The Policy Iteration Algorithm for Average Reward Markov Decision Processes with General State Space

Sean P. Meyn, *Senior Member, IEEE*

***Abstract*— The average cost optimal control problem is addressed for Markov decision processes with unbounded cost. It is found that the policy iteration algorithm generates a sequence of policies which are c-regular (a strong stability condition), where c is the cost function under consideration. This result only requires the existence of an initial c-regular policy and an irreducibility condition on the state space. Furthermore, under these conditions the sequence of relative value functions generated by the algorithm is bounded from below and "nearly" decreasing, from which it follows that the algorithm is always convergent. Under further conditions, it is shown that the algorithm does compute a solution to the optimality equations and hence an optimal average cost policy. These results provide elementary criteria for the existence of optimal policies for Markov decision processes with unbounded cost and recover known results for the standard linear-quadratic-Gaussian problem.**

When these results are specialized to specific applications they reveal new structure for optimal policies. In particular, in the control of multiclass queueing networks, it is found that there is a close connection between optimization of the network and optimal control of a far simpler fluid network model.

***Index Terms*—Howard's algorithm, Markov decision processes, multiclass queueing networks, Poisson equation, policy iteration algorithm.**

I. Introduction

THE POLICY iteration algorithm is a recursive procedure which may be used to construct a solution to the average cost optimal control problem for a Markov decision process (MDP). Surveys on MDP's and on the history of such algorithms may be found in [1], [2], [21], and [39]. This paper presents an analysis of the policy iteration algorithm for general models, providing criteria for convergence of the algorithm and hence also the existence of optimal policies. The scheduling problem for discrete and fluid network models is taken as a particular application of the theory. A strong connection is established between the optimal control of the two network models based upon the general results obtained in this paper.

The question of the existence of average cost optimal policies for MDP's has been studied for a number of years. For a control problem with unbounded cost, some of the weakest conditions for the existence of an optimal policy have been developed by Borkar *et al.* [2], [40], [42]. The idea is to construct a solution to the average cost optimality equation by first considering the β-discounted problem with value function

$$V_\beta(x) = \min_w \mathbf{E}\left[\sum_{t=0}^{\infty} \beta^t c_w(\Phi_t^w) | \Phi_0 = x\right]$$

where c_w is the one-step cost using the policy w, $\mathbf{\Phi}^w$ denotes the resulting state process when the policy w is used, x is the initial condition, and the minimum is with respect to all policies. The difference $h_\beta(x) = V_\beta(x) - V_\beta(\boldsymbol{\alpha})$ is considered, where $\boldsymbol{\alpha}$ is some distinguished state, and one then seeks conditions under which h_β converges as $\beta \uparrow 1$ to a solution to the average cost optimality equations, thereby giving an optimal policy [41]. To make this approach work, in [42], [43], and other papers it is assumed that there is a finite-valued function M on the state space and a constant $N > 0$ such that

$$-N \le h_\beta(x) \le M(x) \tag{1}$$

for all states x and all β sufficiently close to unity. This condition appears to be far removed from the initial problem statement. However, it is shown in [40] that many common and natural assumptions imply (1). Similar results are also reported in [2], and applications to control of queues are developed in [47].

Another approach developed by Hordijk in [22] involves a blanket stability assumption. This is expressed as the uniform Lyapunov drift inequality over all policies w

$$\mathbf{E}[V(\Phi_{t+1}^w) | \Phi_t^w = x] \le V(x) - c_w(x), \qquad x \in S^c \tag{2}$$

where V is a positive function on the state space and S is a finite set, or more generally a compact set. It is now well known that a drift inequality of this form implies c-regularity of the controlled Markov chain and hence also stability of the process in a strong sense [33]. In particular, (2) can be used to establish bounds related to (1) and also a solution to the average cost optimality equations [2].

The present paper proceeds in a manner related to the work of Borkar [2] in that we consider cost functions which are large, perhaps in an average sense, whenever the state is "large." However, rather than consider the question of existence, in this paper we focus on the synthesis of optimal policies through the policy iteration algorithm. The policy iteration algorithm was introduced by Howard in the 1950's [25] (where it was introduced as the policy *improvement* algorithm). This is a natural approach to the synthesis of optimal policies in which a succession of policies $\{w_n\}$ are

Manuscript received October 12, 1995; revised October 1, 1996 and April 1, 1997. Recommended by Associate Editor, J. B. Lasserre. This work was supported in part by the NSF under Grant ECS 940372, the University of Illinois Research Board under Grant Beckman 1-6-49749, and the JSEP under Grant N00014-90-J-1270.

The author is with the Coordinated Science Laboratory and the University of Illinois, Urbana, IL 61801 USA (e-mail: s-meyn@uiuc.edu).

Publisher Item Identifier S 0018-9286(97)08156-7.

constructed together with *relative value functions* $\{h_n\}$ and intermediate steady-state costs $\{\eta_n\}$. In the finite state-space case it is known that the algorithm computes an optimal policy in a finite number of steps, but in the general state-space case, or even in the case of countable state spaces, little is known about the algorithm except in special cases. For instance, the paper [19] again imposes a uniform stability condition, and [12] imposes bounds on the relative value functions which are far stronger than (1). The paper [23] considers general action spaces, but the state space is assumed to be finite.

The "uniform Doeblin condition" described in, for instance [1], has a superficial similarity to the minorization condition A3) assumed in the present paper. In the text [15, ch. 7] the authors use a related minorization condition to ensure convergence of the policy iteration algorithm. However, either of these conditions is equivalent to 1-geometric regularity, which requires the existence of a solution to (2) with V bounded (see [33, Th. 16.0.2 and Sec. 6.2]). For many applications of interest, including network models and state-space models on Euclidean space, the geometry of the problem precludes the existence of a bounded Lyapunov function.

We establish in this paper the following bounds on the relative value functions $\{h_n\}$. These results are based upon extensions of [18], [22], [33], and [37] concerning properties of solutions to Poisson's equation.

1) When properly normalized by additive constants, there are constants b_1, b_2 such that for every state x and iteration n

$$0 \le h_n(x) \le b_1(h_{n-1}(x)+1) \le b_2(h_0(x)+1).$$

2) The function h_{n-1} serves as a stochastic Lyapunov function for the chain $\mathbf{\Phi}^{w_n}$, and hence all of the chains are c_n-regular, as defined in [33, ch. 14], where $c_n = c_{w_n}$ is the one-step cost using the policy w_n.
3) The functions $\{h_n\}$ are "almost" decreasing and hence converge pointwise to a limit h.

These properties hold without any blanket stability condition on the controlled processes, except for the existence of an initial stabilizing policy. The only global requirement is an irreducibility condition on a single compact subset of the state space and a natural unboundedness condition for the cost function.

Our main result, Theorem 4.4, establishes the near monotone convergence 3) and the bounds given in 1). As a corollary to 1) and 3) we find that bounds such as (1) follow directly from the structure of the model and the algorithm and need not be assumed *a priori*. Moreover, in Theorem 4.3, which establishes 2), we see that the policy iteration algorithm constructs recursive solutions to a version of the drift inequality (2). Hence the strong stability assumption implied by (2) is superfluous when working with the policy iteration algorithm because the algorithm automatically generates stabilizing policies. It is only necessary to find an initial stabilizing policy to initiate the algorithm.

Given the strong form of convergence of the algorithm, it is then easy to formulate conditions under which the limit gives rise to an optimal policy. Because we consider general state-space models, the main results apply to a diverse range of systems, including the linear Gaussian model, multiclass queueing networks, and partially observed MDP's. The convergence results and the intermediate bounds obtained are new even in the special case of countable state-space models.

The remainder of the paper is organized as follows. In Section II, we present some general definitions for MDP's and for Markov chains which possess an accessible state. In Section III, we derive convergence results for the policy iteration algorithm in the special case where an accessible state exists for the MDP. These results are then generalized to chains on a general continuous state space in Section IV. In Section V, we give conditions under which the limiting relative value function solves the optimality equation and gives rise to an optimal policy, and in Section VI it is shown how all of these results may be "lifted" to the continuous time framework. In Section VII, we present a detailed application of these results to the scheduling problem for multiclass queueing networks.

II. Markov Chains with and Without Control

A. Markov Chains with an Accessible State

While the major emphasis of this paper is on controlled chains, here we set down notation and basic results in the control-free case. Further details may be found in [33]. We consider a Markov chain $\mathbf{\Phi} = \{\Phi_t : t \in \mathbb{Z}_+\}$, evolving on a state space $\mathbf{X}$, with Borel σ-algebra $\mathcal{B}(\mathbf{X})$. The state space $\mathbf{X}$ is taken to be a general, locally compact, and separable metric space, although in examples we typically confine ourselves to subsets of multidimensional Euclidean space. We use $\mathbf{P}_\mu$ and $\mathbf{E}_\mu$ to denote probabilities and expectations conditional on Φ_0 having distribution μ, and $\mathbf{P}_x$ and $\mathbf{E}_x$ when μ is concentrated at x. The one-step transition function is defined to be P, and we also consider the resolvent defined for $x \in \mathbf{X}$ and $A \in \mathcal{B}(\mathbf{X})$ by

$$K(x, A) := \sum_{t=0}^{\infty} (\tfrac{1}{2})^{t+1} P^t(x, A).$$

The constant $\frac{1}{2}$ is not critical—the point is that $0 < \frac{1}{2} < 1$, so that K is finite valued, and $K(x, A) > 0$ if and only if starting from the state x the process has some possibility of reaching the set A.

In the first part of this paper we assume that there is a state $\boldsymbol{\alpha} \in \mathbf{X}$ which is *accessible* in the sense that $K(x, \boldsymbol{\alpha}) > 0$ for every $x \in \mathbf{X}$. A somewhat stronger condition often holds in practice: suppose that there is a *continuous* function $s : \mathbf{X} \to (0, 1)$ for which the lower bound holds

$$K(x, \boldsymbol{\alpha}) \ge s(x), \qquad x \in \mathbf{X}. \tag{3}$$

Equivalently, $K(x, \boldsymbol{\alpha})$ is bounded below, uniformly for x in any fixed compact subset of $\mathbf{X}$. In this case, all compact sets are *petite*, and consequently the Markov chain is a *T-chain* ([33, ch. 6]—see also Section IV-A below). In the countable state-space case, this assumption is much weaker than irreducibility. The assumption that the chain possesses an

accessible state is relaxed in Section VI, where we consider ψ-*irreducible* chains.

Stability of a Markov chain is frequently defined in terms of the following return times to appropriate subsets of the state space $\mathbf{X}$:

$$\sigma_A = \min(t \geq 0 \colon \Phi_t \in A), \qquad \tau_A = \min(t \geq 1 \colon \Phi_t \in A).$$

Suppose that c is a function on $\mathbf{X}$ with $c \geq 1$. A set $S \in \mathcal{B}(\mathbf{X})$ is called c-*regular* if

$$\sup_{x \in S} \mathbf{E}_x \left[\sum_{t=0}^{\tau_{\boldsymbol{\alpha}}-1} c(\Phi_t) \right] < \infty.$$

The Markov chain itself is called c-regular if the state space $\mathbf{X}$ admits a countable covering by c-regular sets (more general terminology suitable for chains which do not possess an accessible state will be introduced in Section IV). A c-regular chain always possesses a unique invariant probability π such that

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

Moreover, this class of chains satisfies numerous sample path and mean ergodic theorems (see [33, ch. 14–17] and Theorem 4.1 below).

Under mild structural conditions on the Markov chain, c-regularity is equivalent to the following extension of Foster's criterion: the existence of a function $V \colon \mathbf{X} \to \mathbb{R}_+$ and a constant $\kappa \in \mathbb{R}_+$ such that

$$PV(x) := \mathbf{E}[V(\Phi_{t+1}) | \Phi_t = x] \leq V(x) - c(x) + \kappa, \quad x \in \mathbf{X}. \tag{4}$$

One such result is given in Theorem 4.2 below. A closely related equation which is central to the theory of average cost optimal control is *Poisson's equation*

$$Ph(x) := \mathbf{E}[h(\Phi_{t+1}) | \Phi_t = x] = h(x) - c(x) + \eta, \quad x \in \mathbf{X} \tag{5}$$

where $\eta = \pi(c)$, and under certain conditions h is determined by c (up to an additive constant). In addition to its application to optimal control, the existence of a solution to (5) can be used to establish the functional central limit theorem and law of the iterated logarithm for the Markov chain $\boldsymbol{\Phi}$ [33, Th. 17.0.1].

When an accessible state $\boldsymbol{\alpha}$ exists, a solution to Poisson's equation is easily found: define the function h by

$$h(x) = \mathbf{E}_x \left[\sum_{t=0}^{\tau_\alpha - 1} \overline{c}(\Phi_t) \right] \tag{6}$$

where $\overline{c} := c - \pi(c)$. This function may be equivalently expressed $h(x) = \mathbf{E}_x[\sum_{t=0}^{\sigma_\alpha} \overline{c}(\Phi_t)] - \overline{c}(\boldsymbol{\alpha})$. If the chain is c-regular so that the function h is finite valued, it follows from the Markov property that h solves (5). Given the similarity between (4) and (5), and the close connection between (4) and c-regularity, it is not surprising that c-regularity is closely related to the existence of solutions to Poisson's equation. In [18], [33], and [37] it is shown that regularity is a simple approach to obtaining bounds on solutions to Poisson's equation for general state-space chains even when an accessible state does not exist. Results of this kind are further developed in the Appendix.

