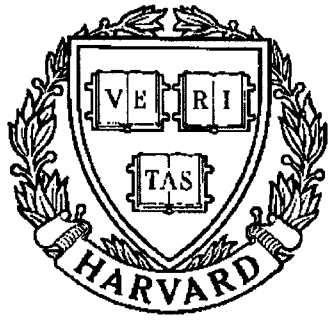
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), the University of Maryland, Harvard University, and Industry

Analysis and Adaptive Control of A Discrete-Time Single-Server Network with Random Routing

by A.M. Makowski and A. Shwartz

TR 89-75r1

ANALYSIS AND ADAPTIVE CONTROL OF A DISCRETE-TIME SINGLE-SERVER NETWORK WITH RANDOM ROUTING

by

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

ABSTRACT

This paper considers a discrete-time system composed of K infinite capacity queues that compete for the use of a single server. Customers arrive in i.i.d batches and are served according to a server allocation policy. Upon completing service, customers either leave the system or are routed instantaneously to another queue according to some random mechanism. As an alternative to simply randomized strategies, a policy based on a stochastic approximation algorithm is proposed to drive a long-run average cost to a given value. The underlying motivation can be traced back to implementation issues associated with constrained optimal strategies.

A version of the ODE method as given by Metivier and Priouret is developed for proving a.s. convergence of this algorithm. This is done by exploiting the recurrence structure of the system under non-idling policies. A probabilistic representation of solutions to an associated Poisson equation is found most useful for proving their requisite Lipschitz continuity. The conditions which guarantee convergence are given directly in terms of the model data. The approach is of independent interest, as it is not limited to this particular queueing application and suggests a way of attacking other similar problems.

Keywords: stochastic approximations, stochastic adaptive control, queueing network.

AMS subject classification: 90B22 Queueing Theory—service models, 90B50 Decision Theory, 93E20 Optimal stochastic control.

Abbreviated title: Adaptive control of a queueing network.

Submitted June 1989. Revised January 1991.

---

[1] Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742, U.S.A. The work of this author was supported partially through ONR Grant N00014-84-K-0614, partially through NSF Grant ECS-83-51836, and partially through United States — Israel Binational Science Foundation Grant BSF 85-00306.

[2] Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa 32000, Israel. The work of this author was supported partially through United States — Israel Binational Science Foundation Grant BSF 85-00306 and partially through a grant from AT&T Bell Laboratories.

I. INTRODUCTION

1.1 Stochastic approximations on Markov chains

In recent years, there has been widespread interest in stochastic approximation algorithms as a means to solve complex engineering problems [5,15]. As a result of the increasing complexity of applications, focus has shifted from the original Robbins-Monro algorithm to more sophisticated versions, and this has led in particular to the study of *projected* stochastic approximation algorithms driven by *Markovian* "noise" or "state" processes.

Such algorithms have the following form. Let $\{\eta(n),\ n = 0,1,\ldots\}$ be the sequence of iterates which take values in a compact convex subset U of $\mathbb{R}^p$, and let $\{X(n),\ n = 0,1,\ldots\}$ be the state process which takes values in some Borel subset S of $\mathbb{R}^K$. They are related by the recursion

$$\eta(0) \in U\ ,\quad \eta(n+1) = \Pi_U\{\eta(n) + a_{n+1} f(\eta(n), X(n+1))\} \qquad n = 0,1,\ldots (1.1)$$

where Π_U denotes the nearest-point projection on U, f is a Borel mapping $U \times S \to \mathbb{R}^p$ and the step size sequence $\{a_{n+1},\ n = 0,1,\ldots\}$ satisfies the conditions $0 < a_n \downarrow 0$, $\sum_{n=0}^{\infty} a_n = \infty$ and $\sum_{n=0}^{\infty} a_n^2 < \infty$. A complete specification of the algorithms (1.1) requires that the conditional probability distribution of $X(n+1)$ given $(X(0), \eta(0), X(1), \ldots, X(n), \eta(n))$ be postulated for each $n = 0,1,\ldots$. For instance, the Markovian dependencies alluded to earlier require

$$P[X(n+1) \in B | X(0), \eta(0), X(1), \ldots, X(n), \eta(n)] = \mu_{\eta(n)}(X(n); B) \qquad n = 0,1,\ldots (1.2)$$

for every Borel subset B of S, where $\{\mu_\eta,\ \eta \in U\}$ is a family of one-step probability transition kernels on S.

The central question in the theory of stochastic approximations is concerned with the convergence properties of the iterate sequence $\{\eta(n),\ n = 0,1,\ldots\}$. For the Robbins-Monro algorithm, direct martingale arguments have been given by Gladyshev [10] to establish a.s. convergence. However, in more complex situations such as (1.2), the direct probabilistic approach does not work, and this failure prompted the development of the so-called ODE method. In most of its forms, the ODE method proceeds in two separate steps. The first step relies on the Kushner-Clark Lemma in order to identify a deterministic ODE, the stability properties of which determine the limit points of $\{\eta(n),\ n = 0,1,\ldots\}$. The second step is probabilistic in nature and depends on the algorithm being considered; its purpose is to show that asymptotically (in the mode of convergence of interest) the output sequence of the original algorithm behaves like the solution to the ODE.

In their monograph [16], Kushner and Clark have given general conditions for successfully completing this second step. In more structured situations [15], Kushner has shown how weak convergence methods—through various tightness properties— pave the way to convergence in probability of the sequence $\{\eta(n),\ n = 0, 1, \ldots\}$. In the Markovian case, Metivier and Priouret [21] have established a.s. convergence by making use of properties of the Poisson equation associated with the transition kernels $\{\mu_\eta,\ \eta \in U\}$ appearing in (1.2). Key to their analysis are various properties of Lipschitz continuity (in η) of the solution to this Poisson equation.

Unfortunately, in all these references, the conditions underlying the second step of the ODE method are given in implicit form and are often hard to verify in specific situations. What seems desirable is a more operational convergence theory where conditions are given *directly* in terms of the model data. For instance, this was done by the authors in the Markovian situation [17] when the state space S is *finite*, in which case (1.2) reduces to

$$P[X(n+1) = y | X(0), \eta(0), X(1), \ldots, X(n), \eta(n)] = p_{X(n)y}(\eta(n)), \quad y \in S \qquad n = 0, 1, \ldots (1.3)$$

for some family $\{P(\eta),\ \eta \in U\}$ of one-step transition probabilities with $P(\eta) \equiv (p_{xy}(\eta))$. A comprehensive convergence theory was developed under the mild condition of Lipschitz continuity for the one-step transition probabilities $\eta \to p_{xy}(\eta)$. This was achieved by using a variant of the approach proposed by Metivier and Priouret, and leads to an a.s. convergence result.

When the state space S is *countably infinite*, the situation is much more difficult and no general results seem available, which guarantee a.s. convergence in terms of *explicit* conditions on the model data. The main technical difficulty in the approach of Metivier and Priouret—used successfully in the finite case [17]— stems from the fact that several quantities of interest are no longer bounded and that the requisite properties of the solution to the Poisson equation are now much harder to obtain. This paper presents arguments for establishing both these smoothness properties and the a.s. convergence of the algorithm. The general framework of interest is described in Section 2, and is couched in the formalism of the theory of Markov decision processes; this is done for notational convenience as will become apparent in later sections. The approach advocated here relies on the recurrence structure of the (controlled) system [19], and on a probabilistic interpretation of the solution to the Poisson equation derived from it [27]. These arguments are developed in the context of an adaptive control problem for a specific queueing system, namely a discrete-time single-server network with random routing, which is described in Section 3. The approach presented here is of much wider applicability and should be of use in analyzing a large class of projected stochastic approximations driven by a Markov chain on a countable state space. The main advantage of

discussing a concrete application rather than the general situation, lies in the fact that the key arguments can then be provided in their simplest form, unencumbered by often confusing technicalities, under *verifiable* conditions given solely in terms of the model data. However, to help the reader apply the ideas proposed here to other situations, each one of the Sections 5–7 ends with an outline of more general technical conditions which permits a development similar to the one given here.

1.2 A time-sharing queueing system

The queueing system considered here is now briefly described; a precise model formulation is available in Section 3: Consider a system composed of K infinite capacity queues that compete for the use of a single server. Time is slotted with the service requirement of each customer corresponding exactly to one time slot. At the beginning of each time slot, the controller gives priority to one of the queues according to some prespecified dynamic priority assignment, and the selected queue is given service attention during that slot. However, due to a variety of reasons ranging from server failure to exogenous interferences, with a positive probability, the service fails, in which case the service of that customer is rescheduled at a later time in accordance with the service allocation policy. When in a given time slot the service succeeds, the customer is either declared serviced and leaves the system at the end of the slot, or is routed to one of the other queues with a fixed probability, depending on both source and destination queues. In the present paper, the failures are assumed generated through *independent Bernoulli* processes, with possibly class-dependent parameters, and this *independently* of the arrival mechanism. New customers may arrive in batches which are modelled as an arbitrary K-dimensional *renewal* process; this captures possible partial correlations between arrivals from different classes in a given slot.

This queueing system and its variants constitute useful models for studying resource allocation issues in several application areas, including computer systems and data networks, and as such they have received a great deal of attention in recent years. Klimov [13] studied a continuous-time version of this system, and proved that a strict priority policy minimizes the discounted cost associated with a cost-per-slot linear in the queue sizes. Tsoucas and Walrand [29] considered an adaptive version of Klimov's problem where the service distributions are unknown.

The case where no routing is allowed has been much studied. For such systems, Sidi and Segall [28] derived the joint equilibrium distribution of the queue size under a fixed priority scheme. Several authors [3,4,8,11] have shown that the μc-rule minimizes a variety of performance measures associated with the aforementioned linear cost structure. In [22], Nain and Ross considered the

situation where several types of traffic, say voice, video and data, compete for the use of a single synchronous communication channel. They formulated this situation as a system of K discrete-time queues and found the service allocation strategy minimizing the long-run average of a linear expression in the queue sizes of $K-1$ customer classes, under the constraint that the long-run average queue size of the remaining customer class does not exceed a certain value. Extending some of the optimality results from Baras, Ma and Makowski [4], they showed that if the constraint can be met, then the optimal policy g is a Markov stationary policy with the following structure: There exist two *static* work-conserving service assignment policies (of which μc-rules are only one description), say $\overline{g}$ and $\underline{g}$, and a scalar η^* in $(0,1)$. At the beginning of each time slot, a coin with bias η^* is flipped, and the policy g implements channel rights according to the outcome via $\overline{g}$ and $\underline{g}$ with probability η^* and $1-\eta^*$, respectively. The bias η^* is determined so as to meet the constraint. This result was extended by Altman and Shwartz to the case where the constraint is also given through a linear combination [1,2].

These results are typical in the broader context of MDPs in that analysis often identifies a policy g of interest which is Markov stationary. In fact, for the problem of minimizing one average cost subject to a constraint on another such cost, an optimal policy which "mixes" two deterministic policies in the manner described above exists under very general conditions as demonstrated by several authors [6, 7, 24]. Unfortunately, this policy may not be readily implementable due either to a lack of knowledge of the actual values of some parameters [14] or to *computational* difficulties inherent to its definition. The situation treated by Nain and Ross is a good case in point, for there non-trivial off-line computations are required in order to actually compute the value of the bias η^*, even if all parameters are known.

1.3 Overview of the paper

This implementation issue provides the motivation for the stochastic approximation studied in this paper. In Section 4, the issue is discussed in the broader context of "steering the cost to a given value", with a view towards applications to constrained optimization [1,2,22]. The problem is now one of finding the bias η^* needed in a simple randomization between two policies $\underline{g}$ and $\overline{g}$ in order to steer a long-run average cost to a given value. The resulting randomized Markov stationary policy—denoted g hereafter—can be implemented by means of a projected stochastic approximation. This algorithm computes on-line estimates of η^* which are then used in a Certainty Equivalence controller α derived from the special form of g. Theorems 4.1 and 4.2 contain the main results concerning the performance of this policy α, namely that the policies α and g yield the same

value for the long-run average cost, and that the iterates $\{\eta(n),\ n = 0, 1, \ldots\}$ converge a.s. under α to the bias value η^*. This improves on earlier results of the authors [25] for the same algorithm in the context of the two-queue system with no routing. There, only convergence in probability was established, albeit under weaker conditions on moments.

The convergence proof for the stochastic approximation algorithm hinges on the availability of bounds on moments of the queue size process which are uniform in the policy, and on the smoothness properties of solutions to an associated Poisson equation [21, 27]. The bounds are obtained in Section 5 by means of renewal arguments which relate the queue size to the recurrence times to the empty state. In Section 6, novel arguments are developed for proving the Lipschitz continuity for solutions to the Poisson equation and for establishing bounds on them. It is appropriate to stress the methodological value of both Sections 5 and 6, in that ideas therein are by no means restricted to the competing queue model or to the randomization of two policies, and can be used *mutatis mutandis* in many other situations. However, the approach was developed here for a stochastic approximation algorithm for a specific model, rather than for general Markov chains with countable state spaces, in order to present the arguments more clearly, unencumbered from technical details and assumptions which often accompany more formal treatments.

The a.s. convergence of the stochastic approximation scheme defining the implementation α is established in Section 7, where the various estimates of the previous sections allow for a rather simple proof. Finally, the cost properties of the policy α are discussed in Section 8 by making use of the convergence of the stochastic approximation and by invoking the results on the "Certainty Equivalence" Principle developed in [26]; the requisite hypotheses of [26] are easily verified for this system with the help of bounds on solutions to the Poisson equation. The paper concludes with an application to the constrained optimization problem discussed by Nain and Ross in [22]. All necessary conditions are verified and the policy α thus constitutes an implementation of the Markov stationary policy which is constrained optimal for this problem.

2. A GENERAL MODEL

This section introduces a general class of projected stochastic approximations driven by Markovian noise. The formalism of the theory of Markov decision processes [REF] was found notationally convenient in defining the class of stochastic approximation schemes of interest. Indeed, this more general framework lends itself naturally to the presentation of more general conditions as done at the end of Sections 5–7.

A few words on the notation and conventions used throughout the paper. The set of all non-negative integers is denoted by $\mathbb{N}$, and $\mathbb{R}$ (resp. $\mathbb{R}_+$) stands for the set of all real (resp. positive real) numbers. The indicator function of a set A is denoted by $I[A]$. Unless stated otherwise, the notation $\lim_n$ and $\overline{\lim}_n$ are understood with n going to infinity.

