Mathematical Methods in Operations Research 2005
Special issue in honor of *Arie Hordijk*

ON ESSENTIAL INFORMATION IN SEQUENTIAL DECISION PROCESSES

Eugene A. Feinberg

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

Received: January 2005 / Revised version: April 2005

Abstract This paper provides sufficient conditions when certain information about the past of a stochastic decision processes can be ignored by a controller. 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.

1 Introduction

The results of this paper are based on the following simple observation. If each state of a controlled stochastic system consists of two coordinates and neither the transition mechanism for the first coordinate nor costs depend on the second coordinate, the controller can ignore the second coordinate values. Theorem 2 and Corollary 4 present the appropriate formulations for discrete-time and for continuous-time jump problems respectively and for various performance criteria including average costs per unit time and total discounted costs. These statements indicate that additional information, presented in the second coordinate, cannot be used to improve the system performance. Though these facts are simple and general, they are useful for the analysis of various particular problems.

This paper is motivated by two groups of applications: (i) controlled queues and (ii) uniformization of Continuous-Time Markov Decision Processes (CTMDP). We illustrate our motivation in the introduction with one of such applications. Additional examples are presented in Section 4.

Consider the problem of routing to parallel queues with the known workload; see the last example "Policies based on queue length and workload" in

** Supported in part by grant DMI-0300121 from the National Science Foundation

Koole [17]. Customers arrive according to a Poisson process into a system consisting of m homogeneous restless queues. The service times of arriving customers are not known at the arrival epochs but the state of the system, that an arrival sees, is known. This state includes the workloads in all m queues. The costs depend only on the workload vector. In particular, the m-dimensional vectors of workloads and numbers of customers in queues are known. We denote them by w and ℓ respectively. Since nothing depends on ℓ, this coordinate was dropped in [17] and the state of the system was reduced to the the workload vector w.

If the service discipline is FIFO or all m queues use deterministic service disciplines, then the sufficiency of using w as the state space follows from the following simple arguments. Let W_{n-1} be the workload vector that the nth arrival sees, $n = 0, 1, \dots$, and W_0 is the null vector. Then the sequence $W_1, \dots, W_n, \dots$ defines the numbers of customers in each of m queues. For the finite-horizon discounted cost criterion studied in [17], Markov policies are optimal. Therefore, the numbers of customers in queues can be ignored by optimal policies.

If the service discipline selects customers randomly, the sequence $W_1, W_2, \dots$ does not define the numbers of customers in queues and the above arguments fail. Consider the following example. Let $m = 1$. The service is non-preemptive and the server selects customers from the queue with equal probabilities. Let an nth arrival see a queue with 2 customers of which one is at the server. The remaining service time of the customer at the server is 1 (everything is in minutes), and the service time of the waiting customer is 3. Thus, the total workload is 4. Let the $(n+1)$st customer arrives in 3 minutes and sees the workload 2. This implies that the service time of the nth arrival is 1. However, the number of customers that $(n+1)$st arrival sees can be either 1 or 2.

The results of this paper imply that even for randomized service disciplines, it is indeed sufficient to know only the workload vector w in the problem considered in [17]. Consider the larger state z which is the set of m strings $(\ell_i, w_{i,1}, \dots, w_{i,\ell_i})$, $i = 1, \dots, m$, where ℓ_i is the number of customers in queue i and $w_{i,1}, \dots, w_{i,\ell_i}$ are their remaining service times. We notice that $(w_{i,1}, \dots, w_{i,\ell_i}) = \emptyset$ when $\ell_i = 0$. Given any service discipline for each of these m queues, it is easy to construct a Markov Decision Process (MDP) with states z for this problem. Each state z defines the workload vector and the numbers of customers in queues as appropriate sums. Therefore, the state z can be rewritten as (w, ℓ, z) and the transition probabilities relevant to w and the costs do not depend on (ℓ, z). Therefore, in view of Theorem 2 and Corollary 4, the coordinates ℓ and z can be dropped and it is indeed sufficient to know only the vector of workloads w.

Now consider the version of same problem from [17] when customers arrive in batches. All the customers in a batch should be directed to the same server. The arguments based on the state expansion to (w, ℓ, z) and on using Theorem 2 and Corollary 4 imply that it is sufficient to know the workload vector w and the queue sizes ℓ can be ignored. The arguments

based on the history of observed workload vectors $W_1, W_2, \ldots$ do not work because this sequence does not define the numbers of customers in queues.

