Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

ON STATE SPACE REDUCTION IN SEQUENTIAL DECISION PROCESSES

Eugene A. Feinberg

Abstract— **This paper provides sufficient conditions when certain information about the past of a stochastic sequential decision processes can be ignored by a controller and when the problem state space can be reduced to a smaller set. We illustrate the results with particular applications to queueing control, control of semi-Markov decision processes with iid sojourn times, and uniformization of continuous-time Markov decision processes.**

I. INTRODUCTION

What information should be used by a controller is one of the central questions in control of stochastic processes and, in particular, in Markov decision processes (MDPs). In the general scheme of MDPs, the history consists of the past states and actions. In this paper, we study the question whether the controller can achieve a better performance by using additional information. We consider a more general model than an MDP. In the model considered in this paper, transition probabilities may depend on the past states and actions and we call such a model a Stochastic Decision Process (SDP). Under general assumptions we show that, if neither the transition probabilities nor the objective criterion depends on the additional information, this information cannot be used to improve the system performance. This allows the controller to reduce the state space of the problem. We also provide such results for continuous-time jump problems.

In addition, we discuss particular applications. We consider control of an $M^X/G/1$ queue with a removable server, control of a Semi-Markov Decision Process (SMDP) with iid time intervals between jumps, uniformized Continuous-Time Markov Decision Processes (CTMDPs), and admission control. For SMDPs with iid sojourn times and uniformized MDPs, we prove that the continuous time parameter can be ignored for the average cost per unit time criterion. Therefore, a uniformized CTMDP is equivalent to the corresponding discrete-time MDPs. If there is a stationary optimal policy for this MDP, this policy is optimal for the original CTMDP because it is optimal for the uniformized CTJMDP and can be implemented in the original CTMDP. For an $M^X/G/1$ queue with a removable server and holding costs depending only on the workload, we show that the knowledge of the numbers of arriving customers in batches and their workloads is not useful when the total workload is known. For the call admission problem, we show that it is possible to merge the classes of arrivals with equal payoffs.

This work was supported in part by grant DMI-0300121 from the National Science Foundation

Department of Applied Mathematics and Statistics; State University of New York; Stony Brook, NY 11794-3600; USA; Eugene.Feinberg@sunysb.edu

The results of this paper provide a rigorous justification of the principle used in applications of MDPs that only minimal possible information should be used in the construction of an MDP for a particular problem. In Section II we prove this principle for discrete-time problems, in Section III we prove it for continuous-time problems, and Section IV deals with applications.

II. DISCRETE-TIME PROBLEMS

Consider a Stochastic Decision Process (SDP) defined by the quadruplet $\{X, A, p, v\}$, where X is the state space, A is the action space, p is the transition kernel, and v is the criterion. We assume that X and A are Borel spaces, i.e. they are isomorphic to measurable subsets of a Polish (in other words, complete separable metric) space; see [3] or [5] for details. Let $H_n = X \times (A \times X)^n$ be the sets of histories up to epoch $n = 0, 1, \ldots$ and let $H = \cup_{0 \le n < \infty} H_n$ be the set of all finite histories. We can also consider the set of infinite histories $H_\infty = (X \times A)^\infty$. The products of the Borel σ-fields on X and A define Borel σ-fields on H_n, $n = 0, 1, \ldots, \infty$, and these σ-fields generate a Borel σ-field on H. Then p is defined as a regular transition probability from $H \times A$ to X, i.e. $p(B|h, a)$ is a Borel function on $H \times A$ for any fixed Borel subset B of X and $p(\cdot|h)$ is a probability measure on A for any pair (h, a), where $h \in H$ and $a \in A$.

A strategy is defined as a regular transition probability from H to A. Therefore, a strategy defines the transition probabilities from H_n to A and the transition kernel p defines the transition probabilities from $H_n \times A$ to X. According to Ionescu Tulcea's theorem [5], any initial probability distribution μ on X and any strategy π define a unique probability measure P^π_μ on H_∞. Following [5], we shall call P^π_μ a strategic measure.

A criterion v is defined as a function of a strategic measure, $v = v(P^\pi_\mu)$. In particular, v can be a numerical function. If p is just a function of (x_n, a_n), the defined SDP becomes a Markov Decision Process (MDP).

