Value Iteration and Optimization of Multiclass Queueing Networks

Rong-Rong Chen and Sean Meyn

ABSTRACT. This paper considers in parallel the scheduling problem for multiclass queueing networks, and optimization of Markov decision processes. It is shown that the value iteration algorithm may perform poorly when the algorithm is not initialized properly. The most typical case where the initial value function is taken to be zero may be a particularly bad choice. In contrast, if the value iteration algorithm is initialized with a stochastic Lyapunov function, then the following hold

(i): A stochastic Lyapunov function exists for each intermediate policy, and hence each policy is *regular* (a strong stability condition).

(ii): Intermediate costs converge to the optimal cost.

(iii): Any limiting policy is average cost optimal.

It is argued that a natural choice for the initial value function is the value function for the associated deterministic control problem based upon a fluid model, or the approximate solution to Poisson's equation obtained from the LP of Kumar and Meyn. Numerical studies show that either choice may lead to fast convergence to an optimal policy.

1991 *Mathematics Subject Classification.* Primary 90B35, 68M20, 90B15, 90C40. Secondary 93E20, 60J20.

Key words and phrases. Multiclass queueing networks; Markov decision processes, optimal control, dynamic programming.

Work supported in part by NSF Grant ECS 940372, and JSEP grant N00014-90-J-1270. This work was completed with the assistance of equipment granted through the IBM Shared University Research program and managed by the Computing and Communications Services Office at the University of Illinois at Urbana-Champaign.

1. Introduction

This paper presents a convergence proof for the value iteration algorithm for general Markov decision processes, and also develops methods for the application of this algorithm to the synthesis of optimal scheduling policies for multiclass queueing networks. The latter results are based upon the close connection between optimization of a network, and optimal control of an associated fluid network model.

Over the past ten years there have been several successful attempts to approximate a network model with a more tractable process to reduce the complexity of the control synthesis problem. The recent paper [**MSS96**] treats the optimal control of a multiclass queueing network by relating this problem to the optimal control of an associated diffusion process in heavy traffic, following the work of [**HW89**]. Methods for translating an optimal policy for the Brownian system model back to an implementable policy for the discrete-stochastic model are introduced in [**Har96**]. In [**Mey97b, Mey96**] it is shown that the value function for the network scheduling problem can be approximated by the value function for an associated fluid limit model. Some heuristics based upon this result are developed in [**Mey97**] to translate a policy for the fluid model back to the original discrete network. The results reported here provide a more exact approach to translating an optimal policy for the fluid model back to the original problem of interest.

We begin with the analysis of a general Markov Decision Process model with one step cost c and state process $\mathbf{\Phi} = \{\Phi(t) : t \geq 0\}$ evolving on a countable state space X. Our goal is to solve the average cost optimal control problem by constructing a stationary policy $\boldsymbol{w}$ with minimal average cost

$$J(\boldsymbol{w}, x) := \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}_x[c(\Phi(t), w(\Phi(t))]. \tag{1.1}$$

Value iteration is perhaps the most common approach in practice to constructing an optimal policy. The idea is to consider the finite time problem with value function

$$V_n(x) = \min \mathsf{E}_x\Big[\sum_{t=0}^{n-1} c(\Phi(t), a(t)) + V_0(\Phi(n))\Big], \tag{1.2}$$

where $\{a(t) : t \geq 0\}$ is a sequence of actions determined by some policy, and the minimum in (1.2) is with respect to all policies. The function $V_0 \colon \mathsf{X} \to \mathbb{R}_+$ is a penalty term - the standard value iteration procedure uses $V_0 \equiv 0$. Letting $\boldsymbol{v}^n$ denote a policy which attains this minimum, it may be assumed without loss of generality that there is a sequence of *state feedback functions* $w^k \colon \mathsf{X} \to \mathcal{A}$, $k \geq 0$, such that for any n, the policy $\boldsymbol{v}^n$ is a Markov policy whose first n actions may be expressed

$$\boldsymbol{v}^n_{[0,n-1]} = (w^{n-1}(\Phi(0)), \ldots, w^0(\Phi(n-1))).$$

The value iteration algorithm is then the standard dynamic programming approach to recursively computing an optimal sequence $(V_n, w^n : n \geq 1)$.

Various convergence proofs and counterexamples have appeared since the early sixties, with most of the general positive results holding in the case of finite state space models only. A thorough survey is found on pages 429–433 of [**Put94**]. In early papers the analyses typically focus on the differential cost function $g_n(x) = V_{n+1}(x) - V_n(x)$ and the normalized value function $h_n(x) = V_n(x) - V_n(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$

is some distinguished state. Under various conditions one may show that as $n \to \infty$, possibly through a subsequence,

$$g_n \to \eta_*, \quad h_n \to h, \quad w^n \to w^*$$

where η_* is the average minimum cost, and the triple (η_*, w^*, h) is a solution of the average cost optimality equation (see (2.1,2.2) below).

Recently there has been a resurgence of interest in understanding the algorithm when the state space is unbounded. The paper [**Cav96**] treats countable state space models where the state space is a single communication class under any stationary policy. Convergence holds under two natural assumptions: the stabilizability condition that the steady state cost is finite for some stationary policy; and the *almost monotone* condition on the cost function of [**Bor91**]. The irreducibility assumption was relaxed in [**CF95**] by imposing a global Lyapunov function condition similar to that of [**Hor77**]. The global Lyapunov function condition is expressed as,

$$\mathsf{E}^{\boldsymbol{w}}[V(\Phi(t+1)) \mid \Phi(t) = x] \le V(x) - c(x, w(x)) + b\mathbb{1}_S(x), \qquad t \in \mathbb{Z}_+, \tag{1.3}$$

where V is a positive function on the state space, $b < \infty$, and S is a finite set, or more generally a compact set. It is assumed in [**CF95**] that there exists a single function V such that (1.3) holds for every Markov policy $\boldsymbol{w}$, where $S = \{\boldsymbol{\theta}\}$ is a singleton. Under this assumption it may be shown that $\mathsf{E}_x^{\boldsymbol{v}}[\tau_\theta]$ is uniformly bounded over all policies, for each initial condition x, where τ_θ is the first return time to the state $\boldsymbol{\theta} \in \mathsf{X}$.

In the paper [**Sen96**], conditions are determined under which the optimal cost η_* is computable through the limit $\eta_* = \lim_{n\to\infty} V_n(x)/n$, $x \in \mathsf{X}$. The analysis is based upon the discounted control problem, and the use of a truncated value function to avoid the difficulties associated with unbounded costs. The paper begins with some implicit bounds on the relative discounted value function for the truncated control problem. These assumptions are related to more readily verifiable conditions such as the near monotone condition of [**Bor91**], and the Lyapunov condition of [**Hor77**]. Hence [**Sen96**] captures some aspects of the results of [**CF95, Cav96**].

None of these contributions are applicable in general for multiclass network models since both the Lyapunov condition and the irreducibility condition fails for many models. A contribution of the present paper is to establish conditions for convergence which are valid in the networks context. Both the assumptions imposed and the methods of analysis are based on the recent treatment of the policy iteration algorithm of [**Mey97b**].

The major contribution of this paper is to resolve a significant drawback to the value iteration approach - it can be extremely slow. On page 385 of [**Put94**] the author writes "In average reward models, value iteration may converge very slowly, and policy iteration may be inefficient in models with many states ..." Indeed, we have applied value iteration to network models with approximately 50,000 states where policy iteration is not directly applicable, and we have found that convergence is slow even for very simple models. The explanation in the network case is easily seen. One is attempting to approximate the relative value function $h(x)$ by the difference $h_n(x) = V_n(x) - V_n(\boldsymbol{\theta})$. When V_0 is taken to be zero, then each approximation h_n is bounded by a linear function of x, and can grow by at most one in each iteration. The actual relative value function h is equivalent to a quadratic on the state space [**Mey97b, Mey97**], so there is a large mismatch between

the two functions whenever the state is large. For this reason, each of the state feedback laws $\{w^n\}$ generated by the value iteration algorithm can actually induce a transient state process $\mathbf{\Phi}$ (see Section 4).

We show in this paper that if the value iteration algorithm is initialized with V_0 equal to a stochastic Lyapunov function satisfying (1.3) for just *one* policy $\boldsymbol{w}$, then the value iteration algorithm constructs recursively solutions to a version of the drift inequality (1.3) for each n. Hence we reach the same conclusion established for the policy iteration algorithm in the companion paper [**Mey97b**]: the strong stability condition (1.3) is superfluous when working with the value iteration algorithm because the algorithm automatically generates stabilizing policies. It is only necessary to find an initial stabilizing policy to initialize the algorithm. Based upon this observation we prove that the intermediate average costs $J(\boldsymbol{w}^n, x)$ are finite for each n, and independent of x; that the average costs $J(\boldsymbol{w}^n, x)$ converge to the optimal cost η_* as $n \to \infty$; and that any limiting policy is average cost optimal.

Some of these ideas have been generalized to the risk sensitive control problemΦin [**BM98**].

In the network optimization problem the relative value function for the optimal policy may be approximated by the value function for the associated fluid control problem. It is thus natural to use the latter value function to initialize the value iteration algorithm. A second approach we consider is based on computing an approximate solution to Poisson's equation through the stability LP of [**KM95**]. Results from numerical experiments show that either choice may lead to fast convergence to an optimal policy. We thus arrive at a new way of using the information gained from solving a deterministic optimization problem to solve the original discrete scheduling problem of interest.

The paper is organized as follows. In the following section we present the main results concerning the convergence of the value iteration algorithm. The assumptions are satisfied for general multiclass queueing networks of the form described in Section 3. Methods for constructing suitable initializations for the value iteration algorithm for the network scheduling problem are described in Sections 4–6. The appendices contain proofs of the main results and some background theory.

2. Value iteration

Consider a general Markov Decision Process whose state space X and action space $\mathcal{A}$ are countable. Detailed treatments of Markov Decision Processes can be found in, for instance [**Put94**]. We present here a bare-bones description of the general model.

Associated with each $x \in \mathsf{X}$ is a non-empty subset $\mathcal{A}(x) \subseteq \mathcal{A}$ whose elements are the admissible actions when the state Φ_t takes the value x at time t. The transitions of the state process $\mathbf{\Phi}$ are governed by the conditional probability distributions $\{P_a(x, y)\}$ which describe the probability that the next state is $y \in \mathsf{X}$ given that the current state is $x \in \mathsf{X}$, and the current action chosen is $a \in \mathcal{A}$. A policy $\boldsymbol{w}$ is a sequence of actions $\{a(t) : t \in \mathbb{Z}_+\}$ which is adapted, that is, $a(t)$ can only depend on the history $\{\Phi(0), \ldots, \Phi(t)\}$. We will consider primarily *Markov* policies of the form $\boldsymbol{w} = \{w^0(\Phi(0)), w^1(\Phi(1)), w^2(\Phi(2)), ...\}$, where for each i the function w^i maps X to $\mathcal{A}$, with $w^i(x) \in \mathcal{A}(x)$ for each x. For a Markov policy $\boldsymbol{w}$ we denote the resulting Markov chain $\mathbf{\Phi}^{\boldsymbol{w}} := \{\Phi^{\boldsymbol{w}}(t) : t \geq 0\}$ - we simply write $\mathbf{\Phi}$ if it is clear from the context which policy has been applied.

A *stationary* policy is a Markov policy for which $w^i = w$ for all i, for some fixed state feedback law w. The action $w(x)$ is applied whenever the state takes the value x, independent of the past and independent of the time-period. We shall write $P_w(x, B) = P_{w(x)}(x, B)$ for the transition law corresponding to a stationary policy $\boldsymbol{w}$. The n-step transition probabilities are denoted

$$P_w^n(x, y) = \mathsf{P}(\Phi^w(n) = y \mid \Phi^w(0) = x), \qquad x, y \in \mathsf{X}.$$

We also use the operator-theoretic notation,

$$P_w^n h\,(x) := \mathsf{E}[h(\Phi^w(n)) \mid \Phi^w(0) = x],$$

where h is any real-valued function on X.

The resolvent kernel is defined for a feedback law w as

$$K_w = \sum_{t=0}^{\infty} 2^{-(t+1)} P_w^t.$$

We will occasionally extend this definition to a Markov policy $\boldsymbol{v} = (v^0, v^1, \dots)$ via

$$K_{\boldsymbol{v}}(x, y) := \mathsf{E}_x^{\boldsymbol{v}}\Big[\sum_{t=0}^{\infty} 2^{-(t+1)} \mathbb{1}(\Phi(t) = y)\Big], \qquad x, y \in \mathsf{X}.$$

We assume that a cost function $c\colon \mathsf{X} \times \mathcal{A} \to [1, \infty)$ is given. The average cost of a particular policy $\boldsymbol{w}$ is, for a given initial condition x, defined as

$$J(\boldsymbol{w}, x) := \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathsf{E}_x^{\boldsymbol{w}}[c(\Phi^w(t), a(t))].$$

A policy $\boldsymbol{w}^*$ is then called *optimal* if $J(\boldsymbol{w}^*, x) \le J(\boldsymbol{w}, x)$ for all policies $\boldsymbol{w}$, and any initial state x.

