Adv. Appl. Prob. **18,** 341–359 (1986)
Printed in N. Ireland

TIME-AVERAGE OPTIMAL CONSTRAINED SEMI-MARKOV DECISION PROCESSES

FREDERICK J. BEUTLER,* *The University of Michigan, Ann Arbor*
KEITH W. ROSS,** *University of Pennsylvania*

Abstract

Optimal causal policies maximizing the time-average reward over a semi-Markov decision process (SMDP), subject to a hard constraint on a time-average cost, are considered. Rewards and costs depend on the state and action, and contain running as well as switching components. It is supposed that the state space of the SMDP is finite, and the action space compact metric. The policy determines an action at each transition point of the SMDP.

Under an accessibility hypothesis, several notions of time average are equivalent. A Lagrange multiplier formulation involving a dynamic programming equation is utilized to relate the constrained optimization to an unconstrained optimization parametrized by the multiplier. This approach leads to a proof for the existence of a semi-simple optimal constrained policy. That is, there is at most one state for which the action is randomized between two possibilities; at all other states, an action is uniquely chosen for each state. Affine forms for the rewards, costs and transition probabilities further reduce the optimal constrained policy to 'almost bang-bang' form, in which the optimal policy is not randomized, and is bang-bang except perhaps at one state. Under the same assumptions, one can alternatively find an optimal constrained policy that is strictly bang-bang, but may be randomized at one state. Application is made to flow control of a birth-and-death process (e.g., an $M/M/s$ queue); under certain monotonicity restrictions on the reward and cost structure the preceding results apply, and in addition there is a simple acceptance region.

CONTROLLED SEMI-MARKOV PROCESSES; DYNAMIC PROGRAMMING; CONSTRAINED OPTIMIZATION; BANG-BANG CONTROL; LAGRANGE MULTIPLIER

1. Introduction

In an earlier work the authors analyzed discrete-time optimal policies for controlled Markov chains [3], the objective being to maximize the average reward by a choice of policy that also met a specified average cost constraint.

Received 21 September 1984; revision received 18 April 1985.

* Postal address: Computer, Information and Control Engineering Program, The University of Michigan, Ann Arbor, MI 48109, USA.

** Postal address: Systems Engineering Department, Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104, USA.

In this paper, we extend our results to a class of continuous-time processes, namely semi-Markov decision processes (hereafter designated SMDP).

We shall study the dynamic maximization of the long-term average reward, subject to a global average-cost constraint, by Lagrangian multiplier techniques. We assume a finite state space, a compact action space, continuity of probabilities and rewards respective to the actions, plus an accessibility condition. At each jump, a control is chosen to parametrize the conditional probabilities for the next jump and sojourn interval. Under reasonable hypotheses, this structure leads to the existence of a constrained optimal policy. The optimal policy is always stationary. It is either non-randomized (i.e., simple), or acts like a simple policy except at one state; whenever the system is in that state, one chooses independently between two actions by the toss of a (biased) coin.

Proof of the above occupies most of Section 2. In Section 3, we introduce linearity assumptions on the reward, cost and transition probabilities. There is then a simple (non-randomized) policy that is optimum under the constraint; moreover, this policy is 'almost bang-bang' in the sense of being bang-bang except possibly at a single state. An alternative optimal policy is bang-bang, but randomization at one state may be necessary. Section 4 presents a further specialization to flow control of an exponential queue (such as an $M/M/s$ queue) for which the linearity assumptions are met. Under natural monotonicity restrictions on the reward and cost structure, there is a flow control acceptance threshold that is easy to calculate and implement.

It may not be immediately obvious why SMDP [15], [19], [7], rather than controlled Markov processes, are of primary interest. After all, Markov processes are often viewed as the most suitable model for the study of communication networks [10], computer operating systems [11], [18], and queueing systems [9]. Indeed, the control of Markov processes with bounded transition rates has been treated by a device (see [4], Section 8.4) that leads back to the simpler case of discrete-time optimization (as in [8], [17], [20], [19]). However, constrained optimization usually calls for a randomized action, which means that the conditional sojourn-time distribution is a mixture of exponentials. Thus, the application of randomized actions destroys the Markov property, and leads in a natural fashion to a semi-Markov process.

Basic to our consideration is an SMDP $\{(X_k, A_k, \tau_k)\}$ where X_k, A_k and τ_k are the state, action and sojourn time for epoch k, respectively. Therefore, if we take $T_0 = 0$ and

$$T_n \triangleq \sum_{k=0}^{n-1} \tau_k, \tag{1.1}$$

then the underlying continuous-time state process $\{Y(t)\}$ is given by

$$Y(t) = X_n \quad \text{for} \quad T_n \leqq t < T_{n+1}.$$

The *action* which is chosen at each jump instant T_k determines the joint distribution of τ_k and X_{k+1}, given X_k. An action a belongs to the *action space* $\mathscr{A}$, which is a compact metric space equipped with a σ-algebra generated by its open sets. To exhibit how the action influences the probabilistic law of the SMDP, first let

$$H_n = (X_0, A_0, \tau_0, \cdots, X_n, A_n, \tau_n) \tag{1.2}$$

be a *history* $\{(X_k, \tau_k)\}$. An action then parametrizes the transition probability in accordance with

$$\begin{aligned} &P[X_{n+1} = y,\ \tau_n \leqq t \mid H_{n-1}, X_n = x, A_n = a] \\ &\qquad = P[X_{n+1} = y,\ \tau_n \leqq t \mid X_n = x, A_n = a]; \end{aligned} \tag{1.3}$$

in other words, the controlled transition probability depends on the past only through the applied action. The right side of (1.3) is called the *controlled semi-Markov kernel,* and is denoted by

$$Q(x, y, t; a) \triangleq P[X_{n+1} = y,\ \tau_n \leqq t \mid X_n = x, A_n = a]. \tag{1.4}$$

It is required that this kernel is a conditional distribution on t in the wide sense (see [5], p. 29 ff.). It is also convenient to introduce the *law of motion* which is defined by

$$P_{xy}(a) \triangleq Q(x, y, \infty, a) = P[X_{n+1} = y \mid X_n = x, A_n = a].$$

The choice of actions is determined by a *policy*. In general, the policy $\boldsymbol{u}$ in the *policy space* (designated U) can be described as $\boldsymbol{u} = \{u_0, u_1, \cdots\}$, where u_k is applied at epoch k. Specifically, $u_k(\cdot \mid X_k = x, H_{k-1})$ is a conditional probability measure in the wide sense over $\mathscr{A}$; this measure is applied to the kernel to yield the conditional joint probability for X_{k+1} and τ_k, given X_k and H_k.