The total expected costs (or rewards) and the average costs (or rewards) per unit time are two important criteria studied in the literature; see [10, page 5] for detailed definitions. Expected total costs can be represented in a form of $v = E^\pi_\mu U(h_\infty)$, where $h_\infty \in H_\infty$ and U is a measurable function on H_∞. Average costs per unit time can be represented as a limiting point of a sequence $v_n = E^\pi_\mu U_n(h_\infty)$, $n = 0, 1, \ldots$, where U_n are measurable functions on H_∞.

We remark that it is natural to consider problems in which action sets depend on the current state or even on the past history, [3], [5], [10], [13], [14], [19], [20]. We do not do

0-7803-9568-9/05/$20.00 ©2005 IEEE

it here because of the following two reasons: (i) simplicity and (ii) the functions U and U_n can be set equal to $-\infty$ on infeasible trajectories for maximization problems and to $+\infty$ for minimization problems.

Now assume that $X = X^1 \times X^2$, where X^1 and X^2 are two Borel spaces. The state of the system is $x = (x^1, x^2)$. In addition, we assume that at each stage $n = 0, 1, \ldots$ the transition kernel p does not depend on the second components of the state space. In other words, the probability

$$\begin{aligned} &p(dx^1_{n+1}|x^1_0, x^2_0, a_0, \ldots, x^1_n, x^2_n, a_n) \\ &= p(dx^1_{n+1}|x^1_0, a_0, \ldots, x^1_n, a_n) \end{aligned} \tag{2.1}$$

does not depend on x^2_i, $i = 0, 1, \ldots, n$. Then it is natural to consider an SDP with the state space X^1, action set A, and transition kernels $p(dx^1_{n+1}|x^1_0, a_0, \ldots, x^1_n, a_n)$. Let $\tilde{P}^\sigma_{\eta^1}$ be a strategic measure for this smaller SDP, where η^1 is an initial probability distribution on X^1 and σ is a policy in the smaller model. Every $\tilde{P}^\sigma_{\eta^1}$ is a probability measure on the space $(X^1 \times A)^\infty$.

Theorem 2.1. *Consider an SDP with the state space $X = X^1 \times X^2$ and let assumption (2.1) hold. For any initial state distribution μ on $X = X^1 \times X^2$ and for any policy π for this SDP, consider a policy σ for the SDP with the state space X^1 defined for all $n = 0, 1, \ldots$ (P^π_μ-a.s.) by*

$$\sigma(da_n|x^1_0 a_0 x^1_1 a_1 \ldots x^1_n) = \frac{P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n da_n)}{P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n)}. \tag{2.2}$$

Then

$$P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots) = \tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 dx^1_1 da_1 \ldots),$$

where μ^1 is the marginal probability measure on X^1 induced by μ, i.e. $\mu^1(C) = \mu(C \times X^2)$ for any measurable subset C of X^1. In other words, $\tilde{P}^\sigma_{\mu^1}$ is the projection of the strategic measure P^π_μ on $(X^1 \times A)^\infty$.

Proof: By Kolmogorov's extension theorem, it is sufficient to prove that for any $n = 0, 1, \ldots$

$$P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n) = \tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n). \tag{2.3}$$

We prove this equality by induction in n. It holds for $n = 0$ because $P^\pi_\mu(x^1_0 \in C) = \tilde{P}^\sigma_{\mu^1}(x^1_0 \in C) = \mu^1(C)$ for any policies π and σ in the corresponding models.

Let (2.3) hold for some n. Then

$$\begin{aligned} &\tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n da_n) \\ &= \tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n)\sigma(da_n|x^1_0 a_0 x^1_1 a_1 \ldots x^1_n) \\ &= P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n da_n), \end{aligned} \tag{2.4}$$

where the first equality follows from the definition of a strategic measure and the second equality follows from (2.2) and (2.3). Since the transition probabilities in the first model do not depend on x^2, we have

$$\begin{aligned} &\tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n da_n dx^1_{n+1}) \\ &= \tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 \ldots dx^1_n da_n)p(dx^1_{n+1}|x^1_0, a_0, \ldots, x^1_n, a_n) \\ &= P^\pi_\mu(dx^1_0 da_0 \ldots dx^1_n da_n)p(dx^1_{n+1}|x^1_0, a_0, \ldots, x^1_n, a_n) \\ &= P^\pi_\mu(dx^1_0 da_0 dx^1_1 da_1 \ldots dx^1_n da_n dx^1_{n+1}), \end{aligned} \tag{2.5}$$

where the first equality follows from the definition of the strategic measure $\tilde{P}^\sigma_{\mu^1}$ and the second equality follows from (2.4). ■