A central concept in this paper is the notion of f-regularity, as developed in [**MT93**], with f equal to some function on the state space. In the present paper the functions of interest take the form $f = c_w$ with $c_w(y) = c(y, w(y))$, $y \in \mathsf{X}$. The definition takes a simple form since the state space is assumed to be countably infinite: the controlled chain is c_w-regular if for some distinguished state $\boldsymbol{\theta} \in \mathsf{X}$,

$$\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta - 1} c_w(\Phi(t))\Big] < \infty, \qquad x \in \mathsf{X},$$

where τ_θ is the first return time to the state $\boldsymbol{\theta}$.

Following [**Mey97b**], a state feedback law w will be called *regular* if the controlled chain is a c_w-regular Markov chain; a policy $\boldsymbol{w}$ is also called regular provided that $\boldsymbol{w}$ is a stationary Markov policy defined by a regular state feedback law. This a highly desirable stability property for the controlled process. If the feedback law w is regular, then necessarily an invariant probability measure π_w exists such that $\pi_w(c_w) < \infty$. Moreover, for a regular w, the resulting cost is $J(\boldsymbol{w}, x) = \sum \pi_w(y) c_w(y)$, independent of x. The following result is a consequence of the f-norm ergodic theorem of [**MT93**, Chapter 14].

THEOREM 2.1. *For any regular feedback law w, there exists a unique invariant probability π_w, and the controlled state process $\boldsymbol{\Phi}$ satisfies $\eta_w := \sum c_w(x)\, \pi_w(x) < \infty$.*

For each initial condition the average cost is equal to η_w, independent of x, and the following limits hold:

$$J(\boldsymbol{w}, x) = \eta_w = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathsf{E}_x[c_w(\Phi(t))] = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} c_w(\Phi(t)), \qquad a.s.\ [\mathsf{P}_x].$$

□

In view of Theorem 2.1, we denote the average cost $J(w) = J(\boldsymbol{w}, x)$ when the policy $\boldsymbol{w}$ is regular.

The construction of an optimal policy typically involves the solution to the following equations

$$\eta_* + h_*(x) = \min_{a\in\mathcal{A}(x)} [c(x,a) + P_a h_*(x)] \tag{2.1}$$

$$w^*(x) = \arg\min_{a\in\mathcal{A}(x)} [c(x,a) + P_a h_*(x)], \qquad x \in \mathsf{X}. \tag{2.2}$$

The equality (2.1), a version of Poisson's equation, is known as the *average cost optimality equation* (ACOE). The second equation (2.2) defines a stationary policy $\boldsymbol{w}^*$ (see e.g. [**Put94, ABF+93, Bor91**] for further discussion).

The *value iteration algorithm*, or *VIA*, is defined inductively as follows. If the value function V_n is given, the action $w^n(x)$ is defined as

$$w^n(x) = \arg\min_{a\in\mathcal{A}(x)} [P_a V_n(x) + c(x,a)], \qquad x \in \mathsf{X}.$$

For each n the following dynamic programming equation is satisfied,

$$V_{n+1}(x) = c_{w^n}(x) + P_{w^n} V_n(x) = \min_{a\in\mathcal{A}(x)} (P_a V_n(x) + c_a(x)),$$

which then makes it possible to compute the next function w^{n+1}.

To simplify notation we denote throughout the remainder of this paper

$$c_n = c_{w^n}; \quad P_n = P_{w^n}; \quad K_n = K_{w^n},$$

and we let E^n denote the expectation operator induced by the stationary policy

$$\boldsymbol{w}^n := (w^n(\Phi(0)), w^n(\Phi(1)), w^n(\Phi(2)), \dots).$$

Suppose that $\boldsymbol{\theta} \in \mathsf{X}$ is some distinguished state, and define

$$h_n(x) = V_n(x) - V_n(\boldsymbol{\theta}); \qquad g_n(x) = V_{n+1}(x) - V_n(x), \qquad x \in \mathsf{X}, n \in \mathbb{Z}_+. \tag{2.3}$$

Then for each n we have the identity $P_n h_n = h_n - c_n + g_n$, which at least superficially resembles the ACOE. We show below that the pair $\{h_n, g_n\}$ does indeed converge to a solution $\{h_*, \eta_*\}$ to (2.1), (2.2) under reasonable conditions on the model and on the initial value function V_0.

We assume that at least one regular policy $\boldsymbol{w}^{-1}$ exists, so that there also exists a function $V_0 \colon \mathsf{X} \to \mathbb{R}_+$ and an $\overline{\eta} < \infty$ such that

$$P_{-1} V_0 \le V_0 - c_{-1} + \overline{\eta}. \tag{2.4}$$

This stabilizability assumption is a generalization of that used in [**Cav96**] and many other papers. If the value iteration algorithm is initialized with this function V_0, then the resulting penalty term in V_n, $n \ge 1$, "regularizes" the intermediate policies to ensure that each is stabilizing, and in the examples considered it also appears to speed convergence.

We assume throughout the paper that there exists a regular, optimal policy $\boldsymbol{w}^*$, a P_{w^*}-invariant probability π_{w^*}, and a relative value function h_* satisfying

the ACOE with $\eta_* = \pi_{w^*}(c_{w^*})$. The quantities $(\boldsymbol{w}^*, \pi_{w^*}, h_*)$ may not be unique, but we fix one such triple throughout the paper in our assumptions and in the analysis to follow. These conditions will be met under the stabilizability part of Assumption (A1), and Assumptions (A2), (A3) below (the conditions of [**Sen86**] may be verified, following the approach of [**Mey97b**]).

Assumptions (A2) and (A3) are related to the near-monotone assumption of [**Bor91**]. We call a function $\underline{c}$ *norm-like* if the sublevel set $\{x : \underline{c}(x) \le b\}$ is a finite subset of X for any finite b. It is assumed in Assumption (A2) below that for any "reasonable" policy $\boldsymbol{w}$, the cost $c_w(x)$ dominates a norm-like function $\underline{c}$ on X. Assumption (A3) then imposes an accessibility condition on the sublevel set $S_0 \subset \mathsf{X}$ defined as

$$S_0 = \{x : \underline{c}(x) \le 2\overline{\eta}\}. \tag{2.5}$$

Although this assumption imposes an accessibility condition on all Markov policies, Assumption (A3) is only used for the optimal policies $\boldsymbol{v}^n$, and the stationary policies $\boldsymbol{w}^n$, $n \in \mathbb{Z}_+$. Assumption (A3) is weaker than the Lyapunov condition of [**CF95**]. It is satisfied for the network scheduling problem described in Section 3, and other storage and routing models found in the operations research area.

Formally, our assumptions are summarized as follows:

(A1): There exists a policy $\boldsymbol{w}^{-1}$, a function $V_0 \colon \mathsf{X} \to \mathbb{R}_+$, and $\overline{\eta} < \infty$ satisfying (2.4) and for the optimal policy $\boldsymbol{w}^*$,

$$\lim_{n\to\infty} \frac{1}{n}\Big(P^n_{w^*} V_0\Big)(x) = 0, \qquad x \in \mathsf{X}.$$

(A2): For each fixed x, the function $c(x, \cdot\,)$ is is norm-like on $\mathcal{A}$, and there exists a norm-like function $\underline{c} \colon \mathsf{X} \to \mathbb{R}_+$ such that for any regular polcy $\boldsymbol{w}$ satisfying $J(w) \le \overline{\eta}$,

$$c_w(x) \ge \underline{c}(x), \qquad x \in \mathsf{X}.$$

(A3): There is a fixed state $\boldsymbol{\theta} \in \mathsf{X}$ and a $\delta > 0$ such that for any Markov policy $\boldsymbol{v}$,

$$K_{\boldsymbol{v}}(x, \boldsymbol{\theta}) \ge \delta \qquad \text{for all } x \in S_0, \tag{2.6}$$

where S_0 is defined in (2.5), and for any action $a \in \mathcal{A}(\boldsymbol{\theta})$,

$$P_a(\boldsymbol{\theta}, \boldsymbol{\theta}) \ge \delta.$$

We will occasionally also assume

(A4): For any regular optimal policy $\boldsymbol{w}$ the controlled chain is irreducible in the usual sense that

$$K_w(x, y) > 0, \qquad x, y \in \mathsf{X}.$$

The main results of this section are summarized in the following two theorems.

THEOREM 2.2. *For the value iteration algorithm initialized with the function V_0, suppose that Assumptions (A1)–(A3) are satisfied. Then*

(i): *For each x the sequences $\{g_n(x)\}$ and $\{h_n(x)\}$ are bounded, and*

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} V_n(\boldsymbol{\theta}) &= \lim_{n\to\infty} g_n(\boldsymbol{\theta}) = \eta_*. \\ \limsup_{n\to\infty} \frac{1}{n} V_n(x) &\le \limsup_{n\to\infty} g_n(x) \le \eta_*, \qquad x \in \mathsf{X}. \end{aligned}$$

(ii): *Each intermediate state feedback law w^n is regular, and each V_n serves as a Lyapunov function for the nth policy:*

$$P_n V_n \le V_n - c_n + \overline{\eta}, \qquad n \ge 0.$$

(iii): *The average cost $J(w^n)$ converges to the optimal cost:*

$$J(w^n) \to \eta_* \qquad \text{as } n \to \infty.$$

(iv): *Any point-wise limit point of the feedback laws $\{w^n : n \in \mathbb{Z}_+\}$ is regular and optimal.*

□

THEOREM 2.3. *Under Assumptions (A1)-(A4) the conclusions of Theorem 2.2 hold, and in addition, as $n \to \infty$,*

$$\begin{aligned} h_n(x) &\to h_*(x) - h_*(\boldsymbol{\theta}) \\ g_n(x) &\to \eta_*, \\ \frac{1}{n} V_n(x) &\to \eta_*, \qquad\qquad x \in \mathsf{X}. \end{aligned}$$

□

PROOF OF THEOREM 2.2. The two limits in (i) follow from Lemma B.9 (ii). The bounds on the two limit supremums follow from Lemma B.9 (i) and the formula

$$\frac{1}{n} \sum_{t=0}^{n-1} g_t(x) = \frac{1}{n} (V_n(x) - V_0(x)).$$

The bound in (ii), and hence also regularity, is established in Proposition B.3.

Result (iii) requires the lower bound $c_w \ge \underline{c}$. From Proposition B.3 and this lower bound we have $\pi_n(\underline{c}) \le \overline{\eta}$ for all n, which shows that the probabilities $\{\pi_n : n \in \mathbb{Z}_+\}$ are tight. From Proposition B.3 and the Comparison Theorem A.1 we have $\pi_n(c_n) \le \pi_n(g_n)$ where $g_n \le \overline{\eta}$. Since the probabilities $\{\pi_n\}$ are tight, for any preassigned $\varepsilon > 0$ there exists a finite set $C \subset \mathsf{X}$ with $\pi_n(C) > 1 - \varepsilon$ for all n. Thus,

$$\begin{aligned} \overline{\eta} - \limsup_{n\to\infty} J(w^n) &\ge \liminf_{n\to\infty} (\overline{\eta} - \pi_n(g_n)) \\ &\ge \liminf_{n\to\infty} \sum_{x\in C} \pi_n(x)(\overline{\eta} - g_n(x)) \\ &\ge (1-\varepsilon)(\overline{\eta} - \eta_*). \end{aligned}$$

Since ε is arbitrary, this shows that $\limsup_{n\to\infty} J(w^n) \le \eta_*$, as desired.

