Optimality of Monotonic Policies for Two-Action Markovian Decision Processes, with Applications to Control of Queues with Delayed Information

Eitan Altman
INRIA, Centre Sophia Antipolis
2004, rte. des Lucioles
B.P. 93
06902 Sophia Antipolis Cedex
FRANCE

Shaler Stidham Jr.
Department of Operations Research
CB 3180, Smith Building
University of North Carolina
Chapel Hill, NC 27599-3180
USA

March 8, 1994;
revised, February 27, 1995 and June 15, 1995

Abstract

We consider a discrete-time Markov decision process with a partially ordered state space and two feasible control actions in each state. Our goal is to find general conditions, which are satisfied in a broad class of applications to control of queues, under which an optimal control policy is monotonic. An advantage of our approach is that it easily extends to problems with both information and action delays, which are common in applications to high-speed communication networks, among others. The transition probabilities are stochastically monotone and the one-stage reward submodular. We further assume that transitions from different states are coupled, in the sense that the state after a transition is distributed as a deterministic function of the current state and two random variables, one of which is controllable and the other uncontrollable. Finally, we make a monotonicity assumption about the sample-path effect of a pairwise switch of the actions in consecutive stages. Using induction on the horizon length, we demonstrate that optimal policies for the finite- and infinite-horizon discounted problems are monotonic. We apply these results to a single queueing facility with control of arrivals and/or services, under very general conditions. In this case, our results imply that an optimal control policy has threshold form. Finally, we show how monotonicity of an optimal policy extends in a natural way to problems with information and/or action delay, including delays of more than one time unit. Specifically, we show that, if a problem without delay satisfies our sufficient conditions for monotonicity of an optimal policy, then the same problem with information and/or action delay also has monotonic (e.g., threshold) optimal policies.

Keywords: Markov Decision Processes, Monotone policies, Delayed information and actions, Submodularity.

1 Introduction

A common approach in the literature on optimal control of queueing systems, as well as related applications in production/inventory, reservoirs, and replacement/maintenance, is to formulate the control problem as a Markovian decision process. The state variable is typically the number of customers or the workload in the system, or a vector of quantities at the nodes in a network of queues. At certain points in time (e.g., arrival or service-completion points) one observes the state of the system and chooses an action, which might be whether to accept or reject an arrival, which class of customer to serve next, how fast to serve or how many servers to use, or which facility or route through the network to assign a customer to. The objective is usually to maximize (minimize) the expected discounted or average benefit (cost) of operating the system over a finite or infinite horizon.

Formulation as a Markovian decision process allows explicit consideration of the trade-off between immediate and expected future benefits when choosing the current action. The optimality equations of dynamic programming allow one (in principle) to compute the optimal value function and an associated optimal control policy by backward recursion (induction on the number of stages, or observation points, remaining in the horizon). More importantly, this approach can often be used as an analytical tool to establish that the objective function at each stage has certain properties (e.g., submodularity) as a function of state and action, which guarantee that an optimal policy has a certain simple form. For example, an optimal policy for acceptance/rejection of arriving customers may be characterized by a *critical number* or *threshold* such that an arrival is accepted iff the number of customers already present is below the threshold. Such a result greatly simplifies the numerical analysis involved in the search for an optimal policy.

When this inductive approach has been used in the literature, it has usually involved showing that the optimal value function has certain convexity, concavity, and/or multimodularity properties, which in turn ensure that the objective function at each stage has the required submodularity or related properties. For surveys of early work in which this approach is applied to problems involving control of a single queueing facility, see Sobel [14], Stidham and Prabhu [16], Crabill, Gross, and Magazine [5], Serfozo [12, 13], Hinderer [8]. For discrete-state problems this approach has been extended to networks of queues, but with limited success (mainly for networks of two nodes) and frequently at the cost of tedious case-by-case analyses. Among the most general and successful applications of this approach are Weber and Stidham [20], Veatch and Wein [19], and Glasserman and Yao [6, 7]. (For surveys of Markovian decision process models for control of networks of queues, see Stidham [15] and Stidham and Weber [17].)

Despite these successes, a significant number of important queueing-control problems have not completely yielded to this convexity/submodularity approach. These include certain problems with a non-discrete state space, some partially observable models, and problems with information and/or action delays. Both these types of delay are common in applications, especially in high-speed communication networks. They are due to the fact that the controller (e.g., a flow controller) may be located quite far away from nodes of the network which have an important influence on the performance measures. The information delay occurs because, by the time the information arrives to the controller, the state of the system may have changed considerably, as the speed at which information travels is non-negligible with respect to transmission times. This is especially true in the case of an end-to-end flow-control mechanism (such as the *TCP/IP*), where the controller can make inferences about the state of the network only after an acknowledgement returns from the destination. The action delay is due to similar physical considerations; if, for example, a flow controller is located sufficiently far from a downstream node through which the traffic flows, then any decision taken by the controller will influence the state of that node only after a time interval that corresponds to the propagation delay.

Altman and Nain [2] have studied a discrete-time model for control of arrivals to a single queueing facility in which arrival information is delayed by one time unit, showing that an optimal policy is characterized by thresholds. Artiges [3] has extended this analysis to a problem in which arriving customers are routed to one of two parallel nodes, again for the case in which information is delayed by one time unit.

In the present paper we study a general Markovian decision model with two actions. By combining the inductive, dynamic-programming technique with a stochastic-ordering approach, we are able to establish submodularity of the objective function at each stage for the finite-horizon model, without the usual intermediate step of proving that the optimal value function is concave or multimodular. Informally speaking, we make the following assumptions:

- the state space is partially ordered;
- there are two actions a available in each state x;
- the one-stage reward function is submodular in (x, a);
- the transition probabilities are stochastically monotone in x for each a;
- transitions from different states are coupled, in the sense that the state after a transition is distributed as a deterministic function of the current state and two random variables, one of which is controllable and the other uncontrollable;
- a pairwise switch of the actions in the next two stages has an effect on the total discounted future reward that is monotonic in the current state.

We state these assumptions more precisely in Section 2. With the possible exception of the last two, they are relatively weak and commonly encountered in the literature on control of queues. The coupling assumption is less standard, but not unduly restrictive. For example, in an arrival-control problem, the controllable random variable might be the number of customers admitted (from a batch) and the uncontrollable random variable might be the number of service completions between successive batch arrivals. The pairwise-switch assumption needs further explanation and illustration, which will be forthcoming in Sections 2 and 3. For now, we observe that it can often be easily verified in special cases in which the actions are *permutable*, in the sense that the sample path after the next two stages is not affected by the pairwise switch. Non-permutability usually is the result of boundary conditions, e.g., the condition that the number of customers or work in a queueing system cannot be negative. In such cases, the pairwise-switch assumption says, intuitively, that the discrepancy in future benefits and costs caused by the non-permutability diminishes as the current state moves away from the boundary.

In many respects, our approach is in the spirit of Johansen and Stidham [9], who analyzed a one-dimensional stochastic input-output system, with a continuous, real-valued state variable. Our model accommodates a multi-dimensional continuous or discrete state space with a partial order and extends in a straightforward way to the cases of both information delay and action delay, each with arbitrary delay intervals. Specifically, we show that the sufficient conditions for monotonicity of an optimal policy without delay directly imply that the same problem with information and/or action delays of an arbitrary number of time units also has monotonic optimal policies. We regard this extension to problems with delay as one of the significant advantages of our approach.

Some of our stochastic assumptions on the model, and our approach for obtaining monotonicity are related to White [21]. A discussion on the relation between our models can be found in [1].

The paper is organized as follows. In Section 2 we present the basic Markov decision model and establish our fundamental result: the optimality of policies that are monotonic (in the partial order on the state space) for the finite- and infinite-horizon discounted problems. These results are illustrated in Section 3 in the setting of a discrete-time model for a single queue, with control of arrivals and/or services. Batch arrivals and bulk service are permitted. In Section 4 we show how the monotonicity of optimal policies extends in a natural way to systems with information and/or action delay.

2 General Model and Basic Results

We consider a discrete-time Markov decision process (*DTMDP*) given by $(\mathcal{X}, \mathcal{A}, q, r, \beta)$, where

(i) $\mathcal{X}$ is the *state space*, assumed to be a standard Borel space.

(ii) $\mathcal{A} = \{0, 1\}$ is the *action space*.

(iii) $q(\cdot; x, a)$ is a probability measure on $\mathcal{X}$ for each $x \in \mathcal{X}$, $a \in \mathcal{A}$, assumed to be a measurable function of x for each $a \in \mathcal{A}$ and Borel subset B of $\mathcal{X}$. That is, $q(B; x, a)$ is the one-step *transition probability* from x into B, if action a is taken.

(iv) $r(x, a)$ is the one-step *reward* earned in state $x \in \mathcal{X}$ if action $a \in \mathcal{A}$ is taken. We assume that $r(x, a)$ is measurable and bounded above as a function of x.

(v) β is the one-step discount factor, $0 < \beta < 1$.

We define a policy $\pi = \{\pi_0, \pi_1, \ldots,\}$ in the usual way; that is, π_t is a conditional probability measure on $\mathcal{A}$ given all the history of actions and states before t, as well as the state at time t. Each (measurable) policy π induces a stochastic process $\{(X_t, A_t), t = 0, 1, \ldots\}$, where X_t and A_t are, respectively, the state and action at time (stage) t. Let P^π and E^π denote the probability measure and the expectation operator, respectively, corresponding to policy π. The

n-stage and infinite-horizon value functions,

$$V_n^\pi(x) \;:=\; E^\pi\left[\sum_{t=0}^{n-1}\beta^t r(X_t,A_t)|X_0=x\right] \;,\; x\in\mathcal{X} \;,\; n\geq 1, \tag{1}$$

$$V^\pi(x) \;:=\; E^\pi\left[\sum_{t=0}^{\infty}\beta^t r(X_t,A_t)|X_0=x\right] \;,\; x\in\mathcal{X} \;, \tag{2}$$

are well defined and bounded above by $M := (1-\beta)^{-1}\sup_{x\in\mathcal{X}}\max\{r(x,0),r(x,1)\} < \infty$. We further define $V_0^\pi(x)=0$ for all x. We assume that $V_n^\pi(x) > -\infty$ for any $x\in\mathcal{X}$ and policy π.

