```
[m, n] = size(A); x = zeros(n, 1); dx = zeros(n, 1);
for k = 1:m
    phi = A(k, :)';
    e = y(k) - phi'*x;
    [R, yb] = newqr(R, phi, e, lam);
    dx = solve(R, yb);
    x = x + dx;
    X(k, :) = x';
end
function [R, yb] = newqr(R, phi, e, lam)
[m, n] = size(R);
BigR = [R zeros(n, 1); phi' e];
for i = 1:n
    BigR = sweep(BigR, i, lam);
end
R = BigR(1:n, 1:n);
yb = BigR(1:n, n + 1);
function A = sweep(A, i, lam)
% Input
%       A - n × n matrix
%       i - The (i, n) rows of the A matrix are swept, with the
%           element A(n, i) being annihilated.
%     lam - The forgetting factor.
% Output
%       A - swept matrix.
[m, n] = size(A);
[c, s, r] = givens(A(i, i)*lam, A(n, i));
A(i, i) = r;
A(n, i) = 0.0;
clam = c*lam;
slam = s*lam;
for k = i + 1:n
    a = A(i, k)*clam + A(n, k)*s;
    A(n, k) = - A(i, k)*slam + A(n, k)*c;
    A(i, k) = a;
end
function [c, s, r] = givens(a, b)
% Input
%       a, b - Elements of a vector for the which b component
%              is to be annihilated by a plane rotation.
% Output
%       c, s - The required transformation.
%          r - The length of the vector.
r = sqrt(a^2 + b^2);
if r < 1.0e-8,
  c = 1; s = 0; r = 1.0e-8;
else
  c = a/r;
  s = b/r;
end
function x = solve(R, b)
% Solves Rx = b for x assuming R is upper right triangular.
% Input
%       b - Column vector of output data.
%       R - right triangular matrix.
% Output
%       x - solution.
[m, n] = size(R);
for k = n :- 1:1,
    sum = 0.0;
    for j = k + 1:n
      sum = sum + R(k, j)*x(j, 1);
    end
    x(k, 1) = (b(k) - sum)/R(k, k);
end
```

References

[1] A. Andrews, "A square root formulation of the Kalman covariance equations," *AIAA J.*, vol. 6, no. 6, 1968.

[2] G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation.* New York: Academic, 1977.

[3] P. E. Gill, W. Murray, and M. H. Wright, *Numerical Linear Algebra and Optimization*, vol. 1. Redwood City, CA: Addison-Wesley, 1991.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, MD: The Johns Hopkins University Press, 1990.

[5] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control.* Englewood Cliffs, NJ: Prentice-Hall, 1984.

[6] P. S. Lewis, "QR-based algorithms for multichannel adaptive least squares Lattice filters," *IEEE Trans. Acoust., Speech, Signal Proc.*, vol. 38, no. 3, pp. 421–431, 1990.

[7] L. Ljung and T. Soderstrom, *Theory and Practice of Recursive Identification.* Cambridge, MA: M.I.T. Press, 1983.

On Constrained Optimization of the Klimov Network and Related Markov Decision Processes

Armand M. Makowski and Adam Shwartz

***Abstract*—We solve a constrained version of the server allocation problem for the Klimov network and establish that the optimal constrained schedule is obtained by randomizing between two fixed priority schemes. This generalizes the work of Nain and Ross in the context of the competing queue problem and also covers the discounted cost case. In order to establish these results, we develop a general framework for optimization under a single constraint in the presence of index-like policies. This methodology is in principle of wider applicability.**

I. Introduction

Consider the discrete-time system of K competing queues with a single Bernoulli server as described in [5] and [8]. For one-step costs which are linear in the queue sizes, it is well known [4], [5], [8] that there exists an optimal policy which is of the strict priority type, and this, under several cost criteria including the discounted and average cost criteria in which case the search for optimal policies reduces to the computation of a few parameters. Let $J_c(\pi)$ and $J_d(\pi)$ be two cost functions associated with the one-step cost functions c and d, when the system is operated under the policy π. A single constraint optimization problem can then be defined as follows:

$$(\boldsymbol{P}_v): \quad \text{Minimize } J_c(\pi) \text{ subject to the constraint } J_d(\pi) \le V$$

Manuscript received April 12, 1992; revised February 21, 1992. This work was supported in part by the National Science Foundation under Grants ECS-83-51836 and CDR-88-03012, and partially through United States—Israel Binational Science Foundation under Grant BSF-85-00306.

