Sādhanā, Vol. 24, Parts 4 & 5, August & October 1999, pp. 339–367. © Printed in India

Algorithms for optimization and stabilization of controlled Markov chains

SEAN MEYN

Coordinated Science Laboratory and The University of Illinois, 1308 W. Main Street, Urbana, IL 61801, USA
e-mail address: s-meyn@uiuc.edu

Abstract. This article reviews some recent results by the author on the optimal control of Markov chains. Two common algorithms for the construction of optimal policies are considered: value iteration and policy iteration.

In either case, it is found that the following hold when the algorithm is properly initialized:

(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.

The network scheduling problem is considered in some detail as both an illustration of the theory, and because of the strong conclusions which can be reached for this important example as an application of the general theory.

Keywords. Multiclass queueing networks; Markov decision processes; optimal control; dynamic programming.

1. Introduction

1.1 *Feedback*

The subject of this article concerns the synthesis of feedback laws for controlled, noisy, nonlinear systems. For any sort of controlled system, such as an automobile steered by a human driver, or the cruise control system on the same vehicle, some sort of feedback will certainly play a prominent role. In this automotive example, the most significant example of feedback involves the driver's eyes which hopefully are following the road. The car's cruise control uses speed measurements to attempt to regulate the actual speed of the vehicle. Even in a modern oven there is a thermosensor which sends measurements to a device whose job is to regulate the temperature to ensure a well-baked cake.

In a control system as simple as the cruise control, one must still be careful in the way that measurements are used in the adjustment of the gas flow to the engine to maintain a constant speed. In aerospace applications such control design requires even greater thought. The design of effective "feedback laws" to utilize available measurements is the subject of

automatic control; an active field in the mathematics and engineering communities for over fifty years.

The control systems considered in this paper may be modelled by the nonlinear state space model

$$\Phi(t+1) = F(\Phi(t), a(t), I(t+1)), \quad t \geq 0, \tag{1}$$

where time is discrete. The sequence $\Phi := \{\Phi(t) : t \geq 0\}$ denotes a *state process* for the system to be controlled, and $a := \{a(t) : t \geq 0\}$ is called the *control sequence*, or *action sequence*, or the *policy*. We assume that controllers are not clairvoyant so that, for any t, the action $a(t)$ depends only upon the variable $\Phi(s)$ for $s \leq t$. The state process takes values in a *state space* X, and the control sequence takes values in an *action space*, denoted $\mathcal{A}$.

In the example of an automobile and driver the variable $\Phi(t)$ will denote the position, velocity etc. for the car to be controlled. The actions $a(t)$ which the driver implements will of course depend upon these variables. The function F describes the dynamics of the system, and the sequence $\{I(t) : t \geq 0\}$ represents noise which is beyond the control of the driver, such as gusts of wind or oil on the road. If this noise is modelled as in i.i.d. sequence (the noise variables $\{I(t) : t \geq 0\}$ are independent, with identical distributions), then the model (1) is known as a *controlled Markov chain, or Markov decision process (MDP).*

The motivation for considering such systems is largely two-fold. First of all, almost any man-made or natural system can be at least approximately modelled by a MDP. Moreover, once one has decided on such a model, there is the hope that simple and effective feedback laws may be computed. We are particularly interested in finding a useful action sequence of the stationary feedback form where $a(t) = w(\Phi(t))$: the *feedback law* w is a static function from the set of states X to the set of actions $\mathcal{A}$

We are interested in control problems of the 'process control' type (such as the cruise control, or the oven). A more relevant example may be found in the area of manufacturing, where today one finds assembly production lines of high complexity. There may be many process-control problems within plants in a facility. However, because it is such an interesting example and because the specialization of the results described here are so striking in this context, in this paper we will consider in some detail the 'meta' control problem of scheduling parts in the system to minimize inventory, and minimize production delay. The control goal is one of regulation, and such a control problem may be conveniently modelled as an MDP.

To show that this scheduling problem fits into our framework, consider a network of the form illustrated in figure 1, composed of d single-server machines, indexed by $\sigma = 1, \ldots, d$. The network is populated by K classes of "customers" (which may in fact be partially assembled parts). After completion of service at a machine, the customer changes class and enters a machine, or leaves the network. It is assumed that there is an exogenous stream of customers of class 1 which arrive to machine $s(1)$. A customer of class k waits in *buffer* K, until it receives service at machine $s(k)$. 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,

$$\Phi(t+1) = \Phi(t) + \sum_{k=0}^{K} I_k(t+1)[e^{k+1} - e^k] w_k(\Phi(t)), \tag{2}$$

where the state process Φ evolves on the state space $X := \mathbb{Z}_+^K$. The vector $e^K \in \mathbb{R}^K$ is the kth basis vector, $e^K = (0, \ldots, 0, 1, 0, \ldots, 0)'$.

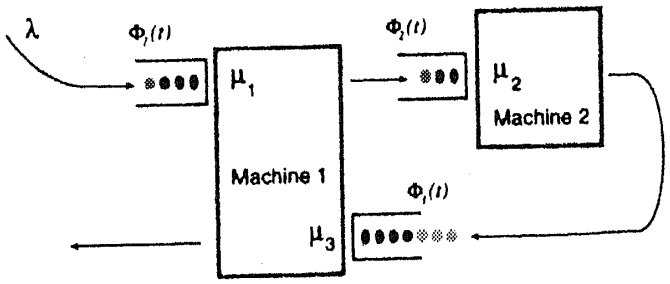

Figure 1. A multiclass network with $d = 2$ and $k = 3$.

The value of $\Phi_i(t)$ is the number of class i customers in the system at time t which await service in buffer i at machine $s(i)$. The function w: $\mathsf{X} \to \mathcal{A} := \{0,1\}^{K+1}$ is the feedback law 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 that $a(t)$ may take on at time t depends on the value x of the system at time t. This set, denoted $\mathcal{A}(x)$, is defined as follows where we write a as a column vector, $a = (a_0, \ldots, a_k') \in \mathcal{A}(x) \subset \{0+1\}^{K+1}$.

(i) $a_0 = 1$, and for any $1 \le i \le K, a_i = 0$ or 1:
(ii) For any $1 \le i \le K, x_1 = 0 \Rightarrow a_i = 0$;
(iii) For any machine $\sigma, 0 \le \sum_{i:s(i)=\sigma} a_i \le 1$;
(iv) For any machine $\sigma, \sum_{i:s(i)=\sigma} a_i = 1$ whenever $\sigma, \sum_{i:s(i)=\sigma} x_i > 0$.

Condition (iv) is the *non-idling* property that a server will always work if there is work to be done.

To regulate this system it is convenient to impose a cost function. Here a natural "cost" is $c(\Phi(t)) = |\Phi(t)| = \sum \Phi_i(t)$, which is the total customer population at time t. We use $|\cdot|$ here to denote the ℓ_1 norm. If this cost criterion can be minimized in some sense, then the system will be well-regulated as desired. The non-idling condition (iv) may then be assumed without any loss of generality since any optimal policy will be non-idling with this cost function.

Suppose that the random variables $\{I(t): t \ge 0\}$ are i.i.d. on $\{0,1\}^{K+1}$. We assume that $\mathrm{P}\{\sum_i I_i(t) = 1\}$, and we define $\mathrm{E}[I_i(t)] = \mu_i$. For $1 \le i \le K, \mu_i$ is interpreted as the service rate for class i customers, and for $i = 0$ we let $\mu_0 := \lambda$, which is interpreted as the arrival rate of customers class 1. It is clear then that the model (2) is an MDP of the form (1). The general results of this paper will be applied to this specific example in § 6.

1.2 *The optimality equations*

The MDP model can be equivalently defined through its transition probabilities. Consider the case where the state space X and the action space $\mathcal{A}$ are both countably infinite. The controlled transition probability P_a is then defined via

$$\begin{aligned} P_a(x,y) &= \mathrm{Prob}\{\Phi(t+1) = y | \Phi(t) = x, \quad a(t) = a\} \\ &= \mathrm{Prob}\{F(x,a,I) = y\}, \quad x,y \in \mathsf{X}, a \in \mathcal{A}, t \ge 0, \end{aligned}$$

where I is a generic noise variable. We consider the constrained control problem where for each x there is a set $\mathcal{A}(x) \subset \mathcal{A}$ such that $a(t) \in \mathcal{A}(x)$ whenever $\Phi(t) = x$.

Our attention is primarily restricted to *Markov policies* where the action $a(t)$ only depends upon $\Phi(t)$. This means that there exists a sequence of feedback laws $\{w^t : t \geq 0\}$ with $a(t) = w^t(\Phi(t))$ for each t; a generalization of the stationary case where $w^t = w^0$ for all t.

We evaluate a policy based upon the average cost criterion:

$$J(x,a) = \limsup_{n\to\infty} \frac{1}{n} \mathsf{E}_x \left[\sum_{t=0}^{n-1} c(\Phi(t), a(t)) \right], \tag{3}$$

where the initial condition $\Phi(0) = x$ is given. The quantity $c(\Phi(t), a(t))$ represents the cost paid at time t; the one-step cost c is a function from $\mathsf{X} \times \mathcal{A}$ to $[1, \infty)$. An *optimal policy* is one that minimizes (3) for any initial condition. It is hoped that an optimal policy does exist, and that it can be taken in the simpler stationary form.

The motivation for considering this average cost control problem is that there is good reason to believe that the optimal solution will take on the relatively simple stationary form. Moreover, it is a natural criterion in settings such as process control, where controlling transients is not so important as achieving good performace in the steady state.

