x)\}$$

is a Borel subset of $X \times A$ containing the graph of a Borel mapping from X to A (the Borel σ-field on $X \times A$ is the minimal σ-field on $X \times A$ containing all sets $Y \times B$ with $Y \in \mathscr{X}$ and $B \in \mathscr{A}$).

(iv) $q(\cdot|x, a)$ is an infinitesimal characteristic that is a nonnegative measure on $(X, \mathscr{X})$ defined for all $x \in X$ and $a \in A$. It is assumed that q satisfies the following conditions: (i) $q(\{x\}|x, a) = 0$ for all x and a, (ii) there is some constant $M < \infty$ such that $q(X|x, a) < M$ for all x and a, and (iii) $q(Y|x, a)$ is a measurable function on $X \times A$ for all $Y \in \mathscr{X}$. Let $q(x, a) = q(X|x, a)$. If an action a is selected at the state x and $q(x, a) = 0$ then x is an absorbing state. If $q(x, a) > 0$ then the sojourn time has an exponential distribution with the intensity $q(x, a)$ and the next state belongs to $Y \in \mathscr{X}$ with probability $q(Y|x, a)/q(x, a)$.

For CTJMDPs it is possible to change actions any time. This capability does not exist in MDPs and SMDPs. It creates additional technical difficulties in constructions of stochastic processes defined by strategies. Starting from the beginning (Miller 1968; Kakumanu 1971, 1977), almost all studies of CTJMDPs considered nonrandomized Markov strategies as the most general class of strategies. For Markov strategies, the decisions are nonrandomized and depend only on the current state and time. Each Markov strategy defines a Markov process on the state space X.

Yushkevich (1977, 1980, 1980a) defined general nonrandomized strategies for which decisions depend on: (a) the previous states and jump epochs and (b) the current state and time.

We note that, given the knowledge of jump epochs in (a), the knowledge of the current time is equivalent to the knowledge of the time passed since the last jump. Yushkevich (1980) constructed appropriate stochastic processes and expectations by reducing a CTJMDP to the SMDP with actions being Borel mappings from $\mathbb{R}_+$ to the sets of available functions. If an action function f is selected when the system jumps to a state x, this function defines control actions as long as the system stays in x. If the time t passed since the last jump is t, the action is $f(t)$. Kitaev (1985) gave an equivalent construction using the results by Jacod (1975). In this section, we introduce these two constructions and explain why they are equivalent.

We consider only nonrandomized strategies in this section. Randomized strategies will be defined in §7 and studied in §§7 and 8. In particular, it will be shown there that nonrandomized strategies are optimal within the broader class of randomized strategies for problems with multiple criteria and constraints.

We describe the approach introduced by Yushkevich (1977, 1980) first. According to this approach, a CTJMDP is an SMDP in which an action that the controller selects when the system jumps to a state x is a Borel function on $\mathbb{R}_+$ with the values in $D(x)$. The value of such a function at t defines the action selected in the CTJMDP after time t has passed since the last jump to the current state x. In this SMDP the controller does not use information about previous actions.

Let $F_n = X \times (\mathbb{R}_+ \times X)^n$ be the set of histories $x_0, \xi_0, x_1, \xi_1, \ldots, x_n$ up to and including the nth jump, where x_i is a state after ith jump and ξ_i is the corresponding sojourn time. Let $F = \bigcup_{0\le n<\infty} F_n$ be the set of histories each of which has a finite number of jumps. A strategy φ is a Borel mapping from $F \times \mathbb{R}_+$ to A such that $\varphi(x_0, \xi_0, x_1, \xi_1, \ldots, x_n, t) \in D(x_n)$. In this definition, t is the time passed since the last jump epoch $t_n = \xi_0 + \xi_1 + \cdots + \xi_{n-1}$. The actual time is $s = t_n + t$.

Now we construct the stochastic processes defined by strategies. Following Yushkevich (1980), we define an SMDP with the state space X and the actions at state $x \in X$ being Borel functions from $\mathbb{R}_+$ to $D(x)$. Because the system does not change states between jumps, $D(x)$ is a valid range for action functions. We call this SMDP *a Yushkevich SMDP* (YSMDP). Strategies for the CTJMDP are nonrandomized strategies in the corresponding YSMDP.

We introduce formal definitions. Let B be a Borel space and Φ_B be the space of measurable functions on $\mathbb{R}_+$ with values in B. Two functions from Φ_B are equal if they are equal almost everywhere on $\mathbb{R}_+$ with respect to the Lebesgue measure on $\mathbb{R}_+$. There is a metric on Φ_B such that this set is Borel (Yushkevich 1980, Lemma 1).

We consider Borel sets $U = \Phi_A$ and $U(x) = \Phi_{D(x)}$, $x \in X$. We consider an SMDP with the state space X, action space U, sets of available actions $U(x)$, and transition kernel

$$(2.1) \qquad Q(t, Y|x, u) = \int_0^t q(Y|x,\, u(s))e^{-\int_0^s q(x,\, u(w))\, dw}\, ds,$$

where $u \in U(x)$, i.e., u is a Borel function from $\mathbb{R}_+$ to $D(x)$. Let $\{X, U, U(x), Q\}$ be, respectively, the state space, the set of all actions, the sets of actions available at $x \in X$, and the transition kernel for the YSMDP.

Let φ be a strategy for the CTJMDP. Then φ is also a strategy for the YSMDP defined by $\varphi(x_0, \xi_0, x_1, \ldots, x_n) = u$, where $u(t) = \varphi(x_0, \xi_0, x_1, \ldots, x_n, t)$, $t \in \mathbb{R}_+$. This interpretation of φ is correct because φ is a measurable mapping from F to U such that $\varphi(x_0, \xi_o, \ldots, x_n) \in U(x_n)$; see Yushkevich (1980, Lemma 3). Because a YSMDP is an SMDP, we can expand in a standard way X to $\bar{X}$ and U to $\bar{U}$; see Appendix A.1. An initial probability measure μ on X and a strategy φ define a probability measure $\mathbb{P}_\mu^\varphi$ on $(\Omega, \mathcal{F})$, where $\Omega = (\bar{X} \times \bar{R}_+)^\infty$ and $\mathcal{F}$ is the Borel σ-field on Ω defined by the products of the Borel σ-fields on X and $\mathbb{R}_+$. So an initial distribution μ and a policy φ define the probability space $(\Omega, \mathcal{F}, \mathbb{P}_\mu^\varphi)$.

Given $(\Omega, \mathscr{F}, \mathbb{P}_x^\varphi)$, we can define a marked point process X_s for which $(\Omega, \mathscr{F})$ is a set of internal histories. For $\omega = (x_0\xi_0x_1\xi_1, \ldots)$ we define $t_0 = 0$, $t_n = t_{n-1} + \xi_n$, $n = 1, 2, \ldots$, and $t_\infty = \lim_{n\to\infty} t_n$. Then $\{x_n, t_n\}$ is a marked point process and Ω is the set of internal histories of the process $X_s(\omega) = x_n I\{t_n \le s < t_{n+1}\} + \bar{x}I\{s \ge t_\infty\}$, $t \in \mathbb{R}_+$.

We can also define a process of actions $A_s(\omega) = \varphi(x_0, \xi_0, x_1, \xi_1, \ldots, x_n, s - t_n)I\{t_n \le s < t_{n+1}\} + \bar{a}I\{s \ge t_\infty\}$. Let a_n be an action that defines transition probabilities at $(n+1)$th jump, $a_n = \varphi(x_0, \xi_0, x_1, \xi_1, \ldots, x_n, \xi_n)$, $n = 0, 1, \ldots$. We remark that another natural notation for a_n would be $A_{t_{n+1}-}$.

An equivalent construction of stochastic processes defined by strategies was introduced by Kitaev (1985). Consider the space $(\Omega, \mathscr{F})$. Let $\mathscr{F}_s = \sigma\{X_s\}$. Then $A_s(\omega)$ is a predictable mapping from $(\Omega \times \mathbb{R}_+)$ to A such that $A_s(\omega) \in D(X_s)$ for all $s \in \mathbb{R}_+$; see Jacod (1975, Lemma 3.3). We define a predictable random measure

$$\nu^\varphi(\omega, ds, dx) = q(dx|X_s, A_s)\, ds, \tag{2.2}$$

where $\nu^\varphi(\omega, 0, dx) = \mu(dx)$. In view of Jacod's (1975) Theorem 3.6, an initial distribution μ and a strategy φ define a unique probability measure $\mathbb{P}_\mu^\varphi$ on $(\Omega, \mathscr{F})$ such that ν is the predictable projection of the random measure defined by $\mathbb{P}_\mu^\varphi$. Jacod's (1975) Proposition 3.1 implies that this measure $\mathbb{P}_\mu^\varphi$ coincides with the measure defined by the transition probabilities (2.1).

Now we define particular classes of strategies for CTJMDPs. Let $\omega_n = x_0, \xi_0, \ldots, \xi_{n-1}, x_n$. A strategy φ is called a (*switching*) *policy* if $\varphi(\omega_n, t)$ does not depend on $\xi_0, \ldots, \xi_{n-1}$. A policy φ is called nonswitching if $\varphi(x_0, \ldots, x_n, t)$ is constant in t for all $x_0, \ldots, x_n$, i.e., actions cannot be changed between jumps. In this case, we consider $\varphi(x_0, \ldots, x_n) = \varphi(x_0, \ldots, x_n, t)$ A nonswitching policy φ is called *a Markov policy* if $\varphi(x_0, \ldots, x_n) = \varphi(x_n, n)$ for all $\omega_n \in \Omega$. A Markov policy φ is called *a stationary policy* if $\varphi(x_n, n) = \varphi(x_n)$. We remark that Markov policies differ from Markov strategies considered in many papers on CTJMDPs as the most general class of possible strategies. For a Markov policy, decisions depend on the current state and the number of jumps that occurred. For a Markov strategy, decisions depend on the current state and time; $\varphi(\omega_n, t) = \varphi(x_n, t_n + t) = \varphi(X_s, s)$. However, we do not consider Markov strategies in this paper because, following Yushkevich (1977, 1980, 1980a) and Kitaev (1985), we have defined stochastic processes for general strategies.

Markov policies do not change actions between jumps. A (switching) policy φ is called a *switching Markov policy* if $\varphi(x_0, \ldots, x_n, t) = \varphi(x_n, n, t)$. For a switching Markov policy, decisions depend on the current state, the number of prior jumps, and the time passed since the last jump. For a switching Markov policy φ and for each couple $(x, n) \in X \times \{0, 1, \ldots\}$, we denote by $R^\varphi(x, n)$ the subset of $\mathbb{R}_+$ on which this function $\varphi(x, n, t)$ is discontinuous in t with (x, n) being fixed (we follow the convention that a function defined on $\mathbb{R}_+$ is continuous at 0 if it is right-continuous at 0).

Let Z^φ be the sets of all couples (x, n) such that the function $\varphi(x, n, t)$ is not constant in t for a switching Markov policy φ. If Z^φ is not finite, we say that a switching Markov policy φ has an infinite order. If Z^φ is finite, we define

$$N^\varphi = \sum_{(x,n)\in Z^\varphi} |R^\varphi(x, n)|$$

as a number of switching times for the switching Markov policy φ, where $|E|$ denotes the cardinality of the set E. We say that a switching Markov policy φ is *a switching Markov policy of order l* if $N^\varphi \le l$. A switching Markov policy of order 0 is a Markov policy.

A switching Markov policy φ is called *a switching stationary policy* if $\varphi(x_n, n, t) = \varphi(x_n, t)$. Let φ be a switching stationary policy. If φ is a switching stationary policy, $R^\varphi(x, n) = R^\varphi(x, m)$ for any $x \in X$ and for any $m, n = 0, 1, \ldots$. Therefore, a switching stationary policy is either a (nonswitching) stationary policy or has an infinite order. For a switching stationary policy φ and for $x \in X$ we define $R^\varphi(x)$ as a subset of $\mathbb{R}_+$ where the function $\varphi(x, t)$ is discontinuous in t for a fixed x. Let $X^\varphi = \{x \in X \mid R^\varphi(x) \neq \varnothing\}$. If X^φ is infinite we say that the policy φ is *an ∞-switching stationary policy.* If X^φ is finite, we say that the policy is *k-switching stationary* if $\sum_{x\in X^\varphi} |R^\varphi(x)| \le k$.

A switching Markov policy π of order m is called *an (m, n)-policy* if there exists a stationary policy φ such that $\pi(x, l, t) = \varphi(x)$ for all $x \in X$, $l \ge n$, and $t \in \mathbb{R}_+$. An (m, n)-policy is called *a strong (m, n)-policy* if there exists an m-switching stationary policy φ^* such that $\bigcup_{n=0}^\infty \bigcup_{t\in\mathbb{R}_+} \{\varphi(x, n, t)\} = \bigcup_{t\in\mathbb{R}_+} \{\varphi^*(x, t)\}$ for all $x \in X$.

We observe that a switching Markov policy is a switching stationary policy in the CTJMDP with the state space $X \times \{0, 1, \dots\}$. Also, a Markov policy of order k in the original CTJMDP is a k-randomized stationary policy in that CTJMDP and vice versa.

A reward structure of a discounted CTJMDP is defined by the following four objects:

(i) positive discount rate α;

(ii) number of constraints K;

(iii) reward rates $\bar{r}_k(x, a)$, $k = 0, \dots, K$; and

(iv) instantaneous rewards $R^*_k(x, a, y)$ earned if the process jumps from state x to state y and action a was chosen at x at jump epoch, $k = 0, \dots, K$.

The functions $\bar{r}_k$ and R^*_k, $k = 0, 1, \dots, K$, are assumed to be measurable and uniformly bounded above.

For an initial distribution μ, a policy φ, and $k = 0, \dots, K$, the expected total discounted rewards are

$$(2.3) \qquad W_k(\mu, \varphi) = \mathbb{E}^\varphi_\mu\left[\sum_{n=0}^\infty e^{-\alpha t_n} R^*_k(x_n, a_n, x_{n+1}) + \int_0^\infty e^{-\alpha t} \bar{r}_k(X_t, A_t)\, dt\right].$$

The expected total discounted rewards can be interpreted as the expected total reward for the YSMDP with the rewards

$$(2.4) \qquad r_k(x, u) = \mathbb{E}\left\{e^{-\alpha\xi_0} R^*_k(x_0, u(\xi_0), x_1) + \int_0^{\xi_0} e^{-\alpha t} \bar{r}_k(x_0, u(t))\, dt \middle| x_0 = x\right\}.$$

A strategy π is called *optimal* for a one-criterion CTJMDP if $W(\mu, \pi) \ge W(\mu, \sigma)$ for all initial distributions μ and for all strategies σ. For a CTJMDP with $(K + 1)$ criteria, we fix an initial distribution μ and numbers $C_1, \dots, C_K$. A strategy π is called *feasible* if $W_k(\mu, \pi) \ge C_k$, $k = 1, \dots, K$. If a feasible strategy exists, the CTJMDP is called *feasible*. A feasible strategy π is called *optimal* if $W_0(\mu, \pi) \ge W_0(\mu, \sigma)$ for all feasible strategies σ.

We sometimes omit indices $k = 0, 1, \dots, K$ for one criterion problems and for multiple criterion problems if the formula holds for all criteria. We assume throughout this paper that $0 \times \infty = 0$. If an initial distribution μ is concentrated at one point x, we substitute μ with x. Thus, we write $W(x, \pi)$, $\mathbb{P}^\pi_x$, and $\mathbb{E}^\pi_x$ instead of $W(\mu, \pi)$, $\mathbb{P}^\pi_\mu$, and $\mathbb{E}^\pi_\mu$ when $\mu(x) = 1$.

