Proceedings of the 30th Conference
on Decision and Control
Brighton, England • December 1991

OPTIMAL INDEX POLICIES FOR MDPS WITH A CONSTRAINT

Armand M. Makowski[1] and Adam Shwartz[2]

ABSTRACT

Many controlled queueing systems possess simple index-type optimal policies, when discounted, average or finite-time cost criteria are considered. This structural results makes the computation of optimal policies relatively simple. Unfortunately, for constrained optimization problems, the index structure of the optimal policies is in general not preserved. As a result, computing optimal policies for the constrained problem appears to be a much more difficult task. We provide a framework under which the solution of the constrained optimization problem uses the same index policies as the non-constrained problem. The method is applicable to the discrete-time Klimov system, which is shown to be equivalent to the open bandit problem.

1. INTRODUCTION

The search for optimal policies in Markov decision processes (MDPs) is usually carried out in two steps. First, one shows that an optimal policy can be found in a small subclass of the admissible policies, e.g., the subclass of stationary policies. Then, one identifies an optimal policy within this subclass.

Let $J_c(\pi)$ and $J_d(\pi)$ be two cost functions, associated with one-step cost functions c and d when the system is operated under the policy π. We consider the constrained optimization problem $(\mathbf{P_V})$: Minimize $J_c(\pi)$ subject to the constraint $J_d(\pi) \leq V$ for some given V. The first such problem was solved for the finite-horizon cost by Derman and Klein [8]. Under the average cost criterion, the existence of optimal stationary policies under multiple constraints was established by Derman and Veinott [9], and by Hordijk and Kallenberg [10], both for finite state S and action space U, including the multi-class case. Under a single class assumption and for a single constraint, the existence of optimal stationary policies which are randomized at a single state was proved by Beutler and Ross [6] for finite S and compact U, and by Sennott [18] for countable S and compact U. Borkar [7] has obtained analogous results under multiple constraints when S is countable and U is compact, and has indicated similar results for other cost criteria. The multiple constraint case for countable S and countable U is treated by Altman and Shwartz [2]. In [3], Altman and Shwartz prove the existence of optimal policies for finite S and U under the discounted and other cost criteria, under multiple constraints; they also present computational algorithms for these optimal policies.

Research supported partially through NSF Grant NSF ECS-83-51836 and partially through United States—Israel Binational Science Foundation Grant BSF 85-00306.

[1] Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, MD 20742, U.S.A.

[2] Electrical Engineering Department, Technion—Israel Institute of Technology, Haifa 32000, Israel. This research was done in part while the author was visiting the Systems Research Center, University of Maryland, College Park, MD 20742, U.S.A.

Unfortunately, except for the finite case and the specific example in [1], there are no efficient methods for the computation of optimal policies. In this note we identify some structural properties which simplify this computation considerably by reducing it to a finite number of policies and to the evaluation of a single parameter in the interval $[0,1]$. We show that these structural assumptions hold, for example, for open bandit processes [12,20,21] and for a single-server network of queues. We deal in particular with the finite, discounted and the average cost criteria.

In Section 2 the general model is introduced. In Section 3 an axiomatic formulation is given, under which constrained-optimal policies retain the index structure. The assumptions of Section 3 are then verified in Section 4 for the cases of finite, discounted and average cost criteria. In Section 5 an equivalence between the discrete-time Klimov problem [11] and arm-acquiring bandits [20] is established; these systems are then shown to fall into the present framework.

A few words on the notation and conventions used in this paper: For any set E endowed with a topology, measurability always means Borel measurability and the corresponding Borel σ-field, i.e., the smallest σ-field on E generated by the open sets of the topology, is denoted by $\mathcal{B}(E)$. Unless otherwise stated, $\lim_n$, $\underline{\lim}_n$ and $\overline{\lim}_n$ are taken with n going to infinity. Moreover, the infimum over an empty set is taken to be ∞ by convention.