2.1 The MDP formulation

To set up the discussion, first consider a MDP (S, U, P) as defined in the literature [REF] where the state space S is a *countable* set—for sake of concreteness it will be convenient to take $S = \mathbb{N}^K$—and the action space U is a *compact convex* subset of $\mathbb{R}^p$. The one-step transition mechanism P is defined through the one-step transition probability functions $U \to \mathbb{R} : u \to p_{xy}(u)$, x, y in S, which are assumed to be *Borel* measurable and to satisfy the standard properties

$$0 \leq p_{xy}(u) \leq 1, \quad \sum\nolimits_y p_{xy}(u) = 1, \quad u \in U,\ x, y \in S. \tag{2.1}$$

The space of probability measures on U (when equipped with its natural Borel σ-field) is denoted by $\mathbb{M}(U)$. An *admissible* control policy π is defined as any collection $\{\pi_n, n = 0, 1, \ldots\}$ of mappings $\pi_n : S \times (U \times S)^n \to \mathbb{M}(U)$ such that for all $n = 0, 1, \ldots$ and every Borel subset B of U, the mapping $S \times (U \times S)^n \to [0, 1] : h_n \to \pi_n(h_n; B)$ is Borel measurable. The collection of all such admissible policies is denoted by $\mathcal{P}$.

The definition of the MDP (S, U, P) postulates the existence of a measurable space $(\Omega, \mathcal{F})$ and of a collection of probability measures $\{P^\pi,\ \pi \in \mathcal{P}\}$ such that the conditions (2.3)–(2.5) below are satisfied. The measurable space $(\Omega, \mathcal{F})$ is chosen large enough to carry the sequences of S-valued rvs $\{X(n),\ n = 0, 1, \ldots\}$ and U-valued rvs $\{U(n),\ n = 0, 1, \ldots\}$, with the interpretation that $X(n)$ denotes the state of the system at time n and $U(n)$ represents the action taken in that state. The feedback information is encoded through the rvs $\{H(n),\ n = 0, 1, \ldots\}$ defined by $H(0) := X(0)$ and

$$H(n) := (X(0), U(0), X(1), \ldots, U(n-1), X(n)); \qquad n = 1, 2, \ldots \tag{2.2}$$

the rv $H(n)$ takes values in $\mathbb{H}_n := S \times (U \times S)^n$, and set $\mathcal{F}_n = \sigma\{H(n)\}$.

To complete the definition, let μ be a fixed probability distribution on S. For every admissible policy π in $\mathcal{P}$, the probability measure P^π is constructed on $(\Omega, \mathcal{F})$ such that under P^π, the rv X_0 has distribution μ, the state transitions are realized according to

$$P^\pi[X(n+1) = y \mid \mathcal{F}_n \vee \sigma\{U(n)\}] = p_{X(n)y}(U(n)), \quad y \in S \qquad n = 0, 1, \ldots \tag{2.3}$$

and the control actions are selected according to

$$P^{\pi}[U(n) \in B \mid \mathcal{F}_n] = \pi_n(B; H(n)) \qquad n = 0, 1, \ldots (2.4)$$

for every Borel subset of U. Consequently,

$$P^{\pi}[X(n+1) = y \mid \mathcal{F}_n] = \int_U p_{X(n)y}(u)\pi_n(du; H(n)), \quad y \in S. \qquad n = 0, 1, \ldots (2.5)$$

The expectation operator associated with π is denoted by E^{π}.

The measurable space $(\Omega, \mathcal{F})$ is often selected to be the so-called canonical space, i.e., Ω is the cartesian product $\Omega := S \times (U \times S)^{\infty}$ endowed with the natural Borel structure inherited from the product topology. However, in many concrete situations, it is more convenient to describe the underlying MDP on a measurable space $(\Omega, \mathcal{F})$ which is somewhat larger than the canonical space. For example, in the setup considered in this paper, additional rvs are needed to encode arrivals, service completions and random routing in the queueing system. In this case the definitions of $\mathbb{H}_n$, $H(n)$ and $\mathcal{F}_n$ are changed accordingly in the obvious way.

Following standard usage, a policy π in $\mathcal{P}$ is said to be a *Markov* or *memoryless* policy if there exists a family $\{g_n,\ n = 0, 1, \ldots\}$ of Borel mappings $g_n : S \to \mathbb{M}(U)$ such that $\pi_n(\cdot; H(n)) = g_n(\cdot; X(n))$ $P^{\pi} - a.s.$ for all $n = 0, 1, \ldots$. In the event the mappings $\{g_n,\ n = 0, 1, \ldots\}$ 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. Under any Markov stationary g, the state process $\{X(n),\ n = 0, 1, \ldots\}$ evolves according to a Markov chain with one-step transition probability matrix $P(g) \equiv (p_{xy}(g))$ given by

$$p_{xy}(g) := \int_U p_{xy}(u)g(du, x), \quad x, y \in S. \qquad (2.6)$$

Finally, a policy π in $\mathcal{P}$ is said to be deterministic or non-randomized policy if there exists a sequence of Borel mappings $\{f_n,\ n = 0, 1, \ldots\}$ such that for each $n = 0, 1, \ldots$, the mapping $f_n : \mathbb{H}_n \to U$ is Borel measurable and the probability measure $\pi_n(\cdot; H(n))$ is a point mass distribution concentrated at $f_n(H(n))$ $P^{\pi} - a.s.$

2.2 The stochastic approximation

Stochastic approximations on Markov chains—as defined by (1.1) and (1.3)—can be interpreted as deterministic policies for the MDP (S, U, P) described earlier. To see this, start with a mapping

$c : S \to \mathbb{R}^p$ and let $\{\eta(n),\ n = 0,1,\ldots\}$ be the sequence of U-valued rvs determined by the recursion

$$\eta(0) \in U\ , \quad \eta(n+1) = \Pi_U\{\eta(n) + a_{n+1}c\big(X(n+1)\big)\}\ . \qquad n = 0,1,\ldots (2.7)$$

As before, Π_U denotes the nearest-point projection on U, and the step size sequence $\{a_{n+1},\ n = 0,1,\ldots\}$ satisfies the usual conditions

$$0 < a_n \downarrow 0, \quad \sum_{n=0}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n^2 < \infty. \tag{2.8}$$

The policy associated with the recursion (2.7) is the deterministic policy $\alpha = \{\alpha_n,\ n = 0,1,\ldots\}$ with the property that for all $n = 0,1,\ldots$, $\alpha_n(\cdot; H(n))$ is the point mass distribution concentrated at $\eta(n)$. That this policy is indeed admissible follows from the fact that for each $n = 0,1,\ldots$, the rv $\eta(n)$ can expressed as a function of the successive states $X(0), X(1), \ldots, X(n)$.

3. THE DISCRETE–TIME KLIMOV MODEL

This section presents in some details the model for the controlled queueing system briefly described in the introduction. First, a few words on the notation and convention in use. Elements of $\mathbb{R}^K$ are always interpreted as $K \times 1$ column vectors, and the k^{th} component of any element x of $\mathbb{R}^K$ is denoted by x_k, $1 \le k \le K$, with a similar convention for rvs. Thus an element x of $\mathbb{R}^K$ can also be written as $(x_1, ..., x_K)'$ (with $'$ denoting transpose), and its norm is given by $|\ x\ | := \sum_{k=1}^K |x_k|$. The elements e and 0 of $\mathbb{R}^K$ are defined as the vectors $e = (1, ..., 1)'$ and $0 = (0, ..., 0)'$ with identical components. The standard basis $\{e^1, ..., e^K\}$ for $\mathbb{R}^K$ is denoted by $\mathcal{B}_K$, while $\mathcal{S}_K$ is the standard simplex defined by

$$\mathcal{S}_K := \{p \in \mathbb{R}^K : \sum_{k=1}^{K} p_k = 1 \quad \text{and} \quad 0 \le p_k \le 1, \quad 1 \le k \le K\}. \tag{3.1}$$

It is plain that $\mathcal{S}_K$ can be identified with $\mathbb{M}(U)$ when $U = \mathcal{B}_K$.

3.1 The basic random variables

The controlled queueing system of interest, the so-called discrete-time Klimov model, will be defined as a MDP. All probabilistic elements are defined on a single sample space Ω equipped with the σ-field of events $\mathcal{F}$. This sample space carries the basic rvs Ξ, $\{U(n),\ n = 0,1,\ldots\}$, $\{A(n),\ n = 0,1,\ldots\}$, $\{B(n),\ n = 0,1,\ldots\}$ and $\{R(n),\ n = 0,1,\ldots\}$ which take values in $\mathbb{N}^K$, $\mathcal{B}_K$,

$\mathbb{N}^K$, $\{0,1\}^K$ and $\{0,1,\ldots K\}^K$, respectively. These quantities have a ready interpretation in the context of the queueing system described in the introduction: The number of customers initially in the k^{th} queue is set at Ξ_k and for each $n = 0, 1, \ldots$, the state of the system is represented by a rv $X(n)$ of integer components with the interpretation that at the beginning of the slot $[n, n+1)$, $X_k(n)$ customers are present in the k^{th} queue, including the one receiving service. The following chain of events occurs:

(i): The control action $U(n)$ is selected with the convention that $U_k(n) = 1$ (resp. $U_k(n) = 0$) if the k^{th} queue is (resp. is not) given service attention during that slot. The fact that $U(n)$ takes values in $\mathcal{B}_K$ guarantees that exactly one queue is given service attention;

(ii): New customers arrive into the system according to the rv $A(n)$ with $A_k(n)$ new customers joining the k^{th} queue;

(iii): A completion of service possibly occurs at the queue that was given service attention during the slot. This is encoded in the binary rv $B(n)$, where $B_k(n) = 1$ (resp. $B_k(n) = 0$) signifies successful completion (resp. abortion) of service for the k^{th} queue conditioned on it being given service attention and non-empty; and

(iv): If a service completion occurs at the queue that was given service attention during the slot, then instantaneously the serviced customer is either transferred to another queue or it leaves the network. This routing decision is implemented through the variable $R(n)$ with the following interpretation. If the service completion occurred at the k^{th} queue, then $R_k(n) = \ell$, $1 \leq \ell \leq K$, means that the serviced customer joins the ℓ^{th} queue while $R_k(n) = 0$ expresses the fact that this customer leaves the system.

As a result of (i)-(iv), the successive system states or queue size vectors form a sequence $\{X(n),\ n = 0, 1, \ldots\}$ of $\mathbb{N}^K$-valued rvs which are generated componentwise through the recursion

$$X_k(n+1) = X_k(n) + A_k(n) - I[X_k(n) \neq 0]U_k(n)B_k(n) + \sum_{\ell=1}^{K} I[X_\ell(n) \neq 0]U_\ell(n)B_\ell(n)I[R_\ell(n) = k]$$

$$1 \leq k \leq K, \quad n = 0, 1, \ldots \quad (3.2)$$

with $X(0) := \Xi$.

At the beginning of each time slot $[n, n+1)$, the decision-maker has knowledge of the rv $H(n)$ which here includes the initial queue sizes Ξ, the past arrival pattern $A(i), 0 \leq i < n$, past decisions

$U(i), 0 \leq i < n$, past service completions $B(i), 0 \leq i < n$ and past routing decisions $R(i), 0 \leq i < n$. The rvs $\{H(n),\ n = 0, 1, \ldots\}$ are thus given recursively by

$$H(0) = \Xi\ , \quad H(n+1) := (H(n), U(n), A(n), B(n), R(n)); \qquad n = 0, 1, \ldots \tag{3.3}$$

The information contained in $H(n)$ is used to generate the control value $U(n)$ implemented in the slot $[n, n+1)$. The selection of this control value is done according to a prespecified mechanism, which may be either deterministic or random.

3.2 The probabilistic structure

Since randomized strategies are allowed, an admissible control policy π is defined as any collection $\{\pi_n,\ n = 0, 1, \ldots\}$ of mappings $\pi_n : \mathbb{H}_n \to \mathcal{S}_K$, with the interpretation that at times $n = 0, 1, \ldots$, the k^{th} queue is given service attention with probability $\pi_n(k; h_n)$ whenever the information vector h_n in $\mathbb{H}_n$ is available to the system controller. Denote the collection of all such admissible policies by $\mathcal{P}$.

Let $q_{\Xi}(\cdot)$ and $q(\cdot)$ be two probability mass distributions on $\mathbb{N}^K$, and fix a service rate vector μ in $(0, 1]^K$. Moreover, let P denote a $K \times K$ substochastic matrix $(p_{k\ell},\ 1 \leq k, \ell \leq K)$, i.e.,

$$0 \leq p_{k\ell} \leq 1 \quad \text{and} \quad \sum_{\ell=1}^{K} p_{k\ell} \leq 1 \quad , \quad 1 \leq k, \ell \leq K \tag{3.4}$$

and set

$$p_{k0} := 1 - \sum_{\ell=1}^{K} p_{k\ell} \quad , \quad 1 \leq k \leq K\ . \tag{3.5}$$

Throughout the discussion, the *non-degeneracy* condition

$$0 < q(0) < 1 \tag{3.6a}$$

and the *finite mean* condition

$$\sum_{a \in \mathbb{N}^K} |a| q(a) < \infty \tag{3.6b}$$

are enforced. It is always assumed that the matrix $I - P$ is *invertible*, a condition which is equivalent to the system being open, i.e., every customer eventually leaves the system with probability one.

The model is now completely specified by *postulating* the existence of a family $\{P^\pi,\ \pi \in \mathcal{P}\}$ of probability measures on the σ-field $\mathcal{F}$ which satisfies the requirements **(R1)–(R3)** below, i.e., for every policy π in $\mathcal{P}$,

(R1): For all x in $\mathbb{N}^K$,

$$P^\pi[\Xi = x] := q_\Xi(x) \ ;$$

(R2): For all a in $\mathbb{N}^K$, b in $\{0,1\}^K$ and r in $\{1,2,\ldots,K\}^K$,

$$\begin{aligned} & P^\pi[A(n) = a,\ B(n) = b,\ R(n) = r \mid \mathcal{F}_n \vee \sigma\{U(n)\}] \\ :=& P^\pi[A(n) = a] P^\pi[B(n) = b] P^\pi[R(n) = r] \\ =& q(a) \cdot \prod_{k=1}^{K} (b_k \mu_k + (1 - b_k)(1 - \mu_k)) \cdot \prod_{k=1}^{K} p_{kr_k} \end{aligned} \qquad n = 0, 1, \ldots$$

where $\mathcal{F}_n = \sigma\{H(n)\}$ with $H(n)$ defined by (3.3).