There are interesting subspaces of U that are not only easier to analyze, but represent policies easier to implement. In particular, a *stationary policy* $\boldsymbol{f}$ in space F is constituted by a probability measure over $\mathscr{A}$ that is conditioned only on the preceding state, and does not depend on the epoch index. More explicitly, we write $m_f(\cdot \mid x)$ to represent the stationary policy $\boldsymbol{f}$; for any measurable set $B \subset \mathscr{A}$

$$m_f(B \mid x) = P_f[A_k \in B \mid X_k = x] = u_k(B \mid X_k = x, H_{k-1}). \tag{1.5}$$

With the definition of a stationary policy in place, we now let

$$Q_f(x, y, t) \triangleq \int_{\mathscr{A}} Q(x, y, t; a) m_f(da \mid x) \tag{1.6}$$

and

$$P_{xy}(\boldsymbol{f}) \triangleq \int_{\mathscr{A}} P_{xy}(a) m_f(da \mid x). \tag{1.7}$$

It is seen from (1.4)–(1.7) that under a stationary policy $\boldsymbol{f}$, the SMDP $\{Y_t\}$ is a semi-Markov process with kernel $Q_f(x, y, t)$, and that the embedded discrete-time process $\{X_k\}$ is a Markov chain with transition matrix $\boldsymbol{P}(\boldsymbol{f})$.

A still more restricted class G is that of *simple* or *non-randomized stationary* policies. These are obtained by specializing the measure m_f to consist of a single atom. In that case, G can be characterized by a simple mapping, namely $\boldsymbol{g}: S \to \mathscr{A}$.

In our earlier paper [3] (see also [21]) we introduced *mixed* policies, whose space is indicated by ${}_{\mathrm{m}}F$. A mixed policy is merely a stationary policy with atoms of mass q and $1-q$ for each $x \in S$. Such a policy may be written symbolically for convenience $\boldsymbol{f}_q = q\boldsymbol{g}_1 + (1-q)\boldsymbol{g}_2$, with $q \in [0, 1]$. The meaning attached to such an $\boldsymbol{f}_q$ is that when the state is x, action $g_1(x)$ is invoked with probability q, and $g_2(x)$ with probability $1-q$.

Finally, we define the class of *semi-simple* policies, which is designated ${}_{\mathrm{s}}F$. We say $\boldsymbol{f} \in {}_{\mathrm{s}}F$ if $\boldsymbol{f} \in F$ is characterized by a measure $m_f(\cdot \mid x)$ possessing the following property: there is a state $x_0 \in S$ for which $m_f(\cdot \mid x_0)$ is described by either one or two atoms; for every other state, $m_f(\cdot \mid x)$ consists of exactly one atom. In other words, there may be one state for which randomization between two actions is required, while the policy is non-random for every other state. It is clear that $G \subset {}_{\mathrm{s}}F \subset {}_{\mathrm{m}}F$.

The system earns a *reward* that accumulates throughout time. The reward consists of two components; one takes on an increment with every jump, and the other is a 'running reward' whose rate is specified. To reflect this, we write

$$C(X_k, A_k) = c_0(X_k, A_k) + \tau_k c(X_k, A_k) \tag{1.8}$$

to denote the incremental reward earned during the sojourn interval $[T_k, T_{k+1})$ when the state is X_k and the action A_k is applied. In terms of the above, we seek to maximize the *average reward* defined by

$$R_x(\boldsymbol{u}) \triangleq \liminf_n \left\{ \frac{E_{\boldsymbol{u}}^x\left[\sum_{k=0}^{n-1} C(X_k, A_k)\right]}{E_{\boldsymbol{u}}^x\left[\sum_{k=0}^{n-1} \tau_k\right]} \right\}, \tag{1.9}$$

in which $E_{\boldsymbol{u}}^x$ is the conditional expectation when the probability measure is

determined by the policy $\boldsymbol{u}$, and the conditioning event is $\{X_0 = x\}$. Ordinarily, it is desired to find the $\boldsymbol{u} \in U$ that yields the supremum of the average reward (1.8) over $\boldsymbol{u} \in U$ for all $x \in S$; however, we have yet to consider the supremum only over those policies that satisfy a constraint we are about to specify. To this end, let the system incur a *cost* called $D(\cdot, \cdot)$ with definition and notation entirely analogous to those of the reward. The *average cost* is given by

$$K_x(\boldsymbol{u}) \triangleq \limsup_n \left\{ \frac{E_{\boldsymbol{u}}^x\left[\sum_{k=0}^{n-1} D(X_k, A_k)\right]}{E_{\boldsymbol{u}}^x\left[\sum_{k=0}^{n-1} \tau_k\right]} \right\}. \tag{1.10}$$

We postpone for the time being any discussion on the measurability of $C(x, \cdot)$ and $D(x, \cdot)$.

Our *constraint* is on the average cost, in the sense that we require

$$K_x(\boldsymbol{u}) \leqq \alpha \tag{1.11}$$

for all x. Then, if U_0 is the subspace of U on which the constraint (1.11) is satisfied, we shall discuss the attainment of

$$R_x = \sup_{\boldsymbol{u} \in U_0} R_x(\boldsymbol{u}). \tag{1.12}$$

Any policy $\boldsymbol{u}$ that attains R_x for each x while simultaneously satisfying the constraint (1.11) is termed a *constrained optimal policy*, or more simply, an *optimal policy*.

2. The constrained optimal policy

In this section, we establish one of the central results of the paper: the existence of an optimal semi-simple policy. First, however, we need to introduce some recurrence and continuity assumptions, and to investigate the long-run average cost and reward for stationary policies.

We begin with some characteristics of the sojourn times τ_n. Let us define

$$\tau(x, a) \triangleq \int_0^\infty \left[1 - \sum_{y \in S} Q(x, y, t; a)\right] dt \tag{2.1}$$

so that clearly

$$\tau(x, a) = E[\tau_n \mid X_n = x, A_n = a]. \tag{2.2}$$

We shall require the following.