2. THE GENERAL MODEL

To set up the discussion, we start with a MDP (S,U,P) as defined in the literature [16, 17, 21]. The state space S and the action space U are assumed to be Polish spaces; the one-step transition mechanism P is defined through a family $(Q(x,u;dy))$ of measurable transition kernels. The state process $\{X_t,\ t=0,1,\ldots\}$ and the control process $\{U_t,\ t=0,1,\ldots\}$ are defined on some measurable space $(\Omega,\mathcal{F})$ (which for sake of concreteness is taken to be the canonical space $S\times(U\times S)^\infty$). The feedback information available to the decision-maker is encoded through the random variables (rvs) $\{H_t,\ t=0,1,\ldots\}$ defined by $H_0 \triangleq X_0$ and by $H_t \triangleq (X_0,U_0,X_1,\ldots,U_{t-1},X_t)$ for all $t=1,2,\ldots$. The rvs X_t, U_t and H_t take values in S, U and $\mathbb{H}_t \triangleq S\times(U\times S)^t$, respectively, for all $t=0,1,\ldots$, and the information σ-field $\mathcal{F}_t$ is defined by $\mathcal{F}_t=\sigma\{H_t\}$.

The space of probability measures on $(U,\mathcal{B}(U))$ is denoted by $\mathbb{M}(U)$. An admissible control policy π is defined as any collection $\{\pi_t,\ t=0,1,\ldots\}$ of mappings $\pi_t:\mathbb{H}_t\to\mathbb{M}(U)$ such that for all $t=0,1,\ldots$ and every Borel subset B of U, the mapping $\mathbb{H}_t\to[0,1]: h_t\to\pi_t(h_t;B)$ is Borel measurable. The collection of all such admissible policies is denoted by $\mathcal{P}$.

Let μ be a given probability measure on $(S,\mathcal{B}(S))$. The definition of the MDP (S,U,P) then postulates the existence of a collection of probability measures $\{\mathbf{P}^\pi,\ \pi\in\mathcal{P}\}$ on $(\Omega,\mathcal{F})$ such that conditions (2.1)–(2.2) below are satisfied: For every admissible policy π in $\mathcal{P}$, the probability measure $\mathbf{P}^\pi$ is constructed so that under $\mathbf{P}^\pi$, the rv X_0 has probability distribution μ, the control actions are selected according to

$$\mathbf{P}^\pi[U_t\in B\,|\,\mathcal{F}_t]=\pi_t(H_t;B),\quad B\in\mathcal{B}(U)$$

$$t=0,1,\ldots \quad (2.1)$$

and the state transitions are realized according to

$$\mathbf{P}^\pi[X_{t+1} \in A | \mathcal{F}_t \vee \sigma\{U_t\}] = Q(X_t, U_t; A), \quad A \in \mathcal{B}(S).$$

$$t = 0, 1, \ldots \quad (2.2)$$

The expectation operator associated with $\mathbf{P}^\pi$ is denoted by $\mathbf{E}^\pi$.

Following standard usage, a policy π in $\mathcal{P}$ is said to be a Markov policy if there exists a family $\{g_t,\ t = 0, 1, \ldots\}$ of Borel mappings $g_t : S \to \mathbb{M}(U)$ such that $\pi_t(\cdot; H_t) = g_t(\cdot; X_t)$ $\mathbf{P}^\pi$–a.s. for all $t = 0, 1, \ldots$. In the event the mappings $\{g_t,\ t = 0, 1, \ldots\}$ are all identical to a given mapping $g : S \to \mathbb{M}(U)$, the Markov policy is termed stationary and is identified with the mapping g itself.

3. A GENERAL CONSTRAINED MDP

We interpret any Borel mapping $c : S \times U \to \mathbb{R}$ as a one-step cost function. In order to avoid unnecessary technicalities we always assume c to be bounded below. In fact, as will be apparent from the discussion, there is no loss of generality in assuming $c \geq 0$, as we do from now on. For any policy π in $\mathcal{P}$, we define $J_c(\pi)$ as the total cost (associated with c) for operating the system under policy π. Several choices are possible and include the long–run average cost

$$J_c(\pi) \triangleq \overline{\lim}_t\, \mathbf{E}^\pi \left[\frac{1}{t+1} \sum_{s=0}^{t} c(X_s, U_s) \right], \quad (3.1)$$

the infinite-horizon β-discounted cost

$$J_c(\pi) \triangleq \mathbf{E}^\pi \left[\sum_{s=0}^{\infty} \beta^s c(X_s, U_s) \right], \quad 0 < \beta < 1 \quad (3.2)$$

and the finite-horizon β-discounted cost

$$J_c(\pi) \triangleq \mathbf{E}^\pi \left[\sum_{s=0}^{T} \beta^s c(X_s, U_s) \right], \quad 0 < \beta \leq 1,\ T = 1, 2, \ldots . \quad (3.3)$$

The definitions (3.1)–(3.3) are all well posed under the non-negativity assumption on c.