This paper is organized in the following way. Section 2 deals with discrete-time problems and Section 3 deals with continuous-time jump problems. In both cases we consider controlled non-Markovian systems because the proofs are the same as for MDPs. We show that if neither the transition mechanism for the first coordinate nor costs depend on the second coordinate, this coordinate can be dropped and therefore the problem can be simplified.

Section 4 deals with additional applications to queueing control and to uniformization, the procedure that reduces CTMDPs to MDPs by using fictitious jumps. For such a reduction, uniformization was introduced by Lippman [18]. It holds for the class of stationary policies [2,4,20,21]. For general policies, durations of fictitious jumps add information to the system. We show that, for problems with arbitrary state spaces and bounded one-step costs, this additional information can be ignored for the average costs per unit time criterion if stationary optimal or nearly optimal policies exist for the corresponding discrete-time MDP.

2 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 the Borel σ-fields on H_n, $n = 0, 1, \ldots, \infty$, and these σ-fields generate the 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 $p(\cdot|h, a)$ is a probability measure on A for any pair (h, a), where $h \in H$ and $a \in A$.

A policy is defined as a regular transition probability from H to A. Therefore, a policy 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 policy π define a unique probability measure P_μ^π on H_∞. Following [5], we shall call P_μ^π a strategic measure.

A criterion v is defined as a numerical function of a strategic measure, $v(\mu, \pi) = v(P_\mu^\pi)$. If p is just a function of (x_n, a_n), the defined SDP becomes an MDP.

The expected total discounted costs and the average costs per unit time are two important criteria studied in the literature. Expected total dis-

counted costs can be represented in the form of

$$v = E^{\pi}_{\mu} U(h_{\infty}), \tag{1}$$

where $h_{\infty} \in H_{\infty}$ and U is a measurable function on H_{∞}. Indeed, let $c(x,a) \geq 0$ be one-step costs. Then formula (1) with $U(h_{\infty}) = \sum_{i=0}^{\infty} \beta^i c(x_i, a_i)$ defines the total expected discounted costs with the discount factor β.

Average costs per unit time can be represented as

$$v(\mu, \pi) = \limsup_{n\to\infty} E^{\pi}_{\mu} U_n(h_{\infty}) \tag{2}$$

with $U_n(h_{\infty}) = n^{-1} \sum_{i=0}^{n-1} c(x_i, a_i)$. We observe that the representation (2) is more general than (1) because (2) becomes (1) when $U_n = U$. In addition to average costs per unit time and to total discounted costs, a large variety of other criteria can be presented in the forms of (1) and (2).

We remark that it is natural to consider problems in which actions sets depend on the current state or even on the past history, [3,5,10,13,14,20, 21]. We do not do 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 $n = 0, 1, \ldots$, transition probabilities on X^1 do not depend on the second component of the state space. In other words,

$$p(dx^1_{n+1} | x^1_0, x^2_0, a_0, \ldots, x^1_n, x^2_n, a_n) = p(dx^1_{n+1} | x^1_0, a_0, \ldots, x^1_n, a_n). \tag{3}$$

For any probability measure P on H_{∞}, consider its projection $\bar{P}$ on $(X_1 \times A)^{\infty}$. The measure $\bar{P}$ is defined by the following probabilities defined on cylinders

$$P(X^1_0 \times A_0 \times \cdots \times X^1_n \times A_n) = P(X^1_0 \times X^2 \times A_0 \times \cdots \times X^1_n \times X^2 \times A_n),$$

where $n = 1, 2, \ldots$, and X^1_i and A^1_i are measurable subsets of X^1 and A respectively, $i = 0, \ldots, n$. For a strategic measure P^{π}_{μ} we note its projection on $(X_1 \times A)^{\infty}$ by $\bar{P}^{\pi}_{\mu}$. Consider the following assumption on the criterion v.

Assumption 1 *$v(\mu, \pi) = v(\mu, \pi')$ for any two policies π and π' such that $\bar{P}^{\pi}_{\mu} = \bar{P}^{\pi'}_{\mu}$.*

If Assumption 1 holds then $v(\mu, \pi) = f(\bar{P}^{\pi}_{\mu})$ for some function f defined on the set of all probability measures on $(X^1 \times A)^{\infty}$. For example, if v is defined by (2) with $U_n(h_{\infty}) = U_n(x^1_0, a_0, x^1_1, a_1, \ldots)$, then Assumption 1 holds. If one-step costs c do not depend on x^2, the average cost per unit time and total discounted cost criteria satisfy Assumption 1.