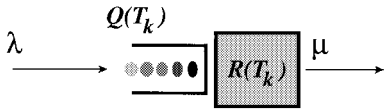

Fig. 1. The $G/G/1$ queue.

Since c is interpreted as a cost function below, it is natural to suppose that this function is *norm-like*—this means that the sublevel set $S_n = \{x \colon c(x) \leq n\}$ is precompact for each n. For a T-chain, c-regularity is then *equivalent* to the existence of a solution to (4) (see Theorem 4.2). Examples of norm-like functions are $c(x) = x^2$ with $\mathbf{X} = (-\infty, \infty)$, or $c(x) = x$ with $\mathbf{X} = [0, \infty)$. However, the function $c(x) = x$ is not norm-like on the state space $\mathbf{X} = (0, \infty)$ since the sublevel set $S_n = (0, n]$ is bounded but not precompact as a subset of $\mathbf{X}$.

A common general state-space model is the $G/G/1$ queue illustrated in Fig. 1. Assume that the arrival stream forms a renewal process and that the service times are i.i.d. with general distribution and finite means. The buffer level is the main quantity of interest. However, to obtain a Markov chain we append the *residual service time* $R(t)$, which is defined to be the remaining service time for the customer in service at time t. Letting $Q(t)$ denote the total number of customers awaiting service at time t, we obtain a Markov chain as follows. Let T_n, $n \geq 0$ denote the arrival times of customers to the queue—we take $T_0 = 0$. Then the process

$$\Phi_n = \begin{pmatrix} Q(T_n) \\ R(T_n) \end{pmatrix}$$

forms a Markov chain on the state space $\mathbf{X} = \mathbb{Z}_+ \times \mathbb{R}_+$. If the arrival rate is less than the service rate ($\rho < 1$), then the system empties with positive probability from any initial condition, and it then follows that (3) holds with $\boldsymbol{\alpha} = 0 \in \mathbf{X}$:

$$K(x, 0) \geq s(x), \qquad x \in \mathbf{X}$$

where the decay rate of the function s depends on the arrival and service distributions. In [31] networks are considered, and the same results are shown to hold true, provided the arrival stream is unbounded. The unboundedness condition can be relaxed for certain classes of networks [5], [17].

If the service times possess a finite-second moment, one can construct a quadratic solution to (4) to establish c-regularity with $c(x) = |x|$. However, for queueing networks, the analysis of an associated fluid model is a more convenient route to establishing regularity of the state process [10], [30].

B. MDP's

We now assume that there is a control sequence taking values in the action space $\mathcal{A}$ which influences the behavior of $\boldsymbol{\Phi}$. The state space $\mathbf{X}$ and the action space $\mathcal{A}$ are assumed to be locally compact separable metric spaces, and we continue to let $\mathcal{B}(\mathbf{X})$ denote the (countably generated) Borel σ-field of $\mathbf{X}$. Associated with each $x \in \mathbf{X}$ is a nonempty and closed subset

$\mathcal{A}(x) \subseteq \mathcal{A}$ whose elements are the admissible actions when the state process Φ_t takes the value x. The set of admissible state–action pairs $\{(x, a)\colon x \in \mathbf{X}, a \in \mathcal{A}(x)\}$ is assumed to be a measurable subset of the product space $\mathbf{X} \times \mathcal{A}$.

The transitions of $\mathbf{\Phi}$ are governed by the conditional probability distributions $\{P_a(x, B)\}$ which describe the probability that the next state is in B for any $B \in \mathcal{B}(\mathbf{X})$ given that the current state is $x \in \mathbf{X}$, and the current action chosen is $a \in \mathcal{A}$. These are assumed to be probability measures on $\mathcal{B}(\mathbf{X})$ for each state–action pair (x, a) and measurable functions of (x, a) for each $B \in \mathcal{B}(\mathbf{X})$. The choice of action when in a state x is governed by a *policy*. A (stationary Markov) policy is simply a measurable function $w\colon \mathbf{X} \to \mathcal{A}$ such that $w(x) \in \mathcal{A}(x)$ for all x. When the policy w is applied to the MDP, then the action $w(x)$ is applied whenever the MDP is in state 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 policy w.

The state process $\mathbf{\Phi}^w := \{\Phi_t^w\colon t \geq 0\}$ of the MDP is, for each fixed policy w, a Markov chain on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$, and we write the t-step transition probabilities for this chain as

$$P_w^t(x, B) = \mathbf{P}(\Phi_t^w \in B | \Phi_0^w = x), \qquad x \in \mathbf{X}, \quad B \in \mathcal{B}(\mathbf{X}) \quad t \in \mathbb{Z}_+.$$

In the controlled case we continue to use the operator-theoretic notation

$$P_w^t h(x) := \mathbf{E}[h(\Phi_t^w) | \Phi_0^w = x].$$

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

$$J(w, x) := \limsup_{N \to \infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{E}_x[c_w(\Phi_t^w)]$$

where $c_w(y) = c(y, w(y))$. A policy w^* will be called *optimal* if $J(w^*, x) \leq J(w, x)$ for all policies w and any initial state x. The policy w is called *regular* if the controlled chain $\mathbf{\Phi}^w$ is a c_w-regular Markov chain. This is a natural and highly desirable stability property for the controlled process: if the policy is regular, then necessarily an invariant probability measure π_w exists such that $\pi_w(c_w) < \infty$. Moreover, for a regular policy, the resulting cost is $J(w, x) = J(w) = \int \pi_w(dy) c_w(y)$, independent of x.

An example of the class of MDP's considered in this paper is the controlled linear system

$$X_{t+1} = AX_t + Bw_t + W_{t+1}, \qquad t \in \mathbb{Z}_+ \tag{7}$$

where W_t, $X_t \in \mathbb{R}^d$ and $w \in \mathbb{R}^p$. The cost c in the linear-quadratic control problem takes the form

$$c(x, w) = \tfrac{1}{2} x^T Q x + \tfrac{1}{2} w^T R w \tag{8}$$

with $Q \geq 0$ and $R > 0$. If $\mathbf{W}$ is i.i.d., then this is an MDP with transition function

$$P_a(x, C) = \mathbf{P}(W_1 + Ax + Ba \in C).$$

The optimization of $J(x, w)$ is known as the linear-quadratic-Gaussian (LQG) problem in the special case where $\mathbf{W}$ is *Gaussian* white noise. Note that the assumption $c \geq 1$ fails in this example. However, c is positive, so that we can add one to the cost function to satisfy the desired lower bound on c and the MDP is essentially unchanged. The linear system (7) will be revisited to illustrate some of the assumptions introduced in the paper.

We assume that there is a large penalty for large control actions or large excursions of the state. This requires that the cost be norm-like. The assumption of norm-like costs has been used extensively in the recent literature; see for instance [1] and [2]. Our main result, Theorem 4.4, relaxes this condition so that the norm-like condition is only required in a time/ensemble-averaged sense. This allows application to the LQG problem where the state weighting matrix Q is not necessarily full rank.

One fruitful approach to finding optimal policies is through the following optimality equations:

$$\eta_* + h(x) = \min_{a \in \mathcal{A}(x)} [c(x, a) + P_a h(x)] \tag{9}$$

$$w^*(x) = \arg\min_{a \in \mathcal{A}(x)} [c(x, a) + P_a h(x)], \qquad x \in \mathbf{X}. \tag{10}$$

Equality (9), a version of Poisson's equation, is known as the *average cost optimality equation* (ACOE). The second equation (10) defines a policy w^*. If a policy w^*, a measurable function h, and a constant η_* exist which solve these equations, then typically the policy w^* is optimal (see for example [1], [2], [20], and [39] for a proof of this and related results).

Theorem 2.1: Suppose that the following conditions hold.

1) The pair (η_*, h) solve the optimality equation (9).
2) The policy w^* satisfies (10), so that

$$c_{w^*}(x) + P_{w^*} h(x) \leq c(x, a) + P_a h(x), \qquad x \in \mathbf{X}, \quad a \in \mathcal{A}(x).$$

3) For any $x \in \mathbf{X}$ and any policy w satisfying $J(w, x) < \infty$

$$\frac{1}{n} P_w^n h(x) \to 0, \qquad n \to \infty.$$

Then w^* is an optimal control and η_* is the optimal cost, in the sense that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[c_{w^*}(\Phi_t^{w^*})] = \eta_*$$

and $J(w, x) \geq \eta_*$ for all policies w and all initial states x. □

We show in this paper that the policy iteration algorithm is an effective approach to establishing the existence of solutions to the optimality equations for general state-space processes. Because it exhibits nearly monotone convergence, in some cases it gives a practical algorithm for the computation of optimal policies even when the state space is not finite. In the special case of network scheduling we find that it gives insight into the structure of optimal policies.

III. The Policy Iteration Algorithm

The policy iteration algorithm (PIA) is a method for successively computing increasingly well-behaved policies for an MDP. The important features of the algorithm can be explained

in a few paragraphs. Suppose that a policy w_{n-1} is given, and assume that h_{n-1} satisfies the Poisson equation

$$P_{n-1}h_{n-1} = h_{n-1} - c_{n-1} + \eta_{n-1}$$

where $P_{n-1} = P_{w_{n-1}}$, $c_{n-1}(x) = c_{w_{n-1}}(x) = c(x, w_{n-1}(x))$, and η_{n-1} is a constant (presumed to be the steady-state cost with this policy). The relative value functions $\{h_n\}$ are not uniquely defined: If h_n satisfies Poisson's equation, then so does $h_n + b$ for any constant b. The main results to follow all apply to specific normalized versions of the relative value functions.

Given h_{n-1}, one then attempts to find an improved policy w_n by choosing, for each x

$$w_n(x) = \arg\min_{a \in \mathcal{A}(x)} [c(x, a) + P_a h_{n-1}(x)]. \tag{11}$$

Once w_n is found, polices $w_{n+1}, w_{n+2}, \cdots$ may be computed by induction, so long as the appropriate Poisson equation may be solved, and the minimization above has a solution.

To analyze the PIA we consider the pair of equations

$$P_n h_n = h_n - \overline{c}_n \tag{12}$$

$$P_n h_{n-1} = h_{n-1} - \overline{c}_n - \gamma_n \tag{13}$$

where $\overline{c}_n = c_n - \eta_n$, and γ_n is *defined* through (13). From the minimization (11) we have

$$c_n + P_n h_{n-1} \le c_{n-1} + P_{n-1} h_{n-1}$$

and from Poisson's equation we have

$$c_{n-1} + P_{n-1} h_{n-1} = h_{n-1} + \eta_{n-1}.$$

Combining these two equations gives the lower bound $\gamma_n(x) \ge \eta_n - \eta_{n-1}$, $x \in \mathbf{X}$. We will show that the sequence $\{\eta_n\}$ converges monotonically, which together with this lower bound shows that the limit inferior of $\{\gamma_n\}$ is bounded from below by zero. Under suitable conditions we also show that the sequence $\{h_n\}$ is bounded from below, and this then gives an upper bound on $\{\gamma_n\}$. Since $\pi_{w_n}(c_n) = \eta_n$, it follows from the Comparison theorem [33, p. 337] that $\pi_{w_n}(\gamma_n) \le 0$. Thus, for large n, the error term γ_n is small, and hence the function h_{n-1} *almost* solves the Poisson equation for P_n. One might then expect that h_n will be close to h_{n-1}. Under mild conditions, this is shown to be true in a very strong sense. In Theorems 3.1 and 4.4 below we establish the following remarkable properties of the PIA.

P1) *Uniform Boundedness from Below*: For some constant $0 < N < \infty$

$$\inf_{x \in \mathbf{X},\, n \ge 0} h_n(x) > -N.$$

P2) *Almost Decreasing Property*: There exists a sequence of functions $\{g_n\colon n \ge 0\}$ such that

$$g_n(x) \le g_{n-1}(x) \le \cdots \le g_0(x), \qquad x \in \mathbf{X}, \quad n \ge 0$$

and for some sequence of positive numbers $\{\alpha_k, \beta_k\}$

$$g_n(x) = \alpha_n h_n(x) + \beta_n, \qquad n \ge 0, \quad x \in \mathbf{X}$$

with $\alpha_k \downarrow 1$, $\beta_k \downarrow 0$ as $k \to \infty$.

These properties together imply that the relative value functions are pointwise convergent to the function $h(x) := \lim_n g_n(x)$.