Now, we consider two Borel mappings $c, d : S \times U \to \mathbb{R}_+$ and for some scalar V, we set

$$\mathcal{P}_V := \{\pi \in \mathcal{P} : J_d(\pi) \leq V\}. \quad (3.4)$$

The corresponding *constrained* optimization problem $(\mathbf{P_V})$ is now formulated as

$$(\mathbf{P_V}) : \quad \text{Minimize} \quad J_c(\cdot) \quad \text{over} \quad \mathcal{P}_V.$$

Implicit in this formulation is the fact that the cost criteria $J_c(\cdot)$ and $J_d(\cdot)$ are of the same type.

For every θ in $[0, 1]$, we define the mapping $c_\theta : S \times U \to \mathbb{R}_+$ by

$$c_\theta(x, u) \triangleq \theta c(x, u) + (1 - \theta) d(x, u), \quad x \in S, u \in U. \quad (3.5)$$

We simplify the notation by using $J_\theta(\pi)$ to denote the total cost associated with c_θ under policy π, whence $J_\theta(\pi) = J_c(\pi)$ for $\theta = 1$ and $J_\theta(\pi) = J_d(\pi)$ for $\theta = 0$. The discussion is given under the following general assumptions **(A1)**, where

(A1) There exists a finite number of Markov stationary policies $g_1, \ldots, g_L$ such that

(A1.a) For each $\ell = 1, \ldots, L$, the mapping $\theta \to J_\theta(g_\ell)$ is continuous on $[0, 1]$; and

(A1.b) The condition

$$\inf_{\pi \in \mathcal{P}} J_\theta(\pi) = \min_{1 \leq \ell \leq L} J_\theta(g_\ell) \triangleq J^*(\theta), \quad \theta \in [0, 1]. \quad (3.6)$$

holds true.

It is plain that under **(A1)**, the mapping $\theta \to J^*(\theta)$ is continuous on $[0, 1]$. As in [1], we call the problem of minimizing $J_\theta(\cdot)$ over the unconstrained set of policies $\mathcal{P}$ the *Lagrangian problem.* We define

$$N(\theta) \triangleq \{\ell \in \{1, \ldots, L\} : J_\theta(g_\ell) = J^*(\theta)\}, \quad \theta \in [0, 1]. \quad (3.7)$$

Using **(A1)** we readily obtain the following properties: For each θ in $[0, 1]$, the index set $N(\theta)$ is always non-empty by virtue of **(A1.b)**, and for each ℓ in $N(\theta)$,

$$\lim_{\tilde\theta \uparrow \theta} J_{\tilde\theta}(g_\ell) = \lim_{\tilde\theta \downarrow \theta} J_{\tilde\theta}(g_\ell) = J^*(\theta). \quad (3.8)$$

Furthermore, if $N(\theta)$ reduces to a singleton, then $N(\theta) = N(\tilde\theta)$ in some open neighborhood of θ.

To proceed, we set

$$n(\theta) \triangleq \min\{n \in N(\theta) : J_d(g_n) = \min_{\ell \in N(\theta)} J_d(g_\ell)\}, \quad \theta \in [0, 1]. \quad (3.9)$$

If $J_0(g_{n(0)}) = J_d(g_{n(0)}) > V$, then the problem $(\mathbf{P_V})$ is not feasible and therefore possesses no solution.

Assuming feasibility from now on, we set

$$\theta^* \triangleq \sup\{\theta \in [0, 1] : J_d(g_{n(\theta)}) \leq V\}. \quad (3.10)$$

If $\theta^* = 0$, then necessarily $J_d(g_{n(0)}) \leq V$, but we may have to entertain the possibility that

$$\min\{J_c(g_\ell) :\ 1 \leq \ell \leq L,\ J_d(g_\ell) \leq V\} > \inf_{\pi \in \mathcal{P}_V} J_c(\pi)$$

since the Lagrangian problem may not provide enough information.

If $\theta^* = 1$, then $(\mathbf{P_V})$ has a solution: Indeed, let $\theta_i \uparrow 1$ in $(0, 1]$ so that $J_d(g_{n(\theta_i)}) \leq V$ for all $i = 1, 2, \ldots$ by the definition of θ^*. A converging subsequence, say $\theta_j \uparrow 1$, can always be selected so that $n(\theta_j) \to n^*$ for some n^* in $\{1, \ldots, L\}$. In fact, we can assert $n(\theta_j) = n^*$ whenever $j \geq j^*$ for some j^*. It is plain that n^* is an element of $N(\theta_j)$ for $j \geq j^*$, whence $J_{\theta_j}(g_{n^*}) = J^*(\theta_j)$. The continuity of $\theta \to J^*(\theta)$ implies that n^* is an element of $N(1)$, and since $J_d(g_{n^*}) \leq V$, we conclude that the policy g_{n^*} solves $(\mathbf{P_V})$.