Finally, result (iv) is immediate from (i) and Fatou's lemma applied to the identity $P_n h_n = h_n - c_n + g_n$ obtained in Proposition B.3. □

PROOF OF THEOREM 2.3. The convergence of h_n follows from Lemma B.11. We may then use the identity $g_n(x) = g_n(\boldsymbol{\theta}) + h_{n+1}(x) - h_n(x)$ and Theorem 2.2 (i) to prove that $g_n(x)$ converges to η^*. The convergence of $V_n(x)/n$ to η^* then follows as in the proof of Theorem 2.2 (i). □

3. Discrete and fluid models for multiclass networks

Consider a network of the form illustrated in Figure 1, composed of d single server stations, indexed by $\sigma = 1, \dots, d$. The network is populated by K classes of customers, and an exogenous stream of customers of class 1 arrive to machine $s(1)$. A customer of class k requires service at station $s(k)$. If the service times

and interarrival times are assumed to be exponentially distributed, then after a suitable time scaling and sampling of the process, the dynamics of the network can be described by the random linear system,

$$\Phi(t+1) = \Phi(t) + \sum_{i=0}^{K} I_i(t+1)[e^{i+1} - e^i] w_i(\Phi(t)), \tag{3.1}$$

where the state process $\mathbf{\Phi}$ evolves on $\mathsf{X} = \mathbb{Z}_+^K$. The random variable $\Phi_i(t)$ denotes the number of class i customers in the system at time t which await service in buffer i at station $s(i)$. The function $w\colon \mathsf{X} \to \{0,1\}^{K+1}$ is the policy to be designed. If $w_i(\Phi(t))I_i(t+1) = 1$, this means that at the discrete time t, a customer of class i is just completing a service, and is moving on to buffer $i+1$ or, if $i = K$, the customer then leaves the system. The set of admissible control actions $\mathcal{A}(x)$ when the state is $x \in \mathsf{X}$ is defined as follows for $a = (a_0, \ldots, a_K)' \in \mathcal{A}(x) \subset \{0,1\}^{K+1}$,

(i): $a_0 = 1$, and for any $1 \le i \le K$, $a_i = 0$ or 1;
(ii): For any $1 \le i \le K$, $x_i = 0 \Rightarrow a_i = 0$;
(iii): For any station σ, $0 \le \sum_{i:s(i)=\sigma} a_i \le 1$.
(iv): For any station σ, $\sum_{i:s(i)=\sigma} a_i = 1$ whenever $\sum_{i:s(i)=\sigma} x_i > 0$.

Condition (iv) is the *non-idling* property that a server will always work if there is work to be done. With the one step cost $c_w(x) = |x| := \sum_i x_i$, the non-idling condition may be assumed without any loss of generality [**Mey97**].

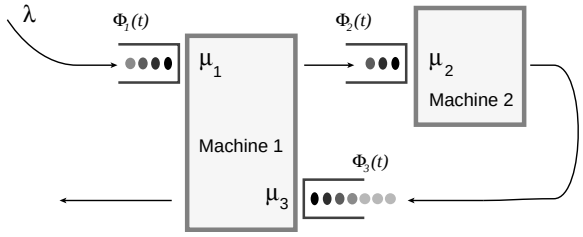

FIGURE 1. A multiclass network with $d = 2$ and $K = 3$.

The random variables $\{I(t) : t \ge 0\}$ are i.i.d. on $\{0,1\}^{K+1}$, with $\mathsf{P}\{\sum_i I_i(t) = 1\} = 1$, and $\mathsf{E}[I_i(t)] = \mu_i$. For $1 \le i \le K$, μ_i denotes the service rate for class i customers, and for $i = 0$ we let $\mu_0 := \lambda$ denote the arrival rate of customers of class 1. For $1 \le i \le K$ we let e^i denote the ith basis vector in $\mathbb{R}^K$, and we set $e^0 = e^{K+1} := 0$. It is evident that these specifications define a Markov Decision Process whose state transition function has the simple form,

$$\begin{aligned} \mathsf{P}_a(x, x + e^{i+1} - e^i) &= \mu_i a_i, \qquad 0 \le i \le K. \\ \mathsf{P}_a(x, x) &= 1 - \sum_0^K \mu_i a_i. \end{aligned}$$

We assume throughout the paper that the usual load conditions are satisfied

$$\rho_\sigma = \sum_{i:s(i)=\sigma} \frac{\lambda}{\mu_i} < 1, \qquad 1 \le \sigma \le d. \tag{3.2}$$

For concreteness we consider exclusively the one step cost $c_w(x) = |x|$. It is now known that for a network model with this cost criterion, regularity of a

stationary policy $\boldsymbol{w}$ is closely connected with the stability of an associated fluid limit model [**Dai95, DM95**]. For each initial condition $\Phi^w(0) = x \neq \boldsymbol{\theta}$ of the controlled chain $\boldsymbol{\Phi}^w$ we construct a continuous time process $\phi^x(t)$ as follows. If $|x|t$ is an integer, set

$$\phi^x(t) = \frac{1}{|x|}\Phi^w(|x|t).$$

For all other $t \geq 0$, define $\phi^x(t)$ by linear interpolation, so that it is continuous and piecewise linear in t. Note that $|\phi^x(0)| = 1$, and that ϕ^x is Lipschitz continuous. The collection of all "fluid limits" is defined by

$$\mathcal{L} := \bigcap_{n=1}^{\infty} \overline{\{\phi^x : |x| > n\}}$$

where the overbar denotes weak closure. The process ϕ evolves on the state space $\mathbb{R}_+^K$. We shall also call $\mathcal{L}$ the *fluid limit model*, in contrast to the *fluid model* which is defined as the set of all continuous solutions to the differential equation

$$\frac{d}{dt}\phi(t) = \sum_{i=0}^{K} \mu_i[e^{i+1} - e^i]u_i(t), \qquad \text{a.e. } t \in \mathbb{R}_+, \tag{3.3}$$

where the function $u(t)$ is analogous to the discrete control, and satisfies similar constraints [**CM91**].

The fluid limit model $\mathcal{L}$ is called *L_p-stable* if

$$\lim_{t\to\infty} \sup_{\phi\in\mathcal{L}} \mathsf{E}[|\phi(t)|^p] = 0.$$

It is shown in [**KM96, Mey97**] that L_2-stability of the fluid limit model is equivalent to a form of c-regularity for the network:

THEOREM 3.1. *The following stability criteria are equivalent for the network under any nonidling, stationary Markov policy $\boldsymbol{w}$.*

(i): *The drift condition holds*

$$P_w V(x) \leq V(x) - |x| + \overline{\eta}, \qquad x \in \mathsf{X}, \tag{3.4}$$

where $\overline{\eta} \in \mathbb{R}_+$, and the function $V : \mathsf{X} \to \mathbb{R}_+$ is equivalent to a quadratic in the sense that, for some $\gamma > 0$,

$$1 + \gamma|x|^2 \leq V(x) \leq 1 + \gamma^{-1}|x|^2, \qquad x \in \mathsf{X}.$$

(ii): *For some quadratic function V,*

$$\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{n=0}^{\sigma_\theta} |\Phi(n)|\Big] \quad \leq \quad V(x), \qquad x \in \mathsf{X}.$$

(iii): *For some quadratic function V and some $\gamma < \infty$,*

$$\frac{1}{N}\sum_{n=1}^{N} \mathsf{E}_x^{\boldsymbol{w}}[|\Phi(n)|] \quad \leq \quad \frac{1}{N}V(x) + \gamma, \quad \textit{for all } x \textit{ and } N \geq 1.$$

(iv): *The fluid limit model $\mathcal{L}$ is L_2-stable.*

(v): *The total cost for the fluid limit $\mathcal{L}$ is uniformly bounded in the sense that*

$$\sup_{\phi\in\mathcal{L}} \mathsf{E}\Big[\int_0^\infty |\phi(\tau)|\, d\tau\Big] < \infty.$$

□

Using Theorem 3.1 it is shown in [**Mey97b**] that the optimal control of a network and the optimal control of the fluid model are also related. As an illustration, consider the three buffer example illustrated in Figure 1. We have taken the parameters

$$\lambda = 0.1429; \quad \mu_1 = 0.3492; \quad \mu_2 = 0.1587; \quad \mu_3 = 0.3492 \tag{3.5}$$

so that $\rho_1 = \lambda/\mu_1 + \lambda/\mu_3 = 9/10$, and $\rho_2 = \lambda/\mu_2 = 9/11$. The optimal policy for the fluid model illustrated in Figure 2 was computed in [**Wei95**]. It can be defined succinctly as

Serve buffer one if and only if $x_3 = 0$, or $x_1 \geq x_3 - 1 + (\mathbb{1}(x_2 = 0))^{-1}$.

The form of the policy is logical: If the second buffer is non-empty, then the last buffer receives exclusive service. When the second buffer x_2 is empty and $x_1 \geq x_3$ then, because service at buffer two is slow, the first buffer releases fluid to avoid starvation at the second machine.

The optimal policy was computed for the stochastic model with the performance index

$$J_n(x) = \frac{1}{n} \sum_{t=1}^{n} \mathsf{E}_x[|\Phi(t)|].$$

To compute the policy numerically we used value iteration, terminated at $n = 7,000$. The buffer levels were truncated so that $x_i < 45$ for all i. This gives rise to a finite state space Markov Decision Process with $45^3 = 91,125$ states. In Figure 3 we see the result of this computation. Again there is a roughly linear or affine switching curve - buffer one is served provided that buffer two is small, and the population at buffer one is reasonably large compared with that at buffer three. The policy illustrated in Figure 3 is closely approximated by the formula

Serve buffer one if and only if $x_3 = 0$, or $x_1 \geq x_3 - 29 + 10\exp(x_2/2)$.

The fluid limit of this approximation is precisely the optimal fluid policy illustrated in Figure 2.

4. Initialization of the VIA

For the optimal scheduling policy the relative value function h_* is equivalent to a quadratic on $\mathsf{X} = \mathbb{Z}_+^K$ in the sense that for some $\gamma > 0$,

$$\gamma|x|^2 \leq \big(h_*(x) - h_*(\boldsymbol{\theta})\big) \leq \gamma^{-1}|x|^2, \qquad x \in \mathsf{X},$$

where $\boldsymbol{\theta}$ is the vector of zeros in X [**Mey97b**]. However, in the standard VIA, for each n

$$\begin{aligned} V_n(x) &= \min \mathsf{E}_x^{\boldsymbol{w}} \Big[\sum_{t=0}^{n-1} |\Phi(t)| \Big] \\ &\leq \min \mathsf{E}_x^{\boldsymbol{w}} \Big[\sum_{t=0}^{n-1} (|\Phi(0)| + t) \Big] \\ &\leq n|x| + \tfrac{1}{2} n^2, \end{aligned}$$

where we are using the skip-free property of the network that $|\Phi(t+1)| \leq |\Phi(t)| + 1$, $t \in \mathbb{Z}_+$. It follows that, for each n, the function $h_n(x) = V_n(x) - V_n(\boldsymbol{\theta})$ is bounded from above by a linear function of x. Hence the approximation $h_n(x) \approx h_*(x)$ is grossly inaccurate for any finite n when the state x is large.

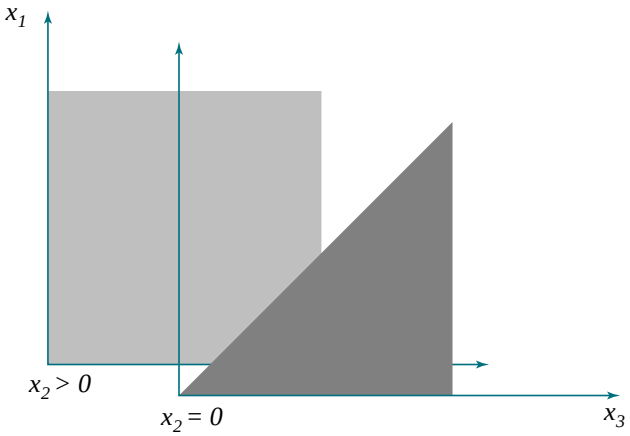

FIGURE 2. Optimal policy for the fluid model with $\rho_2 = 9/10$ and $\rho_1 = 9/11$. In this illustration, the grey regions indicate those states for which buffer three is given exclusive service.

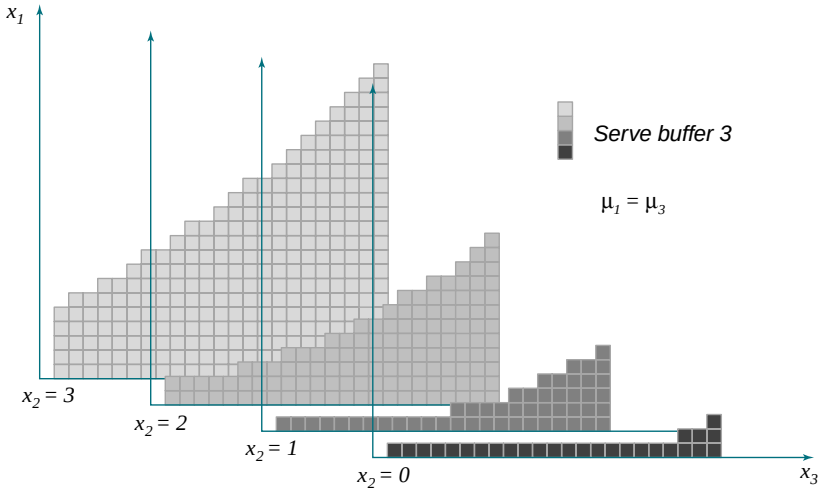

FIGURE 3. Optimal policy for the discrete network. There is an approximately affine switching curve which is similar to the linear switching curve found for the fluid model policy illustrated in Figure 2.

As one might then expect, for any n the actions $w^n(x)$, $x \in \mathsf{X}$, defined by the feedback law w^n may be far from optimal when the state x is large. We illustrate how the feedback laws $\{w^n\}$ generated by the standard VIA can give poor performance with the following two examples.