Hypothesis 2.1. $\tau(x, \cdot)$ is continuous on $\mathscr{A}$, and

$$\sum_{y \in S} Q(x, y, 0; a) < 1 \tag{2.3}$$

for each $x, y \in S$ and each $a \in \mathscr{A}$.

Hypothesis 2.2. The functions $c_0(x, \cdot)$, $c(x, \cdot)$, $d_0(x, \cdot)$ and $d(x, \cdot)$ are all continuous on $\mathscr{A}$.

It follows from the second part of Hypothesis 2.1 that $\tau(x, a) > 0$. Then the assumed continuity, together with the compactness of $\mathscr{A}$ assures a positive lower bound for $\tau(x, a)$, and the same properties also demand a finite upper bound. Therefore,

$$0 < m \leqq \tau(x, a) \leqq M < \infty. \tag{2.4}$$

By virtue of Hypothesis 2.2 and the compactness of $\mathscr{A}$ the costs and rewards are all bounded, and we may therefore suppose them to be non-negative. The continuity will, of course, settle all measurability questions regarding these functions.

The next assumption makes certain that S contains only one positive recurrent class, which to a certain extent does not depend on the choice of policy. Other, less restrictive hypotheses often complicate the results without enhancing insight into the system behavior.

Hypothesis 2.3. For each simple policy $\boldsymbol{g}$, there is a state (take 0 for definiteness) which is accessible from each $x \in S$ for the Markov chain with transition probabilities $P_{xy}(\boldsymbol{g})$.

The above hypothesis implies (see [3], Lemma 2.3) that for each stationary policy $\boldsymbol{f}$ there is an unique probability vector $\boldsymbol{\nu}_f$ that satisfies $\boldsymbol{\nu}_f = \boldsymbol{\nu}_f \boldsymbol{P}_f$; moreover,

$$\nu_f(x) = \lim_n n^{-1} \sum_{k=0}^{n-1} P_f(X_k = y \mid X_0 = x). \tag{2.5}$$

It is convenient to define

$$\pi_f(y) = \frac{\nu_f(y)\tau_f(y)}{\sum_x \nu_f(x)\tau_f(x)} \tag{2.6}$$

in which

$$\tau_f(x) \triangleq \int_{\mathscr{A}} \tau(x, a) m_f(da \mid x). \tag{2.7}$$

With the above hypotheses and facts, one can express the average reward in a way that is more intuitively suggestive than (1.9).

Theorem 2.4. For policies in F, and any x_0, the average reward (1.9) can be written as

$$R(f) = \frac{\sum_x \nu_f(x) C_f(x)}{\sum_x \nu_f(x) \tau_f(x)} \tag{2.8}$$

where $\tau_f(x)$ is as previously defined, and

$$C_f(x) \triangleq \int_{\mathscr{A}} \bar{C}(x, a) m_f(da \mid x), \tag{2.9}$$

with

$$\bar{C}(x, a) \triangleq c_0(x, a) + \tau(x, a) c(x, a). \tag{2.10}$$

Proof. See [14].

With the following notation

$$\hat{P}_{xy}(a) = \begin{cases} \dfrac{m}{\tau(x, a)} P_{xy}(a) & x \neq y \\ 1 - \dfrac{m}{\tau(x, a)} [1 - P_{xx}(a)] & x = y, \end{cases} \tag{2.11}$$

we may state the following result.

Corollary 2.5. For each policy $g \in G$, there is a Markov chain $\{\hat{X}_k\}$ having the same recurrence properties as $\{X_k\}$, and defined on the same space S. For this new Markov chain with transition matrix $\hat{\boldsymbol{P}}(g)$, the associated (unique) probability vector π_g is given by (2.6). Moreover, when $c_0 \equiv 0$, $R(g)$ is furnished by

$$R(g) = \lim_{n \to \infty} n^{-1} \hat{E}_g \left[\sum_0^{n-1} c(\hat{X}_k, g(\hat{X}_k)) \right]. \tag{2.12}$$

Proof. See [14].

Equation (2.11) and the resulting corollary may be inferred from an earlier result in [19].

The remainder of this section is devoted to the proof of optimality of semi-simple policies for constrained SMDP. We begin by introducing the Lagrangian constraint, which appears in terms of the reward $C(\cdot, \cdot)$ as well as the cost $D(\cdot, \cdot)$. In terms of a constraint parameter ω, we consider the

unconstrained optimization for the *parametrized reward*

$$B^{\omega}(x, a) \triangleq C(x, a) - \omega D(x, a). \tag{2.13}$$

The parametrized reward has all the properties of the reward (i.e., boundedness, continuity), except that it need not be non-negative. There is now a *parametrized average reward* $J_x^{\omega}(\boldsymbol{u})$ completely analogous to (1.9), to which Theorem 2.4 is immediately applicable.

Central to the derivation of the optimal parametrized average reward is the following classical result.

Theorem 2.6. For each fixed ω there exists a scalar s^{ω} and a bounded vector $\boldsymbol{h}^{\omega}$ such that the DPE

$$h^{\omega}(x) = \sup_{a \in \mathscr{A}} \left[B^{\omega}(x, a) + \sum_{y \in S} P_{xy}(a) h^{\omega}(y) - s^{\omega} \tau(x, a) \right] \tag{2.14}$$

is satisfied for each $x \in S$. Any policy $\boldsymbol{g}^{\omega} \in G$ specified by

$$g^{\omega}(x) = \arg\sup_{a \in \mathscr{A}} \left[B^{\omega}(x, a) + \sum_{y \in S} P_{xy}(a) h^{\omega}(y) - s^{\omega} \tau(x, a) \right] \tag{2.15}$$

attains

$$J^{\omega} = \sup_{\boldsymbol{u} \in U} J_x^{\omega}(\boldsymbol{u}) \tag{2.16}$$

for all $x \in S$. Moreover, s^{ω} in (2.14) satisfies $s^{\omega} = J^{\omega}$.

Proof. See [15], Theorem 7.6, and [19].