REMARK 2.1. We can consider a nonhomogeneous CTJMDP for which the sets of available actions D, infinitesimal characteristics q, and rewards $\bar{r}_k$, R^*_k, $k = 0, \dots, K$, depend on the jump number. The formal definitions are similar to the homogeneous CTJMDPs but the state parameter x should be replaced with (x, n), $n = 0, 1, \dots$, in the definitions of D, q, $\bar{r}_k$, and R^*_k. A nonhomogeneous CTJMDP is equivalent to a homogeneous CTJMDP with the state space $X \times \{0, 1, \dots\}$. Therefore, the existence of optimal K-switching policies for homogeneous CTJMDPs implies the existence of optimal switching Markov policies of order K for nonhomogeneous CTJMDPs. An important example of a nonhomogeneous CTJMDP is a model with a finite number of jumps (or steps). If $Q(X|x, N, a) = 0$ and $\bar{r}_k(x, N, a) = 0$ for all $a \in D(x, N)$, $x \in X$, and $k = 0, \dots, K$, then we deal with an

N-jump CTJMDP. In this case, the equivalent homogeneous model has the state space $X \times \{0, \ldots, N\}$. Then (2.3) transforms into

$$W_k(\mu, \varphi) = \mathbb{E}_\mu^\varphi\left[\sum_{n=0}^{N-1} e^{-\alpha t_n} R_k^*(x_n, n, a_n, x_{n+1}) + \int_0^{t_N} e^{-\alpha t} \bar{r}_k(X_t, N_t, A_t)\, dt\right],$$

and the terminal rewards $R_k^*(x, N-1, a, y)$ typically do not depend on x and a.

3. Informal descriptions of models and strategies.

In this section, written for the reader's convenience, we provide informal descriptions of the major models and major types of strategies considered in this paper.

SMDP, semi-Markov decision process: a generalization of a discrete-time MDP when sojourn times are not identical to 1. Actions are selected immediately following state transitions. After an action is selected, the sojourn time has a given arbitrary distribution. Though the state of an SMDP remains unchanged between the decision epochs, SMDPs considered in this paper can model the situations when the "underlying state" of the system may change between jumps, as it can take place in control of M/G/1 and GI/M/1 systems; see the paragraph following the definition of rewards r_k in Appendix A.1.

CTJMDP, continuous time Markov decision process: a continuous time version of an MDP in which actions may change any time. However, if the selected actions are constant between jumps, the sojourn times have exponential distributions. The cumulative reward at a state is the sum of two components: a discrete component collected at jump epochs and a continuous component defined by the reward rate that depends only on the current state and action.

ESMDP, exponential semi-Markov decision process: a simplified version of a CTJMDP when actions cannot be changed between jumps and all primitive data are the same as in a CTJMPD. An ESMDP is an SMDP in which the sojourn times, defined by all state-action pairs (x, a), are exponential.

YSMDP, Yushkevich semi-Markov decision process: a construction introduced by Yushkevich (1980). In fact, it is an SMDP with the actions being action-valued functions $f(t)$, $t \in \mathbb{R}_+$, that control a CTJMDP between jumps, where t is the time passed since the last transition.

[CTJMDP], a CTJMDP with the sets of actions being the sets of probability distributions on the sets of actions of the original CTJMDPs. The infinitesimal characteristics and reward rates are the corresponding convex combinations of the appropriate values for the CTJMDP; the rewards collected at jump epochs can be set equal to 0 without loss of generality. It is natural to call [CTJMDP] a convex hull of the original CTJMDP.

[ESMDP], an ESMDP for the [CTJMDP]. It is natural to call [ESMDP] a convex hull of the original ESMDP.

Strategies for SMDPs. *A history* is a finite sequence $x_0 a_0 \xi_0 \ldots x_{n-1} a_{n-1} \xi_{n-1} x_n$ of states, actions, and sojourn times. For a known history, *a strategy* selects an action. In general, actions can be selected randomly and a strategy can be *randomized*. If each trajectory defines an action deterministically then a strategy is called *nonrandomized*.

Policies for SMDPs. If instead of trajectories $x_0 a_0 \xi_0 \ldots x_{n-1} a_{n-1} \xi_{n-1} x_n$ the decision maker observes only $x_0 a_0 \ldots x_{n-1} a_{n-1} x_n$, a strategy is called *a policy*. In other words, the decision maker does not use the information about sojourn times. In particular, if a strategy is a policy, the current time s is unknown to the decision maker. However, the number of jumps that occurred may be known. Similar to strategies, policies can be randomized. We also may consider *nonrandomized policies*.

Because policies ignore the continuous time parameter s, a strategy is a policy if and only if it is a policy in an MDP with the same state and action spaces as in the given SMDP.

In this paper we follow the notation used in Feinberg and Shwartz (1996) for MDPs. Similar to MDPs, we consider *randomized Markov policies*, for which action selections depend only on the current state and the current jump number, and *randomized stationary policies* for which actions depend only on current states. If these policies are nonrandomized, they are called *Markov* or *stationary*, respectively.

A randomized stationary policy is called *m-randomized stationary*, $m = 0, 1, \ldots,$ if the number of additional actions used by randomization procedures is limited by m. A randomized Markov policy is called *a randomized Markov policy of the order m* if the total number of additional actions required by randomization procedures in all state-jump number pairs (x, n) is limited by m. A randomized Markov policy of order m is called *an (m, n)-policy* if it is (nonrandomized) stationary from jump n onwards. An (m, n)-policy is called *a strong (m, n)-policy* if the total number of additional actions used in all states because of its nonstationarity is limited by m. The details on SMDPs can be found in Appendix A.

Strategies for CTJMDPs. *A history* at the moment of the nth jump is $\omega_n = x_0\xi_0 \ldots x_{n-1}\xi_{n-1}x_n$, where $x_0, x_1, \ldots, x_n$ are the states and $\xi_0, \xi_1, \ldots, \xi_{n-1}$ are the sojourn times. If n jumps have occurred prior to the current time, the history is (ω_n, t), where $t \geq 0$ is the time passed since the last jump. *A (nonrandomized) strategy* selects for any history (ω_n, t) an action from the set of actions available at the current state x_n. A strategy is called a *policy* if these selections do not depend on the past sojourn times $\xi_0, \xi_1, \ldots, \xi_{n-1}$. A policy is called *nonswitching* if it does not change actions between jumps. Thus, nonswitching policies are in fact nonrandomized policies in the corresponding ESMDP and MDP. The only minor difference is that the nonrandomized policies in SMDPs and MDPs know the history of selected actions. However, this additional information is unimportant because nonrandomized policies for MDPs in fact do not use it. Indeed, for any nonrandomized policy φ for an MDP that uses the information about the past actions, one can consider a policy ψ that does not, $\psi(x_0, x_i, \ldots, x_n) = \varphi(x_0, \varphi(x_0), x_1, \varphi(x_0, \varphi(x_0), x_1), \ldots, x_n)$. The policies φ and ψ always select the same actions.

A nonswitching policy is called *Markov* if the selected action is a function of the current state and the number of jumps (x_n, n). A (switching) policy is called *switching Markov* if the action selection depends only on the current state, the number of jumps, and the time passed since the last jump (x_n, n, t). If the above functions do not depend on the number of jumps n, then the corresponding policies are called *stationary* and *switching stationary*.

Similar to m-randomized stationary policies for SMDPs, we consider *m-switching stationary policies* for CTJMDPs. In fact, m-switching stationary policies are simpler objects that m-randomized stationary policies. For an m-switching stationary policy φ, the number of points (x, t) in which the function $\varphi(x, t)$ is discontinuous in t is not greater than m. A switching Markov policy φ is called *Markov switching of order m* if the number of points (x, n, t) where the function φ is discontinuous in t is not greater than m. If a switching Markov policy of order m is (nonswitching) stationary from jump n onward, it is called *an (m, n)-policy*. An (m, n)-policy φ is called *a strong (m, n)-policy* if $\varphi(X, \{0, 1, \ldots\}, \mathbb{R}_+) = \psi(X, \mathbb{R}_+)$ for some m-switching stationary policy ψ.

Randomized strategies for CTJMDPs. The definition of randomized strategies uses the convention that convex combinations of actions can be implemented as actions with the infinitesimal characteristics and reward rates being the convex combinations of the original characteristics. In addition, rewards during jumps should be defined for randomized actions. Though this can be done, it is easier to define randomized policies for the models with zero rewards/costs incurred during jumps. Fortunately, in view of Corollary 4.4, by changing reward rates, it is possible to reduce the model with rewards incurred at jumps to the model without rewards incurred at jumps. The answer to the question, whether randomized strategies can be indeed implemented, depends on a particular application. Fortunately, randomized strategies do not outperform nonrandomized strategies and therefore there is no need to consider them in applications when they are not natural; see §7.

A randomized strategy for a CTJMDP is a strategy for [CTJMDP]. Similarly, (nonswitching) randomized Markov policies are (nonswitching) Markov policies for [CTJMDP]. The same is true for randomized stationary policies.

Randomized stationary and randomized Markov policies for a CTJMPD are the same objects as, respectively, stationary and randomized Markov policies for the corresponding ESMDP. They are defined by the same transition probabilities but they have different impacts on the transition mechanisms in the corresponding models. Because they are the same objects, we shall keep for these policies the same definitions we use for them in SMDPs, but we shall indicate that we deal with randomized policies for a CTJMDP. Thus, we shall consider *m-randomized stationary policies* for a CTJMDP, *randomized Markov policies of order m* for a CTJMDP, *randomized* (m, n)*-policies* for a CTJMDP, and *randomized strong* (m, n)*-policies* for a CTJMDP.

For example, let two actions a and b be selected at state x. The jump intensities are $q(x, a) = 1$ and $q(x, b) = 2$. If a randomized stationary policy for the ESMDP selects these actions with the probabilities 0.5, the sojourn time is a mixture of two exponential distributions and its expectation is $3/4 = 0.5/q(x, a) + 0.5/q(x, b)$. If a randomized stationary policy for the CTJMDP selects these actions at each time with the probabilities 0.5, the sojourn time is exponential with the intensity $3/2 = 0.5 * q(x, a) + 0.5 * q(x, b)$, and its expectation is $2/3$. We remark that, as the former example illustrates, the word "exponential" in the abbreviation ESMDP means that sojourn times are exponential for nonrandomized policies. In general, for randomized policies in an ESMDP, these sojourn times are not exponential. They are mixtures of exponential distributions. Contrary to this, the sojourn times are exponential for randomized policies in CTJMDPs.

Figure 1 represents some of the major classes of strategies considered in this paper. Several important classes of policies are not included in Figure 1. For example, K-switching and K-randomized stationary policies (for the ESMDP and for the CTJMDP) are subclasses of switching stationary and appropriate randomized Markov policies; (K, n)-policies for CTJMDPs are subclasses of switching Markov policies; randomized (K, n)-policies are subclasses of randomized Markov policies in CTJMDPs, and (K, n)-policies for ESMDPs are subclasses of randomized Markov policies in ESMDPs. In addition, Markov policies are subsets of each of the following three classes: switching Markov policies, randomized Markov policies in CTJMDPs, and randomized Markov policies in ESMDPs.

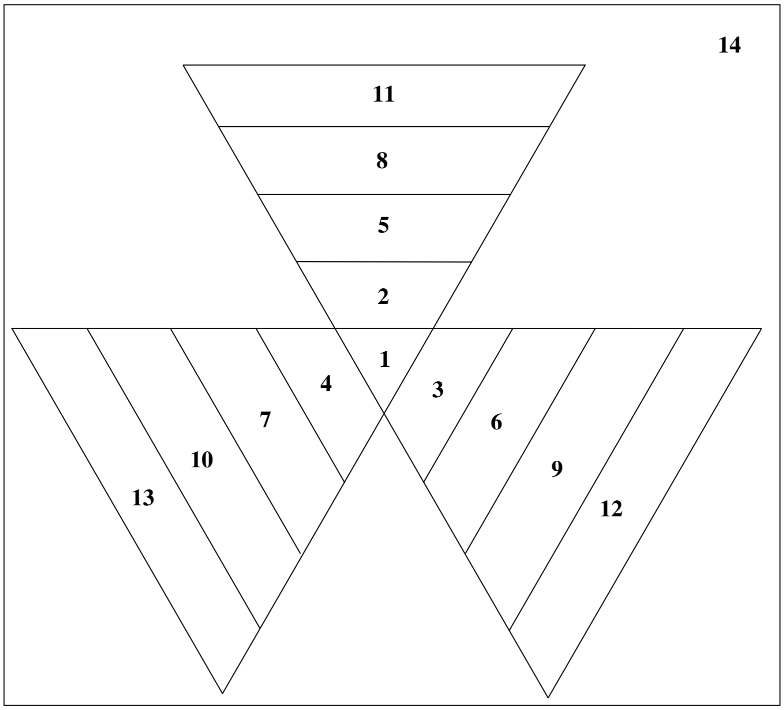
4. Reduction of CTJMDPs to SMDPs.

As explained above, a CTJMDP is in fact a YSMDP. YSMDPs are complicated objects. For example, if a CTJMDP contains finite state and action sets, the corresponding YSMDP contains function-valued action sets. For each state x, this action set is infinite if the original action set $D(x)$ is not a singleton. For a CTJMDP we define the Exponential SMDP (ESMDP), which is an SMDP with the state space X, action space A, sets of available actions $D(x)$, transition kernels

$$(4.1)\qquad Q(t, Y|x, a) = \begin{cases} 0, & \text{if } q(x, a) = 0, \\ q(Y|x, a)(1 - e^{-q(x,a)t})/q(x, a), & \text{otherwise,} \end{cases}$$

discount factor α, and reward functions

$$(4.2)\qquad r_k(x, a) = \frac{\bar{r}_k(x, a) + \int_X R_k^*(x, a, y)q(dy|x, a)}{\alpha + q(x, a)},$$



$1 \subset 2 \subset 5 \subset 8 \subset 11 \subset 14$
$1 \subset 3 \subset 6 \subset 9 \subset 12 \subset 14$
$1 \subset 4 \subset 7 \subset 10 \subset 13 \subset 14$

1. Stationary policies
2. Switching stationary policies
3. Randomized stationary policies in the ESMDP
4. Randomized stationary policies in the CTJMDP
5. Switching Markov policies
6. Randomized Markov policies in the ESMDP
7. Randomized Markov policies in the CTJMDP
8. All (nonrandomized) switching policies
9. All randomized policies in the ESMDP
10. All randomized nonswitching policies in the CTJMDP
11. All (nonrandomized) switching strategies
12. All randomized strategies in the ESMDP
13. All randomized nonswitching strategies in the CTJMDP
14. All (randomized switching) strategies

Figure 1. Some of the Classes of Strategies for CTJMPDs (class 1 is optimal for one-criterion problems; 2, 3, and 4 are optimal for constrained problems; 5, 6, and 7 are optimal for constrained finite-step problems).

where $k = 0, 1, \ldots, K$, $x \in X$, and $a \in D(x)$; see Appendix A for all necessary definitions and facts for SMDPs. We remark that an ESMDP is a much simpler object than a YSMDP. The action sets in a CTJMDP and in the corresponding ESMDP are the same and (2.1, 2.4) transform into (4.1, 4.2) when $u(t) = a = \text{const}$. The only difference between a CTJMDP and the corresponding ESMDP is in the way these processes can be controlled. For example, it is not possible to change actions between jumps in an ESMDP because it is an SMDP.

In this section, for an arbitrary strategy π in a CTJMDP we construct a policy σ in the corresponding ESMDP such that the occupation measures are equal for π and σ. Therefore, the expected total rewards for the corresponding criteria $k = 0, \ldots, K$ are equal for π and σ as well, because reward rates do not depend on occupancy times in states.