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

$$\eta_* + h_*(x) = \min_{a \in \mathcal{A}(x)} [c(x,a) + P_a h_*(x)], \tag{4}$$

$$w^*(x) = \arg\min_{a \in \mathcal{A}(x)} [c(x,a) + P_a h_*(x)], \quad x \in \mathsf{X}. \tag{5}$$

In these equations the transition function P_a is viewed as an infinite-dimensional matrix, and h_* as a compatible column vector. Hence $P_a h_*$ may be viewed as a column vector whose xth entry is

$$P_a h_*(x) := \sum_{y \in \mathsf{X}} P_a(x,y) h_*(y).$$

The equality (4) is known as the *average cost optimality equation* (ACOE). The function h_* is known as a *relative value function*. The second equation (5) defines a stationary policy w^*. If a solution can be found, then typically the stationary policy w^* is optimal (see e.g. Arapostathis *et al* 1993; Borkar 1991; Meyn 1997; Puterman 1994).

The questions addressed in this paper all revolve around these two equations. Is there a solution? If so, how can it be computed? If an algorithm is found, is its convergence guaranteed? How fast? Are the intermediate feedback laws generated by the algorithm useful? The last question is important since we cannot wait forever for whatever algorithm we implement to converge.

There are two popular algorithms to attempt to solve (4). The first is the standard dynamical programming approach known as *value iteration* (VIA). The second is called *policy iteration* (PIA). The idea is this: The equation (4) is a fixed point equation of the form $\mathcal{L}(h) = h$, where h is a function on X, and $g = \mathcal{L}(h)$ is the function defined by

$$g(x) = \min_{a \in \mathcal{A}(x)} \{c(x,a) - \eta_* + P_a h(x)\}.$$

The VIA is then (essentially) the process of successive substitution: given one approximate solution h_k, we compute another, perhaps better approximation via $h_{k+1} = \mathcal{L}(h_k)$.

Although it is frequently convergent, at least locally, the successive approximation method is known to be slow in general since the "slope" of the function $h \to \mathcal{L}(h) - h$ may be small in a neighbourhood of h_*. The standard approach to speeding the convergence of the successive approximation approach is to use derivative information to adjust this undesirable slope – this leads to the Newton–Raphson method, to which the PIA is closely related (Cao 1998; Whittle 1996).

The remainder of this paper is organized as follows. The VIA and PIA are defined formally and their properties are explored in § 4 and 5. Sections 2 and 3 below provide some basic ergodic theory and optimization theory which form a foundation for the remainder of the paper. In the final section of the paper we present applications to the network scheduling problem.

2. A steady state cost?

Since the goal is to find a policy which minimizes the steady state cost given in (3) it is reasonable to first try to understand when the cost $J(x, a)$ can be expected to be finite, and how we can characterize the limit.

We consider in this section the stationary case where a is a Markov policy defined by the feedback law w so that $a(t) = w(\Phi(t)), t \in \mathbb{Z}_+$. We denote by P the transition kernel $P(x, y) = P_w(x, y)$ and similarly drop the dependence of $c(x, a)$ upon the control since the function w is fixed throughout. Hence is this section we let c denote a function on X with $c \geq 1$.

Although we are taking the very special case of countable state space chains, whenever possible we state definitions and results so that extensions to the general state space framework will at least seem plausible. The reader is referred to (Meyn & Tweedie 1992) for an analogous treatment of Markov chains on a general state space.

We assume that the Markov chain Φ with transition law P is ψ-irreducible, as defined by (Meyn & Tweedie 1992). For a countable state space model this means that there is a single communicating class which is reachable, with positive probability, from any initial condition. Equivalently, there exists a state $\theta \in \mathsf{X}$ which is *accessible* in the sense that

$$\sum_{t=0}^{\infty} P^t(x, \theta) = \sum_{t=0}^{\infty} \mathsf{P}\{\Phi(t) = \theta | \Phi(0) = x\} > 0, \quad x \in \mathsf{X}.$$

Throughout the paper we use θ to denote some fixed state which is accessible in this sense. We also assume thoughout that Φ is *aperiodic*, which is equivalent to the requirement that the accessible state θ satisfies $P^n(\theta, \theta) > 0$, for all n sufficiently large.

2.1 *Regularity, Lyapunov functions, and ergodicity*

For this uncontrolled Markov chain the existence of a limit in (3) is guaranteed if the chain is c-regular – a form of stability. For the countable state space chains considered here the definition is simply stated: Φ is c-regular if for some $\theta \in \mathsf{X}$ and every $x \in \mathsf{X}$,

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

where the first return time to the point θ is defined as

$$\tau_\theta = \min(t \geq 1 : \Phi(t) = \theta),$$

set to ∞ if the minimum is over an empty set.

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

$$\pi P(x) := \sum_{y \in \mathsf{X}} \pi(y) P(y, x) = \pi(x), \quad x \in \mathsf{X};$$

$$\pi(c) := \sum_{x \in \mathsf{X}} c(x)\pi(x) < \infty.$$

For c-regular chains we can solve equations of the form

$$Ph = h - c + \eta$$

for the unknown function h with $\eta = \pi(c)$. This is obviously closely connected to the solution of the average cost optimality equation.

A *stochastic Lyapunov function* in the context here is a function $V : \mathsf{X} \to \mathbb{R}_+$ satisfying the drift inequality

$$\mathsf{E}[V(\Phi(t+1))|\mathcal{F}_t] \leq V(\Phi(t)) - c(\Phi(t)) + s(\Phi(t)) \tag{6}$$

where $\mathcal{F}_t = \sigma(\Phi(r) : r \leq t)$, and s is bounded, positive function. When the one step cost $c(\Phi(t))$ is large, then (6) tells us that we expect the value of $V(\Phi(t+1))$ to be correspondingly smaller than $V(\Phi(t))$. Hence there is a negative "drift" towards the center of the state space where the cost is relatively low. The inequality (6) is closely related to both *Foster's criterion* for stability, and an approach known as *Lyapunov's second method.* Hence (6) is called a *Foster-Lyapunov drift inequality.* By the Markov property it may be equivalently expressed algebraically as $PV \leq V - c + s$.

The connections between Lyapunov functions and c-regularity are largely based upon the following general result, which is a consequence of the Comparison Theorem of (Meyn & Tweedie 1992).

Theorem 1. (*Comparison theorem*). *Let Φ be a Markov chain on X satisfying the drift inequality*

$$PV \leq V - c + s. \tag{7}$$

The functions V, c and s take values in $\mathbb{R}_+$.

Then for any stopping time τ

$$\mathsf{E}_x[V(\Phi(\tau))] + \mathsf{E}_x\left[\sum_{t=0}^{\tau-1} c(\Phi(t))\right] \leq V(x) + \mathsf{E}_x\left[\sum_{t=0}^{\tau-1} s(\Phi(t))\right]. \tag{8}$$

□

The comparison theorem provides the bound that is required in the definition of c-regularity provided that the RHS of (8) is finite valued. This of course will depend upon the function s. Define the *resolvent* as the weighted sum

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

A function $s : \mathsf{X} \to [0, 1]$ is called

petite: if there is $\theta \in \mathsf{X}$ such that

$$K(x, \theta) \geq s(x), \quad x \in \mathsf{X}. \tag{9}$$

special: if

$$\sup_{x \in \mathsf{X}} \mathsf{E}_x \left[\sum_{t=0}^{\tau_\theta - 1} s(\Phi(t)) \right] < \infty.$$

In the definition of a special function we do not require that the return time τ_θ be finite-valued.

If $S \subset \mathsf{X}$, and $\delta \mathbb{1}_S$ is petite (special) for some $\delta > 0$, then the set is called petite (special). A geometric trials argument shows that petite sets are always special (Meyn & Tweedie 1992; Nummelin 1984). If the chain is irreducible in the usual sense then obviously every finite set is petite.

If the function s used in Theorem 1 is special then the RHS of (8) will be bounded by $V(x)$ plus a constant, which implies that the chain is c-regular. This observation and renewal theory arguments lead to the following version of the f-Norm Ergodic Theorem of (Meyn & Tweedie 1992).

Theorem 2. *Suppose that Φ is a Markov chain on X satisfying the Foster-Lyapunov drift inequality*

$$PV \leq V - c + \overline{\eta}$$

where $\overline{\eta} > 0$ is a constant; the function c takes values in $[1, \infty)$, and the function V takes values in $\mathbb{R}_+$. Suppose moreover that the set $S = \{x : c(x) \leq 2\overline{\eta}\}$ is petite.

Then

(i) *Φ is a c-regular Markov chain satisfying the bound, for some $b_0 < \infty$,*

$$\mathsf{E}_x \left[\sum_{t=0}^{\tau_\theta - 1} c(\Phi(t)) \right] \leq V(x) + b_0, \quad x \in \mathsf{X};$$

(ii) *There is a unique invariant probability π, and $\pi(c) \leq \overline{\eta}$;*

(iii) *The "mean cost" converges for any x:*

$$\mathsf{E}_x[c(\Phi(t))] \to \pi(c) \quad t \to \infty.$$

□

2.2 *Poisson's equation*

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

$$Ph = h - c + \eta, \tag{10}$$

where $h : \mathsf{X} \to \mathbb{R}$. 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 $s : \mathsf{X} \to [0,1]$, and a probability v on X, the kernel $s \otimes v : \mathsf{X} \times \mathsf{X} \to [0,1]$ is defined as the product $s \otimes v(x.y) = s(x)v(y)$, $x, y \in \mathsf{X}$. Letting v denote the point-mass at θ, the minorization condition (9) may be expressed $K \geq s \otimes v$. Suppose that the Poisson equation has a solution h. Then we have the equivalent formula,

$$Kh = h - K\bar{c},$$

where $\bar{c}(x) = c(x) - \pi(c)$. Since any translation h+const. is also a solution, we may assume that $v(h) = 0$. We then have $(K - s \otimes v)h = h - K\bar{c}$, or

$$(I - (K - s \otimes v))h = K\bar{c}.$$

The inverse of the infinite matrix on the LHS is given by the infinite sum

$$(I - (K - s \otimes v))^{-1} = \sum_{i=0}^{\infty} (K - s \otimes v)^i.$$

Thus we may expect that the solution to Poisson's equation will be expressed by the formula

$$h(x) = \sum_{i=0}^{\infty} (K - s \otimes v)^i K\bar{c}(x), \quad x \in \mathsf{X}, \tag{11}$$

provided the sum is absolutely convergent. This is indeed the case for c-regular chains (Glynn & Meyn 1996; Meyn 1997).