From now on, assume $0 < \theta^* < 1$. Let $\theta_i \downarrow \theta^*$ in $(0, 1)$ and denote by $\overline{n}$ an accumulation point of the sequence $\{n(\theta_i),\ i = 1, 2, \ldots\}$. Similarly, let $\theta_j \uparrow \theta^*$ in $(0, 1)$ and denote by $\underline{n}$ an accumulation point of $\{n(\theta_j),\ j = 1, 2, \ldots\}$ such that $J_d(g_{n(\theta_j)}) \leq V$; if $J_d(g_{n(\theta^*)}) \leq V$, we set $\theta_j = \theta^*$ for all $j = 1, 2, \ldots$, in which case $n(\theta_j) = n(\theta^*) = \underline{n}$. Again, we have $n(\theta_i) = \overline{n}$ and $n(\theta_j) = \underline{n}$ for all i and j large enough. By **(A1.a)**, we see that both $\overline{n}$ and $\underline{n}$ are elements of $N(\theta^*)$, so that the equalities

$$J_{\theta^*}(g_{\underline{n}}) = J_{\theta^*}(g_{\overline{n}}) = J^*(\theta^*) \quad (3.11)$$

must hold. Moreover, it is plain that

$$J_d(g_{\underline{n}}) \leq V \leq J_d(g_{\overline{n}}). \quad (3.12)$$

The first inequality follows by construction and **(A1.a)**, whereas the second inequality results from the construction and (3.9)–(3.10).

Next, we define the policies $\underline{g}$, $\overline{g}$ and $\{g^\eta,\ 0 \le \eta \le 1\}$, as the Markov stationary policies given by

$$\underline{g} \triangleq g_{\underline{n}}, \quad \overline{g} \triangleq g_{\overline{n}} \tag{3.13}$$

and

$$g^\eta \triangleq \eta \underline{g} + (1-\eta)\overline{g}, \quad \eta \in [0,1]. \tag{3.14}$$

Then g^η is the simple randomization between the two policies $\underline{g}$ and $\overline{g}$ with randomization bias η. The identities (3.11)–(3.12) now take the form

$$J_{\theta^*}(\underline{g}) = J_{\theta^*}(\overline{g}) = J^*(\theta^*) \tag{3.15}$$

and

$$J_d(\underline{g}) \le V \le J_d(\overline{g}). \tag{3.16}$$

At this point, we can introduce the condition **(A2)**.

(A2) The mapping $\eta \to J_d(g^\eta)$ is continuous on $[0,1]$.

Lemma 1. *Under* **(A1)–(A2)**, *the equation*

$$J_d(g^\eta) = V, \quad \eta \in [0,1] \tag{3.17}$$

has a solution η^*.

Proof. This is immediate from the fact that the mapping $\eta \to J_d(g^\eta)$ is continuous on $[0,1]$ and from the inequality (3.16) which can written as

$$J_d(g^1) \le V \le J_d(g^0). \tag{3.18}$$

■

We further assume that conditions **(A3)–(A5)** are enforced, where

(A3) The equality

$$J_{\theta^*}(g^\eta) = J_{\theta^*}(\underline{g}), \quad \eta \in [0,1] \tag{3.19}$$

holds;

(A4) The equality

$$J_{\theta^*}(g^{\eta^*}) = \theta^* J_c(g^{\eta^*}) + (1-\theta^*) J_d(g^{\eta^*}) \tag{3.20}$$

holds; and

(A5) For every admissible policy π in $\mathcal{P}$, the inequality

$$J_{\theta^*}(\pi) \le \theta^* J_c(\pi) + (1-\theta^*) J_d(\pi) \tag{3.21}$$

holds.

Theorem 2. *Under* **(A1)–(A5)**, *the policy* g^{η^*} *(where* η^* *is a solution of (3.17)) solves the constrained problem* $(\mathbf{P_V})$ *provided* $\theta^* > 0$.