We consider the finite-horizon control problem:

$$\mathbf{P1(n)}: \qquad \text{find } \pi \text{ that achieves } V_n(x) \;:=\; \sup_\pi V_n^\pi(x) \;,\; x\in\mathcal{X} \;,\; n\geq 1, \tag{3}$$

and the infinite-horizon control problem:

$$\mathbf{P2}: \qquad \text{find } \pi \text{ that achieves } V(x) \;:=\; \sup_\pi V^\pi(x) \;,\; x\in\mathcal{X} \;, \tag{4}$$

The optimal value functions $V_n(x)$ and $V(x)$ are well defined, measurable and bounded above by M. They satisfy the dynamic-programming optimality equations ($x\in\mathcal{X}$):

$$V_{n+1}(x) \;=\; \max_{a=0,1}\{r(x,a)+\beta E[V_n(X_1)|X_0=x,A_0=a]\} \;,\; n\geq 0 \;, \tag{5}$$

$$V(x) \;=\; \max_{a=0,1}\{r(x,a)+\beta E[V(X_1)|X_0=x,A_0=a]\} \;, \tag{6}$$

where $V_0\equiv 0$ and

$$E[f(X_{t+1})|X_t=x,A_t=a] = \int_X q(dy;x,a)f(y) \;,\; x\in\mathcal{X}, a\in\mathcal{A}, t\geq 0 \;,$$

for $f:\mathcal{X}\to\Re$ such that the integral is well defined. A (deterministic Markov) policy that chooses an action a achieving the maximum in (5) in each state x when n stages remain, for each $n\geq 1$, is optimal for the finite-horizon problem. A (deterministic stationary) policy that chooses an action a achieving the maximum in (6) in each state x is optimal for the infinite-horizon problem. (To resolve ties, we shall choose the smaller maximizer, which is known to be measurable). Finally, $V_n(x)\to V(x)$ as $n\to\infty$, for each $x\in\mathcal{X}$.

All these statements follow from standard results in the theory of Markovian decision processes (cf. Schäl [11], Bertsekas [4]). In particular, our model satisfies all the conditions in Schäl [11]. Since rewards are bounded above and the discounting is strict, General Assumption (GA) and Condition (C) in [11] are satisfied. The continuity-compactness conditions on the action space (Condition (A) in [11]) are trivially satisfied, since the action space is finite. Alternatively, one can use Propositions 12 and 13 in Bertsekas [4] (pp. 217-218), noting that Assumption (P) (positive costs) in [4] is satisfied after transforming the problem by subtracting the upper bound from the one-stage rewards $r(x,a)$ (see p. 208 of [4]).

We assume that the state space $\mathcal{X}$ is partially ordered by a relation $\geq$. Our goal is to establish general conditions under which an optimal policy is monotonically decreasing with respect to this partial order. Specifically, for both the finite-horizon and infinite-horizon problems, we wish to show that action 0 is optimal in all states $x' \geq x$ if it is optimal in state x. To this end, we shall need several additional conditions on the *DTMDP*. The first two conditions are fundamental and used throughout the paper.

C1. The transition probability $q(\cdot;x,a)$ is *stochastically monotone* in $x \in \mathcal{X}$ for each $a \in \mathcal{A}$. That is,

$$E[f(X_1)|X_0 = x', A_0 = a] \geq E[f(X_1)|X_0 = x, A_0 = a] \ , \ x \leq x' \ , \ a \in A \ , \tag{7}$$

for all non-decreasing $f : \mathcal{X} \to R$ for which the expectations are well defined.

C2. The one-step reward function $r(x,a)$ is *submodular* in (x,a) (cf. Topkis [18]). That is, $r(x,0) - r(x,1)$ is non-decreasing in $x \in \mathcal{X}$.

Additional conditions will be introduced later as needed.

We consider in most of this section the finite horizon problem. In the end of the section we extend the results to the infinite horizon problem. We shall use induction on the number of remaining stages n to show that an optimal policy is monotonic for the finite-horizon problem. Define ($n \geq 0$, $x \in \mathcal{X}$, $a \in \mathcal{A}$)

$$J_n(x,a) := r(x,a) + \beta E[V_n(X_1)|X_0 = x, A_0 = a] \ . \tag{8}$$

In order to prove that an optimal policy for the n-stage problem is monotonically decreasing in x, it follows from (5) and (8) that it suffices to prove that $J_n(x,a)$ is submodular, that is, that $J_n(x,0) - J_n(x,1)$ is non-decreasing in $x \in \mathcal{X}$. We shall do this by induction on n, using two preliminary results (Lemmas 2.1 and 2.2), the first of which follows.

Lemma 2.1 *Let $n \geq 0$. Suppose that* **C1** *holds, that $J_n(x,a)$ is submodular, and that*

$$\xi(x) := r(x,0) + \beta E[J_n(X_1,1)|X_0 = x, A_0 = 0] - r(x,1) - \beta E[J_n(X_1,0)|X_0 = x, A_0 = 1] \quad (9)$$

is non-decreasing in $x \in \mathcal{X}$. Then $J_{n+1}(x,a)$ is submodular.

Proof: Using (5) and (8), we obtain

$$\begin{aligned}
&J_{n+1}(x,0) - J_{n+1}(x,1)\\
&\quad= r(x,0) + \beta E[V_{n+1}(X_1)|X_0 = x, A_0 = 0]\\
&\qquad -r(x,1) - \beta E[V_{n+1}(X_1)|X_0 = x, A_0 = 1]\\
&\quad= r(x,0) + \beta E[\max\{J_n(X_1,0), J_n(X_1,1)\}|X_0 = x, A_0 = 0]\\
&\qquad -r(x,1) - \beta E[\max\{J_n(X_1,0), J_n(X_1,1)\}|X_0 = x, A_0 = 1]\\
&\quad= r(x,0) + \beta[E[J_n(X_1,0) + \max\{-J_n(X_1,0) + J_n(X_1,1), 0\}|X_0 = x, A_0 = 0]]\\
&\qquad -r(x,1) - \beta[E[J_n(X_1,0) + \max\{-J_n(X_1,0) + J_n(X_1,1), 0\}|X_0 = x, A_0 = 1]]\\
&\quad= r(x,0) + \beta E[J_n(X_1,0)|X_0 = x, A_0 = 0] - r(x,1) - \beta E[J_n(X_1,0)|X_0 = x, A_0 = 1]\\
&\qquad +\beta E[\max\{-J_n(X_1,0) + J_n(X_1,1), 0\}|X_0 = x, A_0 = 0]\\
&\qquad +\beta E[\min\{J_n(X_1,0) - J_n(X_1,1), 0\}|X_0 = x, A_0 = 1]\\
&\quad= r(x,0) + \beta E[J_n(X_1,1)|X_0 = x, A_0 = 0] - r(x,1) - \beta E[J_n(X_1,0)|X_0 = x, A_0 = 1]\\
&\qquad +\beta E[J_n(X_1,0) - J_n(X_1,1)|X_0 = x, A_0 = 0]\\
&\qquad +\beta E[\max\{-J_n(X_1,0) + J_n(X_1,1), 0\}|X_0 = x, A_0 = 0]\\
&\qquad +\beta E[\min\{J_n(X_1,0) - J_n(X_1,1), 0\}|X_0 = x, A_0 = 1]\\
&\quad= r(x,0) + \beta E[J_n(X_1,1)|X_0 = x, A_0 = 0] - r(x,1) - \beta E[J_n(X_1,0)|X_0 = x, A_0 = 1]\\
&\qquad +\beta E[\max\{0, J_n(X_1,0) - J_n(X_1,1)\}|X_0 = x, A_0 = 0]\\
&\qquad +\beta E[\min\{J_n(X_1,0) - J_n(X_1,1), 0\}|X_0 = x, A_0 = 1]\ .
\end{aligned}$$

By hypothesis the first term after the last equality is non-decreasing in x. Condition **C1** and the submodularity of $J_n(x,a)$ imply that the second and third terms are also non-decreasing. Thus $J_{n+1}(x,a)$ is submodular. ∎

If we write the expression $\xi(x)$ in (9) in equivalent form as

$$\begin{aligned}\xi(x) &= E[r(x,0)+\beta r(X_1,1)+\beta^2 V_n(X_2)|X_0=x,A_0=0,A_1=1] \\ &\quad -E[r(x,1)+\beta r(X_1,0)+\beta^2 V_n(X_2)|X_0=x,A_0=1,A_1=0]\ , \qquad (10)\end{aligned}$$

we see that it represents the difference in expected total discounted reward between taking action 0 now and action 1 in the next stage, and taking action 1 now and action 0 in the next stage, assuming in both cases that we then follow an optimal policy for n stages. The hypothesis in Lemma 2.1 is that this difference should be non-decreasing in x. If this hypothesis holds for all $n \geq 0$, then Lemma 2.1 can be used as the basis for an inductive proof that $J_n(x,a)$ is submodular, provided that condition **C2** holds (to start the induction).

To gain an intuitive understanding of the significance of this hypothesis (namely, that $\xi(x)$ is non-decreasing in x), first consider the special case in which the actions $a=0$ and $a=1$ are *permutable*, in the sense that a pairwise switch of the actions taken in stages 0 and 1 does not affect the probability distribution of X_2, given $X_0 = x$. Thus,

$$E[f(X_2)|X_0=x,A_0=0,A_1=1] = E[f(X_2)|X_0=x,A_0=1,A_1=0]\ ,$$

for all $x \in \mathcal{X}$ and all functions $f : \mathcal{X} \to \Re$ such that the expectations are well defined. In this case the expression $\xi(x)$ in (10) reduces to

$$E[r(x,0)+\beta r(X_1,1)|X_0=x,A_0=0,A_1=1] - E[r(x,1)+\beta r(X_1,0)|X_0=x,A_0=1,A_1=0]\ , \qquad (11)$$

and it suffices to verify that the latter is non-decreasing in x. In applications, this can be checked directly from the properties of the one-stage reward r and transition probabilities q.