The advantage of this construction is that it can be applied to Markov chains on a general state space, and may even be extended to continuous time processes, where the kernel K must be replaced with the continuous time version of the resolvent. When X is countable and $\theta \in \mathsf{X}$ is accessible, then a much more transparent construction is

$$h(x) = \mathsf{E}_x \left[\sum_{t=0}^{\tau_\theta - 1} (c(\Phi(t)) - \eta) \right], \quad x \in \mathsf{X}. \tag{12}$$

Evidently, for a c-regular chain the function h is finite-valued.

The papers (Glynn & Meyn 1996; Meyn 1997) use these ideas to establish a sufficient condition for the existence of suitably bounded solutions to Poisson's equation for general state space chains. Theorem 3 specializes this result to the countable state space setting.

Theorem 3. *Suppose that the Markov chain Φ is c-regular, and let*

$$V(x) := \mathsf{E}_x \left[\sum_{t=0}^{\tau_\theta - 1} c(\Phi(t)) \right].$$

Assume that with $\eta = \pi(c)$ the set S defined below is petite

$$S = \{x : Kc(x) \leq \pi(c)\}.$$

Then the function h given in (12) is a solution to Poisson's equation (10), and has the following properties:

(i) The function is bounded from above:

$$h(x) \leq V(x) \quad x \in \mathsf{X}.$$

(ii) It is bounded from below:

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

(iii) It is essentially unique: If g is any finite-valued function which is bounded from below and satisfies

$$Pg \leq g - c + \eta,$$

then $g(x) - g(\theta) = h(x) - h(\theta)$ for almost every $x \in \mathsf{X}[\pi]$, and $g(x) - g(\theta) \geq h(x) - h(\theta)$ for every $x \in \mathsf{X}$. □

The following result is then easily proven since Poisson's equation is a very special case of the drift inequality used in Theorem 2.

Theorem 4. *Suppose that c is unbounded off of petite sets. That is, the sublevel set $S_\eta = \{x : c(x) \leq \eta\}$ is petite for any $\eta > 0$.*
Then the following are equivalent:

(i) *Markov chain Φ is c-regular;*
(ii) *There exists a finite-valued function $V : \mathsf{X} \to \mathbb{R}_+$ and a constant $\overline{\eta} < \infty$ satisfying*

$$PV \leq V - c + \overline{\eta};$$

(iii) *There exists a positive and finite-valued solution h to Poisson's equation.* □

We note that if (ii) holds then the solution h of Poisson's equation can be chosen so that for some constant b_0,

$$0 \leq h(x) \leq V(x) + b_0, \quad x \in \mathsf{X}.$$

This follows from theorem 2.

3. Existence of optimal policies

We now return to the case of a controlled chain with transition function P_a. If a stationary policy defined by a feedback law w is used then the controlled chain has stationary transition probabilities, to be denoted P_w, so that

$$P_w(x, y) = P_{w(x)}(x, y), \quad x, y \in \mathsf{X}.$$

Similarly, we write the resulting one-step cost as c_w so that

$$c_w(x) = c(x, w(x)), \quad x \in \mathsf{X}.$$

A feedback law w is called *regular* if the Markov chain with transition law P_w is c_w-regular, as defined in the previous section. In this case the chain possesses a unique invariant probability, to be denoted π_w, and $\pi_w(c_w) < \infty$.

3.1 *The minimal relative value function*

With a better understanding of Poisson's equation in the control-free case we may now formulate existence criteria for the closely related ACOE (4). To begin we define the

following "cost over one cycle" η_w for a Markov policy w:

$$\eta_w = \min\left\{\eta \geq 1 : \mathsf{E}_\theta^w\left[\sum_{t=0}^{\tau_\theta - 1}(c(\Phi(t), w^t(\Phi(t))) - \eta)\right] \leq 0\right\}. \tag{13}$$

The minimal cyclic cost over all Markov policies will be denoted η_*:

$$\eta_* = \inf\{\eta_w : w \text{ is Markov}\}. \tag{14}$$

When w is a regular feedback law then the resulting stationary policy w satisfies $\eta_w = \pi_w(c_w) = J(x, w), x \in \mathsf{X}$.

The *minimal relative value function* h_* is defined pointwise as

$$h_*(x) = \inf\ \mathsf{E}_x^w\left[\sum_{t=0}^{\tau_\theta - 1}(c(\Phi(t), w^t(\Phi(t))) - \eta_*)\right], \tag{15}$$

where the minimum is over all Markov policies w.

This definition is a generalization of the construction of a solution to Poisson's equation given in the control-free case. To obtain the desired bounds on h_* (we hope that it is finite-valued, and uniformly bounded from below) the following assumptions will be imposed.

(A1): There exists a feedback law w^{-1}, a function $V_0 : \mathsf{X} \to \mathbb{R}_+$, and $\overline{\eta} < \infty$ satisfying

$$P_{w^{-1}} V_0 \leq V_0 - c_{w^{-1}} + \overline{\eta}.$$

(A2): For each fixed $x \in \mathsf{X}$, the set $\mathcal{A}(x)$ is finite.
For each $\eta > 0$ the following sublevel set is finite:

$$S_\eta = \left\{x : \min_{a \in \mathcal{A}(x)} c(x, a) \leq \eta\right\}.$$

(A3): There is a fixed state $\theta \in \mathsf{X}$ such that for any Markov policy w,

$$K_w(x, \theta) > 0, \quad x \in \mathsf{X}.$$

The assumption (A1) means simply that there is at least one regular policy. Condition (A2) asks that large states receive a large penalty through the cost function c. Condition (A3) is a generalization of the petite set condition imposed in the previous section. Using Fatou's lemma one may show that this implies the (apparently) stronger set of conditions which show that the sublevel sets are "uniformly petite" and "uniformly special" over all Markov policies.

PROPOSITION 1
Suppose that (A1)–(A3). *Then*

(i) *For some function* $s : \mathsf{X} \to (0, 1)$ *the following bound holds for all* $x \in \mathsf{X}$ *and all Markov policies* w:

$$K_w(x, \theta) \geq s(x);$$

(ii) *For some non-decreasing function* $B : \mathbb{R}_+ \to \mathbb{R}_+$ *and any Markov policy* w,

$$\sup_{x \in \mathsf{X}} \mathsf{E}_x^w\left[\sum_{t=0}^{\tau_\theta - 1} \mathbb{1}_{S_\eta}(\Phi(t))\right] < B(\eta), \tag{16}$$

where S_η *is defined in assumption* (A2).

Proof. (i) Suppose not. Then there exists some x_0 such that $\inf_w K_w(x_0, \theta) = 0$, and hence there is a sequence of Markov policies w^n for which $K_{w^n}(x_0, \theta) \leq 1/n$. Assume without loss of generality that $w^n \to w^\infty$ pointwise as $n \to \infty$. By Fatou's lemma we then have

$$K_{w^\infty}(x_0, \theta) \leq \liminf_{n \to \infty} K_{w^n}(x_0, \theta) = 0,$$

which contradicts (A3), establishing (i).

(ii) Fix $\eta > 0$ so that S_η is non-empty and find $\delta > 0$ such that $K_w(x, \theta) \geq \delta$ for $x \in S_\eta$, and any Markov policy w. We then have for any such w,

$$\mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} \mathbb{1}_{S_\eta}(\Phi(t)) \right] \leq \delta^{-1} \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} K_{w^{[t]}}(\Phi(t), \theta) \right]$$

where $w^{[t]}$ is the shifted policy

$$w^{[t]} = (w^t, w^{t+1}, \ldots).$$

From the Markov property we can replace the resolvent by a sum over a sample path as follows

$$\begin{aligned}
&\mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} K_{w^{[t]}}(\Phi(t), \theta) \right] \\
&\leq \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} \left(\sum_{k=0}^{\infty} 2^{-k-1} \mathbb{1}(\Phi(t+k) = \theta) \right) \right] \\
&\leq \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} \left(\sum_{k=\tau_\theta - t}^{\infty} 2^{-k-1} \right) \right] \\
&= \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} 2^{-(\tau_\theta - 1)} \right] \\
&\leq \sum_{t=1}^{\infty} 2^{-t} = 1.
\end{aligned}$$

□

3.2 *Properties*

From the uniform bound (16) we then obtain uniform bounds on the minimal relative value function:

PROPOSITION 2

Under (A1)–(A3) *we have* $h_*(\theta) = 0$, *and for some* $b_0 < \infty$

$$-b_0 \leq h_*(x) \leq V_0(x) + b_0, \quad x \in \mathsf{X}.$$

Proof. That $h_*(\theta) = 0$ follows from the minimality of η_* and the definition of h_*.