We may now proceed as in Sections 3 and 4 of [3]. However, we must first be assured that the problem is feasible by the existence of a policy satisfying the average constraint (1.11). Accordingly, we require the following.

Hypothesis 2.7. There exists a $\boldsymbol{g} \in G$ such that

$$K(\boldsymbol{g}) < \alpha. \tag{2.17}$$

In view of the hypothesis and the existence of the DPE solution, all of Section 4 of [3] can be applied without essential change; hence we may conclude that there exists a constrained optimal policy in ${}_{\mathrm{m}}F$ (compare [19]). However, it is more advantageous if randomization is needed for at most one state. That such is actually the case follows from the next theorem.

Theorem 2.8. There exists a constrained optimal policy in ${}_{\mathrm{s}}F$.

Proof. We recall the elements of the proof of Theorem 4.4 in [3]. Following the methodology of the referenced proof, one introduces the constraint

parameter

$$\gamma \triangleq \inf_{\omega} \{\omega \mid K(\boldsymbol{g}^{\omega}) \leqq \alpha\} \tag{2.18}$$

where $\boldsymbol{g}^{\omega}$ is any simple policy that solves

$$h^{\omega}(x) = \sup_{a \in \mathscr{A}} \left[B^{\omega}(x, a) + \sum_{y \in S} P_{xy}(a) h^{\omega}(y) - J^{\omega}\tau(x, a) \right] \tag{2.19}$$

for each $x \in S$. That such a $\gamma < \infty$ exists follows from Hypothesis 2.7, using the argument of Lemma 3.3 in [3].

If there is a $\boldsymbol{g}^{\gamma}$ such that $K(\boldsymbol{g}^{\gamma}) = \alpha$ we are finished, according to Theorem 4.3 of [3]. Otherwise, we consider the policies G^{γ} in G that solve (2.19). As in Theorem 4.4 of [3], we use the compactness of $[0, \beta] \times G^{\beta}$ for any $\beta < \infty$ (see Theorem 3.5 of [3]), together with the continuity of $K(\cdot)$ on G (Lemma 3.4 of [3]) to show that there exist $\underline{\boldsymbol{g}}, \bar{\boldsymbol{g}} \in G^{\gamma}$ such that

$$K(1\underline{\boldsymbol{g}}) = \alpha_0 < \alpha < \alpha^0 = K(\bar{\boldsymbol{g}}). \tag{2.20}$$

In fact, G^{γ} contains not only $\underline{\boldsymbol{g}}$ and $\bar{\boldsymbol{g}}$, but also all elements $\boldsymbol{g}^{(k)} \in G$ of the form

$$g^{(k)}(x) = \begin{cases} \underline{g}(x) & x < k \\ \bar{g}(x) & x \geqq k; \end{cases} \tag{2.21}$$

thus, $\boldsymbol{g}^{(0)} = \bar{\boldsymbol{g}}$, and for consistency we take $\boldsymbol{g}^{(n+1)} = \underline{\boldsymbol{g}}$. That all these $\boldsymbol{g}^{(k)}$ belong to G^{γ} is true because $\underline{\boldsymbol{g}}$ and $\bar{\boldsymbol{g}}$ satisfy (2.19) separately for each coordinate $x \in S$.

Now take m as the first index such that $K(\boldsymbol{g}^{(m)}) < \alpha$. Clearly, $0 < m \leqq n + 1$,

$$K(\boldsymbol{g}^{(m)}) < \alpha < K(\boldsymbol{g}^{(m-1)}), \tag{2.22}$$

and $\boldsymbol{g}^{(m)}(x)$ differs from $\boldsymbol{g}^{(m-1)}(x)$ only for $x = m - 1$. To complete the proof, we proceed to randomize between $\boldsymbol{g}^{(m)}$ and $\boldsymbol{g}^{(m-1)}$ as in the last portion of the proof of Theorem 4.4. in [3]. If the randomization is symbolically written $\boldsymbol{f}_q \triangleq q\boldsymbol{g}^{(m)} + (1 - q)\boldsymbol{g}^{(m-1)}$, we find (just as in [3]) that $J^{\gamma} = J^{\gamma}(\boldsymbol{f}_q)$, and that $K(\boldsymbol{f}_q)$ is continuous in q. It then follows from (2.22) that for some $q \in (0, 1)$ we have $K(\boldsymbol{f}_q) = \alpha$; hence, by Theorem 4.3 of [3], this $\boldsymbol{f}_q$ is optimal. Since $\boldsymbol{f}_q \in {}_{\mathrm{s}}F$, the proof is complete.

3. Linear structures

Theorem 2.8 is extremely useful in providing an indication of the nature of the constrained optimal policy, although Equation (2.14) is generally quite troublesome to solve. Fortunately, some special assumptions consistent with many practical applications lead to simpler optimal policies.

Consider then an exponential queue with service rate μ_x, a controlled input rate a, and a total system capacity of N customers. We find that $\tau(x, a) = [\mu_x + a]^{-1}$ for $0 < x < N$. Note that when a randomized policy is applied, the sojourn distribution is a mixture of two exponentials so that the continuous-time state process is not a Markov process. Further, for $0 < x < N$,

$$P_{xy}(a) = \begin{cases} \dfrac{a}{\mu + a} & y = x + 1 \\ 0 & y = x \\ \dfrac{\mu}{\mu + a} & y = x - 1, \end{cases} \tag{3.1}$$

which means that (2.14) is highly non-linear in a.

While the form of $P_{xy}(a)$ in (3.1) is inconvenient, the same is not true for the discrete-parameter Markov chain with transition probability $\hat{P}_{xy}(a)$ (see (2.11)), which turns out to be linear in a. This suggests a class of applications that take advantage of Corollary 2.5. The idea is to replace constrained maximization of (1.9) via solutions to the DPE (2.14) by the easier approach suitable for a discrete-time system.

However, a careful reading of Corollary 2.5 indicates that the discrete-time optimization is valid only over G; whereas [3] demonstrates that the constrained optimal policy may (and usually does) belong to the larger space F. This suggests that the discrete-time optimization is capable only of finding the best simple policy, which is necessarily sub-optimal. The prospective deficiency can be avoided by some reasonable assumptions which are embodied in the following hypothesis.