In the general, non-commutable case, proving the hypothesis that $\xi(x)$ is non-decreasing in x will be a more complicated task. In fact we shall prove by induction a stronger, sample-path version of the hypothesis, simultaneously with the inductive proof that $J_n(x,a)$ is submodular. To this end, we shall need further conditions on our Markov decision model, in particular a mechanism for coupling the evolution of the process from different starting states.

For $a = 0, 1$, let Y^a be random elements of a partially ordered standard Borel space $\mathcal{Y}$. Denote the partial order on $\mathcal{Y}$ by $\geq$. Let Z be a random element of a standard Borel space $\mathcal{Z}$. We assume that Y^0, Y^1, and Z are mutually independent.

C3. There exists a function $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$, measurable with respect to the product σ-field on $\mathcal{X} \times Y \times \mathcal{Z}$ and non-decreasing in the first argument, such that

$$q(B; x, a) = P\{X_1 \in B | X_0 = x, A_0 = a\} = P\{g(x, Y^a, Z) \in B\} \ , \ x \in \mathcal{X} \ , \ a \in \mathcal{A} \ , \tag{12}$$

for every Borel subset B of $\mathcal{X}$.

Note that condition **C3** implies condition **C1**: the transitions are stochastically monotone. Condition **C3** holds in many applications to queues and related processes, as we shall see. For example, the function g may take the form

$$g(x, y, z) = (x + y - z)^+ \ , \tag{13}$$

where x, y and z are nonnegative real numbers; x is the quantity (e.g., number of customers or workload) in the system, y is the controlled input to the system, and z is the (uncontrolled) output from the system.

In what follows we shall need some properties of *static* policies, that is, deterministic Markov policies that take actions independently of the state of the system. First we introduce some additional notation.

Let $g : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ be given. Define the functions $g_t : \mathcal{X} \times \mathcal{Y}^\infty \times \mathcal{Z}^\infty \to \mathcal{X}$ recursively by

$$\begin{aligned} g_0(x, \mathbf{y}, \mathbf{z}) \ &:= \ x \\ g_1(x, \mathbf{y}, \mathbf{z}) \ &:= \ g(x, y_0, z_0) \\ g_{t+1}(x, \mathbf{y}, \mathbf{z}) \ &:= \ g_t(g_1(x, \mathbf{y}, \mathbf{z}), \theta_1 \mathbf{y}, \theta_1 \mathbf{z}) \ , \ t \geq 1 \ , \end{aligned}$$

where $\mathbf{y} = (y_0, y_1, \ldots, y_t, \ldots)$, $\mathbf{z} = (z_0, z_1, \ldots, z_t, \ldots)$, and θ_t is the shift operator, i.e., $\theta_t \mathbf{y} = (y_t, y_{t+1}, \ldots)$, $\theta_t \mathbf{z} = (z_t, z_{t+1}, \ldots)$, $t \geq 1$. It follows by a straightforward induction on $s \geq 1$ that

$$g_{t+s}(x, \mathbf{y}, \mathbf{z}) = g_s(g_t(x, \mathbf{y}, \mathbf{z}), \theta_t \mathbf{y}, \theta_t \mathbf{z}) \ , \ s, t \geq 1 \ . \tag{14}$$

Note that the function $g_t(x, \mathbf{y}, \mathbf{z})$ depends on $\mathbf{y}$ and $\mathbf{z}$ only through their first t coordinates. We shall sometimes find it convenient to emphasize this fact by writing $g_t(x, \mathbf{y}_{t-1}, \mathbf{z}_{t-1})$ instead of $g_t(x, \mathbf{y}, \mathbf{z})$, where $\mathbf{y}_{t-1} := (y_0, y_1, \ldots, y_{t-1})$, and $\mathbf{z}_{t-1} := (z_0, z_1, \ldots, z_{t-1})$.

In the example above, in which g is given by (13), $g_t(x,\mathbf{y},\mathbf{z})$ represents the quantity in the system at stage t, given $X_0 = x$ and inputs $(y_0,\ldots,y_{t-1})$ and outputs $(z_0,\ldots,z_{t-1})$ in stages $0,\ldots,t-1$.

We shall write $\pi = (a_0, a_1, \ldots, a_t, \ldots)$ to denote the static policy that takes action a_t in stage t, $t = 0,1,\ldots$. Let $\mathbf{Y}^\pi := (Y_0^{a_0}, Y_1^{a_1}, \ldots, Y_t^{a_t}, \ldots)$ be an infinite sequence of independent random elements of $\mathcal{Y}$, with $Y_t^{a_t}$ distributed as Y^{a_t} for each $t = 0,1,\ldots$. Let $\mathbf{Z} := (Z_0, Z_1, \ldots, Z_t, \ldots)$ be an infinite sequence of independent random elements of $\mathcal{Z}$, with Z_t distributed as Z, for each $t = 0,1,\ldots$. We further assume that $\mathbf{Z}$ is independent of $\mathbf{Y}^\pi$. Let $\mathbf{Y}_t^\pi := (Y_0^{a_0}, Y_1^{a_1}, \ldots, Y_t^{a_t})$, and $\mathbf{Z}_t := (Z_0, Z_1, \ldots, Z_t)$, $t \geq 0$. Under condition **C3** it follows by induction on t that

$$
\begin{aligned}
& P^\pi\{X_1 \in B_1, X_2 \in B_2, ..., X_t \in B_t | X_0 = x\} \\
&= \; P\{g_1(x,\mathbf{Y}^\pi,\mathbf{Z}) \in B_1, g_2(x,\mathbf{Y}^\pi,\mathbf{Z}) \in B_1, ..., g_t(x,\mathbf{Y}^\pi,\mathbf{Z}) \in B_t\}, \qquad t \geq 1,
\end{aligned} \tag{15}
$$

for all Borel subsets $B_1, B_2, ..., B_t$ of $\mathcal{X}$. In particular,

$$
P^\pi\{X_t \in B | X_0 = x\} = P\{g_t(x,\mathbf{Y}^\pi,\mathbf{Z}) \in B\} = P\{g_t(x,\mathbf{Y}_{t-1}^\pi,\mathbf{Z}_{t-1}) \in B\} \ , \ t \geq 1 \ , \tag{16}
$$

for every Borel subset B of $\mathcal{X}$.

For a static policy $\pi = (a_0, a_1, \ldots, a_t, \ldots)$, with $x \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}^\infty$, and $\mathbf{z} \in \mathcal{Z}^\infty$, define

$$
R_n^\pi(x,\mathbf{y},\mathbf{z}) := \sum_{t=0}^{n-1} \beta^t r(g_t(x,\mathbf{y},\mathbf{z}), a_t) \ , n \geq 1 \ , \tag{17}
$$

and $R_0^\pi(x) := 0$ for all $x \in \mathcal{X}$. Then it follows that the total discounted reward over n stages, $n \geq 0$, under the static policy π from starting state x, is distributed as $R_n^\pi(x,\mathbf{Y}^\pi,\mathbf{Z})$.

We shall use the following additional conditions.

C4. For all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and all $z, z' \in \mathcal{Z}$,

$$
g_2(x,y,y',z,z') \geq g_2(x,y',y,z,z') \ .
$$

C5. For all $n \geq 0$, $x \in \mathcal{X}$, $y, y' \in \mathcal{Y}$, such that $y' \geq y$, and all $z, z' \in \mathcal{Z}$,

$$
\begin{aligned}
& r(x,0) + \beta r(g(x,y,z),1) + \beta^2 R_n^\pi(g_2(x,y,y',z,z'),\mathbf{y},\mathbf{z}) \\
& \quad -r(x,1) - \beta r(g(x,y',z),0) - \beta^2 R_n^\pi(g_2(x,y',y,z,z'),\mathbf{y},\mathbf{z})
\end{aligned} \tag{18}
$$

is non-decreasing in $x \in \mathcal{X}$, for every static policy $\pi = (a_0, a_1, \ldots, a_t, \ldots)$ and all possible sequences $\mathbf{y} \in \mathcal{Y}^\infty$ and $\mathbf{z} \in \mathcal{Z}^\infty$.

The expression in (18) is a sample-path analogue of the expression $\xi(x)$ in (10) and will be used in the proof that the latter is non-decreasing in x.

Considering the terms in (18) separately, it is readily seen that the following is a sufficient condition for **C5**:

C5'. (i) For all $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and all $z \in \mathcal{Z}$,

$$r(x,0) + \beta r(g(x,y,z),1) - r(x,1) - \beta r(g(x,y',z),0)$$

is non-decreasing in $x \in \mathcal{X}$;

(ii) for every $t \geq 0$, $a \in \{0,1\}$, and all $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and all $z, z' \in \mathcal{Z}$,

$$r(g_t(g_2(x,y,y',z,z'),\mathbf{y},\mathbf{z}),a) - r(g_t(g_2(x,y',y,z,z'),\mathbf{y},\mathbf{z}),a)$$

is non-decreasing in $x \in \mathcal{X}$ for all possible sequences $\mathbf{y} \in \mathcal{Y}^\infty$ and $\mathbf{z} \in \mathcal{Z}^\infty$.

Lemma 2.2 *Let $n \geq 0$, and choose any $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and $z, z' \in \mathcal{Z}$. Suppose that conditions* **C3**-**C5** *hold and that $J_k(x,a)$ is submodular for each $k = 0,1,\ldots,n$. Then*

$$\begin{aligned} & r(x,0) + \beta r(g(x,y,z),1) + \beta^2 V_n(g_2(x,y,y',z,z')) \\ & \quad - r(x,1) - \beta r(g(x,y',z),0) - \beta^2 V_n(g_2(x,y',y,z,z')) \end{aligned} \tag{19}$$

is non-decreasing in $x \in \mathcal{X}$.