The uniformity in the results P1) and P2) requires a subtle analysis of the sequence of processes $\mathbf{\Phi}^{w_n}$. In particular, we consider a "split" chain on an extended state space to derive our main results. It is now well known that for a ψ-irreducible chain, it is possible to construct an "atom" $\breve{\boldsymbol{\alpha}} \in \mathcal{B}^+(\mathbf{X})$ on an extended state space with the property that $P(x, \cdot) = P(y, \cdot)$ for $x, y \in \breve{\boldsymbol{\alpha}}$. In this section only we assume that in fact a singleton $\boldsymbol{\alpha} \in \mathbf{X}$ exists which is accessible for each chain. In Theorem 3.1, we demonstrate the desired properties of the PIA in a simplified setting where the action space is countable and an accessible state exists. A far more general result and a complete proof are given in Section IV. In Section V, we provide conditions which guarantee that the limiting policy w is optimal.

It is convenient to work with the family of resolvent kernels $\{K_n\}$ defined as

$$K_n := \sum_{t=0}^{\infty} (\tfrac{1}{2})^{t+1} P_n^t, \qquad n \ge 0. \tag{14}$$

We initially assume in Assumption 2) below that a condition of the form (3) holds for a continuous function s so that each of the chains $\mathbf{\Phi}^{w_n}$ is a T-chain. Condition 1) is related to the "near-monotone" condition of [2] in the case of countable state-space models.

Theorem 3.1: Suppose that an initial regular policy w_0 exists with average cost $\eta_0 = \pi_{w_0}(c_{w_0})$, and suppose that the MDP satisfies the following.

1) The cost function c is norm-like on the product space $\mathbf{X} \times \mathcal{A}$, and there exists a norm-like function $\underline{c}\colon \mathbf{X} \to \mathbb{R}_+$ such that $c(x, a) \ge \underline{c}(x)$ for any $x \in \mathbf{X}$, $a \in \mathcal{A}(x)$.
2) There is a state $\boldsymbol{\alpha} \in \mathbf{X}$ and a continuous function $s\colon \mathbf{X} \to (0, 1)$ with the following property: if, for any $n \ge 1$, the triplet $(w_{n-1}, h_{n-1}, \eta_{n-1})$ is generated by the PIA with initial policy w_0, then for any policy w_n which satisfies (11)

$$K_n(x, \boldsymbol{\alpha}) \ge s(x), \qquad \text{for all } x \in \mathbf{X}. \tag{15}$$

3) The action space is countable.

Then for each n, the algorithm admits a solution (w_n, h_n, η_n) such that each policy w_n is regular, and the sequence of relative value functions $\{h_n\}$ given by

$$h_n(x) = \mathbf{E}_x \left[\sum_{t=0}^{\tau_\alpha - 1} \overline{c}_n(\Phi_t^{w_n}) \right], \qquad n \ge 0 \tag{16}$$

is finite-valued and satisfies properties P1) and P2).

Proof: We state the proof tersely here since a more general result is proven below. From (13), it is possible to show inductively that h_{n-1} acts as a stochastic Lyapunov function, i.e., a solution to (4), for the policy w_n. One can then deduce from the Comparison Theorem of [33] that the PIA generates regular policies and that the sequence $\{\eta_i\}$ is decreasing.

Let S denote the precompact set

$$S = \{x\colon \underline{c}(x) \le 2\eta_0\}. \tag{17}$$

Given the assumptions of the theorem, there exists a $\delta > 0$ such that $K_n(x, \boldsymbol{\alpha}) \ge \delta$ for any $x \in S$. For any n, by regularity and Theorem A.1 the function h_{n-1} is necessarily bounded from below, and by (13) it satisfies the inequality

$$P_n h_{n-1} \le h_{n-1} - \tfrac{1}{2}\, c_n + \eta_{n-1}\mathbf{1}_S \tag{18}$$

where we have used the fact that $\{x\colon c_n(x) \le \eta_{n-1}\} \subset S$, where the set S is defined in (17). It follows from Dynkin's formula and Fatou's lemma, as in the proof of [33, Proposition 11.3.2], that

$$\mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} c_n(\Phi_t^{w_n})\right] \le 2\Bigg(h_{n-1}(x) - h_{n-1}(\boldsymbol{\alpha}) + \eta_{n-1}\mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} \mathbf{1}_S(\Phi_t^{w_n})\right]\Bigg). \tag{19}$$

The second term in parentheses is eliminated on using $h_{n-1}(\boldsymbol{\alpha}) = 0$. The last term can be bounded using the arguments of the proof of [33, Th. 11.3.11] to obtain

$$\mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} \mathbf{1}_S(\Phi_t^{w_n})\right] \le \mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} K_n(\Phi_t^{w_n}, S)/\delta\right] \le 1/\delta.$$

From these observations and the inequality $\eta_{n-1} \le \eta_0$, (19) may be replaced by

$$\mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} c_n(\Phi_t^{w_n})\right] \le 2[h_{n-1}(x) + \eta_0/\delta]. \tag{20}$$

In particular, this shows that P1) holds with $N = \eta_0/\delta$.

Using the inequality $P_n h_{n-1} \le h_{n-1} - c_n + \eta_{n-1}$ and Dynkin's formula gives an upper bound on h_n

$$\begin{aligned} h_n(x) &= \mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} \bar{c}_n(\Phi_t^{w_n})\right] \\ &\le h_{n-1}(x) - h_{n-1}(\alpha) + (\eta_{n-1} - \eta_n)\mathbf{E}_x[\tau_\alpha] \\ &\le [1 + 2(\eta_{n-1} - \eta_n)]h_{n-1}(x) + 2(\eta_0/\delta)(\eta_{n-1} - \eta_n) \end{aligned}$$

where we have used the inequality $\mathbf{E}_x[\tau_\alpha] \le 2[h_{n-1}(x) + \eta_0/\delta]$, which follows from (20). Thus, for all n and x we have

$$-\eta_0/\delta \le h_n(x) \le (1 + \varepsilon_n)h_{n-1}(x) + (\eta_0/\delta)\varepsilon_n \tag{21}$$

where $\varepsilon_n = 2(\eta_{n-1} - \eta_n)$. The proof of P2) follows on letting

$$g_n(x) = \left(\prod_{t=n+1}^{\infty}(1 + \varepsilon_t)\right)\left(h_n(x) + (\eta_0/\delta)\sum_{t=n+1}^{\infty}\varepsilon_t\right). \tag{22}$$

□

It is well known that the average cost optimal control problem is plagued with counterexamples [1], [11], [39], [41]. It is of some interest then to see why Theorem 3.1 does not fall into any of these traps. Consider first [41, p. 142, counterexamples 1 and 2]. In each of these examples the process, for any policy, is completely nonirreducible in the sense that $\mathbf{P}(\Phi_t^w < \Phi_0^w) = 0$ for all times t and all policies w. It is clear then from the cost structure that the bound (15) on the resolvent cannot hold. A third example is given in [41, Appendix]. Here (15) is directly assumed! However, the cost is not unbounded and is in fact designed to favor large states.

Assumptions 1) and 2) together imply that the center of the state space, as measured by the cost criterion, possesses some minimal amount of irreducibility, at least for the policies $\{w_n\}$. If either the unboundedness condition or the accessibility condition is relaxed, so that the process is nonirreducible on a set where the cost is low, then we see from these counterexamples that optimal stationary policies may not exist.

IV. Convergence for General Processes

To relax the condition that an accessible state exists we consider a split chain for the process with transition function K_n defined in (14). This also allows a relaxation of the norm-like condition on the cost function c. To begin, we must introduce some terminology from the theory of ψ-irreducible chains taken from [33].

A. General State-Space Markov Chains

The Markov chain $\boldsymbol{\Phi}$ with transition function P is called *ψ-irreducible* if the resolvent kernel satisfies K for some measure ψ

$$K(x, A) > 0,\ x \in \mathbf{X} \quad \Longleftrightarrow \quad \psi(A) > 0.$$

We then call ψ a *(maximal) irreducibility measure*. We let $\mathcal{B}^+$ denote the set of $A \in \mathcal{B}(\mathbf{X})$ for which $\psi(A) > 0$. If the chain is ψ-irreducible, then from any initial condition x, the process has a chance of entering any set in $\mathcal{B}^+(\mathbf{X})$ in the sense that $\mathbf{P}_x\{\tau_A < \infty\} > 0$.

We call a set $C \in \mathcal{B}(\mathbf{X})$ *petite* if for some probability ν on $\mathcal{B}(\mathbf{X})$ and $\delta > 0$

$$K(x, A) \ge \delta\nu(A), \qquad x \in C, \quad A \in \mathcal{B}(\mathbf{X}).$$

Equivalently, for a ψ-irreducible chain, the set C is petite if for each $A \in \mathcal{B}^+(\mathbf{X})$, there exists $n \ge 1$ and $\delta > 0$ such that

$$\mathbf{P}_x(\tau_A \le n) \ge \delta, \qquad \text{for any } x \in C. \tag{23}$$

For a ψ-irreducible chain, there always exists a countable covering of the state space by petite sets.

To connect the topology of the state space with measure-theoretic properties of the Markov chain, we typically assume that all compact sets are petite, in which case the Markov chain is called a T-chain. The Markov chain is a ψ-irreducible T-chain if and only if there exists a continuous function $s\colon \mathbf{X} \to (0, 1)$ and a probability measure ν such that

$$K(x, A) \ge s(x)\nu(A), \qquad x \in \mathbf{X}, \quad A \in \mathcal{B}(\mathbf{X}). \tag{24}$$

This lower bound is analogous to (3) and is used to construct the artificial atom $\boldsymbol{\alpha}$ described earlier.

A ψ-irreducible chain is called *Harris* if $\mathbf{P}_x\{\tau_A < \infty\} = 1$ for any $A \in \mathcal{B}^+(\mathbf{X})$ and any $x \in \mathbf{X}$. If, in addition, the chain admits an invariant probability measure π, then the chain is called *positive Harris*.

Suppose that $c\colon \mathbf{X} \to [1, \infty)$ is a function on the state space. For a ψ-irreducible chain, a set $S \in \mathcal{B}(\mathbf{X})$ is called *c-regular* if for any $A \in \mathcal{B}^+(\mathbf{X})$

$$\sup_{x \in S} \mathbf{E}_x \left[\sum_{t=0}^{\tau_A - 1} c(\Phi_t) \right] < \infty.$$

From the characterization in (23) we see that a c-regular set is always petite. The Markov chain is again called c-regular if the state space $\mathbf{X}$ admits a countable covering by c-regular sets. The definition of c-regularity given here generalizes the previous definition which depends upon the existence of an accessible state, and virtually every result can still be proven in this more general context. In particular, a c-regular chain is automatically positive Harris, it always possesses a unique invariant probability π which satisfies $\pi(c) < \infty$, and we have the following consequence of the f-Norm Ergodic theorem of [33, Th. 14.0.1].

Theorem 4.1: Assume that $c\colon \mathbf{X} \to [1, \infty)$ and that $\mathbf{\Phi}$ is c-regular. Then, for any measurable function g which satisfies

$$\sup_{x \in \mathbf{X}} \left(\frac{|g(x)|}{c(x)} \right) < \infty$$

the following ergodic theorems hold for any initial condition:

1) $$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g(\Phi_t) = \pi(g), \qquad \text{a.s. } [\mathbf{P}_x]$$

2) $$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[g(\Phi_t)] = \pi(g).$$

□

Theorem 4.2 shows that the strong c-regularity property is equivalent to the generalization (4) of Foster's criterion.

Theorem 4.2: Assume that $c\colon \mathbf{X} \to [1, \infty)$ is norm-like and that all compact subsets of $\mathbf{X}$ are petite. Then:

1) if there exists a finite, positive-valued solution V to (4), then for each $A \in \mathcal{B}^+(\mathbf{X})$, there exists a $d(A) < \infty$ such that

$$\mathbf{E}_x \left[\sum_{t=0}^{\tau_A} c(\Phi_t) \right] \le 2V(x) + d(A), \qquad x \in \mathbf{X}. \tag{25}$$

Hence, each of the sublevel sets $S_n = \{x\colon V(x) \le n\}$ is c-regular, and the process itself is c-regular;

2) if the chain is c-regular, then for any c-regular set $S \in \mathcal{B}^+(\mathbf{X})$, the function

$$V(x) = \mathbf{E}_x \left[\sum_{t=0}^{\sigma_S} c(\Phi_t) \right], \qquad x \in \mathbf{X} \tag{26}$$

is a norm-like solution to (4).

Proof: The result is essentially known. Bound (4) is equivalent to the drift condition $PV_0 \le V_0 - c + b\mathbf{1}_K$, where K is compact. If (4) holds, we can take $V_0 = 2V$ and $b = 2\kappa$. The result is then an immediate consequence [33, Th. 14.2.3]. □

Consider for example the linear stochastic system

$$X_{t+1} = AX_t + W_{t+1}, \qquad t \in \mathbb{Z}_+ \tag{27}$$

where X_t, $W_t \in \mathbb{R}^d$, and $\mathbf{W}$ is i.i.d. with $W_t \sim N(0, \Sigma)$. Let F by any matrix of suitable dimension satisfying $FF^T = \Sigma$. If F is $d \times q$ for some q, then the *controllability matrix* is the $d \times (dq)$ matrix $\mathcal{C} := [A^{d-1}F | A^{d-2}F | \cdots | AF | F]$, and the pair (A, F) is called *controllable* if the matrix $\mathcal{C}$ has rank d. The process is ψ-irreducible with ψ equal to Lebesgue measure if the pair (A, F) is controllable, since in this case $P^t(x, \cdot)$ is equivalent to Lebesgue measure for any x and any $t \ge d$. By continuity of the model it is easy to check that (24) holds with s continuous and ν equal to normalized Lebesgue measure on an open ball in $\mathbb{R}^d$. We conclude that all compact sets are petite if the controllability condition holds. To find a stochastic Lyapunov function V with $c(x) = x^T Q x$, first solve the Lyapunov equation

$$A^T P A = P - Q.$$

If $P \ge 0$, then $V(x) = x^T P x$ is a solution to (4).