Example 1: a simple queue with controlled service. Here we describe a model which satisfies all of the conditions of [**Cav96**]. Hence with $V_0 \equiv 0$, the VIA does converge to give an optimal policy. However, for each iteration n the state feedback law w^n induces a transient Markov chain $\mathbf{\Phi}$, so that $J(w^n) = \infty$.

The state space is taken as $\mathsf{X} = \mathbb{Z}_+$, and the action space is $\mathcal{A}(x) = \{0,1\}$, $x \neq 0$, with $\mathcal{A}(0) = \{0\}$. There is an arrival rate λ and a variable service rate μ which takes on the small value μ_1, or the larger value $\mu_1 + \mu_2$, depending upon the control. We assume that $\lambda + \mu_1 + \mu_2 = 1$, and define the transition law as follows

$$\begin{aligned} P_a(x, x+1) &= \lambda \\ P_a(x, x-1) &= \mu_1 + \mu_2 a \\ P_a(x, x) &= (1 - \mu_1 - \mu_2 a - \lambda). \end{aligned}$$

That is, if $w(x) = 1$ then a customer receives maximal service during the current time slot when there are x customers in the queue. Assuming that $\rho = \lambda/(\mu_1+\mu_2) < 1$, the feedback law defined as $w^\alpha(x) = 1$, $x \geq \alpha$, is regular for any $\alpha \geq 1$, and the optimal policy is of this form for some α. We assume that $\lambda/\mu_1 > 1$, so that the lasy server defined by $w(x) \equiv 0$ gives rise to a transient chain $\mathbf{\Phi}$.

Suppose that $c_w(x) = (1 + w(x))x$, so that the one-step cost of serving a customer is proportional to the total number of customers in the system. Then for the standard VIA it may be shown inductively that there exists $\{\alpha_n : n \in \mathbb{Z}_+\}$ such that $w^n(x) = 0$ for all $x \geq \alpha_n$. Since no services are initiated when $x \geq \alpha_n$, it follows that the chain $\mathbf{\Phi}^{\boldsymbol{w}^n}$ obtained with the stationary policy $\boldsymbol{w}^n$ is transient for any n, although the policies $\{\boldsymbol{w}^n : n \geq 0\}$ do converge pointwise to give an optimal policy.

To obtain a version of the VIA which generates regular policies, initialize the algorithm with the function $V_0(x) = x^2/(\mu_1 + \mu_2 - \lambda)$. With the feedback law $w^{-1}(x) = \mathbb{1}(x > 0)$ we do have $P_{w^{-1}}V_0 \leq V_0 - c_{w^{-1}} + \overline{\eta}$ for some $\overline{\eta} < \infty$. It follows from Theorem 2.2 that each successive feedback law w^n is regular, and that the policies converge to an optimal policy as $n \to \infty$. In fact, it may be shown directly by induction that $w^n(x) = 1$ for all n and all x sufficiently large, which immediately implies that w^n is regular. We also note that in this case the function $h_n(x) = V_n(x) - V_n(0)$ is equivalent to a quadratic for all $n \geq 0$. Hence, when properly initialized, the VIA returns the correct form of both the optimal policy and the relative value function for large x, and only modifies the intermediate policies for x in a finite subset of X. □

Example 2: The Rybko-Stolyar Model. The next example treats a model introduced independently in the papers [**RS92, KS90**]. Consider the network illustrated in Figure 4 consisting of four buffers and two machines fed by separate arrival streams. It is shown in [**RS92**] that the last buffer-first served policy where buffers 2 and 4 receive priority at their respective machines will make the controlled process $\mathbf{\Phi}$ transient, even when the load conditions (3.2) are satisfied, if the cross-machine condition $\lambda/\mu_2 + \lambda/\mu_4 \leq 1$ is violated.

If the VIA is applied to this model with $V_0 \equiv 0$, then one obtains $V_1(x) = |x|$, and

$$V_2(x) = |x| + \min_w (|x| + 2\lambda - \mu_2 w_2(x) - \mu_4 w_4(x)).$$

The minimizing feedback law w^2 is evidently given by

$$w_2^2(x) = \mathbb{1}(x_2 > 0); \qquad w_4^2(x) = \mathbb{1}(x_4 > 0).$$

We conclude that $J(w^2) \equiv \infty$ when $\lambda/\mu_2 + \lambda/\mu_4 > 1$ since this is precisely the destabilizing policy introduced in [**RS92, KS90**]. □

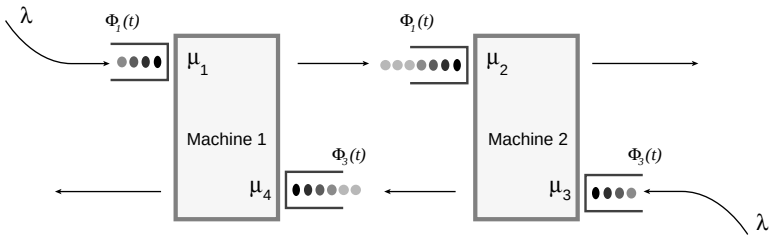

FIGURE 4. A multiclass network with $d = 2$ and $K = 4$.

These examples show that the standard VIA may return policies which are not stabilizing, and we have evidence to suspect that convergence may be slow, given the mismatch between the functions h_n and the limiting relative value function h_*. We have conducted numerical experiments to test the rate of convergence of the VIA for the three-buffer example illustrated in Figure 1 with the parameters defined in (3.5). Figure 5 shows the results from two implementations of the VIA. In the first we consider the standard algorithm with $V_0 \equiv 0$. For comparison purposes we also consider the case $V_0(x) = |x|^2$. This might appear to be a natural choice since it will result in lower values for the terminal cost. However, assumption (A1) is violated in both of these choices for V_0 since the drift inequality (2.4) is violated for any policy.

The vertical axis is the approximate value of the steady state cost $J(w^n)$, where $\boldsymbol{w}^n$ is the stationary policy obtained at the nth iteration of the VIA. Because the problem is large we cannot compute this cost exactly, but instead use

$$J(w^n) \approx \frac{1}{6,600} \sum_{t=1}^{6,600} \mathsf{E}_\theta^{\boldsymbol{w}^n} [|\Phi(t)|].$$

In either case we found that it takes several thousand iterations to reach convergence for this model. The figure shows results from the first 300 steps of the algorithm. Data was saved for n equal to multiples of ten: $n = 10, \ldots, 300$.

In summary, the standard VIA suffers from two potential drawbacks: intermediate policies may perform poorly, and may even give rise to a transient Markov chain, and the convergence of the algorithm can be slow. In the following two sections we propose methods to improve this situation by establishing general methods for constructing a more appropriate initial value function V_0. Although in practice we will never optimize the full infinite dimensional model, the approach described here may be used even in the finite state space truncated model to speed convergence. We see in the next section that for a network with buffer levels truncated to 33, the standard VIA requires thousands of iterations for convergence, while the VIA implemented with an appropriate initial value function converges to the same level of performance in less than twenty steps, even though the chain possesses $33^3 = 35,937$ states.

5. Initialization through the fluid model

From Theorem 3.1 (i) and Theorem 2.2 we conclude that the VIA will converge if the algorithm is initialized with a stationary policy $\boldsymbol{w}^{-1}$ whose fluid model is L_2-stable, since the network model will then satisfy Assumptions (A1)–(A3) when initialized with a solution to (3.4). Assumption (A1) will hold since the relative

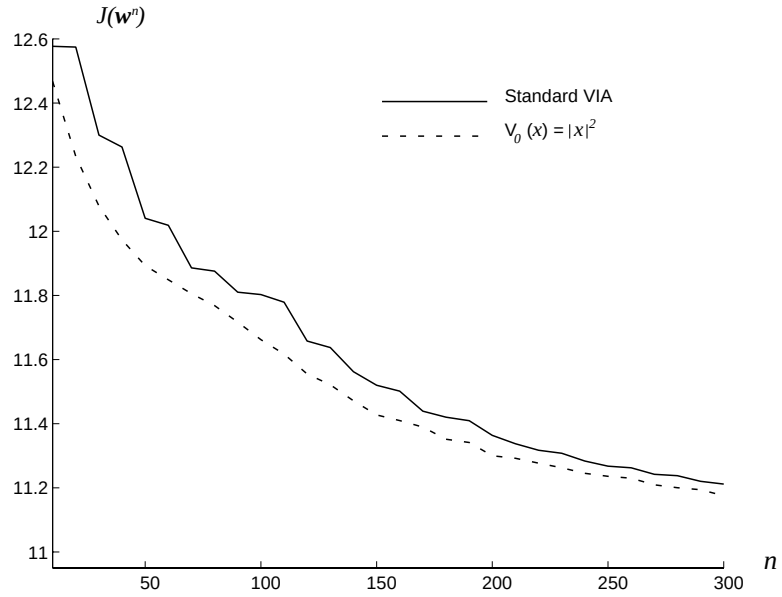

FIGURE 5. Convergence of the VIA with V_0 taken to be zero, and with $V_0(x) = |x|^2$. The vertical axis is an approximation to the steady state cost $J(w^n)$, and the horizontal axis is n, the number of iterations of the algorithm.

value function is equivalent to a quadratic for any such policy, and the accessibility Assumption (A3) holds with $\boldsymbol{\theta}$ equal to the empty state [**Mey97**]. There are many stabilizing policies which may serve as the initial policy $\boldsymbol{w}^{-1}$ (see [**CZ96**]). This leads to several algorithmic approaches to constructing the initial value function V_0 based on the value function for the fluid model. We begin with the following suggestive proposition. The result (ii) is proven in [**Mey97b**]. For the sake of brevity we omit the proof of (i).

PROPOSITION 5.1. If the feedback law w gives rise to a network whose fluid limit model $\mathcal{L}$ is L_2-stable, then

(i): the function V below is a solution to (3.4) for any $b > 1$ and all T sufficiently large.

$$V(x) = b|x|^2 \mathsf{E}_x^{\boldsymbol{w}} \Big[\int_0^T |\phi^x(\tau)|\, d\tau \Big]. \tag{5.1}$$

(ii): There exists a solution h to the Poisson equation $P_w h = h - |x| + \eta_{-1}$, and the function h approximates the value function V as follows,

$$(1 - \varepsilon(|x|, T))V(x) \le h(x) \le (1 + \varepsilon(|x|, T))V(x),$$

where $\varepsilon > 0$ and satisfies $\limsup_{T\to\infty} \limsup_{r\to\infty} \varepsilon(r, T) = 0$.

□

While the function V given in (5.1) is not easily computable in general, if the fluid limit model $\mathcal{L}$ is purely deterministic then we may use the limiting function

$$V(x) = b|x|^2 \int_0^\infty |\phi(\tau)|\, d\tau, \qquad \phi(0) = \frac{x}{|x|}. \tag{5.2}$$

We can also obtain an approximation to (5.1) based upon a finite dimensional linear program [**DEM96**]. If a sufficiently tight approximation to (5.1) is found then one can expect that (3.4) will hold for this approximation. We illustrate this approach with the three buffer example illustrated in Figure 1. We have computed explicitly the function V in (5.2) for two policies: LBFS, and the optimal policy illustrated in Figure 2.

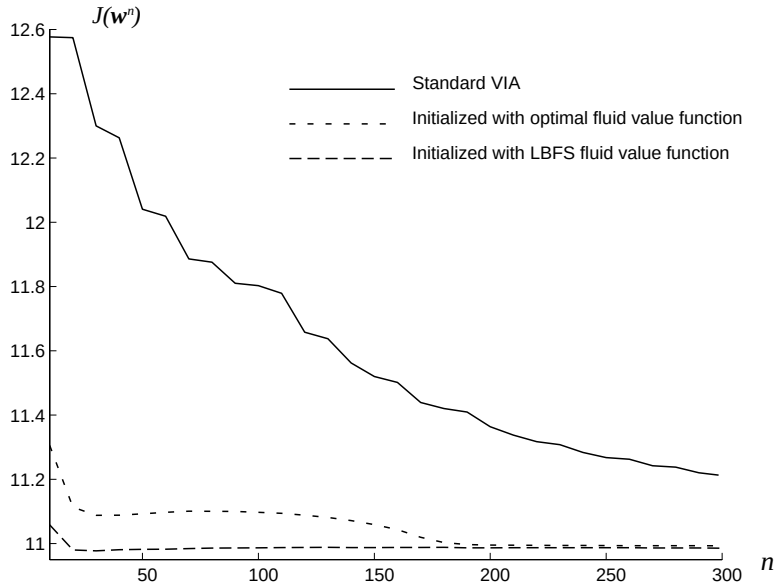

FIGURE 6. Convergence of the VIA with V_0 taken as the value function for the associated fluid control problem. Two value functions are found: one with the optimal fluid policy, and the other taken with the LBFS fluid policy. Both choices lead to remarkably fast convergence. Surprisingly, the "suboptimal" choice using LBFS leads to the fastest convergence of $J(w^n)$ to the optimal cost $\eta_* \approx 10.9$.