We use the same symbol W to denote both the expected total discounted rewards in a CTJMDP and in the corresponding ESMDP. Thus, the interpretation of $W(\mu, \pi)$ depends on whether π is a strategy for a CTJMDP or a policy for an ESMDP. This convention is consistent with the following two facts:

(i) A nonswitching policy for a CTJMDP is a nonrandomized policy for the corresponding ESMDP; in addition, the expected total rewards are equal in these two models for the corresponding criteria; and

(ii) Strategies for a CTJMDP as well as randomized strategies for the corresponding ESMDP are particular cases of randomized strategies for CTJMDPs defined in §7 (the former is obvious, and the latter follows from (8.6) with x_n replaced by ω_n).

For a CTJMDP, we define occupation measures

$$g^{\pi}_{\mu,n}(Y,B)=\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}I\{x_n\in Y, a_n\in B\},\quad n=0,1,\ldots,Y\in\mathscr{X},B\in\mathscr{A}. \tag{4.3}$$

Then for a bounded above (or below) measurable f

$$\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}f(x_n,a_n)=\int_X\int_A f(x,a)g^{\pi}_{\mu,n}(dx,da), \tag{4.4}$$

where (4.3) implies (4.4) for indicator functions and therefore for f. We also define occupation measures

$$G^{\pi}_{\mu,n}(Y,B,Z)=\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}I\{x_n\in Y, a_n\in B, x_{n+1}\in Z\}, \tag{4.5}$$

$$H^{\pi}_{\mu,n}(Y,B)=\mathbb{E}^{\pi}_{\mu}\textstyle\int_{t_n}^{t_{n+1}}e^{-\alpha t}I\{X_t\in Y, A_t\in B\}\,dt, \tag{4.6}$$

where $n=0,1,\ldots,$ $Y,Z\in\mathscr{X}$, $B\in\mathscr{A}$. Similar to (4.4),

$$\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}f(x_n,a_n,x_{n+1})=\int_X\int_A\int_X f(x,a,z)G^{\pi}_{\mu,n}(dx,da,dz), \tag{4.7}$$

$$\mathbb{E}^{\pi}_{\mu}\int_{t_n}^{t_{n+1}}e^{-\alpha t}f(X_t,A_t)\,dt=\int_X\int_A f(x,a)H^{\pi}_{\mu,n}(dx,da) \tag{4.8}$$

for bounded above (or below) functions f.

Lemma 4.1. *If $g^{\sigma}_{\mu,n}=g^{\pi}_{\mu,n}$ and $H^{\sigma}_{\mu,n}=H^{\pi}_{\mu,n}$, $n=0,1,\ldots,$ for two strategies σ and π then $W(\mu,\sigma)=W(\mu,\pi)$.*

Proof. We have from (4.7) and (4.8) that

$$W(x,\pi)=\sum_{n=0}^{\infty}\left[\int_X\int_A\int_X R^*(x,a,z)G^{\pi}_{\mu,n}(dx,da,dz)+\int_X\int_A \bar{r}(x,a)H^{\pi}_{\mu,n}(dx,da)\right]. \tag{4.9}$$

Therefore, if $G^{\sigma}_{\mu,n}=G^{\pi}_{\mu,n}$ and $H^{\sigma}_{\mu,n}=H^{\pi}_{\mu,n}$, $n=0,1,\ldots,$ then $W(\mu,\sigma)=W(\mu,\pi)$. We have

$$\begin{aligned}G^{\pi}_{\mu,n}(Y,B,Z)&=\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}I\{x_n\in Y,a_n\in B\}I\{x_{n+1}\in Z\}\\&=\mathbb{E}^{\pi}_{\mu}\mathbb{E}^{\pi}_{\mu}[e^{-\alpha t_{n+1}}I\{x_n\in Y,a_n\in B\}I\{x_{n+1}\in Z\}|a_n,x_n,t_{n+1}]\\&=\mathbb{E}^{\pi}_{\mu}e^{-\alpha t_{n+1}}I\{x_n\in Y,a_n\in B\}\frac{q(Z|x_n,a_n)}{q(x_n,a_n)}\\&=\int_Y\int_B\frac{q(Z|x,a)}{q(x,a)}g^{\pi}_{\mu,n}(dx,da),\end{aligned} \tag{4.10}$$

where $\frac{0}{0}$ is equal to 0 by definition and the last equality follows from (4.4). Thus, $g^{\sigma}_{\mu,n}=g^{\pi}_{\mu,n}$ implies $G^{\sigma}_{\mu,n}=G^{\pi}_{\mu,n}$. □

The following lemma describes the relationship between measures $g^{\pi}_{\mu,n}$ and $H^{\pi}_{\mu,n}$.

Lemma 4.2. *For every strategy π, initial distribution μ, and $n=0,1,\ldots,$*

$$g^{\pi}_{\mu,n}(Y,B)=\int_Y\int_B q(x,a)H^{\pi}_{\mu,n}(dx,da),\quad Y\in\mathscr{X},B\in\mathscr{A}.$$

To prove Lemma 4.2, we consider the following auxiliary result, which was proved in Feinberg (1994) for $\alpha = 0$.

LEMMA 4.3. *Let q be a nonnegative measurable function on A. Let ξ be a random variable with a distribution $P\{\xi \le t\} = 1 - e^{-\int_0^t q(a(s))ds}$, where $a(t)$ is a measurable mapping from $[0, \infty[$ to $(A, \mathcal{A})$. Let g and H be two measures on A defined by*

$$g(B) = Ee^{-\alpha\xi} I\{a(\xi) \in B\}, \quad B \in \mathcal{A},$$

$$H(B) = E\int_0^{\xi} e^{-\alpha t} I\{a(t) \in B\}\, dt, \quad B \in \mathcal{A}.$$

Then $g(B) = \int_B q(a)H(da)$ for all $B \in \mathcal{A}$.

PROOF. The proof is similar to the proof of Theorem 1 in Feinberg (1994). Let $f(t) = \int_0^t e^{-\alpha s} I\{a(s) \in B\} ds$. This function is nondecreasing and (almost everywhere) differentiable. Therefore, $Ef(\xi) = \int_0^\infty f'(t)P\{\xi > t\}dt$. We have that

$$H(B) = Ef(\xi) = \int_0^\infty e^{-\alpha t} I\{a(t) \in B\} e^{-\int_0^t q(a(s))\, ds}\, dt. \tag{4.11}$$

Straightforward computations imply that

$$g(B) = \int_0^\infty q(a(t)) e^{-\alpha t} I\{a(t) \in B\} e^{-\int_0^t q(a(s))\, ds}\, dt. \tag{4.12}$$

Formulas (4.11) and (4.12) imply the statement of the lemma. □

PROOF OF LEMMA 4.2. From (4.8) we have

$$\int_Y \int_B q(x, a) H^\pi_{\mu, n}(dx, da) = \mathbb{E}^\pi_\mu \int_{t_n}^{t_{n+1}} e^{-\alpha t} q(x_n, A_t) I\{x_n \in Y, A_t \in B\}\, dt. \tag{4.13}$$

This equality implies

$$\begin{aligned} &\int_Y \int_B q(x, a) H^\pi_{\mu, n}(dx\, da) \\ &\quad = \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n \in Y\} \mathbb{E}^\pi_\mu \Big[\int_0^{\xi_n} e^{-\alpha t} q(x_n, A_{t_n + t}) I\{A_{t_n+t} \in B\}\, dt | \omega_n\Big] \\ &\quad = \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n \in Y\} \mathbb{E}^\pi_\mu [e^{-\alpha \xi_n} I\{a_n \in B\} | \omega_n] = g(Y, B), \end{aligned} \tag{4.14}$$

where the second equality follows from Lemma 4.3 and the third equality follows from (4.3). □

COROLLARY 4.4. *For any initial distribution μ and for any strategy π,*

$$W(x, \pi) = \sum_{n=0}^\infty \int_X \int_A \Big[\int_X R^*(x, a, z) q(dz | x, a) + \bar{r}(x, a)\Big] H^\pi_{\mu, n}(dx, da). \tag{4.15}$$

Therefore, if $H^\sigma_{\mu, n} = H^\pi_{\mu, n}$, $n = 0, 1, \ldots,$ for two strategies σ and π then $W(\mu, \sigma) = W(\mu, \pi)$.

PROOF. (4.15) follows from (4.9), (4.10), and Lemma 4.2. □

For a strategy π in a CTJMDP and an initial distribution μ, we define for $n = 0, 1, \ldots,$ the occupation measures

$$\widetilde{M}^\pi_{\mu, n}(Y, B) = \int_Y \int_B H^\pi_{\mu, n}(dx, da)(\alpha + q(x, a)), \quad Y \in \mathcal{X}, B \in \mathcal{A}, \tag{4.16}$$

and

$$\widetilde{m}^\pi_{\mu, n}(Y) = \widetilde{M}^\pi_{\mu, n}(Y, A). \tag{4.17}$$

For the corresponding ESMDP, we define a randomized Markov policy σ such that

$$\sigma_n(B|x) = \frac{\widetilde{M}^{\pi}_{\mu,n}(dx, B)}{\tilde{m}^{\pi}_{\mu,n}(dx)}, \quad B \in \mathcal{A}, \; x \in X. \tag{4.18}$$

We remark that σ_n are defined by (4.18) $\tilde{m}^{\pi}_{\mu,n}$-a.e. We select regular transition probabilities σ_n such that $\sigma(D(x)|x) = 1$ for all $x \in X$. Formula (A.2.5) defines occupation measures $M^{\sigma}_{\mu,n}$ for the ESMDP.

Theorem 4.5. *Let an initial distribution μ be fixed. Let π be an arbitrary strategy for a CTJMDP. Then*

$$M^{\sigma}_{\mu,n} = \widetilde{M}^{\pi}_{\mu,n}, \quad n = 0, 1, \ldots, \tag{4.19}$$

for a randomized Markov policy σ defined by (4.18) *for the corresponding ESMDP and therefore $W(\mu, \sigma) = W(\mu, \pi)$.*

Proof. As follows from (4.15), (4.16), and (4.2),

$$W(\mu, \pi) = \sum_{n=0}^{\infty} \int_X \int_A r(x, a) \widetilde{M}^{\pi}_{\mu,n}(dx, da). \tag{4.20}$$

This formula, the similar formula for SMDPs (A.3.11), and (4.19) imply the second statement of the theorem.

Now we prove (4.19). We observe that for any $n = 0, 1, \ldots$

$$\begin{aligned} \int_Y \int_A H^{\pi}_{\mu,n}(dx, da)(\alpha + q(x,a)) &= \alpha H^{\pi}_{\mu,n}(Y, A) + g^{\pi}_{\mu,n}(Y, A) \\ &= \alpha \mathbb{E}^{\pi}_{\mu} \int_{t_n}^{t_{n+1}} e^{-\alpha t} I\{X_t \in Y\}\, dt + \mathbb{E}^{\pi}_{\mu} e^{-\alpha t_{n+1}} I\{x_n \in Y\} \\ &= \mathbb{E}^{\pi}_{\mu} e^{-\alpha t_n} I\{x_n \in Y\}, \end{aligned} \tag{4.21}$$

where the first equality follows from Lemma 4.2, the second equality follows from definitions (4.6) and (4.3) of H and g, and the third equality is straightforward.

First we show that $m^{\sigma}_{\mu,0} = \tilde{m}^{\pi}_{\mu,0}$. The definition of $M^{\sigma}_{\mu,n}$ implies that $m^{\sigma}_{\mu,0} = \mu$ for any policy σ in any SMDP. We also have that

$$\tilde{m}^{\pi}_{\mu,0}(Y) = \int_Y \int_A (\alpha + q(x, a)) H^{\pi}_{\mu,0}(dx, da) = \mu(Y), \tag{4.22}$$

where the first equality follows from definition (4.17) of $\tilde{m}$; the second one follows from (4.21) and $t_0 = 0$.

Assume that for some $n = 0, 1, \ldots$

$$\tilde{m}^{\pi}_{\mu,n} = m^{\sigma}_{\mu,n}. \tag{4.23}$$

Then

$$M^{\sigma}_{\mu,n}(Y, B) = \int_Y \int_B m^{\sigma}_{\mu,n}(dx) \frac{\widetilde{M}^{\pi}_{\mu,n}(dx, da)}{\tilde{m}^{\pi}_{\mu,n}(dx)} = \widetilde{M}^{\pi}_{\mu,n}(Y, B), \tag{4.24}$$

where the first equality follows from (A.2.9) and (4.18); the second equality follows from (4.23).

Now let (4.19) hold for some $n = 0, 1, \ldots$. Then (A.2.7) and (4.1) imply that

$$\begin{aligned} m^{\sigma}_{\mu,n+1}(Y) &= \int_X \int_A q(Y|x, a) \int_0^{\infty} e^{-(\alpha + q(x,a))t} dt \; M^{\sigma}_{\mu,n}(dx, da) \\ &= \int_X \int_A \frac{q(Y|x, a)}{\alpha + q(x, a)} M^{\sigma}_{\mu,n}(dx, da) = \int_X \int_A q(Y|x, a) H^{\pi}_{\mu,n}(dx, da), \end{aligned} \tag{4.25}$$

where the last equality follows from (4.19) and (4.16).

We have that

$$\tilde{m}^{\pi}_{\mu,n+1}(Y) = \int_Y\int_A H^{\pi}_{\mu,n+1}(dx,da)(\alpha+q(x,a)) \\ = \mathbb{E}^{\pi}_{\mu} e^{-\alpha t_{n+1}} I\{x_{n+1}\in Y\} = G^{\pi}_{\mu,n}(X,A,Y), \tag{4.26}$$

where the first equality is (4.16, 4.17), the second equality is (4.21), and the last is (4.5). We also have from (4.10) and Lemma 4.2 that

$$G^{\pi}_{\mu,n}(X,A,Y) = \int_X\int_A \frac{q(Y|x_n,a_n)}{q(x_n,a_n)} g^{\pi}_{\mu,n}(dx,da) = \int_X\int_A q(Y|x,a)H^{\pi}_{\mu,n}(dx,da). \tag{4.27}$$

Therefore, from (4.25–4.27) we have $m^{\sigma}_{\mu,n+1} = \tilde{m}^{\pi}_{\mu,n+1}$. □

Corollary 4.6. *Consider a strategy π in a CTJMDP.* (i) *For a one-criterion problem, if for any initial state distribution μ a randomized Markov policy σ, defined by* (4.18), *is optimal for the corresponding ESMDP with the initial state distribution μ, then the strategy π is optimal for the CTJMDP when μ is the initial state distribution.* (ii) *For a constrained problem with multiple criteria, if a randomized Markov policy σ defined by* (4.18) *is optimal in the corresponding ESMDP, then the strategy π is optimal in the CTJMDP.*

We remark that if π is stationary or Markov then $\sigma=\pi$ satisfies (4.19) with any μ. In particular, Corollary 4.6 implies the existence of stationary optimal policies in one-criterion problems; see Theorem 6.1(i) below.

5. Equivalent switching policies for CTJMDPs.

As established in the previous section, for each strategy in a CTJMDP there is a policy in the corresponding ESMDP with the same performance vector. We recall that an ESMDP is a particular case of an SMDP when, for nonrandomized strategies, all sojourn times are exponential and transition probabilities do not depend on sojourn times. In view of Theorem 4.5, if for an optimal policy σ for an ESMDP we can construct a policy π in the corresponding CTJMDP such that (4.5) holds (it should hold for all μ for a one-criterion problem), the policy π is optimal for the CTJMDP.