A. M. Makowski is with the Department of Electrical Engineering and the Systems Research Center, University of Maryland, College Park, MD 20742.

A. Shwartz is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9203115.

for some scalar V. When both costs c and d are linear in the queue sizes, under the average cost criterion, Nain and Ross [19] obtained the following optimality result: There exist two fixed priority policies, say g and $\bar{g}$, and a constant η^* in [0, 1] so that at every step, it is optimal to flip a coin with probability η^* for heads, and to use g (respectively $\bar{g}$) if a head (respectively tail) is observed. The optimal randomization bias η^* is selected so as to saturate the constraint.

In view of such results, it is quite natural to inquire whether this structural result for the optimal constrained policy can be extended, say, to cover the following:

i) The situation where the discounted or the finite-time cost criteria are used;

ii) The scheduling problem associated with a natural extension of the competing queue problem, the so-called Klimov system [14], where upon service completion, the customer may either be routed to one of the other queues or leave the system.

More generally, it is certainly of interest to identify conditions under which the solution to a constrained Markov decision process (MDP) does exhibit such a randomized structure. Indeed, once this structural result is established, the search for optimal policies reduces to the identification of the two policies and to the computation of the randomization bias.

We answer i)–ii) in the affirmative in Section IV. In the process, in Section II we develop a more general methodology which applies to systems with "index-like" optimal policies. This is embedded in the only "structural" assumption (A1), which states that for each θ in [0, 1], an optimal policy for the *unconstrained* problem with cost $c + \theta d$ can be found within a given finite set of policies which set is independent of θ in [0, 1]. In Section III, we show that the technical conditions apply to many MDP's, with finite, discounted, and average cost criteria. In Section IV, we establish the equivalence between the discounted Klimov system and open (or arm-acquiring) bandit problems. This establishes the structural result for the Klimov system under the discounted cost criterion, and for all bandits problems under mild conditions. Under slightly stronger assumptions, the structural result also holds for the Klimov system under the average cost criterion.

MDP's under constraints where first solved by Derman and Klein [9] for the finite-horizon cost. When $J(\pi)$ is the average cost criterion and both state space S and action space U are finite, the existence of optimal stationary policies under multiple constraints was established by Derman and Veinott [10]. Hordijk and Kallenberg [13] solved the multiclass 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 [22] for countable S and compact U. Borkar [7] obtained analogous results under multiple constraints when S is countable and U is compact, and indicated similar results for other cost criteria. The multiple constraint case for countable S and countable U is treated by Altman and Shwartz [2]. Frid [11] solved the discounted problem with a single constraint, using the Lagrangian approach. In [3], Altman and Shwartz prove existence of optimal policies for finite S and U under the discounted and other cost criteria, under multiple constraints, and present computational algorithms.

Unfortunately, except for the finite case and the specific example in [1], there are no efficient methods for computing constrained optimal policies. The results mentioned in the previous paragraph establish the existence of an optimal stationary policy which randomizes between some stationary deterministic policies. However, except for the finite case, the search for the two policies to be randomized is over *all* stationary deterministic policies. Our methodology provides conditions under which this search can be restricted to a finite set of policies.

A few words on the notation and conventions used in this note: We consider an MDP (S, U, P) as defined in the literature [20], [21], [25]. Both the state space S and the action space U are Polish spaces, and measurability is taken with respect to their Borel σ-fields $\mathscr{B}(S)$ and $\mathscr{B}(U)$, respectively. The one-step transition mechanism P is described through a family of transition kernels $(Q(x, u; dy))$. The S-valued state process $\{X_t, t = 0, 1, \cdots\}$ and the U-valued control process $\{U_t, t = 0, 1, \cdots\}$ are defined on some measurable space $(\Omega, \mathscr{F})$. The information $(X_0, U_0, X_1, \cdots, U_{t-1}, X_t)$ available at time t is compactly denoted by H_t. We denote the space of probability measures on $\mathscr{B}(U)$ by $\mathbb{M}(U)$, and we write $\mathscr{P}$ for the collection of all (nonanticipative) admissible policies. For every control policy π in $\mathscr{P}$, let $\boldsymbol{P}^\pi$ $(\boldsymbol{E}^\pi)$ denote the probability (expectation operator) induced by π.