(R3): For all e^k in $\mathcal{B}_K$, $1 \leq k \leq K$,

$$P^\pi[U(n) = e^k | \mathcal{F}_n] := \pi_n(k; H_n). \qquad n = 0, 1, \ldots$$

The existence of a sample space $(\Omega, \mathcal{F})$ that carries such a family of probability measures $\{P^\pi,\ \pi \in \mathcal{P}\}$ is easily established via the Kolmogorov Extension Theorem, by taking Ω to be the canonical space $\mathbb{N}^K \times (\mathcal{B}_K \times \mathbb{N}^K \times \{0,1\}^K \times \{0,1,\ldots,K\}^K)^\infty$ equipped with its natural σ-field. This modeling approach was adopted in [25] for a special case of the Markov decision process under consideration; the reader is referred there for additional information.

The reader will readily check that for each policy π in $\mathcal{P}$, the following properties **(P1)–(P4)** hold true under P^π, where

(P1): The $\mathbb{N}^K$-valued rv Ξ and the sequences of rvs $\{A(n),\ n = 0, 1, \ldots\}$, $\{B(n),\ n = 0, 1, \ldots\}$ and $\{R(n),\ n = 0, 1, \ldots\}$ are *mutually independent*;

(P2): The sequences $\{B_k(n),\ n = 0, 1, \ldots\}$ of $\{0,1\}$-valued rvs are *mutually independent i.i.d. Bernoulli* sequences with parameters μ_k, $1 \leq k \leq K$;

(P3): The sequences $\{R_k(n),\ n = 0, 1, \ldots\}$ of $\{0, 1, \ldots, K\}$-valued rvs are *mutually i.i.d.* sequences with

$$P^\pi[R_k(n) = \ell] = p_{k\ell}, \quad 1 \leq k, \ell \leq K \ ; \qquad n = 0, 1, \ldots \quad (3.7)$$

(P4): The $\mathbb{N}^K$-valued rvs $\{A(n),\ n = 0, 1, \ldots\}$ form a sequence of *i.i.d.* rvs with common probability distribution $q(\cdot)$.

For $1 \leq k \leq K$, denote by λ_k the first moment of the sequence $\{A_k(n), \quad n = 0, 1, \ldots\}$ and set $\nu_k = {\mu_k}^{-1}$. For future use, define the *network traffic* coefficient ρ by

$$\rho := \lambda'(I - P)^{-1}\nu \tag{3.8}$$

where $\lambda = (\lambda_1, \ldots, \lambda_K)'$ and $\nu = (\nu_1, \ldots, \nu_K)'$.

A policy π in $\mathcal{P}$ is said to be *non-idling* or *work-conserving* whenever for all $1 \leq k \leq K$, the condition

$$\left[\pi_n(k; H(n)) > 0, \ X(n) \neq 0\right] = \left[\pi_n(k; H(n)) > 0, \ X_k(n) \neq 0\right] \qquad n = 0, 1, \ldots \tag{3.9}$$

holds true P^π-a.s.

4. PROBLEM FORMULATION

Let c denote a mapping $\mathbb{N}^K \to \mathbb{R}$. For any admissible policy π in $\mathcal{P}$, set

$$J(\pi) := \overline{\lim}_n E^\pi\Big[\frac{1}{n+1}\sum_{i=0}^{n} c(X(i))\Big] \tag{4.1}$$

(whenever meaningful) with the usual interpretation that $J(\pi)$ is a measure of system performance when using the policy π.

4.1 Steering the cost

Constrained MDPs lead to optimal stationary policies which randomize between several stationary deterministic policies [6, 7, 23, 24]. Given the constituent deterministic policies, the problem of finding the optimal policy reduces to simultaneously *steering* constraint functionals of the form (4.1) to given values. For simplicity, only the scalar case (arising from a single constraint) is discussed here, in which case the steering problem consists in finding a Markov stationary policy g such that $J(g) = V$ for some given constant V. The discussion is given under the assumption that there exist two Markov (possibly randomized) stationary policies $\underline{g}$ and $\overline{g}$ such that

$$J(\underline{g}) < V < J(\overline{g}). \tag{4.2}$$

For every η in the unit interval [0,1] the policy f^η is the Markov stationary policy obtained by simply randomizing between the two policies $\underline{g}$ and $\overline{g}$ with *bias* η; it is determined through the mapping $f^\eta : \mathbb{N}^K \to \mathcal{S}_K$ where

$$f^\eta(k; x) := \eta\overline{g}(k; x) + (1 - \eta)\underline{g}(k; x), \quad x \in \mathbb{N}^K, \quad 1 \leq k \leq K. \tag{4.3}$$

Note that for $\eta = 1$ (resp. $\eta = 0$) the randomized policy f^η coincides with $\overline{g}$ (resp. $\underline{g}$). Owing to (4.2), if the mapping $\eta \to J(f^\eta)$ is *continuous* on the interval [0,1], then at least one randomized strategy f^{η^*} meets the value V and its corresponding bias value η^* is a solution of the equation

$$J(f^\eta) = V, \quad 0 \leq \eta \leq 1, \tag{4.4}$$

so that the identification $g = f^{\eta^*}$ may take place.

4.2 Implementation issues

Solving the (highly) nonlinear equation (4.4) for the bias value η^* is usually a non-trivial task, even in the simplest of situations [18, 22]. This difficulty is circumvented by proposing alternatives to the policy g which bypass a *direct* solution of the equation (4.4). One possible approach is to design (simple recursive) schemes for estimating the value η^* which solves (4.4) and then to define a so-called "naive feedback" policy $\alpha = \{\alpha_n,\ n = 0, 1, \ldots\}$ via the *Certainty Equivalence Principle.* Such a policy α can be written in the form

$$\alpha_n := \eta(n)\overline{g} + (1 - \eta(n))\underline{g} \qquad n = 0, 1, \ldots (4.5)$$

for some sequence of [0,1]-valued rvs $\{\eta(n),\ n = 0, 1, \ldots\}$ which act as "estimates" for the bias value η^*. It is hoped that the effects of controlling and learning about the system will combine to produce a *consistent* estimation scheme. In such a case, the sequence of estimates $\{\eta(n),\ n = 0, 1, \ldots\}$ converges to the value η^* in some sense, thus providing increasingly better approximations to the appropriate bias value. This policy α will constitute an acceptable *implementation* of g provided $J(\alpha) = J(g)$.

At this point, the reader may wonder as to how such an estimation scheme can be selected. If the function $\eta \to J(f^\eta)$ were *continuous* and *strictly monotone* (necessarily increasing by (4.2)–(4.3)), then the search for η^* could be interpreted as finding the zero of the continuous, strictly monotone function $\eta \to J(f^\eta) - V$, and this brings to mind ideas from the theory of *stochastic approximations* [16]. Here, the Robbins-Monro version of these algorithms suggests that a sequence of bias values $\{\eta(n),\ n = 0, 1, \ldots\}$ be generated through the recursion

$$\eta(0) \in U\ , \quad \eta(n+1) = \Big[\eta(n) + a_{n+1}(V - c(X(n+1))\Big]_0^1 . \qquad n = 0, 1, \ldots (4.6)$$

In (4.6) the notation $[x]_0^1 = 0 \vee (x \wedge 1)$ is used for every x in $\mathbb{R}$, and the sequence of step sizes $\{a_n,\ n = 1, 2, \ldots\}$ satisfies the conditions (2.8).

Lemma 8.2 gives a set of conditions under which the monotonicity property holds.

4.3 The results

This paper is devoted to analyzing the performance of the adaptive policy α defined through (4.5)–(4.6). The main results in this direction are described below and require the additional assumptions **(R4)–(R6)** on the data of the problem, where

(R4): There exists some integer $\gamma \geq 1$ such that for every policy π in $\mathcal{P}$, the moment conditions

$$E^{\pi}[\ |\Xi|^{\gamma}] = \sum_{x \in \mathbb{N}^K} |x|^{\gamma} q_{\Xi}(x) < \infty$$

and

$$E^{\pi}[\ |A(n)|^{\gamma}] = \sum_{a \in \mathbb{N}^K} |a|^{\gamma} q(a) < \infty \qquad n = 0, 1, \ldots$$

hold true;

(R5): There exist an integer $\delta > 0$ and a constant $L > 0$ in $\mathbb{R}$ such that

$$|c(x)| \leq L(1 + |\ x\ |^{\delta}) =: \tilde{c}(|x|), \quad x \in \mathbb{N}^K;$$

and

(R6): The policies $\overline{g}$ and $\underline{g}$ are *non-idling* Markov stationary policies such that (4.2) holds.

The results concerning the policy α are now summarized.

Theorem 4.1. *Assume* **(R1)–(R6)** *to hold with* $\rho < 1$ *and let the integer exponent* γ *in* **(R4)** *and* δ *in* **(R5)** *satisfy the condition*

$$2\delta + 3 \leq \gamma. \tag{4.8}$$

If the mapping $\eta \to J(f^{\eta})$ *is strictly monotone, then*

$$\lim_n \eta(n) = \eta^* \quad P^{\alpha} - \text{a.s.} \tag{4.9}$$

Under these conditions, the system also satisfies a Certainty Equivalence Principle [26] which takes the following form.

Theorem 4.2. *Assume* **(R1)–(R6)** *to hold with* $\rho < 1$ *and let the integer exponent* γ *in* **(R4)** *and* δ *in* **(R5)** *satisfy the condition*

$$\max\{3, 1 + \delta(1 + \epsilon)\} \leq \gamma \tag{4.10}$$

for some $\epsilon > 0$. *If* $\lim_n \eta(n) = \eta^*$ *in probability under* P^α, *then the convergence*

$$J(\alpha) = \lim_t \frac{1}{t+1} \sum_{s=0}^{t} c(X(s)) = J(g) \tag{4.11}$$

takes place in $L^1(\Omega, \mathcal{F}, P^\alpha)$, *so that*

$$J(\alpha) = \lim_t E^\alpha\Big[\frac{1}{t+1} \sum_{s=0}^{t} c(X(s))\Big] = J(g). \tag{4.12}$$

Moreover, for any other mapping $d : \mathbb{N}^K \to \mathbb{R}$, *if there exist an integer* $\delta' > 0$ *and a constant* $L' > 0$ *such that*

$$|d(x)| \leq L'(1 + |\, x \,|^{\delta'}), \quad x \in \mathbb{N}^K \tag{4.13}$$

then both (4.11) and (4.12) hold for the long-run average cost (4.1) associated with d *provided the condition*

$$\max\{3, 1 + \delta'(1 + \epsilon')\} \leq \gamma \tag{4.14}$$

holds for some $\epsilon' > 0$.

The restriction that δ and δ' be integers is not essential but results in some simplifications in the notation. An example where the hypotheses of Theorems 4.1–4.2 hold is given in Section 8.

This section closes with a few facts which are easily derived from the enforced assumptions: Under **(R6)**, the policies $f^\eta, 0 \leq \eta \leq 1$, and α are *all* non-idling since $\overline{g}$ and $\underline{g}$ are non-idling. Moreover, note from (3.2) that

$$X_k(n+1) \leq X_k(n) + A_k(n) + 1, \quad 1 \leq k \leq K. \qquad n = 0, 1, \ldots \tag{4.15}$$

and therefore, by virtue of **(R4)**, $E^\pi[|X(n)|^\gamma] < \infty$ for all $n = 0, 1, \ldots$ under any policy π in $\mathcal{P}$. Since $\delta \leq \gamma$ under either (4.8) or (4.10), it is then immediate from **(R5)** that

$$E^\pi\left[|c(X(n))|\right] < L\big(1 + E^\pi\left[|X(n)|^\delta\right]\big) < \infty \tag{4.16}$$

and therefore $J(\pi)$ is always well defined (and in fact finite by Theorem 5.1 below). A similar argument shows that under the conditions (4.13) and (4.14), the long-run average cost associated with d is also well defined and finite under any policy π in $\mathcal{P}$.

5. MOMENT ESTIMATES

5.1. The bounds

The proofs of Theorems 4.1 and 4.2 require that bounds on moments of the rvs $\{|\ X(n)\ |\ ,\ n = 0,1,\ldots\}$ be available which are uniform over the class of *all* non-idling policies π in $\mathcal{P}$. The derivation of such bounds is given below, and is based on the key observation that the *total* number of customers in the system at any given time n decreases by *at most one* unit in the next time slot $[n, n+1)$, and is therefore bounded above by the number of slots it takes for the queue sizes to empty for the first time after n. This simple fact can be used to advantage when combined to the detailed statistical information obtained by the authors in [19] on the time until the system empties, and leads to the following strong estimates.

Theorem 5.1. *Assume* **(R1)–(R5)** *with* $\rho < 1$. *There exists a single positive constant* K_γ *such that for every non-idling policy* π *in* $\mathcal{P}$, *the moment estimate*

$$\sup_n E^\pi[|\ X(n)\ |^{\gamma-1}] \leq K_\gamma < \infty \tag{5.1}$$

holds true.

Theorem 5.1, the proof of which is presented below, turns out to be a special case of an intermediate result of independent interest given in Theorem 5.4. Before discussing this more general result, it is convenient to notice the following simple and useful consequence of (5.1).

Corollary 5.2. *Assume* **(R1)–(R5)** *with* $\rho < 1$. *Whenever* $\gamma > 2$, *the rvs* $\{|X(n)|,\ n = 0,1,\ldots\}$ *are uniformly integrable under the probability measure* P^π *associated with any non-idling policy* π *in* $\mathcal{P}$.