Corollary 2.1. *Consider an SDP with the state space $X = X^1 \times X^2$ and let assumption (2.1) hold. In addition, let the criterion v be defined as a limiting point of $E^\pi_\mu U_n(x^1_0, a_0, x^1_1, a_1, \ldots)$, where U_n are measurable functions. For an arbitrary policy π in this SDP, consider a policy σ defined by (2.2) in the SDP with the state space X^1. Then $v(\mu, \pi) = v(\mu^1, \sigma)$, where the initial probability μ^1 on X^1 is defined by $\mu^1(B) = \mu(B \times X^2)$ for any measurable subset B of X^1. In other words, if*

$$v(\mu, \pi) = \lim_{n_k \to \infty} E^\pi_\mu U_{n_k}(x^1_0, a_0, x^1_1, a_1, \ldots) \tag{2.6}$$

then

$$v(\mu^1, \sigma) = \lim_{n_k \to \infty} E^\sigma_{\mu^1} U_{n_k}(x^1_0, a_0, x^1_1, a_1, \ldots) = v(\mu, \pi). \tag{2.7}$$

III. Continuous-time jump problems

In the defined SDP, all time intervals between decisions equal 1. In this section, we extend Theorem 2.1 to a more general situation when these intervals may be random and different.

We define a Continuous-Time SDP (CTSDP). A trajectory of a CTSDP is a sequence $x_0, a_0, \tau_0, x_1, a_1, \tau_1, \ldots$, where x_n is the state of the system after jump n, a_n is the action selected after the jump occurred, and τ_n is the time until the next jump. The above definition of an SDP is so general that we can use it to define a CTSDP. We set $\tau_{-1} = 0$ and define a CTSDP $\{X, A, q\}$, as an SDP $\{[0, \infty) \times X, A, q, v\}$, where X is a Borel state space, A is a Borel action space, and q is a transition kernel which is a conditional joint distribution of the sojourn time and the next state. According to this definition, the transition probabilities after the n-th jump are $q(d\tau_n, dx_{n+1}|x_0, a_0, \tau_0, x_1, a_1, \ldots, \tau_{n-1}, x_n, a_n)$. The objective criterion is a function of a strategic measure for this SDP. For SDPs, we consider only initial distributions μ on $[0, \infty) \times X$ with $\mu(0, X) = 1$, i.e. $\tau_{-1} = 0$ with probability 1. Therefore, we interpret μ as a probability measure on X and will not mention τ_{-1} anymore. A CTSDP is called a Semi-Markov Decision Process (SMDP) if the SDP $\{[0, \infty) \times X, A, q\}$ is an MDP. In other words, if the transition kernel q has the form $q(d\tau_n, dx_{n+1}|x_n, a_n)$.

Similarly to the discrete time case, consider an SDP with a Borel state space $X = X^1 \times X^2$ and a Borel action space A. We assume that the joint distributions of τ_n and x^1_{n+1} do

not depend on x_i^2, $i = 0, 1, \ldots, n$, i.e.

$$q(d\tau_n, dx_{n+1}^1 | x_0^1, x_0^2, a_0, \tau_0, x_1^1, x_1^2, a_1, \tau_1, \ldots, x_n^1, x_n^2, a_n) = q(d\tau_n, dx_{n+1}^1 | x_0^1, a_0, \tau_0, x_1^1, a_1, \tau_1, \ldots, x_n^1, a_n). \tag{3.1}$$

Similarly to the discrete time case, we can consider a smaller CTSDP with the state space X^1, action space A, and transition kernel $q(d\tau_n, dx_{n+1}^1 | x_0^1, a_0, \tau_0, x_1^1, a_1, \tau_1, \ldots, x_n^1, a_n)$. Theorem 2.1 implies a similar result for CTSDPs.