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 $\tilde{P}^{\sigma}_{\mu^1}$ be a strategic measure for this smaller SDP, where μ^1 is an initial probability distribution on X^1 and σ is a policy in the smaller model. Every $\tilde{P}^{\sigma}_{\mu^1}$ is a probability measure on the space $(X^1 \times A)^{\infty}$. Let $\tilde{v}(\mu_1, \sigma)$ be a criterion for this SDP. If Assumption 1 holds, we set $\tilde{v}(\mu_1, \sigma) = f(\tilde{P}^{\sigma}_{\mu_1})$.

Theorem 2 *Consider an SDP with the state space $X = X^1 \times X^2$ and let assumption (3) 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 $n = 0, 1, \dots$ (P^π_μ-a.s.) by*

$$\sigma(da_n|x^1_0a_0x^1_1a_1\dots x^1_n) = \frac{P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_n)}{P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_n)}. \tag{4}$$

Then (i)

$$P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots) = \tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots),$$

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 $(X^1 \times A)^\infty$.

(ii) If, in addition, Assumption 1 holds then $\tilde{v}(\mu_1, \sigma) = v(\mu, \pi)$.

Proof By Kolmogorov's extension theorem, to verify (i) it is sufficient to prove that for any $n = 0, 1, \dots$

$$P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_n) = \tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots dx^1_n). \tag{5}$$

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 (5) hold for some n. Then

$$\begin{aligned}\tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_n) &= \tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots dx^1_n)\sigma(da_n|x^1_0a_0x^1_1a_1\dots x^1_n)\\ &= P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_n),\end{aligned} \tag{6}$$

where the first equality follows from the definition of a strategic measure and the second equality follows from (4) and (5). Since the transition probabilities in the first model do not depend on x^2, we have

$$\begin{aligned}&\tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_ndx^1_{n+1})\\ &= \tilde{P}^\sigma_{\mu^1}(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_n)p(dx^1_{n+1}|x^1_0, a_0, \dots, x^1_n, a_n)\\ &= P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_n)p(dx^1_{n+1}|x^1_0, a_0, \dots, x^1_n, a_n)\\ &= P^\pi_\mu(dx^1_0da_0dx^1_1da_1\dots dx^1_nda_ndx^1_{n+1}),\end{aligned} \tag{7}$$

where the first equality follows from the definition of the strategic measure $\tilde{P}^\sigma_{\mu^1}$ and the second equality follows from (6). Statement (i) yields (ii). ■

3 Continuous-time jump problems

In the defined SDP, all time intervals between decisions equal 1. In this section, we extend Theorem 2 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 $x_0, a_0, \tau_0, x_1, a_1, \tau_1, \dots$, 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 $q(d\tau_n, dx_{n+1}|x_0, a_0, \tau_0, x_1, a_1, \dots, \tau_{n-1}, x_n, a_n)$. For this SDP, 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 $q(d\tau_n, dx_{n+1}|x_n, a_n)$.

Since a CTSDP is defined via the corresponding SDP, we have policies and strategic measure also defined for the CTSDP. The objective criterion v is a function of a strategic measure P^π_μ for this CTSDP, i.e. $v(\mu, \pi) = v(P^\pi_\mu)$.

For each $t \ge 0$, consider a measurable function $U_t(h_\infty)$, where $h_\infty = x_0, a_0, \tau_0, x_1, a_1, \tau_1, \dots$. Consider the criterion

$$v(\mu, \pi) = \limsup_{t\to\infty} E^\pi_\mu U_t(h_\infty). \tag{8}$$

The expected average cost per unit time is an important example of a criterion that can be presented in the form (8). Let $c(x, a, t) \ge 0$ 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 $t_{n+1} = t_n + \tau_n$, $n = 0, 1, \dots$. We set $N(t) = \sup\{n = 0, 1, \dots | t_n \le t\}$. The average cost up to time t is $U_t(h_\infty) = L_t/t$, where L_t is the cost up to time t,

$$L_t = \sum_{n=0}^{N(t)-1} c(x_n, a_n, \tau_n) + c(x_{N(t)}, a_{N(t)}, t - t_{N(t)}), \tag{9}$$