Proof. We first note that

$$\begin{aligned} J^*(\theta^*) &= J_{\theta^*}(g^{\eta^*}) && (3.22)\\ &= \theta^* J_c(g^{\eta^*}) + (1-\theta^*) J_d(g^{\eta^*}) && (3.23)\end{aligned}$$

where (3.22) follows from (3.15) and **(A3)**, whereas (3.23) is validated by **(A4)**. On the other hand, we have

$$J_{\theta^*}(\pi) \ge J^*(\theta^*), \quad \pi \in \mathcal{P} \tag{3.24}$$

by virtue of **(A1.b)**, and

$$J_{\theta^*}(\pi) \le \theta^* J_c(\pi) + (1-\theta^*) J_d(\pi), \quad \pi \in \mathcal{P} \tag{3.25}$$

by invoking **(A5)**. By Lemma 1, the policy g^{η^*} is an element of $\mathcal{P}_V$ since $J_d(g^{\eta^*}) = V$ by construction, and upon combining (3.22)–(3.25), we get

$$\theta^* J_c(\pi) + (1-\theta^*) J_d(\pi) \ge J_{\theta^*}(\pi) \ge \theta^* J_c(g^{\eta^*}) + (1-\theta^*) V \tag{3.26}$$

for all π in $\mathcal{P}$. It is now plain from (3.26) that

$$\theta^* J_c(g^{\eta^*}) \le \theta^* J_c(\pi), \quad \pi \in \mathcal{P}_V \tag{3.27}$$

and the result follows since $\theta^* > 0$. ■

Theorem 2 and its proof remain unchanged if **(A2)** is replaced by the conclusion of Lemma 1, namely that

(A2bis) There exists a solution to equation (3.17),

and if, in addition, (3.19) is assumed to hold only for $\eta = \eta^*$. However, **(A2)–(A3)** seem more natural and hold under weak conditions, as established in Section 4.

We conclude this section by noting that the Markovian properties and the specific structure of the cost criterion are not used in the proof of Theorem 2, in that the discussion applies to *any optimization problem* which satisfies conditions **(A1)–(A5)**. The only point which requires special care is the construction of an "interpolated" policy (3.14).

4. THE ASSUMPTIONS

In this section we discuss the assumptions **(A1)–(A5)**; we give concrete and verifiable conditions for several cost criteria. A specific model is analyzed in Section 5.

We focus on three cost criteria, namely the finite-time β-discounted cost criterion (3.3), its infinite-horizon counterpart (3.2) and the long-run average cost criterion (3.1), with the understanding that the discussion and methods apply, *mutatis mutandis*, to other situations as well. However, for the sake of brevity, we shall not elaborate in that direction.

The finite-time cost criterion – Condition **(A2)** holds if the costs are bounded since then the costs are polynomial in η. More generally, the same argument establishes **(A2)** if the costs are merely bounded from below (or from above).

Assumption **(A3)** holds if (3.6) is valid for all initial conditions, since then a backward-induction argument shows that for each η in $[0,1]$, g^η is optimal for the Lagrangian problems. Finally, **(A4)–(A5)** are always valid since under the non-negativity assumption on c and d, the equality

$$J_\theta(\pi) = \theta J_c(\pi) + (1-\theta) J_d(\pi), \quad \theta \in [0,1] \tag{4.1}$$

holds for every admissible policy π in $\mathcal{P}$. Condition **(A1.a)** immediately follows.

The discounted cost criterion – Condition **(A2)** holds if the costs are bounded since then the total discounted cost can be approximated by a finite number of terms in (3.2), and the argument for the finite case applies. More generally, under the same conditions as for the finite cost, the same argument applies provided a finite approximation is valid. This is the case if the tail of the infinite sum is bounded for η in $[0,1]$. This condition holds for all but the most pathological systems.

Assumption **(A3)** holds under rather weak conditions. For example, suppose the action space to be compact and the costs bounded above. Assume further that for each x in S, the mappings $u \to c(x;u)$ and $u \to d(x;u)$ are lower-semi continuous and that the transition kernel $Q(x;\cdot;dy)$ is continuous. Then any policy with actions in the optimal set (determined through the dynamic programming equation) is optimal for the Lagrange problem [17]. This implies that (3.19) holds whenever (3.6) is valid for each initial condition. Note that in this case boundedness from above replaces boundedness from below.

Finally, **(A4)**–**(A5)** always hold since, as in the finite case, (4.1) holds, and condition **(A1.a)** immediately follows.