Corollary 3.1. *Consider a CTSDP with the state space $X = X^1 \times X^2$ and let assumption (3.1) hold. For any initial state distribution μ on $X = X^1 \times X^2$ and for any policy π for the CTSDP with the state space $X = X^1 \times X^2$, consider a policy σ for the CTSDP with the state space X^1 defined (P_μ^π-a.s.) by*

$$\sigma(da_n | x_0^1 a_0 \tau_0 x_1^1 a_1 \tau_1 \ldots x_n^1) = \frac{P_\mu^\pi(dx_0^1 da_0 d\tau_0 dx_1^1 da_1 d\tau_1 \ldots dx_n^1 da_n)}{P_\mu^\pi(dx_0^1 da_0 d\tau_0 dx_1^1 da_1 d\tau_1 \ldots dx_n^1)}, \tag{3.2}$$

$n = 0, 1, \ldots$. *Then*

$$P_\mu^\pi(dx_0^1 da_0 d\tau_0 dx_1^1 da_1 d\tau_1 \ldots) = \tilde{P}_{\mu^1}^\sigma(dx_0^1 da_0 d\tau_0 dx_1^1 da_1 d\tau_1 \ldots), \tag{3.3}$$

where μ^1 is the marginal probability measure X^1 induced by μ, i.e. $\mu^1(C) = \mu(C, X^2)$ for any measurable subset C of X^1. In other words, $\tilde{P}_{\mu^1}^\pi$ is the projection of the strategic measure P_μ^π on $(X^1 \times A \times [0, \infty))^\infty$.

Corollary 3.2. *Consider a CTSDP with the state space $X = X^1 \times X^2$ and let assumption (3.1) hold. In addition, let the criterion v be defined as a limiting point of $E_\mu^\pi U_t(x_0^1, a_0, \tau_0, x_1^1, a_1, \tau_1, \ldots)$, where U_t are measurable functions for each $t \geq 0$. For an arbitrary policy π for this CTSDP, consider a policy σ for the CTSDP with the state space X^1 defined by (3.2). Then $v(\mu, \pi) = v(\mu^1, \sigma)$, where the initial probability μ^1 on X^1 is defined by $\mu^1(B) = \mu(B \times X^2)$ for any measurable subset B of X^1. In other words, if*

$$v(\mu, \pi) = \lim_{t_k \to \infty} E_\mu^\pi U_{t_k}(x_0^1, a_0, \tau_0, x_1^1, a_1, \tau_1, \ldots) \tag{3.4}$$

then

$$v(\mu^1, \sigma) = \lim_{t_k \to \infty} E_{\mu^1}^\sigma U_{t_k}(x_0^1, a_0, \tau_0, x_1^1, a_1, \tau_1, \ldots) = v(\mu, \pi). \tag{3.5}$$

The average cost per unit time is an important example of a criterion that satisfies the conditions of Corollary 3.2. Let $c(x, a, t) \geq 0$ be the cost incurred during time t elapsed since the last jump, where x is the current state and a is the last selected action. Let $t_0 = 0$ and $t_{n+1} = t_n + \tau_n$, $n = 0, 1, \ldots$. We set $N(t) = \sup\{n = 0, 1, \ldots | t_n \leq t\}$. The cumulative cost up to time t is

$$U_t(h_\infty) = \sum_{n=0}^{N(t)-1} c(x_n, a_n, \tau_n) + c(x_{N(t)}, a_{N(t)}, t - t_{N(t)}) \tag{3.6}$$

for any trajectory $h_\infty = x_0, a_0, \tau_0, x_1, a_1, \tau_1, \ldots$. The average cost per unit time is defined as

$$v(\mu, \pi) = \limsup_{t \to \infty} t^{-1} E_\mu^\pi U_t(h_\infty). \tag{3.7}$$

This criterion satisfies (3.4).

IV. Examples of applications

Control of $M^X/G/1$ queues with removable servers. Consider a single-server queue with batch arrivals. The batches arrive according to a Poisson process with a given intensity. At the arrival epoch, the workload in the batch becomes known. Let Y_i be the workload in batch i, where Y_i, $i = 1, 2, \ldots$, are nonnegative iid random variables that are also independent on the arrival process and the state of the queue.