Two experiments were performed to compare the performance of the VIA initialized with these two value functions. The results from two experiments are shown in Figure 6. For comparison, data from the standard VIA shown earlier is also given. We have again taken 300 steps of value iteration, saving data for $n = 10, \dots, 300$. The parameter values $(\lambda, \mu_1, \mu_2, \mu_3)$ are again defined in (3.5). The convergence is exceptionally fast in both experiments. Surprisingly, the "suboptimal" choice using LBFS leads to the fastest convergence.

6. Initialization through a linear program

The computation of (5.2) is currently possible only for very simple models. We present here an alternative initialization of the VIA based upon the stability LP of [**KM95**] which is computationally feasible even for networks with hundreds of buffers.

Since the relative value function for an optimal policy is known to be equivalent to a quadratic, it is natural to attempt to find a quadratic form which gives the required negative drift. Let $\mathcal{I}$ denote the set of ordered pairs $\mathcal{I} = \{(i,j) : 1 \le i,j \le K\}$. For $I \in \mathcal{I}$ and α a vector of dimension $|\mathcal{I}| = K^2$ we define

$$h_I(x) = x_i x_j, \qquad h_\alpha(x) = \sum_{I \in \mathcal{I}} \alpha_I h_I(x).$$

Given a state feedback law w, let $z = z(x)$ denote the vector in $\mathbb{R}^{|\mathcal{I}|}$ whose Ith component is given by $z_I = w_i(x)x_j$, $I \in \mathcal{I}$. For any non-idling policy w, the cost can be expressed in terms of the vector z as $|x| = c^T z$, where $c \in \mathbb{R}^{|\mathcal{I}|}$ is the vector whose Ith component is defined as $c_I = \mathbb{1}(s(i) = s(j))$.

For any I there exist vectors $c^I \in \mathbb{R}^{|\mathcal{I}|}$, $\gamma^I \in \mathbb{R}^K$ and a constant B^I such that

$$P_w h_I\,(x) = h_I(x) - {c^I}^T z + {\gamma^I}^T w(x) + B^I.$$

The vector c^I is defined as follows. Let e^I be the vector that is zero except for a one in the $I = (i,j)$th position, so that ${e^I}^T z = z_I$. Then for $i = j$, $i = j+1$ and $i > j+1$, the vector $c^I = c^{ij}$ is given by

$$\begin{array}{l} 2\mu_{i-1}e^{i-1,i} - 2\mu_i e^{ii}, \\ \mu_{i-1}e^{i-1,i+1} - \mu_i e^{i,i+1} + \mu_i z_{ii} - \mu_{i+1}e^{i+1,i}, \text{ and} \\ \mu_{i-1}e^{i-1,j} - \mu_i e^{ij} + \mu_{j-1}e^{j-1,i} - \mu_j e^{ji} \end{array}$$

respectively. In addition, the non-idling constraint may be expressed as a linear inequality constraint on the variables z through the expression

$$\sum_{i:s(i)=s(k)} z_{i,k} \ge \sum_{i:s(i)=\sigma} z_{i,k}, \qquad 1 \le \sigma \le d,\ 1 \le k \le K.$$

Letting $d^{\sigma,k}$ denote the vector whose (i,j)th entry is defined as

$$d^{\sigma,k}_{i,j} = \begin{cases} -1 & \text{if } s(i) = s(j) \text{ and } j = k; \\ +1 & \text{if } s(i) = \sigma \text{ and } j = k; \\ 0 & \text{otherwise,} \end{cases}$$

this constraint may be expressed ${d^{\sigma,k}}^T z \le 0$, $1 \le \sigma \le d$, $1 \le k \le K$. For $\alpha \in \mathbb{R}_+^{K^2}$ and $\beta \in \mathbb{R}_+^d \times \mathbb{R}_+^K$ define the two vectors

$$c^\alpha := \sum_{I \in \mathcal{I}} \alpha_I c^I, \qquad d^\beta := \sum_{\sigma=1}^{d} \sum_{k=1}^{K} \beta_{\sigma,k} d^{\sigma,k}.$$

The stability LP of [**KM95**] can be interpreted as follows: Find $\alpha \in \mathbb{R}_+^{K^2}$, $\beta \in \mathbb{R}_+^d \times \mathbb{R}_+^K$ such that

$$c^\alpha + d^\beta \ge c, \tag{6.1}$$

where inequalities between vectors are interpreted component-wise. If this lower bound holds, then we obtain the desired Lyapunov drift inequality

$$\begin{aligned} P_w h_\alpha(x) &\le h_\alpha(x) - (c^\alpha + d^\beta)^T x + \gamma^{\alpha T} w(x) + B^\alpha \\ &\le h_\alpha(x) - |x| + \gamma^{\alpha T} w(x) + B^\alpha, \end{aligned}$$

where $\gamma^\alpha = \sum \alpha_I \gamma^I$; $B^\alpha = \sum \alpha_I B^I$.

There can be many values of α, β satisfying the lower bound (6.1). To find a value which is best in an average sense, first note that by the Comparison Theorem A.1 we have the steady state bound

$$\mathsf{E}_{\pi_w}[|x|] \le \mathsf{E}_{\pi_w}[c^{\alpha T} z] \le \mathsf{E}_{\pi_w}[\gamma^{\alpha T} w] + B^\alpha. \tag{6.2}$$

The right hand side is computed in the construction of the performance LP of [**KK94, DBT92**] giving the following formula.

$$\mathsf{E}_{\pi_w}[\gamma_{ij}^T w] + B^{ij} = \begin{cases} -2\lambda, & \text{if } i = j, \\ \lambda, & \text{if } j = i+1, \\ 0, & \text{if } j > i+1. \end{cases}$$

A natural choice of (α, β) is a minimizer of the right hand side of (6.2), subject to the constraint that $c^\alpha + d^\beta \ge c$, which gives the best upper bound $\mathsf{E}_{\pi_w}[|x|] \le \gamma^{\alpha T} \mathsf{E}_{\pi_w}[w] + B^\alpha$ over all such (α, β). This strong connection between the existence of a Lyapunov function, and the existence of a bound on steady state performance is precisely the principle of duality established in [**KM96**].

If vectors (α, β) exist which satisfy these constraints, then the simultaneous Lyapunov condition of Hordijk is also satisfied [**Hor77**]. Unfortunately, in many examples of interest it is not possible to find a single quadratic Lyapunov function suitable for all policies. One such example is the three buffer example given in Figure 1. For this example, the stability LP of [**KM95**] is not feasible for certain service rates, even though the load condition (3.2) is satisfied.

If the feedback law w is specified, then it is often possible to relax some of the constraints on α. For example, for the three-buffer model under the LBFS policy we have $z_{13} = w_1(x) x_3 = 0$, so that the constraint on c^α is relaxed to

$$c_I^\alpha + d_I^\beta \ge c_I, \qquad (I) \ne (1,3).$$

The stability LP was run for this specific example, maintaining the earlier system parameters (3.5), giving the following Lyapunov function for the model:

$$h_{\alpha_1}(x) = x^T Q_1 x; \qquad Q_1 = \begin{pmatrix} 15.9231 & 9.0000 & 9.0000 \\ 9.0000 & 9.0000 & 0 \\ 9.0000 & 0 & 7.5000 \end{pmatrix} \tag{6.3}$$

Another quadratic which solves (3.4) is

$$h_{\alpha_2}(x) = x^T Q_2 x; \qquad Q_2 = \begin{pmatrix} 55.9895 & 31.6456 & 24.3439 \\ 31.6456 & 31.6456 & 0 \\ 24.3439 & 0 & 20.8858 \end{pmatrix} \tag{6.4}$$

The two matrices are almost multiples of one another, $Q_2 \approx 3.5 Q_1$, and in fact the latter choice gives a correspondingly worse upper bound on $\mathsf{E}[|x|]$ through the inequality (6.2).

Two experiments were performed to compare the performance of the VIA initialized with these two quadratic Lyapunov functions. The results are shown in Figure 7. The speed of convergence is not as fast as what was found using the fluid

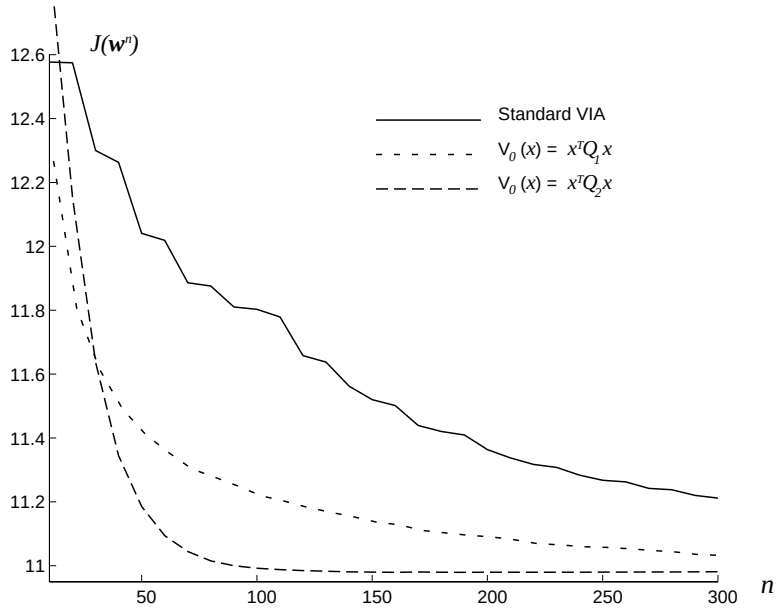

FIGURE 7. Convergence of the VIA with V_0 taken as a quadratic. The quadratic h_{α_1} was found using the stability LP, and the second quadratic h_{α_2} was found through direct calculation. In both cases we see relatively fast convergence of $J(w^n)$ to the optimal cost $\eta_* \approx 10.9$.

model to construct V_0, but we see in Figure 7 that it is significantly faster than the standard algorithm.

7. Conclusions

We have seen that one can obtain excellent steady state performance, even when the time horizon is very short compared with the size of the state space, by adding an appropriate penalty term to the finite horizon cost criterion used in the VIA. For the network scheduling problem two approaches have been described for constructing a useful penalty term. That based upon a fluid model gives the best results in the examples considered, but the simpler approach based upon a quadratic approximation also performs well. Either approach can potentially be used in the online optimization of a large manufacturing system.

For other control problems it will be necessary to have some understanding of the right form for the optimal relative value function in order to initialize the algorithm. One general approach is to apply one step of policy iteration, since the resulting relative value function is a Lyapunov function satisfying (A1) under mild conditions on the process (see [**Mey97b**]).

It is likely that the modified policy iteration algorithm of [**Put94**] can be analyzed using a modification of the proof given in the appendix. We are also considering sample-path based on-line optimization methods using the approaches introduced here.

Appendices

Appendix A. Some stability theory for Markov chains

Here we collect together some general results on Poisson's equation and regularity for a countable state space Markov chain with transition law P and resolvent

$$K = \sum_{i=0}^{\infty} 2^{-(i+1)} P^i.$$

Suppose that c is a function on X with $c \geq 1$. The Markov chain is called c-regular if for some $\boldsymbol{\theta} \in \mathsf{X}$ and every $x \in \mathsf{X}$,

$$\mathsf{E}_x\Big[\sum_{i=0}^{\tau_\theta - 1} c(\Phi(i))\Big] < \infty,$$

where the first entrance time and first return time to the point $\boldsymbol{\theta}$ are defined respectively as $\sigma_\theta = \min(t \geq 0 : \Phi(t) \in \boldsymbol{\theta})$; $\tau_\theta = \min(t \geq 1 : \Phi(t) \in \boldsymbol{\theta})$.

A c-regular chain always possesses a unique invariant probability π such that

$$\pi(c) := \int c(x)\, \pi(dx) < \infty.$$

A set $S \subset \mathsf{X}$ is called *petite* if there is a $\delta > 0$ and $\boldsymbol{\theta} \in \mathsf{X}$ such that

$$K(x, \boldsymbol{\theta}) \geq \delta, \qquad x \in S. \tag{A.1}$$

If the chain is irreducible in the usual sense then every finite set is petite.

The connections between Poisson's equation, the Lyapunov drift (1.3), and regularity are largely based upon the following general result, which is a minor generalization of the Comparison Theorem of [**MT93**].

THEOREM A.1 (Comparison Theorem). *Let $\boldsymbol{\Phi}$ be a Markov chain on X satisfying the drift inequality $PV \leq V - c + s$. The functions s and c take values in $\mathbb{R}_+$, and the function V takes values in $[0, \infty]$ with $V(x_0) < \infty$ for at least one $x_0 \in \mathsf{X}$. Then for any stopping time τ*

$$\mathsf{E}_x[V(\Phi(\tau))] + \mathsf{E}_x\left[\sum_{t=0}^{\tau-1} c(\Phi(t))\right] \quad \leq \quad V(x) + \mathsf{E}_x\left[\sum_{t=0}^{\tau-1} s(\Phi(t))\right].$$

□

The following result is a consequence of the f-Norm Ergodic Theorem of [**MT93**].