A policy π is a Markov policy if there exist Borel mappings $\{g_t, t = 0, 1, \cdots\}$, $g_t: S \to \mathbb{M}(U)$ such that $\pi_t(\cdot; H_t) = g_t(\cdot; X_t)\boldsymbol{P}^\pi$ – a.s. for all t. If $\{g_t, t = 0, 1, \cdots\}$ are all identical to $g: S \to \mathbb{M}(U)$, the Markov policy is termed stationary and is identified with g.

Unless stated otherwise, $\lim_n$, $\underline{\lim}_n$ and $\overline{\lim}_n$ are taken with n going to infinity. The infimum over an empty set is taken to be ∞ by convention.

II. A General Constrained MDP

We interpret any Borel mapping $c: S \times U \to \mathbb{R}$ as a one-step cost function. To avoid unnecessary technicalities we always assume c to be bounded below, and without loss of generality we take $c \geq 0$. For any policy π in $\mathscr{P}$, we define $J_c(\pi)$ as the total cost (associated with c) for operating the system under policy π, with possible choices including the long-run average cost

$$J_c(\pi) \triangleq \overline{\lim}_t \boldsymbol{E}^\pi\left[\frac{1}{t+1}\sum_{s=0}^{t} c(X_s, U_s)\right] \tag{2.1}$$

and the infinite horizon β-discounted cost

$$J_c(\pi) \triangleq \boldsymbol{E}^\pi\left[\sum_{s=0}^{\infty} \beta^s c(X_s, U_s)\right], \qquad 0 < \beta < 1. \tag{2.2}$$

The definitions (2.1) and (2.2) are all well posed under the nonnegativity assumption on c.

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

$$\mathscr{P}_V := \{\pi \in \mathscr{P}: J_d(\pi) \leq V\}. \tag{2.3}$$

The corresponding *constrained* optimization problem $(\boldsymbol{P}_v)$ is formulated as

$$(\boldsymbol{P}_v): \quad \text{Minimize } J_c(\cdot) \text{ over } \mathscr{P}_V.$$

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

$$c_\theta(x, u) \triangleq \theta c(x, u) + (1 - \theta) d(x, u), \qquad x \in S, u \in U. \tag{2.4}$$

To simplify the notation, we denote by $J_\theta(\pi)$ the total cost associated with c_θ, so that $J_\theta(\pi) = J_c(\pi)$ for $\theta = 1$ and $J_\theta(\pi) = J_d(\pi)$ for $\theta = 0$. Assume the following:

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

(A1.a)

$$\inf_{\pi\in\mathscr{P}} J_\theta(\pi) = \min_{1\le l\le L} J_\theta(g_l) \triangleq J^*(\theta), \qquad \theta\in[0,1]. \tag{2.5}$$

(A1.b) For each $l = 1,\cdots,L$, the mapping $\theta \to J_\theta(g_l)$ is continuous on [0, 1].

Under (A1), the mapping $\theta \to J^*(\theta)$ is continuous on [0, 1]. As in [1], the *Lagrangian problem* is defined as the problem of minimizing $J_\theta(\cdot)$ over the unconstrained set of policies $\mathscr{P}$. We define

$$N(\theta) \triangleq \{l\in\{1,\cdots,L\}: J_\theta(g_l) = J^*(\theta)\}, \qquad \theta\in[0,1]. \tag{2.6}$$

By (A1.a), for each θ in [0, 1], the index set $N(\theta)$ is nonempty and (A1.b) implies

$$\lim_{\tilde\theta\uparrow\theta} J_{\tilde\theta}(g_l) = \lim_{\tilde\theta\downarrow\theta} J_{\tilde\theta}(g_l) = J^*(\theta), \qquad l\in N(\theta). \tag{2.7}$$

Furthermore, if $N(\theta)$ is a singleton, then $N(\theta) = \mathrm{N}(\tilde\theta)$ in some open neighborhood of θ. We set

$$n(\theta) \triangleq \min\left\{n\in N(\theta): J_d(g_n) = \min_{l\in N(\theta)} J_d(g_l)\right\}, \qquad \theta\in[0,1]. \tag{2.8}$$

If $J_0(g_{n(0)}) = J_d(g_{n(0)}) > V$, then the problem $(\boldsymbol{P}_v)$ is not feasible and, therefore, possesses no solution. Assuming feasibility from now on, we set

$$\theta^* \triangleq \sup\{\theta\in[0,1]: J_d(g_{n(\theta)}) \le V\}. \tag{2.9}$$

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

$$\min\{J_c(g_l): 1\le l\le L, J_d(g_l)\le V\} > \inf_{\pi\in\mathscr{P}_V}\{J_c(\pi): J_d(\pi)\le V\}$$

since the Lagrangian problem may not provide enough information.