The server can be in one of two states: on and off. If the server is on, its service rate constant and deterministic. The service rate is the amount of workload leaving a nonempty system per unit time. Without loss of generality, we assume that the service rate equals 1. The server can be switched on and off any time. It costs K_0 to switch the server on and K_1 to switch the server off, where these switching costs are nonnegative and at least one of them is positive. If the server is on, the running cost is $r > 0$ and if the server is off, the running cost is zero. The holding cost is a nonnegative function $h(w)$ of the workload w.

The workload can change continuously. However, this problem can be described as an SMDP. The state of this SMDP is a pair (w, g), where w is the current workload, $g = 0$ means that the server is off, and $g = 1$ means that the server is on. Thus, the system state space is $X = [0, \infty) \times \{0, 1\}$, where the first coordinate is the workload and the second coordinate is the state of the server.

Let the initial state of the system $x_0 = (w_0, g_0)$ be given. An action set is defined as $A = [0, \infty)$. The first decision epoch t_0 is 0. Though the workload may change continuously, it is possible to model this system by a process whose states do not change between decision epochs. If action a_n is selected at state $x_n = (w_n, g_n)$ at some decision epoch t_n, the system stays at x_n during the time $\tau_n = \min\{a_n, \xi_n\}$, $a_n \in A$, until the next decision epoch $t_{n+1} = t_n + \tau_n$, where ξ_n is the time until the next batch arrives. The random variables ξ_n are independent and have exponential distributions with the intensity of the Poisson process formed by arriving batches. In addition, if $a_n < \tau_n$, i.e. an arraval did not happened during the selected time a_n, the state of the server g_n changes at the epoch t_{n+1} from 0 to 1 or vice versa. The dynamics of the system can be described by

$$x_{n+1} = (w_{n+1}, g_{n+1}) = \begin{cases} ((w_n - g_n a)^+, (1 - g_n)), & \text{if } \xi_n > a_n; \\ ((w_n - g_n \xi_n)^+ + Z_n, g), & \text{if } \xi_n \leq a_n; \end{cases}$$

where $d^+ = \max\{d, 0\}$ for any number d and Z_n are iid random variable with the same distribution as Y_1.

It is easy to calculate $q(d\tau_n, dx_{n+1}|x_n, a_n)$ by using the explicit definitions of τ_n and x_{n+1} provided above. However, we do not need here the explicit formula for q. The costs incurred during the first u units of time that the system spent at state x_n is

$$c(x_n, a_n, u) = \int_0^u h((w_n - g_n t)^+)dt + r(1 - g_n)u + K_{g_n} I\{u = a_n\},$$

lllwhere $u \le \tau_n$. Consider the average cost per unit time criterion (3.7).

Control of queues with the removable server and known workload has been studied in the literature since 1973 when Balachandran [1] introduced a notion of a D policy that switches the server on when the workload is greater than or equal to D and switches the server off when the system becomes empty. The optimality of D policies under broad conditions was proved in [8], where it was assumed that the controller knows only the workload w and the state of the server.

At the arrival epochs, the controller may also observe the numbers of arrivals in batches and their individual workloads, and use this information to control the system. This formulation leads to the SMDP with the state space $X^1 \times X^2$, where $X^1 = X = [0, \infty) \times \{0, 1\}$ and $X^2 = \cup_{0 \le k < \infty}[0, \infty)^k$ with $[0, \infty)^0 = \emptyset$. The second coordinate of the state space $x^2 = (x^2(1), \dots, x^2(k)) \in X_2$ is the vector of workloads carried by arrivals in a batch with $k = 0, 1, \dots$ items. In particular, $x_n^2 = \emptyset$ means either that there is no arrival at the n-th decision epoch. Corollary 3.2 implies that for any policy π that uses the described additional information there exists a policy σ with the following properties: (i) the expected average costs per unit time (3.7) incurred by σ and π are equal, and (ii) the current and past information about the states of the system that π knows is limited to workloads and the states of the servers. According to [8], D-policies are optimal among policies satisfying (ii). Therefore, D-policies are optimal for $M^X/G/1$ queues in the problem formulation considered in [8] even when the controller takes into an account the numbers of arriving jobs and their individual workloads.