For one-criterion problems, Theorem A.8 and Corollary A.11 imply the existence of optimal stationary and Markov policies. In this case the construction is trivial: we set $\pi=\sigma$. In this section we provide such constructions for constrained problems. We do not know how to construct π for an arbitrary σ or for an arbitrary optimal σ. Fortunately, according to Theorems A.8–A.10, A.12, and Corollary A.11, among optimal policies for SMDPs there exist policies satisfying certain additional properties; e.g., they could be K-randomized stationary. We formulate a weaker form of these properties in Condition 5.1 and provide the construction of π for policies σ that satisfy this condition. In particular, K-randomized stationary policies, strong (K,n)-randomized policies, and Markov policies of order K satisfy Condition 5.1. For a randomized Markov policy σ in an SMDP, we consider the following condition.

Condition 5.1. (i) *for each $x\in X$ and for each $n=0,1,\ldots$ there is a countable or finite subset $A^{\sigma}(x,n)$ of $D(x)$ such that $\sigma_n(A^{\sigma}(x,n)|x)=1$,*

(ii) *there exists a finite or countable set X^{σ} such that for each $n=0,1,\ldots$ and for each $x\in X\backslash X^{\sigma}$ the distribution $\sigma_n(\cdot|x)$ is concentrated at one point.*

For an initial distribution μ and for a randomized Markov policy σ, defined for the corresponding ESMDP and satisfying Condition 5.1, we construct in this section a switching Markov policy π for the CTJMDP such that $W_k(\mu,\pi)=W_k(\mu,\sigma)$, $k=0,\ldots,K$. Therefore, if σ is an optimal randomized Markov policy for the corresponding ESMDP then π is an optimal switching Markov policy for the original CTJMDP. In addition, this construction implies that policies σ and π have the same orders and indices and, if σ is randomized stationary, then π is switching stationary.

We consider a CTJMDP and the corresponding ESMDP. Let σ be a randomized Markov policy defined for this ESMDP and let σ satisfy Condition 5.1. Because we can add decisions that have zero probabilities to $A^\sigma(x,n)$, we can write, without loss of generality, that $A^\sigma(x,n) = \{a(x,n,1), a(x,n,2), \ldots\}$, $i = 1,2,\ldots$, $x \in X$, and $n = 0,1,\ldots$.

We set $S_0(x,n) = 0$,

$$(5.1)\qquad s_i(x,n) = -(\alpha + q(x, a(x,n,i)))^{-1} \ln\left(1 - \frac{\sigma_n(a(x,n,i)|x)}{\sum_{k=i}^{\infty} \sigma_n(a(x,n,k)|x)}\right),$$

where $\frac{0}{0} = 0$ and $\ln 0 = -\infty$. Let $S_i(x,n) = S_{i-1}(x,n) + s_i(x,n)$, $i = 1,2,\ldots$. We define a switching Markov policy π for the CTJMDP by

$$(5.2)\quad \pi(x,n,t) = a(x,n,i) \quad \text{for } x \in X, n = 0,1,\ldots, S_{i-1}(x,n) \le t < S_i(x,n),\ i = 1,2,\ldots.$$

Condition 5.1 implies that $\pi(x,n,t) = \varphi_n(x)$ for $x \in X \backslash X^\sigma$ for a measurable function φ_n. Because the set X^σ is countable and the function $\pi(x,n,t)$ is measurable in t for $x \in X^\sigma$, the function $\pi(x,n,t)$ is measurable in (x,t). We remark that if σ is a randomized stationary policy then $a(x,n,i)$, $s_i(x,n)$, $S_i(x,n)$, and $\pi(x,n,t)$ do not depend on n, and therefore π is a switching stationary policy.

Theorem 5.2. *For a CTJMDP, consider the corresponding ESMDP. Let σ be a randomized Markov policy in this ESMDP such that σ satisfies Condition* 5.1. *Consider the switching Markov policy π in the CTJMDP defined by* (5.2). *Then equality* (4.18) *holds for all $n = 0,1,\ldots$.*

Proof. We fix some $n = 0,1,\ldots$. Let Y be a measurable subset of $X \backslash X^\sigma$. From (4.21) we have

$$(5.3)\qquad \tilde{m}^\pi_{\mu,n}(Y) = \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n \in Y\}.$$

Because $A_t = \varphi_n(x_n)$ when $t_n \le t < t_{n+1}$, we have, similar to (4.21),

$$(5.4)\qquad \widetilde{M}^\pi_{\mu,n}(Y,B) = \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n \in Y\} I\{\varphi_n(x_n) \in B\}.$$

Then for any bounded above measurable function f,

$$(5.5)\qquad \mathbb{E}^\pi_\mu e^{-\alpha t_n} f(x_n) = \int_X f(x) \tilde{m}^\pi_{\mu,n}(dx).$$

From (5.4) and (5.5) we have

$$(5.6)\qquad \widetilde{M}^\pi_{\mu,n}(Y,B) = \int_Y I\{\varphi_n(x_n) \in B\} \tilde{m}^\pi_{\mu,n}(dx).$$

Let $x \in X^\sigma$ and $\tilde{m}^\pi_{\mu,n}(x) > 0$. If (4.18) holds for any x satisfying these two conditions, then the proof is completed.

The definition of σ implies that for all $i = 1,2,\ldots$

$$\begin{aligned}(5.7)\qquad g^\pi_{\mu,n}(x, a(x,n,i)) &= \mathbb{E}^\pi_\mu e^{-\alpha t_{n+1}} I\{x_n = x, S_{i-1} \le \xi_n < S_i\} \\ &= \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n = x\} \mathbb{E}^\pi_\mu [e^{-\alpha \xi_n} I\{S_{i-1} \le \xi_n < S_i\} | x_n = x].\end{aligned}$$

We observe that for $t \in [S_{i-1}, S_i[$

$$(5.8)\qquad \mathbb{P}^\pi_\mu\{\xi_n > t | x_n = x\} = e^{-q(x,a(x,n,i))(t - S_{i-1})} \prod_{k=1}^{i-1} e^{-q(x,a(x,n,k))S_k},$$

and

$$\mathbb{E}^\pi_\mu[e^{-\alpha\xi_n} I\{S_{i-1} \le \xi_n < S_i\} | x_n = x] = \int_{S_{i-1}}^{S_i} e^{-\alpha t} dF(t), \tag{5.9}$$

where $F(t) = 1 - \mathbb{P}^\pi_\mu\{\xi_n > t | x_n = x\}$. Straightforward calculations imply that

$$\mathbb{E}^\pi_\mu[e^{-\alpha\xi_n} I\{S_{i-1} \le \xi_n < S_i\} | x_n = x] = \frac{q(x, a(x, n, i))}{\alpha + q(x, a(x, n, i))} \sigma(a(x, n, i) | x, n). \tag{5.10}$$

From Lemma 4.2, (4.16), (5.7), and (5.10) we have that

$$\widetilde{M}^\pi_{\mu,\,n}(x, a(x, n, i)) = \sigma_n(a(x, n, i) | x) \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n = x\}, \tag{5.11}$$

and

$$\widetilde{m}^\pi_{\mu,\,n}(x) = \mathbb{E}^\pi_\mu e^{-\alpha t_n} I\{x_n = x\}. \tag{5.12}$$

Formulas (5.11) and (5.12) imply (4.18). □

Corollary 5.3. *Consider a CTJMDP. Let μ be an initial measure and σ be a randomized Markov in the corresponding ESMDP. If σ satisfies Condition* 5.1, *then*

$$\widetilde{M}^\pi_{\mu,\,n} = M^\sigma_{\mu,\,n}, \quad n = 0, 1, \dots,$$

for the switching Markov policy π defined by (5.2), *and therefore $W_k(\mu, \pi) = W_k(\mu, \sigma)$, $k = 0, \dots, K$.*

Proof. The corollary follows from Theorems 5.2 and 4.5. □

6. Optimization of CTJMDPs. In this section we describe the structure of optimal strategies for CTJMDPs and algorithms for their computation. We assume in this section that the state space X is finite or countable. We prove the existence of optimal strategies for CTJMDPs that have a structure similar to the structure of optimal policies for SMDPs described in Appendix A. The only significant difference between optimal strategies in CTJMDPs and SMDPs is that switching Markov (or stationary) policies play the same role in CTJMDPs as randomized Markov (or stationary) policies in SMDPs.

The results of this section follow from the results of §§4 and 5 and Appendix A.3. In §4 we proved that for each strategy in a CTJMDP there exists a policy in the corresponding ESMDP such that the corresponding expected total discounted rewards are equal. Appendix A.3 describes the structure of optimal strategies and algorithms for SMDPs. Because an ESMDP is a particular case of an SMDP, Appendix A.3 also describes optimal strategies for ESMDPs.

According to the results presented in Appendix A.3, optimal policies for SMDPs can be selected in a way that they are either randomized stationary or randomized Markov. For a problem with a single criterion, these policies are nonrandomized. For a problem with a finite number of constraints, an optimal randomized stationary policy uses randomized decisions only on a finite subset of states, and an optimal randomized Markov policy does it only on a finite subset of pairs (x, n), where x is a state and n is a jump number. In addition, in each state such an optimal policies uses a finite numbers of actions. The total number of such states or state-jump pairs is limited by the number of constraints. Thus, these optimal policies satisfy Condition 5.1. Theorem 5.2 relates an optimal randomized Markov (or stationary) policy in an ESMDP to a switching Markov (stationary) policy in the corresponding CTJMDP with the same performance vector. Therefore, this policy is optimal for the CTJMDP. In addition, the number of switching points is limited by the number of constraints.

For an ESMDP, formulas (4.1) and (4.2) express the transition kernel Q and expected cumulative discounted rewards r_k. We also have from (A.2.4) and (4.1) that

$$\beta(Y|x,a) = q(Y|x,a)/(\alpha + q(Y|x,a)). \tag{6.1}$$

In particular, $\beta(x,a) = q(x,a)/(\alpha+q(x,a))$, $\beta(y|x,a) = q(y|x,a)/(\alpha+q(x,a))$.

In this section, we consider a CTJMDP that satisfies the compactness and continuity assumptions similar to Assumption A.7 from Appendix A.3. These assumptions are formulated in Theorem 6.1 and they imply that the corresponding ESMDP satisfies Assumption A.7. Corollaries 4.6 and 5.3 correspond the appropriate strategies in CTJMDPs and ESMDPs. Theorems A.8 and A.9 and Corollary A.11 describe the structure of optimal policies in SMDPs and therefore in ESMDPs. These statements imply the following.

Theorem 6.1. *Let a CTJMDP with a finite or countable state space X satisfy the following compactness and continuity assumptions*: (1) *the sets $D(x)$, $x \in X$, are compact,* (2) *for all $x, y \in X$ the functions $q(y|x,a)$ and $q(x,a)$ are continuous in $a \in D(x)$, and* (3) *the functions $\bar{r}_k(x,a)$ and $R^*_k(x,a,y)$ are upper semi-continuous in $a \in D(x)$ and bounded above. Then*

(i) *there exists an optimal stationary policy for a one-criterion CTJMDP and there is an optimal Markov policy for a one-criterion nonhomogeneous* (*in particular, finite-step*) *CTJMDP*;

(ii) *for an infinite horizon problem with multiple criteria and K constraints, there exists a K-switching stationary optimal policy and for some $n = 0, 1, \ldots$ there exists an optimal strong (K, n)-policy*;

(iii) *for a nonhomogeneous problem* (*in particular, for a finite-step problem*) *with multiple criteria and K constraints, there exists an optimal switching Markov policy of order K* (*in this case x should be replaced with (x,n), $n = 0, 1, \ldots,$ in continuity conditions* (1–3) *of this theorem*).

Now we discuss how to compute optimal policies. Let X and A be finite. Theorem A.10 describes the computation of an optimal K randomized stationary policy for an infinite horizon SMDP with K constraints. Theorem A.12 describes the computation of an optimal randomized Markov policy of order K for a finite-step SMDP with K constraints. For a CTJMDP we can consider the corresponding ESMDP. For this ESMDP, Theorems A.10 and A.12 describe the computation of optimal policies. Formula (6.1) provides explicit expressions for functions β that participate in the LP (A.3.25–A.3.28) and in the LP (A.3.30–A.3.34). Theorem 5.2 describes the transformation of a randomized policy for a ESMDP into an optimal switching Markov policy. However, if we have a CTJMDP, we do not need to compute an optimal policy for the ESMDP to get an optimal policy for the CTJMDP.

For a CTJMDP with finite state and action spaces, we consider the LP (A.3.25–A.3.28) for $N = \infty$ and the LP (A.3.30–A.3.34) for $N < \infty$. In view of Theorems A.10 and A.12, a CTJMDP is feasible if and only if the corresponding LP is feasible and, if this LP is feasible, it has an optimal solution. Let u be the optimal basic solution. If $N = \infty$ then $u = \{u_{x,a}\}$. If $N < \infty$ then $u = \{u_{x,n,a}\}$. In both cases u defines an optimal switching strategy. Let $N < \infty$. We provide explicit formulas (6.2) and (6.3) for an optimal solution that follow from (A.3.35) and (5.2). We consider sets $D_u(x,n) = \{a \in D(x,n) \,|\, u_{x,n,a} > 0,$ $x \in X, n = 0, 1, \ldots\}$. If $D_u(x,n) = \varnothing$ we let $\pi(x,n,t) = a$ for all $t \geq 0$, where a is an arbitrary element of $D(x)$. If $D_u(x,n) \neq \varnothing$ we order elements $D_u(x,n)$ in an arbitrary way, $D_u(x,n) = \{a(1), \ldots, a(j(x,n))\}$ and consider π defined by (5.2). In particular, (5.1) can be rewritten as

$$s_i(x,n) = -(\alpha + q(x,a(i)))^{-1} \ln\left(1 - u_{x,n,a(i)} \Big/ \sum_{k=i}^{j(x,n)} u_{x,n,a(k)}\right), \tag{6.2}$$

$$i = 1, \ldots, j(x,n) - 1,$$

and $S_{j(x,n)}(x,n) = s_{j(x,n)}(x,n) = \infty$. Thus, (5.2) transforms into

$$\pi(x,n,t) = a(x,n,i) \quad \text{for } S_{t-1}(x,n) \le t < S_i(x,n), \quad i = 1, \ldots, j(x,n), \tag{6.3}$$

where π is an optimal switching Markov policy of order K.

If $N = \infty$, formulas (6.2) and (6.3) define an optimal K-switching stationary policy. In this case, u is a solution of the LP (A.3.25–A.3.28) and all objects defined in the previous paragraph, including variables in formulas (6.2) and (6.3), do not depend on n.

7. Randomized strategies in CTJMDPs. In this section we define randomized strategies for CTJMDPs and prove that the optimal switching stationary and Markov policies, the existence and structure of which are described in Theorem 6.1, are optimal within the broader class of randomized strategies (Corollary 7.3).

For discrete time MDPs, randomized policies can select actions randomly at each epoch. For example, if a system consists of one state and two actions 0 and 1, a decision maker can select actions 0 and 1 at each epoch $t = 0, 1, \ldots$ independently with probabilities 0.5.

For continuous time $t \in \mathbb{R}_+$, this definition is not applicable. For example, again let the system consist of one state and two actions 0 and 1. At each epoch $t \in \mathbb{R}_+$ we want to select 0 or 1 independently and with probabilities 0.5 and 0.5. If this can be done, consider a stochastic process a_t with independent and identically distributed values and with $P(a_t = 0) = P(a_t = 1) = 0.5$, $t \ge 0$. This process is not measurable (Kallianpur 1980, Example 1.2.5). However, to compute appropriate rewards and transition probabilities, we need to integrate functions of a_t. This example shows that actions cannot be selected randomly and independently at each epoch $t \in \mathbb{R}_+$.