Proof: For $k = 0,1,\ldots,n$, consider the expression

$$\begin{aligned} \phi_k(x) \quad := \quad & r(x,0) + \beta r(g(x,y,z),1) + \beta^2 R_k^\pi(g_2(x,y,y',z,z'),\mathbf{y},\mathbf{z}) \\ + \quad & \beta^{k+2} V_{n-k}(g_k(g_2(x,y,y',z,z'),\mathbf{y},\mathbf{z})) \\ - \quad & r(x,1) - \beta r(g(x,y',z),0) - \beta^2 R_k^\pi(g_2(x,y',y,z,z'),\mathbf{y},\mathbf{z}) \\ - \quad & \beta^{k+2} V_{n-k}(g_k(g_2(x,y',y,z,z'),\mathbf{y},\mathbf{z})) \ , \end{aligned}$$

where $\pi = (a_0, a_1, \ldots, a_t, \ldots)$ is an arbitrary static policy and $\mathbf{y} \in \mathcal{Y}^\infty$ and $\mathbf{z} \in \mathcal{Z}^\infty$ are arbitrary sequences. We shall prove that $\phi_k(x)$ is non-decreasing in x using downward induction on $k = n, n-1, \ldots, 0$. This will establish the desired result, since $\phi_k(x)$ reduces to the expression in (19) for $k = 0$. (We define $g_0(x, \cdot)$ to be equal to x and $R_0^\pi(\cdot)$ to be equal to 0.)

To start the induction, note that $\phi_k(x)$ is non-decreasing in x for $k = n$ by condition **C5**, since $\phi_k(x)$ reduces to the expression in (18) in this case. Let $0 \leq k < n$ and suppose the induction hypothesis holds for $k+1$. Consider an arbitrary static policy π and sequences $\mathbf{y} \in \mathcal{Y}^\infty$ and $\mathbf{z} \in \mathcal{Z}^\infty$. Let $x \in \mathcal{X}$, $x' \in \mathcal{X}$, with $x' > x$. To simplify the notation, let

$$\begin{aligned}
\zeta_1 &:= g_k(g_2(x', y, y', z, z'), \mathbf{y}, \mathbf{z}) \ , \ \zeta_2 := g_k(g_2(x', y', y, z, z'), \mathbf{y}, \mathbf{z}) \ , \\
\zeta_3 &:= g_k(g_2(x, y, y', z, z'), \mathbf{y}, \mathbf{z}) \ , \ \zeta_4 := g_k(g_2(x, y', y, z, z'), \mathbf{y}, \mathbf{z}) \ ,
\end{aligned}$$

and note that

$$\zeta_1 \geq \max\{\zeta_2, \zeta_3\} \geq \min\{\zeta_2, \zeta_3\} \geq \zeta_4 \ , \tag{20}$$

by conditions **C3** and **C4**. Let $\Delta := V_{n-k}(\zeta_1) - V_{n-k}(\zeta_2) - V_{n-k}(\zeta_3) + V_{n-k}(\zeta_4)$, and recall that

$$V_{n-k}(\zeta_i) = \max_{a=0,1} J_{n-k-1}(\zeta_i, a) \ , \ i = 1, 2, 3, 4 \ .$$

Case 1. Suppose the same action a is optimal in both ζ_2 and ζ_3. Then

$$\Delta \geq J_{n-k-1}(\zeta_1, a) - J_{n-k-1}(\zeta_2, a) - J_{n-k-1}(\zeta_3, a) + J_{n-k-1}(\zeta_4, a) \ .$$

Case 2. Suppose that action 0 is optimal in ζ_2 and action 1 in ζ_3. Then

$$\begin{aligned}
\Delta &\geq J_{n-k-1}(\zeta_1, 0) - J_{n-k-1}(\zeta_2, 0) - J_{n-k-1}(\zeta_3, 1) + J_{n-k-1}(\zeta_4, 1) \\
&= J_{n-k-1}(\zeta_1, 0) - J_{n-k-1}(\zeta_2, 0) - J_{n-k-1}(\zeta_3, 0) + J_{n-k-1}(\zeta_4, 0) \\
&\quad + J_{n-k-1}(\zeta_3, 0) - J_{n-k-1}(\zeta_3, 1) - J_{n-k-1}(\zeta_4, 0) + J_{n-k-1}(\zeta_4, 1) \\
&\geq J_{n-k-1}(\zeta_1, 0) - J_{n-k-1}(\zeta_2, 0) - J_{n-k-1}(\zeta_3, 0) + J_{n-k-1}(\zeta_4, 0)
\end{aligned}$$

where the second inequality follows from (20) and the assumption that $J_{n-k-1}(x, a)$ is submodular.

Case 3. Suppose that action 1 is optimal in ζ_2 and action 0 in ζ_3. Then

$$\Delta \geq J_{n-k-1}(\zeta_1, 0) - J_{n-k-1}(\zeta_2, 1) - J_{n-k-1}(\zeta_3, 0) + J_{n-k-1}(\zeta_4, 1)$$

$$
\begin{aligned}
&= J_{n-k-1}(\zeta_1,0) - J_{n-k-1}(\zeta_2,0) - J_{n-k-1}(\zeta_3,0) + J_{n-k-1}(\zeta_4,0) \\
&\quad + J_{n-k-1}(\zeta_2,0) - J_{n-k-1}(\zeta_2,1) - J_{n-k-1}(\zeta_4,0) + J_{n-k-1}(\zeta_4,1) \\
&\geq J_{n-k-1}(\zeta_1,0) - J_{n-k-1}(\zeta_2,0) - J_{n-k-1}(\zeta_3,0) + J_{n-k-1}(\zeta_4,0)
\end{aligned}
$$

where the second inequality follows from (20) and the assumption that $J_{n-k-1}(x,a)$ is submodular.

Thus in all cases we have

$$
\begin{aligned}
\Delta &\geq J_{n-k-1}(\zeta_1,a'_k) - J_{n-k-1}(\zeta_2,a'_k) - J_{n-k-1}(\zeta_3,a'_k) + J_{n-k-1}(\zeta_4,a'_k) \qquad (21) \\
&= r(\zeta_1,a'_k) + \beta E[V_{n-k-1}(g(\zeta_1,Y^{a'_k},Z))] - r(\zeta_2,a'_k) - \beta E[V_{n-k-1}(g(\zeta_2,Y^{a'_k},Z))] \\
&\quad - r(\zeta_3,a'_k) - \beta E[V_{n-k-1}(g(\zeta_3,Y^{a'_k},Z))] + r(\zeta_4,a'_k) + \beta E[V_{n-k-1}(g(\zeta_4,Y^{a'_k},Z))]\ ,
\end{aligned}
$$

where $a'_k := 0$ in Cases 2 and 3 and $a'_k := a$ (the common optimal action in ζ_2 and ζ_3) in Case 1.

Let π' be a static policy that takes the actions $(a_0, a_1, \ldots, a_{k-1})$ specified by policy π in stages $0,1,\ldots,k-1$ and then takes action a'_k in stage k. Note that $R^{\pi'}_{k+1}(\cdot,\mathbf{y},\mathbf{z}) = R^{\pi}_{k}(\cdot,\mathbf{y},\mathbf{z}) + \beta^k r(g_k(\cdot,\ \mathbf{y}_{k-1},\mathbf{z}_{k-1}),a'_k)$ and that $\mathbf{Y}^{\pi'}_k = (\mathbf{Y}^{\pi}_{k-1}, Y^{a'_k}_k)$, where $Y^{a'_k}_k$ is independent of $\mathbf{Y}^{\pi}_{k-1}$ and distributed as $Y^{a'_k}$. It follows from (21) and the definition of ζ_i, $i=1,2,3,4$, that

$$
\begin{aligned}
&\phi_k(x') - \phi_k(x) \\
&\geq r(x',0) + \beta r(g(x',y,z),1) + \beta^2 R^{\pi'}_{k+1}(g_2(x',y,y',z,z'),\mathbf{y},\mathbf{z}) \\
&\quad + \beta^{k+3} E[V_{n-k-1}(g_{k+1}(g_2(x',y,y',z,z'),\mathbf{Y}^{\pi'}_k,\mathbf{Z}_k))|\mathbf{Y}^{\pi}_{k-1} = \mathbf{y}_{k-1},\mathbf{Z}_{k-1}=\mathbf{z}_{k-1}] \\
&- r(x',1) - \beta r(g(x',y',z),0) - \beta^2 R^{\pi'}_{k+1}(g_2(x',y',y,z,z'),\mathbf{y},\mathbf{z}) \\
&\quad - \beta^{k+3} E[V_{n-k-1}(g_{k+1}(g_2(x',y',y,z,z'),\mathbf{Y}^{\pi'}_k,\mathbf{Z}_k))|\mathbf{Y}^{\pi}_{k-1} = \mathbf{y}_{k-1},\mathbf{Z}_{k-1}=\mathbf{z}_{k-1}] \\
&- r(x,0) - \beta r(g(x,y,z),1) - \beta^2 R^{\pi'}_{k+1}(g_2(x,y,y',z,z'),\mathbf{y},\mathbf{z}) \\
&\quad - \beta^{k+3} E[V_{n-k-1}(g_{k+1}(g_2(x,y,y',z,z'),\mathbf{Y}^{\pi'}_k,\mathbf{Z}_k))|\mathbf{Y}^{\pi}_{k-1} = \mathbf{y}_{k-1},\mathbf{Z}_{k-1}=\mathbf{z}_{k-1}] \\
&+ r(x,1) + \beta r(g(x,y',z),0) + \beta^2 R^{\pi'}_{k+1}(g_2(x,y',y,z,z'),\mathbf{y},\mathbf{z}) \\
&\quad + \beta^{k+3} E[V_{n-k-1}(g_{k+1}(g_2(x,y',y,z,z'),\mathbf{Y}^{\pi'}_k,\mathbf{Z}_k))|\mathbf{Y}^{\pi}_{k-1} = \mathbf{y}_{k-1},\mathbf{Z}_{k-1}=\mathbf{z}_{k-1}]\ ,
\end{aligned}
$$

where in each case the expectation is taken with respect to the distribution of $Y_k^{a'_k}$ and Z_k. The induction hypothesis implies that the above expression, with the expectations removed, is non-negative for every realization of $\mathbf{Y}_k^{\pi'} = (\mathbf{Y}_{k-1}^{\pi}, Y_k^{a'_k})$ and $\mathbf{Z}_k$. The desired result, $\phi_k(x') - \phi_k(x) \geq 0$, then follows upon replacing the expectations. ∎

In order to complete the construction of an inductive proof of the monotonicity of an optimal policy based on Lemmas 2.1 and 2.2, it remains to establish conditions under which monotonicity of the sample-path expression in (19), for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and for all $z, z' \in \mathcal{Z}$, implies monotonicity of the expression $\xi(x)$ involving expectations in (10).