Hypothesis 3.1. Let c_0 and d_0 be identically 0, with the running reward and cost functions having the affine forms

$$\begin{aligned} c(x, a) &= \bar{r}(x) + ar(x) \\ d(x, a) &= \bar{k}(x) + ak(x), \end{aligned} \tag{3.2}$$

with $\mathscr{A}$ a closed interval in R^+. Likewise, $\hat{\boldsymbol{P}}$ is affine, viz.

$$\hat{P}_{xy}(a) = \bar{p}_{xy} + ap_{xy}. \tag{3.3}$$

We can—and shall—assume without loss of generality that $\mathscr{A} = [0, \bar{a}]$.

Before exhibiting the role of the hypothesis, we review some applicable discrete-time results. Consider the controlled Markov chain $\{X_n\}$ with law of motion $\hat{P}_{xy}(a)$ furnished by (2.11). From Corollary 2.5, the continuous-time averages are equivalently expressed for any $\boldsymbol{g} \in G$ by

$$J_x^{\omega}(\boldsymbol{g}) = \liminf_n n^{-1} \hat{E}_{\boldsymbol{g}}^x \left[\sum_0^{n-1} b^{\omega}(\hat{X}_k, g(\hat{X}_k)) \right] \tag{3.4}$$

and

$$K(\boldsymbol{g}) = \limsup_n n^{-1}\hat{E}_{\boldsymbol{g}}^x\left[\sum_0^{n-1} d(\hat{X}_k, g(\hat{X}_k))\right], \tag{3.5}$$

where $\boldsymbol{b}^\omega \triangleq \boldsymbol{c} - \omega \boldsymbol{d}$. The discrete-time DPE corresponding to the maximization of (3.4) is (see [15], Theorem 6.17)

$$J^\omega + h^\omega(x) = \sup_{a \in \mathscr{A}} \left\{ b^\omega(x, a) + \sum_y \hat{P}_{xy}(a) h^\omega(y) \right\}. \tag{3.6}$$

Under our accessibility assumption (Hypothesis 2.3) on $\boldsymbol{P}(\boldsymbol{g})$, (3.6) always has at least one solution which entails an $\boldsymbol{h}^\omega$ unique up to a constant. Moreover, the simple policy(ies) $\boldsymbol{g}^\omega \in G$ which maximize the right-hand side of (3.6) are optimal respective to all policies, i.e.,

$$J^\omega(\boldsymbol{g}^\omega) = \sup_{\boldsymbol{u} \in \boldsymbol{U}} J_x^\omega(\boldsymbol{u}) = J^\omega. \tag{3.7}$$

for all $x \in S$.

When the above is applied in the light of Hypothesis 3.1, we find that $\boldsymbol{g}^\omega$ is a bang-bang control. Specifically, we let

$$z^\omega(x) \triangleq r(x) - \omega k(x) + \sum_y p_{xy} h^\omega(y), \tag{3.8}$$

noting that the uniqueness of $\boldsymbol{h}^\omega$ up to a constant, together with $\sum_y p_{xy} = 0$ renders $z^\omega(x)$ unique also. It now follows from (3.6) that the optimal policy $\boldsymbol{g}^\omega$ is specified by

$$g^\omega(x) = \bar{a} \quad \text{if} \quad z^\omega(x) > 0, \tag{3.9a}$$

$$g^\omega(x) = 0 \quad \text{if} \quad z^\omega(x) < 0, \tag{3.9b}$$

and

$$g^\omega(x) = \text{any value in } [0, \bar{a}] \quad \text{if} \quad z^\omega(x) = 0. \tag{3.9c}$$

To obtain the optimal constrained policy, however, we must apply more delicate techniques. Basic to the analysis is the following paraphrase of [3], Theorem 4.3.

Lemma 3.2. Suppose that for some $\omega \geqq 0$ and some $\boldsymbol{g} \in G$ we have

$$J^\omega(\boldsymbol{g}) = J^\omega \quad \text{and} \quad K(\boldsymbol{g}) = \alpha \tag{3.10}$$

for the discrete-time constrained optimization problem described by Equations (3.4) and (3.5). Then the same $\boldsymbol{g}$ also solves the associated constrained SMDP optimization problem.

Proof. The proof of Theorem 4.3 of [3] utilizes monotonicity properties that hold in the continuous-time as well as the discrete-time model. Therefore, that

theorem applies to policies of class G appearing in Theorem 2.6. On the other hand, from Corollary 2.5, there is a one-to-one relation between J^ω, $J^\omega(\boldsymbol{g})$, $K(\boldsymbol{g})$, and $\boldsymbol{g}^\omega$ in the continuous time and associated discrete-time problem described by Equations (3.4) through (3.7). Accordingly, it suffices to obtain (3.10) for the latter.

The above lemma along with Hypothesis 3.1 leads to central results on the non-randomized and/or bang-bang character of optimal policies.

Theorem 3.3. Under Hypothesis 3.1, there exists a simple policy which is an optimal constrained policy for the SMDP problem.

Proof. According to Lemma 3.2, we need only verify the existence of the desired $\boldsymbol{g} \in G$ for the discrete-time problem whose unconstrained optimal policies are given by (3.9). We begin the proof by defining G^γ to be the set of simple policies satisfying the DPE (3.6) when the applicable parameter γ is that of (2.18). From Lemma 3.3. of [3], Hypothesis 2.7 assures that $\gamma < \infty$; since solutions to the DPE must exist for each ω, G^γ is non-trivial.

The proof uses the same notation as Theorem 2.8. As was the case there, if there is a $\boldsymbol{g}^\gamma \in G^\gamma$ such that $K(\boldsymbol{g}^\gamma) = \alpha$, there is nothing further to prove. Otherwise, there exist $\boldsymbol{g}, \bar{\boldsymbol{g}} \in G^\gamma$ satisfying (2.20).