The long-run average cost criterion – Condition **(A2)** was established when the state space S is finite in [13], and for the queueing system discussed in the next section [15]. A general method for verifying **(A2)** is available in [19]. In particular, this condition holds whenever the Markov chain is ergodic under both $\underline{g}$ and $\overline{g}$, provided the costs are integrable under the resulting invariant measures [13].

Condition **(A3)** can be established using dynamic programming arguments, as in the case of the discounted cost, although the requisite conditions are more stringent [17,21]. For some systems (such as the one described in Section 5), **(A3)** can be established by direct arguments [5,15].

Finally, we observe that for every admissible policy π in $\mathcal{P}$, the inequalities

$$\begin{aligned}
J_\theta(\pi) &= \overline{\lim}_t E^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} c_\theta(X_s,U_s)\right] \\
&= \overline{\lim}_t \left\{\theta E^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} c(X_s,U_s)\right]\right. \\
&\qquad \left. +(1-\theta)E^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} d(X_s,U_s)\right]\right\} \\
&\le \theta\overline{\lim}_t E^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} c(X_s,U_s)\right] \\
&\qquad +(1-\theta)\overline{\lim}_t E^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} d(X_s,U_s)\right] \\
&= \theta J_c(\pi) + (1-\theta)J_d(\pi), \quad \theta\in[0,1] \qquad (4.2)
\end{aligned}$$

always hold, so that condition **(A5)** is always satisfied. The validity of **(A4)** is more delicate to establish. In [19], the authors give conditions under which the long-run average cost criterion (3.1) is obtained as a limit under stationary policies. Under these conditions, **(A4)** holds, and **(A1.a)** follows.

5. BANDITS AND QUEUES

The purpose of this section is to show the equivalence between the discrete-time Klimov problem [11,14] and arm-acquiring bandit processes [20]. Continuous-time versions of this result are in [12,21]. Since both systems were discussed in detail elsewhere, we shall give only short, informal descriptions. Throughout this section, the rv ξ and the i.i.d. sequence $\{A(t),\ t=0,1,\ldots\}$ take their values in $\mathbb{N}^K$. We introduce the finiteness assumption

(F) $\quad \mathbf{E}[\xi_k] < \infty$ and $\mathbf{E}[A_k(t)] \triangleq \lambda_k < \infty,\ k=1,2,\ldots,K.$

Arm-acquiring bandits

The formulation in this section is given in the terminology of queueing systems, in order to facilitate the comparison: Customers of type $1,2,\ldots,N$ arrive into the system. A customer of type n can be in one of the states $\{1,2,\ldots,S_n\}$. It is convenient to lump together customers sharing both type and state [20]; we shall say that a customer of type n in state s, $s=1,\ldots,S_n$, resides in queue k, where

$$k=\sum_{j=1}^{n-1} S_j + s$$

and we set $K=\sum_{n=1}^N S_n$. With this convention, the number of customers initially in the system is ξ, and new customers arrive to the queues according to the arrival process $\{A(t),\ t=0,1,\ldots\}$. At most one customer can be served at a time. If a customer from queue k is served at time slot t, then at the end of the slot, with probability $p_{k\ell}$ this customer moves to queue ℓ, $k,\ell=1,\ldots,K$. All other customers do not change state—in other words, they remain at their queues. It is clear that the vector x in $\mathbb{N}^K$, where x_k is the number of customers in queue k, serves as a state for this MDP provided arrival, service completion and routing processes are mutually independent. The action $u=k$ is interpreted as service of queue k, $u=0$ as idle server, with the provision that $x_k=0$ implies $u\ne k$, $k=1,2,\ldots,K$. If a customer in queue k is served, then reward $r(k)$ is obtained. The reward to be maximized is of the discounted type (3.2), and takes the form

$$J_r(\pi;x) \triangleq \mathbf{E}^\pi\left[\sum_{s=0}^{\infty}\beta^s r(U_s)\right], \quad 0<\beta<1$$

which is well defined since r is bounded.

The classical description of the arm-acquiring bandits requires $\sum_\ell p_{k\ell}=1$ for each $k=1,\ldots,K$. However, this restriction is a purely semantic one since the effect of departures from the system can always be captured through the introduction of an absorbing queue with small (negative) reward for service, so that it is never served.