Lemma 2.3 *Fix $n \geq 0$. Assume that the expression in (19) is non decreasing in $x \in \mathcal{X}$, for all $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, and for all $z, z' \in \mathcal{Z}$. If either*
(i) the support of Y^0 is smaller than the support of Y^1 (i.e., any point in the set of values that Y^0 may take, is smaller than or equal to any point in the set of values that Y^1 may take); or
(ii) $\mathcal{Y}$ is the set of real numbers and $Y^0 \leq Y^1$ in the likelihood ratio ordering (see Ross [10], p. 266),
then the expression $\xi(x)$ in (10) is non-decreasing in $x \in \mathcal{X}$ for the above n.

Proof: The proof under condition (i) is straightforward. Assume that condition (ii) holds. Fix any $x' > x \in \mathcal{X}$. Define $\psi : \mathcal{Y}^2 \times \mathcal{Z}^2 \to \mathbb{R}$ as

$$\begin{aligned}
\psi(y, y', z, z') \quad := \quad & r(x', 0) + \beta r(g(x', y, z), 1) + \beta^2 V_n(g_2(x', y, y', z, z')) \\
& -r(x', 1) - \beta r(g(x', y', z), 0) - \beta^2 V_n(g_2(x', y', y, z, z')) \\
& -r(x, 0) + \beta r(g(x, y, z), 1) + \beta^2 V_n(g_2(x, y, y', z, z')) \\
& +r(x, 1) - \beta r(g(x, y', z), 0) - \beta^2 V_n(g_2(x, y', y, z, z'))
\end{aligned}$$

and further define $\nu(y, y') = E[\psi(y, y', Z_0, Z_1)]$,

where Z_0 and Z_1 are independent and each distributed as the generic random variable Z (see definition above (15)). By assumption, $\psi(y, y', z, z') \geq 0$ whenever $y' \geq y$. Since ψ is antisymmetric in the first two arguments, this implies that $\psi(y, y', z, z') \geq \psi(y', y, z, z')$ whenever $y' \geq y$ and hence $\nu(y, y') \geq \nu(y', y)$ whenever $y' \geq y$. Since $Y^0 \leq Y^1$ in the likelihood ratio ordering, it follows by [10] (p. 268, Proposition 8.4.2) that $\nu(Y^0, Y^1) \geq_{st} \nu(Y^1, Y^0)$ and hence $E[\nu(Y^0, Y^1)] \geq E[\nu(Y^1, Y^0)]$, where Y^0 and Y^1 are independent. (Presently we shall need the mutual independence of Y^0 and Y^1, as well as their independence from Z_0 and Z_1, since they will represent in (22) below random variables corresponding to different times, which generate the dynamics of the *DTMDP* as in equation (15).) Since ν is antisymmetric, $E[\nu(Y^0, Y^1)] \geq -E[\nu(Y^0, Y^1)]$, which implies that $E[\nu(Y^0, Y^1)] \geq 0$. The proof is established then by noting that (10), condition **C3**, and (16) imply that

$$\begin{aligned} \xi(x') - \xi(x) &= E[r(x', 0) + \beta r(g(x', Y_0^0, Z_0), 1) + \beta^2 V_n(g_2(x', Y_0^0, Y_1^1, Z_0, Z_1))] \\ &\quad - E[r(x', 1) + \beta r(g(x', Y_0^1, Z_0), 1) + \beta^2 V_n(g_2(x', Y_0^1, Y_1^0, Z_0, Z_1))] \\ &\quad - E[r(x, 0) + \beta r(g(x, Y_0^0, Z_0), 1) + \beta^2 V_n(g_2(x, Y_0^0, Y_1^1, Z_0, Z_1))] \\ &\quad + E[r(x, 1) + \beta r(g(x, Y_0^1, Z_0), 1) + \beta^2 V_n(g_2(x, Y_0^1, Y_1^0, Z_0, Z_1))] \\ &= E[\nu(Y^0, Y^1)] , \end{aligned} \tag{22}$$

where the second equality follows from the fact that Y_0^a and Y_1^a are both distributed as Y^a $(a = 0, 1)$ and from the independence of $\mathbf{Y}^\pi$ and $\mathbf{Z}$. ∎

Combining Lemmas 2.1, 2.2 and 2.3 we obtain:

Theorem 2.1 *Assume* **C2**,**C3**,**C4**,**C5**, *and either (i) or (ii) of Lemma 2.3. Then for all* $n = 0, 1, \ldots,$

(a) $J_n(x, a)$ *is submodular;*

(b) the expression in (9) is non-decreasing in $x \in \mathcal{X}$.

(c) For each $n = 0, 1, \ldots,$ *there exists a deterministic Markov policy* π *that is optimal for* **P1(n)** *such that* π_k *is monotone non-increasing for all* $k = 0, 1, \ldots, n$*; thus if in state* x *at stage* k *it is optimal to choose* $a = 0$*, then* $a = 0$ *is also optimal for all* $x' \geq x$ *at stage* k.

Proof: We obtain (a) and (b) by the following induction argument. To start the induction we note that $J_0(x, a) = r(x, a)$, which is submodular by **C2**. Thus (a) holds for $n = 0$. Now the expression in (19) is non-decreasing in x for n=0 (by **C5**), and hence, by Lemma 2.3, the expression in (9) is non-decreasing in x for $n = 0$. Hence, by part (a) for $n = 0$ and Lemma 2.1, $J_1(x, a)$ is submodular. (Recall that **C3** implies **C1**.)

Now we use Lemma 2.2 for $n = 1$ and conclude that the expression in (19) is non-decreasing in x for $n = 1$. Now Lemma 2.3 implies that the expression in (9) is non-decreasing in x for $n = 1$, and hence, by Lemma 2.1, that $J_2(x, a)$ is submodular. The rest of the induction now follows similarly.

Part (c) follows from the optimality equation (5), the definition of J_n (8), and part (a). ∎

The previous results extend immediately to the infinite-horizon problem with discounting. Define ($x \in \mathcal{X}$, $a \in \mathcal{A}$)

$$J(x, a) := r(x, a) + \beta E[V(X_1)|X_0 = x, A_0 = a] \ . \tag{23}$$

In order to prove that an optimal policy for the infinite-horizon problem is monotonically decreasing in x, it follows from (6) and (23) that it suffices to prove that $J(x, a)$ is submodular.

Theorem 2.2 *Assume* **C2**,**C3**,**C4**,**C5**, *that* $J(x, a) > -\infty$ *for all* x *and* a, *and either (i) or (ii) of Lemma 2.3. Then*
(a) $J(x, a)$ *is submodular;*
(b) There exists a deterministic stationary policy that is optimal for **P2**, *which is monotonic non-increasing; thus if in state* x *it is optimal to choose* $a = 0$, *then* $a = 0$ *is also optimal for all* $x' \geq x$.

Proof: Part (a) follows from part (a) of Theorem 2.1 and the fact that $V_n(x) \to V(x)$ as $n \to \infty$, for all $x \in \mathcal{X}$. Part (b) then follows immediately from part (a). ∎

3 Control of arrivals and service in a single queue

In this section we apply the theory for our general model to a discrete-time queueing system. The state space $\mathcal{X}$ is $\{0, 1, \ldots\}$ and the state variable X_n denotes the number of customers in the system. Since the state space in this case is totaly ordered, any monotonic policy is of a threshold type, i.e., there exists a critical level x such that one action is optimal below x, and the other action above x.

Let Z_n be the uncontrolled netflow to the system at time n, that is, the total number of uncontrolled arrivals minus the number of uncontrolled departures. Similarly, let Y_n be the controlled netflow to the system at time n. The evolution of the state is thus given by

$$X_{n+1} = g(X_n, Y_n, Z_n) = (X_n + Y_n + Z_n)^+ \ , \tag{24}$$

where $(a)^+ := \max\{0, a\}$.

The control mechanism is the following. Let $\{Y_n^0, Y_n^1\}$ be two i.i.d. sequences with $Y_n^1 \geq Y_n^0$ a.s. and Y_n^0 are independent of Y_n^1. We set $Y_n = Y_n^{A_n}$, where $A_n \in \{0, 1\}$ is the action taken at time n. Thus

$$Y_n = Y_n^0 \mathbf{1}\{A_n = 0\} + Y_n^1 \mathbf{1}\{A_n = 1\} \ . \tag{25}$$

This model can correspond to service control or to flow control in a queueing system or to both simultaneously. Here are some examples:

1. Flow control. Consider K input flows, where only the first one is controlled. Let α_n^i be the (random) number of arrivals of flow type i at time n, $i = 1, \dots, K$. At time slot n, a (ranndom) number d_n customers may be served. Then $Y_n^1 = \alpha_n^1$ and $Y_n^0 = 0$. Action $a = 1$ corresponds to accepting the batch that arrives through the first flow, whereas $a = 0$ corresponds to rejecting it. The uncontrolled netflow is given by $Z_n = \sum_{i=2}^K \alpha_n^i - d_n$, Here (d_n), (α_n^1), and $(\sum_{i=2}^K \alpha_n^i)$ are independent i.i.d. sequences.