Now let us examine G^γ in greater detail. This set in G is regarded as a subset of an $(N+1)$-dimensional parallelopiped bounded on each coordinate by 0 and $\bar{a}$. In fact,

$$G^\gamma = A^{(0)} \times A^{(1)} \times \cdots \times A^{(N)} \tag{3.11}$$

where each

$$\begin{aligned} A^{(x)} &= \{\bar{a}\} \quad \text{if} \quad z^\gamma(x) > 0 \\ A^{(x)} &= \{0\} \quad \text{if} \quad z^\gamma(x) < 0 \\ A^{(x)} &= [0, \bar{a}] \quad \text{if} \quad z^\gamma(x) = 0. \end{aligned} \tag{3.12}$$

Since G^γ contains at least two elements, G^γ must include at least one $A^{(x)}$ of the last type. Further, G^γ is a connected compact set containing both $\boldsymbol{g}$ and $\bar{\boldsymbol{g}}$. Then from the continuity of $K(\cdot)$, and from (2.20), there exists at least one $\boldsymbol{g}_0 \in G^\gamma$ for which $K(\boldsymbol{g}_0) = \alpha$. For this element (as for all members of G^γ), we have $J^\gamma(\boldsymbol{g}_0) = J^\gamma$. The proof of the theorem is thus complete.

Corollary 3.4. Among the simple optimal constrained policies there exists at least one such that $g(x) \in (0, \bar{a})$ for at most one $x \in S$, the policy being bang-bang at each of the remaining states.

Proof. According to (3.9), $\boldsymbol{g}^\omega$ can always be chosen to be a bang-bang control, so that the subsequences converging to $\boldsymbol{g}$ and $\bar{\boldsymbol{g}}$ (see again Theorem

4.4 of [3]) are bang-bang, and the same is finally true for $\boldsymbol{g}$ and $\bar{\boldsymbol{g}}$. Define $\boldsymbol{g}^{(m)}$ and $\boldsymbol{g}^{(m-1)}$ as in Theorem 2.8; these are also bang-bang controls. Now $g^{\gamma}(m-1)$ is not uniquely specified, so that we must have $z^{\gamma}(m-1)=0$.

The above facts permit us to construct policies $\boldsymbol{g}_q \in G^{\gamma}$ by the formula

$$\boldsymbol{g}_q = \begin{cases} g^{(m)}(x) & x \neq m-1 \\ \bar{a}q & x = m-1 \end{cases}$$

for any $q \in [0, 1]$. Since $\boldsymbol{g}_0$ is identical with one of $\boldsymbol{g}^{(m)}$ or $\boldsymbol{g}^{(m-1)}$, and $\boldsymbol{g}_1$ is the same as the other, we have that $K(\boldsymbol{g}_q) < \alpha$ for one of $q = 0, 1$, and $K(\boldsymbol{g}_q) > \alpha$ for the other. The continuity of $K(\cdot)$ on G then assures the existence of a $q \in (0, 1)$ such that $K(\boldsymbol{g}_q) = \alpha$; this $\boldsymbol{g}_q$ is then optimal, and is bang-bang except for $g_q(m-1)$.

An alternative optimal policy is bang-bang for every state, but there may be one state where there is a random choice between $\bar{a}$ and 0.

Corollary 3.5. Under Hypothesis 3.1, there exists an optimal constrained bang-bang policy that is randomized at no more than one state.

Proof. Take $\boldsymbol{g}^{(m)}$, $\boldsymbol{g}^{(m-1)}$ and γ as above. Proceeding as in Theorem 2.8, we find a $q \in (0, 1)$ such that $\boldsymbol{f}_q$ is the constrained optimal policy. Since $\boldsymbol{g}^{(m)}$ and $\boldsymbol{g}^{(m-1)}$ are bang-bang controls differing only for $x = m-1$, the proof is complete.

Remark. If, as in (3.1), $\tau(m-1, \bar{a}) \neq \tau(m-1, 0)$, the conditional sojourn time, given the state $m-1$, is gamma distributed; hence, the controlled process is not Markov, but semi-Markov.

4. Optimal flow control

We now apply the theory and results of this paper to the flow control problem described at the beginning of the preceding section. In fact, we examine a more general problem by allowing the action space to be state dependent with $\mathscr{A}_x = [0, \beta_x]$ and $\beta_N = 0$. There is already a substantial literature on unconstrained dynamic flow control of queues (see [20] for a recent survey and bibliography), and we do not intend to duplicate this body of work. Nevertheless, unconstrained optimality is almost trivially obtained from what already appears earlier in this paper, and in any case is required for the study of optimization under a constraint. For the flow control problem, the unconstrained version of the DPE (3.6) becomes

$$R + \mu_x f(x) = \sup_{a \in \mathscr{A}_x} [c(x, a) + af(x+1)] \tag{4.1}$$

with $\mu_0 = 0$ and $\boldsymbol{f}$ related to the $\boldsymbol{h}$ in (4.1) by

$$f(x) = h(x) - h(x-1). \tag{4.2}$$

It is seen that the $\boldsymbol{P}(a)$ for the controlled queue already meets condition (3.3) of Hypothesis 3.1. We now and hereafter assume that the other linearity requirements of Hypothesis 3.1 are satisfied also. Then (4.1) takes on the form

$$R + \mu_x f(x) = \bar{r}(x) + \beta_x [r(x) + f(x+1)]^+, \tag{4.3}$$

where we write $[w]^+ \triangleq \max(0, w)$. The form of the solution to the unconstrained optimization problem is now immediately evident. We may take

$$n^* \triangleq \min \{x : r(x) + f(x+1) \leqq 0\}, \tag{4.4}$$

and $n^* = N$ if the right side of (4.4) fails to exist. According to (4.3), one optimal policy is then specified by $\boldsymbol{g}$ such that

$$g(x) = \begin{cases} \beta_x & x < n^* \\ 0 & x = n^*. \end{cases} \tag{4.5}$$

Note that the values of $g(x)$ for $x > n^*$ are irrelevant, since the state set $\{n^* + 1, \cdots, N\}$ is wholly transient if $g(n^*) = 0$. Thus, we have proved that the unconstrained optimal policy can always be taken bang-bang, and that this policy is completely described by n^*. Moreover, a calculation based on $\boldsymbol{\pi_g}$ obtained from the well-known equilibrium probabilities for the birth–death process indicates that the corresponding average reward is

$$R = R(n^*), \tag{4.6}$$

where we define

$$R(k) \triangleq \frac{\sum_0^{k-1} \beta_x \rho_x r(x) + \sum_0^{k} \rho_x \bar{r}(x)}{\sum_0^{k} \rho_x} \tag{4.7}$$

with $\rho_0 = 1$ and for $j > 0$

$$\rho_j \triangleq \frac{\beta_0 \beta_1 \cdots \beta_{j-1}}{\mu_1 \mu_2 \cdots \mu_j}. \tag{4.8}$$

Because of Hypothesis 3.1, at least one version of the optimal constrained policy is simple, with $g(x)$ either equal to 0 or β_x possibly except for one x, where $g(x) \in (0, \beta_x)$; this follows from Corollary 3.4. Nevertheless, it remains possible that for some n we have $g(x) = \beta_x$ for $x < n$, $0 < g(n) < \beta_n$, and $g(n+1) = \beta_{n+1}$. Such an optimal $\boldsymbol{g}$ does not represent the most desirable

situation as compared to one where $g(n) \in (0, \beta_n)$ is followed by $g(x) = 0$ for all $x > n$.