The lower bound follows from proposition 1 which implies the explicit bound $h_*(x) \geq -\eta_* B(\eta_*), x \in \mathsf{X}$. The upper bound is a consequence of minimality of h_*:

$$h_*(x) \le \mathsf{E}_x^{w^{-1}}\left[\sum_{t=0}^{\tau_\theta-1} c(\Phi(t), w^t(\Phi(t)))\right] \le V_0(x) + \text{const.}$$

□

For any Markov policy $w^0 = (w_0^0, w_1^0, \ldots)$ we can derive the following dynamic programming inequality using the fact that $h_*(\theta) = 0$:

$$\begin{aligned}
&\min_{a\in\mathcal{A}(x)} \{c(x,a) - \eta_* + P_a h_*(x)\} \\
&\le c(x, w_0^0(x)) - \eta_* + \sum_{y\neq\theta} P_{w_0^0}(x,y) h_*(y) \\
&= c(x, w_0^0(x)) - \eta_* + \sum_{y\neq\theta} P_{w_0^0}(x,y)\left\{\min_w \mathsf{E}_y^w\left[\sum_{t=0}^{\tau_\theta-1}(c(\Phi(t), w^t(\Phi(t))) - \eta_*)\right]\right\} \\
&\le c(x, w_0^0(x)) - \eta_* + \sum_{y\neq\theta} P_{w_0^0}(x,y)\left\{\mathsf{E}_y^{w^{0[1]}}\left[\sum_{t=0}^{\tau_\theta-1}(c(\Phi(t), w_{t+1}^0(\Phi(t))) - \eta_*)\right]\right\},
\end{aligned}$$

where in the last inequality we denote $w^{0[1]} = (w_1^0, w_2^0, \ldots)$. From the Markov property we see that

$$\min_{a\in\mathcal{A}(x)} \{c(x,a) - \eta_* + P_a h_*(x)\} \le \mathsf{E}_x^{w^0}\left[\sum_{t=0}^{\tau_\theta-1}(c(\Phi(t), w_t^0(\Phi(t))) - \eta_*)\right].$$

We can minimize over all Markov policies since w^0 was arbitrary to obtain the inequality:

$$\min_{a\in\mathcal{A}(x)} \{c(x,a) - \eta_* + P_a h_*(x)\} \le h_*(x), \quad x \in \mathsf{X}.$$

Suppose that w^* is a minimizing feedback law:

$$c_{w^*}(x) + P_{w^*} h_*(x) = \min_{a\in\mathcal{A}(x)} \{c(x,a) + P_a h_*(x)\} \tag{17}$$

We see then that the Poisson inequality holds

$$c_{w^*}(x) + P_{w^*} h_*(x) \le \eta_* + h_*. \tag{18}$$

It follows that w^* is regular, and from the comparison theorem we have the bound $\pi_{w^*}(c_{w^*}) = \eta_{w^*} \le \eta_*$. By minimality of η_* the equality $\eta_{w^*} = \eta_*$ must hold, and by minimality of h_* and uniqueness of solutions to Poisson's equation the inequality (18) must also be an equality. We summarize these results in the existence theorem.

Theorem 5. *Suppose that* (A1)–(A3) *hold and that* $\eta_* < \infty$. *Then*

(i) *The minimal relative value function is a solution to the ACOE.*
(ii) *Any feedback law* w^* *satisfying* (17) *is optimal over all Markov policies.*
(iii) *Any feedback law* w^* *satisfying* (17) *minimizes the "relative cost": For any Markov policy* w, *and any* x,

$$h_*(x) = \mathsf{E}_x^{w^*}\left[\sum_{t=0}^{\tau_\theta-1}(c_{w^*}(\Phi(t))) - \eta_*)\right] \le \mathsf{E}_x^{w}\left[\sum_{t=0}^{\tau_\theta-1}(c(\Phi(t), w^t(\Phi(t))) - \eta_*)\right].$$

(iv) *If* (h_+, w^+) *is any other solution to* (4) *for which* $\inf_x h_+(x) > -\infty$, *then* w^+ *is regular with unique invariant probability* π_+, *and*

$$h_+(x) - h_+(\theta) = h_*(x), \quad \text{if } \pi_+(x) > 0;$$
$$h_+(x) - h_+(\theta) \geq h_*(x), \quad \text{for all } x.$$

□

4. Value iteration

Now that we know that the ACOE has a solution, the next question we consider is, how do we compute a solution? Recall that the ACOE may be described as a fixed point equation

$$h_* = \mathcal{L}(h_*) = \min_a \{c(\cdot, a) - \eta_* + P_a h_*\}.$$

The first algorithm we shall look at is the value iteration algorithm, or VIA.

4.1 *Successive approximation*

The VIA approach to solving the ACOE was described in the introduction as being "somewhat like" the successive approximation approach to solving a fixed point equation. Suppose that $\tilde{h}_0$ is given as an initial condition. Assume $\tilde{h}_0 : \mathsf{X} \to \mathbb{R}$ with $\inf_{x \in \mathsf{X}} \tilde{h}_0(x) > -\infty$. Then the actual successive approximation algorithm defines a sequence of functions $\{\tilde{h}_n : n \geq 0\}$ inductively as follows: $\tilde{h}_{n+1} = \mathcal{L}(\tilde{h}_n)$, or equivalently,

$$\tilde{h}_{n+1}(x) = \min_{a \in \mathcal{A}(x)} \{c(x, a) - \eta_* + P_a \tilde{h}_n(x)\}, \quad x \in \mathsf{X}, n \geq 0. \tag{19}$$

For any n the function $\tilde{h}_n$ has the interpretation

$$\tilde{h}_n(x) = \min \mathsf{E}_x^w \left[\sum_{t=0}^{n-1} (c(\Phi(t), w^t(\Phi(t))) - \eta_*) + \tilde{h}_0(\Phi(n)) \right],$$

where the minimum is over all Markov policies. Note the similarity between this and the definition (15), which is one motivation for the algorithm. However the difficulty with (19) is that one must know a *priori* the optimal cost η_*.

This turns out to be a minor difficulty since the constant η_* plays a minor role in the recursion (19). On eliminating this constant we arrive at the more easily implemented value iteration algorithm, which may be described as a two step procedure: Given the value function V_n,

Step 1: Compute the feedback law w^n through the minimization

$$w^n(x) = \arg\min_{a \in \mathcal{A}(x)} (c_a + P_a V_n(x)).$$

Step 2: Compute the value function V_{n+1} via

$$V_{n+1} = c_n + PnV_n,$$

with $c_n = c_{w^n}$ and $P_n = P_{w^n}$.

The function $V_0 \geq 0$ is given as an initial condition. When the initial conditions are identical, $\tilde{h}_0 = V_0$, we then see that $V_n(x) - V_n(\theta) = \tilde{h}_n(x) - \tilde{h}_n(\theta)$ for all x and n. Hence the function h_n defined by

$$h_n(x) := V_n(x) - V_n(\theta), \quad x \in \mathsf{X}, \tag{20}$$

is a plausible approximation to the solution of the ACOE when n is large.

The value function V_n has the interpretation

$$V_n(x) = \min \mathsf{E}_x \left[\sum_{t=0}^{n-1} c(\Phi(t), a(t)) + V_0(\Phi(n)) \right], \tag{21}$$

where the minimum is over all policies, and is attained with the Markov policy v^n whose first n actions may be expressed

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

We shall let $w^n := (w^n, w^n, w^n, \ldots)$ denote the stationary policy defined by the nth feedback law.

4.2 *Stability*

The VIA generates a sequence of feedback laws $\{w^n : n \geq 0\}$ which one expects will in some way approximate an optimal feedback law when n becomes large. At the very least we hope that the feedback laws are regular or exhibit some form of stability. However, one cannot even expect the controlled chains to be recurrent unless some additional assumptions are imposed. The following example is taken from (Chen & Meyn 1997).

Consider the network illustrated in figure 2 consisting of four buffers and two machines fed by separate arrival streams. It is shown by (Rybko & Stolyar 1992) that the last buffer-first served policy where buffers 2 and 4 receive priority at their respective machines will make the controlled process Φ transient, even under the usual traffic condition that $\rho < 1$, 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 given by

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

We conclude that $J(w^2) = \infty$ when $\lambda/\mu_2 + \lambda/\mu_4 > 1$ since this is precisely the destabilizing policy introduced by (Kumar & Seidman 1990; Rybko & Stolyar 1992). In fact,

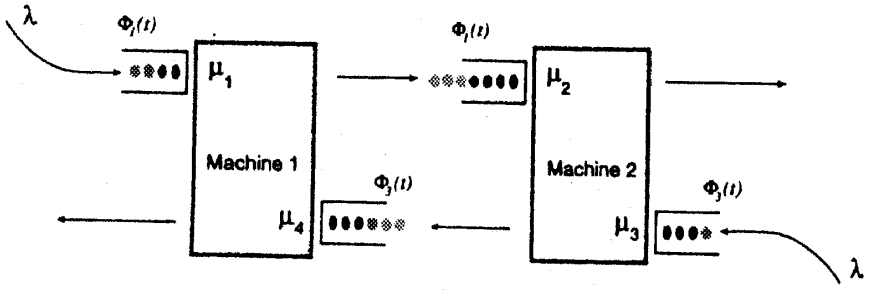

Figure 2. A multiclass network with $d = 2$ and $K = 4$.

not only is the average cost infinite, the feedback law w^2 gives rise to a transient Markov chain. □

The difficulty in this example is that the VIA is initialized with $V_0 \equiv 0$, which is not remotely close to the actual relative value function h_*. Suppose that instead of zero, the initial condition V_0 is chosen as a Lyapunov function for just one policy. In this case all of the feedback laws $\{w^n\}$ are guaranteed to be regular. To see this let V_0 be the function defined in (A1), let $g_n = V_{n+1} - V_n$, set $\overline{\eta}_n = \sup_{x \in X} g_n(x)$, and let $\eta_n = \eta_{w^n}$. We let P_n and c_n denote the one-step transition probability and one-step cost obtained on the nth iteration, and π_n denote the invariant probability for P_n, if one exists:

$$P_n = P_{w^n}, \quad c_n = c_{w^n}, \quad \pi_n = \pi_{w^n}.$$

Theorem 6. *Under* (A1)–(A3), *if the VIA is initialized with the Lyapunov function* V_0 *then for each* $n \in \mathbb{Z}_+$,

(i) $\overline{\eta}_{n+1} \le \overline{\eta}_n \le \overline{\eta}$;
(ii) $P_n V_n \le V_n - c_n + \overline{\eta}_n$;
(iii) *The feedback law* w^n *is regular with* $\eta_n \le \overline{\eta}_n$.

Proof. To see (i) we first note that for all n,

$$P_n g_n = P_n V_{n+1} + c_n - (P_n V_n + C_n) \ge V_{n+2} - V_{n+1} = g_{n+1}. \tag{22}$$

It immediately follows that $\overline{\eta}_n = \sup_x g_n(x) \ge \overline{\eta}_{n+1}$.