SMDPs with iid sojourn times. Consider an SMDP in which the sojourn times τ_n do not depend on states and actions and form a sequence of nonnegative iid random variables. Let the costs c incurred during the first u units of time in state x_n, where $u \le \tau_n$, be nonnegative and satisfy the condition $c(x_n, a_n, u) \le C_1 + C_2 u$ for all $x_n \in X$, $a_n \in A$, where C_1 and C_2 are nonnegative finite constants. The function c is assumed to be measurable. Let $\bar{c}(x, a) = Ec(x, a, \tau_1)$ be the expected total reward until the jump if an action a is selected at a state x. We shall also assume that $0 < \bar{\tau} < \infty$, where $\bar{\tau} = E\tau_1$.

From an intuitive point of view, such an SMDP with average rewards per unit time is essentially an MDP and the knowledge of a real time parameter t is unimportant. We prove this fact by using Corollary 2.1.

Let $t_0 = 0$ and $t_{n+1} = t_n + \tau_n$, $n = 0, 1, \dots$. Consider the cost function U_t defined by (3.6) and the average costs per unit time defined in (3.7).

Since all sojourn times are iid, it is intuitively clear that the rewards do not depend on actual sojourn times. Our immediate goal is to prove that, for any initial distribution μ and for any policy π, the average cost per unit time $v(\mu, \pi)$ can be represented as

$$v(\mu, \pi) = \limsup_{n \to \infty} n^{-1} E_\mu^\pi \sum_{i=0}^{n-1} \bar{c}(x_{t_i}, a_{t_i})/\bar{\tau}. \tag{4.1}$$

To prove (4.1) we first rewrite it in the following form

$$v(\mu, \pi) = \limsup_{n \to \infty} (n\bar{\tau})^{-1} E_x^\pi U_{t_n}. \tag{4.2}$$

Second, we observe that

$$\limsup_{n \to \infty} (n\bar{\tau})^{-1} E_x^\pi U_{t_n} = \limsup_{n \to \infty} (n\bar{\tau})^{-1} E_x^\pi U_{n\bar{\tau}}. \tag{4.3}$$

To prove (4.3), we notice that

$$\frac{|E_\mu^\pi U_{t_n} - E_\mu^\pi U_{n\bar{\tau}}|}{n} \le C_1 \frac{E_\mu^\pi |N_\mu^\pi(n\bar{\tau}) - n|}{n} + C_2 \frac{E_\mu^\pi |t_n - n\bar{\tau}|}{n} \tag{4.4}$$

and the right hand side of (4.4) tends to 0 as $n \to \infty$. The first summand in the right hand side of (4.4) tends to 0 according to [12, Theorem 5.1, p. 54], [12, Theorem 1.1, p. 166], and the fact that a.s. convergence implies convergence in probability. The second summand tends to 0 according to [11, Lemma 13, p. 192]. Thus, (4.3) is proved.

We observe that

$$\limsup_{n \to \infty} (n\bar{\tau})^{-1} E_x^\pi U_{n\bar{\tau}} = \limsup_{t \to \infty} (\bar{\tau}[t/\bar{\tau}])^{-1} E_x^\pi U_{\bar{\tau}[t/\bar{\tau}]} = \limsup_{t \to \infty} t^{-1} E_x^\pi U_{\bar{\tau}[t/\bar{\tau}]}.$$

In addition,

$$0 \le t^{-1}[U_t - U_{\bar{\tau}[t/\bar{\tau}]}] \le t^{-1} C_1 (N(t) - N(t - \bar{\tau})) + C_2 \bar{\tau}/t. \tag{4.5}$$

By taking the expectation in (4.5), setting $t \to \infty$, and applying the renewal theorem, we obtain the equality

$$v(\mu, \pi) = \limsup_{t \to \infty} t^{-1} E_x^\pi U_{\bar{\tau}[t/\bar{\tau}]}.$$

This equality, $t^{-1}\bar{\tau}[t/\bar{\tau}] \to 1$, and (4.3) imply (4.2). Thus, (4.1) is proved.

We consider this CTSDP as an SDP with the state space $X^1 \times X^2$, where $X^1 = X$ and $X^2 = [0, \infty)$. The time parameter $t \in X^2$ affects neither the transition probabilities between states in X^1 nor the objective criterion w. The latter follows from (4.1). Therefore, in view of Corollary 2.1, the policies that do not use the information about sojourn times $\tau_0, \tau_1, \dots$ are as good as policies that use this information.