$h_\infty = x_0, a_0, \tau_0, x_1, a_1, \tau_1, \dots$. Then the expected average cost per unit time is defined by (8). The expected total discounted costs can also be represented in the form of (8) because they can be presented in the form of (1) with $h_\infty = x_0, a_0, \tau_0, x_1, a_1, \tau_1, \dots$. In particular, for the expected total discounted costs

$$U(h_\infty) = \sum_{n=0}^{\infty} e^{-\gamma \sum_{i=0}^{n-1} \tau_i} \int_0^{\tau_n} c(x_n, a_n, t) e^{-\gamma t} dt,$$

where $\gamma > 0$ is the discount rate. Note that (1) is a particular form of (8) with $U_t = U$.

Similarly to the discrete time case, consider a CTSDP 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^2_i, $i = 0, 1, \ldots, n$, i.e.

$$\begin{aligned} &q(d\tau_n, dx^1_{n+1}|x^1_0, x^2_0, a_0, \tau_0, x^1_1, x^2_1, a_1, \tau_1, \ldots, x^1_n, x^2_n, a_n) \\ &\quad = q(d\tau_n, dx^1_{n+1}|x^1_0, a_0, \tau_0, x^1_1, a_1, \tau_1, \ldots, x^1_n, a_n). \end{aligned} \tag{10}$$

Similarly to the discrete-time case, we denote by $\bar{P}$ the projection on $(X^1 \times A \times [0,\infty))^\infty$ of a probability measure P defined on $(X^1 \times A \times [0,\infty))^\infty$. For a strategic measure P^π_μ we denote the corresponding projection by $\bar{P}^\pi_\mu$. The following assumption is similar to Assumption 1.

Assumption 3 *$v(\mu, \pi) = f(\bar{P}^\pi_\mu)$ for some function f defined on the set of probability measures on $(X^1 \times A \times [0,\infty))^\infty$.*

For example, Assumption 3 holds for the criterion (8) when $U_t(h_\infty) = U_t(x^1_0, a_0, \tau_0, x^1_1, a_1, \tau_1, \ldots)$.

Similarly to the discrete time case, we consider a smaller CTSDP with the state space X^1, action space A, and transition kernel $q(d\tau_n, dx^1_{n+1}|x^1_0, a_0, \tau_0, x^1_1, a_1, \tau_1, \ldots, x^1_n, a_n)$. Let $\tilde{P}^\sigma_{\mu^1}$ be a strategic measure for this smaller CTSDP, where μ^1 is an initial probability distribution on X^1 and σ is a policy in the smaller model. Every $\tilde{P}^\sigma_{\mu^1}$ is a probability measure on the space $(X^1 \times A \times [0,\infty))^\infty$. When Assumption 3 holds, we consider the criterion $\tilde{v}(\mu^1, \sigma) = f(\tilde{P}^\sigma_{\mu^1})$. Theorem 2 implies the similar result for CTSDPs.

Corollary 4 *Consider a CTSDP with the state space $X = X^1 \times X^2$ and let assumption (10) 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 for all $n = 0, 1, \ldots$ (P^π_μ-a.s.) by*

$$\sigma(da_n|x^1_0 a_0 \tau_0 x^1_1 a_1 \tau_1 \ldots x^1_n) = \frac{P^\pi_\mu(dx^1_0 da_0 d\tau_0 dx^1_1 da_1 d\tau_1 \ldots dx^1_n da_n)}{P^\pi_\mu(dx^1_0 da_0 d\tau_0 dx^1_1 da_1 d\tau_1 \ldots dx^1_n)}, \tag{11}$$

$n = 0, 1, \ldots$. Then (i)

$$P^\pi_\mu(dx^1_0 da_0 d\tau_0 dx^1_1 da_1 d\tau_1 \ldots) = \tilde{P}^\sigma_{\mu^1}(dx^1_0 da_0 d\tau_0 dx^1_1 da_1 d\tau_1 \ldots), \tag{12}$$

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, $\tilde{P}^\pi_{\mu^1}$ is the projection of the strategic measure P^π_μ on $(X^1 \times A \times [0,\infty))^\infty$.

(ii) If, in addition, Assumption 3 holds then $\tilde{v}(\mu_1, \sigma) = v(\mu, \pi)$.

4 Examples of applications

In addition to the example considered in the Introduction, in this Ssection we apply Theorem 2 and Corollary 4 to the $M^X/G/1$ queue with a removable server and known workload, to the admission control, to SMDPs with iid sojourn times, and to uniformization of CTMDPs.