If $\theta^* = 1$, them $(\boldsymbol{P}_v)$ has a solution: Indeed let $\theta_i \uparrow 1$ in (0, 1] so that $J_d(g_{n(\theta_i)}) \le V$ for all $i = 1,2,\cdots$, 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 $\{1,\cdots,L\}$. In fact, we can assert $n(\theta_j) = n^*$ whenever $j\ge j^*$ for some j^*. It is plain that n^* is an element of $N(\theta_j)$ for $j\ge j^*$, 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 $J_d(g_{n^*})\le V$, we conclude that the policy g_{n^*} solves $(\boldsymbol{P}_v)$.

From now on, assume $0<\theta^*<1$. Let $\theta_i\downarrow\theta^*$ in (0, 1) and denote by $\bar n$ an accumulation point of the sequence $\{n(\theta_i), i = 1,2,\cdots\}$. Similarly, let $\theta_j\uparrow\theta^*$ in (0, 1) such that $J_d(g_{n(\theta_j)}) \le V$ and denote by n an accumulation point of $\{n(\theta_j), j = 1,2,\cdots\}$. Again, $n(\theta_i) = \bar n$ and $n(\theta_j) = \underline{n}$ for all i and j large enough. By (A1.b), both $\bar n$ and $\underline{n}$ are elements of $N(\theta^*)$, so that

$$J_{\theta^*}(g_{\underline{n}}) = J_{\theta^*}(g_{\bar n}) = J^*(\theta^*) \tag{2.10}$$

must hold. Moreover, it is plain that

$$J_d(g_{\underline{n}}) \le V \le J_d(g_{\bar n}). \tag{2.11}$$

The first inequality follows by construction and (A1.b), whereas the second inequality results from the construction and from (2.8) and (2.9).

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

$$\underline{g} \triangleq g_{\underline{n}}, \quad \bar g \triangleq g_{\bar n} \tag{2.12}$$

$$g^\eta \triangleq \eta\underline{g} + (1-\eta)\bar g, \qquad \eta\in[0,1]. \tag{2.13}$$

Then g^η is the simple randomization between the two policies $\underline{g}$ and $\bar g$ with randomization bias η. The identities (2.10) and (2.11) now take the form

$$J_{\theta^*}(\underline{g}) = J_{\theta^*}(\bar g) = J^*(\theta^*) \tag{2.14}$$

$$J_d(\underline{g}) \le V \le J_d(\bar g). \tag{2.15}$$

At this point we introduce assumption (A2):

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

Lemma 1: Under (A1) and (A2) there exists a solution η^* to the equation

$$J_d(g^\eta) = V, \qquad \eta\in[0,1]. \tag{2.16}$$

Proof: This is immediate from the fact that the mapping $\eta - J_d(\mathrm{g}^\eta)$ is continuous on [0, 1] and from the inequality (2.15) which can written as $J_d(g^1)\le V\le J_d(g^0)$. ■

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

(A3) $$J_{\theta^*}(g^\eta) = J_{\theta^*}(\underline{g}), \qquad \eta\in[0,1]. \tag{2.17}$$

(A4) $$J_{\theta^*}(g^{\eta^*}) = \theta^* J_c(g^{\eta^*}) + (1-\theta^*)J_d(q^{\eta^*}). \tag{2.18}$$

(A5) For every admissible policy π in $\mathscr{P}$,

$$J_{\theta^*}(\pi) \le \theta^* J_c(\pi) + (1-\theta^*)J_d(\pi). \tag{2.19}$$

Theorem 2: Under (A1)–(A5), the policy g^{η^*} [where η^* is a solution of (2.16)] solves the constrained problem $(\boldsymbol{P}_v)$ provided $\theta^*>0$.