The bound (ii) is then a simple consequence of the definitions of V_{n+1} and g_n:

$$P_n V_n = V_{n+1} - c_n = V_n - c_n + g_n \le V_n - c_n + \overline{\eta}_n.$$

The final result (iii) then follows from theorem 2. □

Hence it pays to initialize the VIA using a stochastic Lyapunov function. If there is a choice, one should choose the function V_0 so that the constant $\overline{\eta}$ is as small as possible since this represents a global upper bound on the performance of all succeeding feedback laws.

4.3 *Convergence*

Convergence of the VIA is more delicate than simple stability of the intermediate policies. A complete proof of the convergence of the functions $\{h_n\}$ given in (20) may be found (Chen & Meyn 1997) for an initial function V_0 satisfying (A1). The case $V \equiv 0$ is considered (Cavazos-Cadena 1996) under somewhat stronger assumptions on the model. Here we sketch the main ideas of the proof given by (Chen & Meyn 1997).

The first step is

PROPOSITION 3
Under (A1)–(A3) *we have* $h_n(\theta) = 0$ *for all* n, *and for some* $b_0 < \infty$

$$-b_1 \le h_n(x) \le V_0(x) + b_0, \quad x \in X.$$

Proof. The lower bound follows from theorem 6 and the comparison theorem which

together give

$$h_n(x) = V_n(x) - V_n(\theta) \geq \mathsf{E}_x^{w^n}\left[\sum_{t=0}^{\tau_\theta - 1}(c_n(\Phi(t)) - \bar{\eta}_n)\right] \geq -\bar{\eta}\,\mathsf{E}_x^{w^n}\left[\sum_{t=0}^{\tau_\theta - 1}\mathbb{1}_{S_{\bar{\eta}}}(\Phi(t))\right]$$

The RHS is uniformly bounded from below by Proposition 3.1.

For the upper bound use minimality of V_n to obtain the bound

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

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

$$h_n(x) \leq \mathsf{E}_x^0\left[\left(\sum_{t=0}^{n-1} c_0(\Phi(t)) + V_0(\Phi(n))\right)\mathbb{1}(\tau_\theta > n)\right] + \mathsf{E}_x^0\left[\sum_{t=0}^{\tau_\theta - 1} c_0(\Phi(t))\right] + \mathsf{E}_x^0[(V_{n-\tau_\theta}(\theta) - V_n(\theta))\mathbb{1}(\tau_\theta \leq n)]. \tag{23}$$

Each of these three terms may be bounded through regularity of w^0. From the inequality $P_0 V_0 \leq V_0 - c_0 + \bar{\eta}$ and theorem 2 we see that the second term is bounded by $V_0(x)$ plus a constant. The last term is similary bounded since one may show that $g_k(\theta) \geq -b_1, k \geq 0$, for some constant b_1. Hence

$$V_{n-\tau_\theta}(\theta) - V_n(\theta) = -\sum_{k=n-\tau_\theta}^{n-1} g_k(\theta) \leq b_1 \tau_\theta.$$

The first term is bounded using these same ideas and an inductive argument. □

These uniform bounds may then be used to establish a form of "uniform regularity" for the Markov policies $\{v^n : n \geq 0\}$.

Lemma 1. *Under* (A1)–(A3) *there is a fixed constant* b_0 *such that for all* x *and* n,

$$\mathsf{E}_x^{v^n}[n \wedge \tau_\theta] \leq b_0(V_0(x) + 1).$$

Proof. Letting $V(t) = h_{n-t+1}(\Phi(t))$ we can establish the following drift inequality for the non-homogeneous chain

$$\mathsf{E}^{v^n}[V(t+1)|\mathcal{F}_t] \leq V(t) - c_{v_t^n}(\Phi(t)) + \bar{\eta}, \quad 0 \leq t < n.$$

This is similar to the Lyapunov drift inequality (2.1) and may be used in the same way to obtain the desired bound on the mean of τ_θ. The bound is uniform in n because of proposition 1 and proposition 3. □

The next key step is to obtain a uniform bound on

$$\bar{g}(x) := \limsup_{n\to\infty} g_n(x), \quad x \in \mathsf{X}.$$

Lemma 2. Suppose that (A1)–(A3) hold and that the initial value function V_0 satisfies

$$(1/n)P^n_{w^*}V_0(x) \to 0, \quad n \to \infty, \ x \in \mathsf{X}.$$

Then

(i) $\bar{g}(x) \le \eta_*$ *for every* $x \in \mathsf{X}$,
(ii) $\lim_{n\to\infty} g_n(\theta) = \eta_*$.

Proof. There are two steps to the proof. First we show that θ is maximal in the sense that

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

To see this we apply the upper bound $P_n g_n \le g_{n+1}$ from (22) to construct a submartingale under v^n defined as $M(t) = g_{n-t+1}(\Phi(t))$. Letting $\tau = \tau_\theta \wedge n$ then gives

$$\mathsf{E}^{v^n}_x[g_{n-r+1}(\Phi(\tau))] = \mathsf{E}[M(\tau)] \ge M(0) = g_{n+1}(x).$$

Using this bound and the previous lemma giving uniform bounds on the first moment of τ we find that for any x,

$$\bar{g}(\theta) = \limsup_{n\to\infty} \mathsf{E}^{v^n}_x[g_{n-\tau+1}(\Phi(\tau))] \ge \limsup_{n\to\infty} g_{n+1}(x) = \bar{g}(x).$$

Given that θ is maximal we then obtain a bound on $\bar{g}(\theta)$. We can find a subsequence $\{n_i\}$ of $\mathbb{Z}_+$ such that for each t the following limit exists pointwise

$$W_t(x) := \lim_{i\to\infty} h_{n_i-t+1}.$$

Moreover, from the uniform bounds on $\{h_n\}$ we may choose $\{n_i\}$ so that

$$P_{w^*}W_t \ge W_{t-1} - c_{w^*} + \bar{g}(\theta),$$

and so that for some $b_0 < \infty$,

$$-b_0 \le W_t \le V_0 + b_0.$$

By iteration we then obtain

$$P_{w^*}W_t \ge W_{t-n} - \sum_{i=0}^{n-1} P^i_{w^*}c_{w^*} + n\bar{g}(\theta).$$

and from the bounds on $\{W_t\}$ this gives

$$\bar{g}(\theta) \le \frac{1}{n}\left(P^n_{w^*}V_0 + 2b_0 + \sum_{i=0}^{n-1} P^i_{w^*}c_{w^*}\right).$$

Letting $n \to \infty$ we find that the RHS converges to η_*, which gives the desired upper bound $\bar{g}(x) \le \bar{g}(\theta) \le \eta_*, x \in \mathsf{X}$.

To obtain equality when $x = \theta$ involves the identity

$$P_n h_n = h_n - c_n + g_n,$$

which implies that $\pi_n(g_n) \ge \eta_n \ge \eta_*$. An argument based upon Fatou's lemma then shows that $\bar{g}(\theta) \ge \eta_*$. Since we already have the reverse inequality this shows that $\bar{g}(\theta) = \eta_*$.

□

Combining these bounds we obtain the following main result. Part (iii) follows from strict uniqueness of solutions to Poisson's equation which we have established in theorem 3 for the irreducible case only.

Theorem 7. *Suppose that* (A1)–(A3) *hold and that the initial condition* V_0 *satisfies*

$$(1/n)P_{w^*}^n V_0(x) \to 0, \quad n \to \infty, \ x \in \mathsf{X}.$$

Then

(i) *As* $n \to \infty$,

$$g_n(\theta) \to \eta_* \quad \textit{and} \quad (1/n)V_n(\theta) \to \eta_*,$$

where η_* *is the optimal cost;*

(ii) *If* h_∞ *is any point-wise limit of the sequence* $\{h_n : n \geq 0\}$, *then the pair* (h_∞, η_*) *is a solution to the average cost optimality equation* (1.4);

(iii) *Suppose in addition that* Φ^w *is irreducible in the usual sense for any optimal feedback law* w. *Then as* $n \to \infty$,

$$h_n(x) \to h_*(x), \quad x \in \mathsf{X},$$

with h_* *equal to the minimal relative value function.* □

5. Policy iteration

The PIA may be viewed as an optimized form of the VIA. This is seen by considering the Lyapunov drift inequality which holds at the nth state of the VIA under (A1)–(A3)

$$c_n + P_n V_n \leq V_n + \bar{\eta}_n. \tag{24}$$

It is this inequality that ensures regularity of both w^n and w^{n+1}, and implies that η_n and η_{n+1} are finite. Although there are many solutions to (24) in general, for an irreducible chain the solution which minimizes the upper bound $\bar{\eta}_n$ is unique, up to an additive constant. This 'optimal' Lyapunov function is h_n, the solution to Poisson's equation, and the minimal upper bound is η_n, the steady state cost under w^n.

The PIA is then an implementaion of this reasoning: at each stage n the solution to Poisson's equation is used to construct the next policy w^{n+1} instead of the value function V_n. Hence we again arrive at a two-step procedure: Given a feedback law w^n,

Step 1: Find a solution h_n to the Poisson equation

$$c_n + P_n h_n = h_n + \eta_n$$

with $\inf_x h_n(x) > -\infty$.

Step 2: Compute the next policy through the minimization

$$w^{n+1}(x) = \arg\min_{a \in \mathcal{A}(x)} (c_a + P_a h_n(x)).$$

Although similar in form to the VIA, this algorithm is substantially more computationally intensive due to the matrix inversion required in step 1.