It will be shown that a constrained optimal policy of the more desirable type is guaranteed by certain conditions on the behavior of the reward and cost functions. Accordingly, we require the following.

Hypothesis 4.1. Define

$$Q(x) \triangleq \mu_{k+1} r(x) + \bar{r}(x+1) \tag{4.9}$$

and

$$V(x) \triangleq \mu_{x+1} k(x) + \bar{k}(x+1). \tag{4.10}$$

Then Q is strictly decreasing, and V is non-decreasing.

Remark. Note that (4.7) can be written equally well for the average reward J and cost K, so that in particular $J(0) = Q(-1)$ and $K(0) = V(-1)$. We then have the following result.

Lemma 4.2. $J(x_0) = J(x_0 + 1)$ iff $J(x_0) = Q(x_0)$, and J is then strictly decreasing for $x \geqq x_0 + 1$. If $J(x_0) > Q(x_0)$, then J is strictly decreasing for $x \geqq x_0$.

Proof. Write

$$J(x) = \frac{p_1(x)}{p_2(x)} \tag{4.11}$$

where p_1 and p_2 are respectively the numerator and denominator in (4.7). Similarly, let

$$Q(x) = \frac{q_1(x)}{q_2(x)} \tag{4.12}$$

in which

$$q_1(x) = \beta_x \rho_x r(x) + \rho_{x+1} \bar{r}(x+1) \quad \text{and} \quad q_2(x) = \rho_{x+1}. \tag{4.13}$$

Furthermore, we find that

$$J(x+1) = \frac{p_1(x) + q_1(x)}{p_2(x) + q_2(x)}. \tag{4.14}$$

If Hypothesis 4.1 is applied to (4.14), and the elementary properties of fractions taken into account, the lemma follows.

With the aid of the lemma, we show that there is a constrained optimal policy that does not exhibit the anomaly suggested just prior to Hypothesis 4.1. We turn to the Lagrangian formulation with a parametrized reward $\boldsymbol{g}^{\omega}$ as in Section 3. Taking

$$W^{\omega}(x) \triangleq Q(x) - \omega V(x) \tag{4.15}$$

leaves $\boldsymbol{W}^{\omega}$ with the same monotonicity ascribed to $\boldsymbol{Q}$; hence Lemma 4.2 is applicable to the maximum average reward J^{ω} obtained when $\bar{\boldsymbol{r}}$ is replaced by $\bar{\boldsymbol{b}}^{\omega}$ and $\boldsymbol{r}$ by $\boldsymbol{b}^{\omega}$ in the reward function. We also define

$$n^{\omega} \triangleq \min \{x : b^{\omega}(x) + f^{\omega}(x+1) \leqq 0\} \tag{4.16}$$

in analogy to the n^* defined earlier.

The optimal parametrized solution of the DPE (4.3) is now characterized as follows.

Theorem 4.3. Let Hypothesis 4.1 apply, and continue to assume the linearity and accessibility hypotheses. Then for each $\boldsymbol{g}^{\omega}$ satisfying the DPE (4.3), either

$$g^{\omega}(x) = \begin{cases} \beta_x & x < n^{\omega} \\ 0 & x \geqq n^{\omega} \end{cases} \tag{4.17a}$$

or

$$g^{\omega}(x) = \begin{cases} \beta_x & x < n^{\omega} \\ \xi\beta_{n^{\omega}} & x = n^{\omega} \\ 0 & x > n^{\omega} \end{cases} \tag{4.17b}$$

where ξ is any number in $(0, 1)$. In either case,

$$J^{\omega} = J^{\omega}(n^{\omega}). \tag{4.18}$$

Proof. If $b^{\omega}(n^{\omega}) + f^{\omega}(n^{\omega}+1) < 0$, the DPE is satisfied only if $\boldsymbol{g}^{\omega}$ is of the form (4.17a). Hence we need only consider

$$b^{\omega}(n^{\omega}) + f^{\omega}(n^{\omega}+1) = 0. \tag{4.19}$$

In either one of these two cases, one possible $\boldsymbol{g}^{\omega}$ is given by (4.17a), so that (4.18) is true.

In (4.3) with $x = n^{\omega} + 1$, substitute (4.18) and (4.19), so that (4.3) implies $J^{\omega}(n^{\omega}) \geqq W^{\omega}(n^{\omega})$; hence, by Lemma 4.2

$$J^{\omega}(n^{\omega}) > J^{\omega}(n^{\omega}+2), \tag{4.20}$$

with J^{ω} strictly decreasing from $n^{\omega}+1$ on. Let

$$\bar{n}^{\omega} \triangleq \min \{x : x > n^{\omega},\ b^{\omega}(x) + f^{\omega}(x+1) < 0\}. \tag{4.21}$$

Then there exists $\boldsymbol{g}^{\omega}$ solving the DPE such that

$$g^{\omega}(x) = \begin{cases} \beta_x & x < \bar{n}^{\omega} \\ 0 & x = \bar{n}^{\omega}. \end{cases} \tag{4.22}$$

Therefore, $J^{\omega} = J^{\omega}(\bar{n}^{\omega})$. A comparison of this last statement with (4.20) shows that $\bar{n}^{\omega} = n^{\omega} + 1$, from which $g^{\omega}(n^{\omega}+1)$ must be 0.

With the aid of Theorem 4.3, it is now easy to show that the optimal constrained policy takes on one of the forms (4.17).