To avoid these obstacles, another definition of randomized strategies has been developed for CTJMDPs in the literature. We observe that in discrete time, if at a state x we select an action a with probability q and an action b with probability $1-q$, the transition probabilities and one-step expected rewards will be the same as if we select an action c that defines transition probabilities $p(\cdot|x,c) = qp(\cdot|x,a) + (1-q)p(\cdot|x,b)$ and rewards $r(x,c) = qr(x,a) + (1-q)r(x,b)$. In other words, a randomized action is an action with the transition probabilities, and one-step rewards equal the expectations of the corresponding transition probabilities and one-step rewards. These properties can be considered as the definition of randomized actions in discrete time. This definition can be carried over to CTJMDPs. According to this definition of randomized strategies, if at epoch $t \in \mathbb{R}_+$ a probability measure on the action set is selected in a CTJMDP, it means that the decision maker selects an action with the transition intensities, reward intensities, and rewards equal to expectations of the corresponding values with respect to this measure. Hordijk and van der Duyn Schouten (1979) introduced this approach for a particular class of strategies, and Kitaev (1984) expanded it to all strategies.

Now we give formal definitions. For a Borel set B we denote by $\mathcal{P}(B)$ the set of probability measures on B. If one considers the weak topology on $\mathcal{P}(B)$, then this set is Borel as well (Parthasarathy 1967, Chapter II, Theorems 6.2 and 6.5). If B_1 is a measurable subset of B, then $p \in \mathcal{P}(B_1)$ can be interpreted as an element of $\mathcal{P}(B)$ with $p(B \backslash B_1) = 0$. Therefore, $\mathcal{P}(B_1) \subseteq \mathcal{P}(B)$. If B is compact then $\mathcal{P}(B)$ is compact in the weak topology (Parthasarathy 1967, Chapter II, Theorem 6.4). For convenience, if a symbol denotes a generic element of a Borel set, we shall use the bold version of this symbol to denote a probability distribution on this set. For example, we use the notation $b \in B$ and $\mathbf{b} \in \mathcal{P}(B)$. For a measurable function f on B we define

$$\mathbf{f}(\mathbf{b}) = \int_B f(b)\mathbf{b}(db), \quad \mathbf{b} \in \mathcal{P}(B). \tag{7.1}$$

This function is measurable (Bertsekas and Shreve 1978, Corollary 7.29.1).

Consider a CTJMDP defined in §2 by objects $\{X, A, D, q, \alpha, K, \bar r_k, R^*_k\}$. Corollary 4.4 implies that the objective function remains the same if we set R^*_k to be identically equal to zero and add $\int_X R^*(x,a,z)q(dz|x,a)$ to the reward rates $\bar r_k$. Therefore, without loss of generality we assume that $R^*_k \equiv 0$ for all k. Thus, we consider a CTJMDP defined by the objects $\{X, A, D, q, \alpha, K, \bar r_k\}$. We define a CTJMDP defined by the objects $\{X, \mathbf{A}, \mathbf{D}, \mathbf{q}, \alpha, K, \bar{\mathbf{r}}_k\}$, where $\mathbf{A} = \mathcal{P}(A)$, $\mathbf{D}(x) = \mathcal{P}(D(x))$,

$$\mathbf{q}(Y|x,\mathbf{a}) = \int_{D(x)} q(Y|x,a)\mathbf{a}(da),$$

$\mathbf{q}(x,\mathbf{a}) = \mathbf{q}(X|x,\mathbf{a})$, and

$$\bar{\mathbf{r}}_k(x,\mathbf{a}) = \int_{D(x)} \bar r_k(x,a)\mathbf{a}(da), \quad k = 0, \dots, K. \tag{7.2}$$

We call the new CTJMDP a *convex hull* of the original CTJMDP. We denote this new CTJMDP by [CTJMDP]. We remark that if $\mu \in \mathcal{P}(\mathcal{P}(E))$ for a Borel set E then we can consider a measure $\nu \in \mathcal{P}(E)$ such that $\nu(B) = \int_{\mathcal{P}(E)} p(E)\mu(dp)$. In view of this remark, it is easy to observe that, if we consider the convex hull of the convex hull of a CTJMDP, it coincides with the convex hull of a CTJMDP. In other words, [[CTJMDP]] = [CTJMDP].

Any strategy in a [CTJMDP] is called *a randomized strategy* in the original CTJMDP. In other words, a randomized strategy π is defined by a measurable mapping π from $F \times \mathbb{R}_+$ to $\mathbf{A}$ such that $\pi(D(x_n)|\omega_n, t) = 1$ for all $\omega_n = x_0, \xi_0, x_1, \xi_1, \dots, x_n \in F_n$, $n = 0, 1, \dots$. This function can be viewed as a predictable mapping from $\Omega \times \mathbb{R}_+$ to $\mathbf{A}$; Jacod (1975, Lemma 3.3). Similar to (2.2), a randomized strategy π and an initial distribution μ define a predictable random measure ν^π on $\Omega \times \mathbb{R}_+$ by

$$\nu^\pi(\omega, ds, dx) = ds \int_{D(X_s)} q(dx|X_s, a)\pi(da|\omega, s), \tag{7.3}$$

and $\nu^\pi(\omega, 0, dx) = \mu(dx)$, where $\pi(da|\omega, s) = \pi(da|\omega_n, s - t_n)I\{t_n \le s < t_{n+1}\} + \sigma(da)I\{s \ge t_\infty\}$ and σ is an arbitrary probability distribution on A. Then $\mathbb{P}^\pi_x$ is a unique probability measure on $(\Omega, \mathcal{F})$ such that ν^π is a predictable projection of the random measure defined by $\mathbb{P}^\pi_\mu$.

We say that π is a randomized stationary (Markov) policy in a CTJMDP if π is a stationary (Markov) policy in the [CTJMDP]. In other words, a strategy π is *a randomized Markov policy* if $\pi(da|\omega_n, t) = \pi(da|x_n, n)$, $n = 0, \dots$. A strategy π is *a randomized stationary policy* if $\pi(da|\omega_n, t) = \pi(da|x_n)$.

Because CTJMDPs and ESMDPs are defined by the same objects, there is a natural one-to-one correspondence between CTJMDPs and ESMDPs. We denote by [ESMDP] the ESMDP that corresponds to the [CTJMDP].

Let π be a randomized Markov policy in the [ESMDP]. We define a randomized Markov policy σ in the ESMDP by

$$\sigma_n(da|x) = \int_{\mathbf{D}(x)} \frac{\mathbf{a}(da)(\alpha + q(x,a))}{\alpha + \mathbf{q}(x,\mathbf{a})} \pi_n(d\mathbf{a}|x). \tag{7.4}$$

We consider the measures $m^\sigma_{\mu,n}$ and $m^\pi_{\mu,n}$ on X, measure $M^\sigma_{\mu,n}$ on $X \times A$, and measure $M^\pi_{\mu,n}$ on $X \times \mathcal{P}(A)$ defined for the [ESMDP] and ESMDP, respectively, in (A.2.6) and (A.2.5).

Theorem 7.1. *For a randomized Markov policy π in the [ESMDP] consider a randomized Markov policy σ in the ESMDP defined by* (7.4). *Then*

$$m^\pi_{\mu,n} = m^\sigma_{\mu,n} \tag{7.5}$$

for any initial distribution μ and for any $n = 0, 1, \dots$. Furthermore,

$$W_k(\mu, \pi) = W_k(\mu, \sigma), \quad k = 0, \dots, K. \tag{7.6}$$

Proof. First, $m^{\pi}_{\mu,0} = m^{\sigma}_{\mu,0} = \mu$. Let (7.5) hold for some n. The definition of measure M implies that

$$M^{\sigma}_{\mu,n}(dx, da) = \sigma_n(da|x)m^{\sigma}_{\mu,n}(dx), \tag{7.7}$$

and

$$M^{\pi}_{\mu,n}(dx, d\mathbf{a}) = \pi_n(d\mathbf{a}|x)m^{\pi}_{\mu,n}(dx). \tag{7.8}$$

We have that for any $Y \in \mathscr{X}$

$$\begin{aligned} m^{\sigma}_{\mu,(n+1)}(Y) &= \int_X \int_A \frac{q(Y|x,a)}{\alpha + q(x,a)} \sigma_n(da|x) m^{\sigma}_{\mu,n}(dx) \\ &= \int_X \int_A \frac{q(Y|x,a)}{\alpha + q(x,a)} \int_{\mathscr{P}(A)} \frac{\mathbf{a}(da)(\alpha + q(x,a))}{\alpha + \mathbf{q}(x,\mathbf{a})} \pi_n(d\mathbf{a}) m^{\sigma}_{\mu,n}(dx) \\ &= \int_X \int_{\mathscr{P}(A)} (\alpha + \mathbf{q}(x,\mathbf{a}))^{-1} \int_A q(Y|x,a)\mathbf{a}(da)\pi_n(d\mathbf{a}) m^{\sigma}_{\mu,n}(dx) \\ &= \int_X \int_{\mathscr{P}(A)} \frac{\mathbf{q}(Y|x,\mathbf{a})}{\alpha + \mathbf{q}(x,\mathbf{a})} \pi_n(d\mathbf{a}|x) m^{\pi}_{\mu,n}(dx) = m^{\pi}_{\mu,(n+1)}(Y), \end{aligned} \tag{7.9}$$

where the first equality follows from (A.2.7) applied to the ESMDP, (6.1), and (7.7); the second equality follows from (7.4); the third equality follows from the change of integration; in the forth equality we use (7.1), the definition of $\mathbf{q}$, and the induction assumption; the last equality follows from (A.2.7) applied to the [ESMDP], (6.1), and (7.8). Thus (7.5) is proved.

We fix some $k = 0, \ldots, K$. To simplify the notation, we omit index k everywhere in the remainder of this proof. For the ESMDP, if an action a is selected in a state x, the expected total rewards between jumps are described in formula (4.2). Similarly, the expected total rewards for the [ESMDP] are

$$\mathbf{r}(x,\mathbf{a}) = \frac{\bar{\mathbf{r}}(x,\mathbf{a})}{\alpha + \mathbf{q}(x,\mathbf{a})}. \tag{7.10}$$

Because π is a policy in the [ESMDP], formula (A.3.11) implies that

$$W(x,\pi) = \sum_{n=0}^{\infty} \int_X \int_A \mathbf{r}(x,\mathbf{a}) M^{\pi}_{\mu,n}(dx, d\mathbf{a}). \tag{7.11}$$

We have that

$$\begin{aligned} \int_X \int_A r(x,a) M^{\sigma}_{\mu,n}(dx,da) &= \int_X \int_A r(x,a)\sigma_n(da|x) m^{\pi}_{\mu,n}(dx) \\ &= \int_X \int_A \frac{\bar{r}(x,a)}{\alpha + q(x,a)} \int_{\mathscr{P}(A)} \frac{\mathbf{a}(da)(\alpha + q(x,a))}{\alpha + \mathbf{q}(x,\mathbf{a})} \\ &\qquad \cdot \pi_n(d\mathbf{a}|x) m^{\pi}_{\mu,n}(dx) \\ &= \int_X \int_{\mathscr{P}(A)} (\alpha + \mathbf{q}(x,\mathbf{a}))^{-1} \int_A \bar{r}(x,a)\mathbf{a}(da)\pi_n(d\mathbf{a}|x) m^{\pi}_{\mu,n}(dx) \\ &= \int_X \int_{\mathscr{P}(A)} \mathbf{r}(x,\mathbf{a})\pi_n(d\mathbf{a}|x) m^{\pi}_{\mu,n}(dx) \\ &= \int_X \int_{\mathscr{P}(A)} \mathbf{r}(x,\mathbf{a}) M^{\pi}_{\mu,n}(dx, d\mathbf{a}), \end{aligned} \tag{7.12}$$

where the first equality follows from (7.7) and (7.5); the second equality follows from (4.2) and (7.4); the third equality follows from the change of integration; the fourth equality follows from the definition of $\bar{\mathbf{r}}$, (7.2); and the last equality follows from (7.8). Formula (A.3.11) for $W(\mu,\pi)$ applied to the policy σ and formulas (7.11), (7.12) implies the second statement of the theorem. □

We consider single and multiple objective problems for a CTJMDP with the set of strategies expanded to the set of all randomized strategies. For a multiple objective problem, the constants $C_1, \ldots, C_K$ in constraints remain the same as in the original CTJMDP but the set of feasible strategies is expanded.

Corollary 7.2. *Consider a countable state CTJMDP satisfying conditions* (1)–(3) *from Theorem* 6.1. *Let μ be an initial state distribution on X. For any feasible randomized strategy π there exists a feasible switching stationary policy σ such that $W_0(\mu, \sigma) \geq W_0(\mu, \pi)$.*

Proof. A randomized strategy π in the CTJMDP is a (nonrandomized) strategy in the [CTJMDP]. Theorem 4.5 implies that there exists a randomized Markov policy π^1 in the [ESMDP] such that $W_k(\mu, \pi^1) = W_k(\mu, \pi)$, $k = 0, \ldots, K$. Theorem 7.1 implies that there is a randomized Markov policy π^2 in the ESMDP such that $W_k(\mu, \pi^2) = W_k(\mu, \pi^1)$, $k = 0, \ldots, K$. Theorem A.9 implies the existence of a K-randomized stationary policy π^3 such that $W_0(\mu, \pi^3) \geq W_0(\mu, \pi^2)$ and $W_k(\mu, \pi^3) \geq C_k$, $k = 1, \ldots, K$. Theorem 5.2 constructs in the CTJMDP a feasible switching stationary policy σ with the required properties. □

We remark that Corollary 7.2 is also applicable to one-criteria problems. In this case, any policy is feasible. Thus, Corollary 7.2 states that for any randomized strategy in a CTJMDP there exists a better or equal regular strategy. Corollary 7.2 and Theorem 6.1 imply the following result.

Corollary 7.3. *Optimal policies, whose existence is stated in Theorem* 6.1, *are also optimal within the broader class of randomized strategies.*

8. More on randomized policies in CTJMDPs.

In this section we prove the optimality of randomized stationary and Markov policies for a CTJMDP with multiple criteria and constraints (Theorem 8.3). We also prove that optimal policies for an ESMDP with multiple criteria and constraints, the existence and structure of which are stated in Theorem A.9 and Corollary A.11, can be implemented as randomized policies in the corresponding CTJMDP. Therefore, these policies are also optimal for the CTJMDP (Theorem 8.4).

We observe that randomized Markov (stationary) policies for a CTJMDP and for the corresponding ESMDP are defined by the same objects. These objects are transition probabilities from $X \times \{0, 1, \ldots\}$ to A (from X to A) such that $\pi_n(D(x) \mid x) = 1$, $n = 0, 1, \ldots$ $(\pi(D(x) \mid x) = 1)$. The major difference between a randomized Markov policy for a CTJMDP and the corresponding randomized Markov policy for the ESMDP is that they define different transition mechanisms described above (e.g., if a randomized Markov policy for an ESMDP selects an action a with probability p and action b with probability $(1 - p)$, then the corresponding sojourn time ξ has the distribution defined by

$$P\{\xi > t\} = pe^{-q(x,a)t} + (1 - p)e^{-q(x,b)t}.$$

If a randomized Markov policy for a CTJMDP makes the similar selection then the sojourn time ξ is defined by

$$P\{\xi > t\} = e^{-(pq(x,a)+(1-p)q(x,b))t}.$$

Because randomized Markov policies for CTJMDPs and the corresponding ESMDPs are defined by the same transition probabilities, we shall apply the definitions introduced for randomized Markov policies for SMDPs to randomized Markov policies for CTJMDPs. In particular, we shall apply definitions of K-randomized stationary, randomized Markov policies of order K, and strong (K, n)-policies to CTJMDPs. So we shall consider *K-randomized stationary policies* and *randomized Markov policies of order K* for CTJMDPs. To distinguish from switching policies, we keep "randomized" in the description of randomized strong (K, n)-policies in CTJMDPs. So, we shall consider *randomized strong (K, n)-policies* for CTJMDPs.