The discrete-time Klimov problem

Customers of type $1,2,\ldots,K$ arrive to their respective queues according to the arrival process $\{A(t),\ t=0,1,\ldots\}$. The number of customers present at time $t=0$ is given by ξ. The server can attend at most one queue at a time. If the server attends a non-empty queue, say queue k, $k=1,\ldots,K$, during time slot t, then at the end of the slot

- One customer leaves that queue with probability μ_k and, with probability $1-\mu_k$ no customer leaves that queue.
- If a customer left queue k, then with probability $\tilde{p}_{k\ell}$ it joins queue ℓ, $\ell=1,\ldots,K$. It leaves the system with probability $1-\sum_{\ell=1}^K \tilde{p}_{k\ell}$.

For $k,\ell=1,\ldots,K$, we set $p_{k\ell}\triangleq\mu_k\tilde{p}_{k\ell}$ for $\ell\ne k$ and $p_{kk}\triangleq 1-\mu_k(1-\tilde{p}_{kk})$. Using this transformation, the values of μ_k are henceforth taken to be 1. Clearly, if arrival, service completion and routing processes are assumed mutually independent, the dynamics of this system are identical to the dynamics of the corresponding arm-acquiring bandit system.

The state of this system is again the vector x in $\mathbb{N}^K$ where x_k denotes the number of customers in queue k, $k=1,\ldots,K$. The cost for the Klimov problem is defined by

$$c(x,u)=c(x)\triangleq\sum_{k=1}^K c_k x_k, \qquad x\in\mathbb{N}^K$$

for some constants $c_1, \dots, c_K$ (which are usually assumed non-negative). The objective is to *minimize* the discounted cost associated with this one-step cost, viz.,

$$J_c(\pi) \triangleq \mathbf{E}^\pi \left[\sum_{s=0}^{\infty} \beta^s c(X_s) \right], \quad \pi \in \mathcal{P}. \tag{5.1}$$

Following the cost-transformation of [4, 5] it is straightforward to derive the identity

$$J_c(\pi) = \frac{\mathbf{E}[c(\xi)]}{1-\beta} + \frac{\beta}{(1-\beta)^2} c(\lambda) - \frac{\beta}{1-\beta} J_{\tilde{c}}(\pi) \tag{5.2}$$

where $\lambda = (\lambda_1, \dots, \lambda_K)'$ and the one-step cost $\tilde{c}$ is defined by

$$\tilde{c}(x,u) \triangleq \sum_{k=1}^{K} \mathbf{1}[u=k] \tilde{c}_k$$

with

$$\tilde{c}_k \triangleq \left[c_k - \sum_{\ell=1}^{K} p_{k\ell} c_l \right], \quad k = 1, \dots, K \tag{5.3}$$

and action u is defined as in the bandit problem. As a result, for *each fixed* β in $(0,1)$, we have

$$\arg\min J_c(\pi) = \arg\max J_{\tilde{c}}(\pi) \ . \tag{5.4}$$

Observe that the cost function $\tilde{c}$ depends only on the queue being served, and so it is a legitimate cost function for the bandit problem.

The equivalence result

Theorem 3. *Any discrete-time Klimov problem defines an arm-acquiring bandit system with the same dynamics. Under* **(F)**, *they possess the same optimal policies, with costs related by (5.2)–(5.3) (with $r(k) \triangleq \tilde{c}_k$, $k = 1, \dots, K$). Conversely, any arm-acquiring bandit system defines a Klimov problem with the same dynamics. Moreover, Under* **(F)**, *if the vector $r \triangleq (r(1), r(2), \dots, r(K))'$ is in the range of $I - P$, then the cost in the Klimov problem can be defined so as to satisfy the transformation (5.2)–(5.3) (with $\tilde{c}_k \triangleq r(k)$, $k = 1, \dots, K$) and consequently, the same policies are optimal for both systems.*

The proof follows from the preceding discussion, upon observing that if r is in the range of $I - P$ then there is a one-to-one mapping between $(c_1, \dots, c_K)$ and $(\tilde{c}_1, \dots, \tilde{c}_K)$.

Constrained Optimization

The best-known class of problems for which the hypotheses **(A1)–(A5)** hold is the class of arm-acquiring (or open) bandit processes [20] described above. For consistency with the notation of Section 3, we let c and d still denote the two cost functions (although in this case they are independent of x).

Lemma 4. *For the arm-acquiring bandit problem under the discounted cost criterion, conditions* **(A1)–(A5)** *hold.*

Proof. It is well known [20] that the optimal policy for this system possesses an index rule structure. Thus an optimal policy (for any $0 \le \theta \le 1$) chooses only which queue to serve. Therefore such a policy is uniquely determined by an ordering of the queues, where a queue is served only if queues with higher priority are empty. Since there is a finite number $K!$ of such policies, **(A1.b)** follows. Since the costs are bounded and the action space is discrete, the argument in Section 4 now establishes the result. ■

We call the Klimov problem *stable* if $\rho \triangleq \lambda'(I - P)e < 1$ (where e is the element $(1, \dots, 1)'$ of $\mathbb{N}^K$). A policy is called non-idling if $x_k = 0$ implies $u \neq k$.