THEOREM A.2. *Suppose that the conditions of the Comparison Theorem A.1 hold, $c \geq 1$, s is a constant, and the set $S = \{x : c(x) \leq 2s\}$ is petite. Then $\boldsymbol{\Phi}$ is a c-regular Markov chain with a unique invariant probability π. The function V satisfies $V(x) < \infty$ whenever $\pi(x) > 0$.*

The proof of the following result is identical to that of Theorem 11.3.11 of [**MT93**]. For any probability distribution $\boldsymbol{a}$ on $\mathbb{Z}_+$ define the generalization of the resolvent kernel

$$K_{\boldsymbol{a}}(x, y) = \sum_{n=0}^{\infty} \boldsymbol{a}(n) P^n(x, y), \qquad x, y \in \mathsf{X}.$$

We denote the mean of $\boldsymbol{a}$ by $m(a) = \sum n a(n)$.

LEMMA A.3. *For the Markov chain $\mathbf{\Phi}$ on X, suppose that there is a set $S \subseteq \mathsf{X}$, a point $\boldsymbol{\theta} \in \mathsf{X}$, and a constant $\delta > 0$ such that $K_{\boldsymbol{a}}(x, \boldsymbol{\theta}) \geq \delta$ for $x \in S$. Then for all x,*

$$\mathsf{E}_x \left[\sum_{t=0}^{\tau_\theta - 1} \mathbb{1}_S(\Phi(t)) \right] \leq m(\boldsymbol{a})/\delta.$$

□

Suppose that $\mathbf{\Phi}$ is a c-regular Markov chain with invariant probability π, and denote $\eta = \pi(c)$. The Poisson equation is defined as

$$Ph = h - c + \eta, \tag{A.2}$$

where $h \colon \mathsf{X} \to \mathbb{R}$. The computation of h in the finite state space case involves a simple matrix inversion which can be generalized to the present setting provided that the chain is c-regular.

Given a function $s \colon \mathsf{X} \to [0,1]$, and a probability ν on X, the kernel $s \otimes \nu \colon \mathsf{X} \times \mathsf{X} \to [0,1]$ is defined as the product $s \otimes \nu\,(x,y) = s(x)\nu(y)$, $x, y \in \mathsf{X}$. Letting ν denote the point-mass at $\boldsymbol{\theta}$, and $s = \delta \mathbb{1}_S$, the minorization condition (A.1) may be expressed $K \geq s \otimes \nu$. Letting G denote the kernel

$$G = \sum_{t=0}^{\infty} (K - s \otimes \nu)^t,$$

a solution to Poisson's equation may be explicitly written as

$$h(x) = GK\bar{c}\,(x) = \sum_{i=0}^{\infty} (K - s \otimes \nu)^i K\bar{c}\,(x), \tag{A.3}$$

where $\bar{c}(x) = c(x) - \pi(c)$, provided the sum is absolutely convergent [**GM96, Mey97b**].

The paper [**GM96**] uses these ideas to establish the following sufficient condition for the existence of suitably bounded solutions to Poisson's equation. Define the set S by

$$S = \{x : Kc\,(x) \leq \pi(c)\}. \tag{A.4}$$

If the chain is positive recurrent we have $\pi(S) > 0$.

THEOREM A.4. *Suppose that the Markov chain $\mathbf{\Phi}$ is positive recurrent. Assume further that $\eta = \int c(x)\,\pi(dx) < \infty$, and that the set S defined in (A.4) is petite. Then there exists a solution h to Poisson's equation (A.2) which is finite for every $x \in \mathsf{X}$ satisfying $\pi(x) > 0$, and is bounded from below everywhere:*

$$\inf_{x \in \mathsf{X}} h(x) > -\infty.$$

If $\mathbf{\Phi}$ is also c-regular then h can be taken as (A.3), which satisfies the bound

$$h(x) \leq d_1 \mathsf{E}_x \Big[\sum_{t=0}^{\tau_\theta - 1} c(\Phi(t)) \Big], \qquad x \in \mathsf{X},$$

where d_1 is a finite constant. □

Uniqueness of the solution to Poisson's equation is established in [**Mey97b**] using the previous lower bound.

THEOREM A.5. *Suppose that the Markov chain $\mathbf{\Phi}$ is positive recurrent, that $\eta = \pi(c) < \infty$, and assume that S defined in (A.4) is petite. Let g be finite-valued, bounded from below, and satisfy*

$$Pg \le g - c + \eta.$$

Then $\mathbf{\Phi}$ is c-regular and for some constant b,

(i): $g(x) = GK\bar{c}\,(x) + b$ *for almost every* $x \in \mathsf{X}$ $[\pi]$.
(ii): $g(x) \ge GK\bar{c}\,(x) + b$ *for every* $x \in \mathsf{X}$.

□

Closely related is the following

LEMMA A.6. *Suppose that $\mathbf{\Phi}$ is c-regular with invariant probability π, and suppose that $z\colon \mathsf{X} \to \mathbb{R}$ is bounded from below, and is superharmonic: $Pz \le z$. Then*

(i): $z(x) = \pi(z)$ *for almost every* $x \in \mathsf{X}$ $[\pi]$.
(ii): $z(x) \ge \pi(z)$ *for every* $x \in \mathsf{X}$.

□

Appendix B. Convergence of the value iteration algorithm

We present here proofs of our main results. Throughout this section we assume that (A1)–(A3) are satisfied, even when this is not mentioned explicitly.

Central to the analysis is the incremental cost g_n and function h_n defined in (2.3). In the standard version of the VIA where $V_0 = 0$, the functions $\{g_n : n \in \mathbb{Z}_+\}$ are positive-valued for each n, but may be unbounded. In the present case we find that the opposite situation arises. When V_0 is a Lyapunov function for some policy, the functions $\{g_n\}$ are strictly bounded from above, but may be unbounded from below. This is a desirable situation since an upper bound on the sequence $\{g_n : n \in \mathbb{Z}_+\}$ permits us to conclude that the each of the stationary policies $\{\boldsymbol{w}^n : n \in \mathbb{Z}_+\}$ is regular. These results are summarized in Proposition B.3. We first require the following two lemmas.

LEMMA B.1. *Suppose that for the state feedback law w there exists a solution $V\colon \mathsf{X} \to \mathbb{R}_+$ to the inequality*

$$P_w V\,(x) \le V(x) - c_w(x) + \overline{\eta}, \qquad x \in \mathsf{X}. \tag{B.1}$$

Then the controlled chain has the following properties:

(i): *The feedback law is regular, and hence the controlled chain has a unique invariant probability π_w.*
(ii): *There exists a constant B_1 depending only on $\overline{\eta}$ and δ such that*

$$\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta - 1} c_w(\Phi(t))\Big] \le B_1(V(x) + 1) \qquad x \in \mathsf{X}.$$

(iii): $\eta_w = \pi_w(c_w) \le \overline{\eta}$.
(iv): $\pi_w(\boldsymbol{\theta}) \ge \delta \pi_w(S_0) > 0$.
(v): $V(x) - V(\boldsymbol{\theta}) \ge -\overline{\eta}/\delta, \qquad x \in \mathsf{X}$.

PROOF. From (B.1) and the definition of S_0 we obtain the inequality

$$P_w V\,(x) \le V(x) - \tfrac{1}{2} c_w(x) + \overline{\eta}\mathbb{1}_{S_0}(x).$$

Applying the Comparison Theorem A.1 then gives

$$\frac{1}{2}\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta-1} c_w(\Phi(t))\Big] \le V(x) - V(\boldsymbol{\theta}) + \overline{\eta}\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta-1} \mathbb{1}_{S_0}(\Phi(t))\Big]. \tag{B.2}$$

From (A3) the minorization condition (2.6) holds for K_w:

$$K_w(x,\boldsymbol{\theta}) \ge \delta \mathbb{1}_{S_0}(x) \qquad x \in \mathsf{X}. \tag{B.3}$$

Applying Lemma A.3 then gives

$$\mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta-1} c_w(\Phi(t))\Big] \le 2V(x) + 2\overline{\eta}/\delta.$$

This proves (ii) with $B_1 = 2 + 2\overline{\eta}/\delta$. Result (i) and (iii) follow immediately from (ii) and the Comparison Theorem.

To prove (iv) observe that $\pi_w(\bar{c}) \le \eta_w \le \overline{\eta}$. Hence the sublevel set S_0 must have positive π_w-measure. From the inequality (B.3) we can invoke invariance to deduce (iv).

Finally, (v) also follows from (B.2) and Lemma A.3:

$$0 \le V(x) - V(\boldsymbol{\theta}) + \mathsf{E}_x^{\boldsymbol{w}}\Big[\sum_{t=0}^{\tau_\theta-1} \mathbb{1}_{S_0}(\Phi(t))\Big] \le V(x) - V(\boldsymbol{\theta}) + \overline{\eta}/\delta.$$

□

Let $\overline{\eta}_n = \sup_{x\in\mathsf{X}} g_n(x)$ and $\underline{\eta}_n = \inf_{x\in\mathsf{X}} g_n(x)$.

LEMMA B.2. *For each $n \in \mathbb{Z}_+$,*

(i): $P_{n+1}g_n(x) \le g_{n+1}(x)$
(ii): $\underline{\eta}_n \le \underline{\eta}_{n+1}$
(iii): $g_{n+1}(x) \le P_n g_n(x)$
(iv): $\overline{\eta}_{n+1} \le \overline{\eta}_n \le \overline{\eta}$

PROOF. Result (i) follows from the bound $V_{n+1} = P_nV_n + c_n \le P_{n+1}V_n + c_{n+1}$, as shown here:

$$\begin{aligned} g_{n+1} = V_{n+2} - V_{n+1} &\ge V_{n+2} - (P_{n+1}V_n + c_{n+1}) \\ &= P_{n+1}V_{n+1} + c_{n+1} - P_{n+1}V_n - c_{n+1} \\ &= P_{n+1}g_n. \end{aligned}$$

To prove (ii), we apply (i) and the definition of $\underline{\eta}_n$:

$$\underline{\eta}_{n+1} = \inf_{x\in\mathsf{X}} g_{n+1}(x) \ge \inf_{x\in\mathsf{X}} P_{n+1}g_n(x) \ge \inf_{y\in\mathsf{X}} g_n(y) = \underline{\eta}_n.$$

We now prove (iii). First observe that

$$P_nV_{n+1} = P_n(V_n + g_n) = P_nV_n + P_ng_n = V_{n+1} - c_n + P_ng_n.$$

From the definition of $\{V_n\}$ we then have

$$V_{n+2} = P_{n+1}V_{n+1} + c_{n+1} \le P_nV_{n+1} + c_n = V_{n+1} + P_ng_n,$$

from which the result follows. Result (iv) then follows immediately, as in (ii). □

We may now establish the desired stability properties of the VIA under (A1)-(A3).

PROPOSITION B.3. The policy $\boldsymbol{w}^n$ satisfies, for each n,

(i): The following identity holds for all x:

$$P_n V_n(x) = V_n(x) - c_n(x) + g_n(x). \tag{B.4}$$

(ii): The sequence $\{g_n : n \in \mathbb{Z}_+\}$ is uniformly bounded from above:

$$g_n(x) \le \overline{\eta}, \qquad x \in \mathsf{X},\ n \in \mathbb{Z}_+. \tag{B.5}$$

(iii): The chain $\mathbf{\Phi}^n$ is c_n-regular, and there exists a constant depending only on δ and $\overline{\eta}$ such that for each n,

$$\mathsf{E}_x^n\Big[\sum_{t=0}^{\tau_\theta - 1} c_n(\Phi(t))\Big] \le B_1(V_n(x) + 1), \qquad x \in \mathsf{X}.$$

(iv): The stationary policy $\boldsymbol{w}^n$ is regular with unique invariant probability π_n, and the invariant probability satisfies

$$J(w^n) = \pi_n(c_n) \le \overline{\eta}_n.$$

PROOF. Result (i) is essentially the definition of V_n, g_n: For each n,

$$P_n V_n = V_{n+1} - c_n = V_n - c_n + (V_{n+1} - V_n) = V_n - c_n + g_n.$$

Result (ii) follows from Lemma B.2, and (iii) directly from Lemma B.1. Result (iv) follows from (ii), (iii), and the Comparison Theorem applied to (B.4). □

An application of this proposition and Lemma B.1 gives a lower bound on the sequences $\{g_n\}$, $\{h_n\}$:

LEMMA B.4. *For all $n \in \mathbb{Z}_+$,*

$$g_n(\boldsymbol{\theta}) \ge -(\overline{\eta}/\delta) \qquad h_n(x) \ge -(\overline{\eta}/\delta),\ x \in \mathsf{X}.$$

PROOF. The lower bound on h_n follows immediately from Lemma B.1 and Proposition B.3. We then have,

$$-(\overline{\eta}/\delta) \le P_n h_n(\boldsymbol{\theta}) = h_n(\boldsymbol{\theta}) - c_n(\boldsymbol{\theta}) + g_n(\boldsymbol{\theta}) \le g_n(\boldsymbol{\theta}).$$

□

These bounds can now be used to establish a uniform *upper* bound on $\{h_n\}$.