Let π be a randomized Markov policy in the CTJMDP. Then π is also a (nonrandomized) Markov policy in the [ESMDP]. Formula (7.4) transforms into

$$\sigma_n(da|x) = \pi_n(da|x)(\alpha + q(x, a))/(\alpha + \mathbf{q}(x, \pi_n(\cdot|x))). \tag{8.1}$$

Theorem 7.1 implies the following statement.

Corollary 8.1. *For a randomized Markov policy π in a CTJMDP* (*or, equivalently, a Markov policy in the corresponding* [*ESMDP*]), *consider a randomized Markov policy σ defined by* (8.1) *for the ESMDP. Then equalities* (7.5) *and* (7.6) *hold.*

Let σ be a randomized Markov policy for the ESMDP. We define a randomized Markov policy π for the CTJMDP by

$$\pi_n(da|x) = \frac{\sigma_n(da|x)}{\alpha + q(x, a)} \left(\int_{D(x)} \frac{\sigma_n(da|x)}{\alpha + q(x, a)} \right)^{-1}. \tag{8.2}$$

Corollary 8.2. *Consider a CTJMDP. For a randomized Markov policy σ for the corresponding ESMDP consider a randomized Markov policy π for the CTJMDP defined by* (8.2). *Then $W_k(\mu, \pi) = W_k(\mu, \sigma)$, $k = 0, \ldots, K$.*

Proof. We multiply the both sides of (8.2) by $(\alpha + q(x, a))$ and integrate. We get

$$\alpha + \mathbf{q}(x, \pi_n(\cdot|x)) = \left(\int_{D(x)} \frac{\sigma_n(da|x)}{\alpha + q(x, a)} \right)^{-1} = \frac{\pi_n(da|x)(\alpha + q(x, a))}{\sigma_n(da|x)}, \tag{8.3}$$

where the second equality follows from (8.2). Since (8.3) implies (8.1), Corollary 8.1 implies Corollary 8.2. □

In particular, when for some $x \in X$ the measure $\sigma_n(\cdot|x)$ is concentrated on a finite set $D^*(x) = \{a^1, a^2, \ldots, a^l\}$, (8.2) has the following form:

$$\pi_n(a_i|x) = \frac{\sigma_n(a_i|x)}{\alpha + q(x, a_i)} \left(\sum_{j=1}^{l} \frac{\sigma_n(a_j|x)}{\alpha + q(x, a_j)} \right)^{-1}. \tag{8.4}$$

Corollary 7.5, Theorem A.9, and Corollary A.11 applied to the ESMDP imply the following result.

Theorem 8.3. (i) *Consider a countable state CTJMDP with K constraints satisfying conditions* (1)–(3) *from Theorem* 6.1. *If there exists a feasible strategy then for this CTJMDP there exists an optimal K-randomized stationary policy and for some finite $n = 0, 1, \ldots$ there exists an optimal randomized strong (K, n)-policy. If there exists a feasible strategy for a nonhomogeneous CTJMDP with K constraints then there exists an optimal randomized Markov policy of order K for this CTJMDP.*

Proof. We apply formula (8.4) to an optimal policy σ for the ESMDP described in Theorem A.9. The statement for nonhomogeneous CTJMDPs follows from Corollary A.11. □

We remark that if X and A are finite, we can apply (8.4) to the optimal policies computed in formulas (A.3.29) and (A.3.35). This provides us with the explicit algorithm that computes optimal randomized Markov policies of order K when $N < \infty$ and optimal K-randomized stationary policies for homogeneous problems when $N = \infty$.

In conclusion, we consider a CTJMDP and the corresponding ESMDP. We shall show that any randomized Markov policy for the ESMDP can be represented as a randomized policy for the CTJMDP. This implies that optimal policies for the ESMDP described in Theorem A.9 and Corollary A.11 can be viewed as optimal randomized policies for the CTJMDP.

Let σ be a randomized Markov policy for the ESMDP. Then Jacod's (1975) Proposition 3.1 implies that the predictable projection of the random measure of the process defined by σ and by an initial distribution μ has the following form: $\nu^\sigma(0, dx) = \mu(dx)$ and

$$\nu^\sigma(dt, dx) = \sum_{n\geq 0} \frac{\int_{D(x_n)} e^{-q(x_n, a)(t-t_n)} q(dx|x_n, a)\sigma_n(da|x_n)}{\int_{D(x_n)} e^{-q(x_n, a)(t-t_n)}\sigma_n(da|x_n)} 1\{t_n \leq t < t_{n+1}\}\, dt. \tag{8.5}$$

We define by

$$\pi_n(da|x_n, t-t_n) = \frac{e^{-q(x_n, a)(t-t_n)}\sigma_n(da|x_n)}{\int_{D(x_n)} e^{-q(x_n, a)(t-t_n)}\sigma_n(da|x_n)} \tag{8.6}$$

a randomized policy π for the CTJMDP for which the choice of an action depends on the following factors: the current state, the number of jumps, and the time passed after the last jump. We remark that formula (8.6) defines a switching Markov policy in the [CTJMDP].

Formula (7.3) implies that $\nu^\pi = \nu^\sigma$. Therefore, $\mathbb{P}_\mu^\pi = \mathbb{P}_\mu^\sigma$ for any initial distribution μ. It is easy to see that also $W_k(\mu, \pi) = W_k(\mu, \sigma)$, $k = 0, \ldots, K$. So, optimal randomized policies for the ESMDP, described in Theorem (A.9), can be implemented as randomized policies for the CTJMDP. Thus, we have the following statement.

Theorem 8.4. *Consider a countable state CTJMDP with K constraints satisfying conditions* (1)–(3) *from Theorem* 6.1. *If this problem is feasible then optimal policies for the corresponding ESMDP, whose existence is stated in Theorem* A.9 *and Corollary* A.11, *can be implemented by formula* (8.6) *as optimal randomized policies in the original CTJMDP.*

So, we have described three forms of optimal strategies for constrained CTJMDPs: (i) switching policies (Theorem 6.1 and Corollary 7.3), (ii) randomized policies (Corollary 8.2), and (iii) randomized policies which implement optimal randomized policies for the ESMDP (Theorem 8.4). In our opinion, switching policies are the most natural among the described classes of optimal strategies.

9. Conclusion. In this paper, we developed the techniques that reduce discounted CTJMDPs with multiple criteria and constraints to SMDPs and eventually to MDPs. By using these techniques, we have developed the theory of discounted CTJMDPs with multiple criteria and constraints. For feasible problems, the optimal policies can be founded in each of the following three forms: (i) randomized policies for ESMDPs (this is similar to constrained discrete time MDPs), (ii) switching policies for CTJMDPs (this is the most natural form of optimal policies), and (iii) randomized policies for CTJMDPs. The latter may not be natural for some applications.

By considering occupation measures, for an arbitrary strategy for a CTJMDP that changes actions between jumps, Theorem 4.5 constructs an equivalent randomized Markov policy for the ESMDP that does not change actions between jumps. Theorem 5.2 is essentially a converse of Theorem 4.5. However, it constructs equivalent switching policies only for randomized Markov policies that (i) are randomized on at most a countable subset of states and (ii) for each state select actions from a countable subset of actions. We do not know how to construct in a CTJMDP an equivalent switching policy for an arbitrary randomized Markov (or stationary) policy in the ESMDP. For countable state constrained discounted MDPs, Feinberg and Shwartz (1996) proved optimality of K-randomized stationary policies and strong (K, n)-policies, where K is the number of constraints and n is a finite integer. These policies satisfy conditions (i) and (ii). This result and its extension to SMDPs, Theorem A.9, allow us to reduce countable constrained ESMDPs to CTJMDPs. If the existence of optimal K-randomized stationary policies or optimal strong (K, n)-policies were known for constrained discounted MDPs with uncountable state spaces, it would extend

Theorem 6.1 to CTJMDPs with Borel state spaces. The existence of such policies is an open question.

Another important question is what kind of results hold for average rewards per unit time and for undiscounted total rewards. In our opinion, the approach, developed in this paper, is applicable to these criteria under the additional assumption that $\inf_{a\in D(x)} q(x,a) > 0$ for each $x \in X$. For discounted CTJMDPs, the discount rate $\alpha > 0$ plays the role of this assumption. If this assumption fails, the statement similar to Theorem 4.5 does not hold and switching strategies can outperform strategies that change actions between jumps. For example, if $\alpha = 0$ and $\inf_{a\in D(x)} q(x,a) = 0$, it is possible that $g_{\mu,0}(Y,A) < \mu(A)$ and (4.21) does not hold.

In particular, Feinberg (2002a) considered an average reward CTJMDP with K constraints and with finite state and action sets. The standard unichain condition, that every stationary policy defines a Markov chain with one ergodic class, was assumed there. According to Feinberg (2002a), if $q(x,a) > 0$ for all x and a then optimal policies can be selected either among K-randomized stationary policies for ESMDPs or among K-switching stationary policies. If $q(x,a) = 0$ for some x and a then there exists an optimal K-switching stationary policy for a feasible problem and this policy can be better than any K-randomized policy for the ESMDP; Feinberg (2002a, Example 3.1).

Appendix. Semi-Markov decision processes. This appendix contains the results on SMDPs used in this paper. The proofs and additional details can be found in Feinberg (2002).

A.1. Definitions. A probability structure of an SMDP is specified by the four objects $\{X, A, D(x), Q(t, Y|x, a)\}$ where X, A, and D are the same objects as in an CTJMDP and the transition mechanism is defined by a regular transition measure $Q(\cdot|x,a)$ from $X \times A$ into $\mathbb{R}_+ \times X$. It is assumed that (a) $Q(B|x,a)$ is a Borel function on $X \times A$ for any Borel subset $B \subseteq \mathbb{R}_+ \times X$ and (b) $Q(\cdot|x,a)$ is a measure on $\mathbb{R}_+ \times X$ with $Q(\mathbb{R}_+ \times X|x,a) \le 1$ for any $(x,a) \in X \times A$. We denote $Q(t, Y|x,a) = Q([0,t] \times Y|x,a)$ for any $0 \le t < \infty$ and for any Borel $Y \subseteq X$. If action a is selected in state x then $Q(t, Y|x,a)$ is the joint probability that the sojourn time is not greater than $t \in \mathbb{R}_+$ and the next state y is in Y (it is possible that $y = x$ with positive probability).

Let ξ be the sojourn time. Then $P\{\xi \le t\} = Q(t, X|x,a)$. In this appendix, we make the following standard assumption that implies that the system does not have accumulation points:

ASSUMPTION A.1. *There exist $\bar\epsilon > 0$ and $\bar t > 0$ such that $Q(\bar t, X|x,a) < 1 - \bar\epsilon$ for all $x \in X$ and for all $a \in A$.*

Let $\mathbf{H}_n = X \times (A \times \mathbb{R}_+ \times X)^n$, $n = 0, 1, \ldots, \infty$, be the set of all histories up to and including the nth jump. Then $\mathbf{H} = \bigcup_{0\le n<\infty} \mathbf{H}_n$ is the set of all histories that contain a finite number of jumps. The sets $\mathbf{H}_n$, $n = 0, 1, \ldots, \infty$, and $\mathbf{H}$ are endowed with the σ-fields generated by the σ-fields $\mathscr{X}$, $\mathscr{A}$, and $\mathscr{B}(\mathbb{R}_+)$. A (possibly randomized) strategy π is defined as a regular transition probability from $\mathbf{H}$ to A such that $\pi(D(x_n) \mid \omega_n) = 1$ for each $\omega_n = x_0 a_0 \xi_0 \ldots x_{n-1} a_{n-1} \xi_{n-1} x_n \in \mathbf{H}$, $n = 0, 1, \ldots$.

To define a sample space that includes trajectories that have a finite number of jumps over $\mathbb{R}_+$, we add an additional point $\bar x \notin X$ to X and an additional point $\bar a \notin A$ to A. Let $\bar X = X \cup \{\bar x\}$ and $\bar A = A \cup \{\bar a\}$. We also define $D(\bar x) = \{\bar a\}$, $Q((\infty, \bar x)|x,a) = 1 - Q(\mathbb{R}_+ \times X|x,a)$ for $x \in X$, $a \in A$, and $Q((\infty, \bar x)|x,a) = 1$ when either $x = \bar x$ or $a = \bar a$. We have that Q is a regular transition probability from $\bar X \times \bar A$ to $\bar R_+ \times \bar X$, where $\bar R_+ = [0, \infty]$.

Let $\bar{\mathbf{H}}_n = \bar X \times (\bar A \times \bar R_+ \times \bar X)^n$, $n = 0, 1, \ldots, \infty$. We also consider $\mathscr{B}(\bar{\mathbf{H}}_n) = \bar{\mathscr{X}} \times (\bar{\mathscr{A}} \times \mathscr{B}(\bar R_+) \times \bar{\mathscr{X}})^n$, where $\bar{\mathscr{X}} = \sigma(\mathscr{X}, \{\bar x\})$, $\bar{\mathscr{A}} = \sigma(\mathscr{A}, \{\bar a\})$. According to the Ionescu Tulcea theorem (Neveu 1965, Section 5.1), any initial distribution μ on X and any strategy π

define a probability measure on the set $(\overline{\mathbf{H}}_\infty, \mathscr{B}(\overline{\mathbf{H}}_\infty))$. We denote this measure by $\mathbb{P}_\mu^\pi$ and we denote the expectation operator with respect to this measure by $\mathbb{E}_\mu^\pi$.

Let $\mathbf{h}_\infty = (x_0 a_0 \xi_0 x_1 a_1 \xi_1 \ldots)$. We set $t_0 = 0$ and $t_n = t_{n-1} + \xi_{n-1}$, $n = 0, 1, \ldots$. Let $N(t) = \sup\{n \geq 0\colon t_n \leq t\}$. Assumption A.1 implies that $N(t) < \infty$, ($\mathbb{P}_\mu^\pi$-a.s.) for all $t \in \mathbb{R}_+$ and $t_n \to \infty$ ($\mathbb{P}_\mu^\pi$-a.s.) as $n \to \infty$ for all μ and π.

We may consider an SMDP as an object that has two time parameters:

(i) first time parameter is the actual continuous time t, $t = t_n$ at an nth jump epoch;

(ii) the second parameter is the jump number n.

We say that a strategy is a *policy* if at each epoch t_n, $n = 0, 1, \ldots$, the decision does not depend on the times $t_1, \ldots, t_n$. Now we give a formal definition of a policy. Let $H_n = X \times (A \times X)^n$, $n = 0, 1, \ldots, \infty$, and $H = \bigcup_{0 \leq n < \infty} H_n$. A policy π is defined as a transition probability from H to A such that $\pi(D(x_n) \mid h_n) = 1$ for each $h_n = x_0 a_0 \ldots x_{n-1} a_{n-1} x_n \in H$, $n = 0, 1, \ldots$. A *randomized Markov* policy π is defined by a sequence of transition probabilities $\{\pi_n\colon n = 0, 1, \ldots\}$ from X into A such that $\pi_n(D(x) \mid x) = 1$, $x \in X$, $n = 0, 1, \ldots$. A *Markov policy* is defined by a sequence of mappings $\varphi_n\colon X \to A$ such that $\varphi_n(x) \in D(x)$, $x \in X$, $n = 0, 1, \ldots$. A *randomized stationary* policy π is defined by a transition probability π from X into A such that $\pi(D(x) \mid x) = 1$, $x \in X$. A *stationary* policy is defined by a mapping $\varphi\colon X \to A$ such that $\varphi(x) \in D(x)$, $x \in X$.