2. Service control. Again consider K input flows and let α_n^i be the number of arrivals of flow type i at time n, $i = 1, \dots, K$. None of the input flows can be controlled. At time slot n, d_n customers may be served. The service mechanism can be controlled by choosing either to serve or idle. Then $Y_n^0 = -d_n$, $Y_n^1 = 0$, and $Z_n = \sum_{i=1}^K \alpha_n^i$. Thus action 1 correspond to idling, and action 0 corresponds to serving.

Let $h : \mathcal{X} \to \mathbb{R}$, the holding cost, be an arbitrary non-decreasing convex function. The one-step reward is given by $r(x, a) = -h(x) + \gamma a$, where γ is some positive constant.

Theorem 3.1 *(i) For any $n = 0, 1, \ldots$, there exists an optimal policy for* **P1(n)** *such that π_k is a non-increasing threshold policy for all $k = 0, 1, \ldots, n$.*
(ii) There exists an optimal stationary non-increasing threshold policy for **P2**.

Proof: Parts (i) and (ii) follow from Theorem 2.1 (c) and Theorem 2.2 (b), respectively. It suffices to check that the assumptions of the theorems hold. Condition **C2** holds, since $r(x, a)$

is separable. Condition **C3** holds by the definition of the dynamics in (24). Condition (i) of Lemma 2.3 holds by the definition of Y_n. Next we establish **C4**. Indeed, we have for $y' \geq y$,

$$\begin{aligned} g_2(x, y, y', z, z') &= ((x+y+z)^+ + y' + z')^+ = \max(0, y'+z', x+y+z+y'+z') \\ &\geq \max(0, y+z', x+y+z+y'+z') = g_2(x, y', y, z, z') \ . \end{aligned}$$

Note that

$$\begin{aligned} x_t &= g_t(x, y_0, y_1, \ldots, y_{t-1}, z_0, z_1, \ldots, z_{t-1}) \\ &= \max\left\{0, \max_{1\leq j\leq t-1} \sum_{k=j}^{t-1} y_k + z_k, x + \sum_{k=0}^{t-1} y_k + z_k\right\} \end{aligned} \tag{26}$$

is the solution to the recursive (Lindley) equation $x_{t+1} = (x_t + y_t + z_t)^+$.

It remains to check **C5'**. First observe that **C5'**(i) holds by definition of r, and the fact that h is convex and non-decreasing. Fix $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $y' \in \mathcal{Y}$, such that $y' \geq y$, $z, z' \in \mathcal{Z}$, and sequences $\mathbf{y} = (y_0, y_1, \ldots, y_t, \ldots)$, $\mathbf{z} = (z_0, z_1, \ldots, z_t, \ldots)$. Denote

$$\begin{aligned} c_1 &:= \sum_{j=0}^{t-1}(y_j + z_j) + y + y' + z + z', \qquad c_2 := \sum_{j=0}^{t-1}(y_j + z_j) + z', \\ c_3 &:= \max_{0\leq j\leq t-1} \sum_{k=j}^{t-1}(y_k + z_k). \end{aligned}$$

Then we have by (26)

$$g_t(g_2(x, y, y', z, z'), \mathbf{y}, \mathbf{z}) = g_{t+2}(x, y, y', \mathbf{y}, z, z', \mathbf{z}) = \max\{c_1 + x, c_2 + y', c_3, 0\} \tag{27}$$

We show that $g_{t+2}(x, y, y', \mathbf{y}, z, z', \mathbf{z}) - g_{t+2}(x, y', y, \mathbf{y}, z, z', \mathbf{z})$ is non-increasing with x. For $x' \geq x$,

$$\begin{aligned} Q &:= h\left(g_{t+2}(x', y, y', \mathbf{y}, z, z', \mathbf{z})\right) - h\left(g_{t+2}(x', y', y, \mathbf{y}, z, z', \mathbf{z}\right) \\ &\quad -h\left(g_{t+2}(x, y, y', \mathbf{y}, z, z', \mathbf{z}\right) + g_{t+2}(x, y', y, \mathbf{y}, z, z', \mathbf{z}) \\ &= h\left(\max\{c_1 + x', c_2 + y', c_3, 0\}\right) - h\left(\max\{c_1 + x', c_2 + y, c_3, 0\}\right) \\ &\quad -h\left(\max\{c_1 + x, c_2 + y', c_3, 0\}\right) + h\left(\max\{c_1 + x, c_2 + y, c_3, 0\}\right) \leq 0. \end{aligned}$$

To get the last inequality, we distinguish between the following two cases. If the maximizer of $\{c_1 + x', c_2 + y', c_3, 0\}$ is $c_2 + y'$ then it is also the maximizer of $\{c_1 + x, c_2 + y', c_3, 0\}$ since $x' \geq x$. Hence,

$$Q = -h\left(\max\{c_1 + x', c_2 + y, c_3, 0\}\right) + h\left(\max\{c_1 + x, c_2 + y, c_3, 0\}\right) \leq 0$$

since h is non-decreasing. If the maximizer of $\{c_1 + x', c_2 + y', c_3, 0\}$ is not $c_2 + y'$, then we have

$$Q = -h\left(\max\{c_1 + x, c_2 + y', c_3, 0\}\right) + h\left(\max\{c_1 + x, c_2 + y, c_3, 0\}\right) \leq 0,$$

again by the fact that h is non-decreasing, and since $y' \geq y$. This establishes **C5'**. ∎

4 Dynamic control with delay

In this section we consider dynamic control in the presence of two types of delay:
(i) *information delay*: at time t the controller knows about the state of the system only until time $t - N_1$.
(ii) *action delay*: the action taken at time t by the controller influences the state of the system only at time $t + N_2$.
This generalizes the model in Altman and Nain [2], who consider only information delay. Let $N = N_1 + N_2$ denote the total delay.

We shall assume that the transitions are governed by random variables Y^a and Z (where a is the action); more precisely, we assume that the dynamics of the *DTMDP* satisfy assumption **C3**. It is natural to assume that the controller has complete information (without delay) on the control; more precisely, we assume that Y_s (and of course A_s), $s < t$ are known to the controller at time t. By the fact that the state is known only after some delay we shall also understand that the variables Z are available to the decision maker after a delay; indeed, if we could know the Z variables without delay then we could infer the current state of the system, and the problem would reduce trivially to the original non-delayed *DTMDP*.

We define a modified *DTMDP*: $\{\mathcal{W}, \mathcal{A}, \hat{q}, \hat{r}\}$. The state space is $\mathcal{W} = \mathcal{X} \times \hat{\mathcal{Y}}^N$, where $\hat{\mathcal{Y}} := \mathcal{Y} \times \mathcal{A}$. An element of the new modified state space has the form

$$w = (x, \hat{y}_{-N}, \ldots, \hat{y}_{-1}) \,, \tag{28}$$

where $\hat{y}_{-i} = (y_{-i}, a_{-i}) \in \hat{\mathcal{Y}}$, $i = 1, \ldots, N$. The state variable w has the following meaning: x is the last known state, i.e., the state N_1 time units ago, and a_{-i} and y_{-i} are the action a taken and the realization of the random variable Y^a, respectively, i time units ago. The state thus includes a memory (cf. e.g. [2]). We include both a_{-i} and y_{-i} in the new state description so as to ensure that **C2-C5** hold for the modified *DTMDP* (see Lemma 4.1 below). The action space is unchanged.

We define the immediate cost as $\hat{r}(w,a) := r(x, a_{-N})$, for w given by (28).

Remark. An alternative cost function would be

$$\begin{aligned} \tilde{r}(w,a) &:= E[r(X_N, a) | X_0 = x, Y_1 = y_{-N}, \ldots, Y_N = y_{-1}] \\ &= E[r(g_N(x, (y_{-N}, \ldots, y_{-1}), (Z_1, \ldots, Z_N)), a)] \,, \end{aligned} \tag{29}$$

where $Z_1, \ldots, Z_N$ are i.i.d. with $Z_n \sim Z$, $n = 1, \ldots, N$. Note that, for an infinite-horizon problem, a policy is optimal with respect to the cost related to $\hat{r}$ if and only if it is optimal with respect to the cost related to $\tilde{r}$. We elaborate on this observation in the remark at the end of this section, where we also explain the reason for prefering the immediate cost $\hat{r}$.

For $a \in \mathcal{A}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$, let $\hat{y} := (y, a)$, $\hat{z} := z$, and define the state-transition function $\hat{g}$ for the modified *DTMDP* by

$$\hat{g}(w, \hat{y}, \hat{z}) := (g(x, y_{-N}, z), (\hat{y}_{-N+1}, \ldots, \hat{y}_{-1}, \hat{y})) \,,$$

where $w = (x, \hat{y}_{-N}, \ldots, \hat{y}_{-1})$, and g is defined as in **C3** for the original *DTMDP*. The transition probabilities are given by

$$\hat{q}(B; w, a) = P\{\hat{g}(w, (Y^a, a), Z) \in B\} \,,\ w \in \mathcal{W} \,,\ a \in \mathcal{A} \,, \tag{30}$$

for every Borel subset B of $\mathcal{W}$ (cf. (12)).