5.2. Recurrence properties

To formalize the argument outlined earlier, it is necessary to study the recurrence structure of the process $\{X(n),\ n = 0,1,\ldots\}$ under any non-idling policy π in $\mathcal{P}$. To that end consider the rvs $\{\tau_k,\ k = 0,1,2,\ldots\}$ and $\{\sigma_k,\ k = 1,2,\ldots\}$ defined recursively by $\tau_0 = \sigma_1 := 0$, and

$$\tau_{k+1} := \inf\{n > \sigma_{k+1} : X(n) = 0\} \qquad k = 0,1,\ldots (5.2a)$$

and

$$\sigma_{k+1} := \inf\{n > \tau_k : X(n) \neq 0\} \qquad k = 1,2,\ldots (5.2b)$$

with the convention that $\tau_{k+1} = \infty$ (resp. $\sigma_{k+1} = \infty$) whenever the defining set is empty or when $\sigma_{k+1} = \infty$ (resp. $\tau_k = \infty$). Note that these definitions are different from those given in [19] (where $\nu(0)$ denotes the present rv τ_1). For $k = 2,3,\ldots$ the rv τ_k is the time epoch at which the system

empties itself for the $(k-1)^{rst}$ time after τ_1, so that σ_{k+1} is the time epoch when the system becomes again non empty for the first time after τ_k. Moreover, define the rvs $\{\theta_k,\ k=1,2,\ldots\}$ by

$$\theta_{k+1} = \tau_{k+1} - \tau_k \qquad k = 0,1,\ldots \quad (5.3)$$

(with the convention $\infty - \infty = 0$) so that $\theta_1 = \tau_1$. The following proposition summarizes results which were obtained in Sections 4–5 of [19].

Proposition 5.3. *Assume* **(R1)–(R4)** *with* $\rho < 1$. *Under any non-idling policy* π *in* $\mathcal{P}$, *the rvs* $\{\theta_k,\ k=1,2,\ldots\}$ *form a delayed renewal process whose statistics are independent of the policy* π, *with finite means given by*

$$E^\pi[\theta_k|X(0)=x] = \begin{cases} \frac{1}{1-\rho}\cdot x'(I-P)^{-1}\nu + \frac{1}{1-\rho}1[x=0] & \text{if } k=1 \\ \frac{1}{1-q(0)}\cdot\frac{1}{1-\rho} & \text{if } k=2,3,\ldots \end{cases} \quad (5.4)$$

for all x *in* $\mathbb{N}^K$. *Moreover, the rv* θ_2 *possesses finite moments of order* γ, *and for every integer* ℓ, $1 \le \ell \le \gamma$, *there exists a positive constant* C_ℓ *(independent of the policy* π*) such that*

$$E^\pi[\tau_1^\ell|X(0)=x] \le C_\ell(1+|x|^\ell)\ , \quad x \in \mathbb{N}^K\ . \quad (5.5)$$

In view of this result, it is natural to introduce $\mathcal{E}_x$ as the expectation operator with respect to the distribution of τ_1 given that $X(0)=x$ and that *any* non-idling policy is used. Finally, for reference, denote by $G(\cdot)$ the distribution of the rv $\theta_1(=\tau_1)$ and by $F(\cdot)$ the common distribution of the i.i.d. rvs $\{\theta_k,\ k=2,3,\ldots\}$. By definition, the distributions $G(\cdot)$ and $F(\cdot)$ do not coincide.

5.3. A renewal estimate

The (continuous-time) counting process $\{N(t),\ t \ge 0\}$ naturally associated with the sequence $\{\tau_n,\ n=0,1,\ldots\}$ is defined by

$$N(t) := \max\{k \ge 0 : \tau_k \le t\}, \quad t \ge 0 \quad (5.6)$$

with the ready interpretation that $N(t)$ represents the number of times the queue has returned to the empty state by time t. With this notation, the observation made earlier translates into

$$|\ X(n)\ | \le \tau_{N(n)+1} - n. \qquad n = 0,1,\ldots \quad (5.7)$$

Now, for any *monotone non-decreasing* mapping $r : \mathbb{R}_+ \to \mathbb{R}_+$, set

$$R_G(t) := E^{\pi}_{q_\Xi}[r(\tau_{N(t)+1} - t)], \quad t \geq 0. \tag{5.8}$$

The subscripts G and Ξ in (5.8) emphasize the fact that the system is started with an initial queue size Ξ distributed according to the distribution $q_\Xi(\cdot)$. Since the sequence $\{\theta_k,\ k = 2, 3, \ldots\}$ is a *non*-delayed renewal sequence, it is appropriate to define

$$R_F(t) := E^{\pi}[r(\tau_{N(t+\tau_1)+1} - (t + \tau_1))], \quad t \geq 0 \tag{5.9}$$

as this corresponds to a non-delayed renewal process with $G = F$.

The first part of this section is devoted to deriving a bound on the expected values $\{R_G(t),\ t \geq 0\}$ for any non-idling policy π, with a view towards generating (via (5.7)) a bound for the sequence of expected values $\{E^{\pi}[r(\mid X(n) \mid)],\ n = 0, 1, \ldots\}$.

Theorem 5.4. *Assume* **(R1)–(R4)** *with* $\rho < 1$ *and let* π *be an arbitrary non-idling policy in* $\mathcal{P}$. *Under the finite moment assumptions*

$$m_G(r) := \int_0^\infty r(\theta) dG(\theta) < \infty \quad \text{and} \quad m_F(r) := \int_0^\infty r(\theta) dF(\theta) < \infty, \tag{5.10}$$

the condition

$$K_F(r) := \int_0^\infty \int_0^\theta r(\theta - t) dt dF(\theta) = \int_0^\infty \int_t^\infty r(\theta - t) dF(\theta) dt < \infty \tag{5.11}$$

implies

$$\sup_{t \geq 0} R_G(t) = \sup_{t \geq 0} E^{\pi}_{q_\Xi}[r(\tau_{N(t)+1} - t)] < \infty. \tag{5.12}$$

Proof: Let r_G and r_F be the mappings $\mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$r_G(t) := \int_t^\infty r(\theta - t) dG(\theta) \quad \text{and} \quad r_F(t) := \int_t^\infty r(\theta - t) dF(\theta), \quad t \geq 0. \tag{5.13}$$

The finiteness conditions (5.10) translate into $r_G(0) = m_G(r) < \infty$ and $r_F(0) = m_F(r) < \infty$. Since r takes positive values and is monotone non-decreasing, the indefinite integrals entering the definition (5.13) are well defined, and satisfy the inequalities

$$0 \leq \int_t^\infty r(\theta - t) dG(\theta) \leq \int_t^\infty r(\theta - s) dG(\theta) \leq \int_s^\infty r(\theta - s) dG(\theta) \tag{5.14}$$

whenever $0 \leq s \leq t$. As a result, the mapping r_G is well defined and monotone non-increasing. Similar comments hold for r_F.

A standard renewal argument [12, pp. 183] applied to the process $\{r(\tau_{N(t)+1} - t),\ t \geq 0\}$ shows that

$$R_G(t) = \int_0^t R_F(t-\theta)dG(\theta) + \int_t^\infty r(\theta - t)dG(\theta), \quad t \geq 0 \tag{5.15}$$

whence

$$\begin{aligned} R_G(t) &\leq \int_0^t R_F(t-\theta)dG(\theta) + \int_0^\infty r(\theta)dG(\theta) \\ &\leq \sup_{0 \leq s \leq t} R_F(s) + m_G(r), \quad t \geq 0 \end{aligned} \tag{5.16}$$

by the remarks made earlier. This clearly shows that under (5.10), the result (5.12) will hold if the bound

$$\sup_{t \geq 0} R_F(t) < \infty \tag{5.17}$$

can be established.

When $G = F$, the renewal equation (5.15) specializes to

$$R_F(t) = r_F(t) + \int_0^t R_F(t-\theta)dF(\theta), \quad t \geq 0. \tag{5.18}$$

Since the mapping r_F is monotone non-increasing and takes non-negative values, it is therefore integrable as a result of (5.11), whence *directly Riemann integrable* [12, pp. 190-191]. The fact that $0 \leq r_F(t) \leq m_F(r)$ for all $t \geq 0$ implies that R_F is bounded on finite intervals [12, Thm. 4.2, p. 184]. Finally, note that the distribution $F(\cdot)$ has support on $\mathbb{N}$ and is therefore arithmetic, say with span d. All requisite conditions are now in place to apply the Basic Renewal Theorem [12, Thm. 5.5.1, p. 191] to the renewal equation (5.18) to obtain

$$\lim_n R_F(c + nd) = \frac{d}{m_F} \sum_{n=0}^{\infty} r_F(c + nd)\ , \quad c \geq 0 \tag{5.19}$$

where m_F is the first moment of F (which is finite by Proposition 5.3). Since the mapping r_F is non-increasing, it readily follows from (5.19) that for all $c \geq 0$,

$$\lim_n R_F(c + nd) \leq \frac{1}{m_F} \left\{ dr_F(0) + \sum_{n=1}^{\infty} dr_F(nd) \right\}$$

$$\leq \frac{1}{m_F}\left\{dr_F(0) + \int_0^\infty r_F(t)dt\right\}$$

$$= \frac{1}{m_F}\{dm_F(r) + K_F(r)\} < \infty \tag{5.20}$$

where the finiteness of the last bound results from (5.10)–(5.11). In particular,

$$\lim_n R_F(nd+\ell) \leq \frac{1}{m_F}\{dm_F(r) + K_F(r)\} < \infty\ , \qquad \ell = 1,2,\ldots,d \tag{5.21}$$

and therefore

$$\sup_n R_F(n) < \infty. \tag{5.22}$$

Since $N(t)$ in constant on $[n, n+1)$, direct inspection of (5.9) shows that $R_F(t) \leq R_F(n)$ whenever $n \leq t < n+1$ owing to the monotonicity of r, whence $\sup_{t\geq 0} R_F(t) = \sup_n R_F(n)$ and (5.17) is now immediate from (5.22) ■

Proof of Theorem 5.1: Start with the mapping r given by $r(x) = x^{\gamma-1}$ for all $x \geq 0$, and observe that

$$K_F(r) = \int_0^\infty \int_0^\theta r(\theta - t)dtdF(\theta) = \frac{1}{\gamma}\int_0^\infty \theta^\gamma dF(\theta). \tag{5.23}$$

Under **(R4)**, Proposition 5.3 and (5.23) imply the conditions (5.10)–(5.11), and a straightforward application of Theorem 5.4 yields (5.1). ■

5.4. Extensions

Bounds of the form (5.1) are related to the stability of the system under the class of policies of interest, and are typically established through system-specific arguments. In fact, as will become apparent from the discussion in Sections 6–7, (5.1) need only hold for a small number of policies. The methods of the present section can be extended to the general model in the following way. Let Π denote a class of policies under which (5.1) is sought to hold, and consider the conditions **(G1)**–**(G3)** below, where

(G1): There exists a positive constant c_s such that for every policy π in Π,

$$P^\pi[\mid X(n+1)\mid \leq \mid X(n)\mid + c_s] = 0\ . \qquad n = 1,2,\ldots \tag{5.24}$$

Define the rvs $\{\tau_k,\ \sigma_k,\ \theta_k,\ k = 1,2,\ldots\}$ and $\{N(t),\ t \geq 0\}$ as in Section 5.2, and assume all these rvs to be finite a.s. under each policy π in Π. Under **(G1)**, it follows that for every π in Π,

$$\mid X(n)\mid \leq c_s(\tau_{N(n)+1} - n) \qquad P^\pi - \text{a.s.} \qquad n = 1,2,\ldots \tag{5.25}$$

Note that in this general context, $\{N(t),\ t \geq 0\}$ may not be a renewal process, since π need not be a stationary policy. Let ℓ be a positive integer, and introduce conditions **(G2)**–**(G3)** as

(G2): There exists a stationary policy π^* in Π such that for every policy π in Π either

$$E^{\pi}\left[\mid X(n)\mid^{\ell-1}\mid X(0)=x\right] \leq E^{\pi^*}\left[\mid X(n)\mid^{\ell-1}\mid X(0)=x\right],$$

or

$$E^{\pi}\left[\mid \tau_{N(n)+1}-n\mid^{\ell-1}\mid X(0)=x\right] \leq E^{\pi^*}\left[\mid \tau_{N(n)+1}-n\mid^{\ell-1}\mid X(0)=x\right]$$

$$x \in \mathbb{N}^K,\ n=1,2,\ldots$$

and

(G3): For every initial condition x in $\mathbb{N}^K$, there exists a positive constant $C_\ell(x)$ such that

$$E^{\pi^*}\left[\tau_1^{\ell} \mid X(0)=x\right] \leq C_\ell(x).$$

Note that **(G1)**–**(G2)** together imply via (5.25) that for every policy π in Π,

$$E^{\pi}\left[\mid X(n)\mid^{\ell-1}\right] \leq {c_s}^{\ell-1} E^{\pi^*}\left[\mid \tau_{N(n)+1}-n\mid^{\ell-1}\right]. \qquad n=1,2,\ldots \quad (5.26)$$

Assumptions **(G1)**–**(G3)** are natural in queueing systems when conservation laws are available. Note that $\mid\cdot\mid$ could denote *any* norm on $\mathbb{R}^K$, a fact which could be use to advantage when employing conservation laws. The generalization of Theorem 5.1 can now be stated.

Theorem 5.1bis. *Consider the general model. Under* **(G1)**–**(G3)**, *there exists a single constant* K_γ *(with* $\gamma = l$*) such that (5.1) holds for every policy* π *in* Π.

Proof: Under the Markov stationary policy π^*, the process $\{N(t),\ t \geq 0\}$ is a delayed renewal process. The proof of (5.1) with $\pi = \pi^*$ is identical to the proof of Theorem 5.1 and the desired result now follows from **(G2)**. ■

6. ON THE POISSON EQUATION

6.1. The Poisson equation

Fix η in the unit interval $[0,1]$ and denote by P^η (resp. E^η) the probability measure (resp. expectation operator) induced by the policy f^η. Moreover, let P_x^η (resp. E_x^η) denote the (conditional) probability measure (resp. expectation operator) induced by the policy f^η given that $X(0)=x$, with x ranging in $\mathbb{N}^K$.

Recall that under P^η, the rvs $\{X(n),\ n = 0, 1, \ldots\}$ form a time-homogeneous Markov chain over $\mathbb{N}^K$, and let $(P^\eta(x,y))$ denote the corresponding one-step transition probabilities. It is plain from (4.3) that

$$P^\eta(x,y) = \eta P^1(x,y) + (1-\eta)P^0(x,y)\ , \quad x, y \in \mathbb{N}^K \tag{6.1}$$

where $(P^1(x,y))$ (resp. $(P^0(x,y))$) are the one-step transition probabilities under $\overline{g}$ (resp. $\underline{g}$).