For the nonlinear state-space model

$$\Phi_{t+1} = F(\Phi_t, W_{t+1}), \qquad t \in \mathbb{Z}_+$$

the ψ-irreducibility condition can still be verified under a nonlinear controllability condition called *forward accessibility* [33, ch. 7].

B. Stability and Convergence of the PIA

We now return to controlled Markov chains and present our main results. Since these kernels greatly simplify further analysis of the PIA, much of our analysis focuses on $\{K_n\}$ rather than $\{P_n\}$. This is possible because if (12) and (13) hold, we have the analogous pair of equations

$$K_n h_n = h_n - K_n \bar{c}_n \tag{28}$$

$$K_n h_{n-1} \le h_{n-1} - K_n \bar{c}_n + \eta_{n-1} - \eta_n. \tag{29}$$

To invoke the algorithm we must ensure that the required minimum exists. The following condition holds automatically under appropriate continuity conditions. See Theorem A.4 for results in this direction.

A1) For each n, if the PIA yields a triplet $(w_{n-1}, h_{n-1}, \eta_{n-1})$ which solves Poisson's equation

$$P_{n-1} h_{n-1} = h_{n-1} - c_{n-1} + \eta_{n-1}$$

with h_{n-1} bounded from below, then the minimization

$$w_n(x) := \arg\min_{a \in \mathcal{A}(x)} [c(x, a) + P_a h_{n-1}(x)]$$

admits a measurable solution w_n.

Condition A1) may be relaxed by taking a "near minimizer" w_n such that

$$c_n(x) + P_n h_{n-1}(x) \le c(x, a) + P_a h_{n-1}(x) + e_n \qquad x \in \mathbf{X},\ a \in \mathcal{A}(x)$$

where $\{e_n\}$ is a positive, summable sequence. The development to follow then requires only superficial modifications.

Condition A2) relates the average optimality problem with the discounted optimal control problem; this assumption is satisfied if the state dependent cost $V_{1/2}$ is norm-like, where

$V_{1/2}(x)$ denotes the optimal cost for the discounted optimal control problem when the discount factor is one-half, and the initial condition is x.

A2) For each fixed x, the function $c(x, \cdot)$ is norm-like on $\mathcal{A}$, and there exists a norm-like function $\underline{c}\colon \mathbf{X} \to \mathbb{R}_+$ such that for the policies w_n obtained through the PIA

$$\infty > K_n c_n(x) \geq \underline{c}(x), \qquad \text{for any } x \in \mathbf{X}, \quad n \in \mathbb{Z}_+.$$

Under Assumptions A1) and A2), the algorithm produces stabilizing policies recursively. The proof of Theorem 4.3 may be found in the Appendix.

Theorem 4.3: Suppose that A1) and A2) hold and that for some n, the policies $\{w_i\colon i < n\}$ and relative value functions $\{h_i\colon i < n\}$ are defined through the PIA. Suppose moreover that:

1) the relative value function h_i is bounded from below, $i \leq n-1$;
2) all compact sets are petite for the Markov chains $\{\mathbf{\Phi}^{w_i}, i \leq n-1\}$, and for $\mathbf{\Phi}^{w_n}$, where w_n is a policy given in A1).

Then, the PIA admits a solution (w_n, h_n, η_n) such that

1) the relative value function h_n is bounded from below;
2) all of the policies $\{w_i\colon i \leq n\}$ are regular;
3) for all $0 \leq i \leq n$ the constant η_i is the cost at the ith stage, $\eta_i = J(w_i)$, and the costs are decreasing

$$\eta_0 \geq \eta_1 \geq \cdots \geq \eta_n.$$

□

To obtain convergence of the algorithm, we strengthen Assumption 2) of Theorem 4.3 to the following uniform accessibility condition.

A3) There is a fixed probability ν on $\mathcal{B}(\mathbf{X})$, a $\delta > 0$, and an initial regular policy w_0 with the following property. For each $n \geq 1$, if the PIA yields a triplet $(w_{n-1}, h_{n-1}, \eta_{n-1})$ with h_{n-1} bounded from below, then for any policy w_n given in A1)

$$K_n(x, A) \geq \delta\nu(A) \qquad \text{for all } x \in S, \quad n \geq 0, \quad A \in \mathcal{B}(\mathbf{X}) \tag{30}$$

where S denotes the precompact set

$$S = \{x\colon \underline{c}(x) \leq 2\eta_0\}. \tag{31}$$

Under A3), the kernel $K_n - s \otimes \nu$ is positive, where $s\colon \mathbf{X} \to \mathbb{R}_+$ is defined as $s(x) = \delta \mathbf{1}_S(x)$, and

$$(K_n - s \otimes \nu)(x, A) := K_n(x, A) - \delta \mathbf{1}_S(x)\nu(A), \quad x \in \mathbf{X}, \quad A \in \mathcal{B}(\mathbf{X}). \tag{32}$$

Hence the following *potential kernel* is well defined:

$$G_n(x, A) := \sum_{t=0}^{\infty} (K_n - s \otimes \nu)^t(x, A), \qquad x \in \mathbf{X}, \quad A \in \mathcal{B}(\mathbf{X}) \tag{33}$$

although it may take on positive-infinite values. This kernel has the interpretation

$$G_n(x, A) = \mathbf{E}_x\left[\sum_{t=0}^{\sigma_\alpha} \mathbf{1}_A(\Psi_t^{w_n})\right]$$

where $\mathbf{\Psi}^{w_n}$ is a Markov chain with transition function K_n, and τ_α is the first return time to the atom $\breve{\alpha}$ on an extended state space. We also have $G_n s(x) = \mathbf{P}_x\{\sigma_\alpha < \infty\} \leq 1$ [33], [36].

If the assumptions of Theorem 4.3 hold so that the relative value functions are bounded from below, it then follows from [37, Corollary 3.2], Theorem A.1, and Theorem A.3 that the minimal solution h_n to Poisson's equation (28) which is bounded from below, and which satisfies $\nu(h_n) = 0$, is

$$h_n(x) = G_n K_n \bar{c}_n(x) := \iint_{\mathbf{X}\times\mathbf{X}} G_n(x, dy) K_n(y, dz) \bar{c}_n(z). \tag{34}$$

These ideas are used in the Appendix to prove the main result of this paper.

Theorem 4.4: Under A1)–A3), for each n the PIA admits a solution (w_n, h_n, η_n) such that w_n is regular, and the sequence of relative value functions $\{h_n\}$ given in (34) satisfies Properties P1) and P2). □

To give a more concrete interpretation of the assumptions of Theorem 4.4, consider (7). Condition A1), which demands the existence of a minimizing policy, is satisfied because the model is continuous. The assumption of an initial regular policy in A3) is simply stabilizability of (A, B), and the accessibility condition (30) holds if the noise is full rank. We show here that the norm-like condition A2) is implied by the standard observability condition on $(A, \sqrt{Q})$, where Q is the state weighting matrix given in (8).

Assume that $\mathbf{W}$ is i.i.d. and Gaussian with $W_t \sim N(0, \Sigma)$. To verify A2), we show that the value function $V_{1/2}$ for the optimal 1/2-discounted control problem is norm-like on $\mathbf{X}$ and then take $\underline{c}(x) = \frac{1}{2}V_{1/2}(x)$. By definition, $V_{1/2}(x) = 2\min_w \{K_w c_w(x)\}$, and if $w^\star$ achieves this minimum then

$$V_{1/2}(x) = \sum_{t=0}^{\infty} 2^{-n} \mathbf{E}_x^{w^\star}[x_t^T Q x_t + u_t^T R u_t].$$

The optimal policy for the discounted optimal control problem is linear, and consequently it is easy to formulate criteria under which $V_{1/2}$ is norm-like. To see this, let $u = w^\star(x) = -K_\star x$ denote the optimal control for the discounted control problem. Then, the value function $V_{1/2}$ is of the form

$$V_{1/2}(x) = x^T \Lambda x + V_{1/2}(0)$$

where Λ is positive semidefinite. To show that $V_{1/2}$ is norm-like, we must prove that $\Lambda > 0$. The function $V_{1/2}$ satisfies the dynamic programming equation

$$\tfrac{1}{2} P_{w^\star} V_{1/2} = V_{1/2} - c_{w^\star}$$

or equivalently

$$\tfrac{1}{2} V_{1/2}(0) + \tfrac{1}{2} \mathbf{E}_x[x_1^T \Lambda x_1] = x^T \Lambda x + V_{1/2}(0) - x^T[Q + K_\star^T R K_\star]x.$$

Given the closed-loop system description $x_1 = (A-BK_\star)x + W_1$, it follows that $V_{1/2}(0) = 2\,\text{trace}\,(\Lambda\Sigma)$ and that Λ satisfies the Lyapunov equation

$$\tfrac{1}{2}\,A_\star^T \Lambda A_\star + Q_\star = \Lambda$$

where $A_\star = A - BK_\star$, and $Q_\star = Q + K_\star^T R K_\star$. From this it can be shown that observability of $(A, \sqrt{Q})$ is sufficient to guarantee positivity of Λ [29].

To verify A3), suppose the initial policy w_0 is linear so that each subsequent policy is of the form $w_n(x) = -K_n x$, and let $A_n = A - BK_n$ denote the closed-loop system matrix. Then the accessibility condition A3) holds if $\Sigma > 0$, and the steady-state costs $\{\eta_n\}$ are bounded. The boundedness condition holds automatically under A1) and A2) through stability of the algorithm, which is guaranteed by Theorem 4.3. In the nonlinear setting a noise controllability condition implies A3) using the approach of [33, ch. 7].

From these results it follows that Theorem 4.4 recovers known properties of the Newton–Raphson technique applied to the LQG problem. The well-known decreasing property of the solutions $\{\Lambda_n\}$ to the associated Riccati equation is also closely related to P2). The proof of P2) given in the Appendix depends upon the bound

$$h_n(x) \le [1 + 2(\eta_{n-1} - \eta_n)]h_{n-1}(x) + b_n \qquad (35)$$

where b_n is a constant. In the linear case, it can be shown that the relative value function h_n takes the form $h_n(x) = h_n(0) + x^T\Lambda_n x$ whenever the controls are linear. Letting $x \to \infty$, it follows from the previous inequality that

$$\Lambda_n \le [1 + 2(\eta_{n-1} - \eta_n)]\Lambda_{n-1}, \qquad n \ge 1. \qquad (36)$$

It may be shown directly that $\Lambda_n \le \Lambda_{n-1}$ [14], [49], so the bound (35) is not tight in the linear model. However, the semi-decreasing property (36) is sufficient to deduce convergence of the algorithm.

V. Optimality

Now that we know that $\{h_n\}$ is pointwise convergent to a function h, we can show that the PIA yields an optimal policy. Theorem 5.1 is similar to [20, Th. 4.3] which also requires a continuity condition related to A4). Weaker conditions are surely possible for a specific application.

A4) The function c: $\mathbf{X} \times \mathcal{A} \to [1, \infty)$ is continuous, and the functions $(P_a h_n\,(x)$: $n \ge 0)$ and $P_a h\,(x)$ are continuous in a for any fixed $x \in \mathbf{X}$.

Theorem 5.1: Suppose that there exists a regular policy w_0 and that Assumptions A1)–A4) hold. Then:

1) for the policy w_0, the PIA produces a sequence of solutions (w_n, h_n, η_n) such that $\{h_n\}$ is pointwise convergent to a solution h of the optimality equation (9) and any policy w which is a pointwise limit of w_n satisfies (10). Moreover, the costs $\{\eta_n\}$ are decreasing with n;
2) any limiting policy w is c_w-regular so that for any initial condition $x \in \mathbf{X}$

$$J(w) = \eta_w = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[c_w(\Phi_t^w)].$$

Proof: We first establish the upper bound

$$\eta + h(x) \le c_w(x) + P_w h(x), \qquad x \in \mathbf{X}. \qquad (37)$$

Assuming that w is a pointwise limit of $\{w_n\}$, it follows that for each x there is a subsequence $\{n_i(x)\}$ such that $w_{n_i(x)}(x) \to w(x)$ as $i \to \infty$. Observe that from Poisson's equation

$$\begin{aligned}\eta + h(x) - c_w(x) &= \lim_{i\to\infty} P_{n_i(x)} h_{n_i}(x) \\ &= \lim_{i\to\infty} \frac{1}{\alpha_{n_i}} P_{n_i(x)} g_{n_i}(x) - \frac{\beta_{n_i}}{\alpha_{n_i}} \\ &\le \lim_{i\to\infty} P_{n_i(x)} g_{n_{k_0}}(x) \\ &\le P_w g_{n_{k_0}}(x)\end{aligned}$$

where k_0 is arbitrary, and the first inequality is a consequence of the fact that $\{g_n\}$ is decreasing in n. By dominated convergence, it then follows that (37) does hold. Conversely, we have that

$$\begin{aligned}\eta_{n-1} + h_{n-1}(x) - c_n(x) &\ge P_n h_{n-1}(x) \\ &= \frac{1}{\alpha_n} P_n g_{n-1}(x) - \frac{\beta_{n-1}}{\alpha_{n-1}} \\ &\ge \frac{1}{\alpha_n} P_n h(x) - \frac{\beta_{n-1}}{\alpha_{n-1}}.\end{aligned}$$

Letting $n \to \infty$ through the subsequence $\{n_i\}$ then gives by A4)

$$\eta + h(x) \ge c_w(x) + P_w h(x), \qquad x \in \mathbf{X}.$$

Hence, the limit h satisfies $P_w h = h - \eta + c$, where $c(x) = c_w(x)$. That is, (9) is satisfied.