Proof: We first note that

$$J^*(\theta^*) = J_{\theta^*}(g^{\eta^*}) \tag{2.20}$$

$$= \theta^* J_c(g^{\eta^*}) + (1-\theta^*)J_d(g^{\eta^*}) \tag{2.21}$$

where (2.20) follows from (2.14) and (A3), whereas (2.21) is validated by (A4). Now

$$J_{\theta^*}(\pi) \ge J^*(\theta^*), \qquad \pi\in\mathscr{P} \tag{2.22}$$

by virtue of (A1.a), and

$$J_{\theta^*}(\pi) \le \theta^* J_c(\pi) + (1-\theta^*)J_d(\pi), \qquad \pi\in\mathscr{P} \tag{2.23}$$

by invoking (A5). By Lemma 1, the policy g^{η^*} is an element of $\mathscr{P}_V$ since $J_d(g^{\eta^*}) = V$ by construction, and upon combining (2.20)–(2.23), we get

$$\begin{aligned} &\theta^* J_c(\pi) + (1-\theta^*)J_d(\pi)\\ &\quad \ge J_{\theta^*}(\pi)\\ &\quad \ge \theta^* J_c(g\tau^*) + (1-\theta^*)V, \qquad \pi\in P \end{aligned} \tag{2.24}$$

It is now plain from (2.24) that

$$\theta^* J_c(g^{\eta^*}) \le \theta^* J_c(\pi), \qquad \pi\in\mathscr{P}_V \tag{2.25}$$

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 there exists a solution to (2.16), and if, in addition, (2.17) is assumed to hold only for $\eta = \eta^*$. However, (A2) and (A3) seem more natural and hold under weak conditions, as shown in Section III. Moreover, η^* is usually not known and therefore (A2)–(A4) are verified by establishing the conditions for *all* η in [0, 1].

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 (2.13).

In particular, consider the finite horizon β-discounted cost

$$J_c(\pi) \triangleq E^{\pi}\left[\sum_{s=0}^{T} \beta^s c(X_s, U_s)\right], \qquad 0 < \beta \leq 1, T = 1, 2, \cdots. \tag{2.26}$$

The derivation of Lemma 1 and Theorem 2 holds verbatim, provided (A1) holds with the word "stationary" omitted. Since the identification of a policy g with a single function $S \to \mathbb{M}(U)$ does not hod any longer, (2.13) is interpreted naturally as

$$g_t^{\eta} \triangleq n\underline{g}_t + (1 - \eta)\bar{g}_t, \quad \eta \in [0, 1], \qquad t = 0, 1, \cdots. \tag{2.27}$$

III. The Assumptions

In this section, we discuss the assumptions (A1)–(A5); we give concrete and verifiable conditions for several cost criteria. The discussion and methods apply, *mutatis mutandis*, to other situations as well. A specific model is analyzed in Section IV.

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 (2.5) is valid for all initial conditions, since then a backward-induction argument proves that for any η in [0, 1], g^{η} is optimal for the Lagrangian problems. Finally, (A4) and (A5) are always valid since under the nonnegativity assumption on c and d,

$$J_{\theta}(\pi) = \theta J_c(\pi) + (1 - \theta) J_d(\pi), \qquad \theta \in [0, 1] \tag{3.1}$$

for every admissible policy π in $\mathscr{P}$. Condition (A1.b) 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 (2.2) uniformly in η, and the finite case argument 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 is 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 semicontinuous and that the transition kernel function $u \to Q(x; \cdot; dy)$ is weakly continuous (i.e., whenever $c: S \to \mathbb{R}$ is bounded and continuous, the mapping $u \to \int c(y)\, dQ(x, u, dy)$ is continuous on U). Then any policy with actions in the optimal set (determined through the dynamic programming equation) is optimal for the Lagrangian problem [21]. This implies that (2.17) holds whenever (2.5) is valid for each initial condition. Note that in this case boundedness from above replaces boundedness from below.

Finally, (A4) and (A5) always hold since, as in the finite case, (3.1) holds, and condition (A1.b) immediately follows.

The Long-Run Average Cost Criterion: Condition (A2) was established when the state space S is finite in [16], and for the queueing system discussed in the next section in [18]. A general method for verifying (A2) is available in [23]. In particular, this condition holds whenever the Markov chain is ergodic under both g and $\bar{g}$, provided the costs are integrable under the resulting invariant measures [16].

Condition (A3) can be established using dynamic programming arguments, as in the case of the discounted cost, although the requisite conditions are more stringent [21], [25]. For some systems (such as the one described in Section IV), (A3) can be established by direct arguments [5], [18].