The mapping $h : \mathbb{N}^K \to \mathbb{R}$ and the scalar J solve the *Poisson equation* (associated with the policy f^η) with forcing function $c : \mathbb{N}^K \to \mathbb{R}$ if

$$h(x) + J = c(x) + \sum\nolimits_y P^\eta(x,y)h(y), \quad x \in \mathbb{N}^K\,. \tag{6.2}$$

Clearly the solution pair (h, J) to (6.2) depends on η, and it is the purpose of this section to establish its regularity properties with respect to η. This information is essential both for establishing the validity of the Certainty Equivalence Principle [20, 26] and for studying the convergence of the stochastic approximation algorithm (4.6) by the method of Metivier and Priouret [21]. From now on, this dependence of J and $h(x)$ on the bias η is denoted simply by $J(\eta)$ and $h(\eta, x)$ for all x in $\mathbb{N}^K$.

Define the *first return time* to state $x = 0$ as the $\mathcal{F}_n$-stopping time T given by

$$T := \inf\{n > 0 : X(n) = 0\} \tag{6.3}$$

so that $T = \tau_1$ in the notation of Section 5. Set

$$T_\ell(x) := \mathcal{E}_x[T^\ell] = E_x^\eta[T^\ell], \quad x \in \mathbb{N}^K \qquad\qquad \ell = 1, \ldots, \gamma \tag{6.4}$$

where the notation that follows Proposition 5.3 has been used. For easy reference recall the estimate (5.5), valid under (**R1**)–(**R4**), i.e., for each $\ell = 1, \ldots, \gamma$, there exists a positive constant C_ℓ so that

$$T_\ell(x) \le C_\ell(1 + |\ x\ |^\ell), \quad x \in \mathbb{N}^K\,. \tag{6.5}$$

As pointed out already in Section 5, during each slot, at most one customer may leave the system, so that for each $t = 0, 1, \ldots$, $|X(t)|$ is necessarily no larger than the forward recurrence time (expressed in slots) to the empty state, and in particular $|X(0)| \le T$. Since the mapping $x \to \tilde{c}(|x|)$ defined in (**R5**) is a non-decreasing function of $|x|$, it is plain from (6.5) that whenever $\delta + 1 \le \gamma$, the bounds

$$E_x^\eta\left[\sum_{t=0}^{T-1} |c(X(t))|\right] \le E_x^\eta\left[\sum_{t=0}^{T-1} \tilde{c}(|X(t)|)\right] \le \mathcal{E}_x\left[T\tilde{c}(T)\right] < \infty\ , \quad x \in \mathbb{N}^K \tag{6.6}$$

hold, and the definition

$$C(\eta, x) := E_x^{\eta}\Big[\sum_{i=0}^{T-1} c(X(i))\Big], \quad x \in \mathbb{N}^K \tag{6.7}$$

is thus well posed. An explicit expression for a solution to the Poisson equation is available and is now given [9, 27].

Theorem 6.1. *Assume conditions* **(R1)–(R6)** *to hold with* $\rho < 1$ *and* $\delta + 1 \leq \gamma$. *A solution pair* $(h(\eta), J(\eta))$ *to the Poisson equation (6.2) with* $h(\eta, 0) = 0$ *is given by*

$$J(\eta) = \frac{C(\eta, 0)}{T_1(0)} \quad \text{and} \quad h(\eta, x) = C(\eta, x) - J(\eta) T_1(x)\ , \quad x \in \mathbb{N}^K \tag{6.8a}$$

and the equality

$$J(f^{\eta}) = \lim_n\ E^{\eta}\left[\frac{1}{n+1}\sum_{t=0}^{n} c(X(t))\right] = J(\eta) \tag{6.8b}$$

holds true.

In view of (6.8b) and of the ergodic properties of this system under f^{η}, it is plain that $J(\eta)$ is also the expectation of $c(X)$ under the invariant measure corresponding to the policy f^{η}.

6.2. Lipschitz continuity

The representation (6.8) will be put to use in studying the regularity of the solution pair to the Poisson equation (6.2). To simplify the presentation of the main result of this section, set

$$K(x) := \mathcal{E}_x[T^2 \tilde{c}(T)], \quad x \in \mathbb{N}^K\,. \tag{6.9}$$

Theorem 6.2.: *Assume* **(R1)–(R6)** *with* $\rho < 1$ *and* $\delta + 2 \leq \gamma$. *Then for all* x *in* $\mathbb{N}^K$, $K(x) < \infty$ *and the function* $\eta \to C(\eta, x)$ *is Lipschitz continuous on* $[0, 1]$ *with Lipschitz constant* $4K(x)$, *i.e.,*

$$|\, C(\eta, x) - C(\eta', x)\,| \leq 4K(x)\,|\,\eta - \eta'\,|\ , \quad \eta, \eta' \in [0, 1]\,. \tag{6.10}$$

Proof: Fix x in $\mathbb{N}^K$. That $K(x)$ and $\mathcal{E}_x[T\tilde{c}(T)]$ are both finite is plain from (6.5) under the assumption $\delta + 2 \leq \gamma$. The result (6.10) is established below for c non-negative in the form

$$|\, C(\eta, x) - C(\eta', x)\,| \leq 2K(x)\,|\,\eta - \eta'\,|\ , \quad \eta, \eta' \in [0, 1] \tag{6.11}$$

so that the result for a general c is now immediate. Therefore, it suffices to assume c to be non-negative in the remainder of this proof. The arguments proceed in three steps.

Step 1: Fix η in $[0,1]$. Notice that for every $\mathbb{N}^K$-valued sequence $\{x(i),\ i=0,1,\ldots\}$ with $x(0)=x$, the relations

$$P_x^\eta[X(i)=x(i), 1\le i\le m] = \prod_{i=0}^{m-1} P^\eta(x(i),x(i+1)), \qquad m=1,2,\ldots \tag{6.12}$$

hold as a result of the Markov property of the chain $\{X(n),\ n=0,1,\ldots\}$ under P^η. The product form of (6.12) and the linear structure of (6.1) now imply that for each $m=1,2,\ldots$, the mapping $\eta \to P_x^\eta[X(i)=x(i), 1\le i\le m]$ is a polynomial of degree m in η over $[0,1]$ and has derivatives of all orders.

Set $A=[X(i)=x(i), 1\le i\le m]$ in (6.12) and observe that

$$\begin{aligned}\frac{d}{d\eta}P_x^\eta[A] &= \frac{d}{d\eta}\prod_{i=0}^{m-1} P^\eta(x(i),x(i+1)) \\ &= \sum_{t=0}^{m-1}[P^1(x(t),x(t+1)) - P^0(x(t),x(t+1))]\prod_{i=0,i\ne t}^{m-1} P^\eta(x(i),x(i+1)).\end{aligned} \tag{6.13}$$

This suggests defining for every $t=0,1,\ldots$, the policy 0_t (resp. 1_t) as the Markov policy that operates according to f^0 (resp. f^1) at time t, and according to f^η otherwise. With this notation, (6.13) now takes the form

$$\frac{d}{d\eta}P_x^\eta[X(i)=x(i), 1\le i\le m] = \sum_{t=0}^{m-1} P_x^{1_t}[A] - P_x^{0_t}[A]. \tag{6.14}$$

The definition of the policies 0_t and 1_t implies that $P_x^{1_t}[A] = P_x^{0_t}[A]$ whenever $m\le t$, so that (6.14) can also be rewritten as

$$\frac{d}{d\eta}P_x^\eta[X(i)=x(i), 1\le i\le m] = \sum_{t=0}^{n} P_x^{1_t}[A] - P_x^{0_t}[A], \quad m\le n. \tag{6.15}$$

Step 2: To proceed, define

$$C_m(\eta,x) := E_x^\eta\left[1[T\le m]\sum_{t=0}^{T\wedge m-1} c(X(t))\right] = E_x^\eta\left[\sum_{k=1}^{m} 1[T=k]\sum_{t=0}^{k-1} c(X(t))\right] \tag{6.16}$$

for all $m=1,2,\ldots$. The definition of T implies that

$$[T=k] = [X(t)\ne 0, 0<t<k, X(k)=0], \qquad k=1,2,\ldots \tag{6.17}$$

so that

$$C_m(\eta, x) = \sum_{k=1}^{m} \sum_{(x(1),\ldots,x(k))\in\mathcal{Z}_k} P_x^{\eta}[X(i) = x(i), 1 \le i \le k] \sum_{t=0}^{k-1} c(x(t)) \tag{6.18}$$

where the second sum is taken over the set $\mathcal{Z}_k$ given by

$$\mathcal{Z}_k := \{(x(1), x(2), \ldots, x(k)) \in (\mathbb{N}^K)^k : x(i) \neq 0, 1 \le i < k \text{ and } x(k) = 0\}.$$

$$k = 1, 2, \ldots \tag{6.19}$$

By arguments made earlier, it is plain that on the event $[T = k]$, the bounds $|X(t)| \le k$, $0 \le t \le k$, must necessarily hold, and therefore (6.18) reduces to

$$C_m(\eta, x) = \sum_{k=1}^{m} \sum_{(x(1),\ldots,x(k))\in\mathcal{Z}_k'} P_x^{\eta}[X(i) = x(i), 1 \le i \le k] \sum_{t=0}^{k-1} c(x(t)) \tag{6.20}$$

where the *finite* set $\mathcal{Z}_k'$ is given by

$$\mathcal{Z}_k' := \{(x(1), x(2), \ldots, x(k)) \in \mathcal{Z}_k : |x(i)| \le k,\ 1 \le i \le k\}. \qquad k = 1, 2, \ldots \tag{6.21}$$

Hence, in view of remarks made earlier in the proof, the mapping $\eta \to C_m(\eta, x)$ is a polynomial of degree m in η since it is the sum of a finite number of polynomial functions, each one of degree no greater than m.

Since $C_m(\eta, x)$ is a polynomial in η for each $m = 1, 2, \ldots$, the derivative $\dot{C}_m(\eta, x)$ exists in the interval $[0, 1]$. To compute it, differentiate (6.20) and use (6.14)–(6.15) to conclude that

$$\dot{C}_m(\eta, x) = \sum_{t=0}^{m-1} E_x^{1_t} \left[\sum_{s=0}^{T\wedge m-1} 1[T \le m] c(X(s))\right] - E_x^{0_t} \left[\sum_{s=0}^{T\wedge m-1} [1[T \le m] c(X(s))]\right]. \tag{6.22}$$

The very same argument that lead from (6.14) to (6.15) now implies that whenever $0 \le k \le t$, the relation

$$E_x^{1_t} \left[1[T = k] \sum_{s=0}^{k-1} c(X(s))\right] = E_x^{0_t} \left[1[T = k] \sum_{s=0}^{k-1} c(X(s))\right] \tag{6.23}$$

holds. Therefore (6.22) can be rewritten (in the manner of (6.16)) as

$$\dot{C}_m(\eta, x) = \sum_{t=0}^{m-1} E_x^{1_t} \left[\sum_{s=0}^{T\wedge m-1} 1[t < T \le m] c(X(s))\right] - E_x^{0_t} \left[\sum_{s=0}^{T\wedge m-1} 1[t < T \le m] c(X(s))\right]. \tag{6.24}$$

On the other hand,

$$\left|\sum_{t=0}^{m-1} E_x^{1_t}\left[\sum_{s=0}^{T\wedge m-1} 1[t<T\le m]c(X(s))\right]\right| \le \sum_{t=0}^{m-1}\left|E_x^{1_t}\left[1[t<T]\sum_{s=0}^{T\wedge m-1} 1[T\le m]\tilde{c}(T)\right]\right|$$

$$\le \sum_{t=0}^{m-1} \mathcal{E}_x\left[1[t<T\le m]T\tilde{c}(T)\right]$$

$$\le \mathcal{E}_x\left[T^2\tilde{c}(T)\right] \qquad (6.25)$$

by elementary calculations. A similar bound holds for the terms corresponding to the policies 0_t in (6.24). It then follows from (6.24) and (6.25) that the derivative $\dot{C}_m(\eta,x)$ of $C_m(\eta,x)$ is bounded on $[0,1]$ by $2K(x)$, and this uniformly in m, i.e.,

$$\left|\dot{C}_m(\eta,x)\right| \le 2K(x) \qquad m=0,1,\ldots(6.26)$$

for all η in $[0,1]$

Step 3: The easy estimates

$$0 \le C(\eta,x) - C_m(\eta,x) = E_x^{\eta}\left[1[T>m]\sum_{t=0}^{T-1} c(X(t))\right] \le E_x^{\eta}\left[1[T>m]T\tilde{c}(T)\right]$$

$$m=0,1,\ldots(6.27)$$

imply via the Monotone Convergence Theorem that $\lim_m C_m(\eta,x) = C(\eta,x)$ *uniformly* in η since $\mathcal{E}_x[T\tilde{c}(T)] < \infty$. Consequently, with $0 \le \eta < \eta' \le 1$,

$$|C(\eta,x) - C(\eta',x)| = \lim_m |C_m(\eta,x) - C_m(\eta',x)|$$

$$= \lim_m \left|\int_{\eta}^{\eta'} \dot{C}_m(y,x)dy\right| \qquad (6.28)$$

$$\le 2K(x)|\eta-\eta'|$$

upon making use of (6.26), and this establishes (6.11). ■

Note that the estimate (6.27) shows that $C(\eta,x)$ is *continuous* under the weaker condition $\delta+1 \le \gamma$.

6.3. Corollaries

Theorem 6.2 has several useful consequences which are now given in the next few corollaries. The first such corollary is obtained by combining Theorems 6.1 and 6.2 in a straightforward manner; details are left to the interested reader.