LEMMA B.5. *There is a finite constant B_2, independent of n, k or x, such that*

$$h_n(x) \quad \le \quad B_2(V_k(x) + 1), \qquad 0 \le k \le n, x \in \mathsf{X}.$$

PROOF. It is enough to prove the result for $k = 0$ since we may treat the kth step of the algorithm as a new starting point. We have from minimality of V_n, for any $n \in \mathbb{Z}_+$,

$$\begin{aligned} V_n(x) \quad &\le \quad \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + V_0(\Phi(n))\right) \mathbb{1}(\tau_\theta > n)\right] \\ &+ \quad \mathsf{E}_x^0\left[\left(\sum_{t=0}^{\tau_\theta - 1} c_0(\Phi(t)) + V_{n-\tau_\theta}(\boldsymbol{\theta})\right) \mathbb{1}(\tau_\theta \le n)\right], \end{aligned}$$

where E^0 is the expectation operator obtained with the policy $\boldsymbol{w}^0$. Subtracting $V_n(\boldsymbol{\theta})$ from both sides then gives

$$\begin{aligned} h_n(x) &\leq \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + V_0(\Phi(n))\right) \mathbb{1}(\tau_\theta > n)\right] \\ &+ \mathsf{E}_x^0\left[\sum_{t=0}^{\tau_\theta - 1} c_0(\Phi(t))\right] \\ &+ \mathsf{E}_x^0\left[(V_{n-\tau_\theta}(\boldsymbol{\theta}) - V_n(\boldsymbol{\theta}))\, \mathbb{1}(\tau_\theta \leq n)\right]. \end{aligned} \tag{B.6}$$

We now proceed to bound each of these terms. First, letting $L_n(x)$ denote the first term on the right hand side of (B.6), we have

$$\begin{aligned} L_n(x) &:= \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + V_0(\Phi(n))\right) \mathbb{1}(\tau_\theta > n)\right] \\ &= \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + \mathsf{E}^0[V_0(\Phi(n)) \mid \Phi(0), \ldots, \Phi(n-1)]\right) \mathbb{1}(\tau_\theta > n)\right] \\ &\leq \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + (V_0(\Phi(n-1)) - c_0(\Phi(n-1)) + \overline{\eta})\right) \mathbb{1}(\tau_\theta > n)\right] \\ &\leq L_{n-1}(x) + \overline{\eta}\mathsf{P}^{w^0}\{\tau_\theta > n \mid \Phi_0 = x\}. \end{aligned}$$

Hence by iteration we have for all n and x,

$$\begin{aligned} L_n(x) &\leq L_0(x) + \overline{\eta}\mathsf{E}_x^0[\tau_\theta] \\ &= V_0(x) + \overline{\eta}\mathsf{E}_x^0[\tau_\theta]. \end{aligned}$$

Proposition B.3 (iii) combined with this inequality then gives $L_n(x) \leq V_0(x) + B_1(V_0(x) + 1)$.

The second term in (B.6) is also bounded using Proposition B.3 (iii):

$$\mathsf{E}_x^0\left[\sum_{t=0}^{\tau_\theta - 1} c_0(\Phi(t))\right] \leq B_1(V_0(x) + 1).$$

To bound the final term, note that by Lemma B.4, for any $n > \tau_\theta$,

$$V_{n-\tau_\theta}(\boldsymbol{\theta}) - V_n(\boldsymbol{\theta}) = -\sum_{k=n-\tau_\theta}^{n-1} g_k(\boldsymbol{\theta}) \leq (\overline{\eta}/\delta)\tau_\theta.$$

The third term on the right hand side of (B.6) is thus again bounded by Proposition B.3 (iii):

$$\mathsf{E}_x^0\left[(V_{n-\tau_\theta}(\boldsymbol{\theta}) - V_n(\boldsymbol{\theta}))\, \mathbb{1}(\tau_\theta \leq n)\right] \leq (\overline{\eta}/\delta)\mathsf{E}_x^0[\tau_\theta] \leq (\overline{\eta}/\delta)B_1(V_0(x) + 1).$$

Thus each of the expectations on the right hand side of (B.6) is bounded as desired, with $B_2 = 1 + (2 + (\overline{\eta}/\delta))B_1$. □

From the optimality equations we have for all $n \in \mathbb{Z}_+$ and $x \in \mathsf{X}$,

$$P_n h_n(x) = h_{n+1}(x) - c_n(x) + g_n(\boldsymbol{\theta}).$$

This identity together with the bounds already obtained on the sequence $\{h_n\}$ are precisely what is needed to deduce a strong form of stability for the time-inhomogeneous chain $\boldsymbol{\Phi}^{\boldsymbol{v}^n}$.

LEMMA B.6. *There is a constant B_3 dependent only on $\overline{\eta}$ and δ such that for all x and n,*

$$\mathsf{E}_x^{\boldsymbol{v}^{n+1}}[n \wedge \tau_\theta] \le B_3(V_0(x)+1).$$

PROOF. For n fixed denote

$$M(t) = h_{n+1-t}(\Phi(t)) - t\overline{\eta} + \sum_{i=0}^{t-1} \underline{c}(\Phi(i)), \qquad 0 \le t \le n+1.$$

We show here that $(M(t), \mathcal{F}_t)$ is a supermartingale, with $\mathcal{F}_t = \sigma(\Phi_0, \dots, \Phi_t)$, $t \ge 0$. For each $0 \le t \le n$,

$$\begin{aligned}
\mathsf{E}^{\boldsymbol{v}^{n+1}}[M(t+1) \mid \mathcal{F}_t] &= \mathsf{E}^{\boldsymbol{v}^{n+1}}[h_{n-t}(\Phi(t+1)) \mid \mathcal{F}_t] - (t+1)\overline{\eta} + \sum_{i=0}^{t} \underline{c}(\Phi(i)) \\
&= P_{n-t} h_{n-t}(\Phi(t)) - (t+1)\overline{\eta} + \sum_{i=0}^{t} \underline{c}(\Phi(i)) \\
&= h_{n-t+1}(\Phi(t)) - c_{n-t}(\Phi(t)) + g_{n-t}(\boldsymbol{\theta}) - (t+1)\overline{\eta} + \sum_{i=0}^{t} \underline{c}(\Phi(i)) \\
&\le M(t),
\end{aligned}$$

where we have used the bounds $g_t \le \overline{\eta}$, $c_w \ge \underline{c}$. This establishes the supermartingale property. Now let $\tau = \tau_\theta \wedge n$, and apply the optional stopping theorem to obtain the bound

$$\mathsf{E}_x^{\boldsymbol{v}^{n+1}}\Big[h_{n+1-\tau}(\Phi(\tau)) + \sum_{i=0}^{\tau-1} (\underline{c}(\Phi(i)) - \overline{\eta})\Big] = \mathsf{E}_x^{\boldsymbol{v}^{n+1}}[M(\tau)] \le M(0) \le h_{n+1}(x) \le B_2(V_0(x)+1).$$

Since we also have $h_k(x) \ge -\overline{\eta}/\delta$ for all x and k, it follows that

$$\mathsf{E}_x^{\boldsymbol{v}^{n+1}}\Big[\sum_{i=0}^{\tau-1} (\underline{c}(\Phi(i)) - \overline{\eta})\Big] \le B_2(V_0(x)+1) + \overline{\eta}/\delta.$$

Using the definition of S_0 we then obtain

$$\frac{1}{2}\mathsf{E}_x^{\boldsymbol{v}^{n+1}}\Big[\sum_{i=0}^{\tau-1} \underline{c}(\Phi(i))\Big] \le B_2(V_0(x)+1) + \overline{\eta}/\delta + \overline{\eta}\mathsf{E}_x^{\boldsymbol{v}^{n+1}}\Big[\sum_{i=0}^{\tau-1} \mathbb{1}_{S_0}(\Phi(i))\Big].$$

Exactly as in the proof of Lemma A.3 given in Theorem 11.3.11 of [**MT93**] we may deduce via Assumption (A3) that $\mathsf{E}_x^{\boldsymbol{v}^{n+1}}[\sum_{i=0}^{\tau-1} \mathbb{1}_{S_0}(\Phi(i))] \le 1/\delta$. The lemma then follows with $B_3 = 2(B_2 + 2\overline{\eta}/\delta)$. □

For $x \in \mathsf{X}$ let $\bar{g}(x) = \limsup_{n\to\infty} g_n(x)$, and $\underline{g}(x) = \liminf_{n\to\infty} g_n(x)$.

LEMMA B.7.

$$\bar{g}(x) \le \bar{g}(\boldsymbol{\theta}), \qquad x \in \mathsf{X}.$$

PROOF. Let $m(t) = g_{n-t+1}(\Phi^{\boldsymbol{v}^{n+1}}(t))$. The adapted process $(m(t), \mathcal{F}_t)$ is a submartingale since by Lemma B.1,

$$\mathsf{E}^{\boldsymbol{v}^{n+1}}[m(t+1) \mid \mathcal{F}_t] = P_{n-t} g_{n-t}(\Phi(t)) \ge g_{n-t+1}(\Phi(t)) = m(t).$$

From the optional stopping theorem with $\tau = \tau_\theta \wedge n$ we have $\mathsf{E}^{\boldsymbol{v}^{n+1}}[m(\tau)] \geq m(0)$, or

$$\mathsf{E}^{\boldsymbol{v}^{n+1}}[g_{n-\tau+1}(\Phi(\tau))] \geq g_{n+1}(x).$$

For any k define $\bar{g}_k(x) = \sup_{i \geq k} g_i(x)$, so that $\bar{g}_k(x) \to \bar{g}(x)$ as $k \to \infty$. Letting s_n denote the integer part of $n/2$ we then have from the previous bound

$$\begin{aligned} g_{n+1}(x) &\leq \mathsf{E}^{\boldsymbol{v}^{n+1}}[\bar{g}_{s_n}(\Phi(\tau))\mathbb{1}(\tau_\theta < s_n)] + \overline{\eta}\mathsf{E}^{\boldsymbol{v}^{n+1}}[\mathbb{1}(\tau_\theta \geq s_n)] \\ &= \bar{g}_{s_n}(\boldsymbol{\theta})\mathsf{P}_x^{\boldsymbol{v}^{n+1}}(\tau_\theta < s_n) + \overline{\eta}\mathsf{P}_x^{\boldsymbol{v}^{n+1}}(\tau_\theta \geq s_n). \end{aligned}$$

Since $\mathsf{P}_x^{\boldsymbol{v}^{n+1}}(\tau_\theta \geq s_n) \leq (2/n)\mathsf{E}_x^{\boldsymbol{v}^{n+1}}[\tau] \leq (2/n)B_3(V_0(x)+1)$, we may take limit supremums of both sides with respect to n to obtain $\bar{g}(x) \leq \bar{g}(\boldsymbol{\theta})$. □

LEMMA B.8. *If $g_{n_i}(\boldsymbol{\theta}) \to \bar{g}(\boldsymbol{\theta})$, $i \to \infty$, then for any integer t,*

$$g_{n_i - t}(\boldsymbol{\theta}) \to \bar{g}(\boldsymbol{\theta}), \qquad i \to \infty.$$

PROOF. The proof is similar to an argument given in [**Cav96**]. However in the present setting we do not know if the sequence of functions $\{g_n\}$ is bounded from below.