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

$$\begin{aligned} J_{\theta}(\pi) &= \overline{\lim}_t \left\{ \theta E^{\pi}\left[\frac{1}{t+1}\sum_{s=0}^{t} c(X, U\mathrm{s})\right] \right. \\ &\qquad \left. + (1 - \theta) E^{\pi}\left[\frac{1}{t+1}\sum_{s=0}^{t} d(X_s, U_s)\right]\right\} \\ &\leq \theta \overline{\lim}_t E^{\pi}\left[\frac{1}{t+1}\sum_{s=0}^{t} c(X_s, U_s)\right] \\ &\qquad + (1 - \theta)\overline{\lim}_t E^{\pi}\left[\frac{1}{t+1}\sum_{s=0}^{t} d(X_s, U_s)\right] \\ &= \theta J_c(\pi) + (1 - \theta) J_d(\pi), \qquad \theta \in [0, 1] \end{aligned} \tag{3.2}$$

so that condition (A5) is always satisfied. The validity of (A4) is more delicate to establish. In [23], the authors give conditions under which the long-run average cost criterion (2.1) is obtained as a limit under stationary policies. Under these conditions, (A4) holds, and (A1.b) follows.

IV. Bandits and Queues

The purpose of this section is to show the equivalence between the discrete-time Klimov problem [14], [17] and arm-acquiring bandit processes [24]. Continuous-time versions of this result are discussed in [15], [25]. 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 $\{A(t), t = 0, 1, \cdots\}$ taking their values in $\mathbb{N}^K$ are held fixed. We introduce the finiteness assumption

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

Arm-Acquiring Bandits: The formulation is given in the terminology of queueing systems in order to facilitate the comparison: Customers of type $1, 2, \cdots, N$ arrive into the system; a customer of type n can be in one of the states $\{1, 2, \cdots, S_n\}$. It is convenient to lump together customers sharing both type and state [24]; we shall say that a customer of type n in state $s = 1, \cdots, S_M$, resides in queue k, where

$$k = \sum_{j=1}^{n-1} S_j + s \tag{4.1}$$

and where $K = \sum_{n=1}^{N} S_n$. 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 $\{A(t), t = 0, 1, \cdots\}$. At most one customer can be given service attention at a time. If a customer from queue k is served in time slot t, then at the end of the slot, with probability p_{kl} this customer moves to queue l, k, $l = 1, \cdots, K$. All other customers do not change state—in other words, they remain at their queues. The routing rvs are assumed to form an i.i.d. sequence. 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. This together with the assumption on the routing mechanism implies that at each queue, the events that a customer leaves the system can be modeled by i.d.d. Bernoulli rvs with queue-dependent parameter. The action $u = k$ is interpreted as service at queue $k, u = 0$ as idle server, with the provision that $x_k = 0$ implies $u \neq k$, $k = 1, 2, \cdots, K$. If a customer in queue k is served, then reward $r(k)$ is incurred. The reward to be maximized is of the discounted type (2.2), and takes the form

$$J_r(\pi; x) \triangleq E^{\pi}\left[\sum_{t=0}^{\infty} \beta^t r(U_t)\right], \qquad 0 < \beta < 1 \tag{4.2}$$

which is well defined since r is bounded.

The classical description of the arm-acquiring bandits requires $\Sigma_l p_{kl} = 1$ for each $k = 1, \cdots, K$. 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 $1, 2, \cdots, K$ arrive to their respective queues according to the arrival process $\{A(t),\ t = 0, 1, \cdots\}$. The number of customers present at time 0 is given by ξ. The server can attend at most one queue at a time. If the server attends a nonempty queue, say queue k, $k = 1, \cdots, K$, during time slot t, then at the end of the slot the following sequence of events takes place:

> One customer leaves that queue with probability μ_k and, with probability $1 - \mu_k$ no customer leaves that queue; If a customer has left queue k, then with probability $\tilde{p}_{kl}$ it joins queue l, $l = 1, \cdots, K$, and it leaves the system with probability $1 - \Sigma_{l=1}^{K} \tilde{p}_{kl}$.

For $k, l = 1, \cdots, K$, we set $p_{kl} \triangleq \mu_k \tilde{p}_{kl}$ for $l \neq k$ 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. Then clearly, assuming arrival, service completion and routing processes to be mutually independent, the dynamics of this system are equivalent to the dyanmics 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, $k = 1, \cdots, K$. The cost for the Klimov problem is defined by

$$c(x, u) = c(x) \triangleq \sum_{k=1}^{K} c_k x_k, \qquad x \in \mathbb{N}^K, u = 0, 1, \cdots, K$$

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

$$J_c(\pi) \triangleq E^{\pi}\left[\sum_{t=0}^{\infty} \beta^t c(X_t)\right], \qquad \pi \in \mathscr{P}. \tag{4.3}$$