To see that the optimality equation is satisfied, recall from (11) that

$$c_n(x) + P_n h_{n-1}(x) - c(x, a) \le P_a h_{n-1}(x)$$

for all admissible $(x, a) \in \mathbf{X} \times \mathcal{A}$. It follows that for any a

$$c_n(x) + \frac{1}{\alpha_n} P_n h(x) - c(x, a) \le \frac{1}{\alpha_n} P_a g_{n-1}(x).$$

Letting $n \to \infty$, we see that (10) is satisfied. □

Theorem 5.1 still does not address a central issue: does the solution to (9) give rise to an optimal policy? The example on [2, p. 87] shows that some extra conditions are required, even when the cost is norm-like. We now present a new approach which in many examples gives a direct verification of the technical Assumption 3) in Theorem 2.1 and thereby proves that the policy given through the PIA is indeed optimal.

If the controlled chain $\mathbf{\Phi}^w$ is ψ_w-irreducible and the resulting cost $\eta_w := \pi_w(c_w) = \int c_w(x)\pi_w(dx)$ is finite, let S_w denote any fixed c_w-regular set for which $\pi_w(S_w) > 0$. We then define the function

$$V_w(x) = \mathbf{E}_x\left[\sum_{t=0}^{\tau_w - 1} c_w(\Phi_t^w)\right] \qquad (38)$$

where $\tau_w = \tau_{S_w}$. Since $\pi_w(S_w) > 0$, the function V_w is a.e. $[\pi_w]$ finite-valued [33, Th. 14.2.5]. Note that by [33, Th. 14.2.3], the particular c_w-regular set S_w chosen is not

important. If S_w^1 and S_w^2 give rise to functions V_w^1 and V_w^2 of the form (38), then for some constant $\gamma \geq 1$

$$\gamma^{-1}(V_w^1(x)+1) \leq V_w^2(x) \leq \gamma(V_w^1(x)+1), \qquad x \in \mathbf{X}.$$

Theorem 5.2: Suppose the following.

1) The optimality equations (9) and (10) hold for (w^*, h, η_*), with h bounded from below.
2) For any policy w the average cost $K_w c_w$ is finite and norm-like, and all compact sets are petite for the Markov chain $\mathbf{\Phi}^w$.
3) For any policy w there exists some constant $b = b(w) < \infty$ such that

$$|h(x)| \leq b(1+V_w(x)), \qquad x \in \mathbf{X}. \tag{39}$$

Then w^* is optimal in the sense that for any initial condition $x \in \mathbf{X}$, and any policy w

$$\begin{aligned} J(w^*, x) = \eta_* &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[c_{w^*}(\Phi_t^{w^*})] \\ &\leq \pi_w(c_w) \\ &\leq \liminf_{\beta\uparrow 1} (1-\beta) \sum_{t=0}^{\infty} \beta^t \mathbf{E}_x[c_w(\Phi_t^w)] \\ &\leq J(w, x). \end{aligned} \tag{40}$$

Proof: First note that if $J(w, x) = \infty$ for all x, then there is nothing to prove. If not, then since all compact sets are petite, the Markov chain $\mathbf{\Phi}^w$ is a positive recurrent T-chain, with unique invariant probability π_w, and $\eta_w := \pi_w(c_w)$ is finite [33, Th. 14.0.1].

Under the assumptions of the theorem, we can show inductively that

$$P_w^n V_w = V_w - \sum_{t=0}^{n-1} P_w^t (c_w - s_w)$$

where $s_w \geq 0$ and satisfies $\pi_w(s_w) = \pi_w(c_w)$. This function can be written explicitly as

$$s_w(x) = \int_{S_w} P_w(x, dy) \mathbf{E}_y \left[\sum_{t=0}^{\tau_w - 1} c_w(\Phi_t) \right].$$

It follows from [33, Th. 14.0.1] that $P_w^n V_w(x)/n \to 0$ as $n \to \infty$ for a.e. $[\pi_w]$ $x \in \mathbf{X}$. By (39), we also have $P_w^n h(x)/n \to 0$ for such x.

From the optimality equations (9) and (10), we have

$$P_w^n h(x)/n + \frac{1}{n} \sum_{t=0}^{n-1} \mathbf{E}_x[c_w(\Phi_t^w)] \geq h(x)/n + \eta_*, \qquad n \geq 1, \quad x \in \mathbf{X}.$$

Letting $n \to \infty$ proves the inequality

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[c_w(\Phi_t^w)] \geq \eta_* \quad \text{a.e. } x \in \mathbf{X}[\pi_w].$$

It also follows from [33, Th. 14.0.1] that $\eta_w \geq \eta_*$.

To prove the theorem, we must now consider the limit and limit infimum in (40) for *every* x. That

$$\frac{1}{n} \sum_{t=1}^{n} \mathbf{E}_x[c_{w^*}(\Phi_t^{w^*})] \to \eta_*, \qquad x \in \mathbf{X}$$

follows from Theorem 4.1. The limit infimum is more subtle. From the assumptions of the theorem, the function $K_w c_w$ is *unbounded off petite sets* [33]. It then follows as in the proof of [33, Th. 17.1.7] that, for every initial condition, the law of large numbers holds

$$\lim_{\beta\uparrow 1} (1-\beta) \sum_{t=0}^{\infty} \beta^k K_w c_w(\Phi_t^w) = \eta_w \mathbf{1}_H + \infty \mathbf{1}_{H^c} \qquad \text{a.s. } [\mathbf{P}_x]$$

where H is the event that the w-controlled chain $\mathbf{\Phi}^w$ enters its maximal Harris set. On the event H^c we actually have $K_w c_w(\Phi_t^w) \to \infty$ as $t \to \infty$, since then the process visits each petite set only finitely often [33, Th. 9.0.1]. Taking expectations of both sides of this equation and applying Fatou's lemma then gives

$$\liminf_{\beta\uparrow 1} (1-\beta) \sum_{t=0}^{\infty} \beta^t \mathbf{E}_x[K_w c_w(\Phi_t^w)] \geq \eta_w, \qquad x \in \mathbf{X}.$$

It can also be shown using the resolvent equation [33, p. 291] that for any x which satisfies $K_w c_w(x) < \infty$

$$\lim_{\beta\uparrow 1} (1-\beta) \sum_{t=0}^{\infty} \beta^t (\mathbf{E}_x[K_w c_w(\Phi_t^w)] - \mathbf{E}_x[c_w(\Phi_t^w)]) = 0$$

and this completes the proof of the second inequality in (40). The last inequality follows from [39, Lemma 8.10.6]. □

To see how Theorem 5.2 is applied, consider again the linear model (7) with Gaussian noise satisfying $\Sigma = \mathbf{E}[W_t W_t^T] > 0$. For this example we showed above that a solution (w^*, h, η_*) to the ACOE exists with h a quadratic. Suppose that w is any (measurable) nonlinear feedback control. From the assumption that $\Sigma > 0$, the process $\mathbf{\Phi}^w$ is ψ-irreducible, with $\psi =$ Lebesgue measure. The function V_w given in (38) then satisfies the lower bound

$$V_w(x) \geq \sum_{t=0}^{\infty} 2^{-k} P_w^t c_w(x) \geq V_{1/2}(x) - b_w \tag{41}$$

where $V_{1/2}$ is the value function for the discounted problem and b_w is a finite constant. To see this, let $\tau_w = \tau_{S_w}$, and write

$$\begin{aligned} V_{1/2}(x) &\leq \sum_{t=0}^{\infty} 2^{-t} P_w^t c_w(x) \\ &= \mathbf{E}_x \left[\sum_{t=0}^{\tau_w - 1} 2^{-k} c_w(\Phi_t^w) \right] + \mathbf{E}_x \left[\sum_{t=\tau_w}^{\infty} 2^{-k} c_w(\Phi_t^w) \right] \\ &\leq V_w(x) + \mathbf{E}_x \left[2^{-\tau_w} \sum_{t=0}^{\infty} 2^{-k} c_w(\Phi_{\tau_w + t}^w) \right]. \end{aligned}$$

The lower bound (41) on V_w then bound follows from the strong Markov property and regularity of the set S_w. If $(A, \sqrt{Q})$ is observable so that $V_{1/2}$ dominates a positive definite quadratic, then from Theorems 5.1 and 5.2, we deduce that the PIA does yield an optimal policy over the class of all nonlinear feedback control laws.

VI. Continuous-Time Processes

Because the general theory of Harris processes in continuous time is now rather complete, the previous results carry over to this setting with few changes. The reader is referred to [13], [32], and [34] for details on how to extend the theory of [33] to continuous-time processes. We sketch here what is necessary for the understanding of the PIA in continuous time. The text [39] also describes methods for "lifting" continuous-time optimality criteria from existing discrete-time theory. This approach is based upon uniformization, which requires a countable state-space model and some strong bounds on the infinitesimal generator.

Consider a continuous time MDP with (extended) generator $\mathcal{D}_a$ parameterized by the action $a \in \mathcal{A}(x)$. First, suppose that w^* is a policy, and let h solve Poisson's equation in continuous time

$$\mathcal{D}_{w^*} h = -c_{w^*} + \eta_{w^*} \tag{42}$$

where the definitions of c_w and η_w remain the same. The existence of well-behaved solutions to (42) follows under conditions analogous to the discrete time case [18]. Suppose that for any other policy w

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

where we assume that h is in the domain of the generator $\mathcal{D}_w$—this is the only significant technicality in the continuous-time framework. From Poisson's equation (42), the bound above may be written

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

and then by the definition of the extended generator

$$\eta_{w^*} \le \frac{1}{T} \int_0^T \mathbf{E}_x[c_w(\Phi_t^w)]\, dt + \frac{1}{T} \left(P_w^T h(x) - h(x)\right).$$

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

$$\liminf_{T \to \infty} \frac{1}{T} \int_0^T \mathbf{E}_x[c_w(\Phi_t^w)]\, dt \ge \eta_{w^*}$$

provided that

$$\frac{1}{T} P^T h(x) \to 0 \quad \text{as } T \to \infty.$$

Conditions under which this holds may be formulated as in Theorem 5.2.

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

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

where we again adopt the notation used in the discrete-time development. This equation has a solution provided that the chain $\mathbf{\Phi}^{w_{n-1}}$ is c_{n-1}-regular [18]. A new policy w_n is then found which satisfies, for any other policy w

$$c_n + \mathcal{D}_{w_n} h_{n-1} \le c_w + \mathcal{D}_w h_{n-1}.$$

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

$$\mathcal{D}_{w_n} h_{n-1} \le -c_n + \eta_{n-1}.$$

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 $\mathbf{\Phi}^{w_n}$ is c_n-regular, and we have $\eta_n \le \eta_{n-1}$.

Since the policy w_n is so well behaved, one can assert that Poisson's equation

$$\mathcal{D}_n h_n = -\overline{c}_n$$

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 4.4 can then be formulated using almost identical methodology.

VII. Networks

In this section we apply the general results of Sections IV and V to the scheduling problem for multiclass queueing networks using a countable state-space network model with deterministic routing, as may be used in the design of a semiconductor manufacturing plant. It will be clear that the results obtained apply to many other related models in the operations research area.

Consider a network of the form illustrated in Fig. 2, composed of d single server stations, which we index by $\sigma = 1, \cdots, d$. The network is populated by K classes of customers. Class k customers require service at station $s(k)$. An exogenous stream of customers of class 1 arrive to machine $s(1)$. 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

$$\Phi_{t+1} = \Phi_t + \sum_{k=0}^{K} I_{t+1}(k)[e^{k+1} - e^k] w_t(k) \tag{43}$$

where the state process $\mathbf{\Phi}$ evolves on $\mathbf{X} = \mathbb{Z}_+^K$, and $\Phi_t(k)$ denotes the number of class k customers in the system at time t.

The random variables $\{I_n\colon n \ge 0\}$ are i.i.d. on $\{0, 1\}^{K+1}$, with $\mathbf{P}\{\sum_k I_n(k) = 1\} = 1$, and $\mathbf{E}[I_n(k)] = \mu_k$. For $1 \le k \le K$, μ_k denotes the service rate for class k customers. For $k = 0$, we let $\mu_0 := \lambda$ denote the arrival rate of customers of class 1. For $1 \le k \le K$ we let e^k denote the kth basis vector in $\mathbb{R}^K$, and we set $e^0 = e^{K+1} := 0$.