Corollary 6.3. *Under the hypotheses of Theorem 6.2, the functions* $\eta \to J(\eta)$ *and* $\eta \to h(\eta, x)$, *with* x *ranging in* $\mathbb{N}^K$, *are Lipschitz continuous on* $[0,1]$, *i.e., for all* η *and* η' *in* $[0,1]$,

$$|J(\eta) - J(\eta')| \leq 4\frac{K(0)}{T_1(0)} \cdot |\eta - \eta'| \tag{6.29}$$

and

$$|h(\eta, x) - h(\eta', x)| \leq 4K_h(x) \cdot |\eta - \eta'| \ , \quad x \in \mathbb{N}^K \tag{6.30}$$

with

$$K_h(x) := K(x) + \frac{K(0)}{T_1(0)} \cdot T_1(x), \quad x \in \mathbb{N}^K \ . \tag{6.31}$$

The behavior of the Lipschitz constants $K(x)$ and $K_h(x)$, and of the solution $h(\eta, x)$ for $|x|$ large is needed in some of the arguments given in Section 7. The estimates on the Lipschitz constants are given first.

Corollary 6.4. *Assume* **(R1)–(R6)** *with* $\rho < 1$ *and* $\delta + 2 \leq \gamma$. *There exists a positive constant* C *such that*

$$|K(x)| \leq C\left(1 + |x|^{\delta+2}\right) \ , \quad x \in \mathbb{N}^K \tag{6.32a}$$

and

$$|K_h(x)| \leq C\left(1 + |x|^{\delta+2}\right) \ , \quad x \in \mathbb{N}^K \ . \tag{6.32b}$$

Proof: Fix $x \in \mathbb{N}^K$. Note from **(R6)** and (6.9) that

$$\begin{aligned} K(x) &\leq L\mathcal{E}_x\left[T^2(1 + T^\delta)\right] \\ &\leq 2L\mathcal{E}_x\left[T^{\delta+2}\right] \leq 2LC_{\delta+2}\left(1 + |x|^{\delta+2}\right) \end{aligned} \tag{6.33}$$

with the last inequality following from (6.5), so that (6.32a) holds wherever $C \geq 2LC_{\delta+2}$. The inequality (6.32b) is readily obtained from (6.31) upon making use of (6.5) and (6.32a). ■

The growth of solutions to the Poisson equation can now be described.

Corollary 6.5. *Assume* **(R1)–(R6)** *with* $\rho < 1$ *and* $\delta + 1 \leq \gamma$. *There exists a positive constant* B_h *such that*

$$|\, h(\eta, x)\,| \leq B_h\left(1 + |\, x\,|^{\delta+1}\right), \quad x \in \mathbb{N}^K \tag{6.34}$$

for every η *in* $[0,1]$.

Proof: By the remark following the proof of Theorem 6.2, the mapping $\eta \to C(\eta, 0)$ is continuous on $[0,1]$ and therefore bounded there. For each x in $\mathbb{N}^K$, straightforward arguments show that

$$\begin{aligned} | h(\eta, x) | &\leq E_x^\eta \left[\sum_{t=0}^{T-1} | c(X(t)) | \right] + \frac{T_1(x)}{T_1(0)} \cdot \sup_{0 \leq \eta \leq 1} |C(\eta, 0)| \\ &\leq \mathcal{E}_x [T \tilde{c}(T)] + B_1 T_1(x) \end{aligned} \tag{6.35}$$

with

$$B_1 := \frac{1}{T_1(0)} \cdot \sup_{0 \leq \eta \leq 1} |C(\eta, 0)|. \tag{6.36}$$

The passage from (6.35) to (6.34) is validated by the same arguments as the ones given in the proof of Corollary 6.4. ■

Finally, a bound on the moments of the rvs $\{h(\eta(n), X(n+1)),\ n = 0, 1, \ldots\}$ is obtained.

Corollary 6.6. *Assume* **(R1)–(R6)** *with* $\rho < 1$ *and* $r(\delta+1)+1 \leq \gamma$ *for some non-negative integer* r. *Then there exists a positive constant* H_r *such that the bound*

$$\sup_n E^\alpha [|h(\eta(n), X(n+1))|^r] \leq H_r \tag{6.37}$$

holds.

Proof: For every η in $[0,1]$, Corollary 6.5 immediately implies

$$| h(\eta, x) |^r \leq |2B_h|^r \left(1 + | x |^{r(\delta+1)}\right), \quad x \in \mathbb{N}^K \tag{6.38}$$

so that

$$E^\alpha [| h(\eta(n), X(n+1)) |^r] \leq |2B_h|^r \left(1 + E^\alpha [| X(n+1) |^{r(\delta+1)}]\right). \qquad n = 0, 1, \ldots \tag{6.39}$$

The conclusion (6.37) is now obtained from Theorem 5.1 upon selecting $H_r = |2B_h|^r (1 + K_\gamma)$ since $r(\delta+1) \leq \gamma - 1$. ■

6.4. The general model

In [27] the authors have developed a methodology for proving smoothness properties of solutions to the Poisson equation in a fairly general setting. This is done by invoking the recurrence properties of the underlying Markov chain in order to obtain continuity, Lipschitz continuity and differentiability properties. The ideas of the present paper are however amenable to generalization

as follows. Suppose (as in (4.3)) that at each step the stationary policy $\overline{g}$ is used with some probability η while the stationary policy $\underline{g}$ is used with probability $(1-\eta)$. Then, as in (6.1), the one-step transition probabilities take the form

$$p_{xy}(\eta) = \eta p_{xy}(0) + (1-\eta)p_{xy}(1), \quad x, y \in \mathbb{N}^K, \ \eta \in U \ , \tag{6.40}$$

where $p_{xy}(0)$ (resp. $p_{xy}(0)$), x, y in $\mathbb{N}^K$, are the one-step transition probabilities under $\overline{g}$ (resp. $\underline{g}$). Under these assumptions the original MDP collapses to the model where the action space is $U = [0,1]$, and the transitions are realized according to (6.40).

However, the structure (6.40) may arise through a mechanism different from the one outlined above. Given this structure, with some abuse of notation, let η also denote the stationary policy which uses action η at every stage. Fix η in $[0,1]$ and, as in (6.13)–(6.14), let 0_t (resp. 1_t) denote the policy which uses action 0 (resp. action 1) at time t, and otherwise uses action η. Condition **(G4)** then takes the form

(G4): The distribution of the rv T under the policies P^{1_t} and P^{0_t} is stochastically monotone in t, i.e., for all increasing functions $r : \mathbb{R}_+ \to \mathbb{R}$, the mappings $t \to E^{1_t}[r(T)]$ and $t \to E^{0_t}[r(T)]$ are monotone.

As the arguments in (6.41) below reveal, a conditions much weaker than **(G4)** will suffice. Let Π be the collection of policies $\{1_t,\ 0_t;\ t = 1, 2, \ldots; 0 \le \eta \le 1\}$.

Theorem 6.2bis. *Consider the general model and assume conditions* **(G1), (G2), (G4)** *and* **(G3)** *for $l = 1, 2, \ldots, \gamma$ to hold (with $\tau_1 := T$ as in Section 5.4). Then the conclusion of Theorem 6.2 holds.*

Proof: For the sake of simplicity, set $c_s = 1$ in **(G1)** as the extension to the case $c_s \neq 1$ is obvious. Under the conditions **(G1)**–**(G3)**, the proof is almost identical to that of Theorem 6.2, except that **(G1)**, **(G4)** and the monotonicity of $\tilde{c}$ are used in (6.25) to obtain

$$\begin{aligned}
\left| \sum_{t=0}^{m-1} E_x^{1_t} \left[\sum_{s=0}^{T \wedge m-1} 1[t < T \le m] c(X(s)) \right] \right| &\le \sum_{t=0}^{m-1} \left| E_x^{1_t} \left[1[t < T] \sum_{s=0}^{T \wedge m-1} 1[T \le m] \tilde{c}(T) \right] \right| \\
&\le \sum_{t=0}^{m-1} \left| E_x^{1_t} \left[1[t < T] T \tilde{c}(T) \right] \right| \\
&\le \sup_t E_x^{1_t} \left[T^2 \tilde{c}(T) \right] \le C_l(x) \qquad l = 2 + \delta
\end{aligned} \tag{6.41}$$

for all $x \in \mathbb{N}^K$. ■

Corollary 6.3 continues to hold as stated under the hypotheses of Theorem 6.2bis. Furthermore, it is easy to see that the growth estimates of Corollaries 6.4–6.5 continue to hold under the assumption that

$$C_l(x) \le \tilde{C}_l \cdot \left(1 + |\, x \,|^{c_m l}\right), \quad x \in \mathbb{N}^K \tag{6.42}$$

for some positive constants $\tilde{C}_l$ and c_m.

7. CONVERGENCE OF THE STOCHASTIC APPROXIMATIONS

7.1. The ODE method

This section is devoted to proving the convergence of the recursive scheme (4.5)–(4.6) when the policy α is in use. Recall that the mapping $\eta \to J(f^\eta)$ is monotone *increasing.*

The following additional assumption **(R7)** is imposed in order to carry out the analysis.

(R7): The equation

$$J(f^\eta) = V, \qquad 0 \le \eta \le 1 \tag{7.1}$$

has a *unique* solution η^*.

Note that the continuity of $\eta \to J(f^\eta)$ now implies

$$\left[J(f^\eta) - V\right](\eta - \eta^*) > 0, \quad \eta \ne \eta^* \in [0,1]. \tag{7.2}$$

The uniqueness of the solution to (7.1) is tantamount to *local strict* monotonicity and in practice, is often verified by establishing some stronger monotonicity property on $\eta \to J(f^\eta)$ such as **(R7b)** below.

(R7b): The mapping $[0,1] \to \mathbb{R} : \eta \to J(f^\eta)$ is *strictly monotone increasing.*

In Section 8, condition **(R7b)** is shown to hold for a steering problem which arises from a constrained optimization problem.

The proof of Theorem 4.1 given below uses a version of the ODE method which was proposed by Metivier and Priouret in [21]. The arguments combine the deterministic lemma of Kushner and Clark [16] with a probabilistic result based on properties of the Poisson equation (6.2). This key result is given the next proposition, the proof of which is delayed until the second part of the section. To state the result, consider the rvs $\{Y(n),\ n = 0,1,\ldots\}$ given by

$$Y(n) := J(f^{\eta(n)}) - c(X(n+1)) \qquad n = 0,1,\ldots \tag{7.3}$$

and pose

$$m(n,t) := \max\{k > n : \sum_{i=n}^{k-1} a_i \leq t\} \ , \quad t > 0 \ . \qquad n = 0,1,\ldots (7.4)$$

Theorem 7.1 *Assume* **(R1)–(R6)** *with* $\rho < 1$ *and* $2\delta + 3 \leq \gamma$. *For each* $t > 0$, *the convergence*

$$\lim_n \left(\sup_{n \leq k \leq m(n,t)} | \sum_{i=n}^{k} a_i Y(i) | \right) = 0 \qquad P^\alpha - a.s. \tag{7.5}$$

takes place.

Proof of Theorem 4.1. As shown in [16, 21], the convergence (7.5) underlines the P^α-a.s. convergence of the estimates $\{\eta(n),\ n = 0,1,\ldots\}$ to η^*. The reader is invited to consult these references for a complete exposition of the arguments which are now briefly summarized: Interpolate the estimate sequence $\{\eta(n),\ n = 0,1,\ldots\}$ by a piecewise linear function $\eta^{(0)} : [0,\infty) \to \mathbb{R}$ such that $\eta^{(0)}(t_n) = \eta(n)$ at time $t_n = \sum_{i=0}^{n-1} a_i$ for all $n = 0,1,\ldots$ (with $t_0 = 0$). Moreover, define a sequence of left shifts $\{\eta^{(n)}(\cdot),\ n = 0,1,\ldots\}$, i.e., $\eta^{(n)}(t) = \eta^0(t - t_n)$ for all $t \geq 0$, in order to bring the "asymptotic part" of $\{\eta(n),\ n = 0,1,\ldots\}$ back to a neighborhood of the time origin.

Now observe that the recursion (4.6) can be written in the form

$$\eta(n+1) = \Big[\eta(n) + a_{n+1}\big[(V - J(f^{\eta(n)})) + Y(n)\big]\Big]_0^1 \qquad n = 0,1,\ldots (7.6)$$

and that from any convergent subsequence $\{\eta^{(m)}(\cdot),\ m = 0,1,\ldots\}$ a further convergent subsequence $\{\eta^{(m_p)}(\cdot),\ p = 0,1,\ldots\}$ can then be extracted by standard boundedness and equicontinuity arguments. It is then easy to see from Theorem 7.1 that the limit $\eta^*(\cdot)$ along this subsequence, and for that matter the limit of *any* convergent subsequent, satisfies the ODE

$$\dot{\eta}^*(t) = V - J(f^{\eta^*(t)}), \quad t \geq 0 \ , \quad \eta^*(0) \in [0,1] \ . \tag{7.7}$$

Owing to (7.2), this ODE is *asymptotically stable* with a *unique* stable point η^* in $[0,1]$. A simple shifting argument now implies $\eta^*(t) = \eta^*$ for all $t \geq 0$ and this completes the proof. These arguments are standard and are therefore omitted here in the interest of brevity. ■

The remainder of this section is devoted to a proof of (7.5).

7.2. A proof of Theorem 7.1

The Poisson equation (6.2) implies the relations

$$E^{\eta}[h(\eta, X(n+1)) \mid \mathcal{F}_n] = h(\eta, X(n)) + J(\eta) - c(X(n)) \qquad n = 0, 1, \ldots (7.8)$$

for all $0 \leq \eta \leq 1$. It then follows from (6.8b) and (7.3) that

$$\begin{aligned} -Y(n) &= c(X(n+1)) - J(\eta(n)) \\ &= h(\eta(n), X(n+1)) - E^{\eta(n)}[h(\eta(n), X(n+2)) \mid \mathcal{F}_{n+1}] \\ &= Z_n^{(1)} + Z_n^{(2)} + Z_n^{(3)} \qquad n = 0, 1, \ldots (7.9) \end{aligned}$$

with

$$Z_n^{(1)} := h(\eta(n), X(n+1)) - E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n], \qquad (7.10a)$$

$$Z_n^{(2)} := E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n] - E^{\eta(n+1)}[h(\eta(n+1), X(n+2)) \mid \mathcal{F}_{n+1}] \qquad (7.10b)$$

and

$$Z_n^{(3)} := E^{\eta(n+1)}[h(\eta(n+1), X(n+2)) \mid \mathcal{F}_{n+1}] - E^{\eta(n)}[h(\eta(n), X(n+2)) \mid \mathcal{F}_{n+1}] \qquad (7.10c)$$

for all $n = 0, 1, \ldots$. It now suffices to show that

$$\lim_n \left(\sup_{n \leq \ell \leq m(n,t)} \mid \sum_{i=n}^{\ell} a_i Z_i^{(k)} \mid \right) = 0 \qquad P^{\alpha} - a.s. \qquad (7.11)$$

for all $t > 0$ and all $k = 1, 2, 3$.