Theorem 4.4. Let Hypothesis 2.7 hold. Then there is a simple optimal constrained policy defined as follows in terms of integer n and some $\xi \in [0, 1)$:

$$g(x) = \begin{cases} \beta_x & x < n \\ \xi\beta_n & x = n \\ 0 & x > n. \end{cases} \tag{4.23}$$

Proof. In Theorem 3.3 it was shown that the optimal constrained policy is attained by some $\boldsymbol{g} \in G^{\gamma}$. But from Theorem 4.3 each such $\boldsymbol{g}$ is of the form (4.23).

As in Section 3, there is a parallel form for the optimal policy following from Corollary 3.5 and the above. This policy is bang-bang, but possibly randomized at one state.

Corollary 4.5. Under Hypothesis 2.7, there is an optimal control as in (4.23), except that the control is randomized at some state n, at which the flow rate takes on value β_n with probability $q \in [0, 1]$, and 0 with probability $1 - q$.

When Hypothesis 4.1 is satisfied by the relevant service rates, rewards and costs, Theorem 4.4 applies to queues with Poisson arrivals, exponential service distributions at state-dependent rates, and finite storage capacity, as for example the $M/M/s/N$ queue. If the number of customers in the system is less than n, the arrival rate is set at full blast (β_x), and the arrival rate is set to 0 when the number of customers is greater than n. For exactly n customers, the input stream is reduced in intensity to $\xi\beta_n$ according to Theorem 4.4. For an alternative optimal constrained policy, a strict bang–bang action (possibly randomized at one state) can be applied, as described by Corollary 4.5. An efficient algorithm is given in [14] to determine the optimal parameters n and ξ.

We conclude with an example illustrative of the generality of Theorem 4.4.

Example 4.6. Consider the $M/M/1/N$ queueing model with service rate μ and cost

$$d(x, a) = x - va, \tag{4.24}$$

where v is some positive number. By Theorem 2.4, the global constraint then becomes

$$K(\boldsymbol{g}) = \lim_{t\to\infty} t^{-1} E_{\boldsymbol{g}}\left\{\int_0^t [Y(\xi) - vg(Y(\xi))]\, d\xi\right\}. \tag{4.25}$$

With $\alpha = 0$ in (1.11), the constraint on the average cost (4.25) is tantamount to

$$\frac{\lim_{t\to\infty} t^{-1} E_g\left[\int_0^t Y(\xi)\, d\xi\right]}{\lim_{t\to\infty} t^{-1} E_g\left[\int_0^t g(Y(\xi))\, d\xi\right]} \leqq v \tag{4.26}$$

when $g(0) \neq 0$. Observe that, by Little's formula (see e.g. [2]), the ratio appearing in (4.26) may be interpreted as the average time delay. Furthermore, a simple queueing argument indicates that there exists a simple policy $\boldsymbol{g}$ forcing the time delay to be strictly less than v (and hence satisfying Hypothesis 2.7) if and only if $\mu^{-1} < v$. Therefore, if the latter condition is met, and if $c(x, a)$ satisfies Hypotheses 3.1 and 4.1, the conclusions of Theorem 4.4 and Corollary 4.5 apply; either one of the two resulting policies may be chosen according to ease of implementation. This example then generalizes a result of [12], where the reward was specified by $c(x, a) = a$, and the optimization restricted to the class of simple policies G.

References

[1] BERTSEKAS, D. P. (1976) *Dynamic Programming and Stochastic Control.* Academic Press, New York.

[2] BEUTLER, F. J. (1983) Mean sojourn times in Markov queueing networks: Little's formula revisited. *IEEE Trans. Inf. Theory* **29,** 233–241.

[3] BEUTLER, F. AND ROSS, K. (1985) Optimal policies for controlled Markov chains with a constraint. *J. Math. Anal. Appl.* **112,** 236–252.

[4] ÇINLAR, E. (1975) *Introduction to Stochastic Processes.* Prentice-Hall, Englewood Cliffs, NJ.

[5] DOOB, J. L. (1953) *Stochastic Processes.* Wiley, New York.

[6] FELLER, W. (1966) *An Introduction to Probability Theory and its Applications,* Vol. 2. Wiley, New York.

[7] FOX, B. (1966) Markov renewal programming by linear fractional programming. *SIAM J. Appl. Math.* **14,** 1418–1432.

[8] HAJEK, B. (1984) Optimal control of two interacting service stations. *IEEE Trans. Autom. Control* **29,** 491–499.

[9] KELLY, F. P. (1979) *Reversibility and Stochastic Networks.* Wiley, New York.

[10] KLEINROCK, L. (1976) *Queueing Systems Vol. 2: Computer Applications.* Wiley, New York.

[11] KOYABASHI, H. (1978) *Modeling and Analysis: an Introduction to System Performance Evaluation Methodology.* Addison-Wesley, Reading, Mass.

[12] LAZAR, A. A. (1983) The throughput time delay function of an $M/M/1$ queue. *IEEE Trans. Inf. Theory* **29,** 914–918.

[13] MURTY, K. G. (1983) *Linear Programming.* Wiley, New York.

[14] ROSS, K. W. (1985) Constrained Markov Decision Processes with Queueing Applications. Thesis, Computer, Information and Control Engineering Program, University of Michigan.

[15] ROSS, S. M. (1982) *Applied Probability Models with Optimization Applications.* Holden Day, San Francisco.

[16] Ross, S. M. *Stochastic Processes*. Wiley, New York.

[17] Rosberg, Z., Varaiya, J. and Walrand, J. (1982) Optimal control of service in tandem queues. *IEEE Trans. Autom. Control* **27,** 600–610.

[18] Sauer, C. and Chandi, K. (1981) *Computer Systems Performance Modeling*. Prentice-Hall, Englewood Cliffs, NJ.

[19] Schweitzer, P. J. (1971) Iterative solution to the functional equations of undiscounted Markov renewal programming. *J. Math. Anal. Appl.* **34,** 495–501.

[20] Stidham, S., jr (1982) Optimal control of arrivals to queues and networks of queues. 21st IEEE Conf. Decision and Control, Orlando, Fl.

[21] White, D. J. (1972) Dynamic programming and probabilistic constraints. *Operat. Res.* **22,** 654–664.