The sequence $\{w_n\colon n \ge 0\}$ is the control, which takes values in $\{0, 1\}^{K+1}$. We define $w_n(0) \equiv 1$. The set of admissible control actions $\mathcal{A}(x)$ is defined in an obvious manner. For $a \in \mathcal{A}(x)$:

1) for any $1 \le k \le K$, $a_k = 0$ or 1;
2) for any $1 \le k \le K$, $x_k = 0 \Rightarrow a_k = 0$;
3) for any station σ, $0 \le \sum_{k\colon s(k)=\sigma} a_k \le 1$;
4) for any station σ, $\sum_{k\colon s(k)=\sigma} a_k = 1$ whenever $\sum_{k\colon s(k)=\sigma} x_k > 0$.

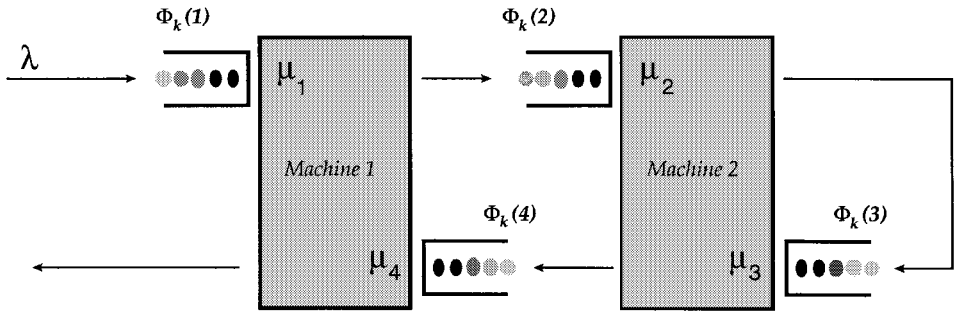

Fig. 2. A multiclass network with $d = 2$ and $K = 4$.

If $a_k = 1$, then buffer k is chosen for service. Condition 2) then imposes the physical constraint that a customer cannot be serviced at a buffer if that buffer is empty. Condition 3) means that only one customer may be served at a given instant at a single machine σ. The nonidling Condition 4) is satisfied by any optimal policy. An inductive proof of this fact may be constructed based upon value iteration [35].

Since the control is bounded, a reasonable cost function is $c(x, a) = c^T x$, where $c \in \mathbb{R}^K$ is a vector with strictly positive entries. For concreteness, we take $c(x, a) = |x| := \sum_k x_k$. Since $\mathcal{A}(x)$ is a finite set for any x, it follows that A2) holds with this cost function.

The transition function has the simple form

$$\mathbf{P}_a(x, x + e^{k+1} - e^k) = \mu_k a_k, \qquad 0 \le k \le K$$
$$\mathbf{P}_a(x, x) = 1 - \sum_0^K \mu_k a_k.$$

The accessibility condition (15) holds where $\boldsymbol{\alpha}$ denotes the empty state $\boldsymbol{\alpha} = (0, \cdots, 0)^T \in \mathbf{X}$, and hence also A3) is satisfied. This follows from the nonidling assumption: if $|x| = m$, and if the total number of buffers in the system is K, then because of the nonidling Assumption 4) the network will be empty at time mK provided that 1) no customers arrive to the network during the time interval $[0, mK]$ and 2) none of the mK services are virtual services. That is, $\sum_{i=1}^K I_i(k+1) w_i^k = 1$, $0 \le k < mK$. The probability of this event is bounded from below by $(1-\lambda)^{mK}\lambda^{mK}$, and hence we have for any nonidling policy $\mathbf{w}$

$$K_w(x, \boldsymbol{\alpha}) \ge (\tfrac{1}{2})^{|x|K+1} P^{|x|K}(x, \boldsymbol{\alpha}) \ge s(x) := (\tfrac{1}{2})^{|x|K+1}(1-\lambda)^{|x|K}\lambda^{|x|K}, \qquad \text{for all } x \in \mathbf{X}.$$

Associated with this network is a *fluid model*. For each initial condition $\Phi(0) = x \ne 0$, we construct a continuous-time process $\phi^x(t)$ as follows. If $|x|t$ is an integer, we set

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

For all other $t \ge 0$, 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. The collection of all "fluid limits" is defined by

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

where the overbar denotes weak closure. The process ϕ evolves on the state space $\mathbb{R}_+^K$ and, for a wide class of scheduling policies, satisfies a differential equation of the form

$$\frac{d}{dt}\phi(t) = \sum_{k=0}^K \mu_k[e^{k+1} - e^k]u_t(k) \tag{44}$$

where the function u_t is analogous to the discrete control and satisfies similar constraints [9].

It is now known that stability of (43) in terms of c-regularity is closely connected with the stability of the fluid model [9], [10], [27], [30]. The fluid model $\mathcal{L}$ is called L_p-stable if

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

It is shown in [27] that L_2-stability of the fluid model is equivalent to a form of c-regularity for the network.

Theorem 7.1—Kumar and Meyn [27]: The following stability criteria are equivalent for the network under any nonidling policy.

i) Drift condition (4) holds for some function V. The function V is equivalent to a quadratic in the sense that, for some $\gamma > 0$

$$1 + \gamma|x|^2 \le V(x) \le 1 + \gamma^{-1}|x|^2, \qquad x \in \mathbf{X}. \tag{45}$$

ii) For some quadratic function V

$$\mathbf{E}_x\left[\sum_{n=0}^{\sigma_\alpha} |\Phi_n|\right] \le V(x), \qquad x \in \mathbf{X}$$

where σ_α is the first entrance time to $\boldsymbol{\alpha} = 0$.

iii) For some quadratic function V and some $c < \infty$

$$\sum_{n=1}^N \mathbf{E}_x[|\Phi_n|] \le V(x) + cN, \qquad \text{for all } x \text{ and } N \ge 1.$$

iv) The fluid model $\mathcal{L}$ is L_2-stable. □

The previous result can be strengthened. If the fluid model is L_2-stable, then in fact $\phi(t) = 0$ for all t sufficiently large [35].

A policy w^* is called *optimal for the fluid model* if for any other policy $\mathbf{w}$ which is L_2-stable

$$\liminf_{T\to\infty} \liminf_{|x|\to\infty} \left(\mathbf{E}_x^w\left[\int_0^T |\phi^x(s)|\, ds\right] - \mathbf{E}_x^{w^*}\left[\int_0^T |\phi^x(s)|\, ds\right]\right) \ge 0.$$

In [35, Th. 3.2(v)] it is shown that L_2-stability is equivalent to uniform boundedness of the total cost. Hence L_2-stability may be assumed without any real loss of generality. If the paths of the fluid limit model are purely deterministic, this form of optimality amounts to minimality of the total cost

$$\int_0^\infty |\phi(\tau)|\, d\tau.$$

Currently, there is much interest in directly addressing methods for the synthesis of optimal policies for fluid network models [7], [16], [38], [48].

Theorem 7.1 is used to establish the following theorem, which shows that policy iteration always converges for this network model if the initial policy is stabilizing. To initiate the algorithm, one can choose one of the known stabilizing policies such as last buffer-first served or first buffer-first served [8], [28], or one may choose a policy which is *optimal* for the fluid model.

Theorem 7.2: If the initial policy w_0 is chosen so that the fluid model is L_2-stable, then

1) the PIA produces a sequence $\{(w_n, h_n, \eta_n)\colon n \geq 0\}$ such that each associated fluid model is L_2-stable. Any policy w^* which is a pointwise accumulation point of $\{w_n\}$ is an optimal average cost policy;
2) for each $n \geq 1$

$$\liminf_{T\to\infty} \liminf_{|x|\to\infty} \left(\mathbf{E}_x^{w_n - 1}\left[\int_0^T |\phi^x(s)|\, ds\right] - \mathbf{E}_x^{w_n}\left[\int_0^T |\phi^x(s)|\, ds\right]\right) \geq 0.$$

Hence if w_0 is optimal for the fluid model, so is w_n for all n;

3) with h equal to the relative value function for any optimal policy w^* generated by the algorithm

$$\limsup_{T\to\infty} \limsup_{|x|\to\infty} \left| \frac{h(x)}{|x|^2} - \mathbf{E}_x^{w^*}\left[\int_0^T |\phi^x(s)|\, ds\right]\right| = 0.$$

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

4) for any policy w whose fluid model is L_2 stable

$$\limsup_{T\to\infty} \limsup_{|x|\to\infty} \left(\frac{h(x)}{|x|^2} - \mathbf{E}_x^{w}\left[\int_0^T |\phi^x(s)|\, ds\right]\right) \leq 0.$$

Hence, the limiting policy w^* is optimal for the fluid model.

Proof: Observe from Theorems 7.1 and A.1 that whenever the fluid model is L_2-stable, the relative value function h is equivalent to a quadratic, in the sense of (45). Moreover, for any policy w, we have the lower bound

$$V_w(x) := \mathbf{E}_x\left[\sum_{t=0}^{\tau_\alpha - 1} c_w(\Phi_t^w)\right] \geq \tfrac{1}{2}|x|^2.$$

This is a consequence of the skip-free property of the network model. From this and Theorems 5.1 and 5.2, we obtain 1).

To see 2), we use the approximation

$$\frac{1}{|x|}\,\mathbf{E}_x^w\left[\sum_{t=0}^{T|x|-1} \frac{|\Phi_t|}{|x|}\right] = \mathbf{E}_x^w\left[\int_0^T |\phi^x(s)|\, ds\right] + o(1) \quad (46)$$

where the term $o(1)$ vanishes as $x \to \infty$. Consider Poisson's equation and the bound (13) together, which when iterated gives for any $T > 0$

$$P_n^{T|x|} h_{n-1}(x) \leq h_{n-1}(x) - \sum_{t=0}^{T|x|-1} \mathbf{E}^{w_n}[|\Phi_t|] + T|x|\eta_{n-1}$$

$$P_n^{T|x|} h_n(x) = h_n(x) - \sum_{t=0}^{T|x|-1} \mathbf{E}^{w_n}[|\Phi_t|] + T|x|\eta_n.$$

We will combine these equations with (46) and take limits, but to do so we must eliminate the term $P_n^{T|x|}h_n(x)/|x|^2$. To show that this converges to zero as $x \to \infty$, $T \to \infty$, apply the upper bound $|h_n(x)| \leq W(x) = b(|x|^2+1)$, where $b < \infty$ is a constant, and use weak convergence to obtain

$$\limsup_{x\to\infty} \frac{P_n^{T|x|}h_n(x)}{|x|^2} \leq \limsup_{x\to\infty} \frac{P_n^{T|x|}W(x)}{|x|^2} \leq b \sup_{\phi\in\mathcal{L}} \mathbf{E}[|\phi(T)|^2].$$

Hence, by L_2-stability of the fluid model, $P_n^{T|x|}\, h_n(x)/|x|^2 \to 0$ as $x \to \infty$, and then $T \to \infty$. Combining the previous equations and using (46) then proves 2).

The proofs of 3) and 4) are similar. Letting h denote the value function for the optimal policy w^*, we have

$$P_{w^*}h(x) = h(x) - |x| + \eta_*$$

and for any policy w

$$P_w h(x) \geq h(x) - |x| + \eta_*$$

where η_* is the optimal steady-state cost. The proof then follows as above by iteration and letting $|x| \to \infty$. □

The exponential assumption here is not crucial. In [10] the case of general distributions is developed, and the analogous regularity results are obtained when a fluid model is L_2-stable.

VIII. Conclusions and Extensions

This paper has introduced several techniques for the analysis of MDP's which suggest further development of MDP's on a general state space.

1) It is likely that duality theory using linear programming formulations may be strengthened using the bounds on solutions to Poisson's equation obtained in this paper (see [20]).
2) Given the uniform lower bounds obtained in Theorem 4.4 and the lower bounds required in the analysis of, for instance [19] and [42], the relationship between discounted and average control problems may be further developed using the techniques presented here.

3) In [3] and [44] conditions are provided which ensure that value iteration generates an optimal policy, and in [4] convergence of this algorithm has been addressed under a global Lyapunov condition of the form (2). Since this Lyapunov condition does not generally hold for simple network examples such as that illustrated in Fig. 2, and the irreducibility condition of [3] is not satisfied in general for network models, it is of interest to see if convergence, or even near monotone convergence can be guaranteed under assumptions similar to Theorem 4.4. Some results in this direction were obtained recently in [6] and [35].
4) Sample-path versions of the PIA such as the actor–critic algorithm have been subject to much analysis in recent years [26], [45]. A generalization of the bounds obtained here to this setting would greatly enhance our understanding of these algorithms.

We are currently investigating the network problem of Section VII to see if some additional insight can be gained in this interesting application.