To initialize the algorithm we only need a feedback law w^0 for which the associated Poisson equation admits a positive solution. That is, we require an initial regular feedback law, and we find in theorem 8 below that with this initialization all subsequent feedback laws will also be regular. This result also establishes bounds on the intermediate solutions to Poisson's equation which are tight enough to ensure convergence. For definiteness we take the specific solution to Poisson's equation given by

$$h_n(x) = \mathsf{E}_x^n\left[\sum_{t=0}^{\tau_\theta - 1}(c_n(\Phi(t)) - \eta_n)\right], \quad x \in \mathsf{X}.$$

Theorem 8. *Under* (A1)–(A3), *suppose that the PIA is initialized with a regular policy, such as* w^{-1}. *Then the PIA generates a sequence of regular feedback laws* $\{w^n : n \geq 0\}$ *and relative value functions* $\{h_n : n \geq 0\}$ *such that*

(i) *The average costs are decreasing:*

$$J(x, w^{n+1}) = \eta_{n+1} \leq \eta_n, \quad n \geq 0, x \in \mathsf{X}.$$

(ii) *The relative value functions are uniformly bounded from below, and "almost" decreasing: There exists* $b_0, b_1 < \infty$ *such that*

$$-b_0 \leq h_{n+1}(x) \leq (1 + b_0(\eta_n - \eta_{n+1}))h_n + b_1(\eta_n - \eta_{n+1}), \quad n \geq 0, x \in \mathsf{X}.$$

(iii) *The sequence* $\{h_n : n \geq 0\}$ *is pointwise convergent to a limit* h_∞ *satisfying, for some* $b_2, b_3 < \infty$,

$$-b_2 - \leq h_\infty(x) \leq b_2 h_0(x) + b_3, \quad x \in \mathsf{X}.$$

Proof. The proof is similar to the proof of theorem 6. To begin, observe that for each n the function h_n serves as a Lyapunov function for the transition function P_{n+1}:

$$c_{n+1} + P_{n+1}h_n \leq h_n + \eta_n. \tag{25}$$

From the comparison theorem and theorem 2 we conclude by induction that each feedback law w^n is regular and that $\eta_{n+1} \leq \eta_n$ for all n.

Also using the comparison theorem and (24) we obtain two important bounds: For some constants b_4, b_5,

$$\mathsf{E}_x^{n+1}[\tau_\theta] \leq b_4 h_n(x) + b_5, \quad n \geq 0, x \in \mathsf{X} \tag{26}$$

and,

$$\mathsf{E}_x^{n+1}\left[\sum_{t=0}^{\tau_\theta - 1}(c_{n+1}(\Phi(t)) - \eta_n)\right] \leq h_n(x) - h_n(\theta) = h_n(x), \quad n \geq 0, \ x \in \mathsf{X}. \tag{27}$$

The first bound requires some additional work using the "uniformly special" property of the sublevel sets of c given in Proposition 1.

Since the LHS of (27) differs from the definition of h_{n+1} only in the use of η_n in place of η_{n+1}, we can combine (26) and (27) to obtain

$$\begin{aligned} h_{n+1}(x) &\leq h_n(x) + (\eta_n - \eta_{n+1})\mathsf{E}_x^{n+1}[\tau_\theta] \\ &\leq (1 + b_4(\eta_n - \eta_{n+1}))h_n(x) + b_5(\eta_n - \eta_{n+1}). \end{aligned}$$

These arguments complete the proof. □

From these bounds we may prove a complement to theorem 7.

Theorem 9. *Suppose that* (A1)–(A3) *hold and that the initial condition* h_0 *satisfies*

$$(1/n)P^n_{w^*}h_0(x) \to 0, \quad n \to \infty, x \in \mathsf{X}.$$

Then, as $n \to \infty$,

$$h_n(x) \to h_\infty(x), \quad \eta_n \to \eta_*,$$

where (h_∞, η_*) *is a solution to the average cost optimality equations, and* η_* *is the optimal cost.* □

6. Networks

We now return to the network scheduling problem described in the introduction. This is given as the nonlinear state space model (2) with countable state space $\mathbb{Z}_+^K$. Viewed as a combinatorial optimization problem the scheduling problem possesses unimaginable complexity. For the idealized model (2) with unbounded buffers and hence countable state space it is unlikely that any optimal scheduling policy will ever be explicitly computable except in trivial cases such as a single queue where there is nothing to schedule.

However, when viewed as a dynamic control problem there is much that can be said, and the general results described in the preceding sections aid a great deal in our understanding of this control problem. We describe here the form of the optimal policy when the system state corresponds to a highly congested network. In this region of the state space the optimal policy approximates the solution to a fluid control problem which attempts to drive the state towards the origin in an optimal fashion.

The fluid model is described by an ordinary differential equation

$$\frac{d}{dt}\phi(t) = \sum_{k=0}^{K} \mu_k[e^{k+1} - e^k]u_k(t), \tag{28}$$

where the function $u(t) \in \mathbb{R}^{K+1}$ is analogous to the discrete control, and satisfies similar constraints (Dai 1995). If (28) is to approximate (2) then the control u must depend on which feedback law w was used for the original discrete model. Suppose, for example, that w is the last buffer-first served (LBFS) policy which gives priority to a class k customer over a class j customer at the same machine if and only if $k \geq j$, then the control u for the fluid model (28) will be the same, though it must be described with greater care: at a given machine σ, if $s(k) = \sigma$ then

$$\sum_{i \geq k : s(i) = \sigma} u_i(t) = 1, \quad \text{whenever} \quad \sum_{i \geq k : s(i) = \sigma} \phi_i(t) > 0.$$

The primary reason for studying (28) is that when properly scaled, this equation describes approximately the behaviour of the network (2) in a transient phase during which the network is highly congested. The scaling is through the initial condition $\boldsymbol{\Phi}(0) = x$ which is assumed to be large: If $|x|t$ is an integer, we set

$$\phi^x(t) = (1/|x|)\boldsymbol{\Phi}(|x|t).$$

For all other $t \geq 0$, we 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.

One can show that for large x the process ϕ^x is an approximate solution to (28) for some choice of u.

In this section we describe qualitatively the consequences of the relationship between (2) and (28) to the network scheduling problem. We will not make many precise statements – for theorems and proofs, the reader is referred to the papers by (Dai 1995; Kumar & Meyn 1996) for connections with stability, and (Chen & Meyn 1997; Meyn 1997) for results pertaining to network optimization and optimization of the fluid model (28).

6.1 *The relative value function*

To understand the minimal relative value function h_* for the network control problem we first must understand regularity in this context. For any feedback law w it may be shown using the skip-free property that $|\Phi(t+1)| \geq |\Phi(t)| - 1$ that

$$h_w(x) := \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} (c(\Phi(t), w^t(\Phi(t))) - \eta_*) \right] \geq \frac{1}{4}|x|^2 - \frac{1}{2} B(2\eta_*),$$

where θ denotes the empty state in which every buffer is empty. Thus the function h_* satisfies the same lower bound.

Provided that the network can be stabilized (that is, the usual traffic condition that $\rho < 1$ holds) for many policies we can obtain a quadratic upper bound

$$h_w(x) \leq b_0(|x|^2 + 1), \quad x \in \mathsf{X},$$

where b_0 is some fixed constant. This bound holds for the last buffer-first served policy, for example. Hence the *minimal* relative value function must be *equivalent to a quadratic* in the sense that for some $\epsilon > 0, b_0 < \infty$,

$$\epsilon|x|^2 - b_0 \leq h_*(x) \leq \epsilon^{-1}|x|^2 + b_0, \quad x \in \mathsf{X}.$$

These observations are a consequence of the following result taken from (Kumar & Meyn 1996; Meyn 1997b). Although this result is stated as a list of "properties", the reader must be warned that many of the results given here are not entirely precise as stated.

Property 1. *The following are equivalent for any feedback law w*:

(i) *There exists a solution V to the Foster–Lyapunov drift criterion* (7) *which is equivalent to a quadratic.*
(ii) *There exists a solution to Poisson's equation* (10) *which is equivalent to a quadratic.*
(iii) *The "total cost" for the fluid model is uniformly bounded: For some* $B_0 < \infty$

$$\int_0^\infty |\phi(t)| \mathrm{d}t < B_0.$$

□

The fluid limit model will be called *stable* if property 6.1 (iii) is satisfied. That is, the total cost for the fluid model is uniformly bounded over all initial conditions satisfying

$|\phi(0)| = 1$. It will be useful to consider the minimal cost for the fluid model given by

$$V_*(x) = \min \int_0^\infty |\phi(t)|dt; \quad \phi(0) = x, x \in \mathsf{X}, \tag{29}$$

where the minimum is over all possible controls u for the fluid model.

The equivalence expressed in property 1 can be refined to show that in fact the fluid model is stable in the sense of (iii) then the solution to Poisson's equation in (ii) can be approximated by the total cost:

$$h(x) \approx \int_0^\infty |\phi(t)|\mathrm{d}t, \quad \text{when } \phi(0) = x, \ |x| \gg 1. \tag{30}$$

This suggestive approximation leads to an interesting implementation of the VIA, and is also used to show that the PIA for the network simultaneously acts as the policy improvement algorithm for the fluid model which computes the value function V_*. These connections will be explored next.

6.2 *Algorithms*

We have seen that the PIA leads to the pair of equations

$$\begin{aligned} P_n h_n &= h_n - c + \eta_n \\ P_{n+1} h_n &\le h_n - c + \eta_n, \end{aligned}$$

where here we are taking $c(x) = |x|$. If h_n is quadratically bounded, then it follows from theorem 2 and the above inequality that h_{n+1} is also quadratically bounded. Hence, from property 1 if the initial feedback law w^0 is chosen so that its fluid model is stable, then all subsequent policies will have stable fluid models.

From these equations we can also show that $h_{n+1}(x)/|x|^2 \le h_n(x).|x|^2 + o(1)$, where the term $o(1)$ is bounded by a constant times $1/|x|$. From (30) we then obtain the following.

Property 2. *If the initial feedback law w^0 is chosen so tht the fluid model is stable, then*

(i) *The PIA produces a sequence $\{(w^n, h_n, \eta_n) : n \le 0\}$ such that each associated fluid model is stable. Any feedback law w^* which is a pointwise accumulation point of $\{w^n\}$ is an optimal average cost policy.*

(ii) *For each $n \le 1$, with identical initial conditions,*

$$\int_0^\infty |\phi^n(s)|\mathrm{d}s \le \int_0^\infty |\phi^{n-1}(s)|\mathrm{d}s$$

Hence if w^0 is optimal for the fluid model, so is w^n for all n.