The reward structure of an SMDP is specified by the three objects $\{\alpha, K, r_k(x, a)\}$, where

(a) $\alpha > 0$ is a discount rate;

(b) $K = 0, 1, \ldots$ is a number of constraints. We omit the index 0 when $K = 0$; and

(c) $r_k(x, a)$ is the expected discounted cumulative reward at state x for criterion $k = 0, \ldots, K$ if action a is selected. We assume that r_k are bounded above Borel functions on $X \times A$. We set $r_k(\bar{x}, \bar{a}) = 0$, $k = 0, \ldots, K$.

We remark that in many applications, a reward in a state is defined via cumulative rewards collected in a state over the period of time passed since the system moved to this state. This reward also includes the reward collected when the system jumps into or out of the state. Then the functions r_k can be computed as the corresponding expectation. Here we are not concerned with a particular form of the primitive entries that define the reward functions because for natural models this form is unimportant for infinite horizon and finite-step horizon problems considered in this paper. In fact, the cumulative rewards can be random variables. In addition, the "underlying state" can change between decision epochs and these changes may affect actual reward rates. For example, in admission control to GI/M/1 queues and in control of service modes of M/G/1 queues, the actual state of the system may change between decision epochs. These decision epochs are arrivals to GI/M/1 queues and departures from M/G/1 queues as well as arrivals to empty M/G/1 queues. However, if the controller cannot react to state changes (departures from GI/M/1 queues and arrivals to M/G/1 queues) between decision epochs, these problems can be models as SMDPs.

Given an initial state distribution μ and a strategy π, the expected total discounted rewards over the infinite horizon are

$$W_k(\mu, \pi) = \mathbb{E}_\mu^\pi \sum_{n=0}^{\infty} e^{-\alpha t_n} r_k(x_n, a_n), \quad k = 0, \ldots, K. \tag{A.1.1}$$

Similar to CTJMDPs, if an initial distribution μ is concentrated at one point x, we substitute μ with x in various objective functions W, measures $\mathbb{P}$, and expectations $\mathbb{E}$. If we consider one criterion or what we write is true for all criteria, we may omit indices $k = 0, 1, \ldots, K$. The definitions of optimal policies are similar to the definitions for CTJMDPs in §2.

We have defined a homogeneous SMDP. In addition, we can consider a nonhomogeneous SMDP when the action sets D, rewards r_k, and transition kernels Q depend on the step number. Such an SMDP can be viewed as a particular case of the definitions given above when the state space X is replaced with the state space $X \times \{0, 1, \dots\}$ of pairs (x, n), where x is a state and n is a jump number. An important particular case is an N-step model with the criterion

$$W_k(\mu, \pi, N) = \mathbb{E}_\mu^\pi \sum_{n=0}^{N} e^{-\alpha t_n} r_k(x_n, n, a_n), \quad k = 0, \dots, K, \tag{A.1.2}$$

where r_k are measurable bounded above functions. For finite-step models, the rewards $h_k(x) = r_k(x, N, a)$ are usually called final or terminal. An important application of finite-step SMDPs is scheduling of a finite number of jobs with random durations; Ross (1983), Pinedo (1995).

A.2. Reduction of SMDPs to MDPs. A discrete time MDP is a particular case of an SMDP when all sojourn times ξ_i are deterministic and equal to 1. In this case, the transition mechanism is defined by transition probabilities $p(dy|x, a)$ instead of transition kernels Q; $p(X|x, a) = 1$. In other words, $Q(t, Y|x, a) = p(Y|x, a)I\{t \ge 1\}$. Because all sojourn times are equal to 1, each strategy in an MDP is a policy and strategic measures are defined on $(H_\infty, \mathscr{B}(H_\infty))$. For MDPs formula (A.1.1) can be written in a simpler form:

$$W_k(\mu, \pi) = \mathbb{E}_\mu^\pi \sum_{n=0}^{\infty} \beta^n r_k(x_n, a_n), \quad k = 0, \dots, K, \tag{A.2.3}$$

where $\beta = e^{-\alpha}$ is a discount factor.

It is well known that for a one-criterion discounted SMDP can be reduced to a discounted MDPs; see Heyman and Sobel (1984, p. 202). Here we provide the reduction for problems with multiple criteria.

We define the regular nonnegative conditional measures on X

$$\beta(Y|x, a) = \int_0^\infty e^{-\alpha t} Q(dt, Y|x, a). \tag{A.2.4}$$

Let $\beta(x, a) = \beta(X|x, a)$. Assumption A.1 implies that $\beta(x, a) \le 1 - \bar{\epsilon}(1 - e^{-\alpha \bar{t}}) < 1$. We can interpret α as an intensity with which the process dies. Then $\beta(x, a)$ is the probability that the process does not die before the next jump. We observe that $\beta(x, a) = 0$ implies that state x is absorbing under action a.

For a strategy π, initial distribution μ, and jumps $n = 0, 1, \dots,$ we define bounded non-negative measures $M_{\mu,n}^\pi$ on $X \times A$ and $m_{\mu,n}^\pi$ on X,

$$M_{\mu,n}^\pi(Y, B) = \mathbb{E}_\mu^\pi \; e^{-\alpha t_n} I\{x_n \in Y, a_n \in B\}, \tag{A.2.5}$$

$$m_{\mu,n}^\pi(Y) = \mathbb{E}_\mu^\pi \; e^{-\alpha t_n} I\{x_n \in Y\}, \tag{A.2.6}$$

where $Y \in \mathscr{B}(X)$ and $B \in \mathscr{B}(A)$. According to formula (3.5) in Feinberg (2002),

$$m_{\mu,(n+1)}^\pi(Y) = \int_X \int_A \beta(Y|x, a) M_{\mu,n}^\pi(dx, da). \tag{A.2.7}$$

Because $m_{\mu,n}^\pi(Y) = M_{\mu,n}^\pi(Y, A)$, we have that m is a projection of M on X. In view of Corollary 7.27.2 in Bertsekas and Shreve (1978), there is a ($m_{\mu,n}^\pi$-a.e.) unique regular transition probability from X to A such that

$$\sigma_n(da|x) = \frac{M_{\mu,n}^\pi(dx, da)}{m_{\mu,n}^\pi(dx)}. \tag{A.2.8}$$

By definition, (A.2.8) is equivalent to

$$M_{\mu,n}^\pi(Y, B) = \int_Y \sigma_n(B|x) m_{\mu,n}^\pi(dx) \tag{A.2.9}$$

for all $Y \in \mathscr{B}(X)$, $B \in \mathscr{B}(A)$. Since $M^{\pi}_{\mu,n}$ is concentrated on graph D then for every $n = 0, 1, \ldots$ we can select a version of σ_n such that $\sigma_n(D(x)|x) = 1$ for all $x \in X$. Then $\sigma = \{\sigma_n\colon n = 0, 1, \ldots\}$ is a randomized Markov policy. Let $R_n(\mu, \pi) = \mathbb{E}^{\pi}_{\mu} e^{-\alpha t_n} r(x_n, a_n)$.

Lemma A.2. *Consider an SMDP. Let π be a strategy and μ be an initial distribution. Then for a randomized Markov policy σ defined by* (A.2.8)

$$M^{\sigma}_{\mu,n} = M^{\pi}_{\mu,n}, \quad n = 0, 1, \ldots. \tag{A.2.10}$$

In addition, $R_n(\mu, \sigma) = R_n(\mu, \pi)$ for all $n = 0, 1, \ldots$ and therefore $W(\mu, \sigma) = W(\mu, \pi)$ for any bounded above Borel reward function r.

Corollary A.3. *Consider an SMDP. Let μ be an initial distribution. Then for any strategy π there exists a policy σ such that* (A.2.10) *holds and therefore $W(\mu, \sigma) = W(\mu, \pi)$ for any bounded above Borel reward function r.*

We notice that Lemma A.2 generalizes to SMDPs the result established for MDPs by Derman and Strauch (1966). According to Derman and Strauch (1966), given an initial distribution, for any policy in an MDP there exists a randomized Markov policy with the equal performance. However, the major difference is that formula (A.2.8) defines different Markov policies σ for different discount rates α while the equivalent policy for MDPs is the same for all discount factors because $\xi_n = 1$.

Given an SMDP, we shall construct an equivalent MDP. Corollary A.3 implies that, in order to establish an equivalency, it is sufficient to compare the performances of policies.

We fix an arbitrary $\bar\beta \in [1 - \bar\epsilon(1 - e^{-\alpha\bar t}), 1[$ and define the transition probabilities $\bar p$ from $\bar X \times A$ to A

$$\bar p(Y|x, a) = \begin{cases} \beta(Y|x, a)/\bar\beta, & \text{if } Y \in \mathscr{B}(X), x \in X; \\ 1 - \beta(Y|x, a)/\bar\beta, & \text{if } Y = \{\bar x\}, x \in X; \\ 1, & \text{if } Y = \{\bar x\}, x = \bar x. \end{cases}$$

Then we consider an MDP with the state set $\bar X$, action set A, sets of available actions $D(x)$, reward functions r_k, $k = 0, \ldots, K$, discount factor $\bar\beta$, and transition probabilities $\bar p$. Let $\overline{\mathbb{P}}$ and $\overline{W}$ respectively denote the strategic measures and expected total discounted rewards for this MDP. The sets of all policies for the original SMDP and this MDP coincide. The following statement demonstrates that this MDP is equivalent to the original SMDP (statement (iii) follows from statement (ii) and Corollary A.4).

Corollary A.4. *Consider an SMDP and let an initial distribution μ and a policy π be given. Then the following statements hold*:

(i) $M^{\pi}_{\mu,n}(Y, B) = \bar\beta^n \overline{\mathbb{P}}^{\pi}_{\mu}(x_n \in Y, a_n \in B)$, *where* $n = 0, 1, \ldots, Y \in \mathscr{B}(X)$, *and* $B \in \mathscr{B}(A)$;

(ii) $W_k(\mu, \pi) = \overline{W}_k(\mu, \pi)$ *for all* $k = 0, \ldots, K$;

(iii) *A policy is optimal for an SMDP if and only if it is optimal for the MDP obtained from that SMDP by adding an absorbing state $\bar x$ to the state space and by replacing the transition kernel Q and discount rate α with the transition probabilities $\bar p$ and discount factor $\bar\beta$.*

A.3. Optimization of SMDPs. Corollary A.4 implies that the optimization of an SMDP is equivalent to the optimization of the MDP introduced before the formulation of Corollary A.4. Dynamic programming and linear programming are two major tools used for MDPs. Though the parameters of an MDP in Corollary A.4 depend on the selection of $\bar\beta$, the corresponding dynamic programming and the corresponding linear programming equations do not depend on the selected value of $\bar\beta$. For example, the optimality (dynamic programming) operator is $T^a f(x) = r(x, a) + \int_X f(y)\beta(dy|x, a)$; see, e.g., Puterman (1994, §11.3.3) for the discrete state space.

For constrained MDPs, the linear programming approach is natural. The main idea of this approach is to replace the finding of an optimal policy with the finding of an optimal occupation measure; see, e.g., Altman (1999), Borkar (2002), and Piunovskiy (1997). This approach can be easily extended to SMDPs.

For an MDP from Corollary A.4, the occupation measure is a measure on $(X\times A, \mathscr{X}\times\mathscr{A})$ defined by $\overline{M}_\mu^\varphi=\sum_{n=0}^\infty \bar{\beta}^n\overline{\mathbb{P}}_\mu^\varphi$ for a policy π and for an initial distribution μ. For an SMDP, we define an occupation measures $M_\mu^\pi=\sum_{n=0}^\infty M_{\mu,n}^\pi$, where π is a strategy and μ is an initial distribution.

For a policy π, Corollary A.4 implies that $M_\mu^\pi=\overline{M}_\mu^\pi$ and

$$W_k(\mu,\pi)=\overline{W}_k(\mu,\pi)=\int_X\int_A r(x,a)M_\mu^\pi(dx,da). \tag{A.3.11}$$

Let $\mathscr{M}_+$ be the set of all measures on $(X\times A,\mathscr{X}\times\mathscr{A})$. For $M\in\mathscr{M}_+$ we denote by m its projection on $(X,\mathscr{X})$, $m(Y)=M(Y\times A)$ for $Y\in\mathscr{X}$.

We fix an initial distribution μ. Lemma 4.6 in González-Hernández and Hernández-Lerma (1999) provides the necessary and sufficient condition for a measure $M\in\mathscr{M}_+$ to be an occupation measure. This result and Corollaries A.3 and A.4 imply that the set of all occupation measures for a given initial distribution μ is the set of all measures satisfying the following two conditions:

$$m(Y)=\mu(Y)+\int_X\int_A\beta(Y\mid x,a)M(dx,da)\quad\text{for all } Y\in\mathscr{X}, \tag{A.3.12}$$

$$M\in\mathscr{M}_+. \tag{A.3.13}$$

In addition, Lemma 4.6 in González-Hernández (1999) and Corollaries A.3 and A.4 imply that if M is a strategic measure for a given initial distribution μ then $M(X\times A)=1/\alpha$ and $M=M_\mu^\varphi$ for a randomized stationary policy φ that satisfies

$$\varphi(B\mid x)=\frac{M(dx,B)}{m(dx)}\quad m\text{-a.e., } B\in\mathscr{A}. \tag{A.3.14}$$

Thus, we have the following result which is similar to Lemma 3.3 in González-Hernández (1999) where MDPs were studied.

Theorem A.5. (i) *A constrained SMDP is feasible if and only if the following LP is feasible*:

$$\text{maximize }\int_X\int_A r_0(x,a)M(dx,da), \tag{A.3.15}$$

subject to (A.3.12, A.3.13) *and*

$$\int_X\int_A r_k(x,a)M(dx,da)\ge C_k,\quad k=1,\ldots,K. \tag{A.3.16}$$

(ii) *An optimal policy exists for a constrained SMDP if and only if the LP* (A.3.12), (A.3.13), (A.3.15), *and* (A.3.16) *has a solution.*

(iii) *If M is a solution of this LP then a randomized stationary policy defined in* (A.3.14) *is optimal.*

For MDPs, Corollary 5.1 in Feinberg and Piunovskiy (2002) provides a sufficient condition for the existence of an optimal policy (see there also explanations at the bottom of p. 107 that one of the conditions listed there always holds for the discounted criterion; see also earlier sufficient conditions by González-Hernández and Hernández-Lerma 2000). For an SMDP, the conditions are almost the same. The only difference is that we need an assumption that $Q(\cdot\mid x,a)$ is weakly continuous in (x,a). This assumption implies the weak continuity of transition probabilities for an MDP considered in Corollary A.4. This weak continuity is required for MDPs to ensure the existence of optimal policies; see González-Hernández and Hernández-Lerma (2000) and Feinberg and Piunovskiy (2002). For finite-step SMDPs, Theorem A.5 implies the following.