$M^X/G/1$ **queue with a removable server**. 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. The server can be turned on and off and switching costs are positive. The holding costs depend only on the workload.

Control of queues with the removable server and known workload has been studied in the literature since 1973 when Balanchandran [1] introduced the 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 for average costs per unit time 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.

Now we consider the situation when customers arrived in batches. The question is whether the information about the numbers of customer in batches and individual service times is useful? The answer is that this information is useless and can be ignored. Therefore, D-policies are optimal for $M^X/G/1$ queues under the conditions formulated in [8].

The correctness of this answer follows from the following arguments. As was shown in [8, Section 3], it is sufficient to make switching decisions only at the arrival epochs and at the epochs when the system becomes empty. Therefore, we can consider an SMDP when decisions are selected only at these epochs. The state of this SMDP is (w, g), where w is the workload and d is the state of the server (on or off). We enlarge the state of the system by adding the coordinate $z = (j, s_1, \dots, s_n)$ containing the number of customers j in the arrived batch and the customer service times. Then neither transitions of (w, g) nor costs depend on z and, in view of Corollary 4, the coordinate z can be dropped from the state description. Thus, D-policies are also optimal for $M^X/G/1$ queues.

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. Suppose that there are m types of customers. A type i customer pays r_i for the service when the customer is admitted, $i = 1, \dots, m$. The types of arriving customers are iid and do not depend on any other events associated with the system. 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's type

becomes 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 on X^1 do not depend on the type of an arrived 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. In view of Corollary 4, the information regarding the arrival type can be ignored if customer's payoff r is known. Therefore, if $r_i = r_j$ for type i and j customers, the customers of these types can be merged into one type of customers. Therefore, the number of different customer types can be reduced to the number of different payoffs r_i.

In fact, it is natural to assume that $r_i \neq r_j$ when $i \neq j$. Miller [19] and Feinberg and Reiman [9] used this assumption. The need to consider the problem with $r_i = r_j$ for $i \neq j$ appears in cases of 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; see [6] for details.

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 $u \leq \tau_n$, be nonnegative and satisfy the condition $c(x_n, a_n, u) \leq C_1 + C_2 u$ 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 cost 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 costs per unit time is essentially an MDP and the knowledge of a real time parameter t is unimportant. We prove this fact by using Theorem 2.

Let $t_0 = 0$ and $t_{n+1} = t_n + \tau_n$, $n = 0, 1, \ldots$. Consider the total cost L_t up to time t defined in (9) and the expected average costs per unit time $v(\mu, \pi)$ defined by (8) with $U_t(h_\infty) = L_t/t$.

Since all sojourn times are iid, it is intuitively clear that the costs do not depend on actual sojourn times. Our immediate goal is to prove that for any initial distribution μ and for any policy π

$$v(\mu, \pi) = \limsup_{n\to\infty} n^{-1} E^\pi_\mu \sum_{i=0}^{n-1} \bar{c}(x_{t_i}, a_{t_i})/\bar{\tau}. \tag{13}$$

To prove (13) we first rewrite it,

$$v(\mu, \pi) = \limsup_{n\to\infty} (n\bar{\tau})^{-1} E^\pi_x L_{t_n}. \tag{14}$$

Second, we observe that

$$\limsup_{n\to\infty}(n\bar{\tau})^{-1}E_x^\pi L_{t_n} = \limsup_{n\to\infty}(n\bar{\tau})^{-1}E_x^\pi L_{n\bar{\tau}}. \tag{15}$$

To prove (15), we notice that $N(t)$ is a renewal process and

$$\frac{|E_\mu^\pi L_{t_n} - E_\mu^\pi L_{n\bar{\tau}}|}{n} \le C_1\frac{E|N(n\bar{\tau})-n|}{n} + C_2\frac{E|t_n - n\bar{\tau}|}{n} \tag{16}$$

and the right hand side of (16) tends to 0 as $n\to\infty$. The first summand in the right hand side of (16) tends to 0 according to [12, Theorem 5.1, p. 54, and 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, (15) is proved.