(iii) *With ϕ^* equal to the fluid solution for any optimal feedback law w^* generated by the algorithm, and h_* equal to the minimal relative value function*

$$h_*(x) \approx |x|^2 \int_0^\infty |\phi^*(s)|\mathrm{d}s, \quad \phi^*(0) = x/|x|, x \gg 1.$$

Hence, when properly normalized, the relative value function approximates the value function for the fluid model control problem.

(iv) *With ϕ^* equal to the fluid solution for any optimal feedback law w^* generated by the algorithm, and ϕ equal to any other fluid model solution with identical initial condition $\phi^*(0) = \phi(0) = x$,*

$$\int_0^\infty |\phi^*(s)|\mathrm{d}s \leq \int_0^\infty |\phi(s)|\mathrm{d}s$$

Hence the limiting feedback law w^* is optimal for the fluid model. □

Property 2 (iii) leads to a candidate function V_0 to inititialize the *value* iteration algorithm. Note that if we by luck chose the initial function V_0 exactly equal to h_* then we would find that $V_n = h_* + n\eta_*$ for each n. While we will never be so lucky in this optimization problem, given (iii) it is very sensible to use the value function V_* given in (29) as an initial condition in the VIA since, at least for large x, this is approximately equal to h_*.

While the function V_* given (29) is not easily computable in general, we can obtain an approximation based upon a finite-dimensional linear program (Humphrey *et al* 1996). If a sufficiently tight approximation to (29) is found then one can expect that (A1) will hold for this approximation.

We illustrate this approach with the three-buffer example illustrated in figure 1. We have computed explicitly the value function V_* for the optimal fluid policy using the optimal fluid policy given (Weiss 1995). For comparison we have also computed the value function $V_0 = V_{\mathrm{LBFS}}$ for the fluid model under the LBFS policy.

Three experiments were performed based upon three different initializations: $V_0 \equiv 0$, which is the standard VIA; $V_0 = V_{\mathrm{LBFS}}$; and $V_0 = V_*$. The results from two experiments are shown in figure 3.

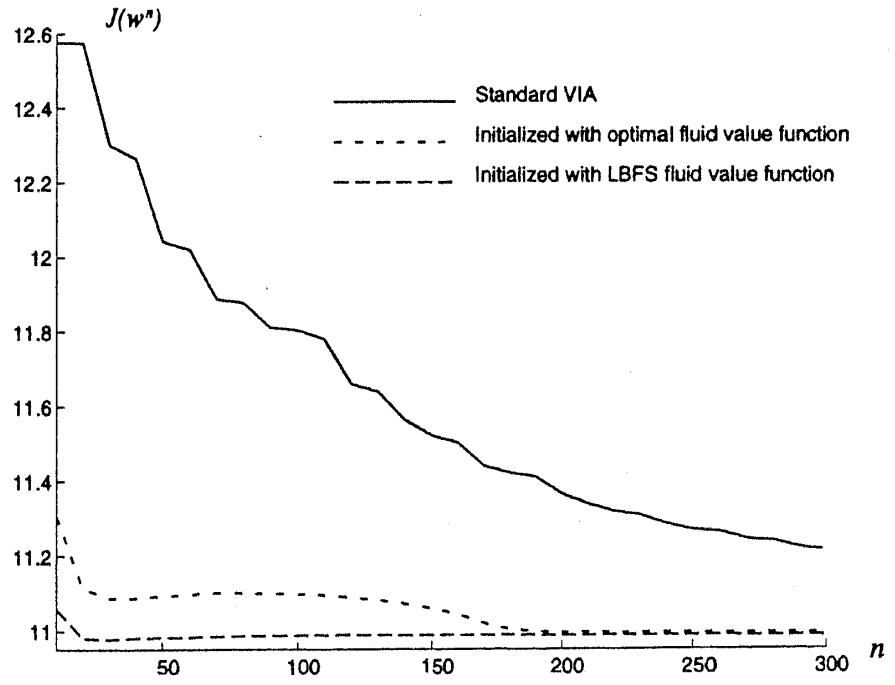

Figure 3. Convergence of the VIA with V_0 taken as the value function for the associated fluid control problem. Two value functions are found: one with the optimal fluid policy, and the other taken with the LBFS fluid policy. Both choices lead to remarkably fast convergence. Surprisingly, the "suboptimal" choice using LBFS leads to the fastest convergence of $J(w^n)$ to the optimal cost $\eta_* \approx 10.9$.

We have taken truncated buffer levels to 33 in the model to obtain a chain with $33^3 = 35{,}937$. In each experiment value iteration was performed for 300 steps, saving data for $n = 10, \ldots, 300$. The parameter values $(\lambda, \mu_1, \mu_2\mu_3)$ are given by [7]. The convergence is slow when $V_0 \equiv 0$, but exceptionally fast in the other two experiments. Surprisingly, the "suboptimal" choice using LBFS leads to the fastest convergence.

In summary, the optimal value function for the fluid model does approximate the minimal relative value function. We see this in a general result, and through the experiment illustrated in figure 3.

The question then is, how can we use this insight in network design? In a real, discrete queueing network with a large number of stations, it is natural to construct a policy which uses the optimal fluid policy to regulate the buffer levls to some desirable, perhaps optimal, mean value. Attempting to maintain some non-zero steady state in this way has two desirable features. One is that when problems arise causing network congestion, the policy will resemble the optimal fluid policy as desired. In normal operating conditions when the state stays near some target level we expect good performance since states are neither too large, nor too small. Small buffer levels are not necessarily desirable since starvation of work to a queue can lead to a loss of resource utilization, burstiness, and even instability.

Some initial experiments on a class of policies based upon these ideas were introduced by (Meyn 1997b), which also includes some positive numerical experiments.

7. Extensions

Many extensions of these results are possible. The general state space case is particularly important in the network problem since it is unfortunate to be forced to assume that service times and interarrival times are exponentially distributed, as we did in the previous section.

Here we describe three situations which lie outside of the class of models which have been considered in this paper. We merely lay some foundation for extending the results of this paper to the more general, or more complex setting. Details may be found in the appropriate references.

7.1 *General state spaces*

The situation where X is uncountable, such as $\mathsf{X} = \mathbb{R}^n$, is typical in many MDPs encountered in practice. The techniques used in this paper can be extended to this more complex setting provided that (A1)–(A3) are modified appropriately.

A complete treatment of the PIA is given by (Meyn 1997) where (A3) is replaced by

(A3′): There is a fixed probability ν on $\mathcal{B}(\mathsf{X})$, a $\delta > 0$, such that for any feedback law w satisfying $\eta_w \leq \eta_0$,

$$K_w(x, A) \geq \delta\nu(A) \quad \text{for all } x \in S,\ n \geq 0,\ A \in \mathcal{B}(\mathsf{X}),$$

where S denotes the sublevel set

$$S = \Big\{x : \min_a c(x, a) \geq 2\eta_0\Big\}. \tag{31}$$

Most of the proofs go through essentially unchanged with the solution to Poisson's equation given by (9) instead of (10) which depends upon an accessible state. Using these ideas it is shown (Meyn 1997a) that the PIA converges for general state space models, and

from this the results of the previous section concerning optimization of networks carry over without modification.

7.2 *Continuous time*

Continuous time models are easily treated using these same techniques, even for general state space models. Again, the main issue is understanding Poisson's equation, which in continuous time takes the form

$$\mathcal{D}_{w^0} h = -c_{w^0} + \eta_{w^0}, \tag{32}$$

where $\mathcal{D}_{w^0}$ denotes the generator of the process under the feedback law w^0. The existence of well-behaved solutions to (32) follows under conditions analogous to the discrete time case (Glynn & Meyn 1996).

To see how the optimality equations appear, and how one can establish existence of an optimal policy, suppose w^* is a feedback law such that for any other feedback law w,

$$c_{w^*} + \mathcal{D}_{w^*} h \leq c_w + \mathcal{D}_w h.$$

If h solves Poisson's equation (32) with $w^0 = w^*$ then the bound above may be written,

$$\eta_{w^*} \leq c_w + \mathcal{D}_w h,$$

and then by the definition of the extended generator,

$$\eta_{w^*} \leq \frac{1}{T}\int_0^T \mathsf{E}_x[c_w(\Phi_t^w)]\mathrm{d}t + \frac{1}{T}(P_w^T h(x) - h(x)).$$

This bound holds for any x, and any $T > 0$. Letting $T \to \infty$, it follows that

$$\liminf_{T\to\infty} \frac{1}{T}\int_0^T \mathsf{E}_x[c_w(\Phi_t^w)]\mathrm{d}t \geq \eta_{w^*},$$

provided that

$$(1/T)P_w^T h(x) \to 0, \quad \text{as } T \to \infty.$$

To construct a solution to Poisson's equation it is simplest to first establish a solution to Poisson's equation for the resolvent defined by

$$R_w = \int_0^\infty e^{-t} P_w^t \mathrm{d}t.$$

Suppose that a solution to the Poisson equation is found for the resolvent:

$$R_w g = g - c_w + \eta.$$

Using the formula $\mathcal{D}_w R_w = R_w - I$ we may set $h = R_w g$ to obtain

$$\mathcal{D}_w h = \mathcal{D}_w R_w g = (R_w - I)g = -c_w + \eta_w.$$

Since stability properties of the continuous time process and the resolvent are essentially equivalent (Down *et al* 1996; Meyn & Tweedie 1992), one can construct solutions to the Poisson equation and the minimal relative value function under assumptions analogous to (A1)–(A3).