APPENDIX A
POISSON'S EQUATION

In this section, we collect together some general results on Poisson's equation for uncontrolled chains on a general state space. We assume throughout that the chain is positive Harris with unique invariant probability π. The function $c\colon \mathbf{X} \to [1, \infty)$ in (5) is assumed to satisfy

$$\pi(c) := \int_{\mathbf{X}} c(x)\pi(dx) < \infty.$$

Define the function $s\colon \mathbf{X} \to [0, 1]$ as $s = \delta \mathbb{1}_S$, where $\delta > 0$, and S denotes the sublevel set

$$S = \{x\colon Kc(x) \le \pi(c)\}. \tag{47}$$

We always have that $\pi(S) > 0$, and hence $\pi(s) > 0$. Under the assumption that S is petite, for $\delta > 0$ sufficiently small the resolvent satisfies the minorization condition $K \ge s \otimes \nu$ as in (32). In this case we let G denote the kernel

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

If the sum $Gc(x) = \sum (K - s \otimes \nu)^t c(x)$ is convergent for each x, which amounts to c-regularity of the process, then a solution to Poisson's equation may be explicitly written as

$$h(x) = GK\overline{c}(x) = \sum_{i=0}^{\infty} (K - s \otimes \nu)^i K\overline{c}(x). \tag{48}$$

The following theorem establishes the existence of suitably bounded solutions to Poisson's equation, given that the process is c-regular.

Theorem A.1: Suppose that the Markov chain $\mathbf{\Phi}$ is positive Harris. Assume further that $\pi(c) = \int c(x)\pi(dx) < \infty$ and that the set S defined in (47) is petite. Then there exists a solution h to Poisson's equation (5) which is finite a.e.$[\pi]$ and is bounded from below everywhere

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

If $\mathbf{\Phi}$ is also c-regular, then h can be chosen so that

$$h(x) \le dV_S(x) := d\mathbf{E}_x\left[\sum_{t=0}^{\tau_S - 1} c(\Phi_t)\right], \qquad x \in \mathbf{X}$$

where d is a finite constant.

Proof: From [33] and [36] we know that the invariant probability for $\mathbf{\Phi}$ may be written

$$\pi(A) = \frac{\int G(x, A)\nu(dx)}{\int G(x, \mathbf{X})\nu(dx)}, \qquad A \in \mathcal{B}(\mathbf{X}).$$

Since $\pi(c) < \infty$ it follows that $G(x, c) < \infty$ for almost every $x \in \mathbf{X}$ $[\nu]$, and as in the proof of [33, Proposition 14.1.2] we know that $G(x, c) < \infty$ for almost every $x \in \mathbf{X}$ $[\pi]$.

To obtain a lower bound, observe that

$$K\overline{c}(x) \ge \begin{cases} 0, & \text{on } S^c \\ -\pi(c), & \text{on } S. \end{cases}$$

It follows that $K\overline{c}(x) \ge -Ns$, where $N = \pi(c)/\delta$. Hence, for any x

$$h(x) := GK\overline{c}(x) \ge -NGs(x) \ge -N.$$

The function h solves Poisson's equation, and the lower bound is thereby established.

To obtain the upper bound when the chain is c-regular, [33, Th. 14.2.3] may be used to show that $GKc(x) \le dV_S(x)$ for a constant d. It follows that the function h given in (48) satisfies the desired upper bound. □

Given the lower bound on h, it is easy to establish uniqueness. Related results are given in [18], [20], [33], and [37]. First we give the following technical result.

Lemma A.2: Suppose that $\mathbf{\Phi}$ is positive Harris with invariant probability π, and suppose that $z\colon \mathbf{X} \to \mathbb{R}$ is bounded from below and is superharmonic: $Pz \le z$. Then

1) $z(x) = \pi(z)$ for almost every $x \in \mathbf{X}$ $[\pi]$;
2) $z(x) \ge \pi(z)$ for every $x \in \mathbf{X}$.

Proof: The superharmonic property implies that $(z(\Phi_t), \mathcal{F}_t)$ is a supermartingale. Since it is bounded from below, it must be convergent. This together with Harris recurrence implies that z is constant a.e., exactly as in the proof of [33, Th. 17.1.5]. We then have by Fatou's lemma, for every x

$$z(x) \ge \liminf_{t \to \infty} \mathbf{E}_x[z(\Phi_t)] \ge \mathbf{E}_x[\liminf_{t \to \infty} z(\Phi_t)] = \pi(z).$$

□

Theorem A.3: Suppose that the Markov chain $\mathbf{\Phi}$ is positive Harris, that $\eta = \pi(c) < \infty$, and assume that S defined in (47) is petite. Let g be finite-valued, be bounded from below, and satisfy

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

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

i) $g(x) = GK\overline{c}(x) + b$ for almost every $x \in \mathbf{X}$ $[\pi]$;
ii) $g(x) \ge GK\overline{c}(x) + b$ for every $x \in \mathbf{X}$.

Proof: Since the set S is petite, we may again choose $\delta > 0$ so small in the definition of $s = \delta \mathbb{1}_S$ so that $K \ge s \otimes \nu$. We also assume without loss of generality that $\nu(g) = 0$. From the inequality which g is assumed to satisfy, we have

$$[I - (K - s \otimes \nu)]g \ge K\overline{c}.$$

Iteration then gives

$$\limsup_{N\to\infty} [g(x) - (K - s \otimes \nu)^N g(x)] \ge GK\overline{c}(x).$$

Since we always have $GK\overline{c}(x) \ge -\pi(c)G(x, S) > -\infty$, the inequality above establishes finiteness of $GK\overline{c}(x)$ for all x. We must also have $G\mathbb{1}(x) < \infty$, and since the function g is bounded from below, it follows that $\liminf_{N\to\infty} (K - s \otimes \nu)^N g(x) \ge 0$. The bound on $GK\overline{c}$ then becomes

$$h(x) := GK\overline{c}(x) \le g(x).$$

Thus, the positive function $z(x) = g(x) - h(x)$ is superharmonic, and the result follows from Lemma A.2. □

We saw in Theorem A.1 that solutions to Poisson's equation may be taken to be positive. Here we show that solutions are continuous under extra conditions. While this result is not used explicitly, it is clear that continuity can greatly simplify the verification of some of the conditions imposed in the paper.

The transition function P is called *Feller* if Pg is a continuous function on $\mathbf{X}$ whenever g is bounded and continuous. It is easy to show that for a Feller chain, if compact sets are c-regular, then the solution h to Poisson's equation may be taken to be lower semicontinuous. This follows from the representation of h through (48), where s is a continuous function with compact support. To obtain continuity requires some stronger assumptions. We call a set S *c-regular of degree two* if for any $A \in \mathcal{B}^+(\mathbf{X})$

$$\sup_{x\in S} \mathbf{E}_x \left[\sum_{t=0}^{\tau_A - 1} (k+1)c(\Phi_t)\right] < \infty.$$

From the relation

$$\mathbf{E}_x \left[\sum_{t=0}^{\tau_A - 1} (k+1)c(\Phi_t)\right] = \mathbf{E}_x \left[\sum_{t=0}^{\tau_A - 1} \mathbf{E}_{\Phi_t} \left[\sum_{i=0}^{\tau_A - 1} c(\Phi_i)\right]\right] \quad (49)$$

it follows that the set S is c-regular of degree two if and only if it is simultaneously c-regular, and V-regular, where V is the function (26) given in Theorem 4.2. That is, the pair of equations are satisfied

$$\begin{aligned} PV_1 &\le V_1 - c + b_1 \mathbb{1}_{S_1} \\ PV_2 &\le V_2 - V_1 + b_2 \mathbb{1}_{S_2} \end{aligned} \quad (50)$$

where $\{b_i\}$ are finite constants, $\{S_i\}$ are petite, and the functions $\{V_i\}$ are finite-valued and positive (see [33, Th. 14.2.3]). When $c \equiv 1$, this condition is equivalent to requiring that $\mathbf{E}_x[\tau_A^2]$ be bounded on S, for $A \in \mathcal{B}^+(\mathbf{X})$. See [46] for further discussion.

We now establish continuity. A similar result may be found in [24] for V-uniformly ergodic chains.

Theorem A.4: Assume that c: $\mathbf{X} \to \mathbb{R}_+$ is norm-like and continuous and that all compact subsets of $\mathbf{X}$ are c-regular of degree two. Assume moreover that

1) the chain has the Feller property;
2) for some compact set $C \in \mathcal{B}^+(\mathbf{X})$, and some continuous function V: $\mathbf{X} \to \mathbb{R}_+$

$$\mathbf{E}_x \left[\sum_{t=0}^{\tau_C} c(\Phi_t)\right] \le V(x), \qquad x \in \mathbf{X};$$

3) PV is a continuous function.

Then there exists a continuous solution h to Poisson's equation (5).

Proof: From regularity of degree two we can obtain solutions to the pair of inequalities analogous to (50):

$$\begin{aligned} KV_1 &\le V_1 - Kc + d_1 s \\ KV_2 &\le V_2 - V_1 + d_2 s \end{aligned}$$

where $V_1 \le d(V + 1)$ for a constant d, and the function V_2 is finite everywhere and uniformly bounded on compact subsets of $\mathbf{X}$. The function s is continuous with compact support, and K satisfies $K \ge s \otimes \nu$.

We can show by induction that the function h_n defined by

$$h_n = \sum_{t=0}^{n-1} (K - s \otimes \nu)^t K\overline{c}, \qquad n \ge 1$$

is a continuous function on $\mathbf{X}$. The proof will be complete when we show that $h_n \to h$ uniformly on compact subsets of x.

From the bounds on $\{V_i\}$ we can show by inversion as above that

$$\begin{aligned} GKc &\le V_1 + d_1 \\ GV_1 &\le V_2 + d_2 \end{aligned}$$

which shows that $GGKc \le V_2 + d_1 + d_2$. The left-hand side can be computed as follows:

$$\begin{aligned} GGKc(x) &:= \left(\sum_{i=0}^{\infty} (K - s \otimes \nu)^i\right) \cdot \left(\sum_{j=0}^{\infty} (K - s \otimes \nu)^j\right) Kc(x) \\ &= \sum_{i,\, j=0}^{\infty} (K - s \otimes \nu)^{i+j} Kc(x) \\ &= \sum_{m=0}^{\infty} (m+1)(K - s \otimes \nu)^m Kc(x). \end{aligned}$$

Using the bound $|\overline{c}| \le (1+\eta)c$, this gives

$$\begin{aligned}|h(x)-h_n(x)| &= \left|\sum_{t=n}^{\infty}(K-s\otimes\nu)^t K\overline{c}(x)\right| \\ &\le (1+\eta)\,\frac{1}{n}\sum_{t=n}^{\infty} t(K-s\otimes\nu)^t Kc(x) \\ &\le (1+\eta)\,\frac{1}{n}\,(V_2(x)+d_1+d_2).\end{aligned}$$

The right-hand side evidently converges to zero uniformly on compact subsets of $\mathbf{X}$, and this proves the theorem. □

Appendix B
Some Technical Proofs

Proof of Theorem 4.3: The main idea is to apply (29), which is a version of (4) if $h_{n-1} \ge 0$. We first prove 1) and 2). Result 3) is then a consequence of 2), (29), and the Comparison Theorem [33].

For any $i \le n$, assume without loss of generality that $h_{i-1} \ge 0$, and let w_i be the policy which attains the minimum

$$w_i(x) = \arg\min_{a\in\mathcal{A}(x)} [c(x,a) + P_a h_{i-1}(x)].$$

Then by (29) the function $V_i = h_{i-1}$ satisfies

$$K_i V_i \le V_i - K_i c_i + \eta_{i-1}$$

a version of (4). Applying Theorem 4.2 we see that 2) holds, and applying Theorem A.1 we obtain 1). □

Proof of Theorem 4.4: To begin, we prove by induction that the policies $\{w_n\}$ are regular using the following induction hypothesis.

> For any $n \ge 0$, the PIA generates policies $\{w_0, \cdots, w_n\}$ such that the policy w_n is regular and a finite-valued solution h_n to Poisson's equation exists which satisfies the lower bound
>
> $$h_n(x) \ge -\frac{\eta_0}{\delta}, \qquad x \in \mathbf{X}.$$

Assume that the induction hypothesis is true for $n-1$. We have from (29) the inequality

$$K_n h_{n-1} \le h_{n-1} - \tfrac{1}{2} K_n c_n + \eta_0 \mathbb{1}_S \tag{51}$$

where S is defined in (31). Since the set S is assumed to be petite in A3), and since h_{n-1} is bounded from below, Theorem 4.2 shows that the Markov chain with transition probability K_n is $K_n c_n$-regular, and it then follows that the policy w_n is regular.

We now obtain the lower bound on h_n. From (34) and the definition of S we have

$$h_n = G_n K_n \overline{c}_n \ge -\eta_n G_n \mathbb{1}_S \ge -\frac{\eta_0}{\delta}\, G_n s.$$

Since $G_n s(x) \le 1$ for any x, this establishes the lower bound and completes the proof of the induction hypothesis.