Lemma 5. *Assume* **(F)** *and that the Klimov problem is stable. Moreover, let $c_k \ge 0$, $k = 1, \dots, K$. (i) If P is diagonal then* **(A1.b)** *holds, where $\{g_\ell, \ \ell = 1, 2, \dots, L\}$ is a collection of strict priority policies; (ii) If $\{g_\ell, \ \ell = 1, 2, \dots, L\}$ is a collection of stationary non-idling policies, then* **(A1.a)** *and* **(A2)–(A5)** *hold.*

Proof. Part (i) is established in [4,5]. Under the conditions in (ii), Makowski and Shwartz [14,19] establish **(A2)**, whereas **(A4)** follows from [19]. As discussed in Section 4, **(A5)** holds, and **(A1.a)** follows from **(A4)**. Finally, under the regularity conditions established in [14], standard dynamic programming techniques yield **(A3)**. ■

REFERENCES

[1] E. Altman and A. Shwartz "Optimal priority assignment: a time sharing approach," *IEEE Trans. Automatic Control* **AC–34**, pp. 1098–1102 (1989).

[2] E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies," *SIAM J. Control and Optimization* **29**, pp. 786–809 (1991).

[3] E. Altman and A. Shwartz, "Adaptive Control of constrained Markov chains: criteria and policies," *Annals of Operations Research* **28**, pp. 101–134 (1991).

[4] J.S. Baras, A.J. Dorsey, and A.M. Makowski, "Two competing queues with linear costs and geometric service requirements: The μc–rule is often optimal," *Advances in Applied Probability* **17**, pp. 186–209 (1985).

[5] J.S. Baras, D.-J. Ma, and A.M. Makowski, "K competing queues with geometric service requirements and linear costs: the μc–rule is always optimal," *Systems & Control Letters* **6**, pp. 173–180 (1985).

[6] F. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *Math. Anal. Appl.* **112**, pp. 236–252 (1985).

[7] V.S. Borkar, "Controlled Markov chains with constraints," Preprint (1989).

[8] C. Derman and M. Klein, "Some remarks on finite horizon Markovian decision models," *Oper. Res.* **13**, pp. 272–278 (1965).

[9] C. Derman and A.F. Veinott, Jr., "Constrained Markov decision chains," *Management Sci.* **19**, pp. 389–390 (1972).

[10] A. Hordijk and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming," *Mathematics of Operations Research* **9**, pp. 276–289 (1984).

[11] G.P. Klimov, "Time sharing systems," *Theory of Probability and Its Applications*; Part I: **19**, pp. 532–553 (1974). Part II: **23**, pp. 314–321 (1978).

[12] T.L. Lai and Z. Ying, "Open bandit processes and optimal scheduling of queueing networks," *Adv. Appl. Prob.* **20**, pp. 447–472 (1988).

[13] D.-J. Ma, A.M. Makowski and A. Shwartz, "Stochastic approximation for finite state Markov chains," *Stochastic Processes and Their Applications* **35**, pp. 27–45 (1990).

[14] A.M. Makowski and A. Shwartz, "Recurrence properties of a discrete-time single-server network with random routing," EE Pub. **718**, Technion, Haifa (Israel) (1989).

[15] A.M. Makowski and A. Shwartz, "Analysis and adaptive control of a discrete-time single-server network with random routing," *SIAM J. Control Opt.*, accepted for publication (1991).

[16] S.M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York (NY) (1984).

[17] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal," *Z. Wahr. verw. Gebiete* **32**, pp. 179-196 (1975).

[18] L. I. Sennott, "Constrained average-cost Markov decision chains," Preprint (1990).

[19] A. Shwartz and A. M. Makowski, "On the Poisson equation for Markov chains," EE Pub. **646**, Technion, Haifa (Israel). Also under revision, *Math. of Operations Research* (1987).

[20] P. Whittle, "Arm acquiring bandits," *The Annals of Probability* **9**, pp. 284–292 (1981).

[21] P. Whittle, *Optimization Over Time; Dynamic Programming and Stochastic Control*, Wiley & Sons, New York (NY) (1982).