The PIA in continuous time is given as follows. Suppose that the policy w^{n-1} is given, and that h_{n-1} satisfies Poisson's equation

$$\mathcal{D}_{n-1}h_{n-1} = -\bar{c}_{n-1},$$

where we again adopt the notation used in the discrete time development. This equation has a solution provided that the chain $\Phi^{w^{n-1}}$ is c_{n-1}-regular (Glynn & Meyn 1996). A new feedback law w^n is then found which satisfies, for any other feedback law w,

$$c_n + \mathcal{D}_{w^n}h_{n-1} \leq c_w + \mathcal{D}_w h_{n-1}.$$

To see when this is an improved policy, let $w = w^{n-1}$: from the inequality above, and Poisson's equation which h_{n-1} is assumed to satisfy, we have

$$\mathcal{D}_{w^n}h_{n-1} \leq -c_n + \eta_{n-1}.$$

This is a stochastic Lyapunov drift inequality, but now in continuous time. Again, if h_{n-1} is bounded from below, if c_n is norm-like, and if compact sets are petite, then the process Φ^{w^n} is c_n- regular, and we have $\eta_n \leq \eta_{n-1}$.

Since the feedback law w^n is so well-behaved, one can assert that Poisson's equation

$$\mathcal{D}_n h_n = -\bar{c}_n,$$

has a solution h_n which is bounded from below, and hence the algorithm can once again begin the policy improvement step. Results analogous to theorem 9 can then be formulated using almost identical methodology (Meyn 1997a).

7.3 *Risk sensitive control*

The techniques described here may also be extended to other control criteria. One which has attracted much recent attention is the "risk sensitive" criterion,

$$R(x,a) = \limsup_{n\to\infty} \frac{1}{n}\log\left(\mathsf{E}_x\left[\exp\left(\sum_{t=0}^{n-1} c(\Phi(t), a(t))\right)\right]\right). \tag{33}$$

The use of the exponential reduces the probability of large excursions of the state since the one-step cost c satisfies condition (A2).

Under certain conditions on the model, in particular, under the assumption of linearity, the controls that optimize (33) are known to be robust to model uncertainty. In general, it may be shown that any stationary policy which gives rise to a finite risk senstive cost will enjoy some attractive properties. In particular, the controlled chain is geometrically ergodic (Borkar & Meyn 1998), which itself implies some degree of robustness to model uncertainty (Glynn & Meyn 1996).

The first step in solving this control problem is to define an appropriate analogue of the relative value function. We first define the cost over one cycle via

$$\Lambda_w := \inf\left\{\Lambda \geq 0 : \mathsf{E}_\theta^w\left[\exp\left(\sum_{t=0}^{\tau_\theta - 1}[c(\Phi(t), w^t(\Phi(t))) - \Lambda]\right)\right] \leq 1\right\}.$$

We then let $\Lambda_* := \inf \Lambda_w$, where the infimum is over all Markov policies, and set $\lambda_* = \exp(\Lambda_*)$. It is shown by (Borkar & Meyn 1998) that $R(x,w) \geq \Lambda_*$ for any Markov policy.

A candidate relative value function and optimal policy are defined respectively as follows: For each $x \in \mathsf{X}$,

$$h_*(x) := \min_w \mathsf{E}_x^w \left[\exp\left(\sum_{t=0}^{\tau_\theta - 1} [c(\Phi(t), w^t(\Phi(t))) - \Lambda_*] \right) \right] \tag{34}$$

$$w^*(x) := \arg\min_{a \in \mathcal{A}} (\exp(c(x,a)) P_a h_*(x)), \tag{35}$$

where in (35) the feedback law w^* is taken to be any solution to the minimization.

With these definitions we find the following

PROPOSITION 4

Suppose that (A1)–(A3) *hold, and suppose that* $\eta_* < \infty$. *Then the function* h_* *satisfies the following bounds*

(i) $h_*(\theta) = 1$;
(ii) $\exp(c_{w^*}(x)) P h_*(x) \le \lambda_* h_*, \quad x \in \mathsf{X}$;
(iii) $\inf_{x \in \mathsf{X}} h_*(x) > 0$.

Property (i) follows from minimality of Λ_*; (ii) follows from dynamic programming arguments, exactly as in the proof of theorem 5 and (iii) follows from proposition 1 (ii) and Jensen's inequality: for any Markov policy,

$$\log \mathsf{E}_x^w \left[\exp\left(\sum_{t=0}^{\tau_\theta - 1} [c(\Phi(t), w^t(\Phi(t))) - \Lambda_*] \right) \right] \ge \mathsf{E}_x^w \left[\sum_{t=0}^{\tau_\theta - 1} [c(\Phi(t), w^t(\Phi(t))) - \Lambda_*] \right]$$
$$\ge -\Lambda_* B(\Lambda_*).$$

This uniform lower bound then gives a uniform lower bound on h_*. □

It is property (ii) that implies the geometric recurrence for the chain under the policy w^*. This inequality combined with the uniform lower bound (iii) implies that h_* acts as a stochastic Lyapunov function for the process, satisfying a bound far stronger than (7), which implies in particular that

$$\mathsf{E}_x^{w^*} [c_{w^*}(\Phi(t))] \to \pi(c),$$

geometrically fast as $t \to \infty$. See (Borkar & Meyn 1998) for other consequences.

The optimality of w^* is proven by iterating (ii), which gives

$$\mathsf{E}_x^{w^*} \left[\exp\left(\sum_{t=0}^{n-1} [c_{w^*}(\Phi(t)) \right) h_*(\Phi(n)) \right] \le \lambda_*^n h_*(x).$$

Using this and the lower bound (iii) shows that $R(x, w^*) = \Lambda_*$ for any $x \in \mathsf{X}_* := \{x : h_*(x) < \infty\}$. The set X_* is all of X if the chain Φ^{w^*} is irreducible.

When assembled together this leads to the following existence theorem for solutions to a dynamic programming equation.

Theorem 10. *Suppose that* (A1)–(A3) *hold and that* $\Lambda_* < \infty$. *Then*

(i) *The function h_* solves the dynamic programming equation*

$$\min_{a\in\mathcal{A}(x)} \exp(c(x,a))P_a h_*(x) = \lambda_* h_*;$$

(ii) *Any feedback law w^* satisfying*

$$\min_{a\in\mathcal{A}(x)} \exp(c(x,a))P_a h_*(x) = \exp(c_{w^*}(x))\, P_{w^*} h_*(x), \quad x \in \mathsf{X}, \tag{36}$$

is optimal over all Markov policies;

(iii) *Any feedback law w^* satisfying* (36) *minimizes the "relative cost": For any Markov policy w,*

$$h_*(x) = \mathsf{E}_x^{w^*}\left[\exp\left(\sum_{t=0}^{\tau_\theta-1}(c_{w^*}(\Phi(t)) - \Lambda_*)\right)\right]$$
$$\le \mathsf{E}_x^{w}\left[\exp\left(\sum_{t=0}^{\tau_\theta-1}(c(\Phi(t), w^t(\Phi(t))) - \Lambda_*)\right)\right].$$

(iv) *It (h_+, w^+) is any other solution to* (4) *for which $\inf_{x\in\mathsf{X}} h_+(x) > -\infty$, then P_{w^+} has a unique invariant probability π_+, and*

$$h_+(x)/h_+(\theta) = h_*(x), \quad \pi_+(x) > 0;$$
$$h_+(x)/h_+(\theta) \ge h_*(x), \quad \textit{forall}(x).$$

□

Given the relative value function h_* and the analogous theory of the "multiplicative Poisson equation" developed by (Balaji & Meyn 1998) one can then construct multiplicative versions of the PIA and VIA, and prove convergence using approaches similar to those presented here.

Work supported in part by NSF Grant ECS 940372; JSEP grant N00014-90-J- 1270; and a Fulbright Research Fellowship. 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.

References

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

Balaji S, Meyn S P 1998 Multiplicative ergodic theorems for an irreducible Markov chain. (submitted)

Borkar V S 1991 *Topics in controlled Markov chains.* Pitman Research Notes in Mathematics Series #240 (London: Longman Scientific & Technical)

Borkar V S, Meyn S P 1998 Risk sensitive optimal control: Existence and synthesis for models with unbounded cost. *SIAM J. Control Optim.* (submitted)

Cao X 1998 The relation among potentials, perturbation analysis, and Markov decision process. *J. Discrete Event Dyn. Syst.* (to appear)

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

Chen R-R, Meyn S P 1997 Value iteration and optimization of multiclass queueing networks. *Queueing Syst.* (to appear)

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

Down D, Meyn S P, Tweedie R L 1996 Geometric and uniform ergodicity of Markov Processes. *Ann. Probab.* 23: 1671–1691

Glynn P W, Meyn S P 1996 A Lyapunov bound for solutions of Poisson's equation. *Ann. Probab.* 24

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

Kumar P R, Meyn S P 1996 Duality and linear programs for stability and performance analysis of queueing networks and scheduling policies. *IEEE Trans. Autom. Control* AC-41: 4–17

Kumar P R, Seidman T I 1990 Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Autom. Control* AC-35: 289–298

Meyn S P, Tweedie R L 1992 *Generalized resolvents and Harris recurrence of Markov processes. Lecture Notes in Mathematics*. Berlin Springer-Verlag

Meyn S P, Tweedie R L 1993 *Markov chains and stochastic stability.* (London: Springer-Verlag)

Meyn S P 1997a The policy improvement algorithm for Markov decision processes with general state space. *IEEE Trans. Autom. Control* AC-42: 191–196

Meyn S P 1997b Stability and optimization of multiclass queueing networks and their fluid models. In *Proceedings of the summer seminar on "The Mathematics of Stochastic Manufacturing Systems"* (Am. Math. Soc.)

Nummelin E 1984 *General irreducible Markov chains and non-negative operators* (Cambridge, MA: University Press)

Puterman M L 1994 *Markov decision processes* (New York, Wiley)

Rybko A N, Stolyar A L 1992 On the ergodicity of stochastic processes describing open queueing networks. *Probl. Peredachi Inf.* 28: 2–26

Weiss G 1995 Optimal draining of a fluid re-entrant line. In *Stochastic networks IMA volumes in Mathematics and its Applications*, (New York: Springer-Verlag) vol. 71, pp 91–103

P Whittle 1996 *Optimisation: Basics and beyond* (Chichester: John Wiley and Sons)