We remark that the assumption that $c(x_n, a_n, u) \le C_1 + C_2 u$, where C_1 and C_2 are constants, for SMDPs with iid sojourn times is similar to the assumption that costs are

bounded in discrete-time MDPs. The case of unbounded costs is important but we do not study it in this paper.

Uniformized Continuous-Time Markov Decision Processes (CTMDPs). A CTMDP is an SMDP with exponential sojourn times independent from the states the system jumps to. In other words, $q(d\tau_n dx_{n+1}|x_n, a_n) = \lambda(x_n, a_n)p(dx_{n+1}|x_n, a_n)$, where (i) $0 \le \lambda(x, a) < K$ for all $x \in X$, $a \in A$, and for some $K < \infty$, and (ii) p is a transition kernel from $X \times A$ into A with the property $p(x|x, a) = 0$ for all $x \in X$. The system incurs two types of costs: (i) the instant costs $c(x_n, a_n, x_{n+1})$ when the system jumps from state x_n to state x_{n+1} and the control a_n is used, and (ii) the continuous costs $C(x_n, a_n)$ incurred per unit time in state x_n if the control a_n is chosen. For simplicity, we assume that the functions c and C are nonnegative and bounded. In addition, we assume that these functions are measurable. Though for CTMDPs it is possible to consider policies that change actions between jumps (see [7], [15], [16]), we do not do it here for the sake of simplicity. In fact, according to the terminology in [7], CTJMDPs considered here are ESMDPs (exponential SMDPs or, more precisely, SMDPs with exponential sojourn times).

Uniformization (see Lippman [17] or monographs [2], [19], [20]) introduces fictitious jumps from states x_n into themselves with intensities $(K - \lambda(x_n, a_n))$. This reduces a CTMDP with jump intensities bounded above by K to a CTMDP with sojourn times being iid exponential random variable with the intensity K. The above results on CTSDPs with iid sojourn times imply that the controller does not benefit from the knowledge of sojourn times in the uniformized CTMDP. Therefore, for the uniformized CTMDP, it is possible to restrict the set of all policies to the policies that do not use any information about sojourn times of the uniformized process. This completes the reduction of the uniformized CTMDP to the corresponding discrete time MDP.

If there is a stationary optimal policy for the corresponding discrete time MDP, this policy is optimal for the original CTMDP. This follows from the following three observations: (i) this stationary policy is optimal for the uniformized CTJMDP, (ii) this stationary policy can be implemented in the original CTMDP, and (iii) any policy in the original CTMDP can be implemented in the uniformized CTMDP.

We remark that the results on the reduction of continuous-time models to discrete time hold also for discounted total rewards. We concentrate on average costs per unit times in this paper because this is a more difficult case than discounting. Though uniformization can also be applied to discounted costs [4, p. 432], discounted CTMDPs and discounted SMDPs can be directly reduced to discrete time discounted MDPs without using uniformization; see [7].

Admission control. Consider a finite queue with a renewal process of arrivals. If this queue contains n customers, the departure time has an exponential distribution with the intensity μ_n. Arriving customers belong to different types. To simplify the problem formulation, suppose that there are m types of customers. A type i customer pays R_i for the service when the customer is admitted, $i = 1, \dots, m$. Let R_i have the cumulative distribution function F_i. The types of arriving customers are iid and do not depend on any other events associated with the system. Given the arrival's type, the possible payoff does not depend on any other events. The service intensity μ_n does not depend on the types of accepted customers.

An arrival can be either accepted or rejected when it is entering the system. If the system is full, the arrival is rejected. An arrival can also be rejected to maximize average rewards per unit time. The arrival type i and the amount R_i are known at the arrival epoch. The question is which arrivals should be rejected to maximize the average rewards per unit time?

By considering arrival epochs as decision epochs, it is easy to formulate this problem as an average reward SMDP with iid sojourn times equal to interarrival times. The state space is $X^1 \times X^2$, where X^1 is the set of pairs (n, r) with n equal to the number of customers that an arrival sees in the system and with r equal to the amount that the arrival is willing to pay if admitted, and X^2 is the arrival type. We observe that transition probabilities do not depend on the type of an arriving customer.In addition, the reward function is $r = r(x^1, x^2) = r((n, r), m) = r$ and therefore the rewards do not depend on the second coordinate $x^2 = m$, which is the customer type.