We introduce the following partial order on $\mathcal{W}$. Let $w = (x, \hat{y}_{-N}, \ldots, \hat{y}_{-1})$, and let $w' = (x', \hat{y}'_{-N}, \ldots, \hat{y}'_{-1})$, where $\hat{y}_{-i} = (y_{-i}, a_{-i})$, $\hat{y}'_{-i} = (y'_{-i}, a'_{-i})$, $i = 1, \ldots, N$. We say $w \leq w'$ if, for any $z_i \in \mathcal{Z}$, $i = 1, \ldots, N$,

$$g_N(x, (y_{-N}, \ldots, y_{-1}), (z_1, \ldots, z_N)) \leq g_N(x', (y'_{-N}, \ldots, y'_{-1}), (z_1, \ldots, z_N)) \,. \tag{31}$$

In other words, $w \le w'$ if the last current (unknown) true state (as defined in the original MDP) is smaller, for every possible realization of $Z = (Z_1, \ldots, Z_N)$, when the current modified state in the new *DTMDP* is w, than when it is w'. Note that the partial order is not influenced by $(a_{-N}, \ldots, a_{-1})$ and $(a'_{-N}, \ldots, a'_{-1})$. We also introduce a partial order on $\hat{\mathcal{Y}} = \mathcal{Y} \times \mathcal{A}$: $\hat{y} = (y, a) \le (y', a') = \hat{y}'$ if $y \le y'$ in the partial order on $\mathcal{Y}$. Note also that this partial order is not influenced by a and a'.

For $y_i \in \mathcal{Y}$, define $\hat{y}_i := (y_i, a_i)$, $i = 0, 1, \ldots$, and let $\hat{\mathbf{y}} := (\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_t, \ldots)$. Similarly, for $z_i \in \mathcal{Z}$, define $\hat{z}_i := z_i$, $i = 0, 1, \ldots$, and let $\hat{\mathbf{z}} := (\hat{z}_0, \hat{z}_1, \ldots, \hat{z}_t, \ldots)$. Analogously to the definition of the state-transition functions g_t, define the functions $\hat{g}_t : \mathcal{W} \times \hat{\mathcal{Y}}^\infty \times \mathcal{Z}^\infty \to \mathcal{W}$ recursively by

$$\begin{aligned} \hat{g}_0(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}) &:= w \\ \hat{g}_1(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}) &:= \hat{g}(w, \hat{y}_0, \hat{z}_0) \\ \hat{g}_{t+1}(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}) &:= \hat{g}_t(\hat{g}_1(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}), \theta_1 \hat{\mathbf{y}}, \theta_1 \hat{\mathbf{z}}) \ , \ t \ge 1 \ . \end{aligned}$$

As in the case of the original *DTMDP*, it follows by induction on $s \ge 1$ that

$$\hat{g}_{t+s}(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}) = \hat{g}_s(\hat{g}_t(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}), \theta_t \hat{\mathbf{y}}, \theta_t \hat{\mathbf{z}}) \ , \ s, t \ge 1 \ .$$

We shall sometimes write $\hat{g}_t(w, \hat{\mathbf{y}}_{t-1}, \hat{\mathbf{z}}_{t-1})$ instead of $\hat{g}_t(w, \hat{\mathbf{y}}, \hat{\mathbf{z}})$, where $\hat{\mathbf{y}}_{t-1} := (\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{t-1})$, and $\hat{\mathbf{z}}_{t-1} := (\hat{z}_0, \hat{z}_1, \ldots, \hat{z}_{t-1})$, since the function $\hat{g}_t(w, \hat{\mathbf{y}}, \hat{\mathbf{z}})$ depends on $\hat{\mathbf{y}}$ and $\hat{\mathbf{z}}$ only through their first t coordinates.

Lemma 4.1 *Assume that* **C2-C5** *hold for the original DTMDP. Then they hold also for the modified DTMDP.*

Proof: Condition **C2** holds trivially for the modified *DTMDP*, since for any w, $w' \in \mathcal{W}$ we have

$$\hat{r}(w', 0) - \hat{r}(w', 1) = 0 = \hat{r}(w, 0) - \hat{r}(w, 1) \ .$$

To show the modified *DTMDP* satisfies condition **C3**, choose some $w = (x, \hat{y}_{-N}, \ldots, \hat{y}_{-1}) \le (x', \hat{y}'_{-N}, \ldots, \hat{y}'_{-1}) = w'$, where $\hat{y}_{-i} = (y_{-i}, a_{-i})$, $\hat{y}'_{-i} = (y'_{-i}, a'_{-i})$, $i = 1, \ldots, N$. Further choose some $\hat{y} = (y, a)$, $\hat{y}' = (y', a')$ such that $\hat{y} \le \hat{y}'$. (That is, $y \le y'$.) Let $z \in \mathcal{Z}$ and $z_i \in \mathcal{Z}$, $i = 1, \ldots, N$. Then

$$\begin{aligned} & g_N(g(x, y_{-N}, z), (y_{-N+1}, \ldots, y_{-1}, y), (z_1, \ldots, z_N)) \\ &\quad = \; g(g_N(x, (y_{-N}, \ldots, y_{-1}), (z, z_1, \ldots, z_{N-1})), y, z_N) \\ &\quad \le \; g(g_N(x', (y'_{-N}, \ldots, y'_{-1}), (z, z_1, \ldots, z_{N-1})), y', z_N) \\ &\quad = \; g_N(g(x', y'_{-N}, z), (y'_{-N+1}, \ldots, y'_{-1}, y'), (z_1, \ldots, z_N)) \, , \end{aligned}$$

where the equalities follow from the definition of the operators $\{g_t, t \ge 1\}$ and the inequality follows from the facts that $w \le w'$ in the sense of the definition (31) and that **C3** holds for the original *DTMDP*. It follows from the definition of $\hat{g}$ that

$$\hat{g}(w, \hat{y}, \hat{z}) \le \hat{g}(w', \hat{y}', \hat{z}) \text{ , for all } \hat{z} \in \mathcal{Z} \text{ ,}$$

in the partial order defined by (31). That is, **C3** holds for the modified *DTMDP*.

To verify **C4** for $N \ge 2$, first note that

$$\begin{aligned} \hat{g}_2(w, \hat{y}, \hat{y}', \hat{z}, \hat{z}') &= (g_2(x, y_{-N}, y_{-N+1}, z, z'), (\hat{y}_{-N+2}, \ldots, \hat{y}_{-1}, \hat{y}, \hat{y}')) \, , \\ \hat{g}_2(w, \hat{y}', \hat{y}, \hat{z}, \hat{z}') &= (g_2(x, y_{-N}, y_{-N+1}, z, z'), (\hat{y}_{-N+2}, \ldots, \hat{y}_{-1}, \hat{y}', \hat{y})) \, , \end{aligned}$$

and hence the inequality $\hat{g}_2(w, \hat{y}, \hat{y}', \hat{z}, \hat{z}') \ge \hat{g}_2(w, \hat{y}', \hat{y}, \hat{z}, \hat{z}')$ can be written equivalently as

$$\begin{aligned} & (g_2(x, y_{-N}, y_{-N+1}, z, z'), (\hat{y}_{-N+2}, \ldots, \hat{y}_{-1}, \hat{y}, \hat{y}')) \\ &\quad \ge \; (g_2(x, y_{-N}, y_{-N+1}, z, z'), (\hat{y}_{-N+2}, \ldots, \hat{y}_{-1}, \hat{y}', \hat{y})) \end{aligned}$$

which follows if

$$\begin{aligned} & g_N(g_2(x, y_{-N}, y_{-N+1}, z, z'), (y_{-N+2}, \ldots, y_{-1}, y, y'), (z_1, \ldots, z_N)) \\ &\quad \ge \; g_N(g_2(x, y_{-N}, y_{-N+1}, z, z'), (y_{-N+2}, \ldots, y_{-1}, y', y), (z_1, \ldots, z_N)) \end{aligned}$$

for arbitrary $(z_1, \ldots, z_N)$. The latter holds true since the original *DTMDP* satisfies condition **C4**, and since

$$\begin{aligned} & g_N(g_2(x, y_{-N}, y_{-N+1}, z, z'), (y_{-N+2}, \ldots, y_{-1}, y, y'), (z_1, \ldots, z_N)) \\ &\quad = \; g_2(g_N(x, (y_{-N}, \ldots, y_{-1}), (z, z', z_1, \ldots, z_{N-2})), y, y', z_{N-1}, z_N) \, . \end{aligned}$$

$$
\begin{aligned}
& g_N(g_2(x, y_{-N}, y_{-N+1}, z, z'), (y_{-N+2}, \ldots, y_{-1}, y', y), (z_1, \ldots, z_N)) \\
= \;& g_2(g_N(x, (y_{-N}, \ldots, y_{-1}), (z, z', z_1, \ldots, z_{N-2})), y', y, z_{N-1}, z_N) \ .
\end{aligned}
$$

For $N = 1$ condition **C4** holds by a similar argument.

Finally we check **C5**. Choose a static policy $\pi = (a_0, a_1, \ldots, a_t, \ldots)$ and define

$$
\hat{R}_n^\pi(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}) := \sum_{t=0}^{n-1} \beta^t \hat{r}(\hat{g}_t(w, \hat{\mathbf{y}}, \hat{\mathbf{z}}), a_t) \ , n \geq 1 \ , \tag{32}
$$

and $\hat{R}_0^\pi(w) := 0$ for all $w \in \mathcal{W}$.