We observe that

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

In view of (15), the last line of equalities implies

$$\limsup_{t\to\infty} t^{-1}E_x^\pi L_{\bar{\tau}[t/\bar{\tau}]} = \limsup_{n\to\infty}(n\bar{\tau})^{-1}E_x^\pi L_{t_n} \tag{17}$$

In addition,

$$0 \le t^{-1}[L_t - L_{\bar{\tau}[t/\bar{\tau}]}] \le t^{-1}C_1(N(t) - N(t-\bar{\tau})) + C_2\bar{\tau}/t. \tag{18}$$

By taking the expectation in (18), setting $t\to\infty$, and applying the renewal theorem, we obtain the equality

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

This equality, (17), and (15) imply (14). Thus, (13) is proved.

We consider this SMDP as an MDP 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 v. The latter follows from (13). Therefore, in view of Theorem 2, 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. Consider the MDP with the same state and action sets as the given SMDP, with the same transition probabilities, and with one-step costs $\bar{c}(x,a)/\tau$. For average costs per unit time, this MDP has the same value function as the original SMDP. In addition, stationary optimal policies for this MDP are optimal for the original SMDP.

We remark that the assumption that $c(x_n,a_n,u)\le C_1 + C_2u$, 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.

Uniformization of Continuous-Time Markov Decision Processes (CTMDPs). A CTMDP is an SMDP with exponential sojourn times

and with transition probabilities that do not depend on these times. 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 cost rates $C(x_n, a_n)$ 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], CTMDPs considered here are ESMDPs (exponential SMDPs or, more precisely, SMDPs with exponential sojourn times).

Uniformization (see Lippman [18] or monographs [2,20,21]) introduces fictitious zero-cost 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 SMDP with sojourn times being iid exponential random variables with the intensity K. The above results on SMDPs with iid sojourn times imply that the controller does not benefit from the knowledge of sojourn times in the uniformized SMDP and the problem can be reduced to an MDP. If this MDP has a stationary optimal policy, this policy is optimal for the original CTMDP. This justifies uniformization for nonstationary policies and without using the fact that there is a stationary optimal policy for the original CTMDP.

We provide additional explanations. Let v be the expected average cost per unit time in the original CTMDP and v^1 and v^2 be the similar criteria for the uniformized SMDP and in the corresponding MDP respectively. This MDP has the same states and actions as the original CTMDP, the transition probabilities $\tilde{p}(y|x, a) = \lambda(x, a)p(y|x, a)/K$ when $y \ne x$, and $\tilde{p}(x|x, a) = 1 - \lambda(x, a)/K$. The one-step costs are $\tilde{c}(x, a) = C(x, a) + \sum_{y \ne x} c(x, a, y)\lambda(x, a)p(y|x, a)$.

Since any policy π for the original CTMDP can be implemented in the uniformized SMDP, $v^1(\mu, \pi) = v(\mu, \pi)$. According to the results on SMDPs with iid sojourn times, there exists a policy σ in the MDP such that $v^2(\mu, \sigma) = v^1(\mu, \pi)$. Thus, if for the MDP there exists a stationary optimal policy, this policy is optimal for the original CTMDP. This is also true for problems with multiple criteria and constraints with a fixed initial distribution μ; see [7] for the constrained problem definition. However, for constrained discrete-time problems, an optimal policy is typically randomized stationary. For an optimal randomized stationary policy φ in the described above MDP, it is possible to construct a policy ψ in the original CTMDP such that $v(\mu, \psi) = v^2(\mu, \varphi)$ and therefore the policy ψ is optimal for the original CTMDP.

We remark that the reduction of continuous-time models to discrete time, by using uniformization, holds also for discounted total costs [18,

4]. However, discounted CTMDPs and discounted SMDPs can be directly reduced to discrete time discounted MDPs without using uniformization; see [7].

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. (2004). Optimality of Randomized Trunk Reservation for a Problem with a Single Constraint. Department of Applied Mathematics and Statistics, SUNY at Stony Brook, http://www.ams.sunysb.edu/∼feinberg/public/FanFeinberg1.pdf .
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. Koole, G. (2205). Routing to parallel homogeneous queues, *Mathematical Methods of Operations Research*, this issue.
18. Lippman, S. A. (1975). Applying a New Device in the Optimization of Exponential Queueing Systems. *Oper. Res.* **2**3:687-710.
19. Miller, B. L. (1969). A queueing reward system with several customer classes. *Management. Sci.* **1**6:235-245.
20. Puterman, M. L. (1994). *Markov Decision Processes.* John Wiley, New York.
21. Sennott, L. I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley, New York.