Therefore, one can use the policies for this problem that do not take into account the customer type. In particular, Miller [18] and Feinberg and Reiman [9] studied Markovian problems when the type i customer payoff r_i is deterministic. Of course, it is natural to consider the situation when $r_i \neq r_j$ for $i \neq j$. However, if $r_i = r_j$ for $i \neq j$, we can merge these customer classes without loss of optimality. This follows from the above results for CTSDPs with iid sojourn times. The need to consider the problem with $r_i = r_j$ for $i \neq j$ appears in problems with multiple criteria and constraints. Even when different classes have different rewards, the method of Lagrangian multipliers may lead to the situation when different classes have equal rewards [6].

V. Conclusions

For discrete-time and continuous-time stochastic sequential decision processes, this paper proves the following natural and simple observation: if the states of the system consist of two coordinates and neither the transition mechanism nor the objective function depend on the second coordinate, this coordinate can me dropped. In other words, the controller cannot benefit from the knowledge of the irrelevant information.

We apply these general results to various particular problems of queueing control and to uniformization of continuous-time Markov Decision Processes and simplify these problems. In fact, this paper was motivated by these applications.

REFERENCES

[1] Balachandran, K. R. (1973). Control policies for a single server system. *Management. Sci.* **1**9:1013-1018.

[2] Bertsekas, D. P. (2001). *Dynamic Programming and Optimal Control*, Second Edition, Scientific, Belmont, MA.

[3] Bertsekas, D. P. and Shreve, S. E. (1978). *Stochastic Optimal Control: The Discrete-Time Case*, Academic Press, New York; republished by Athena Scientific, Belmont, MA 1997.

[4] Cassandras. C. G. (1993). *Discrete Event Systems: Modeling and Performance Analysis.* IRWIN, Boston.

[5] Dynkin, E. B. and Yushkevich, A. A. (1979). *Controlled Markov Processes.* Springer-Verlag, New York.

[6] Fan-Orzechowski, X. and Feinberg, E.A. (2005). Optimal Admission Control for a Markovian Queue Under the Quality of Service Constraint. Submitted to 44th IEEE CDC-ECC Conference.

[7] Feinberg, E.A. (2004). Continuous-time discounted jump-Markov decision processes: a discrete-event approach. *Math. Oper. Res.* **2**9:492-524.

[8] Feinberg, E. A. and Kella, O. (2002). Optimality of D-policies for an $M/G/1$ queue. *Queueing Systems* **4**2:355-376.

[9] Feinberg, E. A. and Reiman, M. I. (1994). Optimality of randomized trunk reservation. *Probability in the Engineering and Informational Sciences* 8:463-489.

[10] Feinberg, E. A. and Shwartz, A., eds. (2002). *Handbook of Markov Decision Processes.* Kluwer, Boston.

[11] Fristedt, B. and Gray, L. (1997). *A Modern Approach to Probability Theory.* Birkhäuser, Boston.

[12] Gut, A. (1988). *Stopped Random Walks. Limit Theorems and Applications.* Springer-Verlag, New York.

[13] Hinderer, K. (1970). *Foundations of Non-Stationary Dynamic Programming with Discrete Time Paramerter*. Springer-Verlag, New York.

[14] Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory*, Mathematical Centre Tracts **51**, Amsterdam.

[15] Kitayev, M. Yu. (1985). Semi-Markov and jump Markov controlled models: average cost criterion. SIAM Theory Probab. Appl. 30:272-288.

[16] Kitayev, M. Yu. and Rykov, V. V. (1995). *Controlled Queueing Systems*, CRC Press, New York.

[17] Lippman, S. A. (1975). Applying a New Device in the Optimization of Exponential Queueing Systems. *Oper. Res.* **2**3:687-710.

[18] MILLER, B. L. (1969). A queueing reward system with several customer classes. *Management. Sci.* **1**6:235-245.

[19] Puterman, M. L. (1994). *Markov Decision Processes.* John Wiley, New York.

[20] Senott, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley, New York.