It is enough to prove the result for $t = 1$. By taking a further subsequence if necessary we may assume that there is a kernel P and a function g such that $P_{n_i-1}(x,y) \to P(x,y)$ and $g_{n_i-1}(x) \to g(x)$ as $i \to \infty$ pointwise. The kernel P is substochastic: $P(x,\mathsf{X}) \leq 1$, $x \in \mathsf{X}$. Using the inequality $P_{n_i-1}g_{n_i-1}(\boldsymbol{\theta}) \geq g_{n_i}(\boldsymbol{\theta})$ and Fatou's Lemma then gives

$$\begin{aligned} \bar{g}(\boldsymbol{\theta}) &\leq \limsup_{i\to\infty} \sum_{y\in\mathsf{X}} P_{n_i-1}(\boldsymbol{\theta}, y) g_{n_i-1}(y) \\ &\leq \sum_{y\in\mathsf{X}} \limsup_{i\to\infty} P_{n_i-1}(\boldsymbol{\theta}, y) g_{n_i-1}(y) \\ &\leq Pg\,(\boldsymbol{\theta}) \leq P(\boldsymbol{\theta},\mathsf{X})\bar{g}(\boldsymbol{\theta}), \end{aligned}$$

where in the last inequality we are using the fact that $\boldsymbol{\theta}$ is maximal. Fatou's lemma is applicable because $\{g_n\}$ is uniformly bounded from above. It follows that $P(\boldsymbol{\theta},\mathsf{X}) = 1$ and that $g(y) = \bar{g}(\boldsymbol{\theta})$ for every $y \in \mathsf{X}$ for which $P(\boldsymbol{\theta}, y) > 0$. Since $P(\boldsymbol{\theta},\boldsymbol{\theta}) > \delta$ by assumption, we conclude that $g(\boldsymbol{\theta}) = \bar{g}(\boldsymbol{\theta})$. Since $g(\boldsymbol{\theta})$ is an arbitrary limit point of the sequence $\{g_{n_i-1}(\boldsymbol{\theta}) : i \geq 0\}$ the conclusion of the lemma follows. □

LEMMA B.9. **(i):** *$\bar{g}(x) \leq \eta^*$ for every $x \in \mathsf{X}$.*
(ii): *$\lim_{n\to\infty} g_n(\boldsymbol{\theta}) = \eta^*$.*

PROOF. We first prove (i). From the previous lemma it is enough to show that $\bar{g}(\boldsymbol{\theta}) \leq \eta^*$. We show that there exists a sequence of functions $\{W_t : t \geq 0\}$ from X to $\mathbb{R}_+$ such that for some $B_4 < \infty$,

$$W_t(x) \leq B_4(V_0(x)+1), \qquad x \in \mathsf{X}, t \in \mathbb{Z}_+; \tag{B.7}$$

$$P_{w^*}W_t\,(x) \geq W_{t-1}(x) - c_{w^*}(x) + \bar{g}(\boldsymbol{\theta}), \qquad x \in \mathsf{X}, t \in \mathbb{Z}_+. \tag{B.8}$$

Given these bounds, we then have by iteration,

$$B_4 + B_4 P_{w^*}^n V_0\,(x) \geq W_0(x) - \sum_{t=0}^{n-1} P_{w^*}^t c_{w^*}\,(x) + n\bar{g}(\boldsymbol{\theta}).$$

Dividing by n and letting $n \to \infty$ then shows that

$$\bar{g}(\boldsymbol{\theta}) \le \lim_{n\to\infty} (1/n) \sum_{t=0}^{n-1} P_{w^*}^t c_{w^*}(x) = \eta^*,$$

as claimed.

To prove that such a sequence exists, first consider the inequality $P_{w^*} h_{n_i - t}(x) \ge h_{n_i - t + 1} - c_{w^*}(x) + g_{n_i - t}(\boldsymbol{\theta})$. Letting $W_t^{(i)}(x) = \overline{\eta} + h_{n_i - t}(x)$, $x \in \mathsf{X}$, we obtain for each i,

$$P_{w^*} W_t^{(i)}(x) \ge W_{t-1}^{(i)}(x) - c_{w^*}(x) + g_{n_i - t}(\boldsymbol{\theta}), \qquad x \in \mathsf{X}, t \in \mathbb{Z}_+.$$

Assume that $\{n_i\}$ is chosen so that $g_{n_i}(\boldsymbol{\theta}) \to \bar{g}(\boldsymbol{\theta})$ as $i \to \infty$. Then by choosing a subsequence if necessary we may find functions $\{W_t\}$ with $W_t^{(i)} \to W_t$ pointwise as $t \to \infty$. Since the functions $\{h_t\}$ are bounded as desired, the inequalities (B.7), (B.8) then follow from Lemma B.8 and the Dominated Convergence Theorem.

To prove (ii), consider any limit point $g(\boldsymbol{\theta})$ of the sequence $\{g_n(\boldsymbol{\theta}\}$. We can assume without loss of generality that there are functions g, h on X, a feedback law w, and that there is a subsequence $\{m_i\}$ of $\mathbb{Z}_+$ with $g_{m_i}(x) \to g(x)$, $h_{m_i}(x) \to h(x)$, $w_{m_i}(x) \to w(x)$, $i \to \infty$, for all $x \in \mathsf{X}$. From Fatou's Lemma we then have $P_w h \le h - c_w + g$, and from the Comparison Theorem A.1 we then have $\pi_w(c_w) \le \pi(g) \le \eta^*$, where the last inequality follows from (i). Since $\pi_w(c_w) \ge \eta^*$ by optimality, it then follows that $g(x) = \eta^*$ for a.e. $x \in \mathsf{X}$ $[\pi_w]$. Lemma B.1 (iv) completes the proof. □

LEMMA B.10. *Under (A1)-(A4) the solution h_* to the ACOE is unique up to an additive constant over all solutions which are bounded from below.*

PROOF. To begin we note that under (A1)-(A3) there is a minimal relative value function given by

$$h_{\min}(x) := \min_{\boldsymbol{w}} \mathsf{E}_x^{\boldsymbol{w}} \Big[\sum_{k=0}^{\tau_\theta - 1} [c(\Phi_k, w_k(\Phi_k)) - \eta_*] \Big]$$

where the minimum is over all Markov policies. As in [**BM98**] we may show that $h_{\min}$ solves the optimality equation

$$\min_{a \in \mathcal{A}(x)} [P_a h_{\min}(x) + c(x,a)] \le h_{\min} + \eta_*.$$

It may be shown that $h_{\min}$ is bounded from below as in Lemma B.1. By Lemma A.5 it then follows that this must be an equality. Let $w_{\min}$ be any optimizing policy in the minimization above so that

$$P_{w_{\min}} h_{\min} = h_{\min} - c_{w_{\min}} + \eta_*.$$

Note that the feedback law $w_{\min}$ must be regular.

If h is *any* solution to the ACOE for which $\inf_{x \in \mathsf{X}} h(x) > -\infty$ then by Lemma A.5 we have for some regular policy $\boldsymbol{w}$,

$$h(x) - h(\theta) = \mathsf{E}_x^{\boldsymbol{w}} \Big[\sum_{k=0}^{\tau_\theta - 1} [c_w(\Phi_k) - \eta_*] \Big]$$

By minimality of $h_{\min}$ it then follows that the function s defined by $s(x) = h(x) - h(\boldsymbol{\theta}) - h_{\min}(x)$, $x \in \mathsf{X}$, is positive-valued. Moreover we have $P_w s \le s$, so by Lemma A.6 the function s must be constant. □

LEMMA B.11. *Under (A1)-(A4),*

$$h_n(x) \to h_*(x) - h_*(\boldsymbol{\theta}), \qquad \text{as } n \to \infty.$$

PROOF. Let h be any pointwise limit of the $\{h_n\}$. The function h is finite valued by Lemma B.5. Then using Fatou's lemma we may find a limiting feedback law w such that $P_w h \leq h - c_w + \eta_*$. By Theorem A.5 and (A4) it follows that this is an equality

$$P_w h = h - c_w + \eta_*. \tag{B.9}$$

Thus by the previous lemma we have $h(x) = h_*(x) - h_*(\boldsymbol{\theta})$. □

References

[ABF+93] A. Arapostathis, V. S. Borkar, E. Fernandez-Gaucherand, M. K. Ghosh, and S. I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: a survey*, SIAM J. Control Optim. **31** (1993), 282–344.

[Bor91] V. S. Borkar, *Topics in controlled Markov chains*, Pitman Research Notes in Mathematics Series # 240, Longman Scientific & Technical, UK, 1991.

[BM98] V.S. Borkar and S.P. Meyn, *Risk sensitive optimal control: Existence and synthesis for models with unbounded cost.*, SIAM J. Control and Optim., 1998. Submitted for publication.

[Cav96] R. Cavazos-Cadena, *Value iteration in a class of communicating Markov decision chains with the average cost criterion*, Technical report, Universidad Autónoma Agraria Anonio Narro, 1996.

[CF95] R. Cavazos-Cadena and E. Fernandez-Gaucherand, *Value iteration in a class of average controlled Markov chains with unbounded costs: Necessary and sufficient conditions for pointwise convergence*, Proceedings of the 34th IEEE Conference on Decision and Control (New Orleans, LA), 1995, pp. 2283–2288.

[CM91] H. Chen and A. Mandelbaum, *Discrete flow networks: Bottlenecks analysis and fluid approximations*, Mathematics of Operations Research **16** (1991), 408–446.

[CZ96] H. Chen and H. Zhang, *Stability of multiclass queueing networks under priority service disciplines*, Technical Note, 1996.

[Dai95] J. G. Dai, *On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid limit models*, Ann. Appl. Probab. **5** (1995), 49–77.

[DBT92] I. Ch. Paschalidis D. Bertsimas and J. N. Tsitsiklis, *Scheduling of multiclass queueing networks: Bounds on achievable performance*, Workshop on Hierarchical Control for Real–Time Scheduling of Manufacturing Systems (Lincoln, New Hampshire), October 16–18, 1992.

[DEM96] J. Humphrey D. Eng and S.P. Meyn, *Fluid network models: Linear programs for control and performance bounds*, Proceedings of the 13th IFAC World Congress (San Francisco, California) (J. Cruz J. Gertler and M. Peshkin, eds.), vol. B, 1996, pp. 19–24.

[DM95] J. G. Dai and S.P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Trans. Automat. Control **40** (1995), 1889–1904.

[GM96] P. W. Glynn and S. P. Meyn, *A Lyapunov bound for solutions of Poisson's equation*, Ann. Probab. **24** (1996).

[Har96] J.M. Harrison, The BIGSTEP approach to flow management in stochastic processing networks, Stochastic Networks Theory and Applications, 57–89, Stochastic Networks Theory and Applications, Clarendon Press, Oxford, UK, 1996, pp. 57–89, F.P. Kelly, S. Zachary, and I. Ziedins (ed.).

[Hor77] A. Hordijk, *Dynamic programming and Markov potential theory*, 1977.

[HW89] J. M. Harrison and L. M. Wein, *Scheduling networks of queues: Heavy traffic analysis of a simple open network*, QUESTA **5** (1989), 265–280.

[KK94] S. Kumar and P. R. Kumar, *Performance bounds for queueing networks and scheduling policies*, IEEE Trans. Automat. Control **AC-39** (1994), 1600–1611.

[KM95] P. R. Kumar and S. P. Meyn, *Stability of queueing networks and scheduling policies*, IEEE Transactions on Automatic Control **40** (1995), no. 2, 251–260.

[KM96] P.R. Kumar and S.P. Meyn, *Duality and linear programs for stability and performance analysis queueing networks and scheduling policies*, IEEE Transactions on Automatic Control **41** (1996), no. 1, 4–17.

[KS90] P. R. Kumar and T. I. Seidman, *Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems*, IEEE Trans. Automat. Control **AC-35** (1990), no. 3, 289–298.

[Mey96] S.P. Meyn, *The policy improvement algorithm: General theory with applications to queueing networks and their fluid models*, 35th IEEE Conference on Decision and Control, Kobe, Japan, December 1996.

[Mey97] S.P. Meyn, *Stability and optimization of multiclass queueing networks and their fluid models*, proceedings of the summer seminar on "The Mathematics of Stochastic Manufacturing Systems", American Mathematical Society, 1997.

[Mey97b] S.P. Meyn, *The policy improvement algorithm for Markov decision processes with general state space*, IEEE Trans. Auto. Control, 1997.

[MSS96] L. F. Martins, S. E. Shreve, and H. M. Soner, *Heavy traffice convergence of a controlled, multiclass queueing system*, SIAM J. Control and Optimization **34** (1996), no. 6, 2133–2171.

[MT93] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, London, 1993.

[Put94] M. L. Puterman, *Markov Decision Processes*, Wiley, New York, 1994.

[RS92] A. N. Rybko and A. L. Stolyar, *On the ergodicity of stohastic processes describing the operation of open queueing networks*, Problemy Peredachi Informatsii **28** (1992), 3–26.

[Sen86] L. I. Sennott, *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Operations Research Letters **5** (1986), 17–23.

[Sen96] L.I. Sennott, *The convergence of value iteration in average cost Markov decision chains*, Operations Research Letters **19** (1996), 11–16.

[Wei95] G. Weiss. Optimal draining of a fluid re-entrant line. In *Stochastic Networks*, volume 71 of *IMA volumes in Mathematics and its Applications*, pages 91–103. Springer-Verlag, N.Y., 1995.

Rong-Rong Chen, University of Illinois, Department of Mathematics, 1409 W. Green St., Urbana, IL 61801

Sean Meyn, Coordinated Science Laboratory and the University of Illinois, 1308 W. Main Street, Urbana, IL 61801, URL `http://black.csl.uiuc.edu:80/~meyn`

Current address: July 15, 1997 – March 15, 1998. Department of Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India