Introduce the following vectors:

$$
\begin{aligned}
\tilde{\mathbf{a}} &:= (a_{-N}, \ldots, a_{-1}, 0, 1, a_0, a_1, \ldots) \ ; \\
\tilde{\mathbf{y}} &:= (y_{-N}, \ldots, y_{-1}, y, y', y_0, y_1, \ldots) \ ; \\
\tilde{\mathbf{z}} &:= (z, z', z_0, z_1, \ldots) \ .
\end{aligned}
$$

We first note that

$$
\begin{aligned}
& \hat{R}_n^\pi(\hat{g}_2(w, \hat{y}, \hat{y}', \hat{z}, \hat{z}'), \hat{\mathbf{y}}, \hat{\mathbf{z}}) \\
= \;& \sum_{t=0}^{n-1} \beta^t \hat{r}(\hat{g}_t(\hat{g}_2(w, \hat{y}, \hat{y}', \hat{z}, \hat{z}'), \hat{\mathbf{y}}, \hat{\mathbf{z}}), a_t) = \sum_{t=0}^{n-1} \beta^t \hat{r}(\hat{g}_{t+2}(w, (\hat{y}, \hat{y}', \hat{\mathbf{y}}), (\hat{z}, \hat{z}', \hat{\mathbf{z}})), a_t) \\
= \;& \sum_{t=0}^{n-1} \beta^t \hat{r}((g_{t+2}(x, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}), \theta_{t+2}\tilde{\mathbf{y}}), a_t) = \sum_{t=0}^{n-1} \beta^t r(g_{t+2}(x, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}), \tilde{a}_{t+2}) \ .
\end{aligned}
$$

A similar expression holds for $\hat{R}_n^\pi(\hat{g}_2(w, \hat{y}', \hat{y}, \hat{z}, \hat{z}'), \hat{\mathbf{y}}, \hat{\mathbf{z}})$. Define

$$
\begin{aligned}
\Delta_n(0,1) &:= \hat{r}(w, 0) + \beta \hat{r}(\hat{g}(w, \hat{y}, \hat{z}), 1) + \beta^2 \hat{R}_n^\pi(\hat{g}_2(w, \hat{y}, \hat{y}', \hat{z}, \hat{z}'), \hat{\mathbf{y}}, \hat{\mathbf{z}}) \\
\Delta_n(1,0) &:= \hat{r}(w, 1) + \beta \hat{r}(\hat{g}(w, \hat{y}', \hat{z}), 0) + \beta^2 \hat{R}_n^\pi(\hat{g}_2(w, \hat{y}', \hat{y}, \hat{z}, \hat{z}'), \hat{\mathbf{y}}, \hat{\mathbf{z}}) \ .
\end{aligned}
$$

Checking **C5** amounts to showing that $\Delta_n(1,0) - \Delta_n(0,1)$ is nondecreasing in w for all n. We first note that for $n \leq N-2$, $\Delta_n(0,1) = \Delta_n(1,0)$. Let

$$
\rho := g_N(x, \tilde{\mathbf{y}}, \tilde{\mathbf{z}}) = g_N(x, (y_{-N}, \ldots, y_{-1}), (z, z', z_0, \ldots, z_{N-3})) \ .
$$

For $n = N - 1$ we have:

$$\begin{aligned}\Delta_n(0,1) - \Delta_n(1,0) &= (\Delta_{N-2}(0,1) + \beta^{N-2} r(\rho, 0)) - (\Delta_{N-2}(1,0) + \beta^{N-2} r(\rho, 1)) \\ &= \beta^{N-2}(r(\rho,0) - r(\rho,1)) \ .\end{aligned}$$

This expression is indeed nondecreasing in w, since ρ is nondecreasing in w by the definition of the ordering (31) and because condition **C2** holds for the original *DTMDP*.

For $n > N - 1$ we have:

$$\begin{aligned}\Delta_n(0,1) = \Delta_{N-2}(0,1) &+ \beta^{N-2}[r(\rho,0) + \beta r(g_1(\rho, y, z_{N-2}), 1) \\ &+ \beta^2 R_{n-N+2}(g_2(\rho, y, y', z_{N-2}, z_{N-1}), \theta_{N+2}\tilde{\mathbf{y}}, \theta_{N+2}\tilde{\mathbf{z}})] \ , \\ \Delta_n(1,0) = \Delta_{N-2}(1,0) &+ \beta^{N-2}[r(\rho,1) + \beta r(g_1(\rho, y', z_{N-2}), 0) \\ &+ \beta^2 R_{n-N+2}(g_2(\rho, y', y, z_{N-2}, z_{N-1}), \theta_{N+2}\tilde{\mathbf{y}}, \theta_{N+2}\tilde{\mathbf{z}})] \ .\end{aligned}$$

In this case $\Delta_n(0,1) - \Delta_n(1,0)$ is indeed nondecreasing in w, since ρ is nondecreasing in w and condition **C5** holds for the original *DTMDP*. ∎

As an immediate corollary we get:

Theorem 4.1 *Assume that* **C2-C5** *hold for the original* DTMDP *and either (i) or (ii) of Lemma 2.3 holds for the original* DTMDP*. Then there exists an optimal Markov deterministic policy π for the finite-horizon modified* DTMDP *such that π_k is monotonic in $\mathcal{W}$ for all $k = 0, 1, ..., n$; thus if at state w it is optimal to choose $a = 0$ at stage k, then $a = 0$ is also optimal at all $w' \geq w$ at stage k. Moreover, there exists a deterministic stationary policy that is optimal for the infinite-horizon modified* DTMDP*, which is monotonic in $\mathcal{W}$; thus if at state w it is optimal to choose $a = 0$, then $a = 0$ is also optimal for all $w' \geq w$.*

As an application of Theorem 4.1, we may consider the problem considered in Theorem 3.1. It follows that under the assumptions of that theorem, the optimal policy is monotonic also in the case of the delayed information structure of this section.

Remark. If the immediate cost were defined as in (29) then Lemma 4.1 need not hold. More precisely, **C5** need not hold for the modified *DTMDP*. To see where things go wrong, it suffices to concentrate on the case $n = 0$. We have

$$
\begin{aligned}
\Delta_0(0,1) &= \tilde{r}(w,0) + \beta\tilde{r}(\hat{g}(w,\hat{y},\hat{z}),1) \\
&= E[r(g_N(x,(y_{-N},\ldots,y_{-1}),(Z_0,\ldots,Z_{N-1})),0) \\
&\quad +\beta r(g_{N+1}(x,(y,y_{-N},\ldots,y_{-1}),(z,Z_0,\ldots,Z_{N-1})),1)] \\
&\neq E[r(g_N(x,(y_{-N},\ldots,y_{-1}),(Z_0,\ldots,Z_{N-1})),0) \\
&\quad +\beta r(g(g_N(x,(y_{-N},\ldots,y_{-1}),(Z_0,\ldots,Z_{N-1})),y,z),1)] \ ,
\end{aligned}
$$

since y and $(y_{-N},\ldots,y_{-1})$, and z and $(Z_0,\ldots,Z_{N-1})$, need not commute. This fact makes it impossible to compare $\Delta_0(0,1)$ to $\Delta_0(1,0)$, and hence to deduce that **C5** holds for the modified *DTMDP* from the fact that it holds for the original one.

Acknowledgement. We are grateful to Professor Karl Hinderer for his constructive and very helpful comments on an earlier version of this paper.

REFERENCES

[1] E. Altman, S. Foss, E. Riehl, and S. Stidham. Optimality of monotonic policies for two-action markovian decision processes, including information and action delays. Technical Report No. UNC/OR/TR-94-2, Department of Operations Research, University of North Carolina, Chapel Hill, 1994.

[2] E. Altman and P. Nain. Closed-loop control with delayed information. INRIA Report No. 1638, PERFORMANCE '92, Newport, USA., 1992.

[3] D. Artiges. Routing to parallel servers with delay. Research Report, INRIA, Sophia Antipolis, France, 1993.

[4] D. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models.* Prentice Hall, Englewood Cliffs, New Jersey, 1987.

[5] T. Crabill, D. Gross, and M. Magazine. A classified bibliography of research on optimal design and control of queues. *Operations Research*, 25:219–232, 1977.

[6] P. Glasserman and D. Yao. Monotone optimal control of permutable GSMPs. *Mathematics of Operations Research*, 19:449–476, 1994.

[7] P. Glasserman and D. Yao. *Monotone Structure in Discrete-Event Systems.* Wiley, New York, 1994.

[8] K. F. Hinderer. On the structure of solutions of stochastic dynamic programs. In M. Iosifescu, editor, *Proceedings of the Seventh Conference on Probability Theory*, pages 173–182, Bucharest, 1984. Editura Academiei Republicii Socialiste România.

[9] S.G. Johansen and S. Stidham. Control of arrivals to a stochastic input-output system. *Advances in Applied Probability*, 12:972–999, 1980.

[10] S.M. Ross. *Stochastic Processes.* John Wiley, New York, 1983.

[11] M. Schäl. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *Z. Wahrscheinlichkeitstheorie verw. Gerb.*, 32:179–196, 1975.

[12] R. Serfozo. Monotone optimal policies for Markov decision processes. In R. Wets, editor, *Stochastic Systems, II: Optimization*, volume 6, pages 202–215, Amsterdam, New York, 1976. North-Holland. Mathematical Programming Studies.

[13] R. Serfozo. Optimal control of random walks, birth and death processes, and queues. *Advances in Applied Probability*, 13:61–83, 1981.

[14] M.J. Sobel. Optimal operation of queues. In A.B. Clarke, editor, *Mathematical Methods in Queueing Theory*, volume 98, pages 145–162, Berlin, 1974. Springer-Verlag. Lecture Notes in Economics and Mathematical Systems.

[15] S. Stidham. Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control*, 30:705–713, 1985.

[16] S. Stidham and N.U. Prabhu. Optimal control of queueing systems. In A.B. Clarke, editor, *Mathematical Methods in Queueing Theory*, volume 98, pages 263–294, Berlin, 1974. Springer-Verlag. Lecture Notes in Economics and Mathematical Systems.

[17] S. Stidham and R. Weber. A survey of Markov decision models for control of networks of queues. *Queueing Systems: Theory and Applications*, 13:291–314, 1993.

[18] D. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26:305–321, 1978.

[19] M. Veatch and L. Wein. Monotone control of queueing and production/inventory systems. *Queueing Systems: Theory and Applications*, 12:391–408, 1992.

[20] R. Weber and S. Stidham. Control of service rates in networks of queues. *Advances in Applied Probability*, 24:202–218, 1987.

[21] C.C. White. Monotone control laws for noisy, countable-state Markov chains. *European Journal of Operational Research*, 5:124–132, 1980.