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

$$J_c(\pi) = \frac{Ec(\xi)}{1-\beta} + \frac{\beta}{(1-\beta)^2} c(\lambda) - \frac{\beta}{1-\beta} J_{\tilde{c}}(\pi), \qquad \pi \in \mathscr{P}. \tag{4.4}$$

where the one-step cost $\tilde{c}$ is defined by

$$\tilde{c}(x, u) \triangleq \sum_{k=1}^{K} 1[u = k] \tilde{c}_k, \quad \tilde{c}_k \triangleq \left[c_k - \sum_{l=1}^{K} p_{kl} c_l\right], \qquad k = 1, \cdots, K \tag{4.5}$$

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

$$\arg\min J_c(\pi) = \arg\max J_{\tilde{c}}(\pi). \tag{4.6}$$

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

The Equivalence Result: We have the following theorem:

Theorem 3: Any discrete-time Klimov problem defines an arm-acquiring bandit system with the same dynamics. Under $(\boldsymbol{F})$, they possess the same optimal policies, with costs related by (4.4) and (4.5) (with $r(k) \triangleq \tilde{c}_k$, $k = 1, \cdots, K$). Conversely, any arm-acquiring bandit system defines a Klimov problem with the same dynamics. Moreover, Under $(\boldsymbol{F})$, if the vector $r \triangleq (r(1), r(2), \cdots, r(K))'$ is in the range of $I - P$, then the cost in the Klimov problem can be defined so as to satisfy the transformation (4.4) and (4.5) (with $\tilde{c}_k \triangleq r(k)$, $k = 1, \cdots, K$) 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 $(c_1, \cdots, c_K)$ and $(\tilde{c}_1, \cdots, \tilde{c}_K)$.

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 [24] described above. For consistency with the notation of Section II, we let c and d denote the two cost functions (which are here 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 [24] that the optimal policy for this system possesses an index-structure. Thus an optimal policy (for any $0 \leq \theta \leq 1$) 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.a) follows. Since the costs are bounded and the action space is discrete, the argument in Section III now establishes the result. ■

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

Lemma 5: Consider the average-cost case. Assume $(\boldsymbol{F})$ and that the Klimov problem is stable. Moreover, let $c_k \geq 0$, $k = 1, \cdots, K$. i) If $\{g_l,\ l = 1, 2, \cdots, L\}$ is a collection of stationary nonidling policies, then (A1.b) and (A2)–(A5) hold; ii) If P is diagonal than (A1.a) holds, where $\{g_l,\ l = 1, 2, \cdots, L\}$ is a collection of strict priority policies.

Proof: Under the conditions in i), Makowski and Shwartz [17], [23] establish (A2), whereas (A4) follows from [23]. As discussed in Section III, (A5) holds, and (A1.b) follows from (A4). Finally, under the regularity conditions established in [17], standard dynamic programming techniques yield (A3). Part ii) is established in [4], [5] ■

When P is diagonal, Theorem 2 now implies the existence of an optimal policy which randomizes between two strict priority policies, and we recover the results of [19]. In general, if we strengthen $(\boldsymbol{F})$ to require finite second moments, then [17], [18] establish that for every stationary nonidling policy π, the average cost $J_c(\pi)$ of (2.1) is obtained as a limit. From general results on MDP's there exists an optimal stationary policy for the average Lagrangian problem. Since the costs are positive, sample path arguments imply that this policy can be assumed nonidling. A standard Tauberian theorem [12] now implies that for each stationary nonidling policy, the average cost is the limit of the normalized discounted cost. Since (A1.a) holds in the discounted case (Lemma 4) where $g_1,\cdots,g_L$ are strict priority policies, (A1.a) holds also for the average problem under the above conditions. Theorem 2 now implies the existence of an average cost optimal policy which randomizes between two strict priority policies.

Thus the result of Nain and Ross [19] extends to the Klimov problem, and this under both the discounted and the average cost criteria.

References

[1] E. Altman and A. Shwartz, "Optimal priority assignment: A time sharing approach," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1098–1102. 1989.

[2] E. Altman and A. Shwartz, "Markov decision problems and state-action frequencies," *SIAM J. Contr. Optimiz.*, vol. 29, pp. 786–809, 1991.

[3] E. Altman and A. Shwartz, "Adaptive control of constrained Markov chains: Criteria and policies," *Ann. Oper. Res.*, vol. 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," *Adv. App. Prob.*, vol. 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," *Syst, Contr. Lett.*, vol. 6, pp. 173–180, 1985.