This will be done in three steps. To facilitate the presentation, define the rvs $\{S_n^{(k)},\ n = 0, 1, \ldots\}$ by

$$S_n^{(k)} := \sum_{i=0}^{n-1} a_i Z_i^{(k)} \qquad n = 1, 2, \ldots (7.12)$$

for $k = 1, 2, 3$, with $S_0^{(1)} = S_0^{(2)} = S_0^{(3)} = 0$.

Step 1: The rvs $\{Z_n^{(1)},\ n = 0, 1, \ldots\}$ form a $(P^{\alpha}, \mathcal{F}_n)$ martingale-difference, whence $\{S_n^{(1)},\ n = 0, 1, \ldots\}$ is a zero-mean $(P^{\alpha}, \mathcal{F}_n)$-martingale. Routine calculations show that

$$\sup_n E^{\alpha}[\mid S_n^{(1)} \mid^2] = \sup_n E^{\alpha}\left[\sum_{i=0}^{n-1} a_i^2 \mid Z_i^{(1)} \mid^2\right] \qquad (7.13)$$

$$\leq \sup_n E^{\alpha}\left[|h(\eta(n), X(n+1))|^2\right] \cdot 4 \sum_{i=0}^{\infty} a_i^2 \qquad (7.14)$$

$$\leq 4H_2 \cdot \sum_{i=0}^{\infty} a_i^2. \qquad (7.15)$$

The passage from (7.14) to (7.15) uses the estimate (5.37) given in Corollary 6.6 (with $r = 2$ since $2\delta + 2 \leq \gamma - 1$). It is plain from (4.7) that the left handside of (7.13) is finite, and the $(P^\alpha, \mathcal{F}_n)$-martingale $\{S_n^{(1)},\ n = 0, 1, \ldots\}$ is thus uniformly integrable under P^α. By the Martingale Convergence Theorem, the rvs $\{S_n^{(1)},\ n = 0, 1, \ldots\}$ converge a.s. under P^α (to an a.s. finite limit), in which case they form a Cauchy sequence P^α-a.s. and (7.11) follows for $k = 1$.

Step 2: For $k = 2$, note first the relations

$$\begin{aligned} S_{\ell+1}^{(2)} - S_n^{(2)} &= \sum_{i=n}^{\ell} a_i Z_i^{(2)} \\ &= -\sum_{i=n}^{\ell} (a_{i-1} - a_i) E^{\eta(i)}[h(\eta(i), X(i+1)) \mid \mathcal{F}_i] \qquad (7.16) \\ &\quad + a_{n-1} E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n] - a_\ell E^{\eta(\ell+1)}[h(\eta(\ell+1), X(\ell+2)) \mid \mathcal{F}_{\ell+1}] \end{aligned}$$

valid for all $0 \leq n < \ell$. Define the rvs $\{K_n,\ n = 0, 1, \ldots\}$ by

$$K_n := E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n] \qquad n = 0, 1, \ldots (7.17)$$

and set

$$B_r = \sup_n E^\alpha [\mid K_n \mid^r]. \qquad r = 1, 2, \ldots (7.18)$$

It is clear from (6.37) (with $r = 1, 2$) and Jensen's inequality that $B_1 \leq H_1 < \infty$ and $B_2 \leq H_2 < \infty$.

With this notation, (7.16) can be rewritten as

$$\mid S_{\ell+1}^{(2)} - S_n^{(2)} \mid \leq a_{n-1}|K_n| + \sum_{i=n}^{\ell} (a_{i-1} - a_i)|K_i| + a_\ell |K_{\ell+1}| \qquad (7.19)$$

for all $0 \leq n < \ell$ since $a_n \downarrow 0$. If the rvs $\{S_n,\ n = 1, 2 \ldots\}$ and $\{R_n,\ n = 0, 1, \ldots\}$ are now defined by

$$S_n = \sum_{i=1}^{n} (a_{i-1} - a_i)|K_i| \qquad n = 1, 2, \ldots (7.20)$$

and

$$R_n = \sum_{i=0}^{n} |a_i|^2 |K_{i+1}|^2, \qquad n = 0, 1, \ldots (7.21)$$

then (7.19) becomes

$$\mid S_{\ell+1}^{(2)} - S_n^{(2)} \mid \leq a_{n-1}|K_n| + \mid S_\ell - S_{n+1} \mid + a_\ell |K_{\ell+1}| \,, \quad 0 \leq n \leq \ell. \qquad (7.22)$$

The definition (7.20) implies

$$E^{\alpha}[S_n] \leq B_1 \sum_{i=0}^{n} (a_{i-1} - a_i) = B_1(a_0 - a_n) \leq B_1 a_0. \qquad n = 0, 1, \ldots (7.23)$$

Since $S_n \leq S_{n+1}$, the limit $S_\infty := \lim_n S_n$ exists and therefore $E^{\alpha}[S_\infty] \leq B_1 a_0$ by using the Monotone Convergence Theorem on (7.23). Consequently, $S_\infty < \infty$ P^{α}-a.s. and the rvs $\{S_n,\ n = 0, 1, \ldots\}$ form a Cauchy sequence P^{α}-a.s., i.e.,

$$\lim_n \sup_{\ell > n} \mid S_\ell - S_{n+1} \mid = 0 \quad P^{\alpha} - a.s. \qquad (7.24)$$

To handle the first and last terms of (7.22), observe that $R_n \leq R_{n+1}$, hence the limit $R_\infty := \lim_n R_n$ exists and satisfies

$$E^{\alpha}[R_\infty] \leq B_2 \sum_{i=0}^{\infty} a_i^2 < \infty \qquad (7.25)$$

by virtue of the Monotone Convergence Theorem. Consequently, $\lim_n R_n = R_\infty < \infty$ P^{α}-a.s., whence $\lim_n a_{n-1}|K_n| = 0$ P^{α}-a.s. or equivalently

$$\lim_n \sup_{\ell \geq n} a_{\ell-1}|K_\ell| = 0 \quad P^{\alpha} - a.s. \qquad (7.26)$$

by the Cauchy convergence criterion. Making use of (7.24) and (7.26) readily leads (via (7.22)) to the conclusion (7.11) for $k = 2$.

Step 3: For $k = 3$, observe that (7.8) and the estimates of Corollary 6.3 readily yield the estimates

$$\begin{aligned} & \mid E^{\eta}[h(\eta, X(n+1)) \mid \mathcal{F}_n] - E^{\tilde{\eta}}[h(\tilde{\eta}, X(n+1)) \mid \mathcal{F}_n] \mid \\ = & \mid h(\eta, X(n)) - h(\tilde{\eta}, X(n)) + J(\eta) - J(\tilde{\eta}) \mid \\ \leq & 4\tilde{K}(X(n)) \cdot \mid \eta - \tilde{\eta} \mid \qquad n = 0, 1, \ldots (7.27) \end{aligned}$$

for all η and $\tilde{\eta}$ in $[0, 1]$, where

$$\tilde{K}(x) := K(x) + 2\frac{K(0)}{T_1(0)} T_1(x), \quad x \in \mathbb{N}. \qquad (7.28)$$

The recursion (4.6) implies

$$\mid \eta(n+1) - \eta(n) \mid \leq a_{n+1} \mid V - c(X(n+1)) \mid \qquad n = 0, 1, \ldots (7.29)$$

and the inequality

$$| Z_n^{(3)} | \leq 4a_{n+1}Q(X(n+1)) \qquad n = 0, 1, \ldots (7.30)$$

is now obtained from (7.27), upon setting

$$Q(x) := \tilde{K}(x)\big(V + |c(x)|\big), \quad x \in \mathbb{N}^K . \tag{7.31}$$

Under **(R5)**, with the help of (6.5) and (6.32a), it is a simple exercise to check that

$$Q(x) \leq C\big(1 + |x|^{2\delta+2}\big), \quad x \in \mathbb{N}^K \tag{7.32}$$

for some positive constant C. Consequently,

$$E^\alpha\Big[\sum_{i=0}^{n} a_i |Z_i^{(3)}|\Big] \leq C \cdot \sum_{i=0}^{n} a_i^2 E^\alpha\Big[1 + |X(i+1)|^{2\delta+2}\Big] \tag{7.33}$$

$$\leq C\big(1 + K_\gamma\big) \cdot \sum_{i=0}^{\infty} a_i^2 \qquad n = 0, 1, \ldots (7.34)$$

where the passage from (7.33) to (7.34) is a simple consequence of (5.1) (since $2\delta + 2 \leq \gamma - 1$). Now, in exactly the same way as in Step 2 of the proof, this uniform bound (7.34) implies

$$\lim_n \sup_{\ell \geq n} \Big(\sum_{i=n}^{\ell} a_i |Z_i^{(3)}|\Big) = 0 \qquad P^\alpha - a.s. \tag{7.35}$$

and (7.11) obviously holds for $k = 3$. ■

7.3. The general model

The results of this section rely on boundedness and smoothness properties of solutions to the Poisson equation, but the structure of the proof is otherwise quite general. In fact, consider a set of stationary policies, parameterized by $\eta \in [0, 1]$ and let $(J(\eta),\ h(\eta, x))$ denote the solution to the Poisson equation under policy η. Such parameterization may arise as in Section 6.4, but for the purposes of the present section this is immaterial. Suppose that the properties **(i)–(iii)** below can be established, as was done for the queueing model under consideration, where

(i): For each η in $[0, 1]$, $J(\eta)$ equals the cost under η,

(ii): For each η in $[0, 1]$, $x \to h(\eta, x)$ is at most polynomial in x,

(iii): For each x in $\mathbb{N}^K$, $\eta \to h(\eta, x)$ and $\eta \to J(\eta)$ are Lipschitz in η, where the Lipschitz constant of $h(\eta, x)$ is at most polynomial in x.

Then the conclusions of Theorem 7.1 (and hence the conclusions of Theorem 4.1) hold with proofs almost unchanged, provided appropriate bounds on moments of the state process are available. Condition **(ii)** is obtained in Corollary 6.6 and its generalizations, and validates Steps 1–2 in Section 7.2 above. Conditions **(iii)** allows the argument in Step 3 to be carried through and the only changes required involve the constants and the exponents δ, γ and r.

8. CONVERGENCE OF THE ADAPTIVE POLICY AND APPLICATIONS

This final section contains a proof of Theorem 4.2, as well as the discussion of an application that arises in constrained optimization.

8.1. A proof of Theorem 4.2.

The proof follows from general results obtained by the authors on the Certainty Equivalence Principle when specialized to "simply randomized" policies [26]. First note that the (assumed) convergence $\lim_n \eta(n) = \eta^*$ in probability under P^α, when combined to Theorem 7.2 of [26], implies the key convergence condition (C) [Ibid., Section 4]. Consequently, the convergence (4.11)–(4.12) follows from Theorem 3.1bis in [Ibid.] provided the hypotheses of Theorems 4.2 and 6.1bis of [Ibid.] are satisfied. These hypotheses consist in the tightness of the rvs $\{X(t),\ t = 0, 1, \ldots\}$ under P^α and of bounds on the moments of the rvs $\{c(X(t)), h(\eta^*, X(t)),\ t = 0, 1, \ldots\}$ under various policies. It is easy to check that these conditions are all implied by the following condition:

There exist $\epsilon > 0$ and a positive constant C_ϵ such that for every non-idling policy π in $\mathcal{P}$, the bounds

$$\sup_t E^\pi\big[|X(t)|^{1+\epsilon}\big] \le C_\epsilon, \tag{8.1}$$

$$\sup_t E^\pi\big[|c(X(t))|^{1+\epsilon}\big] \le C_\epsilon \tag{8.2}$$

and

$$\sup_t E^\pi\big[|h(\eta^*, X(t))|^{1+\epsilon}\big] \le C_\epsilon \tag{8.3}$$

hold.

Observe that by virtue of Theorem 5.1, the bound (8.1) readily holds whenever $1 + \epsilon \le \gamma - 1$. By assumption, c is of polynomial growth with rate δ, so that (8.2) holds if $\delta(1 + \epsilon) \le \gamma - 1$ by the remark made earlier. To obtain the third bound (8.3), observe from (6.34) that for every $\epsilon > 0$,

$$|h(\eta^*, X(n))|^{1+\epsilon} \le |2B_h|^{1+\epsilon}\big(1 + |X(n)|^{(\delta+1)(\epsilon+1)}\big), \qquad n = 0, 1, \ldots \tag{8.4}$$

and (8.3) follows with $(1+\epsilon)(1+\delta) \leq \gamma - 1$ by again making use of Theorem 5.1. Consequently (8.1)–(8.3) will hold provided ϵ is chosen positive such that $1 + (1+\delta)(1+\epsilon) \leq \gamma$.

An identical analysis applies for the long-run average cost associated with d; details are left to the interested reader. ■

8.2 An application to constrained optimization

Consider the following situation discussed by Nain and Ross in [22]. Several types of traffic, say voice, video and data, compete for the use of a single resource (or server). The performance requirements for this system are defined by the minimization of a weighted average of the number of video and data packets subject to the constraint that the average number of voice packets waiting for service does not exceed V. This situation can be modelled by a system of K competing queues with $P = 0$. For a precise definition of the performance measures, set

$$c(x) := x_K \quad \text{and} \quad d(x) := \sum_{k=1}^{K-1} d_k x_k \quad x \in \mathbb{N}^K \tag{8.5}$$

where $d_1, \ldots, d_{K-1}$ are positive constants. Denote by $J_c(\pi)$ (resp. $J_d(\pi)$) the long-run average cost (4.1) associated with the cost c (resp. d) when using the policy π in $\mathcal{P}$. The constrained optimization problem (P_V) is then formulated as

$$(P_V): \qquad \text{Minimize } J_d(\cdot) \text{ over } \mathcal{P}_V \tag{8.6}$$

where $\mathcal{P}_V := \{\pi \in \mathcal{P} : J_c(\pi) \leq V\}$.