To obtain an upper bound on h_n, first note from (51) and the assumption that $\int h_{n-1}\, d\nu = 0$ that

$$[I - (K_n - s\otimes\nu)]h_{n-1} \ge \tfrac{1}{2} K_n c_n - \frac{\eta_0}{\delta}\, s.$$

Applying $\sum_{t=0}^{N-1}(K_n - s\otimes\nu)^t$ to both sides of this inequality allows us to invert the kernel $[I-(K_n - s\otimes\nu)]$

$$\limsup_{N\to\infty}(h_{n-1} - (K_n - s\otimes\nu)^N h_{n-1}) \ge \tfrac{1}{2} G_n K_n c_n - \frac{\eta_0}{\delta}\, G_n s.$$

Using the crude bound $(K_n - s\otimes\nu)^N h_{n-1} \ge -\eta_0/\delta$ together with the bound $G_n s \le 1$ shows that $G_n K_n c_n(x)$ is everywhere finite. Since $K_n c_n$ is norm-like and h_{n-1} is bounded from below, it then follows that $\liminf_{N\to\infty}(K_n - s\otimes\nu)^N h_{n-1} \ge 0$ so that the previous bound gives

$$0 \le G_n K_n c_n \le 2h_{n-1} + 2\,\frac{\eta_0}{\delta}. \tag{52}$$

Using (52) we may now obtain an upper bound on $h_n = G_n K_n \overline{c}_n$. Inequality (29) can be written

$$[I - (K_n - s\otimes\nu)]h_{n-1} \ge K_n \overline{c}_n - (\eta_{n-1} - \eta_n).$$

By c_n-regularity of $\mathbf{\Phi}^{w_n}$, we can apply G_n to both sides of this identity to give

$$\begin{aligned}h_{n-1} &\ge G_n K_n \overline{c}_n - (\eta_{n-1}-\eta_n) G_n \mathbb{1} \\ &= h_n - (\eta_{n-1} - \eta_n) G_n \mathbb{1}\end{aligned} \tag{53}$$

where $\mathbb{1}$ is the function on $\mathbf{X}$ which is identically one. The justification for this inversion follows as in the derivation of (52). Since $G_n \mathbb{1} \le G_n K_n c_n \le 2(h_{n-1} + (\eta_0/\delta))$ by (52), this gives

$$h_n(x) \le [1 + 2(\eta_{n-1} - \eta_n)]h_{n-1}(x) + 2(\eta_0/\delta)(\eta_{n-1} - \eta_n).$$

Thus, we recover the bound (21) obtained in the special case where an accessible state exists. To prove P2), we again define $\{g_n\}$ through (22). It is clear that $\{g_n\}$ and $\{h_n\}$ are related by additive and multiplicative constants as required in P2). To conclude, we now establish that $\{g_n\}$ is monotone decreasing and bounded from below.

From the bound (21)

$$\begin{aligned}g_n(x) &\le \left(\prod_{t=n+1}^{\infty}(1+\varepsilon_t)\right)\Bigg((1+\varepsilon_n)h_{n-1}(x) + (\eta_0/\delta)\varepsilon_n \\ &\qquad\qquad + (\eta_0/\delta)\sum_{t=n+1}^{\infty}\varepsilon_t\Bigg) \\ &\le \left(\prod_{t=n+1}^{\infty}(1+\varepsilon_t)\right)(1+\varepsilon_n) \\ &\quad\cdot\left(h_{n-1}(x) + (\eta_0/\delta)\sum_{t=n}^{\infty}\varepsilon_t\right) \\ &= g_{n-1}(x).\end{aligned}$$

We also have from the lower bound on h_n given in (21)

$$g_n \ge -\frac{\eta_0}{\delta}\prod_{t=0}^{\infty}(1+\varepsilon_t) \ge -\frac{\eta_0}{\delta}\exp(2(\eta_0 - \eta)), \qquad n \in \mathbb{Z}_+.$$

Hence, $g_n(x) \downarrow g(x) > -\infty$ as $n \to \infty$ for each x, and this and the form of g_n proves P2). □

ACKNOWLEDGMENT

The research for this paper was begun while the author was visiting O. Herńandez-Lerma at the Centro de Investigacion del IPN, Mexico City, and the author wishes to thank him for sharing his unpublished work. He is also grateful for his hospitality and explanations of the state of the art in Markov decision theory.

The author also wishes to thank L. Sennott of Illinois State University, E. Fernández-Gaucherand of the University of Arizona, R.-R. Chen of the University of Illinois, and the anonymous reviewers who provided many useful suggestions and references during the revision of this manuscript.

REFERENCES

[1] 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. Contr. Optim.*, vol. 31, pp. 282–344, 1993.

[2] V. S. Borkar, "Topics in controlled Markov chains," in *Pitman Res. Notes in Math. Series # 240.* UK: Longman Scientific & Technical, 1991.

[3] R. Cavazos-Cadena, "Value iteration in a class of communicating Markov decision chains with the average cost criterion," Univ. Autónoma Agraria Anonio Narro, Tech. Rep., 1996.

[4] 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," *J. Appl. Probability*, vol. 33, pp. 986–1002, 1996.

[5] F. Charlot and A. Nafidi, "Irréducibilité, petits ensembles, et stabilité des réseaux de Jackson généralisés," Univ. de Rouen UFR des Sciences, Tech. Rep., 1996.

[6] R.-R. Chen and S. P. Meyn, "Value iteration and optimization of multiclass queueing networks," to be published.

[7] J. Humphrey, D. Eng, and S. P. Meyn, "Fluid network models: Linear programs for control and performance bounds," in *Proc. 13th IFAC World Congr.*, J. Cruz, J. Gertler, and M. Peshkin, Eds., San Francisco, CA, 1996, vol. B, pp. 19–24.

[8] J. Dai and G. Weiss, "Stability and instability of fluid models for certain re-entrant lines," *Math. Ops. Res.*, vol. 21, no. 1, pp. 115–134, Feb. 1996.

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

[10] J. G. Dai and S. P. Meyn, "Stability and convergence of moments for multiclass queueing networks via fluid limit models," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 1889–1904, Nov. 1995.

[11] R. Dekker, "Counterexamples for compact action Markov decision chains with average reward criteria," *Comm. Statist.-Stoch. Models*, vol. 3, pp. 357–368, 1987.

[12] C. Derman, "Dunumerable state MDP's," *Ann. Amth. Statist.*, vol. 37, pp. 1545–1554, 1966.

[13] D. Down, S. P. Meyn, and R. L. Tweedie, "Geometric and uniform ergodicity of Markov processes," *Ann. Probab.*, vol. 23, no. 4, pp. 1671–1691, 1996.

[14] M. Duflo, *Méthodes Récursives Aléatoires.* Masson, 1990.

[15] E. B. Dynkin and A. A. Yushkevich, "Controlled Markov processes," in volume Grundlehren der mathematischen Wissenschaften 235 of *A Series of Comprehensive Studies in Mathematics.* New York: Springer-Verlag, 1979.

[16] D. Bertsimas, F. Avram, and M. Ricard, "Fluid models of sequencing problems in open queueing networks: An optimal control approach," Massachusetts Inst. Technol., Tech. Rep., 1995.

[17] S. Foss, "Ergodicity of queueing networks," *Siberian Math. J.*, vol. 32, pp. 183–202, 1991.

[18] P. W. Glynn and S. P. Meyn, "A Lyapunov bound for solutions of Poisson's equation," *Ann. Probab.*, vol. 24, Apr. 1996.

[19] O. Hernández-Lerma and J. B. Lasserre, "Policy iteration for average cost Markov control processes on Borel spaces," IPN, Departamento de Matematicas, Mexico, and LAAS-CNRS, France, Tech. Rep., 1995; *Acta Applicandae Mathematicae*, to be published.

[20] ———, *Discrete Time Markov Control Processes I.* New York: Springer-Verlag, 1996.

[21] O. Hernández-Lerma, R. Montes-de-Oca, and R. Cavazos-Cadena, "Recurrence conditions for Markov decision processes with Borel state space: A survey," *Ann. Operations Res.*, vol. 28, pp. 29–46, 1991.

[22] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, 1977.

[23] A. Hordijk and M. L. Puterman, "On the convergence of policy iteration," *Math. Ops. Res.*, vol. 12, pp. 163–176, 1987.

[24] A. Hordijk, F. M. Spieksma, and R. L. Tweedie, "Uniform stability conditions for general space Markov decision processes," Leiden Univ. and Colorado State Univ., Tech. Rep., 1995.

[25] R. A. Howard, *Dynamic Programming and Markov Processes.* New York: Wiley, 1960.

[26] V. R. Konda and V. S. Borkar, "Learning algorithms for Markov decision processes," Indian Inst. Sci., Bangalore, Tech. Rep., 1996.

[27] P. R. Kumar and S. P. Meyn, "Duality and linear programs for stability and performance analysis queueing networks and scheduling policies," *IEEE Trans. Automat. Contr.*, vol. 41, pp. 4–17, Jan. 1996.

[28] S. Kumar and P. R. Kumar, "Fluctuation smoothing policies are stable for stochastic re-entrant lines," in *Proc. 33rd IEEE Conf. Decision Contr.*, Dec. 1994.

[29] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems.* New York: Wiley-Intersci., 1972.

[30] S. P. Meyn, "Transience of multiclass queueing networks via fluid limit models," *Ann. Appl. Probab.*, vol. 5, pp. 946–957, 1995.

[31] S. P. Meyn and D. Down, "Stability of generalized Jackson networks," *Ann. Appl. Probab.*, vol. 4, pp. 24–148, 1994.

[32] S. P. Meyn and R. L. Tweedie, "Generalized resolvents and Harris recurrence of Markov processes," *Contemporary Math.*, vol. 149, pp. 227–250, 1993.

[33] ———, *Markov Chains and Stochastic Stability.* London: Springer-Verlag, 1993.

[34] ———, "Stability of Markovian processes III: Foster–Lyapunov criteria for continuous time processes," *Adv. Appl. Probab.*, vol. 25, pp. 518–548, 1993.

[35] S. P. Meyn, "Stability and optimization of multiclass queueing networks and their fluid models," in *Proc. Summer Seminar on "The Math. Stochastic Manufacturing Syst.,"* Amer. Math. Soc., 1997.

[36] E. Nummelin, *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge: Cambridge Univ. Press, 1984.

[37] ———, "On the Poisson equation in the potential theory of a single kernel," *Math. Scand.*, vol. 68, pp. 59–82, 1991.

[38] J. Perkins, "Control of push and pull manufacturing systems," Ph.D. thesis, Univ. Illinois, Urbana, IL, Sept. 1993; Tech. Rep. UILU-ENG-93-2237 (DC-155).

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

[40] R. K. Ritt and L. I. Sennott, "Optimal stationary policies in general state space Markov decision chains with finite action set," *Math. Ops. Res.*, vol. 17, no. 4, pp. 901–909, Nov. 1993.

[41] S. M. Ross, "Applied probability models with optimization applications," in *Dover Books on Advanced Mathematics*, 1992; republication of the work first published by Holden-Day, 1970.

[42] L. I. Sennott, "A new condition for the existence of optimal stationary policies in average cost Markov decision processes," *Ops. Res. Lett.*, vol. 5, pp. 17–23, 1986.

[43] ———, "Average cost optimal stationary policies in infinite state Markov decision processes with unbounded cost," *Ops. Res.*, vol. 37, pp. 626–633, 1989.

[44] ———, "The convergence of value iteration in average cost Markov decision chains," *Ops. Res. Lett.*, vol. 19, pp. 11–16, 1996.

[45] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," Massachusetts Inst. Technol., Cambridge, MA, Tech. Rep. LIDS-P-2322, Mar. 1996: also *IEEE Trans. Automat. Contr.*.

[46] P. Tuominen and R. L. Tweedie, "Subgeometric rates of convergence of f-ergodic Markov chains," *Adv. Appl. Probab.*, vol. 26, pp. 775–798, 1994.

[47] R. Weber and S. Stidham, "Optimal control of service rates in networks of queues," *Adv. Appl. Probab.*, vol. 19, pp. 202–218, 1987.

[48] G. Weiss, "On the optimal draining of re-entrant fluid lines," Georgia Inst. Technol. Technion, Tech. Rep., 1994.

[49] P. Whittle, *Risk-Sensitive Optimal Control.* Chichester, NY: Wiley, 1990.

Sean P. Meyn (S'85–M'87–SM'95) received the B.A. degree in mathematics Summa Cum Laude from the University of California, Los Angeles, in 1982 and the Ph.D. degree in electrical engineering from McGill University in 1987.

He completed a two-year Postdoctoral Fellowship at the Australian National University in Canberra and is now an Associate Professor in the Department of Electrical and Computer Engineering and a Research Associate Professor in the Coordinated Science Laboratory at the University of Illinois, Urbana. He is coauthor (with R. L. Tweedie) of the monograph *Markov Chains and Stochastic Stability* (London: Springer-Verlag, 1993).

Dr. Meyn has served on the editorial boards of several journals in the systems and control and applied probability areas. Also, he was a University of Illinois Vice Chancellor's Teaching Scholar in 1994 and received jointly with Tweedie the 1994 ORSA/TIMS Best Publication in Applied Probability Award. For the 1997 academic year, he is a Visiting Professor at the Indian Institute of Science, Bangalore, under a Fulbright Research Award.