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

[7] V. S. Borkar, "Controlled Markov chains with constraints," *Sandha —Indian Aca. Sci. J. Eng.*, vol. 15, pp. 405–413, 1990.

[8] C. Buyukkov, P. Varaiya, and J. Walrand, "The $c\mu$-rule revisited," *Adv. Appl. Prob.*, vol. 17, pp. 237–238, 1985.

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

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

[11] E. B. Frid, "On optimal strategies in control problems with constraints," *Theory Prob. Appl.*, vol. 17, pp. 188–192, 1972.

[12] D. P. Heyman and M. J. Sobel, *Stochastic Methods in Operations Research II: Stochastic Optimization.* New York: McGraw-Hill, 1984.

[13] A. Hordijk and L. C. M. Kallenberg, "Constrained undiscounted stochastic dynamic programming," *Math. Oper. Res.*, vol. 9, pp. 276–289, 1984.

[14] G. P. Klimov, "Time sharing systems," *Theory Prob. Appl.*, Part I, vol. 19, pp. 532–553, 1974; Part II, vol. 23, pp. 314–321, 1978.

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

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

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

[18] ——, "Stochastic approximations and adaptive control of a discrete-time single server network with random routing," *SIAM J. Contr. Opt.*, vol. 30, 1992.

[19] P. Nain and K. W. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Auto. Contr.*, vol. AC-31, pp. 883–888, 1986.

[20] S. M. Ross, *Introduction to Stochastic Dynamic Programming.* New York: Academic, 1984.

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

[22] L. I. Sennott, "Constrained average-cost Markov decision chains," preprint, 1990.

[23] A. Shwartz and A. M. Makowski, "On the Poisson equation for Markov chains," *EE Pub.*, vol. 646, Technion, under revision, *Math. Oper. Res.*, 1987.

[24] P. Whittle, "Arm-acquiring bandits," *Ann. Probability*, vol. 9, pp. 284–292, 1981.

[25] ——, *Optimization Over Time, Part II; Dynamic Programming and Stochastic Control.* New York: Wiley 1982.

Polynomial LQG Regulator Design for General System Configurations

A. Casavola and E. Mosca

Abstract—**This note deals with a polynomial-equation approach to the linear quadratic Gaussian (LQG) regulation problem for a general system configuration. The solution is given in terms of a left-spectral factorization plus a pair of bilateral Diophantine equations. The resulting control-design procedure is based on an innovations representation of the system. This can be obtained from a physical description by solving, via polynomial equations, a minimum mean-square error (MMSE) filtering problem. The use in cascade of the above two procedures allows one to generalize previous polynomial design results to general system configurations.**

I. Introduction

This note deals with the polynomial-equation approach to the linear quadratic Gaussian regulation (LQGR) problem. The latter is sometimes referred to as the "standard" H_2 optimal-control problem [1]. The system configuration considered here is general and comprises all possible control-system configurations as special cases.

Recently, a transfer-matrix Wiener–Hopf approach to the general LQGR problem was considered in [2] and [3] for the discrete- and continuous-time cases respectively. These solutions, however, suffer from the fact that the controller, not being obtained in irreducible form, is susceptible to exhibit unstable hidden modes, whenever the system is open-loop unstable. In this case, in fact, owing to numerical inaccuracy, stability of the closed-loop system may be lost [2].

The main goal of this note is to explore if the above difficulties can be overcome by using the polynomial-equation approach first introduced by [4]. A possible way to address the problem is to use an innovations representation of the system, obtainable from the physical description by solving a minimum mean-square error filtering (MMSE) problem. This, in turn, can be solved via polynomial equations as well [5] and [6]. Consequently, the whole polynomial LQGR design for general system configurations explicitly involves a two-stage procedure reminiscent of the certainty-equivalence property of stochastic dynamic programming for the LQGR. This approach addresses a more general setting and yields a simpler design method than the direct route used in [7] and [8].

The outline of the note is as follows. In Section II, the polynomial approach to the general LQGR problem is formulated, and conditions are given for its solvability. The polynomial

Manuscript received March 4, 1991; revised January 21, 1992. This work was supported in part by the Italian MURST and CNR.

The authors are with the Dipartimento di Sistemi e Informatica, Università di Firenze, Via di S. Marta, 3-50139 Florence, Italy.

IEEE Log Number 9203116.

0018-9286/93$03.00