Assume the problem to be feasible and non-trivial, i.e., $\mathcal{P}_V$ is non-empty and the policies which minimize J_d are not in $\mathcal{P}_V$. In that case, Nain and Ross [22] showed that there exist two strict priority policies $\overline{g}$ and $\underline{g}$ and a bias η^* satisfying the equation

$$J_c(f^\eta) = V, \quad \eta \text{ in } [0,1] \tag{8.7}$$

such that f^{η^*} defined through (4.3) is optimal. While the policies $\overline{g}$ and $\underline{g}$ can be found explicitly, the determination of η^* is a difficult task since for $0 < \eta < 1$ the evaluation of $J_c(f^\eta)$ requires solving a Riemann-Hilbert problem. That this computational difficulty can be circumvented by making using of a stochastic approximation–based policy is the content of the following.

Theorem 8.1 *Assume* **(R1)**-**(R5)** *with* $\rho < 1$ *and* $\gamma \geq 5$. *The scheme (4.5)–(4.6) solves the constrained optimization problem* (P_V) *provided it is feasible.*

Proof: Nain and Ross [22, Thm. 3.1, pp. 885-886] showed that if the problem is feasible and non-trivial, then there exist Markov stationary policies $\overline{g}$ and $\underline{g}$ such that (8.7) has at least one solution. In fact, both policies are fixed priority policies with $\underline{g}$ giving highest priority to queue K, and $\overline{g}$ giving lowest priority to queue K, while the relative priorities of the other queues are otherwise identical. Moreover, the mapping $\eta \to J_d(f^\eta)$ is monotone non-decreasing. It is shown in Lemma 8.2 below that this mapping is in fact strictly monotone increasing. When $\gamma \geq 5$, the conditions of Theorems 4.1 and 4.2 are readily verified with $\delta = 1$. Hence, $\lim_n \eta(n) = \eta^*$ P^α-a.s. so that $J_c(\alpha) = J_c(f^{\eta^*}) = V$ and $J_d(\alpha) = J_d(f^{\eta^*})$, i.e., α is a policy in $\mathcal{P}_V$ and is thus also constrained optimal.

If the problem is trivial, i.e., $J_c(\overline{g}) \leq V$, then $\overline{g}$ solves (P_V). In that case, the same arguments imply that $\lim_n \eta(n) = 1$ $P^\alpha - a.s.$, and optimality follows. ■

In the case $K = 2$, the two policies $\overline{g}$ and $\underline{g}$ are necessarily the fixed priority rules for queue 1 and 2, respectively. In this case, the adaptive policy does not assume any prior information on the statistics of the system, provided **(R1)**-**(R5)** hold with $\gamma \geq 5$. In this case, the optimality was obtained by Shwartz and Makowski [25] under a slightly weaker assumption (namely $\gamma \geq 3$), but the convergence (4.10) was only in probability.

This section concludes with the following monotonicity result which was needed in the proof of Theorem 8.1.

Lemma 8.2. *Under* **(R1)**-**(R5)** *the mapping* $\eta \to J_c(f^\eta)$ *is strictly monotone increasing on* $[0, 1]$.

Proof: It is plain from (6.8) that proving the strict monotonicity of $\eta \to J_c(f^\eta)$ is equivalent to proving the same for $\eta \to C(\eta, 0)$. Fix η in $[0, 1]$ and recall the definition (4.3) of the policy f^η.

The representation (6.22) of the derivative of $C_m(\eta, 0)$ can be written in the form

$$\begin{aligned}\dot{C}_m(\eta,0) &= \sum_{t=0}^{m-1} E_0^{1_t}\left[\sum_{\ell=1}^{m} 1[T=\ell]\sum_{s=0}^{\ell-1} 1[T\leq m]X_K(s)\right] - E_0^{0_t}\left[\sum_{\ell=1}^{m} 1[T=\ell]\sum_{s=0}^{\ell-1} 1[T\leq m]X_K(s)\right] \\ &= \sum_{t=0}^{m-1}\sum_{\ell=t+1}^{m}\left[E_0^{1_t}\left[1[T=\ell]\sum_{s=0}^{\ell-1} X_K(s)\right] - E_0^{0_t}\left[1[T=\ell]\sum_{s=0}^{\ell-1} X_K(s)\right]\right] \qquad (8.8)\end{aligned}$$

where (6.23) was used. If it were possible to show bounds of the form

$$\Delta(\ell,t,s) := E_0^{1_t}\left[1[T=\ell]X_K(s)\right] - E_0^{0_t}\left[1[T=\ell]X_K(s)\right] \geq \epsilon(\ell,t,s) \qquad (8.9)$$

with $\epsilon(\ell, t, s) \geq 0$ for all $0 \leq s < \ell$ and $0 \leq t < \ell$, and $\epsilon(\ell, t, s) > 0$ for at least one such triple (ℓ, t, s), then necessarily for *some* m, $0 < \dot{C}_m(\eta, 0) \leq \dot{C}_{m+1}(\eta, 0) \leq \ldots$ and the strict monotonicity would follow from the second equality in (6.28).

Fix t and ℓ such that $0 \leq t < \ell$. It is easy to see that $\Delta(\ell, t, s) = 0$ whenever $0 \leq s \leq t < \ell$, so that only the case $0 < t < s$ has to be considered in order to prove (8.9). This is done by the following coupling arguments.

Let $\hat{P}$ be a probability measure on $(\Omega, \mathcal{F})$ under which **(P1)**–**(P4)** hold and $X(0) = 0$. Moreover, let $\{\beta(n),\ n = 0, 1, \ldots\}$ be a sequence of i.i.d. Bernoulli rvs with parameter η which is also independent of the rvs $\{A(n), B(n),\ n = 0, 1, \ldots\}$ under $\hat{P}$.

The key point of the proof is to construct on Ω a pair of processes $\{X^0(n),\ n = 0, 1, \ldots\}$ and $\{X^1(n),\ n = 0, 1, \ldots\}$ such that (i) $\{X^0(n),\ n = 0, 1, \ldots\}$ (resp. $\{X^1(n),\ n = 0, 1, \ldots\}$) under $\hat{P}$ is statistically indistinguishable from $\{X(n),\ n = 0, 1, \ldots\}$ under $P_0^{0_t}$ (resp. $P_0^{1_t}$), and (ii) a simple comparison leads to (8.9). To that end, for each $i = 0, 1$, define the process $\{X^i(n),\ n = 0, 1, \ldots\}$ by the recursion

$$X_k^i(n+1) = X_k^i(n) + A_k(n) - I[X_k^i(n) \neq 0] U_k^i(n) B_k^i(n)\ , \quad 1 \leq k \leq K \qquad n = 0, 1, \ldots \quad (8.10)$$

with $X^i(0) = 0$, where the sequences $\{U^i(n),\ n = 0, 1, \ldots\}$ and $\{B^i(n),\ n = 0, 1, \ldots\}$ still need to be specified.

For $i = 0, 1$, the control actions $\{U^i(n),\ n = 0, 1, \ldots\}$ are defined by

$$U^i(n) = \beta(n)\, \overline{g}(X^i(n)) + (1 - \beta(n))\, \underline{g}(X^i(n))\ , \quad n \neq t \qquad (8.11a)$$

$$U^0(t) = \underline{g}(X^0(t)) \qquad (8.11b)$$

$$U^1(t) = \overline{g}(X^1(t)) \qquad (8.11c)$$

so that the rvs $\{X^0(n),\ n = 0, 1, \ldots\}$ (resp. $\{X^1(n),\ n = 0, 1, \ldots\}$) are governed by the policy 0_t (resp. 1_t).

Only the service sequences $\{B^i(n),\ n = 0, 1, \ldots\}$, $i = 0, 1$, need to be specified. First, set $B^0(n) \equiv B(n)$ for all $n = 0, 1, \ldots$ and observe from the construction (8.10)–(8.11) that the distribution of $\{X^0(n),\ n = 0, 1, \ldots\}$ under $\hat{P}$ obviously coincides with the distribution of $\{X(n),\ n = 0, 1, \ldots\}$ under $P_0^{0_t}$. The construction of the process $\{B^1(n),\ n = 0, 1, \ldots\}$ is somewhat more involved, and is done below. In order to facilitate the coupling argument, the actual service duration of each customer will be defined in such a way so as to have identical length (for *each* ω in Ω) in both processes. To do this, the rvs $B^1(n)$ are defined in (8.12)–(8.14) so that the number of unsuccessful services experienced by each customer is identical in both systems. Set

$$B^1(n) := B(n) \qquad n = 0, 1, \ldots, t-1 \quad (8.12)$$

and observe from (8.10) that in order to determine the process $\{X^1(n),\ n = 0,1,\ldots\}$, it suffices to provide the values of $B_k^1(n)$ at times n such that $U^1(n) = e^k$, $1 \le k \le K$. For all $i = 0,1$ and $1 \le k \le K$, set

$$\tau_k^i(1) := \min\{n \ge t : U^i(n) = e^k\} \tag{8.13b}$$

$$\tau_k^i(\ell) := \min\{n > \tau_k^i(\ell-1) : U^i(n) = e^k\}, \qquad \ell = 2,3,\ldots \tag{8.13b}$$

and define

$$B_k^1(\tau_k^1(\ell)) := B_k(\tau_k^0(\ell)), \quad 1 \le k \le K \qquad \ell = 1,2,\ldots \tag{8.14}$$

With these definitions, the actual number of times each customer is served is identical in both systems, while the sequences $\{B(n),\ n = 0,1,\ldots\}$ (under $P_0^{1,t}$) and $\{B^1(n),\ n = 0,1,\ldots\}$ (under $\hat{P}$) are statistically indistinguishable. Consequently, the distribution of $\{X^1(n),\ n = 0,1,\ldots\}$ under $\hat{P}$ coincides with the distribution of $\{X(n),\ n = 0,1,\ldots\}$ under $P_0^{1,t}$. Moreover, by construction (with the notation of (6.3)), it is easy to see that $T^0 = T^1$ and $X_K^0(n) \le X_K^1(n)$ for all $n = 0,1,\ldots$ $\hat{P}$ a.s., whence

$$\Delta(\ell,t,s) = \hat{E}\left[1[T^1 = \ell]\left(X_K^1(s) - X_K^0(s)\right)\right] \ge 0. \tag{8.15}$$

Finally, for $s = t+1$, observe that on the event

$$A := [T^0 = \ell] \cap [X_K^0(t) \ne 0] \cap [X_k^0(t) \ne 0 \text{ for some } k = 1,2,\ldots,K-1] \cap [B_K(t) = 1], \tag{8.16}$$

the equality $X_K^1(t+1) - X_K^0(t+1) = 1$ holds, and that $\hat{P}[A] > 0$. Consequently,

$$\hat{E}\left[1[T^1 = \ell]\left[X_K^1(s) - X_K^0(s)\right]\right] := \epsilon(\ell,t,t+1) \ge \hat{P}[A] > 0 \tag{8.17}$$

and the result is established. ■

REFERENCES

[1] A. Altman and A. Shwartz, "Optimal priority assignment with general constraints," in *Proceedings of the 24th Allerton Conference on Communications, Control and Computing*, pp. 1147-1148, Monticello (IL) (1987).

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

[3] J.S. Baras, A.J. Dorsey and A.M. Makowski, "Two computing queues with geometric service requirements and linear costs: the μc-rule is often optimal," *Adv. Appl. Prob.* **17**, pp. 186-209 (1985).

[4] 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).

[5] A. Benveniste, M. Metivier and P. Priouret, *Algorithmes Adaptatifs et Approximations Stochastiques*, Masson, Paris (France) (1987).

[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. Buyyukkoc, P.P. Varaiya, and J. Walrand, "The $c\mu$-rule revisited," *Adv. Appl. Prob.* **17**, pp. 234-235 (1985).

[9] C. Derman and A. F. Veinott, Jr. "A solution to a countable system of equations arising in Markovian decision processes," *Ann. Math. Stat.* **38**, pp. 582-584 (1967).

[10] E. G. Gladyshev, "On Stochastic Approximation," *Theo. Prob. Appl.* **10**, pp. 275-278 (1965).

[11] J. M. Harrison, "A priority queue with discounted linear costs," *Operations Res.* **23**, pp. 270-282 (1975).

[12] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*, Academic Press, New York (NY) (1974).

[13] 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).

[14] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM J. Control Opt.* **23**, pp. 329-380 (1985).

[15] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge (MA) (1984).

[16] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*, Applied Mathematical Sciences **26**, Springer-Verlag, Berlin (1978).

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

[18] A.M. Makowski and A. Shwartz, "Implementation issues for Markov decision processes," pp. 323-337 in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, W. Fleming and P.-L. Lions, Eds., The IMA Volumes in Mathematics and Its Applications **10**, Springer-Verlag, New York (NY) (1988).

[19] 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).

[20] P. Mandl, "Estimation and control in Markov chains," *Adv. Appl. Prob.* **6**, pp. 40-60 (1974).

[21] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Info. Theory* **IT-30**, pp. 140-150 (1984).

[22] P. Nain and K. W. Ross, "Optimal priority assignment with hard constraint," *IEEE Trans. Auto. Control* **AC-31** pp. 883-888 (1986).

[23] K. W. Ross, *Constrained Markov Decision Processes with Queueing Applications*, Ph.D. Thesis, Computer, Information and Control Engineering, University of Michigan, Ann Arbor (MI) (1985).

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

[25] A. Shwartz and A.M. Makowski, "An optimal adaptive scheme for two competing queues with constraints," pp. 515-532 in *Analysis and Optimization of Systems*, A. Bensoussan and J.-L. Lions Eds., Springer-Verlag Lecture Notes in Control and Information Sciences **83** (1986).

[26] A. Shwartz and A. M. Makowski, "Comparing policies in Markov decision processes: Mandl's Lemma revisited," *Mathematics of Operations Research* **15** pp. 155–174 (1990).

[27] A. Shwartz and A. M. Makowski, "On the Poisson equation for Markov chains," *Mathematics of Operations Research*, under revision (1987).

[28] M. Sidi and A. Segall, "Structured priority queueing systems with applications to packet-radio networks," *Performance Evaluation* **3**, pp. 265-275 (1983).

[29] P. Tsoucas and J. Walrand, "Optimal adaptive server allocation in a network," *Systems & Control Letters* **7**, pp. 323-327 (1986).