Corollary A.6. *Let N be a finite integer.*

(i) *A constrained N-step SMDP is feasible if and only if the following LP*:

$$\text{(A.3.17)} \qquad \text{maximize} \sum_{n=0}^{N-1} \int_X \int_A r_0(x, n, a) M_n(dx, da),$$

subject to

$$\text{(A.3.18)} \qquad m_n(Y) = \mu(Y) + \int_X \int_A \beta(Y|x, a) M_{n-1}(dx, da) \quad \textit{for all } Y \in \mathscr{X},$$

with $M_{-1}(B) \equiv 0$ and $m_{-1} = \mu$,

$$\text{(A.3.19)} \qquad M_n \in \mathscr{M}_+ \quad \textit{and} \quad m_n(Y) = M_n(Y, A), \qquad n = 0, \ldots, N-1,$$

$$\text{(A.3.20)} \qquad \sum_{n=0}^{N-1} \int_X \int_A r_k(x, n, a) M(dx, da) \ge C_k, \quad k = 1, \ldots, K$$

is feasible.

(ii) *An optimal policy exists for a constrained N-step SMDP if and only if the LP* (A.3.17)–(A.3.20) *has a solution. In addition, if M is a solution of this LP then a randomized Markov policy*

$$\text{(A.3.21)} \qquad \varphi_n(B|x) = \frac{M_n(dx, B)}{m_n(dx)} \quad m_n\text{-a.e.}, \quad n = 1, \ldots, N, \; B \in \mathscr{A},$$

is optimal.

For countable state MDPs, Feinberg and Shwartz (1996) proved that the number of randomization procedures that an optimal policy may use is limited by the number of constraints. This result, which is still open for uncountable MDPs, is important for CTJMDPs because it implies that the number of switching points may be limited by the number of constraints in CTJMDPs. Therefore, until the end of the appendix, we shall consider countable models. We shall assume that the following conditions hold.

Assumption A.7. (a) *X is countable or finite*;

(b) *$D(x)$ is compact for each $x \in X$*;

(c) *for all $x, y \in X$, functions $\beta(y|x, a)$ and $\beta(x, a)$ are continuous in $a \in D(x)$* (*the latter condition is not needed if X is finite*);

(d) *functions $r_k(x, a)$ are uniformly bounded above and for each $x \in X$ they are upper semi-continuous in $a \in D(x)$.*

We remark that if $Q(\cdot|x, a)$ are weakly continuous in $a \in D(x)$ and Assumption A.1 holds, then Assumption A.7(c) holds.

For nonhomogeneous SMDPs considered in the end of §A.1, we also assume Assumption A.7 with x being replaced by (x, n) in conditions (b)–(d). For finite-step homogeneous models, conditions (b)–(d) imply the corresponding conditions when the states x are replaced with (x, n).

First we consider an SMDP with one criterion ($K = 0$). Let Π be the set of all strategies, $v(x) = \sup\{W(x, \pi) | \pi \in \Pi\}$ be the value function and $A^c(x) = \{a \in D(x) | v(x) = r(x, a) + T^a v(x)\}$, $x \in X$, be the sets of conserving actions. Similar to Theorem 4.2(i) in Feinberg and Shwartz (1996), $A^c(x)$ are nonempty and compact. Theorem A.8 follows from Corollary A.4 and from the results for discounted MDPs (Feinberg and Shwartz 1996, Theorem 4.2).

Theorem A.8. *Consider a one-criterion SMDP with a finite or countable state space X. A randomized stationary policy π is optimal if and only if $\pi(A^c(x)|x) = 1$ for all $x \in X$. Therefore, if Assumption* A.7 *holds then there exists an optimal stationary policy.*

To compute an optimal policy, one can apply the standard techniques, such as policy iteration, value iteration, and linear programming algorithms, to the MDP described in Corollary A.4. For a finite-step SMDP, an optimal policy can be computed by the standard value iteration (dynamic programming) algorithm.

Now we consider a discounted SMDP with multiple criteria. To characterize optimal policies, we need additional definitions similar to those introduced in Feinberg and Shwartz (1996) for MDPs.

We say that a policy is *discrete* if for each history $h_n = x_0 a_0 \ldots x_n$ the measure $\pi(\cdot|h_n)$ is concentrated on a countable set. We say that a randomized Markov policy π has *index* m if it is discrete and

$$\text{(A.3.22)} \qquad \sum_{x\in X}\left[\left|\bigcup_{n=0}^{\infty}\{a\in D(x)|\pi_n(a|x)>0\}\right|-1\right]\le m.$$

If a randomized stationary policy has an index m, then this policy is called *m-randomized* stationary. For a randomized stationary policy π, (4.1) is equivalent to

$$\text{(A.3.23)} \qquad \sum_{x\in X}[|\{a\in D(x)|\pi(a|x)>0\}|-1]\le m.$$

We say that a randomized Markov policy π is a randomized Markov policy of order m if it is discrete and

$$\text{(A.3.24)} \qquad \sum_{n=0}^{\infty}\sum_{x\in X}[|\{a\in D(x)|\pi_n(a|x)>0\}|-1]\le m.$$

For N-step models, ∞ should be replaced by $(N-1)$ in (A.3.22) and (A.3.24).

If a randomized Markov policy has the index m, it uses no more than m additional actions than a stationary policy. A randomized Markov policy of the order m uses no more than m additional actions than a Markov policy.

A randomized Markov policy π is called *an (m, n)-policy* if it is of the order m, and from the step n onwards it coincides with a stationary policy. The last property means that there is a stationary policy φ such that $\pi_i(\varphi(x)|x) = 1$ for all $x \in X$ and for all $i \geq n$. If an (m, n)-policy has the index m then it is called *a strong (m, n)-policy*. The following theorem follows from Theorem 2.1 in Feinberg and Shwartz (1996) and Corollary A.4.

Theorem A.9. *Let the state space X be finite or countable and let Assumption* A.7 *holds. If an SMDP is feasible, then*
(i) *there exists an optimal K-randomized stationary policy*; *and*
(ii) *for some finite $n = 0, 1, \ldots$ there exists an optimal strong (K, n)-policy.*

If A is finite or countable, the LP (A.3.12), (A.3.13), (A.3.15), and (A.3.16) can be rewritten in the following form

$$\text{(A.3.25)} \qquad \text{maximize} \sum_{y\in X}\sum_{a\in D(y)} r_0(y,a)u_{y,a},$$

subject to

$$\text{(A.3.26)} \qquad \sum_{a\in D(y)} u_{y,a} - \sum_{z\in X}\sum_{a\in D(z)} \beta(y|z,a)u_{z,a} = \mu(y), \quad y\in X,$$

$$\text{(A.3.27)} \qquad \sum_{y\in X}\sum_{a\in D(y)} r_k(y,a)u_{y,a} \ge C_k, \quad k=1,\ldots,K,$$

$$\text{(A.3.28)} \qquad u_{y,a}\ge 0, \quad y\in X,\ a\in D(y).$$

Theorem A.5 and the results on discounted MDPs (Feinberg and Shwartz 1995, Theorem 4.2) imply the following statement.

Theorem A.10. (i) *Let the state space X be finite or countable and Assumption* A.7 *holds. If the LP* (A.3.25)–(A.3.28) *is feasible, it has an optimal solution.*

(ii) *Let X and A be finite. If u is an optimal basic solution of the LP* (A.3.25)–(A.3.28) *then the formula*

$$\varphi(a|y) = \begin{cases} \dfrac{u_{y,a}}{\sum_{b\in D(y)} u_{y,b}}, & \text{if } \sum\limits_{b\in D(y)} u_{y,b} > 0, \\ 1\{a = a(y)\}, & \text{otherwise}, \end{cases} \tag{A.3.29}$$

where $y \in X$ and $a(y)$ is an arbitrary element of $D(y)$, defines an optimal K-randomized stationary policy φ.

Consider a nonhomogeneous SMDP. Theorem A.9 implies the following.

Corollary A.11. (i) *For a one-criterion nonhomogeneous SMDP there exists an optimal Markov policy.* (ii) *If a nonhomogeneous SMDP with K criteria is feasible, then there exists an optimal randomized Markov policy of order K.*

Now we consider a finite-step SMDP. If X and A are finite or countable, the LP (A.3.17)–(A.3.20) can be rewritten in the following form:

$$\text{maximize} \sum_{y\in X}\sum_{n=0}^{N-1}\sum_{y\in D(y)} r_0(y, n, a)u_{y,n,a}, \tag{A.3.30}$$

subject to

$$\sum_{a\in D(y)} u_{y,0,a} = \mu(y), \quad y \in X, \tag{A.3.31}$$

$$\sum_{a\in D(y)} u_{y,n,a} - \sum_{z\in X}\sum_{a\in D(z)} \beta(y|z, n, a)u_{z,n-1,a} = 0, \quad n = 0, \dots, N-1, \ \ y \in X, \tag{A.3.32}$$

$$\sum_{y\in X}\sum_{n=0}^{N-1}\sum_{a\in D(y)} r_k(y, n, a)u_{y,n,a} \ge C_k, \quad k = 1, \dots, K, \tag{A.3.33}$$

$$u_{y,n,a} \ge 0, \quad y \in X, \ \ n = 0, \dots, N-1, \ \ a \in D(y). \tag{A.3.34}$$

Corollary A.6 and the results on finite horizon MDPs (Feinberg and Shwartz 1995, §4.1) imply the following result.

Theorem A.12. *Let the state space be finite or countable and let Assumption* A.7 *holds. Consider a finite-step SMDP.* (i) *If this LP is feasible, it has an optimal solution.* (ii) *If the state and action sets X and A are finite and u is an optimal basic solution of the LP* (4.21)–(4.25), *then the formula*

$$\varphi_n(a|y) = \begin{cases} \dfrac{u_{y,n,a}}{\sum_{b\in D(y)} u_{y,n,b}}, & \text{if } \sum\limits_{b\in D(y)} u_{y,n,b} > 0, \\ 1\{a = a(y)\}, & \text{otherwise}, \end{cases} \tag{A.3.35}$$

where $n = 0, 1, \dots, N-1$, $y \in X$, and $a(y) \in D(y)$ are arbitrary, defines an optimal randomized Markov policy φ of order K.

Acknowledgments. This research was partially supported by NSF Grants DMI-9500746, DMI-9908258, and DMI-0300121. The author thanks Martin L. Puterman for helpful suggestions. The author is grateful to the associate editor and referee for their valuable comments. In particular, §§3 and 9 and Figure 1 were added following the associate editor's suggestions.

References

Altman, E. 1999. *Constrained Markov Decision Processes.* Chapman & Hall/CRC, Boca Raton, FL.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control, Volume II.* Athena Scientific, Belmont, MA.

Bertsekas, D. P., S. E. Shreve. 1978. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, New York. (Republished by Athena Scientific, 1997.)

Borkar, V. S. 2002. Convex analytic methods in Markov decision processes. E. A. Feinberg, and A. Shwartz, eds. *Handbook of Markov Decision Processes: Methods and Applications.* Kluwer, Boston, MA, 347–375.

Cassandras, C. G. 1993. *Discrete Event Systems.* IRWIN, Boston, MA.

Derman, C., R. E. Strauch. 1966. A note on memoryless rules for controlling sequential control processes. *Ann. Math. Statist.* **37** 276–278.

Dynkin, E. B., A. A. Yushkevich. 1979. *Controlled Markov Processes.* Springer, Berlin.

Feinberg, E. A. 1994. A generalization of "expectation equals reciprocal of intensity" to non-stationary exponential distributions. *J. Appl. Probab.* **31** 262–267.

Feinberg, E. A. 2002. Constrained discounted semi-Markov decision processes. Z. How, J. A. Filar, and A. Chen, eds. *Markov Processes and Controlled Markov Chains*. Kluwer, Dordrecht, The Netherlands, 233–244.

Feinberg, E. A. 2002a. Constrained finite continuous-time Markov decision processes with average rewards. *Proc. 2002 IEEE Conference on Decisions and Control*, Las Vegas, December 2002. IEEE Control Systems Society, Piscataway, NJ, 3805–3810.

Feinberg, E. A., A. B. Piunovskiy. 2002. Nonatomic total reward Markov decision processes with multiple criteria. *J. Math. Anal. Appl.* **273** 93–111.

Feinberg, E. A., A. Shwartz. 1995. Constrained Markov decision models with weighted discounted rewards. *Math. Oper. Res.* **20** 302–320.

Feinberg, E. A., A. Shwartz. 1996. Constrained discounted dynamic programming. *Math. Oper. Res.* **21** 922–945.

Gonzáles-Hernández, J., O. Hernández-Lerma. 1999. Envelopes of sets of measures, tightness, and Markov control processes. *Appl. Math. Optim.* **40** 377–392.

Hernández-Lerma, O., J. Gonzáles-Hernández. 2000. Constrained Markov control processes in Borel spaces. *Math. Meth. Oper. Res.* **52** 271–285.

Heyman, D. P., M. J. Sobel. 1984. *Stochastic Models in Operations Research, Volume II: Stochastic Optimization.* McGraw-Hill, New York.

Hordijk, A., F. A. van der Duyn Schouten. 1979. Discretization procedures for continuous time decision processes. *Trans. Eighth Prague Conf. on Inform. Theory, Statistical Decision Functions, Random Processes, 1978, Volume C*. Academia, Prague, Czechoslovakia, 143–154,

Jacod, J. 1975. Multivariate point processes: Predictable projections, Radon-Nikodym derivatives, representation of martingales. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **31** 235–253.

Kakumanu, P. 1971. Continuously discounted Markov decision models with countable state and action space. *Ann. Math. Statist.* **42** 919–926.

Kakumanu, P. 1977. Relation between continuous and discrete time Markovian decision problems. *Naval Res. Logist. Quart.* **24** 431–439.

Kallianpur, G. 1980. *Stochastic Filtering Theory.* Springer, New York.

Kitaev, Yu. M. 1985. Semi-Markov and jump Markov controlled models: Average cost criterion. *SIAM Theory Probab. Appl.* **30** 272–288.

Kitaev, Yu. M., V. V. Rykov. 1995. *Controlled Queueing Systems.* CRC Press, New York.

Lippman, S. A. 1975. Applying a new device in the optimization of exponential service systems. *Oper. Res.* **23** 687–710.

Miller, B. L. 1968. Finite state continuous time Markov decision processes with a finite planning horizon. *SIAM J. Control* **6** 266–280.

Miller, B. L. 1968a. Finite state continuous time Markov decision processes with an infinite planning horizon. *J. Math. Anal. Appl.* **22** 552–569.

Neveu, J. 1965. *Mathematical Foundations of the Calculus of Probability.* Holden-Day, San Francisco, CA.

Parthasarathy, K. R. 1967. *Probability Measures on Metric Spaces.* Academic Press, New York.

Pinedo, M. 1995. *Scheduling: Theory, Algorithms, and Systems.* Prentice Hall, Englewood Cliffs, NJ.

Piunovskiy, A. B. 1997. *Optimal Control of Random Sequences in Problems with Constraints.* Kluwer, Boston, MA.

Piunovskiy, A. B. 1997a. A controlled jump discounted model with constraints. *SIAM Theory Probab. Appl.* **42** 51–72.

Puterman, M. L. 1994. *Markov Decision Processes.* John Wiley, New York.

Ross, S. M. 1983. *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

Sennott, L. I. 1999. *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley, New York.

Serfozo, R. F. 1979. An equivalence between continuous and discrete time Markov decision processes. *Oper. Res.* **27** 616–620.

Yushkevich, A. A. 1977. Controlled Markov models with countable state space and continuous time. *SIAM Theory Probab. Appl.* **22** 215–235.

Yushkevich, A. A. 1980. On reducing a jump controllable Markov model to a model with discrete time. *SIAM Theory Probab. Appl.* **25** 58–69.

Yushkevich, A. A. 1980a. Controlled jump Markov models. *SIAM Theory Probab. Appl.* **25** 244–266.

Yushkevich, A. A., E. A. Feinberg. 1979. On homogeneous Markov models with continuous time and finite or countable state space. *SIAM Theory Probab. Appl.* **26** 156–161.